Gene-gene interaction in Genome Wide Association Studies

Mathieu Emily

31 mars 2017
Toulouse
Séminaire MIAT

Statistiques à Agrocampus Ouest

- **5 enseignants-chercheurs** :
  - Enseignement à Agrocampus Ouest
  - Recherche au sein de l'IRMAR (Equipe de Statistique)
- **1 Ingénieur**
- Actuellement **2 doctorants**

Principales activités

- **Enseignement** à Agrocampus Ouest
  - Licence, Spécialisation et Master en **Science des données**
  - Livres : Analyse de données avec R, Statistique avec R, Analyse factorielle simple et multiple, Statistique générale
  - MOOC : Analyse de données (sur FUN), Sensométrie

Principales activités

- **Enseignement** à Agrocampus Ouest
  - Licence, Spécialisation et Master en **Science des données**
  - Livres : Analyse de données avec R, Statistique avec R, Analyse factorielle simple et multiple, Statistique générale
  - MOOC : Analyse de données (sur FUN), Sensométrie
- **Recherche** à l'IRMAR
  - Principaux thèmes statistiques :
    - Analyse factorielle
    - Modélisation en grande dimension
    - Données manquantes
  - Principaux domaines d'application :
    - Etudes consommateurs et sensorielles
    - Génomique, protéomique
    - Données d'ERP (Event Related Potentials)

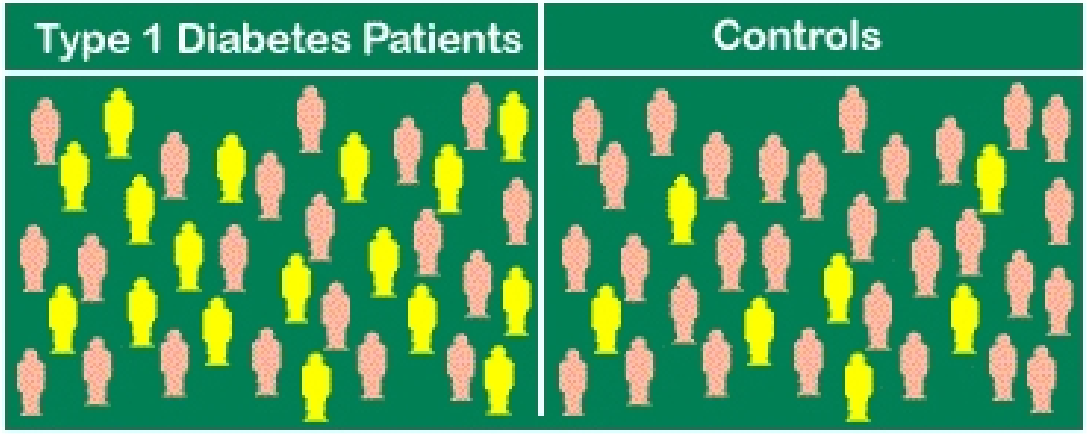
Principales activités

- **Enseignement** à Agrocampus Ouest
  - Licence, Spécialisation et Master en **Science des données**
  - Livres : Analyse de données avec R, Statistique avec R, Analyse factorielle simple et multiple, Statistique générale
  - MOOC : Analyse de données (sur FUN), Sensométrie
- **Recherche** à l'IRMAR
  - Principaux thèmes statistiques :
    - Analyse factorielle
    - Modélisation en grande dimension
    - Données manquantes
  - Principaux domaines d'application :
    - Etudes consommateurs et sensorielles
    - Génomique, protéomique
    - Données d'ERP (Event Related Potentials)
- **Autres activités**
  - Packages R : FactoMineR, SensoMineR, FAMT, missMDA, GeneGeneInteR
  - Congrès : useR!2009, CARME 2011, JSTAR'11, Sensometrics 2012, missDATA 2015, StatLearn'16, JSTAR'16

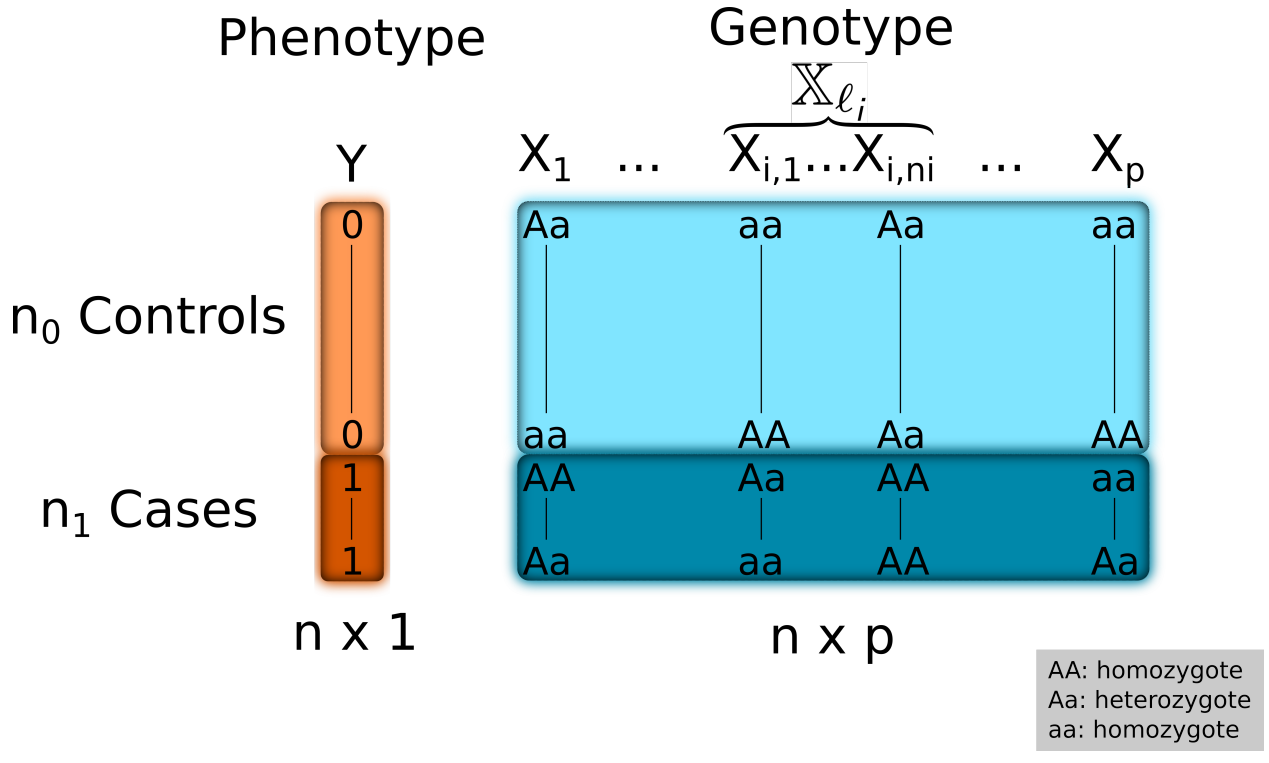
Outline

Genome-wide association studies (GWAS)

- **Case/control** studies
  - Detection of **differences in allelic frequencies** between cases and controls individuals
  - Genotyping of individuals from both populations



A typical dataset

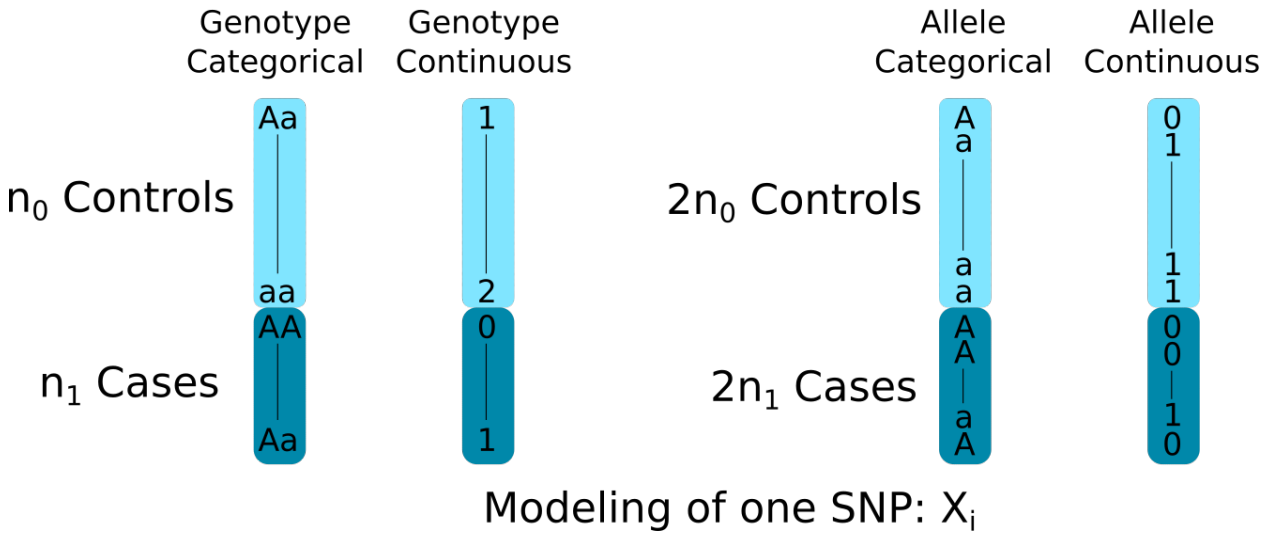


- $p \approx 500,000$
- $n \approx 1,000$

Biological question: Can we predict a phenotype given the genotype?

Statistical modeling of a single SNP

- Various **modeling** of a SNP (X_i) can be proposed:



Modeling of one SNP: X_i

- **Mathematical reasons** drive the choice for a statistical modeling
  - Impact on the biological interpretation of the results

Single-locus testing

- **Contingency table**:

	$Y=0$	$Y=1$	Total
X=AA	n_0^{AA}	n_1^{AA}	n^{AA}
X=Aa	n_0^{Aa}	n_1^{Aa}	n^{Aa}
X=aa	n_0^{aa}	n_1^{aa}	n^{aa}
	n_0	n_1	n

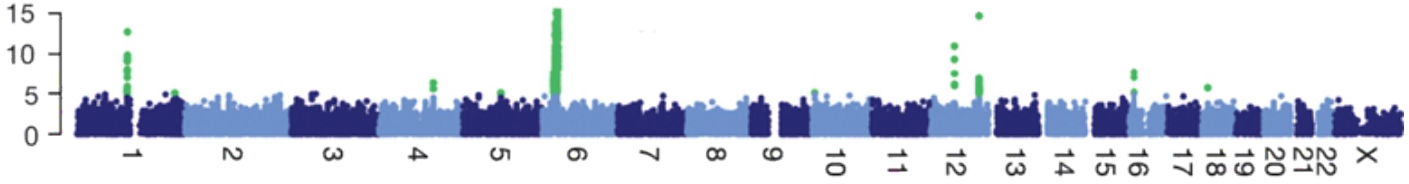
- **Cochran-Armitage** test of trend:

$$CA = \frac{T}{\sqrt{\tilde{T}}} = \frac{\sum_{i=1}^{3} t_i (n^i{}_0 n_1 - n_1^i n_0)}{\sqrt{\frac{n_0 n_1}{n}\left(\sum_{i=1}^{3} t_i^2 n^i (n - n^i) - 2\sum_{i=1}^{2}\sum_{j=i+1}^{3} t_i t_j n^i n^j\right)}}$$

  - $t = [0, 1, 1]$ optimal to test whether allele a is recessive over allele A
  - $t = [0, 1, 2]$ locally optimal to test whether alleles a and A are codominant
    - Linear trend in the frequencies
    - Very close to a Student test with X_i as a discrete variable

Single-locus testing

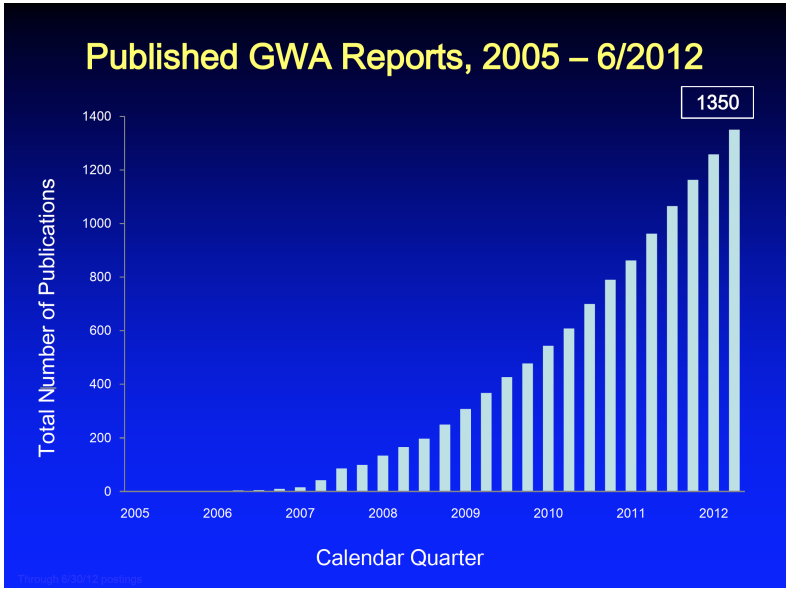
- **Visualisation** with a Manhattan plot:

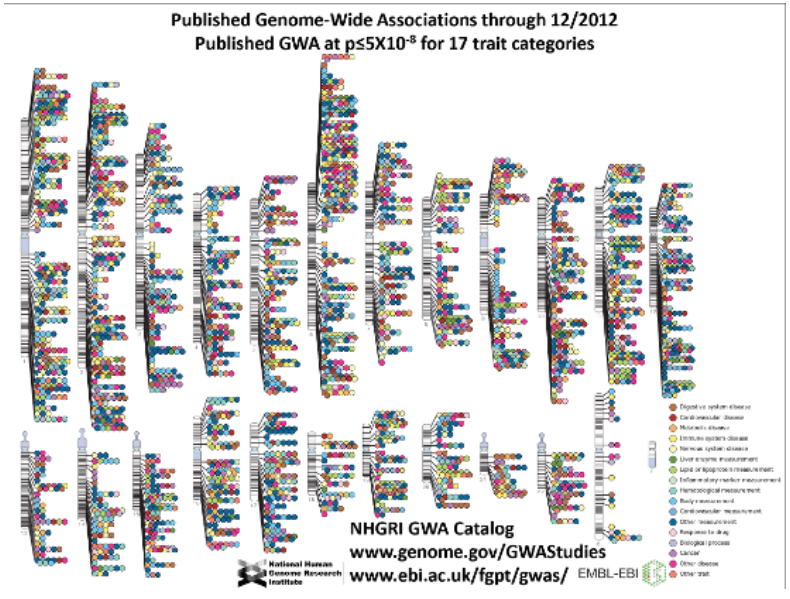


- Potential issues:
  - **Multiple testing issue**:
    - Significance level: 5×10^{-8} estimated using simulation
  - **Computational issue**: very fast with adapted software (PLINK)
  - **Interpretation** of the results

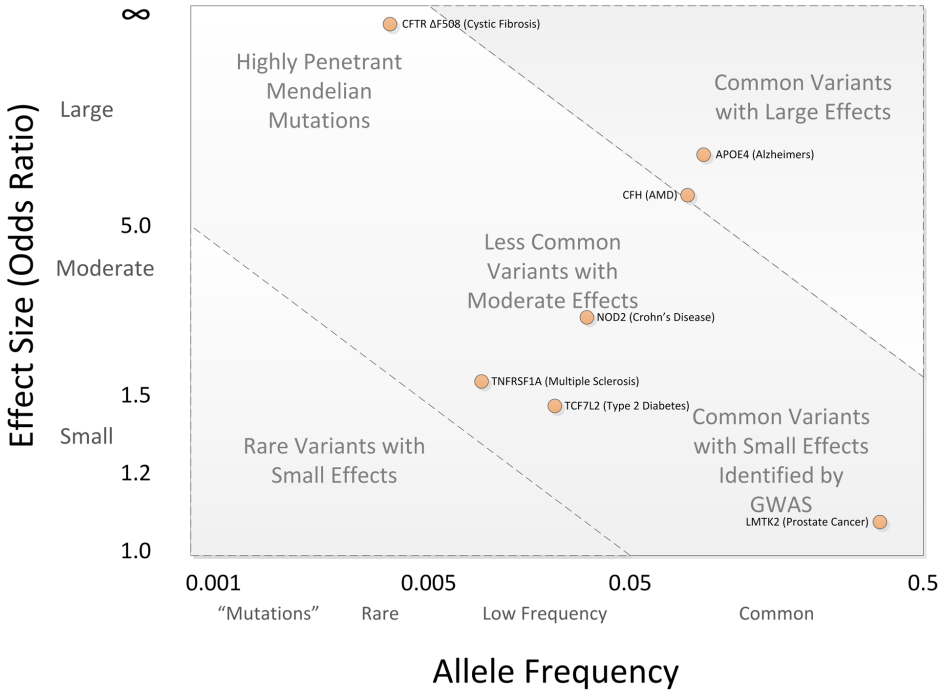
A success story?...yes

- **Since 2005**, a lot of variants has been found in susceptibility to various complex diseases: prostate cancer, Crohn's disease, etc...





A success story?...yes and no



- GWAS typically identify **common variants with small effect sizes**, lower right part of the graph (Bush WS, Moore JH, *PLoS Comput Biol*, 2012)

A success story?...no

TABLE 1

From the following article:
Finding the missing heritability of complex diseases
Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll & Peter M. Visscher
Nature **461**, 747-753(8 October 2009)
doi:10.1038/nature08494

back to article

Table 1. Estimates of heritability and number of loci for several complex traits.

Figures & Tables index | Next table

Disease	Number of loci	Proportion of heritability explained	Heritability measure
Age-related macular degeneration[72]	5	50%	Sibling recurrence risk
Crohn's disease[21]	32	20%	Genetic risk (liability)
Systemic lupus erythematosus[73]	6	15%	Sibling recurrence risk
Type 2 diabetes[74]	18	6%	Sibling recurrence risk
HDL cholesterol[75]	7	5.2%	Residual* phenotypic variance
Height[15]	40	5%	Phenotypic variance
Early onset myocardial infarction[76]	9	2.8%	Phenotypic variance
Fasting glucose[77]	4	1.5%	Phenotypic variance

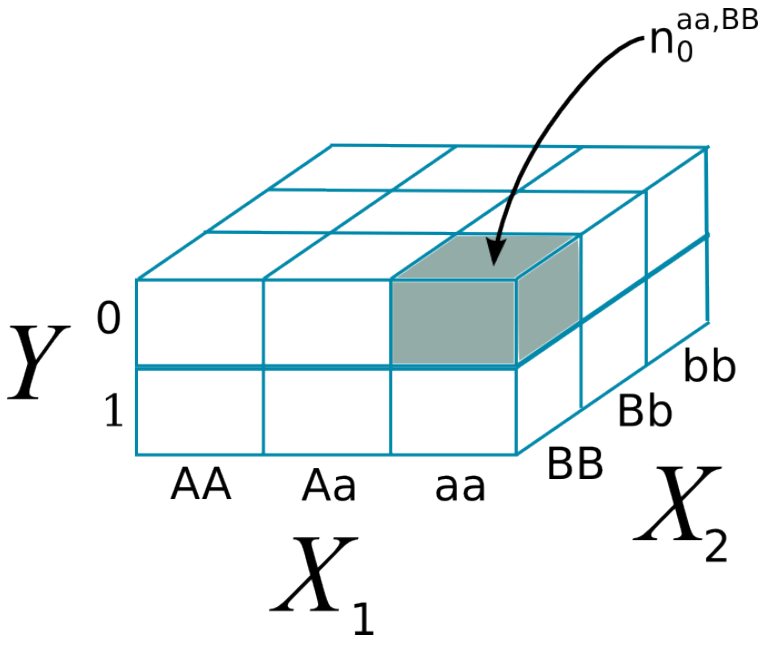
***Residual is after adjustment for age, gender, diabetes.**

- GWAS has generated new challenges in **the quest of missing heritability!**
  - **Next "natural" step**: testing for interaction

Outline

Context

- **Biological** interaction (epistasis) and heritability in GWAS *(Cordell, 2009)*.



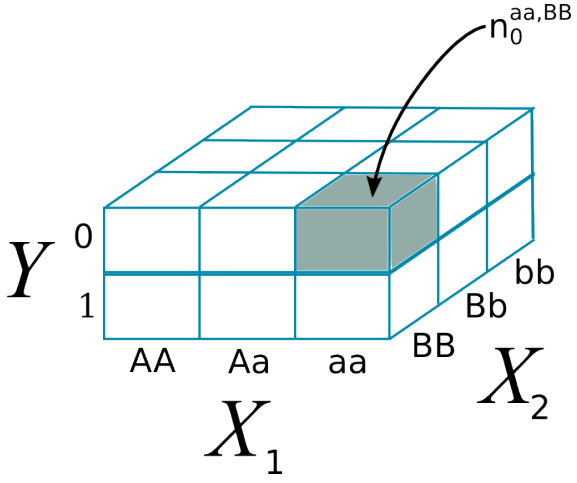
- Regression problem?

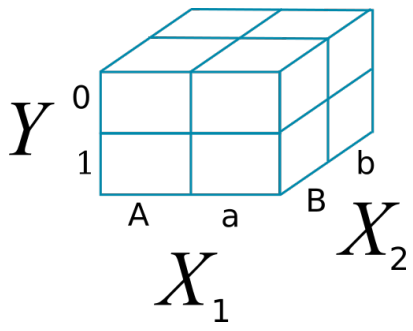
Logistic regression

- Allelic version (PLINK "gold standard" test *(Purcell et al., 2007)*)

$$\text{logit}\left(\mathbb{P}[Y=1|(X_1,X_2)=(x_1,x_2)]\right)=\alpha+\beta\mathbb{I}_{x_1=A}+\gamma\mathbb{I}_{x_2=B}+\delta\mathbb{I}_{(\mathbf{x_1},\mathbf{x_2})=(\mathbf{A},\mathbf{B})}$$

  - $\mathcal{H}_0 : \delta = 0$ vs $\mathcal{H}_1 : \delta \neq 0$
  - χ^2 à 1 un degré de liberté





Logistic regression

- Allelic version (PLINK "gold standard" test *(Purcell et al., 2007)*)

$$\text{logit}\left(\mathbb{P}[Y=1|(X_1,X_2)=(x_1,x_2)]\right)=\alpha+\beta\mathbb{I}_{x_1=A}+\gamma\mathbb{I}_{x_2=B}+\delta\mathbb{I}_{(\mathbf{x_1},\mathbf{x_2})=(\mathbf{A},\mathbf{B})}$$

  - $\mathcal{H}_0 : \delta = 0$ vs $\mathcal{H}_1 : \delta \neq 0$
  - χ^2 à 1 un degré de liberté

- **Genotype-based** version of the test:

$$\begin{aligned}\text{logit}\left[P(Y=1|(X_1,X_2)=(x_1,x_2)]\right) &= \alpha+\sum_{i=1}^{2}\beta_i\mathbb{I}_i(x_1)+\sum_{i=1}^{2}\gamma_i\mathbb{I}_i(x_2)\\ &\quad+\sum_{i=1}^{2}\sum_{j=1}^{2}\delta_{\mathbf{i},\mathbf{j}}\mathbb{I}_{(i,j)}(x_1,x_2)\end{aligned}$$

  - **Statistical hypothesis** can be written as

$$\mathcal{H}_0 \quad : \quad \delta_{i,j}=0 \ \ \forall(i,j)\in\{1,2\}^2 \text{ vs } \mathcal{H}_1 : \exists(i,j)\in\{1,2\}^2 \ \ \delta_{i,j}\neq 0$$

  - Under $\mathcal{H}_0$: χ^2 with **4 dof**

Other regression-based methods

- **T_{IH}** *(Wu et al., 2010; Ueki and Cordell, 2012)*
  - Statistics:

$$\mathbf{T_{IH}} = \frac{(\lambda(\theta_A) - \lambda(\theta_N))^2}{\nu_A + \nu_N}$$

    where:
    - $\lambda(\theta_i)$ is the log Odds Ratio of alleles
    - ν_i the variance of $\lambda(\theta_i)$
  - $\mathcal{H}_0 : T_{IH} = 0$ with $T_{IH} \sim_{\mathcal{H}_0} \chi^2_{1dof}$

Other regression-based methods

- **T_{IH}** *(Wu et al., 2010; Ueki and Cordell, 2012)*
  - Statistics:

$$\mathbf{T_{IH}} = \frac{(\lambda(\theta_A) - \lambda(\theta_N))^2}{\nu_A + \nu_N}$$

    where:
    - $\lambda(\theta_i)$ is the log Odds Ratio of alleles
    - ν_i the variance of $\lambda(\theta_i)$
  - $\mathcal{H}_0 : T_{IH} = 0$ with $T_{IH} \sim_{\mathcal{H}_0} \chi^2_{1dof}$

- **BOOST** *(Wan et al., 2010)*
  - Model: log-linear model for counts:

$$\log\left(n_k^{i,j}\right) = \lambda + \lambda_i^{X_1} + \lambda_j^{X_2} + \lambda_k^{Y} + \lambda_{ij}^{X_1X_2} + \lambda_{ik}^{X_1Y} + \lambda_{jk}^{X_2Y} + \lambda_{\mathbf{ijk}}^{\mathbf{X_1X_2Y}}$$

  - $\mathcal{H}_0 : [\lambda_{00k}^{X_1X_2Y}, \lambda_{10k}^{X_1X_2Y}, \lambda_{01k}^{X_1X_2Y}, \lambda_{11k}^{X_1X_2Y}] = [0, 0, 0, 0]$
    - Deviance-based test with 4 dof

Limit of regression-based approach

- Epistasis is a **departure from independence** of the effects of X_1 and X_2 in the way that they combine to cause Y *(Cordell, 2002)*.
- **Statistical** interaction is a **deviation from the additivity** of the marginal effects of X_1 and X_2 on Y *(Agresti, 2013)*.
- Limitation: **Statistical interaction $\not\Rightarrow$ Biological interaction**

Challenge

How can we formalize a biologically relevant statistical hypothesis?

Our approach

- Our approach IndOR: **Ind**ependent **O**dds **R**atio *(Emily, 2012)*

- IndOR is based on: "Epistasis is a departure from **independence**"
  - Under the null hypothesis (no epistasis), cases and controls share the **same amount of dependency**

$$\mathcal{H}_0 : \frac{\mathbb{P}\left[(X_1, X_2) = (x_1, x_2) | Y = 1\right]}{\mathbb{P}[X_1 = x_1 | Y = 1]\mathbb{P}[X_2 = x_2 | Y = 1]} = \frac{\mathbb{P}[(X_1, X_2) = (x_1, x_2) | Y = 0]}{\mathbb{P}[X_1 = x_1 | Y = 0]\mathbb{P}[X_2 = x_2 | Y = 0]}$$

$$\mathcal{H}_1 : \frac{\mathbb{P}\left[(X_1, X_2) = (x_1, x_2) | Y = 1\right]}{\mathbb{P}[X_1 = x_1 | Y = 1]\mathbb{P}[X_2 = x_2 | Y = 1]} \neq \frac{\mathbb{P}[(X_1, X_2) = (x_1, x_2) | Y = 0]}{\mathbb{P}[X_1 = x_1 | Y = 0]\mathbb{P}[X_2 = x_2 | Y = 0]}$$

- Using Bayes formula and by considering (AA, BB) as the **baseline genotype**, we have under $\mathcal{H}_0$:

$$\frac{OR(x_i, x_j)}{OR(x_i)OR(x_j)} = 1 \Longleftrightarrow \varphi_{(\mathbf{x_i}, \mathbf{x_j})} = \log\left(\frac{\mathbf{OR(x_i, x_j)}}{\mathbf{OR(x_i)OR(x_j)}}\right) = \mathbf{0}$$

Definition of the statistic IndOR

- Let us introduce the **4-dimensional** vector Φ, such as:

$$\Phi = [\varphi_{(Aa,Bb)}; \varphi_{(aa,Bb)}; \varphi_{(Aa,bb)}; \varphi_{(aa,bb)}]$$

- Statistical hypothesis can be written as:

$$\mathcal{H}_0 : \Phi = [0,0,0,0] \text{ and } H_1 : \Phi \neq [0,0,0,0]$$

- To test for $\mathcal{H}_0$, we defined our **Wald statistic**, IndOR, as follows:

$$IndOR = \Phi V_\Phi^{-1} \Phi^t$$

  where V_Φ^{-1} is the inverse of the variance-covariance of Φ.

- Under $\mathcal{H}_0$ we have:

$$IndOR \sim \chi^2(4d.f.)$$

Estimation of Φ and V_Φ

- Φ is evaluated using **maximum likelihood estimators**

$$\widehat{\varphi_{(x_i,x_j)}} = \log\left(\frac{n_1^{x_i x_j}}{n_0^{x_i x_j}}\frac{n_0^{AABB}}{n_1^{AABB}}\right) - \log\left(\frac{n_1^{x_i}}{n_0^{x_i}}\frac{n_0^{AA}}{n_1^{AA}}\right) - \log\left(\frac{n_1^{x_j}}{n_0^{x_j}}\frac{n_0^{BB}}{n_1^{BB}}\right)$$

Estimation of Φ and V_Φ

- Φ is evaluated using **maximum likelihood estimators**

$$\widehat{\varphi_{(x_i,x_j)}} = \log\left(\frac{n_1^{x_i x_j}}{n_0^{x_i x_j}}\frac{n_0^{AABB}}{n_1^{AABB}}\right) - \log\left(\frac{n_1^{x_i}}{n_0^{x_i}}\frac{n_0^{AA}}{n_1^{AA}}\right) - \log\left(\frac{n_1^{x_j}}{n_0^{x_j}}\frac{n_0^{BB}}{n_1^{BB}}\right)$$

- Estimation of V_Φ:
  - Assumption: **multinomial distribution** for counts:

$$[N_i^{AABB}, \ldots, N_i^{aabb}] \sim \mathcal{M}ult(p_i^{AABB}, \ldots, p_i^{aabb})$$

  - Delta method:

$$\log\left(N_i^{x_a,x_b}\right) \quad \approx \quad \log(n_i p_i^{x_a,x_b}) + \sqrt{\frac{(1-p_i^{x_a,x_b})}{n_i p_i^{x_a,x_b}}}\delta_i^{x_a,x_b}$$

    - $\delta_i^{x_a,x_b} \sim \mathcal{N}(0,1)$ and explicit covariance between $\delta_i^{x_a,x_b}$ and $\delta^{x'_a,x'_b}$.
  - The variance-covariance structure for N_{ij} can then be estimated and **plugged in the formulation of V_Φ**.

Estimation of Φ and V_Φ

- Φ is evaluated using **maximum likelihood estimators**

$$\widehat{\varphi_{(x_i,x_j)}} = \log\left(\frac{n_1^{x_i x_j}}{n_0^{x_i x_j}}\frac{n_0^{AABB}}{n_1^{AABB}}\right) - \log\left(\frac{n_1^{x_i}}{n_0^{x_i}}\frac{n_0^{AA}}{n_1^{AA}}\right) - \log\left(\frac{n_1^{x_j}}{n_0^{x_j}}\frac{n_0^{BB}}{n_1^{BB}}\right)$$

- Estimation of V_Φ:
  - Assumption: **multinomial distribution** for counts:

$$[N_i^{AABB}, \ldots, N_i^{aabb}] \sim \mathcal{M}ult(p_i^{AABB}, \ldots, p_i^{aabb})$$

  - Delta method:

$$\log\left(N_i^{x_a,x_b}\right) \approx \log(n_i p_i^{x_a,x_b}) + \sqrt{\frac{(1-p_i^{x_a,x_b})}{n_i p_i^{x_a,x_b}}}\delta_i^{x_a,x_b}$$

    - $\delta_i^{x_a,x_b} \sim \mathcal{N}(0,1)$ and explicit covariance between $\delta_i^{x_a,x_b}$ and $\delta^{x'_a,x'_b}$.
  - The variance-covariance structure for N_{ij} can then be estimated and **plugged in the formulation of V_Φ**.

Derivation of a **closed-form expression** for IndOR

Validation protocol

- A **three-steps** protocol.

Validation protocol

- A **three-steps** protocol.

  Step 1: Evaluation under the null hypothesis

Validation protocol

- A **three-steps** protocol.

  Step 1: Evaluation under the null hypothesis

  Step 2: Evaluation under the alternative hypothesis

Validation protocol

- A **three-steps** protocol.
  - Step 1: Evaluation under the null hypothesis
  - Step 2: Evaluation under the alternative hypothesis
  - Step 3: Application to real dataset(s)

Validation protocol

- A **three-steps** protocol.

  Step 1: Control of the **type-I error** under three situations
  - (1) No effect, (2) one marginal single effect, (3) two marginal effects

Validation protocol

- A **three-steps** protocol.

Step 1: Control of the **type-I error** under three situations

- (1) No effect, (2) one marginal single effect, (3) two marginal effects

Step 2: **Power study** under various scenarios

- Objective: Comparison with existing methods:
  - PLINK, BOOST and T_{IH}
- Results:
  - IndOR $\approx T_{IH}$ for classical disease models (Recessive and/or Dominant)
  - IndOR has more power to detect biological models.
  - Better results for IndOR when:
    (1) $\sharp$ controls $>$ $\sharp$ cases
    (2) X_1 and X_2 are linked

Validation protocol

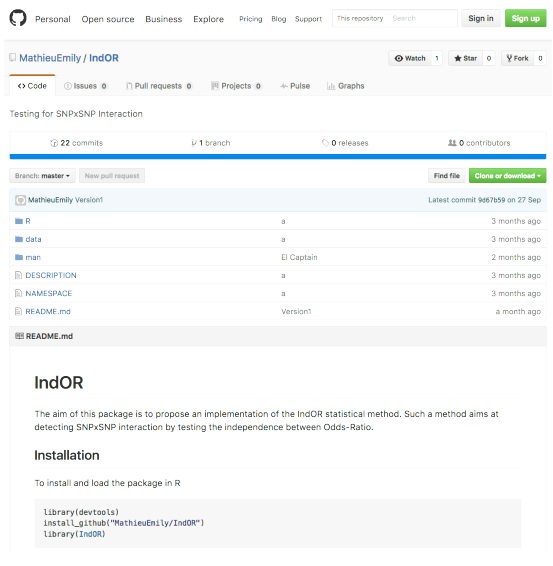
- A **three-steps** protocol.

Step 1: Control of the **type-I error** under three situations
  - (1) No effect, (2) one marginal single effect, (3) two marginal effects

Step 2: **Power study** under various scenarios
  - Objective: Comparison with existing methods:
    - PLINK, BOOST and T_{IH}
  - Results:
    - IndOR $\approx T_{IH}$ for classical disease models (Recessive and/or Dominant)
    - IndOR has more power to detect biological models.
    - Better results for IndOR when:
      (1) $\sharp$ controls $>$ $\sharp$ cases
      (2) X_1 and X_2 are linked

Diffusion of the method:

`https://github.com/MathieuEmily/IndOR/`



Validation protocol

Step 3: Study of a landmark dataset (Wellcome Trust Case Control Consortium, 2007)

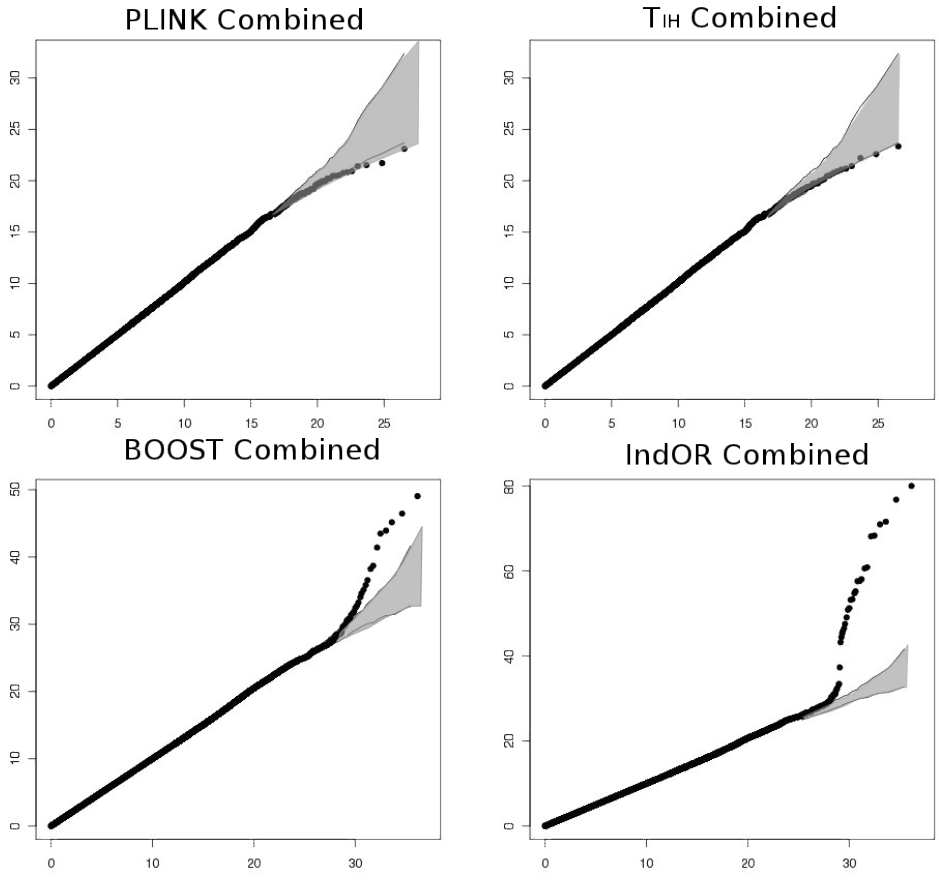
- $n = 17,000$ individuals
- **p=3,000,000** SNP pairs.

Results

- **Scalable** method
- Potential **findings**:
  - Two SNP pairs associated with Crohn's disease

$(X_1, X_2) =$ (rs6496669, rs434157)

$(X_1, X_2) =$ (rs9009, rs2830075)



QQ plots for PLINK, T_{IH}, BOOST and IndOR (Emily, 2012)

Limitations of SNP-SNP interaction

- Significance level?
  - Correlation among SNPs
- Biological interpretation:
  - The SNP is not the functional unit of the genome

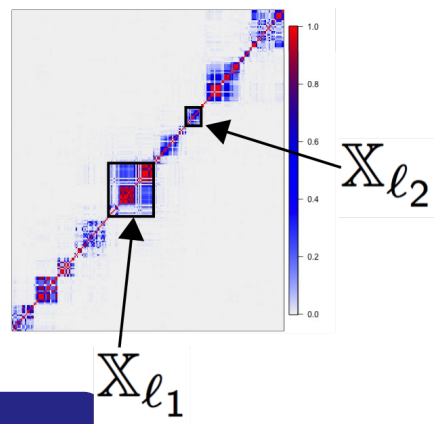
$\Rightarrow$ Accounting for the block structure of the genome

Outline

Context

- **Block-level testing** allows to:
  - Account for the block structure of the genome *(Huang et al., 2011)*
  - Characterize functional level (gene) *(Phillips, 2008)*
  - Facilitate the biological interpretation of findings *(Neale and Sham, 2004)*

- Detecting **interaction between two blocks** is challenging:
  - Let consider two blocks of variables $\mathbb{X}_{\ell_1}$ and $\mathbb{X}_{\ell_2}$:
    $\mathbb{X}_{\ell_1} = \left[X_{1,\ell_1}, \ldots, X_{p_{\ell_1},\ell_1}\right]$
    $\mathbb{X}_{\ell_2} = \left[X_{1,\ell_2}, \ldots, X_{p_{\ell_2},\ell_2}\right]$
    where each $X_{i,j}$ is discrete with values in $\{0; 1; 2\}$.



Challenge

How can we test for an association between Y and the interaction between $\mathbb{X}_{\ell_1}$ and $\mathbb{X}_{\ell_2}$?

Formalism

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_{n_c} \\ y_{n_c+1} \\ \vdots \\ y_{n_c+n_d} \end{bmatrix} \quad \mathbb{X}_{\ell_1} = \begin{bmatrix} \mathbf{X_1^c} \\ \\ \mathbf{X_1^d} \end{bmatrix} = \begin{bmatrix} x^1_{11} & \cdots & x^1_{1p_{\ell_1}} \\ \vdots & \ddots & \vdots \\ x^1_{n_c1} & \cdots & x^1_{n_cp_{\ell_1}} \\ x^1_{(n_c+1)1} & \cdots & x^1_{(n_c+1)p_{\ell_1}} \\ \vdots & \ddots & \vdots \\ x^1_{(n_c+n_d)1} & \cdots & x^1_{(n_c+n_d)p_{\ell_1}} \end{bmatrix}$$

$$\mathbb{X}_{\ell_2} = \begin{bmatrix} \mathbf{X_2^c} \\ \\ \mathbf{X_2^d} \end{bmatrix} = \begin{bmatrix} x^2_{11} & \cdots & x^2_{1p_{\ell_2}} \\ \vdots & \ddots & \vdots \\ x^2_{n_c1} & \cdots & x^2_{n_cp_{\ell_2}} \\ x^2_{(n_c+1)1} & \cdots & x^2_{(n_c+1)p_{\ell_2}} \\ \vdots & \ddots & \vdots \\ x^2_{(n_c+n_d)1} & \cdots & x^2_{(n_c+n_d)p_{\ell_2}} \end{bmatrix}$$

PCA-based method *(Li et al., 2009)*

- **PCA** first performed on each set of SNPs $\mathbb{X}_{\ell_1}$ and $\mathbb{X}_{\ell_2}$
- **Likelihood ratio test** between two logistic models:

$$\text{logit}\left(\mathbb{P}\left[Y=1|PC^1_{X_1}\dots PC^{n_1}_{X_1}, PC^1_{X_2}\dots PC^{n_2}_{X_2}\right]\right) = \beta_0 + \sum_{i=1}^{n_1} PC^i_{X_1} + \sum_{j=1}^{n_2} PC^j_{X_2} + \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} PC^i_{X_1} PC^j_{X_2}$$

and $\mathcal{M}_{No}$ in Equation 27:

$$\text{logit}\left(\mathbb{P}\left[Y=1|PC^1_{X_1}\dots PC^{n_1}_{X_1}, PC^1_{X_2}\dots PC^{n_2}_{X_2}\right]\right) = \beta_0 + \sum_{i=1}^{n_1} PC^i_{X_1} + \sum_{j=1}^{n_2} PC^j_{X_2}$$

- n_1 and n_2 are chosen to retrieve a predefined amount of information (70%)

U-like statistics

$$U = \frac{z_d - z_c}{\sqrt{\mathbb{V}(z_d - z_c)}}$$

1. $z_d = \frac{1}{2}(\log(1 + r_d) - \log(1 - r_d))$: Fisher transformation of r_d the **maximum canonical correlation coefficient** between $\mathbf{X_1^d}$ and $\mathbf{X_2^d}$.
   - Under $\mathcal{H}_0 :\sim \mathcal{N}(0, 1)$
   - $\mathbb{V}(z_d - z_c)$ estimated by **boostrap**
   - *(Peng et al., 2010)*

U-like statistics

$$U = \frac{z_d - z_c}{\sqrt{\mathbb{V}(z_d - z_c)}}$$

1. $z_d = \frac{1}{2}\left(\log(1 + r_d) - \log(1 - r_d)\right)$: Fisher transformation of r_d the **maximum canonical correlation coefficient** between $\mathbf{X_1^d}$ and $\mathbf{X_2^d}$.
   - Under $\mathcal{H}_0 :\sim \mathcal{N}(0, 1)$
   - $\mathbb{V}(z_d - z_c)$ estimated by **boostrap**
   - *(Peng et al., 2010)*
2. $z_d = \frac{1}{2}\left(\log(1 + kr_d) - \log(1 - kr_d)\right)$: **kernelized version**.
   - Under $\mathcal{H}_0 :\sim \mathcal{N}(0, 1)$
   - $\mathbb{V}(z_d - z_c)$ estimated by boostrap
   - *(Larson et al., 2014)*

U-like statistics

$$U = \frac{z_d - z_c}{\sqrt{\mathbb{V}(z_d - z_c)}}$$

1. $z_d = \frac{1}{2}(\log(1 + r_d) - \log(1 - r_d))$: Fisher transformation of r_d the **maximum canonical correlation coefficient** between $\mathbf{X_1^d}$ and $\mathbf{X_2^d}$.
   - Under $\mathcal{H}_0 :\sim \mathcal{N}(0, 1)$
   - $\mathbb{V}(z_d - z_c)$ estimated by **boostrap**
   - *(Peng et al., 2010)*
2. $z_d = \frac{1}{2}(\log(1 + kr_d) - \log(1 - kr_d))$: **kernelized version**.
   - Under $\mathcal{H}_0 :\sim \mathcal{N}(0, 1)$
   - $\mathbb{V}(z_d - z_c)$ estimated by boostrap
   - *(Larson et al., 2014)*
3. $z_d = \beta_d$: path coefficient between $\mathbf{X_1^d}$ and $\mathbf{X_2^d}$ in a **Partial Least Square Path Modeling (PLSPM)**.
   - Distribution unknown under $\mathcal{H}_0$
   - Significance tested by permutations
   - *(Zhang et al., 2013)*

Composite Linkage Disequilibrium (CLD) *(Rajapakse et al., 2012)*

- CLD is based on the **normalized quadratic distance** (NQD):

$$\delta^2 = \mathrm{tr.}\Big((\tilde{D} - \tilde{C})W^{-1}(\tilde{D} - \tilde{C})W^{-1}\Big)$$

with:

$$\tilde{D} = \begin{bmatrix} W_{11} & D_{12} \\ D_{21} & W_{22} \end{bmatrix} \qquad \tilde{C} = \begin{bmatrix} W_{11} & C_{12} \\ C_{21} & W_{22} \end{bmatrix}$$

where D and C are **marginal estimates** of the covariance matrix:

$$D = \mathrm{Cov}\left(\mathbf{X_1^d}, \mathbf{X_2^d}\right) = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} \qquad C = \mathrm{Cov}\left(\mathbf{X_1^c}, \mathbf{X_2^c}\right) = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

and W the **pooled estimate** of the covariance matrix:

$$W = \frac{n_c C + n_d D}{n_c + n_d} = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$$

Gene-Based Information Gain Method (GBIGM) *(Li et al., 2015)*

- GBIGM is based on the **information gain rate** $\Delta R_{1,2}$:

$$\Delta R_{1,2} = \frac{\min(H_1, H_2) - H_{1,2}}{\min(H_1, H_2)} \tag{1}$$

where H_1, H_2, $H_{1,2}$ are the conditional entropies (given $\mathbf{Y}$) of $\mathbf{X_1}$, $\mathbf{X_2}$ and the pooled SNP set $(\mathbf{X_1}, \mathbf{X_2})$ respectively.

- Assuming that $H(.)$ is the classical **entropy function**, we have:

$$\begin{aligned} H_1 &= H(\mathbf{Y}, \mathbf{X_1}) - H(\mathbf{X_1}) && (2)\\ H_2 &= H(\mathbf{Y}, \mathbf{X_2}) - H(\mathbf{X_2}) && (3)\\ H_{1,2} &= H(\mathbf{Y}, \mathbf{X_1}, \mathbf{X_2}) - H(\mathbf{X_1}, \mathbf{X_2}) && (4) \end{aligned}$$

Other point-of-view

- Existing methods: **comparison of covariance structures** in cases and controls

Other point of view

Signal detection approach

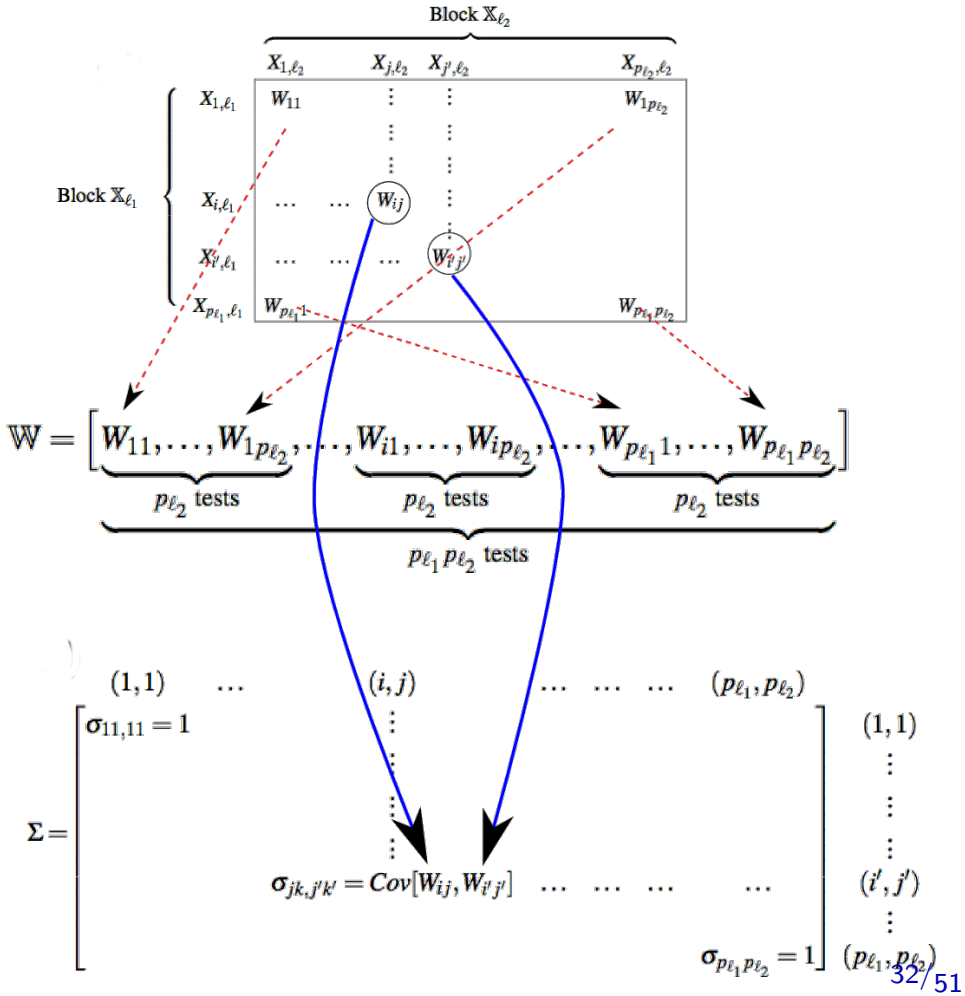
Our approach

- **A Gene-based GEne-Gene interActTiOn test: AGGrEGATOr** *(Emily, 2016)*

- Let $\beta_3^{i,j}$ be the **interaction coefficient**:
  $\text{logit}\left[\mathbb{P}(Y=1|X_{i,\ell_1}=x_1, X_{j,\ell_2}=x_2)\right]$
  $= \beta_0^{i,j} + \beta_1^{i,j}x_1 + \beta_2^{i,j}x_2 + \beta_3^{i,j}x_1x_2$

- For **each pair** $(X_{i,\ell_1}, X_{j,\ell_2})$, let define:

$$W_{ij} = \frac{\widehat{\beta_3^{i,j}}}{\widehat{\sigma\left(\widehat{\beta_3^{i,j}}\right)}}$$



minP: a block-based interaction test

- **Block-level** statistical hypothesis:

$$\mathcal{H}_0 : \quad \forall 1 \leq i \leq p_{\ell_1} \text{ and } \forall 1 \leq j \leq p_{\ell_2}, \quad W_{ij} = 0,$$
$$\mathcal{H}_1 : \quad \exists (i,j) \text{ where } 1 \leq i \leq p_{\ell_1} \text{ and } 1 \leq j \leq p_{\ell_2}, \quad W_{ij} \neq 0.$$

- Under $\mathcal{H}_0$, we have:

$$\mathbb{W} = [W_{11}, \ldots, W_{p_{\ell_1} p_{\ell_2}}] \sim \mathcal{N}(\mathbf{0}, \Sigma),$$

where $\mathcal{N}(\mathbf{0}, \Sigma)$ is the **multivariate normal density** with mean $\mathbf{0}$, the $p_{\ell_1} \times p_{\ell_2}$ null vector, and **covariance matrix** Σ.

- Σ can be decomposed as:

$$\Sigma = [\sigma_{(i,j),(i',j')}]_{\substack{i=1\ldots p_{\ell_1}; j=1\ldots p_{\ell_2} \\ i'=1\ldots p_{\ell_1}; j'=1\ldots p_{\ell_2}}}$$

is a $(p_{\ell_1} \times p_{\ell_2}) \times (p_{\ell_1} \times p_{\ell_2})$ symmetric matrix where:

$$\sigma_{(i,j),(i',j')} = Cov(W_{ij}, W_{i',j'})$$

.

minP: a block-based interaction test

- The **minP** approach:

$$\text{minP} = 1 - \mathbb{P}\Big[\max(|Z_1|, |Z_2|, \ldots, |Z_{p_{\ell_1} p_{\ell_2}}|) < W_{\max}\Big] \tag{5}$$

with:

- $W_{\max}$ be the **maximum of the absolute values** for the observed statistics:

$$W_{\max} = \max\{|W_{11}|, \ldots, |W_{p_{\ell_1} p_{\ell_2}}|\}$$

- $\mathbb{Z} = [Z_1, \ldots, Z_{p_{\ell_1} p_{\ell_2}}]$ be a **multivariate Gaussian random** vector:

$$\mathbb{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

minP: a block-based interaction test

- The **minP** approach:

$$\text{minP} = 1 - \mathbb{P}\Big[\max(|Z_1|, |Z_2|, \ldots, |Z_{p_{\ell_1} p_{\ell_2}}|) < W_{\max}\Big] \tag{5}$$

with:

  - $W_{\max}$ be the **maximum of the absolute values** for the observed statistics:

$$W_{\max} = \max\{|W_{11}|, \ldots, |W_{p_{\ell_1} p_{\ell_2}}|\}$$

  - $\mathbb{Z} = [Z_1, \ldots, Z_{p_{\ell_1} p_{\ell_2}}]$ be a **multivariate Gaussian random** vector:

$$\mathbb{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

- Computation of Equation (5) is based on a **numerical integration** knowing Σ *(Conneely and Boehnke, 2007)*

minP: a block-based interaction test

- The **minP** approach:

$$\text{minP} = 1 - \mathbb{P}\Big[\max(|Z_1|, |Z_2|, \ldots, |Z_{p_{\ell_1} p_{\ell_2}}|) < W_{\max}\Big] \tag{5}$$

with:

  - $W_{\max}$ be the **maximum of the absolute values** for the observed statistics:

$$W_{\max} = \max\{|W_{11}|, \ldots, |W_{p_{\ell_1} p_{\ell_2}}|\}$$

  - $\mathbb{Z} = [Z_1, \ldots, Z_{p_{\ell_1} p_{\ell_2}}]$ be a **multivariate Gaussian random** vector:

$$\mathbb{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

- Computation of Equation (5) is based on a **numerical integration** knowing Σ *(Conneely and Boehnke, 2007)*

- **Estimation of Σ** based on the dependency between variables.

  - Let $r_{i,i'} = \frac{p_{ii'} - p_i p_{i'}}{\sqrt{p_i(1-p_i)p_{i'}(1-p_{i'})}}$ be the correlation *(Hill and Robertson, 1968)*.
  - We proposed that:

$$\sigma_{(i,j),(i',j')} = Cov(W_{ij}, W_{i'j'}) \approx r_{i,i'} r_{j,j'}$$

Evaluation

Step 1: Control of the **type-I error**: robust to:

- the presence of marginal effects
- different patterns of dependencies within blocks

Evaluation

Step 1: Control of the **type-I error**: robust to:

- the presence of marginal effects
- different patterns of dependencies within blocks

Step 2: Power study under various scenarios

- Objectives:
  - Comparison with existing methods (PCA, CCA, KCCA, CLD, PLSPM, GBIGM)
  - Investigation under real pattern of dependencies
- Results:
  - AGGrEGATOr is the most powerful method when one causal pair is associated
  - Acceptable power in presence of multiple causal pairs

Evaluation

Step 1: Control of the **type-I error**: robust to:

- the presence of marginal effects
- different patterns of dependencies within blocks

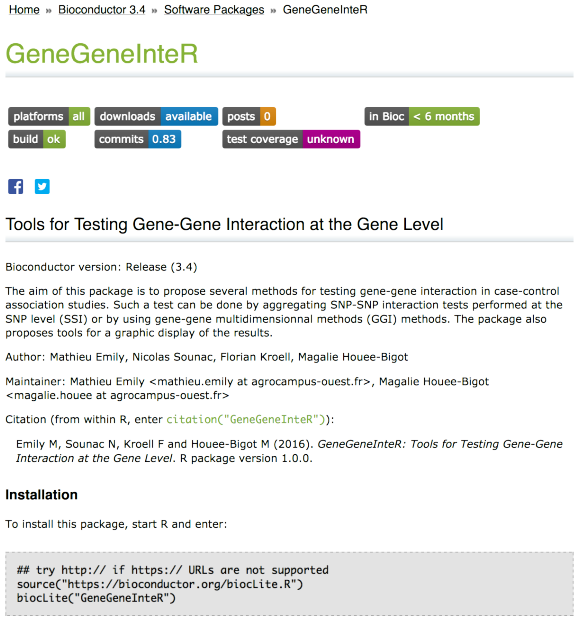
Step 2: Power study under various scenarios

- Objectives:
  - Comparison with existing methods (PCA, CCA, KCCA, CLD, PLSPM, GBIGM)
  - Investigation under real pattern of dependencies
- Results:
  - AGGrEGATOr is the most powerful method when one causal pair is associated
  - Acceptable power in presence of multiple causal pairs

Diffusion of the method:

`http://bioconductor.org/packages/GeneGeneInteR/`

Emily et al., (2016)



Validation protocol

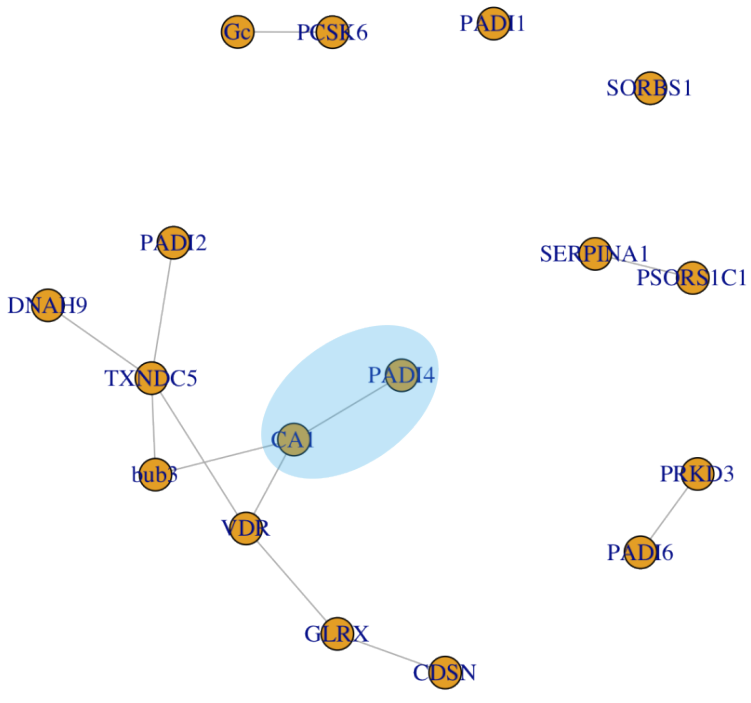
Step 3: Study of the **GSE39428 dataset** related to Rheumatoid Arthritis

(Chang et al., 2013)

- $n_0 = 163$ controls and $n_1 = 266$ cases
- Analysis of **131 block pairs** from 17 blocks (genes).

Results

- Identification of **12 significant gene pairs**
- **Replication** of 1 gene pair in the WTCCC dataset:
  - *CA1 - PADI4*



Interaction network for GSE39428 (Emily et al. 2016)

Outline

Description of the dataset

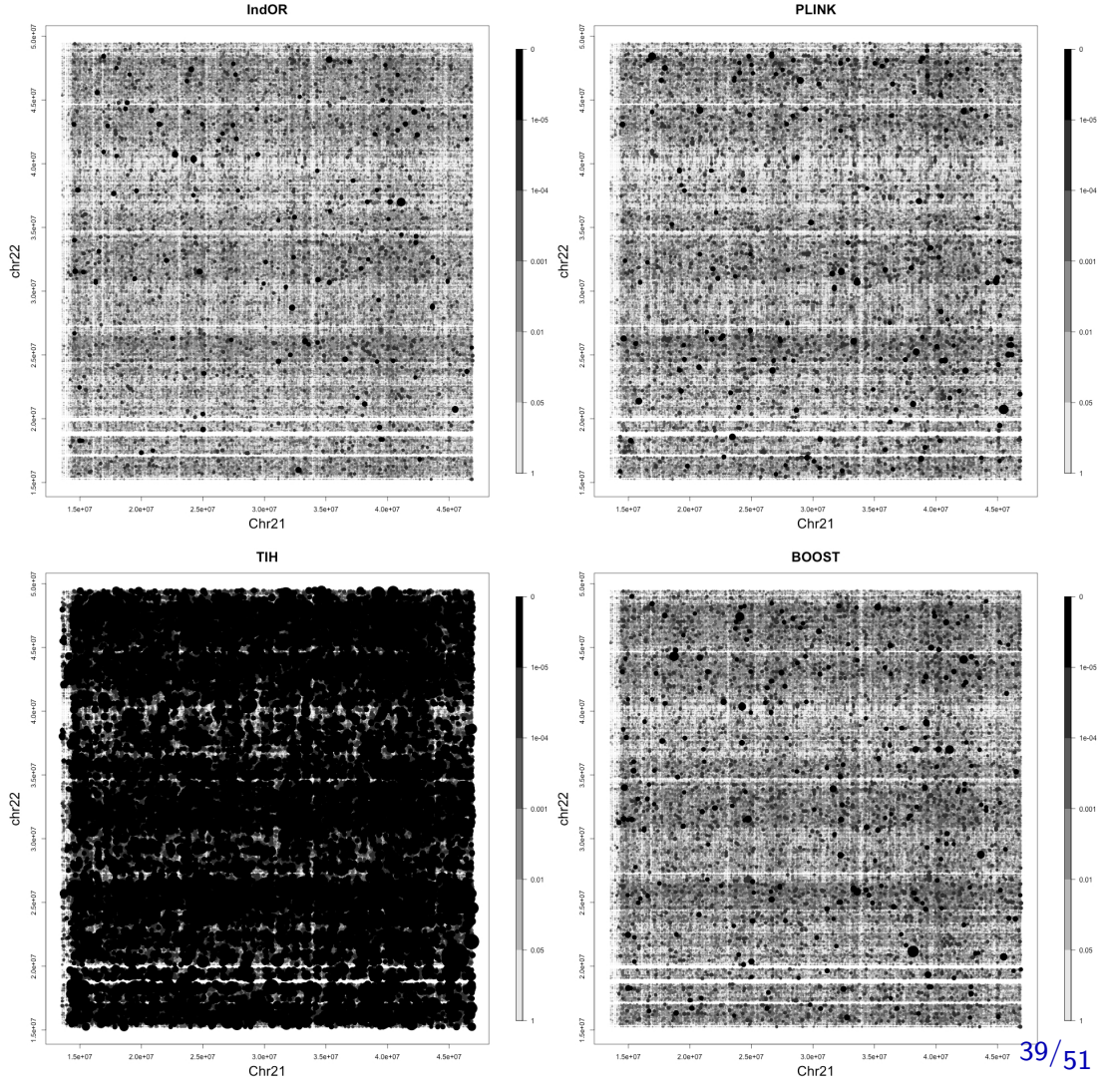
- Study of the interaction between Chr 21 and Chr 22 in susceptibility with Crohn's disease
- 4518 SNPs (Chr. 21) and 3708 SNPs (Chr. 22)
  - $> \mathbf{16,750,000}$ SNP-SNP tests
- 233 genes (Chr. 21) and 438 genes (Chr. 22)
  - $> \mathbf{100,000}$ Gene-Gene tests

Challenge

How can we **visualize** such amount of tests?

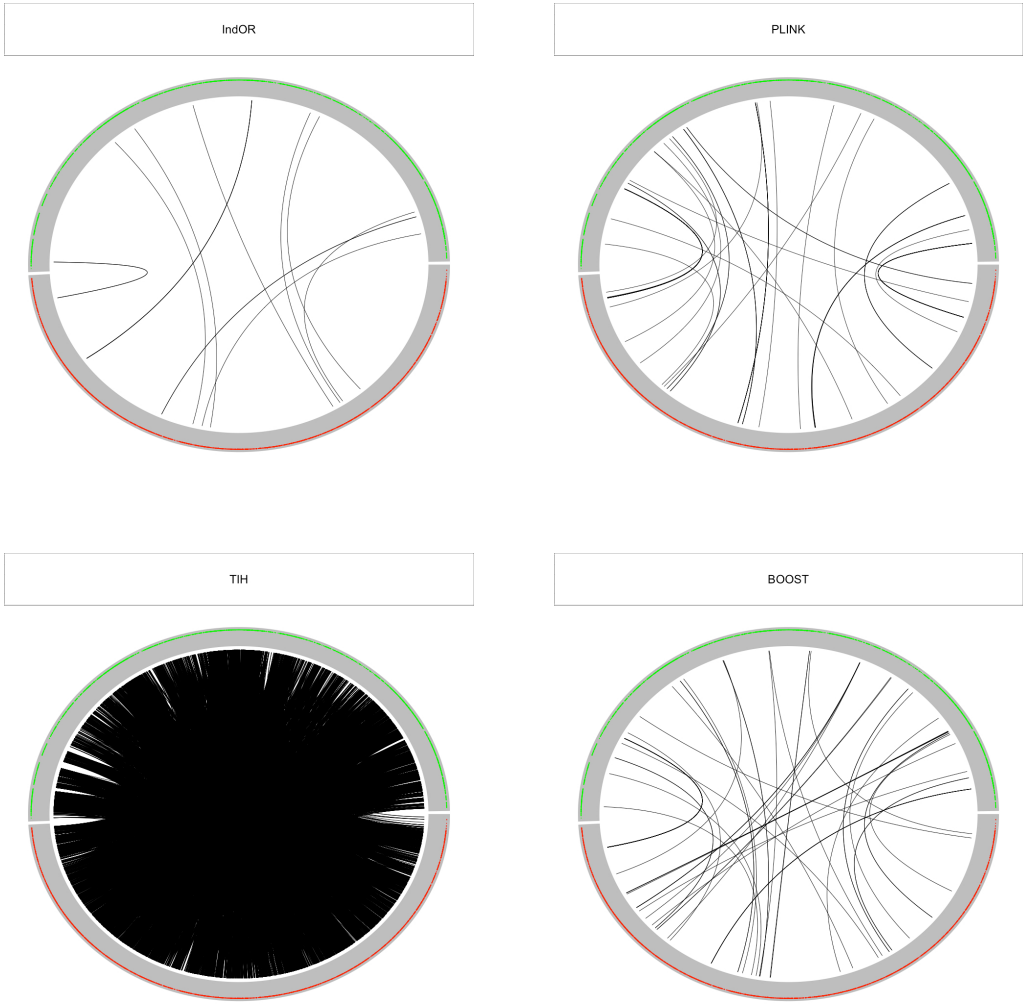
Heatmap visulisation for SNP-SNP interaction

- IndOR, PLINK and BOOST: No SNP pairs are significative after BH correction
- TIH: issue regarding the control of the type-I error



Circos visualisation for SNP-SNP interaction

- Significance threshold: 10^{-7}



Biological interpretation for SNP-SNP interaction

	BB	Bb	bb
AA	398	471	199
Aa	500	713	200
aa	166	229	84

Controls

	BB	Bb	bb
AA	235	376	80
Aa	331	438	200
aa	106	154	59

Cases

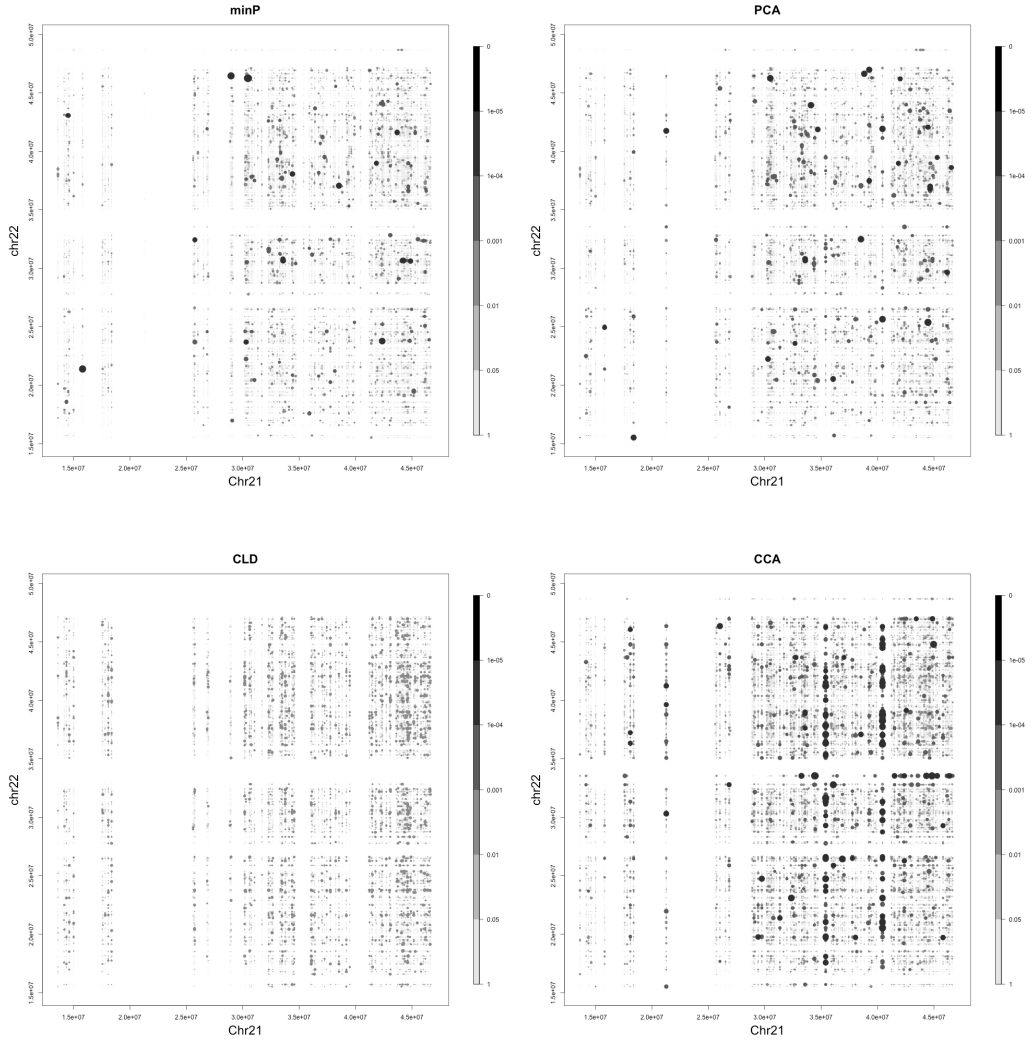
	BB	Bb	bb
AA	1.00	**1.35**	**0.68**
Aa	1.12	1.04	**1.69**
aa	1.08	1.14	1.19

Odds ratio relative to AA/BB

SNP1	SNP2	PLINK	Logit	TIH	BOOST	IndOR	nNA
SNP_A-4215965 41101820	SNP_A-1971082 36986784	1.59	39.92	4.49	39.92	40.09	70
		0.11	4.5×10^{-8}	0.03	4.5×10^{-8}	4.1×10^{-8}	

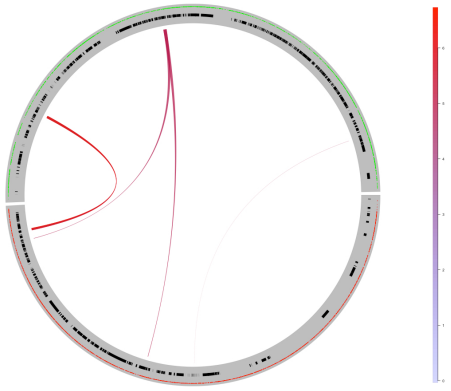
Heatmap visulisation for Gene-based interaction

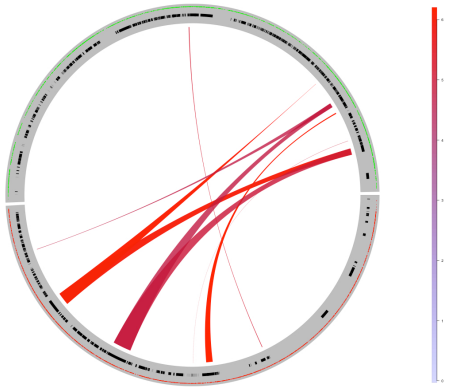
- Impact of the marginal effect for CCA

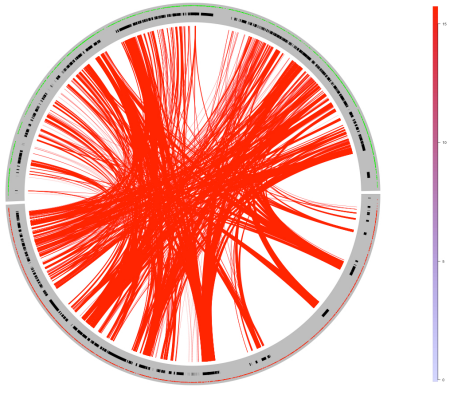


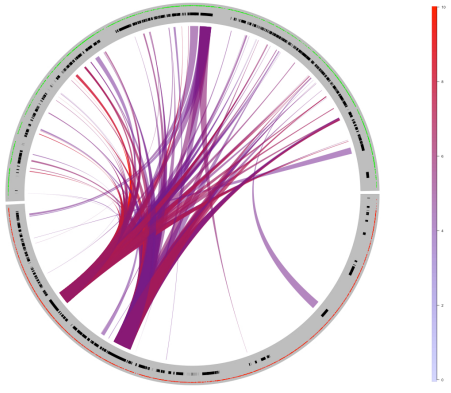
Circos visualisation for Gene-based interaction

- Significance threshold: 10^{-5}









Outline

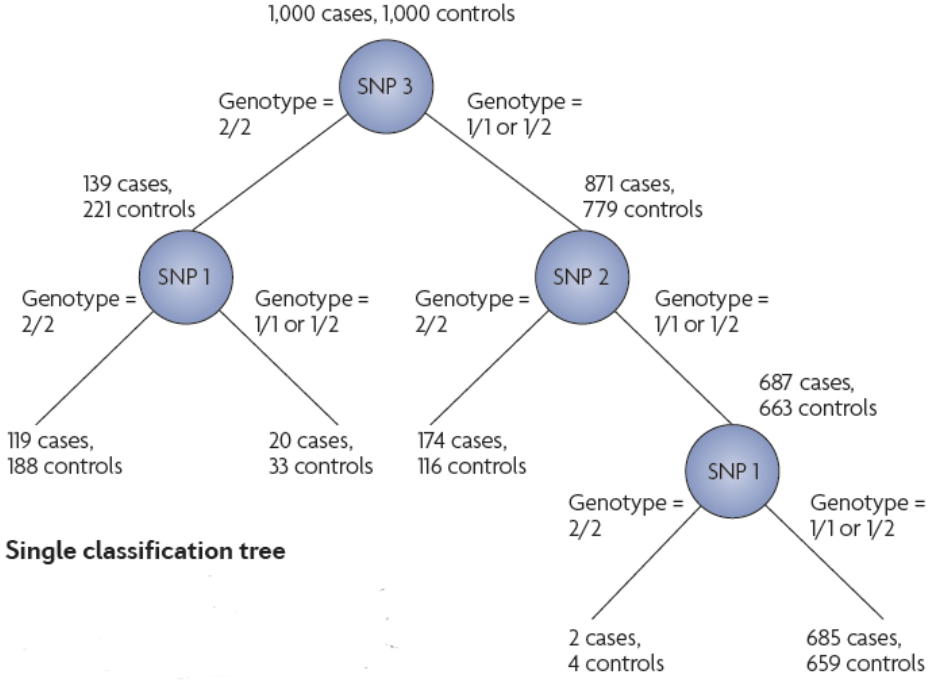
Other point-of-views

- Extensive research have been made to propose "heuristic" approaches that can detect **non-linear interaction** as well as **higher-order interaction**.
- **Data mining methods**:
  - MDR: Multifactor dimensionality reduction (Ritchie *et al.*, AJHG, 2001)
- **Machine learning methods**:
  - CART, Random Forest (Random jungle, Schwartz *et al.*, 2008)
  - TunedRelieF (Moore *et al.*, 2007)
- All these methods are known to suffer from **overfitting**.

MDR

- Rather than testing for interaction, MDR seeks to identify **combinations of loci that influence a disease outcome**, possibly by interactions or by main effects. (measure of heterogeneity)
- MDR has been used to identify potential interacting loci in several diseases:
  - breast cancer, type 2 diabetes, rheumatoid arthritis...
  - To date **no replication has confirmed those results**
- User-friendly software: `http://www.epistasis.org/software.html`
- Drawbacks
  - Too easy? (not necessarily a drawback)
  - **Sensitive to unbalanced case/control ratio**

Recursive partitioning



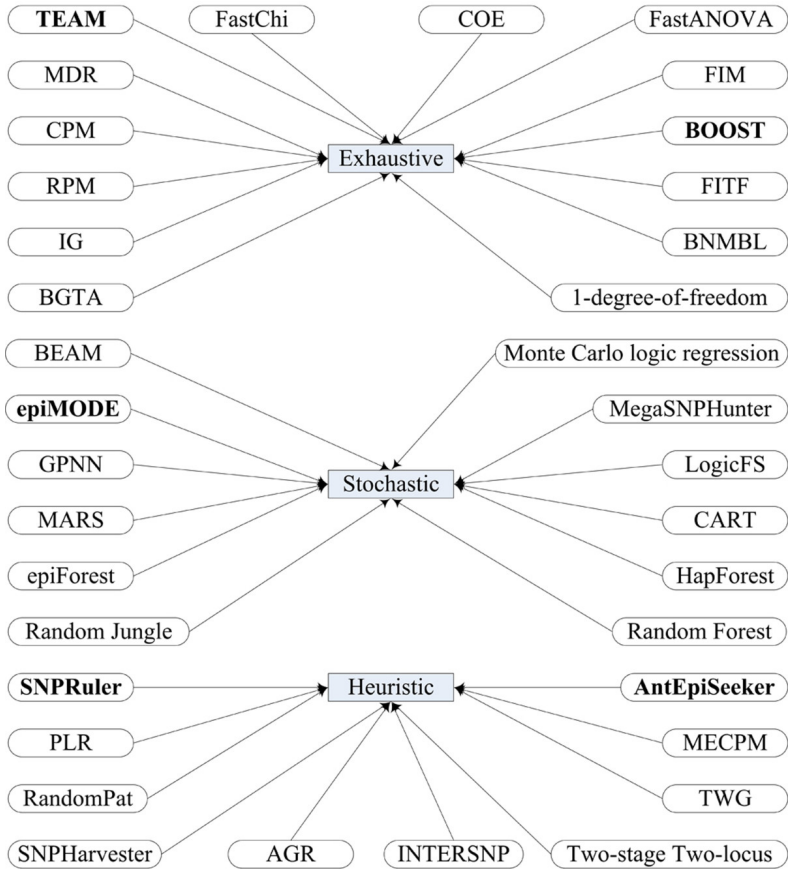
Single classification tree

- Splitting is made using **Gini impurity criterion**
- Substantial improvements in classification accuracy can result from growing an ensemble of trees: **random forests**
- `http://randomjungle.com`

TunedRelieF

- Uses a **measure of proximity between individuals**
- Allows for determining the **nearest neighbours** of each individual from within their own phenotype class and from within the opposite phenotype class.
- Implemented in the **MDR java applet**.

A large number of methods...



Shang *et al.* BMC Bioinformatics, 2011.

Take home messages

- Two **genetic levels** for Gene-Gene interaction:
  - SNP level
  - Gene (or SNP-set) level
- Statistical methods are **complementary**
- The **challenge** of detecting interaction:
  - The number of possible interaction models is huge
  - Computation performance vs. statistical significance

Thank you for your attention!