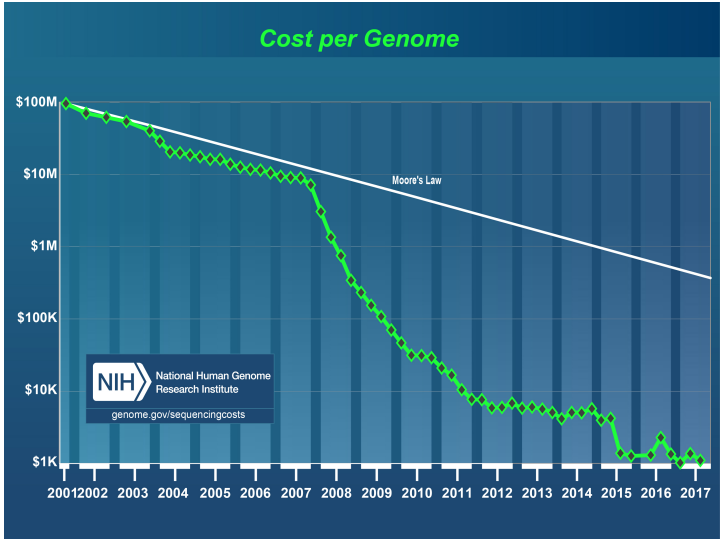
Factorization Matrix Methods For Genome-wide Association Testing

Kevin Caye[1], Olivier Michel[2], Olivier Francois[1]

[1] TIMC-IMAG, [2] GIPSA-lab

Big data in genomic



Genomic data

- The DNA is long sequence of nucleotide : A, C, T, G (3 billion for Human being)
- Data are single-nucleotide polymorphism called SNPs (10 millions for humain species)
- There are 2 alleles for each locus.

$$\text{ADNs} \begin{cases} \cdots & G & A & T & C & C & \cdots \\ \cdots & G & A & A & C & C & \cdots \\ \cdots & G & A & A & C & C & \cdots \\ \cdots & G & A & T & C & C & \cdots \\ \cdots & G & A & T & C & C & \cdots \end{cases}$$

Figure – **SNPs illustration** The nucleotyde differing between the DNA sequences is a SNP.

Genomic matrix

Data are put into a matrix of size $n \times p$:

$$\mathbf{Y} = \begin{bmatrix} 0 & 1 & 2 & 2 & \cdots & \cdots & \cdots \\ 1 & 1 & 0 & 1 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \cdots \\ 0 & 0 & 2 & 0 & \cdots & \cdots & \cdots \end{bmatrix}$$

Figure – **Genomic matrix illustration.** Each entry of the matrix is number of time a mutant allele is observed for a given idividual and locus.

Association testing

Goal

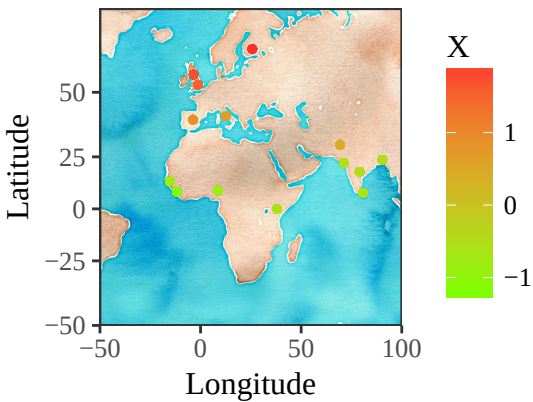
We want to detect SNPs correlated to a variable of interest.

$$\begin{bmatrix} 0 & 1 & 2 & 2 & \cdots & \cdots & 0 & \cdots \\ 1 & 1 & 0 & & \cdots & \cdots & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \vdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \vdots & \cdots \\ 0 & 0 & 2 & 0 & \cdots & \cdots & 1 & \cdots \end{bmatrix} \sim \begin{bmatrix} 0.2 \\ 1.5 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}$$

Example of variable of interest

- Disease : diabetes, celiac disease (dubois 2010)
- Phenotype : the size (wood et al. 2010)
- Environment : the temperature (Frichot et al. 2013)

Genome wide association study with environmental variable



Genome data come from the 1000Genome project (1000Genome Consortium 2015)

- 1409 individuals from 14 populations
- 5397214 SNPs

Variable of interest

- climatic data from WordClim DataBase
- first principal component

We want to identify SNPs associated with the climate

Linear regression model

Linear model for each SNP $\mathbf{Y}_j$

$$\mathbf{Y}_j = \mathbf{X}\beta_j + \mathbf{E_j},$$

where

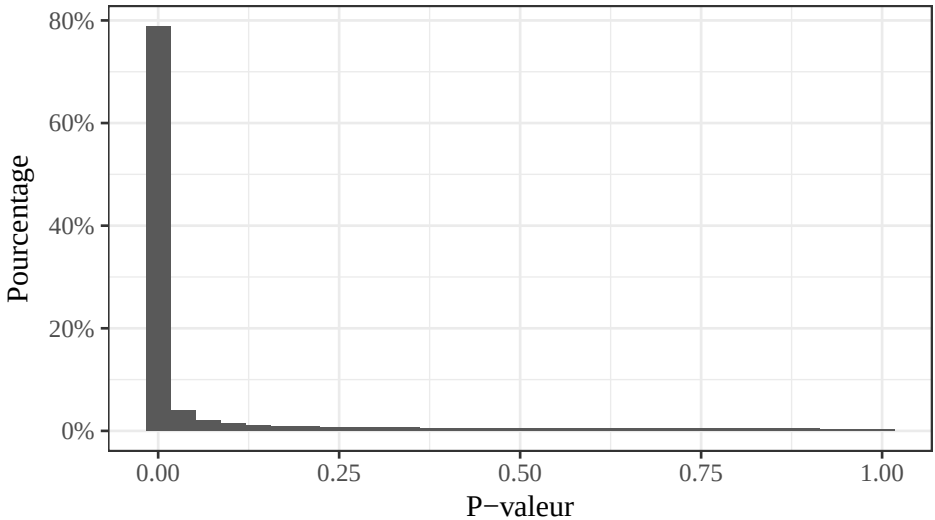
- β_j stand for the effect of the variable of interest $\mathbf{X}$ on the variable $\mathbf{Y}_j$
- $\mathbf{E}_j$ is the residual noise

We want to detect locus for which we can reject the null hypothesis

$$H_0 : \beta_j = 0$$

We use the Student's t-test

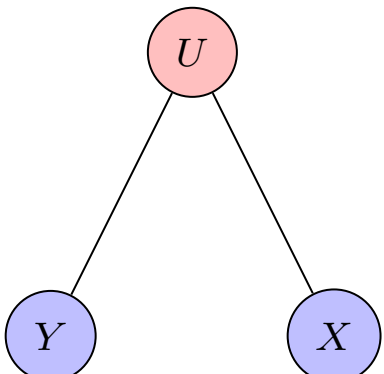
Histogram of *p*-valeurs genome-environment association testing



Section 2

Confounders correction in association testing

Latent factor mixed model (LFMM)



$$\mathbf{Y} = \mathbf{X}\mathbf{B}^T + \mathbf{U}\mathbf{V}^T + \mathbf{E}$$

where

- $\mathbf{U}$ latent factor matrix of size $n \times K$
- $\mathbf{V}$ latent factor effect matrix $p \times K$
- $\mathbf{B}$ is the effect of the variable of interest $\mathbf{X}$ on $\mathbf{Y}$ of size $p \times 1$
- $\mathbf{E}$ is the residual error matrix of size $n \times p$

Estimation method for regression model with latent factors

Méthode	Modèle	Algorithme	Référence
sva-twostep	ACP et régression linéaire	moindres carrés ordinaire et SVD	Leek and Storey 2007
sva-irw	*weighted*-ACP et régression linéaire	moindres carrés ordinaire et *weighted*-SVD	Leek and Storey 2008
cate	analyse factorielle et régression linéaire	EM ou SVD et estimation des moindres carrés généralisée	Wang et al. 2018
ridgeLFMM	factorisation matricielle avec régularisation L_2	SVD et estimation des moindres carrés régularisée en norme L_2	
lassoLFMM	factorisation matricielle avec régularisation L_1	*soft-thresholded* SVD et estimation des moindres carrés régularisée en norme L_1	

L2 regularized least-squares estimates

Loss function

$$\mathcal{L}_{\text{ridge}}(\mathbf{U}, \mathbf{V}, \mathbf{B}) = \frac{1}{2} \left\| \mathbf{Y} - \mathbf{U}\mathbf{V}^T - \mathbf{X}\mathbf{B}^T \right\|_F^2 + \frac{\lambda}{2} \|\mathbf{B}\|_2^2$$

Estimates

1. Compute

$$\hat{\mathbf{U}}\hat{\mathbf{V}}^T = \sqrt{\mathbf{P}_\lambda}^{-1} \text{svd}_K(\sqrt{\mathbf{P}_\lambda}\mathbf{Y})$$

where

$$\mathbf{P}_\lambda = \mathbf{Id}_n - (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{Id}_n)^{-1}\mathbf{X}^T\mathbf{X}$$

2. Compute

$$\hat{\mathbf{B}}^T = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{Id}_d)^{-1}\mathbf{X}^T(\mathbf{Y} - \hat{\mathbf{U}}\hat{\mathbf{V}}^T),$$

L2 regularized least-squares estimates

If $\lambda \to 0$

- $P_\lambda = Id_n$ - $(X^T X)^{-1} X^T X$
- P_λ is not invisible
- **U** et **V** are computed on the residual of the linear regression of **Y** by **X**

Si $\lambda \to \infty$

- $P_\lambda = Id_n$
- **U** et **V** is given by the SVD of rank K

L2 regularized least-squares estimates

Theorem 1

Let $\lambda > 0$. The estimates $\hat{\mathbf{U}}$, $\hat{\mathbf{V}}$ and define a global mimimum of the penalized loss function $\mathcal{L}_{\text{ridge}}$.

Idea of the proof

$$\mathcal{L}_{\text{ridge}}(\mathbf{U}, \mathbf{V}, \mathbf{B}) \geq \mathcal{L}_{\text{ridge}}(\mathbf{U}, \mathbf{V}, (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{Id}_d)^{-1}\mathbf{X}^T(\mathbf{Y} - \mathbf{U}\mathbf{V}^T)) = \frac{1}{2}\left\|\sqrt{\mathbf{P}_\lambda}(\mathbf{Y} - \mathbf{U}\mathbf{V}^T)\right\|_F^2$$

L1 regularized least-squares estimates

Loss function

$$\mathcal{L}_{\text{lasso}}(\mathbf{W}, \mathbf{B}) = \frac{1}{2} \left\| \mathbf{Y} - \mathbf{W} - \mathbf{X}\mathbf{B}^T \right\|_F^2 + \mu \|\mathbf{B}\|_1 + \gamma \|\mathbf{W}\|_*$$

where

- $\mathbf{W}$ is the latent matrix such that
$$\mathbf{W} = \mathbf{U}\mathbf{V}^T$$
- $\|\mathbf{W}\|_*$ is the nuclear norm equal to the sum of the matrix $\mathbf{W}$ eigen values.

L1 regularized least-squares estimates

block-coordinate descent algorithm

Initialize

$$\hat{\mathbf{W}}_{t=0} = 0$$
$$\hat{\mathbf{B}}_{t=0} = 0$$

Then alternate

1. Compute $\hat{\mathbf{B}}_t$ the optimum of

$$\mathcal{L}'_{\text{lasso}}(\mathbf{B}) = \frac{1}{2}||(\mathbf{Y} - \hat{\mathbf{W}}_{t-1}) - \mathbf{X}\mathbf{B}^T||_F^2 + \mu||\mathbf{B}||_1 \quad (1)$$

2. Compute $\hat{\mathbf{W}}_t$ the optimum of

$$\mathcal{L}''_{\text{lasso}}(\mathbf{W}) = \frac{1}{2}||(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_t^T) - \mathbf{W}||_F^2 + \gamma||\mathbf{W}||_* \quad (2)$$

L1 regularized least-squares estimates

Theorem 1

The block-coordinate descent algorithm for estimating the L1 regularized parameters converge toward a global minimum of the loss function $\mathcal{L}_{\text{lasso}}$.

The proof rely on a work of Tseng et al. (2001)

The main hypothesis are :

- $\mathbf{W}, \mathbf{B} \mapsto \left\|\mathbf{Y} - \mathbf{W} - \mathbf{X}\mathbf{B}^T\right\|_F^2$ is convex and differentiable (the term of attach to data in $\mathcal{L}_{\text{lasso}}$)
- $\mathbf{B} \mapsto \|\mathbf{B}\|_1$ is continuous and convex (regularization term in $\mathcal{L}_{\text{lasso}}$)
- $\mathbf{W} \mapsto \|\mathbf{W}\|_*$ is continuous and convex (regularization term in $\mathcal{L}_{\text{lasso}}$)

hypothesis testing corrected for confunders (Price et al. 2006)

For each explained variable $\mathbf{Y}_j$

$$\mathbf{Y}_j = \hat{\mathbf{U}}\gamma_j^T + \mathbf{X}\beta_j + \mathbf{E_j},$$

where $\hat{\mathbf{U}}$ is an estimates of the latent variable matrix.
We test the following hypothesis

$$H_0 : \beta_j = 0$$

We want to the list $\Gamma = \{1, .., J\}$ such that

$$p(\beta_j = 0 | j \in \Gamma) = T$$

where T is the expected false discovery rate (FDR).
We used the q-valeur (Storey et al. 2003)

Section 3

Methods Comparison On Simulations

Methods comparison on dataset simulated from 1000Genomes dataset

Simulated Dataset

- Compute the K first principal components

$$\mathbf{Y} = \mathbf{U}\mathbf{V}^T + \mathbf{E}$$

- Simulate $\mathbf{U}'$ and $\mathbf{X}'$ by controlling the correlation.
- Then create a new matrix

$$\mathbf{Y}' = \mathbf{U}'\mathbf{V}^T + \mathbf{X}'\mathbf{B}'^T + \mathbf{E}$$

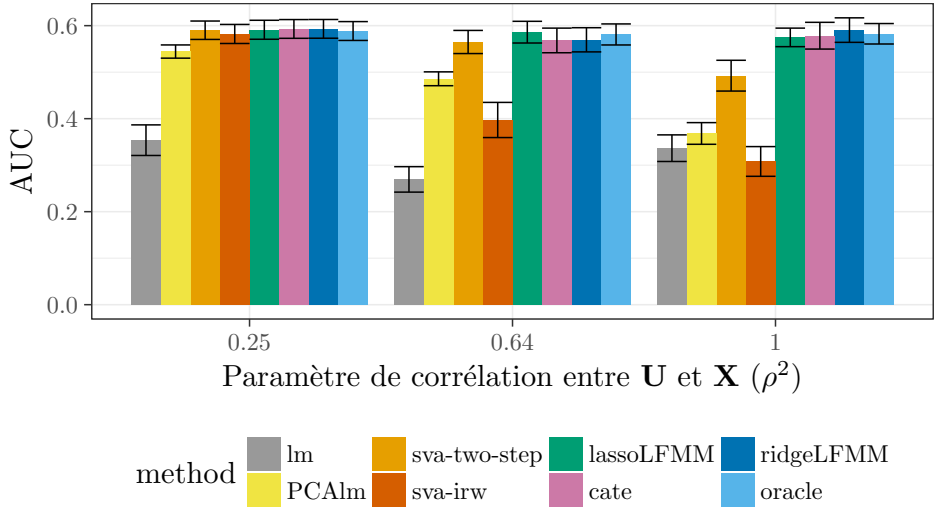
We compared the following method

- lm
- lmPCA
- sva-twostep
- sva-irw
- cate
- oracle
- ridgeLFMM
- lassoLFMM

Criteria

- AUC : Area under the curve (1 - FDR) $\times$ recall (or power)

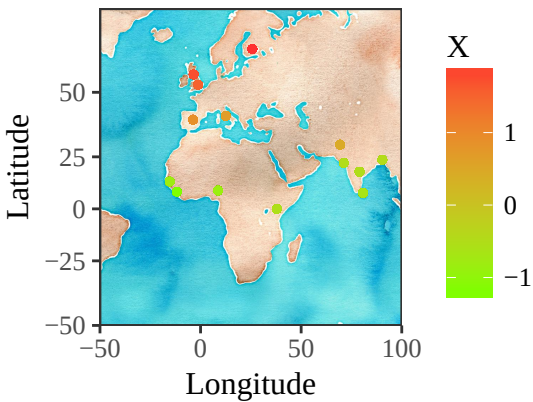
Result of methods comparison on simulated dataset



Section 4

Methods Comparison on True Dataset

Gene-environment association study (GEA)



Genome data come from the 1000Genome project (1000Genome Consortium 2015)

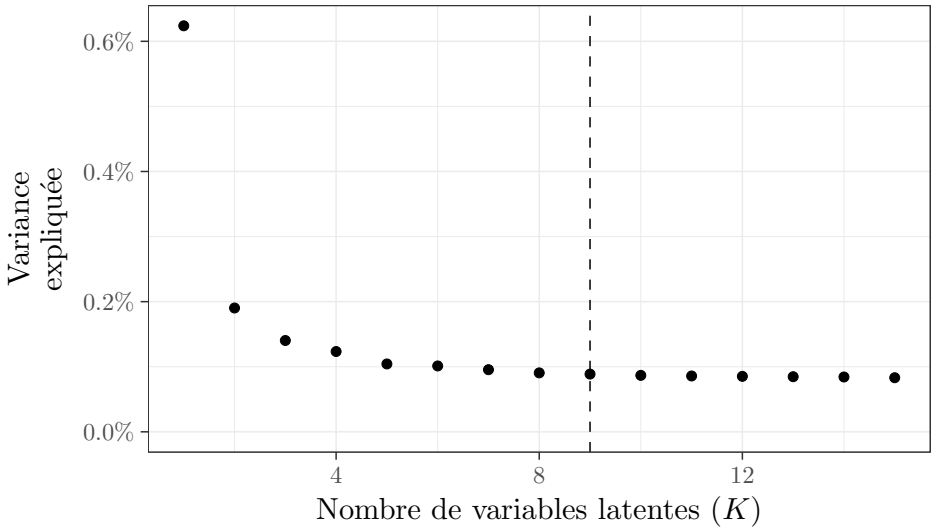
- 1409 individuals from 14 populations
- 5397214 SNPs

Variable of interest

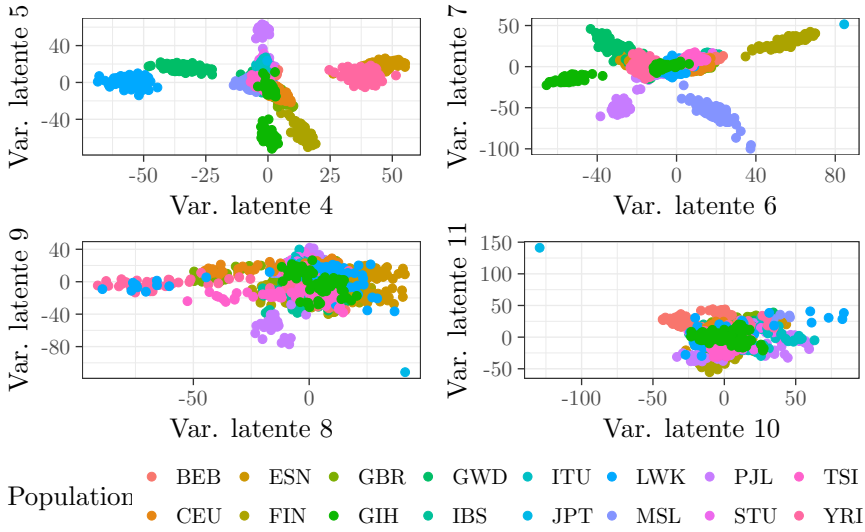
- climatic data from WordClim DataBase
- first principal component

We want to identify SNPs associated with the climate

Choice of K for the gene-environment association study



Choice of K for the gene-environment association study



Results of the gene-environment association study

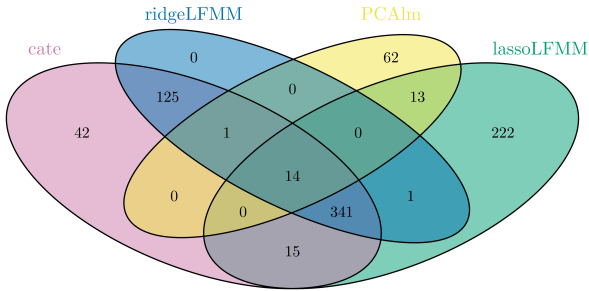
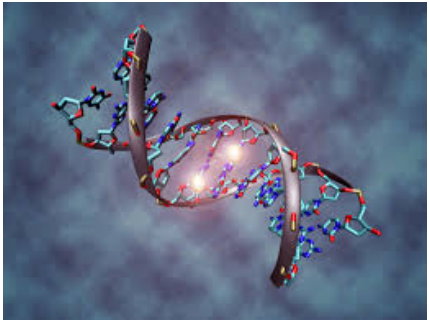


Figure – Venn diagram for an expected false discovery rate of 1%.

Results of the gene-environment association study

SNPs	Détecté par les méthodes	Description du phénotype
rs10908907	ridgeLFMM, cate	Alcoholism (heaviness of drinking)
rs10496731	lassoLFMM	Body Height
rs2472297	ridgeLFMM, cate, lassoLFMM	Caffeine metabolism
rs2256175	ridgeLFMM, cate, lassoLFMM	Cholesterol total
rs2472297	ridgeLFMM, cate, lassoLFMM	Coffee consumption (cups per day)
rs2278544, rs2322659	lassoLFMM	Congenital lactase deficiency
rs4954218	ridgeLFMM, cate, lassoLFMM	Corneal structure
rs882300	ridgeLFMM, cate, lassoLFMM	Electrocardiographic traits
rs882300	ridgeLFMM, cate, lassoLFMM	Electrocardiography
rs2256175	ridgeLFMM, cate, lassoLFMM	Giant cell arteritis
rs2256175, rs6085576, rs2104012, rs1983716, rs2853977	ridgeLFMM, cate, lassoLFMM	Height
rs6430549	ridgeLFMM, cate, lassoLFMM	Hematocrit
rs2278544, rs2322659	lassoLFMM	Lactose intolerance
rs882300	ridgeLFMM, cate, lassoLFMM	Multiple sclerosis
rs1123848	ridgeLFMM, cate, lassoLFMM	Neuroblastoma
rs17158483	lassoLFMM	Obesity-related traits

Association study between DNA methylation level and the rheumatoid arthritis (EWAS)



(Wikipedia)

Dataset (Liu et al. 2013)

- **Y** contains methylation level for 485577 DNA location for 699 individuals (354 case and 335 control)
- **X** stand for the rheumatoid arthritis

Confounding factors

- celular composition
- age
- gender
- tabacco consummation

Goal

Find the methylation sites associated with the rheumatoid arthritis

Results of the EWAS

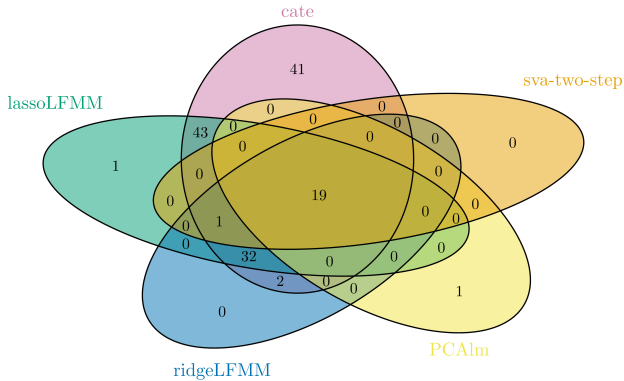


Figure – Venn diagram for an expected false discovery rate of 1 %.

Methylation sites found out in other studies (Rahmani et al. 2016, Zou et al. 2014) (EWAS)

ID	Chr	Position	Gene	PCAlm	lassoLFMM	cate	ridgeLFMM
cg16411857	**16**	**57023191**	**NLRC5**	**9.2e-13**	**2.4e-12**	**6.6e-12**	**5.3e-12**
cg07839457	**16**	**57023022**	**NLRC5**	**1.9e-11**	**4.5e-11**	**1.1e-10**	**9.7e-11**
cg05428452	**6**	**32712979**	**HLA-DQA2**	**5.4e-11**	**4.6e-11**	**8.5e-11**	**8.8e-11**
cg02508743	8	56903623	LYN	2.9e-08	2.7e-08	2.7e-08	2.8e-08
cg20821042	**6**	**32709158**	**HLA-DQA2**	**6.5e-08**	**6.1e-08**	**9.6e-08**	**1.0e-07**
cg13081526	6	32449961		1.5e-07	1.2e-07	2.0e-07	2.2e-07
cg18052547	6	32552547	HLA-DRB1	1.8e-07	1.8e-07	3.0e-07	3.1e-07
cg25372449	**6**	**32490350**	**HLA-DRB5**	**2.5e-07**	**2.6e-07**	**4.5e-07**	**4.6e-07**
cg02030958	13	110386267		4.0e-07	7.8e-08	6.0e-08	1.1e-07
cg16171858	3	58472734		4.6e-07	1.6e-07	2.7e-08	3.8e-08
cg03280622	8	145023013	PLEC1	4.7e-07	5.0e-09	5.8e-09	3.8e-08
cg24150157	19	51891210	LIM2	6.2e-07	3.1e-07	1.6e-07	2.1e-07
cg26244575	12	76354015		6.9e-07	2.7e-09	5.0e-10	4.2e-09
cg05370853	6	32606634	HLA-DQA1	7.1e-07	3.0e-07	3.3e-07	4.4e-07
cg14989316	10	80757927	LOC283050	7.3e-07	6.1e-08	7.8e-08	2.1e-07
cg17360552	6	32725332	HLA-DQB2	8.1e-07	6.1e-07	1.1e-06	1.2e-06
cg01373248	3	18480297	SATB1	8.1e-07	1.4e-07	1.1e-07	2.5e-07
cg26164488	2	64440295		9.3e-07	3.5e-09	1.6e-09	1.4e-08
cg05874806	2	102350276	MAP4K4	1.1e-06	1.1e-06	4.7e-07	5.6e-07

Section 5

Conclusions

lfmm

- Two new methods to estimate the confunction factor for correcting association studies
- Theorical convergence of the algorithms
- Same power that oracle on simulation
- On true dataset methods based on the latent factor mixed model discover more associations
- On true dataset association discovered can be different between methods
- We deliver a R package `lfmm` which implement these methods

Thank you for your attention !

Choix du nombre de variables latentes

On projette $\mathbf{Y}$ sur l'espace orthogonal à $\mathbf{X}$ en prenant $\lambda = 0$

$$\mathbf{D}_0\mathbf{Q}^T\mathbf{Y} = \mathbf{D}_0\mathbf{Q}^T\mathbf{U}\mathbf{V}^T + \mathbf{D}_0\mathbf{Q}^T\mathbf{E}.$$

On calcule les valeurs singulières pour visualiser la variance expliquée par chaque variable latente (scree plot).