



Efficient multi-locus biomarker discovery

Chloé-Agathe Azencott

Center for Computational Biology (CBIO)
Mines ParisTech – Institut Curie – INSERM U900
PSL Research University, Paris, France

October 5, 2017 – M2BI

<http://cazencott.info> chloe-agathe.azencott@mines-paristech.fr @cazencott

Precision Medicine

- **Adapt** treatment to the **(genetic) specificities** of the patient.
E.g. Trastuzumab for HER2+ breast cancer.



Precision Medicine

- ▶ **Adapt** treatment to the **(genetic) specificities** of the patient.

E.g. Trastuzumab for HER2+ breast cancer.

- ▶ **Data-driven** biology/medicine

Identify similarities between patients that exhibit similar phenotypes.



Precision Medicine

- ▶ **Adapt** treatment to the **(genetic) specificities** of the patient.
E.g. Trastuzumab for HER2+ breast cancer.
- ▶ **Data-driven** biology/medicine
Identify similarities between patients that exhibit similar phenotypes.

Biomarker discovery = Feature Selection



1. Genome-Wide Association Studies

Common disease common variant hypothesis

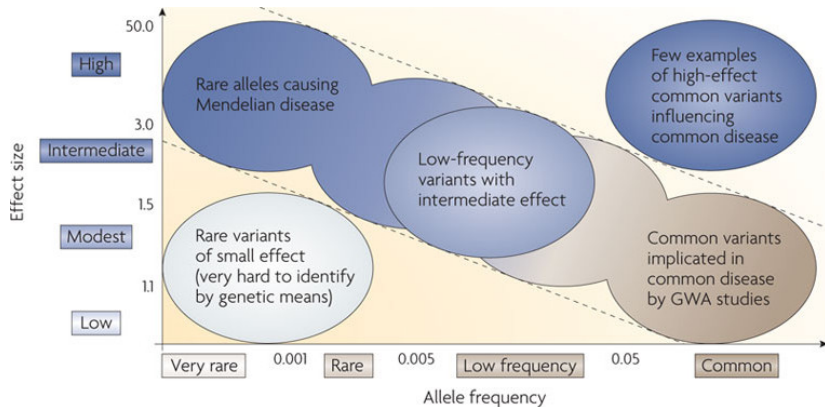
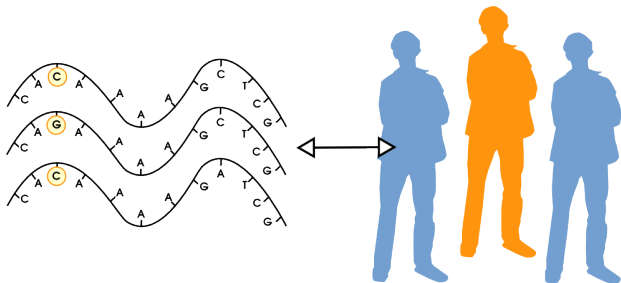
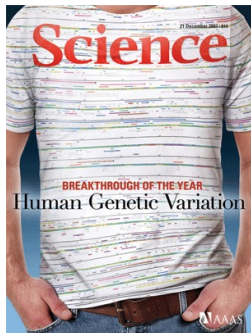


Image source: [antonarakis2010]

GWAS: Genome-Wide Association Studies



Which genomic features explain the phenotype?

$p = 10^5 - 10^7$ Single Nucleotide Polymorphisms (SNPs)

$n = 10^2 - 10^4$ samples

[pennisi2007]

Guilt by association

Use **Linkage disequilibrium**

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



Image: Francis Collins 2008

SNP genotyping

SNP microarray



- ▶ Cheap
- ▶ Small data files
- ▶ 500K – 5M loci captured.

Whole-Genome Sequencing



- ▶ More expensive
- ▶ Large data files
- ▶ ~ 3B loci captured (incl. rare variants).

Quality control

► Missing data

- **Discard** SNPs with high missingness rate
- **Impute** missing SNPs, e.g. with IMPUTE2 [**howie2009**].

► 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 $2p \times (1 - p)$
 - aa has frequency $(1 - p) \times (1 - p)$
- Compare observed with expected with a χ^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	$N \times p^2$ 1499	$N \times 2p(1 - p)$ 209	$N \times (1 - p)^2$ 7

- The value of the χ^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.

Qualitative GWAS

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

- **Contingency table**

	AA	Aa	aa
Cases			
Ctrls			

	0	1
Cases	a	b
Ctrls	c	d

Statistical tests: χ^2 , Cochran-Armitage trend test, etc.

- **Logistic regression**

$$\text{logit}(p(y|X)) = \beta_0 + \beta_1 X$$

- **Odds-ratio**

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

Quantitative GWAS

- ▶ **Linear regression**

$$y = \beta_0 + \beta_1 X$$

- ▶ p-value: Is $\hat{\beta}_1$ significantly different from 0?

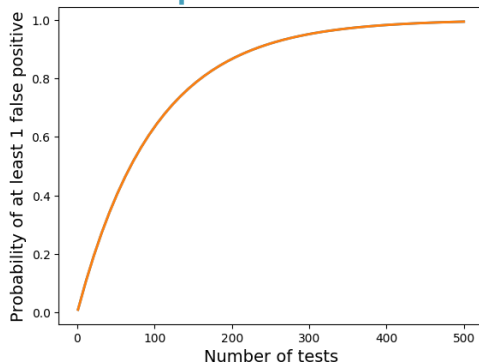
Wald test: compare $\frac{\hat{\beta}_1^2}{\text{Var}(\hat{\beta}_1)}$ to a χ^2 distribution.

- ▶ **Effect size:** β_1 .

Multiple Hypothesis Testing

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

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



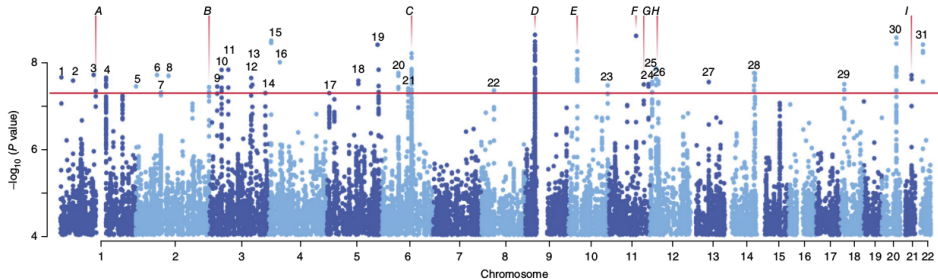
- ▶ Controlling **Family-Wise Error Rate (FWER)**

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

FP = number of false positives (Type I errors)

- ▶ **Bonferroni** correction: $\alpha \rightarrow \frac{\alpha}{p}$

Manhattan plots

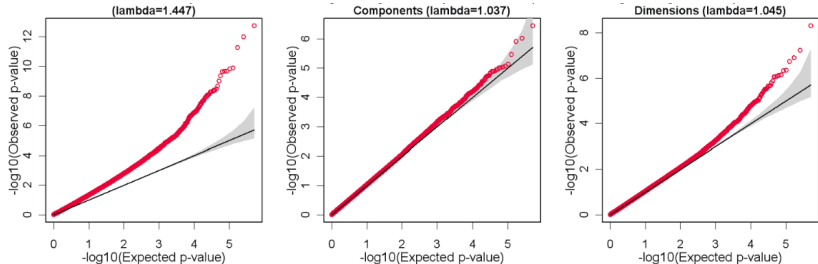


- ▶ $-\log_{10}(\text{p-value})$ for each SNP
- ▶ SNP ordered by genomic position (chr+bp)
- ▶ significance line.

Ref: [bonder2016]

Q-Q plots

- ▶ Most SNPs not associated $\Rightarrow$ p-values **uniformly distributed**.
- ▶ **Quantile-quantile plot:**
 - y-axis: $-\log_{10}$ observed order statistics of p-value
 - x-axis: $-\log_{10}$ expected order statistics of uniform
- ▶ **Genomic inflation factor** λ : regression coefficient of the univariate regression line.



A close-up photograph of two brass keys hanging from a silver metal ring. The keys are positioned diagonally across the frame, with the heads of the keys at the top and the blades pointing downwards. The background is a plain, light-colored surface.

Replication in an independent cohort is key.

Interpretation

- ▶ **Association does not imply causation.**
 - The functional SNP might be one in **LD** with the associated one
 - Eliminate possible **confounders** (sex, environment, population structure)
 - **Population structure**
 - SNP near lactase gene is associated with multiple sclerosis (MS)
 - MS is more prevalent in populations of Northern European ancestry
 - Correction with **principal components**: EIGENSTRAT [price2006]
 - **Linear mixed models**: FastLMM [lippert2011]
- ▶ SNP “mode of action”:
 - modify the **protein** (non-synonymous mutations, etc.)
 - regulates the **expression of a gene**
 - nearby: **cis-eQTL**
 - faraway: **trans-eQTL**.

GWAS discoveries

- ▶ **The GWAS Catalog** <https://www.ebi.ac.uk/gwas/>
- ▶ **Clinical benefits:** Ankylosing spondylitis
 - Role of interleukine 17 pathway
 - Consentyx (**secukinumab**), approved January 15, 2016.
- ▶ SNPs associated with **drug resistance**
in e.g. *Mycobacterium tuberculosis*, *Staphylococcus aureus*,
Streptococcus pneumoniae or HIV.

Ref: [visscher2012; manolio2013; power2017]

Missing heritability

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

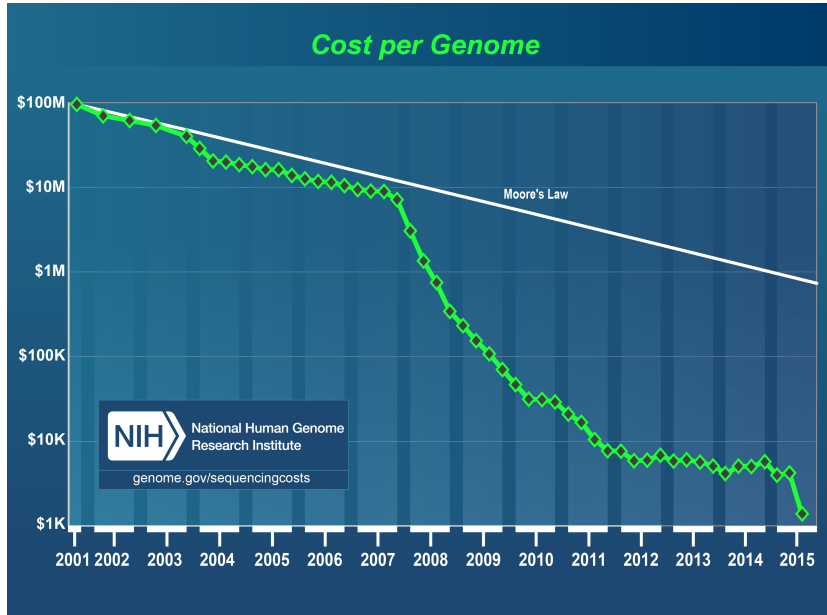
Many possible reasons:

- 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.

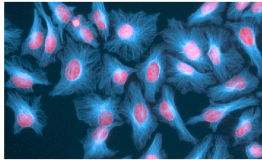
Ref: [manolio2009]

2. Feature selection in high dimension with few samples

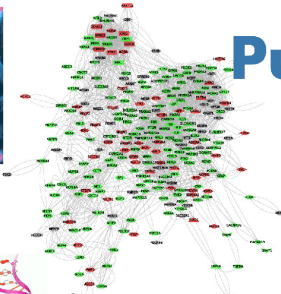
Sequencing costs



Big data!



phenome



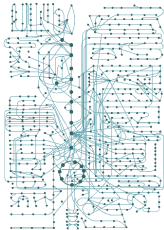
interactome



publications



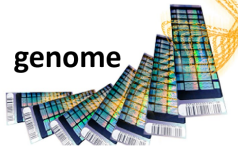
transcriptome



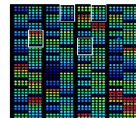
metabolome



methylome



genome



proteome

Image sources: [ajc1@ flickr](#); [Zlir'a@wikimedia](#)

Big data!



THE CANCER GENOME ATLAS

National Cancer Institute

National Human Genome Research Institute

A banner for the Cancer Genomics Hub. It features a blue background with a glowing DNA double helix on the left and a glowing orange structure on the right. The text "Cancer Genomics Hub" is in white, and "A resource of the National Cancer Institute" is in a smaller white font below it.

Cancer Genomics Hub
A resource of the National Cancer Institute

A banner for the 1000 Genomes project. It has a dark background with a pattern of colorful, glowing human chromosomes. The text "1000 Genomes" is in yellow, and "A Deep Catalog of Human Genetic Variation" is in white below it.

1000 Genomes

A Deep Catalog of Human Genetic Variation

A banner for DREAM Challenges. It features a blue background with glowing, wavy lines and starburst patterns, suggesting a network or data visualization. The text "DREAM Challenges" is in white.

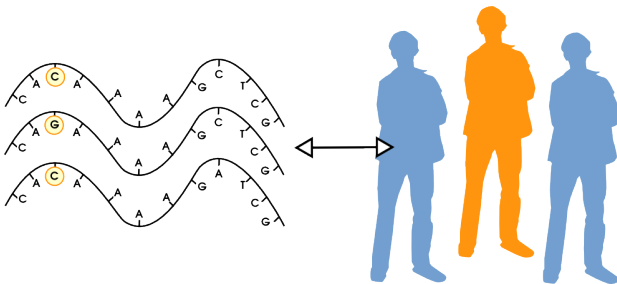
DREAM Challenges

SAY BIG DATA



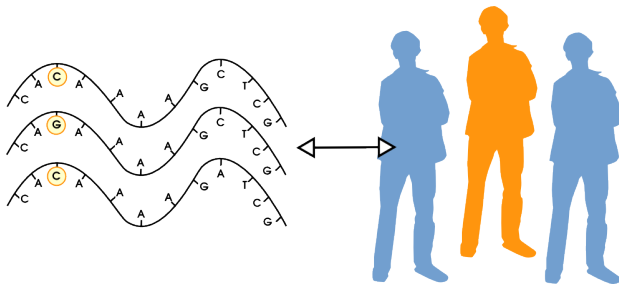
ONE MORE TIME

GWAS: Genome-Wide Association Studies



Which genomic features explain the phenotype?

GWAS: Genome-Wide Association Studies

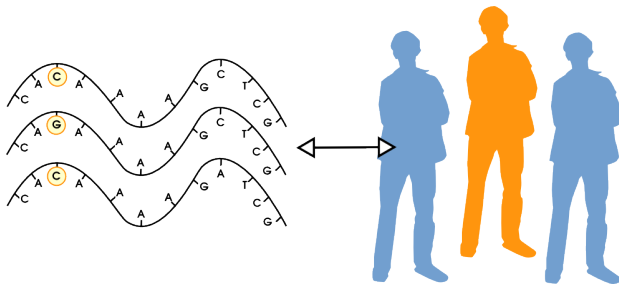


Which genomic features explain the phenotype?

$p = 10^5 - 10^7$ Single Nucleotide Polymorphisms (SNPs)

$n = 10^2 - 10^4$ samples

GWAS: Genome-Wide Association Studies



Which genomic features explain the phenotype?

$p = 10^5 - 10^7$ Single Nucleotide Polymorphisms (SNPs)

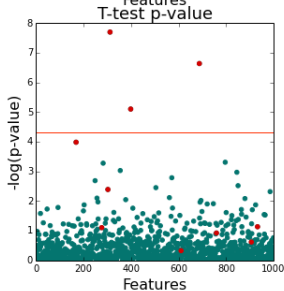
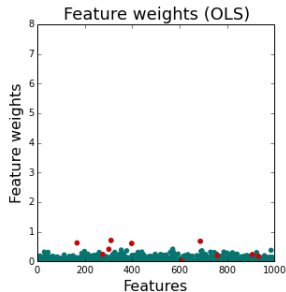
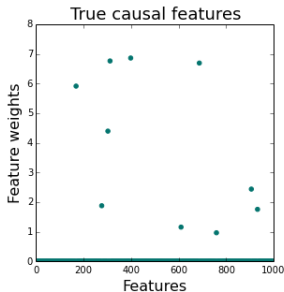
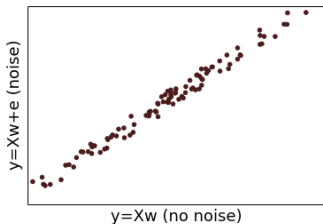
$n = 10^2 - 10^4$ samples

High-dimensional (large p)

Low sample size (small n)

Large p, small n data

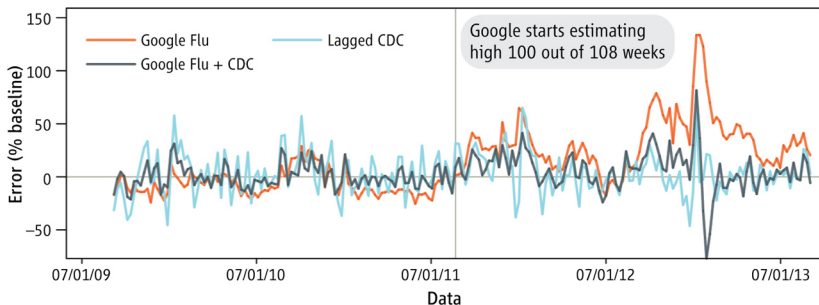
Simulation: $n=100$, $p=1000$, 10 causal features, $y = Xw + e$.



Google Flu Trends

D. Lazer, R. Kennedy, G. King and A. Vespignani. **The Parable of Google Flu: Traps in Big Data Analysis.** Science 2014 [lazer2014]

- ▶ **p = 50 million** search terms
- ▶ **n = 1152** data points



- ▶ Predictive search terms include terms related to high-school basketball.

3. Machine learning approaches

A detour: a few principles of machine learning.

A collection of tools is arranged on a white background. In the foreground, a silver adjustable wrench is open, showing its internal mechanism. To its left is a hammer with a silver head and a black handle. Above the wrench is a screwdriver with a silver shaft and a black handle. To the right of the screwdriver are two small silver pins or screws. In the background, a hex key and a regular screw are visible. The word "Reducing p" is written in a bold, blue, sans-serif font across the middle of the image.

Reducing p

Regularization

Control the number of parameters

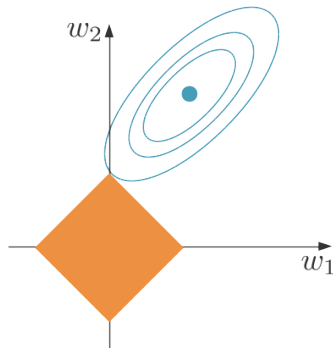
- ▶ limit **model complexity**
- ▶ avoid **overfitting**

▶ Least-squares regression

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2}_{\text{loss}}$$

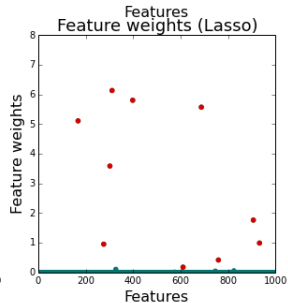
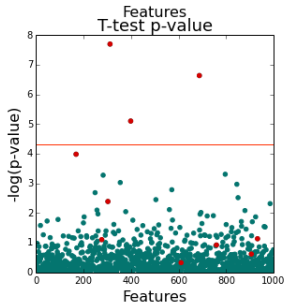
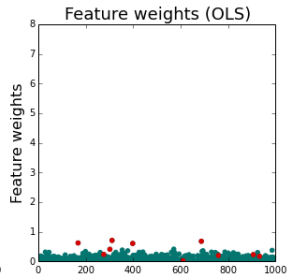
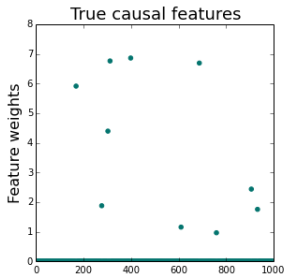
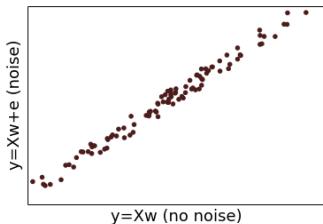
▶ **Lasso** [ani1994]

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2}_{\text{loss}} + \lambda \underbrace{\|\mathbf{w}\|_1}_{\text{sparsity}}$$



Large p, small n data

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



Integrating prior knowledge

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

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

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



Original feature space



Constrained feature space

Elephant image by Danny Chapman @ Flickr.

Structured sparsity

- **Lasso:** Selecting variables in high dimension.

$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2}_{\text{loss}} + \lambda \underbrace{||\mathbf{w}||_1}_{\text{sparsity}}$$



- **Lasso+:** Structured regularizer.

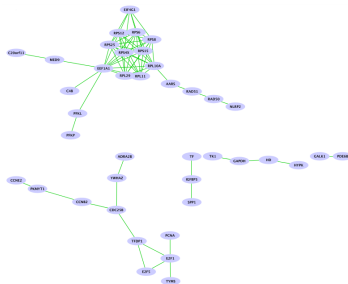
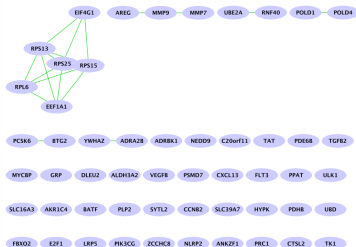
$$\arg \min_{\mathbf{w} \in \mathbb{R}^p} \underbrace{||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2}_{\text{loss}} + \lambda \underbrace{\Omega(\mathbf{w})}_{\text{structure}}$$

Gene selection with the Graph lasso

icml2009

Group Lasso with Overlaps and Graph Lasso

Laurent Jacob, Guillaume Obozinski and Jean-Philippe Vert



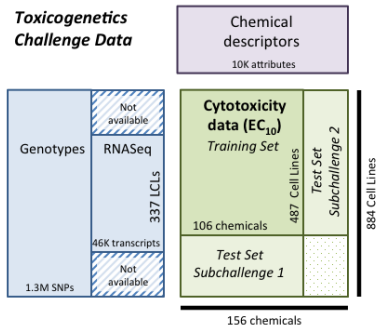
Ref: [jacob2009]



Increasing n

Toxicogenetics / Pharmacogenomics

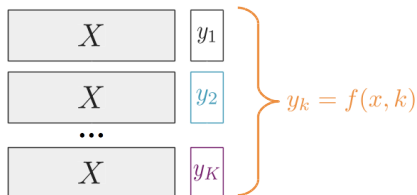
Tasks (phenotypes) = chemical compounds



F. Eduati, L. Mangravite, et al. (2015) **Prediction of human population responses to toxic compounds by a collaborative competition.** Nature Biotechnology, 33 (9), 933–940 doi: 10.1038/nbt.3299

Multi-task approaches

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



Multi-task approaches

- **Multitask Lasso** [Obozinski et al. 2006]

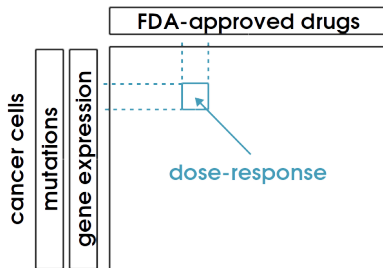
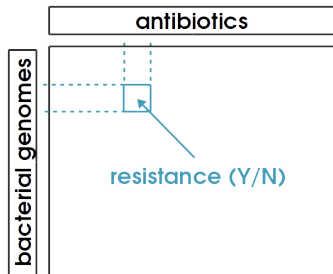
$$\arg \min_{\beta \in \mathbb{R}^{T \times p}} \underbrace{\sum_{t=1}^T \frac{1}{n_t} \sum_{m=1}^{n_t} \left(y_{tm} - \sum_{j=1}^p \beta_{tj} x_{tmj} \right)^2}_{\text{loss}} + \underbrace{\lambda \sum_{j=1}^p \|\beta_j^2\|_2}_{\text{task sharing}}$$

- Selects the same features across tasks
- Controls weights magnitude (ℓ_2 shrinkage).

Applications

Tasks = chemicals (large L)

- ▶ Antibiotic resistance
- ▶ Cancer chemotherapy



Open questions

► Robustness/stability

- Interpret correlated features?
- Recover the same features when the data changes slightly?

► Complex interaction patterns

- Limited to additive or quadrative effects
- High-order features?
- Set Covering Machines?

► Statistical significance

- Computing p-values?
- Correcting for multiple hypotheses?
- Controlling FDR?

Practical session

`https://tinyurl.com/TODO`

References