Laboratoire de Mathématiques et Modélisation d'Évry

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

Alia Dehman, Pierre Neuvial, Christophe Ambroise

Laboratoire de Mathématiques et Modélisation d'Evry (LaMME)
Université d'Évry Val d'Essonne

27-06-2014

Sommaire

Sommaire

Contexte

- Identification des variations génétiques associées à un trait phénotypique dans une population donnée.
- **Trait phénotypique :** qualitative ou quantitative
  Variables explicatives : des marqueurs biologiques - *Single Nucleotide Polymorphisms (SNP)*
- **Le modèle de régression :**

$$y_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \epsilon_i \quad , i = 1, \ldots, n$$

  - n : le nombre d'individus
  - p : le nombre de SNP
  - Y_i : la variable à expliquer pour l'individu i
  - $X_{.j}$: le génotype du SNP j (peut prendre des valeurs 0, 1 ou 2)

Parcimonie et grande dimension

Parcimonie : Seul un "petit" sous-ensemble de SNP est réellement associé au phénotype.

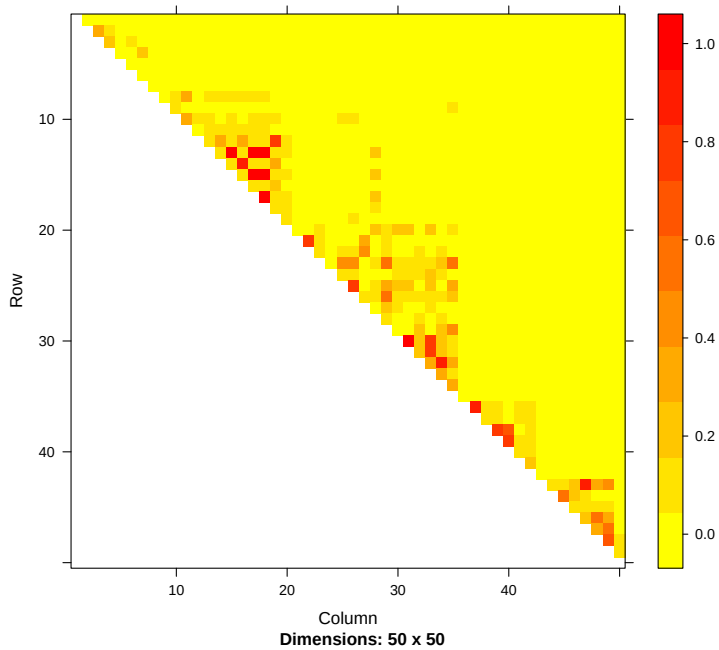
$$Card\{j, \beta_j \neq 0\} \ll p$$

Grande dimension : Plusieurs milliers de marqueurs contre quelques centaines d'observations.

$$p \gg n$$

Dépendance spatiale : LD

- Coefficients r^2 entre les **50 premiers SNP** du Chromosome 22 (Dalmasso et al. 2008)
- Présence de blocs de déséquilibre de liaison
- Il s'agit de blocs transmis intacts d'une génération à l'autre



Block-Wise Approach using Linkage Disequilibrium (BALD)

1. Classification des SNP en groupes adjacents et en déséquilibre de liaison en utilisant la similarité du LD.
2. Estimation d'un nombre de groupes optimal à l'aide de la statistique Gap.
3. Sélection des blocs associés au phénotype à l'aide du Group Lasso (Yuan et. al., 2005) :

$$\hat{\beta}^{GL} = \arg\min_{\beta} \sum_{i} \left(y_i - \mathbf{X}_{i.}\beta\right)^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g}||\beta_{\mathbf{g}}||_2.$$

Sommaire

Simulations

Génération de **génotypes de SNP** avec une structure de groupes et de **phénotypes continus** associés à ces génotypes.

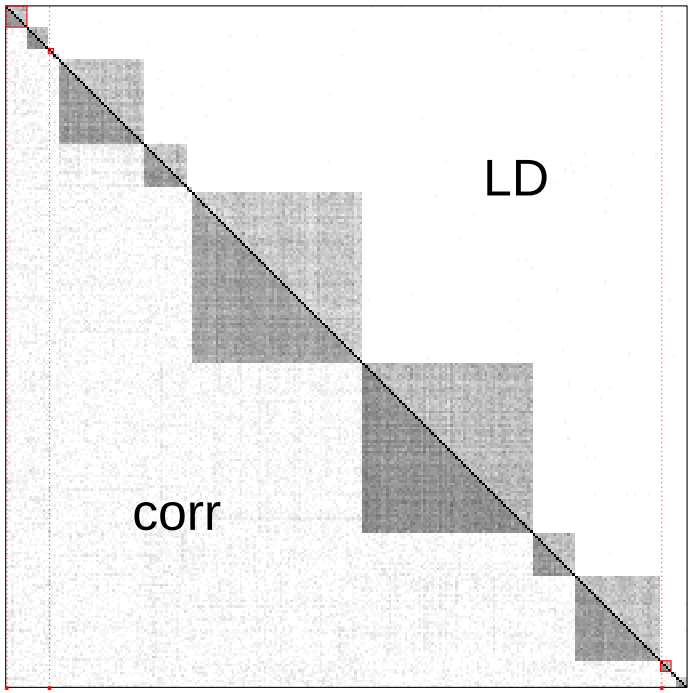
p marqueurs

phénotype associé

$$\text{n individus} \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix} \qquad \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

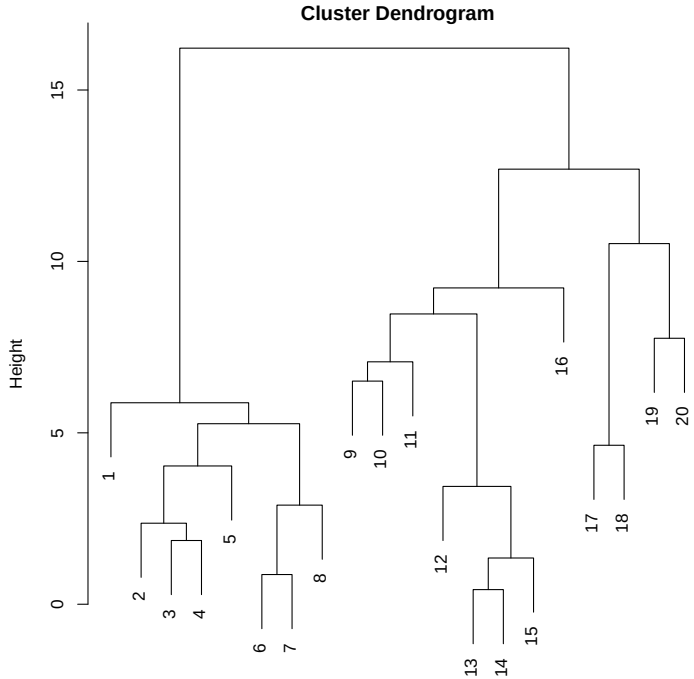
Simulations

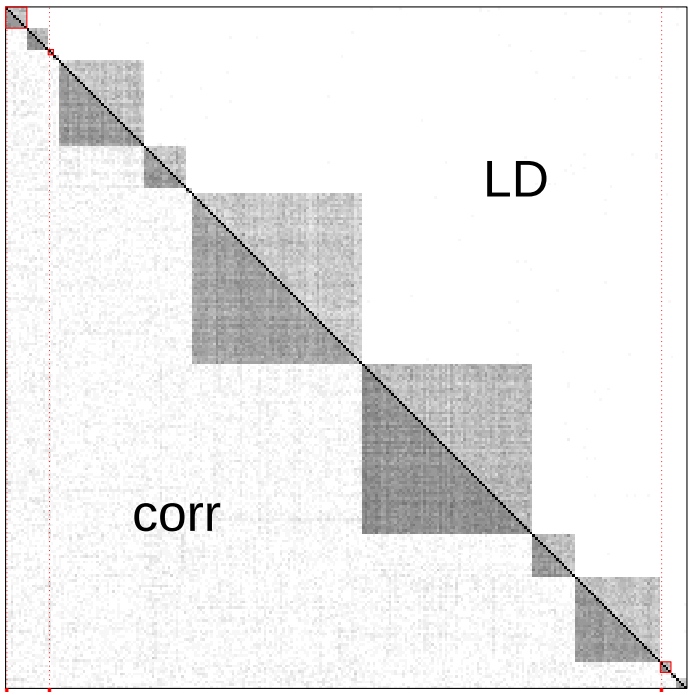
```
betas ← simBeta(blockSizes,
sig.blocks, nb.per.block)

sim ← simulation(n, betaSNP,
blockSizes, corr, R2)
```



Clustering hiérarchique avec contrainte d'adjacence

```
tree ← cWard(X, h, sim=simR2)
```



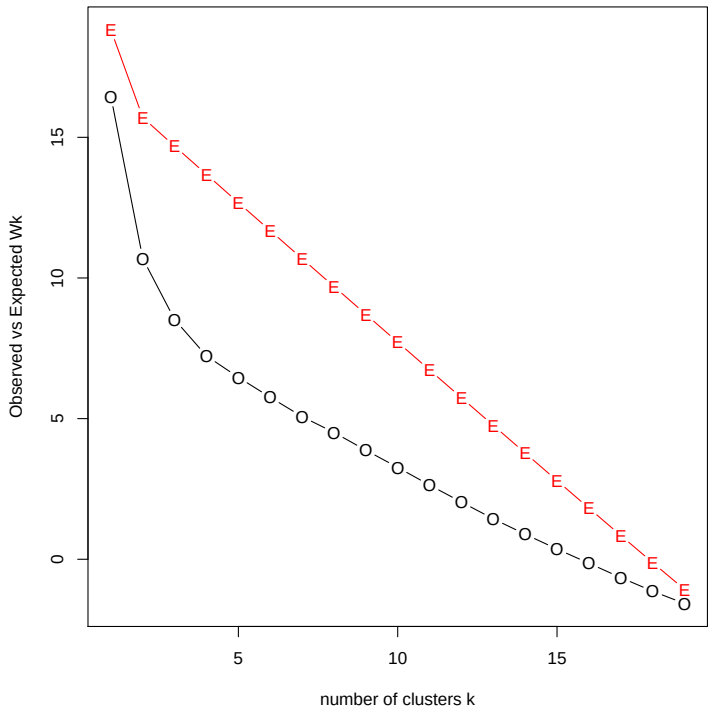


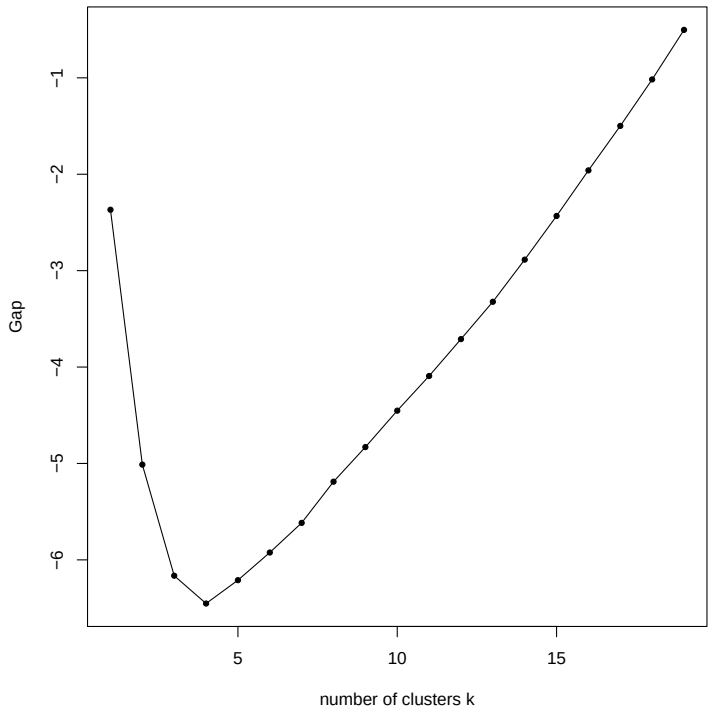
Implémentation

	rioja	**cWard**
Type d'entrée	matrice de dissimilarités $p \times p$	matrice des génotypes $n \times p$
Complexité en temps	$\mathcal{O}(np^2)$	$\mathcal{O}(np^2)$
Complexité en mémoire	$\mathcal{O}(p^2)$	$\mathcal{O}(np)$

Sélection du nombre de groupes : statistique Gap

```
gapS ← gapStatistic(X, min.nc=1, max.nc=p-1, B=500)
```





Sélection des groupes associés

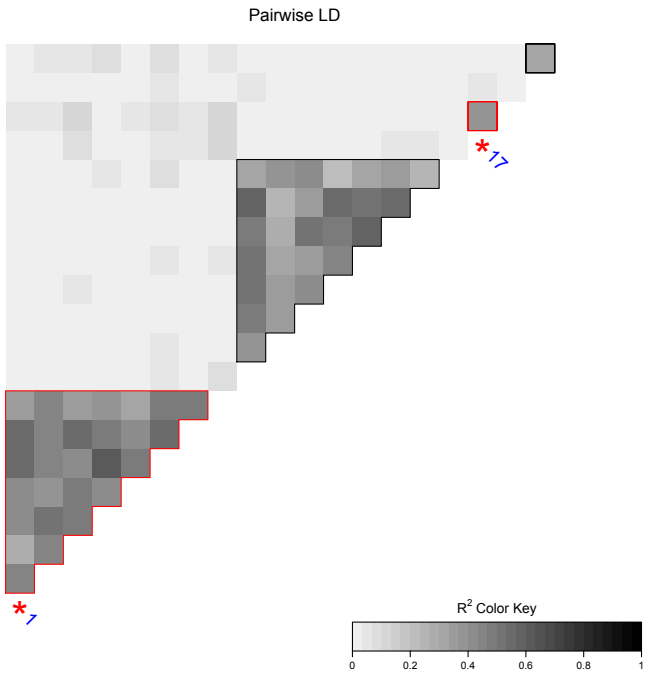
```
selGL ← select(method="groupLasso", X=X, y=y,
nlambda=100, groups=groups)
```

Ou bien : les 3 étapes sont implémentées en une seule fonction :

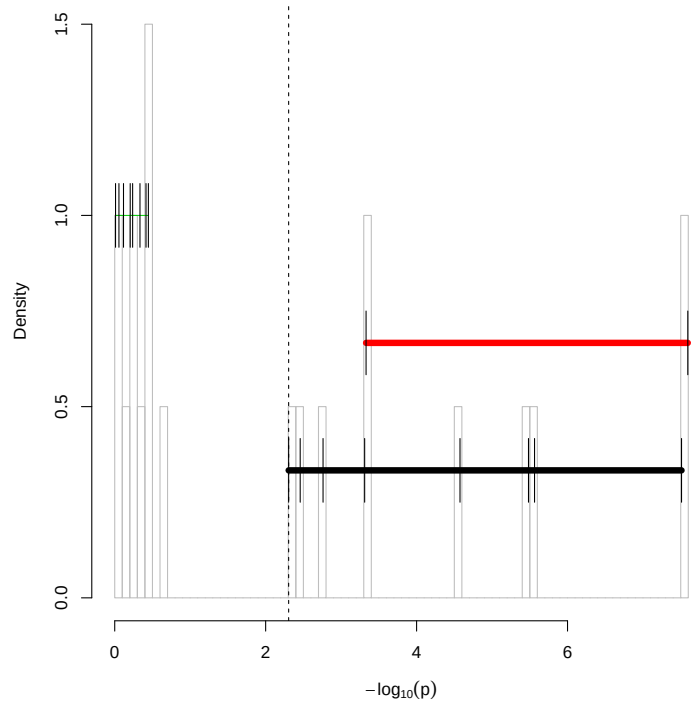
```
grplassoCWard(X, y, nlambda=100, groups=NULL )
```

Représentations graphiques des résultats

```
plotHeatmap(X, blockSizes,
selBlocks=c(1,3),
snpNames=as.character(1:20),
snpStar=c("1","17"))
```




```
plotGroupsGL(coefsGL,
nbGroup=3, pvals)
```



Sommaire

Conclusion et perspectives

Pour résumer :

- Mise en place de simulations à vérité connue.
- Paramètres interprétables.
- Implémentation et comparaison des performances de l'approche par blocs de LD avec d'autres approches.

Perspectives :

- Optimiser l'implémentation de la fonction `cWard`.
- Avoir un seuil de significativité des blocs sélectionnés par Group Lasso.

Merci à Pierre Neuvial, Christophe Ambroise et Cyril Dalmasso.

Merci pour votre attention !