Sparse Multiple Correspondence Analysis for selection of Single Nucleotide Polymorphisms

Anne BERNARD

QFAB Bioinformatics, University of Queensland, Brisbane

December 2^{nd}, 2015

Motivation

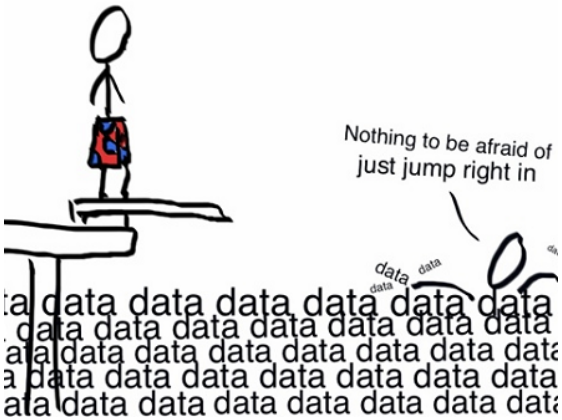
- Genomics becoming common place in many fields
- Require sophisticated multivariate techniques to analyse high dimentional data

$\Rightarrow$ **Variable selection** and **dimensionality reduction** necessary to obtain simpler structures and interpret results

High-dimensional data

Analysis of the data structure and observation of a possible natural separation between individuals depending on their human genetic heritage.

One data set $\mathbf{X}$ $(I \times J) \Rightarrow$ Unsupervised multivariate analysis



Continuous data: PCA
Categorical data: MCA

In case of high dimensional data ($I \gg J$): results difficult to interpret.

Solution: Use/Develop appropriate statistical methods to **select relevant variables** and **facilitate interpretation** of the results.

Case study: Genes potentially involved in skin aging

To date, no Genome Wide Association Study (GWAS) has sought links with skin aging.

$\Rightarrow$ In 2010 establishment of a GWAS research project by the CE.R.I.E.S. (funded by Chanel) to identify genes potentially involved in facial skin aging

Case study: Genes potentially involved in skin aging

Aim

Study the impact of genetic on expression of skin aging

How

By finding potential links between

- **Single Nucleotide Polymorphism** (SNPs) and
- Severity of **signs of skin aging**, given the age of the individuals and any aggravating factors (smoking status, sun exposure...)

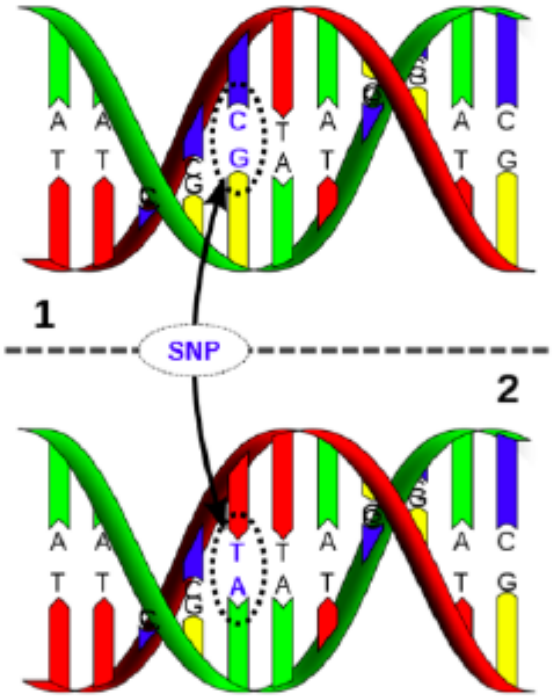
Material

- 502 caucasian women of the SU.VI.MAX cohort living in the Paris area (aged 44-70 years)
- Data as covariates : age, Body Mass Index, smoking habits, hormonal status, lifetime sun exposure
- Digital images of the face taken for each participants

- Blood sample taken for DNA extraction and genetic analysis with Illumina Omni1 chips (1 million of SNPs)

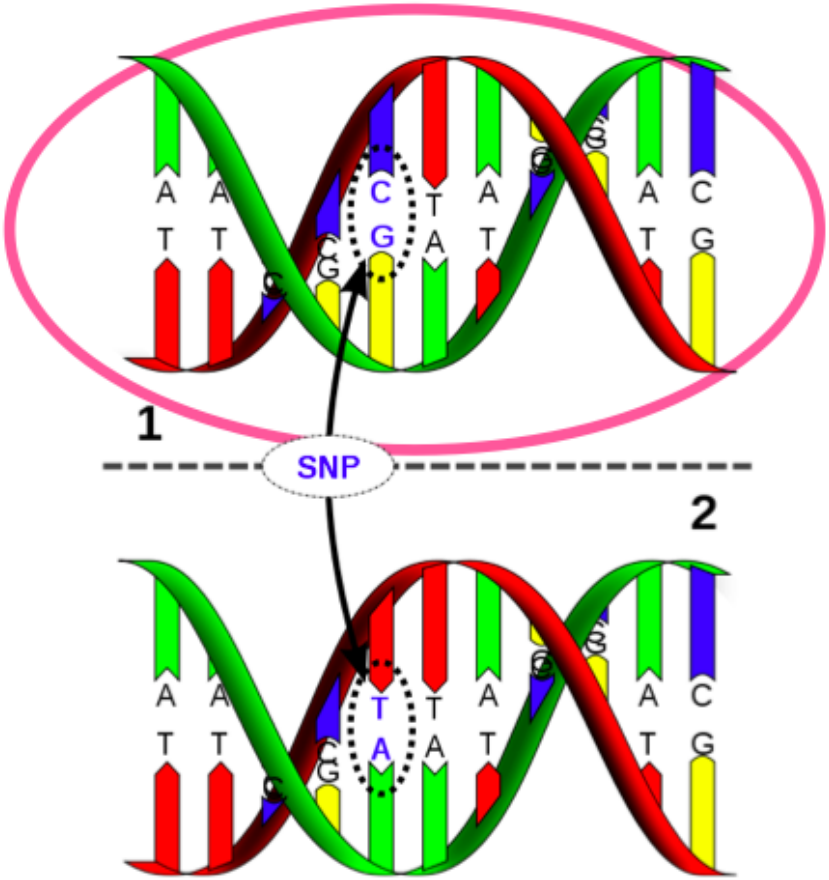
What are SNPs?

- Most common type of genetic variation
- Replacement of one nucleotide by another one
  - A, T, G, C



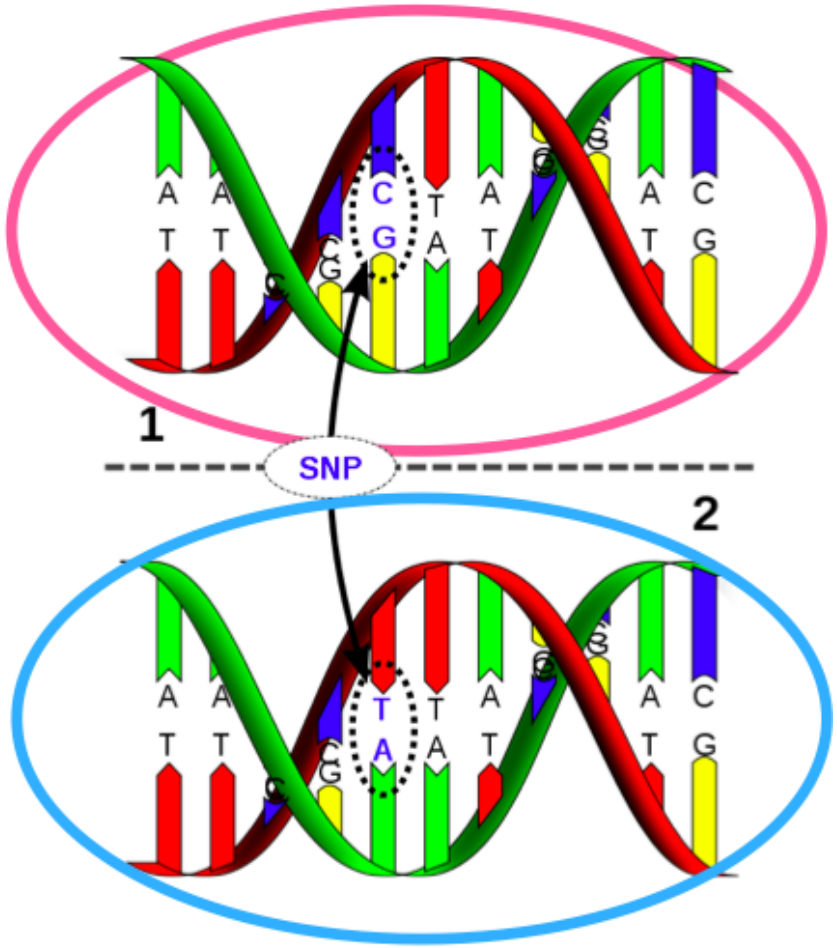
What are SNPs?

Chromosome from Mother



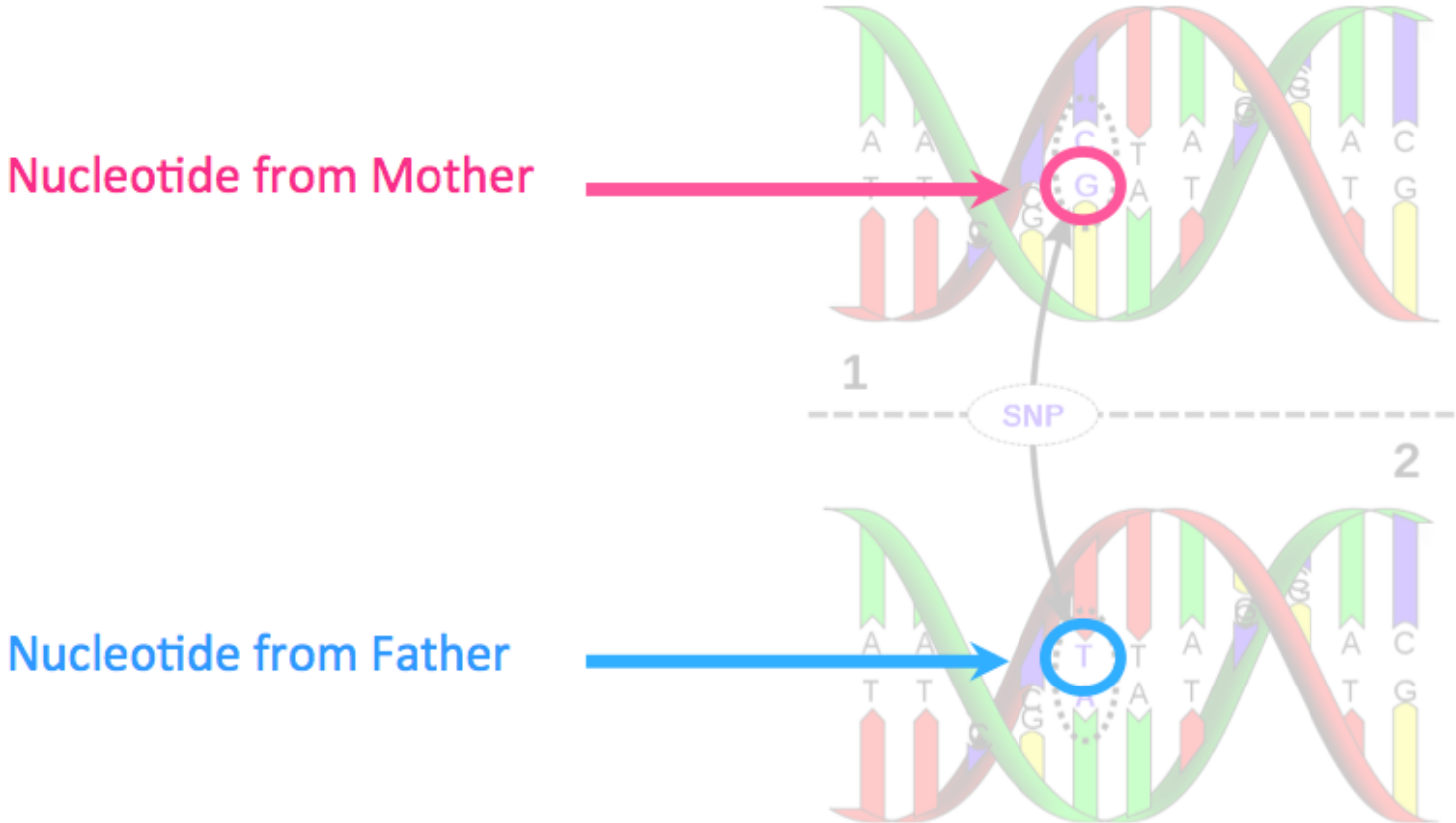
What are SNPs?

Chromosome from Mother

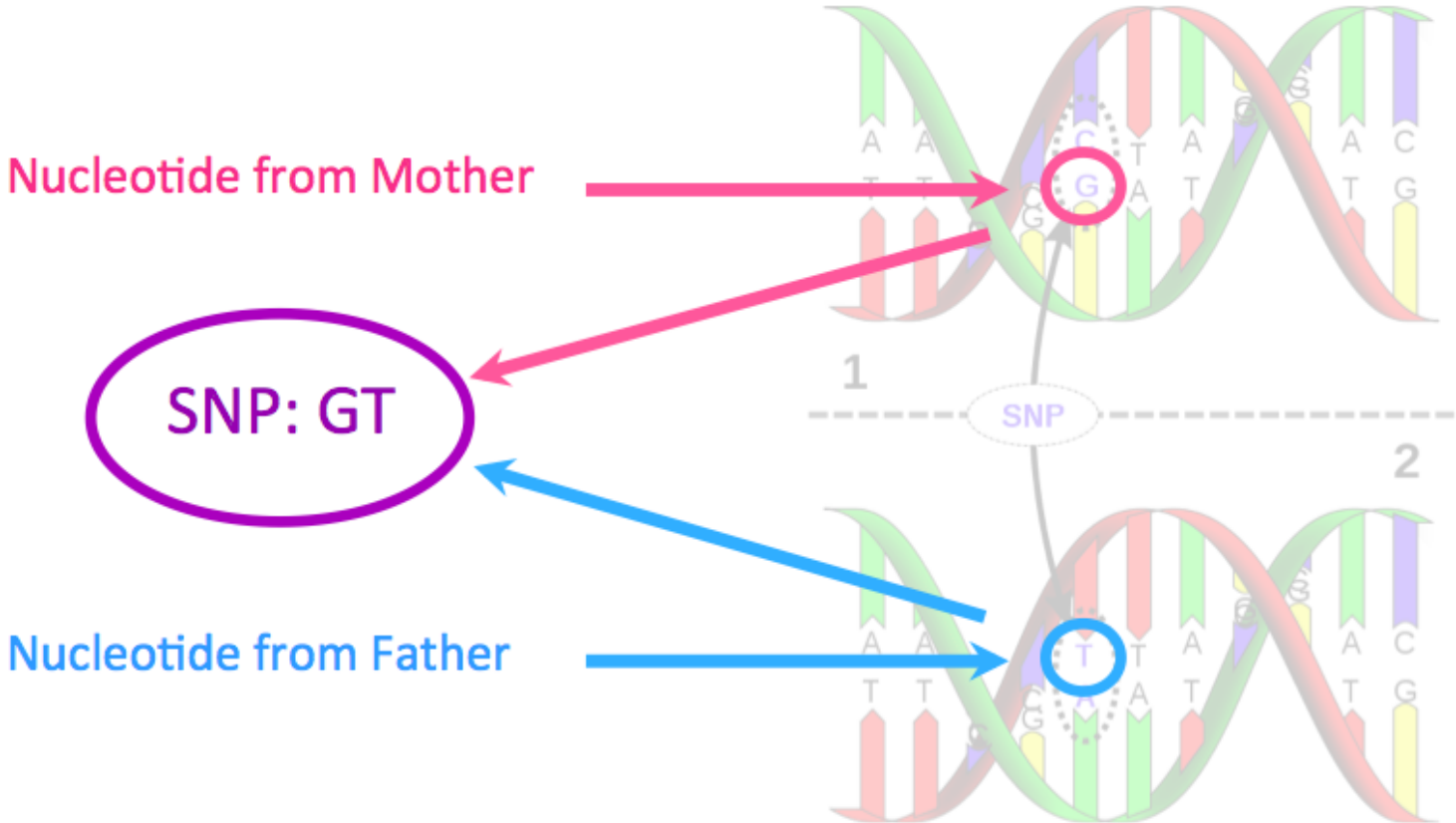


Chromosome from Father

What are SNPs?



What are SNPs?



SNP Coding

ID	SNP.1	SNP.2
AK	GG	AA
JD	GT	CC
MR	GT	AC
GB	GG	CC
NH	TT	AC
AL	GG	CC
DO	TT	AC
JM	TT	CC
ED	TT	AC
CB	TT	CC
CF	GG	AC
OD	TT	CC
DM	TT	AC
NS	GG	CC
JR	GT	AA

Original Coding

SNP Coding

ID	SNP.1	SNP.2
AK	GG	AA
JD	GT	CC
MR	GT	AC
GB	GG	CC
NH	TT	AC
AL	GG	CC
DO	TT	AC
JM	TT	CC
ED	TT	AC
CB	TT	CC
CF	GG	AC
OD	TT	CC
DM	TT	AC
NS	GG	CC
JR	GT	AA

Original Coding

GG	GT	TT	CC	AC	AA	ID
1	0	0	0	0	1	AK
0	1	0	1	0	0	JD
0	1	0	0	1	0	MR
1	0	0	1	0	0	GB
0	0	1	0	1	0	NH
1	0	0	1	0	0	AL
0	0	1	0	1	0	DO
0	0	1	1	0	0	JM
0	0	1	0	1	0	ED
0	0	1	1	0	0	CB
1	0	0	0	1	0	CF
0	0	1	1	0	0	OD
0	0	1	0	1	0	DM
1	0	0	1	0	0	NS
0	1	0	0	0	1	JR

Nominal Coding

SNP Coding

ID	SNP.1	SNP.2
AK	GG	AA
JD	GT	CC
MR	GT	AC
GB	GG	CC
NH	TT	AC
AL	GG	CC
DO	TT	AC
JM	TT	CC
ED	TT	AC
CB	TT	CC
CF	GG	AC
OD	TT	CC
DM	TT	AC
NS	GG	CC
JR	GT	AA

Original Coding

GG	GT	TT	CC	AC	AA	ID
1	0	0	0	0	1	AK
0	1	0	1	0	0	JD
0	1	0	0	1	0	MR
1	0	0	1	0	0	GB
0	0	1	0	1	0	NH
1	0	0	1	0	0	AL
0	0	1	0	1	0	DO
0	0	1	1	0	0	JM
0	0	1	0	1	0	ED
0	0	1	1	0	0	CB
1	0	0	0	1	0	CF
0	0	1	1	0	0	OD
0	0	1	0	1	0	DM
1	0	0	1	0	0	NS
0	1	0	0	0	1	JR

SNP1 SNP2

SNP Coding

ID	SNP.1	SNP.2
AK	GG	AA
JD	GT	CC
MR	GT	AC
GB	GG	CC
NH	TT	AC
AL	GG	CC
DO	TT	AC
JM	TT	CC
ED	TT	AC
CB	TT	CC
CF	GG	AC
OD	TT	CC
DM	TT	AC
NS	GG	CC
JR	GT	AA

Original Coding

GG	GT	TT	CC	AC	AA	ID
1	0	0	0	0	1	AK
0	1	0	1	0	0	JD
0	1	0	0	1	0	MR
1	0	0	1	0	0	GB
0	0	1	0	1	0	NH
1	0	0	1	0	0	AL
0	0	1	0	1	0	DO
0	0	1	1	0	0	JM
0	0	1	0	1	0	ED
0	0	1	1	0	0	CB
1	0	0	0	1	0	CF
0	0	1	1	0	0	OD
0	0	1	0	1	0	DM
1	0	0	1	0	0	NS
0	1	0	0	0	1	JR

Nominal Coding

SNP Coding

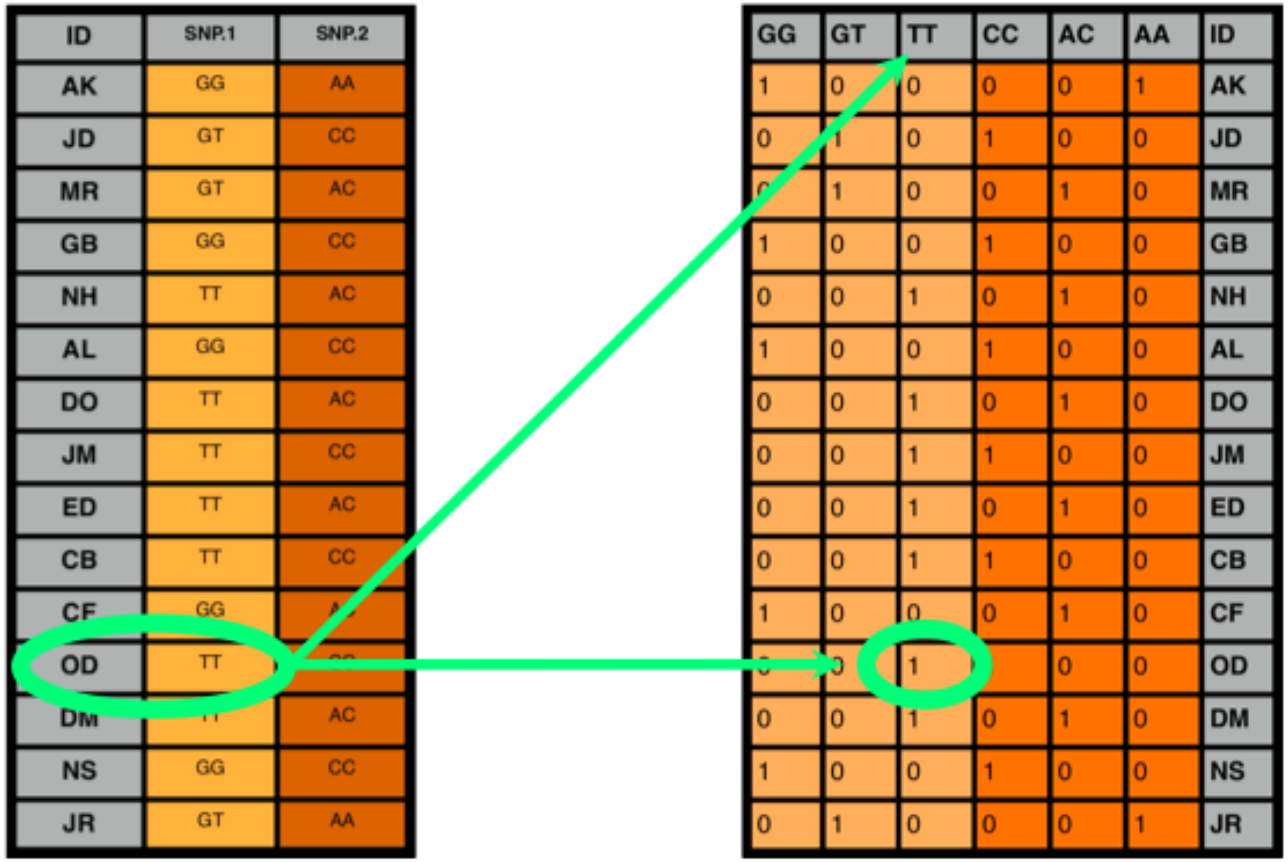
ID	SNP.1	SNP.2
AK	GG	AA
JD	GT	CC
MR	GT	AC
GB	GG	CC
NH	TT	AC
AL	GG	CC
DO	TT	AC
JM	TT	CC
ED	TT	AC
CB	TT	CC
CF	GG	
OD	TT	CC
DM		AC
NS	GG	CC
JR	GT	AA

Original Coding

GG	GT	TT	CC	AC	AA	ID
1	0	0	0	0	1	AK
0	1	0	1	0	0	JD
0	1	0	0	1	0	MR
1	0	0	1	0	0	GB
0	0	1	0	1	0	NH
1	0	0	1	0	0	AL
0	0	1	0	1	0	DO
0	0	1	1	0	0	JM
0	0	1	0	1	0	ED
0	0	1	1	0	0	CB
1	0	0	0	1	0	CF
0	0	1	1	0	0	OD
0	0	1	0	1	0	DM
1	0	0	1	0	0	NS
0	1	0	0	0	1	JR

Nominal Coding

SNP Coding



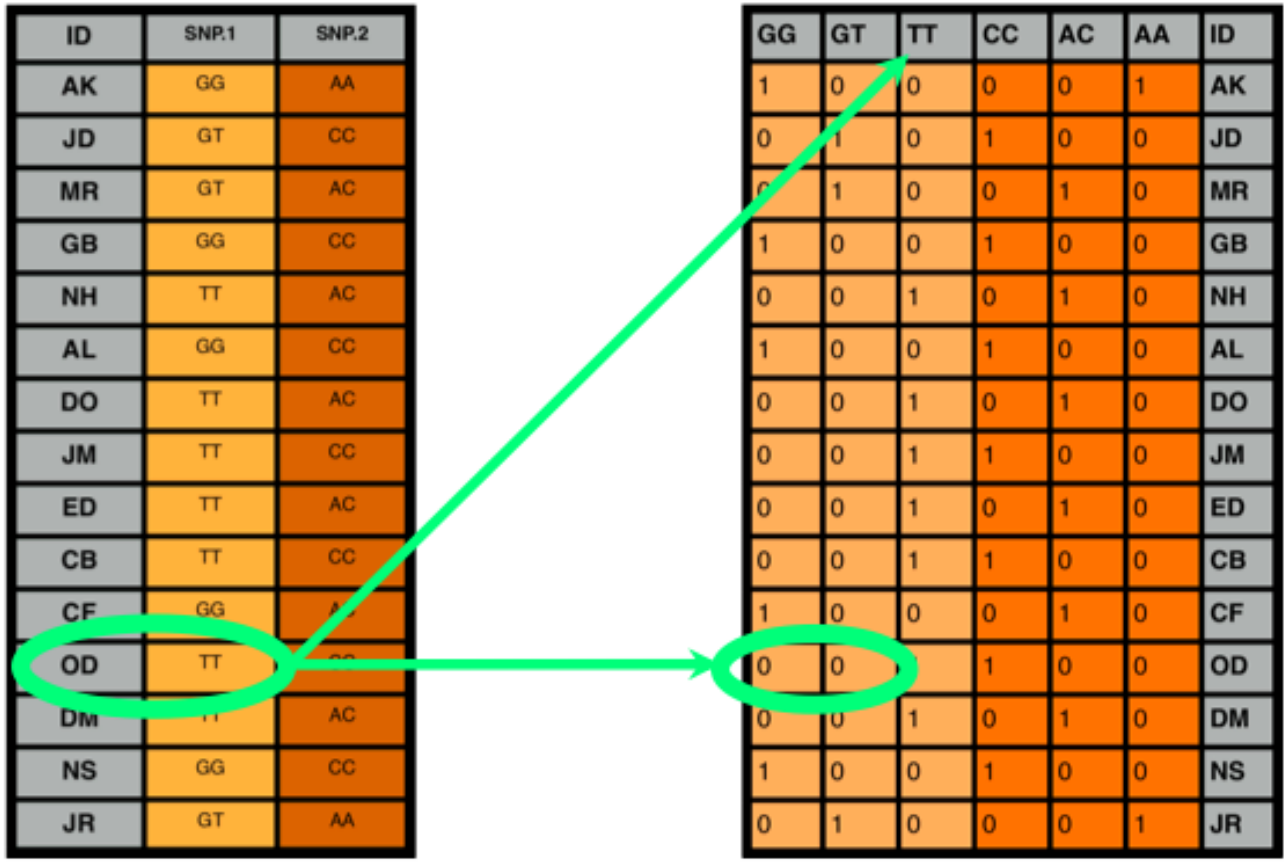
ID	SNP.1	SNP.2
AK	GG	AA
JD	GT	CC
MR	GT	AC
GB	GG	CC
NH	TT	AC
AL	GG	CC
DO	TT	AC
JM	TT	CC
ED	TT	AC
CB	TT	CC
CF	GG	
OD	TT	
DM		AC
NS	GG	CC
JR	GT	AA

GG	GT	TT	CC	AC	AA	ID
1	0	0	0	0	1	AK
0	1	0	1	0	0	JD
0	1	0	0	1	0	MR
1	0	0	1	0	0	GB
0	0	1	0	1	0	NH
1	0	0	1	0	0	AL
0	0	1	0	1	0	DO
0	0	1	1	0	0	JM
0	0	1	0	1	0	ED
0	0	1	1	0	0	CB
1	0	0	0	1	0	CF
0	0	1		0	0	OD
0	0	1	0	1	0	DM
1	0	0	1	0	0	NS
0	1	0	0	0	1	JR

Original Coding

Nominal Coding

SNP Coding



ID	SNP.1	SNP.2
AK	GG	AA
JD	GT	CC
MR	GT	AC
GB	GG	CC
NH	TT	AC
AL	GG	CC
DO	TT	AC
JM	TT	CC
ED	TT	AC
CB	TT	CC
CF	GG	
OD	TT	
DM		AC
NS	GG	CC
JR	GT	AA

GG	GT	TT	CC	AC	AA	ID
1	0	0	0	0	1	AK
0	1	0	1	0	0	JD
0	1	0	0	1	0	MR
1	0	0	1	0	0	GB
0	0	1	0	1	0	NH
1	0	0	1	0	0	AL
0	0	1	0	1	0	DO
0	0	1	1	0	0	JM
0	0	1	0	1	0	ED
0	0	1	1	0	0	CB
1	0	0	0	1	0	CF
0	0		1	0	0	OD
0	0	1	0	1	0	DM
1	0	0	1	0	0	NS
0	1	0	0	0	1	JR

Original Coding

Nominal Coding

Case study: Material

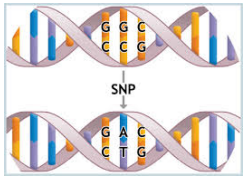
Variables to be explained

Phenotypes 

- Sagging score [0-10]

and others (global photoageing, wrinkling score, lentigines score)

Explained variables

Genetic data 

795 063 SNPs analysed

Targeted set of 537 SNPs

- "Candidate GWAS"
- 1611 disjunctive columns
- 537 blocks

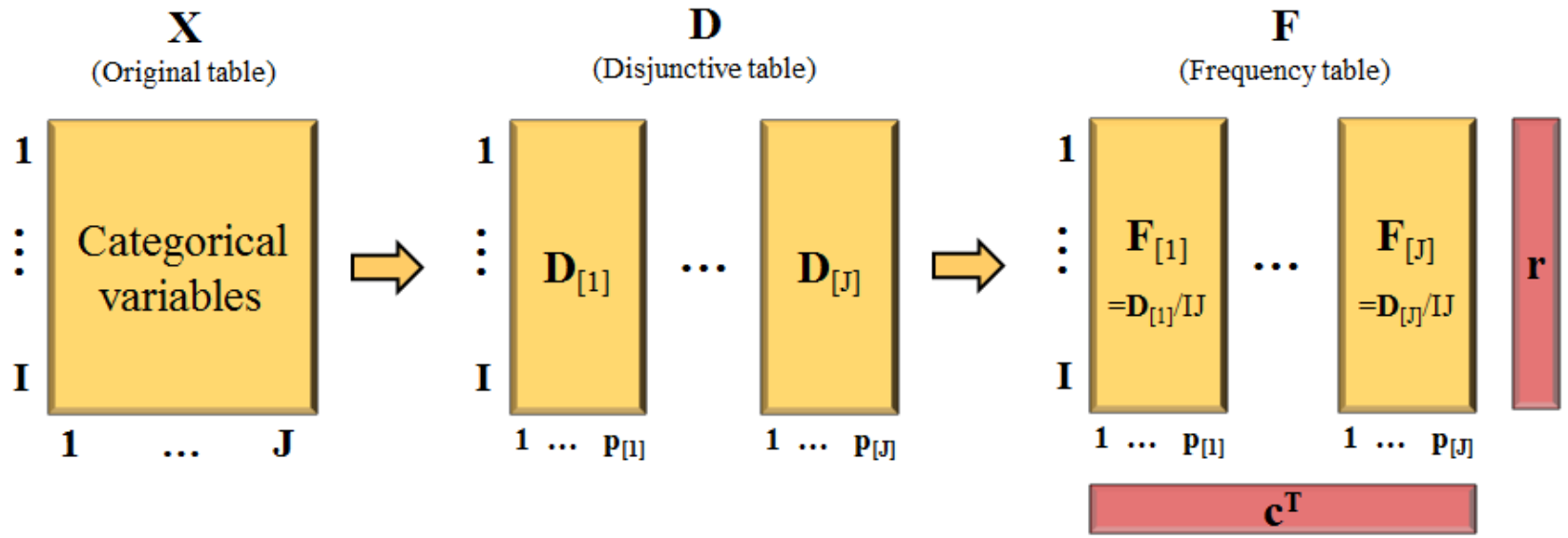
Application on SNPs data: Exploratory analysis

Step 1 Visualization of links between SNPs and between samples using **Multiple Correspondence Analysis** (MCA)

Step 2 Selection of the most important SNPs for a component using the sparse extension of MCA to select variables: **Sparse MCA**

Multivariate Exploratory Methods

When **X** matrix of categorical variables



$\mathbf{r} = \mathbf{F1}$

$\mathbf{c} = \mathbf{F}^T\mathbf{1}$

$\mathbf{p}_{[j]}$ =Nb of modalities of variable j

Multiple Correspondence Analysis

MCA via Generalized SVD of F

$$\mathbf{F} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T \qquad \text{with } \mathbf{P}^T\mathbf{M}\mathbf{P} = \mathbf{Q}^T\mathbf{W}\mathbf{Q} = \mathbf{I}$$

where $\mathbf{F} = [\mathbf{F}_{[1]}|\ldots|\mathbf{F}_{[j]}|\ldots|\mathbf{F}_{[J]}]$ and $\mathbf{Q} = [\mathbf{Q}_{[1]}|\ldots|\mathbf{Q}_{[j]}|\ldots|\mathbf{Q}_{[J]}]$

In the case of PCA: $\mathbf{M} = \mathbf{W} = \mathbf{I}$
In the case of MCA: $\mathbf{M} = \mathbf{D_r}^{-1}$ $\mathbf{W} = \mathbf{D_c}^{-1}$

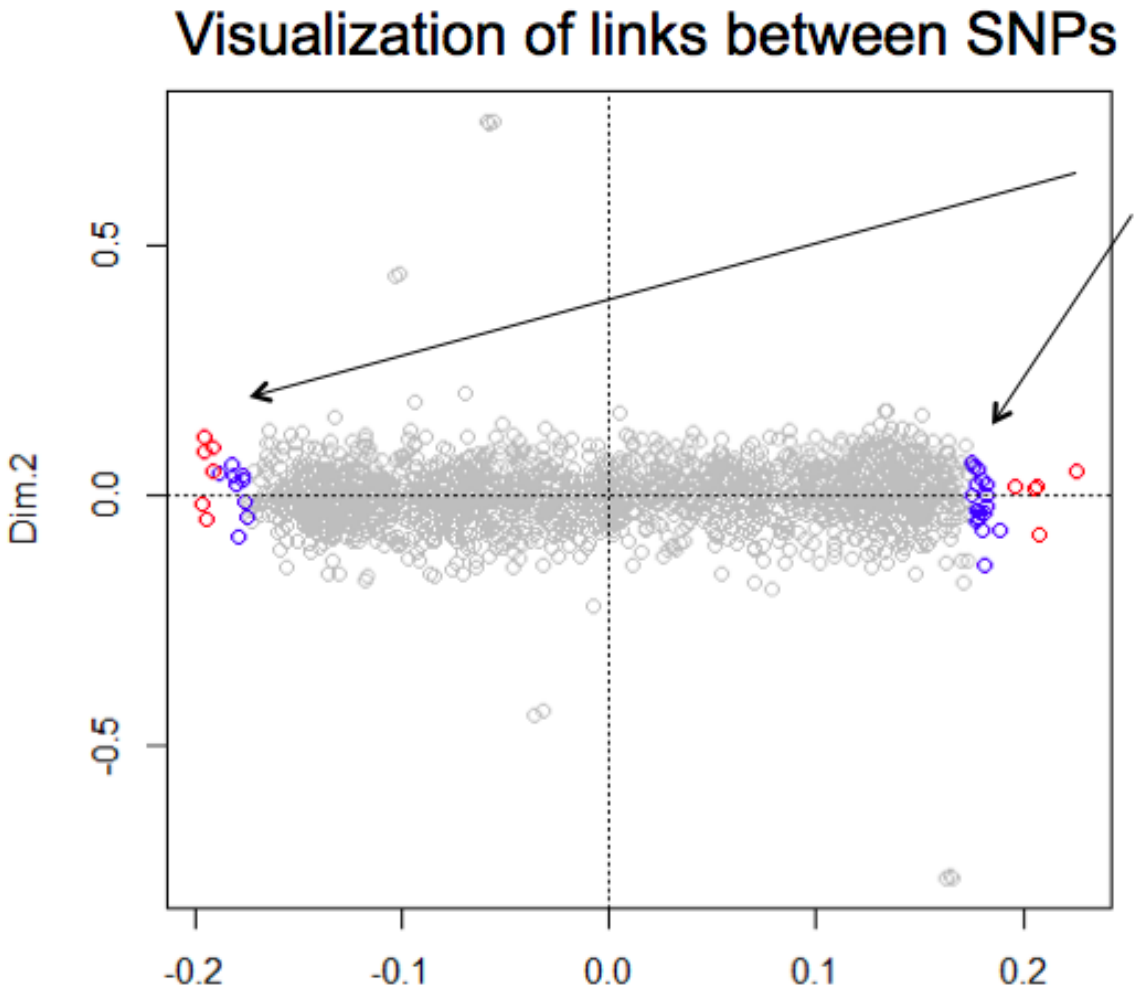
GSVD as low rank approximation of matrices

MCA can be seen as the solution of

$$\min_{\tilde{\mathbf{p}},\tilde{\mathbf{q}}} \|\mathbf{F} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T\|^2_{\mathbf{W}} \qquad \tilde{\mathbf{p}}^T\mathbf{M}\tilde{\mathbf{p}} = \tilde{\mathbf{q}}^t\mathbf{W}\tilde{\mathbf{q}} = \mathbf{1} \qquad (1)$$

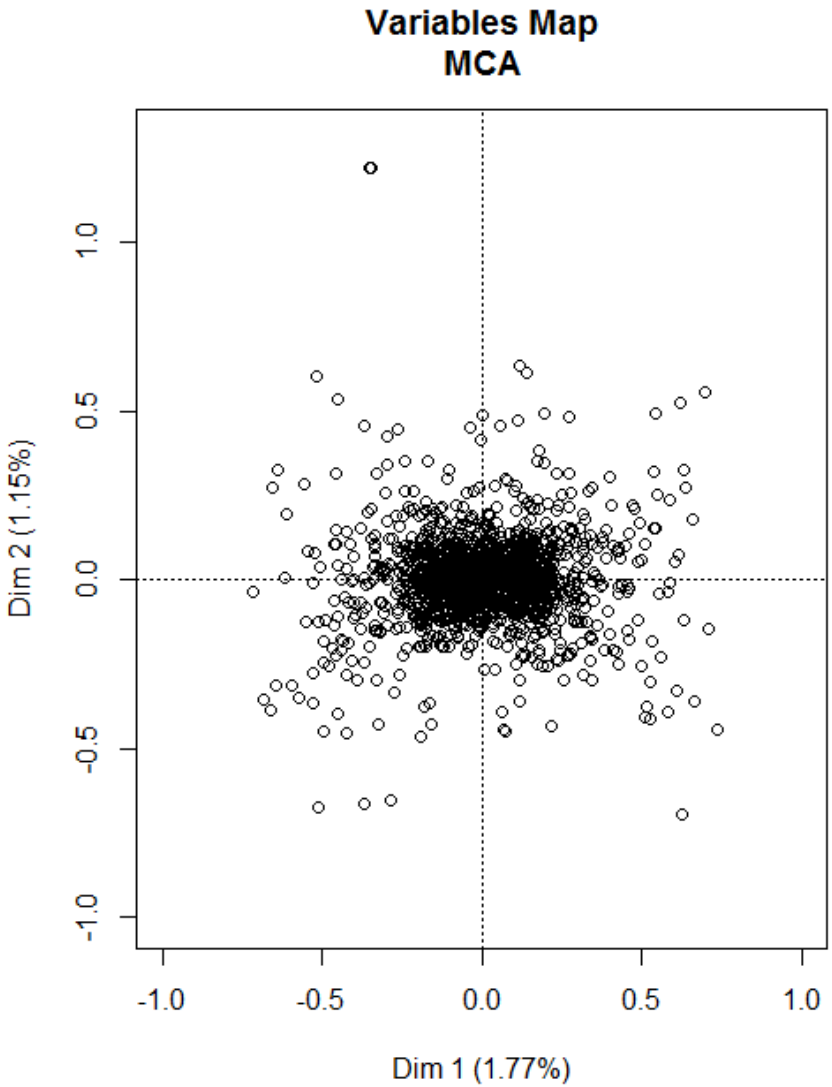
with $\mathbf{F}^{(1)} = \tilde{\mathbf{p}}\tilde{\mathbf{q}}$ the best rank-one matrix approximation of $\mathbf{F}$

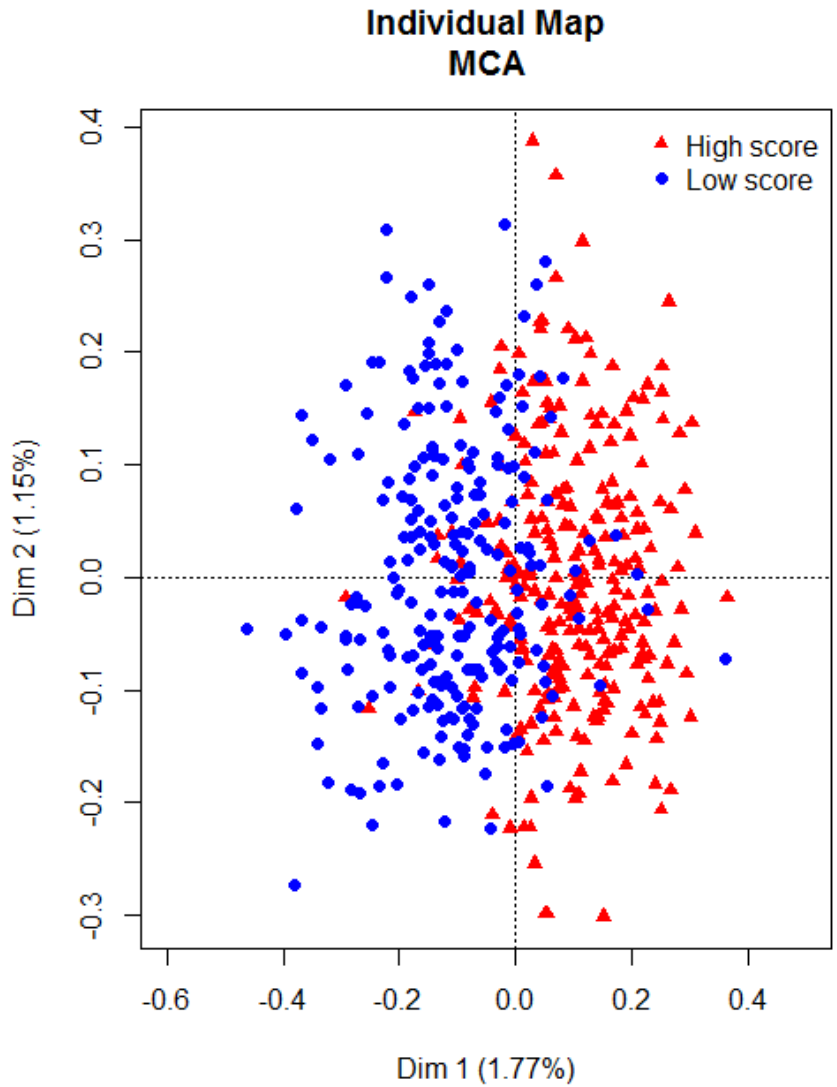
Application on SNPs data: MCA analysis



- The SNPs the most contributing to the first axis
- 2 SNPs are close if individuals have the same genetic
- Too many SNPs → we want to select the most relevant ones

Application on SNPs data: sparse MCA





Application on SNPs data

Step 1 Visualization of links between SNPs and between samples using **Multiple Correspondence Analysis** (MCA)

Step 2 Selection of the most important SNPs for a component using the sparse extension of MCA to select variables: **Sparse MCA**

From MCA to sparse MCA

Challenge

To facilitate interpretation of MCA results

$\Rightarrow$ Select the most contributing SNPs on each axis (easier visualisation of relationship between SNPs and phenotype)

How?

Constraints imposed in the MCA problem to set coefficients to zero

From MCA to sparse MCA

ID	SNP.1	SNP.2
AK	GG	AA
JD	GT	CC
MR	GT	AC
GB	GG	CC
NH	TT	AC
AL	GG	CC

GG	GT	TT	CC	AC	AA	ID
1	0	0	0	0	1	AK
0	1	0	1	0	0	JD
0	1	0	0	1	0	MR
1	0	0	1	0	0	GB

SNP1 SNP2

Selection of **1 column** in the original table (categorical variable **X**)

=

Selection of **a block of indicator variables** in the complete disjunctive table

From MCA to sparse MCA

Sparse MCA via GSVD

$$\min_{\tilde{\mathbf{p}},\tilde{\mathbf{q}}} \|\mathbf{F} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T\|^2_{\mathbf{W}} \qquad \tilde{\mathbf{p}}^T\mathbf{M}\tilde{\mathbf{p}} = \tilde{\mathbf{q}}^t\mathbf{W}\tilde{\mathbf{q}} = \mathbf{1} \qquad (2)$$

From MCA to sparse MCA

Sparse MCA via GSVD

+regularization penalty function applied on $\tilde{\mathbf{q}}$

$$\min_{\tilde{\mathbf{p}},\tilde{\mathbf{q}}} \|\mathbf{F} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T\|^2_{\mathbf{W}} + P_\lambda(\tilde{\mathbf{q}}) \qquad \tilde{\mathbf{p}}^T\mathbf{M}\tilde{\mathbf{p}} = \tilde{\mathbf{q}}^t\mathbf{W}\tilde{\mathbf{q}} = \mathbf{1} \qquad (2)$$

P_λ is a penalty function with tuning regularization parameter λ

From MCA to sparse MCA

Sparse MCA via GSVD

+regularization penalty function applied on $\tilde{\mathbf{q}}$

$$\min_{\tilde{\mathbf{p}},\tilde{\mathbf{q}}} \|\mathbf{F} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T\|^2_{\mathbf{W}} + P_\lambda(\tilde{\mathbf{q}}) \qquad \tilde{\mathbf{p}}^T\mathbf{M}\tilde{\mathbf{p}} = \tilde{\mathbf{q}}^t\mathbf{W}\tilde{\mathbf{q}} = \mathbf{1} \qquad (2)$$

P_λ is a penalty function with tuning regularization parameter λ

$\Rightarrow$ Use the **Group LASSO penalization**

$$P_\lambda(\beta) = \lambda \sum_{k=1}^{K} \sqrt{J_{[k]}} \|\beta_{[k]}\|_2$$

$J_{[k]}$: number of variables in block k

λ: penalty parameter to determine (cross validation, ad hoc approach,...)

From MCA to sparse MCA

Sparse MCA via GSVD

+regularization penalty function applied on $\tilde{\mathbf{q}}$

$$\min_{\tilde{\mathbf{p}},\tilde{\mathbf{q}}} \|\mathbf{F} - \tilde{\mathbf{p}}\tilde{\mathbf{q}}^T\|^2_{\mathbf{W}} + P_\lambda(\tilde{\mathbf{q}}) \qquad \tilde{\mathbf{p}}^T\mathbf{M}\tilde{\mathbf{p}} = \tilde{\mathbf{q}}^t\mathbf{W}\tilde{\mathbf{q}} = \mathbf{1} \qquad (2)$$

P_λ is a penalty function with tuning regularization parameter λ

$\Rightarrow$ Use the **Group LASSO penalization**

$$P_\lambda(\beta) = \lambda \sum_{k=1}^{K} \sqrt{J_{[k]}} \|\beta_{[k]}\|_2$$

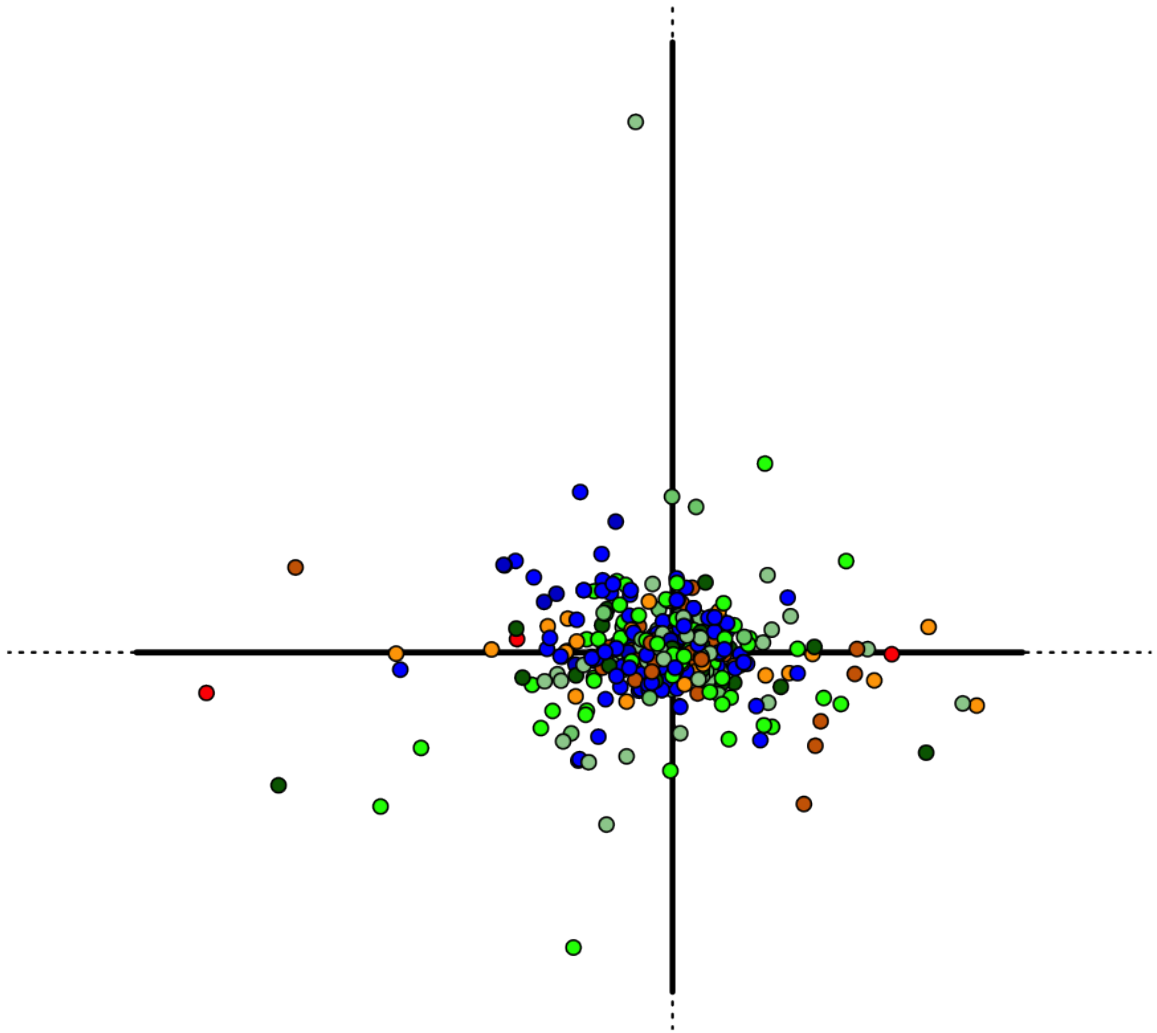
$J_{[k]}$: number of variables in block k

λ: penalty parameter to determine (cross validation, ad hoc approach,...)

Result: Entire blocks of dummy variables are selected or removed

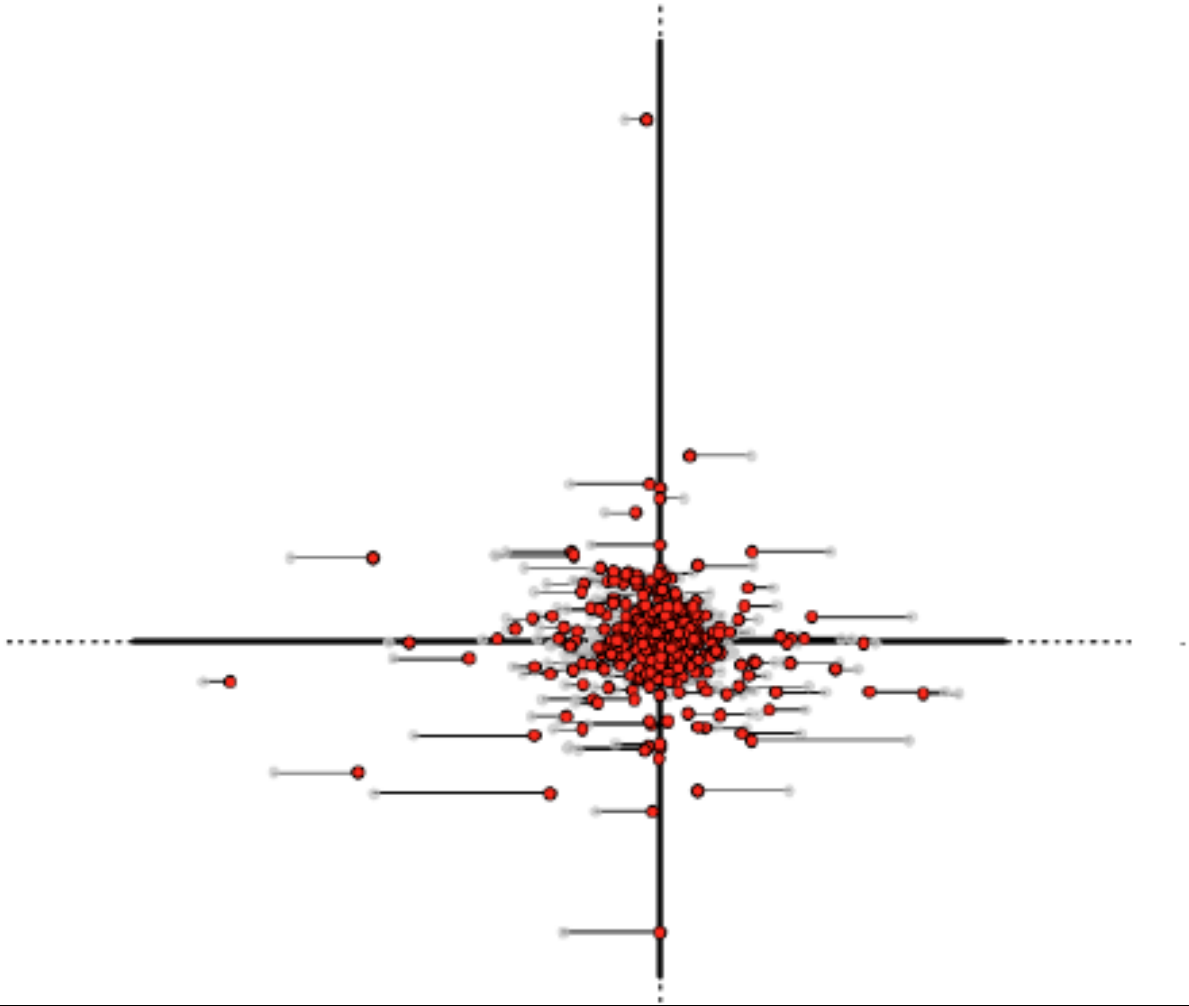
Penalty parameter influence

Tunning parameter $\lambda = 0 \Rightarrow$ sparse MCA=MCA



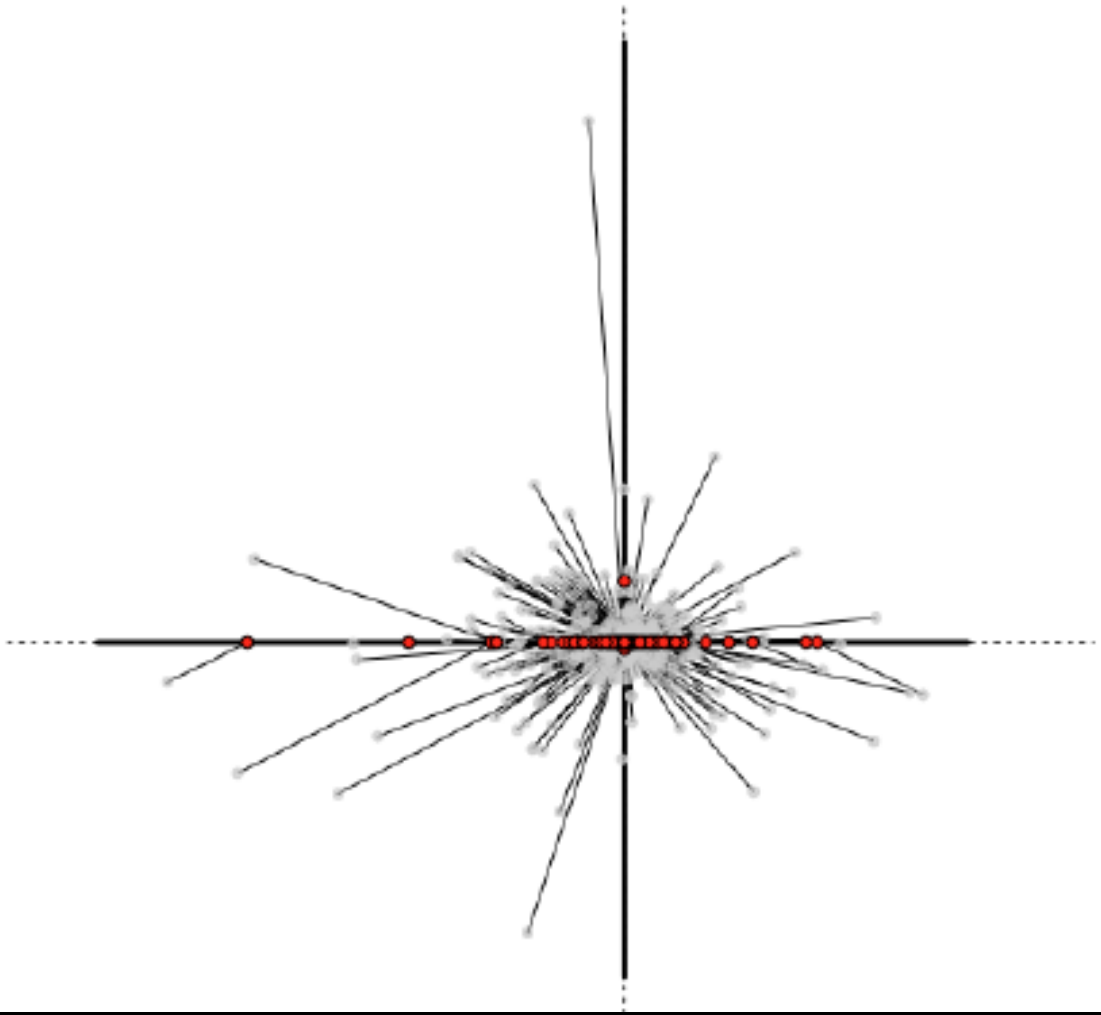
Penalty parameter influence

λ increasing $\Rightarrow$ some loadings set to 0



Penalty parameter influence

λ increasing again $\Rightarrow$ more loadings set to 0



Case study: Genes potentially involved in skin aging

Sagging

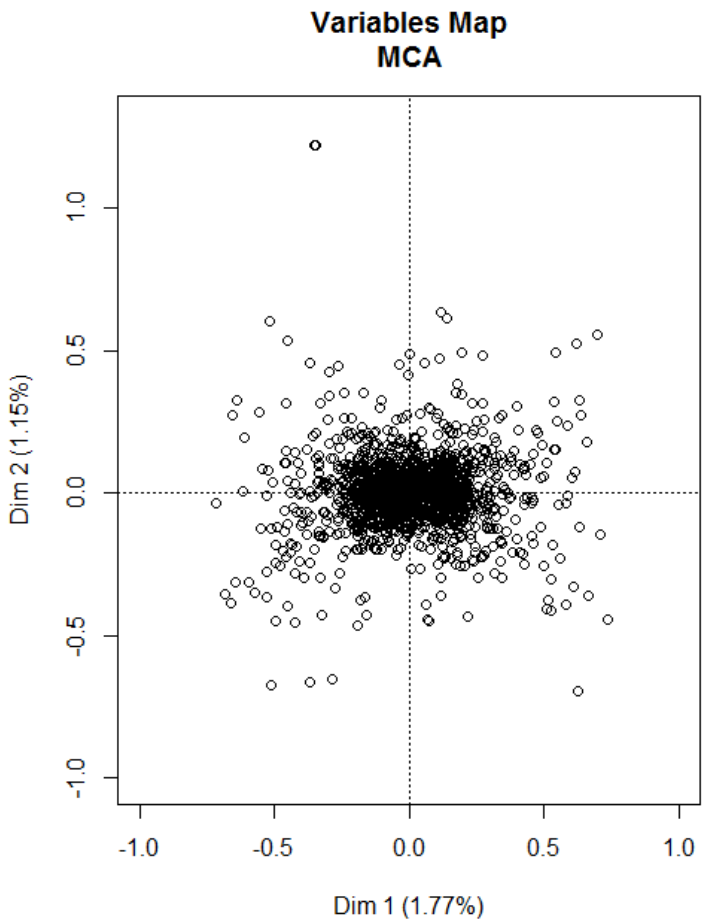
537 targeted SNPs
≈ 498 genes

Selection with Sparse MCA

142 SNPs
≈ 112 genes

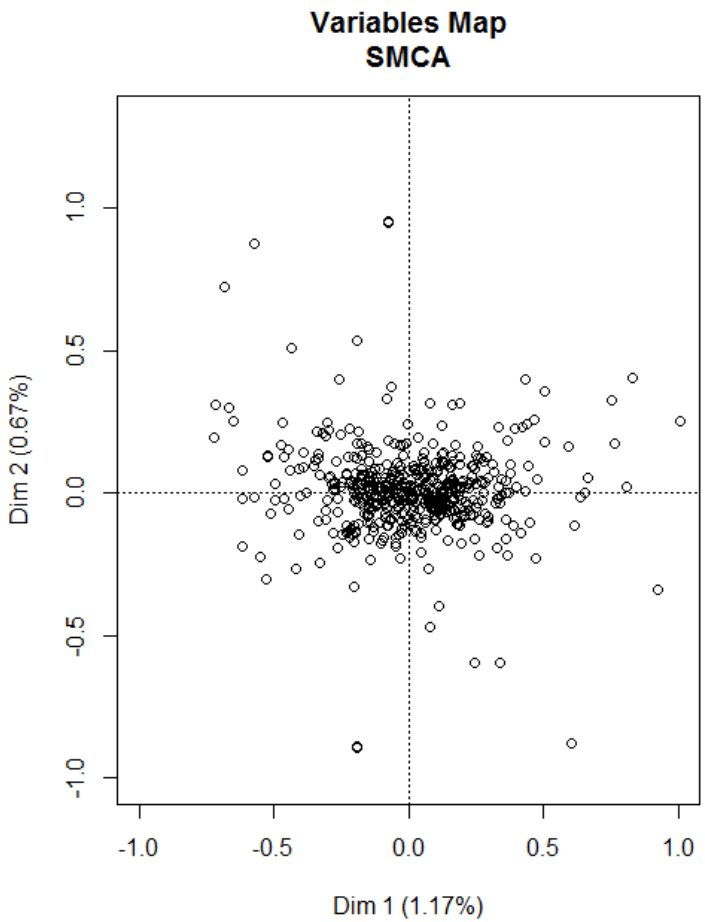
Application on SNPs data: sparse MCA

Before selection



537 SNPs

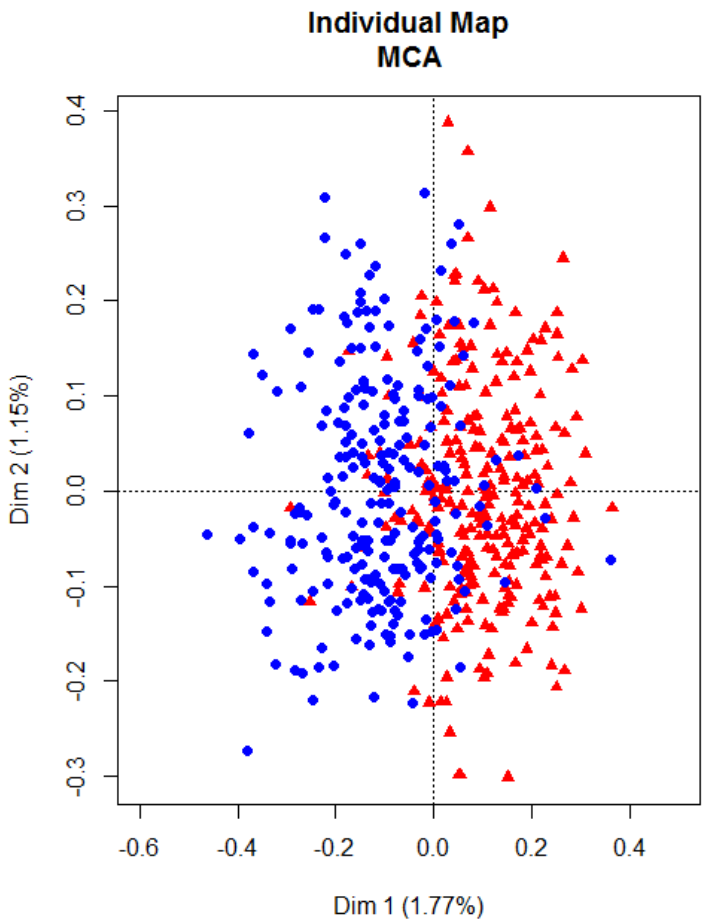
After selection



142 SNPs

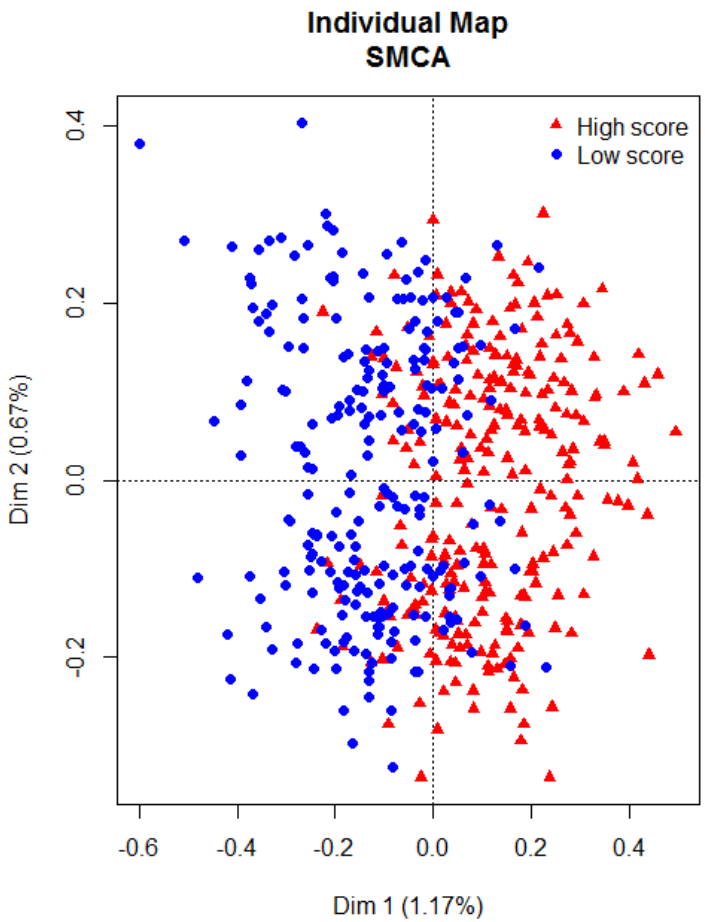
Application on SNPs data: sparse MCA

Before selection



537 SNPs

After selection



142 SNPs

Application on SNPs data: sparse MCA

Comparison of the loadings

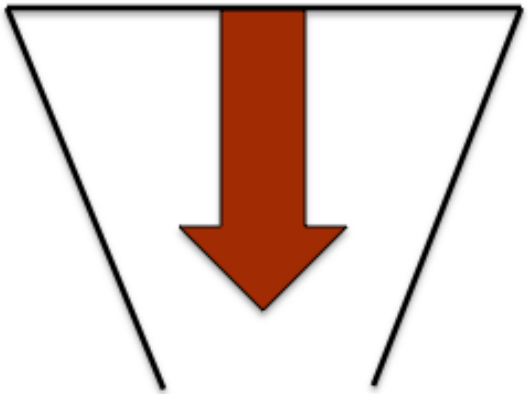
SNPs	MCA		SMCA	
	Comp1	Comp2	Comp1	Comp2
SNP1.AA	-0.078	0.040	-0.092	0.102
SNP1.AG	-0.014	-0.027	-0.022	-0.053
SNP1.GG	0.150	-0.002	0.132	-0.003
SNP2.AA	-0.082	0.041	**0.000**	**0.000**
SNP2.AG	-0.021	-0.025	**0.000**	**0.000**
SNP2.GG	-0.081	0.040	**0.000**	**0.000**
SNP3.CC	-0.004	0.050	**0.000**	0.083
SNP3.CG	0.016	0.021	**0.000**	0.042
SNP3.GG	-0.037	-0.325	**0.000**	-0.432
SNP4.AA	0.149	-0.003	0.050	**0.000**
SNP4.AG	-0.016	-0.025	-0.002	**0.000**
SNP4.GG	-0.081	0.040	-0.100	**0.000**
...	...	...	...	...
Nb non-zero loadings	1554	1554	172	108
Variance (%)	1.14	0.63	0.32	0.16
Cumulative variance (%)	1.14	1.77	0.32	0.48

Application on SNPs data: sparse MCA

Sagging 

537 targeted SNPs

≈ 498 genes



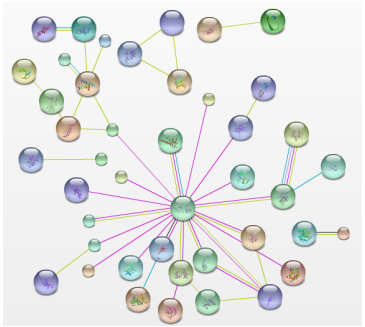
Selection with Sparse MCA

142 SNPs

≈ 112 genes

Conclusion of the case study

- Selection of relevant SNPs that explain the most important variability between indivuals
- Results generating new biological hypotheses to be further investigated
- Further analyses: detection of interactions, pathway analysis



- Selected SNPs implicated in biological pathways such as Map-Kinase (cell growth factor) linked with skin aging

General conclusions

- Unsupervised method to select categorical variables
- Produce sparse loading structures
  $\rightarrow$ easier interpretation of the results
- Powerful in a context of variable selection in high dimension issues
  $\rightarrow$ reduce noise as well as computation time
- **Research in progress:** Extension of Sparse MCA to select variables within a block
  $\rightarrow$ sparsity at both group and individual feature levels

References

Bernard, A., Guinot, C. and Saporta, G. (2012), Sparse principal component analysis for multiblock data and its extension to sparse multiple correspondence analysis, Proceedings Compstat 2012, 99-106,

Jolliffe, I. T. et al. (2003), A modified principal component technique based on the LASSO, *Journal of Computational and Graphical Statistics*, 12, 531–547,

Lê Cao, K.-A., Gonzalez I. and Dejean S (2009). integrOmics: an R package to unravel relationships between two omics data sets, Bioinformatics 25(21):2855–2856,

Shen, H. and Huang, J. (2008), Sparse principal component analysis via regularized low rank matrix approximation, *Journal of Multivariate Analysis*, 99, 1015–1034,

Yuan, M. and Lin, Y. (2006), Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B*, 68, 49–67,

Zou, H., Hastie, T. and Tibshirani, R. (2006), Sparse Principal Component Analysis, *Journal of Computational and Graphical Statistics*, 15, 265–286.

Acknowledgements

Work in collaboration with:

Pr. Gilbert Saporta *CNAM, laboratoire CEDRIC, Paris, France*

Dr. Christiane Guinot *Université Francois Rabelais, département d'informatique, Tours, France*

Pr. Hervé Abdi *Department of Brain and Behavioral Sciences, The University of Texas at Dallas, Richardson, TX, USA*

Pr. Arthur Tenenhaus *SUPELEC, Gif-sur-Yvette, France*