Modélisation pangénomique du déséquilibre de liaison à l'aide de réseaux bayésiens hiérarchiques latents et applications

Par

Raphaël Mourad

Une thèse déposée à l'Ecole polytechnique de l'Université de Nantes
ED 503-137
Spécialité : Informatique
Discipline : Bioinformatique et génétique statistique
pour obtenir le titre de

Docteur en Sciences

1 octobre 2011

Jury :

Rapporteurs :	Louis Wehenkel	Professeur	Université de Liège
	Céline Rouveirol	Professeur	Université Paris-Nord
Directeur de thèse :	Philippe Leray	Professeur	Polytech'Nantes
Co-encadrants :	Christine Sinoquet	Maître de Conférence	Université de Nantes
	Jean-Jacques Schott	Directeur de recherche	Université de Nantes
Président :	Florence d'Alché-Buc	Professeur	Université d'Evry
Examinateurs :	David Balding	Professeur	University College London
Invité :	Christian Dina	Ingénieur	Université de Nantes

Editée en LaTeX 2_ε.

A ma famille et à mes amis.

Résumé

Les récentes technologies génomiques à haut-débit ont ouvert la voie aux études d'association visant la caractérisation systématique à l'échelle du génome des facteurs génétiques impliqués dans l'apparition des maladies génétiques complexes, telles que l'asthme et le diabète. Dans ces études, le déséquilibre de liaison (linkage disequilibrium, LD) reflète l'existence de dépendances complexes au sein des données génétiques et joue un rôle central, puisqu'il permet une localisation précise des facteurs génétiques. Néanmoins, la haute complexité du LD, ainsi que la dimension élevée des données génétiques, constituent autant de difficultés à prendre en compte. Les travaux de recherche réalisés au cours de cette thèse se sont placés dans cette perspective.

La contribution des travaux de recherche présentés est double, puisqu'elle est à la fois théorique et appliquée. Sur le plan théorique, nous avons proposé une nouvelle approche de modélisation du LD. Elle est basée sur le développement d'un modèle issu du domaine de l'intelligence artificielle et de l'apprentissage automatique, la forêt de modèles hiérarchiques à classes latentes (FMHCL). Les nouveautés les plus significatives introduites sont la possibilité de prendre en compte la nature floue du LD et de hiérarchiser les différents degrés de LD. Un nouvel algorithme d'apprentissage supportant le passage à l'échelle, nommé CFHLC, a été développé et décliné en deux versions : la première nécessitant le découpage du génome en fenêtres contiguës pour résoudre le problème de passage à l'échelle, et la seconde (CFHLC+), plus récente et évoluée, résolvant le problème au moyen d'une fenêtre glissante sur le chromosome. A l'aide d'un jeu de données réelles, la comparaison de la méthode CFHLC avec des méthodes concurrentes a montré qu'elle offre une modélisation plus fine du LD. En outre, l'apprentissage sur des données présentant des patrons de LD variés a démontré la capacité de la FMHCL a reproduire fidèlement la structure du LD. Enfin, l'analyse empirique de la complexité de l'apprentissage a montré la linéarité en temps lorsque le nombre de variables à traiter augmente.

Sur le plan appliqué, nous avons exploré deux pistes de recherche : la recherche de causalités et la visualisation synthétique et intuitive du LD. D'une part, une étude systématique de la capacité des FMHCL à la recherche de causalités est illustrée dans le contexte de la génétique d'association. Ce travail a établi les bases du développement de nouvelles méthodes de recherche dédiées à la découverte de facteurs génétiques causaux pour les études d'association à l'échelle du génome. D'autre part, une méthode a été développée pour la visualisation synthétique et intuitive du LD

adaptée aux trois principales situations que peut rencontrer le généticien : la visualisation du LD de courte distance, de longue distance et dans un contexte pangénomique. Cette nouvelle méthode apporte des atouts majeurs qui sont les suivants : (i) le LD par paires (deux variables) et le LD multilocus (deux variables ou plus) sont simultanément visualisés, (ii) le LD de courte distance et le LD de longue distance sont facilement distingués, et (iii) l'information est synthétisée de manière hiérarchique.

Abstract

Recent high-throughput genomic technologies opened the way for association studies aiming at the genome-wide characterization of genetic factors involved in complex genetic diseases, such as asthma and diabetes. In these studies, linkage disequilibrium (LD) reflects the existence of complex dependences in genetic data and plays a central role, since it ensures a precise localization of genetic factors. Nevertheless, the high complexity of LD, as well as the large dimension of genetic data, represents strong difficulties to consider. Research works of this PhD were carried out in this context.

The contribution of research works presented here is twofold, since it is both theoretical and applied. On the theoretical side, we proposed a new approach of LD modeling. It is based on the development of a model coming from artificial intelligence and machine learning, the forest of hierarchical latent class models (FHLCM). The most significant contributions introduced are the ability of taking into account the fuzzy nature of LD and organizing into a hierarchy the multiple LD degrees. A novel scalable learning algorithm, named CFHLC, was developed in two versions : the first requires to split genome into contiguous windows to resolve the scalability issue, and the second (CFHLC+), more recent and advanced, implements a sliding window on chromosome. Using a real dataset, the comparison of the CFHLC method with others revealed that the former offers a more accurate modeling of LD. Besides, learning on data showing varying LD patterns showed the ability of FHLCM to faithfully reproduce the LD structure. Finally, the empirical analysis of learning complexity showed linearity in time when the number of variables to process increases.

On the applied side, we explored two research avenues : causal discovery and global and intuitive visualization of LD. On the one hand, a systematic study of the ability of FHLCM for causal discovery is illustrated in the context of genetic association. This work established the basis of the development of novel methods for causal genetic factor identification in genome-wide association studies. On the other hand, a method was developed for the global and intuitive visualization of LD into three main contexts that geneticist can meet : visualization of short-range, long-range and genome-wide LD. This new method brings several assets as follows : (i) both pairwise LD (two variables) and multilocus LD (more than two variables) are simultaneously displayed, (ii) short-range and long-range LD are easily distinguished, and (iii) information is summarized in a hierarchical manner.

Remerciements

Tout d'abord je remercie Christine Sinoquet et Philippe Leray pour m'avoir guidé lors de ces trois années de doctorat. Après une année plus ou moins perdue à rechercher vainement une thèse (ne possédant pas de Master recherche, ni de Master pro d'ailleurs), je suis tout fraîchement arrivé en première année dans l'équipe. A cette époque, je ne connaissais pratiquement rien à l'informatique, à part quelques bribes de programmation C, tout au plus. Et pourtant, mes encadrants ont su m'aider et croire en moi, et je leur en suis très reconnaissant. Il faut dire que la volonté et l'expertise de Christine, combinées aux conseils et au recul de Philippe ne pouvaient que me tirer vers le haut. Je remercie également mon encadrant côté "bio", Jean-Jacques Schott, et Christian Dina qui ont pu assurer la pertinence de nos développements bioinformatiques grâce à leurs nombreux conseils en génétique. Tout ceci s'est déroulé dans la bonne humeur, ce qui m'a permis de passer des années inoubliables... Je n'ai aucun regret, sauf celui de ne pouvoir continuer une année de thèse en plus :), bien que je sois en même temps très impatient de découvrir l'après-thèse.

Je tiens à remercier les membres du jury, Céline Rouveirol et Louis Wehenkel, pour m'avoir fait l'honneur de leur présence et m'avoir permis grâce à leurs remarques et leurs conseils de perfectionner ce mémoire de thèse. Je remercie aussi Florence d'Alché-Buc pour m'avoir fait l'honneur de présider mon jury de thèse, ainsi que David Balding pour son rôle d'examinateur.

Je suis très reconnaissant envers toute l'équipe COD (Connaissance et Décision), en particulier Pascale, qui s'est toujours montrée très à l'écoute envers les doctorants, et Sylvie pour sa grande aide et sa bonne humeur. J'en profite aussi pour remercier mon laboratoire, le LINA, et le département informatique de l'Ecole Polytechnique de l'Université de Nantes.

Je remercie les organisateurs du projet de Bioinformatique Ligérienne (BIL) sans lequel je n'aurais pu réaliser cette thèse, car la recherche ça coûte cher :(.

Je remercie particulièrement ma famille, en particulier, ma mère, mon père, mon frère, ma marraine et mon parrain, ainsi que tous mes amis qui sont très nombreux et qui se reconnaîtront facilement. Que dire pour finir sinon que mes collègues, stagiaires, thésards et autre post-docs vont beaucoup me manquer!! Même si j'en suis moins sûr concernant la mafia roumaine, Claudia, Teddy et Vlad (dont je ne citerais pas les noms de famille par peur de représailles). Je vais regretter aussi Zohra et

ses pâtisseries tunisiennes, Thomas, le porte-parole de VM matériaux, ainsi que les petits jeunes, Zineddine, Yasin et Nicolas, qui ne savent pas encore très bien ce qui les attend...

Liste des publications au 26.09.2011

Revues internationales

- Raphaël Mourad, Christine Sinoquet et Philippe Leray. Probabilistic graphical models for genetic association studies. *Briefings in bioinformatics*, 2011, à paraître.
- Raphaël Mourad, Christine Sinoquet et Philippe Leray. A hierarchical Bayesian network approach for linkage disequilibrium modeling and data-dimensionality reduction prior to genome-wide association studies. *BMC Bioinformatics*, 12 :16, 2011.

Conférences internationales

- Raphaël Mourad, Christine Sinoquet et Philippe Leray. Learning hierarchical Bayesian networks for genome-wide association studies. *19th International Conference on Computational Statistics (COMPSTAT)*, pages 549–556, 2010.

Conférences nationales

- Vittorio Perduca, Raphaël Mourad, Christine Sinoquet and Gregory Nuel. Waffect : a method to simulate case-control samples in genome-wide association studies. *Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM)*, 2011.
- Vittorio Perduca, Raphaël Mourad, Christine Sinoquet and Gregory Nuel. Waffect : a method to simulate case-control samples in genome-wide association studies. *Résumé 4th Paris workshop on Genomic Epidemiology*, 2011.
- Thomas Morisseau, Christine Sinoquet, Raphaël Mourad, Philippe Leray et Christian Dina. GWAS-AS : assistance for a thorough evaluation of advanced algorithms dedicated to genome-wide association studies. *Poster Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM)*, 2010.

Rapports de recherche

- Raphaël Mourad, Christine Sinoquet et Philippe Leray. Forests of hierarchical latent models for association genetics. LINA, Research Report, hal-00503013, 2010.

– Raphaël Mourad, Christine Sinoquet et Philippe Leray. Learning a forest of hierarchical Bayesian networks to model dependencies between genetic markers. LINA, Research Report, hal-00444087, 2010.

Divers

– Vittorio Perduca, Raphaël Mourad, Christine Sinoquet and Gregory Nuel. Waffect : a method to simulate case-control samples in genome-wide association studies. *Journées de Statistique (JDS)*, 2011.
– Raphaël Mourad, Christine Sinoquet et Philippe Leray. Hierarchical Bayesian networks applied to association genetics. *MODGRAPHII, journée satellite de JOBIM*, 2010.
– Raphaël Mourad, Christine Sinoquet et Philippe Leray. Réseaux bayésiens hiérarchiques avec variables latentes pour la modélisation des dépendances entre SNP : une approche pour les études d'association pangénomiques. *XVIIèmes Rencontres de la Société Francophone de Classification (SFC)*, pages 25–29, 2010.
– Raphaël Mourad, Christine Sinoquet et Philippe Leray. Modélisation des dépendances entre marqueurs génétiques à l'aide de réseaux bayésiens avec variables latentes : une approche pour les études d'association pangénomiques. *Session poster é-EGC2010, école d'hiver d'EGC*, 2010.
– Raphaël Mourad, Christine Sinoquet et Philippe Leray. Modélisation des dépendances locales entre SNP à l'aide d'un réseau bayésien. *XVIèmes Rencontres de la Société Francophone de Classification (SFC)*, pages 169–172, 2009.
– Raphaël Mourad, Christine Sinoquet et Philippe Leray. A Bayesian network approach to model local dependencies among SNPs. *MODGRAPH, journée satellite de JOBIM*, 2009.

Table des matières

Table des figures

Liste des tableaux

1 Introduction

Sommaire

1.1 Contexte

Parmi l'ensemble des maladies que nous pouvons rencontrer chez l'Homme, certaines sont influencées par des facteurs génétiques et sont ainsi appelées "génétiques" [24]. 17370 maladies génétiques humaines ont été recensées en 2007 [59]. Celles-ci peuvent impliquer des gènes ou des zones non géniques de l'acide désoxyribonucléique (ADN). Les maladies génétiques se subdivisent en deux grandes catégories : maladies monogéniques et maladies multifactorielles.

Les maladies de type monogénique, comme la mucoviscidose ou la phénylcétonurie, sont causées par la mutation d'un seul gène. Par exemple, la phénylcétonurie est due à une mutation dans la région du gène de la phénylalanine hydroxylase, une enzyme essentielle au bon fonctionnement de l'organisme [80]. L'analyse génétique de ce type de maladie s'est révélée relativement aisée et le progrès des connaissances a été considérable dans ce domaine [25]. Cela provient du fait que, pour ce genre de maladie, un seul facteur génétique est en cause. De plus, il est généralement le principal facteur de la maladie, l'influence de l'environnement demeurant souvent minime. Ainsi, la part des facteurs génétiques dans le déterminisme de la maladie, appelée aussi héritabilité, est souvent très importante [79]. Par exemple, la phénylcétonurie montre une héritabilité avoisinant 100% [41]. En conséquence, les études sur ces maladies parviennent facilement à détecter la corrélation très forte qui existe entre le gène et la maladie.

Les maladies multifactorielles sont causées conjointement par un ensemble de facteurs génétiques et environnementaux. Ces maladies sont nombreuses : les diabètes de types 1 et 2, l'hypertension artérielle, l'asthme, la polyarthrite rhumatoïde, la maladie de Crohn, certains cancers et certaines maladies mentales comme la schizophrénie, pour ne citer que quelques exemples. De plus, elles sont assez communes dans la population, ce qui en fait un enjeu majeur de santé publique. Le diabète de type 2 est un bon exemple de maladie multifactorielle [7, 20]. Il atteint des individus possédant des antécédents familiaux ainsi qu'un mode de vie particulier comme l'inactivité physique [8]. L'obésité est aussi connue comme un facteur à risque dans l'apparition de la maladie. L'étude des facteurs génétiques impliqués dans les maladies multifactorielles a débuté par les analyses de liaison sur des familles [105], mais ces dernières ont été le plus souvent infructueuses [70]. En effet, les analyses de liaison présentent de nombreux inconvénients : une localisation très imprécise de la mutation ou des mutations responsables, une faible sensibilité aux effets modestes des facteurs génétiques et une focalisation sur un sous-type de la maladie.

Les méthodes d'association basées sur l'étude de populations ont été proposées afin de résoudre ce problème [17]. Ces méthodes d'association sont fondées sur l'existence de dépendances statistiques, appelées déséquilibre de liaison (linkage disequilibrium, LD), généralement observées entre positions (ou loci) proches sur l'ADN (distances inférieures à 100 kilobases). Cette caractéristique biologique assure une localisation fine des mutations causales non observées à l'aide de marqueurs génétiques adjacents

(ce point important sera détaillé dans le chapitre 2, section 2.4.1, page 20). Récemment, les nouvelles technologies génomiques à haut débit ont permis d'accéder à la variabilité du génome entier à l'aide de centaines de milliers (voire plus d'un million) de marqueurs génétiques : les polymorphismes d'un seul nucléotide (single nucleotide polymorphism, SNP) figurent parmi les plus utilisés. Ces technologies ont ouvert la voie aux études d'association pangénomiques (EAP) [62] vouées à la localisation systématique des mutations causales sur le génome humain.

Dans ce contexte, un certain nombre de grands projets de recherche ont vu le jour pour la caractérisation systématique et approfondie de la variabilité génétique dans les populations humaines. Citons notamment le projet HapMap qui a été mis en œuvre en trois volets [96–98], et plus récemment, le projet 1000 génomes [95].

1.2 Objectif

Les recherches réalisées lors de la préparation de cette thèse de doctorat sont nées de la volonté de développer de nouvelles méthodes bioinformatiques pour l'analyse de données, en amont des études d'association pangénomiques, par exemple pour la réduction de dimension des données. Plus particulièrement, l'accent a été porté sur la modélisation du déséquilibre de liaison qui joue un rôle central dans les études d'association. Face à la très grande quantité de données, tant au niveau du nombre de variables (centaines de milliers) que du nombre d'individus (quelques milliers) à traiter, les méthodes à concevoir doivent faire preuve à la fois de performance et de capacité à passer à l'échelle. Dans ce contexte, l'application de nouvelles approches basées sur les modèles graphiques probabilistes (MGP) est apparue comme une piste de recherche intéressante, grâce à leur capacité à analyser un grand volume de données dans un cadre incertain, tout en assurant une finesse de modélisation dans les systèmes complexes.

Nos travaux se sont focalisés sur deux grands axes complémentaires : (i) la modélisation pangénomique du déséquilibre de liaison, et (ii) l'utilisation de cette modélisation, pour d'une part, la découverte de causalité, et d'autre part, la visualisation du déséquilibre de liaison.

1.3 Enjeux

Les enjeux de ces travaux sont nombreux. La modélisation pangénomique du déséquilibre de liaison permet de caractériser systématiquement les patrons de dépendance existants entre SNP, ouvrant ainsi la voie au généticien désirant, par exemple, comprendre la structure observée du LD dans le génome humain.

Outre cet aspect théorique, les applications suivantes de la modélisation du LD à l'aide de MGP sont attendues :

- le prétraitement des données, notamment par la réduction de leur dimension,
- la localisation précise des mutations causales à travers la connaissance de la structure de dépendances entre SNP,
- une visualisation intuitive de la structure du LD grâce à la nature graphique des MGP,
- la possibilité de simuler des données de SNP.

Les travaux de thèse présentés dans ce mémoire ont exploré les trois premiers points.

1.4 Plan du manuscrit

Ce manuscrit s'articule principalement autour de cinq grands chapitres, décrits ci-après. Chaque chapitre est généralement construit de la manière suivante : énoncé introductif des points abordés, développement et enfin conclusion. Ce manuscrit est conçu de manière à ce qu'un lecteur, qu'il soit statisticien, informaticien ou généticien, puisse le comprendre aisément. C'est la raison pour laquelle les deux premiers chapitres introduisent les notions élémentaires de génétique d'une part et les MGP d'autre part. La théorie relative à la modélisation pangénomique du LD à l'aide de MGP puis les applications sont ensuite développées dans les deux chapitres suivants. Enfin, le manuscrit s'achève sur la conclusion relative aux travaux réalisés et pose les fondements des perspectives de recherche.

Le chapitre 2 introduit les notions élémentaires de génétique et de biologie en général, dans l'optique de mieux appréhender le LD et son rôle dans le contexte des études d'association pangénomiques. Afin de guider le lecteur non-généticien et non-biologiste de surcroît, les notions les plus simples de biologie moléculaire et cellulaire sont présentées et illustrées, telles que la cellule, le génome et l'ADN. Ensuite, les préceptes fondamentaux de génétique sont introduits : les mutations, les recombinaisons, les notions d'haplotype et de génotype, le déséquilibre de liaison, les SNP et leur génotypage à haut débit. Dans le cadre des maladies génétiques complexes, l'épidémiologie de ces maladies, la relation entre les marqueurs génétiques et la maladie, et les études d'association pangénomiques sont abordées. Ce premier chapitre se termine sur une présentation des nombreux défis que soulèvent les études d'association pangénomiques et sur les raisons de développer de nouveaux outils provenant du monde de l'intelligence artificielle, tels que les modèles graphiques probabilistes.

Le chapitre 3 introduit les principaux concepts relatifs aux modèles graphiques probabilistes. A titre de rappel, un ensemble de définitions simples est d'abord fourni. D'une part, les définitions concernent les concepts d'indépendance marginale et conditionnelle en théorie des probabilités, et les concepts d'entropie, d'entropie conditionnelle et d'information mutuelle en théorie de l'information. D'autre part, certaines définitions relatives à la théorie des graphes sont rappelées. Ensuite, deux grandes

familles de MGP sont présentées : les réseaux bayésiens (RB) et les réseaux de Markov (RM). Seule la première famille, les réseaux bayésiens, est abordée de manière approfondie, car faisant l'objet des travaux de doctorat présentés dans ce mémoire. Un exemple simple et pratique de mise en œuvre des RB est ensuite employé afin de donner l'intuition de ces modèles au lecteur non familier. Ensuite, les trois grandes problématiques associées à l'utilisation des RB sont présentées : l'inférence probabiliste, l'apprentissage de paramètres et enfin l'apprentissage de structure. L'accent est porté sur l'apprentissage de paramètres dans le cadre de données incomplètes (*e.g.* présence de variables latentes) et sur l'apprentissage de structure. Ces deux types d'apprentissage seront essentiels pour comprendre l'algorithme d'apprentissage de modèle proposé dans ce travail de recherche. En conclusion, sont exposées et discutées les motivations de l'application des modèles graphiques probabilistes au traitement des données, en amont des études d'association pangénomiques.

Le chapitre 4 réunit l'ensemble du travail relatif à la modélisation pangénomique du LD. Les enjeux et le contexte de la modélisation du LD sont d'abord introduits. Un état de l'art sur l'application des MGP à la modélisation du LD est présenté. Ensuite, le choix d'un nouveau modèle - la forêt de modèles hiérarchiques à classes latentes (FMHCL) - est discuté et argumenté, notamment au travers de ses nombreux avantages pour la modélisation du LD, tels que la possibilité de prendre en compte la nature floue du LD et de hiérarchiser les différents degrés de LD. Un état de l'art général sur l'apprentissage d'un modèle très proche, le modèle hiérarchique à classes latentes, est présenté. Un nouvel algorithme d'apprentissage, nommé CFHLC (Construction of Forests of Hierarchical Latent Class models), est proposé et détaillé. Une seconde version (CFHLC+) est ensuite présentée. Elle offre l'avantage de ne plus nécessiter le découpage du génome en fenêtres. Notre approche de modélisation est ensuite évaluée sur des données simulées et réelles. Nous réalisons la comparaison avec des approches concurrentes basées sur un modèle en blocs (SNP contigus) ou encore basées sur un modèle en clusters (SNP non contigus), notamment en nous appuyant sur le jeu de données bien connu de Daly *et al.* [19]. Le passage à l'échelle de l'algorithme aux données pangénomiques est aussi empiriquement prouvé. Ce troisième chapitre s'achève sur la récapitulation des nombreuses perspectives d'application de la modélisation proposée.

Le chapitre 5 aborde deux applications de la modélisation proposée : (i) la recherche de causalités, et (ii) la visualisation synthétique et intuitive du LD dans des contextes variés. La partie (i) se focalise sur une étude systématique de la capacité des FMHCL à la découverte de causalités et est illustrée dans le contexte de la génétique d'association. Ce travail établit les bases du développement de nouvelles méthodes de recherche dédiées à la découverte de SNP causaux pour les études d'association pangénomiques. La partie (ii) propose une méthode originale pour la visualisation synthétique et intuitive du LD adaptée aux trois principales situations que peut rencontrer le généticien : la visualisation du LD de courte distance, de longue distance et dans un contexte pangénomique. Chacune des parties s'achève sur une conclusion des travaux de recherches et sur la présentation des perspectives.

Le dernier chapitre récapitule l'ensemble des contributions apportées lors de ces travaux de recherche, mais ouvre également sur différentes perspectives scientifiques relatives aux possibilités d'amélioration de la modélisation du LD proposée, à l'application à la découverte de SNP causaux et enfin au développement d'un outil de visualisation du LD.

2

Le déséquilibre de liaison dans le contexte des études d'association pangénomiques

Sommaire

2.1 Introduction

L'objectif de ce chapitre est d'introduire les notions élémentaires de génétique et de biologie en général, dans l'optique de mieux appréhender le déséquilibre de liaison et son rôle dans le contexte des études d'association pangénomiques. Afin de guider le lecteur non-généticien et non-biologiste de surcroît, les notions les plus simples de biologie moléculaire et cellulaire sont présentées et illustrées, telles que la cellule, le génome et l'ADN. Ensuite, les préceptes fondamentaux de génétique sont introduits : les mutations, les recombinaisons, les notions d'haplotype et de génotype, le déséquilibre de liaison, les mutations au niveau d'un seul nucléotide (single nucleotide polymorphism, SNP) et leur génotypage à haut débit. Dans le cadre des maladies génétiques complexes, l'épidémiologie de ces maladies, la relation entre les marqueurs génétiques et la maladie, et les études d'association pangénomiques sont abordées. Ce premier chapitre se termine sur une présentation des nombreux défis que soulèvent les études d'association pangénomiques et sur les raisons de développer de nouveaux outils provenant du monde de l'intelligence artificielle, tels que les modèles graphiques probabilistes.

2.2 Préceptes de biologie moléculaire et cellulaire

La cellule est l'unité structurale, fonctionnelle et reproductrice de tous les êtres vivants (à l'exception des virus) [3]. Elle est le siège de la plupart des activités métaboliques. L'organisme humain serait constitué de plus de 10^{13} cellules eucaryotes. Les cellules présentent une taille moyenne de 10 microns. Celles-ci sont regroupées en tissus cellulaires, puis en organes, et constituent enfin l'organisme entier.

A l'intérieur de la cellule, le noyau est le siège de l'immense majorité du patrimoine génétique, appelé aussi génome. Dans une certaine mesure, le génome influence le développement des individus et de leurs caractéristiques, appelées aussi traits ou phénotypes. A titre d'exemple, citons la taille d'un individu et la couleur de ses yeux comme caractéristiques extérieures, et sa pression artérielle et sa capacité à métaboliser une molécule pharmaceutique comme caractéristiques physiologiques. Le support de l'information génétique est la molécule d'acide désoxyribonucléique (ADN). Cette dernière est composée de deux brins complémentaires et antiparallèles (*i.e.* en sens opposés) se faisant face et formant une double hélice (voir figure 2.1a). Chaque brin est un polymère (ou séquence) de nucléotides. Chaque nucléotide est composé de 3 molécules : une molécule d'acide phosphorique, une molécule de désoxyribose et une base azotée. Quatre bases différentes existent : l'adénine (A), la guanine (G), la cytosine (C) et la thymine (T). Ainsi, le génome constitue-t-il un code formé à partir d'un alphabet à 4 lettres. Chez l'Homme, la taille du génome est importante, elle est d'environ $3,4$ milliards de paires de bases, ce qui lui confère une grande complexité.

Les molécules d'ADN associées à des protéines, notamment des histones, forment des chromosomes au sein du noyau de la cellule (voir figure 2.1b). L'organisme humain

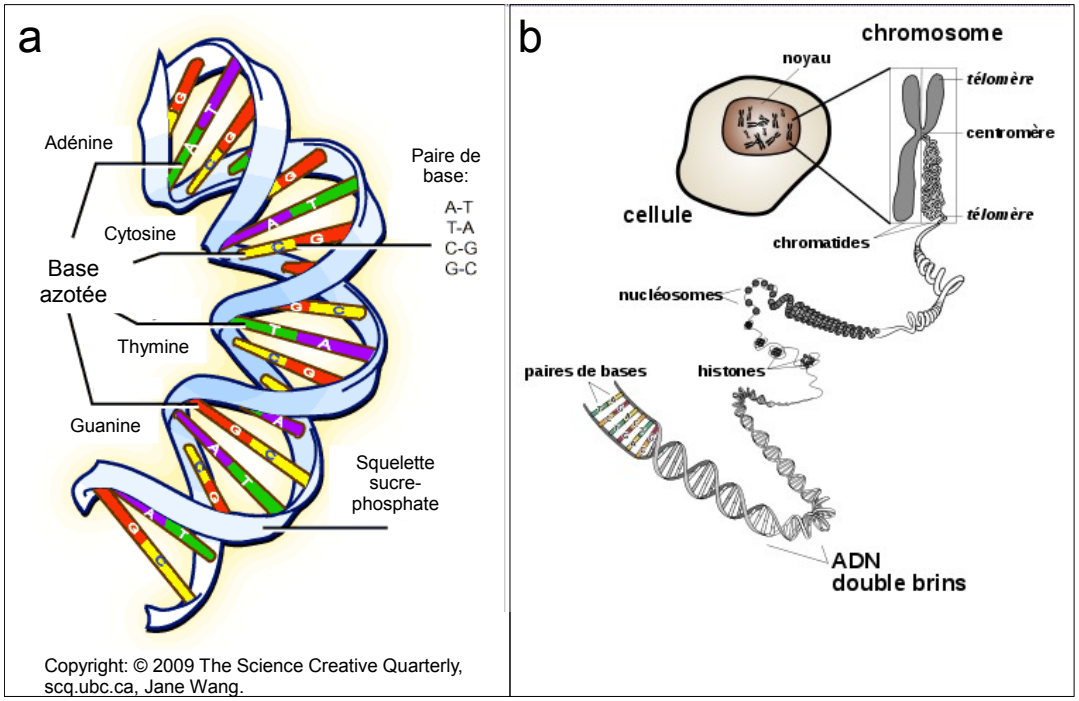

FIGURE 2.1: a) Molécule d'acide désoxyribonucléique (ADN). b) De l'acide désoxyribonucléique au chromosome.

est diploïde, c'est-à-dire qu'il possède deux copies de chaque chromosome (appelées chromosomes homologues). Les chromosomes sont au nombre de 23 paires : 22 paires de chromosomes dits autosomes et une paire de chromosomes sexuels. Ces derniers sont identiques chez la femme (XX) mais différents chez l'homme (XY).

En ce qui concerne le génome humain, les récentes estimations révèlent un nombre de gènes compris entre 20000 et 25000 [78]. Le gène est l'unité d'information génétique. Un gène code généralement pour la formation d'un acide ribonucléique (ARN) et d'une protéine, et par ce biais, il participe à l'élaboration du phénotype. Les gènes ne représentent qu'une partie du génome (moins de 30%) et les régions codantes des gènes, appelées exons, n'occupent même pas 3% du génome. Les gènes ne sont pas uniformément distribués dans le génome. Des domaines riches en gènes alternent avec des régions pauvres. Les autres régions sont composées de séquences non codantes dont les fonctions n'ont pas encore été tout à fait éclaircies. Ces régions sont appelées aussi régions non géniques : elle joueraient de nombreux rôles tels que des rôles structurels et régulateurs [43].

2.3 Préceptes de génétique

2.3.1 Les mutations

L'existence des mutations provient de la non-fiabilité du système héréditaire impliquant la molécule d'ADN. En effet, la capacité de l'ADN à stocker, répliquer, transmettre et coder l'information n'est pas parfaite [45]. Les mutations résultantes de ces erreurs sont à l'origine de la variabilité génétique des organismes vivants, générant de nouveaux allèles (*i.e.* nouvelles versions) pour les gènes et façonnant le génome. Cette variabilité génétique introduit une variabilité phénotypique chez les êtres vivants et leur permet de s'adapter aux nouvelles conditions de leur environnement, et sur le long terme, d'évoluer vers de nouvelles espèces. L'exemple le plus emblématique ce phénomène est certainement celui du phalène du bouleau [1] dont la forme sombre s'est répandue autour des villes industrielles d'Angleterre au cours du XIX^{e} siècle, en raison de la pollution atmosphérique par les résidus de combustion du charbon. Cependant, à plus court terme, ces mêmes mutations sont aussi à l'origine de la mort cellulaire, de cancers et de maladies génétiques.

En outre, les mutations permettent l'analyse génétique. La variabilité introduite par les mutations permet aux généticiens d'identifier et de localiser les facteurs génétiques impliqués dans les maladies génétiques. Ainsi, les mutations, telles que celles survenant au niveau d'un seul nucléotide (*i.e.* SNP), peuvent servir de "marqueurs génétiques". Cet aspect des mutations sera détaillé dans la section 2.4.1, page 20.

Une mutation est définie comme une modification de la séquence d'ADN du génome. Tout changement nucléotidique est considéré comme mutation. Les mutations sont diverses : substitution d'un nucléotide, délétion ou insertion d'une ou plusieurs paires de bases, modification majeure du chromosome. Les mutations peuvent avoir différents impacts sur l'organisme. Une mutation peut affecter un phénotype (*i.e.* un caractère). Si elle en affecte plusieurs, elle est appelée mutation pléiotrope. Dans le cas où elle n'induit aucune modification du phénotype, il s'agit d'une mutation silencieuse.

Les mutations peuvent survenir à l'intérieur d'un gène, et plus particulièrement, dans les régions codant la protéine (exons). Elles peuvent aussi être localisées à l'extérieur du gène, par exemple, dans les régions régulatrices. C'est la raison pour laquelle les mutations peuvent avoir des effets divers sur le phénotype.

2.3.2 Les recombinaisons

Les gènes se distribuent le long des chromosomes. *A priori*, lors de la transmission du matériel génétique à la descendance, les allèles de chaque gène devraient être

1. La phalène du bouleau (*Biston betularia* L.) est un insecte de l'ordre des lépidoptères, de la famille des géométridés. C'est un papillon nocturne des régions tempérées. Il est souvent cité comme exemple d'adaptation à l'évolution de son milieu naturel par mutation puis sélection naturelle.

transmis en un seul bloc du fait du lien physique existant entre eux sur le chromosome. Cependant, il est observé qu'un échange de séquences d'ADN a lieu entre les chromosomes homologues lors de leur appariement en prophase de méiose 1.

Le crossing-over, que nous appellerons aussi plus simplement "recombinaison", désigne le phénomène par lequel apparaissent des combinaisons génétiques nouvelles, dans une cellule ou un individu. Ces combinaisons seront différentes de celles observées chez les cellules ou individus parentaux. La figure 2.2a illustre la recombinaison entre deux chromosomes homologues. Trois gènes α, β et γ sont étudiés. Chez le parent, le premier chromosome, coloré en bleu (foncé), possède la séquence ABC tandis que le second chromosome, coloré en jaune (clair) possède la séquence abc. Après recombinaison entre le gène α et le gène β, la séquence du chromosome 1 est Abc et celle du chromosome 2 est aBC.

Les recombinaisons sont des phénomènes aléatoires. Le nombre de recombinaisons entre deux gènes d'un même chromosome est proportionnel à la distance physique les séparant sur l'ADN. Il est alors possible de calculer un second type de distance, différent de la distance physique. Cette seconde distance est appelée distance génétique et s'estime à partir du taux de recombinaison. Son unité de mesure est le centiMorgan (cM).

Le taux de recombinaison mesure le pourcentage de séquences recombinantes dans la descendance. Il varie de 0 (lorsque les deux séquences sont totalement liées génétiquement) à $0,5$ (lorsque les deux séquences ne sont pas du tout liées).

2.3.3 Génotype et haplotype

L'haplotype se définit par une séquence de loci, qu'ils soient gènes ou marqueurs génétiques, sur un même chromosome. Le génotype, quant à lui, correspond à la composition allélique (*i.e.* en termes de nombre d'allèles) de chaque locus de la séquence. Pour illustrer la différence entre haplotype - ou génotype phasé - et génotype, prenons l'exemple suivant présenté en figure 2.2b. Dans cet exemple, nous considérons une séquence de 3 loci répartis sur un chromosome. L'observation du génotype $AABbCC$ ne fournit pas la connaissance de la phase gamétique, *i.e.* elle ne permet pas d'établir la distinction entre les deux combinaisons de paires d'haplotypes possibles : $\frac{ABC}{AbC}$ et $\frac{AbC}{ABC}$ (haplotype maternel au-dessus et paternel en dessous). Actuellement, seule la connaissance du génotype est accessible à faible coût grâce aux nouvelles techniques de génotypage à haut débit. Le génotypage des haplotypes est, quant à lui, très long et coûteux. Pour connaître la phase gamétique, des méthodes statistiques d'inférence des haplotypes à partir des génotypes ont été développées [13, 82], pour n'en citer que quelques unes.

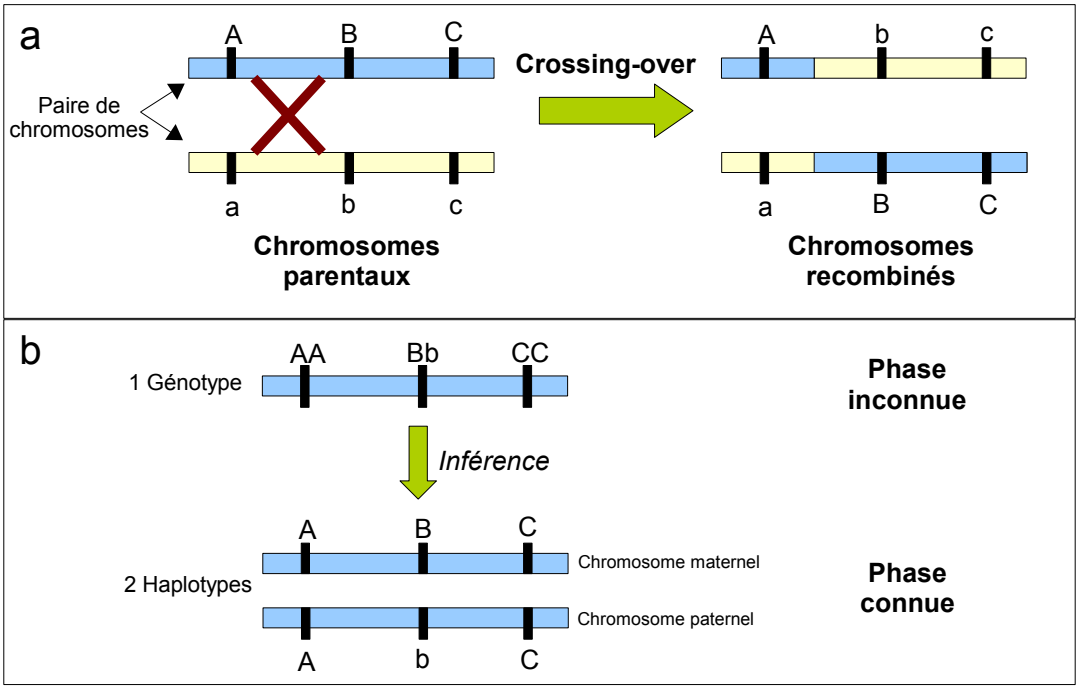

FIGURE 2.2: a) Schéma d'une recombinaison entre deux chromosomes homologues. b) Illustration de la différence entre génotype et haplotype.

2.3.4 Évolution des haplotypes

Les haplotypes du génome humain tirent leurs origines des mécanismes moléculaires de la reproduction sexuée et de l'histoire évolutive de notre espèce. Au fil des générations, de nouveaux haplotypes apparaissent à partir des haplotypes ancestraux par le biais des mutations des séquences d'ADN et des recombinaisons ayant lieu entre eux. Les nouveaux haplotypes se répandent dans la population par le biais des phénomènes démographiques et évolutifs comme la consanguinité, la dérive génétique ou la sélection naturelle [77]. Pour une introduction plus exhaustive à l'évolution des haplotypes, le lecteur peut se référer à Réf.[71].

Les phénomènes de recombinaison exercent une grande influence sur la variabilité des haplotypes. Prenons l'exemple de la figure 2.3. Nous observons la formation des haplotypes actuels dans une population à partir de deux haplotypes ancestraux. Si une mutation est présente sur l'haplotype ancestral 1 (en rouge) et absente de l'haplotype ancestral 2 (en bleu), alors une partie de la descendance possédera la mutation. Qui plus est, les individus possédant cette mutation présenteront de fortes chances que la région chromosomique voisine de la mutation soit identique à celle correspondante sur le chromosome ancestral 1. Ceci s'explique par le fait que la probabilité que des évènements de recombinaison se soient produits entre la mutation et sa région avoisinante est très faible. Ce constat illustre le concept de déséquilibre de liaison entre loci proches qui sont transmis de manière corrélée au fil des générations. Nous pouvons

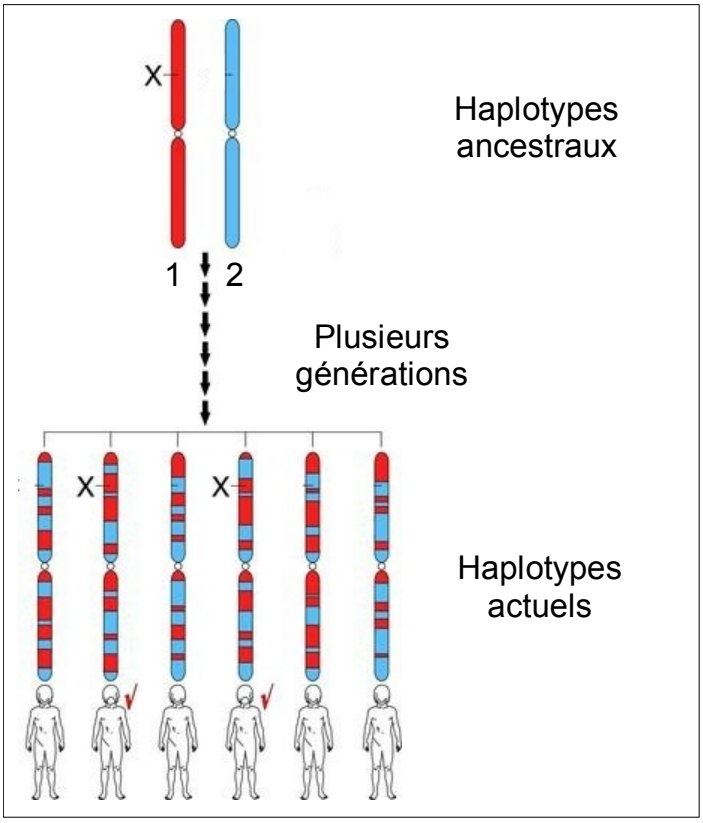

FIGURE 2.3: Évolution des haplotypes avec transmission d'une mutation.
La mutation est symbolisée par X. *Copyright 2009 http ://www.hapmap.org/.*

aussi nous apercevoir que le déséquilibre de liaison permet de tracer une mutation à l'aide de marqueurs proches (donc en déséquilibre de liaison avec la mutation), et fonde ainsi l'approche des études dites "d'association". Le secteur disciplinaire qui met en œuvre de telles études se nomme la génétique d'association.

2.3.5 Le déséquilibre de liaison

Formellement, le déséquilibre de liaison (linkage disequilibrium, LD) est l'association non aléatoire entre les allèles de deux loci (ou plus) à l'intérieur d'une population donnée [57]. Le LD est estimé généralement à partir d'haplotypes, mais peut aussi s'estimer à partir de génotypes. L'estimation du LD sera abordée dans la suite de la section. Les généticiens [66, 75] classent le LD en deux catégories : l'un très fréquent et s'observant sur de courtes distances, *i.e.* < 10 kilobases (kb), et l'autre plus rare et s'observant sur de longues distances ($> 100kb$).

A l'origine, le LD provient du lien physique existant entre les loci présents sur un même chromosome et dont les allèles sont transmis de générations en générations. Ce

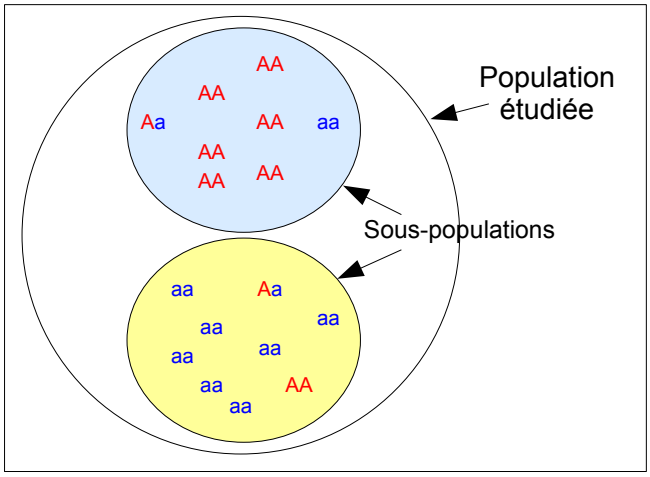

FIGURE 2.4: Illustration du phénomène de stratification de la population.
Pour le gène α, deux allèles A et a existent. L'allèle A est plus fréquent dans la sous-population présentée en haut de la figure que dans la sous-population présentée en bas de la figure.

phénomène est appelé "liaison génétique". De nombreux facteurs peuvent accroître ou diminuer le LD. Le facteur le plus important est la recombinaison génétique qui casse le lien physique entre loci d'un même chromosome. Avec le temps, *i.e.* au fur et à mesure des générations, des évènements de recombinaison entraînent la disparition du LD entre deux loci. Étant de nature aléatoire, le nombre d'évènements de recombinaison entre deux loci va dépendre de la distance les séparant. Plus les loci sont proches, plus le taux de recombinaison sera faible, et plus le LD diminuera lentement. Ce phénomène a notamment été modélisé par Malécot [54]. Un autre facteur important est le phénomène de stratification des populations, nommé aussi structure de la population, dans le contexte de la génétique d'association. Ce facteur crée un LD artefactuel entre deux loci, dû à la présence de plusieurs sous-populations présentant des fréquences alléliques différentes (voir figure 2.4), à l'intérieur de la population étudiée. Par exemple, il est souvent à l'origine du LD observé entre loci éloignés de plus de 500 *kb*, ou parfois observé entre chromosomes différents. D'autres facteurs influent sur le LD mais de manière généralement moins significative, comme la dérive génétique, la sélection, la consanguinité, la mutation, le flux de gènes et la taille de l'échantillon étudié [47, 75].

L'analyse du LD sur les données haplotypiques a révélé l'existence de zones présentant de très fortes corrélations, appelées blocs haplotypiques ou haploblocs, séparées par de courtes zones montrant de faibles corrélations [96–98]. Les travaux précurseurs de Daly *et al.* [19] sur une séquence de 500 *kb* provenant de la région 5q31 chez 256 patients atteints de la maladie de Crohn a révélé le faible nombre d'haplotypes différents à l'intérieur des blocs. Par exemple, pour 7 blocs s'étendant sur une distance de 92 *kb* et contenant 31 SNP, seulement 4 des 231 haplotypes possibles ont été observés pour 94% des chromosomes. Ces blocs mesurent en moyenne de 10 à 20 *kb*, mais peuvent varier de quelques kbs à plusieurs centaines de kbs. Cette structure en

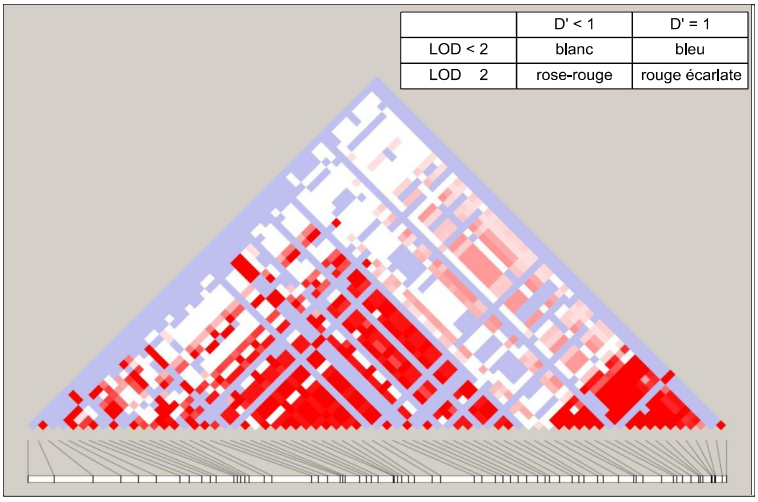

FIGURE 2.5: Carte triangulaire de chaleur du déséquilibre de liaison (LD) d'une séquence réelle de 500 *kb*.
Génome humain, chromosome 1, *région [10 000 kb - 10 500 kb]. Le LD est évalué à l'aide de la matrice triangulaire des dépendances (mesurées par le LOD et le* D'*) pour chaque paire de SNP. Le LOD et le* D' *sont des mesures du LD qui seront détaillées à la fin de la section. Pour une paire de SNP donnée, plus la couleur est rouge (sombre), plus les SNP sont en LD.*

blocs s'explique par le fait que les évènements de recombinaison ne sont pas distribués de manière homogène sur une échelle fine : le taux de recombinaison peut varier grandement d'une position chromosomique à une autre. Les blocs de LD sont donc des régions à faibles taux de recombinaison séparées par des points chauds (hotspots, en anglais) de recombinaison. Cette structure en blocs peut être aisément visualisée à l'aide de la carte triangulaire dite de "chaleur" (CTC) du LD par paires. Dans la figure 2.5, on peut discerner 2 haploblocs (triangles rouges) sur une séquence de 500 *kb*. Les mesures de LD comme le D' et le LOD seront présentées par la suite dans la section.

Un aspect intéressant du LD est qu'il réduit le nombre de marqueurs nécessaires pour capturer la grande majorité de la variation génétique [47]. Ainsi, il est possible de ne génotyper qu'un nombre réduit de SNP afin de réaliser des études d'association. Le projet HapMap est issu en partie de cette observation. La notion de SNP sera détaillée en section 2.3.6, page 17.

Le projet international HapMap est né d'un partenariat de scientifiques et d'organismes de financement regroupant le Canada, la Chine, le Japon, le Nigeria, le Royaume-Uni et les États-Unis. L'objectif est de développer une ressource publique afin d'aider les chercheurs à découvrir les gènes associés aux maladies humaines et à la variabilité de la réponse aux médicaments. Le consortium du projet s'est fixé les objectifs suivants [61] :

– créer une base de données publique des variations génotypiques et haplotypiques

des SNP pour 4 populations humaines : 30 trios (composés d'un individu et de ses deux parents) d'Ibadan au Nigéria (YRI), 30 trios de résidents de l'Utah ayant des origines du nord-ouest de l'Europe (CEU), 44 individus non apparentés de Tokyo au Japon (JPT) et 45 individus Han de Beijing en Chine (CHB),
- génotyper toutes les 5000 paires de bases en moyenne un SNP commun, *i.e.* présentant une fréquence de 5% dans la population,
- permettre la détermination d'un ensemble de tagSNPs (SNP étiquettes) : l'identification de ces tagSNPs est destinée à aider le généticien lors de la recherche des mutations causales, en réduisant par exemple le nombre de tests d'association SNP-maladie à réaliser et en améliorant ainsi la puissance de détection (capacité à détecter) des mutations causales,
- aider au développement de technologies de génotypage, d'outils d'analyse et d'études génétiques sur les maladies complexes.

Les mesures de LD sont des mesures d'association qui quantifient l'écart entre les fréquences haplotypiques observées et celles attendues sous hypothèse d'indépendance entre les allèles (hypothèse nommée équilibre de liaison) [57]. Nous nous focaliserons sur les mesures de LD les plus couramment utilisées : D, D', r^2 et LOD. Ce sont des mesures réalisées sur des paires de loci.

Prenons l'exemple suivant : deux loci α et β possèdent chacun deux allèles, A et a, et B et b, respectivement. Soit p_A, p_a, p_B et p_b les fréquences alléliques et p_{AB}, p_{Ab}, p_{aB} et p_{ab} les fréquences haplotypiques. Une mesure très simple du LD consiste à évaluer la différence entre la fréquence observée d'un haplotype donné et celle attendue sous hypothèse d'indépendance entre les loci α et β. Cette mesure s'appelle le coefficient de déséquilibre de liaison (D) et sa formule est la suivante [23] :

$$D = p_{AB} - p_A\ p_B = p_{ab} - p_a\ p_b. \tag{2.1}$$

Plus D est élevé, plus les loci sont en déséquilibre de liaison. Des mesures standardisées ont été proposées afin de disposer de valeurs comprises entre -1 et 1. D' et r^2 sont les plus usuelles. D' a été introduit par Lewontin [50] et sa formule est la suivante :

$$D' = \frac{D}{D_{max}}$$

avec

$$D_{max} = \begin{cases} min(p_A\ p_b; p_a\ p_B) & \text{si } D > 0 \\ min(p_a\ p_b; p_A\ p_B) & \text{si } D < 0. \end{cases}$$

D' a la propriété de prendre les valeurs -1 et 1 lorsqu'un ou plusieurs des haplotypes sont absents dans la population. Prenons l'exemple présent dans le tableau 2.1.

Dans cet exemple, $D = -0,0593$ et $D' = -1$. Nous observons qu'un des quatre haplotypes est absent de la population : on parle alors de déséquilibre complet. Ainsi, D' n'indique pas si le premier locus porte toute l'information du second. De surcroît, des valeurs de D' inférieures à 1 (ou supérieures à -1) n'ont pas d'interprétation précise quant au niveau de dépendance existant entre les loci, et dépendent de la

	b	B
a	0.1029	0.0719
A	0.8252	0

Tableau 2.1: Exemple de déséquilibre complet.

taille de l'échantillon. C'est la raison pour laquelle r^2 est préféré à D' car il mesure la quantité d'information que fournit un locus sur l'autre. Sa formule est la suivante :

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

Avec r^2, une valeur de 1 signifie que le premier locus porte toute l'information du second locus, et *vice versa*. Dans ce cas, seuls deux des quatre haplotypes possibles sont observés. Cette situation est appelée déséquilibre parfait. Prenons l'exemple présent dans le tableau 2.2.

	b	B
a	0.5763	0
A	0	0.4237

Tableau 2.2: Exemple de déséquilibre parfait.

Ici, $D = 0,2442$, $D' = 1$ et $r^2 = 1$, alors que précédemment r^2 était égal à $0,6044$. Enfin, une autre mesure de LD fréquemment employée est le logarithme du rapport des chances (log of odds), noté LOD. Celui-ci évalue le logarithme décimal du rapport entre la vraisemblance sous hypothèse de LD et la vraisemblance sous hypothèse d'équilibre de liaison.

Dans la pratique, nous n'avons généralement accès qu'à la connaissance des génotypes lors des études populationnelles et nous ne pouvons donc pas mesurer directement le déséquilibre de liaison sur les haplotypes. Afin de contourner ce problème, les méthodes d'inférence d'haplotypes peuvent être employées.

2.3.6 Les SNP et leur génotypage à haut débit

Les SNP désignent les variations d'un seul nucléotide sur un site particulier du génome, entre individus d'une même espèce et pour une population donnée. Les généticiens emploient les SNP comme marqueurs génétiques de sites particuliers du génome. Prenons l'exemple des deux séquences en Figure 2.6.

Séquence 1 :	AATCGAT
Séquence 2 :	AATGGAT
	*

FIGURE 2.6: Exemple de variation d'un seul nucléotide (SNP).

Sur le tableau 2.6, le nucléotide en position 4 diffère entre les deux séquences : la séquence 1 présente la lettre C tandis que la séquence 2 présente la lettre G. Ainsi, nous venons d'identifier un SNP : c'est un nucléotide qui peut varier entre les différentes séquences présentes dans la population étudiée. En biologie, nous utilisons le terme de polymorphisme de nucléotide (SNP : single nucleotide polymorphism).

Chez l'Homme, il existerait une dizaine de millions de SNP couvrant le génome entier (parmi les 3,4 milliards de paires de bases), comme l'a montré le projet HapMap [96–98]. Les SNP constituent les variations les plus fréquentes dans le génome humain : environ 90%. Il a été estimé en moyenne qu'un nucléotide sur 1200 varie d'une personne à l'autre. Par ailleurs, les SNP apparaissent sur le génome toutes les 100 à 300 paires de bases en moyenne. Les SNP se présentent, dès lors, comme des outils très fins de mesure des variations génétiques chez l'Homme. Le génotypage permet de réaliser ces mesures. Pour un SNP, les quatre versions alléliques (ou allèles) A, T, G et C peuvent exister. Cependant, ne sont généralement considérées que les deux versions les plus fréquentes dans la population d'étude. En outre, l'être humain est un organisme diploïde, ce qui signifie qu'il possède deux allèles pour chaque SNP. Ainsi, le génotype d'un SNP fournit une mesure à trois modalités. Par exemple, si les deux allèles les plus fréquents d'un SNP sont A et T, les génotypes possibles sont les suivants : $\frac{A}{A}$, $\frac{A}{T}$, $\frac{T}{A}$ et $\frac{T}{T}$. Les techniques actuelles ne permettent pas de distinguer le génotype $\frac{A}{T}$ du génotype $\frac{T}{A}$, et ainsi, ces derniers ne forment qu'une même classe.

Des méthodes à haut débit utilisant la technologie des puces à ADN ont été développées par les industriels pour le génotypage des SNP (de manière analogue aux puces d'expression génique). Ainsi, il est désormais possible de génotyper sur une puce plus d'une centaine de milliers de marqueurs pour un individu donné, ce qui a ouvert la voie au développement des études pangénomiques, c'est-à-dire à l'échelle du génome. En outre, la technologie évolue rapidement et la dernière puce développée par Affimetrix®, nommée Genome-Wide Human SNP Array 6.0, permet de génotyper plus de 900000 SNP. Une puce est constituée d'une plaque de quelques centimètres carrés, sur laquelle sont déposés, selon un quadrillage, des centaines de milliers de spots d'ADN. Chacun de ces spots contient une quantité infime d'ADN, de l'ordre du picomole (10^{-12} mole), servant de sonde pour chaque SNP. La méthode de génotypage est illustrée en figure 2.7. Le principe est simple : l'ADN nucléaire est d'abord extrait des cellules étudiées (*e.g.* cellules sanguines). Puis, l'ADN nucléaire est fragmenté au moyen d'enzymes de restriction. Les fragments obtenus sont alors amplifiés (reproduits en de nombreux exemplaires) par la méthode de réaction en chaîne par polymérase (polymerase chain reaction, PCR), permettant ainsi d'obtenir de grandes

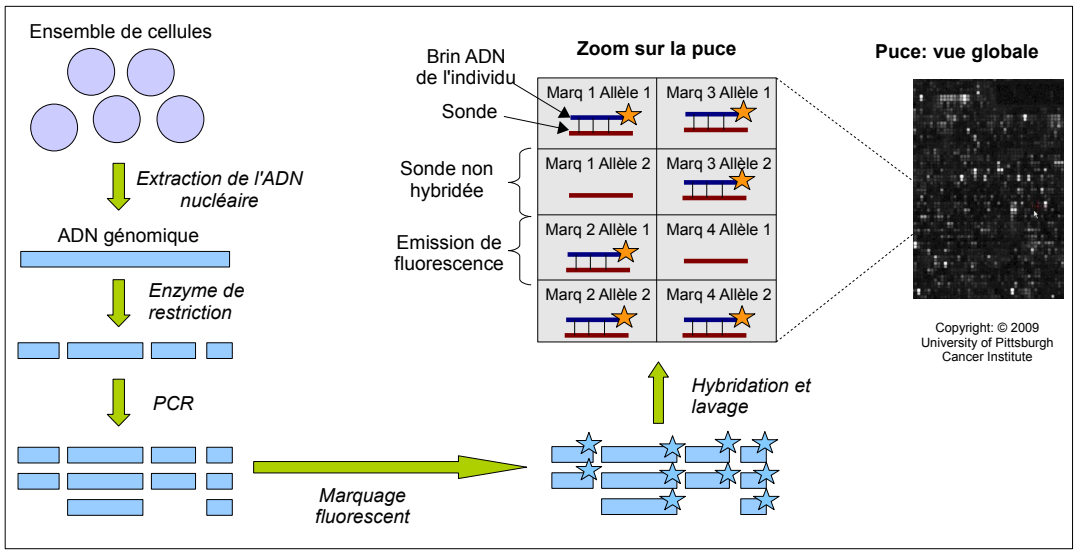

FIGURE 2.7: Génotypage à haut-débit à l'aide d'une puce Affymetrix®.
Toutes les étapes ne sont pas détaillées. PCR : réaction en chaîne par polymérase ; Marq : marqueur.

quantités d'ADN. Ensuite, les fragments sont marqués à l'aide d'une molécule fluorescente avant d'être déposés sur la puce. Les simples brins d'ADN sont capables de s'hybrider avec leur sonde complémentaire si cette dernière existe sur la puce. La puce est ensuite lavée ; l'ADN n'ayant pu s'hybrider est ainsi éliminé. Chaque fois qu'un brin d'ADN de l'individu a pu s'hybrider avec une sonde de la puce, un signal fluorescent est émis, en cet endroit de la puce. Enfin, la puce est scannée et une photographie est prise. Nous pouvons déduire alors, pour chaque SNP, si l'individu possède 0, 1 ou 2 copies de l'allèle considéré selon que le signal est soit, d'intensité 0, x ou $2x$, respectivement. Ces puces permettent de répondre à diverses problématiques comme la découverte de mutations causales dans le cadre des études d'association, que nous verrons plus en détail dans la section 2.4.3, page 23, le diagnostic de maladies génétiques [84], ou encore, l'étude de la variabilité génétique des populations humaines [14, 92]. Il faut souligner que de nouvelles technologies, dites de "reséquençage", ont été récemment développées afin d'accéder directement à la séquence génomique des individus [95, 111].

2.4 Épidémiologie des maladies multifactorielles

Les maladies multifactorielles, aussi appelées maladies complexes ou communes, sont définies comme les maladies génétiques causées conjointement par un ensemble de facteurs génétiques et environnementaux (voir section 1.1, chapitre 1, page 2). L'étude de l'ensemble de ces facteurs, et par extension l'ensemble de ces facteurs, est

nommée étiologie. L'hypothèse la plus communément admise pour expliquer ces maladies est l'hypothèse "maladie commune-variants communs" (common disease-common variant, CDCV) [38]. Elle postule que l'accumulation de diverses causes fréquentes dans la population, ayant chacune un effet faible sur la maladie, détermine l'apparition de la maladie. La seconde hypothèse est l'hypothèse "maladie commune-variants rares" (common disease-rare variant, CDRV) [38]. A l'inverse, selon cette seconde hypothèse, les variants génétiques ne sont présents que dans une faible portion de la population, mais présentent des effets importants. La première hypothèse a constitué le paradigme scientifique pour le développement des études d'association pangénomiques visant l'élucidation des mécanismes génétiques des maladies multifactorielles. Du fait du succès encore mitigé des études d'association pangénomiques, le choix entre la première et la seconde hypothèse anime un vif débat. Parmi les diverses causes génétiques, doivent être distingués les effets additifs des facteurs génétiques, les interactions gène-gène (nommé épistasie) et les interactions gène-environnement. Au sein des causes environnementales, il existe à la fois des effets additifs et des effets d'interaction.

Comme Josué Feingold l'a montré [25], les chercheurs rencontrent de nombreuses difficultés lors de l'étude de ce type de maladies. Par exemple, les résultats de nombreuses études de liaison ou d'association n'ont pu être confirmés par d'autres études. Certaines causes possibles de ces échecs ont été envisagées. Par exemple, l'analyse d'un grand nombre de marqueurs génétiques a pour conséquence d'entraîner statistiquement la détection, parfois importante, de résultats faussement positifs. Sur le plan épidémiologique, certains échantillons de malades présentent des différences au niveau de la gravité de la maladie et de son stade évolutif, ce qui peut mener à des résultats contradictoires. Enfin, un dernier facteur à prendre en considération est l'hétérogénéité génétique des maladies multifactorielles, *i.e.* le fait que différentes combinaisons de facteurs génétiques peuvent induire l'apparition de la même maladie. L'hétérogénéité génétique peut rendre difficile l'interprétation des études d'association, puisque le modèle génétique de la maladie n'est pas le même pour tous les individus. Cela peut aboutir à des conclusions peu concordantes concernant les facteurs génétiques impliqués dans la maladie.

2.4.1 Les marqueurs génétiques et leur relation à la maladie

Nous venons de voir qu'un ensemble de facteurs génétiques influence les maladies multifactorielles. Différents types de mutations peuvent avoir été à l'origine de la maladie. Les SNP n'en représentent qu'un type, mais il en existe bien d'autres comme les insertions, les délétions et les inversions de séquences sur l'ADN.

Alors pourquoi n'employer que les SNP si ces derniers ne représentent qu'une partie des mutations sur l'ADN ? En fait, pour les généticiens, les SNP ne représentent pas seulement des mutations sur l'ADN, ce sont aussi des marqueurs génétiques qui permettent de "baliser" le génome afin d'en étudier la variation. Dans le cadre des

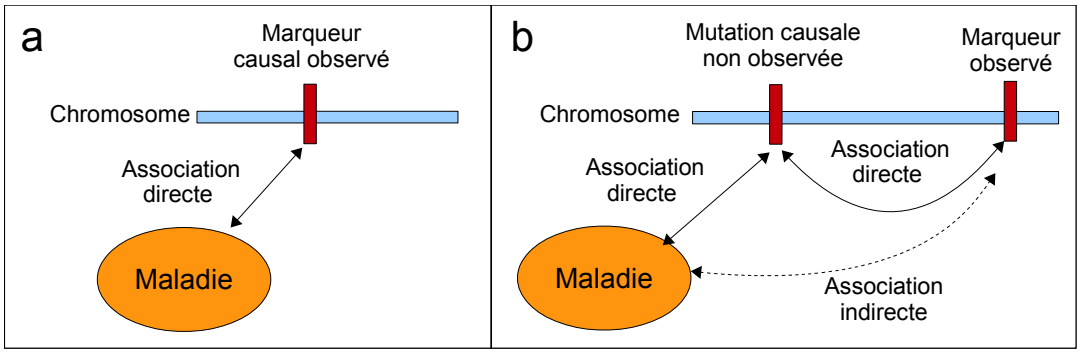

FIGURE 2.8: a) Association directe maladie-marqueur observé. b) Association indirecte maladie-marqueur non observé.

études de génétique d'association, le déséquilibre de liaison existant entre loci proches joue un rôle central dans la localisation des mutations causales ne faisant pas partie du lot de SNP observés. Il existe deux types de situations auxquelles le généticien peut être confronté : (i) la dépendance directe entre une mutation portée par le SNP observé et le phénotype (figure 2.8a), et (ii) la dépendance indirecte entre un SNP observé et le phénotype, via la mutation causale non observée, en LD avec le SNP observé (figure 2.8b). Dans ces deux situations, le généticien est capable de localiser la mutation causale avec grande précision, pourvu que la région chromosomique étudiée soit balisée avec suffisamment de SNP. Leur abondance, le fait que les SNP soient génétiquement stables et peuvent être génotypés facilement et à bas coût grâce aux nouvelles technologies génomiques à haut débit ont justifié l'emploi des SNP lors des études d'association pangénomiques. Dans le cas de l'étude du génome humain, un million de SNP est suffisant pour assurer une bonne couverture de la variabilité génétique. Ceci explique en partie pourquoi les études pangénomiques nécessitent le traitement de volumes de données considérables et représentent un défi de taille pour les généticiens et bioinformaticiens.

Les approches génétiques classiques reposent sur l'étude de la liaison génétique à l'intérieur de familles, appelées aussi pedigrees. Par exemple, lorsqu'un SNP et la mutation causale de maladie sont proches sur le chromosome, alors la liaison génétique assurera une forte dépendance entre eux par leur co-transmission pendant un certain nombre de générations. En fait, plus deux marqueurs sont proches sur le chromosome, moins il est probable qu'une recombinaison survienne entre eux, et ainsi, plus il est probable qu'ils soient transmis simultanément à la descendance. Les deux loci seront alors associés statistiquement dans la descendance. Malheureusement, ce type d'approche soulève de nombreux problèmes pour la localisation des mutations responsables des maladies multifactorielles (voir section 1.1, chapitre 1, page 2, 3ème paragraphe).

Les approches plus récentes sont basées sur le LD existant entre loci très proches et apparaissent comme une solution pour l'étude des maladies multifactorielles. En

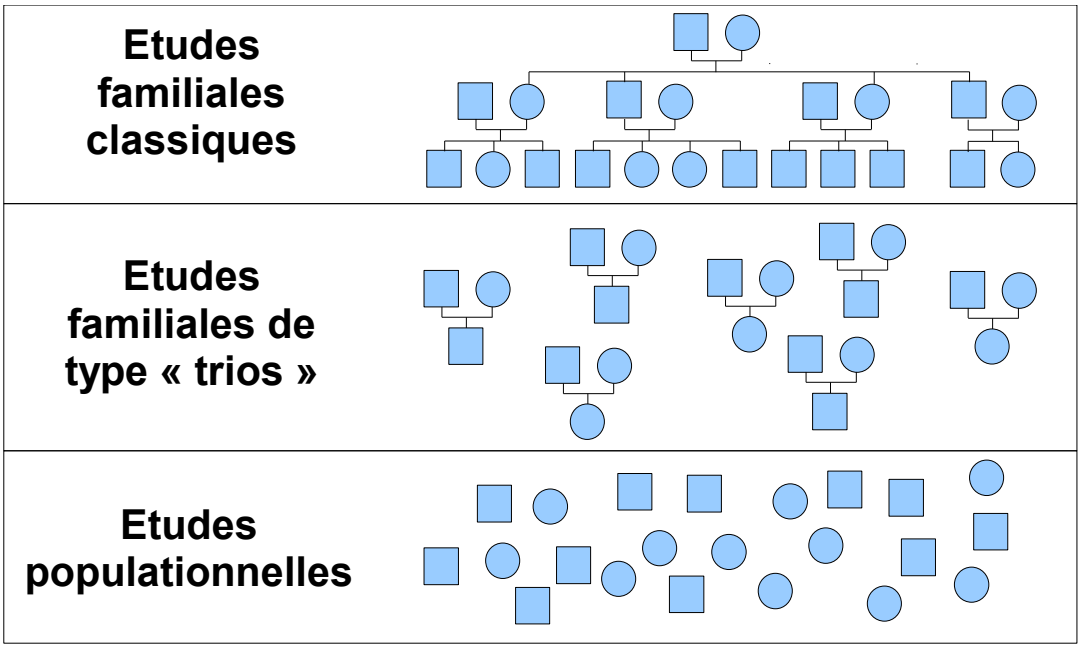

FIGURE 2.9: Les différents dispositifs expérimentaux permettant la cartographie des loci de susceptibilité.
Un carré représente un individu de sexe masculin tandis qu'un cercle représente un individu de sexe féminin.

effet, si la liaison génétique s'exerce généralement sur des distances de quelques mégabases (Mb), le LD, quant à lui, s'exerce entre marqueurs beaucoup plus proches, en moyenne 10 kb [97].

2.4.2 Les dispositifs expérimentaux

Dans le domaine de la cartographie des loci de susceptibilité prédisposants aux maladies multifactorielles, trois grandes classes de dispositifs expérimentaux ont vu le jour [79] : les études familiales classiques, les études familiales de type "trios", et les études populationnelles (figure 2.9).

Les études familiales classiques étudient des familles dont certains individus sont atteints de la maladie. Les études familiales de type "trios", quant à elles, ne se focalisent pas sur une seule famille mais plutôt sur un ensemble dont seulement l'individu malade, son père et sa mère sont étudiés. Les études populationnelles analysent une population d'individus non apparentés composée d'individus malades (cas) et d'individu sains (témoins). Chacun de ses dispositifs présente des avantages et des inconvénients. Premièrement, les familles sont difficiles à recruter mais leur étude possède l'avantage de se focaliser sur une seule forme de la maladie et n'est donc pas affectée par les problèmes d'hétérogénéité génétique. Deuxièmement, le recrutement des trios est plus simple, mais l'approche manque de puissance (*i.e.* de capacité à détecter

les facteurs génétiques). Troisièmement, le recrutement des populations est rapide et simple, et l'approche populationnelle permet d'étudier les interactions gène-gène et gène-environnement. Cependant, cette dernière approche est plus sujette à la présence d'artefacts dans les résultats, par exemple, du fait de la stratification des populations. De plus, cette approche est rendue compliquée par l'hétérogénéité génétique de la maladie.

2.4.3 Les études d'association pangénomiques

Le travail réalisé dans le cadre de la présente thèse s'inscrit dans les études populationnelles, et plus particulièrement celles qui portent sur la cartographie du génome complet (*i.e.* les méthodes pangénomiques). Lors de ces études, un ensemble d'individus sains et d'individus atteints par la maladie est génotypé. L'idée sur laquelle repose ces études est très simple. Elle part du principe que la découverte des facteurs génétiques causaux peut être réalisée en comparant les fréquences alléliques des SNP entre les deux populations d'individus. Lorsque l'allèle d'un SNP est (légèrement) plus fréquent chez les patients atteints, alors il pourra être conclu que cet allèle (ou une mutation non observée très proche) exerce une influence sur l'étiologie de la maladie. Contrairement aux études gènes-candidats qui ciblent un ensemble de gènes potentiels [93], les études d'association pangénomiques investiguent une grande partie du génome sans aucun *a priori* sur l'identité ou la localisation des loci impliqués. Cette approche représente une stratégie impartiale, non dirigée et exploratoire. Elle présente l'avantage de permettre la découverte de nouvelles connaissances sur la position des facteurs causaux.

Malgré leur simplicité de mise œuvre, en comparaison des autres approches, les études d'association pangénomiques (EAP) présentent un certain nombre de difficultés, notamment au niveau statistique. L'analyse de ces données génétiques de grande dimension est complexe [5, 51, 58, 62]. La recherche simple d'association entre chaque marqueur et la variable indicatrice sain/malade se révèle difficile car elle entraîne la découverte d'un très grand nombre de faux positifs (SNP statistiquement associés à la maladie mais en réalité non impliqués dans son étiologie). Une solution simple consiste à réaliser un seuillage des associations afin de ne récupérer que les plus significatives, mais cela a pour effet de diminuer sensiblement la puissance statistique. Par ailleurs, les combinaisons de marqueurs génétiques et de certains facteurs environnementaux sont susceptibles de jouer un rôle important dans l'apparition de la maladie, ce qui engendre une explosion combinatoire dans les analyses à réaliser. Enfin, il faut souligner la difficulté de manipuler et de traiter des données aussi volumineuses.

2.5 Conclusion

Dans ce chapitre d'introduction au déséquilibre de liaison (LD) dans le contexte des études d'association pangénomiques (EAP), nous avons souligné principalement

la complexité de la structure du LD et le rôle majeur de ce dernier dans les EAP. C'est la raison pour laquelle, la modélisation fine du LD devrait garantir une analyse plus aisée lors des EAP, par exemple en réduisant la forte dimension des données. Des outils ont déjà été développés dans cet objectif, dont certains proviennent du monde de l'intelligence artificielle et de l'apprentissage automatique, comme les modèles graphiques probabilistes [30, 44, 68] qui seront développés plus en détail dans la suite du document. Dans cette problématique, une difficulté importante à surmonter demeure la mise en œuvre des méthodes d'apprentissage des modèles graphiques probabilistes dans le contexte de données de très grande dimension, comme celles issues des EAP.

3

Les modèles graphiques probabilistes

Sommaire

3.1 Introduction

L'objectif de ce chapitre est d'introduire les principales notions dans le domaine des modèles graphiques probabilistes (MGP). A titre de rappel, un ensemble de définitions simples est d'abord fourni. D'une part, les définitions concernent les concepts d'indépendance marginale et conditionnelle en théorie des probabilités, et les concepts d'entropie et d'information mutuelle en théorie de l'information. D'autre part, certaines définitions relatives à la théorie des graphes sont rappelées. Ensuite, deux grandes familles de MGP sont présentées : les réseaux bayésiens (RB) et les réseaux de Markov (RM). Seule la première famille, les réseaux bayésiens, est abordée de manière approfondie. Un exemple simple et pratique de mise en œuvre des RB est ensuite employé afin de donner l'intuition de ces modèles au lecteur non familier. Ensuite, les trois grandes problématiques associées à l'utilisation des RB sont présentées : l'inférence probabiliste, l'apprentissage de paramètres et enfin l'apprentissage de structure. L'accent est porté sur l'apprentissage de paramètres dans le cadre de données incomplètes (*e.g.* présence de variables latentes) et sur l'apprentissage de structure. Les deux types d'apprentissage seront essentiels pour mieux comprendre l'algorithme d'apprentissage de modèles proposé dans ce travail de recherche. En conclusion, les motivations de l'application des modèles graphiques probabilistes au traitement des données issues des études d'association pangénomiques sont exposées et discutées.

3.2 Préceptes

3.2.1 Théorie des probabilités

Considérons $\mathbf{X} = \{X_1, ..., X_n\}$, un ensemble de n variables aléatoires discrètes. Dans les MGP, deux propriétés sont centrales : l'indépendance marginale et l'indépendance conditionnelle.

Définition 1

L'indépendance marginale entre deux variables X_i et X_j, notée $X_i \perp X_j$, est définie par rapport à la distribution de probabilité jointe $P(X_i, X_j)$ de la façon suivante :

$$P(X_i, X_j) = P(X_i)\ P(X_j).$$

Une non-égalité implique que X_i et X_j sont marginalement dépendantes.

Définition 2

L'indépendance conditionnelle entre deux variables X_i et X_j sachant un sous-ensemble de variables $\boldsymbol{S} \subseteq \boldsymbol{X} \backslash \{X_i, X_j\}$, notée $X_i \perp X_j | \boldsymbol{S}$, est :

$$P(X_i, X_j | \boldsymbol{S}) = P(X_i | \boldsymbol{S})\ P(X_j | \boldsymbol{S}).$$

Une non-égalité implique que X_i et X_j sont conditionnellement dépendantes sachant $\boldsymbol{S}$.

3.2.2 Théorie de l'information

Définition 3

L'entropie d'une variable X_i est définie de la manière suivante :

$$\mathcal{H}(X_i) = -\sum_{x_i} p_{x_i} \; log \; p_{x_i},$$

où p_{x_i} est la probabilité de chaque état x_i de X_i, et où la somme inclut tous les états possibles. De la même manière, on peut calculer l'entropie d'un ensemble de variables en sommant sur tous les états possibles de leur loi jointe. L'entropie est une mesure de l'incertitude associée à une variable ou plusieurs variables.

Définition 4

L'entropie conditionnelle d'une variable X_i sachant une variable X_j est définie de la manière suivante :

$$\mathcal{H}(X_i|X_j) = \mathcal{H}(X_i, X_j) - \mathcal{H}(X_j),$$

où $\mathcal{H}(X_i, X_j)$ est l'entropie jointe de X_i et X_j. L'entropie conditionnelle mesure l'entropie restante provenant de la variable X_i, si l'on connait parfaitement la variable X_j.

Définition 5

L'information mutuelle entre deux variables X_i et X_j est :

$$\mathcal{I}(X_i; X_j) = \mathcal{H}(X_i) - \mathcal{H}(X_i|X_j) = \mathcal{H}(X_j) - \mathcal{H}(X_j|X_i).$$

L'information mutuelle entre deux variables mesure le degré de dépendance entre elles. Si $\mathcal{I}(X_i; X_j) = 0$, alors les deux variables sont indépendantes.

3.2.3 Théorie des graphes

L'idée sous-jacente à la théorie des graphes est de proposer un outil pour traiter des problèmes mettant en œuvre des ensembles sur lesquels sont définies des relations binaires, quelle que soit la nature de ces relations. De nombreuses notions de graphes sont nécessaires à la compréhension des modèles graphiques probabilistes, et c'est la raison pour laquelle nous allons les présenter. Dans la suite, nous ne considérerons que les graphes finis (*i.e.* composés d'un ensemble fini de nœuds et d'arêtes).

Définition 6

Soit $V = \{v_1, ..., v_n\}$ un ensemble fini non vide. Un graphe G sur V (sans boucles) est défini par la donnée du couple $G = \{V, E\}$ où $E \subset \{(u, v) \parallel u, v \in V$,et $u \neq v\}$. V est alors nommé l'ensemble des nœuds de G.

Définition 7

Soit un graphe $G = \{V, E\}$. Pour tout élément $(u, v) \in E$,

- *(u, v) est une arête (notée $(u - v)$) si et seulement si $(v, u) \in E$,*

– (u, v) *est un arc (noté* $(u \to v)$*) si et seulement si* $(v, u) \notin E$.

Définition 8

Un graphe $G = \{V, E\}$ *est un graphe orienté (noté* $\overrightarrow{G}$*) si et seulement si tous les éléments de* E *sont des arcs. De la même façon, un graphe* $G = \{V, E\}$ *est un graphe non orienté (noté* $\overline{G}$*) si et seulement si tous les éléments de* E *sont des arêtes. Un graphe est mixte s'il ne possède aucune de ces deux propriétés.*

Définition 9

Dans un graphe orienté $\overrightarrow{G} = \{V, E\}$*, un chemin est une séquence d'arcs* $(e_i)_{i \in \{1,...,p\}}$ *vérifiant la propriété suivante : l'origine de tout arc* e_i *est l'extrémité de l'arc* e_{i-1} *précédent, dans la séquence. Un circuit est un chemin dont l'extrémité du dernier arc est l'origine du premier.*

Définition 10

Dans un graphe quelconque $G = \{V, E\}$*, une chaîne* $(e_i)_{i \in \{1,...,p\}}$ *est une séquence d'arêtes vérifiant la propriété suivante : pour tout* $i \in 2, ..., p-1$*, l'une des extrémités d'une arête* e_i *est une extrémité de l'arête* e_{i-1} *précédente ; l'autre extrémité de* e_i *est une extrémité de l'arête suivante* e_{i+1}*. Un cycle est une chaîne dont une extrémité du dernier arc est une extrémité du premier.*

Définition 11

Soit un graphe $G = \{V, E\}$*,* $\forall W \subset V$*,* $\forall F \subset E$ *:*

– $(W, E \cap W \times W)$ *est un sous-graphe de* G*,*
– (V, F) *est un graphe partiel de* G*.*

Définition 12

Soit un graphe $G = \{V, E\}$ *:*

– *connexité :* G *est connexe si et seulement si pour tout* $u, v \in V, u \neq v$*, il existe une chaîne entre* u *et* v*.*
– *connexité forte :* G *est fortement connexe si et seulement si pour tout* $u, v \in V, u \neq v$*, il existe un chemin entre* u *et* v*.*
– *graphe complet :* G *est complet si et seulement si* $\forall u, v \in V, u \neq v, (u - v) \in E$*.*

Définition 13

Les composantes connexes d'un graphe sont les sous-graphes connexes maximaux (c'est-à-dire de cardinal maximal en termes de nombre d'arêtes). De même, les cliques d'un graphe sont les sous-graphes complets maximaux.

Définition 14

Un graphe $G = \{V, E\}$ *est un arbre si et seulement s'il est connexe et sans cycle. Un graphe* G *est une arborescence si et seulement si* G *est un arbre et possède une unique racine, à partir de laquelle il existe un chemin unique vers tous les autres nœuds. Une forêt est un graphe dont toutes les composantes connexes sont des arbres.*

Définition 15

Dans un graphe orienté, un nœud v *est parent d'un nœud* u *si et seulement s'il existe un arc* $(v \to u)$*. Le parent est alors noté* Pa_u*. Le nœud* u *est alors appelé enfant du nœud* v*. Dans un graphe non orienté, un nœud* v *est voisin d'un nœud* u *si et seulement s'il existe une arête* $(v - u)$*.*

Définition 16

Un nœud z est ancêtre d'un nœud u si et seulement s'il existe un chemin de z vers u. Le nœud u est alors appelé descendant du nœud z.

3.3 Introduction aux modèles graphiques probabilistes

Notre étude se place dans le cadre d'un raisonnement statistique sur des données multivariées, impliquant un ensemble de variables aléatoires discrètes. Les SNP sont des variables quantitatives discrètes (0, 1 ou 2), et peuvent ainsi être considérés comme variables qualitatives ou quantitatives. Le phénotype est de nature purement qualitative puisqu'il indique si l'individu est atteint ou non de la maladie génétique multifactorielle.

Les MGP sont issus à la fois de la théorie des probabilités et de la théorie des graphes [67]. D'une part, le raisonnement probabiliste fournit le cadre incertain avec la notion de variables aléatoires, de dépendances et d'indépendances probabilistes entre les variables aléatoires. D'autre part, les graphes assurent la modélisation de la structure des relations de dépendance entre les variables. Dans un MGP, chaque nœud du graphe correspond à une variable aléatoire et chaque arête (ou arc) représente une relation de dépendance entre les variables. Le graphe encode la façon dont la distribution de probabilité jointe peut être décomposée en un ensemble de distributions de probabilité de dimensions respectives plus petites. Ainsi, les MGP offrent un cadre de représentation compacte de la distribution de probabilité jointe en réduisant le nombre de paramètres libres requis (*i.e.* degrés de liberté), assurant ainsi une meilleure robustesse [1].

Différentes sous-classes de MGP existent, dont les plus courantes sont les réseaux de Markov (RM) et les réseaux bayésiens (RB). Les RM sont des graphes non orientés, tandis que les réseaux bayésiens sont des graphes orientés sans circuit (GOSC). Contrairement aux RM, les RB présentent la particularité de pouvoir représenter des relations de causalité (*i.e.* cause-effet) entre les variables [53]. Les MGP généralisent un grand nombre de modèles, ce qui en fait un cadre d'étude très puissant. Par exemple, les modèles de Markov et les modèles de Markov cachés, deux outils indispensables en analyse de séquences, peuvent être considérés comme des cas particuliers de RB.

Formellement, les modèles graphiques probabilistes sont des graphes $G(\mathbf{X}, \mathbf{E})$ à l'intérieur desquels l'ensemble de nœuds $\mathbf{X} = \{X_1, ..., X_n\}$ représente n variables aléatoires, et l'ensemble des arêtes $\mathbf{E}$ capture les dépendances entre ces variables (*i.e.* la structure). Les variables peuvent être soit observées (*i.e.* mesurées), soit latentes (*i.e.* non mesurées).

1. La robustesse est la capacité des paramètres estimés à ne pas être modifiés par une petite modification dans les données

Dans ce chapitre introductif aux MGP, nous nous focaliserons sur les RB. En effet, ces derniers ont fait l'objet des travaux de recherche présentés dans ce mémoire de doctorat. Les RM seront quant à eux présentés de manière beaucoup plus succincte, dans le but de mieux appréhender l'état de l'art qui couvrira à la fois les RB et les RM.

3.4 Réseaux bayésiens

3.4.1 Introduction

Afin d'introduire les réseaux bayésiens, nous allons prendre l'exemple suivant (à titre purement didactique). Imaginons que nous étudiions le mécanisme génétique à l'origine de la couleur de peau, notée C. Le mécanisme héréditaire est considéré comme monogénique, c'est-à-dire qu'un seul gène influence la couleur de peau. Nous avons identifié le gène, noté G. Considérons que l'allèle (*i.e.* la version du gène, ou modalité en langage statistique) à l'origine de la couleur foncée soit dominant. Ainsi sa présence en une seule copie suffit pour qu'un sujet ait la peau foncée. Supposons que le phénomène ait été étudié sur un certain nombre d'individus. Nous avons ainsi pu calculer les probabilités pour le gène G de porter l'allèle donnant la couleur foncée ou de ne pas le porter, et les probabilités qu'a un individu d'avoir la peau foncée ou non sachant la nature de son gène. Nous pouvons modéliser cet état de connaissances à l'aide d'un RB très simple (voir figure 3.1).

Sur la figure 3.1, un arc relie le gène G à la couleur C de la peau. Cet arc représente le lien probabiliste existant entre les deux variables : le gène G et la couleur C de la peau sont dépendants. Un tableau de probabilités est associé à chacune des variables. Par exemple, *a priori*, nous savons que 40% des individus possèdent l'allèle de la couleur foncée dans la population étudiée. Ensuite, 90% des individus possédant l'allèle de la couleur foncée présentent une peau foncée. Par ailleurs, 90% des individus n'ayant pas l'allèle de la couleur foncée présentent une peau de couleur claire.

Le RB modélise de façon équivoque la relation de dépendance existant entre les variables : d'une part, il représente cette dépendance selon la théorie des graphes à l'aide d'un arc, et d'autre part, il la matérialise selon la théorie des probabilités avec la table de probabilités conditionnelles $P(C|G)$. Cet exemple demeure très basique au regard des possibilités offertes par les RB. Une des propriétés les plus importantes des RB est qu'ils permettent de modéliser des relations d'indépendance et de dépendance conditionnelles entre les variables. Reprenons l'exemple précédent et admettons que nous avons maintenant identifié le produit protéique P du gène G. De plus, nous savons que la couleur de la peau peut être influencée par des facteurs non génétiques comme l'usage d'une crème bronzante, noté B, et que l'action de B ne repose pas sur la protéine P. Reconstruisons le RB à partir de cet état de connaissances (voir figure 3.2).

Dans ce nouveau RB, la couleur C de la peau devient indépendante du gène G

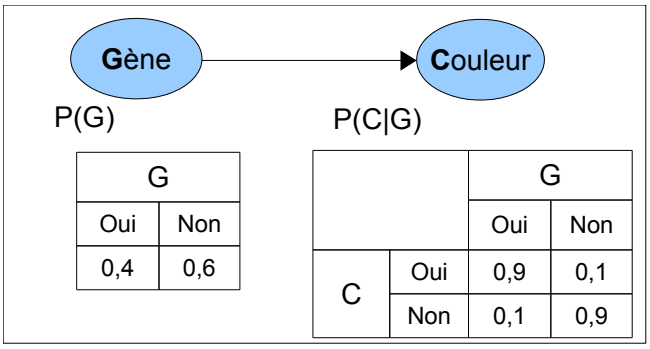

FIGURE 3.1: Réseau bayésien modélisant l'influence du gène G sur la couleur de peau C.

sachant la protéine P. En effet, la connaissance de la variable G devient inutile lorsque nous connaissons l'état de la variable P afin de prédire C. Cet exemple illustre l'indépendance conditionnelle. Une propriété plus étonnante peut s'illustrer par le fait que la variable P devient dépendante de B sachant C : c'est la dépendance conditionnelle, qui est modélisée graphiquement par une structure en V, nommée V-structure. Cette propriété est beaucoup moins intuitive mais nous pouvons la comprendre de la manière suivante. Si nous observons une personne à la peau foncée et que nous savons qu'elle a utilisé de la crème bronzante, nous allons avoir tendance à penser qu'il est plus probable que la couleur de sa peau ait été causée par l'utilisation de la crème bronzante plutôt que par le fait que naturellement sa peau ait un teint foncé, *i.e.* que la protéine P soit présente dans sa peau. Ainsi nous voyons que les deux variables, usage de crème bronzante B et présence de la protéine P, deviennent dépendantes lorsque nous connaissons la couleur de la peau de la personne, alors qu'*a priori*, nous savons que les deux variables ne sont pas liées biologiquement.

La factorisation de la loi jointe joue un rôle essentiel dans les RB (et dans les MGP en général). Elle permet de décomposer d'une manière simple la distribution de probabilité jointe en un produit de distributions de probabilité conditionnelle de chaque nœud X_i conditionnellement à ses parents Pa_{X_i} dans le graphe :

$$P(\mathbf{X}) = \Pi_{i=1}^{n} P(X_i|Pa_{X_i}). \tag{3.1}$$

Dans l'exemple précédent, la distribution de probabilité jointe (DPJ) se décompose de la manière suivante :

$$P(G, P, C, B) = P(G)\ P(P|G)\ P(B)\ P(C|P, B).$$

La connaissance de la DPJ a de nombreuses applications. Elle permet notamment de réaliser des opérations d'inférence probabiliste, de simuler des données et d'évaluer la qualité d'un modèle grâce au calcul de la vraisemblance des données (la vraisemblance sera présentée en section 3.4.4.1, page 34). Par exemple, l'inférence probabiliste permet de calculer la probabilité d'avoir mis de la crème bronzante sachant certaines informations dans le RB. *A priori*, la probabilité d'avoir mis de la crème bronzante

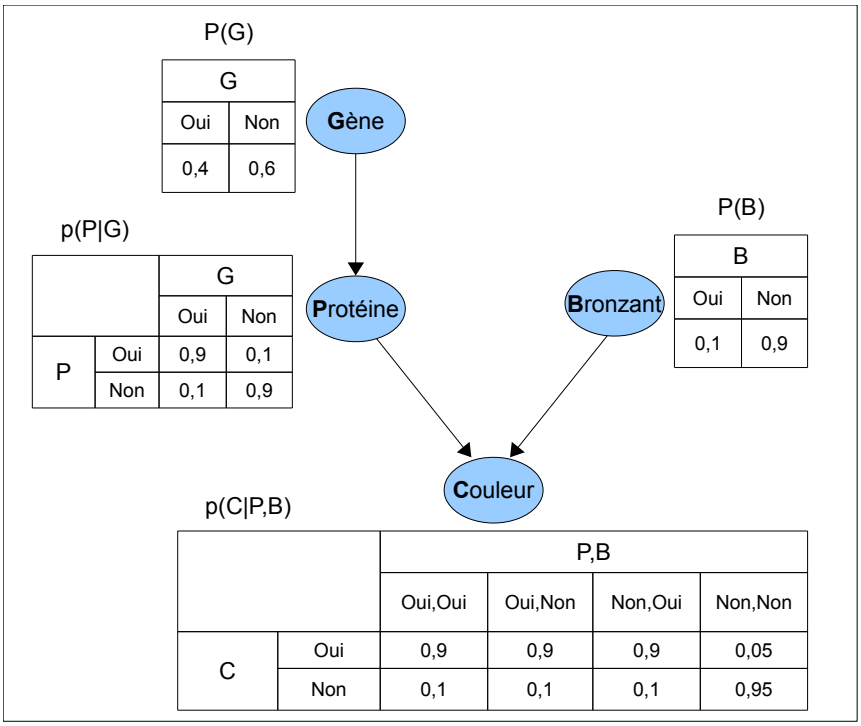

FIGURE 3.2: Réseau bayésien modélisant l'influence du gène G et de la crème bronzante B sur la couleur de peau C.
P indique la présence de la protéine issue de l'expression du gène G.

est $P(B = oui) = 0,1$. Si l'on considère le cas où la personne a la peau foncée, alors la probabilité devient $P(B = oui|C = oui) = 0,197$. Cette probabilité est environ dix fois moins importante si la personne a la peau claire : $P(B = oui|C = non) = 0,018$.

Pour notre problématique de modélisation du déséquilibre de liaison, la DPJ est inconnue et représente ainsi l'objectif à atteindre. Le moyen, pour y parvenir, consiste à utiliser des méthodes d'apprentissage automatique à partir de données. Pour cela, deux étapes majeures sont à distinguer :

- l'apprentissage de la structure, *i.e.* des dépendances marginales et conditionnelles encodées par le graphe du RB,
- l'apprentissage des paramètres, *i.e.* des distributions de probabilité conditionnelle ou marginale associées à chaque nœud. Dans la plupart des situations, l'apprentissage de structure est une étape préliminaire à l'apprentissage de paramètres, et se révèle la plus difficile.

Par souci didactique, dans la suite du mémoire, nous présenterons dans l'ordre : la définition des RB, l'inférence probabiliste, l'apprentissage de paramètres et enfin l'apprentissage de structure. Nous avons délibérément choisi de ne pas présenter certains aspects des RB qui ne seront pas employés dans la suite du mémoire.

3.4.2 Définition

Définition 17

Un réseau bayésien est défini par un graphe orienté sans circuit (GOSC), $G(\boldsymbol{X}, \boldsymbol{E})$, et un ensemble de paramètres θ. L'ensemble des nœuds $\boldsymbol{X} = \{X_1, ..., X_n\}$ représente n variables aléatoires et l'ensemble des arcs $\boldsymbol{E}$ capture les dépendances entre ces variables. L'ensemble θ réunit les distributions de probabilité conditionnelle $\theta_i = [\mathbb{P}(X_i|Pa_{X_i})]$ où Pa_{X_i} sont les parents du nœud X_i. Si un nœud ne possède pas de parents, alors il est décrit par une distribution de probabilité a priori.

3.4.3 Inférence probabiliste

D'un point de vue intuitif, l'inférence probabiliste (IP) dans un RB consiste à propager une ou plusieurs informations certaines au sein de ce réseau, pour en déduire ensuite la façon dont sont modifiées les croyances (plus ou moins certaines) concernant les autres nœuds [67]. L'IP nécessite la connaissance de la structure (*i.e.* du graphe) et des paramètres du RB (*i.e.* des tables de probabilités associées aux nœuds). D'un point de vue probabiliste, l'IP consiste à estimer la probabilité d'un évènement quelconque sachant certaines connaissances dans le RB. L'IP présente l'avantage de prendre en compte l'incertitude présente dans les connaissances. Pour illustrer l'IP, reprenons l'exemple du RB de la figure 3.2. Grâce à l'IP, nous pouvons déterminer que la probabilité *a priori* d'avoir la peau bronzée, notée $P(C = oui)$, est égale à 0.456. Maintenant, si l'on sait que la personne a utilisé de la crème bronzante, alors la probabilité *a posteriori* d'avoir la peau bronzée, notée $P(C = oui|B = oui)$, est égale à 0.9.

L'IP repose sur le calcul de la probabilité *a posteriori* de l'évènement A sachant l'observation e (ou évidence), selon le théorème de Bayes :

$$p(A|e) = \frac{p(e|A)\ p(A)}{p(e)}.$$

Dans la pratique, l'IP est beaucoup plus complexe. L'IP a été prouvée NP-difficile (non polynomial difficile) dans le cas général [18].

Diverses méthodes ont été proposées pour la mise en œuvre de l'IP dans les RB. Parmi les méthodes exactes, deux classes principales ont vu le jour : (i) les méthodes de propagation de messages étendus par des algorithmes de coupe (ou de conditionnement) [74], et (ii) les méthodes utilisant des regroupements de nœuds [39, 40, 48]. Les premières réalisent une propagation des messages le long des arcs du GOSC, tandis que les secondes opèrent au préalable une transformation du GOSC en arbre de jonction pour ensuite appliquer une version simplifiée de la propagation de messages. Pour une introduction plus détaillée à ces méthodes, le lecteur peut se référer à Réf.[67].

3.4.4 Apprentissage de paramètres

L'apprentissage de paramètres (AP) est une étape clé lors de la construction d'un RB. Étant donné une structure, l'AP consiste à estimer de la manière la plus précise les paramètres du modèle à partir de données. L'AP est un sujet vaste et complexe, c'est pourquoi nous nous restreindrons aux méthodes les plus utilisées dans le cadre des RB.

3.4.4.1 Données complètes

Dans le cas le plus simple, l'AP est réalisé sur des données complètes, *i.e.* pour lesquelles aucune donnée n'est manquante ou bien pour lesquelles toutes les variables du RB correspondant sont observées (aucune variable non observée). Si l'on ne considère pas de connaissance *a priori* sur les paramètres (approche fréquentiste), alors il est possible d'estimer les paramètres de manière très simple à l'aide de la méthode du maximum de vraisemblance (MV). La méthode du MV est simple : elle cherche à maximiser la probabilité que les données observées aient été générées par un modèle dont la structure est fixée. On considère un ensemble de données d'apprentissage $D = \{\mathbf{x}^1, ..., \mathbf{x}^m\}$ comprenant m observations pour les n variables de l'ensemble $\mathbf{X} = \{X_1, ..., X_n\}$. Une observation est une réalisation de ces n variables aléatoires : pour $i \in 1, ..., m$, $\mathbf{x}^i = \{x_1^i, ..., x_n^i\}$. On appelle vraisemblance du modèle, notée $L(\theta)$, au vu des observations D d'un échantillon de taille m indépendamment et identiquement distribué (i.i.d.) selon la DPJ de paramètres θ, le terme :

$$L(\theta) = P(D|\theta) = L(\mathbf{x}^1, ..., \mathbf{x}^m|\theta) = P(\mathbf{x}^1|\theta) \times ... \times P(\mathbf{x}^m|\theta) = \Pi_{i=1}^m P(\mathbf{x}^i|\theta).$$

La méthode du maximum de vraisemblance consiste alors à identifier les paramètres θ^{MV} maximisant $L(\theta)$, appelé maximum de vraisemblance. Dans la pratique, le maximum de la log-vraisemblance, $log\ L(\theta)$, est utilisé car il permet de simplifier les calculs et revient à identifier les mêmes paramètres. Dans le cas des données discrètes qui nous préoccupe, l'estimation par la méthode du maximum de vraisemblance est triviale et revient à calculer la fréquence d'apparition d'un évènement :

$$\hat{P}(X_i = x_k|Pa_{X_i} = x_j) = \hat{\theta}_{i,j,k}^{MV} = \frac{N_{i,j,k}}{\sum_k N_{i,j,k}}$$

où $N_{i,j,k}$ est le nombre d'évènements pour lesquels la variable X_i est dans l'état x_k et ses parents Pa_{X_i} sont dans la configuration x_j.

Dans de nombreuses situations, il peut être plus intéressant de réaliser un apprentissage bayésien. Contrairement à l'approche fréquentiste (*e.g.* maximum de vraisemblance), l'approche bayésienne prend en compte un certain nombre d'*a priori* sur les paramètres. Par exemple, l'apprentissage bayésien [10] permet d'apporter des connaissances *a priori* sur le phénomène étudié, indépendamment des données d'apprentissage. L'approche bayésienne consiste à calculer la distribution *a posteriori* des

paramètres $P(\theta|D)$ [76]. Les paramètres $\hat{\theta}^{MAP}$ maximisant la probabilité *a posteriori* (MAP) peuvent alors être employés :

$$\hat{\theta}^{MAP} = argmax_\theta \ log \ P(\theta|D).$$

La distribution *a posteriori* $P(\theta|D)$ est calculée à l'aide de la formule de Bayes incorporant des aprioris sur les paramètres :

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

où $P(\theta|D)$ s'exprime comme le produit de la vraisemblance marginalisée $P(D|\theta)$ (où θ est ici une variable aléatoire) et d'une probabilité *a priori* sur les paramètres $P(\theta)$, divisé par une constante de normalisation ($P(D)$ est constant quel que soit le modèle traité et n'est pas pris en compte lors du calcul). Une approche alternative au maximum *a posteriori* consiste à calculer l'espérance *a posteriori* des paramètres à inférer.

3.4.4.2 Données incomplètes

Dans le cas où certaines données sont manquantes, ou bien des variables non observées (appelées aussi variables latentes) sont présentes dans le modèle, l'AP est bien plus complexe et long car il nécessite dans la plupart des situations un apprentissage basé sur des méthodes itératives. L'algorithme le plus utilisé dans ce cas est l'algorithme expectation-maximization (EM) [22], car il est simple à mettre en œuvre et efficace dans la majorité des situations. Il peut être employé à la fois dans le cadre fréquentiste et dans le cadre bayésien. Nous le présenterons en détail sous la perspective fréquentiste.

Soit $log \ P(D|\theta) = log \ P(D_o, D_m|\theta)$ la log-vraisemblance des données complètes. D_o et D_m correspondent aux données observées et aux données manquantes, respectivement. En considérant un modèle de référence θ^*, il est possible d'estimer la distribution de probabilité des données manquantes $P(D_m|\theta^*)$, et ainsi de calculer $Q(\theta;\theta^*)$ l'espérance de la log-vraisemblance des données complètes :

$$Q(\theta;\theta^*) = E_{\theta^*}[log \ P(D_o, D_m|\theta)].$$

$Q(\theta;\theta^*)$ est l'espérance de la vraisemblance d'un ensemble de paramètres θ quelconque, calculée en employant la distribution de données manquantes $P(D_m|\theta^*)$. Dans le cadre des réseaux bayésiens, l'équation précédente peut se réécrire de la façon suivante :

$$Q(\theta;\theta^*) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N^*_{ijk} \ log \ \theta_{ijk} \tag{3.2}$$

où $N^*_{ijk} = E_{\theta^*}[N_{ijk}] = N \times P(X_i = x_k, Pa_{X_i} = x_j|\theta^*)$ est obtenu par inférence dans le réseau de paramètres θ^* si les valeurs de $\{X_i, Pa_{X_i}\}$ ne sont pas complètement

mesurés, et par simple comptage sinon. N est le nombre total d'observations.

L'algorithme EM est très simple et consiste à itérer deux étapes. Soient $\theta^t = \{\theta^t_{ijk}\}$ les paramètres du modèle à l'itération t. A l'étape $t+1$, deux calculs sont mis en œuvre :

- calcul de l'espérance : estimation des N^*_{ijk} de l'équation 3.2 à partir des paramètres de référence θ^t,
- maximisation : identification des paramètres θ^{t+1} qui maximisent Q :

$$\theta^{t+1}_{ijk} = \frac{N^*_{ijk}}{\sum_k N^*_{ijk}}.$$

Ces deux étapes sont répétées tant qu'il est possible d'augmenter la valeur de Q.

L'algorithme converge généralement vers un maximum local. C'est pourquoi de nombreuses méthodes ont été développées pour pallier ce problème. La plus simple consiste à relancer l'algorithme en tirant aléatoirement de nouvelles valeurs pour les paramètres de départ (réinitialisation aléatoire). De nombreuses variantes de l'algorithme EM ont été proposées, notamment pour simplifier le problème de l'étape de maximisation (EM généralisé) [22]. L'algorithme EM peut aussi s'appliquer au cadre bayésien. Il suffit pour cela de remplacer le maximum de vraisemblance de l'étape de maximisation par le maximum *a posteriori* ou l'espérance *a posteriori*.

3.4.5 Apprentissage de structure

Dans un grand nombre de situations, la structure du RB est inconnue et il est nécessaire de la déterminer avant de réaliser l'apprentissage des paramètres. C'est le cas, par exemple, en génétique d'association, où les relations entre variables (SNP et phénotype) ne sont pas (ou sont seulement partiellement) connues *a priori*. Malheureusement, l'apprentissage de la structure (AS) est un problème NP-difficile [16] et demeure l'étape la plus complexe lors de l'apprentissage d'un RB. C'est pourquoi la plupart des algorithmes développés sont des heuristiques conçues dans le but de surmonter la difficulté due à la très grande dimension de l'espace de recherche des structures. Le nombre r de structures possibles est superexponentiel en fonction du nombre de variables n :

$$r(n) = \begin{cases} 1 & , n = 0 \text{ ou } 1 \\ \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-1)} \, r(n-i) & , n > 1. \end{cases}$$

Afin de donner un ordre de grandeur, pour 5 et 10 variables, les valeurs de r sont égales à 29281 et à 4.2×10^{18}, respectivement.

La complexité de l'AS a motivé le développement de deux grandes familles d'algorithmes [34] : les méthodes par contraintes (MC) et les méthodes basées sur un score (MS). Les MC mettent en œuvre une recherche d'indépendances conditionnelles entre

les variables, à l'aide de tests statistiques tels que le χ^2 conditionnel et le coefficient de corrélation partielle de Pearson. A titre d'exemple, nous pouvons citer l'algorithme PC [88] qui part du graphe complet non orienté (toutes les arêtes sont présentes), puis les retire au fur et à mesure que des indépendances sont découvertes. Ensuite, pour orienter le graphe, l'identification des dépendances conditionnelles permet de repérer les V-structures. A l'issue de ces deux étapes, le graphe obtenu est un graphe partiellement orienté qu'il faut finir d'orienter tout en ne rajoutant pas de nouvelles V-structures.

Les MS cherchent à optimiser un score évaluant la pertinence d'une structure par rapport aux données observées, lors d'une recherche locale à travers l'espace des GOSC. Les scores, tels que le critère d'information d'Akaike (Akaike information criterion, AIC) [2] et le critère d'information bayésien (Bayesian information criterion, BIC) [83], peuvent être décomposés en deux termes : le premier évalue l'adéquation du modèle aux données (maximum de vraisemblance), tandis que le second pénalise le modèle en considérant sa dimension de façon à empêcher tout surapprentissage [2]. La formule du score BIC est la suivante :

$$BIC(M, D) = log\ P(D|\theta^{MV}, M) - \frac{1}{2} dim(M)\ log\ N$$

avec M le modèle, $dim(M)$ la dimension de M et N le nombre d'observations. Un des algorithmes les plus simples, la recherche gloutonne [15], consiste à partir d'une structure donnée, à explorer le voisinage de cette structure à l'aide d'opérateurs d'addition, de suppression et d'inversion d'arc dans le GOSC. L'algorithme s'arrête lorsque le score n'augmente plus et la structure courante est alors renvoyée.

D'autres classes de méthodes plus récentes ont vu le jour. Notamment, les méthodes hybrides combinent une première phase de recherche d'indépendances conditionnelles pour guider la phase suivante de recherche gloutonne dans l'espace des structures. Ces méthodes hybrides tirent parti des avantages des méthodes par contraintes, rapides, afin de restreindre ensuite la recherche locale. Ces méthodes engendrent ainsi un gain important en terme de temps de calcul. Par exemple, l'algorithme max-min hill climbing (MMHC) [106] permet de surpasser un grand nombre de méthodes lorsqu'un nombre important de variables est considéré (100 k variables). Des études de simulation évaluant et comparant les différents algorithmes d'apprentissage en termes de performance et de temps de calcul peuvent être trouvées dans Réf.[106, 112, 115].

Des méthodes ont aussi été développées lorsque des valeurs sont manquantes ou des variables latentes sont présentes. Nous verrons en détail cette situation dans le prochain chapitre, consacré aux contributions algorithmiques des travaux présentés dans le cadre de cette thèse.

2. Le surapprentissage désigne le fait qu'un modèle possédant un trop grand nombre de paramètres s'ajustera très bien aux données sur lequel il aura été appris mais aura de la peine à généraliser les caractéristiques de ces données.

3.5 Réseaux de Markov

Les réseaux de Markov (RM) sont des graphes non orientés. Dans ces modèles, la DPJ peut être factorisée en cliques du graphe G à l'aide de la formule suivante :

$$P(\mathbf{X}) = \frac{1}{Z}\Pi_{C \in cl(G)}\ \phi(C)$$

où $cl(G)$ est l'ensemble des cliques maximales de G, les fonctions ϕ sont les paramètres (souvent nommés potentiels) et Z est la constante de normalisation, généralement impossible à calculer dans la pratique.

Les RM permettent de prendre en compte certaines dépendances que les RB ne peuvent pas représenter, les dépendances dans un cycle. Par contre, contrairement aux RB, le RM ne modélisent pas les dépendances marginales (V-structure).

Certaines classes particulières de RM sont très intéressantes, tels que les modèles ayant une structure en arbre et les RM décomposables. D'une part, les modèles à structure en arbre permettent, par exemple, de réaliser l'inférence probabiliste en temps linéaire. Ils sont en outre très simples à apprendre. D'autre part, l'utilisation de RM décomposables (RMD) simplifie le calcul de la DPJ. Les RMD sont décomposables en cliques et séparateurs de cliques. Un séparateur de cliques est un ensemble de nœuds dont la suppression a pour effet de déconnecter deux cliques ou plus. Pour les RMD, la décomposition de la DPJ est la suivante :

$$P(\mathbf{X}) = \frac{\Pi_{C \in cl(G)} P(C)}{\Pi_{S \in sep(G)} P(S)} \tag{3.3}$$

où $sep(G)$ est l'ensemble des séparateurs de cliques de G, et $P(C)$ et $P(S)$ sont les DPJ des cliques et des séparateurs, respectivement. Grâce à cette décomposition simple de la DPJ, il est possible de calculer rapidement la vraisemblance d'un modèle et ainsi d'évaluer ce dernier grâce à un score, tel que le score BIC. Dans la suite du mémoire, nous nous concentrerons sur les RMD qui ont rencontré plus de succès que les RM pour la modélisation du déséquilibre de liaison (LD) et des données issues des études d'association pangénomiques (EAP).

3.6 Conclusion

Dans ce chapitre d'introduction aux modèles graphiques probabilistes, nous avons présenté deux classes très répandues de modèles, d'une part les réseaux bayésiens et d'autre part les réseaux de Markov. En outre, nous avons montré leur utilité sur un exemple simple de génétique. Pour l'apprentissage de ces modèles, deux étapes sont fondamentales : l'apprentissage de la structure, *i.e.* des dépendances entre variables, et l'apprentissage des paramètres, *i.e.* des distributions locales de probabilité. La première étape est la plus complexe dans la situation où toutes les données sont observées.

En effet, dans cette situation, les paramètres sont rapidement appris par maximum de vraisemblance (ou maximum a posteriori) qui met en œuvre un simple comptage des observations, alors que l'apprentissage de structure nécessite un parcours dans l'espace superexponentiel des graphes orientés sans circuit. Dans la situation où toutes les données ne sont pas observées, par exemple en présence de variables latentes, l'étape d'apprentissage de paramètres nécessite aussi l'emploi de longs calculs (*e.g.* algorithme EM). Malgré ces inconvénients, nous verrons dans le prochain chapitre que les modèles avec variables latentes présentent un certain nombre d'avantages pour la modélisation du LD qui a motivé les travaux de cette thèse.

4

Modélisation pangénomique du déséquilibre de liaison à l'aide de forêts de modèles hiérarchiques à classes latentes

Sommaire

4.1 Introduction

La modélisation du déséquilibre de liaison (linkage disequilibrium, LD) représente un enjeu majeur en génétique statistique car elle joue un rôle important lors de la cartographie des loci de susceptibilité à la maladie et aide à mieux appréhender l'histoire des populations humaines [75]. Par exemple, la réduction de la dimension des données [73] et l'inférence haplotypique [1] sont des applications de la modélisation du LD en cartographie des loci. Comme nous avons vu dans le chapitre 2, section 2.3.5, page 13, la structure du LD est complexe. Elle est constituée de blocs de tailles variées présentant une faible diversité haplotypique. De plus, l'existence de dépendances entre loci non contigus et le fait que les frontières de ces blocs ne soient pas précisément définies sur le génome ont pour effet d'accroître la complexité de la structure en blocs du LD. Outre l'aspect de la complexité, il faut prendre en compte le caractère incertain inhérent aux données génétiques. En effet, ces données peuvent être partiellement ou complètement non observées (phase manquante). De plus, elles peuvent être sujettes à des erreurs de mesure, telles que les erreurs de génotypage.

Les raisons précédentes justifient l'utilisation des modèles graphiques probabilistes (MGP) qui constituent un cadre d'étude approprié pour la modélisation fine du LD. Les MGP permettent par exemple d'estimer les fréquences haplotypiques avec beaucoup de précision, tout en intégrant l'incertitude des données. Cependant le passage à l'échelle des méthodes basées sur les MGP demeure une point problématique majeur à prendre en considération lors de l'apprentissage à partir de données de très grandes dimensions, telles que les données issues des études d'association pangénomiques (EAP).

Ce chapitre présente l'ensemble des travaux relatifs à la modélisation pangénomique du LD. Les enjeux et le contexte de la modélisation du LD sont d'abord introduits. Un état de l'art sur l'application des MGP à la modélisation du LD est présenté. Ensuite, le choix d'un nouveau modèle - la forêt de modèles hiérarchiques à classes latentes - est discuté et argumenté, notamment au travers de ses nombreux avantages pour la modélisation du LD, tels que la possibilité de prendre en compte la nature floue du LD et de hiérarchiser les différents degrés de LD. Un état de l'art général sur l'apprentissage d'un modèle très proche, le modèle hiérarchique à classes latentes, est présenté. Un nouvel algorithme d'apprentissage nommé CFHLC (Construction of Forests of Hierarchical Latent Class models) est proposé et détaillé. Une seconde version (CFHLC+) est ensuite présentée. Elle offre l'avantage de ne plus nécessiter le découpage du génome en fenêtres. Notre approche de modélisation est ensuite évaluée sur des données simulées et réelles. A l'aide du jeu de données de Daly *et al.* [19], cette approche est comparée à un certain nombre d'approches concurrentes basées sur une modélisation en blocs (SNP contigus) ou basées sur une modélisation en clusters (SNP non contigus). Par ailleurs, le passage à l'échelle est démontré pour des données pangénomiques. La conclusion de ce troisième chapitre porte sur les perspectives d'amélioration de la modélisation proposée et sur ses champs d'application possibles.

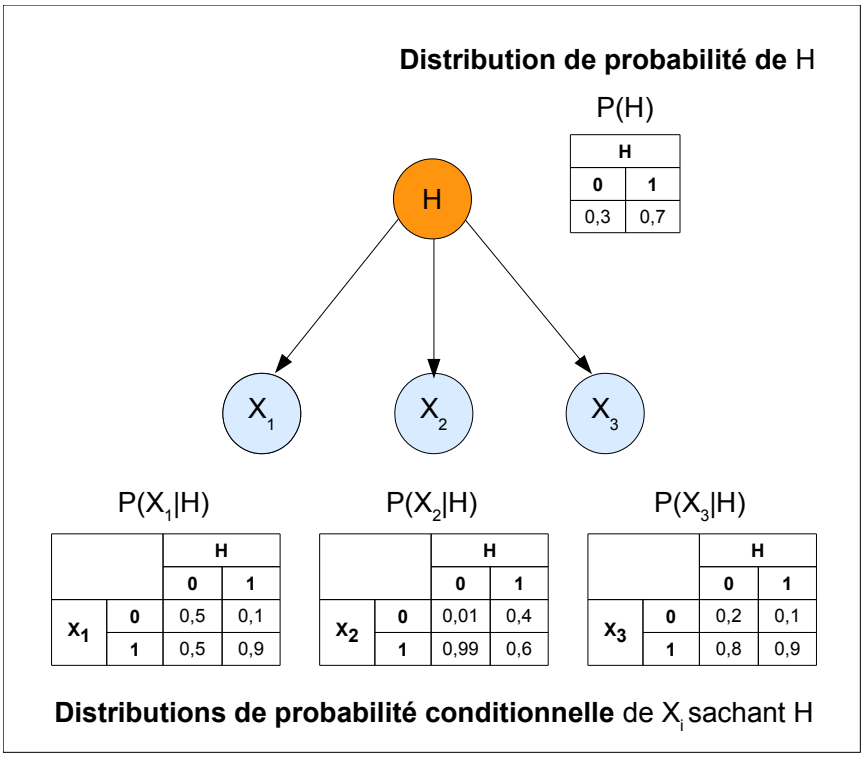

FIGURE 4.1: Modèle à classes latentes.
Les variables observées et les variables latentes sont représentées par des nœuds bleus (clairs) et rouges (foncés), respectivement.

4.2 État de l'art

Un certain nombre de méthodes basées sur les MGP ont été proposées afin de modéliser le LD. Ces méthodes peuvent être principalement classées en fonction du type de modèle employé, de leur capacité ou non à passer à l'échelle et de l'application à la génétique visée en aval de la modélisation du LD. Dans le cadre de ce travail de thèse de doctorat, un état de l'art sur l'application des MGP à la modélisation du LD a été publié dans Réf.[65].

4.2.1 Différents modèles de déséquilibre de liaison

A priori, la direction des arcs dans les MGP (comme dans les RB par exemple) paraît arbitraire dans le cas de la modélisation du LD. En effet, elle ne possède pas de signification génétique particulière, n'impliquant pas de relation causale entre SNP. En effet, il est difficile d'imaginer qu'un SNP puisse représenter un facteur causal d'un autre SNP. C'est pourquoi les RM, des modèles basés sur des graphes non orientés, apparaissent de prime abord comme des outils de choix pour modéliser le LD. Dans cette optique, Thomas et Camp [102] réalise un apprentissage de la structure basé

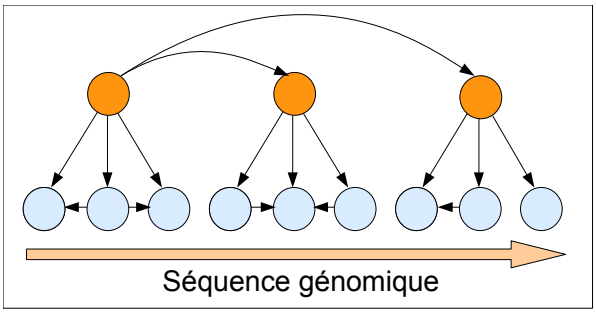

FIGURE 4.2: Ensemble de modèles à classes latentes augmenté par des dépendances entre SNP et des dépendances entre variables latentes, proposé par Nefian [68].
Les SNP et les variables latentes sont représentés par des nœuds bleus (clairs) et rouges (foncés), respectivement.

sur le score. Lors du parcours de l'espace des graphes non orientés, seuls les graphes décomposables sont considérés, ce qui simplifie le calcul du score (voir chapitre 3, section 3.5, page 38), mais présente l'inconvénient d'engendrer une perte de temps de calcul lors de la proposition de solutions non décomposables. En outre, Thomas et Camp incorporent la position chromosomique des loci comme information *a priori* dans la formule du score. Cette idée se base sur le fait que le LD décroît naturellement avec la distance physique existant entre les SNP (modèle de Malécot [54]). Malgré cet *a priori*, la méthode de Thomas et Camp réussit à identifier des dépendances complexes entre SNP non adjacents.

Les RB ont aussi été proposés pour la modélisation du LD. Deux travaux [49, 108] ont suggéré que la direction des arcs pourrait être utile pour le généticien, notamment pour la sélection d'un sous-ensemble de SNP indépendants et hautement informatifs (*i.e.* tagSNPs). Une approche standard, la recherche gloutonne, qui réalise une recherche locale basée sur un score, est implémentée dans la méthode BNTagger [49]. L'espace des GOSC est exploré à l'aide d'opérateurs d'addition, de suppression et d'inversion d'arc. Afin d'éviter les maxima locaux, la réinitialisation aléatoire ou le recuit simulé représentent des solutions simples mais efficaces. L'emploi de méthodes plus sophistiquées d'apprentissage de la structure, telles que les algorithmes génétiques combinés à une recherche locale, s'est révélé performant pour trouver rapidement une structure optimale [108].

Une famille particulière de RB, les RB avec VL, offre aussi de nombreuses possibilités. L'intérêt des VL est qu'elles capturent des dépendances complexes entre les VO. Elles peuvent aussi être utilisées à la place des VO pour la réduction de dimension des données. Parmi cette famille, le modèle à classes latentes (MCL) a été largement employé. Un MCL est défini comme un ensemble de VO X_i ayant toutes pour parent une même et unique VL H (voir figure 4.1). Dans ce modèle, chaque classe latente (*i.e.* chaque modalité de la VL) représente un cluster (probabiliste) des différentes configurations possibles prises par les VO. L'hypothèse d'indépendance locale, inhérente

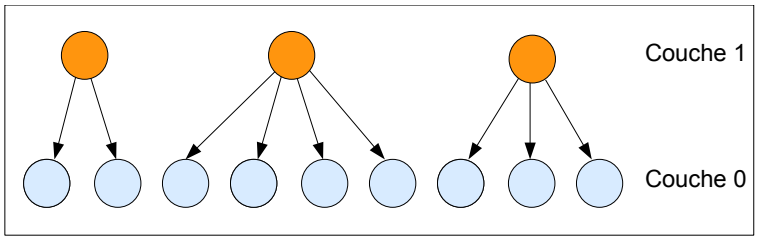

FIGURE 4.3: Ensemble de modèles à classes latentes, proposé par Zhang et Ji [116].
Voir figure 4.2 pour la nomenclature des nœuds.

à ce modèle, stipule que les VO sont toutes indépendantes conditionnellement à la VL.

Par exemple, Nefian [68] modélise le LD à l'aide d'un ensemble de modèles à classes latentes augmenté par des dépendances entre SNP et des dépendances entre VL (voir figure 4.2). Cette modélisation présente l'avantage de prendre en compte à la fois les dépendances à l'intérieur des blocs de SNP (représentées par les MCL) et les dépendances entre blocs (représentées par les liens entre les VL). L'apprentissage du modèle s'appuie sur un découpage arbitraire de la séquence en petites fenêtres de taille fixe (6 SNP). Pour chaque fenêtre, un MCL est créé. Ensuite, Nefian applique l'algorithme structural expectation-maximization (SEM) afin d'apprendre les dépendances entre SNP et les dépendances entre VL : à chaque étape de l'algorithme SEM, les données des VL sont d'abord complétées à l'aide du modèle courant, dans le but de pouvoir ensuite calculer le score des modèles dans le voisinage du modèle courant. Nefian emploie des opérateurs d'ajout/retrait d'arcs entre VL et entre VO pour le parcours du voisinage des structures. Malheureusement, le manque de flexibilité de la méthode de Nefian demeure un inconvénient important.

Assez similaire au modèle de Nefian, le modèle implémenté dans le programme HaploBlock (http ://bioinfo.cs.technion.ac.il/haploblock/) peut être considéré comme un ensemble de blocs, chacun modélisé par un MCL, et reliés par une chaîne de Markov qui prend en compte les dépendances entre blocs adjacents [29, 30]. L'atout de ce modèle est qu'il met en œuvre un certain nombre de concepts de génétique des populations comme la mutation, le goulot d'étranglement, la dérive génétique et les points chauds de recombinaison (pour une introduction aux concepts de génétique des populations, se référer à Réf.[69]).

Zhang et Ji, quant à eux, proposent une approche plus souple pour la modélisation du LD. Elle constitue une amélioration directe du modèle implémenté dans le logiciel Gerbil [44], basé sur une structure en blocs. Pour cela, un ensemble de MCL est appris (voir figure 4.3) à l'aide de l'algorithme SEM [116]. L'algorithme SEM développé met en œuvre un opérateur de réaffectation d'un SNP d'un cluster vers un autre, et intègre un recuit simulé afin d'augmenter la probabilité de convergence vers une solution globale. A la différence des travaux précédents, cette méthode peut partitionner la séquence en clusters de tailles variées de SNP non nécessairement adjacents. Malgré

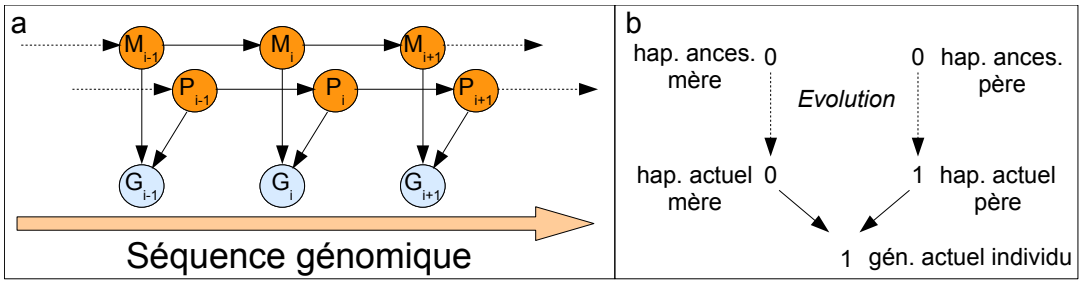

FIGURE 4.4: Le modèle de Markov caché utilisé par Scheet et Stephens [82] : a) représentation global du modèle, b) illustration pour un SNP.
a) Les nœuds M_i et P_i sont les haplotypes ancestraux maternels et paternels au SNP i, respectivement. Le nœud G_i est le génotype actuel observé au SNP i. b) Pour les haplotypes, les symboles 0 *et* 1 *représentent l'allèle originel et l'allèle mutant, respectivement. Pour les génotypes, les symboles* 0, 1, 2 *représentent le nombre d'allèles mutants. Hap. : haplotype ; Géno. : génotype ; Ances. : ancestral.*

cette amélioration, un inconvénient réside dans le fait que le nombre de MCL (*i.e.* de clusters) doit être spécifié au préalable.

4.2.2 Passage à l'échelle

Aucune des quatre méthodes mentionnées précédemment ne permet le passage à l'échelle avec le nombre de SNP : elles ne peuvent traiter plus d'un millier de SNP, ce qui les restreint à l'analyse de petites régions chromosomiques. Pour résoudre ce problème, plusieurs méthodes ont été proposées : les premières sont basées sur un processus markovien, tandis que les secondes contraignent l'apprentissage de la structure par une distance physique maximale.

Les processus markoviens constituent des outils naturels pour le traitement de données séquentielles, telles que les données génétiques. Dans cette perspective, les modèles de Markov cachés (MMC) présentent l'avantage de manipuler des états latents, comme la phase manquante dans les données génotypiques. En se basant sur cette idée, Scheet et Stephens [82] proposent un modèle où les états latents correspondent à des clusters haplotypiques, interprétés comme des haplotypes ancestraux. Le modèle est présenté en figure 4.4a, en suivant la représentation des RB. Les arcs verticaux peuvent être considérés comme des mutations apparues entre les haplotypes ancestraux (maternels et paternels) et le génotype actuel observé, alors que les arcs horizontaux sont des évènements de recombinaison survenus entre deux SNP contigus. Ce modèle est implémenté dans le programme bien connu fastPHASE, disponible à l'adresse Internet http ://stephenslab.uchicago.edu/software.html. La grande force de ce modèle réside dans le fait qu'il peut prendre en compte à la fois la structure en blocs du LD et la diminution graduelle de ce dernier avec la distance séparant les SNP. En outre, fastPHASE est précis et capable de manipuler de grands jeux de

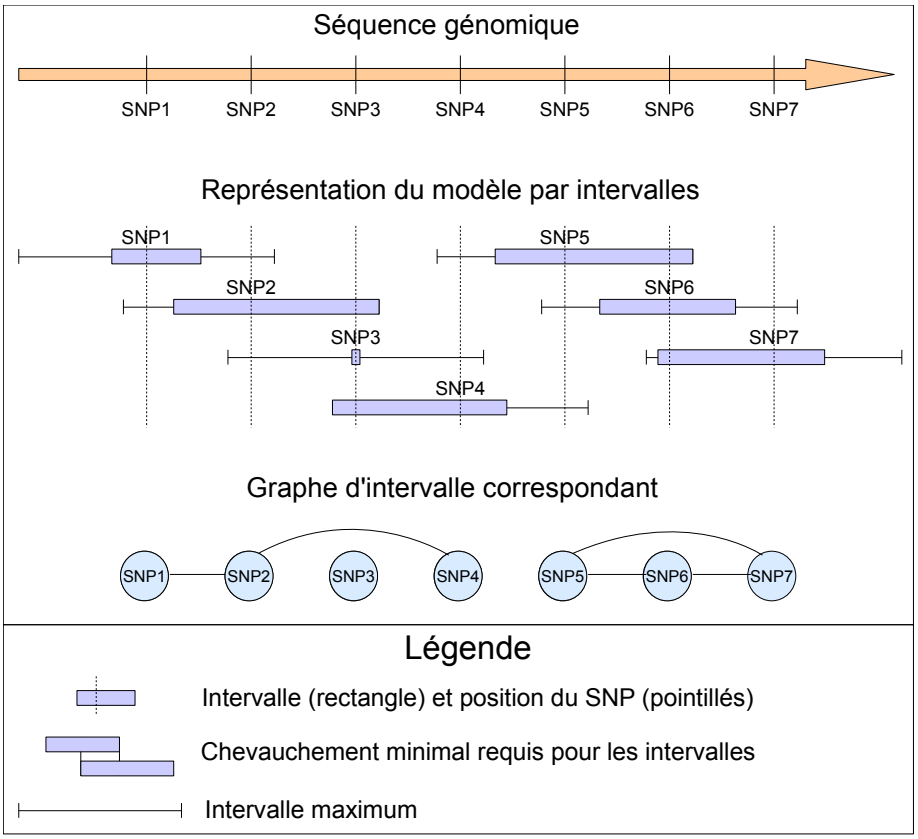

FIGURE 4.5: Graphe d'intervalle (GI) utilisé par Thomas [100].
Voir figure 4.2 pour la nomenclature des nœuds.

données (un millier d'individus et des centaines de milliers de SNP). Le logiciel fastPHASE a été considéré comme l'outil de référence pour l'inférence pangénomique des haplotypes jusqu'à l'apparition de la méthode de Browning et Browning implémentée dans la suite logicielle BEAGLE [11, 13].

Browning [12] propose l'emploi du modèle de Markov de longueur variable (MMLV) qui s'adapte automatiquement à l'étendue du LD entre les SNP présents sur le chromosome. Dans le MMLV, la longueur de la mémoire du processus n'est pas constante et peut varier le long de la chaîne selon le contexte, permettant ainsi de prendre en compte la nature multilocus du LD. Par exemple, la longueur de la mémoire sera plus importante dans les régions de fort LD que dans les régions de faible LD. En comparaison du MMC mentionné précédemment, le MMLV présente l'avantage de ne pas nécessiter de spécification préalable de la structure du modèle, telle que le nombre d'haplotypes ancestraux, et l'apprentissage peut être réalisé au moyen d'heuristiques rapides. La suite logicielle BEAGLE implémente cet algorithme. Elle est disponible à l'adresse Internet http ://faculty.washington.edu/browning/beagle/beagle.html.

Afin de surmonter le problème de la grande dimension des données, une autre

stratégie consiste à restreindre l'espace de recherche des MGP par une contrainte portant sur la distance physique séparant les SNP. Suivant cette idée, deux méthodes ont été développées. Verzilli *et al.* [107] ont conçu une approche "Markov chain Monte Carlo" (MCMC) afin d'échantillonner dans l'espace des cliques disjointes, un sous-espace des graphes non orientés décomposables. Deux opérateurs simples de parcours sont utilisés : la fusion et la scission de deux cliques choisies aléatoirement dans le graphe courant. La distance physique maximale entre SNP pouvant appartenir à une même clique, ainsi que la taille maximale des cliques, sont fixées, assurant ainsi le passage à l'échelle de la méthode. L'utilisation de ces contraintes ainsi que le parcours d'un espace plus petit que celui des RM décomposables - l'espace des cliques disjointes - assure un apprentissage très rapide. Pour donner un ordre d'idée : l'algorithme ne nécessite que 5 minutes pour traiter 10^5 SNP et 268 individus. Cette méthode est implémentée dans le package R graphminer et est disponible à l'adresse Internet http ://homepages.lshtm.ac.uk/encdcver/.

Thomas [100] a proposé d'employer les graphes d'intervalle, une sous-classe des graphes non orientés décomposables. Un graphe d'intervalle (GI) est un graphe pour lequel les nœuds représentent des intervalles localisés sur une même ligne et pour lequel les arêtes connectent les paires d'intervalles se chevauchant. Les GI sont intrinsèquement décomposables. La décomposabilité en cliques de petite taille constitue la clé pour le passage à l'échelle de l'apprentissage des RM. Elle assure un calcul simple et local de la vraisemblance, évite une perte de temps importante liée à la proposition de solutions non décomposables, et permet l'emploi d'algorithmes d'inférence performants tels que l'algorithme forward-backward (appelé aussi message passing dans l'arbre de jonction) [48]. Outre ces atouts calculatoires, les GI apparaissent intuitivement comme des outils particulièrement adaptés pour la modélisation du LD, puisque les intervalles peuvent être interprétés comme le LD présent autour d'un locus. Les GI sont facilement contraints à l'aide des positions physiques des SNP, reflétant le fait que des corrélations fortes auront davantage de chance d'être observées entre SNP proches sur le chromosome. Le modèle est présenté en figure 4.5.

En théorie, une structure de données en arbre peut être utilisée pour stocker et mettre à jour en temps logarithmique les intervalles inférés par échantillonnage MCMC (ici nous parlons de complexité en fonction du nombre de SNP à traiter). Cependant, en pratique, la construction de l'arbre se révèle superlinéaire en temps lorsque de grands jeux de données sont traités (plus de 10^4 SNP). Pour résoudre ce problème, une nouvelle restriction a été introduite [101] : les intervalles ne sont pas autorisés à dépasser une largeur maximale de part et d'autre de la position du SNP. Très récemment, l'emploi de GI est devenu obsolète avec le développement du modèle d'Abel et Thomas [1] qui se base sur un nouvel algorithme d'échantillonnage dans l'espace général des graphes décomposables [103, 104]. Pour ce modèle, la linéarité de l'algorithme d'apprentissage a aussi été atteinte au moyen de contraintes portant sur les graphes : seuls les SNP séparés par une distance inférieure à une distance maximale peuvent être dépendants. Cette contrainte est en fait identique à celle employée pour l'apprentissage des GI. Afin d'estimer le modèle (GI ou graphe décomposable)

<table>
<tr><th>Modèle</th><th>Variables</th><th>Restrictions</th><th>Passage à l'échelle</th><th>Paradigme</th><th>Objectif</th><th>Logiciel</th><th>Référence</th></tr>
<tr><td>RM</td><td rowspan="2">allèles</td><td>décomposabilité</td><td rowspan="5">non</td><td rowspan="4">fréquentiste</td><td>modélisation du LD</td><td>HapGraph</td><td>(Thomas et Camp, 2004)</td></tr>
<tr><td>RB</td><td rowspan="3">aucune</td><td>sélection de tagSNPs</td><td>-</td><td>(Lee et Shatkay, 2006)</td></tr>
<tr><td rowspan="2">RB avec VL</td><td rowspan="2">allèles
&
clusters
haplotypiques</td><td>modélisation du LD</td><td>-</td><td>(Villanueva et Maciel, 2010)</td></tr>
<tr><td>modélisation du LD</td><td>-</td><td>(Nefian, 2006)</td></tr>
<tr><td>inférence de blocs haplotypiques</td><td>-</td><td>(Zhang et Ji, 2009)</td></tr>
<tr><td>RB avec VL (MMC)</td><td>allèles
&
clusters
haplotypiques</td><td>ordre physique des SNP</td><td rowspan="4">oui</td><td>fréquentiste</td><td>inférence d'haplotypes</td><td>fastPHASE</td><td>(Scheet et Stephens, 2006)</td></tr>
<tr><td>RB (MMLV)</td><td>clusters haplotypiques</td><td rowspan="2">fréquentiste</td><td rowspan="2">étude d'association pangénomique</td><td>Beagle</td><td>(Browning, 2006)</td></tr>
<tr><td>RM</td><td>génotypes</td><td>distance physique & décomposabilité</td><td>bayésien</td><td>graphminer</td><td>(Verzilli et al., 2006)</td></tr>
<tr><td rowspan="2">RM</td><td rowspan="2">allèles</td><td>distance physique & graphe d'intervalle</td><td>fréquentiste</td><td rowspan="2">inférence d'haplotypes</td><td>IntervalLD</td><td>(Thomas, 2009a)
(Thomas, 2009b)</td></tr>
<tr><td>distance physique & décomposabilité</td><td>fréquentiste</td><td>FitGMLD</td><td>(Abel et Thomas, 2011)</td></tr>
</table>

TABLEAU 4.1: Comparaison des méthodes basées sur des modèles graphiques probabilistes pour la modélisation du déséquilibre de liaison. *RM : réseau de Markov ; RB : réseau bayésien ; MMC : modèle de Markov caché ; MMLV : modèle de Markov de longueur variable ; VL : variable latente ; LD : déséquilibre de liaison.*

et d'inférer les haplotypes, une approche "diviser pour régner" considère une fenêtre glissante le long du chromosome. A l'intérieur de la fenêtre, l'algorithme forward-backward est employé. Grâce à ce moyen, les complexités en temps et en mémoire sont linéaires.

4.2.3 Applications

Les applications des MGP à la modélisation du LD sont multiples. Elles vont de la sélection de tagSNPs à l'inférence des haplotypes. Par exemple, dans BNTagger, une heuristique recherche les tagSNPs, qui peuvent être employés de différentes manières dans le cadre des EAP. Par exemple, les tagSNPs servent à réduire la dimension des données pour la recherche d'associations. Ils sont aussi très utiles pour l'imputation [1] de données de SNP non génotypés. Ainsi l'on peut génotyper qu'un nombre réduit de SNP et imputer ensuite les SNP restants pour assurer une localisation plus précise des SNP causaux. Pour la sélection de tagSNPs, BNTagger surclasse plusieurs méthodes de l'état de l'art comme Eigen2htSNP [52] et STAMPA [31].

La méthode d'Abel et Thomas [1], quant à elle, peut s'appliquer à : (i) l'inférence pangénomique d'haplotypes, (ii) la simulation d'haplotypes pour les individus d'un pedigree de structure connue lorsque les haplotypes fondateurs présentent du LD, et (iii) le calcul de la distribution *a posteriori* des allèles pour chacun des loci. Par exemple, l'inférence d'haplotypes à partir de données génotypiques est une problématique majeure pour le généticien. Les haplotypes représentent l'information génétique des chromosomes. Leur étude aide à mieux comprendre les processus évolutifs des populations humaines. Outre cet aspect théorique, il est important d'identifier les

1. L'imputation consiste à inférer des données manquantes à l'aide d'une méthode statistique.

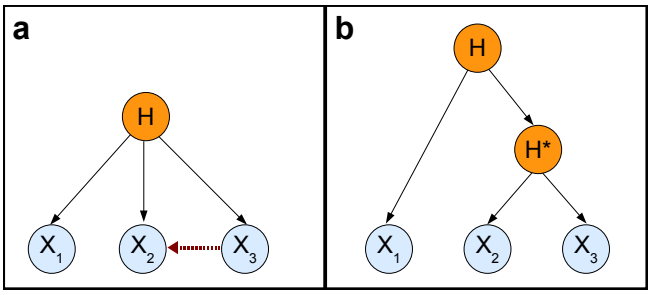

FIGURE 4.6: a) Dépendance locale (non prise en compte) dans un modèle à classes latentes. b) Modélisation de la dépendance locale à l'aide d'un modèle hiérarchique à classes latentes. *Voir figure 4.2 pour la nomenclature des nœuds.*

haplotypes portant la mutation causale à l'origine de la maladie. En qui concerne l'inférence d'haplotypes pour de très grands jeux de données (*e.g.* 10^3 individus et 10^5 SNP), la méthode d'Abel et Thomas obtient des résultats légèrement moins précis que ceux de l'algorithme le plus performant implémenté dans la suite logicielle BEAGLE [13].

Nous venons de présenter différentes méthodes basées sur les MGP et visant à modéliser le LD. Ces différentes méthodes de modélisation sont résumées dans le Tableau 4.1. Nous avons comparé les familles de modèles, les variables analysées, les restrictions imposées lors de l'apprentissage, la capacité de la méthode à passer à l'échelle, le paradigme d'apprentissage statistique et l'objectif de la méthode.

4.3 Forêt de modèles hiérarchiques à classes latentes

Les travaux sur le développement d'un nouveau modèle, la forêt de modèles hiérarchiques à classes latentes, dédié à la modélisation du LD ont été publiés dans Réf.[63, 64].

4.3.1 Introduction et motivations

Le modèle hiérarchique à classes latentes (MHCL) a été proposé comme généralisation du MCL [113]. Il présente l'avantage de ne plus être basé sur une hypothèse souvent infirmée par les données, l'hypothèse d'indépendance locale des VO. Ce phénomène est illustré en figure 4.6a. Une dépendance locale est observée entre les VO X_2 et X_3 conditionnellement à la VL H (flèche en pointillés). Cette dépendance locale peut être modélisée à l'aide d'une VL supplémentaire H^* reliant les VO X_2 et X_3 au reste du modèle.

Le MHCL est défini comme un arbre orienté dont les feuilles sont des VO et les nœuds restants sont des VL, organisées en couches multiples (voir figure 4.7). Outre

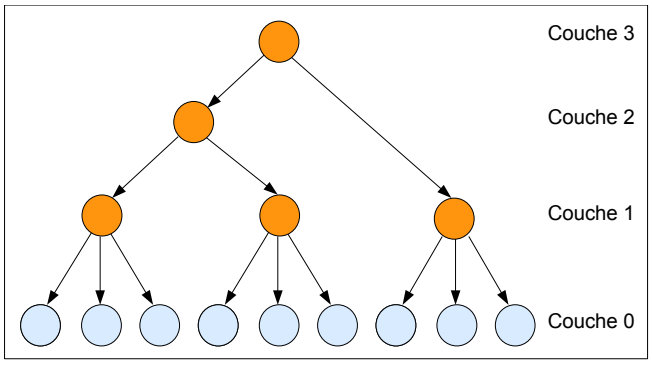

FIGURE 4.7: Modèle hiérarchique à classes latentes (MHCL).
Voir figure 4.2 pour la nomenclature des nœuds.

sa capacité à prendre en compte la dépendance locale entre VO, le MHCL offre de nombreux avantages pour l'analyse du LD. Premièrement, il permet une modélisation précise et souple du LD. D'une part, la structure en blocs, dans laquelle les SNP ne sont regroupés que s'ils sont adjacents sur l'ADN, est remplacée par une structure en clusters, moins contraignante, car elle n'oblige pas les SNP à être contigus. Dans ce modèle, un cluster est défini comme un ensemble de SNP (*i.e.* feuilles dans l'arbre) "subsumés"[2] par une VL. D'autre part, la structure hiérarchique du modèle peut être vue comme une forme de classification ascendante hiérarchique des SNP. Cette structure présente ainsi l'originalité de prendre en compte la nature floue des frontières définissant les clusters. En effet, cette hiérarchie définit différents degrés de dépendances entre SNP, ce qui limite l'aspect arbitraire du découpage en clusters, lequel aboutit nécessairement à un seuillage des dépendances.

Deuxièmement, ce modèle peut fournir une interprétation biologique du LD dans le cas de l'analyse de données haplotypiques. En effet, une VL peut être vue comme un clustering (probabiliste) des haplotypes multilocus définis par les VO subsumées par la VL. Ainsi, chaque état (ou classe) de la VL peut être considéré comme un cluster d'haplotypes. Or il a été montré que dans une situation de recombinaison ancestrale limitée, des haplotypes similaires tendent à partager une parenté ancestrale commune [62]. Cette situation est très vraisemblable pour les VL de bas niveau dans l'arbre, supposées couvrir de petites régions chromosomiques montrant un fort LD et un faible taux de recombinaison. Dans la situation inverse, *i.e.* où une VL connecte des SNP éloignés (*e.g.* plus de 100 *kb*), la VL pourra être interprétée comme un effet de la structure de la population [75].

Une amélioration triviale du MHCL, mais qui se révèle plus adaptée pour le LD, consiste à ne plus considérer une structure en arbre mais une structure en forêt, *i.e.* un ensemble d'arbres (voir figure 4.8). Dans cette nouvelle structure définissant un nouveau modèle, la forêt de modèles hiérarchiques à classes latentes (FMHCL), les SNP

2. Nous définissons, dans un arbre ou une forêt, la subsomption comme la relation descendant-ancêtre.

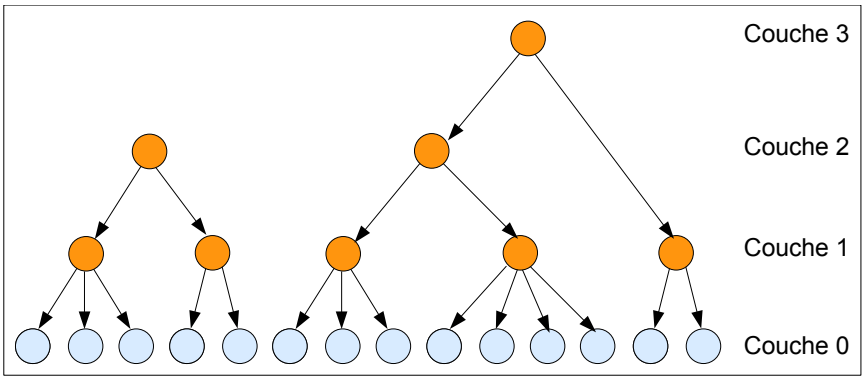

FIGURE 4.8: Forêt de modèles hiérarchiques à classes latentes (FMHCL).
Voir figure 4.2 pour la nomenclature des nœuds.

ne sont plus contraints à être dépendants les uns des autres, que ce soit directement (via une VL) ou indirectement (via une chaîne de VL). Cette particularité assure une meilleure modélisation puisque, dans la majorité des situations, le LD est inexistant entre SNP très éloignés. La figure 4.9 illustre ce phénomène. Elle représente le LD par paires de SNP pour une séquence de 2 Mb. En outre, l'emploi de la FMHCL autorise un meilleur passage à l'échelle lors de l'apprentissage du modèle puisqu'il est possible de contraindre la recherche des dépendances entre SNP par la distance physique les séparant.

Les avantages de la modélisation du LD par une FMHCL sont nombreux. Tout d'abord, ce modèle permet de réduire la dimension des données génétiques, ce qui est avantageux dans le contexte des études pangénomiques. En effet, les VL peuvent être employées à la place des SNP pour la recherche d'associations avec la maladie. De plus, la structure du modèle peut guider la découverte d'associations causales, *i.e.* des associations impliquant des SNP causaux (ou les SNP les plus proches des mutations causales non observées). En effet, il est envisageable d'employer, par exemple, des tests d'indépendance conditionnelle mesurant la dépendance SNP-maladie conditionnellement à la VL parente du SNP. Étant donné que, selon la structure du modèle, un SNP est indépendant de tous les autres SNP conditionnellement à sa VL parentale, alors il est possible de mesurer le lien direct existant entre ce dernier et la maladie, assurant ainsi l'identification d'associations causales. Des méthodes plus complexes de découverte d'associations causales sont possibles, notamment à travers le parcours en profondeur de la forêt. Outre la recherche d'associations, la FMHCL représente un outil particulièrement adéquat pour la visualisation de la structure spatiale du LD. En effet, sa nature graphique et sa structure hiérarchique autorisent une visualisation synthétique et intuitive des dépendances entre SNP.

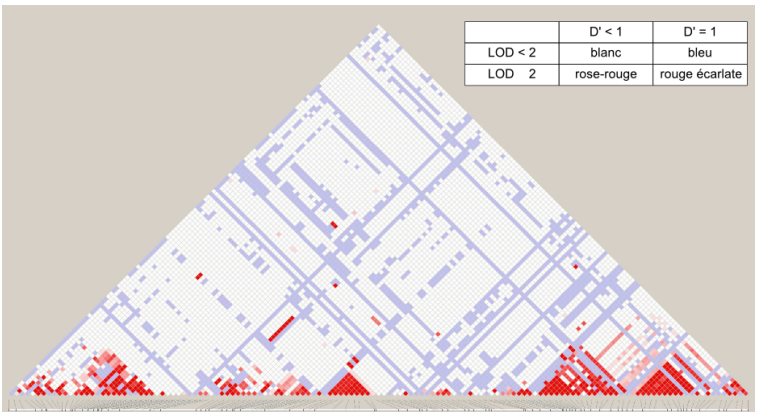

FIGURE 4.9: Carte triangulaire de chaleur du D'/LOD d'une séquence de 2 Mb provenant du génome humain.

4.3.2 État de l'art sur l'apprentissage de modèles hiérarchiques à classes latentes

Aucune méthode antérieure au travail présenté dans ce mémoire n'a été conçue pour l'apprentissage de FMHCL à partir de données. Néanmoins, l'apprentissage des MHCL - modèles très proches - a fait l'objet d'un certain nombre de travaux de recherche. Dans la suite, nous présentons l'état de l'art des MHCL.

Généralement, l'apprentissage des RB nécessite d'abord la mise en œuvre d'une étape d'inférence de structure (G) avant l'inférence des paramètres du modèle (θ). Dans le cas des MHCL, cet apprentissage est plus complexe, puisque, d'une part, le nombre et la cardinalité des VL doivent en plus être appris, et d'autre part, la frontière entre apprentissage de structure et apprentissage de paramètres est plus mince car les deux étapes sont généralement imbriquées. Nous verrons que les approches développées pour l'apprentissage diffèrent principalement selon les points suivants : (i) l'apprentissage de la structure ; (ii) la détermination de la cardinalité des VL ; (iii) l'apprentissage des paramètres ; (iv) le passage à l'échelle ; (v) l'application.

Deux grands types d'approches ont été mises en œuvre : les méthodes basées sur un score et les méthodes basées sur le clustering de variables. Pour le premier type, l'algorithme SEM permet d'optimiser successivement θ conditionnellement à G ($\theta \mid G$) et G conditionnellement à θ ($G \mid \theta$). Afin de parcourir l'espace des structures, la recherche gloutonne à l'aide du critère BIC a été appliquée avec succès [114]. Pour cela, l'espace des structures et des cardinalités des VL est exploré à l'aide de quatre opérateurs : ajout ou suppression d'une VL, et ajout ou suppression d'un état à une VL.

L'apprentissage par clustering de variables offre une alternative qui consiste à tester statistiquement la dépendance entre variables et à regrouper celles qui sont dépendantes. L'idée sous-jacente est qu'il est probable que les dépendances observées

entre les variables proviennent de l'influence d'une ou plusieurs variables non observées, c'est-à-dire de variables latentes. Dans cette perspective, la classification hiérarchique de variables apparaît comme une solution adéquate pour inférer la structure des MHCL. En effet, si l'on considère que la hiérarchie inférée comme un arbre, alors chaque noeud interne de la hiérarchie peut être considéré comme une variable latente de l'arbre. En se fondant sur cette idée, Wang *et al.* [110] commencent par construire un arbre binaire par classification ascendante hiérarchique (CAH) classique ; ensuite, ils appliquent des opérations de régularisation et de simplification de la structure, de façon à ce qu'éventuellement plus de deux nœuds puissent être subsumés par une VL. L'approche de Hwang *et al.*, quant à elle, restreint la recherche à celle des arbres binaires [37]. Afin de construire l'arbre, une méthode similaire à la CAH est employée. D'abord, une partition des VO par paires est réalisée, selon le critère de la maximisation de l'information mutuelle. Chaque cluster de deux variables identifié définit ensuite un MCL (le MCL a été détaillé en section 4.2.1, page 44). Deuxièmement, les paramètres sont appris par EM pour chaque MCL. Les valeurs manquantes des VL peuvent ainsi être imputées, et les VL sont alors considérées comme de nouvelles VO lors d'une prochaine étape de partitionnement. L'arbre est alors construit par itération de ces deux étapes (partitionnement, imputation des valeurs manquantes des VL).

L'apprentissage des paramètres en présence de VL requiert au préalable la détermination de leur cardinalité, *i.e.* leur nombre d'états (ou de classes). La méthode la plus simple consiste à ignorer cette étape de détermination et à assigner arbitrairement une valeur pour la cardinalité. Suivant cette idée, Hwang *et al.* [37] choisissent de fixer la cardinalité des VL à deux dans les arbres binaires dont les VO sont elles-mêmes binaires. Dans cette situation, fixer la cardinalité à une valeur égale à 2 n'est pas dénué de bon sens. En effet, cette stratégie réduit raisonnablement la cardinalité de la distribution de probabilité jointe des deux variables enfants subsumées par la VL, en passant d'une valeur de 4 (2×2) à une valeur de 2. Cependant, dans le cas général où les VO ne sont pas binaires, cette stratégie présente plusieurs défauts : d'une part, une cardinalité trop faible entraînera une perte d'information conséquente lors du processus de subsomption des VO ; d'autre part, une cardinalité trop élevée engendrera un surapprentissage du modèle et un coût calculatoire plus élevé lors de l'apprentissage des paramètres. Wang *et al.* [110] proposent une étape de régularisation afin de réduire la cardinalité d'une VL, sachant celles des voisins (variables enfants et parents). Considérons une VL Y dont les voisins sont Z_i. Après régularisation, la cardinalité sera la suivante :

$$|Y| = \frac{\Pi_{i=1}^{k}|Z_i|}{max_{i=1}^{k}|Z_i|}. \tag{4.1}$$

Il est aussi envisageable d'employer une approche par recherche gloutonne. En fixant une valeur de départ pour la cardinalité, celle-ci est incrémentée ou décrémentée jusqu'à atteindre un optimum local pour un score donné. Cette approche présente l'avantage d'être plus efficace pour retrouver la cardinalité optimale, mais de loin beaucoup plus gourmande en calculs, puisqu'elle nécessite plusieurs exécutions de l'algorithme

EM, qui suit lui-même une procédure itérative. Il faut noter qu'il est préférable dans la pratique d'initialiser la cardinalité à une valeur faible (*e.g.* 2) puis d'incrémenter, plutôt que de procéder de manière inverse.

Généralement, l'algorithme EM est employé pour l'apprentissage de paramètres en présence de VL ou de données manquantes, mais il présente l'inconvénient de nécessiter des temps de calculs importants et ne garantit pas la convergence vers un optimum global. Afin d'accélérer cet algorithme, Hwang *et al.* [37] ont implémenté une heuristique basée sur l'imputation déterministe partielle des valeurs manquantes des VL. Pour chaque MCL composé de deux variables enfants E_i et E_j, l'heuristique est la suivante : tous les individus présentant la configuration la plus fréquente de $\{E_i, E_j\}$ se voient assigner une valeur donnée (*e.g.* 0) pour la VL. De manière analogue, les individus caractérisés par la seconde configuration la plus fréquente se voient assigner une autre valeur (*e.g.* 1). Après que les données aient été partiellement imputées, l'algorithme EM est exécuté pour l'apprentissage des paramètres. Finalement, bien que de nombreuses méthodes autres qu'EM existent pour l'apprentissage de modèles à variables latentes, telles que les méthodes de descente de gradient [87], ces dernières n'ont pas été employées dans le cas des MHCL. Cela s'explique probablement par la plus grande simplicité de mise en œuvre de l'algorithme EM, par comparaison avec les autres méthodes.

L'approche de Hwang *et al.* est la seule capable de passer à l'échelle pour le traitement de données de grande dimension. Lors d'une application à un jeu de données d'expression génique, plus de 6000 gènes et 60 individus ont pu être traités. Malgré cela, la série de restrictions imposées par la méthode de Hwang *et al.*, telles que l'emploi d'arbres binaires et de variables binaires, représente un inconvénient majeur pour notre objectif de modélisation fine et flexible du LD. Le manque de méthode d'apprentissage dédiée à la FMHCL, ainsi que l'absence de méthode performante pour le MHCL nous ont amenés à développer notre propre méthode.

4.3.3 Algorithme CFHLC

Notre algorithme d'apprentissage, nommé Construction of Forests of Hierarchical Latent Class models (CFHLC), implémente une approche hiérarchique proche de celle de Hwang *et al.* [37]. Cependant CFHLC présente un avantage majeur : il lève la double restriction arbre binaire-variables binaires. Au lieu de construire un MCL pour chaque paire de variables dépendantes, les variables sont regroupées en ensembles pouvant contenir plus de deux variables. Ainsi il est possible de créer des arbres qui ne sont pas binaires. La cardinalité des VL, quant à elle, est déterminée sachant le nombre de variables du MCL auquel elle appartient. De ce fait, les VL créées peuvent avoir une cardinalité supérieure à 2.

Deux variantes de CFHLC ont été proposées. Dans la première, pour des raisons de passage à l'échelle, un découpage arbitraire de la séquence chromosomique en

larges fenêtres a été proposé, tandis que dans la seconde version, l'étape de découpage n'est plus nécessaire. Nous commencerons par exposer la première variante, puis nous présenterons les améliorations ayant permis de développer la seconde variante.

4.3.3.1 Principe

Notre algorithme peut traiter à la fois des données génotypiques (phase inconnue) et des données haplotypiques (phase connue), et *a fortiori* tout type de données discrètes. Bien que notre algorithme soit une méthode d'inférence de valeurs manquantes des VL, elle ne réalise pas de phasage de données génotypiques. L'algorithme prend en entrée une matrice $\mathscr{D}_{\mathbf{X}}$ définie sur un domaine fini discret, par exemple $\{0,\ 1,\ 2\}$ pour les données génotypiques ou $\{0,\ 1\}$ pour les données haplotypiques de chaque SNP. $\mathscr{D}_{\mathbf{X}}$ décrit un ensemble de n individus et de p variables $\mathbf{X} = \{X_1, ..., X_p\}$. L'algorithme CFHLC retourne une FMHCL, c'est-à-dire une structure de forêt G et θ, les paramètres du modèle. Deux espaces de recherche sont explorés : l'espace des forêts orientées, et l'espace probabilisé. De plus, l'ensemble des variables latentes $\mathbf{H}$ de la FMHCL est calculé, ainsi que la matrice de données imputées associées $\mathscr{D}_{\mathbf{H}}$.

Afin de manipuler des données de grande dimension, notre approche combine deux stratégies. La première est simple et consiste à découper les données pangénomiques en régions contiguës. Dans le cas de la modélisation du LD, ce découpage, bien que restrictif, n'est pas dénué de fondement biologique, puisque la majorité des dépendances s'observe entre SNP proches sur l'ADN. Ensuite, une FMHCL est apprise pour chaque fenêtre. A l'intérieur d'une fenêtre, l'apprentissage est réalisé par subsomption à l'aide d'une procédure ascendante hiérarchique : (i) à chaque étape d'agglomération, un partitionnement en cliques est employé afin d'identifier des cliques de variables ; (ii) ensuite, chaque clique est susceptible d'être subsumée par une VL, constituant ainsi un MCL. Pour chaque MCL, l'apprentissage des paramètres et l'imputation des données manquantes (pour la variable latente) sont réalisés. Le schéma global de la méthode est présenté en figure 4.10.

4.3.3.2 Partitionnement en cliques de variables

Martin et VanLehn [56] ont proposé d'associer une VL à chaque clique maximale du graphe non orienté des dépendances par paires de variables (voir figure 4.11). Cependant, la recherche de telles cliques est un problème NP-complexe. De plus, contrairement à l'objectif de Martin et VanLehn, notre tâche est de découvrir des cliques non chevauchantes, puisqu'à l'intérieur d'une FMHCL, une variable ne peut avoir qu'un seul parent. Ce sont les raisons pour lesquelles l'emploi d'une méthode de partitionnement en cliques de variables apparaît plus appropriée que l'approche de Martin et VanLehn. En effet, elle permet d'identifier des cliques non chevauchantes de manière efficace et simple à partir de la connaissance des dépendances par paires de variables.

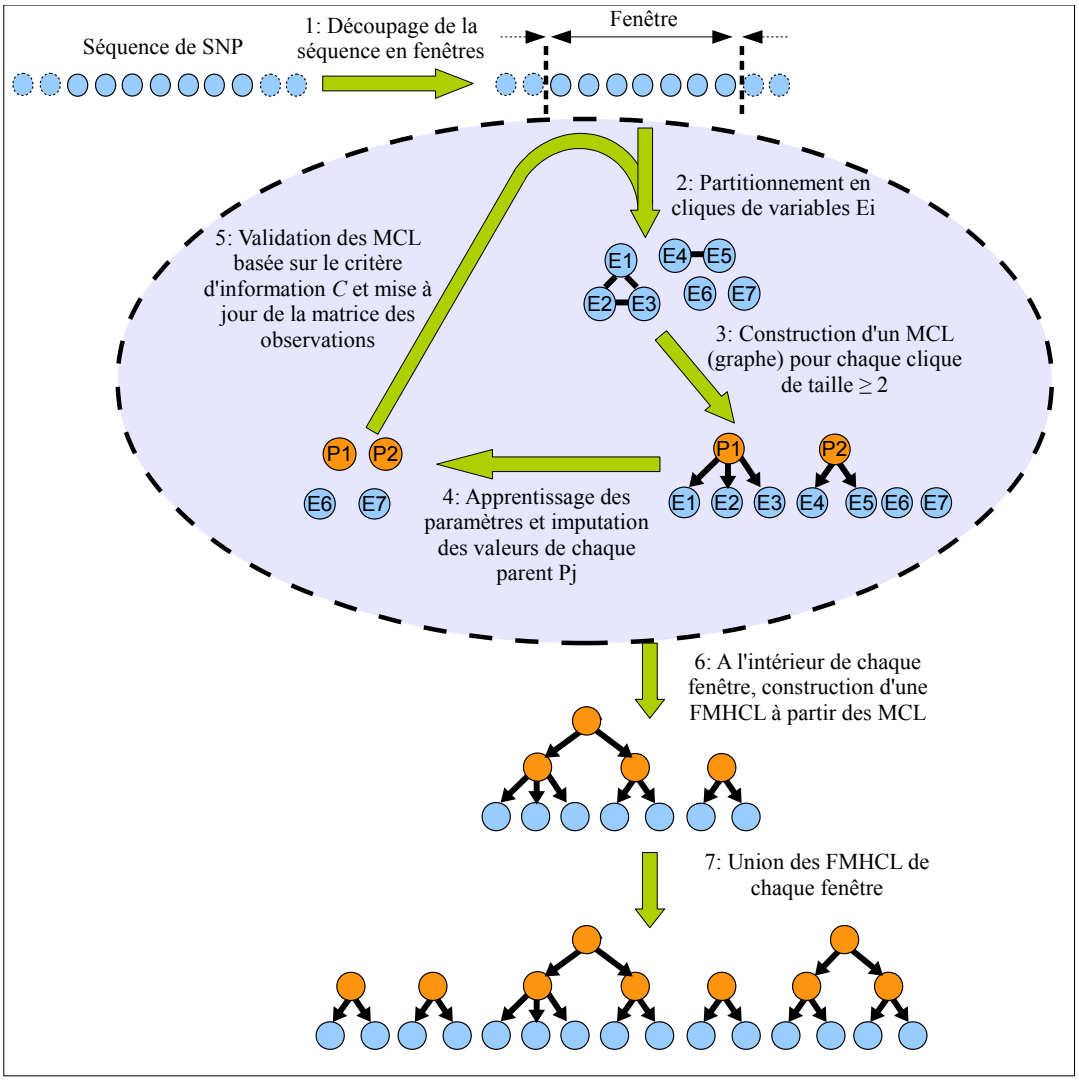

FIGURE 4.10: Schéma de l'algorithme CFHLC.

Un algorithme de partitionnement en cliques de variables, à la fois performant et permettant le passage à l'échelle, a été proposé par Ben-Dor *et al.* [9]. Cet algorithme, nommé Cluster Affinity Search Technique (CAST), a été conçu pour l'identification de clusters de gènes pour des données d'expression de grande dimension. En entrée, CAST requiert la matrice d'adjacence du graphe non orienté des dépendances entre variables. Afin de la calculer simplement à partir des données, nous proposons d'employer la matrice des informations mutuelles des paires de SNP. Ensuite, cette matrice est seuillée au moyen d'un seuil t_{MI}. Le quantile des valeurs de la matrice est employé comme seuil. L'emploi du quantile, plutôt qu'une autre valeur arbitraire fixée par l'utilisateur, a son importance. Pour mieux le comprendre, il faut savoir que la distribution des informations mutuelles évolue d'une étape à une autre de CFHLC puisque des VO peuvent être remplacées par des VL. L'emploi du quantile permet

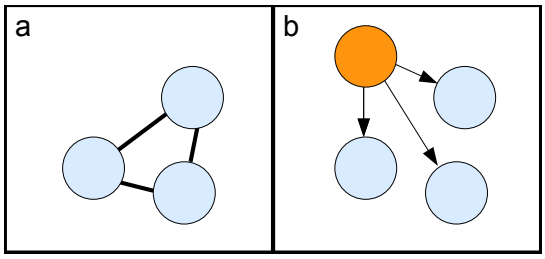

FIGURE 4.11: Introduction d'une variable latente à (a) une clique maximale du graphe non orienté des dépendances entre variables, afin de former (b) un modèle à classes latentes. *Voir figure 4.2 pour la nomenclature des nœuds.*

alors la découverte des cliques les plus solides (dont les variables sont les plus dépendantes), relativement à la distribution des informations mutuelles entre les variables. Ensuite, CAST construit les clusters de manière gloutonne, *i.e.* les uns après les autres. Les auteurs de CAST définissent une mesure d'affinité $a(x)$ d'un élément x comme la somme des valeurs de similarité (0 ou 1 dans la matrice d'adjacence) entre x et les éléments présents dans le cluster courant $C_{courant}$. x est un élément de forte affinité s'il vérifie l'inégalité $a(x) \geq t_{CAST}|C_{courant}|$, avec t_{CAST}, un seuil de similarité spécifié par l'utilisateur. Sinon, x est considéré comme un élément de faible affinité. Pour résumer, CAST alterne entre ajout d'éléments de forte affinité à $C_{courant}$, et retrait d'éléments de faible affinité. Lorsque le processus s'est stabilisé, $C_{courant}$ est fermé. Le processus traite alors un nouveau cluster, et ainsi de suite.

4.3.3.3 Détermination de la cardinalité des variables latentes

Une étape difficile est de choisir - idéalement d'optimiser - la cardinalité de chaque VL. Ce problème ne peut se résoudre par une recherche gloutonne du fait de la grande dimension des données. Bien que la méthode de Wang *et al.* présente l'avantage d'être très rapide (voir section 4.3.2, paragraphe 4, page 54), dans notre contexte, elle est inapplicable. Prenons l'exemple d'un MHCL appris à partir de données de SNP. La première couche du MHCL contient la majorité des VL du modèle. Dans ce cas, une VL de la première couche pourra peut-être subsumer plus d'une dizaine de variables enfants (*i.e.* des SNP). Comme la cardinalité est égale à 3 pour tous les SNP (trois génotypes possibles), la cardinalité de la VL demeurera très élevée même après régularisation. Par exemple, pour un MCL contenant 10 SNP, la cardinalité après régularisation sera : $|H| = 3^{10}/3 = 3^9 = 19683$ (voir équation 4.1, page 55). Pour ne pas attribuer une valeur faible à la cardinalité de manière arbitraire (voir section 4.3.2, page 54), nous avons proposé de l'estimer à partir d'une information disponible, le nombre d'enfants à l'intérieur du MCL. La raison est la suivante : pour un nombre d'enfants restreint, plus ce nombre est élevé, plus grande est la cardinalité de la table de contingence de ces derniers, et plus grand est attendu le nombre de combinaisons observées sur les individus (cardinalité de la table de contingence, en ne comptant

pas les valeurs nulles). C'est la raison pour laquelle nous proposons d'estimer la cardinalité de la VL comme une fonction du nombre de ses enfants. Nous avons choisi d'employer une fonction affine, mais des études plus poussées sur la détermination de la meilleure fonction mériteraient d'être réalisées. Afin de limiter la complexité du modèle, une valeur maximale est fixée.

4.3.3.4 Apprentissage des paramètres et imputation

L'apprentissage des paramètres est réalisé étape après étape de partitionnement en cliques. A l'étape i, l'apprentissage des paramètres consiste simplement à créer un MCL pour chaque clique de plus de deux variables, puis à apprendre les paramètres à l'aide d'une procédure EM classique (voir section 3.4.4.2, chapitre 3, page 35). Cette procédure prend en entrée la cardinalité de la VL H et renvoie la distribution de probabilité du modèle, *i.e.* une distribution de probabilité *a priori* de H, et des distributions de probabilité conditionnelle des enfants $\mathbf{X}$ ($\mathbf{X} = \{X_1, ..., X_n\}$) conditionnellement à H. Après imputation des VL, de nouvelles données sont disponibles afin d'alimenter la nouvelle étape de construction de la FMHCL : chaque VL identifiée pendant l'étape i sera considérée comme VO durant l'étape $i + 1$.

Il faut noter que l'imputation de données des VL peut être réalisée de différentes façons. Pour chaque individu ℓ, il est possible d'échantillonner un ensemble de points de la VL H de manière à refléter au mieux la distribution du modèle, à l'aide de la probabilité suivante :

$$P(H = c|\mathbf{x}^\ell) = \frac{\Pi_{i=1}^{n}\ P(x_i^\ell|H = c)\ P(H = c)}{\sum_{c=1}^{k}\ \Pi_{i=1}^{n}\ P(x_i^\ell|H = c)\ P(H = c)},$$

avec $\mathbf{x}^\ell = \{x_1^j, ..., x_n^j\}$ l'ensemble des valeurs, pour l'individu ℓ, des variables enfants $\{X_1, ..., X_n\}$ de la VL, et k le nombre de classes de la VL H. Cependant, cette approche présente l'inconvénient de nécessiter une grande quantité de mémoire. Une autre manière de procéder consiste à considérer la probabilité d'appartenance d'un individu à une classe c comme un poids, et de construire la FMHCL sur des données pondérées, mais cela nécessite aussi un surplus de mémoire, qui est critique lors du passage à l'échelle sur des données pangénomiques. C'est la raison pour laquelle nous avons privilégié la solution la plus simple qui consiste à employer la valeur la plus probable dans la distribution :

$$c^* = argmax_c\{P(H = c|\mathbf{x}^\ell)\}.$$

4.3.3.5 Contrôle de la perte d'information

Contrairement à la méthode de Hwang *et al.* [37] qui vise principalement la compression de données, un contrôle de la perte d'information lors du processus de subsomption est requis dans notre cas : à chaque étape i, chaque VL candidate H qui

ne véhicule pas suffisamment d'information sur ses enfants doit être invalidée. Cela a pour conséquence de reconsidérer ces enfants comme des nœuds isolés lors de la prochaine étape $i+1$.

L'information mutuelle $\mathcal{I}(X;H)$ mesure la quantité d'information d'une variable enfant X véhiculée par la variable parent H (et *vice versa*). La définition de l'information mutuelle est rappelée en section 3.2.2, chapitre 3, page 27. Cette mesure d'information mutuelle peut être normalisée pour être ramenée à un critère plus intuitif - un ratio compris entre 0 et 1 - :

$$\frac{\mathcal{I}(X;H)}{\min\left(\mathcal{H}(X),\ \mathcal{H}(H)\right)}.$$

En calculant une moyenne sur les informations mutuelles $\mathcal{I}(X_i;H)$ obtenues pour chaque couple enfant-parent $\{X_i;H\}$ d'un MCL, nous obtenons un mesure globale, notée $\mathcal{C}$, de la quantité d'information (exprimée en pourcentage) qu'une VL véhicule sur ses enfants :

$$\mathcal{C} = \frac{1}{n}\sum_i \frac{\mathcal{I}(X_i;H)}{\min\left(\mathcal{H}(X_i),\ \mathcal{H}(H)\right)},$$

où n est le nombre d'enfants du MCL.

4.3.3.6 Pseudocode de CFHLC

Le pseudocode de CFHLC est présenté dans les algorithmes 1 et 2. Les paramètres d'entrée sont au nombre de 7. La taille de la fenêtre s spécifie le nombre de SNP contigus par fenêtre. Le seuil minimal t est employé afin de limiter la perte d'information mesurée par le critère $\mathcal{C}$ mentionné précédemment. Les paramètres a, b et $card_{max}$ servent à calculer la cardinalité de chaque VL à l'aide d'une fonction affine. Finalement, le paramètre *AlgoPartitionnementCliques* introduit une flexibilité dans le choix de la méthode de partitionnement en cliques de variables.

A l'intérieur de chaque fenêtre i, une procédure de classification ascendante hiérarchique adaptée est initialisée à partir d'une première couche de modèles univariés. Un modèle univarié est construit pour chaque VO de la fenêtre courante $\mathscr{W}_i$ (lignes 6 à 8). Un partitionnement en cliques est ensuite réalisé (ligne 12). Chaque clique contenant au moins 2 nœuds est sujette à l'apprentissage d'un MCL (lignes 19 et 20) suivi par sa validation (lignes 23 à 28). De manière à simplifier l'apprentissage de la FMHCL, la cardinalité de la VL est estimée au moyen d'une fonction affine du nombre de nœuds de la clique (ligne 19). L'algorithme *apprentissage_modèle_à_classes_latentes* est intégré dans ce cadre général (ligne 20). Après validation au moyen du seuil t (lignes 23 et 24), le MCL est utilisé afin d'enrichir la FMHCL associée à la fenêtre courante $\mathscr{W}_i$ (ligne 25) : un processus de fusion permet de connecter le nœud additionnel (correspondant à la VL) à ses nœuds enfants, eux-mêmes déjà présents dans la FMHCL en cours de construction ; les distributions *a priori* des enfants sont remplacées par les

distributions conditionnellement à la VL. La VL nouvellement créée, H_{j_k}, est ajoutée à l'ensemble de VL, alors que ses données imputées (pour tous les individus), $\mathscr{D}[H_{j_k}]$, sont sauvegardées (ligne 26). Dans $\mathscr{W}_i$, les variables de $\mathscr{C}\ell_{j_k}$ sont maintenant remplacées par la VL correspondante ; la matrice de données $\mathscr{D}[\mathscr{W}_i]$ est mise à jour en conséquence (ligne 27). En revanche, les nœuds de cliques invalidées sont considérés comme isolés pour la prochaine étape. La construction de la FMHCL est arrêtée lorsque chaque clique identifiée est réduite à un singleton (ligne 13) ou lorsque plus aucune clique de taille supérieure ou égale à 2 n'est validée (ligne 31). Finalement, la collection de forêts (graphes) est successivement augmentée par les forêts construites à l'intérieur de chaque fenêtre (ligne 36). En parallèle, grâce à l'hypothèse d'indépendance entre les fenêtres, la distribution jointe de la FMHCL est simplement calculée comme le produit des distributions associées à chacune des fenêtres (ligne 36).

Algorithme 1 CFHLC

INPUT :
$\mathbf{X}$, un ensemble de p variables observées ($\mathbf{X} = X_1, ..., X_p$),
$\mathscr{D}_{\mathbf{X}}$, les observations correspondantes pour n individus,
$\mathbf{s}$, la taille de la fenêtre,
$\mathcal{C}$, le critère permettant d'estimer la perte d'information lors de la construction de la FMHCL,
$\mathbf{t}$, un seuil employé afin de contraindre la perte d'information (critère $\mathcal{C}$),
$\mathbf{a}, \mathbf{b}$ et $\mathbf{card_{max}}$, les paramètres employés afin d'estimer la cardinalité des VL,
AlgoPartitionnementCliques, un algorithme dédié au partitionnement en cliques de variables.

OUTPUT :
GOSC et θ, le graphe orienté sans circuit et les paramètres de la FMHCL construite, respectivement,
$\mathbf{H}$, l'ensemble de variables latentes identifiées lors de la construction ($\mathbf{H} = \{H_1, ..., H_m\}$),
$\mathscr{D}_{\mathbf{H}}$, les données correspondantes imputées pour les n individus.

$nbw \leftarrow p/s$ /* Calcul du nombre de fenêtres contiguës */
$GOSC \leftarrow \emptyset$; $\theta \leftarrow \emptyset$; $\mathbf{H} \leftarrow \emptyset$; $\mathscr{D}_{\mathbf{H}} \leftarrow \emptyset$

pour $i = 1$ **à** nbw
  /* Construction de la couche 0 */
  $\mathscr{W}_i \leftarrow \{X_{(i-1)\times s+1}, ..., X_{i\times s}\}$; $\mathscr{D}[\mathscr{W}_i] \leftarrow \mathscr{D}[(i-1)\times s + 1 : i \times s)]$
  $\{\cup_{j\in\mathscr{W}_i} GOSC_{univ_j}, \cup_{j\in\mathscr{W}_i}\theta_{univ_j}\} \leftarrow apprentissage\_modèles\_univariés(\mathscr{W}_i)$
  $GOSC_i \leftarrow \cup_{j\in\mathscr{W}_i} GOSC_{univ_j}$; $\theta_i \leftarrow \cup_{j\in\mathscr{W}_i}\theta_{univ_j}$

  étape $\leftarrow 1$
  tant que vrai
    $\{\mathscr{C}\ell_1, ..., \mathscr{C}\ell_{\#c}\} \leftarrow partitionnement(\mathscr{W}_i, \mathscr{D}[\mathscr{W}_i], AlgoPartitionnementCliques)$
    si toutes les cliques $\mathscr{C}\ell_q$ sont des singletons **alors sortir boucle fin si**

    $\{\mathscr{C}\ell_{j_1}, ..., \mathscr{C}\ell_{j_{\#c_2}}\} \leftarrow identification\_cliques\_de\_taille\_strictement\_supérieure\_à\_un(\mathscr{C}\ell_1, ..., \mathscr{C}\ell_{\#c})$
    $nbCliquesValides \leftarrow 0$

    pour k = 1 **à** $\#c_2$
     $card_{VL} \leftarrow min(a \times nombre\_de\_variables(\mathscr{C}\ell_{j_k}) + b, card_{max})$
     $\{GOSC_{j_k}, \theta_{j_k}, H_{j_k}, \mathscr{D}[H_{j_k}]\} \leftarrow apprentissage\_modèle\_à\_classes\_latentes(\mathscr{C}\ell_{j_k}, \mathscr{D}[\mathscr{C}\ell_{j_k}], card_{VL})$

     /* Validation de la clique courante - voir section 4.3.3.5, page 60 */
     si $(\mathcal{C}(GOSC_{j_k}, \mathscr{D}[\mathscr{C}\ell_{j_k}] \cup \mathscr{D}[H_{j_k}]) \geq t)$
       $incr(nbCliquesValides)$
       $GOSC_i \leftarrow fusion\_structures(GOSC_i, GOSC_{j_k})$; $\theta_i \leftarrow fusion\_paramètres(\theta_i, \theta_{j_k})$
       $\mathbf{H} \leftarrow \mathbf{H} \cup H_{j_k}$; $\mathscr{D}_{\mathbf{H}} \leftarrow \mathscr{D}_{\mathbf{H}} \cup \mathscr{D}[H_{j_k}]$
       $\mathscr{D}[\mathscr{W}_i] \leftarrow (\mathscr{D}[\mathscr{W}_i] \setminus \mathscr{D}[\mathscr{C}\ell_{j_k}]) \cup \mathscr{D}[H_{j_k}]$; $\mathscr{W}_i \leftarrow (\mathscr{W}_i \setminus \mathscr{C}\ell_{j_k}) \cup H_{j_k}$
     fin si
    fin pour

    si $(nbCliquesValides = 0)$ **alors sortir boucle fin si**

    $incr(étape)$
  fin tant que

  $GOSC \leftarrow GOSC \cup GOSC_i$; $\theta \leftarrow \theta \times \theta_i$
fin pour

Algorithme 2 apprentissage_modèle_à_classes_latentes

INPUT :
$\mathscr{C}\ell_u$: une clique contenant au moins deux nœuds,
$\mathscr{D}[\mathscr{C}\ell_u]$: les observations correspondantes pour n individus,
$card_{VL}$: la cardinalité de la variable latente à créer.

OUTPUT :
un modèle à classes latentes décrit par :
$GOSC_u$ et θ_u, la structure et les paramètres du modèle à classes latentes, respectivement,
H_u, une variable latente,
$\mathscr{D}[H_u]$, les données imputées pour la variable latente (pour n individus).

1: $H_u \leftarrow création\_variable\_latente()$
2: $GOSC_u \leftarrow construire\_structure\_du\_modèle\_à\_classes\_latentes(H_u, \mathscr{C}\ell_u)$
3: $\theta_u \leftarrow exécuter\_EM(GOSC_u, \mathscr{D}[\mathscr{C}\ell_u], card_{VL})$
4: $\mathscr{D}[H_u] \leftarrow imputer\_données(\theta_u, \mathscr{D}[\mathscr{C}\ell_u])$

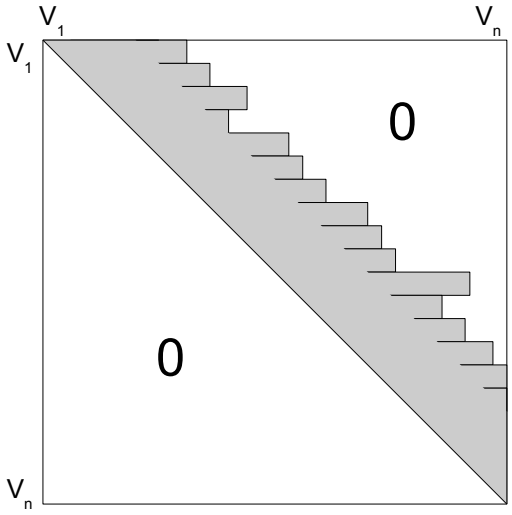

FIGURE 4.12: Exemple de matrice creuse des dépendances entre les variables, calculées dans l'algorithme CFHLC+.
Les dépendances ne sont calculées qu'entre variables V_i (SNP ou variables latentes) séparées par une distance maximale. Les dépendances calculées sont représentées en foncé, tandis que les dépendances non calculées sont représentées par des zéros. Cette matrice est symmétrique, seules les dépendances de la diagonale supérieure (ou inférieure) sont calculées.

4.3.3.7 Algorithme CFHLC+

Nous avons développé une seconde version de l'algorithme CFHLC, nommée CFHLC+, qui ne nécessite plus de découpage préalable du génome pour l'analyse de données pangénomiques. Au lieu de réaliser un découpage en fenêtres qui empêche la modélisation des dépendances existant entre fenêtres adjacentes, une contrainte simple peut être employée : les dépendances ne sont calculées qu'entre variables (SNP ou VL) séparées par une distance physique maximale sur le chromosome. Contrairement aux SNP, les VL ne possèdent pas de position physique sur le chromosome. Pour résoudre ce problème, pour une VL, la moyenne des positions des SNP subsumés est employée. Cette contrainte physique amène à considérer une matrice creuse de dépendances par paires (figure 4.12). Cette matrice creuse présente une bande diagonale dans laquelle les dépendances sont calculées. Dans les zones denses en SNP, plus de dépendences sont calculées entre les variables que dans les zones peu denses. Cela explique pourquoi d'une variable à une autre, la largeur de la bande peut varier. Seules les valeurs de la bande sont stockées en mémoire. CAST, l'algorithme de partitionnement en cliques, a été réimplémenté afin de manipuler de larges matrices creuses.

Cette nouvelle version, plus flexible, peut être vue comme une approche par fenêtre glissante. En fixant une distance maximale suffisamment large (*e.g.* 0.1-5 Mb), il est désormais possible de capturer le LD à longue distance dans le contexte des EAP. En outre, dans cette version, il a été inclus une étape de réinitialisations aléatoires

multiples de l'algorithme EM afin d'augmenter la probabilité de convergence de l'apprentissage de paramètres vers un optimum global. CFHLC+ a été expérimenté pour la visualisation du LD dans le chapitre 5, section 5.3, page 95. Le passage à l'échelle a été étudié dans la section 4.4.6, page 78.

4.4 Résultats expérimentaux et discussion

4.4.1 Implémentation

Les algorithmes CFHLC et CFHLC+ ont été développés en langage C++, à l'aide de la librairie ProBT conçue pour les réseaux bayésiens (http :// bayesian-programming.org). Nous avons intégré à l'intérieur des deux algorithmes une implémentation C++ de l'algorithme CAST en nous basant sur l'implémentation en langage JAVA de Ben Fry (http ://benfry.com/clustering/). En ce qui concerne la visualisation des graphes, le logiciel Tulip (http ://tulip.labri.fr/TulipDrupal/) a été privilégié, de manière à assurer une certaine qualité de représentation et un passage à l'échelle lors de la manipulation de grands graphes (comportant plus de 100 k nœuds).

Pour les expérimentations présentées dans ce chapitre, CFHLC a été exécuté sur un ordinateur personnel standard (3.8 GHz, 3.3 Go de RAM). CFHLC+, quant à lui, a été exécuté sur un serveur Intel Xeon X7460 (2.66GHz, 120 Go de RAM dont seulement quelques Go ont été nécessaires). Le code de CFHLC+ est disponible en version Windows 32 bits à l'adresse Internet https ://sites.google.com/site/-raphaelmouradeng/home/programs.

4.4.2 Protocole expérimental

La performance de notre algorithme a été évaluée sur des données génétiques réelles et simulées, haplotypiques (phasées) et génotypiques (non phasées).

Pour l'analyse des données réelles, le jeu de données bien connu de Daly *et al.* [19] a été employé. Il est disponible à l'adresse Internet http ://www-genome.wi.mit.edu/-humgen/IBD5/index.html. Ce jeu de données se compose de 129 trios (deux parents et un enfant). Pour chaque individu, 103 SNP ont été génotypés dans la région $5q31$ (chromosome 5) et couvrent 617 kb.

En ce qui concerne l'analyse des données simulées, deux programmes ont été employés : HAPGEN et HAPSIMU. A l'aide de HAPGEN (http ://www.stats.ox.ac.uk/ ∼marchini/software/gwas/hapgen.html), nous avons généré des données haplotypiques pour 2000 individus non apparentés (*i.e.* 4000 haplotypes). Les séquences simulées couvrent une région d'une centaine de kb, contenant environ 20-30 SNP. Les haplotypes utilisés comme références proviennent de la deuxième phase du projet HapMap (HapMap II) et concernent les habitants des États-Unis issus d'ancêtres ayant vécu

dans le nord et l'ouest de l'Europe (CEU) (http ://hapmap.ncbi.nlm.nih.gov/). Cinq séquences montrant des degrés de LD variables ($médiane(r^2)$ variant de 0.007 à 0.5) ont été générées.

HAPSIMU (http ://l.web.umkc.edu/liujian/) a été utilisé afin de simuler des données génotypiques dans un contexte plus proche des études d'association pangénomiques, caractérisées par une structure de la population. HAPSIMU emploie comme références les haplotypes CEU et les haplotypes concernant les habitants d'Ibadan au Nigéria (YRI, HapMap II). Nous avons fait varier le nombre de variables observées. Pour cela, nous avons considéré les trois échantillons suivants : 1 k, 10 k et 100 k SNP (2000 individus pour chaque échantillon). Pour chaque échantillon, 10 jeux de données ont été générés.

4.4.3 Modélisation du déséquilibre de liaison

4.4.3.1 Données réelles

En employant le jeu de données de Daly *et al.*, notre objectif est d'évaluer la façon dont la FMHCL obtenue par CFHLC est capable de s'ajuster à la structure réelle des données biologiques. En outre, CFHLC est comparé à quatre autres méthodes de modélisation du LD.

Données haplotypiques *versus* données génotypiques

Deux FMHCL ont été comparées, l'une apprise à partir de données haplotypiques, et l'autre à partir de données génotypiques. Les graphes correspondants sont présentés en figures 4.13 et 4.14, respectivement. Globalement, les deux graphes sont similaires : la majorité des SNP qui sont connectés par une VL dans le graphe des données haplotypiques (GDH) sont aussi connectés par une VL dans le graphe des données génotypiques (GDG), *e.g.* SNP1, SNP4 et SNP6. De plus, une part importante de ces SNP partage un parent commun dans les deux graphes : par exemple, dans le GDH et le GDG, nous observons que SNP61 et SNP65 sont liés par une VL appartenant à la couche 1. Ainsi, en termes de structures, l'apprentissage de la FMHCL à partir de données génotypiques fournit des résultats très proches de celle obtenue lors de l'apprentissage à partir de données haplotypiques, malgré l'incertitude liée au manque de connaissance de la phase gamétique. Cependant, en moyenne, nous observons que les SNP du GDG sont plus connectés : 8 composantes connexes sont seulement identifiées dans le GDH, alors que le GDG comporte 15 composantes connexes. Par exemple, les deux arbres encadrés 1 et 5 du GDH (figure 4.13) sont connectés par une VL de haut niveau dans le GDG, formant l'arbre 1 (figure 4.14).

Structure globale

Nous nous attendons à ce que le graphe de la FMHCL reflète la structure en blocs haplotypiques : de grands blocs de SNP adjacents et corrélés, séparés par des points

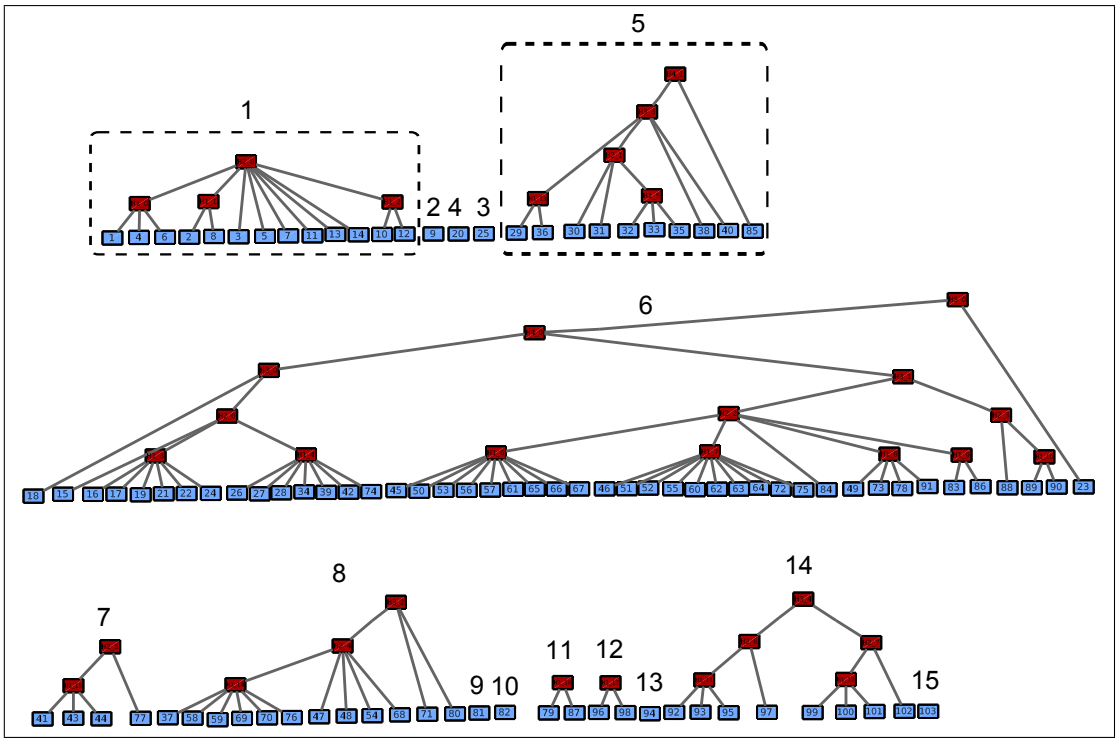

FIGURE 4.13: Graphe de la forêt de modèles hiérarchiques à classes latentes apprise à partir des données haplotypiques de Daly *et al.* [19].
Les variables observées sont numérotées de 1 *à* 103*, alors que les variables latentes sont notées "Hℓ_i" où* ℓ *spécifie le numéro de couche et* i *énumère les variables appartenant à la même couche. Les paramètres de CFHLC sont* $a = 0.2$, $b = 2$, $card_{max} = 20$, $t_{CAST} = 0.95$, $t_{MI} = quantile_{MI}(0.95)$ *et* $t = 0.3$ *(pour une description des paramètres, voir section 4.3.3.6, page 61).*

chauds de recombinaison. Tout d'abord, nous observons sur les deux graphes que la position physique des SNP influence effectivement leur connexion, puisque les SNP proches tendent à être connectés par une VL appartenant à une couche basse, alors que les SNP distants sont généralement liés par une VL de haut niveau. Néanmoins, de fortes dépendances entre SNP distants sont aussi observées, *e.g.* entre SNP26 et SNP74, et entre SNP49 et SNP91 (voir figure 4.13, arbre 6 et figure 4.14, arbre 3). Cette caractéristique révèle que la structure du LD n'est pas exclusivement dominée par des influences spatiales et justifie notre approche de structure en clusters (au lieu de la structure classique en blocs). En outre, les graphes montrent que les clusters identifiés sont cohérents vis-à-vis de la réalité biologique, *i.e.* la variation des taux de recombinaison inférés à l'aide du logiciel PHASE v2.1 [91] le long de la séquence étudiée (voir figure 4.15). En effet, la plupart des sous-arbres racinés dans une VL de bas niveau couvrent des régions présentant un faible taux de recombinaison (TR). 68% et 94% des VL de la couche 1 couvrent des segments chromosomiques montrant des TR en deçà de 4 cM/Mb et 9 cM/Mb, respectivement. La même tendance est observée pour 44% et 66% des VL de la couche 2, respectivement. Ces résultats montrent la

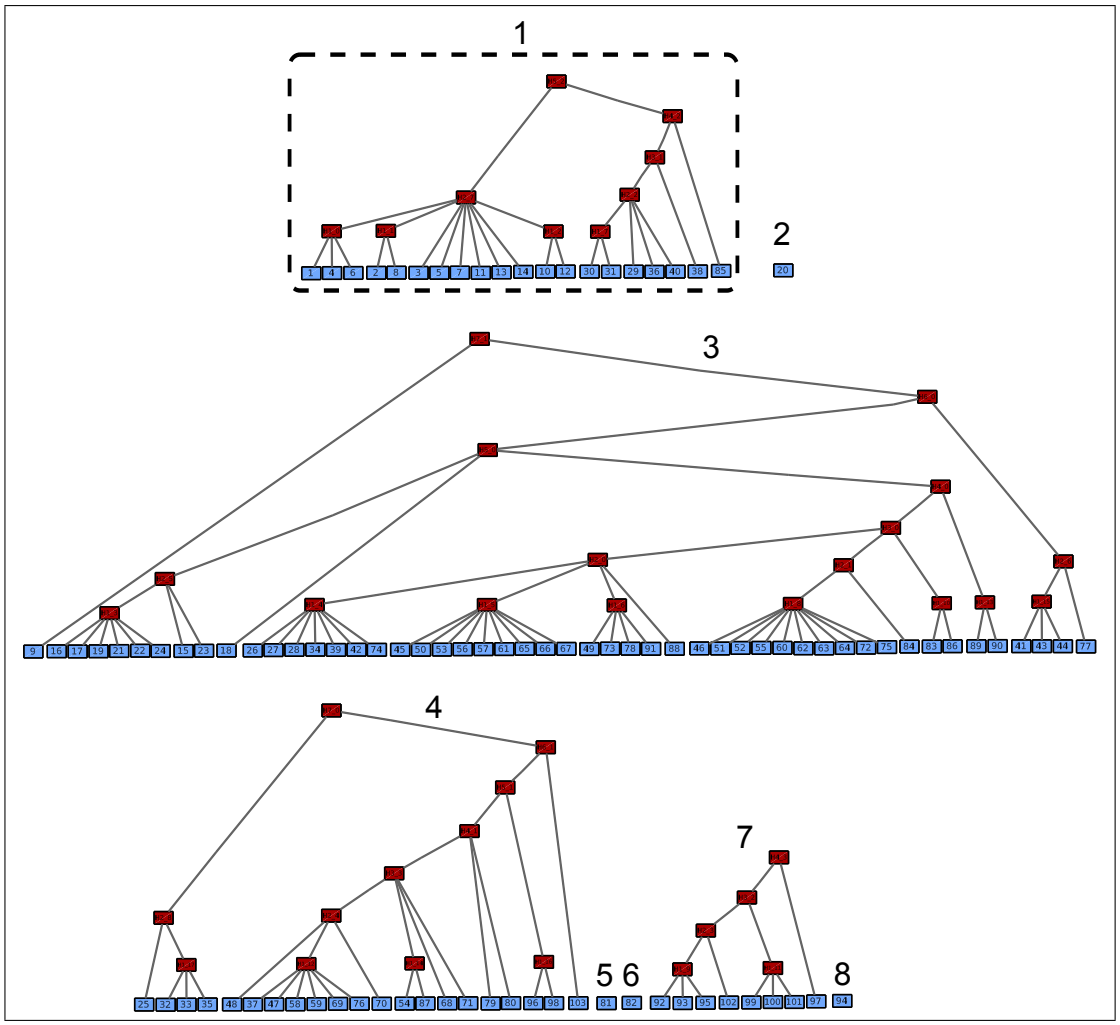

FIGURE 4.14: Graphe de la forêt de modèles hiérarchiques à classes latentes apprise à partir des données génotypiques de Daly *et al.* [19].
Pour la nomenclature des nœuds et les paramètres de CFHLC, voir figure 4.13.

pertinence de l'interprétation, au moins partielle, des VL de bas niveau en tant que clustering d'haplotypes ancestraux, lorsque les données analysées sont des haplotypes (voir section 4.3.1, page 51).

Structure en blocs *versus* structure en clusters

Nous avons comparé la structure obtenue par CFHLC avec les sorties de quatre autres méthodes conçues pour la modélisation du LD : la méthode de Daly *et al.* [19], le logiciel Gerbil [44], le programme HaploBlock [29] et l'algorithme de Zhang *et al.* [116]. La méthode de Daly met en œuvre un modèle de Markov caché. Les autres méthodes sont détaillées dans la section 4.2.1, page 44, dédiée à l'état de l'art des MGP conçus pour la modélisation du LD. Les trois premières méthodes partitionnent la

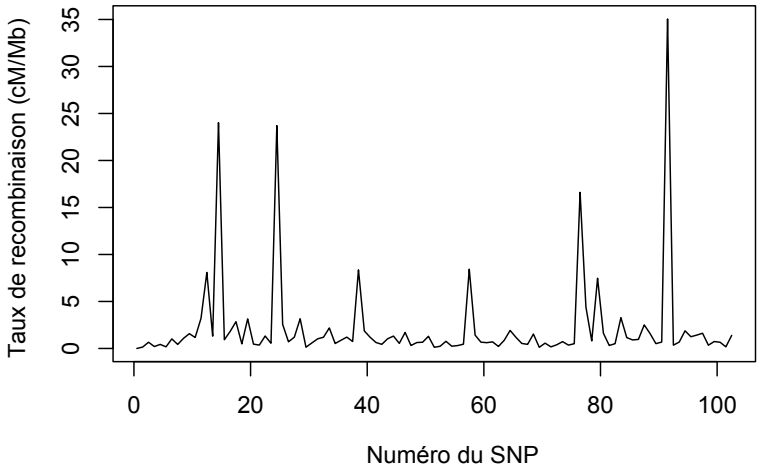

FIGURE 4.15: Taux de recombinaison (cM/Mb) inférés à l'aide du logiciel PHASE v2.1 à partir des données génotypiques de Daly *et al.* [19].

séquence en blocs de SNP contigus, tandis que la dernière méthode construit des clusters (non chevauchants) de SNP non contigus. Nous rappelons que CFHLC génère une classification hiérarchique de SNP non contigus.

Dans la figure 4.16, nous comparons les structures en blocs et en clusters obtenues à l'aide des quatre méthodes mentionnées précédemment et de CFHLC. En dépit du fait que ces méthodes s'attaquent à la modélisation du LD de manières différentes, des similitudes sont observées (voir lignes en pointillés en figure 4.16). Par exemple, le dernier bloc identifié par la méthode de Daly *et al.*, Gerbil et l'algorithme de Zhang *et al.* (ligne 6) est aussi inféré par notre méthode (ligne 31). De légères différences sont observées au niveau des deux premiers blocs résultant de la méthode de Daly *et al.* et Gerbil, qui ne forment qu'un seul bloc pour l'algorithme de Zhang *et al.* (ligne 8) et CFHLC (ligne 15). La majorité des divergences entre notre méthode et les autres résulte de sa capacité à prendre en compte la délimitation floue des clusters. Cette aptitude est bien illustrée par la zone centrale de la séquence (SNP26-SNP74), qui présente deux points chauds de recombinaison faible (entre les SNP39 et SNP40, et entre les SNP58 et SNP59). En effet, on observe que le découpage en clusters dans cette zone centrale n'est pas net et que différents degrés de LD sont présents. Une autre différence avec les autres méthodes consiste en la présence de SNP groupés par les autres méthodes, mais non regroupés par CFHLC (*e.g.* SNP9, SNP20 et SNP25 dans notre modèle).

Comparaison de performances

Les temps d'exécution, les taux de réduction de dimension et les taux de compression d'entropie sont présentés dans le tableau 4.2. Les résultats révèlent que Gerbil est l'algorithme testé le plus rapide, avec un temps d'exécution de 40 *s*. Néanmoins, CFHLC et l'algorithme de Zhang *et al.*, qui apprennent des modèles plus complexes

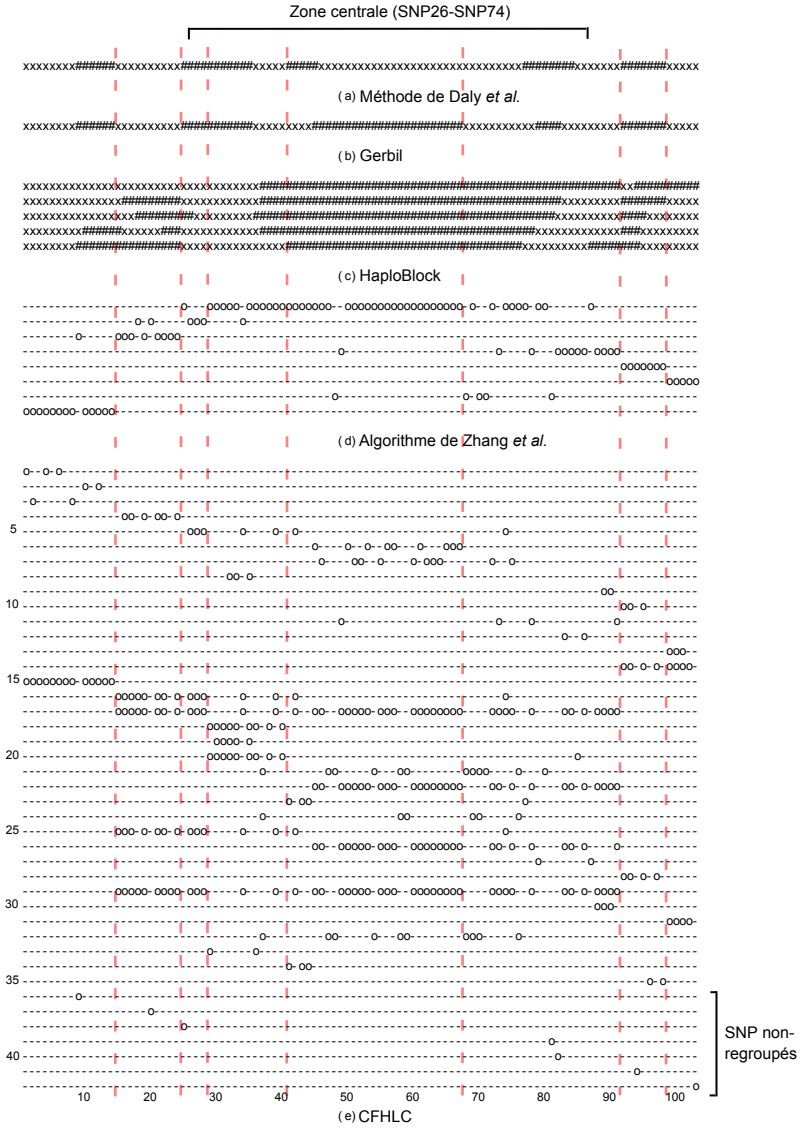

FIGURE 4.16: Comparaison des résultats de cinq méthodes conçues afin de modéliser le déséquilibre de liaison, pour le jeu de données de Daly *et al.* [19].
Partitions de SNP contigus (blocs) inférés par (a) la méthode de Daly et al., (b) le logiciel Gerbil et (c) le programme HaploBlock. La sous-figure (c) présente cinq sorties différentes produites par la méthode non déterministe d'HaploBlock. Les blocs sont représentés par des séquences de x alternant avec des séquences de #. Partitions de SNP non contigus (clusters) inférés par (d) l'algorithme de Zhang et al. et (e) la méthode CFHLC. La sous-figure (d) montre une partition de SNP, alors que la sous-figure (e) présente une hiérarchie. Le symbole o de la $i^{ème}$ ligne et $j^{ème}$ colonne indique que le $j^{ème}$ SNP appartient au $i^{ème}$ cluster. Les lignes en pointillés soulignent les similitudes entre les cinq méthodes. Les paramètres de CFHLC sont : $a = 0.2$, $b = 2$, $card_{max} = 20$, $t_{CAST} = 0.95$, $t_{MI} = quantile_{MI}(0.95)$ et $t = 0.3$ (pour une description des paramètres, voir section 4.3.3.6, page 61).

Algorithme	Temps d'exécution	TRD	TCE
Méthode de Daly *et al.*	–	0.107	0.313
Gerbil	40 s	0.107	0.300
HaploBlock	158 mn	0.066	0.241
Algorithme de Zhang *et al.*	168 s	0.078	0.229
CFHLC	84 s	0.146	0.231

TABLEAU 4.2: Comparaison des temps d'exécution, des taux de réduction de dimension (TRD) et des taux de compression d'entropie (TCE) entre CFHLC et quatre autres approches, pour le jeu de données de Daly *et al.* : méthode de Daly *et al.* [19], logiciel Gerbil [44], programme HaploBlock [29] et algorithme de Zhang *et al.* [116].
Les trois derniers programmes ont été exécutés sur un ordinateur standard. Comme nous n'avions pas accès au programme de Daly et al., nous avons seulement pu comparer le taux de réduction de dimension et le taux de compression d'entropie obtenus avec les autres méthodes.

(*i.e.* clusters de SNP au lieu de blocs), réalisent leur tâche en un temps raisonnable, 84 s et 168 s, respectivement. En comparaison des autres méthodes, HaploBlock est la méthode la plus lente, avec un temps d'exécution de 155 mn, dû à la grande complexité de la méthode d'apprentissage basée sur la génétique des populations.

Pour ces trois méthodes de partitionnement en blocs, nous définissons le taux de réduction de dimension (TRD) comme le ratio du nombre de blocs sur le nombre de SNP. En ce qui concerne l'algorithme de Zhang *et al.*, le TRD est défini comme le ratio du nombre de clusters sur le nombre de SNP. Dans le cas de CFHLC, nous considérons que l'information de chaque arbre de la FMHCL peut être synthétisée par sa racine, apportant ainsi la meilleure réduction de dimension. Ainsi, le TRD est défini comme le nombre de racines dans la forêt entière divisé par le nombre de SNP. HaploBlock génère le plus petit nombre de blocs avec une moyenne de 6.8 (TRD de 0.066), alors que l'algorithme de Zhang *et al.* partitionne la séquence en 8 clusters (TRD de 0.078), et la méthode de Daly *et al.* et Gerbil identifient tous les deux 11 blocs (TRD de 0.107). CFHLC présente la plus faible réduction de dimension avec 15 arbres (TRD de 0.146), due à la présence de 7 SNP non regroupés : SNP9, SNP20, SNP25, SNP81, SNP82, SNP94 et SNP103.

Comme alternative à la mesure de réduction de dimension, nous définissons un taux de compression d'entropie (TCE) comme le ratio de la somme des entropies des blocs (ou clusters) dans la partition à l'entropie sous hypothèse d'absence de structure (*i.e.* la somme des entropies individuelles des SNP). Nous observons maintenant une différence de classement entre les méthodes. Nous remarquons que CFHLC et l'algorithme de Zhang *et al.*, qui apprennent tous les deux des modèles de clusters, fournissent les meilleures (*i.e.* les plus basses) valeurs de TCE (autour de 0.23), alors que HaploBlock, Gerbil et la méthode de Daly *et al.* montrent des valeurs de TCE de 0.241, 0.3 et 0.313, respectivement. La supériorité des résultats de TCE obtenus par les modèles de clusters s'explique par l'absence d'une contrainte sur la proximité physique des SNP, présente dans les modèles de blocs. En outre, le critère TCE ne

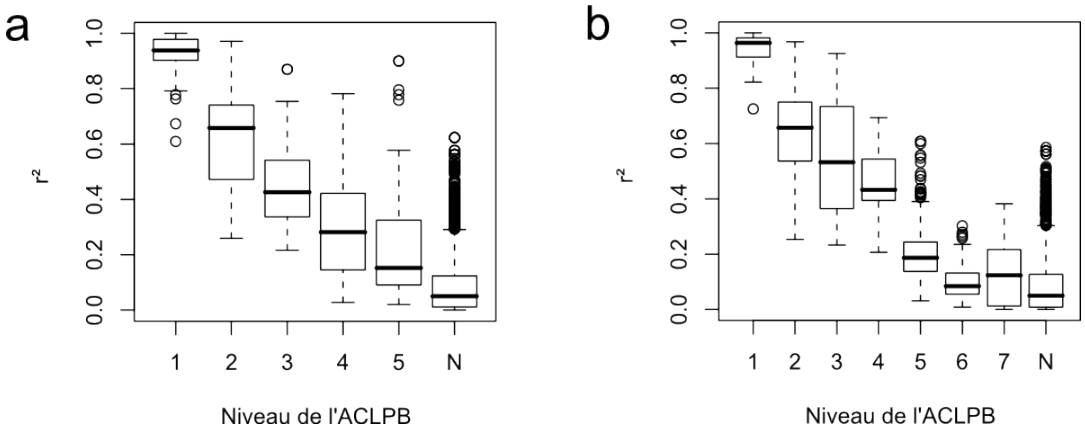

FIGURE 4.17: Relation entre le coefficient de corrélation au carré r^2 pour chaque paire de SNP et le niveau de leur ancêtre commun le plus bas (ACLPB) dans la forêt, pour le jeu de données de Daly *et al.* (a) Données haplotypiques, (b) Données génotypiques.
N indique la situation où les deux SNP considérés n'appartiennent pas au même arbre. Les paramètres de CFHLC sont $a = 0.2$, $b = 2$, $card_{max} = 20$, $t_{CAST} = 0.95$, $t_{MI} = quantile_{MI}(0.95)$ *et* $t = 0.3$ *(pour une description des paramètres, voir section 4.3.3.6, page 61).*

pénalise pas la méthode CFHLC, puisque les SNP non regroupés contribuent relativement peu à la quantité d'information totale.

Niveau des variables latentes

Dans la section 4.3.1, page 51, nous avions avancé que les couches de la FMHCL pourraient décrire différents degrés de LD. Afin de le vérifier, nous analysons la relation entre le coefficient de corrélation au carré r^2 pour chaque paire de SNP et le niveau de leur ancêtre commun le plus bas (ACLPB) dans la forêt. Notons qu'ici l'ACLPB est une notion de la théorie des graphes. Elle ne présente pas de lien avec la notion d'ACLPB utilisée en phylogénie. Nous aurons d'ailleurs l'occasion d'utiliser à nouveau le concept d'ACLPB au chapitre 5, section 5.3.3.1, page 101. Les figures 4.17a et 4.17b montrent cette relation pour des données haplotypiques et génotypiques, respectivement. Partant de valeurs élevées $[0.9 - 1.0]$, r^2 diminue de façon presque linéaire lorsque le niveau de l'ACLPB augmente. Ainsi, nous concluons que la structure en couches de la FMHCL reflète fidèlement les différences de degré de LD. Ces résultats nous amènent à comparer visuellement la carte triangulaire de chaleur du r^2 par paires de SNP à la carte triangulaire de chaleur du niveau de l'ACLPB pour des données haplotypiques ou génotypiques (voir figure 4.18). Dans ce but, le code couleur employé dans la carte triangulaire de chaleur du r^2 (CTCR) est réutilisé pour la carte triangulaire de chaleur du niveau (CTCN) de l'ACLPB, de la manière suivante. Dans la CTCR, l'intensité de la couleur de chaque cellule varie en fonction de la valeur du r^2. Comme nous pouvons calculer la valeur médiane du r^2 pour chaque niveau de l'ACLPB, il est possible d'attribuer une couleur à chaque cellule de la CTCN, en réemployant la même échelle de couleur que celle de la CTCR.

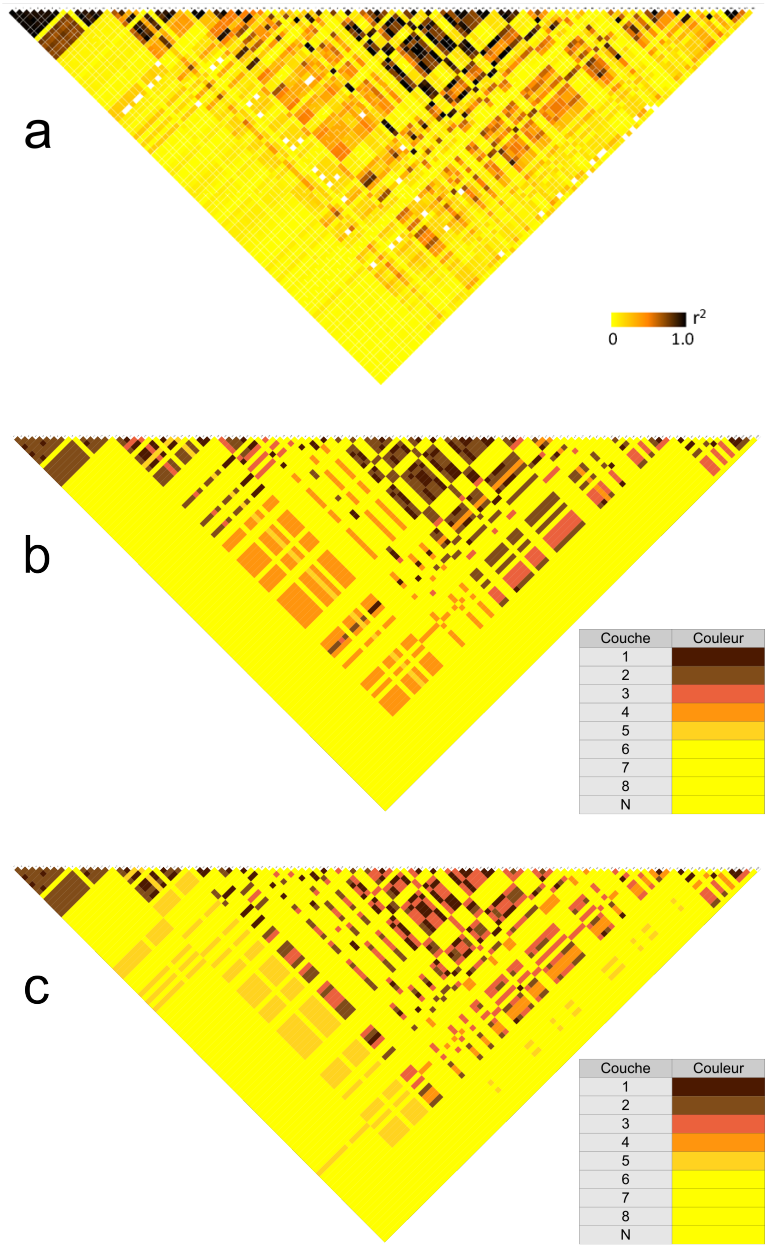

FIGURE 4.18: Carte triangulaire de chaleur du r^2 par paires de SNP *versus* carte triangulaire de chaleur du niveau de l'ancêtre commun le plus bas (ACLPB) dans la forêt, pour le jeu de données de Daly *et al.*. (a) Carte triangulaire de chaleur du r^2 par paires de SNP (cette carte provient de Réf.[116]). (b) Carte triangulaire de chaleur du niveau de l'ACLPB (données haplotypiques). (c) Carte triangulaire de chaleur du niveau de l'ACLPB (données génotypiques). *Pour les détails concernant N et les paramètres de CFHLC, voir figure 4.17.*

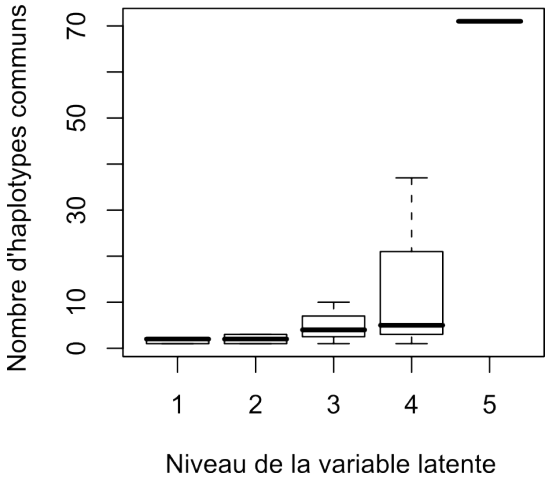

FIGURE 4.19: Relation entre le nombre d'haplotypes communs et le niveau de la variable latente (VL), pour le jeu de données de Daly *et al.*
Pour chaque VL, les haplotypes observés sont définis par les variables observées (VO), i.e. les valeurs des feuilles de l'arbre enraciné par la VL. L'ensemble des haplotypes communs est défini comme le plus petit sous-ensemble d'haplotypes observés couvrant au moins 75% *de l'échantillon. La diversité haplotypique est mesurée par le nombre médian d'haplotypes communs observés au niveau ℓ. Pour les détails concernant les paramètres de CFHLC, voir figure 4.17.*

La correspondance entre la CTCR et la CTCN montre clairement la capacité de la FMHCL à modéliser de manière précise les différents degrés de LD : la majorité des dépendances à l'intérieur de la CTCR sont présentes dans la CTCN, et avec une intensité comparable. En outre, nous remarquons que la modélisation des données génotypiques donne des résultats très proches de celle obtenue avec des données haplotypiques.

La diversité haplotypique est généralement très faible à l'intérieur des blocs (et *a fortiori* à l'intérieur des clusters) haplotypiques. En outre, dans notre modèle hiérarchique, la diversité haplotypique devrait être plus élevée à l'intérieur d'un cluster dont le niveau de la VL le subsumant est élevé. Afin de vérifier ceci, nous nous sommes basés sur les clusters identifiés en figure 4.16(e). Pour une VL de niveau ℓ, la diversité haplotypique est calculée comme le nombre d'haplotypes communs observés dans les feuilles de l'arbre enraciné dans cette VL. La figure 4.19 montre que la diversité haplotypique médiane demeure très basse (en-dessous de 6) pour les quatre premières couches et augmente fortement jusqu'à environ 70 pour la cinquième couche. Ce résultat confirme que la diversité haplotypique augmente avec le niveau de la VL.

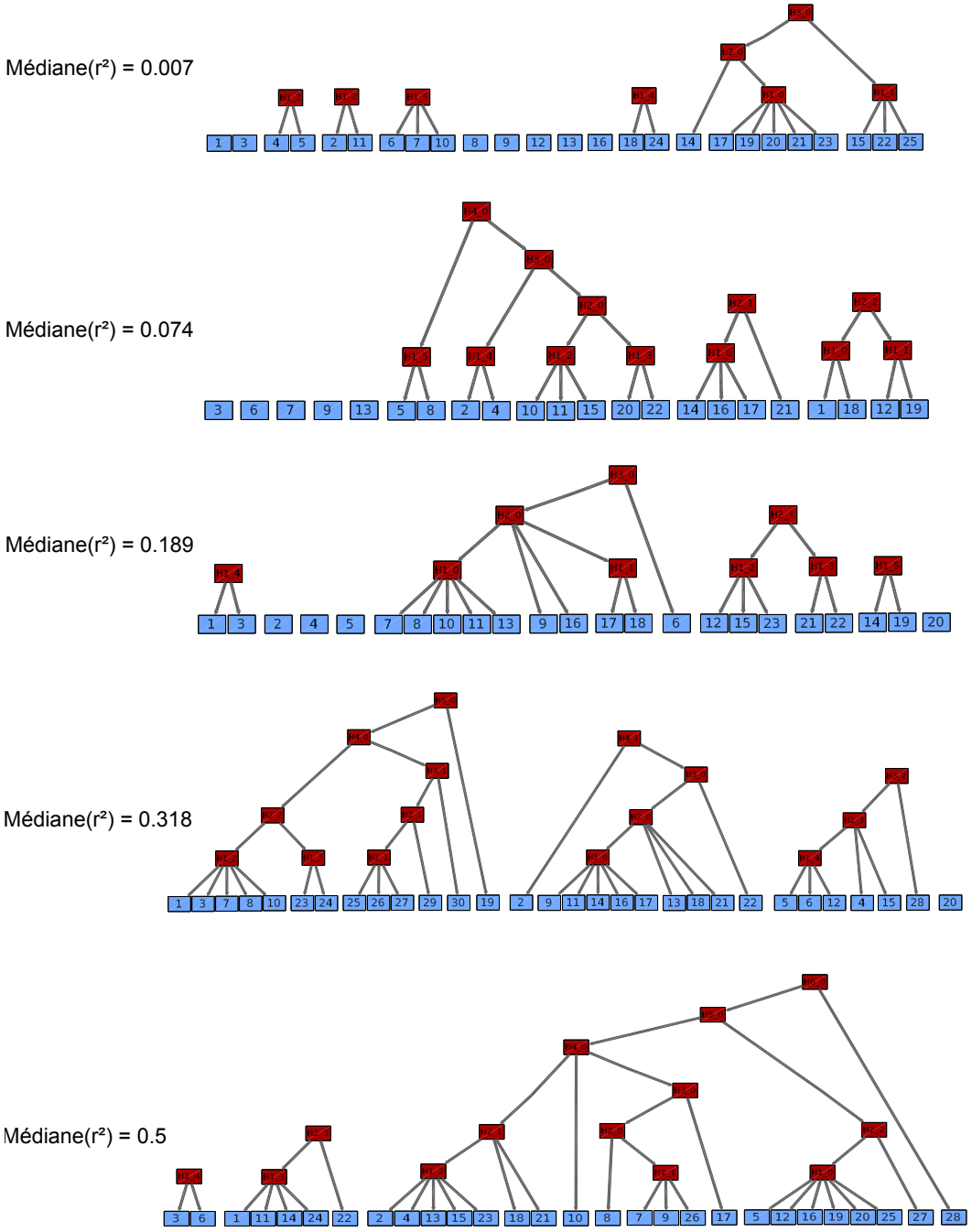

FIGURE 4.20: Influence du degré de déséquilibre de liaison sur la structure de la forêt de modèles hiérarchiques à classes latentes.
Cinq séquences montrant des degrés de LD variables ont été employées afin d'apprendre les forêts de modèles hiérarchiques à classes latentes. Les paramètres de CFHLC sont : $a = 0.2$, $b = 2$, $card_{max} = 20$, $t_{CAST} = 0.95$, $t_{MI} = quantile_{MI}(0.95)$ *et* $t = 0.6$ *(pour une description des paramètres, voir section 4.3.3.6, page 61).*

4.4.3.2 Données simulées

Finalement, l'impact de la variation du degré de LD a été étudié. Pour cela, nous avons généré des données haplotypiques à l'aide du logiciel HAPGEN. Cinq séquences, montrant des degrés de LD variables ($médiane(r^2)$) allant de 0.007 à 0.5, ont été employées pour l'apprentissage des FMHCL. La figure 4.20 montre les forêts obtenues. Les forêts révèlent que l'augmentation du degré de LD entraîne une plus grande connectivité ainsi qu'un plus grand nombre de couches. En effet, lorsque la médiane du coefficient r^2 est égale à 0.007, 11 composantes connexes sont identifiées et la plus haute VL appartient à la troisième couche. A l'inverse, dans le cas où la médiane du coefficient r^2 est égale à 0.5, la forêt est seulement composée de 3 composantes connexes et la plus haute VL appartient à la sixième couche. Ainsi, nous concluons que la méthode CFHLC peut traiter des séquences présentant des degrés de LD variés et générer des structures reflétant fidèlement le LD.

4.4.4 Passage à l'échelle de CFHLC pour des études d'association pangénomiques

Le passage à l'échelle a été étudié à l'aide de données simulées par HAPSIMU (voir annexe A.1, page 117). Dans le cas le plus problématique (100 k SNP), seulement 15 heures sont nécessaires lorsque la taille de la fenêtre s est fixée à 100. Pour le même jeu de données, dans les cas où $s = 200$ et $s = 600$, les temps d'exécution sont de 20.5 h et 62.5 h, respectivement. En ce qui concerne le cas 10 k SNP, les temps d'exécution sont de 1.3 h, 2 h et 5.8 h pour $s = 100$, $s = 200$ et $s = 600$, respectivement.

4.4.5 Analyse de CFHLC

Un ensemble d'expérimentations sur l'algorithme CFHLC est présenté en annexes A.2 à A.9. L'annexe A.2, page 118, souligne l'effet de la taille de la fenêtre sur le temps d'exécution. En annexe A.3, page 119, la réduction de dimension des données est évaluée comme le nombre de VL à l'intérieur de chaque couche de la forêt. Les annexes A.4 à A.7 (pages 120 à 123) montrent l'influence de la taille de la fenêtre sur le nombre de racines de la forêt, le nombre de VL, le nombre de couches et la distribution des VL entre les couches, respectivement. Les annexes A.8 et A.9 (pages 124 et 125) analysent comment l'information diminue lorsque le numéro de la couche augmente.

Un résultat important est que CFHLC peut assurer une réduction de dimension des données de plus de 80% (voir annexe A.4, page 120). En ce qui concerne la complexité spatiale de la méthode, la FMHCL entière n'est jamais stockée dans la RAM, puisque chaque fenêtre peut être sauvegardée indépendamment sur le disque dur.

4.4.6 Passage à l'échelle de CFHLC+

Les variations de la complexité temporelle empirique de CFHLC+ en fonction du nombre de SNP traités et de la taille de la fenêtre sont montrées en annexes B.1 et B.2, respectivement. D'une part, nous observons que la complexité temporelle empirique de CFHLC+ est presque linéaire avec le nombre de SNP traités. D'autre part, elle est aussi proche de la linéarité avec la taille de la fenêtre considérée.

4.5 Conclusion

La contribution de ce travail est multiple : (i) un nouveau cadre d'analyse du déséquilibre de liaison (linkage disequilibrium, LD) a été proposé, et testé sur des données simulées et réelles. Ce cadre s'est révélé pertinent pour modéliser finement les dépendances entre SNP et pour réduire la dimension des données ; (ii) CFHLC, un algorithme conçu pour l'apprentissage de la forêt de modèles hiérarchiques à classes latentes (FMHCL), a été développé et s'est montré efficace pour l'analyse de données pangénomiques.

En comparaison des travaux de Verzilli *et al.* [107], notre méthode présente l'avantage d'offrir une modélisation plus précise du LD grâce à un modèle hiérarchique, ainsi qu'une synthèse de l'information génétique au travers des variables latentes (VL). De plus, en comparaison avec l'algorithme de Zhang et Ji [116], notre approche ne nécessite pas la spécification *a priori* du nombre de clusters, et peut de surcroît capturer différents niveaux de dépendance. A l'heure actuelle, elle est la seule permettant de prendre en compte la nature floue du LD. Du point de vue de l'apprentissage de la FMHCL, notre algorithme est le premier à permettre de traiter des données de grandes dimensions, tout en ne restreignant pas l'inférence du modèle à des arbres binaires et à des variables binaires, comme le fait l'approche de Hwang *et al.* [37]. L'étude sur des données simulées a pu montrer que notre algorithme est capable de refléter le LD dans un grand nombre de situations allant d'un faible à un fort LD global.

Bien que notre algorithme ait été conçu dans le cadre des EAP, nous devons souligner qu'il peut aussi s'appliquer à tout type de données présentant des dépendances spatiales, en particulier pour des données séquentielles. C'est la raison pour laquelle il serait intéressant d'évaluer la capacité de notre algorithme à pouvoir être appliqué à plusieurs problématiques différentes.

Dans le contexte des études pangénomiques, les applications de notre méthode sont multiples :

– Grâce à sa structure hiérarchique latente, la FMHCL pourrait être employée comme un outil simple et efficace pour la recherche de relations de causalité. En effet, les couches multiples de la FMHCL représentent différents degrés de

synthèse de l'information (*i.e.* réduction de dimension). Or la connexion entre les variables du modèle par relation de subsomption permet de réaliser un parcours depuis les causes générales (couches hautes) jusqu'aux causes spécifiques (couches basses), et *vice versa*. D'autre part, la structure d'indépendance conditionnelle encodée par le modèle permet de construire des tests d'indépendance conditionnelle afin d'isoler les causes les unes des autres. Ainsi, la découverte de causalité pourrait être guidée par la structure du modèle.

– La nature graphique et hiérarchique de la FMHCL permettrait le développement d'outils de visualisation intuitive et synthétique des dépendances entre SNP. Le LD par paires pourrait être perçu à travers le niveau de l'ancêtre commun le plus bas reliant deux SNP dans la forêt. En effet, nous avons observé que ce niveau est corrélé au LD par paires (voir figure 4.17, section 4.4.3, page 73). Le LD multilocus, quant à lui, peut être interprété comme un ensemble de SNP subsumés par une VL. En outre, il est possible de calculer une mesure de LD multilocus à l'aide de la distribution jointe de la FMHCL.

– Les VL de haut niveau sont supposées représenter l'effet de la structure de la population sur le LD de longue distance. Si cela était confirmé, il serait intéressant d'employer ces variables afin de mieux comprendre les mélanges de sous-populations.

Lors de ce travail de thèse de doctorat, nous avons choisi de poursuivre le premier et le second axe de recherche appliquée, *i.e.* la découverte de causalité et la visualisation du LD à l'aide de FMHCL. Les travaux réalisés sont présentés dans le chapitre suivant.

4.6 Perspectives

De nombreuses perspectives sont envisageables afin de poursuivre ces travaux de modélisation du LD, notamment au niveau de l'apprentissage du modèle.

Les méthodes basées sur le clustering spectral [109] représentent probablement une approche bien plus puissante pour le partitionnement de variables en cliques que celle implémentée dans l'algorithme CAST. Le clustering spectral a connu un grand succès grâce à sa simplicité de mise en œuvre et à son efficacité pour regrouper les variables corrélées entre elles, étant donnée une matrice de similarité par paires de variables. Une approche basée sur la régularisation, telle que le graphical Lasso [26], offre une autre alternative. Elle consiste à inférer une matrice creuse des dépendances entre variables. Par comparaison à la méthode employée dans CFHLC, l'un des avantages du graphical Lasso est qu'il utilise une mesure de dépendance, le coefficient de corrélation partielle, qui possède une propriété plus forte que l'information mutuelle

par paires de variables. En effet, le coefficient de corrélation partielle mesure l'indépendance conditionnelle, tandis que l'information mutuelle par paires de variables mesure l'indépendance marginale. Pour l'instant, cette approche n'a été appliquée qu'aux données d'expression génique. La dimension des données d'expression génique est d'un ordre de grandeur important (*e.g.* 100 individus, 5000 variables) mais qui demeure beaucoup plus faible que celui des données d'EAP.

L'algorithme CFHLC réalise un apprentissage des paramètres de manière locale en utilisant les paramètres appris pour chaque MCL constituant la FMHCL (voir section 4.3.3.6, page 61). L'ajout d'une dernière étape d'apprentissage global de la FMHCL devrait permettre de disposer d'une meilleure modélisation des dépendances entre variables. Bien que cet apprentissage global par EM soit très coûteux en temps de calcul, il est possible d'initialiser les paramètres avec ceux appris localement, ce qui devrait assurer une convergence de l'algorithme bien plus rapide.

Dans notre modélisation, nous ne nous sommes pas intéressés à inférer la phase gamétique des données génotypiques. Nous avons postulé que nous la connaissions *a priori*. Néanmoins, une extension éventuelle de notre modèle pourrait consister à inférer la phase. Par exemple, les variables latentes de bas niveau pourraient représenter des paires d'haplotypes ancestraux paternels et maternels.

Dans la problématique plus générale de modélisation du LD à l'aide de modèles graphiques probabilistes (non nécessairement FMHCL), une piste prometteuse consisterait à poursuivre les travaux de Verzilli *et al.* [107]. En effet, leur approche présente l'avantage d'être très rapide et efficace pour traiter des données pangénomiques. Elle consiste à générer à l'aide d'un échantillonneur de Gibbs et selon un schéma de MCMC un ensemble de réseaux de Markov. En outre, elle intègre une recherche performante d'associations avec la maladie selon un cadre bayésien qui est très avantageux pour les EAP [90], car il permet l'intégration simple de connaissances *a priori* telles que l'espérance du nombre de SNP causaux. Malheureusement, le schéma MCMC proposé est très restrictif : il consiste à ne parcourir que les graphes composés d'un ensemble de cliques disjointes, afin d'assurer la décomposabilité du graphe, un point essentiel qui assure un calcul simple de la vraisemblance. Cependant, avec le développement récent des algorithmes de Thomas et Green [103, 104], il est désormais possible d'échantillonner directement dans la classe générale des graphes décomposables. Ainsi, la réimplémentation de la méthode de Verzilli *et al.* en intégrant la méthode de Thomas et Green pour l'échantillonnage de graphes décomposables mériterait d'être envisagée.

5 Applications

Sommaire

5.1 Introduction

Ce chapitre aborde deux applications de la modélisation du LD proposée dans le troisième chapitre : (i) la recherche de causalités et (ii) la visualisation synthétique et intuitive du LD dans des contextes variés. La partie (i) se focalise sur une étude systématique de la capacité des FMHCL à la recherche de causalités et est illustrée dans le contexte de la génétique d'association. Ce travail établit les bases du développement de nouvelles méthodes dédiées à la découverte de SNP causaux pour les études d'association pangénomiques. La partie (ii) propose une méthode originale pour la visualisation synthétique et intuitive du LD adaptée aux trois principales situations que peut rencontrer le généticien : la visualisation du LD de courte distance, de longue distance et dans un contexte pangénomique. Les perspectives de recherche sont discutées, notamment celles évoquant la nécessité de développer un logiciel intégré équipé d'une interface graphique conviviale afin d'apprendre les FMHCL, de les visualiser et de lancer des analyses d'association génétique.

5.2 Recherche de causalités : illustration sur le cas de la génétique d'association

5.2.1 Introduction

Les FMHCL sont naturellement amenées à être employées comme outils simples et efficaces pour la recherche de relations de causalité grâce à leur structure hiérarchique latente. Un point important à comprendre est le fait que l'orientation des arcs dans les FMHCL ne représente pas une relation de causalité entre les variables. Par contre, la structure de dépendance encodée par la FMHCL permet de construire des tests d'indépendance conditionnelle afin d'isoler les causes les unes des autres. D'autre part, les multiples couches des FMHCL regroupent différents degrés de synthèse de l'information (*i.e.* différents degrés de réduction de dimension). Or, grâce aux relations de subsomption, la connexion entre les variables du modèle permet de réaliser aisément un parcours depuis les causes générales (couches hautes) jusqu'aux causes spécifiques (couches basses), et *vice versa*. Ainsi la découverte de causalité peut être guidée par la structure du modèle.

Dans les études d'association pangénomiques, l'objectif est d'identifier la combinaison de facteurs génétiques impliqués dans l'apparition de la maladie génétique complexe. L'analyse des données issues de ces études consiste à trouver les SNP impliqués dans la maladie et peut être vue comme un problème de découverte de causalité. L'approche naïve est de tester l'association entre chaque SNP et la maladie, mais cela implique la réalisation d'un grand nombre de tests. Ceci a pour conséquence d'engendrer un fort taux de faux positifs, ainsi qu'une perte de puissance lors de la recherche des SNP causaux [5]. Pour résoudre ces problèmes, un grand nombre d'approches a été développé : les tests d'association haplotypique (les haplotypes étant préalablement inférés) [81] et les tests d'association réalisés avec des blocs de SNP [73] ou avec

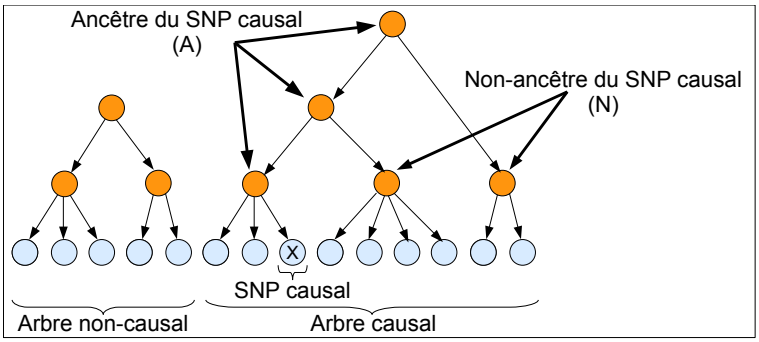

FIGURE 5.1: Illustration des termes spécifiques à notre approche de découverte de causalité : ancêtre du SNP causal, non-ancêtre du SNP causal, arbre causal et arbre non causal.
Les SNP et les variables latentes sont représentés par des nœuds bleus (clairs) et rouges (foncés), respectivement.

des tagSNPs [32] (pour plus de références, voir [51]). Des méthodes moins classiques, provenant des domaines de l'intelligence artificielle et de l'apprentissage automatique, représentent des solutions prometteuses, notamment grâce à leur capacité à traiter des données volumineuses. Dans ce contexte, les méthodes basées sur les MGP ont rencontré un certain succès [33, 55, 84, 99, 107].

Dans le chapitre précédent, les FMHCL ont été employées afin de modéliser le LD à partir de données des EAP. La découverte de causalités dans ce contexte apparaît comme un bon exemple d'application. Mais avant d'employer les FMHCL pour la découverte de SNP causaux, il est judicieux de démontrer au préalable la capacité de la FMHCL apprise par CFHLC pour la découverte de causalité, et plus particulièrement, pour la capture d'associations génétiques indirectes (AGI), que nous définissons ci-après. Dans cet objectif, nous avons réalisé une étude systématique sur des données génétiques réelles et simulées visant à étudier la capacité des VL de la FMHCL à la capture d'AGI. Lors de cette étude préliminaire, nous nous sommes concentrés sur le cas le plus simple : un seul SNP causal influence la maladie. Sous l'hypothèse d'additivité des effets des SNP causaux (*i.e.* interactions SNP-SNP inexistantes), les résultats obtenus peuvent être aisément généralisés au cas où plusieurs SNP sont impliqués dans l'apparition de la maladie.

Nous nommons association génétique indirecte, une dépendance statistique entre un nœud ancêtre du SNP causal (en abrégé, A) de la FMHCL et la maladie. Cette dépendance s'explique par le fait qu'un A a tendance à capturer l'information du SNP causal qu'il subsume (*i.e.* dont il est ancêtre dans le graphe). La recherche d'association génétique indirecte constitue la clé de notre méthode de découverte de causalité dans les études de génétique d'association : l'identification des A permet de cibler les marqueurs causaux potentiels, puisque ces derniers sont les feuilles des arbres enracinés dans les A (voir figure 5.1 qui clarifie la signification des termes spécifiques employés).

5.2.2 Matériel et méthodes

5.2.2.1 Protocole expérimental

Afin d'évaluer la capacité des VL à capturer les AGI, nous avons simulé des données génotypiques (phase inconnue) et phénotypiques pour différents scénarios. En particulier, nous avons fait varier la fréquence de l'allèle mineur (FAM) du SNP simulé causal, le risque relatif génotypique (RRG) et le modèle de la maladie. Nous avons répliqué chaque scénario 100 fois. Des données ont été simulées avec le logiciel HAPGEN (http ://www.stats.ox.ac.uk/∼marchini/software/gwas/hapgen.html) en employant comme référence les haplotypes provenant de la deuxième phase du projet HapMap et concernant les habitants des États-Unis issus d'ancêtres ayant vécu dans le nord et l'ouest de l'Europe (CEU) (http ://hapmap.ncbi.nlm.nih.gov/). Les données génotypiques ont été générées pour 1000 contrôles (individus sains) et 1000 cas (individus malades), et couvrent une région de 1.5 Mb contenant 100 SNP. Parmi les SNP générés, HAPGEN simule un SNP causal du phénotype (cas/contrôle). L'intervalle de variation de la FAM du SNP causal est spécifié à [0.1-0.2], [0.2-0.3] ou [0.3-0.4]. Différents RRG sont considérés : 1.4, 1.6 ou 1.8. Le modèle de la maladie est soit dominant, récessif, additif ou multiplicatif. La combinaison des conditions précédentes amène à tester $3 \times 3 \times 4$ scénarios. Un contrôle qualité de routine a été appliqué sur les données génotypiques : les SNP ayant une FAM inférieure à 0.05 et les SNP déviant de l'équilibre de Hardy-Weinberg [1] avec une p-value inférieure à 0.001 ont été supprimés.

5.2.2.2 Évaluation des associations génétiques indirectes

Afin d'évaluer l'association entre une variable X de la FMHCL (VO ou VL) et le phénotype Y, nous avons employé des tests d'indépendance. Nous n'avons pas utilisé de tests classiquement utilisés en génétique d'association, comme le test de tendance de Cochran-Armitage, car ces derniers ne peuvent être appliqués à la recherche d'association avec les VL. En effet, pour l'étude de données génotypiques, les VL sont des variables discrètes dont les états ne représentent pas la composition allélique des SNP mais plutôt des clusters probabilistes de génotypes multilocus (voir chapitre 4, section 4.3.1, page 51). Ainsi, il n'est pas possible de distinguer entre effet additif, dominant, récessif ou multiplicatif du facteur génétique.

C'est la raison pour laquelle nous avons choisi un test d'indépendance non spécifique. Nous avons choisi d'employer le test du G^2 que nous détaillons ci-après. Soit deux variables qualitatives X et Y pour lesquelles nous souhaitons savoir si elles sont dépendantes ou indépendantes. Posons les hypothèses statistiques :

$$\begin{array}{ll} -H_0 : P(X,Y) = P(X)\ P(Y) & \textit{indépendance,} \\ -H_1 : P(X,Y) = P(X|Y)\ P(Y) = P(Y|X)\ P(X) & \textit{dépendance.} \end{array}$$

1. L'équilibre de Hardy-Weinberg postule qu'il y a un équilibre de la fréquence des allèles et des génotypes au cours des générations.

	Y_1	$\cdots$	Y_j	$\cdots$	Y_p	total
X_1	N_{11}	$\cdots$	N_{1j}	$\cdots$	N_{1p}	$N_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
X_i	N_{i1}	$\cdots$	N_{ij}	$\cdots$	N_{ip}	$N_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
X_q	N_{q1}	$\cdots$	N_{qj}	$\cdots$	N_{qp}	$N_{q.}$
total	$N_{.1}$	$\cdots$	$N_{.j}$	$\cdots$	$N_{.p}$	$N_{..}$

TABLEAU 5.1: Table de contingence $\mathcal{T}$.

La statistique du G^2 est calculée à partir de la table de contingence $\mathcal{T}$ sur les données observées $\{X, Y\}$ (voir tableau 5.1). La formule du G^2 est la suivante :

$$G^2 = 2 \sum_{ij} N_{ij} \ln \left(\frac{f_{ij}}{f_{i.}\ f_{.j}} \right),$$

avec $f_{ij} = \frac{N_{ij}}{N_{..}}$, $f_{i.} = \frac{N_{i.}}{N_{..}}$ et $f_{.j} = \frac{N_{.j}}{N_{..}}$.

Sous l'hypothèse H_0, le G^2 tend asymptotiquement vers une loi du χ^2 à $(q-1) \times (p-1)$ degrés de liberté. La probabilité de rejeter H_0 (et donc d'accepter H_1) alors que H_0 est vraie mesure le risque de se tromper en considérant que X et Y sont dépendants. Cette probabilité est nommée p-value. Plus la p-value faible, plus nous pouvons conclure que les deux variables sont dépendantes.

Notre choix du test du G^2 au lieu d'un test plus connu comme le test du χ^2 s'explique par le fait que le premier correspond au test du ratio de vraisemblance [2] (TRV), tandis que le second est seulement une approximation du TRV. Pour des échantillons de taille raisonnable, les tests du G^2 et du χ^2 fournissent les mêmes résultats. Mais pour des échantillons de petite taille (inférieure à 300 individus), comme c'est le cas pour le jeu de données réelles analysées, de grandes différences sont attendues lorsque des p-values faibles sont considérées.

Nous avons comparé les p-values obtenues, après avoir testé successivement l'association entre le phénotype Y et le SNP causal, les ancêtres du SNP causal et les non-ancêtres du SNP causal (en abrégé N) dans le graphe de la FMHCL. Pour une meilleure visualisation des résultats, les valeurs de $-log_{10}$(p-value) sont utilisées. Des valeurs proches de 0 révèlent une indépendance. Ces valeurs augmentent avec la force de la dépendance.

Pour mesurer la significativité des associations, nous avons implémenté une procédure de permutation conçue pour le calcul du taux d'erreur de type I par test (noté

2. Le test du ratio de vraisemblance teste le ratio de la vraisemblance sous hypothèse H_0 à la vraisemblance sous hypothèse H_1.

Algorithme 3 Procédure de permutation $(X, D_X, Y, D_Y, n_p, \alpha)$

ENTRÉE :
$\mathbf{X}, \mathbf{D_X}$: un ensemble de n_v variables candidates (observées ou latentes) $X = X_1, ..., X_{n_v}$ et les données correspondantes observées ou imputées pour n individus,
$\mathbf{Y}, \mathbf{D_Y}$: une variable cible Y et les données correspondantes observées pour n individus,
$\mathbf{n_p}$: le nombre de permutations,
α : le taux d'erreur de type I lors de tests multiples.

SORTIE :
$\{\alpha'(\mathbf{1}), ..., \alpha'(\mathbf{n}_\ell)\}$: l'ensemble des taux d'erreur de type I par test calculés pour les couches 1 à n_ℓ, respectivement.

pour $\ell = 1$ **à** n_ℓ $\quad distrib_{minPValues}(\ell) \leftarrow \emptyset$ $\quad$ **fin pour**

pour $p = 1$ **à** n_p
  $D_{Y_p} \leftarrow permuterEtiquettes(D_Y)$
  pour $\ell = 1$ **à** n_ℓ
    $pValues(p,\ \ell) \leftarrow \emptyset$
    pour chaque variable X_v dans la couche ℓ
      $pValue_{p,\ \ell,\ v} \leftarrow réaliserTestAssociation(D_{X_v}, D_{Y_p})$
      $pValues(p,\ \ell) \leftarrow pValues(p,\ \ell) \cup pValue_{p,\ \ell,\ v}$
    fin pour
    $distrib_{minPValues}(\ell) \leftarrow distrib_{minPValues}(\ell) \cup min_{X_v}(pValues(p,\ \ell))$
  fin pour
fin pour

pour $\ell = 1$ **à** n_ℓ
  $\alpha'(\ell) \leftarrow quantile(distrib_{minPValues}(\ell), \alpha)$
fin pour

α'), de manière à contrôler le taux d'erreur de type I lors de tests multiples (noté α) à 5%, par exemple. α' fournit un seuil de significativité par test. α, quant à lui, contrôle la probabilité d'observer un faux positif, ou plus, parmi toutes les hypothèses considérées lorsque plusieurs tests d'association sont réalisés. Pour une valeur α donnée, plus le nombre de variables à tester est grand, plus la valeur de α' est faible. Un des avantages de la structure hiérarchique de notre modèle est qu'il présente moins de variables dans les couches hautes. Ainsi, une augmentation de α' avec le niveau de la couche est attendue. C'est pour cela que notre procédure de permutation a été adaptée au calcul d'un α' spécifique à chaque couche.

Dans la procédure de permutation standard [28], les valeurs de Y (le phénotype) sont permutées un certain nombre de fois. Un ensemble de permutations $\mathcal{P}$ est généré (généralement, $|\mathcal{P}| = 1000$). Pour chaque permutation D_{Y_p}, un test d'association entre Y et chaque variable candidate X_v est réalisé et le maximum des statistiques ($max(T)$) obtenues sur tous les tests est sauvegardé. A chaque permutation, est ainsi associée une unique statistique ($max(T)$). Lorsque le nombre de permutations est suffisamment grand, la distribution de $max(T)$ représente une bonne approximation empirique de la distribution sous hypothèse nulle qui stipule qu'aucun X_v n'est associé à Y. Par ailleurs, dans notre modèle, les VL peuvent présenter différentes cardinalités. Ainsi les tests d'association présentent des degrés de liberté différents. Il est

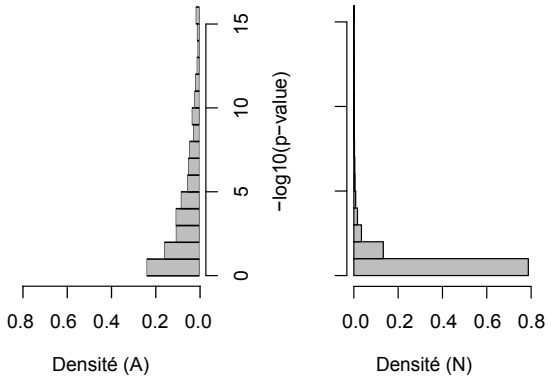

FIGURE 5.2: Histogramme des valeurs de $-log_{10}$(p-value) provenant des tests d'association entre le phénotype et les nœuds ancêtres du SNP causal (A), et entre le phénotype et les nœuds non-ancêtres du SNP causal (N).
*Les résultats regroupent tous les scénarios : FAM (0.1-0.2, 0.2-0.3 et 0.3-0.4), RRG (*1.4, 1.6 *et* 1.8*) et modèle de la maladie (dominance, récessivité, additivité et multiplicativité). Pour une description des scénarios, voir section 5.2.2.1, page 84.*

alors impossible de comparer les statistiques entre elles. C'est la raison pour laquelle nous avons adapté la procédure de permutation standard à cette caractéristique. L'adaptation est simple : au lieu de nous baser sur la distribution des maximum des statistiques, nous employons la distribution des minimum des p-values. En effet, les p-values peuvent être comparées, étant donné que leur calcul prend en compte le degré de liberté.

La nouvelle procédure de permutation est décrite dans l'algorithme 3. Cette procédure consiste à réaliser n_p permutations pour la variable Y. Pour chaque permutation D_{Y_p}, et chaque couche de la FMHCL ℓ, un test d'association est réalisé entre chaque variable X_v appartenant à la couche de la FMHCL et la variable Y (ligne 7). Ensuite, pour chaque permutation, la p-value minimale sur toutes les variables appartenant à la couche ℓ est identifiée et participera à la distribution des p-values minimales pour cette couche (ligne 10). Étant donné un taux d'erreur α, cette distribution permet alors d'extraire le seuil α'^{ℓ} correspondant (ligne 14). Cette valeur de α'^{ℓ}, spécifique à chaque couche, doit être comparée avec la p-value résultant du test d'association entre la variable X_v (appartenant à la couche ℓ) et Y. Ainsi l'association est testée au seuil de significativité spécifique à chaque couche qui est corrigé pour les tests multiples.

5.2.3 Résultats expérimentaux et discussion

Dans la suite, nous nous focalisons sur les quatre premières couches du modèle (couches 0-3) et ne considérons pas les autres couches, pour lesquelles nous ne disposons pas de suffisamment de résultats à analyser (ces couches montrent généralement très peu de variables ou ne sont pas présentes dans le modèle).

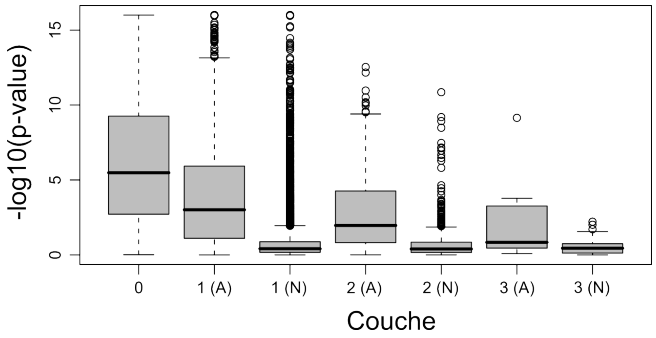

FIGURE 5.3: Boîte à moustaches des valeurs de $-log_{10}$(p-value) pour les différentes couches de la forêt de modèles hiérarchiques à classes latentes, résultant des tests d'association entre le phénotype et les nœuds ancêtres du SNP causal (A), et entre le phénotype et les nœuds non-ancêtres du SNP causal (N).
La couche 0 *se réfère aux tests d'association entre le phénotype et le SNP causal. Voir figure 5.2 pour les explications concernant les scénarios.*

5.2.3.1 Données simulées

Tests d'indépendance marginale

La figure 5.2 compare les histogrammes des valeurs de $-log_{10}$(p-value) résultant des tests d'associations de Y avec les A et les N, respectivement. La comparaison de ces deux histogrammes révèle une grande dissimilarité entre les deux distributions. La majorité (70%) des valeurs de $-log_{10}$(p-value) associées aux A est plus grande que 1, alors que c'est le cas pour seulement 19% des N. En effet, nous observons que de grandes valeurs de $-log_{10}$(p-value) (*e.g.*, plus grandes que 5) sont fréquentes pour les A et très rares pour les N. Le test des rangs de Wilcoxon, réalisé pour comparer les deux distributions, montre une p-value inférieure à 10^{-16}, ce qui confirme que les p-values des A et des N suivent deux distributions différentes.

La figure 5.3 décrit plus précisément les valeurs de $-log_{10}$(p-value) observées pour les différentes couches de la FMHCL dans le cas des tests associés aux A et N. La couche 0 se réfère aux tests d'association entre le phénotype et le SNP causal, et sert de valeur de référence. Dans cette figure, nous observons que la force de l'association avec les A décroît lentement lorsque le numéro de la couche augmente. En revanche, les associations avec les N présentent constamment des valeurs de $-log_{10}$(p-value) en deçà de 0.4, correspondant à des p-values supérieures à 0.4. Bien que les N présentent un certain nombre de faux positifs (moins de 10% ont une p-value inférieure à 0.01), ces résultats montrent clairement une tendance générale : les associations indirectes sont capturées par les A tandis que la majorité des N ne montre pas d'association.

La figure 5.4 souligne la tendance générale des valeurs de $-log_{10}$(p-value) observées pour les A et les N, et compare la médiane des valeurs obtenues pour chaque couche à la valeur du seuil de significativité α' correspondant à cette couche (voir

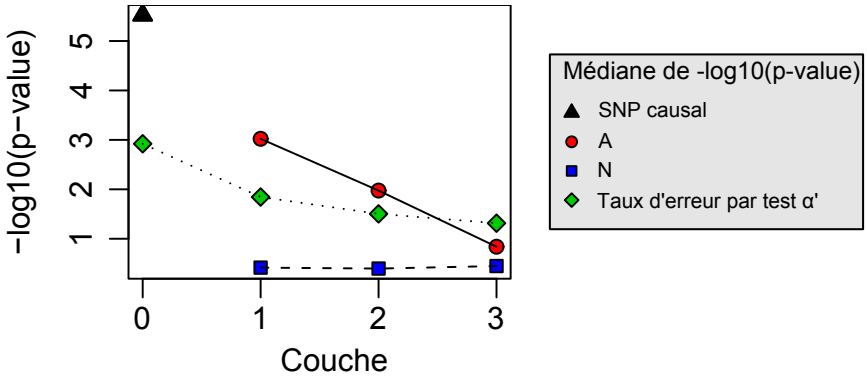

FIGURE 5.4: Médiane des valeurs de $-log_{10}$(p-value) pour les différentes couches de la forêt de modèles hiérarchiques à classes latentes, résultant des tests d'association entre le phénotype et les nœuds ancêtres du SNP causal, et entre le phénotype et les nœuds non-ancêtres du SNP causal.
Pour la définition du taux d'erreur α', voir section 5.2.2.2, page 84. A : ancêtres du SNP causal ; N : non-ancêtres du SNP causal. Voir figure 5.2 pour les explications concernant les scénarios et voir figure 5.3 pour les explications concernant la couche 0.

	Pourcentage
p(NA)	21%
p(NE)	79%
p(NA | FP)	73%
p(NE | FP)	27%
p(FP | NA)	**16%**
p(FP | NE)	**1.6%**

TABLEAU 5.2: Tableau récapitulatif des pourcentages de faux positifs (FP) pour les nœuds non-ancêtres du SNP causal et localisés dans l'arbre causal (NA), et pour les nœuds non-ancêtres du SNP causal et non localisés dans l'arbre causal (NE).
*A = vrais positifs ; N = vrais négatifs + faux positifs ; N = NA (*21%*) + NE (*79%*).*

section 5.2.2.2, page 84). Cette figure révèle que, jusqu'à la seconde couche, des associations significatives sont identifiées pour les A. A l'inverse, en ce qui concerne les N, les médianes des valeurs de $-log_{10}$(p-value) sont plus petites que les valeurs de seuils $-log_{10}(\alpha')$ correspondantes, et ceci, pour toutes les couches. En nous focalisant sur la distribution des N, nous observons que le pourcentage de p-values inférieures à α' (faux positifs) est égal à 4.7%. L'existence de faux positifs peut s'expliquer en partie par la présence de dépendances indirectes entre le SNP causal et les NA, *i.e.* les N localisés dans l'arbre causal. L'arbre causal est l'arbre contenant le SNP causal (voir figure 5.1). A l'opposé, aucun faux positif n'est attendu pour les NE, qui sont les N présents dans les arbres non causaux. En fait, plus de 73% des faux positifs sont des NA, alors qu'ils ne représentent seulement que 21% des N. Plus précisément, le taux de faux positifs dans les NA est de 16.07% alors qu'il est 10 fois inférieur dans les NE. Les différents pourcentages sont résumés dans le tableau 5.2. Ainsi nous

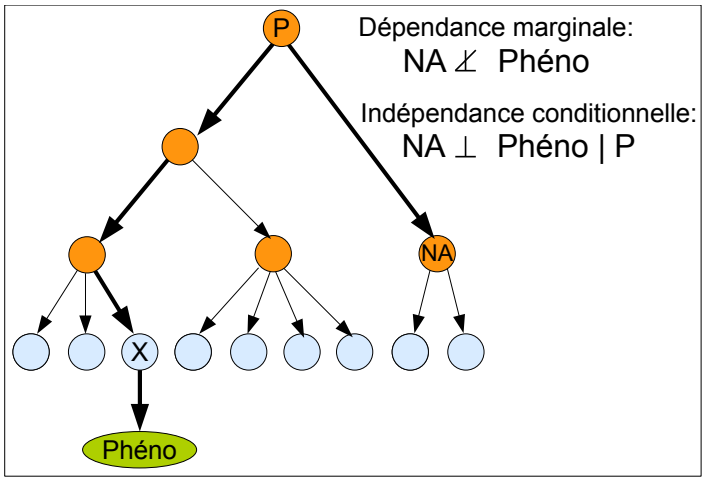

FIGURE 5.5: Illustration de la dépendance indirecte entre non-ancêtres du SNP causal localisés dans l'arbre causal (NA) et le phénotype : dépendance marginale et indépendance conditionnellement au SNP causal.
Phéno : phénotype ; X : SNP causal ; NA : non-ancêtre du SNP causal localisés dans l'arbre causal ; P : parent de NA.

pouvons conclure qu'une importante partie des faux positifs est due à l'existence de dépendances indirectes entre les NA et le phénotype (voir figure 5.5).

Tests d'indépendance conditionnelle

Les tests d'indépendance conditionnelle représentent une solution permettant de ne pas prendre en compte les dépendances indirectes entre le phénotype et les NA, qui passent par le SNP causal. Il est possible de réaliser un test entre le phénotype et un NA conditionnellement à son parent, car le chemin entre le phénotype et le NA passant par le SNP causal est séparé par le parent du NA dans le graphe de la FMHCL (voir figure 5.5). En employant ce test pour tous les NA, nous observons que la plupart des p-values sont plus élevées (donc moins significatives). En effet, plus de 99% des p-values sont au-delà des seuils de significativité calculés par permutation. Ce bon résultat doit être modéré car les p-values des tests conditionnels ont généralement tendance à être beaucoup plus élevées que celles des tests simples, du fait d'un degré de liberté plus grand. Par exemple, considérons une VL de cardinalité égale à 3 ayant une variable parent de même cardinalité. La variable phénotype possède une cardinalité égale à 2. Le degré de liberté est égal à seulement 2 (*i.e.* 2×1) pour le test non conditionnel entre la VL et le phénotype, tandis qu'il est de 6 (*i.e.* $2 \times 1 \times 3$) pour le test conditionnel. Malgré ce phénomène, un résultat intéressant est observé : le taux de faux positifs dans les NA (qui est égal à 1%) est maintenant seulement le double de celui obtenu dans les NE. Nous rappelons qu'en utilisant des tests non conditionnels, le ratio entre le taux de faux positifs chez les NA et le taux de faux positifs chez les NE est d'environ 10. Ainsi, bien que les tests conditionnels entraînent le calcul de p-values plus grandes, ils permettent dans le même temps de réduire le

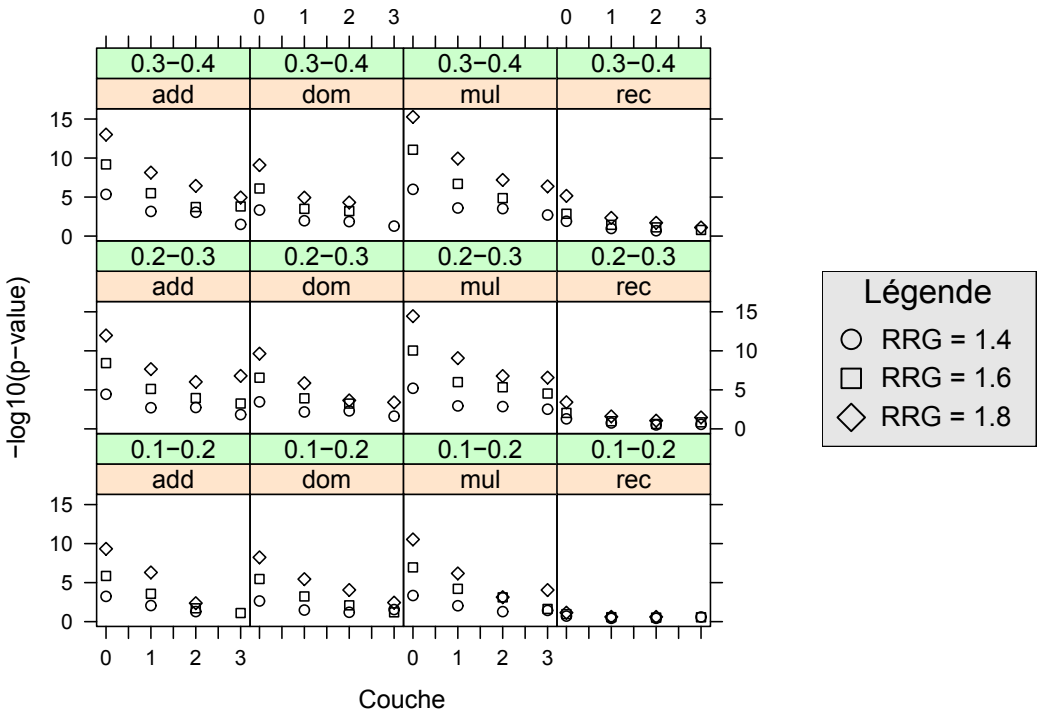

FIGURE 5.6: Médiane des valeurs de $-log_{10}$(p-value) pour les différentes couches de la forêt de modèles hiérarchiques à classes latentes, résultant des tests d'association entre le phénotype et les nœuds ancêtres du SNP causal.
Les fenêtres représentent les différents scénarios testés. En haut de chaque fenêtre, l'intervalle de la FAM et le modèle de la maladie (additivité, dominance, multiplicativité ou récessivité) sont indiqués. Les trois symboles employés se réfèrent aux risques relatifs génotypiques (RRG) considérés pour le SNP causal simulé (voir légende). La couche 0 *se réfère aux tests d'association entre le phénotype et le SNP causal (en considérant les* 100 *réplications).*

taux de faux positifs d'un facteur 5, du fait de la prise en compte des dépendances indirectes entre phénotype et NA.

Comparaison sous différentes configurations génétiques

Nous avons évalué le comportement des VL pour différentes configurations génétiques : intervalle de la FAM ([0.1-0.2], [0.2-0.3] ou [0.3-0.4]), RRG (1.4, 1.6 ou 1.8), et modèle de la maladie (additivité, dominance, multiplicativité ou récessivité). Comme précédemment, nous avons comparé les résultats des tests d'association obtenus pour les A et pour les N. Ces résultats sont présentés en figures 5.6 et 5.7. Globalement, des tendances similaires sont observées pour tous les scénarios : la force de l'association diminue de manière constante de la première couche à la quatrième ; dans le cas des N, la majorité des résultats montre l'absence d'association, quelle que soit la couche étudiée.

Lorsque l'on considère la situation la plus facile pour la détection d'association

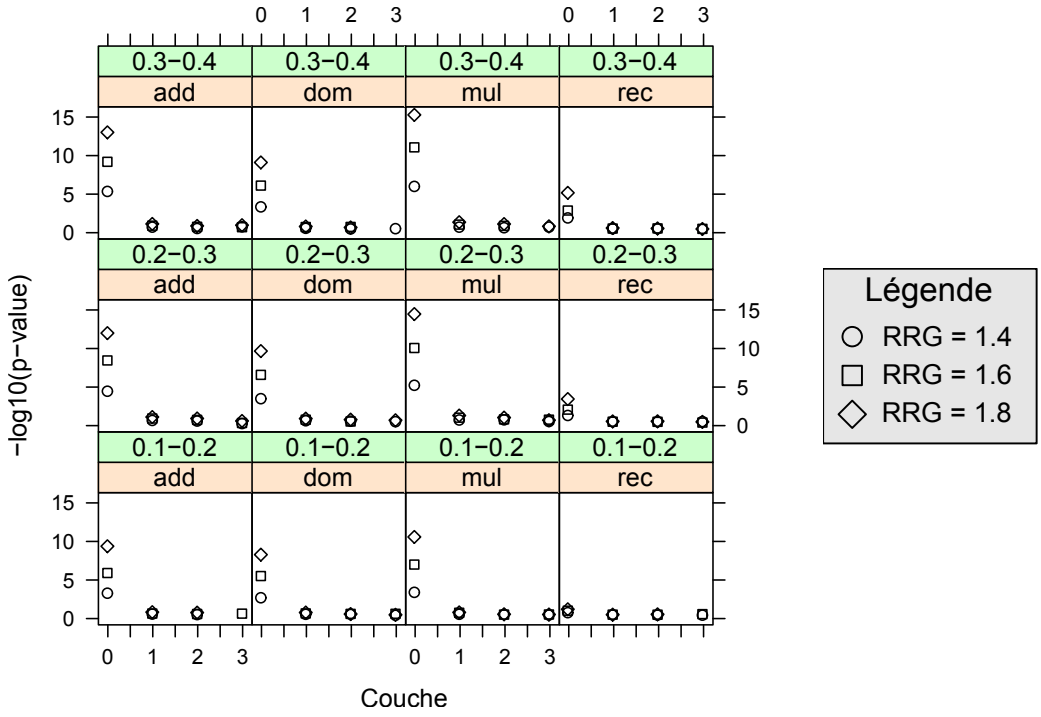

FIGURE 5.7: Médiane des valeurs de $-log_{10}$(p-value) pour les différentes couches de la forêt de modèles hiérarchiques à classes latentes, résultant des tests d'associations entre le phénotype et les nœuds non-ancêtres du SNP causal.
Voir figure 5.6 pour la description des paramètres et les explications concernant la couche 0.

(intervalle de FAM = [0.3-0.4], RRG = 1.8 et modèle multiplicatif), sur toutes les couches, les A présentent de fortes associations ($-log_{10}$(p-value) > 7). Pour une configuration moins idéale, mais plus plausible, (intervalle de FAM = [0.2-0.3], RRG = 1.6 et modèle additif), la médiane des valeurs de $-log_{10}$(p-value) calculées pour les A décroît de 8.3 pour la couche 0, pour atteindre 4.6, 3.2 et 2.2 pour les couches 1, 2 et 3, respectivement. Au contraire, lorsque le modèle est récessif, l'association avec le SNP causal est basse et les A ne peuvent rien capturer (des résultats similaires sont obtenus avec la plupart des méthodes conçues pour la découverte d'associations génétiques). En ce qui concerne les N, des associations nulles sont observées quelle que soit la configuration.

5.2.3.2 Données réelles

Nous avons évalué notre approche de recherche de causalité sur des données génotypiques réelles concernant la région 22q13 (chromosome 22) de 890 *kb* flanquant le gène *CYP2D6* qui a un rôle dans le métabolisme lié à certaines molécules pharmaceutiques [36]. Ce jeu de données est composé de 32 SNP génotypés pour 268 individus

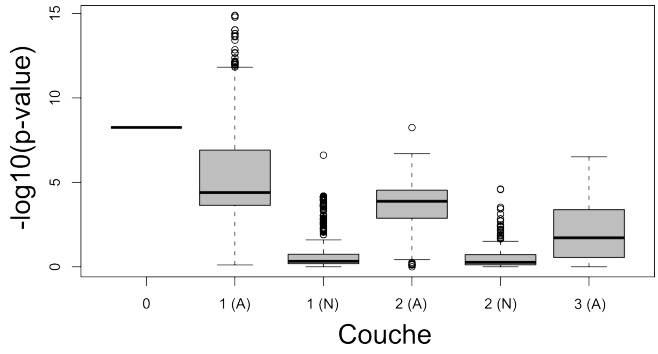

FIGURE 5.8: Boîte à moustaches des valeurs de $-log_{10}$(p-value) pour les différentes couches de la forêt de modèles hiérarchiques à classes latentes (FMHCL), résultant de tests d'association entre le phénotype et les nœuds ancêtres du SNP causal (A), et entre le phénotype et les nœuds non-ancêtres du SNP causal (N), sur données réelles.
La couche 0 *se réfère aux tests d'association entre le phénotype et le SNP causal (SNP* 19*). Dans la couche* 3*, aucun N n'est observé dans les FMHCL.*

et a été récupéré à partir du package R graphminer (http ://homepages.lshtm.ac.uk/-encdcver/) développé par Verzilli *et al.* [107]. Cette région chromosomique a été utilisée dans plusieurs études comme banc d'essai pour les méthodes de cartographie fine basées sur le LD. Le SNP 19 (position 550 *kb*) est le plus proche marqueur du gène *CYP2D6* (positionné à 525.3 *kb*) [94]. C'est la raison pour laquelle nous considérons le SNP 19 comme le marqueur causal lors de nos expérimentations.

Afin de prendre en compte la nature stochastique de notre algorithme (initialisation aléatoire des paramètres lors de l'algorithme EM), nous présentons les résultats de 1000 exécutions. Chacune des exécutions dure en moyenne 5.4 *s* sur un ordinateur personnel standard (3 GHz, 3.3 *Go* RAM). En moyenne, sur les 1000 FMHCL (1000 réplications), les pourcentages de nœuds sont distribués de la façon suivante : 82.62% dans le couche 0, 16.89% dans la couche 1, 0.39% dans la couche 2 et 0.10% dans la couche 3. La figure 5.8 montre les valeurs de $-log_{10}$(p-value) pour les tests d'association sur les A et les N. De façon similaire aux résultats obtenus avec les données simulées, nous observons que les A réussissent à capturer les associations génétiques indirectes, en particulier dans la couche 1, avec une valeur médiane de 4.5, correspondant à des p-values inférieures à 5.10^{-5}. Dans les autres couches, la force d'association est plus faible, mais demeure relativement élevée comme dans la couche 2 montrant une valeur médiane de 4, équivalent à une p-value de 10^{-4}. Comme précédemment avec les données simulées, lorsque nous nous focalisons sur les N, nous observons très peu d'associations. La majorité des p-values (plus de 80%) est supérieure à 0.01.

5.2.4 Conclusion

En nous basant sur l'analyse de données simulées et réelles, nous avons montré la capacité des variables latentes (VL) de la forêt de modèles hiérarchiques à classes

latentes (FMHCL) à être utilisées pour la recherche de causalité, notamment dans le domaine de la génétique humaine. Une capture efficace des associations génétiques indirectes est réalisée grâce à deux propriétés majeures : (i) les nœuds ancêtres du SNP causal réussissent à capturer les associations indirectes avec le phénotype ; (ii) à l'inverse, les nœuds non-ancêtres du SNP causal montrent globalement de très faibles associations. En d'autres termes, il est possible de distinguer entre vraies et fausses associations génétiques indirectes.

5.2.5 Perspectives

Le présent travail demeure une étude préliminaire de l'utilisation de la FMHCL pour la recherche de causalité. La principale perspective de recherche consiste à développer une méthode basée sur ce modèle pour la découverte de SNP causaux dans les études d'association pangénomiques. Pour cela, plusieurs pistes de recherche sont à explorer :

- Pour mesurer l'association entre une VL et le phénotype, nous avons utilisé les données imputées de la VL en choisissant la classe la plus probable (voir chapitre 4, section 4.3.3.4, page 60). Une mesure plus précise de l'association serait d'utiliser, pour chaque individu, la probabilité d'appartenance pour chacune des classes (données pondérées). En outre, au lieu d'employer les probabilités d'appartenance calculées lors de l'apprentissage local de la FMHCL par CFHLC (ce qui est réalisé au niveau de chaque modèle à classe latente), il est possible de calculer ces probabilités par inférence probabiliste sur la loi jointe. Le calcul d'inférence sera efficace, car il est de complexité linéaire avec le nombre de nœuds dans l'arbre.

- Pour la recherche d'association, l'approche que nous avons utilisée est fréquentiste. Elle met en œuvre le calcul de p-values. Cependant de nombreuses études [35, 85, 90, 94, 117] tendent à montrer la supériorité de l'approche bayésienne sur l'approche fréquentiste pour les tests d'association dans le contexte des études d'association pangénomiques. En effet, dans le cadre fréquentiste, la correction des p-values pour les tests multiples mène à calculer des seuils de significativité très stricts, lorsque le nombre de SNP considéré est important (les seuils sont d'autant plus bas qu'il y a de tests à réaliser). Au contraire, dans l'approche bayésienne, le seuil de significativité ne dépend plus du nombre de SNP à tester, mais de l'espérance *a priori* du nombre de SNP causaux. De ce fait, il n'y a plus besoin de réaliser de calcul de seuils de significativité par permutation.

- La structure hiérarchique latente du modèle pourrait être employée pour réaliser une découverte des facteurs causaux selon une approche descendante (des couches hautes vers les couches basses), afin d'identifier les SNP causaux. L'idée est qu'il est probable, lorsqu'on observe une VL associée au phénotype, qu'il existe un chemin partant de cette VL et dans lequel toutes les variables (VL

et SNP) sont aussi associées au phénotype. Le parcours de ce chemin, qui peut être réalisé au moyen d'un parcours en profondeur, devrait guider la découverte de SNP causaux.

5.3 Visualisation pangénomique du déséquilibre de liaison

5.3.1 Introduction et état de l'art

L'étude du LD représente un sujet de recherche majeur en statistique génétique. Par exemple, le LD joue un rôle fondamental dans la cartographie des gènes réalisée par étude d'association, mais son analyse permet aussi d'apporter une meilleure compréhension de l'histoire des populations humaines. Par exemple, les goulots d'étranglement [3], la sélection naturelle et les migrations sont des exemples d'évènements évolutifs qui peuvent être inférés à l'aide de modèles coalescents qui analysent les haplotypes [89].

A l'interface entre les statistiques et l'intelligence artificielle, la fouille de données est le processus d'extraction de connaissances à partir de données [21]. La fouille de données permet de formuler des hypothèses pertinentes à tester et offre des outils complémentaires aux méthodes statistiques conventionnelles. La visualisation de données, que l'on peut considérer comme une branche de la fouille de données, a pour objectif de fournir des outils efficaces et intuitifs afin d'afficher et de synthétiser l'information la plus pertinente sous-jacente aux données [86]. La visualisation de données a été appliquée avec succès en bioinformatique [4], par exemple en génomique et en protéomique.

Le projet international HapMap [98], et plus récemment le projet international 1000 génomes [95], ont fédéré de nombreux efforts afin de caractériser de façon approfondie la variation de la séquence du génome dans les populations humaines. Dans ce contexte, l'application de méthodes de visualisation à l'analyse des patrons de LD s'est révélée essentielle, plus particulièrement pour mettre en relief la structure complexe du LD en blocs [19]. La méthode la plus simple, mais aussi la plus populaire, est la carte triangulaire de chaleur (CTC), construite par exemple par le logiciel Haploview [6]. La CTC est la matrice triangulaire des dépendances par paires de SNP, pour laquelle l'intensité de la couleur indique la force du LD de chaque cellule de la matrice. La CTC affiche généralement la mesure de Lewontin D' ou le coefficient de corrélation au carré r^2. Une autre mesure de dépendance, le ratio de la mesure D' au logarithme du rapport des chances (LOD), est également employée comme mesure standard par Haploview. Dans la CTC, les blocs de LD sont visuellement apparents. Néanmoins, la CTC présente l'inconvénient de n'afficher que les dépendances par paires, empêchant la visualisation des patrons multilocus. Une autre méthode populaire consiste à présenter les taux de recombinaison par cartographie fine calculés le long de la séquence chromosomique. Pour cela, PHASE [91], une méthode basée sur

3. Facteur extrinsèque à une population qui vient diminuer son effectif.

le coalescent [4], peut être utilisée afin d'estimer le taux de recombinaison pour chaque paire de SNP adjacents sur la séquence. Cette approche identifie les points chauds de recombinaison et donne une idée de la structure en blocs du LD, mais nécessite de long calculs. Les techniques de visualisation plus avancées, telles que les "isometric blocks" et les "bifurcation plots" [27], ou le "textile plot" [46], peuvent prendre en compte le LD multilocus. Par exemple, le textile plot emploie un algorithme proche de l'analyse en composantes principales. La stratégie du textile plot consiste à rechercher une configuration géométrique optimale pour les variables et les individus dans un espace linéaire de dimension plus faible que celle des données initiales.

Grâce à la nature graphique et hiérarchique de la FMHCL, l'emploi de cette dernière constitue une voie prometteuse pour le développement d'outils de visualisation intuitif et synthétique des dépendances entre SNP. C'est la raison pour laquelle nous allons appliquer la FMHCL aux trois principales situations que peut rencontrer le généticien : la visualisation du LD de courte distance, de longue distance et dans un contexte pangénomique.

5.3.2 Matériel et méthodes

5.3.2.1 Apprentissage de la FMHCL

Pour les expérimentations sur la visualisation du LD, nous avons employé l'algorithme CFHLC+ qui ne nécessite plus le découpage du génome en fenêtres contiguës de taille constante. Ce nouvel algorithme contraint la recherche des dépendances entre SNP à l'aide de leur position physique sur l'ADN, et correspond en fait à une approche par fenêtre glissante. Fixer la taille de la fenêtre entre 0.1 et 5 Mb représente une stratégie raisonnable afin de capturer le LD de longue distance dans le contexte des EAP.

5.3.2.2 Déséquilibre de liaison multilocus

Puisque la FMHCL modélise le LD multilocus [5], il est possible de calculer une valeur de multilocus LD à partir de la loi jointe de la FMHCL. Notamment, le LD multilocus peut être calculé simplement pour chaque sous-arbre de la FMHCL (*i.e.* pour chaque cluster de LD).

Nothnagel *et al.* ont proposé d'employer l'entropie (voir section 3.2.2, chapitre 3, page 27) afin d'évaluer le LD multilocus dans les blocs haplotypiques [72]. Ils construisent une mesure de LD multilocus comme la différence d'entropies entre équilibre de liaison (S_E) et déséquilibre de liaison (S_B) :

$$\Delta S = S_E - S_B.$$

4. Généalogie ancestrale de la séquence.
5. LD présent entre plusieurs loci.

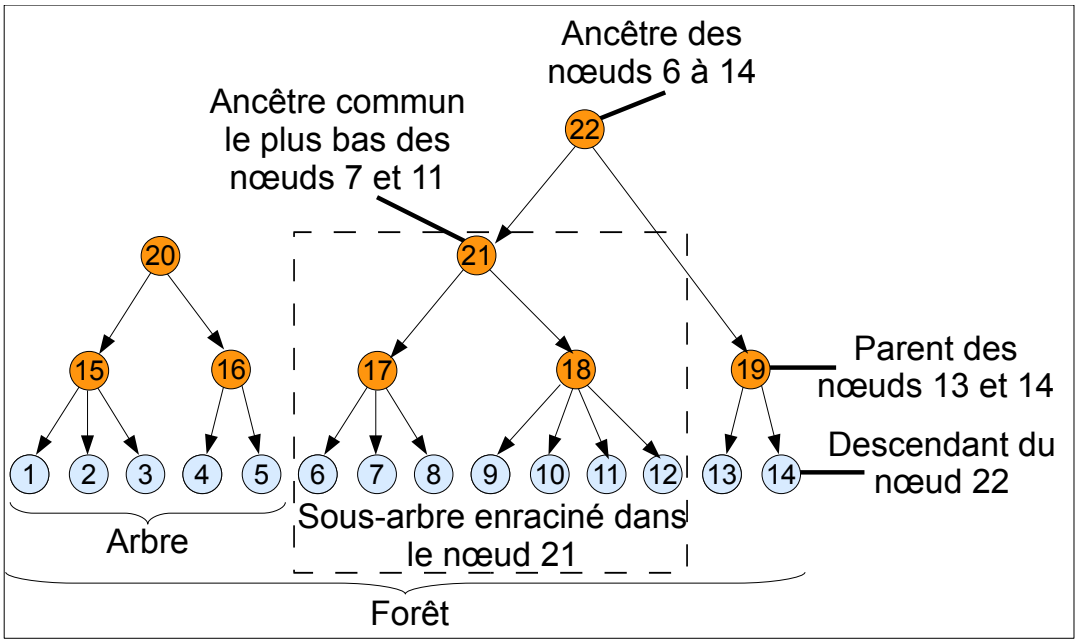

FIGURE 5.9: Illustration des termes spécifiques à notre méthode de visualisation : sous-arbre, arbre, forêt, parent, descendant, ancêtre et ancêtre commun le plus bas.
Les SNP et les variables latentes sont représentées par des nœuds bleus (clairs) et rouges (foncés), respectivement.

Ensuite, la mesure ΔS est normée dans l'intervalle $[0, 1]$:

$$\epsilon = \frac{\Delta S}{S_E} = 1 - \frac{S_B}{S_E}.$$

En suivant la même idée, nous avons conçu une mesure de LD multilocus basée sur la FMHCL apprise. Dans la suite, nous n'allons considérer que les sous-arbres composés d'une VL de la FMHCL et de tous ses descendants, ce qui correspond en fait à des MHCL (*e.g.* le sous-arbre enraciné dans le nœud 21 en figure 5.9). Prenons l'exemple d'un sous-arbre sur un ensemble de n nœuds observés et latents $\mathbf{V} = \{V_1, ..., V_n\}$. $\mathbf{V}$ est composé d'un ensemble de m nœuds observés $\mathbf{X} = \{X_1, ..., X_m\}$ et d'un ensemble de p nœuds latents $\mathbf{H} = \{H_1, ..., H_p\}$. $\mathbf{H}$ peut aussi se diviser en deux : l'ensemble des $p - 1$ nœuds non-racines $\mathbf{N} = \{N_1, ..., N_{p-1}\}$, et le nœud racine R.

L'entropie sous hypothèse d'équilibre de liaison, *i.e.* en l'absence de VL, est simplement la somme des entropies individuelles des VO :

$$S_E = \sum_{j=1}^{m} \mathcal{H}(X_j).$$

L'entropie sous hypothèse de déséquilibre de liaison (modélisé par le sous-arbre de la FMHCL) est l'entropie de la loi jointe des VO, $\mathcal{H}(\mathbf{X})$. Pour éviter un coût

calculatoire important lié à la marginalisation du modèle sur les variables latentes **H**, nous choisissons d'utiliser à la place l'entropie de la loi jointe du modèle :

$$\begin{aligned} S_B &= \mathcal{H}(\mathbf{X}, \mathbf{N}, R) \\ &= \sum_{j=1}^{m} \mathcal{H}(X_j | Pa_{X_j}) + \sum_{k=1}^{p-1} \mathcal{H}(N_k | Pa_{N_k}) + \mathcal{H}(R). \end{aligned}$$

Cette entropie de la loi jointe du modèle est supérieure ou égale à l'entropie de la loi jointe marginalisée des VO :

$$\mathcal{H}(\mathbf{X}, \mathbf{N}, R) \geq \mathcal{H}(\mathbf{X}) \tag{5.1}$$

Sachant les entropies S_E et S_B, ΔS peut s'écrire de la manière suivante :

$$\begin{aligned} \Delta S = &\sum_{j=1}^{m} \mathcal{H}(X_j) - \sum_{j=1}^{m} \mathcal{H}(X_j | Pa_{X_j}) \\ &- \sum_{k=1}^{p-1} \mathcal{H}(N_k | Pa_{N_k}) - \mathcal{H}(R). \end{aligned} \tag{5.2}$$

Étant donné l'équation 5.1, il en résulte que notre différence d'entropie ΔS calculée à partir de la loi jointe du modèle est inférieure ou égale à ΔS que l'on aurait calculée à partir de la loi jointe marginalisée des VO.

Grâce à la formule de l'information mutuelle (section 3.2.2, chapitre 3, page 27), l'équation 5.2 est reformulée comme suit :

$$\begin{aligned} \Delta S = &\sum_{j=1}^{m} \Big(\mathcal{H}(X_j) - \mathcal{H}(X_j | Pa_{X_j}) \Big) \\ &- \sum_{k=1}^{p-1} \Big(\mathcal{H}(N_k) - \mathcal{I}(N_k; Pa_{N_k}) \Big) \\ &- \mathcal{H}(R) \\ = &\sum_{j=1}^{m} \mathcal{I}(X_j; Pa_{X_j}) + \sum_{k=1}^{p-1} \mathcal{I}(N_k; Pa_{N_k}) \\ &- \sum_{k=1}^{p-1} \mathcal{H}(N_k) - \mathcal{H}(R) \\ = &\sum_{i=1}^{n} \mathcal{I}(V_i; Pa_{V_i}) - \sum_{j=1}^{p} \mathcal{H}(H_j). \end{aligned} \tag{5.3}$$

Nous observons que ΔS se compose de deux termes : (i) le premier est souvent employé afin d'évaluer l'ajustement d'un RB sans VL aux données, lorsque la structure

du modèle est un arbre ou une forêt, alors que (ii) le second peut être vu comme un terme de pénalisation spécifique à ces modèles à variables latentes. Ce terme de pénalisation est la somme des entropies individuelles des VL, et permet ainsi de prendre en compte l'augmentation de complexité due à l'incorporation de VL dans le modèle. En effet, l'entropie d'une VL croît avec le nombre de ses classes (ou états) d'une part, et avec l'uniformité de sa distribution, d'autre part.

Finalement, de la même façon que la mesure de Nothnagel *et al.*, ΔS est normée :

$$\epsilon = \frac{\sum_{i=1}^{n} \mathcal{I}(V_i, Pa_{V_i}) - \sum_{j=1}^{p} \mathcal{H}(H_j)}{\sum_{k=1}^{m} \mathcal{H}(X_k)}.$$

Dans de rares situations, la valeur de ϵ peut être légèrement inférieure à 0 (à cause du terme de pénalisation). Dans ce cas, la valeur de ϵ est seuillée à 0. Un calcul efficace des ϵ peut être réalisé en partant de la première couche de la FMHCL et en finissant dans la dernière couche.

5.3.2.3 Tracé de graphe et visualisation

Le tracé, c'est-à-dire le placement des nœuds, et la visualisation (ou affichage) du graphe des FMHCL représentent une étape importante pour notre méthode de représentation graphique du LD [42]. Pour cet objectif, nous proposons une solution simple qui offre un vue claire et intelligible de la structure spatiale du LD. Grâce à la nature hiérarchique des FMHCL, il est possible d'implémenter un tracé simple et intuitif : les nœuds sont placés le long du chromosome et couche par couche. Les SNP sont placés le long du chromosome en utilisant leur ordre physique sur la séquence. Les VL sont positionnées en employant les ordres physiques calculés en moyennant sur ceux des SNP subsumés. Chaque couche est positionnée et ordonnée le long d'un axe perpendiculaire au chromosome.

En ce qui concerne la visualisation, seul un petit nombre de logiciels a été développé afin de manipuler de grands graphes tels que ceux issus de la modélisation pangénomique du LD par des FMHCL. Pour visualiser les graphes, nous avons choisi le logiciel Tulip (http ://tulip.labri.fr/TulipDrupal/) qui est un outil convivial, capable de traiter plus d'un million de nœuds, et qui permet la navigation grâce à un ensemble d'opérations géométriques variées ainsi que l'extraction de sous-graphes et la mise en avant de résultats obtenus après filtrage. Afin de visualiser le LD multilocus (ϵ) pour les sous-arbres de la FMHCL, nous proposons de lier l'intensité de la couleur de la VL subsumant le sous-arbre à la force du LD. La valeur précise du LD est aussi affichée à l'intérieur du nœud représentant la VL.

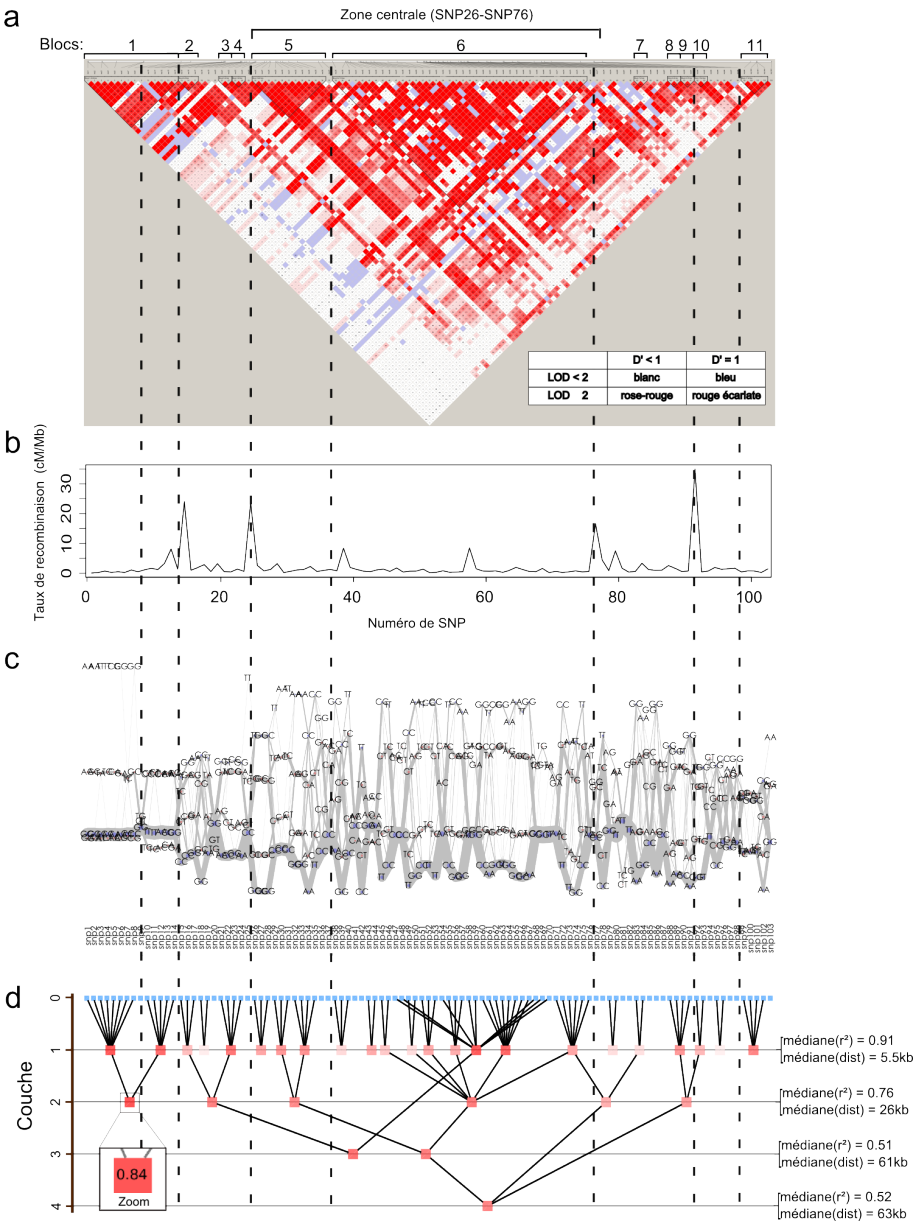

FIGURE 5.10: Comparaison des méthodes de visualisation du déséquilibre de liaison de courte distance, appliquées au jeu de données de Daly *et al.* [19] : a) carte triangulaire de chaleur du D'/LOD pour les blocs de LD calculés par Haploview v4.2, b) taux de recombinaison inférés avec PHASE v2.1, c) textile plot et d) forêt de modèles hiérarchiques à classes latentes affichée avec Tulip.
Pour toutes les paires de SNP possédant leur ancêtre commun le plus bas (ACLPB) dans la couche ℓ, la médiane des r^2 et la médiane des distances calculées (kb) ont été affichées (d) section droite). Le zoom montre la mesure du déséquilibre de liaison multilocus relative à une variable latente. Les lignes en pointillés soulignent les tendances communes aux quatre méthodes.

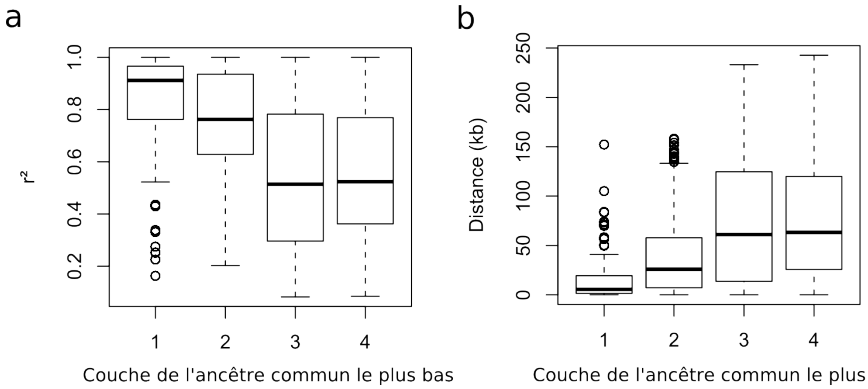

FIGURE 5.11: Relation entre le niveau de l'ancêtre commun le plus bas (ACLPB) des SNP et a) la médiane des valeurs de r^2, et b) la médiane des distances, calculées pour chaque paire de SNP.
Par exemple, la boîte à moustaches de gauche dans a) correspond aux paires de SNP dont l'ACLPB appartient à la couche 1.

5.3.3 Résultats expérimentaux et discussion

5.3.3.1 Déséquilibre de liaison de courte distance

Nous illustrons la visualisation du LD de courte distance sur le jeu de données de Daly *et al.* [19]. Ce jeu de données fournit un bon exemple de patrons complexes de LD présentant des degrés de LD variés. Il comprend 129 trios, chacun composé de deux parents et d'un enfant. Pour chaque individu, 103 SNP ont été génotypés dans la région $5q31$ (chromosome 5) et couvrent 617 *kb*.

Comparaison générale

Notre méthode basée sur la FMHCL est comparée avec deux approches classiques - la CTC de D'/LOD construite avec Haploview v4.2 et le graphique relatif aux taux de recombinaison (GTR) par cartographie fine inférés par PHASE v2.1 - et la méthode la plus avancée du point de vue de la visualisation, le textile plot. Les résultats sont présentés en figure 5.10. En dépit du fait que ces méthodes abordent la visualisation du LD de façons différentes, des tendances communes apparaissent : la plupart des SNP sont regroupés en blocs similaires entre les méthodes (voir lignes en pointillés). Dans la séquence de Daly *et al.*, Haploview a inféré 11 blocs de LD, qui sont encadrés en noir sur la CTC (voir figure 5.10a). En outre, nous observons un certain nombre de dépendances entre SNP non contigus, par exemple, entre les SNP 26 et 28 du bloc 5 et les SNP des blocs 7, 8, 9 et 10. Le GTR indique quatre points chauds de recombinaison aux positions SNP14-SNP15, SNP24-SNP25, SNP76-SNP77 et SNP91-SNP92, montrant des valeurs au delà de 10 cM/Mb (voir figure 5.10b). Ces points chauds de recombinaison définissent 4 grands blocs qui sont partiellement en adéquation avec ceux obtenus par la CTC, comme le soulignent les lignes en pointillés.

Dans le textile plot, plus la dispersion des génotypes est grande entre un homozygote (AA) et l'autre homozygote (BB) sur l'axe vertical, plus le SNP considéré est en LD avec les autres SNP de la séquence (voir figure 5.10c). Les résultats sont similaires entre textile plot, CTC et GTR. Dans le textile plot, les dispersions des génotypes sont élevées à l'intérieur des blocs de LD et faibles à l'extérieur. Outre la mise en évidence des blocs de LD, il est possible de distinguer le LD parfait ($r^2 = 1$ et $D' = 1$), comme entre le SNP1 et le SNP2, du LD complet ($r^2 < 1$ et $D' = 1$), comme observé entre le SNP2 et le SNP3. En effet, pour le premier couple de SNP, il n'y a pas de ligne reliant un homozygote au côté opposé de l'homozygote (*e.g.* GG vers AA entre le SNP1 et le SNP2), alors qu'il y en a pour la seconde paire de SNP. De plus, le textile plot offre une autre fonctionnalité absente de la CTC et du GTR. Le textile plot permet la visualisation des fréquences de génotypes multilocus : plus la ligne reliant deux génotypes élémentaires est épaisse, plus la fréquence du génotype bi-locus correspondant est élevée. Par exemple, nous observons le génotype multilocus le plus fréquent en bas du textile plot.

Le graphe de la FMHCL offre un autre point de vue du LD (voir figure 5.10d). Cette vue est assez similaire à celle de la CTC car la méthode se concentre aussi sur les dépendances entre variables. La méthode a été utilisée sur des données non phasées, mais des résultats similaires sont obtenus avec des données phasées (voir section 4.4.3, page 67). Dans le graphe, les nœuds feuilles sont des SNP (nœuds bleus), tandis que les autres nœuds (nœuds rouges) capturent les patrons de LD multilocus. Un arc entre deux nœuds (latents ou observés) indique une dépendance entre eux. Grâce au concept d'ACLPB pour chaque paire de SNP (voir figure 5.9, page 97), il est possible de tirer parti de la structure hiérarchique du modèle afin d'évaluer le LD par paires. Nous rappelons que l'ACLPB est défini pour deux nœuds v et w comme le nœud le plus bas dans l'arbre (ou la forêt) qui a à la fois v et w pour descendants. Dans la forêt, le niveau de l'ACLPB relatif à deux SNP représente la force du LD par paires existant entre eux. Les niveaux d'ACLPB correspondent à différents degrés de LD par paires et à différentes distances entre SNP (voir figure 5.11). Dans la première couche de la figure 5.11, il y a 27 ACLPB fournissant des valeurs médianes pour le coefficient r^2 et la distance de 0.91 et $5.5kb$, respectivement. Dans la seconde couche, il y a 8 ACLPB et les valeurs médianes pour le r^2 et la distance sont de 0.76 et 26 kb, respectivement. Dans les deux dernières couches, les médianes de r^2 et de la distance sont plus basses (environ 0.5 et 60 kb). Ainsi, le niveau d'un ACLPB dans le graphe véhicule une information essentielle sur une paire de SNP : plus le niveau est haut, plus le LD par paires est faible, et plus la distance séparant les SNP est grande. Les ACLPB de bas niveau représentent un LD par paires fort et de courte distance, alors que les ACLPB de haut niveau correspondent à un LD par paires faible et de longue distance. Par comparaison avec la CTC, le graphe de la FMHCL offre deux avantages pour l'affichage du LD par paires : (i) un LD non significatif entre deux SNP est reflété par l'absence de chemin les reliant, et (ii) les degrés de LD par paires sont catégorisés et visualisés hiérarchiquement par niveaux d'ACLPB. Ainsi, ces caractéristiques permettent de réduire le bruit inhérent à la stochasticité des données et d'offrir une vue du LD par paires, d'une façon plus intelligible.

LD multilocus

En plus du LD par paires, le graphe de la FMHCL présente le LD multilocus, une mesure complémentaire. Dans le graphe de la FMHCL, un cluster de LD (groupe de SNP non nécessairement adjacents) est facilement visualisé, puisqu'il est simplement représenté par une VL (nœud rouge) subsumant des nœuds feuilles (nœuds bleus correspondant à des SNP). Pour chaque cluster, l'intensité de la couleur de la VL est proportionnelle à la force de LD multilocus ϵ. Comme dans la CTC, cette représentation offre une vue globale du LD en un seul affichage. Nous observons que les distributions du LD par paires et du LD multilocus ne sont pas similaires. Nous rappelons que la distribution du LD multilocus correspond à la distribution de l'intensité des couleurs des VL, alors que la distribution du LD par paires peut être appréhendée à travers le niveau de l'ACLPB pour chaque paire de SNP. Notamment, le degré de LD multilocus ne dépend pas du niveau de la VL, contrairement au LD par paires. La distribution du LD multilocus de la FMHCL est similaire à celle du textile plot, qui calcule aussi un LD multilocus. Par exemple, sur la figure 5.10a, le premier bloc de LD inféré avec Haploview est composé de deux plus petits blocs (séparés par la première ligne en pointillés en partant de la gauche) de LD plus fort ($\epsilon = 0.83$ et 0.73, respectivement). Ces deux petits blocs peuvent être aisément visualisés dans le textile plot et la FMHCL. Le premier petit bloc montre le plus grand LD multilocus à la fois dans le textile plot et la FMHCL. Un avantage de la FMHCL sur le textile plot est que nous pouvons facilement voir le LD fort restant entre ces deux petits blocs (qui sont tous les deux représentés par une VL dans la couche 1), car les deux petits blocs sont connectés par une VL supplémentaire dans la couche 2. Des dépendances plus complexes sont observées dans la grande région centrale SNP26-SNP76, avec la présence de VL dans les couches 3 et 4. Cela illustre bien le fait que la structure hiérarchique des FMHCL peut aisément s'adapter à la nature floue des frontières des clusters de LD.

Comparaison du nombre d'éléments graphiques

Finalement, nous comparons le nombre d'éléments graphiques (NEG) entre les quatre méthodes, afin d'évaluer leur capacité à synthétiser l'information. Dans la CTC, le NEG est égal à $(n(n-1))/2$, où n est le nombre de SNP. Pour le GTR, le NEG est de $n-1$, *i.e.* le nombre de valeurs de taux recombinaison (calculées pour chaque paire de SNP adjacents). En ce qui concerne le textile plot, le NEG est compris entre $(n-1) \times 3$ et $(n-1) \times 3 \times 3$, car il y existe $n-1$ paires de SNP adjacents et 3×3 connexions possibles entre les génotypes (AA, AB et BB) pour chaque paire. Dans le graphe de la FMHCL, le NEG est égal à $sn + se$, qui est la somme du nombre de nœuds et du nombre d'arcs. Le NEG varie en fonction de la complexité de la structure de la FMHCL. Il est compris entre n ($sn = n$, $se = 0$) et $4n - 3$ ($sn = 2n - 1$, $se = 2n - 2$). Évidemment, cette comparaison est simpliste car elle ne prend pas en compte le fait que les différents méthodes ne véhiculent pas la même quantité d'information. Néanmoins, il est clairement démontré que, mis à part le GTR

qui fournit la plus grand synthèse d'information, le graphe de la FMHCL offre la vue la plus synthétique. En particulier, la comparaison du graphe de la FMHCL avec la méthode la plus similaire, la CTC, révèle que la synthèse d'information est plus élevée dans le premier (complexité linéaire) que dans le second (complexité quadratique).

5.3.3.2 Déséquilibre de liaison de longue distance

Afin d'illustrer la visualisation en présence de LD de longue distance, nous avons choisi d'étudier le complexe majeur d'histocompatibilité (CMH), une large région chromosomique incluant une famille de gènes codant pour les molécules du CMH. Le CMH joue une rôle essentiel dans le système immunitaire et auto-immunitaire. Un LD de longue distance a été observé dans la région du CMH [46, 60]. Ce LD pourrait s'expliquer par des phénomènes de balayage sélectif [6] (selective sweeps) et par l'histoire des populations (dérive génétique), mais il y existe aussi des éléments indiquant l'influence importante de la recombinaison. Afin d'étudier le LD de longue distance, nous nous sommes concentrés sur la région $[22Mb - 40Mb]$ présente sur le chromosome 6 et qui englobe le CMH. Bien que cette région contienne 14292 SNP, nous avons préféré sélectionner seulement 358 d'entre eux dans le contexte d'une démonstration hors-ligne de notre méthode. Il est toutefois possible d'analyser les 14292 SNP couvrant la région du CMH mais la visualisation d'un si grand graphe nécessite la navigation à l'intérieur du graphe à l'aide d'un logiciel tel que Tulip. Nous avons travaillé sur les 117 individus provenant de la troisième phase du projet HapMap et concernant les habitants des États Unis issus d'ancêtres ayant vécu dans le nord et l'ouest de l'Europe (CEU). Nous avons utilisé les données génotypiques (non phasées).

La carte chromosomique, la carte triangulaire de chaleur, le textile plot et le graphe de la FMHCL sont présentés en figure 5.12. Les trois méthodes de visualisation indiquent la présence d'un fort LD couvrant plusieurs mégabases dans le région centrale (située à l'intérieur des lignes en pointillés). Cette région à fort LD est entourée de régions de faible LD. Dans le textile plot, la grande région de fort LD est révélée par de fortes dispersions des génotypes, pour la plupart des SNP. Dans le graphe de la FMHCL, nous observons de grands arbres et de multiples couches dans la région centrale. En comparaison du textile plot, le graphe de la FMHCL apporte une information supplémentaire : les différentes couches permettent la distinction entre LD de courte distance et LD de longue distance. Le premier est représenté par les ACLPB de bas niveau tandis que le second est mis en avant par les ACLPB de haut niveau. Dans la FMHCL, les dépendances entre SNP distants sont facilement observables. Ce n'est pas le cas avec le textile plot, pour lequel les SNP sont ordonnés le long du chromosome. Pour résoudre ce problème, Kumasaka *et al.* emploient un algorithme de clustering hiérarchique des variables afin de réarranger les positions des SNP le long de l'axe horizontal, permettant de révéler le LD entre SNP distants. Néanmoins, l'inconvénient est que l'ordre physique des SNP est perdu.

6. Désigne la réduction de la variabilité de la région chromosomique entourant une mutation ayant subi une importante pression sélective positive.

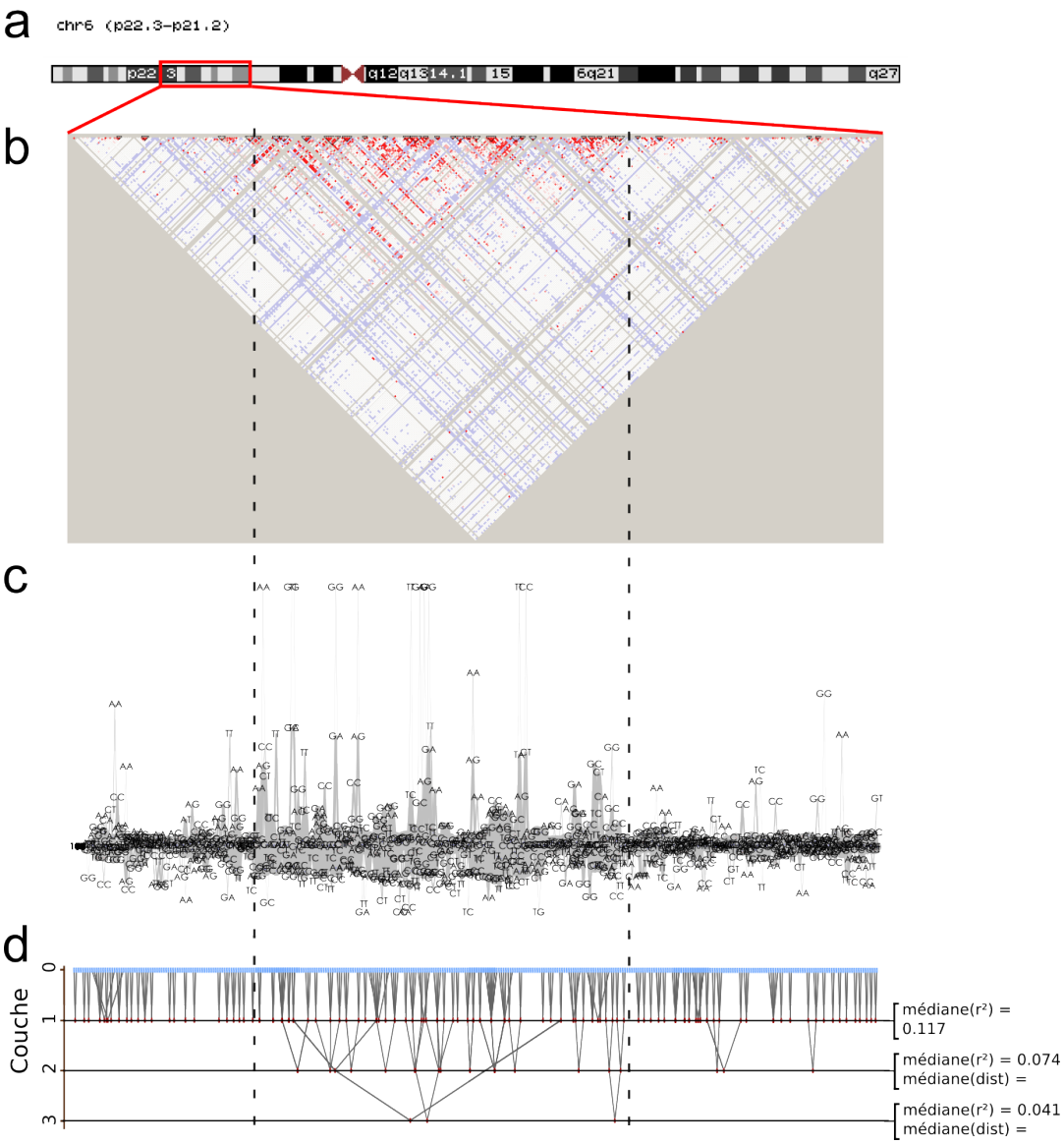

FIGURE 5.12: Visualisation du déséquilibre de liaison de longue distance pour la région $[22Mb - 40Mb]$, chromosome 6, entourant le complexe majeur d'histocompatibilité : a) carte chromosomique construite avec UCSC Genome Browser, b) carte triangulaire de chaleur du D'/LOD construite avec Haploview v4.2, c) textile plot et d) forêt de modèles hiérarchiques à classes latentes visualisée à l'aide de Tulip.
Pour chaque couche ℓ, la médiane des valeurs de r^2 (resp. distances) est calculée sur toutes les paires de SNP ayant leur ancêtre commun le plus bas dans la couche ℓ.

5.3.3.3 Déséquilibre de liaison pangénomique

La visualisation du LD pangénomique peut être réalisée par apprentissage de la FMHCL à l'aide CFHLC+, puis navigation dans l'arbre correspondant sous Tulip. Nous illustrons la visualisation du LD pour le chromosome 1 dans la population CEU (voir section 5.3.3.2, page 104). Le jeu de données se compose de 117 individus (génotypes non phasés) et de 101100 SNP. L'apprentissage de la FMHCL a été contraint par une distance physique maximale entre SNP (ou VL) de 100 kb. CFHLC+ a été exécuté sur un ordinateur personnel standard (3.8 GHz, 3.3 Go de RAM). Seulement 9 heures et 1.1 GB ont été nécessaires pour construire la FMHCL relative à tout le chromosome.

En figure 5.13, le graphe de la FMHCL a été dessiné. La navigation dans le graphe à l'aide de zooms successifs permet de modifier la résolution de la visualisation. Lorsque la fonction de zoom est inactive, le chromosome est simplement représenté par une ligne bleue. Néanmoins, si nous zoomons une première fois sur le graphe, la structure globale devient apparente. Dans la seconde vue, le LD de longue distance entre SNP séparés par 50-100 kb est aisément visualisé. Dans la troisième fenêtre, il devient possible de distinguer chaque cluster de LD dans le graphe (*i.e.* chaque sous-arbre de la FMHCL), afin de voir le nombre de couches et le degré de connectivité de la FMHCL. Le degré de LD par paires est révélé par le niveau d'ACLPB, tandis que le degré de LD multilocus est montré par l'intensité de la couleur des VL. Finalement, si nous zoomons encore, nous pouvons voir la position des SNP et la valeur précise du LD multilocus, qui sont écrites à l'intérieur des nœuds bleus (SNP) et des nœuds rouges (VL), respectivement.

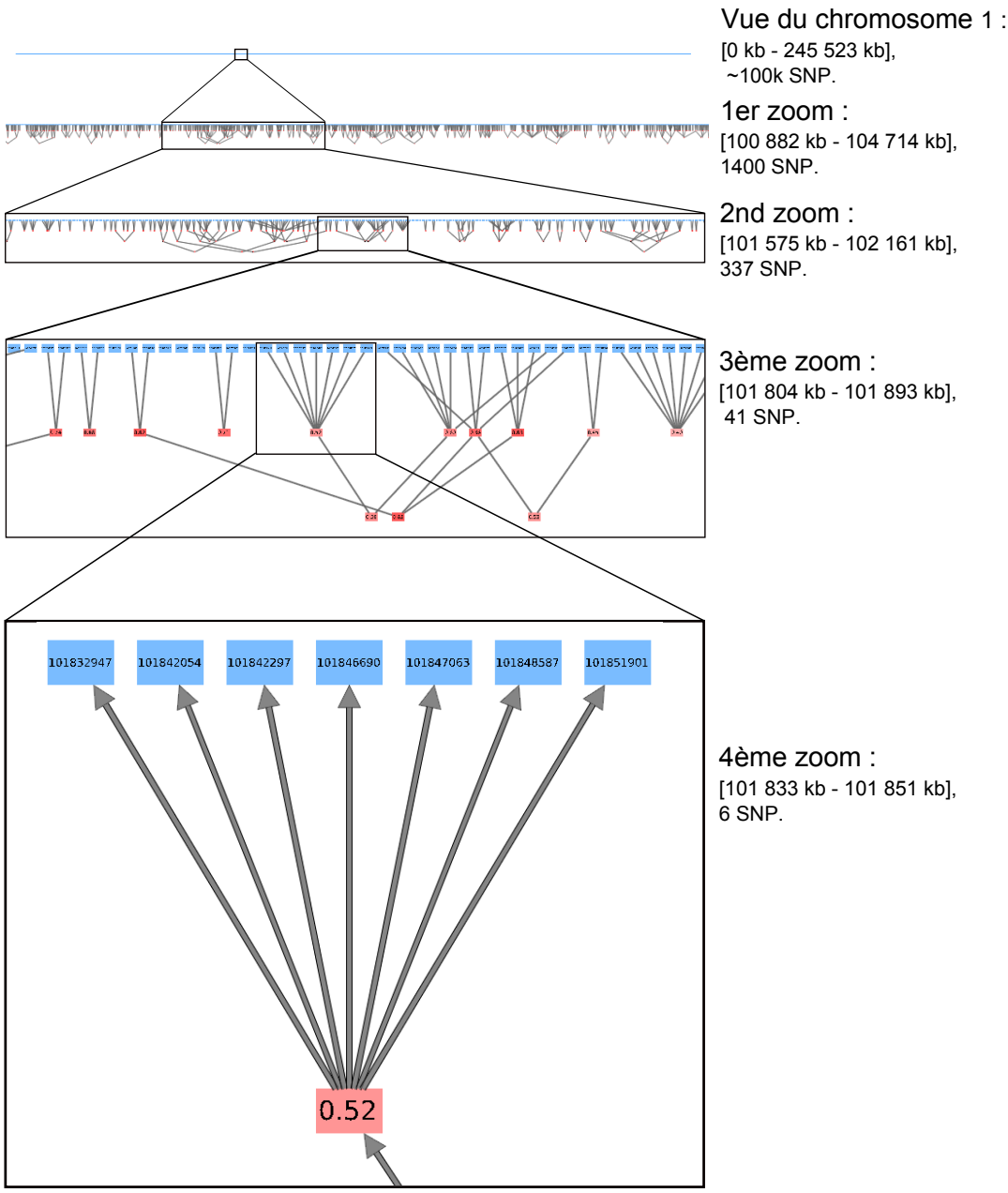

FIGURE 5.13: Visualisation pangénomique du déséquilibre de liaison (LD) pour le chromosome 1 : navigation à l'aide de zooms successifs à l'intérieur du graphe de la FMHCL.
Les positions des SNP sont affichées à l'intérieur des nœuds bleus. Le degré du LD multilocus relatif aux sous-arbres enracinés dans les variables latentes peut être lu à l'intérieur des nœuds rouges.

5.3.4 Conclusion

Notre méthode basée sur la FMHCL pour la visualisation de la structure spatiale du LD s'est révélée capable de fournir une vue synthétique et intuitive dans les trois contextes principaux : l'analyse de LD de courte distance, de longue distance et de LD pangénomique. Notre approche se concentre sur les dépendances entre variables, et c'est la raison pour laquelle elle est très similaire à la CTC. Néanmoins, elle dispose de nombreux avantages vis-à-vis de la CTC : (i) le LD par paires et le LD multilocus sont tous deux visualisés, (ii) l'information est hiérarchiquement structurée, permettant une distinction intelligible entre LD de courte distance et LD de longue distance, et (iii) l'information la plus pertinente est véhiculée, puisque seul le LD significatif est montré et le degré de LD par paires est catégorisé en différents niveaux. En comparaison du textile plot, notre outil de visualisation montre plusieurs inconvénients, contrebalancés par de nombreux avantages. Par exemple, bien que le graphe de la FMHCL ne permette pas de distinguer entre LD complet et LD parfait, ni de montrer les fréquences des génotypes, il identifie clairement le LD de longue distance sans le besoin d'un réarrangement de l'ordre des SNP sur la séquence, comme requis par la méthode du textile plot. En fait, le textile plot et le graphe de la FMHCL sont des approches complémentaires pour l'étude de la structure du LD.

5.3.5 Perspectives

Les perspectives de recherche concernent plusieurs aspects. Tout d'abord, toute l'information portée par la FMHCL n'a pas été employée dans l'approche de visualisation développée. Par exemple, les distributions de probabilité conditionnelle et *a priori* apprises par CFHLC+ pourraient autoriser un aperçu de la distribution des génotypes multilocus, et surtout celui de la distribution des clusters de génotypes (modélisés par les états d'une VL). Enfin, il serait intéressant pour le généticien de disposer d'un logiciel intégré équipé d'un interface conviviale, tel que proposé par Haploview ou par le logiciel du textile plot, afin d'apprendre les FMHCL, de les visualiser et de lancer des analyses d'association génétique, en mode interactif.

6
Conclusion et Perspectives

L'avènement des nouvelles technologies génomiques à haut débit, notamment les puces de génotypage de marqueurs génétiques SNP, a constitué un tournant décisif pour l'analyse génétique des maladies multifactorielles, telles que le diabète et l'asthme. Cette analyse constitue un enjeu de santé publique majeur (diagnostic, prévention et thérapie de ces maladies). Dans ce contexte, les études d'association pangénomiques ont émergé comme approches pertinentes pour la localisation précise des facteurs génétiques impliqués dans l'apparition de ces maladies. Malheureusement, la forte complexité des données provenant de ces études, combinée à leur grand volume, constitue encore un frein important pour la découverte des facteurs génétiques. L'existence des dépendances entre SNP, appelées aussi déséquilibre de liaison (LD), représente une des clés du succès de ces études. Réduction de dimension des données, visualisation du LD, découverte de causalité, étude de la structure de la population sont par exemple autant d'applications différentes possibles, en aval, de la modélisation de ces dépendances. Bien que le projet HapMap ait révélé une structure simple en blocs des dépendances, une analyse plus approfondie révèle que la structure est en fait bien plus complexe. Les travaux de recherche présentés dans ce mémoire sont nés de la volonté de développer de nouveaux outils bioinformatiques offrant une modélisation fine du LD, tout en garantissant un passage à l'échelle des données pangénomiques. Ce chapitre porte naturellement sur les conclusions des travaux de recherche réalisés, mais il ouvre aussi sur les principales perspectives scientifiques offertes.

6.1 Conclusion

Les travaux abordés par ce mémoire s'inscrivent à la confluence de plusieurs domaines scientifiques différents : la biologie (génétique), les mathématiques (statistiques) et l'informatique (intelligence artificielle). C'est la raison pour laquelle,

pour faciliter la compréhension de l'ouvrage et la dissémination des idées présentées, nous avons porté une attention particulière à introduire les notions fondamentales de chaque discipline à des non-spécialistes, dans les deux premiers chapitres.

Nous avons d'abord introduit le LD dans le contexte des études d'association pangénomiques. Pour cela, les concepts essentiels de biologie moléculaire et cellulaire, puis les notions fondamentales de génétique ont été introduits. Ensuite, dans le cadre des maladies génétiques complexes, les études d'association pangénomiques, ainsi que les nombreux défis qu'elles soulèvent, ont été présentés. Enfin, nous avons conclu sur les raisons de développer de nouveaux outils provenant du monde de l'intelligence artificielle, tels que les modèles graphiques probabilistes.

Ensuite, les principaux concepts relatifs aux modèles graphiques probabilistes (MGP) ont été présentés. A titre de rappel, les définitions fondamentales en théorie des probabilités, théorie de l'information et en théorie des graphes ont été introduites. Puis deux grandes familles de MGP ont été abordées : les réseaux bayésiens (RB), faisant l'objet des travaux de cette thèse, et les réseaux de Markov. Les RB ont été d'abord présentés de manière simple et intuitive, puis définis plus formellement. Les trois grandes problématiques de l'utilisation de ces modèles ont ensuite été détaillées : inférence probabiliste, apprentissage de paramètres et apprentissage de structure. Enfin, nous avons terminé en évoquant les motivations de l'application des MGP au traitement des données génétiques, notamment en amont des études d'association pangénomiques.

Les contributions de ce manuscrit concernent deux grands axes : d'une part, un axe théorique focalisé sur la modélisation pangénomique du LD à l'aide de MGP, et d'autre part, un axe appliqué qui découle naturellement de cette modélisation.

Sur le plan théorique, nous avons proposé un nouveau cadre d'analyse pangénomique du LD basé sur la modélisation à l'aide de forêts de modèles hiérarchiques à classes latentes (FMHCL). Les FMHCL présentent un certain nombre d'avantages. Parmi les plus importants, nous pouvons citer : la capacité des FMHCL à traiter directement des données sans inférer les haplotypes, ce qui permet de limiter de façon importante les temps de calculs, la hiérarchisation des différents degrés de LD par paires et multilocus, l'interprétation biologique des variables latentes et la prise en compte de la nature floue du LD. La comparaison avec d'autres méthodes (méthode de Daly *et al.*, Gerbil, HaploBlock et méthode de Zhang *et al.*), sur données réelles, a révélé que les FMHCL offrent une modélisation plus fine des dépendances entre SNP. En outre, les FMHCL se sont montrées capables de reproduire fidèlement les dépendances dans des contextes de patrons de LD variés, sur des données simulées. Pour l'apprentissage des FMHCL à partir de données pangénomiques, nous avons développé un nouvel algorithme permettant le passage à l'échelle. Cet algorithme combine plusieurs approches efficaces en apprentissage : construction locale et ascendante du modèle, découverte de dépendances entre variables pour l'identification de variables latentes et emploi de l'algorithme expectation-maximization pour l'inférence de ces

variables latentes. Deux versions de l'algorithme ont été proposées : la première nécessite le partitionnement du chromosome en larges fenêtres contiguës, tandis que la dernière, plus avancée, applique une fenêtre glissante le long du chromosome. Les deux versions se sont montrées capables de traiter des données pangénomiques et l'analyse empirique de leur complexité a montré la linéarité en temps lorsque le nombre de SNP augmente.

Sur le plan appliqué, deux axes de recherche ont été privilégiés : la recherche de causalité et la visualisation du LD. Les premiers travaux constituent une étude préliminaire systématique visant à étudier la capacité des FMHCL pour la recherche de causalité. Nous avons analysé cette capacité dans le contexte de la génétique d'association. Les bases du développement de nouvelles méthodes de recherche dédiées à la découverte de SNP causaux ont été établies. Pour cet objectif, nous avons défini le concept d'association génétique indirecte qui constitue la clé de notre méthode de recherche de causalité. Dans le graphe, l'identification des ancêtres du SNP causal permet de cibler les régions causales potentielles, et plus précisément, les SNP causaux potentiels. Lors d'études sur des données simulées et réelles, nous avons montré qu'une capture efficace des associations génétiques indirectes est réalisée grâce à deux propriétés majeures : (i) les nœuds ancêtres du SNP causal réussissent à capturer les associations indirectes avec le phénotype ; (ii) à l'inverse, les nœuds non-ancêtres du SNP causal montrent globalement de très faibles associations.

Nous avons aussi développé une nouvelle représentation visuelle du LD. Cette représentation est fondée sur la nature graphique et hiérarchique de la FMHCL et a permis le développement d'un outil de visualisation intuitif et synthétique des dépendances spatiales entre SNP. L'outil s'est révélé efficace pour les trois principales situations que peut rencontrer le généticien : l'analyse de LD de courte distance, de longue distance et de LD pangénomique. En comparaison de la carte triangulaire de chaleur, une méthode classique de visualisation (implémentée par exemple dans Haploview), notre outil a montré les atouts suivants : (i) le LD par paires et le LD multilocus sont tous deux visualisés, (ii) l'information est hiérarchiquement structurée, permettant une distinction intelligible entre LD de courte distance et LD de longue distance, et (iii) l'information la plus pertinente est véhiculée, puisque seul le LD significatif est montré et le degré de LD par paires est catégorisé en différents niveaux. La comparaison de notre outil avec la méthode plus sophistiquée du textile plot montre que les deux approches sont complémentaires pour l'étude de la structure du LD. Par exemple, notre outil permet de visualiser beaucoup plus simplement le LD de longue distance, alors qu'il ne permet pas de distinguer le LD complet du LD parfait comme le fait le textile plot.

6.2 Perspectives

Modélisation du déséquilibre de liaison

A l'issue des travaux présentés dans ce mémoire de thèse, un certain nombre de perspectives s'offrent à nous. Les premières concernent la modélisation pangénomique du LD à l'aide de FMHCL. Un point délicat lors de l'emploi de cette modélisation demeure l'apprentissage à partir de données volumineuses. Le partitionnement en cliques de SNP à l'aide de l'algorithme CAST pourrait être remplacé par une méthode plus précise. Par exemple, le clustering spectral est considéré comme une méthode très efficace et simple à mettre en œuvre pour regrouper des variables corrélées entre elles. Le graphical Lasso représente une autre alternative. Il est fondé sur l'estimation régularisée du coefficient de corrélation partiel entre SNP. Cette estimation évalue l'indépendance conditionnelle entre SNP dans les réseaux creux. Elle est plus pertinente que l'information mutuelle par paires de SNP car elle permet de ne pas prendre en compte les dépendances indirectes entre SNP. Outre la découverte de cliques de SNP, une étape d'apprentissage global des paramètres de la FMHCL pourrait être ajoutée à la fin de CFHLC. En effet, dans l'algorithme CFHLC, l'apprentissage des paramètres est réalisé de manière locale en utilisant les paramètres appris pour chaque modèle à classes latentes. Si l'on utilise les paramètres appris localement pour initialiser l'apprentissage global, alors ce dernier pourrait être réalisé en un temps raisonnable et autorisant le passage à l'échelle.

Par ailleurs, dans notre modélisation, nous ne nous sommes pas intéressés à inférer la phase gamétique des données génotypiques. Nous avons postulé que nous la connaissions *a priori*. Néanmoins, une extension éventuelle de notre modèle pourrait consister à inférer la phase.

Bien que nos travaux se soient focalisés sur les FMHCL, l'utilisation d'autres classes de MGP mériterait aussi d'être explorée. Par exemple, l'approche de Verzilli *et al.* est performante car elle présente l'avantage d'être très rapide et d'intégrer une étape supplémentaire de localisation efficace de SNP causaux. Cette approche met en œuvre le parcours de l'espace des graphes composés de cliques disjointes, afin d'assurer la décomposabilité du graphe et de permettre ainsi un calcul simple de la vraisemblance. Cependant, depuis les travaux récents de Thomas et Green, il est désormais possible d'échantillonner directement et efficacement dans l'espace général des graphes décomposables. Ainsi l'intégration de l'échantillonneur de Thomas et Green dans l'approche de Verzilli *et al.* constitue une piste prometteuse pour la modélisation du LD combinée à la recherche d'associations génétiques.

Découverte de causalité

Au cours de cette thèse, une étude systématique des FMHCL pour la recherche de causalité dans les études de génétique d'association a été menée mais elle ne représente qu'une étape préliminaire dans le développement d'outils de recherche d'associations.

Pour cet objectif, plusieurs points sont à développer. Premièrement, grâce à la structure en forêt du modèle, il est possible d'inférer de manière efficace (*i.e.* en temps linéaire) les valeurs des variables latentes en utilisant la loi jointe du modèle complet (*i.e.* de la FMHCL), et non plus simplement en employant les distributions des modèles à classes latentes apprises localement. Deuxièmement, en comparaison de l'approche fréquentiste employée lors de ces travaux, le paradigme bayésien offre de nombreux atouts pour la recherche d'association car il permet d'incorporer facilement de nombreuses connaissances *a priori*, telles que l'espérance *a priori* du nombre de SNP causaux. Troisièmement, la structure hiérarchique latente du modèle pourrait être utilisée pour réaliser une recherche des facteurs causaux selon une approche descendante (des couches hautes vers les couches basses), afin d'identifier les SNP causaux. L'idée est qu'il est probable, lorsqu'on observe une VL associée au phénotype, qu'il existe un chemin partant de cette VL et dans lequel toutes les variables (VL et SNP) soient aussi associées au phénotype. Le parcours de ce chemin, qui peut être réalisé au moyen d'un parcours en profondeur (depth first search), devrait guider la découverte de SNP causaux.

Visualisation des dépendances

Notre méthode basée sur les FMHCL pour la visualisation du LD s'est révélée capable de fournir une vue synthétique et intuitive des dépendances spatiales entre SNP. Néanmoins, il reste plusieurs améliorations à apporter. Tout d'abord, toute l'information portée par les FMHCL n'a pas été employée dans l'approche de visualisation développée. Par exemple, les distributions de probabilité conditionnelle et *a priori* apprises par CFHLC pourraient fournir un aperçu de la distribution des génotypes multilocus, et surtout celui de la distribution des clusters de génotypes (modélisés par les états d'une VL). Par ailleurs, afin d'assurer une utilisation simple pour l'utilisateur, la méthode de visualisation développée devrait être intégrée dans un logiciel équipé d'une interface conviviale, tel que proposé par Haploview, afin d'apprendre le modèle, de le visualiser et de lancer en mode interactif des analyses de génétique d'association.

Outre les applications développées lors de ces travaux de doctorat, l'utilisation des FMHCL pour l'étude de l'effet de la structure de la population sur le LD mériterait d'être explorée grâce aux variables latentes présentes dans les couches de haut niveau. Par exemple, il pourrait être ainsi possible d'analyser les mélanges de sous-populations dans les études d'association. Enfin, le développement des FMHCL ne devrait pas être cantonné exclusivement au domaine de la génétique. En effet, ces modèles pourraient représenter des outils généraux pour l'analyse de données spatialement structurées, telles que les données spectrales ou certaines données d'écologie des populations par exemple.

Annexes

A

Analyse et passage à l'échelle de CFHLC

A.1 Temps d'exécution moyen *versus* nombre de variables

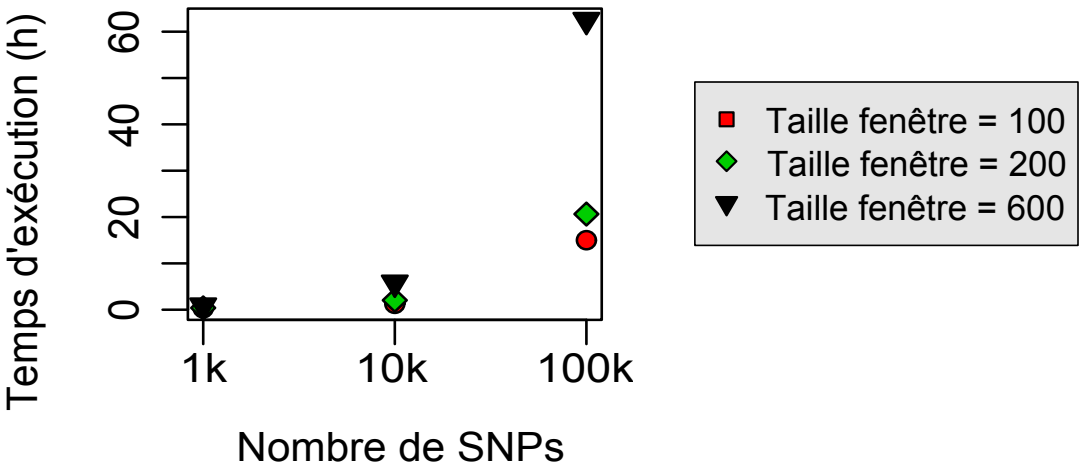

FIGURE A.1: Temps d'exécution moyen *versus* nombre de variables. Les paramètres de CFHLC sont : $s = 100$ SNP, $a = 0.2$, $b = 2$, $card_{max} = 20$, $t_{CAST} = 0.95$, $t_{MI} = quantile_{MI}(0.5)$ *et* $t = 0.5$ (pour la description des paramètres, voir section 4.3.3.6, chapitre 4, page 61).

Cette annexe décrit le temps d'exécution moyen de CFHLC en fonction du nombre de SNP à traiter. Dans le cas le plus problématique (100 k SNP), seulement 15 heures sont nécessaires lorsque la taille de la fenêtre s est fixée à 100. Pour le même jeu de données, dans les cas où $s = 200$ et $s = 600$, les temps d'exécution sont de 20.5 h et 62.5 h, respectivement. En ce qui concerne le cas 10 k SNP, les temps d'exécution sont de 1.3 h, 2 h et 5.8 h pour $s = 100$, $s = 200$ et $s = 600$, respectivement.

A.2 Influence de la taille de la fenêtre sur le temps d'exécution

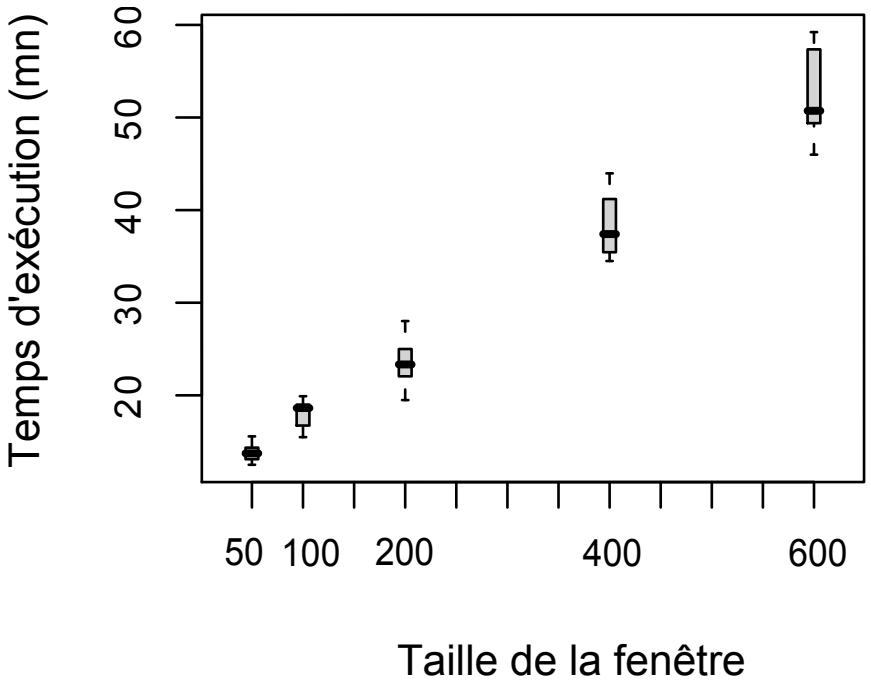

FIGURE A.2: Influence de la taille de la fenêtre sur le temps d'exécution. Les paramètres de CFHLC sont : $a = 0.2$, $b = 2$, $card_{max} = 20$, $t_{CAST} = 0.95$, $t_{MI} = quantile_{MI}(0.5)$ *et* $t = 0.5$ (pour la description des paramètres, voir section 4.3.3.6, chapitre 4, page 61).

Cette annexe décrit plus précisément l'influence de la taille de la fenêtre sur le temps d'exécution. De manière assez surprenante, nous observons une tendance linéaire du temps d'exécution. Il s'agit en fait de deux phénomènes qui agissent en sens contraires. Le premier augmente le nombre de modèles à classes latentes (MCL) à apprendre : plus la taille de la fenêtre est élevée, plus il y aura de couches dans la forêt de modèles hiérarchiques à classes latentes (FMHCL). Le second phénomène entraîne une diminution du nombre de MCL à apprendre : pour des paramètres de CAST constant (seuil d'information mutuelle t_{MI} et seuil de similarité t_{CAST}), plus le nombre de variables à partitionner est grand, plus les clusters sont grands. De grands clusters ont tendance à diminuer le nombre de MCL de la FMHCL.

A.3 Nombre de variables par couche dans la forêt de modèles hiérarchiques à classes latentes

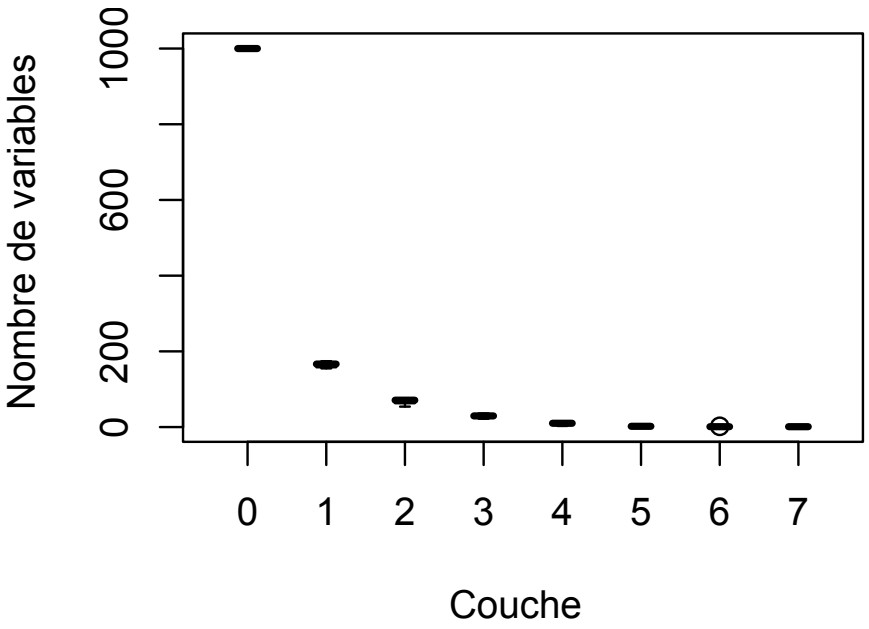

FIGURE A.3: Nombre de variables par couche dans la forêt de modèles hiérarchiques à classes latentes. 1000 SNP traités. Les paramètres de CFHLC sont : $s = 100$ SNP, $a = 0.2$, $b = 2$, $card_{max} = 20$, $t_{CAST} = 0.95$, $t_{MI} = quantile_{MI}(0.5)$ *et* $t = 0.5$ (pour la description des paramètres, voir section 4.3.3.6, chapitre 4, page 61).

Nous observons une diminution importante du nombre de variables par couche, lorsque le niveau de la couche augmente. 64% des variables latentes (VL) sont présentes dans la première couche. La diminution du nombre de VL entre la première et la seconde couche est de 60%.

A.4 Influence de la taille de la fenêtre sur le nombre de racines

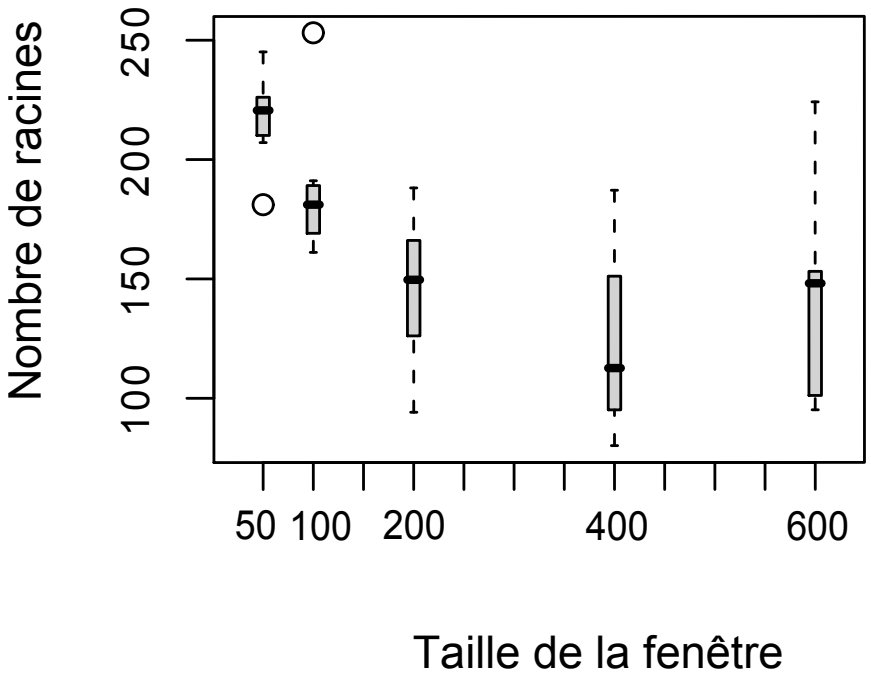

FIGURE A.4: Influence de la taille de la fenêtre sur le nombre de racines. 1000 SNP traités. Les paramètres de CFHLC sont : $a = 0.2$, $b = 2$, $card_{max} = 20$, $t_{CAST} = 0.95$, $t_{MI} = quantile_{MI}(0.5)$ *et* $t = 0.5$ (pour la description des paramètres, voir section 4.3.3.6, chapitre 4, page 61).

Cette annexe met en évidence la diminution du nombre de variables par couche, de la plus basse à la plus élevée. Cette diminution se traduit par un plus petit nombre de variables dont l'association est à tester vis-à-vis de la maladie : de 1000 variables observées à moins de 200 racines dans le cas $s = 100$. Dans ce cas, l'algorithme CFHLC permet une réduction du nombre de variables de plus de 80%.

A.5 Influence de la taille de la fenêtre sur le nombre de variables latentes

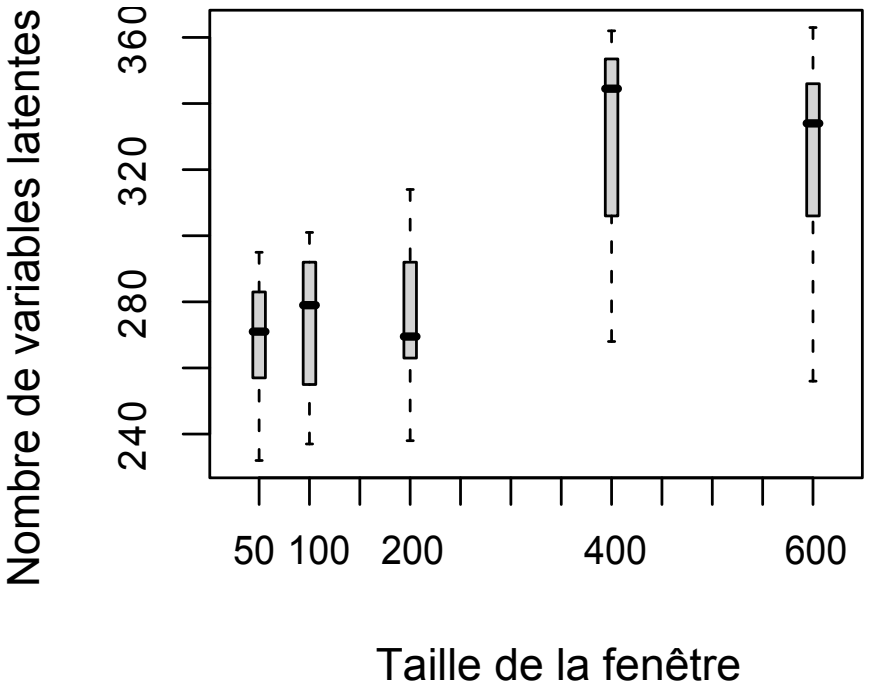

FIGURE A.5: Influence de la taille de la fenêtre sur le nombre de variables latentes. 1000 SNP traités. Les paramètres de CFHLC sont : $a = 0.2$, $b = 2$, $card_{max} = 20$, $t_{CAST} = 0.95$, $t_{MI} = quantile_{MI}(0.5)$ *et* $t = 0.5$ (pour la description des paramètres, voir section 4.3.3.6, chapitre 4, page 61).

Le nombre de variables latentes (VL) augmente avec la taille de la fenêtre. Cette augmentation provient du fait qu'un plus grand nombre de dépendances peut être pris en compte dans les couches supérieures. En moyenne, autour de 270 VL réparties dans 5 à 6 couches sont observées pour le cas $s = 100$, alors que 340 VL et 8 couches sont identifiées pour le cas $s = 600$.

A.6 Influence de la taille de la fenêtre sur le nombre de couches

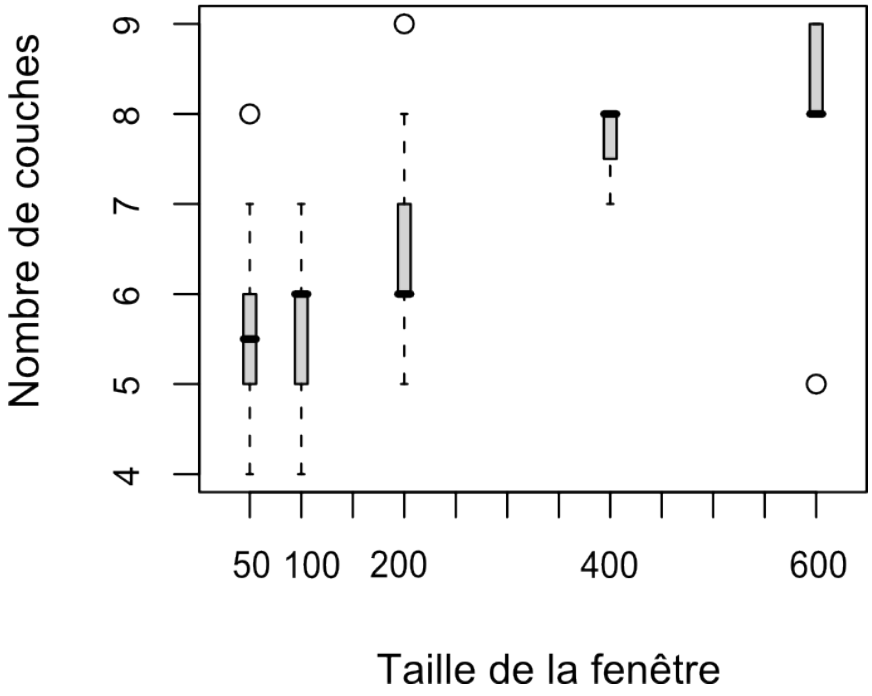

FIGURE A.6: Influence de la taille de la fenêtre sur le nombre de couches. 1000 SNP traités. Les paramètres de CFHLC sont : $a = 0.2$, $b = 2$, $card_{max} = 20$, $t_{CAST} = 0.95$, $t_{MI} = quantile_{MI}(0.5)$ *et* $t = 0.5$ (pour la description des paramètres, voir section 4.3.3.6, chapitre 4, page 61).

De la même manière que pour le nombre de variables latentes, le nombre de couches augmente avec la taille de la fenêtre.

A.7 Influence de la taille de la fenêtre sur le nombre de variables latentes par couche

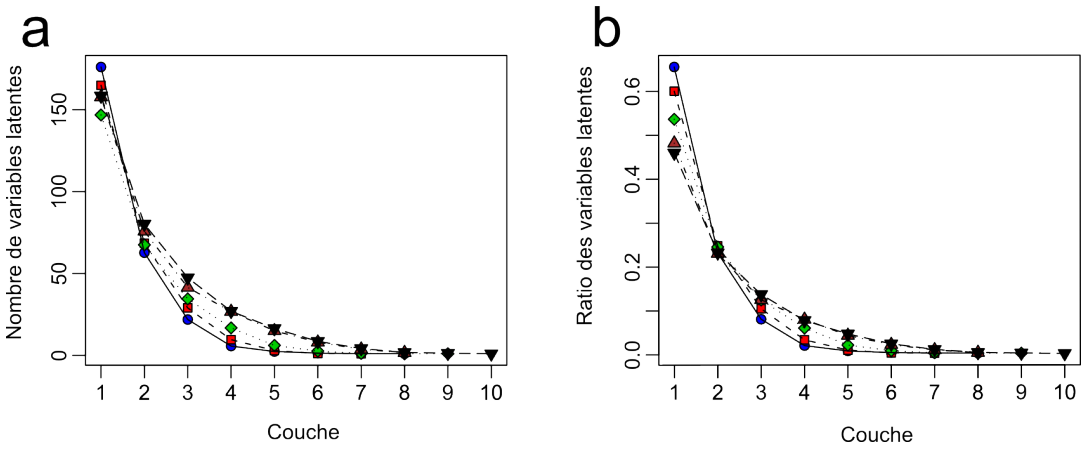

FIGURE A.7: Influence de la taille de la fenêtre sur le nombre de variables latentes (VL) et sur le ratio de VL par couche (ratio = nombre de VL par couche sur le nombre total de variables). (a) Nombre moyen de VL par couche de la forêt de modèles hiérarchiques à classes latentes (FMHCL). (b) Ratio moyen de VL par couche de la FMHCL. 1000 SNP traités. Les paramètres de CFHLC sont : $s = 100$ SNP, $a = 0.2$, $b = 2$, $card_{max} = 20$, $t_{CAST} = 0.95$, $t_{MI} = quantile_{MI}(0.5)$ *et* $t = 0.5$ (pour la description des paramètres, voir section 4.3.3.6, chapitre 4, page 61).

Cette annexe souligne l'influence de la taille de la fenêtre sur (a) le nombre de VL et sur (b) le ratio du nombre de VL d'une couche donnée au nombre total de VL. Nous observons que pour toutes les couches, exceptée la première, le nombre (et le ratio) de VL est plus grand lorsque la taille de la fenêtre est plus importante. Cela s'explique de la manière suivante : lorsque le nombre de variables observées augmente, *i.e.* lorsque la taille de la fenêtre s'accroît, pour des paramètres constants de l'algorithme de partitionnement, le nombre de clusters identifiés est plus faible (plus grands clusters).

A.8 Influence de la taille de la fenêtre sur l'information mutuelle normalisée moyenne

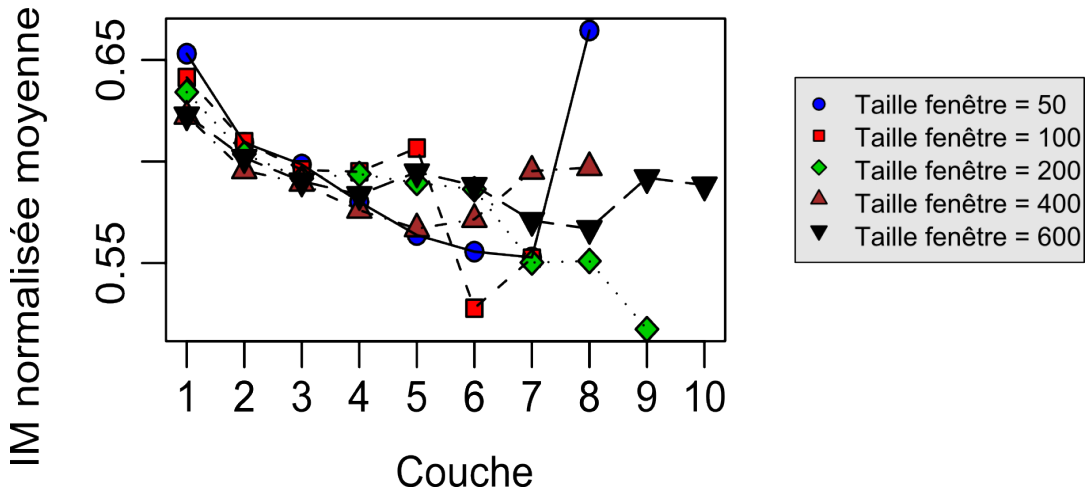

FIGURE A.8: Influence de la taille de la fenêtre sur l'information mutuelle normalisée moyenne ($\mathcal{C}$) pour chaque couche de la forêt de modèles hiérarchiques à classes latentes. 1000 SNP traités. Les paramètres de CFHLC sont : $a = 0.2$, $b = 2$, $card_{max} = 20$, $t_{CAST} = 0.95$, $t_{MI} = quantile_{MI}(0.5)$ *et* $t = 0.5$ (pour la description des paramètres, voir section 4.3.3.6, chapitre 4, page 61).

Cette annexe montre comment l'information diminue lorsque le numéro de la couche augmente. Pour le cas $s = 600$, les quatre premières couches montrent des valeurs moyennes du critère d'information mutuelle $\mathcal{C}$ voisines de 0.62, 0.60, 0.59 et 0.58. Dans les couches plus hautes, $\mathcal{C}$ est au moins égal à 0.52 et 0.56 pour les cas $s = 100$ et $s = 600$, respectivement. Ainsi, nous pouvons constater que la perte d'information est contrôlée.

A.9 Influence des paramètres a et b sur l'information mutuelle normalisée moyenne

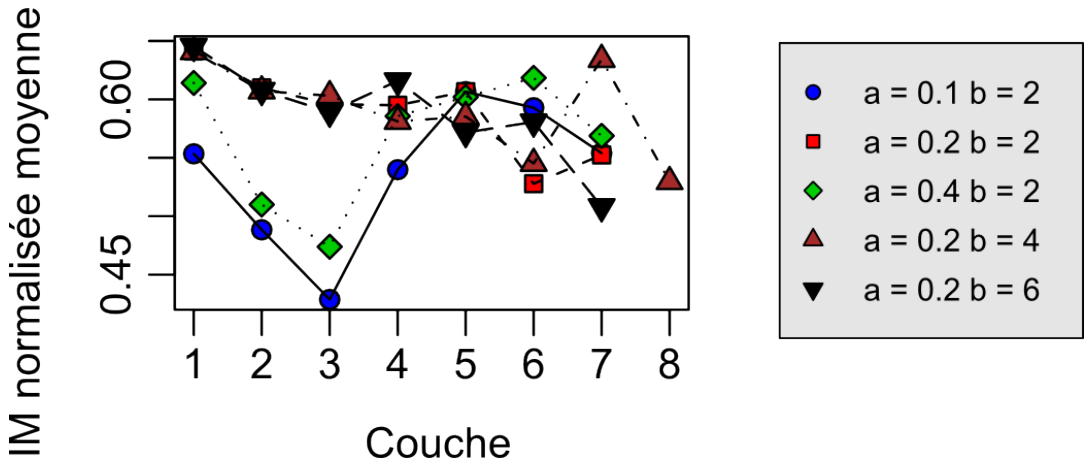

FIGURE A.9: Influence des paramètres a et b sur l'information mutuelle normalisée moyenne, par chaque couche. 1000 SNP traités. Les paramètres de CFHLC sont : $s = 100\ SNP$, $card_{max} = 20$, $t_{CAST} = 0.95$, $t_{MI} = quantile_{MI}(0.5)$ *et* $t = 0.5$ (pour la description des paramètres, voir section 4.3.3.6, chapitre 4, page 61).

De façon peu surprenante, on observe que l'information mutuelle augmente avec les paramètres a et b, car les variables ayant de grandes cardinalités peuvent capturer plus d'information sur leur nœuds enfants.

B

Passage à l'échelle de CFHLC+

B.1 Temps d'exécution *versus* nombre de SNP traités

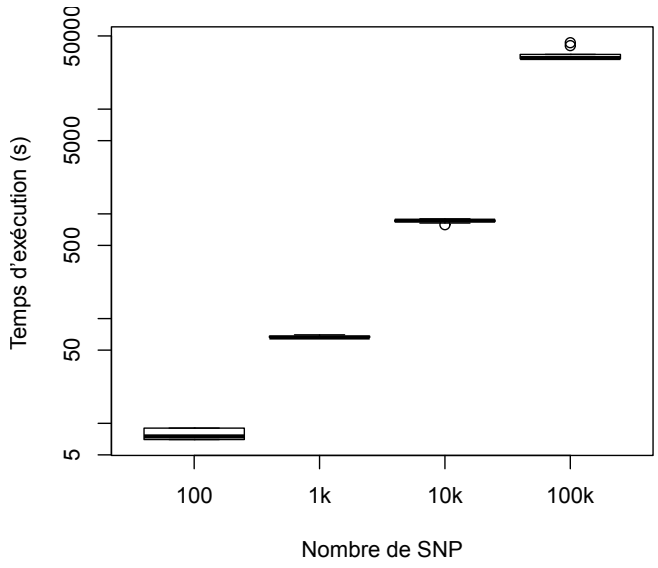

FIGURE B.1: Temps d'exécution *versus* nombre de SNP traités.
Les données traitées proviennent de la population CEU et couvrent le chromosome 1. Les paramètres de CFHLC+ sont : $s = 100kb$, $a = 0.3$, $b = 2$, $card_{max} = 10$, $t_{CAST} = 0.95$, $t_{MI} = quantile_{MI}(0.9)$ *et* $t = 0.1$ *(pour une description des paramètres, voir section 4.3.3.6, chapitre 4, page 61). Le nombre de réinitialisations aléatoires de l'algorithme EM est de 10.*

Le temps d'exécution de CFHLC+ est linéaire avec le nombre de SNP traités.

B.2 Temps d'exécution *versus* taille de la fenêtre

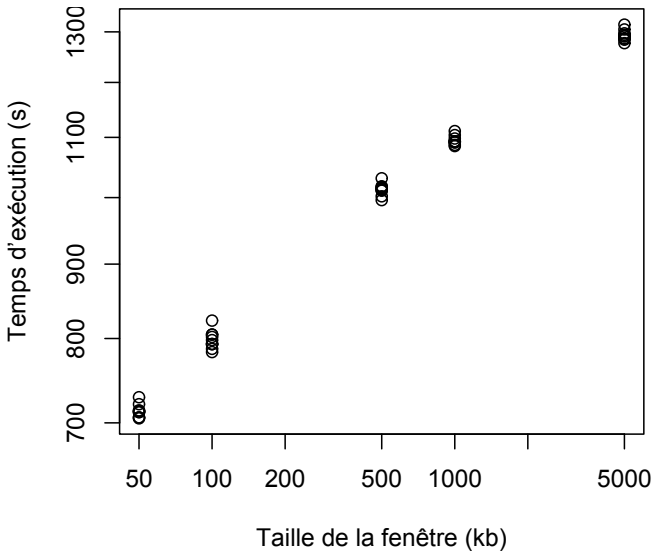

FIGURE B.2: Temps d'exécution *versus* taille de la fenêtre.
Les données traitées proviennent de la population CEU et couvrent le chromosome 1. Voir figure B.1 pour la description des paramètres.

Le temps d'exécution de CFHLC+ est linéaire avec la taille de la fenêtre.

Liste des abréviations

A	Nœud ancêtre du SNP causal
ACLPB	Ancêtre commun le plus bas
ADN	Acide désoxyribonucléique
AGI	Association génétique indirecte
AIC	Akaike information criterion (critère d'information d'Akaike)
AP	Apprentissage des paramètres
ARN	Acide ribonucléique
AS	Apprentissage de structure
BIC	Bayesian information criterion (critère d'information bayésien)
CAH	Classification ascendante hiérarchique
CDCV	Common disease-common variant (maladie commune-variants communs)
CDRV	Common disease-rare variant (maladie commune-variants rares)
CFHLC	Construction of Forests of Hierarchical Latent Class model
CTC	Carte triangulaire de chaleur
CTCN	Carte triangulaire de chaleur du niveau de l'ACLPB
CTCR	Carte triangulaire de chaleur du r^2
EAP	Étude d'association pangénomique
EM	Expectation-maximization
FMHCL	Forêt de modèles hiérarchiques à classes latentes
FP	Faux positif
GDG	Graphe des données génotypiques
GDH	Graphe des données haplotypiques
GI	Graphe d'intervalle
GOSC	Graphe orienté sans circuit
GTR	Graphique relatif aux taux de recombinaison

IP	Inférence probabiliste
LD	Linkage disequilibrium (déséquilibre de liaison)
MC	Méthode basée sur les contraintes
MCL	Modèle à classes latentes
MCMC	Markov chain Monte Carlo
MGP	Modèle graphique probabiliste
MHCL	Modèle hiérarchique à classes latentes
MMC	Modèle de Markov caché
MS	Méthode basée sur le score
MV	Maximum de vraisemblance
N	Nœud non-ancêtre du SNP causal
NA	Nœud non-ancêtre du SNP causal, présent dans l'arbre causal
NE	Nœud non-ancêtre du SNP causal, présent dans l'arbre non-causal
NEG	Nombre d'éléments graphiques
RB	Réseau bayésien
RM	Réseau de Markov
RMD	Réseau de Markov décomposable
RRG	Risque relatif génotypique
SEM	Structural expectation-maximization
SNP	Single nucleotide polymorphism (polymorphisme d'un seul nucléotide)
TCE	Taux de compression d'entropie
TRD	Taux de réduction de dimension
TRV	Test du ratio de vraisemblance
VL	Variable latente
VO	Variable observée

Glossaire

ADN (acide désoxyribonucléique) Molécule support de l'information génétique, 8

Algorithme EM (expectation-maximization) Algorithme pour l'apprentissage de paramètres d'un modèle probabiliste en présence de variables latentes ou de données manquantes, 35

Allèle Version d'un gène, 9

Analyse de liaison Analyse génétique basée sur la recombinaison génétique, 2

Apprentissage de paramètres Apprentissage statistique des paramètres d'un modèle probabiliste, 32

Apprentissage de structure Apprentissage statistique de la structure d'indépendance conditionnelle d'un modèle graphique probabiliste, 32

Apprentissage de structure basé sur des contraintes Désigne les méthodes d'apprentissage de structure basées sur l'identification d'indépendances conditionnelles à l'aide de tests statistiques sur les données, 36

Apprentissage de structure basé sur un score Désigne les méthodes parcourant localement l'espace des structures en cherchant à maximiser un score, 36

Association génétique indirecte Désigne une dépendance statistique entre un nœud ancêtre du SNP causal de la FMHCL et le phénotype, 83

Bayésien Se réfère aux méthodes statistiques basées à la fois sur les données fournies par l'expérience et sur des connaissances *a priori*, 34

BIC (Bayesian information criterion) Score permettant d'évaluer la pertinence d'un modèle sur des données. Il réalise un compromis entre ajustement du modèle aux données et complexité du modèle, 37

Bloc haplotypique Ensemble de loci contigus et fortement corrélés, 14

Cardinalité Dans le cas d'une variable discrète, la cardinalité est le nombre d'états différents qu'elle peut prendre, 55

Carte triangulaire de chaleur Désigne, dans le contexte du LD, la matrice triangulaire des dépendances par paires de SNP, pour laquelle l'intensité de la couleur indique la force du LD de chaque cellule de la matrice, 14, 95

Classification ascendante hiérarchique Classification dans laquelle tous les éléments à regrouper forment initialement des classes individuelles, puis sont rassemblés en classes de plus en plus grandes, 54

Clustering Classification d'éléments en groupes homogènes, 52

Distribution de probabilité jointe Distribution de probabilité d'un ensemble de variables, 29

Déséquilibre de liaison Association non aléatoire entre les allèles de deux loci (ou plus) pour une population donnée, 13

Déséquilibre de liaison multilocus Déséquilibre de liaison observé entre plus de deux loci, 97

Déséquilibre de liaison par paires Déséquilibre de liaison observé entre deux loci, 14

Entropie Mesure du désordre dans un système, 96

Etude d'association pangénomique Analyse génétique basée sur le déséquilibre de liaison et visant à la localisation systématique des facteurs génétiques sur le génome, 23

Etude populationnelle Étude sur une population d'individus non apparentés, 22

Faux positif Dans une étude d'association, un faux positif est un SNP (ou une variable, en générale) qui est considéré comme associé au phénotype, alors qu'en réalité il ne l'est pas, 88

Fréquentiste Se réfère aux méthodes statistiques basées uniquement sur les données fournies par l'expérience, 34

Graphe (théorie) Objet mathématique employé afin de traiter des problèmes mettant en œuvre des ensembles sur lesquels sont définies des relations binaires, quelle que soit la nature de ces relations, 27

Graphe décomposable Graphe pouvant être décomposé en cliques et séparateurs de cliques, 38

Gène Unité d'information génétique, 9

Génotype Composition allélique d'un locus ou d'un ensemble de loci d'une séquence, 11

Génétique d'association Analyse génétique basée sur le déséquilibre de liaison, 13

Haplotype Séquence de loci, qu'ils soient gènes ou marqueurs, sur un chromosome, 11

Haplotype ancestral Désigne l'haplotype qui est l'ancêtre (évolutivement) de la population d'haplotypes étudiés, 47

Homozygote Individu possédant le même allèle sur les deux chromosomes homologues, 101

Imputation Désigne le fait de déterminer les valeurs des données manquantes à l'aide des données observées, 54

Indépendance locale Hypothèse selon laquelle les variables observées sont indépendantes conditionnellement à la(aux) variable(s) latente(s). C'est une hypothèse inhérente au modèle à classes latentes, 45

Inférence probabiliste Estimation de la probabilité d'un évènement quelconque sachant certaines connaissances dans le modèle graphique probabiliste, 31

LD de courte distance LD s'observant sur de courtes distances, *i.e.* < 10 kb, 96

LD de longue distance LD s'observant sur de longues distances, *i.e.* > 100 kb, 96

Maladie génétique multifactorielle Maladie génétique influencée conjointement par un ensemble de facteurs génétiques et environnementaux, 2

Markov chain Monte Carlo Classe de méthodes d'échantillonnage à partir d'une distribution de probabilité. Ces méthodes sont basées sur un parcours de l'espace des solutions suivant une chaîne de Markov, 48

Matrice d'adjacence Pour un graphe fini G à n nœuds, la matrice d'adjacence est une matrice de dimension $n \times n$ dont l'élément non diagonal a_{ij} est le nombre d'arêtes reliant le nœud i au nœud j, 57

Modèle graphique probabiliste Modèle probabiliste dont la structure d'indépendance conditionnelle est encodée à l'aide d'un graphe, 29

Modèle hiérarchique à classes latentes Modèle probabiliste discret possédant plusieurs variables latentes. Dans le cadre des réseaux bayésiens à variables latentes, ce modèle est défini par un arbre dont les feuilles sont des variables observées et les nœuds restants sont des variables latentes, 51

Modèle à classes latentes Modèle probabiliste discret possédant une variable latente. Dans le cadre des réseaux bayésiens à variables latentes, ce modèle est défini comme un ensemble de variables observées ayant toutes pour parent une même et unique variable latente, 45

Mutation Toute altération de la séquence du patrimoine génétique d'un individu, 10

P-value Désigne la probabilité de rejet de la loi H_0 sachant que H_0 est vraie, lors d'un test statistique. Pour les tests d'indépendance, plus la p-value est faible, plus la dépendance est considérée comme forte, 85

Partitionnement en cliques Désigne le fait de partitionner un graphe en cliques disjointes de nœuds, 57

Phase gamétique Connaissance de la séquence des allèles sur le chromosome, 11

Point chaud de recombinaison Zone du génome généralement petite et présentant un fort taux de recombinaison, 14

Processus markovien Processus stochastique dans lequel la prédiction du futur ne dépend que du présent et d'une partie limitée du passé, 47

Recombinaison Phénomène par lequel apparaissent des combinaisons génétiques nouvelles, dans une cellule ou un individu, et différentes de celles observées chez les cellules ou individus parentaux, 11

Risque relatif génotypique (RRG) Mesure le rapport de probabilités d'être atteint d'une maladie entre le génotype présentant une ou deux copies de l'allèle muté et le génotype n'en présentant pas. Le RRG dépend du modèle de la maladie (additif, dominant, récessif ou multiplicatif), 83

Réseau bayésien Modèle graphique probabiliste dont les indépendances conditionnelles sont encodées par un graphe orienté sans circuit, 29

Réseau de Markov Modèle graphique probabiliste dont les indépendances conditionnelles sont encodées par un graphe non orienté, 29

SNP Polymorphisme d'un seul nucléotide, *i.e.* tout changement de base au niveau d'un nucléotide de la séquence d'ADN, 17

Structure de la population Désigne le fait que la population étudiée soit constituée d'un ensemble de sous-populations de natures génétiques différentes, 13

Structure en blocs Structure correspondant à un partitionnement en groupes (appelés blocs) de SNP contigus sur l'ADN, 51

Structure en clusters Structure correspondant à un partitionnement en groupes (appelés clusters) de SNP non-nécessairement contigus sur l'ADN, 51

Subsomption Dans un arbre ou une forêt, la subsomption désigne la relation descendant-ancêtre, 82

Variable discrète Variable qui ne prend qu'un nombre fini (ou dénombrable) de valeurs, 29

Variable latente Variable non mesurée, 4, 26

Vraisemblance Probabilité d'observer les données sachant le modèle. Le maximum de la vraisemblance est utilisé pour évaluer l'ajustement d'un modèle probabiliste aux données observées, 31

Bibliographie

[1] H. J. Abel and A. Thomas. Accuracy and computational efficiency of a graphical modeling approach to linkage disequilibrium estimation. *Statistical Applications in Genetics and Molecular Biology*, 10(1) :5, January 2011.

[2] H. Akaike. Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, 22(1) :203–217, December 1970.

[3] B. Alberts. *Molecular biology of the cell.* Garland Science, 2002.

[4] F. Azuaje and J. Dopazo, editors. *Data analysis and visualization in genomics and proteomics.* Wiley-Blackwell, 2005.

[5] D. J. Balding. A tutorial on statistical methods for population association studies. *Nature Genetics*, 7 :781–790, October 2006.

[6] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly. Haploview : analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2) :263–265, January 2005.

[7] J. C. Barrett, D. G. Clayton, P. Concannon, B. Akolkar, J. D. Cooper, H. A. Erlich, C. Julier, G. Morahan, J. Nerup, C. Nierras, V. Plagnol, F. Pociot, H. Schuilenburg, D. J. Smyth, H. Stevens, J. A. Todd, N. M. Walker, S. S. Rich, and Type 1 Diabetes Genetics Consortium. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics*, 41(6) :703–707, June 2009.

[8] I. Barroso. Genetics of type 2 diabetes. *Diabetic Medecine*, 22(5) :517–535, May 2005.

[9] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. In *Proceedings of the Third Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pages 33–42, 1999.

[10] J. Bernardo and A. F. M. Smith. *Bayesian theory.* John Wiley, 1994.

[11] B. L. Browning and S. R. Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, 84(2) :210–223, February 2009.

[12] S. R. Browning. Multilocus association mapping using variable-length Markov chains. *The American Journal of Human Genetics*, 78(6) :903–913, June 2006.

[13] S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5) :1084–1097, November 2007.

[14] R. T. Brumfield, P. Beerli, D. A. Nickerson, and S. V. Edwards. The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, 18(5) :249–256, May 2003.

[15] D. Chickering and C. Boutilier. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3 :507–554, 2002.

[16] D. Chickering, D. Geiger, and D. Heckerman. Learning Bayesian networks : search methods and experimental results. In *Proceedings of Fifth Conference on Artificial Intelligence and Statistics*, pages 112–128, 1995.

[17] D. Clayton. *Handbook of statistical genetics*, volume 2, chapter Population Association, pages 1216–1237. Wiley Interscience, 3rd edition, 2007.

[18] G. F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2-3) :393–405, March 1990.

[19] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29(2) :229–232, October 2001.

[20] D. Daneman. Type 1 diabetes. *The Lancet*, 367(9513) :847–858, March 2006.

[21] H. M. David J. Hand and P. Smyth. *Principles of data mining.* MIT Press, 2001.

[22] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39 :1–38, 1977.

[23] B. Devlin and N. Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29(2) :311–322, September 1995.

[24] J. Feingold, M. Fellous, and M. Solignac. *Principes de génétique humaine.* Hermann, 1998.

[25] J. Feingold. Maladies multifactorielles : un cauchemar pour le généticien. *Médecines Sciences*, 21(11) :927–933, November 2005.

[26] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3) :432–441, July 2008.

[27] B. Fry. *Computational information design*. PhD thesis, Massachusetts Institute of Technology, 2004.

[28] P. I. Good. *Permutation, parametric, and bootstrap tests of hypotheses*. Springer, 3rd edition, December 2004.

[29] G. Greenspan and D. Geiger. High density linkage disequilibrium mapping using models of haplotype block variation. *Bioinformatics*, 20(Suppl 1) :137–144, March 2004.

[30] G. D. Greenspan and D. Geiger. Modeling haplotype block variation using Markov chains. *Genetics*, 172(4) :2583–2599, April 2006.

[31] E. Halperin, G. Kimmel, and R. Shamir. Tag SNP selection in genotype data for maximizing SNP prediction accuracy. *Bioinformatics*, 21(Suppl 1), June 2005.

[32] B. Han, H. M. Kang, M. S. Seo, N. Zaitlen, and E. Eskin. Efficient association study design via power-optimized tag SNP selection. *Annals of Human Genetics*, 72(6) :834–847, November 2008.

[33] B. Han, M. Park, and X. W. Chen. A Markov blanket-based method for detecting causal SNPs in GWAS. *BMC Bioinformatics*, 11(Suppl 3) :S5, April 2010.

[34] D. Heckerman. A tutorial on learning with Bayesian networks. In D. E. Holmes and L. C. Jain, editors, *Innovations in Bayesian Networks*, volume 156 of *Studies in Computational Intelligence*, chapter 3, pages 33–82. Springer Berlin Heidelberg, 2008.

[35] C. J. Hoggart, J. C. Whittaker, M. De Iorio, and D. J. Balding. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics*, 4(7) :e1000130, July 2008.

[36] L. K. Hosking, P. R. Boyd, C. F. Xu, M. Nissum, K. Cantone, I. J. Purvis, R. Khakhar, M. R. Barnes, U. Liberwirth, K. Hagen-Mann, M. G. Ehm, and J. H. Riley. Linkage disequilibrium mapping identifies a 390 kb region associated with CYP2D6 poor drug metabolising activity. *The Pharmacogenomics Journal*, 2(3) :165–175, January 2002.

[37] K.-B. Hwang, B.-H. Kim, and B.-T. Zhang. Learning hierarchical Bayesian networks for large-scale data analysis. In *The 13th International Conference on Neural Information Processing (ICONIP)*, pages 670–679, 2006.

[38] S. K. Iyengar and R. C. Elston. *Linkage disequilibrium and association mapping : analysis and applications*, volume 376, chapter The genetic basis of complex traits : rare variants or "common gene, common disease"?, pages 71–84. Humana Press, 2007.

[39] F. V. Jensen, S. L. Lauritzen, and K. G. Olesen. Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quaterly*, 4(4) :269–282, 1990.

[40] F. V. Jensen. *Introduction to Bayesian networks.* Springer-Verlag New York, 1996.

[41] J. Joseph. *The missing gene : psychiatry, heredity, and the fruitless search for genes.* Algora Pub, 2006.

[42] M. Jünger and P. Mutzel. *Graph drawing software (mathematics and visualization).* Springer, 1st edition, October 2003.

[43] A. Khajavinia and W. Makalowski. What is "junk" DNA, and what is it worth ? *Scientific American*, 296(5) :104, May 2007.

[44] G. Kimmel and R. Shamir. Gerbil : genotype resolution and block identification using likelihood. *Proceedings of the National Academy of Sciences of the United States of America*, 102(1) :158–162, December 2004.

[45] W. S. Klug, M. R. Cummings, and C. A. Spencer. *Génétique.* Pearson Education, 2006.

[46] N. Kumasaka, Y. Nakamura, and N. Kamatani. The textile plot : a new linkage disequilibrium display of multiple-single nucleotide polymorphism genotype data. *PLoS ONE*, 5(4) :e10207, April 2010.

[47] C. D. Langefield and T. E. Fingerlin. Association methods in human genetics. *Methods in Molecular Biology*, 404 :431–456, 2007.

[48] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(2) :157–224, 1988.

[49] P. H. Lee and H. Shatkay. BNTagger : improved tagging SNP selection using Bayesian networks. *Bioinformatics*, 22(14) :e211–e219, July 2006.

[50] R. C. Lewontin. The interaction of selection and linkage. *Genetics*, 49(1) :49–67, January 1964.

[51] Y. Liang and A. Kelemen. Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases. *Statistics Surveys*, 2 :43–60, 2008.

[52] Z. Lin and R. B. Altman. Finding haplotype tagging SNPs by use of principal components analysis. *The American Journal of Human Genetics*, 75(5) :850–861, November 2004.

[53] D. Madigan and J. York. Bayesian graphical models for discrete data. *International Statistical Review*, 63(2) :215–232, August 1995.

[54] G. Malécot. *Les mathématiques de l'hérédité*. Masson et Cie, 1948.

[55] A. Malovini, A. Nuzzo, F. Ferrazzi, A. A. Puca, and R. Bellazzi. Phenotype forecasting with SNPs data through gene-based Bayesian networks. *BMC Bioinformatics*, 10(Suppl 2) :S7, February 2009.

[56] J. Martin and K. Vanlehn. Discrete factor analysis : learning hidden variables in Bayesian network. Technical report, Department of Computer Science, University of Pittsburgh, 1995.

[57] G. Mc Vean. *Handbook of statistical genetics*, volume 2, chapter Linkage disequilibrium, recombination and selection, pages 909–944. Wiley Interscience, 3rd edition, 2007.

[58] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn. Genome-wide association studies for complex traits : consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5) :356–369, May 2008.

[59] V. McKusick. Mendelian inheritance in man and its online version, OMIM. *The American Journal of Human Genetics*, 80(4) :588–604, April 2007.

[60] M. M. Miretti, E. C. Walsh, X. Ke, M. Delgado, M. Griffiths, S. Hunt, J. Morrison, P. Whittaker, E. S. Lander, L. R. Cardon, D. R. Bentley, J. D. Rioux, S. Beck, and P. Deloukas. A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. *The American Journal of Human Genetics*, 76(4) :634–646, April 2005.

[61] A. Montpetit and F. Chagnon. La carte d'haplotype du génome humain : une révolution en génétique des maladies à hérédité complexe. *Médecine Sciences*, 22(12) :1061–1067, December 2006.

[62] A. P. Morris and L. R. Cardon. *Handbook of statistical genetics*, volume 2, chapter Whole genome association, pages 1238–1263. Wiley Interscience, 3rd edition, 2007.

[63] R. Mourad, C. Sinoquet, and P. Leray. Learning hierarchical Bayesian networks for genome-wide association studies. In *19th International Conference on Computational Statistics (COMPSTAT)*, pages 549–556, 2010.

[64] R. Mourad, C. Sinoquet, and P. Leray. A hierarchical Bayesian network approach for linkage disequilibrium modeling and data-dimensionality reduction prior to genome-wide association studies. *BMC Bioinformatics*, 12 :16, 2011.

[65] R. Mourad, C. Sinoquet, and P. Leray. Probabilistic graphical models for genetic association studies. *Briefings in bioinformatics*, 2011.

[66] J. C. Mueller. Linkage disequilibrium for different scales and applications. *Briefings in bioinformatics*, 5(4) :355–364, December 2004.

[67] P. Naïm, P.-H. Wuillemin, P. Leray, O. Pourret, and A. Becker. *Réseaux bayésiens*. Eyrolles, 3 edition, 2007.

[68] A. V. Nefian. Learning SNP dependencies using embedded Bayesian networks. In *IEEE Computational Systems Bioinformatics conference (CSB)*, 2006.

[69] C. Neuhauser. *Handbook of statistical genetics*, volume 2, chapter Mathematical models in population genetics, pages 843–877. Wiley Interscience, 3rd edition, 2007.

[70] C. Newton-Cheh and J. N. Hirschhorn. Genetic association studies of complex traits : design and analysis issues. *Mutation Research*, 573(1-2) :54–69, June 2005.

[71] M. Nordborg. *Handbook of statistical genetics*, volume 2, chapter Coalescent theory, pages 843–877. Wiley Interscience, 3rd edition, 2007.

[72] M. Nothnagel, R. Fürst, and K. Rohde. Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Human Heredity*, 54(4) :186–198, January 2002.

[73] C. Pattaro, I. Ruczinski, D. M. Fallin, and G. Parmigiani. Haplotype block partitioning as a tool for dimensionality reduction in SNP association studies. *BMC Genomics*, 9 :405, August 2008.

[74] J. Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann, September 1988.

[75] J. K. Pritchard and M. Przeworski. Linkage disequilibrium in Humans : models and data. *The American Journal of Human Genetics*, 69(1) :1–14, July 2001.

[76] C. P. Robert. *The Bayesian choice : from decision-theoretic foundations to computational implementation*. Springer Texts in Statistics. Springer Verlag, 2nd edition, June 2007.

[77] N. A. Rosenberg and M. Nordborg. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, 3(5) :380–390, May 2002.

[78] K. R. Rosenbloom, T. R. Dreszer, M. Pheasant, G. P. Barber, L. R. Meyer, A. Pohl, B. J. Raney, T. Wang, A. S. Hinrichs, A. S. Zweig, P. A. Fujita, K. Learned, B. Rhead, K. E. Smith, R. M. Kuhn, D. Karolchik, D. Haussler, and W. J. Kent. ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Research*, 38(Suppl 1) :D620–625, January 2010.

[79] F. Rousseau and N. Laflamme. Génétique moléculaire humaine : des maladies monogéniques aux maladies complexes. *Médecine Sciences*, 19(10) :950–954, October 2003.

[80] C. N. Sarkissian, A. Gámez, and C. R. Scriver. What we know that could influence future treatment of phenylketonuria. *Journal of Inherited Metabolic Disease*, 32(1) :3–9, August 2008.

[81] D. J. Schaid. Evaluating association of haplotypes with traits. *Genetic Epidemiology*, 27(4) :348–364, December 2004.

[82] P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data : applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4) :629–644, April 2006.

[83] G. Schwartz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2) :461–464, March 1978.

[84] P. Sebastiani, M. F. Ramoni, V. Nolan, C. T. Baldwin, and M. H. Steinberg. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *Nature Genetics*, 37(4) :435–440, March 2005.

[85] B. Servin and M. Stephens. Imputation-based analysis of association studies : candidate regions and quantitative traits. *PLoS Genetics*, 3(7) :e114, July 2007.

[86] S. Simoff, M. H. Böhlen, and A. Mazeika, editors. *Visual data mining : theory, techniques and tools for visual analytics*. Springer, 2008.

[87] J. Snyman. *Practical mathematical optimization : an introduction to basic optimization theory and classical and new gradient-based algorithms*. Springer, 2005.

[88] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*, volume 81 of *Lecture Notes in Statistics*. Springer-Verlag, 1993.

[89] M. Stephens. *Handbook of statistical genetics*, volume 2, chapter Inference under the coalescent, pages 878–908. Wiley Interscience, 3rd edition, 2007.

[90] M. Stephens and D. J. Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10) :681–690, October 2009.

[91] M. Stephens and P. Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *The American Journal of Human Genetics*, 76(3) :449–462, March 2005.

[92] A. C. Syvänen. Accessing genetic variation : genotyping single nucleotide polymorphisms. *Nature Reviews Genetics*, 2(12) :930–942, December 2001.

[93] H. K. Tabor, N. J. Risch, and R. M. Myers. Candidate-gene approaches for studying complex genetic traits : practical considerations. *Nature Reviews Genetics*, 3(5) :391–397, May 2002.

[94] I. Tachmazidou, C. J. Verzilli, and M. D. Iorio. Genetic association mapping via evolution-based clustering of haplotypes. *PLoS Genetics*, 3(7) :e111, July 2007.

[95] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319) :1061–1073, October 2010.

[96] The International HapMap Consortium. The international HapMap project. *Nature*, 426(6968) :789–796, December 2003.

[97] The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063) :1299–1320, October 2005.

[98] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164) :851–861, October 2007.

[99] A. Thomas. Characterizing allelic associations from unphased diploid data by graphical modeling. *Genetic Epidemiology*, 29(1) :23–35, July 2005.

[100] A. Thomas. Estimation of graphical models whose conditional independence graphs are interval graphs and its application to modelling linkage disequilibrium. *Computational Statistics & Data Analysis*, 53(5) :1818–1828, March 2009.

[101] A. Thomas. A method and program for estimating graphical models for linkage disequilibrium that scale linearly with the number of loci, and their application to gene drop simulation. *Bioinformatics*, 25(10) :1287–1292, May 2009.

[102] A. Thomas and N. J. Camp. Graphical modeling of the joint distribution of alleles at associated loci. *The American Journal of Human Genetics*, 74(6) :1088–1101, April 2004.

[103] A. Thomas and P. Green. Enumerating the decomposable neighbors of a decomposable graph under a simple perturbation scheme. *Computational Statistics & Data Analysis*, 53(4) :1232–1238, February 2009.

[104] A. Thomas and P. J. Green. Enumerating the junction trees of a decomposable graph. *Journal of Computational and Graphical Statistics*, 18(4) :930–940, December 2009.

[105] E. A. Thompson. *Handbook of statistical genetics*, volume 2, chapter Linkage analysis, pages 1141–1167. Wiley Interscience, 3rd edition, 2007.

[106] I. Tsamardinos, Brown, and A. Constantin. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1) :31–78, October 2006.

[107] C. J. Verzilli, N. Stallard, and J. C. Whittaker. Bayesian graphical models for genome-wide association studies. *The American Journal of Human Genetics*, 79(1) :100–112, May 2006.

[108] E. Villanueva and C. D. D. Maciel. Modeling associations between genetic markers using Bayesian networks. *Bioinformatics*, 26(18) :i632–i637, September 2010.

[109] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4) :395–416, December 2007.

[110] Y. Wang, N. L. Zhang, and T. Chen. Latent tree models and approximate inference in Bayesian networks. *Journal of Artificial Intelligence Research*, 32 :879–900, August 2008.

[111] D. A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y.-J. Chen, V. Makhijani, G. T. Roth, X. Gomes, K. Tartaro, F. Niazi, C. L. Turcotte, G. P. Irzyk, J. R. Lupski, C. Chinault, X.-z. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D. M. Muzny, M. Margulies, G. M. Weinstock, R. A. Gibbs, and J. M. Rothberg. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189) :872–876, April 2008.

[112] R. Yehezkel and B. Lerner. Bayesian network structure learning by recursive autonomy identification. *The Journal of Machine Learning Research*, 10 :1527–1570, July 2009.

[113] N. L. Zhang. Hierarchical latent class models for cluster analysis. *The Journal of Machine Learning Research*, 5 :697–723, June 2004.

[114] N. L. Zhang and T. Kocka. Efficient learning of hierarchical latent class models. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 585–593, 2004.

[115] N. L. Zhang, Y. Wang, and T. Chen. Discovery of latent structures : Experience with the coil challenge 2000 data set*. *Journal of Systems Science and Complexity*, 21(2) :172–183, 2007.

[116] Y. Zhang and L. Ji. Clustering of SNPs by a structural EM algorithm. In *International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*, pages 147–150, 2009.

[117] Y. Zhang and J. S. Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, 39(9) :1167–1173, September 2007.