Outils mathématiques pour la prévention des maladies complexes

Félix Balazard

23 octobre 2014

1 Introduction

Une maladie complexe est une maladie qui résulte de l'interaction de nombreux gènes ainsi que de facteurs environnementaux. On les appelle complexes en opposition aux maladies génétiques simples ou mendéliennes qui résultent de la mutation d'un seul gène, mutation souvent récessive comme pour la mucoviscidose ou la maladie de Tay-Sachs. Les maladies complexes sont plus communes regroupant le cancer, les maladies cardio-vasculaires, la maladie d'Alzheimer, le diabète de type 1 ou de type 2.

Avec le développement récent des technologies de séquençage, l'information génétique est devenue de plus en plus accessible. Les études d'associations sur toute la largeur du génome ou GWAS ont constitué un premier pas important dans la compréhension de la génétique des maladies complexes. Il est important de comprendre ce que sont ces études GWAS et ce qu'elles ne sont pas. Il s'agit de génotyper des malades et de les comparer à des témoins afin de trouver les régions du génome associées à la maladie. Le verbe génotyper peut induire en erreur, il ne s'agit pas de séquencer entièrement le génome, base par base. Seuls les variants communs où la fréquence de l'allèle mineur est supérieur à 5% dans la population sont obtenus. On appelle un tel nucléotide un SNP pour Single Nucleotide Polymorphism. C'est une approche plus grossière que le séquençage complet qui a été choisi pour des problèmes de coût mais qui peut se justifier scientifiquement. Étant donné les mécanismes de transmission de l'ADN d'une génération à l'autre, et notamment le cross-over qui se produit lors de la méiose, la connaissance d'un SNP dans une zone va nous donner une information sur la zone qui l'entoure. On appelle cela le LD pour Linkage Desequilibrium. Un SNP est corrélé avec toutes les mutations présentes autour de lui, corrélation qui décroit avec la distance. Même si on ne connaît pas la mutation causale de la maladie, il suffit qu'il y ait un SNP en LD avec cette mutation pour qu'on puisse savoir par une étude GWAS que ce SNP, et donc la zone qui l'entoure, est associé à la maladie. Les études GWAS ont pour but de découvrir des SNPs associés avec la maladie étudiée afin de mieux comprendre les causes de la maladie ou de trouver des cibles pour de nouveaux médicaments. On y teste donc chaque SNP un par un grâce à un test du χ^2 et on regarde ceux qui sont significatifs pour la correction de Bonferroni (correction pour les tests multiples). Comme on teste un grand nombre de SNPs, il faut avoir de grands échantillons pour avoir quelque chose de significatif.

Notre objectif est un peu différent puisque nous cherchons à évaluer le risque génétique de chaque personne. Si les études GWAS ont obtenu des résultats nouveaux d'association entre des zones du génomes et des maladies, elles ont déçu car elles n'ont pas permis d'identifier une partie importante de la prédisposition génétique qu'on sait exister grâce aux études familiales. Ce phénomène a été appelé l'héritabilité

manquante. Nous proposons une explication pour l'héritabilité manquante qui est l'importance de l'information de phase, le fait que, pour chaque paire de chromosome, nous avons deux chromosomes homologues distincts qui agissent indépendamment.

L'objectif final de développer un tel indicateur est de pouvoir faire des plans de prévention contre les maladies complexes. Il est facile de comprendre que cela permettra de cibler des mesures de prévention sur une population à haut risque mais cela permettra aussi d'évaluer différentes mesures de prévention. La méthodologie pour ce nouveau genre d'essais cliniques devra s'appuyer sur l'estimation du risque génétique. Un a priori important pour réaliser un tel plan de prévention est de génotyper la population, ce qui deviendra bientôt faisable économiquement avec la chute des coûts de séquençage qu'on observe aujourd'hui. On peut aussi envisager de génotyper les parents proches de malade afin de savoir si ils sont effectivement à risque accru de développer la maladie.

La section 2 exposera un modèle très simple de maladie.

La question à laquelle répondra la section 3 est comment passer de l'échantillon de l'épidémiologue où on compare des malades à des témoins au risque dans la population générale.

Dans la section 4, nous expliquerons les grandes problématiques de l'apprentissage statistique puis nous regarderons les différentes méthodes d'estimation du risque génétique qui ont été essayées jusqu'à présent.

Dans la section 5, on proposera d'estimer le risque après avoir récupéré l'information biologiquement importante qu'est l'information de phase.

Dans la section 6, nous réfléchirons à ce que peuvent nous apporter les études familiales et nous nous pencherons sur la notion d'héritabilité.

2 Risque génétique : un modèle naturel

Considérons une population d'individus (par exemple la population de la France) ayant chacun un génome qui est une variable aléatoire G qui est toute l'information génétique d'une personne et un environnement E qui est toute l'information environnementale d'une personne. Du point de vue mathématique, il n'y a pas de différence conceptuelle entre G et E mais du point de vue de l'épidémiologie, ces deux composantes sont très différentes puisqu'on peut modifier l'environnement d'une personne alors que son génome est fixe.

Nous nous intéressons à une maladie complexe qui fait interagir des facteurs environnementaux et génétiques. Il faut souligner que le processus de déclenchement de la maladie est profondément aléatoire. Deux jumeaux monozygotiques qui ont donc les mêmes gênes et presque le même environnement ont tous les deux le diabète de type 1 dans seulement 42% des cas où au moins l'un des deux est atteint. Nous modélisons donc la maladie par une variable aléatoire associée à un individu (G, E) que l'on note M et qui est une variable de Bernoulli de paramètre $p(G, E)$. Quand $M = 1$ on dit que l'individu est malade et quand $M = 0$ qu'il est sain. Pour g et e un génome et un environnement donné,

$$p(g, e) = \mathbb{P}(M = 1 | G = g, E = e)$$

est la probabilité qu'un individu au génome g et à l'environnement e ait la maladie. On s'y réfèrera avec le vocabulaire de l'épidémiologie comme risque global absolu (génétique et environnemental). On note $\mathbb{P}$ la loi jointe de M,G et E et $\mathbb{E}$ l'espérance sous cette loi. Afin de poser les choses rigoureusement, on peut dire par exemple que $\mathbb{P}$ est la distribution empirique de G, E et M dans la population concernée. Le but de l'épidémiologie des maladies complexes est de connaître p. On a la relation suivante

$$\mathbb{E}[M] = \mathbb{P}(M = 1) = \mathbb{E}[p(G, E)]$$

et cette quantité s'interprète évidemment comme la prévalence de la maladie qui vaut pour le cas du diabète de type 1 en France, 0,38%. On s'intéressera également au risque global (environnemental et génétique) relatif défini par

$$r(g,e) = \frac{p(g,e)}{\mathbb{E}[M]} = \frac{\mathbb{P}(M=1|G=g,E=e)}{\mathbb{P}(M=1)}$$

qui s'interprète de la manière suivante : un individu qui a un génome g et un environnement e a $r(g,e)$ fois plus de chance d'avoir la maladie qu'un individu tiré au hasard dans la population. Il est intéressant de travailler avec un risque relatif plutôt qu'avec un risque absolu dans le cas d'une maladie rare. Dans ce cas là, le risque absolu restera faible et il est difficile pour beaucoup de personnes de se représenter de faibles probabilités et de comprendre les enjeux qu'elles représentent. Si on dit à une personne qu'elle a 1% de chance de développer une maladie, elle ne prendra pas très au sérieux les recommandations qu'on lui fait, alors que si on lui dit qu'elle a trois fois plus de chance de développer la maladie que la moyenne, elle conçoit mieux l'enjeu.

Une parenthèse doit être faite ici pour préciser un point de vocabulaire. En épidémiologie, le risque relatif est le rapport de la probabilité de maladie chez les sujets exposés à un facteur à celle chez les sujets non exposés. Ici, le facteur étant potentiellement très précis, le dénominateur se réduit au risque dans la population. On pourrait aussi parler de risque par rapport à la moyenne de la population.

Oublions pour le moment E et concentrons-nous sur la partie génétique. On peut définir $p(g)$ qui est la probabilité qu'un individu ayant le génome g développe la maladie et qu'on appellera risque génétique absolu. On peut écrire

$$p(g) = \mathbb{E}[p(G,E)|G=g].$$

On définit de même que précédemment le risque génétique relatif

$$r(g) = \frac{p(g)}{\mathbb{E}[M]} = \frac{\mathbb{P}(M=1|G=g)}{\mathbb{P}(M=1)}.$$

Cette section peut sembler triviale. Elle s'oppose en fait à un modèle commun dans ce cadre qui est issu de l'adaptation des méthodes de la génétique quantitative (notamment de sélection animale) au cas binaire d'une maladie. Ce modèle consiste à dire que pour un trait binaire comme une maladie, il existe un trait continu sous-jacent normalement distribué dans la population qui, si il dépasse un seuil, provoquera la maladie. Ce trait est somme d'une composante génétique et d'une composante environnementale. Ce modèle permet notamment de calculer l'héritabilité d'un trait binaire, c'est-à-dire la proportion de variance phénotypique expliqué par la génétique comme exposé dans l'article [4]. Il implique cependant que toutes les personnes partageant le même environnement et le même génome soient malades ou que toutes soient saines. Il est trop déterministe du point de vue de l'auteur. Ce modèle introduit des complications artificielles. Une critique acerbe en est faite dans [9].

3 De l'échantillon de l'épidémiologue à la population générale

Maintenant, intéressons-nous à l'approche de l'épidémiologue. Il recrute dans sa cohorte un maximum de malades et il recrute aussi des témoins non malades. Formalisons ceci par une variable S pour sélection avec $S=1$ si l'épidémiologue recrute la personne et $S=0$ sinon. Il prend (tirage sans remise) une certaine portion de la population malade qu'on note γ. Il tire ensuite sans remise des témoins dans la population saine. Il n'y a pas de raison qu'il y est autant de patients que de témoins. On note β le rapport du nombre

de témoins sur le nombre de patients. Un individu tiré au hasard dans l'échantillon de l'épidémiologue à une plus grande probabilité d'être malade qu'un individu tiré au hasard dans la population générale. M, G et E n'ont pas la même loi que dans la population générale, c'est-à-dire sous $\mathbb{P}$. Dans ce qui va suivre, on va voir qu'il est possible d'estimer ce qui se passe dans la population générale grâce à l'échantillon de l'épidémiologue. On note la mesure de probabilité dans ce dernier

$$\hat{\mathbb{P}} = \mathbb{P}(\cdot|S=1).$$

On peut définir le risque génétique relatif dans l'échantillon de l'épidémiologue par

$$\alpha(g) = \frac{\hat{\mathbb{P}}(M=1|G=g)}{\hat{\mathbb{P}}(M=0|G=g)}.$$

C'est une quantité qu'on peut évaluer en utilisant toutes les techniques modernes d'apprentissage statistique et qu'on va maintenant relier au risque génétique relatif dans la population générale.

$$\alpha(g) = \frac{\hat{\mathbb{P}}(M=1|G=g)}{\hat{\mathbb{P}}(M=0|G=g)} = \frac{\hat{\mathbb{P}}(M=1,G=g)}{\hat{\mathbb{P}}(M=0,G=g)} = \frac{\mathbb{P}(M=1,G=g|S=1)}{\mathbb{P}(M=0,G=g|S=1)}$$
$$= \frac{\mathbb{P}(M=1,G=g,S=1)}{\mathbb{P}(M=0,G=g,S=1)} = \frac{\mathbb{P}(M=1,G=g)}{\mathbb{P}(M=0,G=g)} \times \frac{\mathbb{P}(S=1|M=1)}{\mathbb{P}(S=1|M=0)}$$

car S conditionnellement à M est indépendant de G. On a $\mathbb{P}(S=1|M=1)=\gamma$ et $\mathbb{P}(S=1|M=0) = \frac{\beta\gamma\mathbb{E}[M]}{1-\mathbb{E}[M]}$ (tirage sans remise chez les témoins). Il vient :

$$p(g) = \beta\mathbb{P}(M=0|G=g)\alpha(g) = \beta\frac{\mathbb{E}[M]}{1-\mathbb{E}[M]}\alpha(g)(1-p(g))$$

et on obtient finalement

$$p(g) = \frac{\mathbb{E}[M]\beta\alpha(g)}{1+\mathbb{E}[M](\beta\alpha(g)-1)}.$$

On a donc besoin de connaître avec une certaine précision la prévalence de la maladie en plus de l'échantillon de l'épidémiologue. Pour une maladie comme le diabète de type 1 où l'incidence a augmenté de manière significative au fil des années, il est sans doute plus approprié de ne pas utiliser la prévalence réelle de la maladie mais plutôt la prévalence qu'induit l'incidence actuelle. On peut aussi imaginer corriger pour des variables de confusions comme le sexe ou l'ethnicité en considérant $\mathbb{E}[M]$ comme un a priori bayésien.

Dans l'expression précédente, on voit apparaître la quantité $\beta\alpha(g)$. De plus, si $p(g) << 1$, alors

$$r(g) = \beta\alpha(g)$$

Ceci nous permet d'estimer $r(g)$ quantité définie dans la population générale grâce à une quantité définie dans l'échantillon de l'épidémiologue. Cette approximation est valide pour une maladie rare et faiblement déterminée génétiquement. De plus l'approximation devient moins bonne uniquement pour les personnes au risque le plus élevé. Pour ces personnes, le risque est surévalué ce qui n'est pas très grave. L'intérêt de cette approximation est que quand elle est valide, il n'est pas nécessaire de connaître avec précision la prévalence de la maladie. Si on a pris autant de cas que de témoins, la correction en β disparait et le risque relatif dans l'échantillon est égal au risque relatif dans la population.

4 Apprentissage statistique : notions et applications à la génétique des maladies complexes

En apprentissage statistique, on a des variables d'entrées X_j qui seront ici des SNPs et une variable de sortie Y qui, ici, est notre variable M. On nous donne un certain nombre d'observations indexées par i et puis on veut prédire la variable Y pour une nouvelle observation. Pour faire cela, on choisit un modèle, c'est-à-dire un ensemble de fonctions des X_j, par exemple les fonctions linéaires en X_j, et ensuite, on cherche la fonction dans notre modèle qui minimise l'erreur empirique dans l'échantillon. La fonction d'erreur peut prendre diverses formes (somme du carré des résidus pour la régression linéaire) mais il est important qu'elle soit convexe afin qu'on puisse la minimiser de manière efficace. Naïvement, on peut penser que prendre une classe de fonctions très grande va nous permettre d'avoir une erreur empirique moindre. Mais si l'on fait cela, on va faire du sur apprentissage. Toutes les particularités de notre échantillon se retrouveront dans notre modèle qui n'aura aucune capacité prédictive sur de nouvelles données. C'est pour ceci qu'il faut toujours mesurer la qualité d'un estimateur sur un échantillon de test qui n'a pas servi à élaborer le modèle.

Quand on a un grand nombre de variables, les prédicteurs classiques que sont la régression linéaire ou la régression logistique deviennent instables ou ne sont plus définis. Il est alors nécessaire d'ajouter une pénalisation dans la fonction d'erreur qu'on minimise qui va contrôler la complexité du modèle. Cette pénalisation introduit un nouveau paramètre λ qui quantifie la complexité de l'estimateur. On a donc besoin d'estimer ce paramètre et pour cela, on fait de la validation croisée. On sépare l'ensemble d'entrainement en K parties et on entraine l'estimateur sur les $K-1$ premières parties pour différentes valeur de λ et on regarde la performance de ces estimateurs sur la Kème partie. On recommence K fois en changeant à chaque fois l'ensemble de validation. On choisit finalement le λ qui minimise la moyenne des K erreurs. Une pénalisation très populaire pour la régression est la pénalisation $L1$ car elle induit de la "sparsité" c'est-à-dire qu'elle pousse à mettre des coefficients nuls à certaines variables. On appelle la régression linéaire ou logistique avec pénalisation $L1$ LASSO pour Least Absolute Shrinkage and Selection Operator.

En génétique, le nombre de variables étant très grand, entre la centaine de milliers et plusieurs millions, on a parfois recours à de la présélection. On ne prend que les variables passant un test du χ^2 pour un seuil moins exigeant que celui de Bonferroni avant d'y appliquer par exemple une régression Lasso. Il faut faire alors très attention de bien faire la présélection à chaque étape de la validation croisée car sinon on a sélectionné des variables significatives aussi pour la partie de validation ce qui mène à des estimations surévaluées de la qualité de l'estimateur.

Une approche qu'on trouve souvent dans la littérature est de n'utiliser que les SNPs qui ont été prouvés être liés à la maladie. Ce qui veut dire uniquement les SNPs passant la correction de Bonferroni. C'est par exemple ce que faisait 23andMe. C'est une stratégie qui n'est pas optimale car elle est trop conservative. Une telle stratégie exclut des SNPs apportant de l'information de peur d'inclure des faux positifs. Cette approche est critiquée dans [13, 3].

L'article [13] est l'un des premiers articles à estimer le risque génétique en prenant en compte les leçons de l'apprentissage statistique. Il utilise un support vecteur machine ainsi qu'une régression logistique pénalisée pour classifier patients atteints du diabète de type 1 et témoins. Il utilise pour évaluer la qualité de son estimateur un indicateur indépendant de la proportion de cas et de témoins dans l'échantillon, la courbe ROC (Receiver Operating Curve). Cette courbe se relie facilement aux notions épidémiologiques de sensibilité et de spécificité. La qualité d'un classifieur peut être résumée par l'aire sous la courbe ROC ou AUC (Area Under Curve) qui vaut entre 0,5 et 1. Plus l'AUC est proche de 1 et plus le classifieur est

de bonne qualité. Il obtient aini pour le diabète de type 1, un AUC de 0,84. Dans les articles [14, 1], des AUC comparables sont obtenus pour la maladie de Crohn, la colite ulcérative et la maladie céliaque par les mêmes méthodes.

Beaucoup de ce qui se fait dans le domaine de la prédiction de risque génétique ne tient pas compte de ces remarques. L'interpretome, présenté dans [8], est une initiative permettant d'interpréter en ligne son génome. Le but principal de ce projet est le respect du caractère privé des données génétiques. En utilisant des méthodes modernes d'informatique, les données de l'utilisateur ne quittent pas son navigateur. L'interpretome propose bien un moyen de calculer son risque pour divers maladie mais il s'appuie sur l'article [10] et revient en fait à multiplier les différents Odd Ratios pour les différents SNPs disponibles. Cela revient en fait à faire une régression logistique multivariée en prenant pour coefficient pour chaque variable, le coefficient obtenu dans la régression logistique univariée. C'est la seule méthode envisageable si on a pas de base de données centralisée pour faire les analyses. Si on veut pouvoir effectuer des analyses statistiques de bonnes qualités, il est inévitable de centraliser au moins en partie les données afin de pouvoir faire des analyses statistique.

5 Utilisation des haplotypes dans l'estimation du risque génétique

Les données utilisées dans ces études sont des données issues des nouvelles technologies de séquençage. Toutefois, une information importante est perdue dans la manière même dont se présente ces données. Chaque être humain a 23 paires de chromosomes. Les données utilisées disent que pour tel SNP sur le chromosome 6, le génotype est A/C ce qui veut dire que sur un des chromosomes de la paire à cette position, il y a un A et que sur l'autre, à la même position, il y a un C. Maintenant, si on s'intéresse à un deuxième SNP sur la même paire de chromosomes et que le génotype y est G/T, on ne peut alors pas dire si le A est avec le C ou le G sur le même chromosome. Cette information manquante est l'information de phase et la séquence sur chaque chromosome de la paire est l'haplotype. On trouvera une exposition de l'importance de cette information pour tout phénomène biologique dans [11]. Des méthodes probabilistes pour retrouver cette information de phase existent [2] et obtiennent des précisions correctes. Cependant, jusqu'à présent, ces méthodes étaient surtout utilisées pour ensuite imputer les SNPs voisins grâce à des données de référence comme celles fournies par le projet HapMap ou le projet 1000 genomes.

Dans le cadre des études GWAS, typiquement, chaque SNP fait l'objet d'un test individuel d'association avec la maladie. L'information de phase n'a alors aucun intérêt. Mais par contre, dès qu'il s'agit de combiner plusieurs SNPs, cette information est cruciale.

On trouve tout de même dans la littérature un exemple d'estimation de risque génétique s'appuyant sur les haplotypes dans [7]. Cet article porte sur la maladie de Crohn. L'approche de l'article consiste à imputer les haplotypes de 2 SNPs consécutifs et ensuite à dire pour chaque individu si cet haplotype était présent 0, 1 ou 2 fois. Ainsi le problème est ramené au même problème de classification que quand on travaille avec des SNPs, on a juste multiplié par 4 le nombre de variable. Les auteurs de cet article en utilisant une méthodologie assez rudimentaire ont obtenu des résultats satisfaisants bien que faibles comparés à ceux obtenus par [14]. Cependant, le phénomène le plus intéressant qui se produit dans l'article est que la performance de l'estimateur sur l'ensemble de validation continue d'augmenter jusqu'à inclure 7000 haplotypes alors que les meilleurs modèles dans [14, 13, 1] implique au plus 1000 SNPs. Ceci suggère que mettre les données sous cette forme peut permettre de capturer des effets génétiques de plus faible ampleur.

Une fois que l'on a mis les données génétiques sous la forme d'haplotypes, c'est-à-dire que l'on a récupéré l'information de phase, on a doublé le nombre de variables. Cependant, ces variables ont une structure particulière puisqu'elles consistent en deux fois les mêmes variables mais qui peuvent prendre des valeurs différentes. De plus l'information que l'on peut espérer récupérer dans cette démarche se trouve dans les interactions en *cis* c'est-à-dire sur le même chromosome. On sort ainsi du cadre classique de l'apprentissage statistique et il faut donc au moins en adapter les méthodes si ce n'est en inventer des plus adaptées. C'est le cœur de notre recherche. Le but principal est d'améliorer les capacités prédictives du score de risque génétique. Il est important aussi de mettre en évidence si c'est possible des interactions entre différents haplotypes afin d'améliorer la compréhension biologique de la genèse des maladies.

6 Risque familial et risque génétique

Bien avant que l'on connaisse la structure en hélice de l'ADN, la génétique en tant que discipline existait déjà. Elle se fondait sur l'étude des lignées. Le but des études familiales est encore aujourd'hui de comprendre quelle est la part génétique et environnementale dans une maladie. Dans [5] qui est une étude familiale sur toute la population finlandaise, la concordance du diabète de type 1 entre jumeaux monozygotiques, c'est-à-dire la proportion de paires de jumeaux étant tous les deux diabétiques parmi toutes les paires ayant au moins un jumeau atteint, est de $42,9\%$ alors que la concordance pour jumeaux dizygotiques est de seulement $7,4\%$. L'intérêt d'étudier des jumeaux est qu'ils partagent essentiellement le même environnement (ceci est vrai pour une maladie juvénile comme le diabète de type 1, faux pour une maladie comme Alzheimer) et exactement le même génome pour les vrais jumeaux mais seulement la moitié de leur génome pour les faux jumeaux. Ceci permet de voir la différence de concordance qu'induise les différences génétiques indépendamment de l'environnement.

On souhaite quantifier le caractère génétique ou environnemental d'un trait. C'est dans ce but qu'a été inventé le concept d'héritabilité dont on trouvera une revue détaillée dans [12]. C'est un concept défini au début pour des traits quantitatifs comme le poids d'un poulet à 14 jours de l'éclosion ou la taille d'un humain adulte. C'est la proportion de variance phénotypique expliquée par la variance génétique dans un population. Ce concept a ensuite été adapté à des traits binaires comme la présence ou l'absence d'une maladie. La manière dont cela est fait est détaillée dans [4] qui date de 1950. Ils considèrent qu'il y a un trait quantitatif sous-jacent distribué normalement qui provoque la maladie quand il dépasse un certain seuil.

L'auteur n'est pas sûr que ce concept soit vraiment intéressant pour un trait binaire. Il est par contre clair que les études familiales doivent permettre de fixer une borne supérieure à atteindre sur la performance d'un indicateur de risque génétique. Il est ainsi assez facile d'interpréter la concordance pour les jumeaux monozygotiques. Deux jumeaux monozygotiques partagent G et E, on peut voir le deuxième jumeau comme une variable indépendante du premier et de même loi que le premier. Si on note C_{mz}, le taux de concordance entre jumeaux monozygotiques, on a

$$C_{mz} = \mathbb{P}(M' = 1 | M = 1) = \frac{\mathbb{E}[p(G,E)^2]}{\mathbb{E}[p(G,E)]}$$

où M' est de même loi que M. Ce qui permet de calculer facilement la variance de $p(G,E)$ pour le diabète en Finlande en fonction de la prévalence. Ceci nous dit aussi que la moyenne de $p(G,E)$ chez les malades est de $0,42$.

On trouve dans [6] une comparaison de plusieurs modèles de prédisposition génétique permettant de relier AUC et héritabilité au sens large, c'est-à-dire en incluant le risque au frère et sœur comme mesure

de l'héritabilité. Une approche sans hypothèse supplémentaire pour évaluer l'héritabilité expliquée par un score de prédisposition est de simuler des génomes de frères et de voir quel est leur risque prédit. On peut remarquer que pour faire cela, on a besoin d'avoir l'information de phase puisque, à recombinaison près, ce sont les chromosomes entiers qui sont partagées avec probabilité $\frac{1}{2}$. Nous implémenterons cette approche dans notre travail.

Références

[1] G. Abraham, J. Tye-Din, O. Bhalala, J. Kowalczyk, A.and Zobel, and M. Inouye. Accurate and robust genomic prediction of celiac disease using statistical learning. *PLoS genetics*, 10(2) :e1004137, 2014.

[2] S. Browning and B. Browning. Haplotype phasing : existing methods and new developments. *Nature Reviews Genetics*, 12(10) :703–714, 2011.

[3] G. de Los Campos, D. Gianola, and D. B. Allison. Predicting genetic predisposition in humans : the promise of whole-genome markers. *Nature Rev. Genet.*, 11(12) :880–886, 2010.

[4] E. R. Dempster and I. M. Lerner. Heritability of threshold characters. *Genetics*, 35(2) :212, 1950.

[5] V. Hyttinen, J. Kaprio, L. Kinnunen, M. Koskenvuo, and J. Tuomilehto. Genetic liability of type 1 diabetes and the onset age among 22,650 young finnish twin pairs a nationwide follow-up study. *Diabetes*, 52(4) :1052–1055, 2003.

[6] L. Jostins and J. Barrett. Genetic risk prediction in complex disease. *Human molecular genetics*, 20(R2) :R182–R188, 2011.

[7] J Kang, S Kugathasan, M. Georges, H. Zhao, and J. Cho. Improved risk prediction for crohn's disease with a multi-locus approach. *Human molecular genetics*, 20(12) :2435–2442, 2011.

[8] K.J. Karczewski, T.P. Robert, P. Cordero, N. P. Tatonetti, J. T. Dudley, K. Salari, M. Snyder, R. B. Altman, and S. K. Kim. Interpretome : a freely available, modular, and secure personal genome interpretation engine. World Scientific.

[9] K. Mitchell. What is complex about complex disorders? *Genome Biology*, 13(1) :237, 2012.

[10] A.A. Morgan, R. Chen, and A.J. Butte. Likelihood ratios for genome medicine. *Genome Med*, 2(5) :30–30, 2010.

[11] R. Tewhey, V. Bansal, A. Torkamani, E. Topol, and N. Schork. The importance of phase information for human genomics. *Nature Reviews Genetics*, 12(3) :215–223, 2011.

[12] P. M. Visscher, W. G. Hill, and N. R. Wray. Heritability in the genomics era concepts and misconceptions. *Nature Rev. Genet.*, 9(4) :255–266, 2008.

[13] Z. Wei, K. Wang, H. Qu, H. Zhang, J. Bradfield, C. Kim, E. Frackleton, C. Hou, J. T. Glessner, R. Chiavacci, et al. From disease association to risk assessment : an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet.*, 5(10) :e1000678, 2009.

[14] Z. Wei, W. Wang, J. Bradfield, J. Li, C. Cardinale, E. Frackelton, C. Kim, F. Mentch, K. Van Steen, P. Visscher, et al. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *The American Journal of Human Genetics*, 92(6) :1008–1012, 2013.