FLPCA: A Fused Lasso PCA-based approach to identify influential markers in differentiated populations from dense SNP data

Denis Laloë, Julien Chiquet, Florence Jaffrezic, Mathieu M. Gautier

FLPCA: A Fused Lasso PCA-based approach to identify influential markers in differentiated populations from dense SNP data

IBC meeting
Firenze, July 2014

Denis Laloë, Julien Chiquet, Florence Jaffrézic, Mathieu Gautier

July 8, 2014

The context

Genetic structure of a population

- Natural / Artificial Selection
- Isolation, drift

Markers : SNPs

- Usually biallelic markers
- Throughout the genome
- Mapping SNPs to genes

Geometric Data Analysis

- Duality Diagram: Space of individuals vs Space of markers
- Typological Value
- Modelling through instrumental variables

Subset of influential markers

- Magnitude
- Spatial structure (Linkage Disequilibrium)

SNP Haplotype data

$$\mathbf{X} = [\delta_i^j] = \begin{pmatrix} SNP_1 & SNP_2 & \dots & SNP_k \\ 0 & 1 & \dots & 0 \\ 0 & 0 & \dots & 1 \\ . & . & \dots & . \\ 1 & 0 & \dots & 0 \end{pmatrix}$$

Duality diagram

Dray and Dufour, 2007

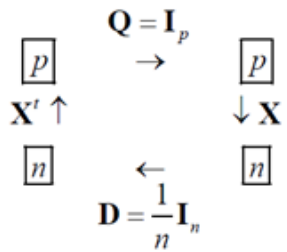
Maximisation of the correlation between variables and components

Variables

V = X'X/*n*

X'XA=AΔ
A'A=I
Principal components

Coordinates of variables
C=X'B



Diagonalisation

X'X **XX'**

same non-null eigenvalues

$\lambda_1 > \lambda_2 > \ldots > \lambda_r > 0$

$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$

Transition formulae
$XA\Delta^{-0,5} = B$
$X'B\Delta^{-0,5} = A$

Maximisation of the individuals dispersion

Individus

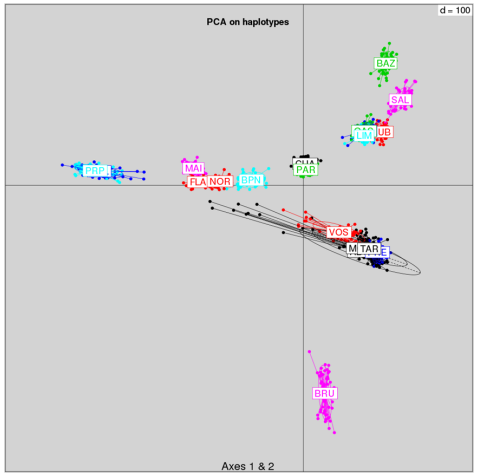
W = XX'/*n*

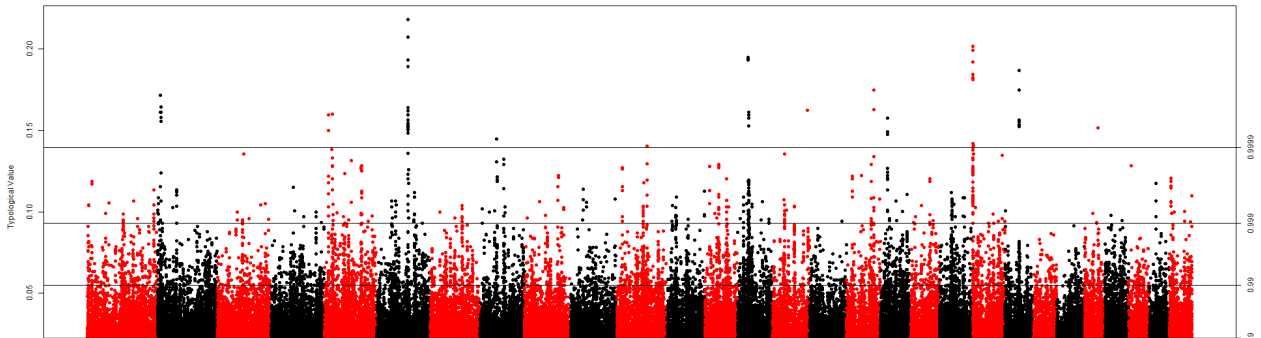
XX'B=BΔ
B'B=I
Principal axes

Coordinates of individuals
L=XA

Typological Values

- **Haplotypes**
  - coordinates of haplotypes along the ith axis $\mathbf{y}^{[i]}$
- **SNPs**
  - coordinates of SNPs along the ith component $\mathbf{c}^{[i]}$
  - $\mathbf{c}_{ij}^2$: Typological value of SNP_j with the ith component
    - R^2 of the model $y^{[i]} = \mu + SNP_j + \epsilon$
    - *Fst* (Laloë and Gautier, 2011)





Problem formulation : the Fused Lasso Signal Approximator (FLSA)

(Tibshirani *et al*,2005; Hoefling, 2010)

- $\mathbf{y} = (y_1, ...y_n)$ an ordered vector of data
- identification of consecutive points with high and constant values.
- FLSA solution

$$\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2) = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \|\mathbf{y} - \boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_{i=1}^{n} |\beta_i| + \lambda_2 \sum_{i=1}^{n-1} |\beta_{i+1} - \beta_i|. \right\}$$

  - λ_1 controls the level of sparsity
  - λ_2 controls the level of smoothness

Adaptive FLSA

A two step procedure (Rinaldi, 2009).

1. Fusion step
   1.1 Fit the FLSA model with $\lambda_1 = 0$, i.e., $\hat{\boldsymbol{\beta}}(0, \lambda_2)$.
   1.2 For the partition $\mathcal{B} = \{B_1, \ldots, B_J\}$ of J blocks (or segments) associated to $\hat{\boldsymbol{\beta}}(0, \lambda_2)$, compute:

$$\tilde{\boldsymbol{\beta}} = \sum_{j=1}^{J} \bar{y}_j \mathbf{1}_{B_j}, \quad \bar{y}_j = \text{card}(B_j)^{-1} \sum_{j=1}^{J} y_j,$$

2. Adaptive step
   Fit the following weighted lasso problem:

$$\hat{\boldsymbol{\beta}}^{\text{AFL}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \left\| \boldsymbol{\beta} - \tilde{\boldsymbol{\beta}} \right\|_2^2 + \lambda_1 \sum_{i=1}^{n} w_i |\beta_i| \right\}, \quad w_i = \sum_{j=1}^{J} \frac{\mathbf{1}_{B_j}}{\sqrt{\text{card}(B_j)}}$$

Hard thresholding : $\text{hard}(x; \lambda) = x \cdot \mathbf{1}_{\{|x| > \lambda\}}$

Model selection

- Cross Validation : Construction of fold not obvious in the case of ordered data
- Penalized Criterion

Penalized Criteria

- IC penalized criterion of the form

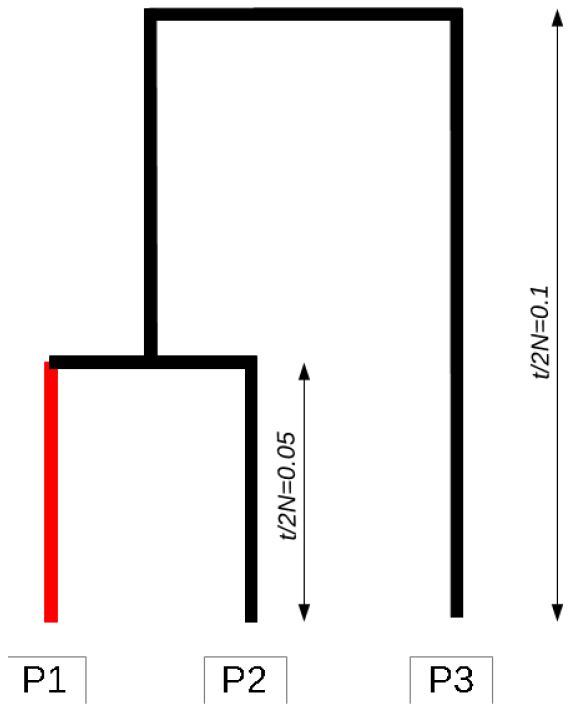
$$\mathrm{IC}(\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)) = \frac{1}{2n} \left\| \mathbf{y} - \hat{\boldsymbol{\beta}} \right\|_2^2 + \frac{\sigma^2}{n} \mathrm{pen}(\mathrm{df}),$$

  where
  - $\mathrm{pen}(\cdot)$ is a function that penalized the number of parameters of the model, df (number of segments different from zero).
  - σ^2 estimated by the plug-in estimator of P Hall (Lebarbier, 2005)
- Criteria
  - AIC: $\mathrm{pen}(\mathrm{df}) = 2 \cdot \mathrm{df}$
  - BIC: $\mathrm{pen}(\mathrm{df}) = \log(n) \cdot \mathrm{df}$
  - BML (*for Birgé, Massart, Lebarbier*; Lebarbier, 2005) : $\mathrm{pen}(\mathrm{df}) = \mathrm{df} * (2 * log(n/\mathrm{df}) + 5)$

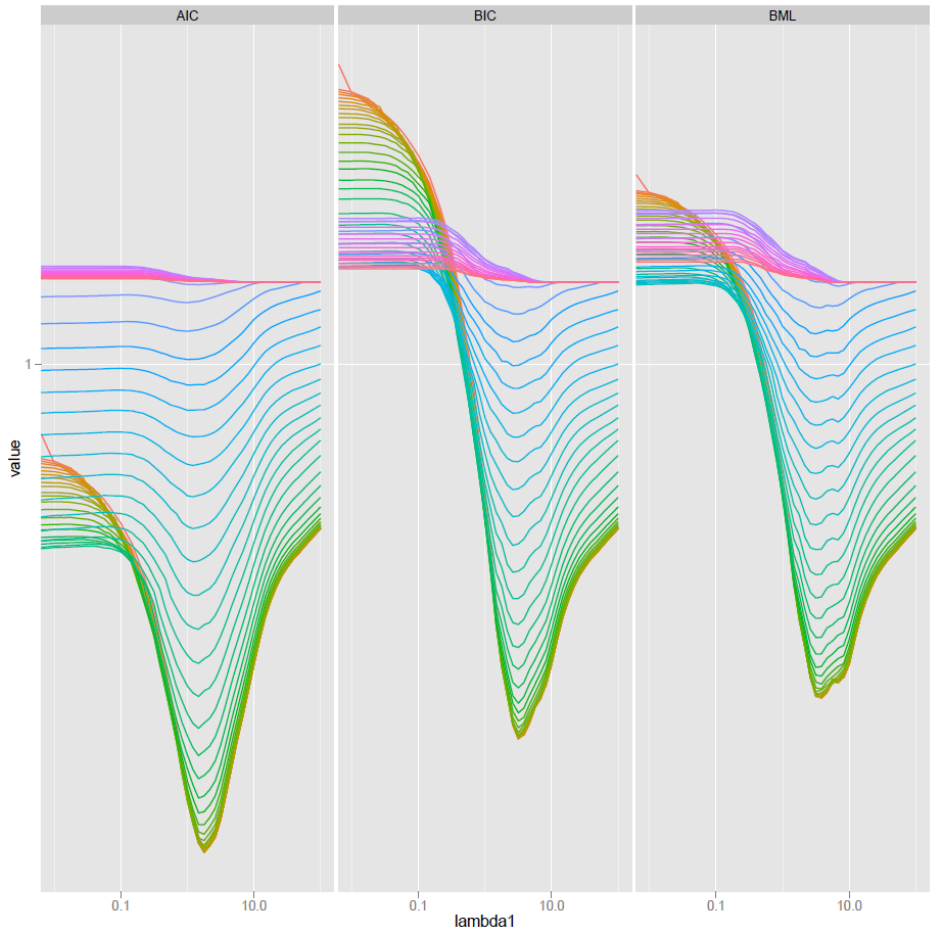
Simulations

- Three populations
- Five 5 Mb-chromosomes
- 1000 SNPs per chromosome
- 1 causal variant at position 2.5 Mb of chromosome1

  driven to fixation via selection in population P1
- Program msms coalescent simulator *Ewing and Hermison, 2010*
- 100 simulations
- First PC : Highest node P1-P2 *vs* P3

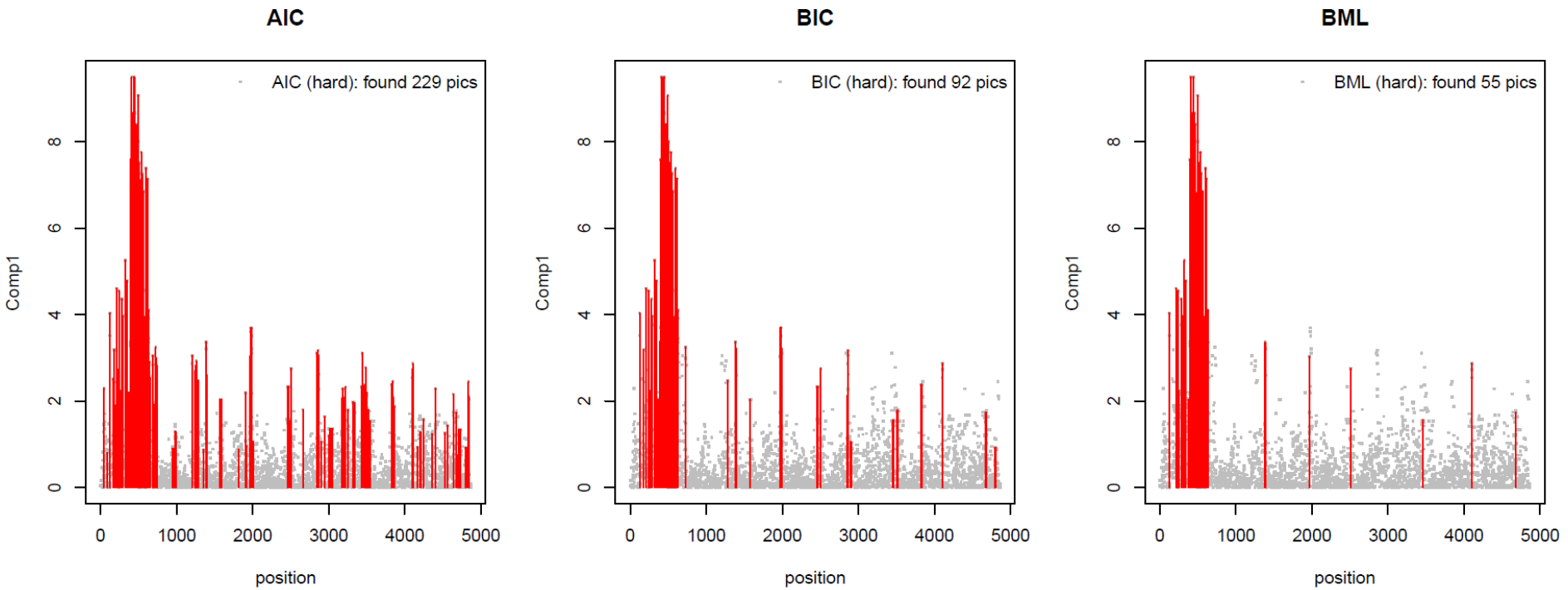


A simulation example. Parameter optimization



Choice of λ_1 and λ_2 according to the penalization criterion

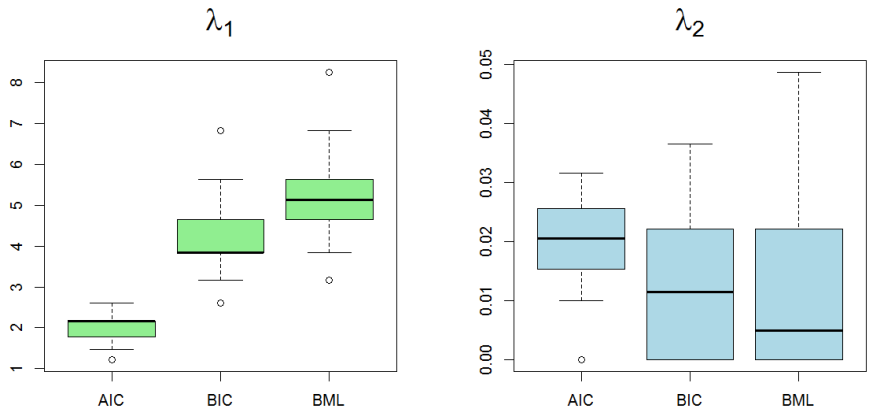
A simulation example. Selected regions



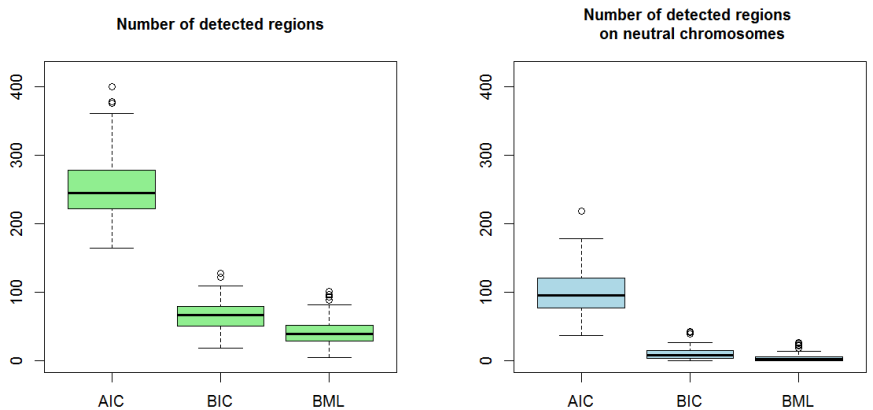
Selected SNPs according to the penalization criterion

Comparison of penalization criteria

1. Choice of parameters



2. Selected regions

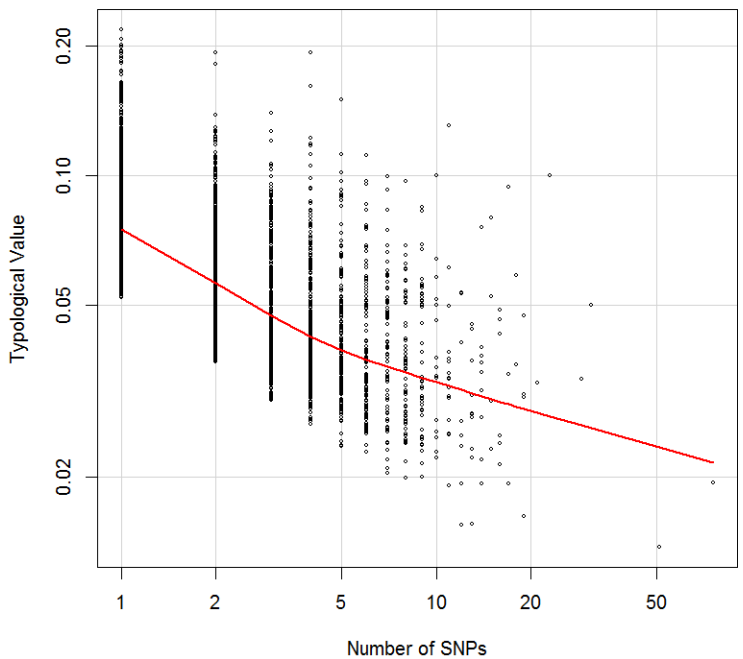


French bovine data

- 600 animals from 20 French dairy and beef cattle breeds
- HD 770k SNP array
- PCA on haplotypes
- a between **beef** vs **dairy** analysis

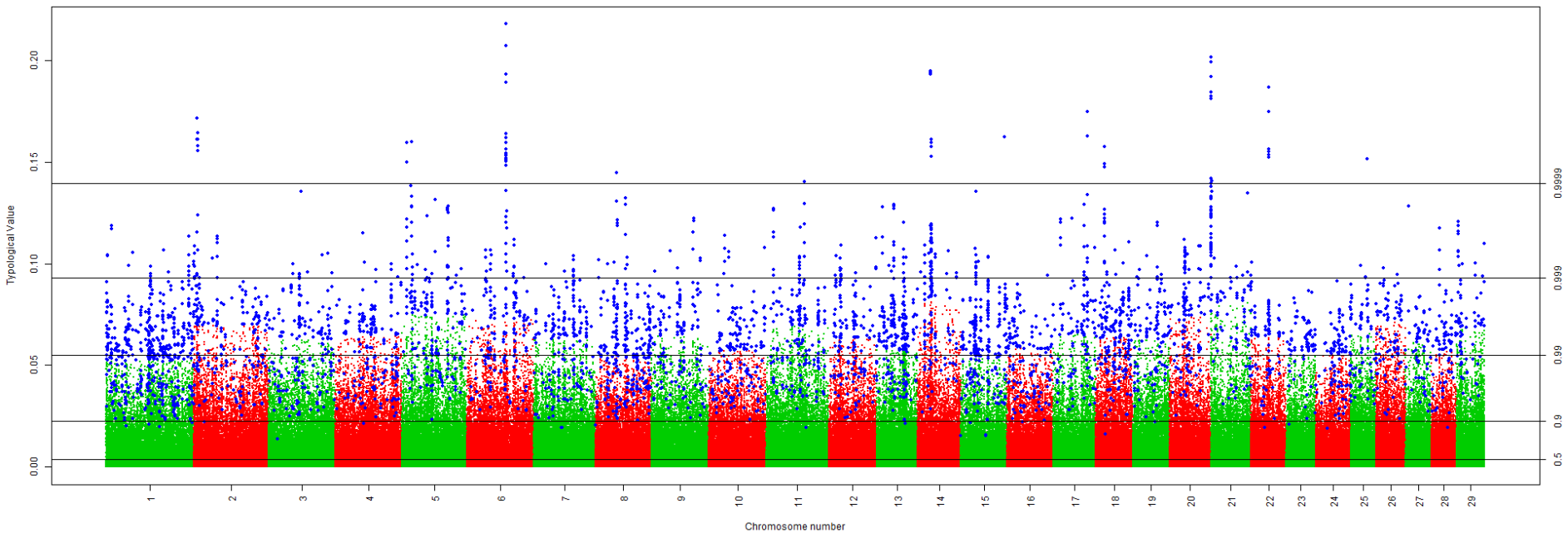
Results

- 10675 selected markers (out of 656152): 1.6%
- Accounts for spatial structure, selected regions containing from 1 to 76 SNPs ($\mu = 2.5$)



Magnitude of selected SNPs according to the block length

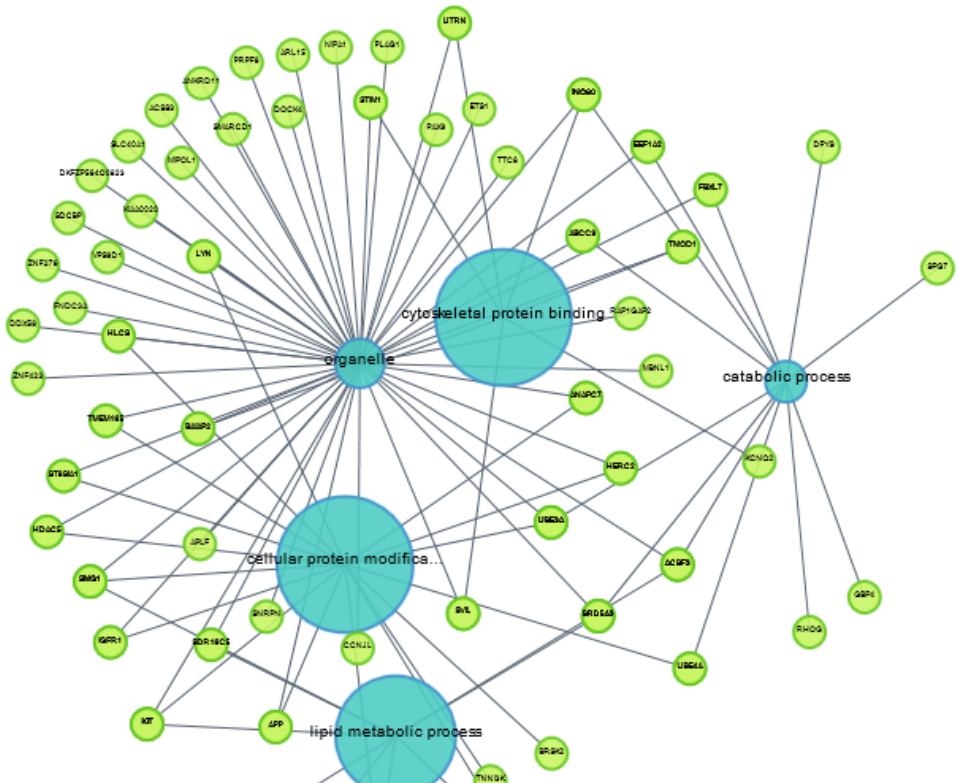
Results



Manhattan plot of typological values. Selected SNPs

Results

- 1067 genes containing selected markers
- Enrichment analysis on the first 100 genes
  - Cytoskeletal protein binding (Muscle)
  - Lipid metabolism process (Milk/Muscle)



Conclusions and Future Work

- Selection of 1.6 % markers
- Accounts for spatial structure, selected regions containing from 1 to 76 SNPs ($\mu = 2.5$)
- Sensitivity to penalization criterion
- Sensitivity to parameter tuning

- Stability selection strategy (Meinshausen and Bühlmann, 2010 ; Yang et al, 2011)
  - Reducing false positives
  - Reducing the effect of parameter tuning

Acknowledgements

Bovine data provided by

- ANR project GEMBAL

Funded by

- ANR project EpiGrani
- Métaprogramme INRA-ACCAF project GALIMED

References

- Dray, S. and Dufour, A-B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22(4).
- Ewing, G. and Hermisson, J. (2010). A coalescent simulation program including recombination, demographic structure and selection at a single locus.*Bioinformatics*, 26(16): 2064-2065.
- Hoefling, H.(2010) A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics* 19(4):984-1006.
- Laloë, D. and Gautier, M. (2011). On the genetic interpretation of Between-Group PCA on SNP data. http://hal.archives-ouvertes.fr/hal-00661214.
- Lebarbier, E.(2005) Detecting multiple change-points in the mean of gaussian process by model selection.*Signal Processing*, 85:717-736
- Meinshausen N. and Bühlmann, P. (2010). Stability selection.*Journal of the Royal Statistical Society Series B*, 72(4) pages 417-473
- Rinaldo, A. (2009). Properties and refinement of the fused lasso. *Ann. Stat., 37(5B):2922-2952*
- Tibshirani, R. *et al.* (2005) Sparsity ans smoothness via the fused lasso.*Journal of the Royal Statistical Society Series B*, pages 91-108.
- Thioulouse, J. (2011) Simultaneous Analysis of a Sequence of Paired Ecological Tables: A Comparison of Several Methods. *Annals of Applied Statistics*,5(4):2300-2325
- Yang, C. *et al.* (2011) Identifying disease-associated SNP clusters via contiguous outlier detection.*Bioinformatics*, 27(18):2578-2585.