Sélection de marqueurs pour la détection d'interactions de gènes

Chloé Friguet, Mathieu Emily

Sélection de marqueurs biologiques pour la détection d'interaction de gènes

Chloé Friguet[1] & Mathieu Emily[2]

[1] *chloe.friguet@univ-ubs.fr, LMBA - Univ. de Bretagne-Sud / IUT de Vannes*
[2] *mathieu.emily@univ-rennes2.fr, IRMAR - Univ. Rennes 2*

Résumé. Nous proposons une nouvelle méthode de sélection de marqueurs biologiques permettant la détection d'interaction de gènes en association avec le développement de maladies complexes. A partir d'un ensemble de marqueurs, notre méthode extrait un sous-ensemble de marqueurs qui caractérise de façon optimale la variabilité de la totalité des couples de marqueurs. Nous quantifions la corrélation d'un couple quelconque de marqueurs par un couple de marqueurs sélectionnés par l'information mutuelle normalisée. La faisabilité de notre méthode a été démontrée sur un ensemble de jeu de données simulé à partir d'un jeu de données de référence. De plus, la comparaison de notre méthode avec les stratégies existantes démontre d'une part la puissance de notre méthode et d'autre part un meilleur contrôle de la proportion de faux positifs.

Mots-clés. Etude d'association pangénomique, interaction de gènes, information mutuelle.

Abstract. We propose a novel method for tagging single nucleotide polymorphism, or SNP, in genome-wide association studies. The aim of our method is to select a set of tag-SNPs that optimally represent the total set of all pairs of SNPs. The correlation between two pairs of SNPs is measured by the normalized mutual information. To demonstrate its feasibility, we apply our method to a set of simulated datasets obtained from a reference panel of individuals. Furthermore, the comparison of our method with existing tagging strategies proved that, on the one hand our method is powerful and that, on the other hand, it significantly decreases the proportion of false discovery.

Keywords. Genome-wide association studies, Gene-gene interaction, mutual information.

1 Introduction

Le séquençage du génome humain combiné à la finalisation du projet HapMap (The International HapMap Project, 2003) ont permis le développement des études d'association à l'échelle du génome. Ces études ont pour objectif la détection de différences de fréquence d'allèle entre une population d'individus malades et une population d'individus sains. Chaque individu est représenté par une variable réponse, que nous supposerons binaire et

qui caractérise le phénotype "malade" ou "sain". De plus, nous disposons pour chaque individu de K mesures de génotypes évaluées par K marqueurs nucléotidiques, répartis le long du génome. Du point de vue statistique, le but des études d'association est de détecter une association significative entre l'ensemble des mesures de marqueurs et le phénotype.

Depuis leur mise en place, les études d'association sont un grand succès et ont permis la détection d'un très grand nombre de gènes statistiquement associés au développement de différentes maladies (Johnson et O'Donnell, 2009). Cependant, ces associations statistiques peinent à se traduire en associations biologiques qui permettraient, à terme, le développement de traitements pour les maladies étudiées. Le problème d'héritabilité manquante, pour lequel les interactions entre les marqueurs joue un rôle fondamental, est souvent avancé comme la cause du manque d'interprétation biologique des résultats (Zuka *et al.*, 2012). La détection d'interaction entre plusieurs marqueurs par une analyse sur tout le génome reste toutefois infructueuse. Même dans le cas d'interaction entre deux marqueurs, le manque de puissance s'avère être une limite importante.

Nous nous intéressons dans ce travail à l'effet de la sélection des marqueurs sur la puissance des tests d'interaction. En effet, la structure par blocs du génome génère de la corrélation entre les marqueurs. Ainsi, pour des raisons financières et statistiques, les marqueurs réellement évalués sont sélectionnés pour être le plus représentatif de l'ensemble des marqueurs. En pratique, la sélection des marqueurs consiste à conserver un sous-ensemble de marqueurs, appelés tag-SNP, tel que chaque marqueur est corrélé à un tag-SNP à un niveau prédéfini (en général $r^2 \geq 0.8$ où r est le coefficient de corrélation). Bien qu'efficace pour la détection d'association simple, cette stratégie n'est pas adaptée à la détection d'interaction : une paire de marqueurs n'est pas nécessairement bien représentée par la paire des tag-SNPs associés.

Dans cet article, nous proposons une stratégie de sélection de marqueurs adaptée à la détection d'association entre un phénotype et l'interaction d'une paire de marqueurs. Notre stratégie, appelée *EpiTag*, sélectionne un ensemble de tag-SNPs, tel que chaque paire de marqueurs est corrélée à une paire de tag-SNPs à un niveau prédéfini. Le niveau de corrélation entre deux paires de marqueurs est évaluée par l'information mutuelle normalisée (Strehl et Ghosh, 2002). Nous présenterons tout d'abord les détails de la méthode proposée. Puis nous comparerons les performances de puissance de notre méthode avec la stratégie couramment utilisée par la méthode *Tagger* et une stratégie sans sélection.

2 Modèle statistique

2.1 Notations

Nous posons Y la variable aléatoire caractérisant le phénotype. Y suit une loi de Bernoulli, où $Y = 1$ code pour le phénotype "malade" et $Y = 0$ pour le phénotype "sain". Soit X_i, $i = 1 \dots K$, la variable aléatoire représentant le génotype du $i^{\text{ème}}$ marqueur. En con-

sidérant que chaque marqueur est génotypique, X_i suit une loi multinomiale de paramètres (p_0, p_1, p_2), où p_0 (resp. p_1, p_2) est la probabilité que X_i soit homozygote majeur (resp. héréozygote, homozygote mineur).

Si les marqueurs sont répartis en deux régions, nous écrirons $X_i^1,\ i = 1 \dots K_1$, l'ensemble des variables modélisant les marqueurs de la première région et $X_i^2,\ i = 1 \dots K_2$ pour la seconde région.

2.2 Test du rapport de vraisemblance

Pour tester l'effet de l'interaction de deux marqueurs sur un phénotype, nous utiliserons un test de rapport de vraisemblance entre deux modèles logistiques. Ce test se formalise de la façon suivante pour la paire de marqueurs (X_i, X_j) :

$$\mathrm{LRT}(X_i, X_j) = D(\mathcal{M}_0(X_i, X_j)) - D(\mathcal{M}_1(X_i, X_j)) \sim_{\mathcal{H}_0} \chi^2(4)$$

où D est la déviance calculée pour les modèles concurents $\mathcal{M}_0$ et $\mathcal{M}_1$ suivants:

$$\mathcal{M}_0 : \mathrm{logit}\left[P(Y = 1|(X_i, X_j) = (x_i, x_j)\right] = \beta_0 + \beta_1 \mathbb{1}_{x=1}(x_i) + \beta_2 \mathbb{1}_{x=2}(x_i) + \beta_3 \mathbb{1}_{x=1}(x_j) + \beta_4 \mathbb{1}_{x=2}(x_j)$$

$$\begin{aligned}\mathcal{M}_1 : \mathrm{logit}\left[P(Y = 1|(X_i, X_j) = (x_i, x_j)\right] = {} & \beta_0 + \beta_1 \mathbb{1}_{x=1}(x_i) + \beta_2 \mathbb{1}_{x=2}(x_i) + \beta_3 \mathbb{1}_{x=1}(x_j) + \beta_4 \mathbb{1}_{x=2}(x_j) \\ & + \beta_5 \mathbb{1}_{x=1}(x_i)\mathbb{1}_{x=1}(x_j) + \beta_6 \mathbb{1}_{x=2}(x_i)\mathbb{1}_{x=1}(x_j) \\ & + \beta_7 \mathbb{1}_{x=1}(x_i)\mathbb{1}_{x=2}(x_j) + \beta_8 \mathbb{1}_{x=2}(x_i)\mathbb{1}_{x=2}(x_j)\end{aligned}$$

Sous l'hypothèse nulle de non-association, la statistique LRT suit asymptotiquement une loi du χ^2 à 4 degrés de liberté.

$$\mathrm{LRT}(X_i, X_j) \sim_{\mathcal{H}_0} \chi^2(4)$$

2.3 Sélection de tag-SNPs - *EpiTag*

Soit $X^1 = (X_1^1, \dots, X_{K_1}^1)$ et $X^2 = (X_1^2, \dots, X_{K_2}^2)$ deux ensembles de variables aléatoires. Pour sélectionner le sous-ensemble de variables le plus représentatif de l'ensemble des couples de variables $\left(X_i^1, X_j^2\right)$ pour $i = 1 \dots K_1$ et $j = 1 \dots K_2$, nous proposons une nouvelle méthode appelée *EpiTag*. La méthode *EpiTag* a pour objectif de déterminer deux sous-ensembles $T^1 = (X_{t_1^1}, \dots, X_{t_{T_1}^1})$ et $T^2 = (X_{t_1^2}, \dots, X_{t_{T_2}^2})$, où $t_i^1 \in \{1, \dots, K_1\}$ et $t_i^2 \in \{1, \dots, K_2\}$, tels que pour un seuil de couverture τ fixé on a:

$$\boxed{\begin{array}{c}\forall (i,j) \in \{1, \dots, K_1\} \quad \times \quad \{1, \dots, K_2\}, \exists (r,s) \in \{1, \dots, T_1\} \times \{1, \dots, T_2\} / \\ \\ NMI\left((X_i, X_j), (X_r, X_s)\right) > \tau\end{array}} \quad (1)$$

où

$$NMI\big[(X_i,X_j),(X_r,X_s)\big] = \frac{I\big[(X_i,X_j),(X_r,X_s)\big]}{\sqrt{H(X_i,X_j)H(X_r,X_s)}}$$

avec

$$\begin{aligned} I\big[(X_i,X_j),(X_r,X_s)\big] &= \sum_{(x_i,x_j,x_r,x_s)\in\{0,1,2\}^4} p_{(i,j,r,s)} \log\left(\frac{p_{(i,j,r,s)}}{\mathbb{P}\big[(X_i,X_j)=(x_i,x_j)\big]\mathbb{P}\big[(X_r,X_s)=(x_r,x_s)\big]}\right) \\ H(X_i,X_j) &= I\big[(X_i,X_j),(X_i,X_j)\big] \\ p_{(i,j,r,s)} &= \mathbb{P}\big[(X_i,X_j,X_r,X_s)=(x_i,x_j,x_r,x_s)\big] \end{aligned}$$

Aisni, $NMI\big[(X_i,X_j),(X_r,X_s)\big]$ est l'information mutuelle normalisée entre le couple (X_i,X_j) et le couple (X_r,X_s) (Strehl et Ghosh, 2002). I et H correspondent respectivement à l'information mutuelle et à l'entropie (Kullback, 1959).

L'implémentation de la méthode *EpiTag* s'inspire de celle de la méthode *Tagger*, couramment utilisée pour la conception des puces à ADN (de Bakker, 2005). Elle consiste à sélectionner itérativement les tag-SNPs de chaque région. A chaque itération, nous sélectionnons le couple le plus informatif comme étant celui qui représente le plus de couples à un seuil de couverture τ fixé. En posant f la fonction suivante :

$$\forall (i,j)\in U_1\times U_2,\quad f(X_i,X_j) = \sum_{(k,\ell)\in U_1\times U_2} \mathbb{1}_{NMI\left[(X_i,X_j),(X_k,X_\ell)\right]\geq\tau}$$

où $U_1 \subseteq \{1,\ldots,K_1\}$ est l'ensemble des variables X_i^1 non sélectionnées à l'étape courante et $U_2 \subseteq \{1,\ldots,K_2\}$ celui des variables X_i^2 non sélectionnées à l'étape courante. Ainsi à l'étape courante, le couple de variables sélectionnées (T_1,T_2) est défini de façon suivante:

$$(T_1,T_2) = \underset{(X_i,X_j)\in U_1\times U_2}{\operatorname{argmax}} f(X_i,X_j)$$

La sélection des tag-SNPs s'arrête lorsque l'ensemble des couples de variables (X_i,X_j) est représenté par au moins un couple de tag-SNPs.

3 Etude de puissance

3.1 Simulation d'un jeu de données

Afin dévaluer la puissance d'*EpiTag*, nous avons comparé les performances de notre méthode sur un jeu de données composé de deux régions chromosomiques positionnées respectivement sur les chromosomes 1 et 2. Pour respecter la structure du génome, nous avons utilisé le programme de simulation GWASimulator (Li et Li, 2008) en prenant

comme population de référence les individus d'origine européenne (CEU) du projet HapMap Phase 3 (International HapMap Consortium, 2003).

Après filtrage des données, nous avons retenu 22 marqueurs sur le chromosome 1 (entre les positions 2,100,000 et 2,200,000) et 22 marqueurs sur le chromosome 2 (entre les positions 1,100,000 et 1,200,000). Le phénotype de chaque individu est obtenu par simulation du modèle suivant :

$$\text{logit}(\mathbb{P}[Y = 1|(X_i, X_j) = (x_i, x_j)]) = \alpha + \alpha\theta\mathbb{1}_{x\geq 1}(x_i)\mathbb{1}_{x\geq 1}(x_j)$$

où X_i et X_j sont les marqueurs causaux, *i.e.* responsable du développement du phénotype $Y = 1$. Le modèle de maladie ainsi simulé est un modèle seuil pour lequel un individu qui possède au moins une copie de l'allèle mineure pour les deux marqueurs causaux a un risque plus élevé de développer le phénotype. Cependant, le risque reste constant si l'individu possède plus de copies des allèles à risque, modélisant un phénomène d'interaction entre les marqueurs causaux. L'évaluation de la puissance a été réalisée sur un échantillon simulé de 1,000 individus 'sains' et 1,000 individus 'malades'.

3.2 Etude comparative

Notre méthode *EpiTag* a été comparée d'une part à la méthode classique *Tagger* utilisée de façon indépendante sur les deux régions et d'autre part à une stratégie sans sélection de marqueurs, pour laquelle nous supposons que tous les marqueurs sont utilisés dans l'étude.

Pour estimer la puissance, nous avons effectué 1,000 simulations de jeux de données. Pour chaque simulation, une paire de SNP causaux a été choisie au hasard. Nous pouvons alors simuler le phénotype comme précisé dans le paragraphe précédent. Pour chaque stratégie comparée (Sans sélection, *Tagger* avec $r^2 > 0.8$, *Tagger* avec $r^2 > 0.9$, *Tagger* avec $r^2 > 0.95$, *EpiTag* avec $NMI > 0.8$ et $NMI > 0.7$), nous avons testé l'ensemble des paires de marqueurs composées d'un marqueur sélectionné dans la première région et d'un marqueur sélectionné dans la deuxième région. Nous avons ensuite appliquer une correction de Benjamini-Hochberg pour tenir compte de la multiplicité des tests.

Les résultats obtenus, résumés dans le tableau 1, montrent que notre méthode *EpiTag* est le meilleur compromis entre puissance et proportion de faux positifs. En effet, avec le seuil $NMI \geq 0.8$, la puissance obtenue est la plus forte et la proportion de faux positifs est significativement plus petit que dans le cas de non-sélection. Si on diminue encore le seuil de représentativité ($NMI \geq 0.7$), la puissance diminue légèrement mais la proportion de faux positifs s'abaissent fortement.

Nous constatons également que la puissance est quasiment nulle dans le cas où l'on utilise la méthode *Tagger* avec des seuils de 0.8 et 0.9. Enfin, la stratégie la plus puissante est la stratégie sans sélection. Cependant, dans ce cas, la proportion de fausse découverte est très importante.

Méthode	Seuil	Nombre de tests	Puissance	Proportion de faux positifs
Sans sélection		484	0.47	0.45
Tagger	$r^2 \geq 0.95$	252	0.40	0.35
Tagger	$r^2 \geq 0.90$	165	0.0	0.28
Tagger	$r^2 \geq 0.80$	126	0.0	0.27
EpiTag	$NMI \geq 0.8$	121	0.47	0.38
EpiTag	$NMI \geq 0.7$	74	0.45	0.29

Table 1: Puissance de détection et proportion de faux positifs estimée à partir de 1,000 simulations pour les six méthodes comparées.

4 Conclusion

Dans cet article, nous avons proposé une méthode de sélection de variables favorisant la détection d'effet de l'interaction entre deux variables et une variable réponse. L'application de notre méthode, *EpiTag*, à la sélection de marqueurs biologiques dans le cadre des études d'association a montré un gain important, notamment dans le contrôle des faux positifs.

Nos résultats ont également démontré que la méthode actuellement utilisée (*Tagger* avec un seuil $r^2 \geq 0.8$) est très peu puissante. Par cette constatation, nous pouvons penser que l'absence de détection d'interaction dans les études d'association s'explique en partie par la conception des puces à ADN. L'utilisation de notre méthode *EpiTag* ouvre la voie à la découverte de nouveaux mécanismes régissant l'étiologie de maladies complexes.

Bibliographie

[1] International HapMap Consortium (2003) The International HapMap Project, *Nature*, **426**:789-796.
[2] Johnson, A. D. et O'Donnell, C. J. (2009) An Open Access Database of Genome-wide Association Results, *BMC Medical Genetics*, **10**.
[3] Zuka, O., Hechtera, E., Sunyaev, S. R. et Landera, E. S. (2012) The mystery of missing heritability: Genetic interactions create phantom heritability, *PNAS*, **109**:1193-1198.
[4] Kullback S. (1959) *Information theory and statistics*, John Wiley and Sons, NY.
[5] Strehl, A. and Ghosh, J. (2002) Cluster ensembles - a knowledge reuse framework for combining multiple partitions *Journal of Machine Learning Research*, **3**: 583-617.
[6] de Bakker, P. I. W. , Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J. and Altshuler, D. (2005) Efficiency and power in genetic association studies. *Nature Genetics*, **37**: 1217-1223.
[7] Li, C. et Li, M. (2008) GWAsimulator: a rapid whole-genome simulation program, *Bioinformatics* **24(1)**: 140-142.