Statistical approaches to detect epistasis in Genome Wide Association Studies

Virginie Stanislas
Supervisors: Cyril Dalmasso and Christophe Ambroise

Laboratoire de Mathématiques et Modélisation d'Évry

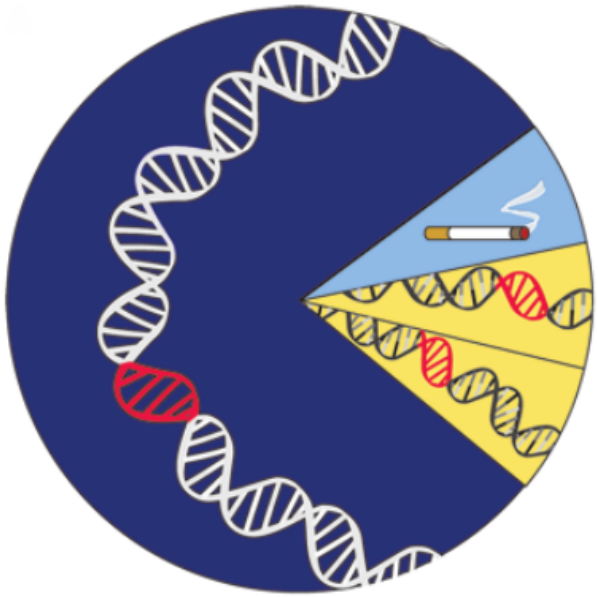
December 18th, 2017

Summary

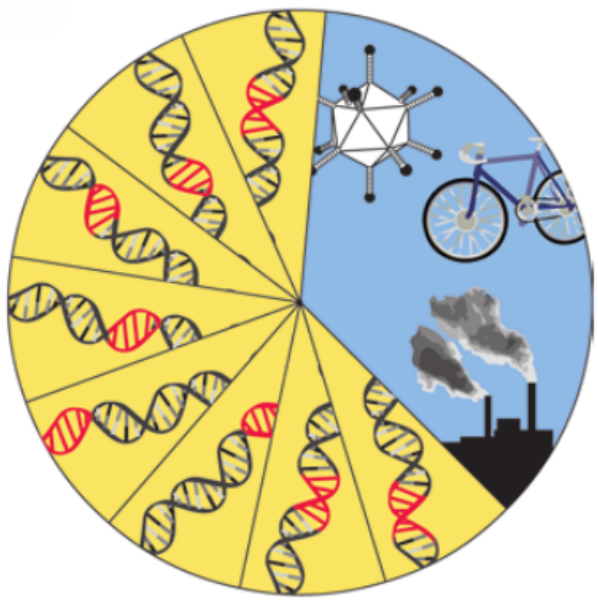
Summary

Complex diseases

Monogenic disease



Complex disease

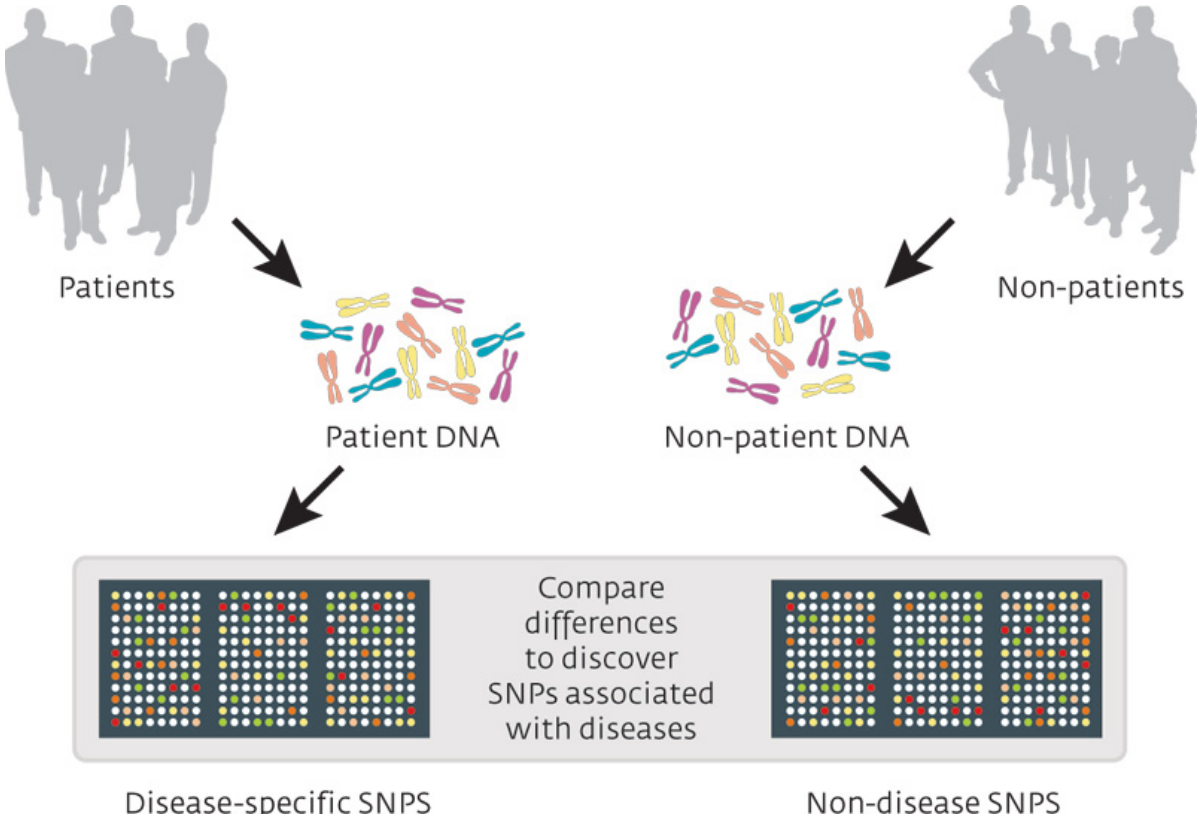


Manolio et al. J Clin Invest. 2008 ;118(5):1590-1605.

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 $Y_i \in \mathbb{R}$)

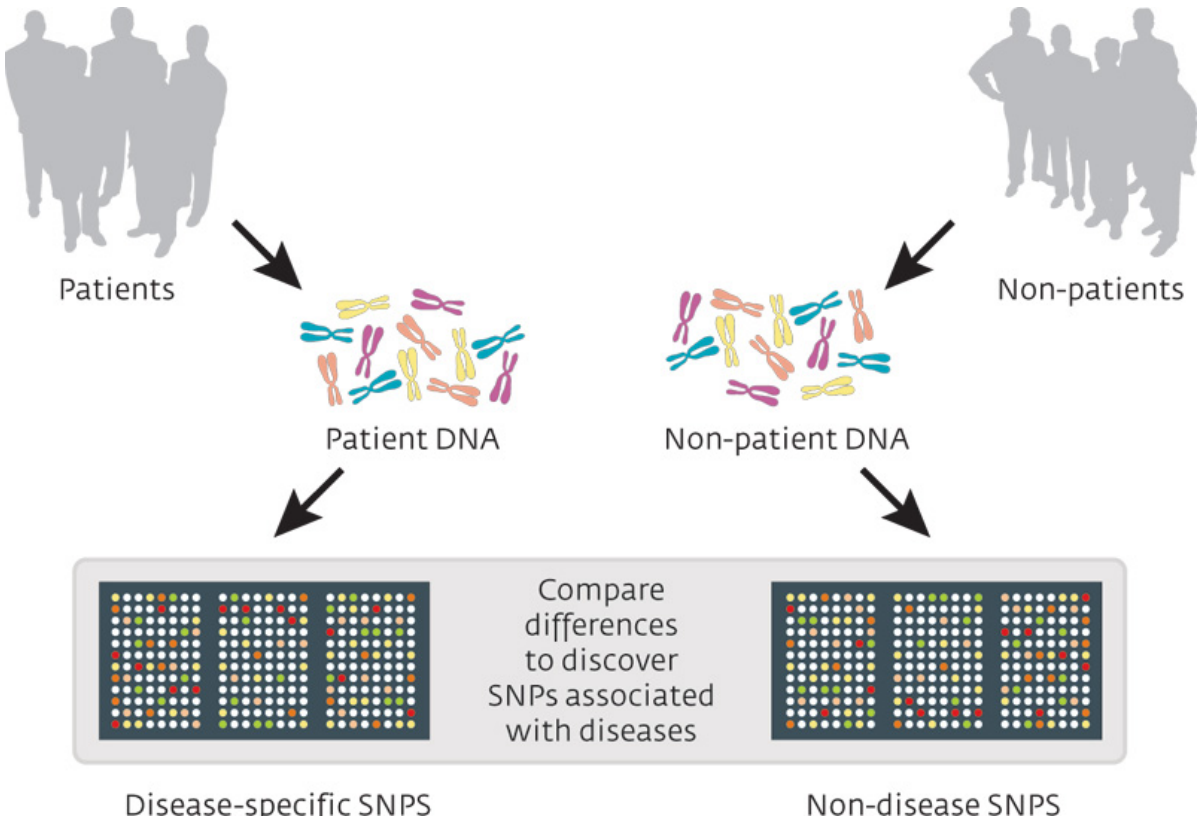


©Pasieka, Science Photo Library

Genome-Wide Association Studies

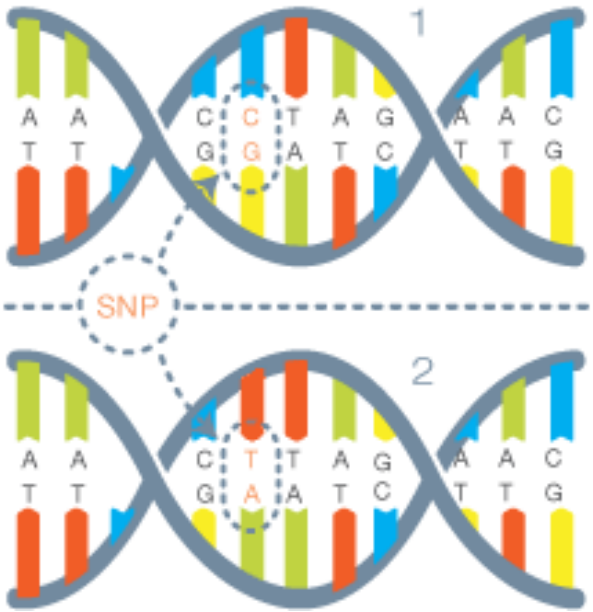
GWAS characteristics:

- **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

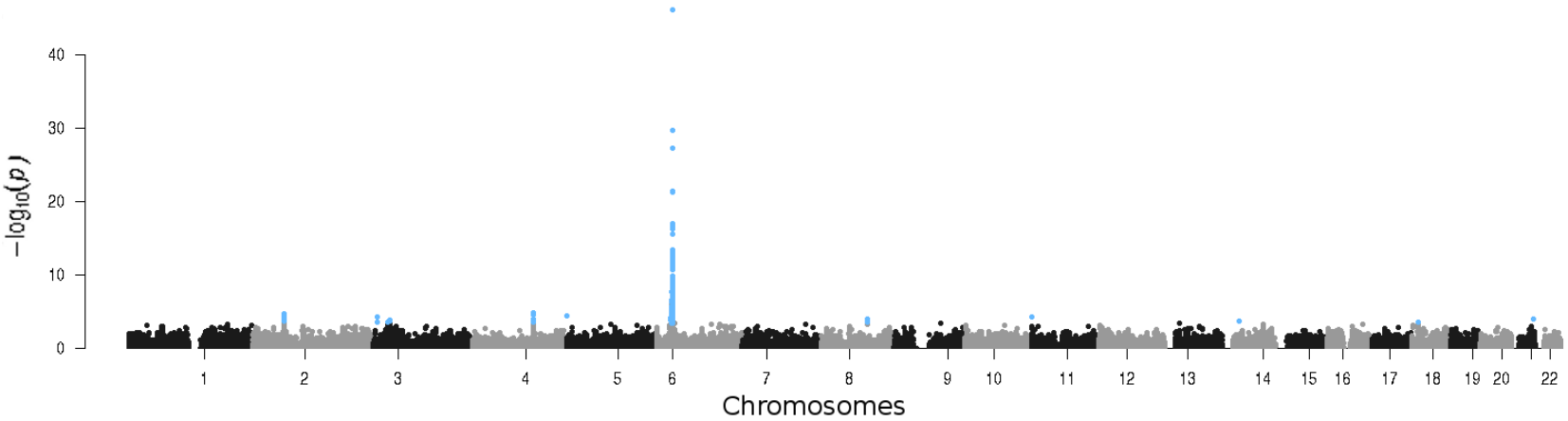


https://genomainternational.com

Genome-Wide Association Studies

SNP analysis

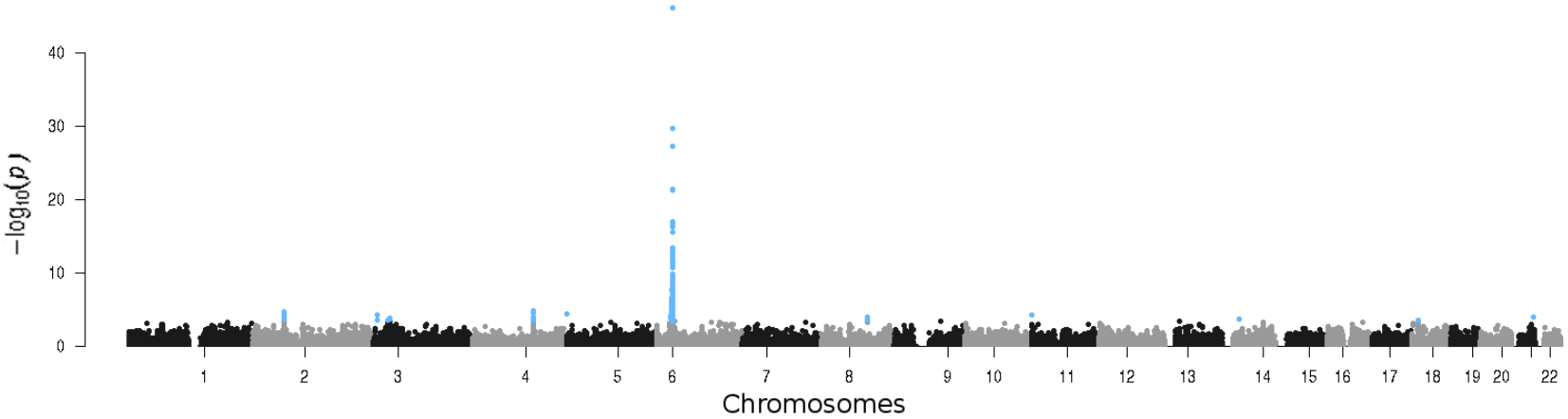
Differences between cases and controls at a specific SNP



Genome-Wide Association Studies

SNP analysis

Differences between cases and controls at a specific SNP



GWAS limits:

- Reproductibility
- Genetic factors missing

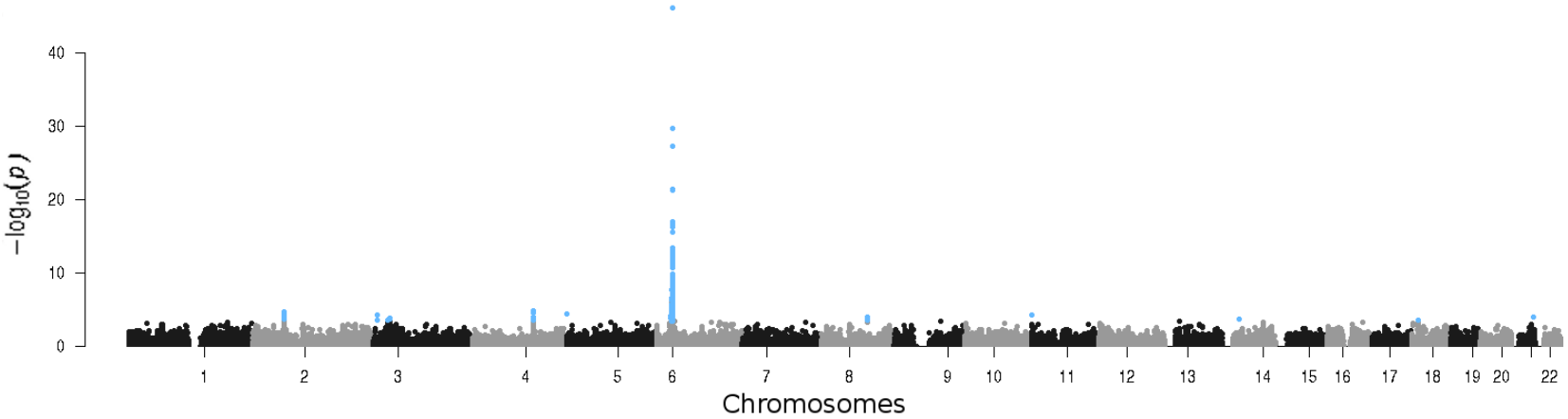
Factors:

- High dimension ($p \gg n$)
- Small effects

Genome-Wide Association Studies

SNP analysis

Differences between cases and controls at a specific SNP



GWAS limits:

- Reproductibility
- Genetic factors missing
- Missing heritability

Missing heritability factors:

- Non consideration of rare variants (MAF $< 0.1\%$)
- Non consideration of structural variants (insertion, deletion, copy numbers...)
- Incorrect estimation measure of heritability
- Complex structure of genetic data

Epistasis - Definition

Epistasis: *Interaction of alleles effects from different markers*

locus 1 \ locus 2	bb	bB	BB
aa	0	0	0
aA	0	1	1
AA	0	1	1

Epistasis - Definition

Epistasis: *Interaction of alleles effects from different markers*

locus 1 \ locus 2	bb	bB	BB
aa	0	0	0
aA	0	1	1
AA	0	1	1

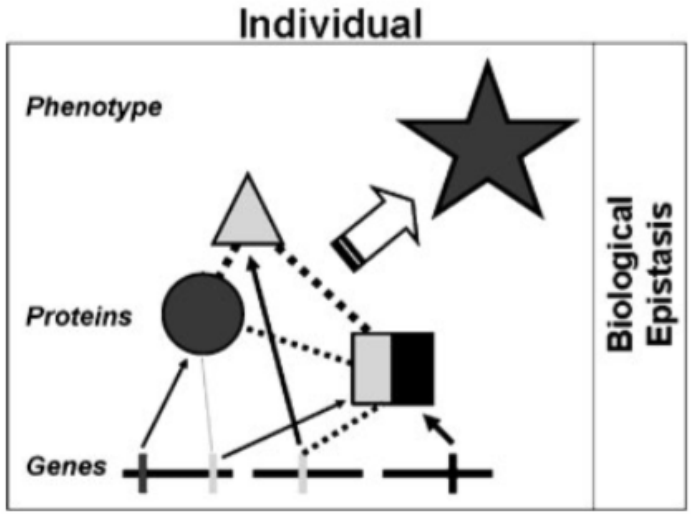
Different definitions according to disciplines with two major distinctions:

Biological epistasis

Statistical epistasis

Epistasis - Definition

Biological epistasis:
Physical interaction at the individual level

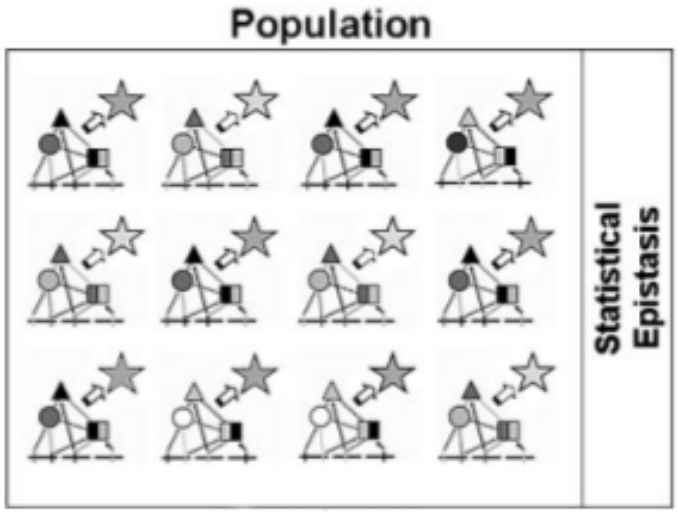


Moore & Williams Bioessays 2005 ;27(6):637-646.

Epistasis - Definition

Statistical epistasis:

Deviation from additive effects of genetic variants at the population level

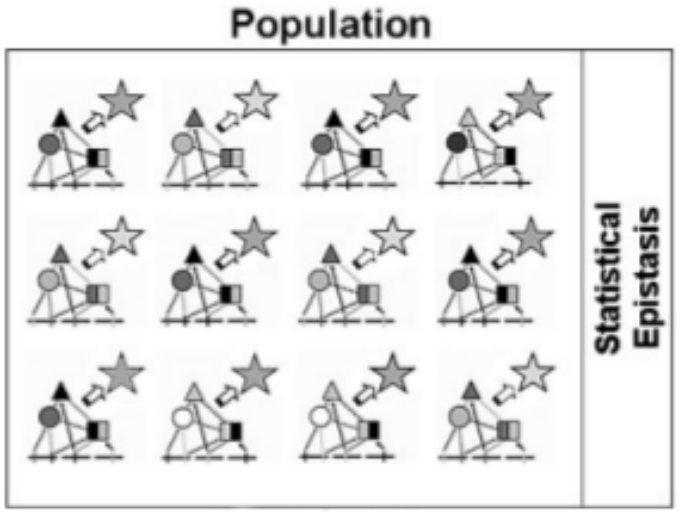


Moore & Williams Bioessays 2005 ;27(6):637-646.

Epistasis - Definition

Statistical epistasis:

Deviation from additive effects of genetic variants at the population level



Moore & Williams Bioessays 2005 ;27(6):637-646.

A possible model:

$$\text{logit}[P(\mathbf{y} = 1|\mathbf{x}_1, \mathbf{x}_2)] = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_1\mathbf{x}_2$$

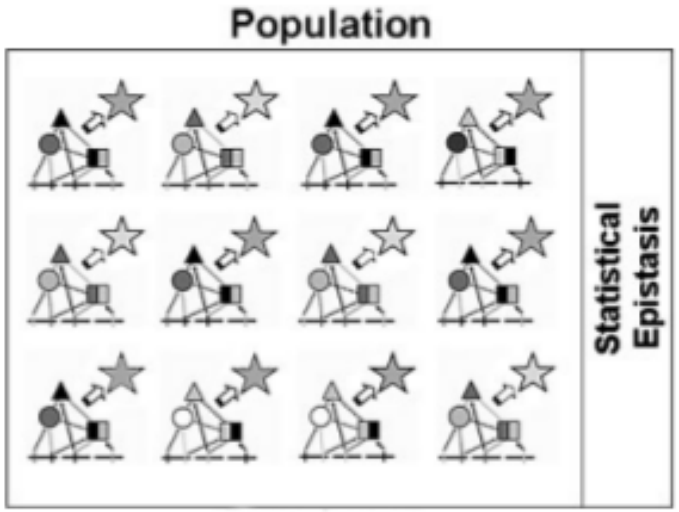
with

- $\mathbf{y}$ a binary phenotype
- $\mathbf{x}_1, \mathbf{x}_2$ the individual effect of both markers

Epistasis - Definition

Statistical epistasis:

Deviation from additive effects of genetic variants at the population level



Moore & Williams Bioessays 2005 ;27(6):637-646.

A possible model:

$$\text{logit}[P(\mathbf{y} = 1|\mathbf{x}_1, \mathbf{x}_2)] = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_1\mathbf{x}_2$$

with

- $\mathbf{y}$ a binary phenotype
- $\mathbf{x}_1, \mathbf{x}_2$ the individual effect of both markers

Epistasis - Challenges to detect it

Methodological

Epistasis - Challenges to detect it

Methodological

➔ 5×10^{11} pairwise interactions to investigate for a GWAS with 10^6 SNPs

Epistasis - Challenges to detect it

Methodological

➔ 5×10^{11} pairwise interactions to investigate for a GWAS with 10^6 SNPs

➔ Curse of dimensionality

Epistasis - Challenges to detect it

Methodological

➔ 5×10^{11} pairwise interactions to investigate for a GWAS with 10^6 SNPs

➔ Curse of dimensionality

➔ Correlation (linkage disequilibrium):

- between observed markers

Epistasis - Challenges to detect it

Methodological

➔ 5×10^{11} pairwise interactions to investigate for a GWAS with 10^6 SNPs

➔ Curse of dimensionality

➔ Correlation (linkage disequilibrium):

- between observed markers
- between observed and causal markers

Epistasis - Challenges to detect it

Methodological

➔ 5×10^{11} pairwise interactions to investigate for a GWAS with 10^6 SNPs

➔ Curse of dimensionality

➔ Correlation (linkage disequilibrium):

- between observed markers
- between observed and causal markers

➔ Distinction between marginal and interaction effects

Epistasis - Challenges to detect it

Methodological

➔ 5×10^{11} pairwise interactions to investigate for a GWAS with 10^6 SNPs

➔ Curse of dimensionality

➔ Correlation (linkage disequilibrium):

- between observed markers
- between observed and causal markers

➔ Distinction between marginal and interaction effects

Interpretation

➔ Moving from statistical estimate of epistasis to biological epistasis

Epistasis - Challenges to detect it

Methodological

➔ 5×10^{11} pairwise interactions to investigate for a GWAS with 10^6 SNPs

➔ Curse of dimensionality

➔ Correlation (linkage disequilibrium):

- between observed markers
- between observed and causal markers

➔ Distinction between marginal and interaction effects

Interpretation

➔ Moving from statistical estimate of epistasis to biological epistasis

Epistasis is ubiquitous in human biology
Investigation indispensable to understand genetic data

Epistasis - Challenges to detect it

Methodological

➔ 5×10^{11} pairwise interactions to investigate for a GWAS with 10^6 SNPs

➔ Curse of dimensionality

➔ Correlation (linkage disequilibrium):

- between observed markers
- between observed and causal markers

➔ Distinction between marginal and interaction effects

Interpretation

➔ Moving from statistical estimate of epistasis to biological epistasis

Epistasis is ubiquitous in human biology
Investigation indispensable to understand genetic data

Large number of approaches proposed

Epistasis - A variety of methods

Epistasis - Scale of interactions

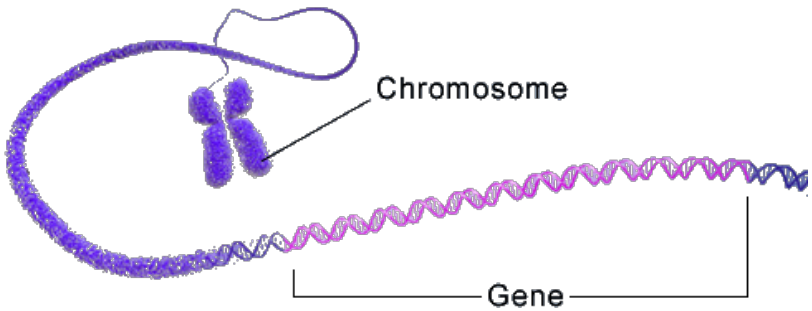
Existing methods: ➔ mainly SNP x SNP
➔ some at a group scale

Group definition:

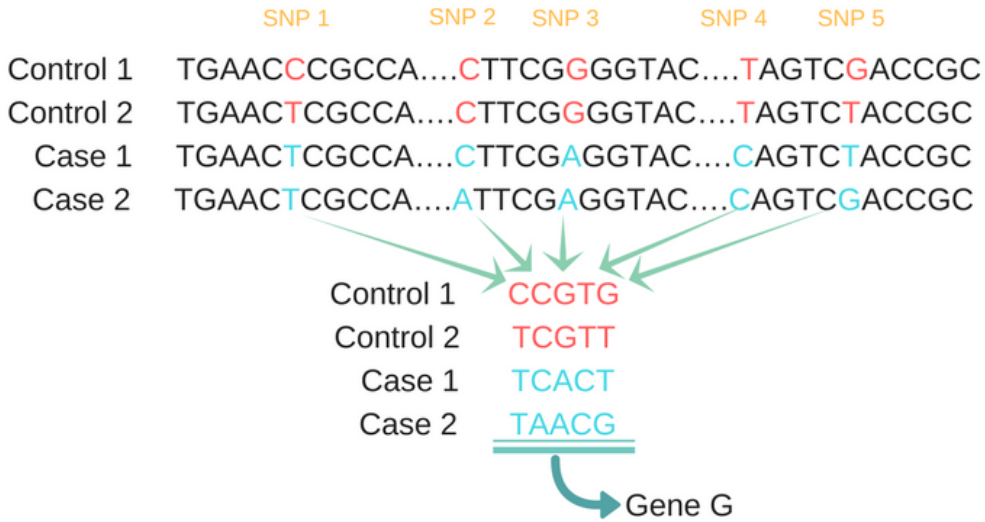
➔ genes
➔ haplotypes
➔ ...

Advantages of group scale approaches:

➔ genetic effects more detectable
➔ reduce the number of variables
➔ consideration of the correlation
➔ results biologically interpretable



U.S. National Library of Medicine



Epistasis - Gene scale methods

Gene level test outside a regression framework:

- Aggregating interaction tests
- Co-association tests

Epistasis - Gene scale methods

Gene level test outside a regression framework:

- Aggregating interaction tests
- Co-association tests

Gene level regression based approaches:

$$\left.\begin{array}{l} PCA \\ PLS \\ Kernel \end{array}\right\} + \text{logistic regression} \textit{(He 2011, Li 2009, Zhang 2008, Wang T 2009)}$$

Epistasis - Gene scale methods

Gene level test outside a regression framework:

- Aggregating interaction tests
- Co-association tests

Gene level regression based approaches:

$$\left.\begin{array}{l} PCA \\ PLS \\ Kernel \end{array}\right\} + \text{penalized regression} \textit{(D'Angelo 2009, Wang X 2014)}$$

Epistasis - Gene scale methods

Gene level test outside a regression framework:

- Aggregating interaction tests
- Co-association tests

Gene level regression based approaches:

$$\left.\begin{array}{l} PCA \\ PLS \\ Kernel \end{array}\right\} + \text{penalized regression} \textit{(D'Angelo 2009, Wang X 2014)}$$

Objectives: To develop a new gene scale method that:

➔ considers a more accurate definition of interaction variables,
➔ is applicable to numerous genes,
➔ resorts to a group penalty

Summary

Group modeling approach

	$SNP_{1,1}$	..	SNP_{1,p_1}	..	$SNP_{G,1}$	..	SNP_{G,p_G}	Pheno
Ind_1	1		0		0		1	y_1
Ind_2	0		0		2		1	y_2
.	2		1		1		2	.
.	0		1		0		0	.
Ind_i	0		2		1		0	y_i
	$\underbrace{\quad}_{gene_1}$				$\underbrace{\quad}_{gene_G}$			

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

model:

$$g(E[\boldsymbol{y}|\boldsymbol{X}]) = \underbrace{\sum_{g}\sum_{p_g}\beta_{g,p_g}\boldsymbol{X}_{g,p_g}}_{\text{Main effects}}$$

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

Group modeling approach

	$SNP_{1,1}$	..	SNP_{1,p_1}	..	$SNP_{G,1}$	..	SNP_{G,p_G}	Pheno
Ind_1	1		0		0		1	y_1
Ind_2	0		0		2		1	y_2
.	2		1		1		2	.
.	0		1		0		0	.
Ind_i	0		2		1		0	y_i
	$\underbrace{\quad}_{gene_1}$				$\underbrace{\quad}_{gene_G}$			

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

r, s two genes

model:

$$g(E[\boldsymbol{y}|\boldsymbol{X}]) = \underbrace{\sum_{g}\sum_{p_g}\beta_{g,p_g}\boldsymbol{X}_{g,p_g}}_{\text{Main effects}} + \underbrace{\sum_{r,s}\gamma_{r,s}\boldsymbol{Z}_{r,s}}_{\text{Interaction effects}}$$

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

q: # of interaction variables for a couple

Interaction variables construction:

Based on literature proposal:

methods	criteria	interaction term
Principal Component analysis (PCA)	$\mathrm{var}(\boldsymbol{X}_r\boldsymbol{v})$ and $\mathrm{var}(\boldsymbol{X}_s\boldsymbol{v})$	$\sum_{j=1}^{q}\sum_{k=1}^{q}\gamma_{jk}^{rs}\boldsymbol{T}_j^r\boldsymbol{T}_k^s$
Partial Least Square (PLS)	$\mathrm{cov}^2(\boldsymbol{Y}\boldsymbol{X}_r\boldsymbol{c},\boldsymbol{X}_s\boldsymbol{w})$	$\sum_{j=1}^{q}\gamma_j^{rs}\boldsymbol{T}_j^{rs}$
Canonical Correlation Analysis (CCA)	$\mathrm{cor}(\boldsymbol{X}_r\boldsymbol{a},\boldsymbol{X}_s\boldsymbol{b})$	$\sum_{j=1}^{q}\gamma_j^{rs}\boldsymbol{U}_j^r\boldsymbol{V}_j^s$

Interaction variables construction:

Based on literature proposal:

methods	criteria	interaction term
Principal Component analysis (PCA)	$\mathrm{var}(\boldsymbol{X}_r\boldsymbol{v})$ and $\mathrm{var}(\boldsymbol{X}_s\boldsymbol{v})$	$\sum_{j=1}^{q}\sum_{k=1}^{q}\gamma_{jk}^{rs}\boldsymbol{T}_j^r\boldsymbol{T}_k^s$
Partial Least Square (PLS)	$\mathrm{cov}^2(\boldsymbol{Y}\boldsymbol{X}_r\boldsymbol{c},\boldsymbol{X}_s\boldsymbol{w})$	$\sum_{j=1}^{q}\gamma_j^{rs}\boldsymbol{T}_j^{rs}$
Canonical Correlation Analysis (CCA)	$\mathrm{cor}(\boldsymbol{X}_r\boldsymbol{a},\boldsymbol{X}_s\boldsymbol{b})$	$\sum_{j=1}^{q}\gamma_j^{rs}\boldsymbol{U}_j^r\boldsymbol{V}_j^s$

Original proposition: Gene-Gene Eigen Epistasis (G-GEE)
We consider $f_{\boldsymbol{u}}(\boldsymbol{X}^r,\boldsymbol{X}^s)$ to represent the interaction between genes r,s.
We can choose $f_{\boldsymbol{u}}(\boldsymbol{X}^r,\boldsymbol{X}^s)$ following two conditions:

Interaction variables construction:

Based on literature proposal:

methods	criteria	interaction term
Principal Component analysis (PCA)	$\mathrm{var}(\boldsymbol{X}_r\boldsymbol{v})$ and $\mathrm{var}(\boldsymbol{X}_s\boldsymbol{v})$	$\sum_{j=1}^{q}\sum_{k=1}^{q}\gamma_{jk}^{rs}\boldsymbol{T}_j^r\boldsymbol{T}_k^s$
Partial Least Square (PLS)	$\mathrm{cov}^2(\boldsymbol{Y}\boldsymbol{X}_r\boldsymbol{c}, \boldsymbol{X}_s\boldsymbol{w})$	$\sum_{j=1}^{q}\gamma_j^{rs}\boldsymbol{T}_j^{rs}$
Canonical Correlation Analysis (CCA)	$\mathrm{cor}(\boldsymbol{X}_r\boldsymbol{a}, \boldsymbol{X}_s\boldsymbol{b})$	$\sum_{j=1}^{q}\gamma_j^{rs}\boldsymbol{U}_j^r\boldsymbol{V}_j^s$

Original proposition: Gene-Gene Eigen Epistasis (G-GEE)

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

We can choose $f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s)$ following two conditions:

criteria
→ $\mathrm{cov}^2((\boldsymbol{X}_r, \boldsymbol{X}_s), f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s))$
→ $\mathrm{cov}^2(\boldsymbol{y}, f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s))$

Interaction variables construction:

Based on literature proposal:

methods	criteria	interaction term
Principal Component analysis (PCA)	$\mathrm{var}(\boldsymbol{X}_r\boldsymbol{v})$ and $\mathrm{var}(\boldsymbol{X}_s\boldsymbol{v})$	$\sum_{j=1}^{q}\sum_{k=1}^{q}\gamma_{jk}^{rs}\boldsymbol{T}_j^r\boldsymbol{T}_k^s$
Partial Least Square (PLS)	$\mathrm{cov}^2(\boldsymbol{Y}\boldsymbol{X}_r\boldsymbol{c}, \boldsymbol{X}_s\boldsymbol{w})$	$\sum_{j=1}^{q}\gamma_j^{rs}\boldsymbol{T}_j^{rs}$
Canonical Correlation Analysis (CCA)	$\mathrm{cor}(\boldsymbol{X}_r\boldsymbol{a}, \boldsymbol{X}_s\boldsymbol{b})$	$\sum_{j=1}^{q}\gamma_j^{rs}\boldsymbol{U}_j^r\boldsymbol{V}_j^s$

Original proposition: Gene-Gene Eigen Epistasis (G-GEE)

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

We can choose $f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s)$ following two conditions:

criteria	methods
$\rightarrow \mathrm{cov}^2((\boldsymbol{X}_r, \boldsymbol{X}_s), f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s))$	G-GEE$_{c1}$
$\rightarrow \mathrm{cov}^2(\boldsymbol{y}, f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s))$	G-GEE$_{c2}$

Interaction variables construction: G-GEE$_{c1}$

We set: $f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s) = \boldsymbol{F}^{rs}\boldsymbol{u}$ with $\boldsymbol{F}^{rs} = \{X^r_{ij} X^s_{ik}\}^{j=1\cdots,p_r;k=1,\cdots,p_s}_{i=1\cdots n}$

$$\hat{\boldsymbol{u}} = arg \min_{\boldsymbol{u},\|\boldsymbol{u}\|=1} \hat{\mathrm{cov}}^2(\boldsymbol{X}, \boldsymbol{F}^{rs}\boldsymbol{u})$$

$$\text{with } \boldsymbol{X} = (\boldsymbol{X}^r, \boldsymbol{X}^s)$$

$$\min_{\boldsymbol{u},\|\boldsymbol{u}\|=1} ||\hat{\mathrm{cov}}[\boldsymbol{F}^{rs}\boldsymbol{u}, \boldsymbol{X}]||^2 = \min_{\boldsymbol{u},\|\boldsymbol{u}\|=1} \boldsymbol{u}^T \boldsymbol{F}^{rs\,T} \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{F}^{rs}\boldsymbol{u}$$

$\boldsymbol{u}$: eigen vector associated to the smallest eigenvalue of $\boldsymbol{F}^{rs\,T}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{F}^{rs}$

We then obtain for each couple (r, s) → $\boldsymbol{Z}^{rs} = \boldsymbol{F}^{rs}\boldsymbol{u}$

Interaction variables construction: G-GEE$_{c2}$

We set: $f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s) = \boldsymbol{F}^{rs}\boldsymbol{u}$ with $\boldsymbol{F}^{rs} = \{X^r_{ij} X^s_{ik}\}^{j=1\cdots,p_r;k=1,\cdots,p_s}_{i=1\cdots n}$

$$\hat{\boldsymbol{u}} = arg \max_{\boldsymbol{u},\|\boldsymbol{u}\|=1} \hat{\mathrm{cov}}^2(\boldsymbol{y}, \boldsymbol{F}^{rs}\boldsymbol{u})$$

$$\max_{\boldsymbol{u},\|\boldsymbol{u}\|=1} ||\hat{\mathrm{cov}}[\boldsymbol{F}^{rs}\boldsymbol{u}, \boldsymbol{y}]||^2 = \max_{\boldsymbol{u},\|\boldsymbol{u}\|=1} \boldsymbol{u}^T \boldsymbol{F}^{rs\,T} \boldsymbol{y}\boldsymbol{y}^T \boldsymbol{F}^{rs}\boldsymbol{u}$$

$\boldsymbol{u}$: eigen vector associated to the largest eigenvalue of $\boldsymbol{F}^{rs\,T}\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{F}^{rs}$

$$\boldsymbol{u} = \boldsymbol{F}^{rs\,T}\boldsymbol{y}$$

We then obtain for each couple (r, s) → $\boldsymbol{Z}^{rs} = \boldsymbol{F}^{rs}\boldsymbol{u} = \boldsymbol{F}^{rs}\boldsymbol{F}^{rs\,T}\boldsymbol{y}$

Interaction variable modeling approaches comparison

methods	criteria	interaction term
G-GEE$_{c1}$	$\mathrm{cov}^2(\boldsymbol{X}, f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s))$	$\boldsymbol{F}^{rs}\boldsymbol{u}\gamma^{rs}$
G-GEE$_{c2}$	$\mathrm{cov}^2(\boldsymbol{Y}, f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s))$	$\boldsymbol{F}^{rs}\boldsymbol{u}\gamma^{rs}$
PCA	$\mathrm{var}(\boldsymbol{X}_r\boldsymbol{v})$ and $\mathrm{var}(\boldsymbol{X}_s\boldsymbol{v})$	$\sum_{j=1}^{q}\sum_{k=1}^{q}\gamma_{jk}^{rs}\boldsymbol{T}_j^r\boldsymbol{T}_k^s$
PLS	$\mathrm{cov}^2(\boldsymbol{Y}\boldsymbol{X}_r\boldsymbol{c}, \boldsymbol{X}_s\boldsymbol{w})$	$\sum_{j=1}^{q}\gamma_j^{rs}\boldsymbol{T}_j^{rs}$
CCA	$\mathrm{cor}(\boldsymbol{X}_r\boldsymbol{a}, \boldsymbol{X}_s\boldsymbol{b})$	$\sum_{j=1}^{q}\gamma_j^{rs}\boldsymbol{U}_j^r\boldsymbol{V}_j^s$

Interaction variable modeling approaches comparison

methods	criteria	interaction term
G-GEE$_{c1}$	$\mathrm{cov}^2(\boldsymbol{X}, f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s))$	$\boldsymbol{F}^{rs}\boldsymbol{u}\gamma^{rs}$
G-GEE$_{c2}$	$\mathrm{cov}^2(\boldsymbol{Y}, f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s))$	$\boldsymbol{F}^{rs}\boldsymbol{u}\gamma^{rs}$
PCA	$\mathrm{var}(\boldsymbol{X}_r\boldsymbol{v})$ and $\mathrm{var}(\boldsymbol{X}_s\boldsymbol{v})$	$\sum_{j=1}^{q}\sum_{k=1}^{q}\gamma_{jk}^{rs}\boldsymbol{T}_j^r\boldsymbol{T}_k^s$
PLS	$\mathrm{cov}^2(\boldsymbol{Y}\boldsymbol{X}_r\boldsymbol{c}, \boldsymbol{X}_s\boldsymbol{w})$	$\sum_{j=1}^{q}\gamma_j^{rs}\boldsymbol{T}_j^{rs}$
CCA	$\mathrm{cor}(\boldsymbol{X}_r\boldsymbol{a}, \boldsymbol{X}_s\boldsymbol{b})$	$\sum_{j=1}^{q}\gamma_j^{rs}\boldsymbol{U}_j^r\boldsymbol{V}_j^s$

$$g(E[\boldsymbol{y}|\boldsymbol{X}]) = \sum_{g}\sum_{p_g}\beta_{g,p_g}\boldsymbol{X}_{g,p_g} + \sum_{r,s}\gamma_{r,s}\boldsymbol{Z}_{r,s}$$

Coefficients estimation

Group LASSO regression

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

Coefficients estimation

Group LASSO regression

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

Limits of the groupLASSO regression:

- $P(S^* \subset \hat{S}) \underset{n \to +\infty}{\longrightarrow} 1$ but $|\hat{S}| \gg |S^*|$
- Difficult to compute p-value or confidence interval

Coefficients estimation

Group LASSO regression

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

Limits of the groupLASSO regression:

- $P(S^* \subset \hat{S}) \underset{n \to +\infty}{\longrightarrow} 1$ but $|\hat{S}| \gg |S^*|$
- Difficult to compute p-value or confidence interval

Adaptive-Ridge Cleaning *Becu JM et al., 2017*

Coefficients estimation: Adaptive-Ridge Cleaning

Setting: $\boldsymbol{H\theta} = \boldsymbol{X}\beta + \boldsymbol{Z}\gamma$
Split randomly $\boldsymbol{H}$ in two subsets $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ of size $n/2$

Coefficients estimation: Adaptive-Ridge Cleaning

Setting: $\boldsymbol{H\theta} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma}$
Split randomly $\boldsymbol{H}$ in two subsets $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ of size $n/2$

First stage: Screening on $\boldsymbol{H}_1$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \left(\sum_i -\log L(y_i; \boldsymbol{H}_{1i}\boldsymbol{\theta}) + \lambda \left[\sum_g \sqrt{p_g} ||\boldsymbol{\theta}^g||_2 \right] \right)$$

Coefficients estimation: Adaptive-Ridge Cleaning

Setting: $\boldsymbol{H\theta} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma}$
Split randomly $\boldsymbol{H}$ in two subsets $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ of size $n/2$

First stage: Screening on $\boldsymbol{H}_1$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \left(\sum_i -\log L(y_i; \boldsymbol{H}_{1i}\boldsymbol{\theta}) + \lambda \left[\sum_g \sqrt{p_g} ||\boldsymbol{\theta}^g||_2 \right] \right)$$

→ $\hat{S}$: support of the selected groups

Coefficients estimation: Adaptive-Ridge Cleaning

Setting: $\boldsymbol{H\theta} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma}$
Split randomly $\boldsymbol{H}$ in two subsets $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ of size $n/2$

First stage: Screening on $\boldsymbol{H}_1$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \left(\sum_i -\log L(y_i; \boldsymbol{H}_{1i}\boldsymbol{\theta}) + \lambda \left[\sum_g \sqrt{p_g} ||\boldsymbol{\theta}^g||_2 \right] \right)$$

→ $\hat{S}$: support of the selected groups

Second stage: Cleaning on $\boldsymbol{H}_2$

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \ ; \ \theta_j = 0 \text{ if } j \notin \hat{S}}{\text{argmin}} \left(\sum_i -\log L(y_i; \boldsymbol{H}_{2i}\boldsymbol{\theta}) + \mu \left[\sum_g \sum_{j \in g} \frac{\lambda \sqrt{p_g}}{||\hat{\boldsymbol{\theta}}^g||_2} \theta_j^2 \right] \right)$$

Coefficients estimation: Adaptive-Ridge Cleaning

Significance of $\tilde{\theta}$: Permutation test based on a Fisher test approach

Coefficients estimation: Adaptive-Ridge Cleaning

Significance of $\tilde{\theta}$: Permutation test based on a Fisher test approach

$$F_g = \frac{\sum_i (y_i - \hat{y}_i^{\omega})^2 - \sum_i (y_i - \hat{y}_i^{\Omega})^2}{\sum_i (y_i - \hat{y}_i^{\Omega})^2}$$

With:
$\hat{y}^{\omega}$: predicted values obtained without the group g
$\hat{y}^{\Omega}$: predicted values using all groups $g \in \hat{S}$

Coefficients estimation: Adaptive-Ridge Cleaning

Significance of $\tilde{\theta}$: Permutation test based on a Fisher test approach

$$F_g = \frac{\sum_i (y_i - \hat{y}_i^{\omega})^2 - \sum_i (y_i - \hat{y}_i^{\Omega})^2}{\sum_i (y_i - \hat{y}_i^{\Omega})^2}$$

$$F_g^* = \frac{\sum_i (y_i - \hat{y}_i^{\omega})^2 - \sum_i (y_i - \hat{y}^{\Omega *}{}_i)^2}{\sum_i (y_i - \hat{y}^{\Omega *}{}_i)^2}$$

With:
$\hat{y}^{\omega}$: predicted values obtained without the group g
$\hat{y}^{\Omega}$: predicted values using all groups $g \in \hat{S}$
$\hat{y}^{\Omega *}$: predicted values using all groups $g \in \hat{S}$ on the matrix $\boldsymbol{H}^*$ of permuted elements for columns corresponding to group g

Coefficients estimation: Adaptive-Ridge Cleaning

Significance of $\tilde{\theta}$: Permutation test based on a Fisher test approach

$$F_g = \frac{\sum_i (y_i - \hat{y}_i^{\omega})^2 - \sum_i (y_i - \hat{y}_i^{\Omega})^2}{\sum_i (y_i - \hat{y}_i^{\Omega})^2}$$

$$F_g^* = \frac{\sum_i (y_i - \hat{y}_i^{\omega})^2 - \sum_i (y_i - \hat{y}^{\Omega *}{}_i)^2}{\sum_i (y_i - \hat{y}^{\Omega *}{}_i)^2}$$

With:
$\hat{y}^{\omega}$: predicted values obtained without the group g
$\hat{y}^{\Omega}$: predicted values using all groups $g \in \hat{S}$
$\hat{y}^{\Omega *}$: predicted values using all groups $g \in \hat{S}$ on the matrix $\boldsymbol{H}^*$ of permuted elements for columns corresponding to group g

Empirical p-values:

$$p_g = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\{F_g \leq F_g^{*b}\}}$$

with B the number of permutations

Summary

Simulations design: Genotype

Completely simulated genotype:

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

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

Genotype from real data:

From a real data set composed of 763 individuals and 63,340 SNPs structured in 7216 genes.

Simulations design: Phenotype

from Wang X et al., 2014:

$$g(E[\boldsymbol{y}_i|(\boldsymbol{X}_i, \boldsymbol{Z}_i)]) = \beta_0 + \sum_g \beta_g \left(\sum_{k \in \mathcal{C}} X_{ik}^g \right) + \sum_{rs} \gamma_{rs} \left(\sum_{(j,k) \in \mathcal{C}^2} X_{ij}^r X_{ik}^s \right)$$

PCA model:

$$g(E[\boldsymbol{y}_i|(\boldsymbol{X}_i, \boldsymbol{Z}_i)]) = \beta_0 + \sum_g \beta_g \left(\sum_{k \in \mathcal{C}} X_{ik}^g \right) + \sum_{rs} \gamma_{rs} C_{i1}^r C_{i1}^s.$$

Simulations design

Adjustment of the strength of association for continuous outcomes

- ϵ_i generated from $\mathcal{N}(0, \sigma^2)$
- σ^2 determined from R^2 coefficient

We note $\boldsymbol{H\theta} = [\boldsymbol{X}, \boldsymbol{Z}] \begin{bmatrix} \beta \\ \gamma \end{bmatrix}$, and $R^2 = \dfrac{\sum(\boldsymbol{H_i\theta} - \bar{y})^2}{\sum(\boldsymbol{H_i\theta} + \epsilon_i - \bar{y})^2}$

We can determined an expression for σ^2

$$\sigma^2 = \frac{(1 - R^2)\sum(\boldsymbol{H_i\theta} - \bar{y})^2}{R^2(n-2)}$$

Simulations studies

First comparison: PCA, PLS and CCA
Choosing the parameters

Second comparison: with G-GEE$_{c1}$ and G-GEE$_{c2}$
Using completely simulated genotype
Using genotype from a real data set

Third comparison: Case-control methods

Fourth comparison: Investigation of new interaction variable definitions

First comparison: methods issued of the literature

Design: Completely simulated genotype
Continuous phenotype from Wang X et al., 2014

Parameters:

- Correlation among SNPs ρ
- MAF values of causal SNPs
- Values of β and γ
- Number of components
- R^2
- Number of genes
- Number of SNPs by genes
- Number of causal SNPs by causal genes
- Number of subjects
- Marginal or/and interaction effects

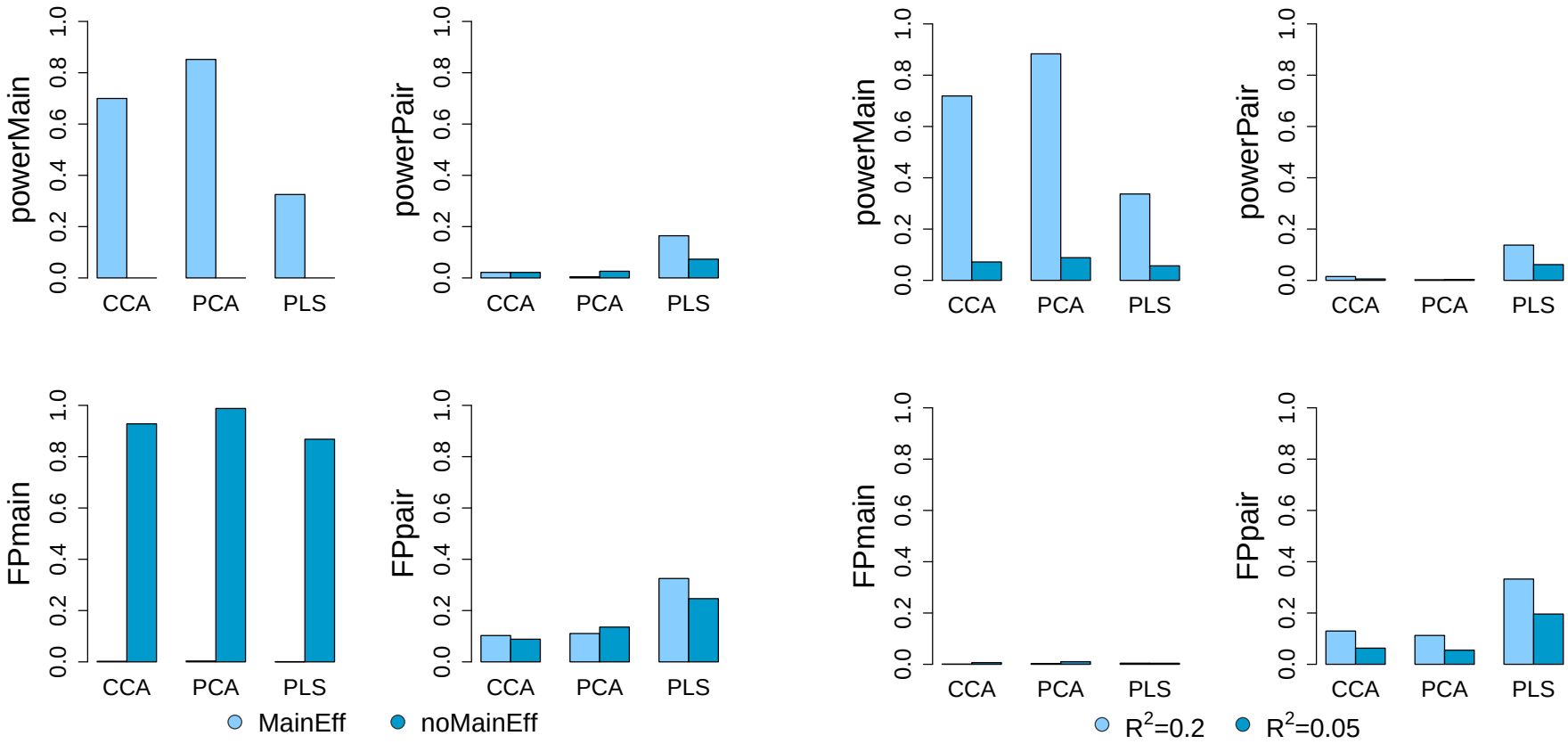
First comparison: methods issued of the literature

Design: Completely simulated genotype
Continuous phenotype from Wang X et al., 2014

Parameters:

- Correlation among SNPs ρ
- MAF values of causal SNPs
- Values of β and γ
- Number of components
- R^2
- Number of genes
- Number of SNPs by genes
- Number of causal SNPs by causal genes
- Number of subjects
- Marginal or/and interaction effects

First comparison: methods issued of the literature



First comparison: methods issued of the literature

Parameters:

- $\rho = 0.8$
- MAF $= 0.2$
- $\beta = \gamma = 2$
- Number of components =2
- R^2
- Number of genes=6
- Number of SNPs by genes=6
- Number of causal SNPs by causal genes=2
- Number of subjects=600
- Marginal or/and interaction effects

Simulations studies

First comparison: PCA, PLS and CCA
Choosing the parameters

Second comparison: with G-GEE$_{c1}$ and G-GEE$_{c2}$
Using completely simulated genotype
Using genotype from a real data set

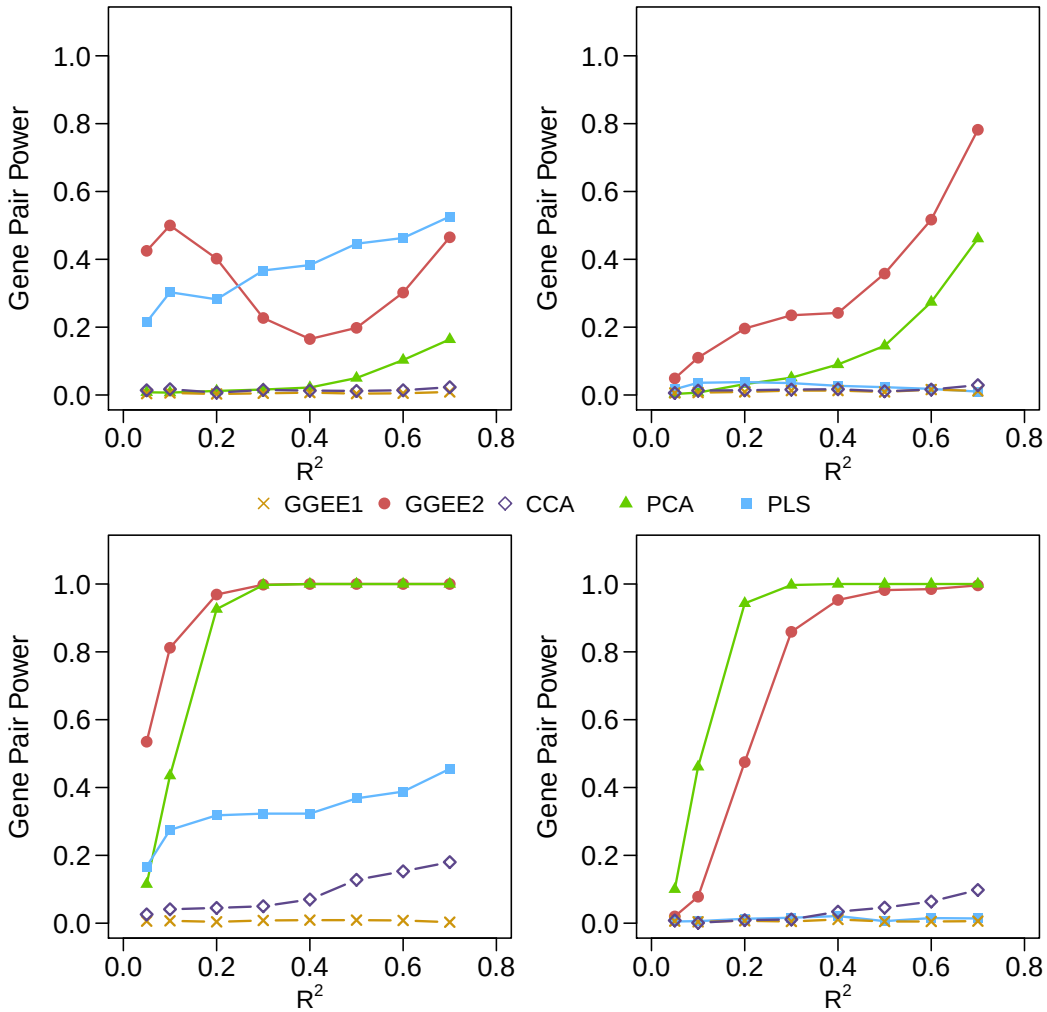
Third comparison: Case-control methods

Fourth comparison: Investigation of new interaction variable definitions

Second comparison: G-GEE and simulated genotypes

Wang model

PCA model

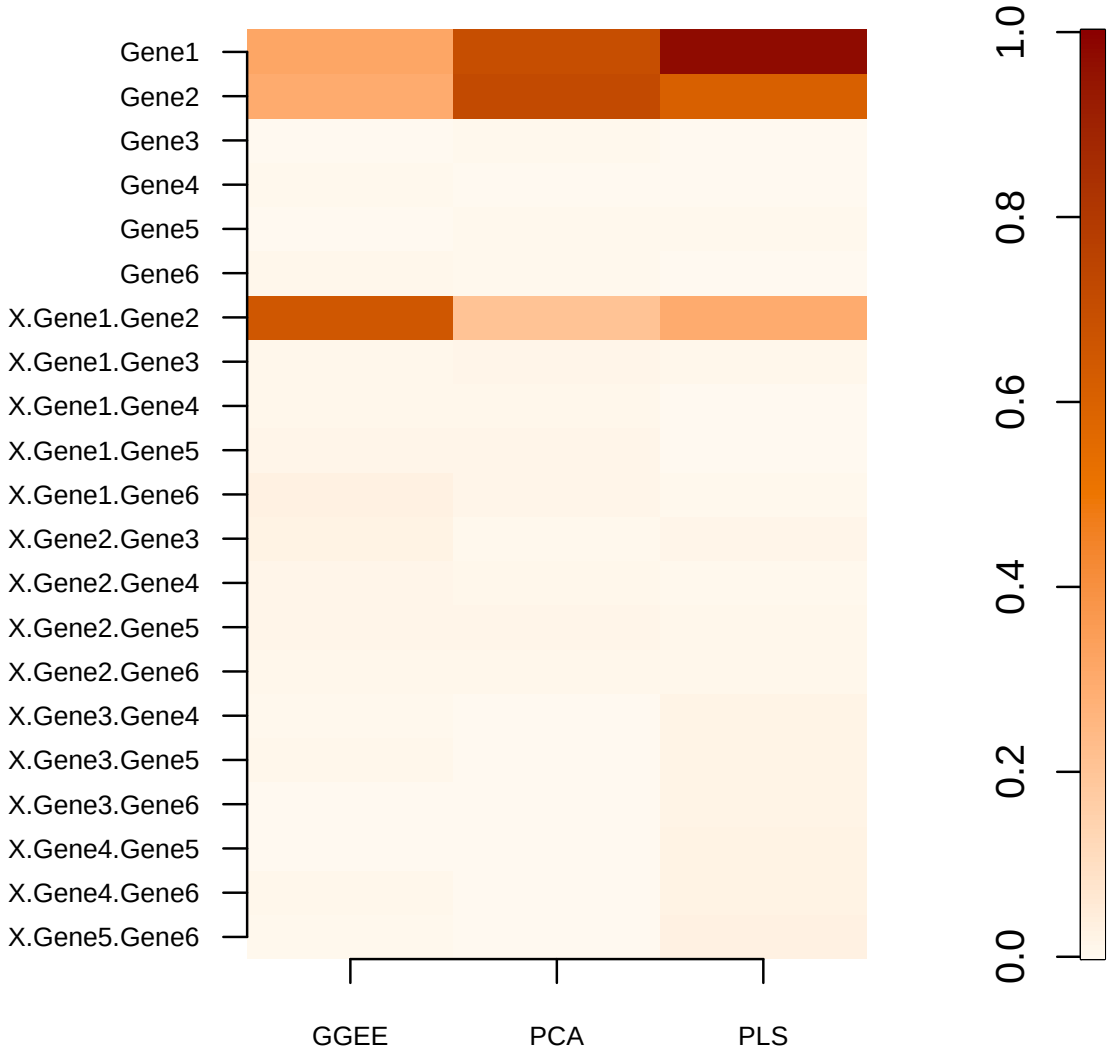


→ Main effects:
gene 1
gene 2

→ Interaction effects:
gene 1 × gene 2

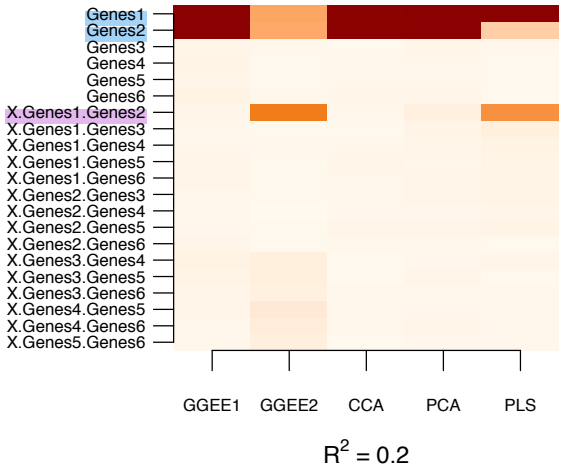
→ Interaction effects:
gene 3 × gene 4

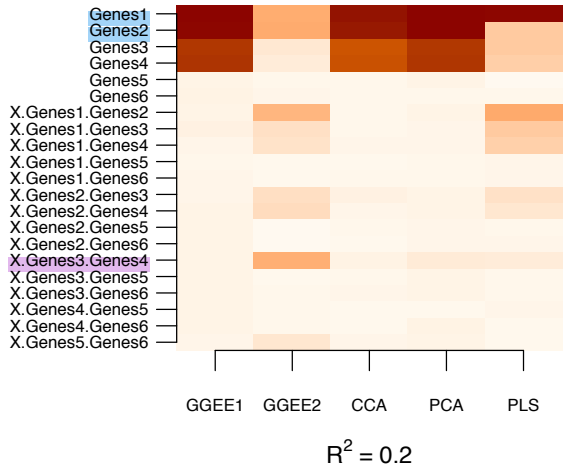
Discoveries matrix - an example



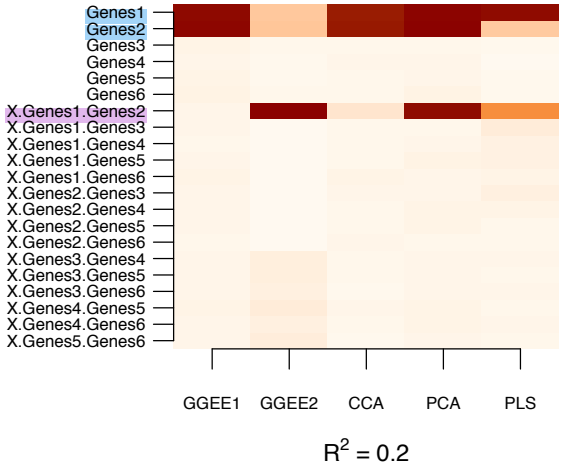
Second comparison: G-GEE and simulated Genotypes - $R^2 = 0.2$

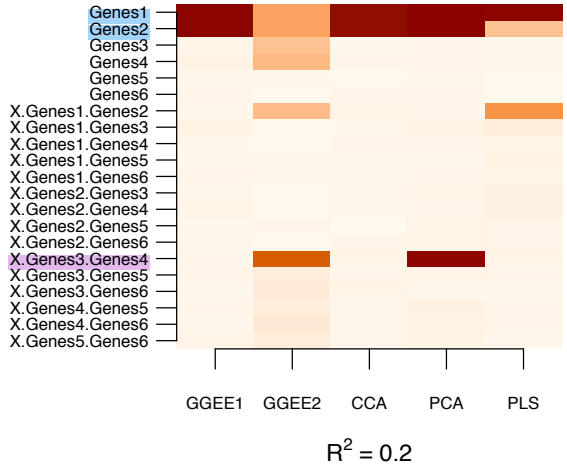
Wang model





PCA model





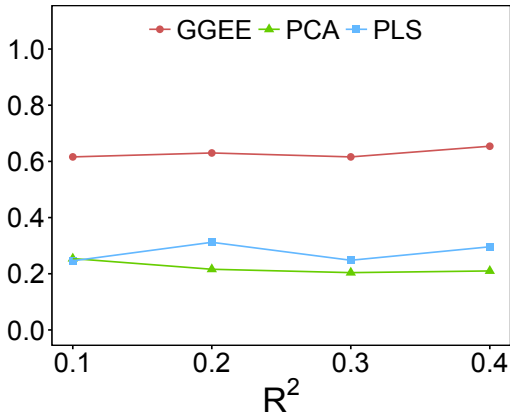
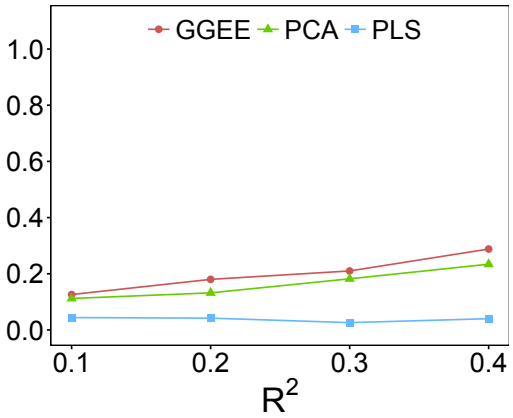
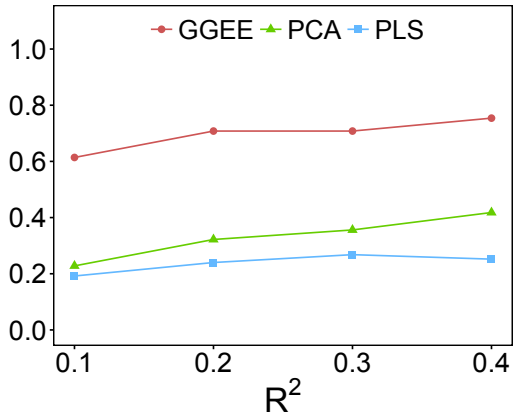
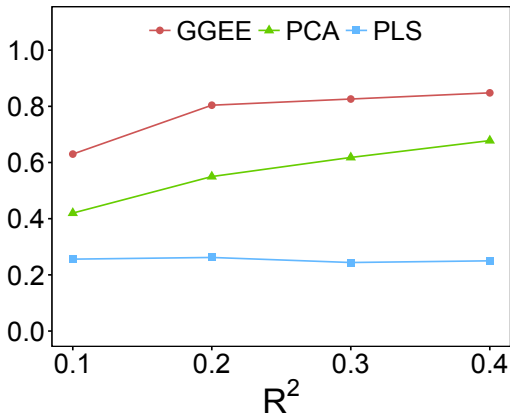
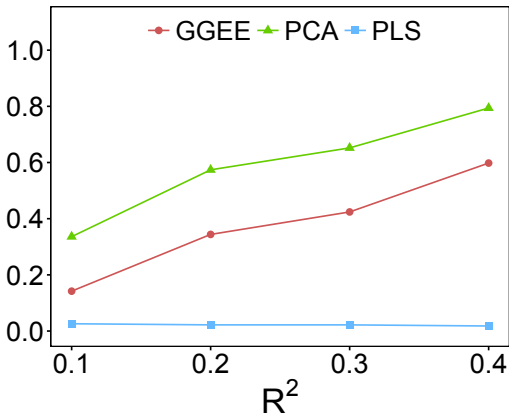
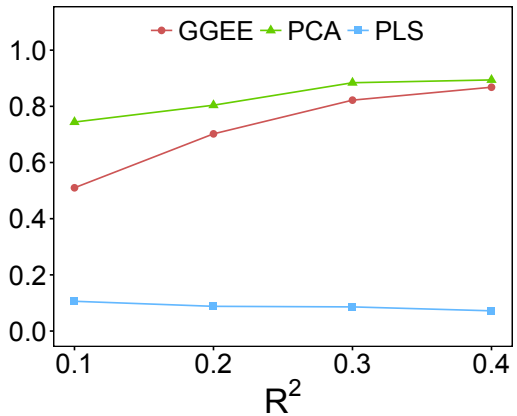
→ Main effects:
gene 1
gene 2

→ Interaction effects:
gene 1 x gene 2

→ Interaction effects:
gene 3 x gene 4

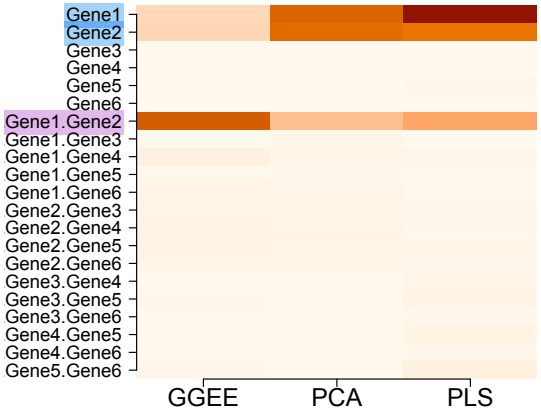
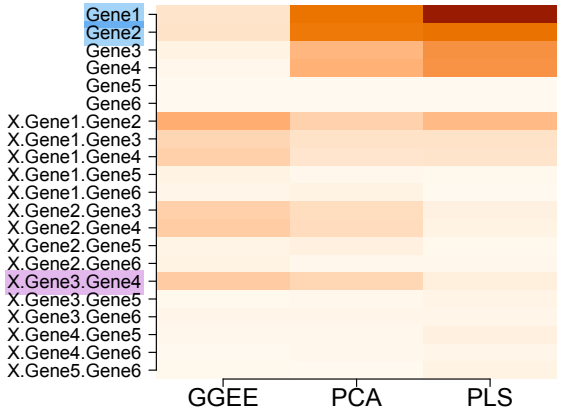
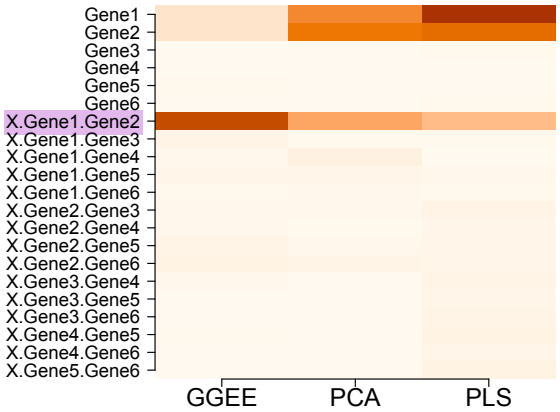
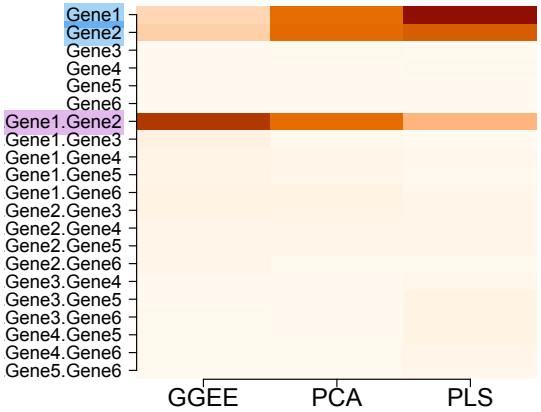
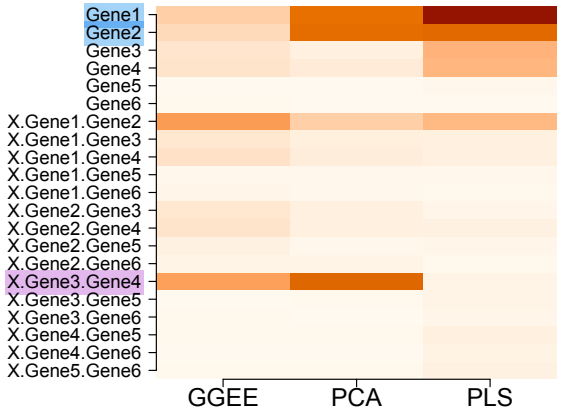
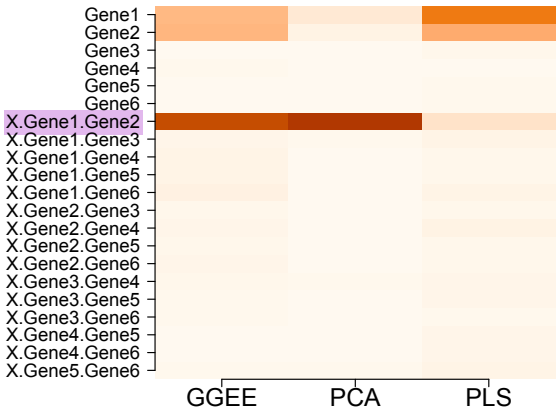
Second comparison: G-GEE and real genotypes

Settings

	Main effects: gene 1, gene 2 Interaction effects: gene 1 × gene 2	Main effects: gene 1, gene 2 Interaction effects: gene 3 × gene 4	Main effects: - Interaction effects: gene 1 × gene 2
Wang simulation model			
PCA simulation model			

Second comparison: G-GEE and real genotypes - $R^2 = 0.2$

Settings

	Main effects: gene 1, gene 2 Interaction effects: gene 1 × gene 2	Main effects: gene 1, gene 2 Interaction effects: gene 3 × gene 4	Main effects: - Interaction effects: gene 1 × gene 2
Wang simulation model			
PCA simulation model			

Simulations studies

First comparison: PCA, PLS and CCA
Choosing the parameters

Second comparison: G-GEE$_{c1}$ and G-GEE$_{c2}$
Using completely simulated genotype
Using genotype from a real data set

Third comparison: Case-control methods

Fourth comparison: Investigation of new interaction variable definitions

Third comparison: Case-control methods

Methods defined outside a regression framework

- Aggregating tests
  → minP (Emily et al. ,2016)
  → GATES (Li et al., 2011)

- Co-association test
  → PLSPM (Zhang et al., 2013)

- LD based test
  → CLD (Rajapakse et al., 2012)

- Entropy based method
  → GBIBM (Li et al., 2015)

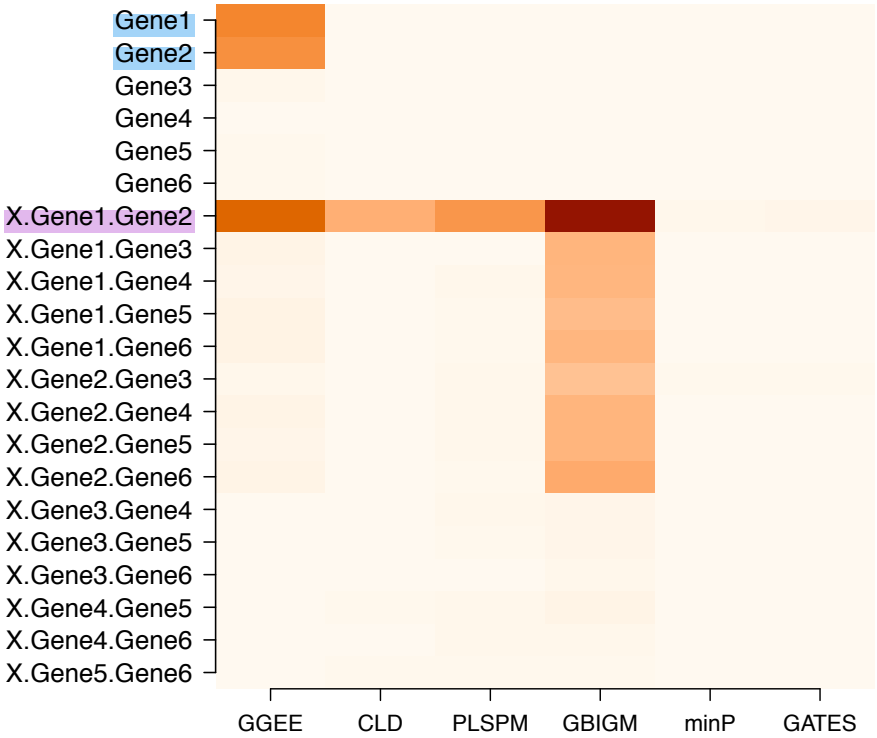
Package R: GeneGeneInteR (Emily et al. ,2017)

Third comparison: Case-control methods

Design:

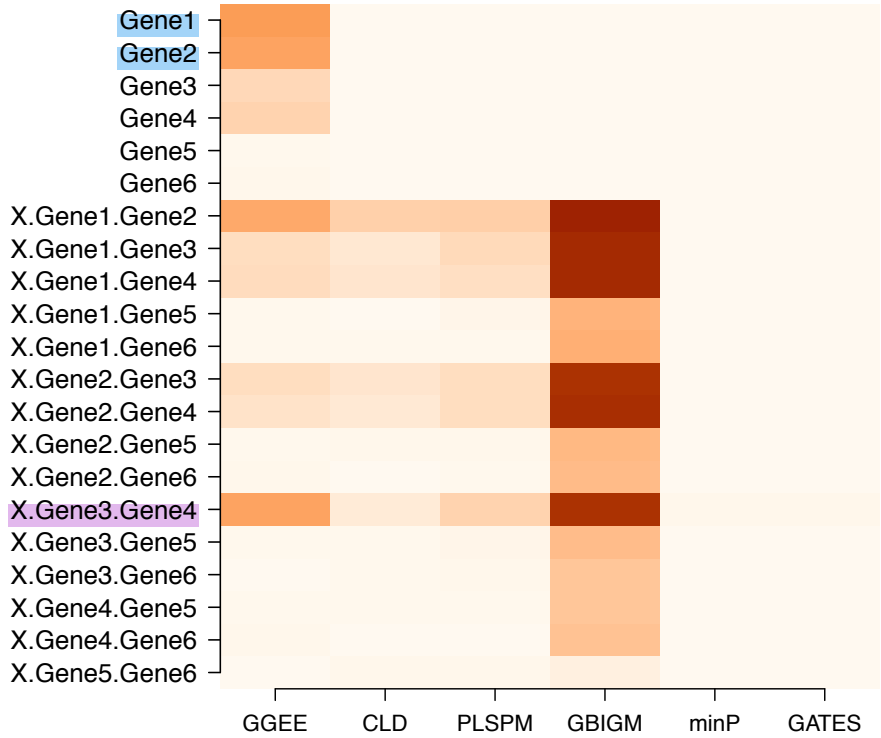
Real Genotypes

Continuous phenotype simulation from *Wang X et al., 2014*:



Main effects:
gene 1
gene 2

Interaction effects:
gene 1 × gene 2
-



Main effects:
gene 1
gene 2

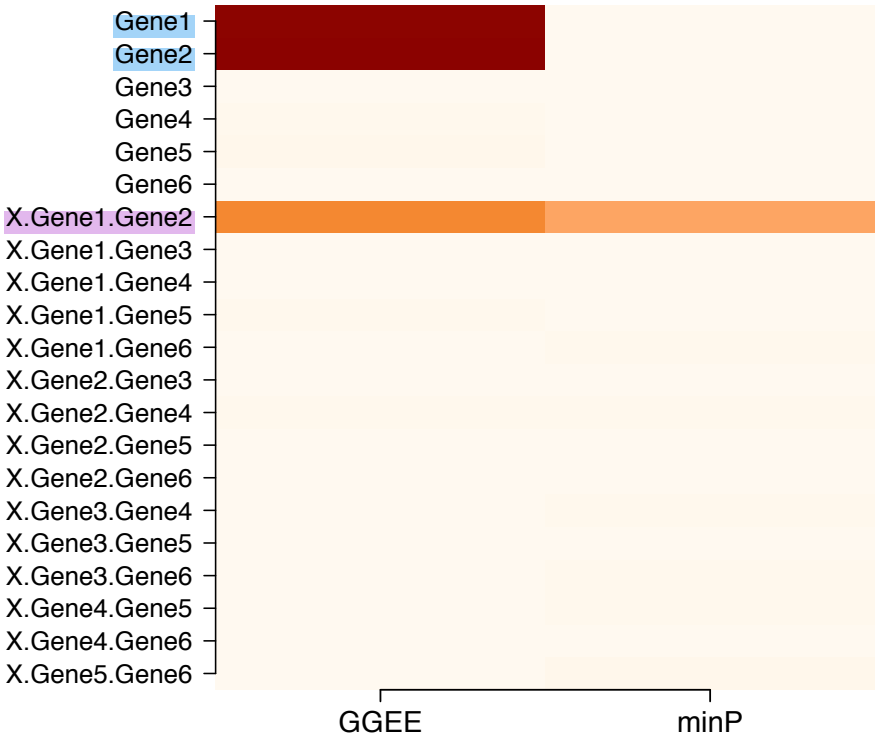
Interaction effects:
gene 3 × gene 4
-

Third comparison: Case-control methods

Design:

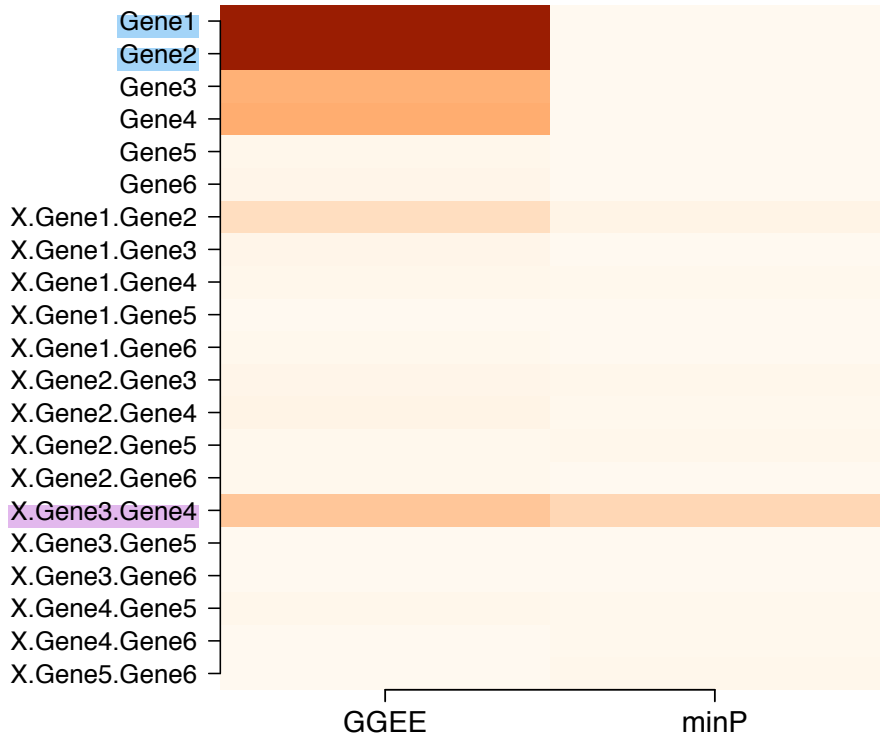
Completely simulated Genotypes

Continuous phenotype simulation from *Wang X et al., 2014*:



Main effects:
gene 1
gene 2

Interaction effects:
gene 1 × gene 2
-



Main effects:
gene 1
gene 2

Interaction effects:
gene 3 × gene 4
-

Simulations studies

First comparison: PCA, PLS and CCA
Choosing the parameters

Second comparison: with G-GEE$_{c1}$ and G-GEE$_{c2}$
Using completely simulated genotype
Using genotype from a real data set

Third comparison: Case-control methods

Fourth comparison: Investigation of new interaction variable definitions

Fourth comparison: Machine Learning based approaches

With G-GEE$_{c2}$, we looked for:

$$\hat{\boldsymbol{u}} = arg \max_{\boldsymbol{u}, \|\boldsymbol{u}\|=1} cov^2(\boldsymbol{y}, f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s))$$

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

We now find new functions $f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s)$ that maximized the criteria:

$$E_{X^r,X^s,Y}[(\boldsymbol{y} - f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s))^2]$$

With the following non parametric approaches:

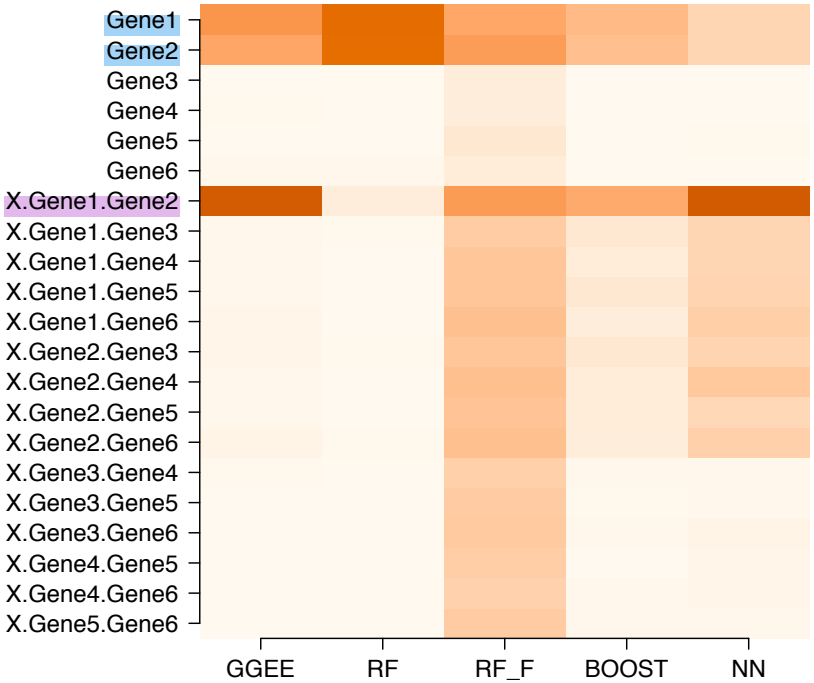
- Random Forests
- Boosting
- SVM
- Neural Network

Fourth comparison: Machine Learning based approaches - $R^2 = 0.4$

Design:

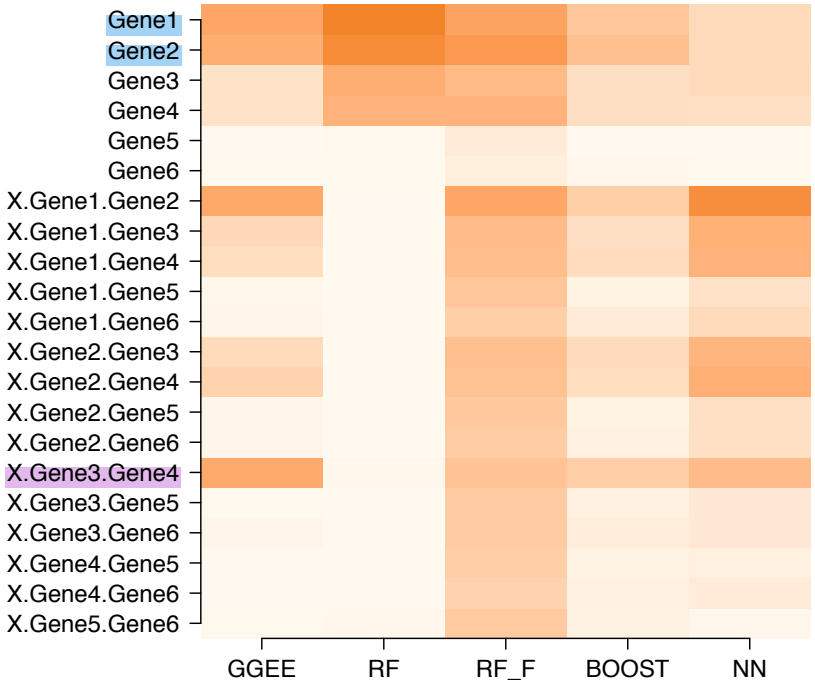
Real Genotypes

Continuous phenotype simulation from *Wang X et al., 2014*:



Main effects:
gene 1
gene 2

Interaction effects:
gene 1 x gene 2
-



Main effects:
gene 1
gene 2

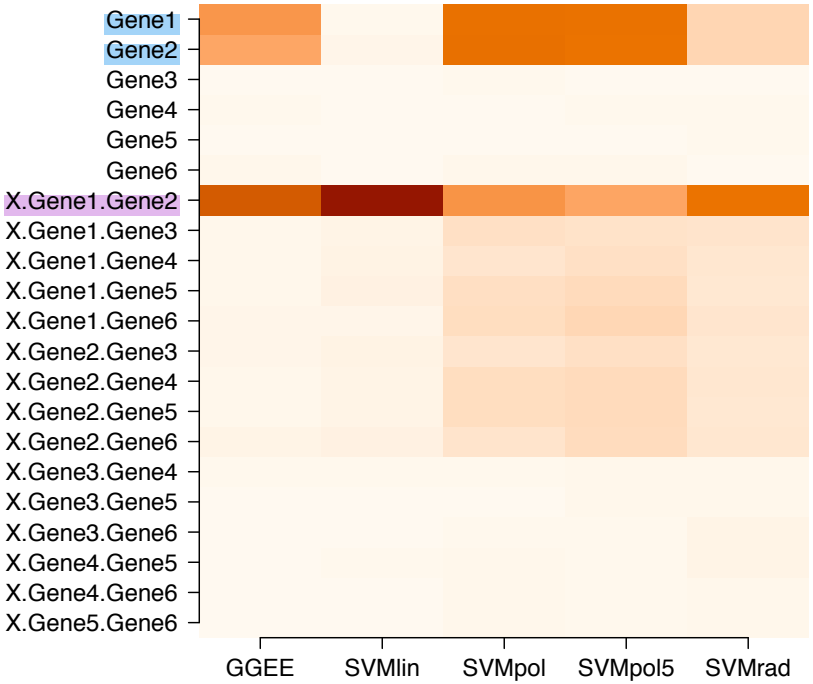
Interaction effects:
gene 3 x gene 4
-

Fourth comparison: Machine Learning based approaches - $R^2 = 0.4$

Design:

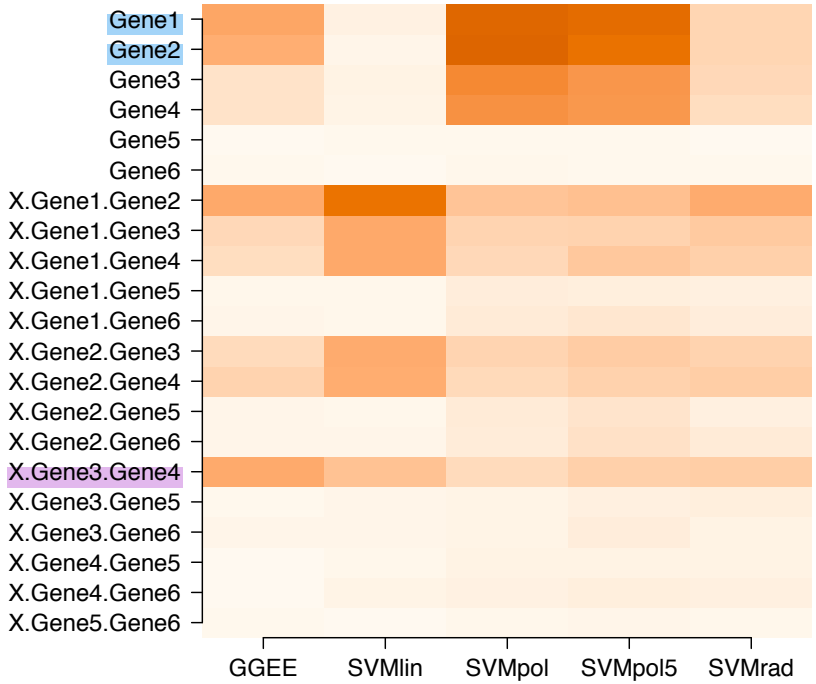
Real Genotypes

Continuous phenotype simulation from *Wang X et al., 2014*:



Main effects:
gene 1
gene 2

Interaction effects:
gene 1 × gene 2
-



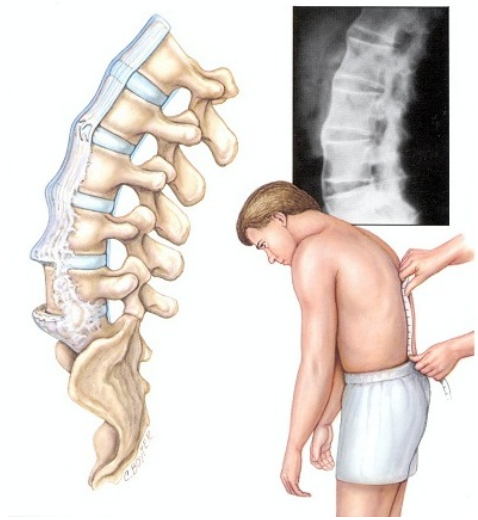
Main effects:
gene 1
gene 2

Interaction effects:
gene 3 × gene 4
-

Summary

Ankylosing Spondylitis

Chronic inflammatory disease of the axial skeleton



http://b4tea.com/

Epidemiology:

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

Risk factors:

- Strong genetic component (heritability >90%)
- Importance of HLA complex

HLA complex:

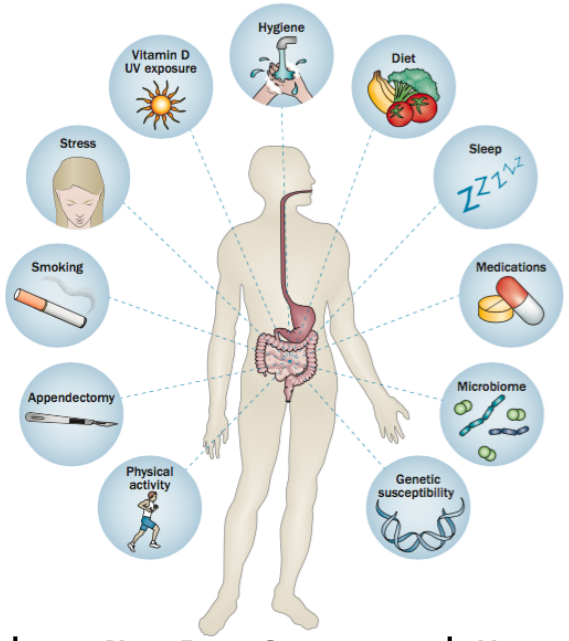
- Localized on chromosome 6
- Regroup about 200 genes
- Coding the immunity system
- Antigen HLA-B27 : associated to SPA

Crohn's Disease

Form of chronic inflammation bowel disease

Epidemiology:

- Prevalence: 10-30 per 100, 000 (Europe and North America)
- More common in the industrialized world
- Median onset of disease: 30 years



Ananthakrishnan, Nat. Rev. Gastroenterol. Hepatol 2015

Multiple risk factors:

- Environmental
- Microbiota
- Genetic

Genetic factors:

→ NOD2, first identified mutation

→ Potential interactions:

- NOD2 and TLR proteins
- NOD2 and CTLA4
- IL23R and CTLA4
- NOD2 and IBD5
- IBD5, ATGL16L1 and IL23R

Quality controls and filtering

Markers filtering:

- SNP call rate $\leq 95\%$
- MAF $\leq 5\%$
- Deviation from Hardy Weinberg Equilibrium in controls ($p < 1 \times 10^{-5}$)
- Duplicates
- SNPs not belonging to one unique gene

Subject filtering:

- Sample call rate $\leq 93\%$
- Duplicates

Ankylosing Spondylitis

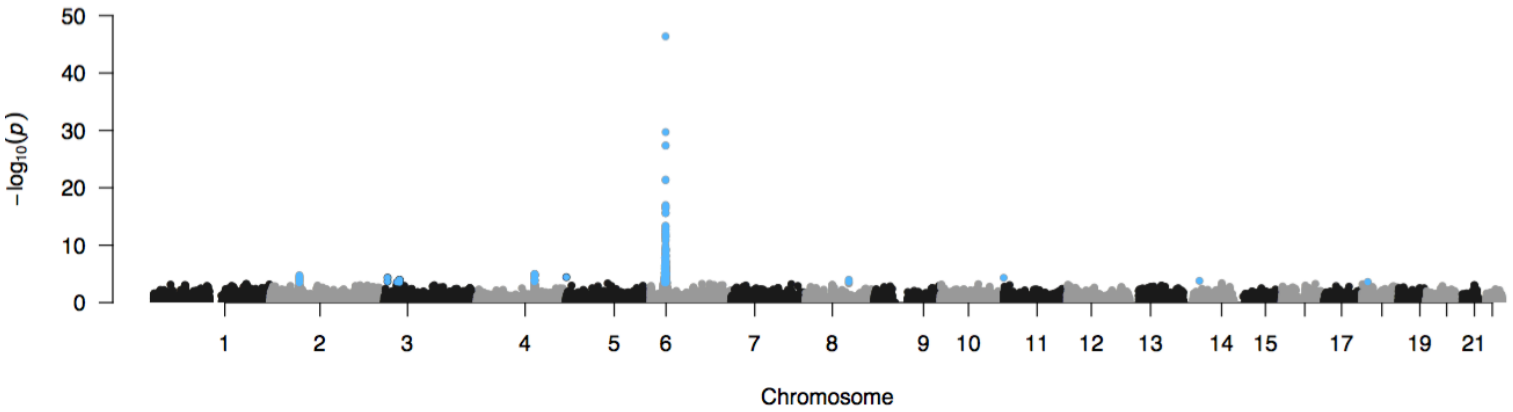
Data set: International Genetics of Ankylosing Spondylitis study

401 cases
357 controls
6 611 genes
51 287 SNPs

Ankylosing Spondylitis

Data set: International Genetics of Ankylosing Spondylitis study

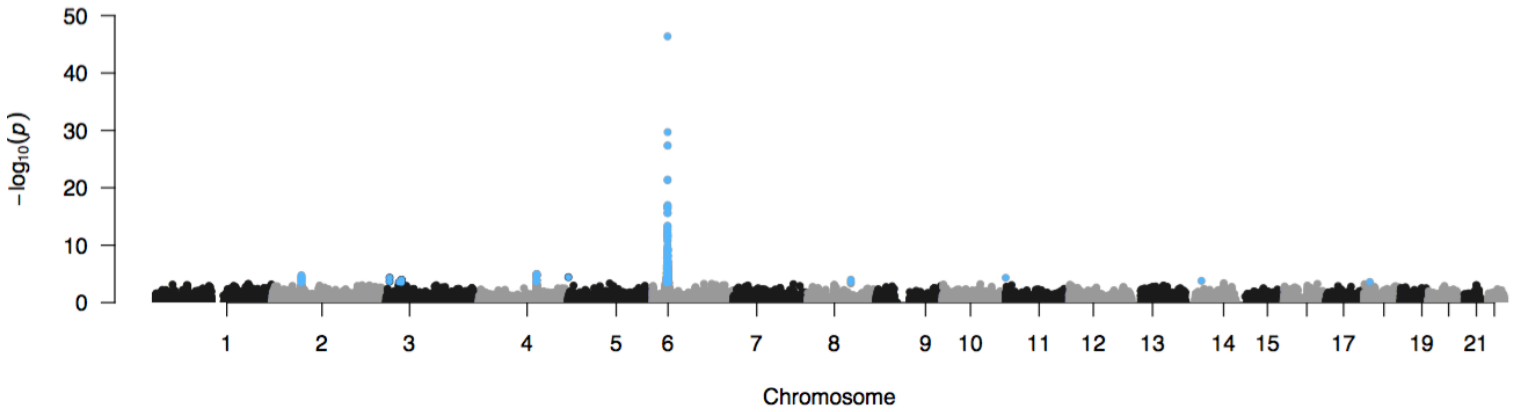
401 cases
357 controls
6 611 genes
51 287 SNPs



Ankylosing Spondylitis

Data set: International Genetics of Ankylosing Spondylitis study

401 cases
357 controls
6 611 genes
51 287 SNPs

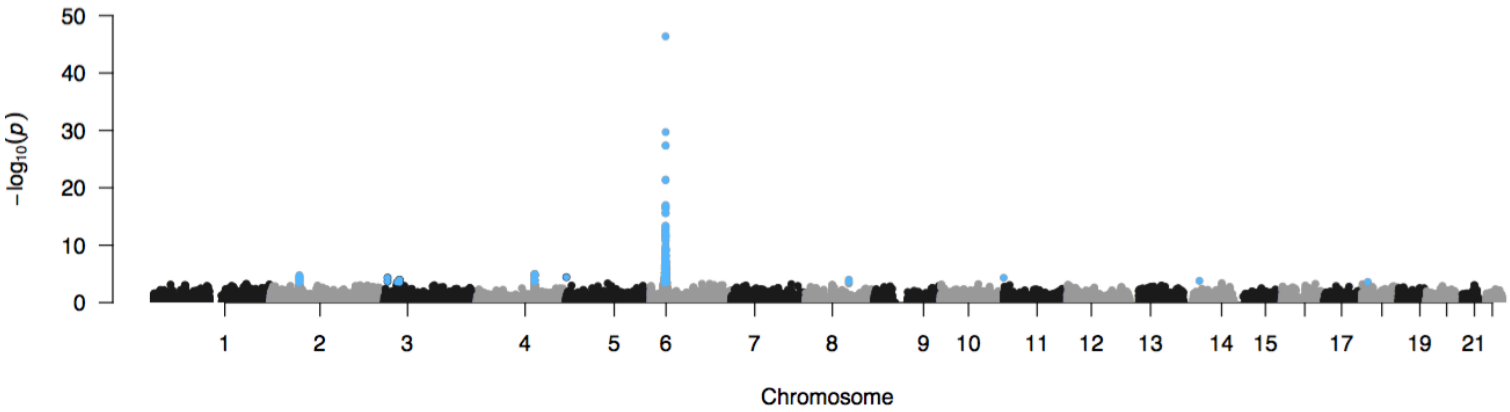


➔ 29 known genes
➔ 62 genes from an univariate analysis
➔ **91 genes to investigate**

Ankylosing Spondylitis

Data set: International Genetics of Ankylosing Spondylitis study

401 cases
357 controls
6 611 genes
51 287 SNPs



➔ 29 known genes
➔ 62 genes from an univariate analysis
➔ **91 genes to investigate**

	Significant results
G-GEE	NKX2-3 x HCG27
PLS	HLA-B HCP5 HLAB x HCG27
PCA	HLA-B EOMES x HCP5 IL1R2 x MICB ZFP57 x LOC101929772 TRIM31 x HCG26

Crohn's Disease

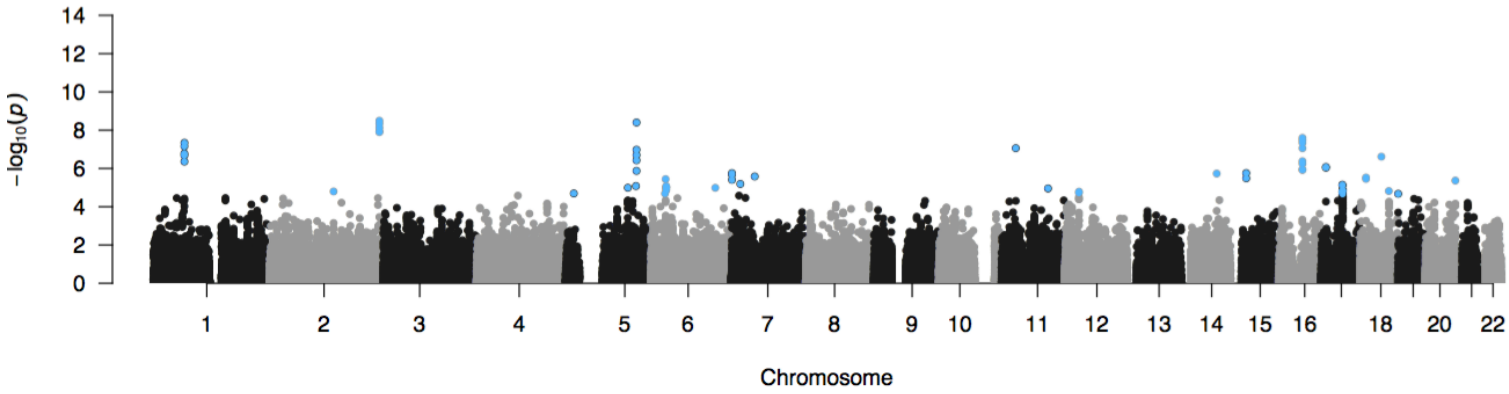
Data set: Wellcome Trust Case-Control Consortium

1938 cases
1500 controls
17 304 genes
140 487 SNPs

Crohn's Disease

Data set: Wellcome Trust Case-Control Consortium

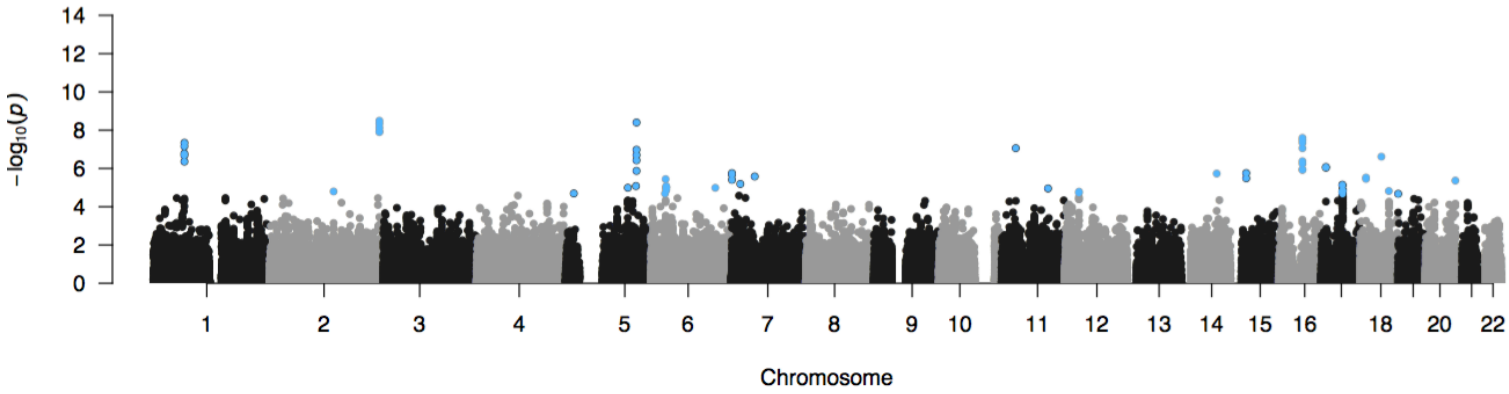
1938 cases
1500 controls
17 304 genes
140 487 SNPs



Crohn's Disease

Data set: Wellcome Trust Case-Control Consortium

1938 cases
1500 controls
17 304 genes
140 487 SNPs

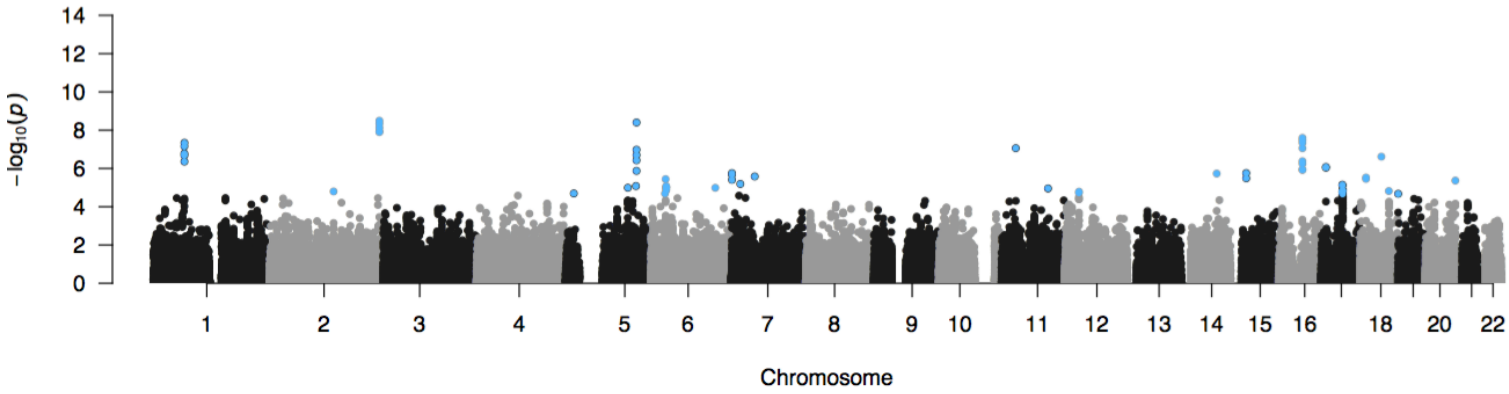


➔ 72 known genes
➔ 60 genes from an univariate analysis
(22 known)
➔ **110 genes to investigate**

Crohn's Disease

Data set: Wellcome Trust Case-Control Consortium

1938 cases
1500 controls
17 304 genes
140 487 SNPs



→ 72 known genes
→ 60 genes from an univariate analysis (22 known)
→ **110 genes to investigate**

	Significant results
G-GEE	LOC105369715 x STAT1 STAT1 x CD6
PLS	IFNGR1 x SBNO2 IRGM x NOD2
PCA	IRGM LOC101929544 x TLR4 BATF x IL10

Summary

Conclusions and perspectives

Contributions:

➔ Proposition of a new Group LASSO framework

➔ Proposition of an original interaction modeling

Pubication, software and presentations:

➔ Package G-GEE available on Github

➔ *Stanislas, V., Dalmasso, C., and Ambroise, C. (2017). Eigen-Epistasis for detecting gene-gene interactions. BMC Bioinformatics, 18(1):54.*

➔ 4 talks and 3 posters in international conferences

Conclusions and perspectives

Limitations:

- ➔ Number of SNPs by genes to analyze
- ➔ Computation costs for estimation coefficients
- ➔ Choice of the genes to consider
- ➔ Confusion phenomenon
- ➔ Sensitive to group definition

Conclusions and perspectives

Perspectives:

- ➔ Explore new $f_{\boldsymbol{u}}(\boldsymbol{X}^r, \boldsymbol{X}^s)$ definitions
- ➔ Optimization of the computational cost of $\boldsymbol{F}^{rs}$
- ➔ New selection of the parameter λ
- ➔ Using another penalization regression framework
- ➔ Gene selection using biological knowledge
- ➔ Investigate other grouping definitions

Thank you for your attention !