Feature selection in high dimension for precision medicine

Chloé-Agathe Azencott

Center for Computational Biology (CBIO)

Mines ParisTech – Institut Curie – INSERM U900

PSL Research University, Paris, France

March 21, 2017 – MACARON Workshop

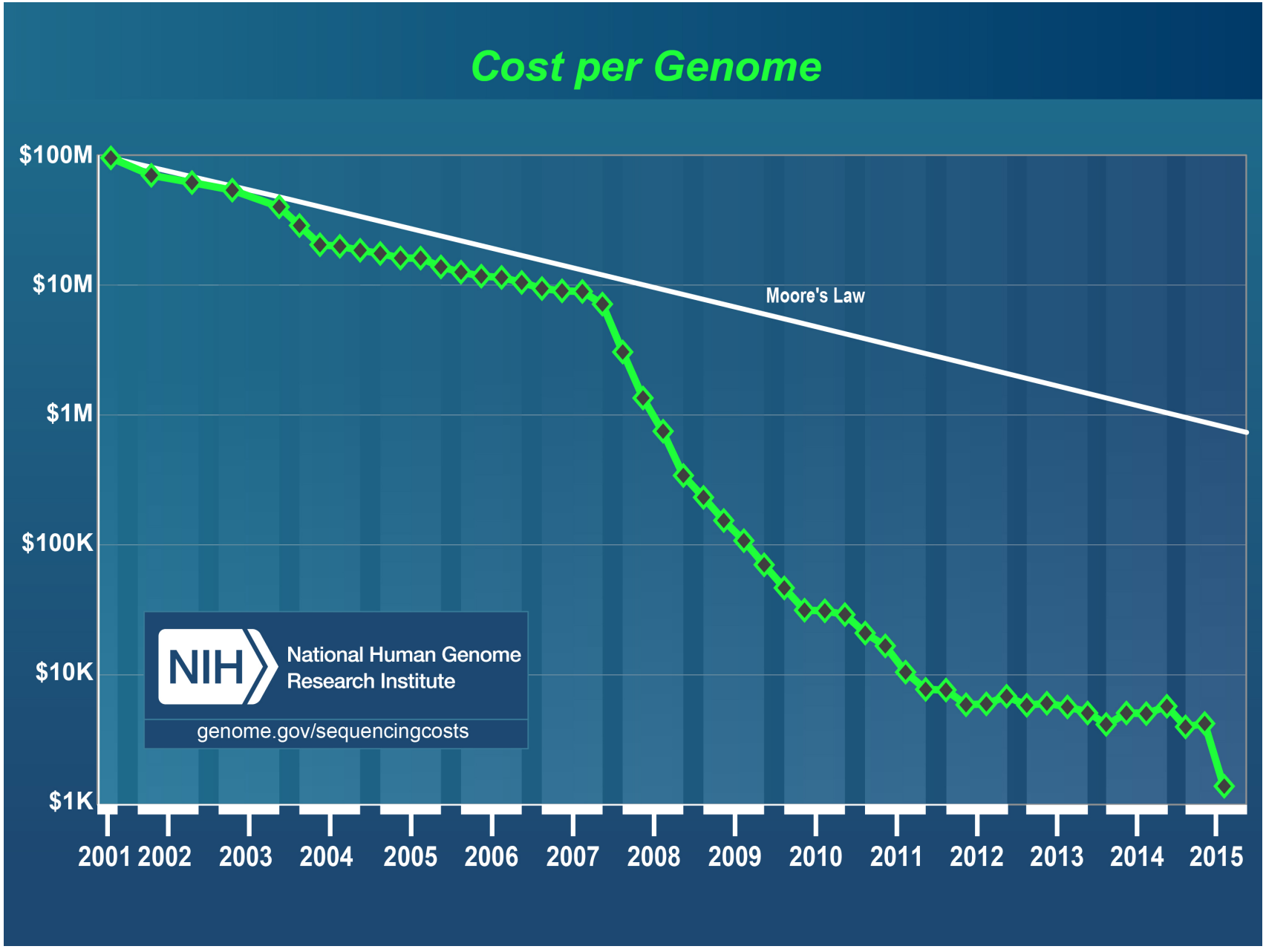
http://cazencott.info chloe-agathe.azencott@mines-paristech.fr @cazencott

Precision Medicine

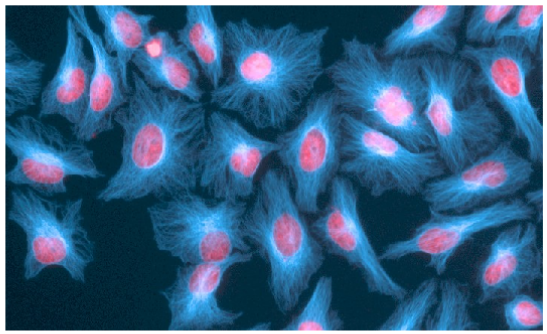
- Treatment **adapted to the (genetic) specificities** of the patient.
  E.g. Trastuzumab for HER2+ breast cancer.
- **Data-driven** biology/medicine

  Identify similarities between patients that exhibit similar susceptibilities / prognoses / responses to treatment.

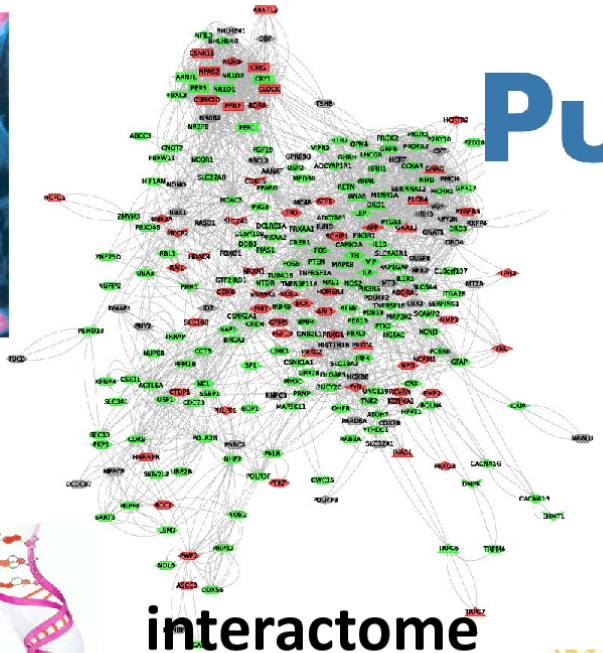
Sequencing costs



Big data!



phenome

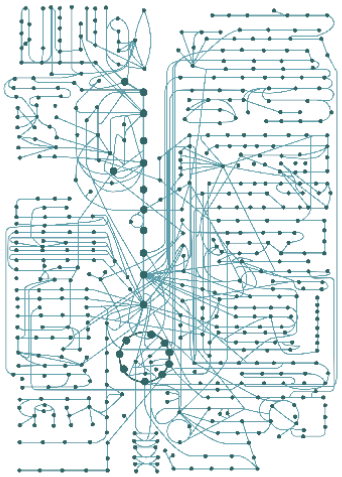


interactome

PubMed

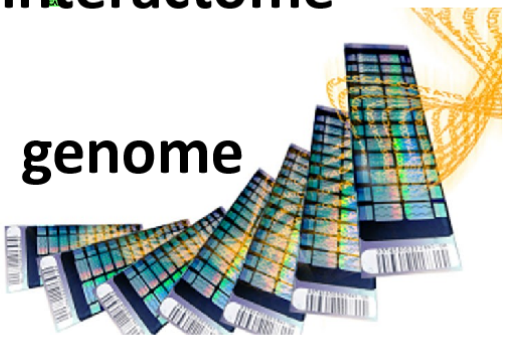
publications

transcriptome

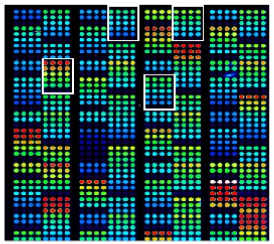


metabolome

methylome



genome



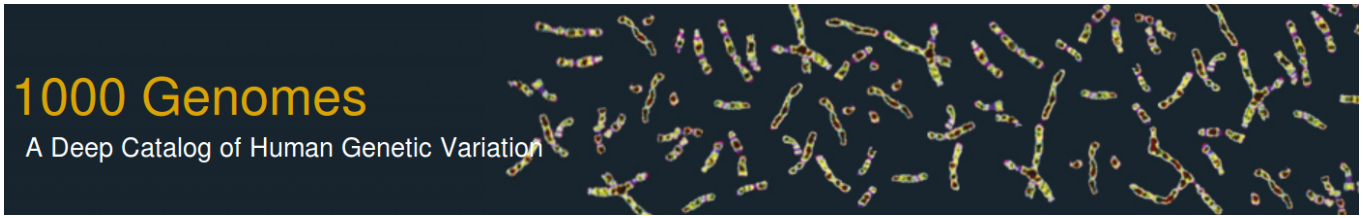
proteome

Image sources: ajc1@ flickr; Zlir'a@wikimedia

Big data!


NIH
THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

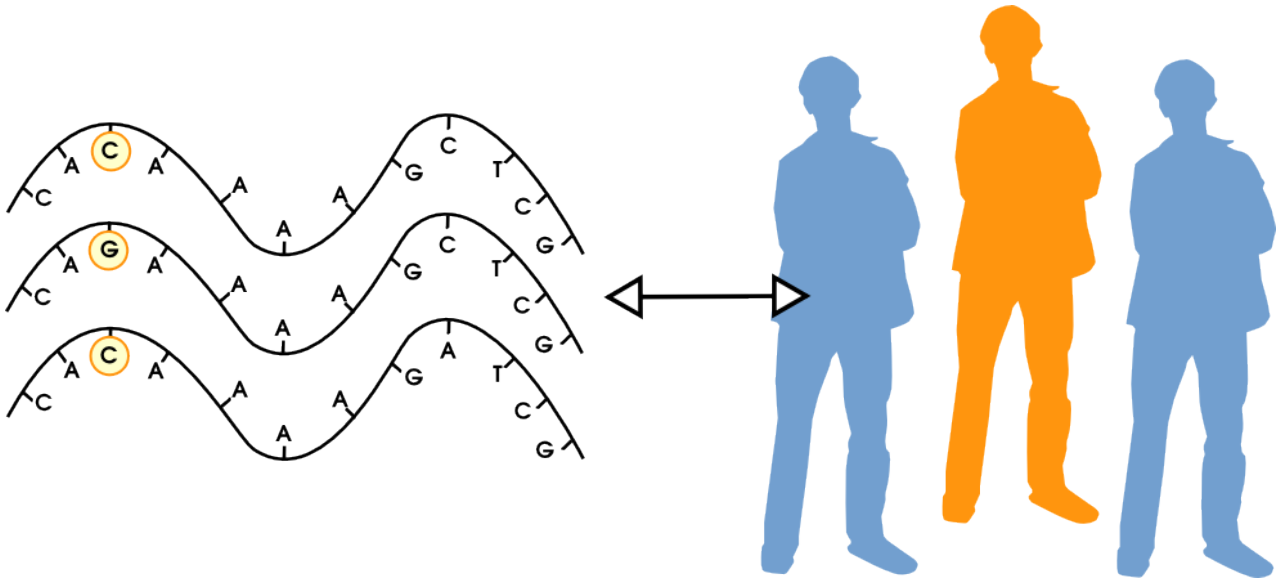

Cancer Genomics Hub
A resource of the National Cancer Institute


1000 Genomes
A Deep Catalog of Human Genetic Variation


DREAM Challenges


SAY BIG DATA
ONE MORE TIME

GWAS: Genome-Wide Association Studies



Which genomic features explain the phenotype?

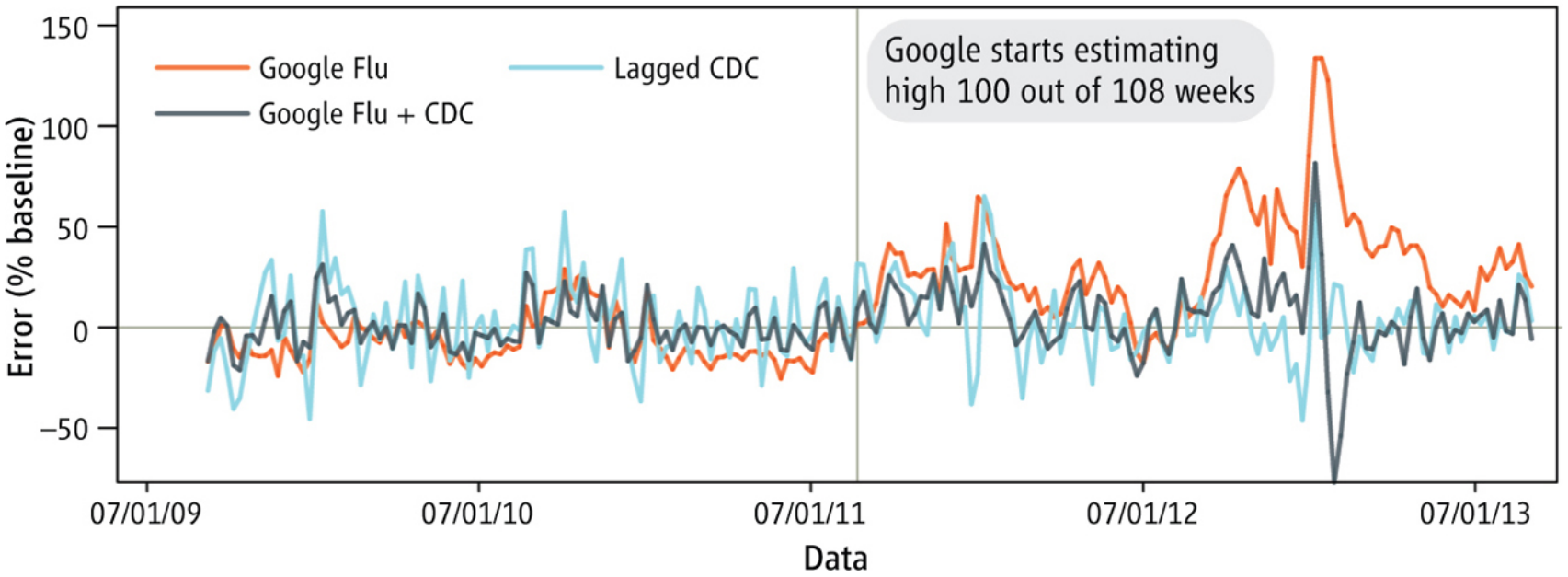
$$p = 10^5 - 10^7 \text{ Single Nucleotide Polymorphisms (SNPs)}$$
$$n = 10^2 - 10^4 \text{ samples}$$

- High-dimensional (large p)
- Low sample size (small n)

Google Flu Trends

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

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



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

Is extracting information from this data doomed from the start?

GWAS successes

Multiple sclerosis

Nature Genetics **41**, 824 - 828 (2009)
Published online: 14 June 2009 | doi:10.1038/ng.396

Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20

The Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene)[1]

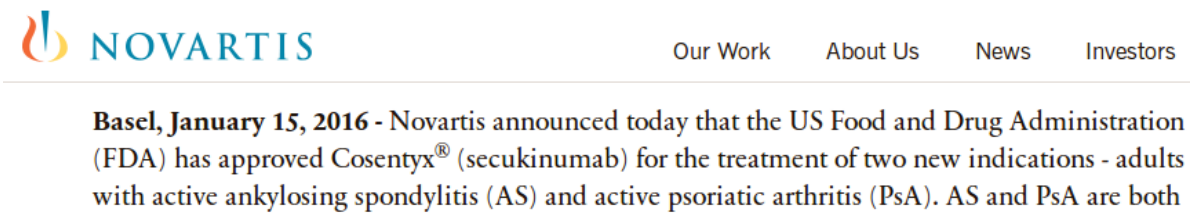
HaemGen consortium

Nature Genetics **41**, 1182 - 1190 (2009)
Published online: 11 October 2009 | doi:10.1038/ng.467

A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium

Nicole Soranzo[1,2,45], Tim D Spector[2,45], Massimo Mangino[2,45], Brigitte

Ankylosing spondylitis



P. Visscher, M. Brown, M. McCarthy, J. Yang. **Five years of GWAS discovery.** AJHG 2012.

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 effets of **multiple SNPs.**

Integrating prior knowledge

Use additional data and prior knowledge to constrain the feature selection procedure.

- **Consistant** with previously established knowledge
- More easily **interpretable**
- **Statistical power.**

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

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

Regularized relevance

Set $\mathcal{V}$ of p variables.

- **Relevance score** $R : 2^{\mathcal{V}} \to \mathbb{R}$

  Quantifies the importance of any subset of variables for the question under consideration.

  Ex : correlation, HSIC, statistical test of association.

- **Structured regularizer** $\Omega : 2^{\mathcal{V}} \to \mathbb{R}$

  Promotes a sparsity pattern that is compatible with the constraint on the feature space.

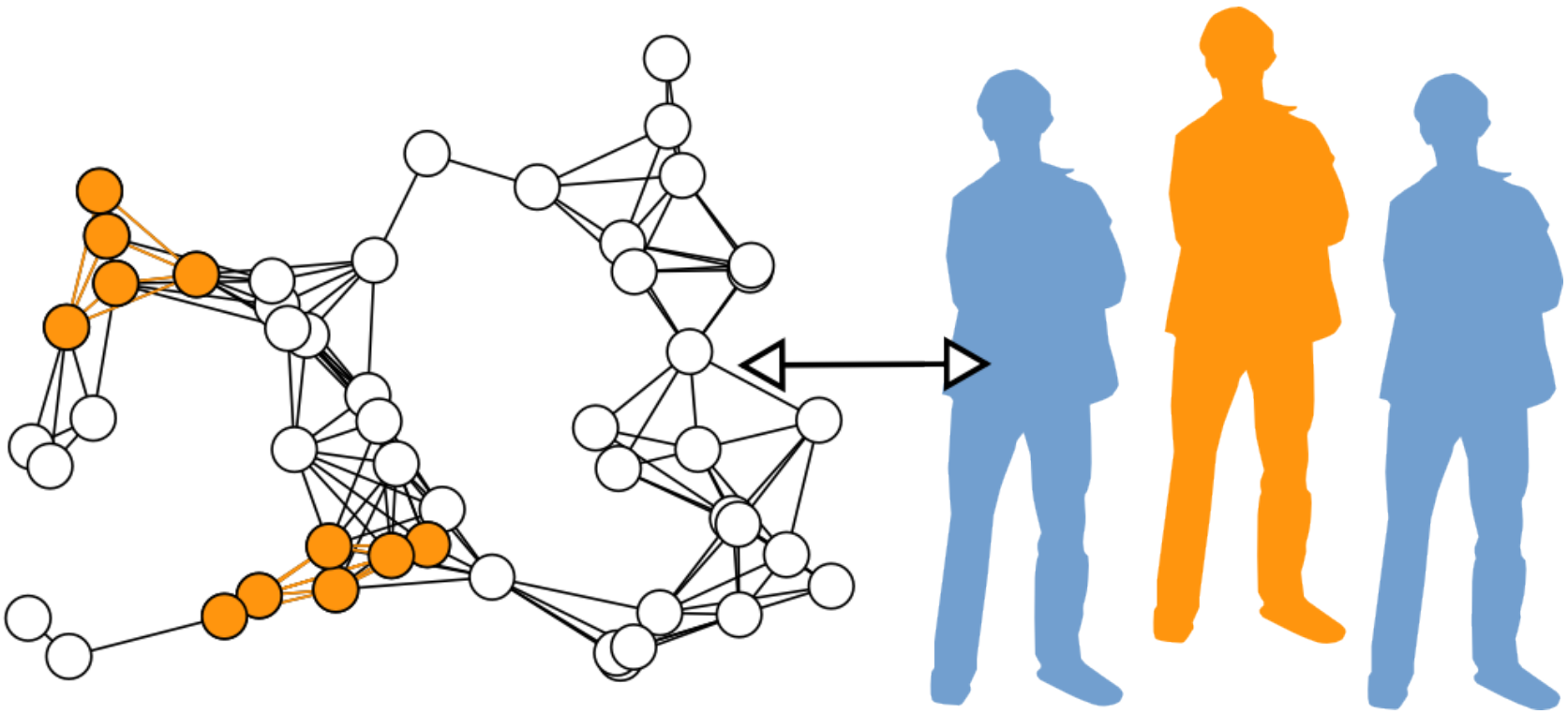
  Ex : cardinality $\Omega : \mathcal{S} \mapsto |\mathcal{S}|$.

- **Regularized relevance**

$$\underset{\mathcal{S} \subseteq \mathcal{V}}{\arg\max}\, R(\mathcal{S}) - \lambda \Omega(\mathcal{S})$$

Network-guided multi-locus GWAS

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



Network-guided GWAS

- **Additive test of association** SKAT [Wu et al. 2011]

$$R(\mathcal{S}) = \sum_{i \in \mathcal{S}} c_i \qquad c_i = (\mathbf{G}^\top (\mathbf{y} - \mu))_i^2$$

- **Sparse Laplacian regularization**

$$\Omega : \mathcal{S} \mapsto \sum_{i \in \mathcal{S}} \sum_{j \notin \mathcal{S}} W_{ij} + \alpha |\mathcal{S}|$$

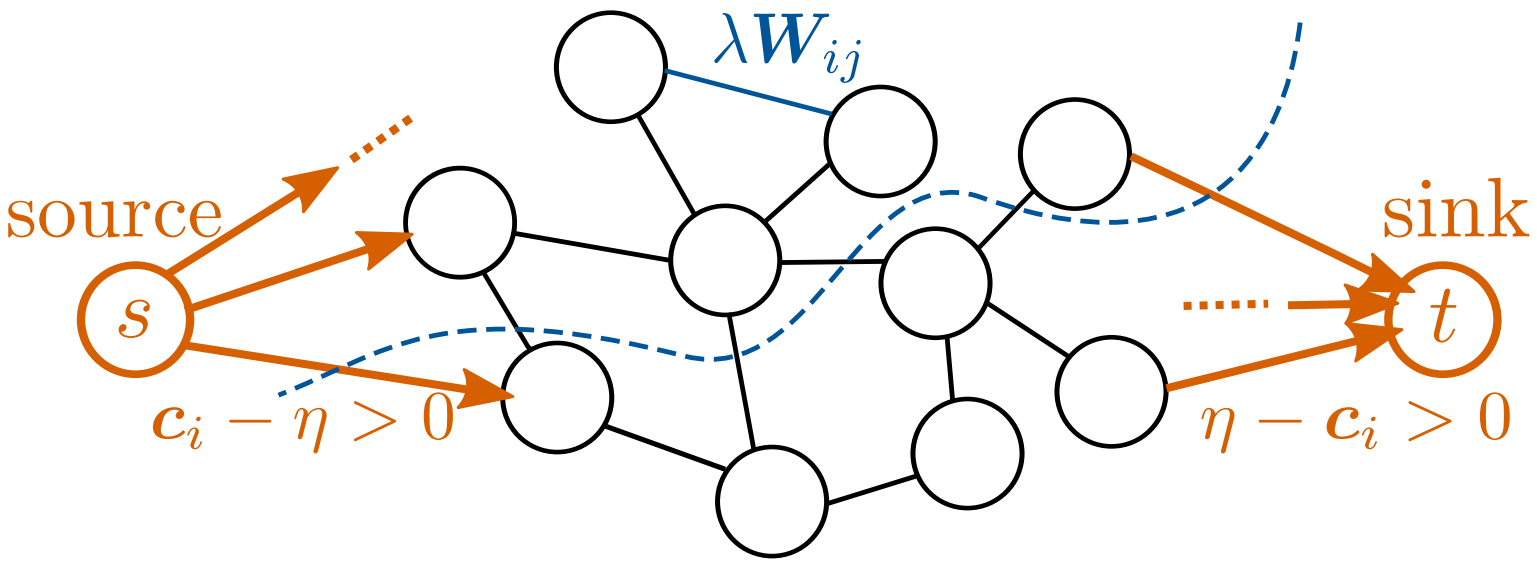
- **Regularized maximization of** R

$$\underset{\mathcal{S} \subseteq \mathcal{V}}{\arg\max} \underbrace{\sum_{i \in \mathcal{S}} c_i}_{\text{association}} - \underbrace{\eta\,|\mathcal{S}|}_{\text{sparsity}} - \underbrace{\lambda \sum_{i \in \mathcal{S}} \sum_{j \notin \mathcal{S}} W_{ij}}_{\text{connectivity}}$$

Minimum cut reformulation

The graph-regularized maximization of score $Q(*)$ is equivalent to a s/t-min-cut for a graph with adjacency matrix $\mathbf{A}$ and two additional nodes s and t, where $\mathbf{A}_{ij} = \lambda \mathbf{W}_{ij}$ for $1 \leq i, j \leq p$ and the weights of the edges adjacent to nodes s and t are defined as

$$\mathbf{A}_{si} = \begin{cases} c_i - \eta & \text{if } c_i > \eta \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \mathbf{A}_{it} = \begin{cases} \eta - c_i & \text{if } c_i < \eta \\ 0 & \text{otherwise} \end{cases}.$$



SConES: Selecting **Con**nected **E**xplanatory **S**NPs.

Comparison partners

- **Univariate linear regression** $y_k = \alpha_0 + \beta \mathbf{G}_k^i$
- **Lasso**

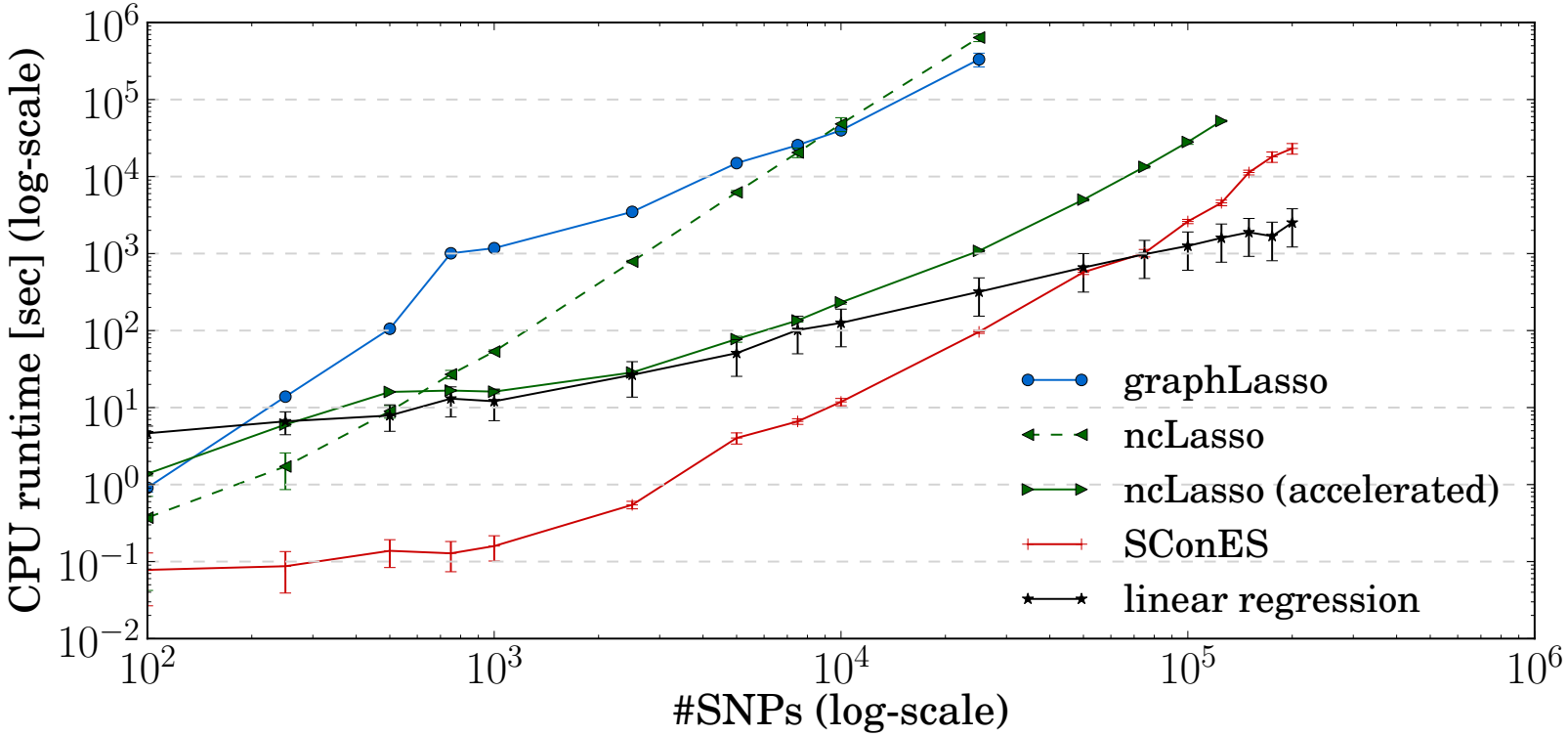
$$\underset{\beta \in \mathbb{R}^p}{\arg\min} \quad \underbrace{\frac{1}{2} ||\mathbf{y} - \mathbf{G}\beta||_2^2}_{\text{loss}} + \underbrace{\eta \ ||\beta||_1}_{\text{sparsity}}$$

- **Feature selection with sparsity and connectivity constraints**

$$\underset{\beta \in \mathbb{R}^p}{\arg\min} \quad \underbrace{\mathcal{L}(\mathbf{y}, \mathbf{G}\beta)}_{\text{loss}} + \underbrace{\eta \ ||\beta||_1}_{\text{sparsity}} + \underbrace{\lambda \, \Omega(\beta)}_{\text{connectivity}}$$

  - **ncLasso**: network connected Lasso [Li and Li, Bioinformatics 2008]
  - Overlapping group Lasso [Jacob et al., ICML 2009]
    - **groupLasso**: E.g. SNPs near the same gene grouped together
    - **graphLasso**: 1 edge = 1 group.

Runtime



$n = 200$ exponential random network (2 % density)

Experiments: Performance on simulated data

- Arabidopsis thaliana genotypes

  n=500 samples, p=1 000 SNPs
  TAIR **Protein-Protein Interaction data** $\sim 50.10^6$ edges

- Higher **power** and lower **FDR** than comparison partners

  except for groupLasso when groups = causal structure

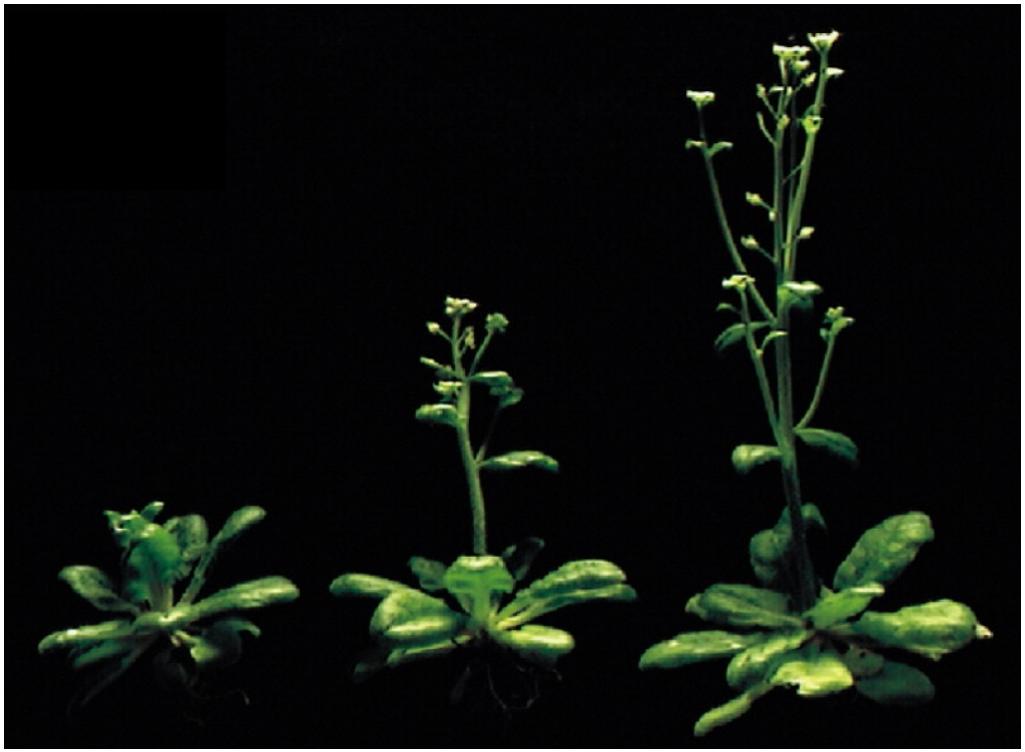
- Fairly robust to **missing edges**
- Fails if network is **random.**

Arabidopsis thaliana flowering time

17 flowering time phenotypes
[Atwell et al., Nature, 2010]

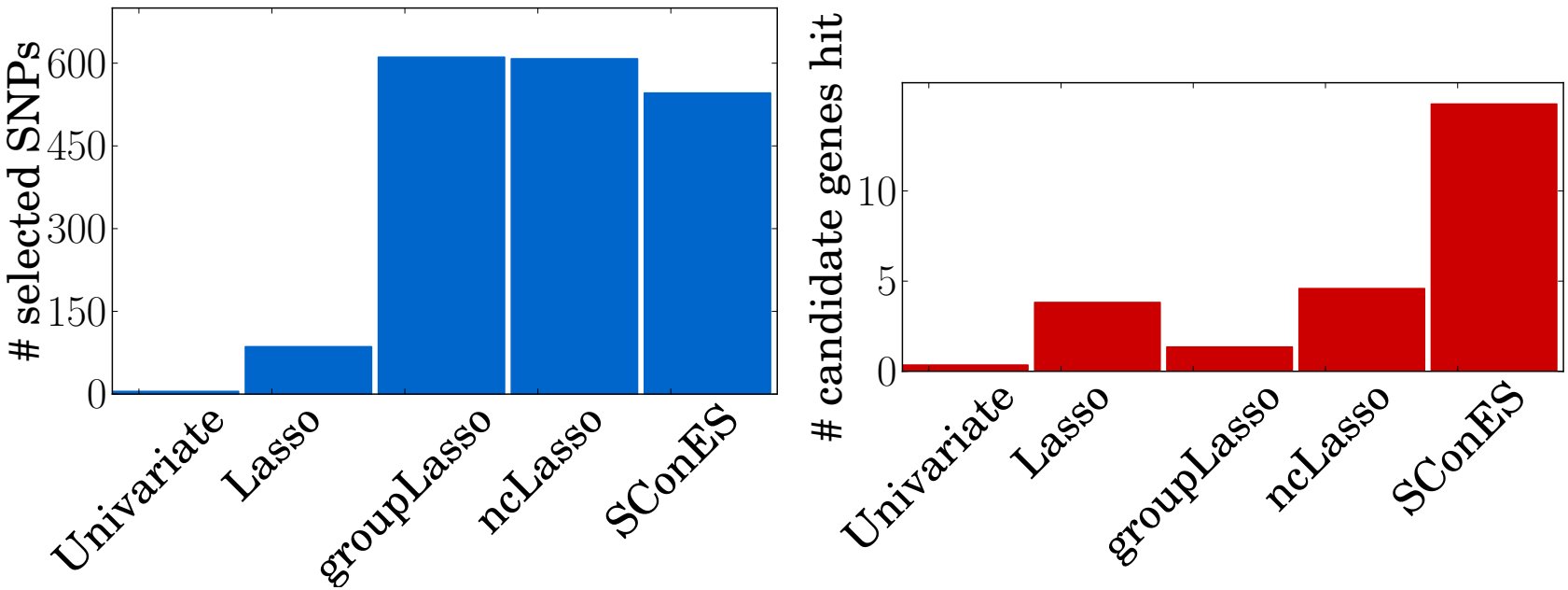
$p \sim 170\,000$ SNPs
(after MAF filtering)
$n \sim 150$ samples

165 **candidate genes**
[Segura et al., Nat Genet 2012]



Correction for **population structure**: regress out PCs.

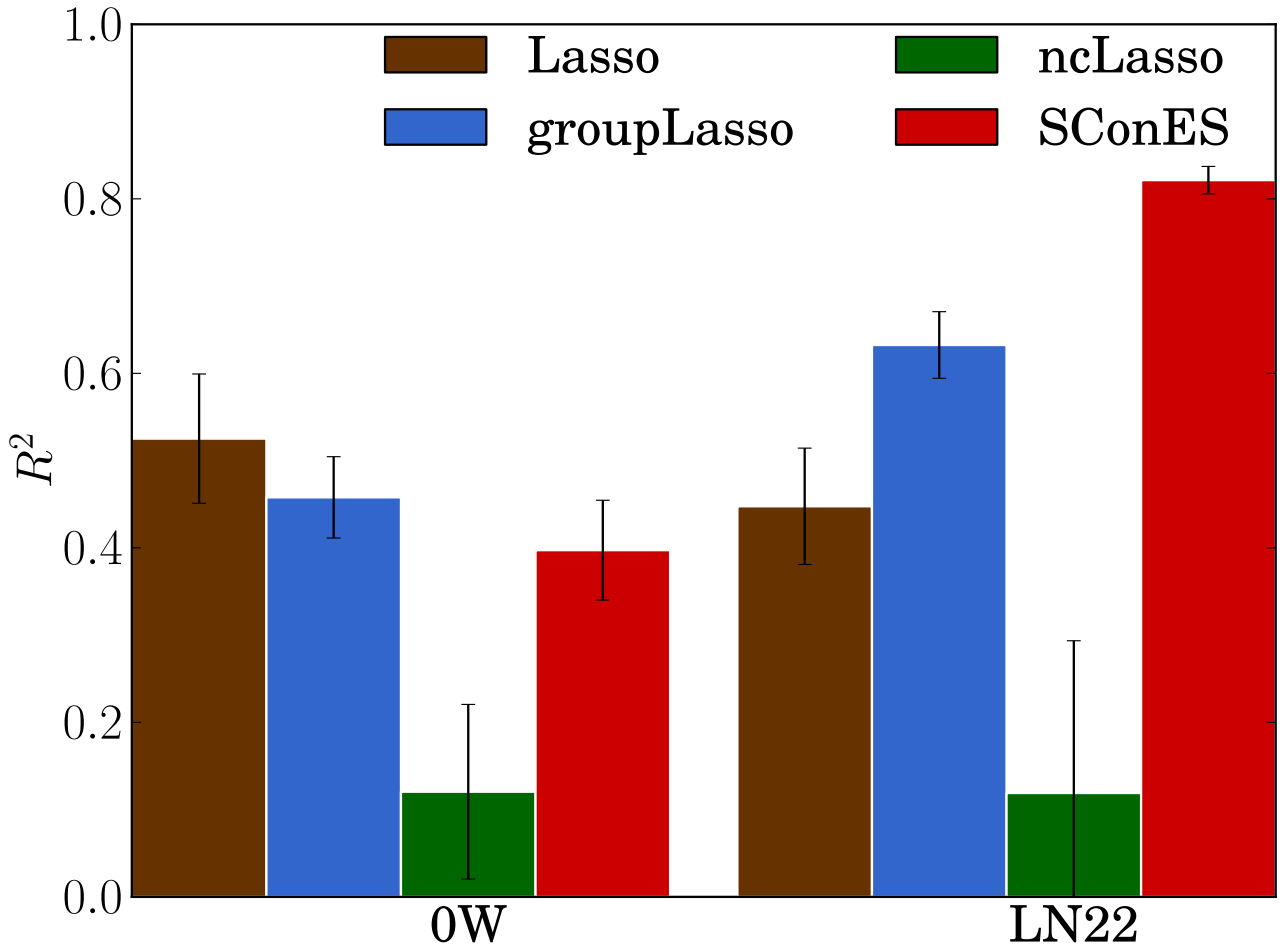
Arabidopsis thaliana flowering time



- SConES selects **about as many SNPs** as other network-guided approaches but **detects more candidates.**

Arabidopsis thaliana flowering time

Predictivity of selected SNPs



SConES: Selecting Connected Explanatory SNPs

- selects connected, explanatory SNPs;
- incorporates large networks into GWAS;
- is efficient, effective and robust.

C.-A. Azencott, D. Grimm, M. Sugiyama, Y. Kawahara and K. Borgwardt (2013) **Efficient network-guided multi-locus association mapping with graph cuts**, Bioinformatics 29 (13), i171–i179 doi:10.1093/bioinformatics/btt238

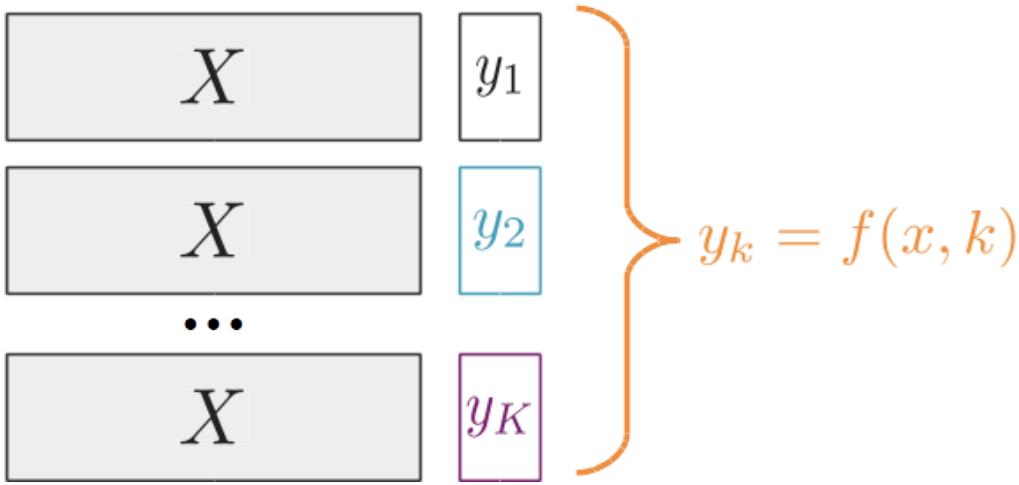
`https://github.com/chagaz/scones`

`https://github.com/chagaz/sfan`

`https://github.com/dominikgrimm/easyGWASCore`

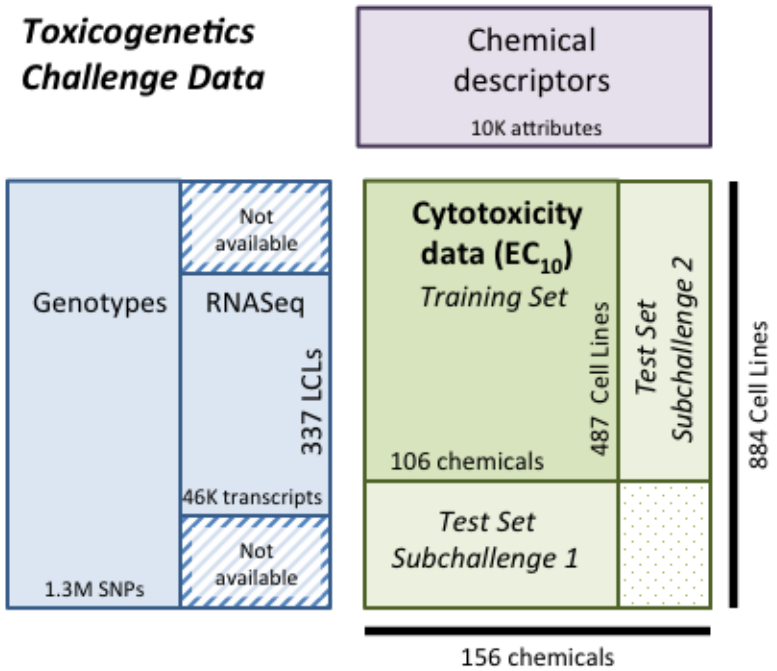
Multi-trait GWAS

Increase sample size by **jointly** performing GWAS for **multiple related phenotypes**



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-SConES

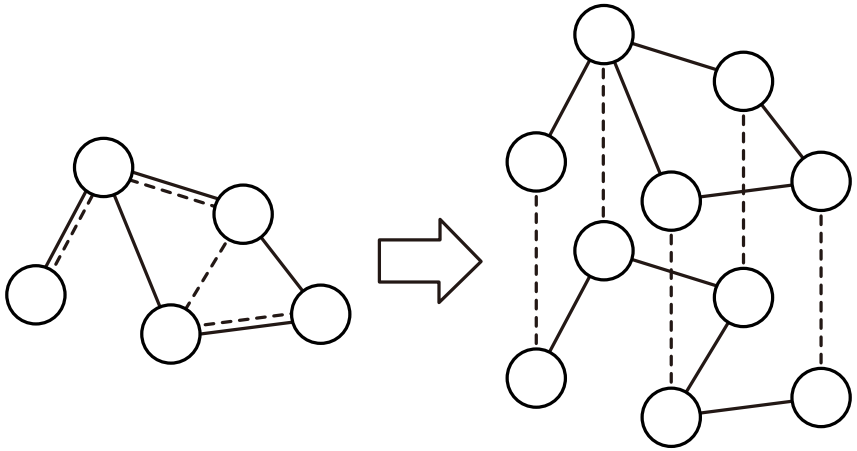
T **related phenotypes.**

- Goal: obtain **similar sets of features** on related tasks.

$$\underset{\mathcal{S}_1,\ldots,\mathcal{S}_T \subseteq \mathcal{V}}{\arg\max} \sum_{t=1}^{T} \left(\sum_{i \in \mathcal{S}} c_i - \eta \left| \mathcal{S} \right| - \lambda \sum_{i \in \mathcal{S}} \sum_{j \notin \mathcal{S}} W_{ij} - \underbrace{\mu \left| \mathcal{S}_{t-1} \,\Delta\, \mathcal{S}_t \right|}_{\text{task sharing}} \right)$$

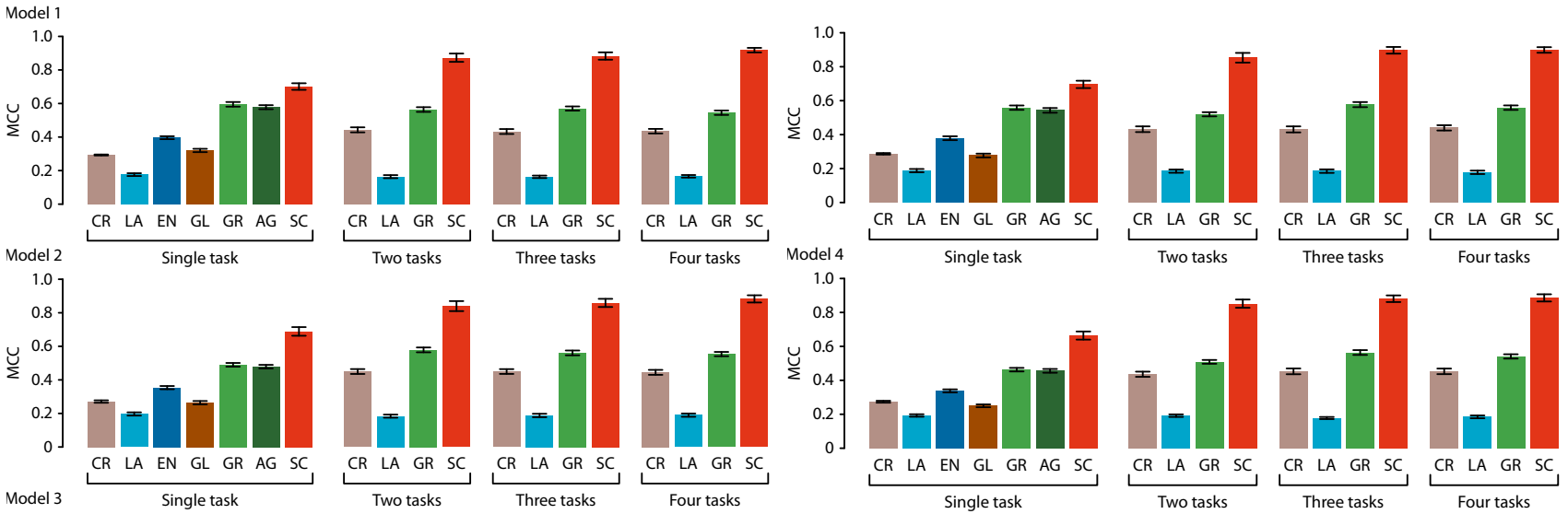
$$\mathcal{S} \,\Delta\, \mathcal{S}' = (\mathcal{S} \cup \mathcal{S}') \setminus (\mathcal{S} \cap \mathcal{S}') \qquad \text{(symmetric difference)}$$

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



Multi-SConES: Multiple related tasks

Simulations: retrieving causal features



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 doi:10.1137/1.9781611973440.23

https://github.com/mahito-sugiyama/Multi-SConES

https://github.com/chagaz/sfan

Leveraging similarity between tasks

Use **prior knowledge** about the **relationship** between the tasks: $\Omega \in \mathbb{R}^{T \times T}$

$$\underset{\mathcal{S}_1,\ldots,\mathcal{S}_T \subseteq \mathcal{V}}{\arg\max} \sum_{t=1}^{T} \left(\sum_{i \in \mathcal{S}} c_i - \eta \left| \mathcal{S} \right| - \lambda \sum_{i \in \mathcal{S}} \sum_{j \notin \mathcal{S}} W_{ij} - \underbrace{\mu \sum_{u=1}^{T} \sum_{i \in \mathcal{S}_t \cap \mathcal{S}_u} \Omega_{tu}^{-1}}_{\text{task sharing}} \right)$$

Can also be mapped to a meta-network.

Code: `http://github.com/chagaz/sfan`

Multiplicative Multitask Lasso with Task Descriptors

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

$$\underset{\beta \in \mathbb{R}^{T \times p}}{\arg\min} \underbrace{\mathcal{L}\left(y_m^t, \sum_{i=1}^{p} \beta_i g_{mi}^t\right)}_{\text{loss}} + \underbrace{\lambda \sum_{i=1}^{p} ||\beta_i||_2}_{\text{task sharing}}$$

- **Multilevel Multitask Lasso** [Lozano and Swirszczw, 2012]

$$\underset{\theta \in \mathbb{R}_+^p, \gamma \in \mathbb{R}^{T \times p}}{\arg\min} \underbrace{\mathcal{L}\left(y_m^t, \sum_{i=1}^{p} \theta_i \gamma_i^t g_{mi}^t\right)}_{\text{loss}} + \underbrace{\lambda_1 ||\theta||_1}_{\text{sparsity}} + \underbrace{\lambda_2 \sum_{i=1}^{p} \sum_{t=1}^{T} |\gamma_i^t|}_{\text{task sharing}}$$

- **Multiplicative Multitask Lasso with Task Descriptors**

$$\underset{\theta \in \mathbb{R}_+^p, \alpha \in \mathbb{R}^{p \times L}}{\arg\min} \underbrace{\mathcal{L}\left(y_m^t, \sum_{i=1}^{p} \theta_i \left(\sum_{l=1}^{L} \alpha_{il} d_l^t\right) g_{mi}^t\right)}_{\text{loss}} + \underbrace{\lambda_1 ||\theta||_1}_{\text{sparsity}} + \underbrace{\lambda_2 \sum_{i=1}^{p} \sum_{l=1}^{L} |\alpha_{il}|}_{\text{task sharing}}$$

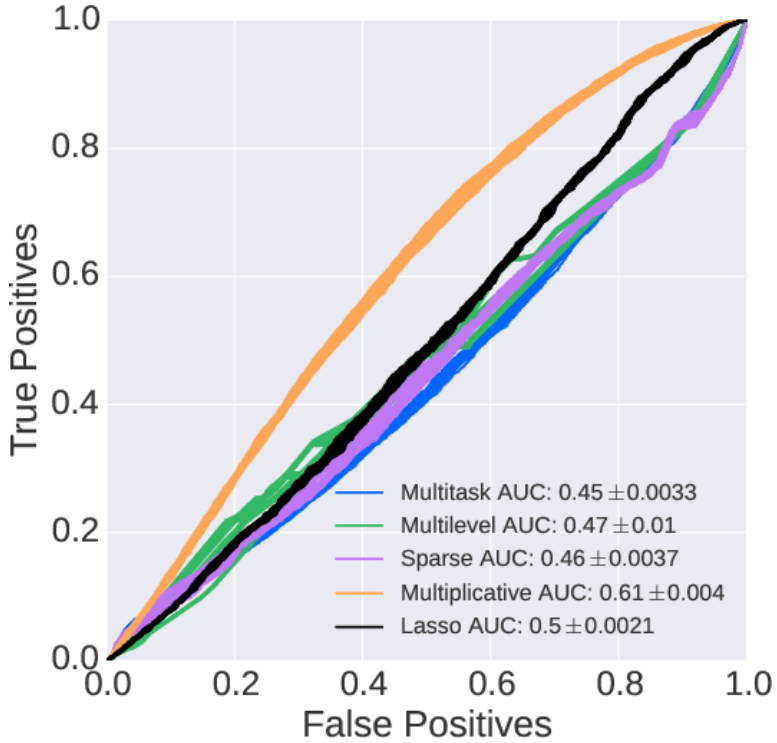
Multiplicative Multitask Lasso with Task Descriptors

$$\underset{\theta \in \mathbb{R}_+^p, \alpha \in \mathbb{R}^{p \times L}}{\arg\min} \underbrace{\mathcal{L}\left(y_m^t, \sum_{i=1}^{p} \theta_i \left(\sum_{l=1}^{L} \alpha_{il} d_l^t\right) g_{mi}^t\right)}_{\text{loss}} + \underbrace{\lambda_1 ||\theta||_1}_{\text{sparsity}} + \underbrace{\lambda_2 \sum_{i=1}^{p} \sum_{l=1}^{L} |\alpha_{il}|}_{\text{task sharing}}$$

- On **simulations:**
  - **Sparser** solution
  - Better **recovery of true features** (higher PPV)
  - Improved **stability**
  - Better **predictivity** (RMSE).

Multiplicative Multitask Lasso with Task Descriptors

- Making predictions for tasks for which you have **no data**.



V. Bellón, V. Stoven, and C.-A. Azencott (2016) **Multitask feature selection with task descriptors**, PSB.

https://github.com/vmolina/MultitaskDescriptor

Limitations of current approaches

- **Robustness/stability**

  Recovering the same SNPs when the data changes slightly.

- **Complex epistasis patterns**
  - Limited to additive or quadrative effects
  - Some work on e.g. random forests + importance score.

- **Statistical significance**
  - Computing p-values
  - Correcting for multiple hypotheses.

https://github.com/chagaz/

CBIO: Víctor Bellón, Yunlong Jiao, Véronique Stoven, Athénaïs Vaginay, Nelle Varoquaux, Jean-Philippe Vert, Thomas Walter.

MLCB Tübingen: Karsten Borgwardt, Aasa Feragen, Dominik Grimm, Theofanis Karaletsos, Niklas Kasenburg, Christoph Lippert, Barbara Rakitsch, Damian Roqueiro, Nino Shervashidze, Oliver Stegle, Mahito Sugiyama.

MPI for Intelligent Systems: Lawrence Cayton, Bernhard Schölkopf.

MPI for Developmental Biology: Detlef Weigel.

MPI for Psychiatry: André Altmann, Tony Kam-Thong, Bertram Müller-Myhsok, Benno Pütz.

Osaka University: Yoshinobu Kawahara.

source: http://www.flickr.com/photos/wwworks/