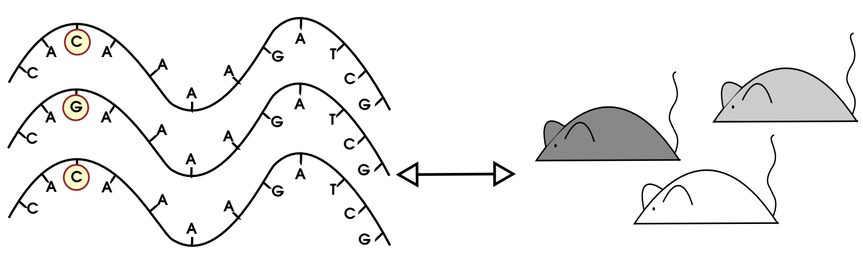
Efficiently mapping phenotypes to networks of genetic loci

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

[1]Machine Learning and Computational Biology Research Group, Max Planck Institutes Tübingen (Germany)
[2]ISIR, Osaka University (Japan), [3]ZBIT, Eberhard Karls Universität Tübingen (Germany)

Unterstützt von / Supported by
Alexander von Humboldt Stiftung/Foundation

Genome-Wide Association Studies (GWAS)



- 250 000 – 10 000 000 Single Nucleotide Polymorphisms (SNPs)
- 100 – 10 000 subjects

→ **Which SNPs explain the phenotype?**

For each SNP **individually**:

$$\underbrace{y_i}_{\text{Phenotype}} = \alpha_0 + \underbrace{\alpha^\top X_i}_{\text{Covariates}} + \underbrace{\beta g_i}_{\text{Genetics}} + \epsilon_i$$

Missing heritability

Single-locus GWAS **fail to explain most of the heritability** of most complex traits

- Rare variants, undetected SNPs with small effect size, ...
- **Joint effects** of multiple SNPs

Multiple-Locus GWAS

Find *multiple* SNPs that *jointly* explain the phenotype

Issue: search space grows exponentially with the number of such SNPs

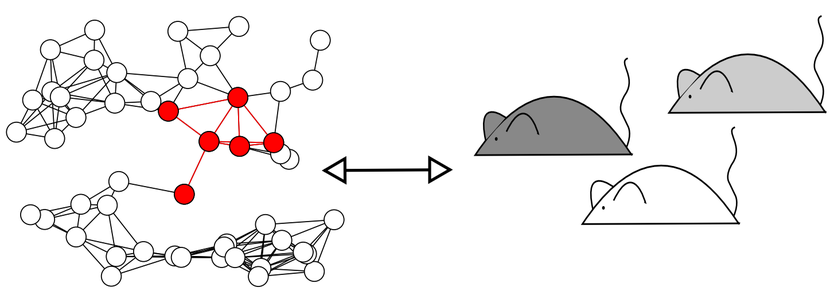
- **Reduce the search space**
  - **Use prior knowledge**
  - **Space-pruning & sampling techniques**

→ Limited to **binary / discrete phenotypes / genotypes** and **pairs of SNPs**

- **GPGPU approaches**

→ Hardly scale to more than **pairs of SNPs**

Networks of SNPs



Selection of graph-structured features

- **Structured sparsity** [Huang et al. 2009]
- **Network-constrained lasso** [Li et al. 2008]
- **Overlapping group-lasso** [Jacob et al. 2009]

→ Do not **scale** to hundreds of thousands of SNPs

- **Path-coding penalties** [Mairal and Yu 2012]

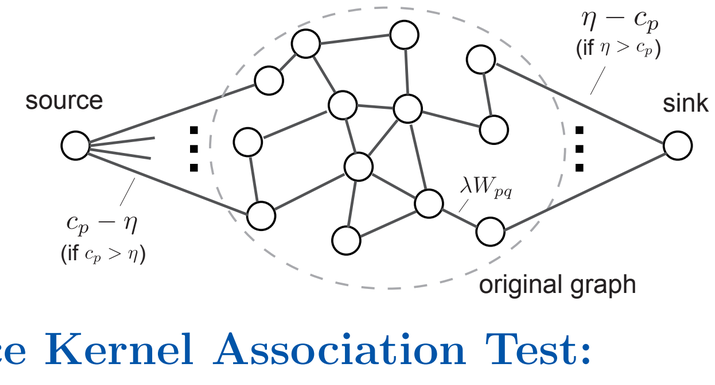
→ **Directed acyclic graphs** only

SOS: Subnetworks of Optimal SNPs

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

$$\arg\max_{\boldsymbol{f}\in\{0,1\}^n} \quad \underbrace{\boldsymbol{c}^\top \boldsymbol{f}}_{\substack{\text{linear}\\ \text{association}^*}} - \underbrace{\lambda\, \boldsymbol{f}^\top \boldsymbol{L} \boldsymbol{f}}_{\text{connectivity}} - \underbrace{\eta\ ||\boldsymbol{f}||_1}_{\text{sparsity}}$$

Max-flow solution:



***Sequence Kernel Association Test:**

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

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

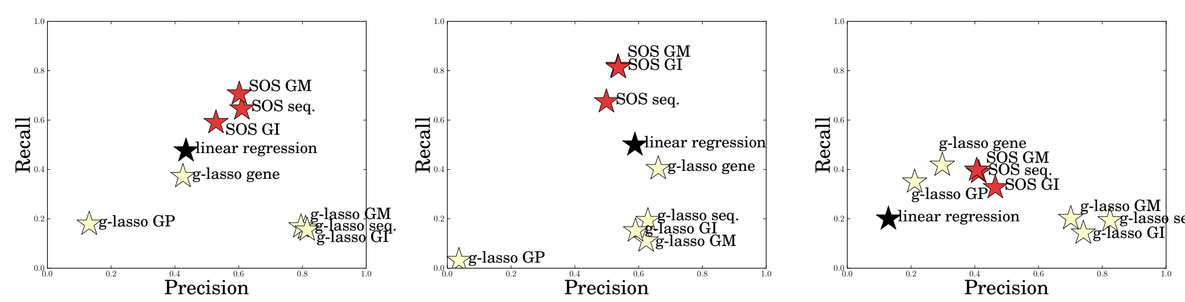
Variance component test statistic:

$$Q^{\mathcal{S}} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^\top \boldsymbol{K}^{\mathcal{S}} (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \qquad \boldsymbol{K}^{\mathcal{S}} = \boldsymbol{G}^{\mathcal{S}} \boldsymbol{G}^{\mathcal{S}\top}$$

$$Q(\boldsymbol{f}^{\mathcal{S}}) = \boldsymbol{c}^\top \boldsymbol{f}^{\mathcal{S}} \qquad \boldsymbol{f}_p^{\mathcal{S}} = \begin{cases} 1 & \text{if } p \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases}$$

$$\boldsymbol{c}_p = (\boldsymbol{G}^\top (\boldsymbol{y} - \hat{\boldsymbol{\mu}}))_p^2$$

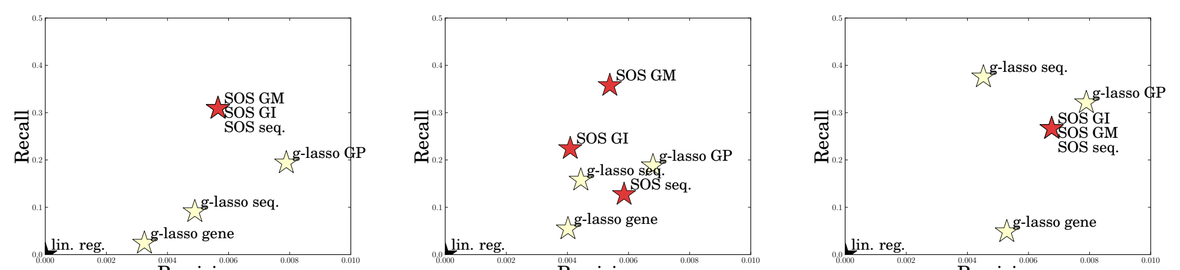
Results on simulated data



Causal SNPs in a sequence | Causal SNPs in a gene | Causal SNPs in a pathway

"GM": SNPs near the same gene are connected; "gene": groups of SNPs near the same gene; "GI": SNPs near either of 2 interacting genes are connected; "GP": groups of SNPs near either of 2 interacting genes. 1,000 SNPs, 500 samples, 10 10-fold cv.

Arabidopsis thaliana flowering time



4W | FTGH | LN22

$\sim 250,000$ SNPs, ~ 800 samples, 10-fold cv. g-lasso GM, GI did not complete after 3 days.

Prec/recall wrt candidate genes.

Further research topics

- Generalization to **submodular penalties**
- Extension to **multiple phenotypes**
- Extension to **epistatic effects**
- Definition of the **SNP network**

References

- C.-A. Azencott, D. Grimm, Y. Kawahara, K. Borgwardt. A min-cut solution to mapping phenotypes to networks of genetic markers. *arXiv:1211.2315* 2012
- J. Huang et al. Learning with structured sparsity. *JMLR* 2011
- L. Jacob et al. Group lasso with overlap and graph lasso. *ICML* 2009
- C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* 2008
- J. Mairal and B. Yu. Path-coding penalties for directed acyclic graphs. *arXiv:1205.0079* 2012
- M. C. Wu et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011