

# Network-guided feature selection in high-dimensional genomic data

Chloé-Agathe Azencott & Vivien Goepp

Center for Computational Biology (CBIO)

Mines ParisTech - Institut Curie - INSERM U900

PSL Research University & Pr[AI]rie, Paris, France

April 1, 2021 – ML in Genomics

<http://cazencott.info>

chloe-agathe.azencott@mines-paristech.fr

@cazencott

# About Chloé

**Janvier 2020    Habilitation à Diriger des Recherches**

**Depuis 2019    Chaire tremplin** à Pr Alrie  
Institut 3IA.

**Depuis 2013    Chargée de recherche** puis **Maîtresse-assistante** au CBIO

**2011–2013    Chargée de recherche post-doctorante**

Machine Learning for Computational Biology, MPI Tübingen (Allemagne).

**2005–2010    Doctorante**

Institute for Genomics and Bioinformatics, University of California Irvine (USA).

**2002–2005    Double diplôme Ingénieur & Master**

École Nationale Supérieure des Télécommunications de Bretagne  
Informatique et Mathématiques.

# About me

## 2020–2021 Post-doctorant

Center for Computational Biology, Mines Paristech & Institut Curie.

## 2016–2019 Doctorant

Université Paris-Descartes (*Université de Paris*).

## 2013–2016 Double diplôme Ingénieur & Master

Supélec

Master in Statistics

# What is this course about?

- ▶ Genome-wide association studies (GWAS)
- ▶ *Problem:* What genome loci are associated with a disease?
- ▶ Applied to *complex* diseases.

# Common disease common variant hypothesis

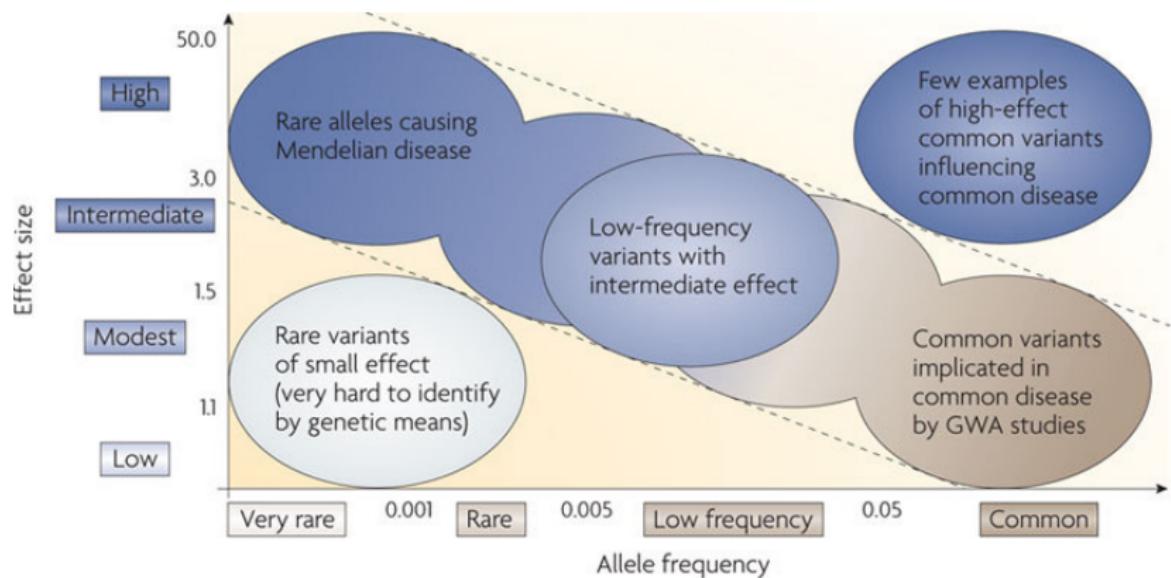
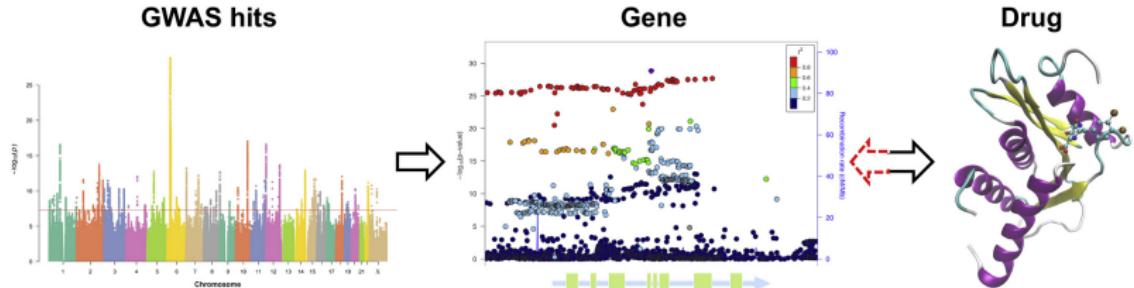


Image source: [Ant+10]

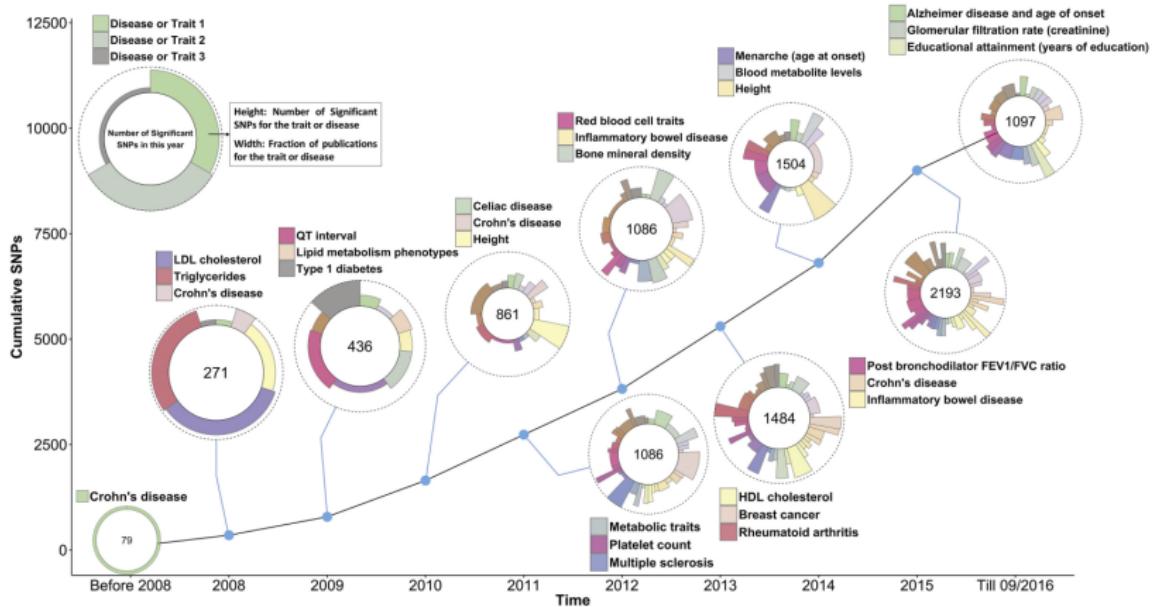
# From GWAS discoveries to drugs



Trait	Gene with GWAS hits	Known or candidate drug
Type 2 Diabetes	<i>SLC30A8/KCNJ11</i>	ZnT-8 antagonists/Glyburide
Rheumatoid Arthritis	<i>PADI4/IL6R</i>	BB-Cl-amidine/Tocilizumab
Ankylosing Spondylitis(AS)	<i>TNFR1/PTGER4/TYK2</i>	TNF-inhibitors/NSAIDs/fostamatinib
Psoriasis(Ps)	<i>IL23A</i>	Risankizumab
Osteoporosis	<i>RANKL/ESR1</i>	Denosumab/Raloxifene and HRT
Schizophrenia	<i>DRD2</i>	Anti-psychotics
LDL cholesterol	<i>HMGCR</i>	Pravastatin
AS, Ps, Psoriatic Arthritis	<i>IL12B</i>	Ustekinumab

Ref: [visscher2017]

# GWAS SNP-trait discovery timeline



Ref: [visscher2017]

# Genome-Wide Association Studies (GWAS)

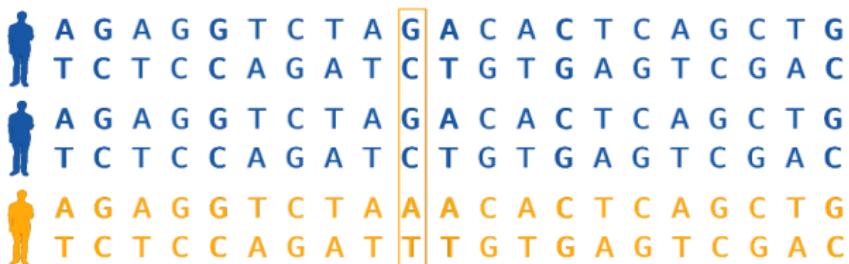


Image courtesy V. Bellón.

Which SNPs explain the phenotype?

$p = 10^5 - 10^7$  SNPs

Single Nucleotide Polymorphisms

Data:  $(X, y) \in \{0, 1, 2\}^{n \times p} \times \mathcal{Y}^n$

$n = 10^3 - 10^5$  samples

$\mathcal{Y} = \{-1, 1\}$  or  $\mathbb{R}$

# Guilt by association

Use **Linkage disequilibrium**

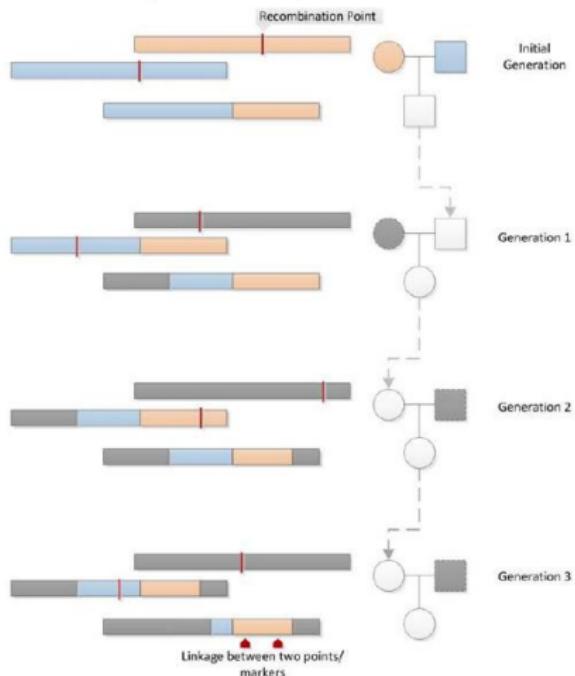
500 000 – 1M **tag SNPs** cover the entire genome.



Image: Francis Collins 2008

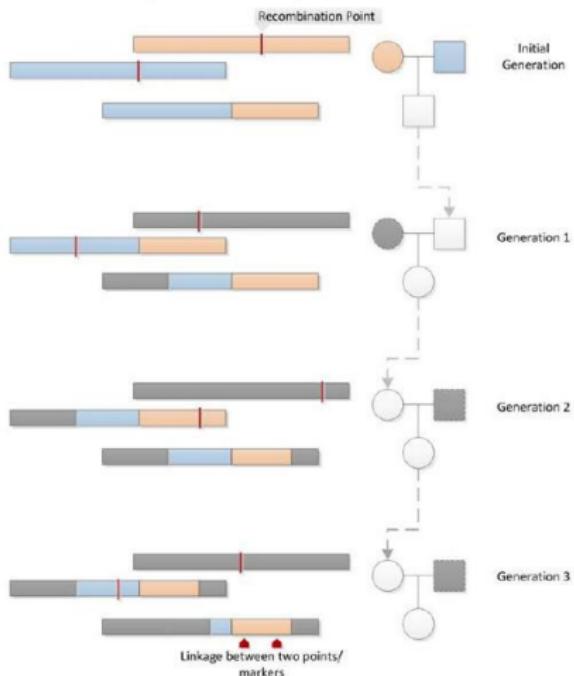
# Linkage disequilibrium

Linkage Within A Family

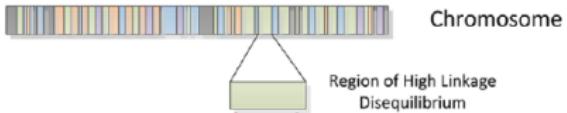


# Linkage disequilibrium

Linkage Within A Family



## Indirect Association



# SNP genotyping

## SNP microarray



## Whole-Genome Sequencing



- ▶ Cheap
- ▶ Small data files
- ▶ 500K – 5M loci captured.
- ▶ More expensive
- ▶ Large data files
- ▶ ~ 3B loci captured (incl. rare variants).



# **GWAS data** in practice

# Quality control

- ▶ **Missing data**
  - **Discard** SNPs with high missingness rate
  - **Impute** missing SNPs, e.g. with IMPUTE2 [howie2009] or BEAGLE [browning2018].
- ▶ **MAF: Minor Allele Frequency**

Focus on **common** variants with MAF > 0.01 or 0.05

  - Common disease common variant hypothesis
  - Rare variants are more likely to be technical artifacts
  - Lack of statistical power.

# Quality control

## ► Hardy-Weinberg equilibrium

- Assume random mating.
- If A has frequency  $p$ , and a has frequency  $(1-p)$ :
  - AA has frequency  $p \times p$
  - Aa/aA has frequency  $2 p \times (1-p)$
  - aa has frequency  $(1-p) \times (1-p)$
- Compare observed with expected with a  $\chi^2$ -test.

$$p = \frac{\#AA + \#Aa/2}{\#AA + \#Aa + \#aa}$$

- Only keep SNPs for which observed and expected are **not significantly different**.

# Quality control

- ▶ **Hardy-Weinberg equilibrium**

- Example: Consider a SNP that has genotype GG for 1 500 individuals, GA for 210 individuals and AA for 5 individuals.

# Quality control

## ► Hardy-Weinberg equilibrium

- Example: Consider a SNP that has genotype GG for 1 500 individuals, GA for 210 individuals and AA for 5 individuals.
- The total number of individuals is  $N = 1\,500 + 210 + 5 = 1715$ .
- The frequency  $p$  of major allele G is

$$p = \frac{1\,500 + 0.5 \times 210}{N} = 93.5\%$$

- The expected and observed frequencies for each genotype are hence:

	GG	GA	AA
Observed	1 500	210	5
Expected	$N \times p^2$ 1499	$N \times 2p(1 - p)$ 209	$N \times (1 - p)^2$ 7

# Quality control

	GG	GA	AA
Observed	1 500	210	5
Expected	$N \times p^2$	$N \times 2 p (1-p)$	$N \times (1-p)^2$
	1499	209	7

- The value of the  $\chi^2$  test statistic is

$$\sum_{i=GG,GA,AA} \frac{(O_i - E_i)^2}{E_i} = 0.576$$

- The corresponding p-value (1df) is 0.44, which is not significant at  $\alpha = 0.05$ .  
Hence the null is not rejected, and the SNP is considered to be in HWE.

# Encoding SNPs

AA, Aa and aa must be represented as numbers.

- ▶ **Allelic dosage / Codominance** model:

AA = 0      Aa = 1      aa = 2

- ▶ **Dominance** model:

AA = 0      Aa = 1      aa = 1

- ▶ **Recessive** model:

AA = 0      Aa = 0      aa = 1

- ▶ **Dummy encoding:**

AA = 0, 0      Aa = 0, 1      aa = 1, 1

Covers all above models, but with twice as many variables.

# Qualitative GWAS

Binary phenotype, i.e. case/controls encoded as 1/0.

- ▶ Is the SNP significantly associated with the phenotype?
- ▶ Contingency table

	AA	Aa	aa
Cases			
Ctrls			

	0	1
Cases	a	b
Ctrls	c	d

Statistical tests:  $\chi^2$ , Cochran-Armitage trend test, etc.

- ▶ Logistic regression

$$\text{logit}(P(\text{case}|x)) = \beta_0 + \beta_1 x$$

Is  $\hat{\beta}_1$  significantly different from 0?

Wald test: compare  $\frac{\hat{\beta}_1^2}{\text{Var}(\hat{\beta}_1)}$  to a  $\chi^2$  distribution.

# Interlude – Fitting a regression

- **Univariate linear regression:**

- **Model:**  $f(x) = \beta_0 + \beta_1 x$
- **Data:**  $n$  samples  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- **Fitting** = finding  $\beta_0$  and  $\beta_1$ : minimize the sum of squared errors  
**Least squares fit** (Gauss/Legendre)

$$\hat{\beta}_0, \hat{\beta}_1 = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - f(x_i))^2$$

$$\hat{\beta}_0, \hat{\beta}_1 = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

- Can be solved **analytically**.

# Interlude – Fitting a regression

- ▶ **Multivariate linear regression:**
  - ▶ **Model:**  $f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
  - ▶ **Data:**  $n$  samples  $((x_{11}, x_{12}, \dots, x_{1p}), y_1), \dots, ((x_{n1}, x_{n2}, \dots, x_{np}), y_n)$
  - ▶ **Fitting** = finding  $\beta_0, \beta_1, \dots, \beta_p$ : minimize the sum of squared errors  
**Least squares fit** (Gauss/Legendre)

$$\hat{\beta}_0, \hat{\beta}_1 = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^n \left( y_i - \left( \beta_0 + \sum_{j=1}^p \beta_j x_{ij} \right) \right)^2$$

- ▶ Can be solved **analytically** or by **gradient descent**.

# Interlude – Fitting a regression

- ▶ **Multivariate logistic regression:**
  - ▶ **Model:**  $f(x) = \text{logistic}(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$ 
    - ▶  $\text{logistic}(u) = \frac{1}{1+\exp(-u)}$  transforms a number between  $-\infty$  and  $+\infty$  into a number between 0 and 1.
    - ▶  $f$  models the probability that  $y = 1$ .
  - ▶ **Data:**  $n$  samples  
 $((x_{11}, x_{12}, \dots, x_{1p}), y_1), \dots, ((x_{n1}, x_{n2}, \dots, x_{np}), y_n)$
  - ▶ **Fitting** = finding  $\beta_0, \beta_1, \dots, \beta_p$ : minimize the sum of **logistic** errors

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \sum_{i=1}^n \log(1 + \exp(-y_i f(x_i)))$$

$y_i = -1$  for controls, 1 for cases

- ▶ Can be solved by **gradient descent**.

# Qualitative GWAS

Binary phenotype, i.e. case/controls encoded as 1/0.

- ▶ What is the **effect** of the SNP on the phenotype?
- ▶ **Contingency table**

	AA	Aa	aa
Cases			
Ctrls			

	0	1
Cases	a	b
Ctrls	c	d

- ▶ **Odds-ratio**

$$\frac{\underbrace{\frac{P(0|\text{case})}{P(0|\text{ctrl})}}_{\text{odds of 0 in cases}}}{\underbrace{\frac{P(1|\text{case})}{P(1|\text{ctrl})}}_{\text{odds of 1 in cases}}} = \frac{ad}{bc}$$

# Quantitative GWAS

- ▶ Is the SNP **significantly associated** with the phenotype?
- ▶ **Linear regression**

$$y = \beta_0 + \beta_1 x$$

Is  $\hat{\beta}_1$  significantly different from 0?

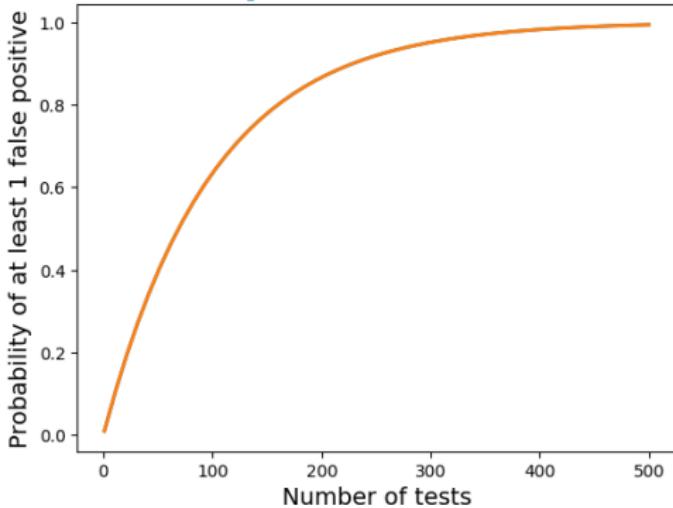
**Wald test:** compare  $\frac{\hat{\beta}_1^2}{\text{Var}(\hat{\beta}_1)}$  to a  $\chi^2$  distribution.

- ▶ What is the **effect** of the SNP on the phenotype?
- ▶ **Effect size:**  $\hat{\beta}_1$ .

# Multiple Hypothesis Testing

- ▶ Probability of having **at least one false positive**:

- For **one** test:  $\alpha$
- For **m** tests:  $1 - (1 - \alpha)^m$



- ▶ Controlling **Family-Wise Error Rate (FWER)**

$$\text{FWER} = P(\#\text{FP} \geq 1)$$

FP = number of false positives (Type I errors)

- ▶ **Bonferroni** correction: reject  $\mathcal{H}_0$  if  $p \leq \frac{\alpha}{m}$

# Controlling False Discovery Rate

- ▶ Instead of ensuring that  $P(\#\text{FP} \geq 1) \leq \alpha$ , control the **false discovery rate**:

$$\mathbb{E} \left( \frac{\#\text{FP}}{\#\text{predicted pos}} \right) \leq \alpha$$

- ▶ More power but also larger number of Type I errors than FWER procedures.

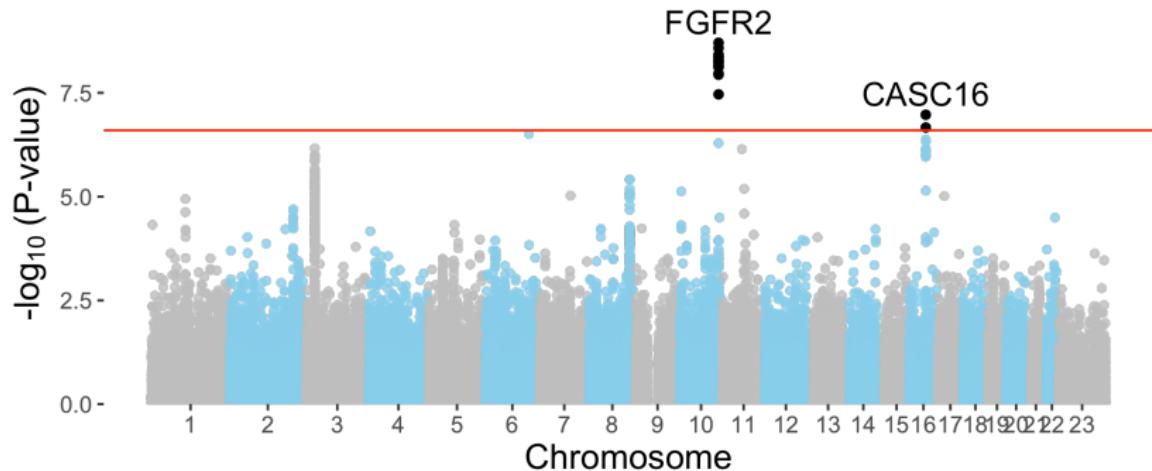
- ▶ **Benjamini-Hochberg:**

- ▶ Order all  $m$  p-values:  $p_1 \leq p_2 \leq \dots \leq p_m$
- ▶ Find the largest  $k$  such that  $p_k \leq \frac{k}{m}\alpha$
- ▶ Reject  $\mathcal{H}_0$  for tests  $1, 2, \dots, k$

- ▶ **Benjamini-Yekutieli**

- ▶ When tests are **not independent**
- ▶ Find the largest  $k$  such that  $p_k \leq \frac{k}{m \sum_{i=1}^m \frac{1}{i}}\alpha$

# Manhattan plots



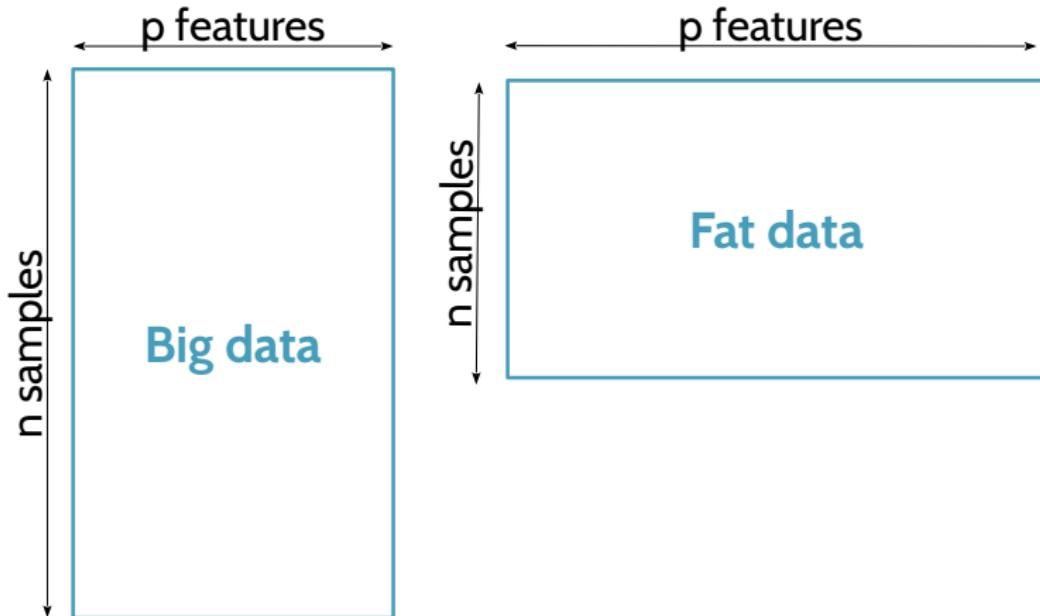
- ▶ -log<sub>10</sub>(p-value) for each SNP
- ▶ SNP ordered by genomic position (chr+bp)
- ▶ significance line.

The background of the slide shows a dark auditorium or lecture hall. In the foreground, there are several rows of red theater-style seats facing towards the right. Behind the seats, a large, closed, light-colored curtain covers the back wall.

# Large p, small n data

## A simple example

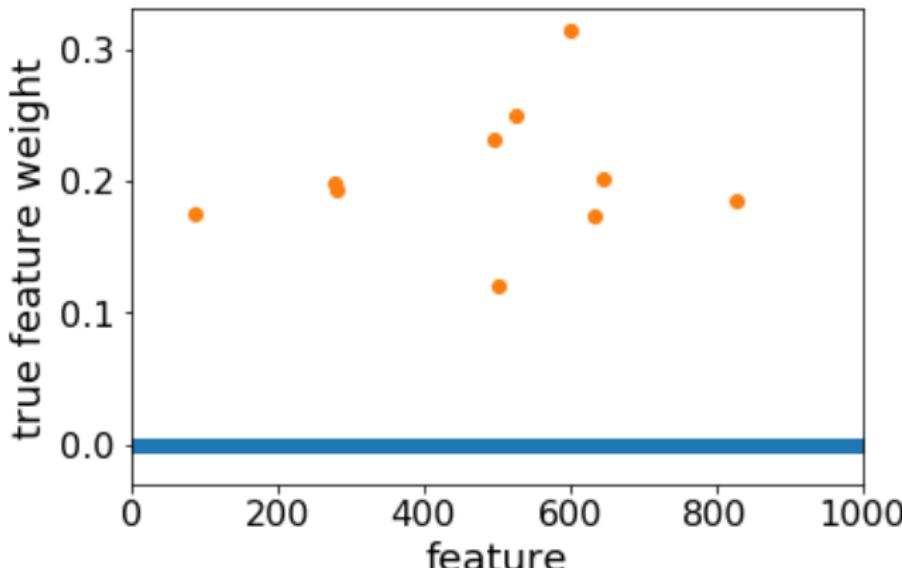
# High-dimensional data with low sample size



# Large p, small n data

**Simulation:** n=150, p=1000, 10 causal features.

$$y = \sum_{j=1}^p w_j x_j + \epsilon$$



# Simulation

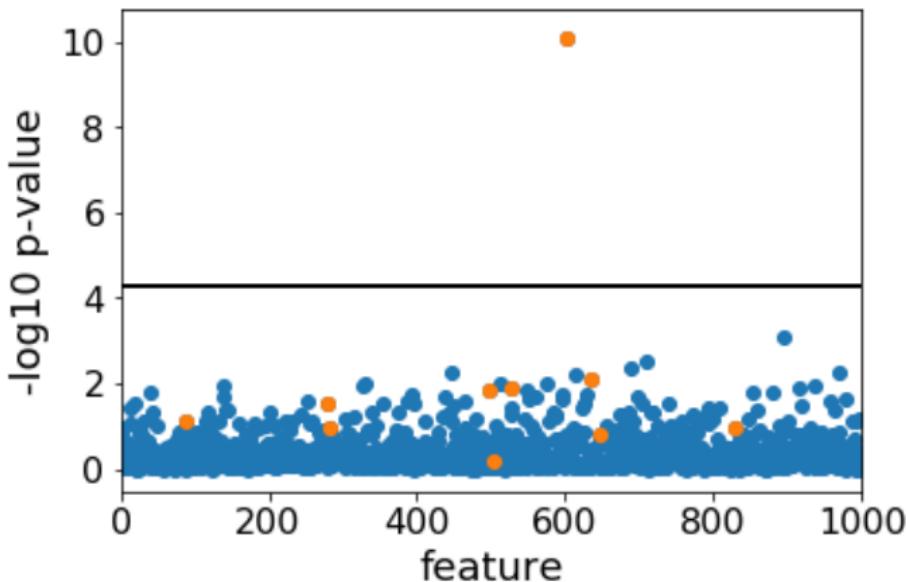
**t-test:** For each feature  $x_j$ ,

- fit  $y \sim w_j x_j + b_j$
- test whether  $w_j \neq 0$ .

# Simulation

**t-test:** For each feature  $x_j$ ,

- fit  $y \sim w_j x_j + b_j$
- test whether  $w_j \neq 0$ .

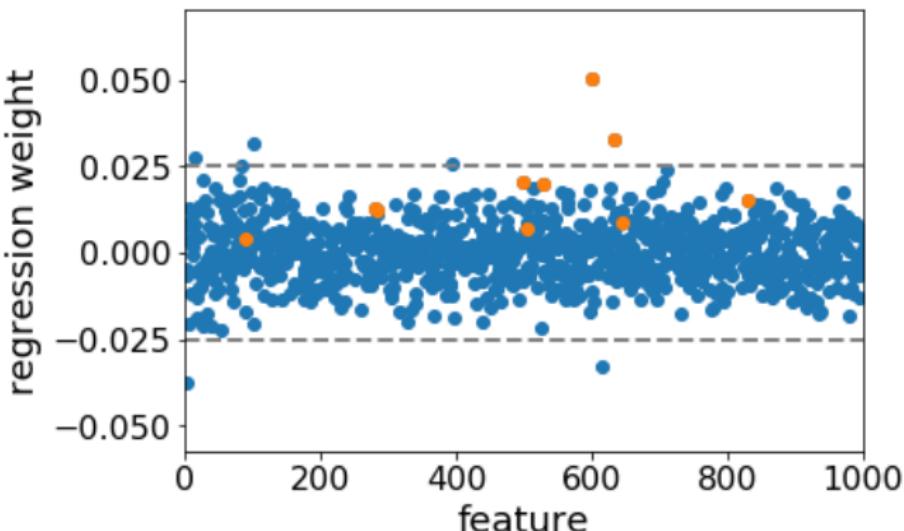


# Simulation

**Linear regression:** Fit  $y \sim \sum_{j=1}^p w_j x_j + b$ .

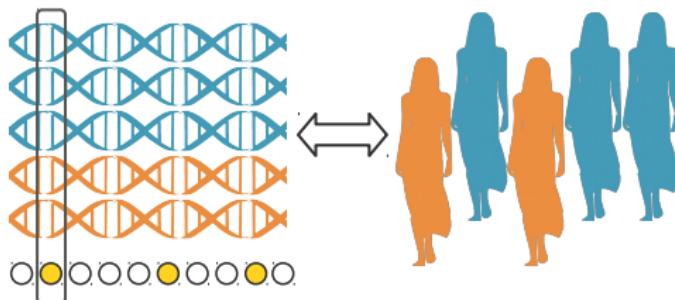
# Simulation

**Linear regression:** Fit  $y \sim \sum_{j=1}^p w_j x_j + b$ .

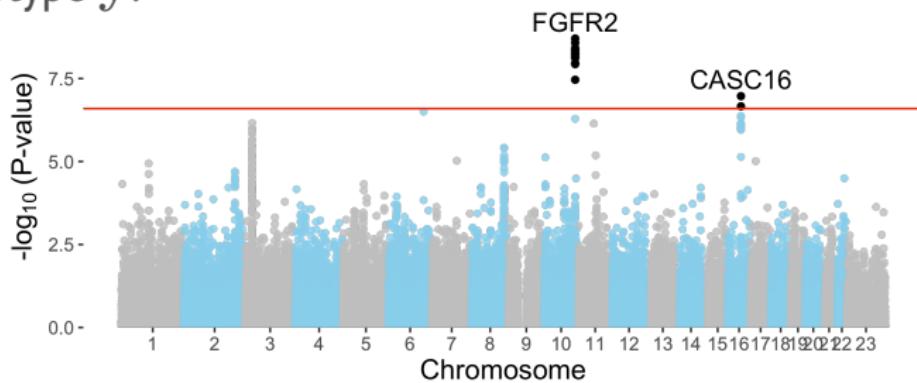


# Statistical challenges in GWAS

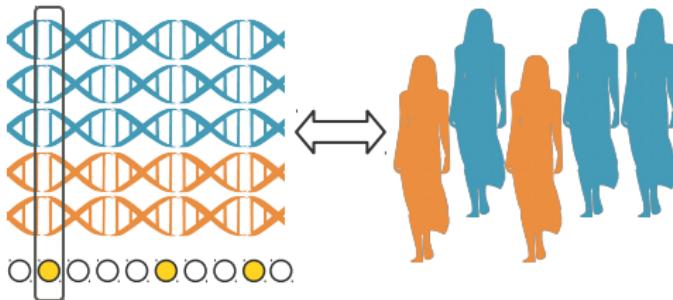
# State-of-the-art: Statistical tests



- ▶ **Statistical test** of association between **each SNP**  $x_j$  and the phenotype  $y$ .



# State-of-the-art: Statistical tests

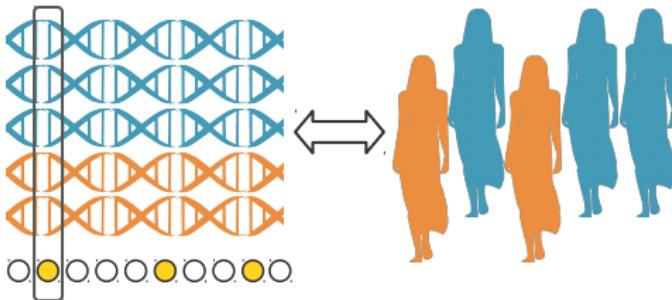


- ▶ **Statistical test** of association between **each SNP**  $x_j$  and the phenotype  $y$ .

## Limitations:

- ⌚ Lack of **statistical power**;

# State-of-the-art: Statistical tests



- ▶ **Statistical test** of association between **each SNP**  $x_j$  and the phenotype  $y$ .

## Limitations:

- (⌚) Lack of **statistical power**;
- (⌚) Consider SNPs **independently from each other**.

# Missing heritability

GWAS **fail to explain** most of the **inheritable variability** of complex traits.

Many possible reasons:

- non-genetic / non-SNP factors
- heterogeneity of the phenotype
- rare SNPs
- weak effect sizes
- **few samples in high dimension ( $p \gg n$ )**
- joint effects of **multiple SNPs**.

# Molecular signatures stability

- ▶ **Stability (robustness):** find similar answers on different data sets linked to the same biological question.
- ▶ **Example:** predicting apparition of distant metastasis at 5 years in breast cancer.
  - **2001:** a signature of **456** genes

[Sør+01]

# Molecular signatures stability

- ▶ **Stability (robustness):** find similar answers on different data sets linked to the same biological question.
- ▶ **Example:** predicting apparition of distant metastasis at 5 years in breast cancer.
  - **2001:** a signature of **456** genes [Sør+01]
  - **2002:** a signature of **70** genes [VV+02]

# Molecular signatures stability

- ▶ **Stability (robustness):** find similar answers on different data sets linked to the same biological question.
- ▶ **Example:** predicting apparition of distant metastasis at 5 years in breast cancer.
  - **2001:** a signature of **456** genes [Sør+01]
  - **2002:** a signature of **70** genes [VV+02]
  - Overlap: **17** genes [ED+05]

# Molecular signatures stability

- ▶ **Stability (robustness):** find similar answers on different data sets linked to the same biological question.
- ▶ **Example:** predicting apparition of distant metastasis at 5 years in breast cancer.
  - **2001:** a signature of **456** genes [Sør+01]
  - **2002:** a signature of **70** genes [VV+02]
  - Overlap: **17** genes [ED+05]
  - Most **random** signatures of 70 genes predict outcome well [VDD11]

# Contributions

## Feature selection in high-dimensional genomic data

1. Using **biological networks** to integrate **prior knowledge**.
2. Considering **multiple related phenotypes** at once.
3. Modeling **nonlinearities**.

# Contributions

## Feature selection in high-dimensional genomic data

1. Using **biological networks** to integrate **prior knowledge**.
2. Considering **multiple related phenotypes** at once.
3. Modeling **nonlinearities**.

(But we won't talk about this today.)



# Machine learning for GWAS

# Reducing p

# Integrating prior knowledge

## ► Regularization

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{\sum_{i=1}^n \left( y^i - \sum_{j=1}^p w_j x_{ij} \right)^2}_{\text{loss}}$$

# Integrating prior knowledge

## ► Regularization

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \sum_{i=1}^n \underbrace{\left( y^i - \sum_{j=1}^p w_j x_{ij} \right)^2}_{\text{loss}}$$



# Integrating prior knowledge

## ► Regularization

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{\sum_{i=1}^n \left( y^i - \sum_{j=1}^p w_j x_{ij} \right)^2}_{\text{loss}} + \lambda \underbrace{\Omega(w_1, w_2, \dots, w_p)}_{\text{regularizer}}$$



# Integrating prior knowledge

## ► Regularization

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{\sum_{i=1}^n \left( y^i - \sum_{j=1}^p w_j x_{ij} \right)^2}_{\text{loss}} + \lambda \underbrace{\Omega(w_1, w_2, \dots, w_p)}_{\text{regularizer}}$$



# Integrating prior knowledge

## ► Regularization

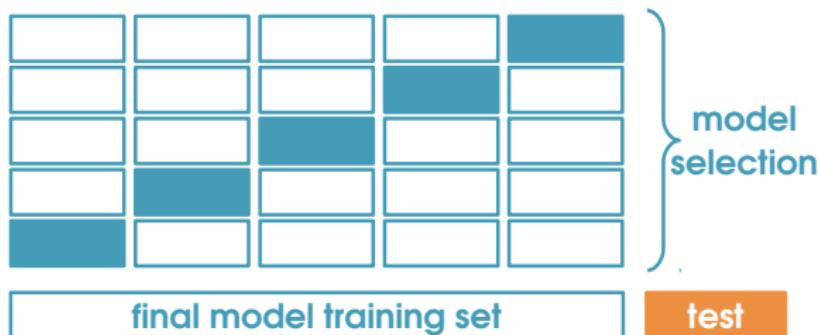
$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{\sum_{i=1}^n \left( y^i - \sum_{j=1}^p w_j x_{ij} \right)^2}_{\text{loss}} + \lambda \underbrace{\Omega(w_1, w_2, \dots, w_j)}_{\text{regularizer}}$$

# Integrating prior knowledge

- **Regularization**

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{\sum_{i=1}^n \left( y^i - \sum_{j=1}^p w_j x_{ij} \right)^2}_{\text{loss}} + \lambda \underbrace{\Omega(w_1, w_2, \dots, w_j)}_{\text{regularizer}}$$

- Set the **regularization hyperparameter** by **cross-validation**



# Integrating prior knowledge

- **Regularization**

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{\sum_{i=1}^n \left( y^i - \sum_{j=1}^p w_j x_{ij} \right)^2}_{\text{loss}} + \lambda \underbrace{\Omega(w_1, w_2, \dots, w_j)}_{\text{regularizer}}$$

- **Prior knowledge:** relatively few features are relevant.

# Integrating prior knowledge

- **Regularization**

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{\sum_{i=1}^n \left( y^i - \sum_{j=1}^p w_j x_{ij} \right)^2}_{\text{loss}} + \lambda \underbrace{\Omega(w_1, w_2, \dots, w_j)}_{\text{regularizer}}$$

- **Prior knowledge:** relatively few features are relevant.
- **Lasso** [Tib94]

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{\sum_{i=1}^n \left( y^i - \sum_{j=1}^p w_j x_{ij} \right)^2}_{\text{loss}} + \lambda \underbrace{\sum_{j=1}^p |w_j|}_{\text{sparsity}}$$

# Integrating prior knowledge

- **Regularization**

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{\sum_{i=1}^n \left( y^i - \sum_{j=1}^p w_j x_{ij} \right)^2}_{\text{loss}} + \lambda \underbrace{\Omega(w_1, w_2, \dots, w_j)}_{\text{regularizer}}$$

- **Prior knowledge:** relatively few features are relevant.
- **Lasso** [Tib94]

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{\sum_{i=1}^n \left( y^i - \sum_{j=1}^p w_j x_{ij} \right)^2}_{\text{loss}} + \lambda \underbrace{\sum_{j=1}^p |w_j|}_{\text{sparsity}}$$

- **Sparsity:** many features are assigned a weight of 0.  
They can be removed from the model.

# Simulation

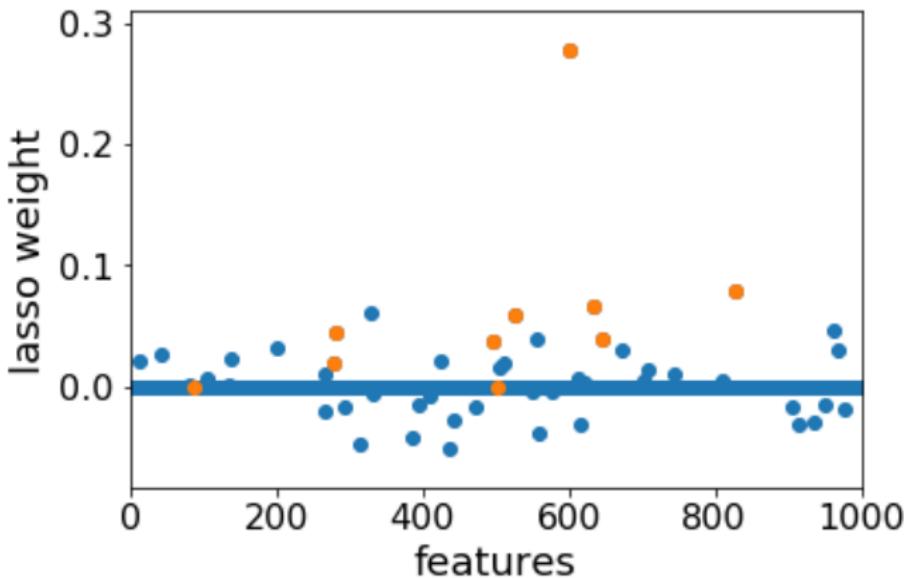
**Lasso regression:** Minimize

$$\sum_{i=1}^n \left( y^i - \sum_{j=1}^p w_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |w_j|$$

# Simulation

Lasso regression: Minimize

$$\sum_{i=1}^n \left( y^i - \sum_{j=1}^p w_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |w_j|$$



# Stability

- ▶ Lasso tends to be **unstable**:
  - ▶ Randomly picks one of several correlated variables.
  - ▶ Different results on similar data sets.
- ▶ **Elastic net** combines  $\ell_2$  shrinkage with lasso [ZH05]

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \sum_{i=1}^n \left( y^i - \sum_{j=1}^p w_j x_{ij} \right)^2 + \lambda \left( (1-\alpha) \sum_{j=1}^p |w_j| + \alpha \sum_{j=1}^p w_j^2 \right)$$

- ▶ **Stability selection** [MB10]
  - ▶ Repeat on multiple bootstrap samples of the data.
  - ▶ Only keep the features that are selected often.

But what about **p-values?**

# But what about p-values?

- ▶ Do you really want a p-value?

# But what about p-values?

- ▶ Do you really want a p-value?
- ▶ What a p-value is: the probability to observe a value as extreme as the one you obtain, under the null.

# But what about p-values?

- ▶ Do you really want a p-value?
  - ▶ What a p-value is: the probability to observe a value as extreme as the one you obtain, under the null.
  - ▶ What a p-value is not: all-powerful magic, nor biological evidence.

# But what about p-values?

- ▶ **Do you really want a p-value?**
  - ▶ What a p-value is: **the probability to observe a value as extreme** as the one you obtain, under the null.
  - ▶ What a p-value is not: **all-powerful magic**, nor **biological evidence**.  
“The p-value was never intended to be a substitute for scientific reasoning.” [WL+16]

# But what about p-values?

- ▶ **Do you really want a p-value?**
    - ▶ What a p-value is: **the probability to observe a value as extreme** as the one you obtain, under the null.
    - ▶ What a p-value is not: **all-powerful magic**, nor **biological evidence**.  
“The p-value was never intended to be a substitute for scientific reasoning.” [WL+16]
- [Ioa05 ; Nuz14 ; Hea+15 ; Hol18]

# But what about p-values?

- ▶ **Do you really want a p-value?**
  - ▶ What a p-value is: **the probability to observe a value as extreme** as the one you obtain, under the null.
  - ▶ What a p-value is not: **all-powerful magic**, nor **biological evidence**.  
“The p-value was never intended to be a substitute for scientific reasoning.” [WL+16]
- ▶ For the **lasso**, possible but computationally intensive [Loc+14; Lee+16]

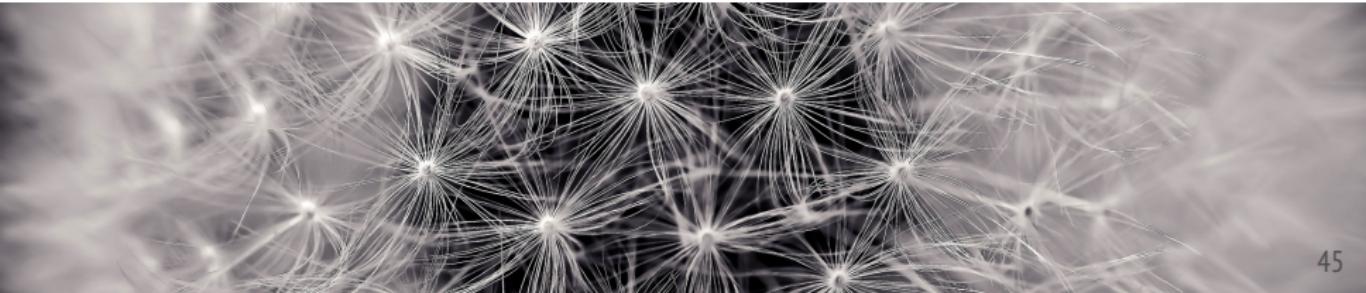
# Integrating prior knowledge

Use prior knowledge as a **constraint** on the selected features

- **Consistant** with previously established knowledge
- Increases **interpretability** and **statistical power**.

Prior knowledge can be represented as **structure**:

- Linear structure of the DNA
- **Groups:** e.g. pathways
- **Networks:** molecular, 3D structure.

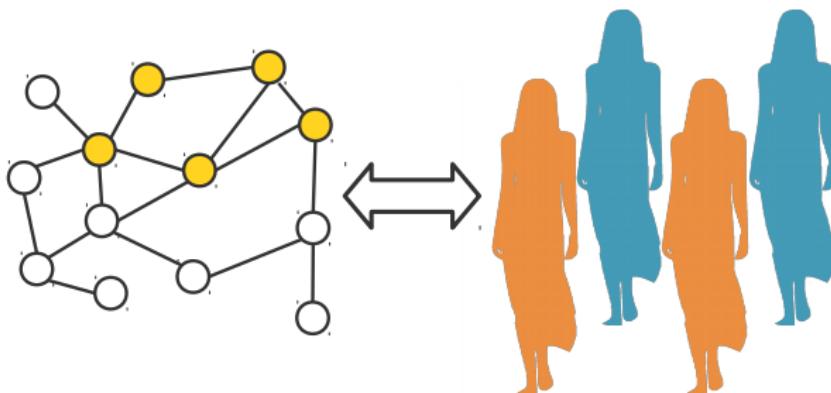


# Network view of complex diseases

- ▶ Biology emerges from the **interplay** of multiple entities.  
DNA, RNA, proteins, metabolites + environment
- ▶ **Biological networks:**
  - ▶ **Nodes:** genes, proteins, etc.
  - ▶ **Edges:** biological relationships between two nodes.
  - ▶ E.g. **protein-protein interaction networks, gene regulatory networks, gene co-expression networks**, etc.
- ▶ **Biological networks** help understanding disease
  - ▶ Understand mutations **in their genomic context**.
  - ▶ Multiple ways of producing the same symptoms.
  - ▶ **Local hypothesis:** genes involved in disease interact with each other.  
[VCB11; Bar+12; Fur13; Cow+17; Hua+18]

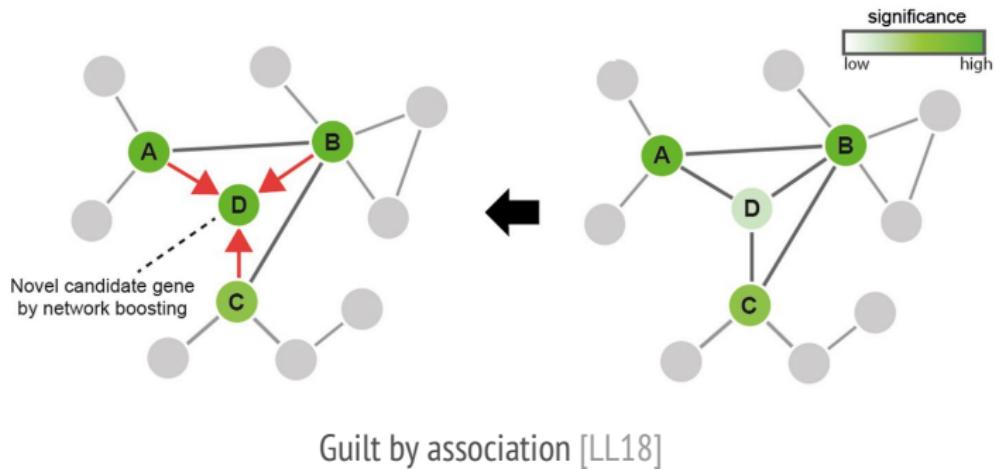
# Network-guided biomarker discovery

- Goal: Find a **set of explanatory features** compatible with a **given network** structure.



# Network-guided GWAS

- ▶ Map SNPs to genes according to genomic position.
- ▶ Combine SNP p-values into gene p-values.
- ▶ Find networks enriched in genes with low p-values.



# Finding high-scoring modules in PPI networks

## Transform SNP p-values into gene p-values

- ▶ **Map SNP to genes:** position on the genomic sequence.
- ▶ **Aggregate p-values:** VEGAS2 [MM15]

## Find high-scoring modules

- ▶ **dmGWAS:** greedy “seed and extend” heuristic [JZ14]
- ▶ **heinz:** Prize-Collecting Steiner Tree Problem [Dit+08]
- ▶ **HotNet2:** based on a heat diffusion process [Lei+15]
- ▶ **LEAN:** focus on star subnetworks [Gwi+17]

# dmGWAS

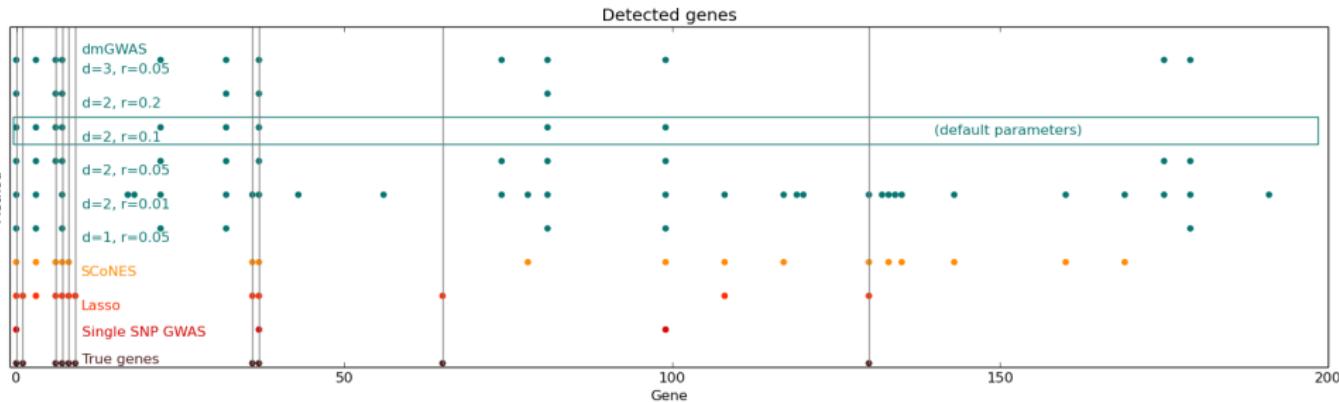
- ▶ Use **biological networks** to analyze the output of a GWAS
- ▶ Find **modules** (subnetworks) enriched in small p-values
- ▶ SNPs p-values → genes p-values
  - 20kb window, min p-value
- ▶ module z-score:  $Z(\mathcal{S}) = \frac{\sum_{i \in \mathcal{S}} z_i}{\sqrt{|\mathcal{S}|}}$
- ▶ greedy search strategy:
  - each gene = seed
  - add neighbor  $i$  (within distance  $d$ ) if  $Z(\mathcal{S} \cup i) \geq Z(\mathcal{S}) \times (1 + r)$
  - keep modules with more than 5 nodes
  - $Z(\mathcal{S}) \rightarrow Z_N(\mathcal{S}) = \frac{Z(\mathcal{S}) - \mu}{\sigma}$  (compare to random modules of size  $|\mathcal{S}|$ )
  - only keep top 1% of modules (according to  $Z_N$ )
  - only keep modules that are significantly associated with the phenotype.

# dmGWAS

**Simulation:**  $n=100$ ,  $p=1000$ , 10 causal SNPs,  $y = Xw + e$ .

SNPs belong to 200 genes, connected on a Barabási-Albert small-world network.

- Ideal lasso situation (simulated according to linear model).
- dmGWAS:
  - which genes are selected **depends a lot on the parameters**
  - **high FDR**, rather **low power**.



# Integrating prior network knowledge

- ▶ **Network-constrained lasso**

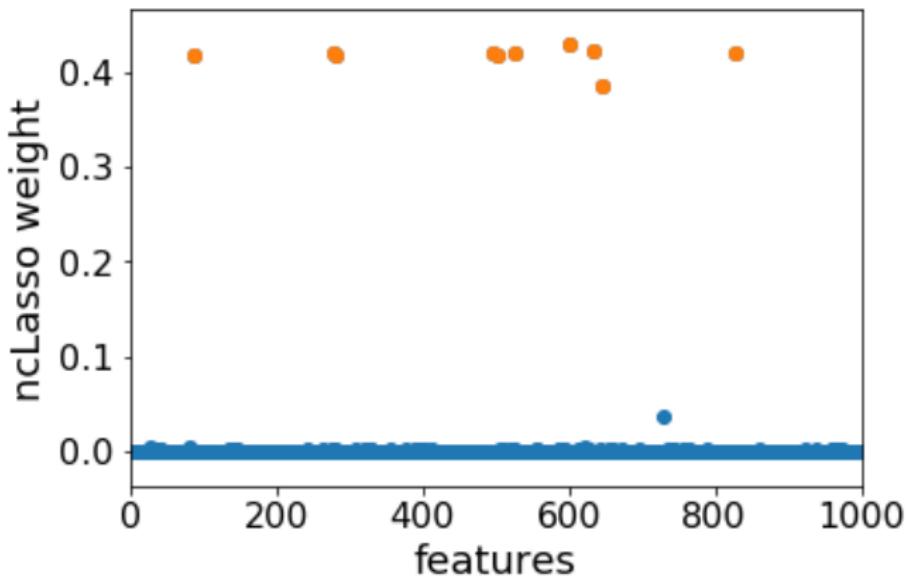
[LL08 ; LL10]

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{\sum_{i=1}^n \left( y^i - \sum_{j=1}^p w_j x_{ij} \right)^2}_{\text{loss}} + \lambda \underbrace{\sum_{j=1}^p |w_j|}_{\text{sparsity}} + \eta \underbrace{\sum_{j=1}^p \sum_{k=1}^p w_j L_{jk} w_k}_{\text{connectivity}}$$

- ▶ **Graph Laplacian**  $L$  ensures  $w$  varies **smoothly** on the network.

# Simulation

## Network-constrained lasso



# Integrating prior network knowledge

- **Regularized relevance** Set  $\mathcal{V}$  of  $p$  variables.

$$\arg \max_{\mathcal{S} \subseteq \mathcal{V}} \underbrace{R(\mathcal{S})}_{\text{relevance}} - \lambda \underbrace{\Omega(\mathcal{S})}_{\text{regularizer}}$$

- **Network-regularized relevance**

$$\arg \max_{\mathcal{S} \subseteq \mathcal{V}} \underbrace{R(\mathcal{S})}_{\text{relevance}} - \lambda \underbrace{|\mathcal{S}|}_{\text{sparsity}} - \eta \underbrace{\sum_{j \in \mathcal{S}} \sum_{k \notin \mathcal{S}} W_{jk}}_{\text{connectivity}}$$

# Integrating prior network knowledge

- **Regularized relevance** Set  $\mathcal{V}$  of  $p$  variables.

$$\arg \max_{\mathcal{S} \subseteq \mathcal{V}} \underbrace{R(\mathcal{S})}_{\text{relevance}} - \lambda \underbrace{\Omega(\mathcal{S})}_{\text{regularizer}}$$

- **Network-regularized relevance**

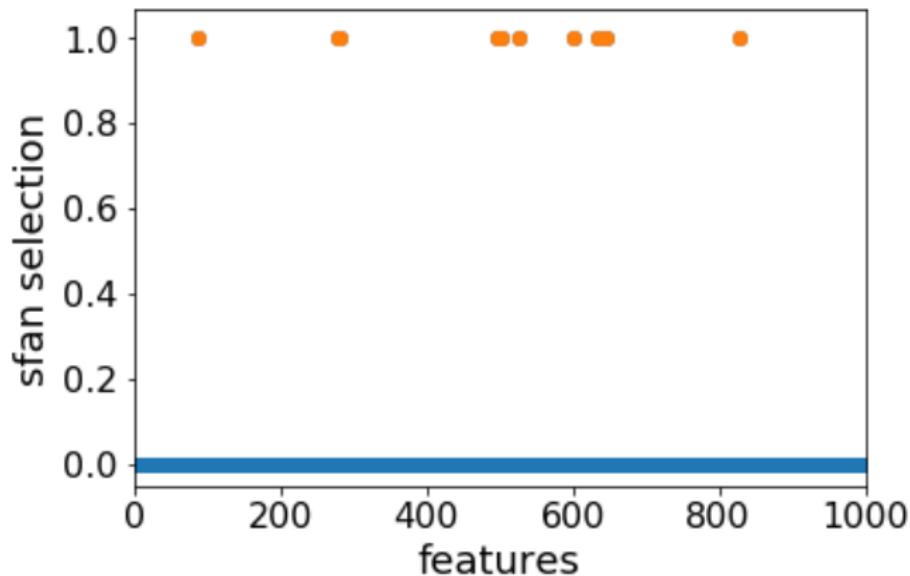
$$\arg \max_{\mathcal{S} \subseteq \mathcal{V}} \underbrace{R(\mathcal{S})}_{\text{relevance}} - \lambda \underbrace{|\mathcal{S}|}_{\text{sparsity}} - \eta \underbrace{\sum_{j \in \mathcal{S}} \sum_{k \notin \mathcal{S}} W_{jk}}_{\text{connectivity}}$$

## SConES: Selecting Connected Explanatory SNPs.

C.-A. Azencott, D. Grimm, et al. **Efficient network-guided multi-locus association mapping with graph cuts.** Bioinformatics 2013  
<https://github.com/chagaz/scones> Bioconductor/martini

# Simulation

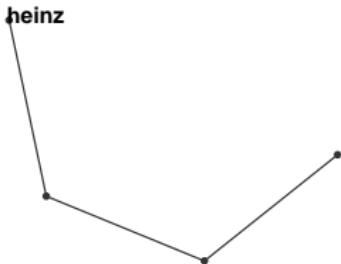
## SConES



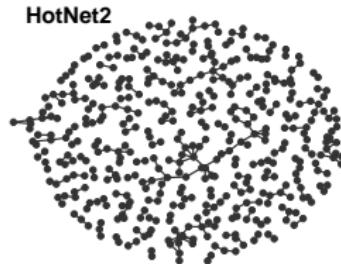
# Other network-based methods



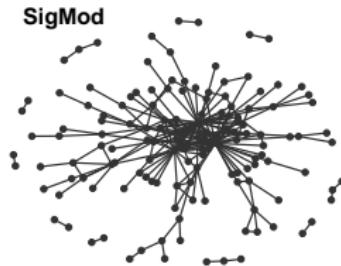
LEAN



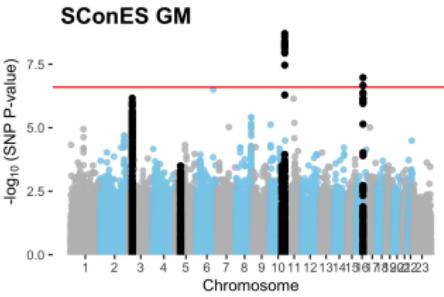
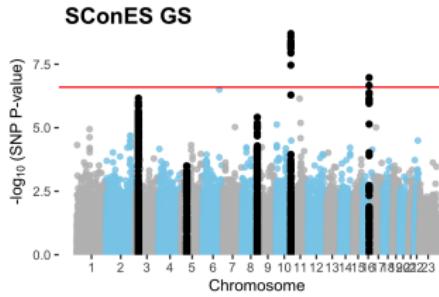
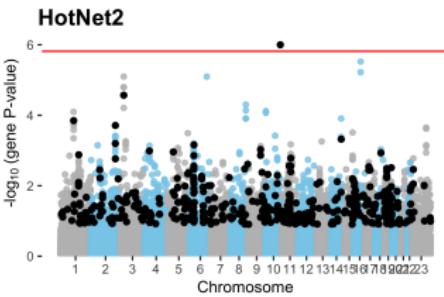
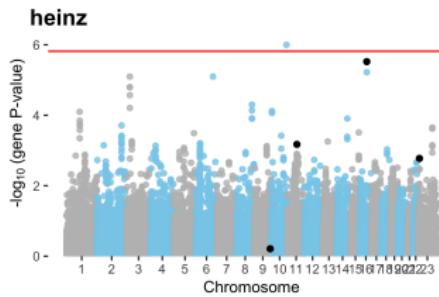
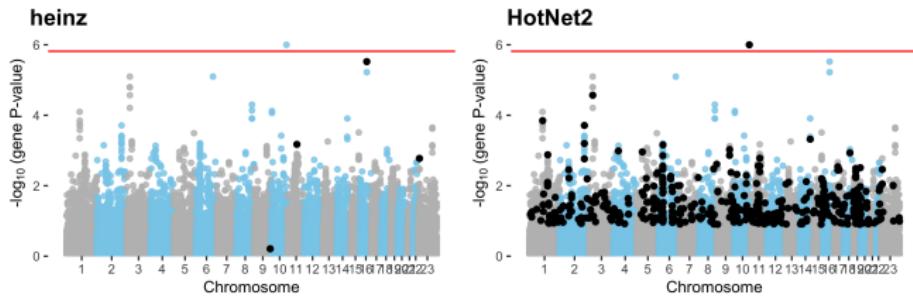
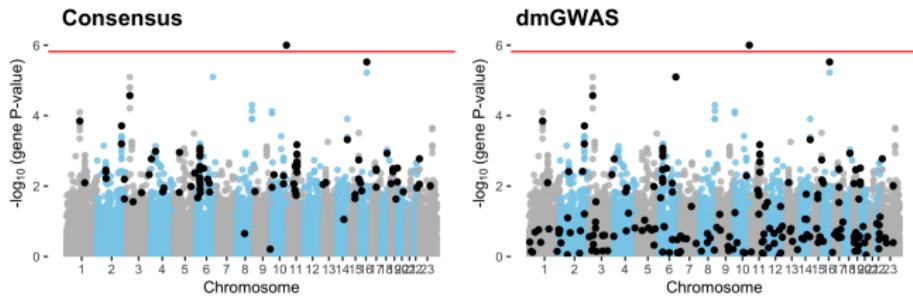
SConES GI



SigMod



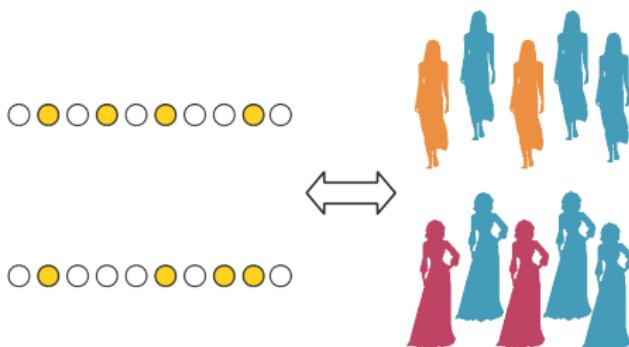
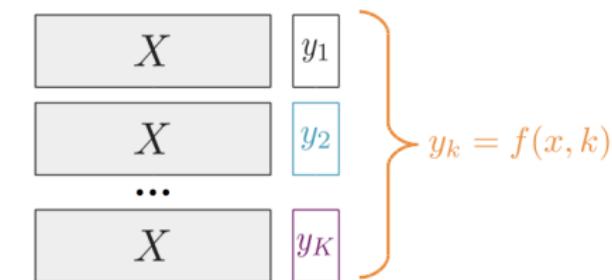
H. Climente-González et al. (2020). **Combining network-guided GWAS to discover susceptibility mechanisms for breast cancer**, BioRxiv.



# Increasing n

# Multi-task approaches

Increase sample size by **jointly** performing feature selection  
for **multiple related phenotypes**

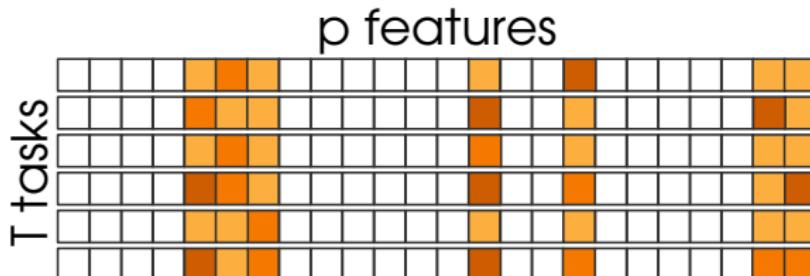


# Multitask Lasso [OTJ06]

- $T$  related phenotypes  $\mathbf{y}_1 \in \mathcal{Y}^{n_1}, \dots, \mathbf{y}_T \in \mathcal{Y}^{n_T}$

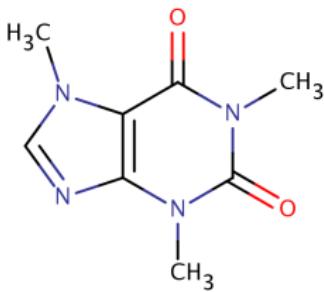
$$\arg \min_{\beta \in \mathbb{R}^{T \times p}} \underbrace{\sum_{t=1}^T \frac{1}{n_t} \sum_{i=1}^{n_t} \left( y_{ti} - \sum_{j=1}^p w_{tj} x_{tij} \right)^2}_{\text{loss}} + \lambda \underbrace{\sum_{j=1}^p \sqrt{\sum_{t=1}^T w_{tj}^2}}_{\text{task sharing}}$$

- Selects the **same features across tasks**
- Controls weights magnitude ( $\ell_2$  shrinkage).



# Task relatedness

- ▶ Tasks that are “more similar” should share more features.
- ▶ E.g. response to treatment.
- ▶ Chemical compounds representation and similarity:
  - ▶ Using the molecular graph
  - ▶ 3D structure, physico-chemical descriptors, etc.



C.-A. Azencott et al. (2007) One- to four-dimensional kernels for virtual screening and the prediction of physical, chemical and biological properties JCIM

# Multi-SConES

- ▶  $T$  related phenotypes  $\mathbf{y}_1 \in \mathcal{Y}^{n_1}, \dots, \mathbf{y}_T \in \mathcal{Y}^{n_T}$
- ▶  $T$  SNP-SNP networks  $W^t \in \mathbb{R}^{p \times p}$
- ▶ SNPs:  $X \in \{0, 1, 2\}^{n \times p}$   $n = \sum_{t=1}^T n_t$
- ▶ Goal: obtain similar sets of features on related tasks.

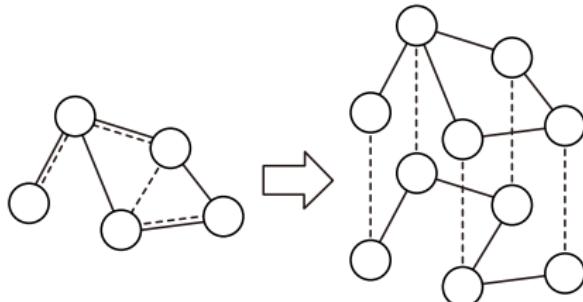
$$\begin{aligned} \arg \max_{\mathcal{S}_1, \dots, \mathcal{S}_T \subseteq \mathcal{V}} \sum_{t=1}^T & \left( \sum_{j \in \mathcal{S}_t} c_j^t - \eta |\mathcal{S}_t| - \lambda \sum_{j \in \mathcal{S}_t} \sum_{k \notin \mathcal{S}_t} W_{jk}^t \right) \\ & - \underbrace{\mu \sum_{t=1}^T \sum_{u=t+1}^T |\mathcal{S}_t \Delta \mathcal{S}_u|}_{\text{task sharing}}. \end{aligned}$$

$$\mathcal{S} \Delta \mathcal{S}' = (\mathcal{S} \cup \mathcal{S}') \setminus (\mathcal{S} \cap \mathcal{S}') \quad (\text{symmetric difference})$$

# Multi-SConES

$$\begin{aligned} \arg \max_{\mathcal{S}_1, \dots, \mathcal{S}_T \subseteq \mathcal{V}} \sum_{t=1}^T & \left( \sum_{j \in \mathcal{S}_t} c_j^t - \eta |\mathcal{S}_t| - \lambda \sum_{j \in \mathcal{S}_t} \sum_{k \notin \mathcal{S}_t} W_{jk}^t \right) \\ & - \mu \underbrace{\sum_{t=1}^T \sum_{u=t+1}^T |\mathcal{S}_t \Delta \mathcal{S}_u|}_{\text{task sharing}}. \end{aligned}$$

- ▶ Can be reduced to single-task by building a **meta-network**.



# Leveraging similarity between tasks

**Task similarity:**  $\Sigma \in \mathbb{R}^{T \times T}$ .

$$\begin{aligned} \arg \max_{\mathcal{S}_1, \dots, \mathcal{S}_T \subseteq \mathcal{V}} & \sum_{t=1}^T \left( \sum_{j \in \mathcal{S}_t} c_j - \eta |\mathcal{S}_t| - \lambda \sum_{j \in \mathcal{S}_t} \sum_{k \notin \mathcal{S}_t} W_{jk}^t \right) \\ & - \mu \underbrace{\sum_{t=1}^T \sum_{u=1}^T |\mathcal{S}_t \Delta \mathcal{S}_u| \Sigma_{tu}}_{\text{task sharing}}. \end{aligned}$$

Can also be reformulated as a maxflow/mincut problem.

## Multi-SConES: Selecting Connected Explanatory SNPs across Multiple related phenotypes

- ☺ selects connected, explanatory SNPs;
- ☺ benefits from using multiple related tasks;
- ☺ only scales up to small number of tasks.
  - M. Sugiyama, C.-A. Azencott, D. Grimm, Y. Kawahara and K. Borgwardt (2014) Multi-task feature selection on multiple networks via maximum flows, SIAM ICDM, 199–207
  - <https://github.com/mahito-sugiyama/Multi-SConES>
  - <https://github.com/chagaz/sfan>

# Feature selection in genomic data ( $p \gg n$ )

## 1. Using **biological networks** to integrate **prior knowledge**.

### SConES

- ▶ Using constraints built from biological networks helps feature selection.
- ▶ Flexibility is required as they are noisy and incomplete.

# Feature selection in genomic data ( $p \gg n$ )

## 1. Using **biological networks** to integrate **prior knowledge**.

### SConES

- ▶ Using constraints built from biological networks helps feature selection.
- ▶ Flexibility is required as they are noisy and incomplete.

## 2. Considering **multiple related phenotypes** at once.

### Multi-SConES, MMLD

- ▶ Jointly selecting features for related tasks alleviates the  $p \gg n$  issue.
- ▶ Using task descriptors can improve both selection and prediction.



**Where next?**

# GWAS questions

- ▶ **Linkage disequilibrium**
  - Pick one SNP per block? Combine SNPs within the same block?
  - How do you define a block?

# GWAS questions

- ▶ **Linkage disequilibrium**
    - Pick one SNP per block? Combine SNPs within the same block?
    - How do you define a block?
  - ▶ **Population structure**
    - Separate the data in homogeneous populations?
    - Incorporate population as covariates?
    - Use multitask / multidomain approaches?
- Work in progress with A Nouira.

# GWAS questions

- ▶ **Linkage disequilibrium**
  - Pick one SNP per block? Combine SNPs within the same block?
  - How do you define a block?
- ▶ **Population structure**
  - Separate the data in homogeneous populations?
  - Incorporate population as covariates?
  - Use multitask / multidomain approaches?  
Work in progress with A Nouira.
- ▶ **SNP-SNP network construction**
  - **SNP-to-gene mapping:** genomic coordinates? known eQTLs? 3D info?  
D. Duroux et al. (2020). Interpretable network-guided epistasis detection, BioRxiv
  - Underlying **gene-gene networks:** PPIN? Regulatory networks?  
Work in progress with H Climente-González, D Duroux, K Van Steen.

# Stable feature selection

- ▶ **Stability**
    - How do you **enforce** stability?
    - Measure stability at the **level** of SNPs? Genes? Pathways?
- Work in progress with A Nouira.

# Multiview feature selection

- ▶ **Multiomics:**
  - Jointly consider SNPs and gene expression, methylation patterns, etc.
  - Integration through **SNP-to-gene mapping?**
- ▶ **Multimodality:** select features on
  - Lab results.
  - Time series: patient trajectories, accelerometer data.
  - Free-form medical text.
  - Images.

Work in progress with A Recanati, NM Mbaye, E Dumas, RT2 lab.
- ▶ **Data privacy**
  - Learning from **federated** data sets without compromising privacy.

Work in progress with A Recanati.

# Funding

- ▶ **Agence Nationale pour la Recherche**  
**SCAPHE** (JCJC ANR-18-CE45-0021-01) 2019–2021.  
**PR[AI]RIE** Springboard Chair 2020–2023.
- ▶ **Alexander von Humboldt Stiftung**  
**Postdoctoral fellowship 2011–2013.**
- ▶ **Deutsche Forschungsgemeinschaft**  
**LaLa (DFG 164930347) 2013.**
- ▶ **European Research Council**  
**MLPM2012 (FP7-PEOPLE-2012-ITN 316861) 2013–2016.**  
**IC-3i-PhD (H2020-MSCA-COFUND-2014 666003) 2016–2019.**  
**MLFPM2018 (H2020-MSCA-ITN-2018 813533) 2019–2022.**
- ▶ **Sancare**
- ▶ **Sanofi-Adventis**

# Thank You

firstname.name@mines-paristech.fr

cazencott.info & goepp.github.io

# References |

- [AL11] David H. Alexander et Kenneth Lange. "Stability selection for genome-wide association". In : *Genetic Epidemiology* 35.7 (2011), p. 722-728.
- [Ant+10] Stylianos E. Antonarakis et al. "Mendelian disorders and multifactorial traits: the big divide or one for all?" In : *Nature Reviews Genetics* 11.5 (2010), p. 380-384.
- [AR15] Samuel J. Aronson et Heidi L. Rehm. "Building the foundation for genomics in precision medicine". In : *Nature* 526.7573 (2015), p. 336-342.
- [Aze+13] Chloé-Agathe Azencott et al. "Efficient network-guided multi-locus association mapping with graph cuts". In : *Bioinformatics* 29.13 (2013), p. i171-i179.
- [Aze16] Chloé-Agathe Azencott. "Network-guided biomarker discovery". In : *Machine Learning for Health Informatics*. Lecture Notes in Computer Science 9605. Springer International Publishing, 2016.
- [Aze18] Chloé-Agathe Azencott. "Machine learning and genomics: precision medicine vs. patient privacy". In : *Philosophical Transactions of the Royal Society A* (2018).
- [Bar+12] Fredrik Barrenäs et al. "Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms". In : *Genome Biology* 13 (2012), R46.
- [BR17] Rina Foygel Barber et Aaditya Ramdas. "The p-filter: multilayer false discovery rate control for grouped hypotheses". In : *J. R. Stat. Soc. B* 79.4 (2017), p. 1247-1268.

# References II

- [Brz+17] Damian Brzyski et al. "Controlling the Rate of GWAS False Discoveries". In : *Genetics* 205.1 (2017). PMID: 27784720, p. 61-75.
- [BSA16] Victor Bellon, Véronique Stoven et Chloé-Agathe Azencott. "Multitask feature selection with task descriptors". In : *Pacific Symposium on Biocomputing*. T. 21. 2016, p. 261-272.
- [CA17] Héctor Climente et Chloé-Agathe Azencott. *martini: GWAS incorporating networks in R*. 2017. url : <https://bioconductor.org/packages/devel/bioc/html/martini.html>.
- [CH+08] Gerda Claeskens, Nils Lid Hjort et al. "Model selection and model averaging". In : *Cambridge Books* (2008).
- [Che+18] Po-Hsuan Cameron Chen et al. *An Augmented Reality Microscope for Real-time Automated Detection of Cancer*. Rapp. tech. Google AI Healthcare, 2018.
- [Cla+08] Robert Clarke et al. "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data". In : *Nature Reviews Cancer* 8.1 (2008), p. 37-49.
- [Cow+17] Lenore Cowen et al. "Network propagation: a universal amplifier of genetic associations". In : *Nature Reviews Genetics* 18.9 (2017), p. 551-562.
- [Dit+08] Marcus T. Dittrich et al. "Identifying functional modules in protein-protein interaction networks: an integrated exact approach". In : *Bioinformatics* 24.13 (2008), p. i223-i231.

# References III

- [DP15] Nicoletta Dessì et Barbara Pes. "Stability in Biomarker Discovery: Does Ensemble Feature Selection Really Help?" In : *Current Approaches in Applied Artificial Intelligence*. Lecture Notes in Computer Science 9101. DOI: 10.1007/978-3-319-19066-2\_19. 2015, p. 191-200.
- [Dro+16] Alexandre Drouin et al. "Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons". In : *BMC Genomics* 17 (2016), p. 754.
- [ED+05] Liat Ein-Dor et al. "Outcome signature genes in breast cancer: is there a unique set?" In : *Bioinformatics* 21.2 (2005), p. 171-178.
- [Edu+15] Federica Eduati et al. "Opportunities and limitations in the prediction of population responses to toxic compounds assessed through a collaborative competition". In : *Nature Biotechnology* 33.9 (2015), p. 933-940. doi : doi:10.1137/1.9781611973440.23.
- [Est+17] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In : *Nature* 542.7639 (2017), p. 115.
- [Fur13] Laura I. Furlong. "Human diseases through the lens of network biology". In : *Trends in Genetics* 29.3 (2013), p. 150-159.
- [GBG16] Laura Gay, Ann-Marie Baker et Trevor A Graham. "Tumour cell heterogeneity". In : *F1000Research* 5 (2016).

# References IV

- [Gra+15] G'Sell Max Grazier et al. "Sequential selection procedures and false discovery rate control". In : *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.2 (2015), p. 423-444.
- [GVB13] Levi A. Garraway, Jaap Verweij et Karla V. Ballman. "Precision oncology: an overview". In : *J. Clin. Oncol.* 31.15 (2013), p. 1803-1805.
- [Gwi+17] Frederik Gwinner et al. "Network-based analysis of omics data: the LEAN method". In : *Bioinformatics* 33.5 (2017), p. 701-709.
- [Hea+15] Megan L Head et al. "The extent and consequences of p-hacking in science". In : *PLoS biology* 13.3 (2015), e1002106.
- [Hol18] Susan Holmes. "Statistical proof? The problem of irreproducibility.". In : *Bulletin (New Series) of the American Mathematical Society* 55.1 (2018).
- [Hua+18] Justin K. Huang et al. "Systematic evaluation of molecular networks for discovery of disease genes". In : *Cell Systems* 6.4 (2018), 484-495.e5.
- [Ioa05] John PA Ioannidis. "Why most published research findings are false". In : *PLoS medicine* 2.8 (2005), e124.
- [JOV09] L. Jacob, G. Obozinski et J.-P. Vert. "Group lasso with overlap and graph lasso". In : *ICML*. 2009, p. 433-440.

# References V

- [JZ14] Peilin Jia et Zhongming Zhao. "Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives". In : *Human Genetics* 133.2 (2014), p. 125-138.
- [Laz+14] David Lazer et al. "The parable of Google Flu: traps in big data analysis". In : *Science* 343.6176 (2014), p. 1203-1205.
- [Lee+16] J. D. Lee et al. "Exact post-selection inference, with application to the lasso". In : *The Annals of Statistics* 44.3 (2016), p. 907-927.
- [Lei+15] Mark D. M. Leiserson et al. "Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes". In : *Nature Genetics* 47 (2015), p. 106-114.
- [Li11] C. Li. "Personalized medicine – the promised land: are we there yet?" In : *Clinical Genetics* 79.5 (2011), p. 403-412.
- [LL08] C. Li et H. Li. "Network-constrained regularization and variable selection for analysis of genomic data". In : *Bioinformatics* 24.9 (2008), p. 1175-1182.
- [LL10] Caiyan Li et Hongzhe Li. "Variable selection and regression analysis for graph-structured covariates with an application to genomics". In : *The annals of applied statistics* 4.3 (2010), p. 1498-1516.
- [LL18] Tak Lee et Insuk Lee. "araGWAB: Network-based boosting of genome-wide association studies in *Arabidopsis thaliana*". In : *Scientific reports* 8.1 (2018), p. 1-6.

# References VI

- [Loc+14] Richard Lockhart et al. "A significance test for the lasso". In : *Annals of statistics* 42.2 (2014), p. 413.
- [MB10] Nicolai Meinshausen et Peter Bühlmann. "Stability selection". In : *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4 (2010). issn : 1467-9868.
- [Mio+16] Riccardo Miotto et al. "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records". In : *Scientific reports* 6 (2016), p. 26094.
- [MM15] Aniket Mishra et Stuart Macgregor. "VEGAS2: software for more flexible gene-based testing". In : *Twin Research and Human Genetics* 18.1 (2015), p. 86-91.
- [NB16] Sarah Nogueira et Gavin Brown. "Measuring the stability of feature selection". In : *Machine Learning and Knowledge Discovery in Databases*. Lecture Notes in Computer Science 9852. Springer International Publishing, 2016, p. 442-457.
- [Nie+15] Clément Niel et al. "A survey about methods dedicated to epistasis detection". In : *Bioinformatics and Computational Biology* (2015), p. 285.
- [Nuz14] Regina Nuzzo. "Scientific method: statistical errors". In : *Nature News* 506.7487 (2014), p. 150.
- [Oks+14] Sebastian Okser et al. "Regularized machine learning in the genetic prediction of complex traits". In : *PLoS genetics* 10.11 (2014), e1004754.

# References VII

- [OTJ06] Guillaume Obozinski, Ben Taskar et Michael I. Jordan. *Multi-task feature selection*. Rapp. tech. UC Berkeley, 2006.
- [Pen07] Elizabeth Pennisi. "Human Genetic Variation". In : *Science* 318.5858 (2007), p. 1842-1843.
- [Raj+18] Alvin Rajkomar et al. "Scalable and accurate deep learning with electronic health records". In : *NPJ Digital Medicine* 1.1 (2018), p. 18.
- [Sch15] Nicholas J. Schork. "Personalized medicine: Time for one-person trials". In : *Nature News* 520.7549 (2015), p. 609.
- [SL12] Grzegorz Swirszcz et Aurelie C. Lozano. "Multi-level lasso for sparse multi-task regression". In : *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. 2012, p. 361-368.
- [Sli+19] Lotfi Slim et al. "kernelPSI: a post-selection inference framework for nonlinear variable selection". In : *Proceedings of the Thirty-Sixth International Conference on Machine Learning (ICML)*. T. 97. 2019, p. 5857-5865.
- [Sny+14] Alexandra Snyder et al. "Genetic basis for clinical response to CTLA-4 blockade in melanoma". In : *New England Journal of Medicine* 371.23 (2014), p. 2189-2199.

# References VIII

- [Sør+01] Therese Sørlie et al. "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications". In : *Proceedings of the National Academy of Sciences* 98.19 (2001), p. 10869-10874.
- [SS17] Hon-Cheong So et Pak C Sham. "Improving polygenic risk prediction from summary statistics by an empirical Bayes approach". In : *Scientific Reports* 7 (2017), p. 41262.
- [Sug+14] Mahito Sugiyama et al. "Multi-task feature selection on multiple networks via maximum flows". In : *SIAM ICDM*. 2014, p. 199-207.
- [Suz+17] Shinya Suzumura et al. "Selective inference for sparse high-order interaction models". In : *PMLR*. 2017, p. 3338-3347.
- [Ter+16] Aika Terada et al. "LAMPLINK: detection of statistically significant SNP combinations from GWAS data". In : *Bioinformatics* (2016), btw418.
- [Tib94] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In : *J. R. Stat. Soc.* 58 (1994), p. 267-288.
- [VCB11] Marc Vidal, Michael E. Cusick et Albert-László Barabási. "Interactome networks and human disease". In : *Cell* 144.6 (2011), p. 986-998. (Visité le 05/02/2020).

# References IX

- [VDD11] David Venet, Jacques E. Dumont et Vincent Detours. "Most random gene expression signatures are significantly associated with breast cancer outcome". In : *PLoS Computational Biology* 7.10 (2011), e1002240.
- [VV+02] Laura J Van't Veer et al. "Gene expression profiling predicts clinical outcome of breast cancer". In : *nature* 415.6871 (2002), p. 530.
- [WL+16] Ronald L Wasserstein, Nicole A Lazar et al. "The ASA's statement on p-values: context, process, and purpose". In : *The American Statistician* 70.2 (2016), p. 129-133.
- [YL06] Ming Yuan et Yi Lin. "Model selection and estimation in regression with grouped variables". In : *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1 (2006), p. 49-67.
- [ZH05] Hui Zou et Trevor Hastie. "Regularization and variable selection via the elastic net". In : *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), p. 301-320.