Association
mathématique
du Québec

ISM
Institut des sciences mathématiques

De la statistique à la génétique : identifier les gènes responsables de maladies complexes

Aurélie Labbe,
Université McGill
aurelie.labbe@mcgill.ca

Résumé

Les récentes avancées dans l'exploration du génome humain ouvrent des perspectives de recherche encore inégalées au niveau de la compréhension et du traitement des « maladies complexes », dont les cancers et les maladies cardiovasculaires ou neuro psychiatriques. La tâche est difficile, car ce que nous considérons comme une seule maladie engendre souvent une multitude de symptômes différents d'un individu à l'autre, causés par un réseau complexe d'interactions entre une multitude de gènes et notre exposition à certains environnements. La statistique et les mathématiques font aujourd'hui partie intégrante du processus analytique qui conduit à la détection de variantes génétiques causant certaines maladies. Nous montrons ici comment les mathématiques et la statistique sont utilisées dans le domaine de la génétique, quels sont les défis auxquels nous faisons face et à quel point nous avons besoin de nouvelles méthodes pour analyser des données de plus en plus nombreuses et complexes.

1 Qu'est-ce que la statistique génétique ?

La statistique génétique est une branche récente de la statistique qui consiste à développer des outils d'analyse de données génomiques provenant de diverses plateformes moléculaires. L'objectif principal de ces méthodes est l'identification de facteurs de risque génétiques ou de déterminants génétiques de maladies. Nous nous intéressons ici au cas particulier de maladies complexes, causées par une interaction de facteurs environnementaux, comportementaux, psychologiques, sociaux, sanitaires, et bien sûr génétiques. La plupart des formes de cancer, le diabète, les maladies psychiatriques ou les maladies cardio-vasculaires sont des exemples de maladies complexes pouvant mettre en jeu des centaines de gènes simultanément. Les maladies complexes s'opposent aux maladies dites mendéliennes, c'est-à-dire celles qui respectent la

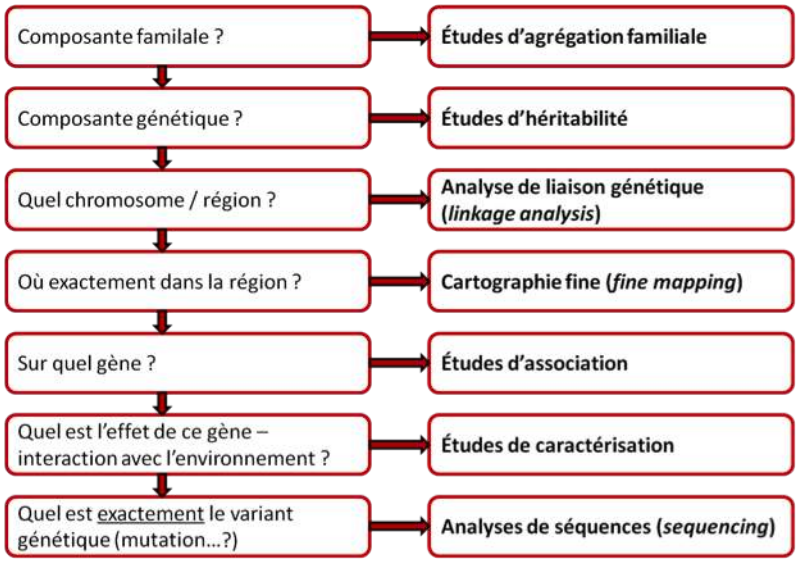


FIGURE 1 – Questions de recherche et analyses

forme dominante ou récessive mise en lumière dans les travaux de Gregor Mendel (1822–1884). Les maladies mendéliennes ont une cause exclusivement génétique et ne font intervenir qu'un seul gène dont le mode de transmission familial est très simple ; c'est le cas de l'albinisme, par exemple.

Les questions de recherche cliniques et les méthodes d'analyse statistique correspondantes sont illustrées à la figure 1. Avant d'identifier les facteurs de risque génétiques d'une maladie, une étude d'agrégation familiale est nécessaire afin d'évaluer si la maladie étudiée présente ou non une composante familiale (et donc peut-être héréditaire). Si tel est le cas, une étude d'héritabilité évaluera l'ampleur de la composante génétique et une analyse de liaison identifiera un ou plusieurs *loci*, c'est-à-dire des régions génomiques assez larges potentiellement liées à la maladie. Les études d'association permettent de raffiner la résolution de ces résultats en identifiant un ou plusieurs *loci* plus précis sur le génome, associés à la maladie. Les études de caractérisation doivent ensuite être effectuées pour évaluer l'effet de ces *loci* sur la maladie, ainsi que pour déterminer la façon dont ces *loci* interagissent entre eux et avec les facteurs environnementaux. Des études de séquençage sont alors nécessaires afin d'identifier précisément quels sont les *loci* causant la maladie.

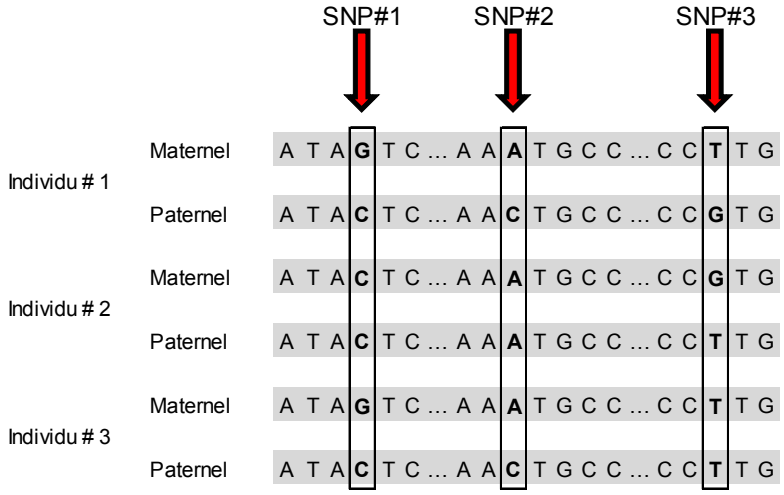
La statistique génétique est un domaine de recherche très récent ayant connu une croissance exponentielle au cours des 15 dernières années. Victime de son succès, ce domaine évolue aussi très rapidement avec la quantité phénoménale de données génétiques générées de nos jours par

les laboratoires de recherche. Les technologies évoluent aussi à un rythme effréné, donnant aux statisticiens un sentiment constant d'urgence, car le besoin de méthodes permettant l'analyse de nouvelles plateformes est en augmentation constante. La littérature dans le domaine est aussi complexe, à la frontière entre la génomique et les mathématiques. Il n'est d'ailleurs pas rare de voir des articles méthodologiques écrits par des chercheurs du domaine de la santé ou, à l'inverse, des articles cliniques écrits par des statisticiens. Le vocabulaire et la façon d'écrire étant très différents dans ces deux domaines, il est parfois difficile de s'y retrouver. De nombreux logiciels gratuits implantant les méthodes statistiques publiées sont aussi disponibles, mais leur utilisation est aux risques et périls des utilisateurs, car les tests ne sont malheureusement pas très poussés.

Le volume des données avec lesquelles les statisticiens doivent travailler constitue le principal défi à relever dans le développement de méthodes d'analyses. En effet, une étude typique d'association sur tout le génome regroupe des milliers d'individus pour lesquels les données de génotypes sont obtenues sur un million de marqueurs génétiques. Les données modernes de séquençage sont encore plus volumineuses, bien que le nombre d'individus se limite encore typiquement à quelques centaines. Les étapes permettant de générer ces données sont nombreuses et font intervenir des notions avancées de biologie moléculaire et de bio-informatique. Bien qu'il ne soit pas nécessaire de connaître en détail ces processus biologiques, travailler dans le domaine de la statistique génétique nécessite un apprentissage de la biologie moléculaire et de la génétique qui se fait généralement de façon informelle au fil du temps (bref, « sur le tas »).

2 Quelques petits rappels indispensables

Il est important de se rappeler que tout individu possède un programme génétique reçu de ses deux parents. Ce phénomène de transmission parentale est à la base de l'hérédité. L'analyse, la compréhension et le contrôle de ce programme génétique et de cette hérédité constituent la base de la génétique. L'ADN, contenu dans le noyau de toutes les cellules d'un individu, contient sous forme codée toutes les informations nécessaires à la vie d'un organisme, du plus simple au plus complexe. Durant le cycle de reproduction des cellules, la molécule d'ADN se condense sous la forme de chromosomes et un gène représente une séquence d'ADN. La molécule d'ADN a une structure hélicoïdale (c'est-à-dire à double hélice) dont les unités sont appelées nucléotides. Ces nucléotides sont formés de quatre types : adéine (A), guanine (G), thymine (T) et cytosine (C). L'ADN peut alors être grossièrement représenté par une longue séquence de lettres (environ 3 milliards), dans un alphabet de quatre lettres. Puisque chaque individu possède deux copies d'ADN (une copie reçue de son père et une copie reçue de sa mère), le code ADN d'un individu pourrait donc être représenté par deux longues séquences de lettres.



Individu	Génotype SNP#1	Génotype SNP#2	Génotype SNP#2
1	G/C	A/C	T/G
2	C/C	A/A	G/T
3	G/C	A/C	T/T

FIGURE 2 – Exemple de trois SNPs sur trois individus

3 Avec quelles données les statisticiens travaillent-ils ?

On estime que le génome humain (une copie de l'ADN) contient environ 3 milliards de bases (A, T, G, C). Il faut toutefois savoir que 99.9% de l'ADN entre deux individus sélectionnés aléatoirement dans la population est identique. Lorsqu'une étude génétique est faite pour identifier les régions de l'ADN associées à une maladie d'intérêt, l'idée est donc évidemment d'aller extraire chez des sujets les seuls *loci* (« localisation ») de l'ADN qui varient entre individus. Ces *loci* sont communément appelés « variants génétiques ». Même si l'ADN varie peu d'un individu à l'autre, le nombre de variants génétiques est assez grand, puisqu'un nucléotide (A, T, G, C) sur 1200 en moyenne varie d'un individu à l'autre. Il existe plusieurs types de variation de l'ADN entre individus : une personne peut avoir des nucléotides supplémentaires à un endroit, ou au contraire une délétion d'un segment d'ADN. Mais les différences sur un seul nucléotide dans une région sont de loin le type de variation génétique le plus fréquent. Ce type de variant génétique est appelé SNP (*Single Nucleotide Polymorphism*) et est illustré à la figure 2. Chaque « orthographe » d'une région de l'ADN est appelée « allèle », et l'ensemble des allèles d'un individu sur ses deux copies d'ADN (maternel et paternel) constitue son génotype. Par exemple, l'individu 1 de la Figure 2 a le génotype GC sur le SNP1 illustré.

Bien évidemment, les données illustrées à la Figure 2 doivent être converties en format numérique avant d'être analysées. Il existe plusieurs façons de coder des génotypes numériquement. Prenons le cas du SNP1 dans la Figure 2, par exemple. Supposons aussi que l'allèle C de ce SNP soit l'allèle le moins fréquent dans la population à ce SNP (aussi appelé allèle mineur). Le génotype peut alors être codé en sommant le nombre d'allèles mineurs pour chaque individu. Par exemple, les individus 1, 2 et 3 ont respectivement 1, 2 et 1 allèle(s) mineur(s) au SNP1. Les génotypes sont donc codés 1, 2, 1 pour ces trois individus. Ce type de code est appelé modèle « additif ». Il est aussi possible de coder les génotypes selon un modèle « récessif », dans lequel la variable de génotype numérique est une variable dichotomique signalant la présence de deux allèles mineurs dans le génotype. Selon ce modèle, les génotypes du SNP1 de la figure 2 sont donc respectivement codés 0, 1, 0 pour les trois individus.

4 Exemple d'une étude d'association génétique : sujets indépendants

Les études d'association permettent d'évaluer statistiquement l'association entre les génotypes à un ou plusieurs SNP et un trait relié à la maladie étudiée. Ce trait mesuré peut être dichotomique (atteint, non atteint) ou bien continu (taux de sucre dans le contexte du diabète, par exemple). Nous considérons ici un exemple dans lequel on cherche à évaluer si le SNP rs9939609, localisé sur le gène FTO (chromosome 16), est associé à l'obésité. Le SNP rs9939609 a deux allèles dans la population : T et A. La fréquence de l'allèle A (allèle mineur) est de 45% chez les Caucasiens, 52% chez les Africains et 14% chez les Asiatiques. Supposons aussi que l'indice de masse corporelle (IMC) soit mesuré sur un échantillon de sujets dont les génotypes au SNP rs9939609 sont aussi disponibles (ceci requiert un échantillon de sang ou de salive). Dans un premier temps, considérons l'association entre le SNP rs9939609 et le trait dichotomique indiquant la présence d'obésité chez un sujet (défini tel que l'IMC est supérieur à 30). Les données peuvent alors être représentées sous la forme du tableau 1.

Un simple test statistique du khi-deux permet de tester s'il y a une association entre le statut d'obésité et le génotype. Dans notre exemple, un tel test donne un seuil observé $p \approx 0.002$, ce qui nous permet de conclure que le SNP rs9939609 semble être associé à la présence de l'obésité.

Statut	**Génotype au SNP rs9939609**		
	TT	AT	AA
Obèse	67	138	37
Non obèse	358	435	115

Tableau 1 – Exemple de données génétiques sur 1150 sujets utilisées en étude d'association entre un SNP et une maladie.

Lorsque l'on souhaite prendre en compte certaines covariables dans le modèle, telles que le sexe ou l'origine ethnique des sujets (nous avons vu que les fréquences d'allèles peuvent varier beaucoup entre les populations), on peut faire appel à un modèle de régression logistique, soit

$$\ln\left\{\frac{\Pr(Y_i=1)}{1-\Pr(Y_i=1)}\right\} = \beta_0 + \beta_1 Geno_i + \beta_2 Sexe_i + \beta_3 Ethnic_i + \epsilon_i,$$

où Y_i est le statut du sujet i (obèse = 1, non-obèse = 0), $Geno_i$ est son génotype au SNP (codé selon un modèle additif, dominant ou récessif) et $Ethnic_i$ est son origine ethnique. Quant à ϵ_i, il s'agit d'un terme aléatoire rendant compte de variations individuelles non contrôlées. Dans ce modèle, le paramètre β_1 représente l'association entre le SNP et le trait étudié.

Il est aussi possible d'utiliser directement le trait continu IMC dans notre analyse. Dans ce cas, un modèle de régression linéaire pourra être employé, à savoir :

$$IMC_i = \beta_0 + \beta_1 Geno_i + \beta_2 Sexe_i + \beta_3 Ethnic_i + \epsilon_i,$$

où IMC_i représente l'IMC du sujet i.

Dans les exemples précédents, les individus participant à l'étude ont été sélectionnés de façon indépendante et, a priori, ne sont pas liés entre eux. Un autre type de plan expérimental couramment utilisé dans les études génétiques est un devis familial, dans lequel des familles entières sont échantillonnées de façon non aléatoire, souvent parce que plusieurs membres de la famille souffrent d'une même maladie. Dans ce type de devis, bien sûr, les individus ne sont pas indépendants les uns des autres au sein des familles et la plupart des méthodes statistiques couramment utilisées ne sont pas applicables, l'hypothèse d'indépendance des données étant violée dans les modèles de régression. Le cas de données familiales requiert donc le développement de méthodes spécifiques à ce contexte.

5 Le cas d'études familiales

Un devis expérimental couramment utilisé est l'échantillonnage de trios (père, mère, enfant), dans lesquels « l'enfant » (qui peut aussi être un adulte) est affecté par la maladie. Considérons l'exemple de la figure 3, dans lequel cinq trios sont échantillonnés et génotypés sur un SNP à trois allèles notés A, B et C. Le test le plus simple d'association familiale entre un SNP et une maladie (rappelons que tous les « enfants » échantillonnés sont atteints par la maladie étudiée) s'appelle le TDT (*Transmission Disequilibrium Test*). Le principe est d'évaluer s'il existe une sur-transmission d'un allèle en particulier, des parents à leur enfant affecté. La transmission parentale dans ce petit jeu de données peut être résumée comme au tableau 2.

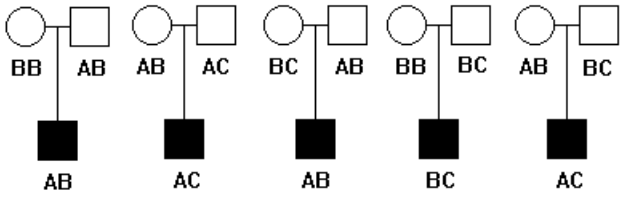


FIGURE 3 – Exemple de cinq trios échantillonnés et génotypés sur un SNP à trois allèles notés A, B et C

Allèle	**A**	**B**	**C**	**Total**
Transmis	4	3	3	10
Non transmis	1	8	1	10
Total	5	11	4	20

TABLEAU 2 – Transmissions parentales des allèles du SNP étudié.
Par exemple, l'allèle A du SNP est transmis quatre fois sur cinq d'un parent à son enfant.

Le TDT consiste essentiellement à faire un test du khi-deux, dans lequel on évalue si les transmissions observées dévient de l'hypothèse nulle de transmission indépendante. En effet, si le SNP étudié n'est pas associé à la maladie, chaque parent devrait transmettre un de ses deux allèles à son enfant affecté avec probabilité 1/2. En revanche, si nous observons une déviation de cet équilibre – l'allèle A étant transmis aux enfants affectés dans beaucoup plus que dans 50% des cas, par exemple – cela peut signifier que cet allèle est associé à la maladie.

Ce test du TDT peut aussi être généralisé au cas de traits continus (dans l'exemple précédent, le trait étudié était dichotomique puisque nous ne considérions que le statut d'affecté des enfants). Notons Y_i le trait (continu) du sujet i (l'IMC, par exemple) et G_i sa variable de génotype. Notons aussi G_{im} et G_{ip} les variables de génotype de la mère et du père du sujet i au SNP étudié. Le modèle permettant de tester l'association entre le trait continu et le SNP s'écrit comme suit :

$$Y_i = \alpha + \beta Z_i + \epsilon_i, \quad Z_i = G_i - E_{H_0}\left[G_i | G_{im}, G_{ip}\right].$$

L'idée de ce modèle est de tester l'association entre Y_i et une variable Z_i mesurant la déviation de transmission observée chez le sujet i par rapport à l'hypothèse d'indépendance de transmission. Cette association est mesurée grâce au paramètre β du modèle. Plus précisément, la variable Z_i représente la différence entre le génotype observé chez l'enfant et ce qu'il aurait dû être, sachant le génotype de ses deux parents, si la probabilité de transmission de chacun des deux allèles était de 50% chez les parents (ce qui signifie que le SNP n'est pas associé à la maladie, tel que spécifié par l'hypothèse nulle). Le test d'association entre le trait et le SNP teste l'hypothèse $\beta = 0$. Ce modèle peut être généralisé au cas de familles de toutes tailles (et pas seulement des trios) mais nous n'entrerons pas dans les détails ici, ces modèles étant très complexes.

Avant de clore la discussion, la figure ci-dessous illustre la façon dont les données familiales sont représentées et codées dans les fichiers de données.

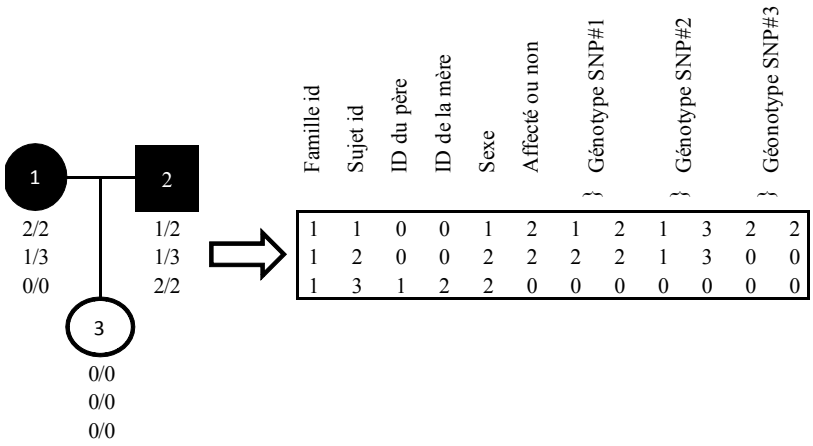


FIGURE 4 – Exemple d'une famille (trio) génotypée sur trois SNPs.

Noter que le lien familial entre les individus d'une même famille est établi dans le fichier grâce aux variables d'identification (ID) du père et de la mère.

6 Les études d'association sur tout le génome (GWA)

Les deux sections précédentes ont illustré comment on pouvait tester l'association entre un SNP et un trait. Or, notre ADN contient des millions de SNPs et les chercheurs sont souvent intéressés à évaluer l'association entre le trait et chacun des SNPs sur le génome. Ces genres d'études sont appelées études d'association sur tout le génome ou GWA (*Genomewide Association*). L'historique des GWA est assez récent, tout comme l'historique des études génétiques d'association en général. Par exemple, le premier gène de maladie mendélienne (maladie non-complexe) n'a été découvert qu'en 1983 dans le cadre de la maladie de Huntington. Le projet Hapmap, permettant l'identification de millions de SNP sur le génome humain, a été complété entre 2002 et 2007 et la publication de la toute première GWA date de 2002. Entre 2005 et 2007, les plateformes de génotype à haut débit permettant de génotyper des centaines de milliers de SNP simultanément sont devenues plus accessibles aux chercheurs (en terme de coût) et les études de GWA ont commencé à devenir de plus en plus populaires.

Conceptuellement, les études de GWA nécessitent la répétition d'une analyse d'association entre un SNP et un trait, et ce des centaines de milliers de fois (pour chaque SNP). Cependant, de telles analyses présentent de nombreux défis statistiques. Un des tout premiers défis identifiés

par les statisticiens est lié à la multiplicité des tests d'hypothèse. En effet, dans de telles analyses, des centaines de milliers de tests doivent être effectués puisque l'association doit être testée pour chaque SNP individuellement. Ce problème de multiplicité des tests est bien connu en statistique.

Avec le développement de la statistique génétique, le développement de méthodes statistiques permettant de contrôler le taux d'erreur est devenu un domaine de recherche en soi. En particulier, au lieu de contrôler l'erreur de type I pour chaque test, la procédure de rejet de l'hypothèse nulle doit être adaptée afin de contrôler le taux de faux positifs (c'est-à-dire le taux de SNPs faussement identifiés comme étant associés au trait) sur les milliers de tests. En général, pour une étude classique de GWA, la valeur du seuil de rejet de l'hypothèse nulle est de l'ordre de 10^{-8}. Hélas, la puissance nécessaire à l'obtention de tels seuils observés n'est pas toujours présente, au grand dam des chercheurs cliniciens. Un autre défi important provient du fait que les SNPs testés ne sont pas indépendants les uns des autres. Cela a plusieurs conséquences, notamment au niveau des procédures de tests multiples qui doivent prendre en compte cette dépendance. En effet, plus les SNPs sont proches l'un de l'autre sur le génome, plus ils sont corrélés ou « liés », c'est-à-dire transmis ensemble d'un parent à l'enfant. Cette corrélation peut être mesurée de diverses façons en utilisant des mesures de déséquilibre de liaison. Nous n'entrerons pas ici dans les détails.

7 Conclusion

Même si elle en est encore à ses débuts, la statistique génétique est en pleine effervescence. À ce jour, cependant, les découvertes cliniques ne sont malheureusement pas encore au rendez-vous. En effet, plusieurs variants génétiques ont été identifiés pour un grand nombre de maladies complexes, mais l'amplitude des effets génétiques sur les traits étudiés reste très faible. Par exemple, des études récentes montrent qu'une vingtaine de variants génétiques ont été identifiés comme étant associés au diabète de type 2, mais ces *loci* expliquent seulement 6% de l'héritabilité de cette maladie. Il est donc très probable que les maladies complexes résultent d'une interaction entre des centaines de *loci*, tous ayant des effets marginaux assez faibles.

La génomique se tourne maintenant vers les variants rares, c'est-à-dire les SNP ayant une fréquence d'allèle mineur de moins de 1% dans la population. L'hypothèse sous-jacente est que ces variants rares pourraient avoir un effet génétique plus important que les variants dits « communs ». Cependant, l'étude des variants rares présente elle aussi de nombreux défis statistiques. En effet, l'inférence est compliquée par le fait que de tels variants sont présents chez très peu de sujets dans un échantillon. Plusieurs méthodes statistiques sont en cours de développement sur le sujet.

Nous avons essayé de donner ici un bref survol de ce qu'est la statistique génétique et des défis

auxquels les statisticiens sont confrontés. Nous n'avons fait qu'effleurer le sujet ; les thèmes abordés ne sont pas exhaustifs. Pour de plus amples détails, les lecteurs pourront consulter [1, 2, 3, 4] ou [5].

Références

[1] Balding, D.J. (2006). A tutorial on statistical methods for population association studies. *Nature Review Genetics*, 7, 781–791.

[2] Bansal, V., Libiger, O., Torkamani, A. & Schork, N.J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, 11, 773–785.

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

[4] Spielman R.S., McGinnis, R.E. & Ewens, W.J. (1993). Transmission test for linkage disequilibrium : the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics*, 52, 506–16.

[5] Visscher, P., Hill, W.G. & Wray, N.R. (2008). Heritability in the genomics era—concepts and misconceptions. *Nature Review Genetics*, 9, 255–266.