BALD : Etude d'association par blocs de déséquilibre de liaison

A. Dehman[a], **P. Neuvial**[a] et **C. Ambroise**[a]

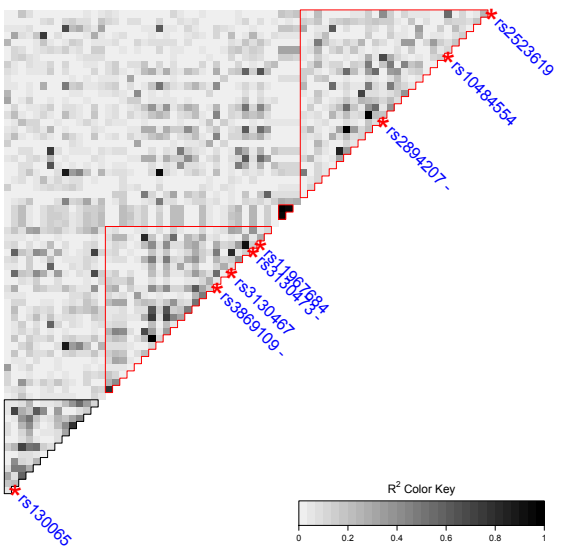
[a]Laboratoire de Mathématiques et Modélisation d'Evry
CNRS UMR 8071 - Université d'Evry val d'Essonne - USC INRA
23, Boulevard de France, 91037 Evry cedex
{alia.dehman, pierre.neuvial, christophe.ambroise}@genopole.cnrs.fr

Mots clefs : Bioinformatique, Etudes d'association génome entier, Déséquilibre de liaison, Group Lasso.

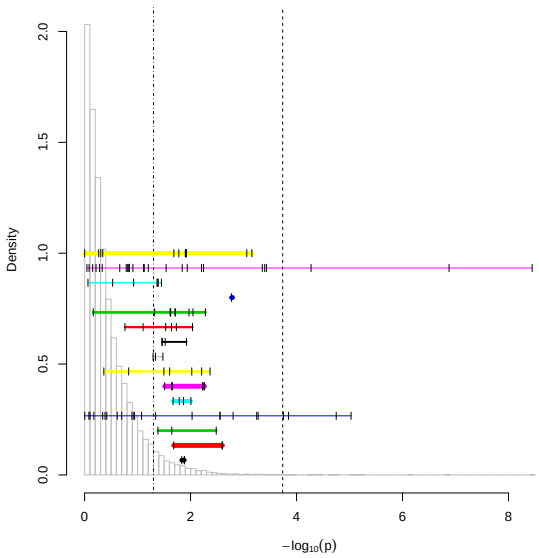
Les résultats d'études d'associations génome entier (GWAS pour Genome-Wide Association Studies) ne permettent d'expliquer qu'une petite partie du risque génétique associé à une maladie notamment parce que ces études n'intègrent pas de connaissances biologiques à priori sur la structure de dépendance entre marqueurs. Cette forte dépendance peut, entre autres, être liée au phénomène de déséquilibre de liaison (LD pour Linkage Disequilibrium) qui crée une structure de groupes entre marqueurs. Nous avons mis en place une méthode d'identification de groupes de marqueurs adjacents le long du génome, qui combine l'inférence des blocs de LD et la sélection d'un sous-ensemble de blocs associés au phénotype d'intérêt [2].

Le package `BALD` (pour Blockwise Approach using Linkage Disequilibrium) permet :

1. la génération de données GWAS structurées, c'est-à-dire des génotypes de SNPs avec une structure de groupe le long du génome ainsi que des phénotypes continus associés à ces génotypes. Les paramètres de la simulation incluent le nombre d'individus, les tailles des groupes de variables, le niveau de corrélation à l'intérieur des groupes, l'intensité de l'association, et le coefficient de détermination du modèle ;
2. l'inférence des blocs de LD à partir d'une matrice de génotypes, et ce en 2 étapes :
   (a) classification des SNP en groupes adjacents et en déséquilibre de liaison, à l'aide d'une classification ascendante hiérarchique utilisant la similarité induite par le LD et la méthode de regroupement de Ward [6]. Plutôt que de requérir le calcul initial de la matrice de similarité entre paires de marqueurs (comme c'est le cas d'une implémentation existante dédiée à un autre contexte applicatif [3]), notre implémentation prend en argument la matrice des génotypes. La complexité en mémoire est donc linéaire (et non plus quadratique) en le nombre de marqueurs ;
   (b) estimation d'un nombre de groupes optimal à l'aide de la statistique Gap [5] ;
3. l'utilisation de plusieurs méthodes de régression existantes via une interface unifiée :
   – régression univariée ;
   – régression multivariée pénalisée (Lasso [4], Elastic-Net [8]) ;
   – régression pénalisée avec pénalité groupée (Group Lasso [7]). La structure de groupes peut être donnée ou estimée à l'aide de la fonction décrite en 2 ;
4. l'évaluation des performances des différentes méthodes en termes de sélection de variables à l'aide de courbes ROC ;
5. la représentation graphique des résultats (Figure 1).



(a) Représentation des mesures de LD entre paires de 68 SNPs. Les SNPs représentés avec une étoile rouge (*) correspondent aux marqueurs sélectionnés par la méthode univariée. Les groupes sélectionnés par le Group Lasso sont délimités par un contour rouge.



(b) Chaque groupe sélectionné par le Group Lasso est représenté par un segment horizontal de couleur allant de la plus petite à la plus grande p-valeur univariée du groupe. Les segments verticaux noirs indiquent les p-valeurs de chaque SNP dans ces groupes.

FIGURE 1 – Représentations graphiques produites par le package `BALD` pour l'analyse des résultats d'une étude d'association sur le VIH [1].

Références

[1] C. Dalmasso, W. Carpentier, L. Meyer, C. Rouzioux, C. Goujard, M.-L. Chaix, O. Lambotte, V. Avettand-Fenoel, S. Le Clerc, L. D. de Senneville, et al. Distinct genetic loci control plasma HIV-RNA and cellular HIV-DNA levels in HIV-1 infection : the ANRS Genome Wide Association 01 study. *PloS one*, 3(12) :e3907, 2008.

[2] A. Dehman, C. Ambroise, and P. Neuvial. Performance of a blockwise approach in variable selection using linkage disequilibrium information. submitted, Apr. 2014.

[3] S. Juggins. *rioja : Analysis of Quaternary Science Data*, 2012. R package version 0.8-5.

[4] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[5] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 63(2) :411–423, 2001.

[6] J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301) :236–244, 1963.

[7] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68(1) :49–67, 2005.

[8] H. Zou and T. Hastie. regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(2) :301–320, 2005.