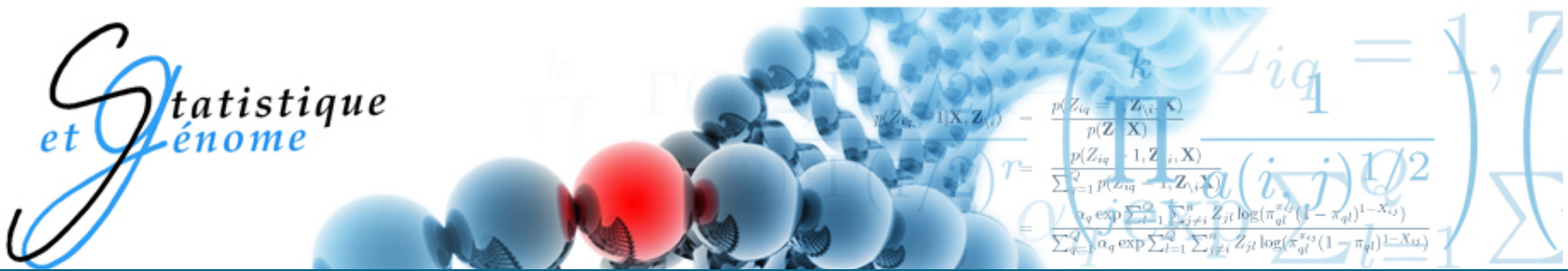
Laboratoire de Mathématiques et Modélisation d'Évry

LaMME

Régression en grande dimension et épistasie par blocs pour les études d'association

V. Stanislas, C. Dalmasso, C. Ambroise

Laboratoire de Mathématiques et Modélisation d'Évry
"Statistique et Génome"



Summary

1. GWAS and Block of linkage desequilibrium
   - Genome Wide Association Studies
   - Blocks of linkage desiquilibrium
   - Hierachical Clustering with Adjacency Constraints
   - How to improve ?
   - Some computation times
2. Epistasis
   - The Gene-Gene Eigen Epistasis Modeling approach
   - Simulations
3. Application
   - Ankylosing Spondylitis
   - First results

Sommaire

1. GWAS and Block of linkage desequilibrium
   - Genome Wide Association Studies
   - Blocks of linkage desiquilibrium
   - Hierachical Clustering with Adjacency Constraints
   - How to improve ?
   - Some computation times
2. Epistasis
   - The Gene-Gene Eigen Epistasis Modeling approach
   - Simulations
3. Application
   - Ankylosing Spondylitis
   - First results

Single-Nucleotide Polymorphism Data

- 90 % of human genetic variation,
- In human genom, SNP with allelic frequency greater than 1 % are present every 300 base pairs (in average)
- 2 SNP among 3 substitute cytosine with thymine

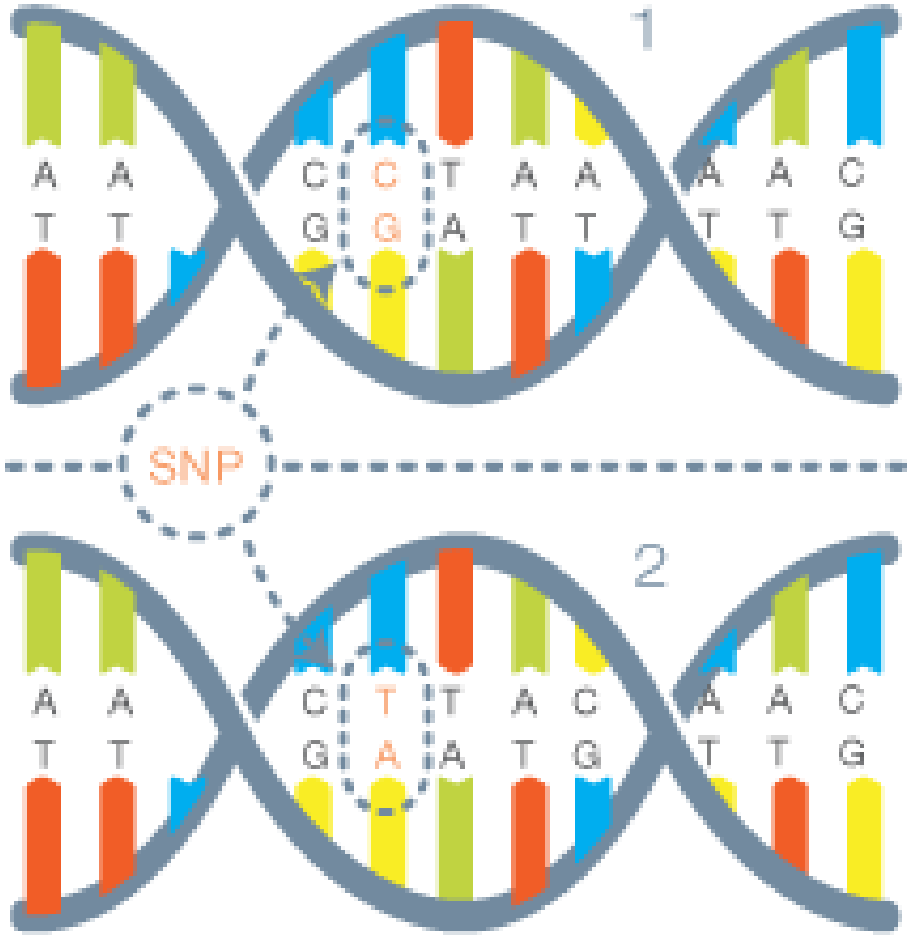


Figure: SNP (wikipedia)

SNP Data

Four possible C/T configurations

$X_{ij} \in \{0, 1, 2\}$ (Individual i at locus j)

Mother	Father C	Father T
C	0	1
T	1	2

High-dimension

p markers

$$\text{n individuals} \begin{pmatrix} \mathbf{x}_{11} & \cdots & \cdots & \mathbf{x}_{1p} \\ \vdots & & & \vdots \\ \mathbf{x}_{n1} & \cdots & \cdots & \mathbf{x}_{np} \end{pmatrix}$$

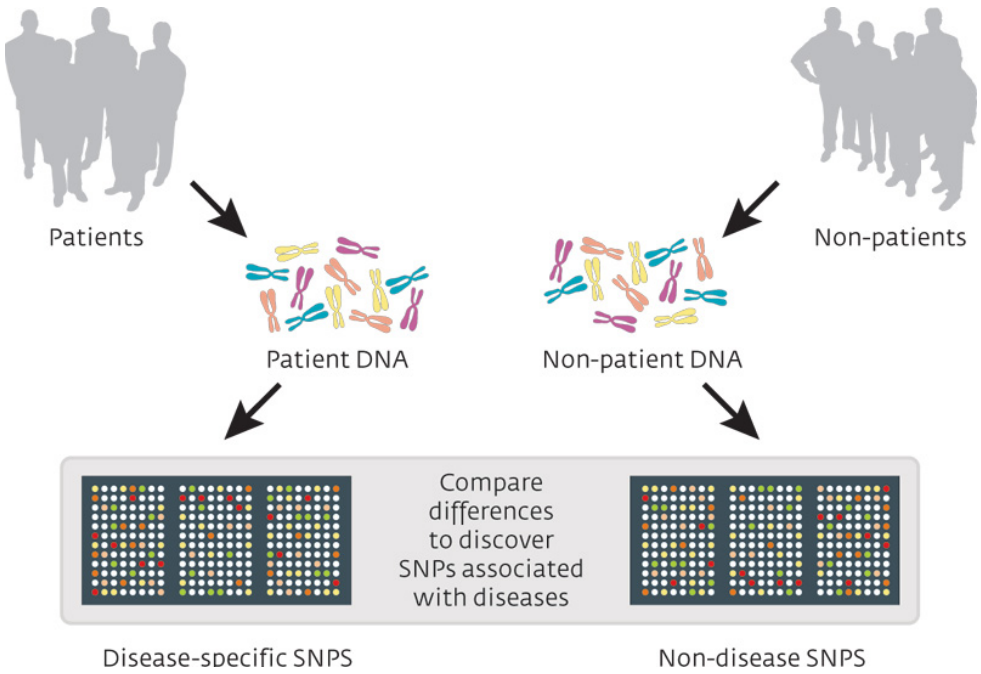
$\approx$ up to few million of SNPs can be genotyped !

$\Rightarrow$ number of variables $\gg$ number of individuals ($p \gg n$)

Genome-Wide Association Studies

GWAS characteristics :

- **Objective :** find associations between genetic markers ($SNP_{i,j} \in \{0, 1, 2\}$) and a phenotypic trait ($Y_i \in \{0, 1\}$ or $Y_i \in \mathbb{R}$)



©Pasieka, Science Photo Library

Genetic markers ➔ SNP

http ://www.siriusgenomics.com/technology/

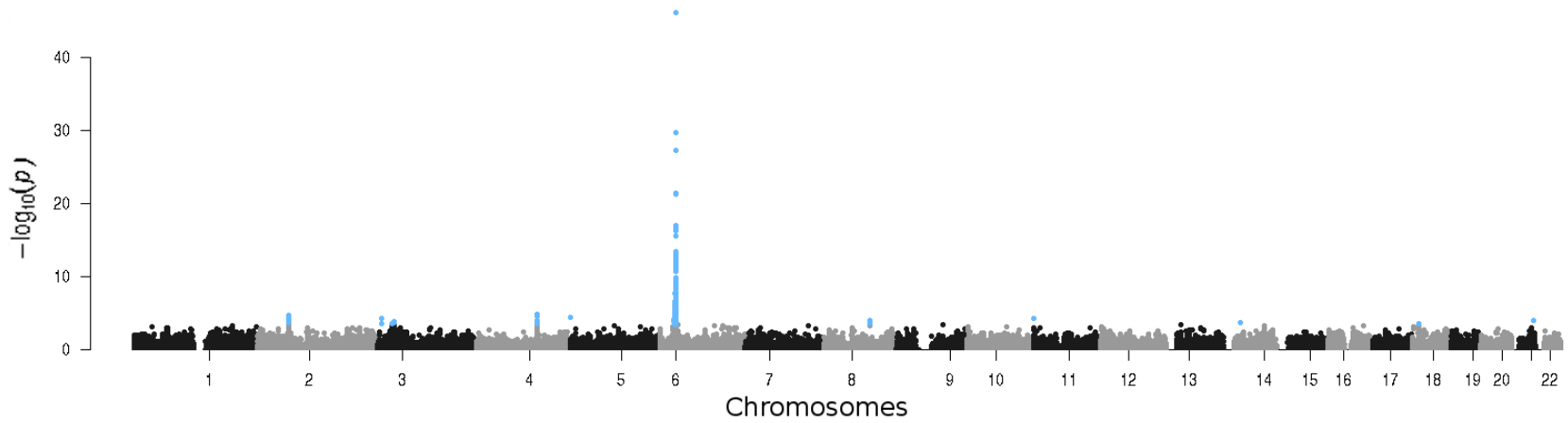
- **Generalized Linear Model**

$$g(E[Y_i|x_i]) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \quad , i = 1, \ldots, n$$

  - n : number of individuals
  - p : number of covariates
  - Y_i : response for the individual i
  - $x_{.j}$: observations for covariate j (coded in 0, 1 or 2)

Genome-Wide Association Studies

- **SNP analysis**
  Differences between cases and controls at a specific SNP



- **GWAS limits :**
  - Reproductibility
  - Heritability
- **Data particularities :**
  - Structuration
  - High dimension (p » n)
  - Small effects

The LD measures

Linkage Desiquilibrium

- non-random association of alleles at two or more loci
- depends on the difference between observed allelic frequencies and those expected from a independent randomly distributed model.

Computation

- Z_j the indicator of the presence of minor allele for SNP j.
- $Z_j \sim \mathcal{B}(p_j)$

$$D(j,k) = p_{jk} - p_j p_k = E[Z_j Zk] - p_j p_k = cov(Z_j, Z_k)$$

$$r^2(j,k) = corr(Z_j, Z_k)$$

ou

$$D'(j,k) = D(j,k)/Dmax(j,k)$$

How to estimate LD ?

snp	vv	vV	VV
uu	a	b	c
uU	d	e	f
UU	g	h	i

snp	v	V
u	α	β
U	γ	δ

Only the genotype data table is observed

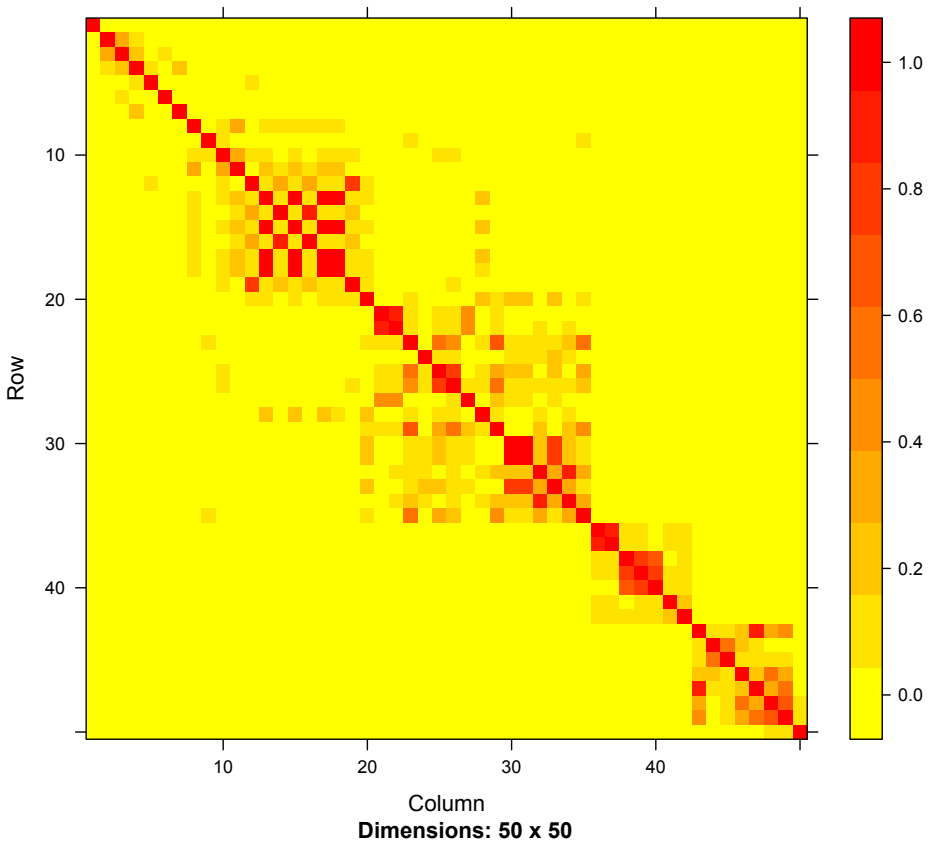
- α, β, γ, δ are estimated
- a system of equations. e.g : $\alpha = 2a + b + d + pe$

with p the "probability" of the haplotype (uv, UV).

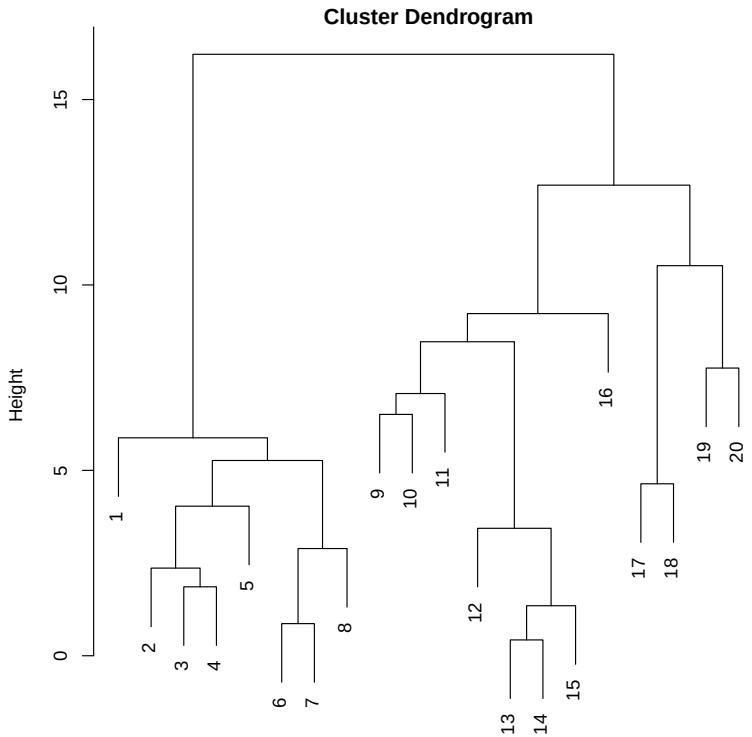
$\Rightarrow$ estimating p, then $(\alpha, \beta, \gamma, \delta)$ and finally $D = p_{UV} - p_U p_V$.

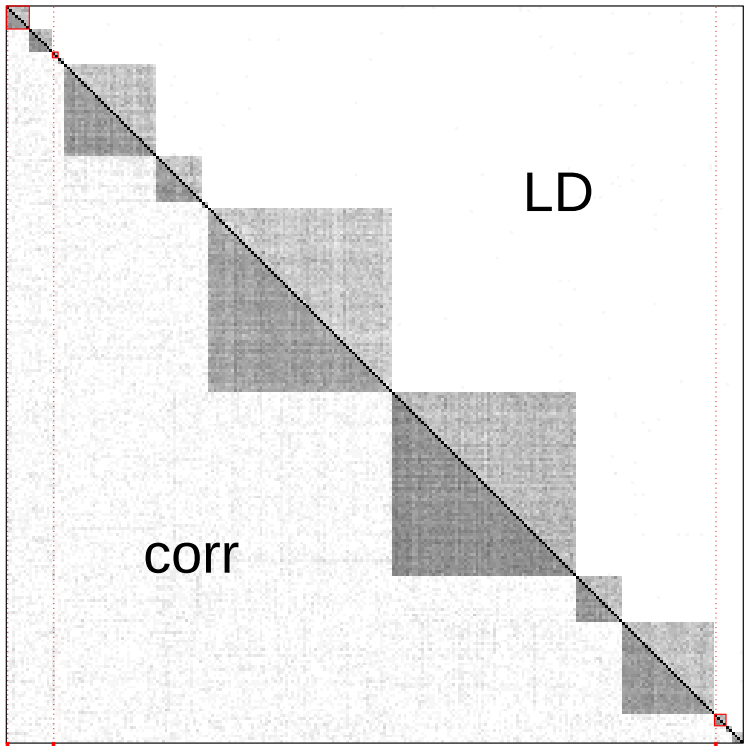
The LD block structure

- the r^2 coefficients among the **50 first SNPs** of the Chromosome 22 (Dalmasso et al. 2008)
- LD structured in blocks



Hierachical Clustering with Adjacency Constraints



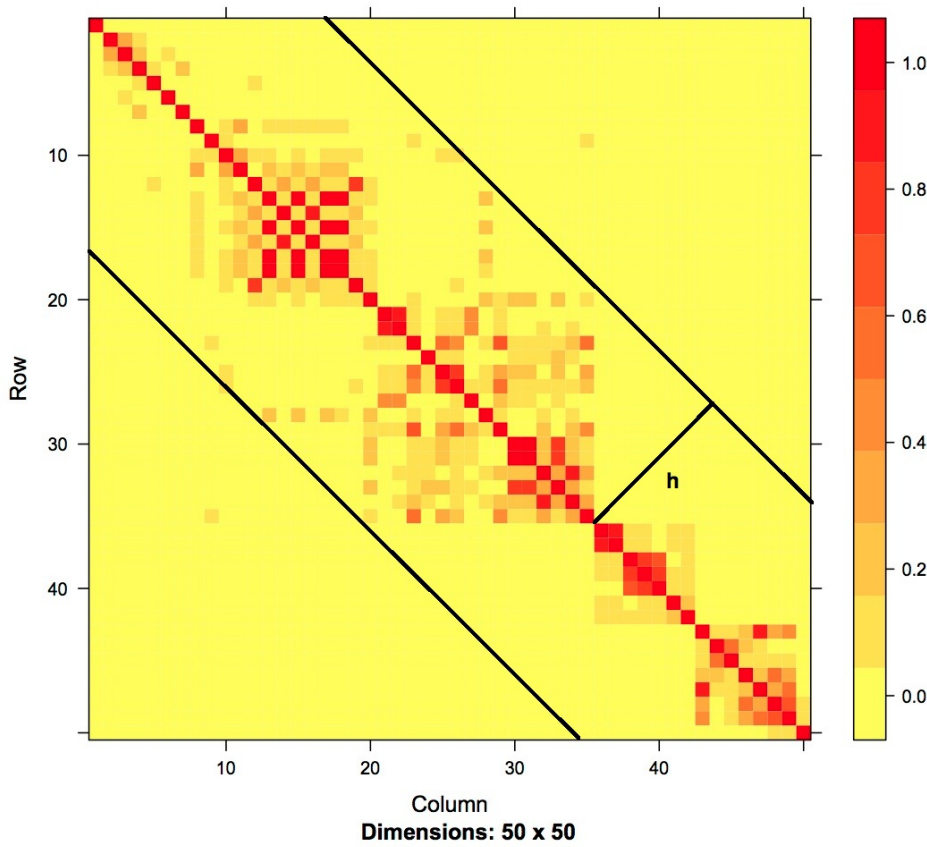


Block-Wise Approach using Linkage Disequilibrium (BALD)

1. Hierarchical clustering of the SNPs with adjacency constraint and using the LD similarity.
2. Estimation of the optimal number of groups using the Gap statistic (Tibshirani et. al., 2001).

The h-band

- All coefficients outside the band "h" are null
- a $p \times h$ similarity matrix



$\Rightarrow$ a hierarchical clustering with adjacency constraint

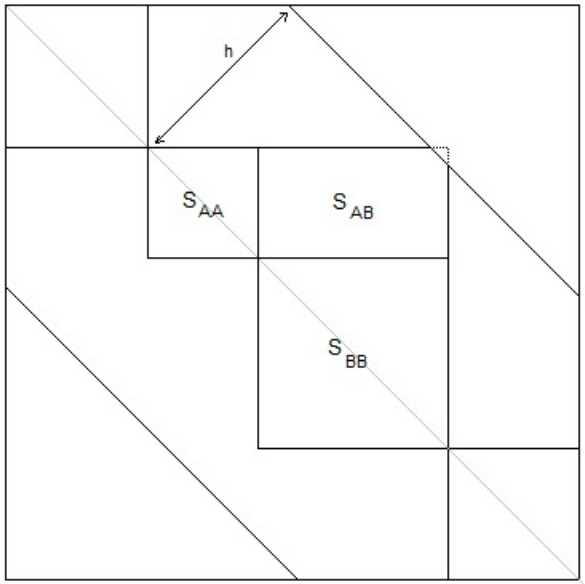
A pseudocode

Data: $\mathbf{X} \in \{0,1,2\}^{n \times p}$, Sim
$\mathcal{C} \leftarrow \{C_i = \{\mathbf{X}_{.i}\}, i \in 1,\ldots,p\}$ `/* clusters = singletons */` ;
$D \leftarrow \{1 - Sim(\mathbf{X}_{.i}, \mathbf{X}_{.(i+1)}), i \in 1,\ldots,p-1\}$;
for $step = 1$ ***to*** $p-1$ **do**
  $i^\star \leftarrow \text{argmin}_{i \in \{1,\ldots,p-step\}} D(C_i, C_{i+1})$;
  $\mathcal{C} \leftarrow \mathcal{C} \setminus \{C_{i^\star}, C_{i^\star+1}\} \cup \{C_{i^\star} \cup C_{i^\star+1}\}$;
  $d_1 \leftarrow D(C_{i^\star-1}, C_{i^\star} \cup C_{i^\star+1})$;
  $d_2 \leftarrow D(C_{i^\star} \cup C_{i^\star+1}, C_{i^\star+2})$;
  $D \leftarrow D \setminus \{D(C_{i^\star-1}, C_{i^\star}), D(C_{i^\star}, C_{i^\star+1})\} \cup \{d_1, d_2\}$;
end

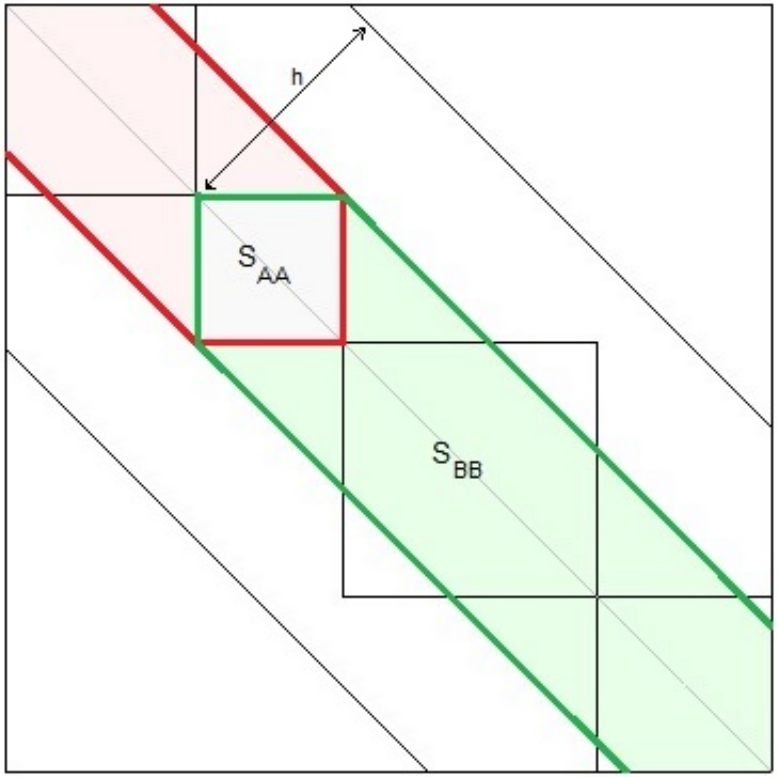
The Ward's distance

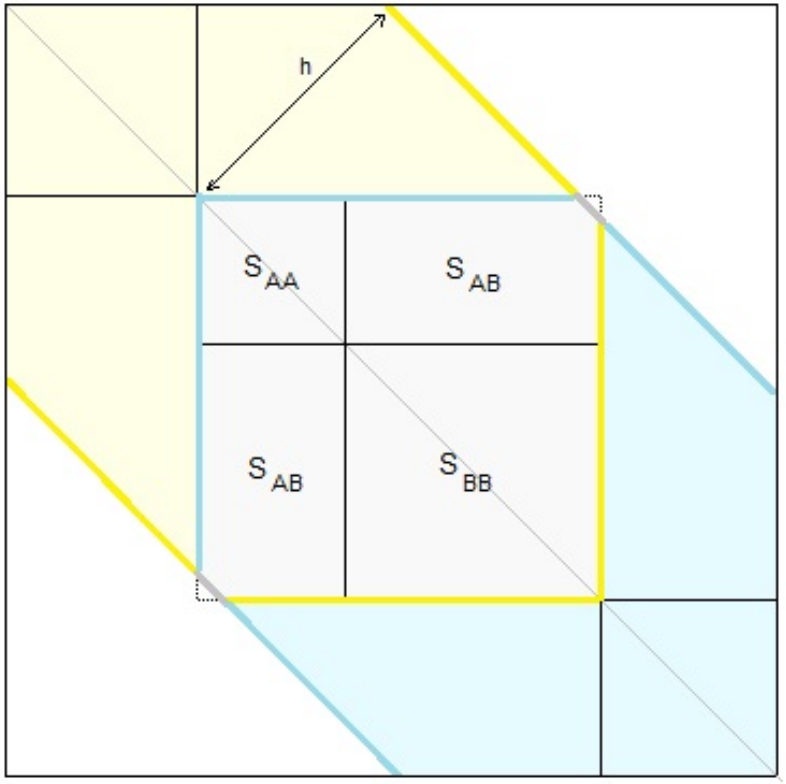
Ward Constrained Hierarchical Clustering

$$d(A,B) = \frac{n_A n_B}{n_A + n_B}\left(\frac{1}{n_A^2}S_{A,A} + \frac{1}{n_B^2}S_{B,B} - \frac{2}{n_A n_B}S_{A,B}\right)$$



The pencils' trick : Calculating S_{AA} and S_{AB}





The "pencils"

Assessing S_{AA}, S_{BB} and S_{AB} requires the calculation of sums of LD measures within *pencil-shaped areas* defined by :

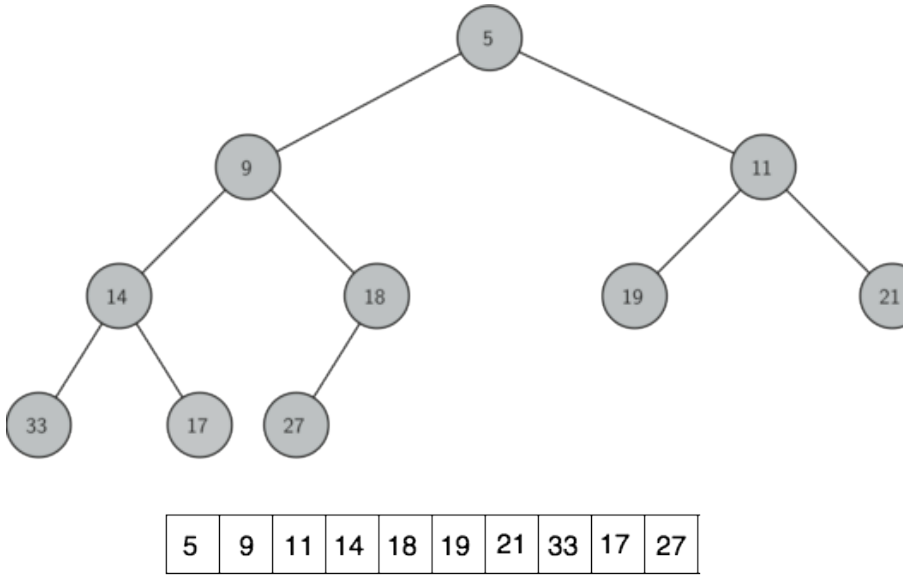
- direction : `right` or `left`
- depth : `hLoc`
- end point : `lim`

$\Rightarrow$ Two arrays of sizes $p \times h$ for storing the pencils sums.

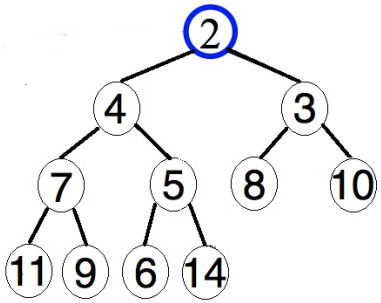
The binary min-heap

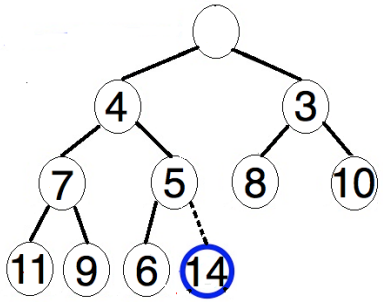
- All nodes are either **less than or equal** to each of its children.

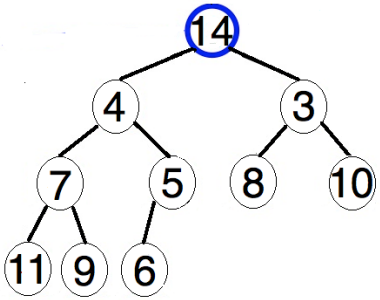
- Uniquely represented by storing its level order traversal in an array. Given a position i :
  - Parent$(i) = \lfloor i/2 \rfloor$
  - Left$(i) = 2i$
  - Right$(i) = 2i + 1$

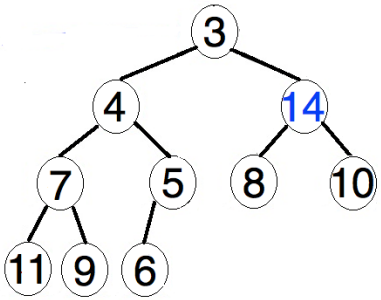


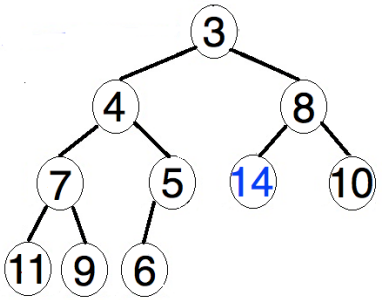
DeleteMin





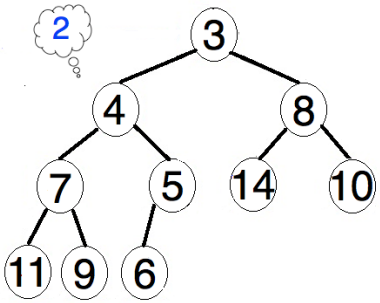


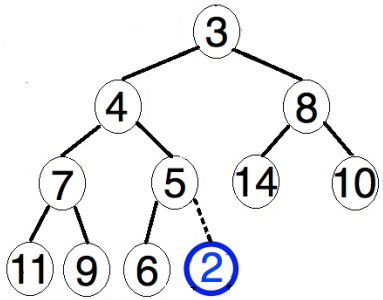


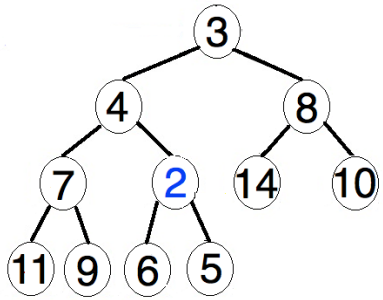


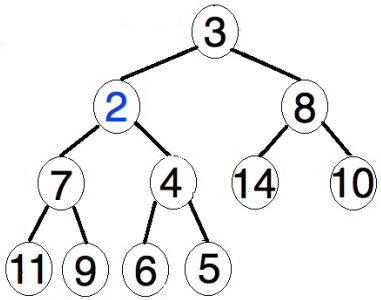
Time complexity : $\mathcal{O}(log(p))$

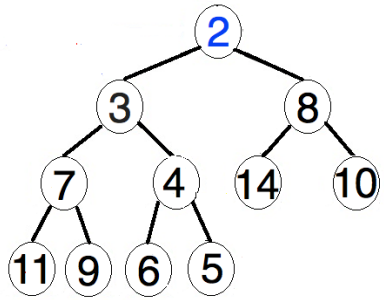
InsertHeap







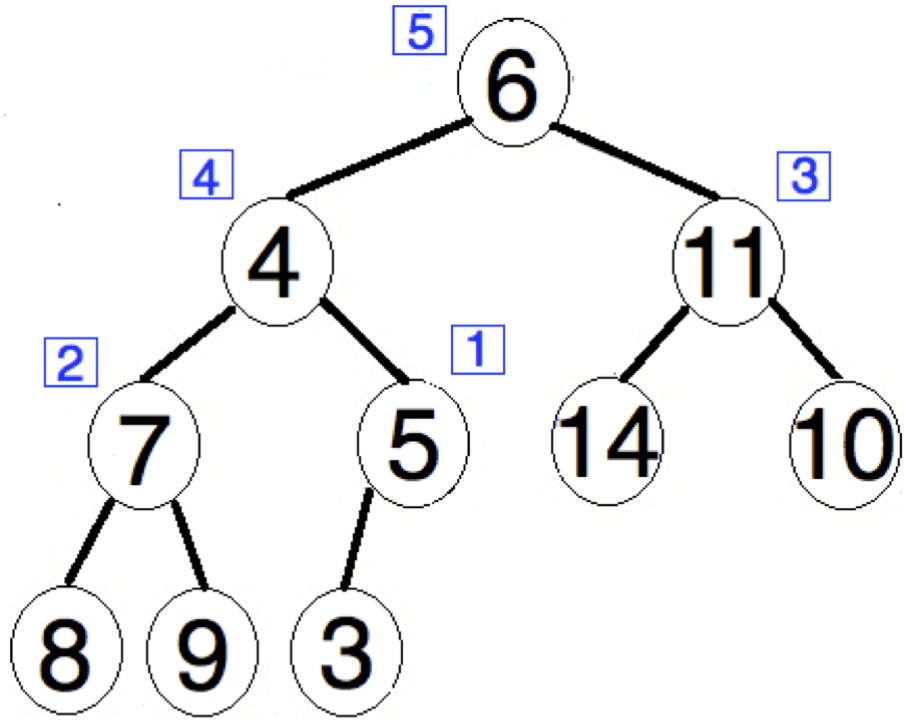




Time complexity : $\mathcal{O}(log(p))$

BuildHeap

Data: An array A
Result: A min-heap H
for $i = \lfloor length(A)/2 \rfloor$ ***down to*** 1
do
  PercolDown(A, i);
end



6	4	11	7	5	14	10	8	9	3

Time complexity : $\mathcal{O}(plog(p))$

Time complexity of some operations

	findMin	**insert**	**deleteMin**
unordered array	$\mathcal{O}(p)$	$\mathcal{O}(1)$	$\mathcal{O}(p)$
binary heap	$\mathcal{O}(1)$	$\mathcal{O}(log(p))$	$\mathcal{O}(log(p))$

cWard in seconds...

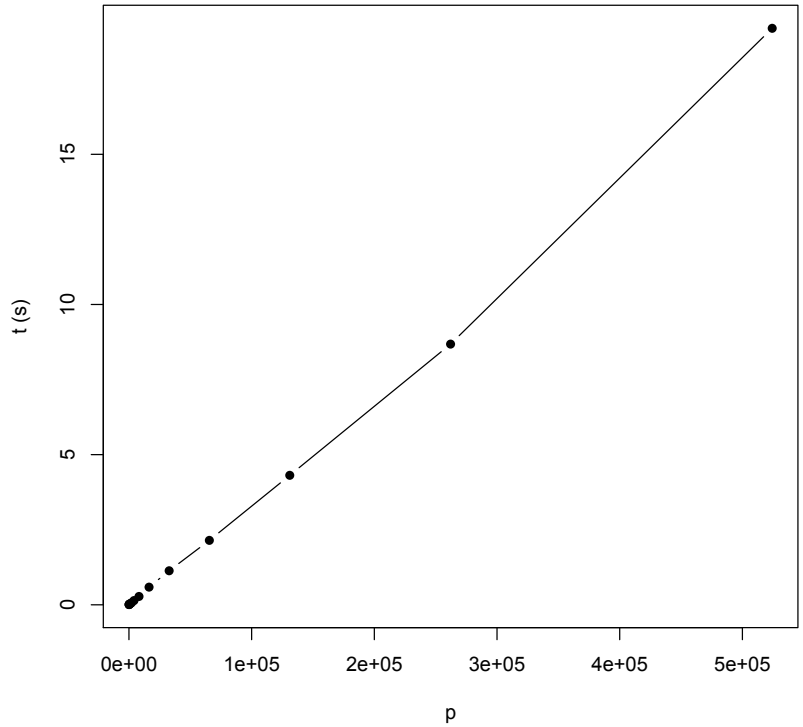


Figure: The mean computation time t versus the number of markers p for the cWard algorithm applied to randomly sampled SNP matrices. $N = 100$, $h = 30$ and t is averaged across 50 simulation runs.

Compared to a former implementation

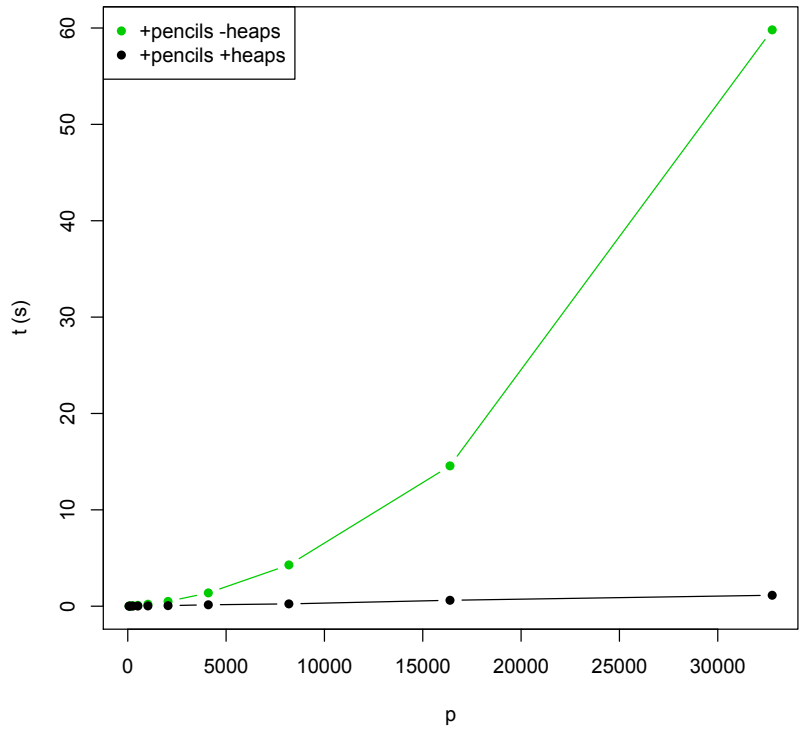


Figure: The mean computation time t versus the number of markers p for the cWard algorithm and an implementation without heaps. t is averaged across 20 simulation runs.

Scalable Hierarchical Clustering with pencils and binary heap

To sum up :

1. A $\mathcal{O}(p^2)$ algorithm does not scale for GWA studies.
2. The Ward distance written in a simple way.
3. Space complexity of $\mathcal{O}(ph)$ by using the pencils' trick.
4. Time complexity of :

$$\underbrace{\mathcal{O}(ph)}_{\text{2 } p\times h \text{ arrays of pencils' sums}} + \underbrace{\mathcal{O}(plog(p))}_{\substack{\text{building the heap and} \\ \text{insert/delete heaps' operations within the loop}}}$$

Ongoing work :

- Currently implemented with a genotype matrix as input.
  $\Rightarrow$ can be generalized to any band similarity matrix.

Sommaire

1. GWAS and Block of linkage desequilibrium
   - Genome Wide Association Studies
   - Blocks of linkage desiquilibrium
   - Hierachical Clustering with Adjacency Constraints
   - How to improve ?
   - Some computation times
2. Epistasis
   - The Gene-Gene Eigen Epistasis Modeling approach
   - Simulations
3. Application
   - Ankylosing Spondylitis
   - First results

Epistasis

Definition

Interaction of alleles effects from different markers

Existing methods

- mainly SNP x SNP
- some at the block (gene) scale

Advantages of gene (or block) scale approaches

- results biologically interpretable
- genetic effects may be easier to detect
- reduce the number of variables

Epistasis - Gene scale methods

Existing gene scale methods :

Two or few genes

- PCA + logistic regression *(He et al. 2011, Li et al. 2009, Zhang et al. 2008)*
- PLS + logistic regression *(Wang T et al. 2009)*

For a larger number of genes

- PCA + LASSO *(D'Angelo et al. 2009)*
- PCA + pathway-guided penalized regression *(Wang X et al. 2014)*

Epistasis - Gene scale methods

Existing gene scale methods :

Two or few genes

- PCA + logistic regression *(He et al. 2011, Li et al. 2009, Zhang et al. 2008)*
- PLS + logistic regression *(Wang T et al. 2009)*

For a larger number of genes

- PCA + LASSO *(D'Angelo et al. 2009)*
- PCA + pathway-guided penalized regression *(Wang X et al. 2014)*

Objectives : To develop a new gene scale method

➔ considers a more accurate definition of interaction variables,
➔ is applicable with with genes,
➔ takes into account the group structure

Group modeling approach

	$SNP_{1,1}$	..	SNP_{1,p_1}	..	$SNP_{G,1}$	..	SNP_{G,p_G}	Pheno
Ind_1	1		0		0		1	1
Ind_2	0		0		2		1	0
.	2		1		1		2	1
.	0		1		0		0	0
Ind_i	0		2		1		0	0
	$\underbrace{\quad\quad\quad}_{gene_1}$				$\underbrace{\quad\quad\quad}_{gene_G}$			

We note $SNP_{1,1} = X_{1,1}$

Group modeling approach

	$SNP_{1,1}$	..	SNP_{1,p_1}	..	$SNP_{G,1}$	..	SNP_{G,p_G}	Pheno
Ind_1	1		0		0		1	1
Ind_2	0		0		2		1	0
.	2		1		1		2	1
.	0		1		0		0	0
Ind_i	0		2		1		0	0
	$\underbrace{\quad\quad\quad}_{gene_1}$				$\underbrace{\quad\quad\quad}_{gene_G}$			

We note $SNP_{1,1} = X_{1,1}$

r, s two genes

Group modeling approach

Model

$$Y_i = \underbrace{\beta_0 + \sum_g \beta_g \left(\sum_{k \in \mathcal{C}} X_{ik}^g \right)}_{\text{Main effects}} + \epsilon_i$$

$$\beta = \left(\underbrace{\beta_{1,1}, \beta_{1,2}, \cdots, \beta_{1,p_1}}_{gene_1}, \cdots, \underbrace{\beta_{G,1}, \cdots, \beta_{G,p_G}}_{gene_G} \right)^T$$

Group modeling approach

Model

$$Y_i = \underbrace{\beta_0 + \sum_g \beta_g \left(\sum_{k \in \mathcal{C}} X_{ik}^g \right)}_{\text{Main effects}} + \underbrace{\sum_{r,s} \gamma_{r,s} Z_i^{r,s}}_{\text{Interaction effects}} + \epsilon_i$$

$$\beta = \left(\underbrace{\beta_{1,1}, \beta_{1,2}, \cdots, \beta_{1,p_1}}_{gene_1}, \cdots, \underbrace{\beta_{G,1}, \cdots, \beta_{G,p_G}}_{gene_G} \right)^T \qquad \gamma = \left(\gamma_{12}, \cdots \underbrace{\gamma_{1G}}_{\gamma_{1G,1}, \cdots, \gamma_{1G,q}} \cdots, \gamma_{(G-1)G} \right)$$

q : # of interaction variables for a couple

Interaction variable : Gene-Gene Eigen Epistasis (G-GEE)

We consider $f_{\boldsymbol{u}}(\boldsymbol{X}_i^r, \boldsymbol{X}_i^s)$ to represent the interaction between genes r, s.

$$\hat{\boldsymbol{u}} = arg \max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} cor(\boldsymbol{y}, f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s))$$

Eigen Epistasis

$f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s) = \boldsymbol{W}^{rs}\boldsymbol{u}$ with $\boldsymbol{W}^{rs} = \{X_{ij}^r X_{ik}^s\}_{i=1\cdots n}^{j=1\cdots,p_r;k=1,\cdots,p_s}$

$$\max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} ||\hat{cor}[\boldsymbol{W}^{rs}\boldsymbol{u}, \boldsymbol{y}]||^2 == \max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} \boldsymbol{u}^T \boldsymbol{W}^{rs\,T} \boldsymbol{y}\boldsymbol{y}^T \boldsymbol{W}^{rs}\boldsymbol{u}$$

$\boldsymbol{u}$: only eigen vector associated of $\boldsymbol{W}^{rs\,T}\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{W}^{rs}$
For each couple (r, s) → $\boldsymbol{Z}^{rs} = \boldsymbol{W}^{rs}\boldsymbol{u}$

Coefficients estimation

Group LASSO regression

$$(\hat{\beta}, \hat{\gamma}) = \underset{\beta,\gamma}{\operatorname{argmin}} \sum_i (y_i - \boldsymbol{X}_i\beta - \boldsymbol{Z}_i\gamma)^2 + \lambda \left(\sum_g \sqrt{p_g} ||\beta^g||_2 + \sum_{rs} \sqrt{p_r p_s} ||\gamma^{rs}||_2 \right)$$

Limits of the groupLASSO regression :

- Difficult to compute p-value or confidence interval

Coefficients estimation

Group LASSO regression

$$(\hat{\beta}, \hat{\gamma}) = \underset{\beta,\gamma}{\operatorname{argmin}} \sum_i (y_i - \boldsymbol{X}_i\beta - \boldsymbol{Z}_i\gamma)^2 +$$
$$\lambda \left(\sum_g \sqrt{p_g} ||\beta^g||_2 + \sum_{rs} \sqrt{p_r p_s} ||\gamma^{rs}||_2 \right)$$

Limits of the groupLASSO regression :

- Difficult to compute p-value or confidence interval

Adaptive-Ridge Cleaning *Becu JM, 2015*

- Use of a specific penalty for group LASSO
- Permutation test based on Fisher test approach for each group $P_k = \frac{1}{R}\#\{F_k^* \geq F_k\}$

Interaction variable modeling approaches comparison

methods	criteria	$Z_i^{rs\,T}\gamma^{rs}$
G-GEE	$cor(\boldsymbol{Y}, f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s))$	$\boldsymbol{W}^{rs}\boldsymbol{u}\gamma^{rs}$
PCA	$var(\boldsymbol{G}_r\boldsymbol{v})$ and $var(\boldsymbol{G}_s\boldsymbol{v})$	$\sum_{j=1}^{q}\sum_{k=1}^{q}\gamma_{jk}^{rs}\boldsymbol{C}_j^r\boldsymbol{C}_k^s$
CCA	$cor(\boldsymbol{G}_r\boldsymbol{a}, \boldsymbol{G}_s\boldsymbol{b})$	$\sum_{j=1}^{q}\gamma_j^{rs}\boldsymbol{A}_j^r\boldsymbol{B}_j^s$
PLS	$cov(\boldsymbol{Y}\boldsymbol{G}_r\boldsymbol{c}, \boldsymbol{G}_s\boldsymbol{w})$	$\sum_{j=1}^{q}\gamma_j^{rs}\boldsymbol{T}_j^{rs}$

Simulation design

Genotype :

$\boldsymbol{X}_i \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ a block diagonal correlation matrix
($\rho = 0.8$ for two SNPs in the same gene)

$\boldsymbol{MAF}_j \sim \mathcal{U}[0.05, 0.5]$ with $MAF_j = 0.2$ if j causal SNP

Continuous phenotype simulated under two different schemes :

➔ from Wang X et al., 2014 :

$$Y_i = \beta_0 + \sum_g \beta_g \left(\sum_{k \in \mathcal{C}} X_{ik}^g \right) + \sum_{rs} \gamma_{rs} \left(\sum_{(j,k) \in \mathcal{C}^2} X_{ij}^r X_{ik}^s \right) + \epsilon_i \quad (1)$$

➔ PCA model :

$$Y_i = \beta_0 + \sum_g \beta_g \left(\sum_{k \in \mathcal{C}} X_{ik}^g \right) + \sum_{rs} \gamma_{rs} C_{i1}^r C_{i1}^s + \epsilon_i. \quad (2)$$

Simulations design

Scenarios :

We consider 600 subjects and 6 SNPs by gene

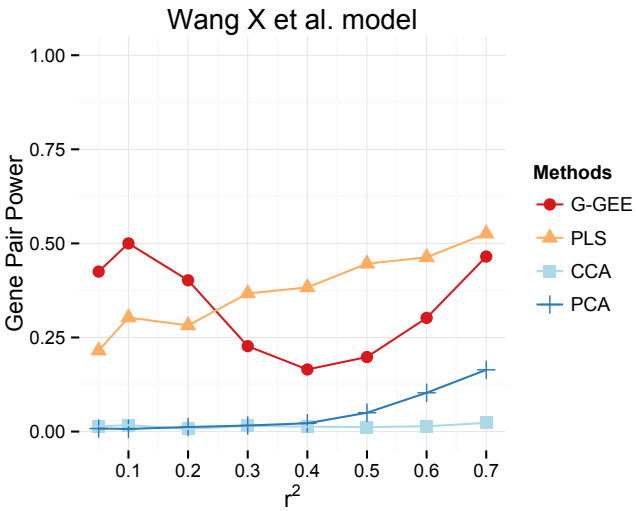
→First scenario on 6 genes, two settings :

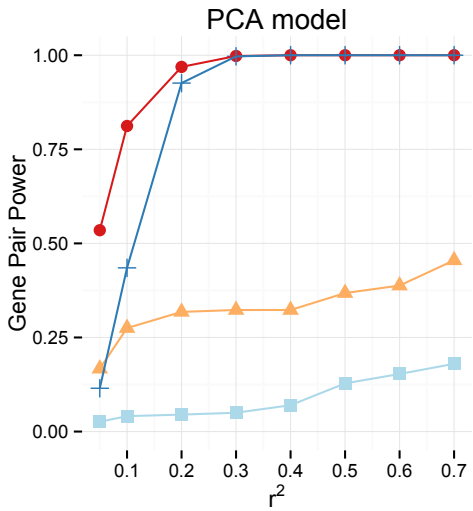
- same genes for main and interaction effects,
- different genes for main and interaction effects.

→Second scenario on 25 genes, one setting :

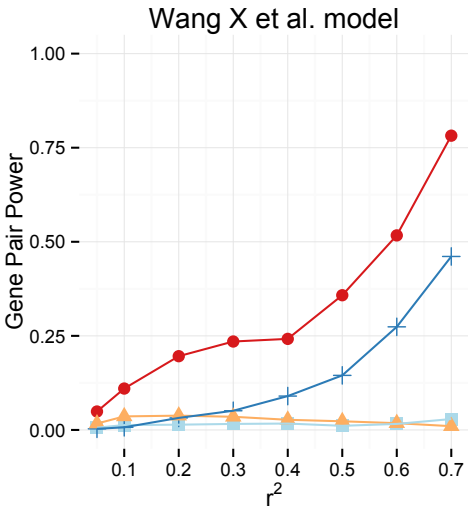
- different genes for main and interaction effects.

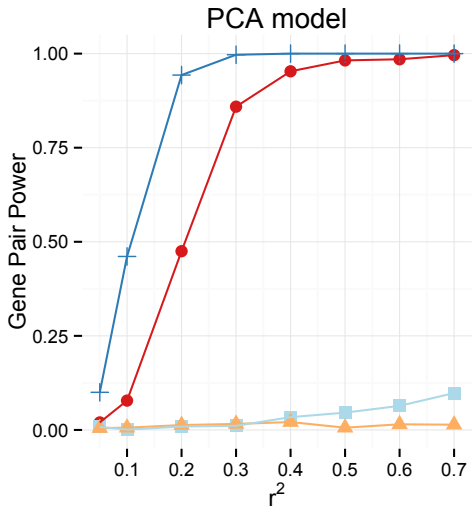
Simulations results - First scenario on 6 genes





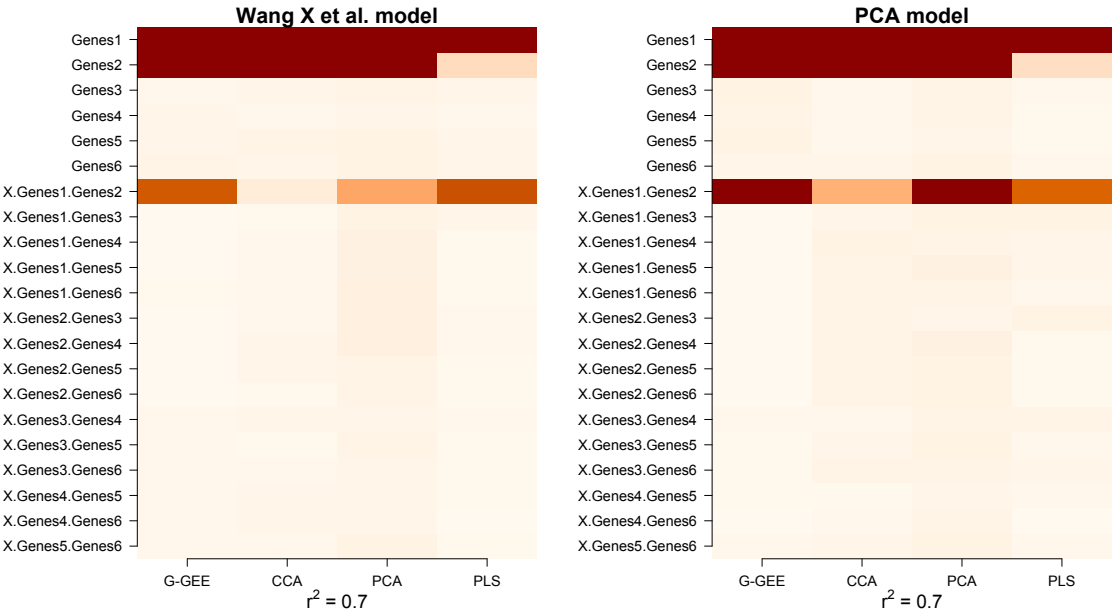
➔ Main effects :
gene 1
gene 2
➔ Interaction effects :
gene 1 x gene 2

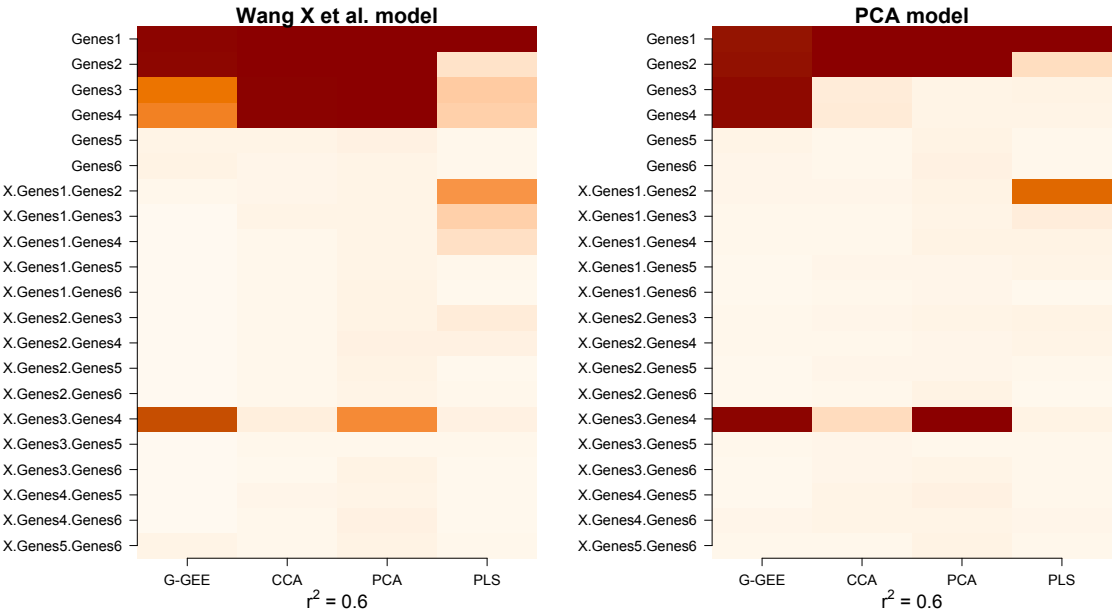




➔ Main effects :
gene 1
gene 2
➔ Interaction effects :
gene 3 x gene 4

Simulations results - First scenario on 6 genes

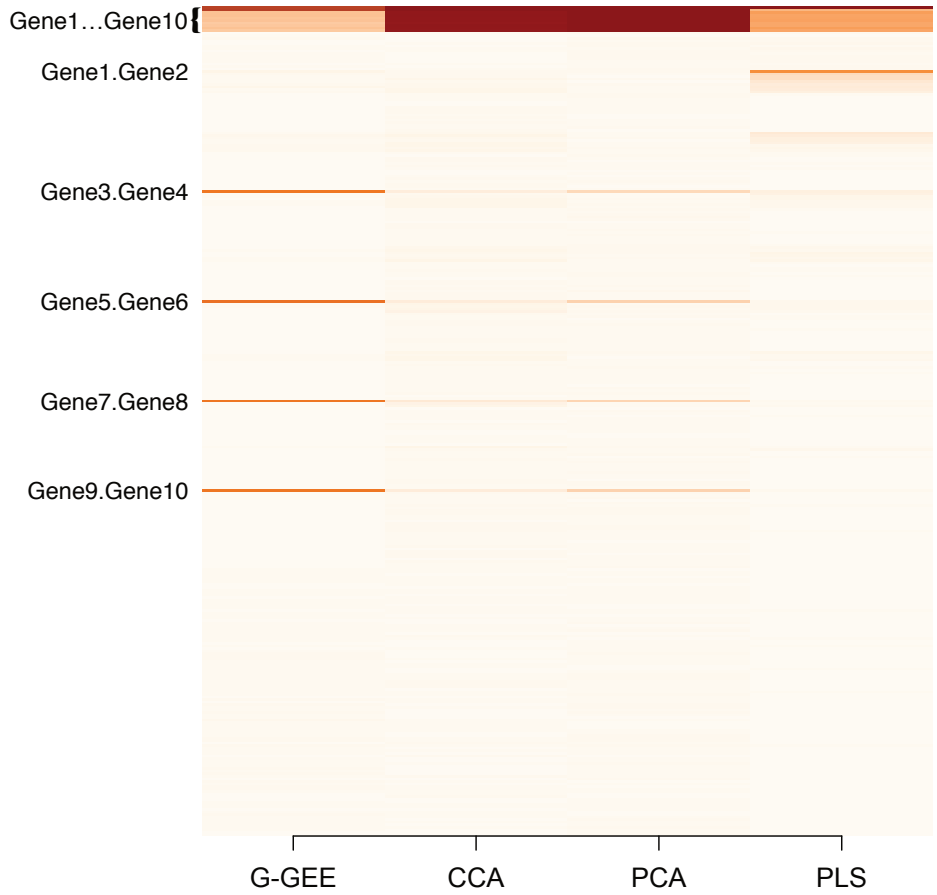




- Main effects :
  - gene 1
  - gene 2
- Interaction effects :
  - gene 1 x gene 2

- Main effects :
  - gene 1
  - gene 2
- Interaction effects :
  - gene 3 x gene 4

Simulations results - Second scenario on 25 genes



➔ Main effects :

gene 1

gene 2

➔ Interaction effects :

gene 3 x gene 4

gene 5 x gene 6

gene 7 x gene 8

gene 9 x gene 10

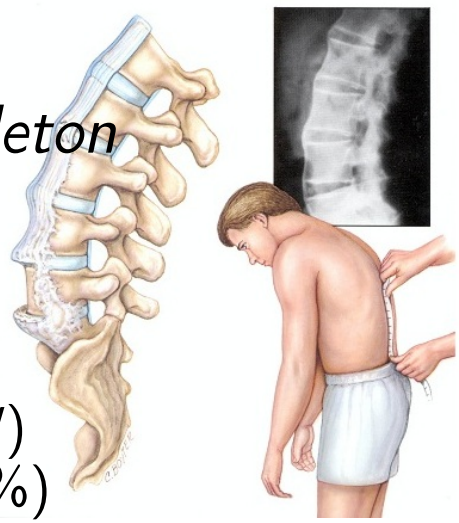
Figure: Wang X et al. model, $r^2 = 0.7$

Sommaire

1. GWAS and Block of linkage desequilibrium
   - Genome Wide Association Studies
   - Blocks of linkage desiquilibrium
   - Hierachical Clustering with Adjacency Constraints
   - How to improve ?
   - Some computation times
2. Epistasis
   - The Gene-Gene Eigen Epistasis Modeling approach
   - Simulations
3. Application
   - Ankylosing Spondylitis
   - First results

Ankylosing Spondylitis

Chronic inflammatory disease of the axial skeleton



Epidemiology :

- Age at first symptoms : 20 - 30 years
- Sexe : predominance for men (sex ratio 2M :1W)
- Prevalence : depend of populations (0.1% - 1.4%)

Right etiology unknown :

- Environmental factors ?
- Genetic factors ?
  ➔ Importance of HLA complex

HLA complex :

- Localized on chromosome 6
- Regroup about 200 genes
- Coding the immunity system
- Antigen HLA-B27 : associated to SPA

➔ Effect from other gene in HLA group ?

➔ Outside HLA group ?

Known genes

Table I Summary of ankylosing spondylitis-susceptibility genes identified by genome-wide association studies

RUNX3	Runt-related transcription factor 3
IL23R	Interleukin 23 receptor
IL12Rβ2	Interleukin 12 receptor, β2
GRP25	G-protein-coupled receptor 25
KIF21B	Kinesin family member 21B
PTGER4	Prostaglandin E receptor 4 (subtype EP_4)
ERAP1	Endoplasmic reticulum aminopeptidase 1
ERAP2	Endoplasmic reticulum aminopeptidase 2
LNPEP	Leucyl/cystinyl aminopeptidase
IL12B	Interleukin 12B
CARD9	Caspase recruitment-domain family member 9
LTβR	Lymphotoxin β-receptor (TNFR superfamily, member 3)
TNFRSF1A	Tumor-necrosis factor-receptor superfamily member 1A
NPEPPS	Aminopeptidase puromycin-sensitive
TBκBP1	TNFR-associated factor family member-associated nuclear factor-κB-binding kinase 1-binding protein
TBX21	T-box 21
IL6R	Interleukin 6 receptor
FCGR2A	Fc fragment of immunoglobulin G, low-affinity IIa, receptor (CD32)
UBE2E3	Ubiquitin-conjugating enzyme E2E 3
GPR35	G-protein-coupled receptor 35
NKX2-3	NK2 homeobox 3
ZMIZ1	Zinc finger, MIZ type-containing 1
SH2B3	Src homology 2B adaptor protein 3
GPR65	G-protein-coupled receptor 65
IL27	Interleukin 27
SULT1A1	Sulfotransferase family cytosolic 1A
TYK2	Tyrosine kinase 2
ICOSLG	Inducible T-cell costimulator ligand
EOMES	Eomesodermin
IL7R	Interleukin 7 receptor
BACH2	BTB and CNC homology 1, basic leucine-zipper transcription-factor 2

Abbreviation: CD, classification determinant.

Tsui et al., 2014 : The genetic basis of ankylosing spondylitis : new insights into disease pathogenesis, The Application of Clinical Genetics :7 105-115

➔ 29 susceptibility genes identified by GWAS

Results

	Significant results
G-GEE	HLA-B x SULT1A1 IL23R x ERAP2
PLS	HLA-B EOMES x BACH2
PCA	HLA-B
CCA	-

Conclusions and perspectives

The G-GEE method :

➔ Takes into account the gene structure of data

➔ Can be applied on a large number of genes

➔ Uses a specific interaction modeling approach

Ankylosing Spondylitis :

➔ Identification of potential interactions to discuss with doctors

➔ HLA-B effect

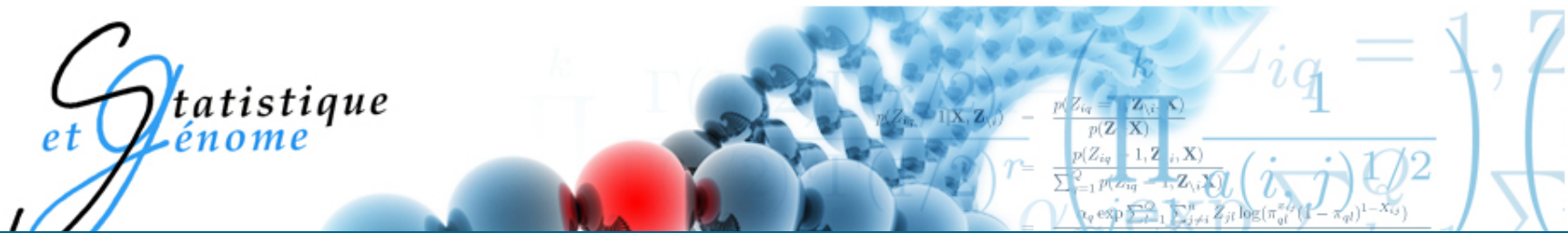
Perspectives :

➔ Explore new $f_{\boldsymbol{u}}(\boldsymbol{X}_i^r, \boldsymbol{X}_i^s)$ definition

➔ Additional simulations on larger data set

➔ New applications on other data set

Thank you for your attention !



→ **Adaptive-Ridge Cleaning**

specific penalty for group LASSO : $\dfrac{\lambda}{\sqrt{|k(j)| \sum_{m \in k(j)} \hat{\theta}_m^2}}$