Combinaison de tests dépendants en études d'association pangénomiques

Florian Hébert [1] & Mathieu Emily [2] & David Causeur [3]

[1] *florian.hebert@agrocampus-ouest.fr*
[2] *mathieu.emily@agrocampus-ouest.fr*
[3] *david.causeur@agrocampus-ouest.fr*

Agrocampus Ouest (IRMAR, UMR CNRS 6625), 65 rue de Saint-Brieuc, 35000 Rennes

Résumé. Les études d'association pangénomiques visent à déterminer des liaisons entre une maladie et des marqueurs génétiques. Pour cela, des tests d'indépendance sont réalisés entre la maladie et chaque marqueur. Le contrôle du taux de faux positifs est d'autant plus difficile que le nombre de marqueurs est très grand et leur dépendance marquée par une structure en blocs. Nous proposons de combiner d'abord les tests au niveau de chaque bloc par une méthode adaptée à la dépendance, puis de corriger les p-values combinées obtenues à la première étape pour contrôler le taux de faux positifs. Une méthode de combinaison de tests basée sur la décorrélation des statistiques est introduite, dont les performances dans un cadre de forte corrélation et de signal *sparse* sont intéressantes.

Mots-clés. GWAS, tests multiples, décorrélation

Abstract. Genome-wide association studies aim at determining associations between a disease and genetic markers. To this end, independence tests are realized between the disease and each marker. Controlling the false discovery rate is challenging because the number of markers is large and their dependence show a strong block structure. We propose to combine first the tests at the level of each block using a suitable method, and then correct the combined p-values obtained at the first step to control the false discovery rate. A combination method based on whitening the tests statistics is also introduced, showing interesting performance in a strong correlation and sparse signal framework.

Keywords. GWAS, multiple testing, whitening

1 Introduction, problème et méthodes existantes

Les études d'association pangénomiques cas-témoins (GWAS) ont pour but d'identifier des associations entre un phénotype binaire et un ensemble de marqueurs génétiques. Le phénotype, noté Y, modélise la présence ($Y = 1$) ou l'absence ($Y = 0$) d'une maladie d'intérêt. Les marqueurs génétiques ou SNPs (Single Nucleotide Polymorphisms) sont

des variables qualitatives à trois modalités $X_i \in \{0,1,2\}, i = 1,\ldots,p$. Une stratégie d'analyse simple marqueur, pour laquelle l'association de chaque X_i avec Y est testée de façon séquentielle et indépendante, a permis l'identification d'un très grand nombre de marqueurs associés à un nombre important de maladies complexes [10]. De plus, il a été montré que les GWAS permettaient la détection de signaux relativement faibles et impliquant des marqueurs fréquents [2]. Dans le cas contraire où les marqueurs impliqués sont rares, la faible puissance de la stratégie simple marqueur s'explique notamment par une grande valeur de p ($p \approx 500\ 000$), *i.e.* plusieurs centaines de milliers de marqueurs sont testés individuellement. Pour contrôler l'erreur de type I, une correction très stricte pour les tests multiples est appliquée, limitant fortement la détection de marqueurs rares. Pour pallier ce problème, il a été proposé de combiner l'association de plusieurs marqueurs en un seul test par une approche dite SNP-set [11], qui présente l'avantage d'intégrer la structure de dépendance existante entre les X_i, liée notamment au concept de déséquilibre de liaison.

Comme illustré sur la figure 1, le génome exhibe en effet une structure de dépendance par blocs. Ainsi, avant de passer à l'échelle du génome entier, il est assez naturel de développer des méthodes statistiques à l'échelle d'un bloc. Dans la suite, nous considérerons un bloc $\mathbb{X} = [X_1,\ldots,X_m]$ de m variables. Pour chaque variable, un test individuel d'association est réalisé et l'ensemble des statistiques de test peut se résumer dans le vecteur de statistiques $\mathcal{Z} = (Z_1,\ldots,Z_m)'$ où Z_i est la statistique de test d'association du i-ème marqueur, correspondant par exemple à une statistique de Cochran-Armitage [8]. Nous ferons l'hypothèse supplémentaire que $\mathcal{Z}$ suit une loi normale multivariée :

$$\mathcal{Z} \sim \mathcal{N}_m(\mu, \Sigma)$$

où $\mu = (\mu_1,\ldots,\mu_m)'$ est un vecteur de dimension m et Σ est la matrice de corrélation associée, de taille $m \times m$. La détection d'une association à l'échelle d'un bloc revient à un problème de détection de signal pour lequel les hypothèses suivantes sont testées :

$$\begin{cases} \mathcal{H}_0 : & \forall i \in \{1,\ldots,m\}, \mu_i = 0 \\ \mathcal{H}_1 : & \exists i \in \{1,\ldots,m\}, \mu_i \neq 0 \end{cases}. \tag{1}$$

Dans le contexte des études d'association, il a récemment été proposé de considérer le problème de détection de signal dans un bloc par des approches de combinaison de p-values. En pratique, la méthode minP [3] fait partie des méthodes les plus utilisées en GWAS. Cette méthode consiste à calculer

$$P_{minP}(\mathcal{Z}) = \mathbb{P}\left[\max_{1\leq i\leq m}(|T_i|) \geq \max_{1\leq i\leq m}(|z_i|)\right]$$

avec z la réalisation observée de $\mathcal{Z}$ et $\mathcal{T} = (T_1,\ldots,T_m)' \sim \mathcal{N}_m(0,\Sigma)$. Mais cette méthode souffre de limites computationnelles dues à l'utilisation de l'intégration numérique ; il

n'est en effet pas possible de combiner plus de 1000 tests à la fois, et déconseillé de le faire pour plus de 500 ou 600 tests. Elle a par ailleurs été définie sans hypothèse sur le signal. Cependant, dans le contexte des GWAS, la détection de certains types de signaux reste un défi majeur. Par exemple, la puissance de détection d'un signal *sparse* et faible, correspondant à une faible proportion de μ_i non nuls, et au fait que les μ_i non nuls sont petits, est limitée avec la méthode minP. Pour pallier ce problème, nous proposons une méthode originale de combinaison de p-values tenant compte de la dépendance entre les tests avant le calcul des p-values, en s'appuyant sur un principe de décorrélation des statistiques de test [5] adapté aux tests d'association.

Après une description de la méthode proposée à la Section 2, nous proposons une évaluation de notre méthode sur des données simulées (Section 3).

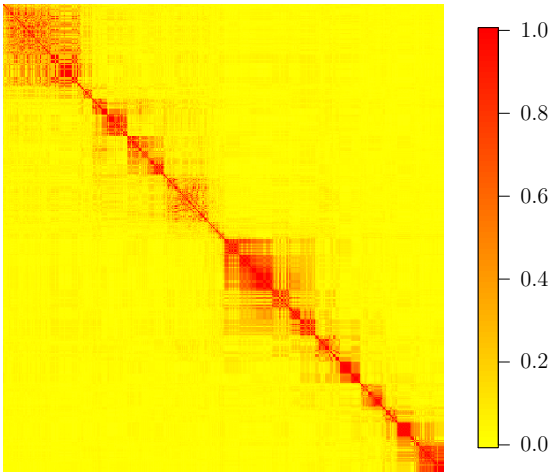

FIGURE 1 – Structure de dépendance (déséquilibre de liaison) entre SNPs ; le croisement de la ligne i et de la colonne j est une mesure de la dépendance entre les SNPs i et j

2 Méthode proposée : décorrélation de statistiques de test

La méthode que nous proposons consiste à décorréler les statistiques de test en multipliant $\mathcal{Z}$ par une matrice de décorrélation, puis à appliquer sur le vecteur décorrélé une méthode de combinaison de tests indépendants.

2.1 Matrice de décorrélation

Pour tout vecteur aléatoire X de matrice de corrélation Σ, il est possible de définir un vecteur $U = \Omega X$ dont les composantes sont décorrélées si Ω est une matrice telle que $\Omega'\Omega = \Sigma^{-1}$. Une telle matrice Ω est appelée matrice de décorrélation (*whitening matrix*, [5]). Trois choix naturels possibles pour Ω peuvent être considérés :

$$\Omega_{ZCA} = P\Lambda^{-1/2}P', \qquad \Omega_{PCA} = \Lambda^{-1/2}P' \qquad \text{et} \qquad \Omega_C = C$$

avec P et Λ la matrice des vecteurs propres et la matrice diagonale des valeurs propres de Σ, et C la factorisation de Cholesky de Σ^{-1}, qui est triangulaire supérieure. Le choix d'une matrice de décorrélation optimale pour notre problème reste une question difficile [5]. Dans la suite, nous utiliserons la matrice Ω_{ZCA}, notamment pour sa définition naturelle, étant donné qu'il s'agit de la matrice racine carrée inverse de Σ, notée $\Sigma^{-1/2}$.

2.2 Décorrélation de statistiques de test

Etant donnée une matrice de décorrélation Ω, pour un vecteur $X \sim \mathcal{N}_m(\mu, \Sigma)$, on a $\Omega X \sim \mathcal{N}_m(\Omega\mu, I_m)$. En appliquant ce résultat au vecteur de statistiques de test, il vient que $\mathcal{Z}^* = \Omega\mathcal{Z} \sim \mathcal{N}_m(\Omega\mu, I_m)$. Une p-value combinée peut alors être calculée par une méthode de combinaison de tests indépendants appliquée à $\mathcal{Z}^*$. Nous utiliserons la méthode de Tippett [9] ; en notant $p^*_{(1)}$ la plus petite p-value observée sur $\mathcal{Z}^*$, la p-value combinée est définie par

$$P_\Omega(\mathcal{Z}) = 1 - (1 - p^*_{(1)})^m.$$

La méthode de Tippett compare la plus petite p-value à sa distribution théorique sous l'hypothèse nulle ; en ce sens, elle peut être vue comme une version initiale de la méthode minP dans un cadre de tests indépendants. La décorrélation permet ici de se passer d'utiliser l'intégration numérique pour le calcul de la p-value, contrairement à l'application directe de la méthode minP au vecteur $\mathcal{Z}$.

2.3 Estimation de Σ

En pratique, la matrice de corrélation Σ de $\mathcal{Z}$ est inconnue. Toutefois, elle peut être estimée sur les X_i car la dépendance des statistiques de test est complètement héritée de la dépendance entre les X_i. Σ est donc estimée par l'estimateur des moments $\widehat{\Sigma}$ calculé sur les X_i, en les considérant en tant que variables quantitatives discrètes. En pratique, un estimateur $\widehat{\Omega}$ de la matrice de décorrélation Ω sera calculé sur $\widehat{\Sigma}$.

3 Evaluation de la méthode proposée

Nous évaluons à présent la performance de la méthode proposée et la comparons à celle de la méthode minP sur des données simulées.

3.1 Procédé de simulation

3.1.1 Simulation des SNPs et modélisation de la corrélation

Un ensemble de 1000 matrices indépendantes notées $\mathbb{X}_i$ de $m = 10$ SNPs et $n = 200$ individus ont été générées (package `R GenOrd`, [1]), chacune d'abord selon une structure d'équicorrélation, puis d'autocorrélation, de paramètre ρ dans chaque cas. La matrice de corrélation Σ correspondant à ces structures est définie respectivement par

$$\Sigma_{ij} = \begin{cases} 1 & \text{si } i = j \\ \rho & \text{sinon} \end{cases} \qquad \text{et} \qquad \Sigma_{ij} = \begin{cases} 1 & \text{si } i = j \\ \rho^{|i-j|} & \text{sinon} \end{cases}.$$

3.1.2 Sparsité de l'association

Pour chaque matrice de SNPs, un phénotype Y est généré selon le modèle

$$\begin{cases} \text{logit}(\mathbb{P}[Y = 1|\mathbb{X}_i = x]) = \sum_{j=1}^{m} \beta_j \mathbb{1}\{x_j = 2\} \\ \beta = \Sigma^{-1}\xi,\ \xi = (0, \ldots, 0, \delta, 0, \ldots, 0)' \end{cases}. \tag{2}$$

Cette définition de β est liée à la notion de vecteur potentiel (*potential vector*) dans le cadre des modèles graphiques gaussiens [6] et assure que le vecteur moyen μ du vecteur de statistiques de test $\mathcal{Z}$ aura une forme proche de celle de ξ et sera donc *sparse*. Les vecteurs de statistiques de test sont ensuite calculés, puis les p-values combinées par la méthode minP, qui semble être la méthode existante la plus adaptée, et par la méthode proposée (en décorrélant par la matrice Ω_{ZCA}).

3.2 Résultats

Les courbes de puissance estimée en fonction de δ sont représentées en figure 2. Il peut être remarqué que la méthode proposée est plus puissante, surtout pour une corrélation importante, au vu de la grande différence entre les courbes en trait plein. D'autres situations ont été étudiées, montrant dans chaque cas une supériorité de l'approche proposée pour une définition d'un signal *sparse* comme proposé dans le modèle (2).

4 Conclusion

Nous proposons une méthode de détection de signal basée sur la décorrélation de statistiques de test. La méthode proposée montre des performances intéressantes lorsque la dépendance est forte et le signal *sparse*. Elle s'inscrit par ailleurs dans la construction d'une approche à double échelle s'appuyant sur la structure par blocs du génome (voir figure 1). En supposant que les blocs sont connus et indépendants, nous pouvons envisager de combiner notre approche de détection de signal à l'échelle d'un bloc à un principe de correction pour les tests multiples afin d'intégrer l'ensemble des blocs du génome.

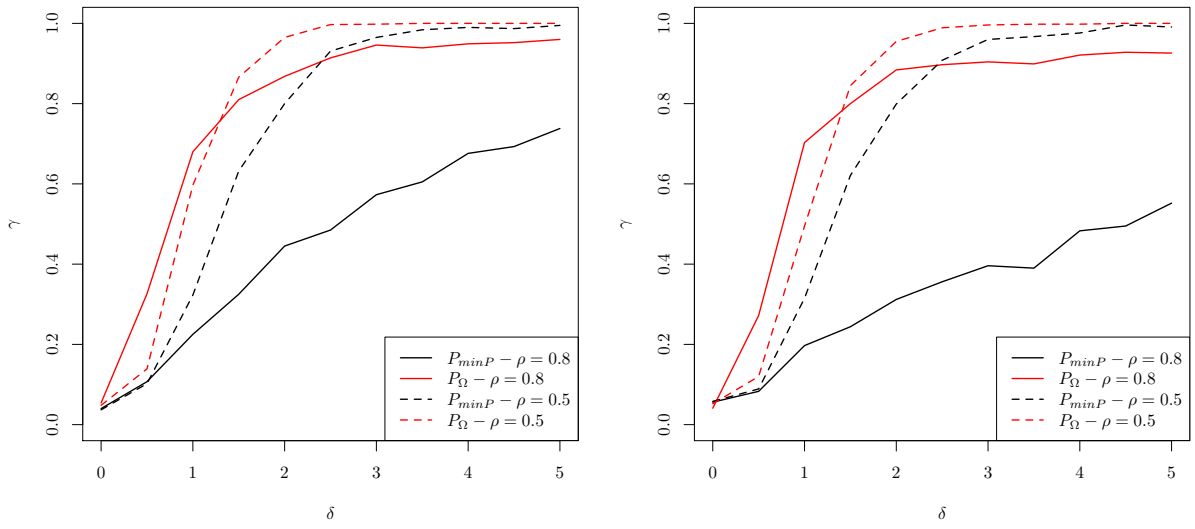

FIGURE 2 – Comparaison de la puissance de la méthode proposée à la méthode minP (équicorrélation à gauche, autocorrélation à droite)

Bibliographie

[1] Barbiero, A. et Ferrari, P. (2014). GenOrd : Simulation of ordinal and discrete variables with given correlation matrix and marginal distributions, R package version 1.2.0.

[2] Bush, W. S. et Moore, J. H. (2012). Genome-wide association studies. *PLoS Computational Biology*, 8(12) :e1002822.

[3] Conneely, K. N. et Boehnke, M. (2007). So many correlated tests, so little time ! Rapid adjustment of p-values for multiple correlated tests. *The American Journal of Human Genetics*, 81(6) :1158–1168.

[4] Friguet, C., Kloareg, M., et Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104(488) :1406–1415.

[5] Kessy, A., Lewin, A., et Strimmer, K. (2017). Optimal whitening and decorrelation. *The American Statistician*, (just-accepted).

[6] Liu, Y. et Willsky, A. (2013). Learning gaussian graphical models with observed or latent fvss. In *Advances in Neural Information Processing Systems*, pages 1833–1841.

[7] Ma, L., Clark, A. G., et Keinan, A. (2013). Gene-based testing of interactions in association studies of quantitative traits. *PLoS Genetics*, 9(2) :e1003321.

[8] Sasieni, P. D. (1997). From genotypes to genes : doubling the sample size. *Biometrics*, 53(4) :1253–1261.

[9] Tippett, L. H. C. (1931). *The Methods of Statistics*. Williams & Norgate Ltd, London.

[10] Welter, D. et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*, 42(D1) :D1001–D1006.

[11] Wu, M. et al. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6) :929–942.