Exploitation de l'association pléiotropique à l'aide de la sélection de variables exploitant la structure de groupe

P.E. Sugier (1,2), M. Sutton (3), T. Truong (2), B. Liquet (1,4)

1) Laboratoire de Mathématiques et de leurs Applications de Pau, Université de Pau et des Pays de l'Adour, UMR CNRS 5142, E2S-UPPA, France
2) CESP (Center for Research in Epidemiology and Population Health), INSERM, Paris-Saclay University, Paris-Sud University , Villejuif, France
3) Queensland University of Technology Centre for Data Science, Brisbane, QLD 4000, Australia
4) Department of Mathematics and Statistics, Macquarie University, Sydney, Australia

Introduction : Les analyses d'association pangénomiques (GWAS) se concentrent sur les tests d'associations d'un phénotype avec plusieurs millions de marqueurs génétiques (SNPs) testés de manière indépendante. Ces analyses ont permis d'identifier individuellement des centaines de milliers de SNPs associés avec de nombreux phénotypes complexes, comme les cancers, permettant ainsi de mieux comprendre comment les variants génétiques communs peuvent affecter les maladies humaines. L'une des découvertes majeures de l'ère GWAS est que la pléiotropie – le fait qu'un même gène influence plusieurs phénotypes – est très répandue dans les maladies complexes chez l'Homme. Nous pouvons tirer parti de ce phénomène pour intégrer des sources de données multiples dans une analyse conjointe, afin d'identifier de nouveaux variants génétiques pléiotropes. De plus, l'incorporation d'information biologique connue telle que la structure de groupe des données génétiques (gène ou voie biologique) peut améliorer la puissance statistique de détection de nouveaux signaux et l'interprétation biologique. Cela a montré un certain succès dans les GWAS, mais n'a pas été largement exploré dans le contexte de la pléiotropie.

Méthode : Nous avons développés des méthodes de régression jointes pour des variables réponses binaires afin d'explorer des effets pléiotropiques [1]. Ces approches incorporent des termes de pénalités de façon astucieuse pour induire de la sélection structurée de variable au niveau du groupe de variables (gène ou voie biologique) et des variables (SNP). L'estimation des differents modèles exploite l'algorithme ADMM (Alternating Direction Method of Multipliers).

Résultats : Une étude de simulation a montré que la méthode SGMT que nous avons développé était capable de détecter plus de vrais signaux que la méthode ASSET [2] tout en conservant un taux de fausses découvertes similaire. Afin d'illustrer notre approche, nous avons appliqué nos méthodes à l'étude des effets génétiques partagés entre le cancer de la thyroïde et le cancer du sein dans des voies biologiques candidates. Ces analyses ont permis de détecter plusieurs

effects pleiotropes associés à ces deux cancers : 11 associations au niveau des SNPs appartenants à 6 différentes voies biologiques. Aucune association n'avait été détectée à l'aide de la méthode ASSET. Un package R regroupant ces méthodes est disponible : https://github.com/matt-sutton/SGMT.

1. Sutton M, Sugier PE, Truong T, Liquet B. Leveraging pleiotropic association using sparsegroup variable selection in genomics data. *En soumission*.
2. Bhattacharjee S, Rajaraman P, Jacobs KB, et al. A Subset-Based Approach Improves Power and Interpretation for the Combined Analysis of Genetic Association Studies of Heterogeneous Traits. *The American Journal of Human Genetics* 2012;90(5):821–835.