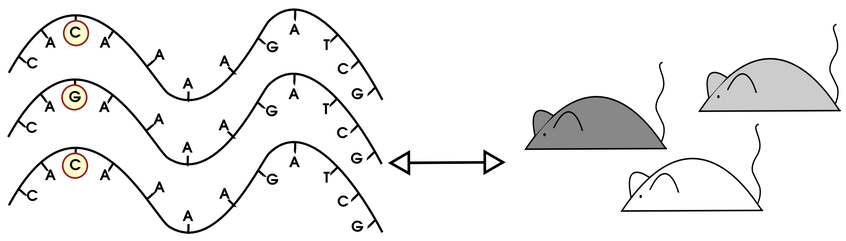
Efficient network-guided multi-locus association mapping with graph cuts

Chloé-Agathe Azencott[1], Dominik Grimm[1], Mahito Sugiyama[1], Yoshinobu Kawahara[2] and Karsten Borgwardt[1,3]

[1]Machine Learning and Computational Biology Research Group
Max Planck Institute for Intelligent Systems and Max Planck Institute for Developmental Biology, Tübingen,
[2] ISIR, Osaka University, [3]ZBIT, Universität Tübingen

Genome-Wide Association Studies (GWAS)



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

Which SNPs explain the phenotype?

For each SNP **individually**:

$$\underbrace{y_i}_{\text{phenotype}} = \alpha_0 + \underbrace{\boldsymbol{\alpha}^\top \mathbf{X}_i}_{\text{covariates}} + \underbrace{\beta_i g_i}_{\text{genetics}} + \epsilon_i$$

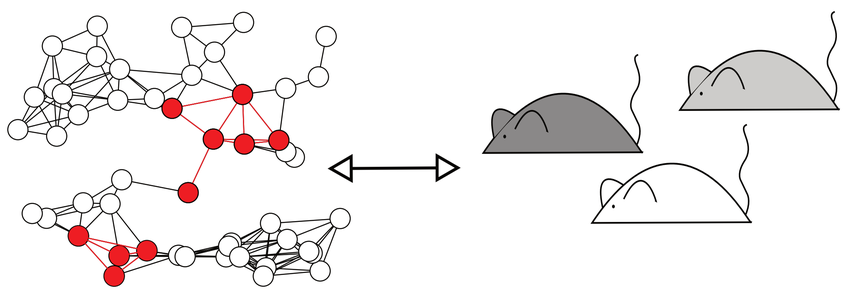
Missing heritability

GWAS **fail to explain most of the heritability** of most complex traits.

- Rare SNPs, small effect sizes, environmental / epigenetic factors
- **Joint effects** of multiple SNPs

Find a *small* number of SNPs, *connected* in a given network, that *jointly* explain the phenotype.

Networks of SNPs



- **ncLasso**: Network Connected LASSO [1]
- **groupLasso**, **graphLasso**: Overlapping Group LASSO [2]
- **Structured sparsity penalty** [3]
- **Path-coding penalties for DAGs** [4]

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

SConES: Selecting Connected Explanatory SNPs

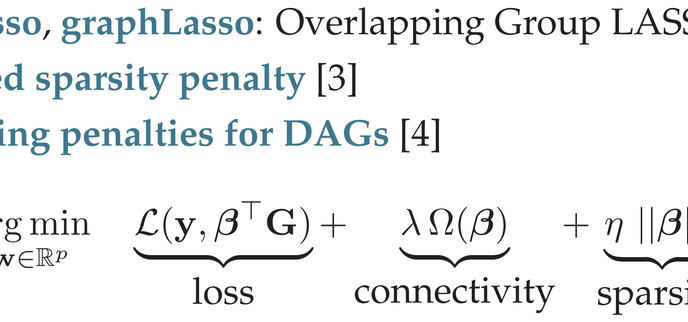
$$\underset{\mathbf{f}\in\{0,1\}^n}{\arg\max} \quad \underbrace{\mathbf{c}^\top \mathbf{f}}_{\substack{\text{linear} \\ \text{association*}}} - \underbrace{\lambda\, \mathbf{f}^\top \mathbf{L} \mathbf{f}}_{\text{connectivity}} - \underbrace{\eta\, ||\mathbf{f}||_0}_{\text{sparsity}}$$

$$f_i = \begin{cases} 1 & \text{if } i \text{ selected} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Laplacian: } \mathbf{L} = \mathbf{D} - \mathbf{W}$$

$$\mathbf{f}^\top \mathbf{L} \mathbf{f} = \sum_{i\sim j} (f_i - f_j)^2$$

Max-flow solution



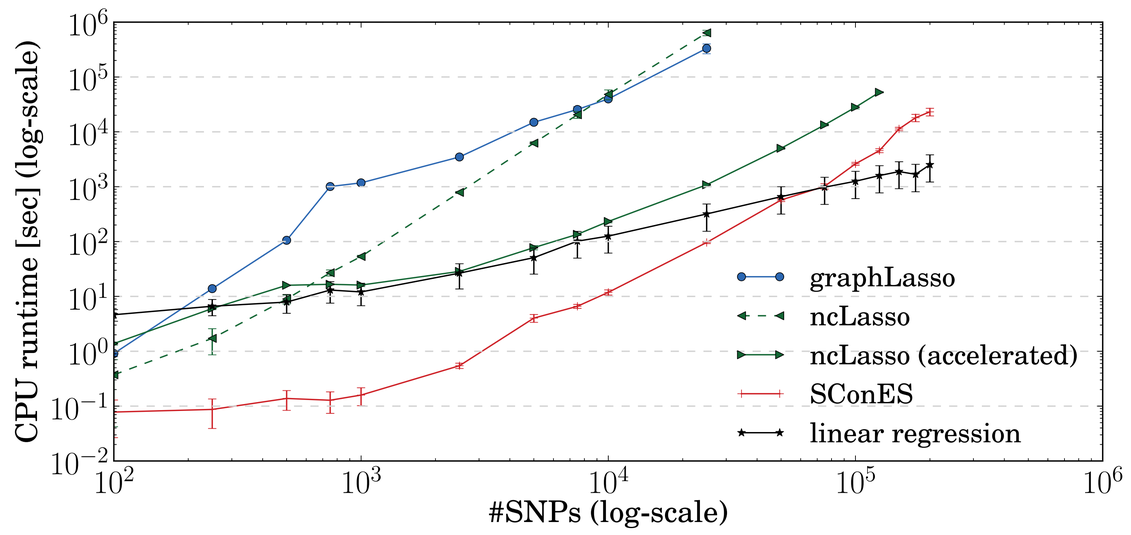
***Sequence Kernel Association Test** [5]

$$\underbrace{y_i}_{\text{phenotype}} = \alpha_0 + \underbrace{\boldsymbol{\alpha}^\top \mathbf{X}_i}_{\text{covariates}} + \underbrace{\boldsymbol{\beta}^\top \mathbf{G}_i^{\mathcal{S}}}_{\substack{\text{genotypic} \\ \text{variation}}} + \epsilon_i$$

$H_0 : \boldsymbol{\beta} = 0 \quad \rightarrow \hat{\mu} = \hat{\alpha}_0 + \mathbf{X}_i \hat{\boldsymbol{\alpha}}$

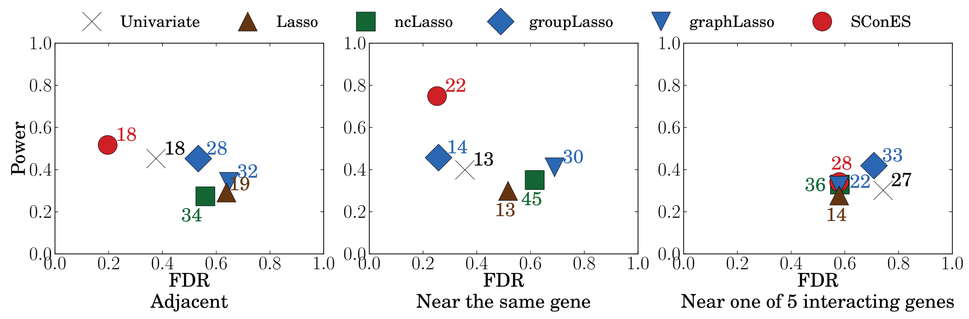
Variance component test statistic: $Q(\mathbf{f}) = \mathbf{c}^\top \mathbf{f}$ $\qquad \mathbf{c}_i = (\mathbf{G}^\top (\mathbf{y} - \hat{\mu}))_i^2$

Runtime



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

Simulations

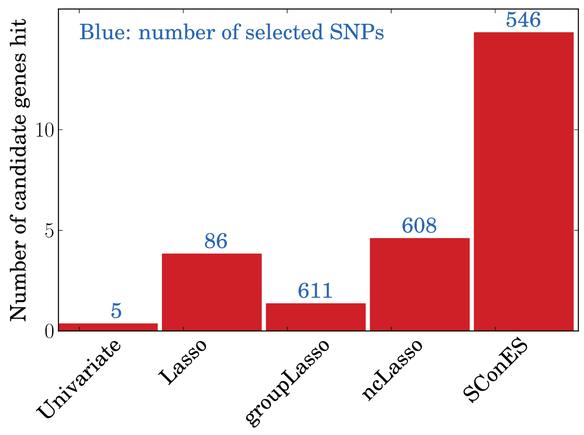


1,000 SNPs, 500 samples, 10 10-fold cv. Numbers denote number of selected SNPs.

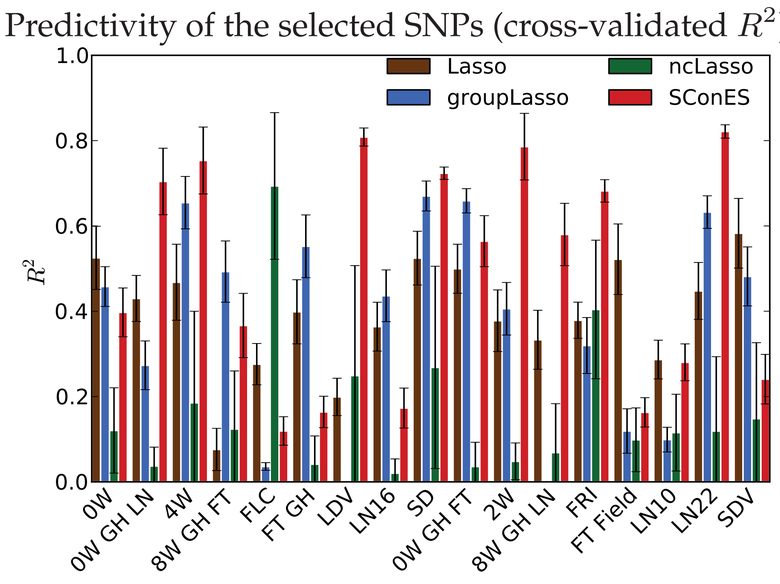
- High **power**
- Low **FDR** (False Discovery Rate)
- Robust to **missing edges**

Arabidopsis thaliana flowering time

Candidate genes hit



Predictivity of the selected SNPs (cross-validated R^2)



Future work

- Multiple phenotypes
- More models of association
- p-values
- SNP network
  - Other regularizers
  - Define the network
  - Learn the network

References

[1] C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

[2] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *ICML*, pages 433–440, 2009.

[3] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *ICML*, pages 417–424, New York, NY, USA, 2009.

[4] J. Mairal and B. Yu. Path coding penalties for directed acyclic graphs. In *NIPS OPT*, 2011.

[5] M. C. Wu, S. Lee, and et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *AJHG*, 89(1):82–93, 2011.

Contact & Code

chloe-agathe.azencott@tuebingen.mpg.de

http://agkb.is.tuebingen.mpg.de

@cazencott

@AGKBorgwardt