Multi-locus genome-wide association studies

Chloé-Agathe Azencott

CBIO, Mines ParisTech – Institut Curie – INSERM U900, Paris (France)

October 21, 2016 – Krupp Symposium

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

Multi-locus genome-wide association studies

Chloé-Agathe Azencott

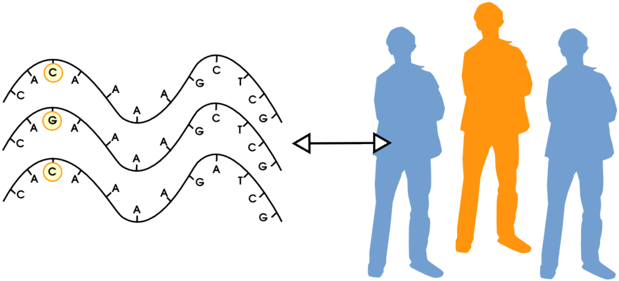
Machine Learning & Computational Biology Research Group
Max Planck Institute for Intelligent Systems &
Max Planck Institute for Developmental Biology

Tübingen (Germany)

March 2011–November 2013

Unterstützt von / Supported by
Alexander von Humboldt
Stiftung / Foundation

Genome-Wide Association Studies

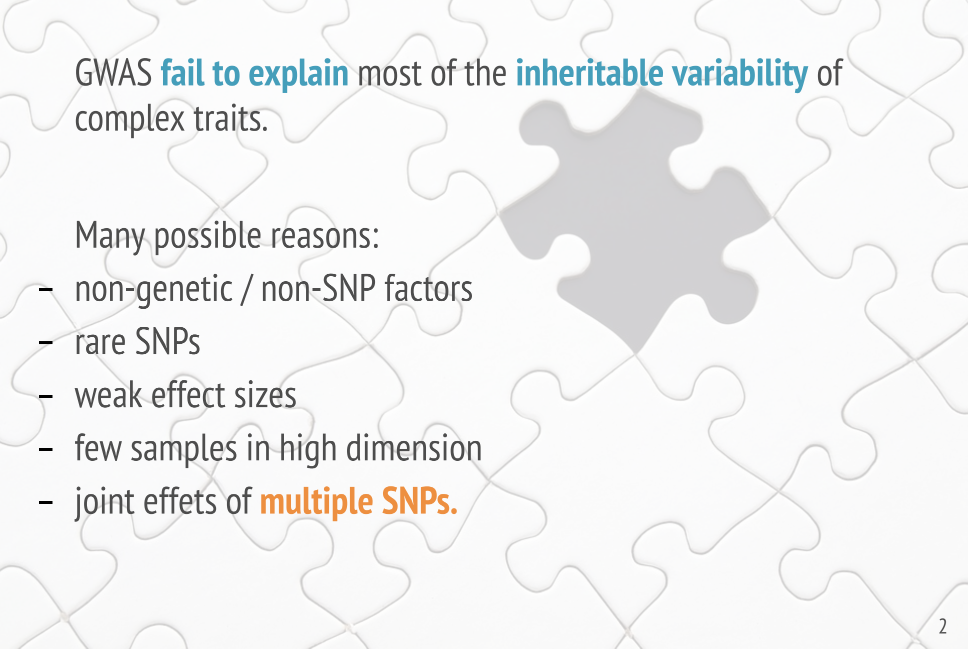


Which regions of the genome explain the phenotype?

Feature selection in **high dimension.**

- Technological advances:
  p = 10^5 – 10^7 Single Nucleotide Polymorphisms (SNPs)
  n = 10^2 – 10^4 samples.
- **Methodological** advances?

Missing heritability



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

Many possible reasons:

- non-genetic / non-SNP factors
- rare SNPs
- weak effect sizes
- few samples in high dimension
- joint effets of **multiple SNPs.**

Multi-locus GWAS

- **Epistasis:** known **synergetic effects** between genes
  - Enhance/suppress **cancer mutations** [Ashworth et al. 2011]

    Loss of VHL (tumor supressor) causes cellular senescense, unless Retinoblastoma (another tumor supressor) is also inactivated.
  - **Working memory** related brain activation [Tan et al. 2007]

    GRM3 adverse effect on prefrontal engagement only in presence of one variant of COMT.

→ Map **pairs of SNPs** to the phenotype.

Search space

10^{12} – 10^{14} SNP pairs

Computational burden → use **Graphical Processing Units**

IC 1101 (largest known galaxy) – Hubble Space Telescope.

EPIBLASTER

- **Difference in correlation** between SNPs:

$$\Delta_{(\text{SNP}_1,\text{SNP}_2)} = \left(\frac{1}{n_{\text{cases}}} \sum_{i \text{ case}} \text{SNP}_1^{(i)} \text{SNP}_2^{(i)} - \frac{1}{n_{\text{ctrls}}} \sum_{i \text{ ctrl}} \text{SNP}_1^{(i)} \text{SNP}_2^{(i)} \right)^2$$

- Limited to **qualitative phenotypes.**

T. Kam-Thong, D. Czamara, et al. (2011). **EPIBLASTER – Fast exhaustive two-locus epistasis detection strategy using graphical processing units.** European Journal of Human Genetics, 19 (4), 465–471 doi:10.1038/ejhg.2010.196

`http://www.psych.mpg.de/2046236/EPIBLASTER.zip`

EpiGPUHSIC

- Extend to **quantitative phenotypes** using the **Hilbert-Schmidt Independence Criterion**

$$\Delta_{(\mathsf{SNP}_1,\mathsf{SNP}_2)} = \left(\sum_i \mathsf{SNP}_1^{(i)}\mathsf{SNP}_2^{(i)}\mathsf{Phenotype}^{(i)}\right)^2$$

- Does not account for **main effects.**

T. Kam-Thong, B. Pütz, B. Müller-Myhsok, and K. M. Borgwardt. (2011) **Epistasis detection on quantitative phenotypes by exhaustive enumeration using GPUs.** Bioinformatics, 27 (13), i214–221 doi:10.1093/bioinformatics/btr218

`http://www.psych.mpg.de/2046246/EpiGPUHSIC.zip`

GLIDE

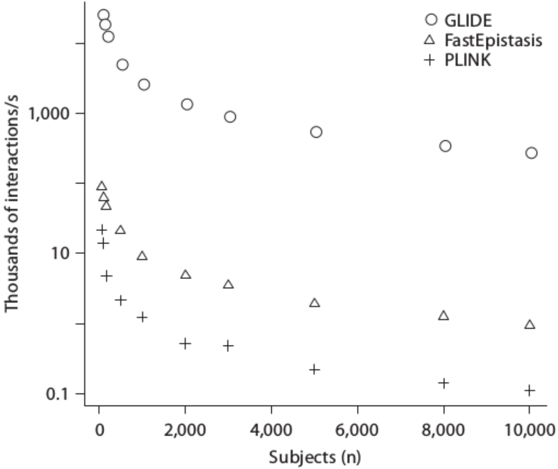
GPU-based **li**near regression for the **d**etection of **e**pistasis

$$\text{Phenotype} = \alpha\,\text{SNP}_1 + \beta\,\text{SNP}_2 + \gamma\,\text{SNP}_1 \times \text{SNP}_2 + \delta$$

- Is γ signficantly different from 0? $\rightarrow$ **t-test.**

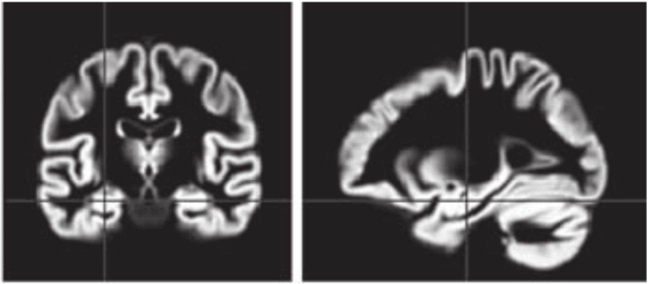
Runtime Performance

Synthetic data: 1 000 subjects, 5 000 SNPs
NVIDIA GTX 580 ($\sim$ \$450 in 2011)



Hippocampus Volume Epistasis Detection

- **GWAS study**: 567 genotyped subjects, about 10^6 SNPs



Hippocampus Volume Epistasis Detection

- **Single-locus** GWAS
  - 20 SNPs with significant main effects
  - 14 associated with hippocampal morphology and brain maturation
    $\rightarrow$ explain **18% of the variance**
- **Two-locus** GWAS
  - Runtime $\approx$ 3 days on a single GPU
  - 20 pairs with lowest p-values (2.6 10^{-13} – 2.6 $^{-11}$)
    - No significant main effects
      $\rightarrow$ 8 independent pairs, explain **40% of the variance**
- **Together** explain **50% of the variance.**

GLIDE

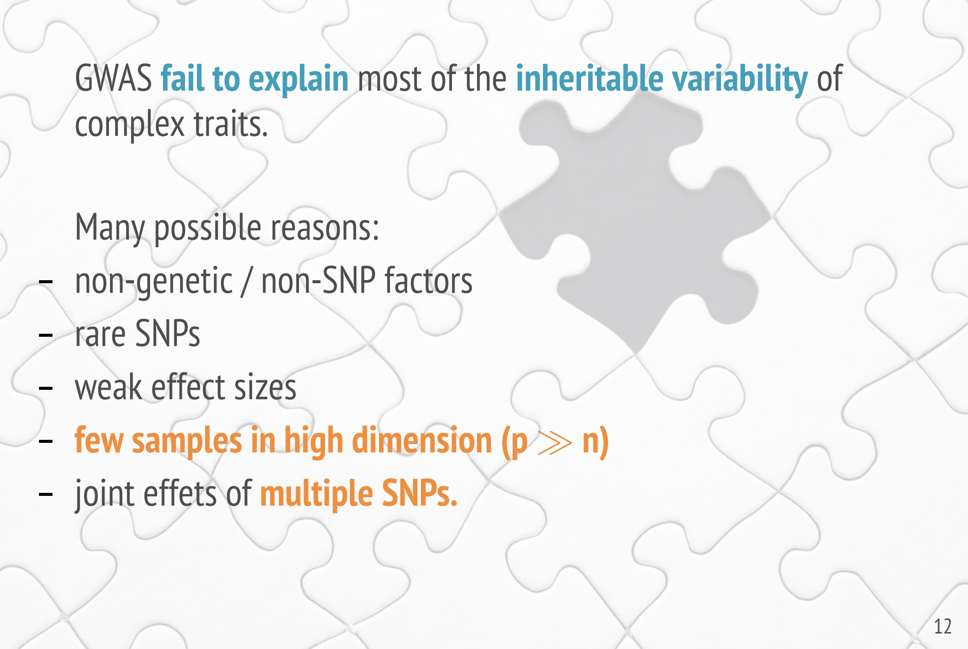
- Both phenotype and genotype can be **continuous**
- **Main effects** are accounted for.

T. Kam-Thong, C.-A. Azencott, L. Cayton, B. Pütz, A. Altmann, N. Karbalai, P. G. Sämann, B. Schölkopf, B. Müller-Myhsok, and K. M. Borgwardt. (2012) **GLIDE: GPU-Based Linear Regression for Detection of Epistasis.** Human Heredity, 73 (4), 220–236 doi: 10.1159/000341885

https://github.com/BorgwardtLab/GLIDE

https://github.com/chagaz/glide-scripts

Missing heritability



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

Many possible reasons:

- non-genetic / non-SNP factors
- 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.

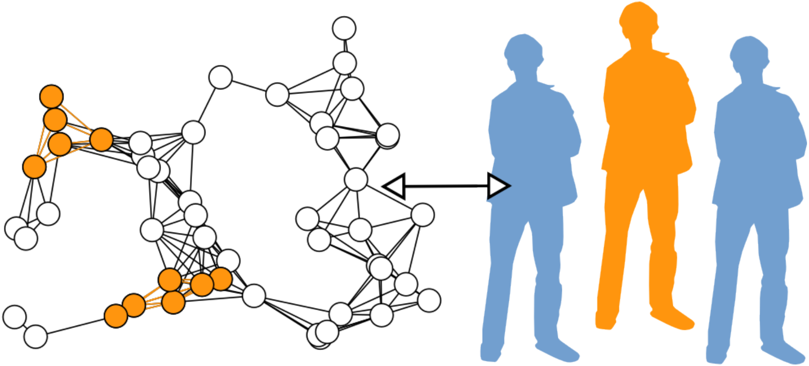
- **Consistant** with previously established knowledge
- More easily **interpretable**
- **Statistical power.**

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

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

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]

$$R(\mathcal{S}) = \sum_{i \in \mathcal{S}} c_i$$

- **Laplacian regularization**

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

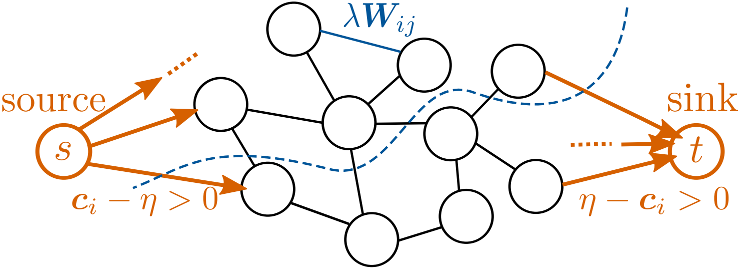
- **Regularized maximization of R**

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

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

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


$\lambda \boldsymbol{W}_{ij}$
source
s
$\boldsymbol{c}_i - \eta > 0$
sink
t
$\eta - \boldsymbol{c}_i > 0$

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

Comparison partners

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

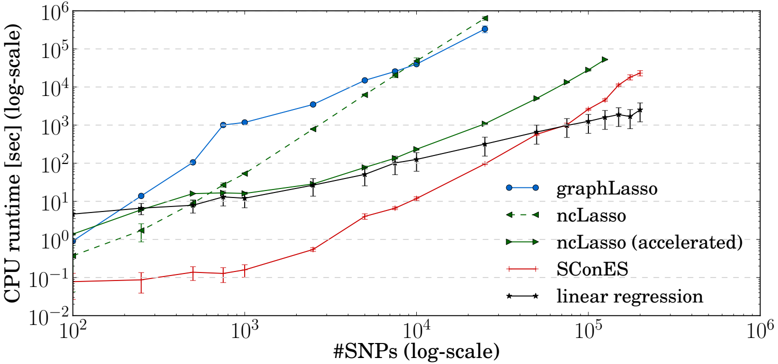
$$\underset{\beta \in \mathbb{R}^p}{\arg\min} \quad \underbrace{\frac{1}{2} ||\mathbf{y} - \mathbf{G}\beta||_2^2}_{\text{loss}} + \underbrace{\eta \; ||\beta||_1}_{\text{sparsity}}$$

- **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

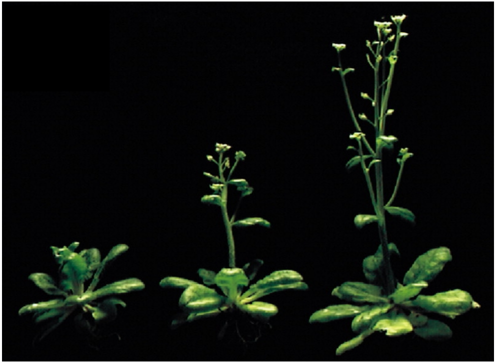
- Fairly robust to **missing edges**
- Fails if network is **random.**

Arabidopsis thaliana flowering time

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

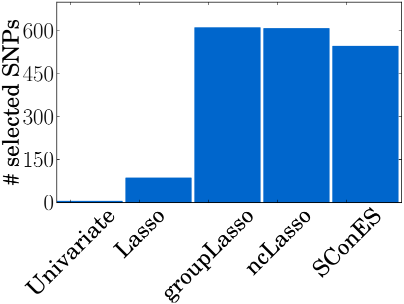
$p \sim$ 170 000 SNPs
(after MAF filtering)
$n \sim$ 150 samples

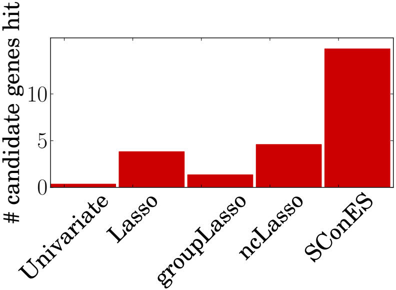
165 **candidate genes**
[Segura et al., Nat Genet 2012]



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

Arabidopsis thaliana flowering time

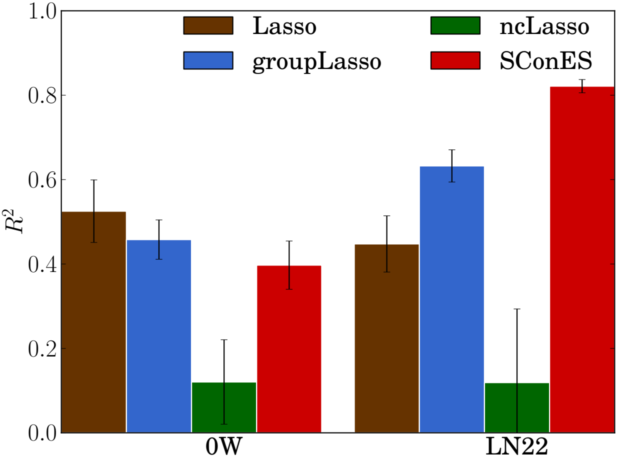




- SConES selects **about as many SNPs** as other network-guided approaches but **detects more candidates.**

Arabidopsis thaliana flowering time

Predictivity of selected SNPs



SConES: Selecting Connected Explanatory SNPs

- **s**elects **con**nected, **e**xplanatory **S**NPs;
- 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

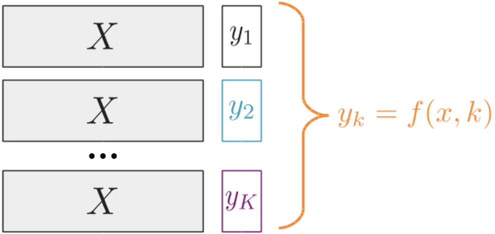
https://github.com/chagaz/scones

https://github.com/chagaz/sfan

https://github.com/dominikgrimm/easyGWASCore

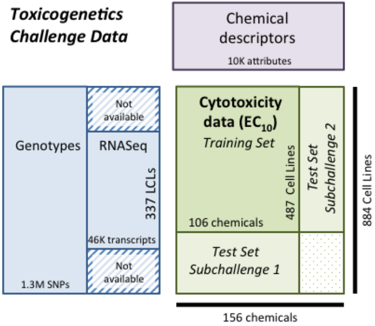
Multi-trait GWAS

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

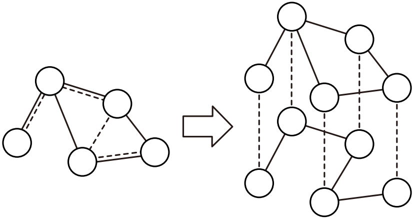
T **related phenotypes.**

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

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

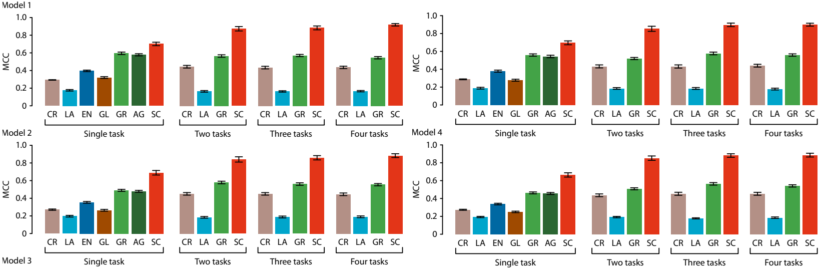
$\mathcal{S} \,\Delta\, \mathcal{S}' = (\mathcal{S} \cup \mathcal{S}') \setminus (\mathcal{S} \cap \mathcal{S}')$ (symmetric difference)

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



Multi-SConES: Multiple related tasks

Simulations: retrieving causal features



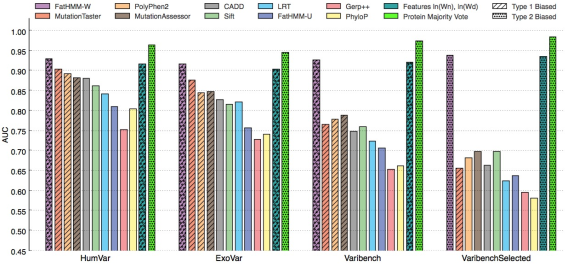
M. Sugiyama, C.-A. Azencott, D. Grimm, Y. Kawahara and K. Borgwardt (2014) **Multi-task feature selection on multiple networks via maximum flows**, SIAM ICDM, 199–207 doi:10.1137/1.9781611973440.23

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

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

SNP pathogenicity

- **SNP deleteriousness prediction tools** → prior knowledge?
- Tools are **unreliable** due to circularity issues in their evaluation:
  - Overlapping training and evaluation sets
  - Gene-level confounding



D. Grimm, C.-A. Azencott, et al. (2015) **The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity.** Human Mutation, 36 (5), 513–523 doi:10.1002/humu.22768

https://github.com/dominikgrimm/pathogenicity

Limitations of current approaches

- **Robustness/stability**

  Recovering the same SNPs when the data changes slightly.

- **Complex epistasis patterns**
  - Limited to additive or quadrative effects
  - Work on random forests + importance score [Yoshida, Stephan].

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

https://github.com/BorgwardtLab/

MLCB Tübingen: Fabian Aicheler, Karsten Borgwardt, Aasa Feragen, Udo Gieraths, Dominik Grimm, Theofanis Karaletsos, Niklas Kasenburg, Limin Li, Chrisoph Lippert, Felipe Llinares-López, Barbara Rakitsch, Damian Roqueiro, Nino Shervashidze, Carl-Johann Simon-Gabriel, Oliver Stegle, Mahito Sugiyama, Valeri Velkov.

MPI for Intelligent Systems: Lawrence Cayton, Bernhard Schölkopf.

MPI for Developmental Biology: Detlef Weigel.

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

Broad Institute: Verneri Anttila, Mark Daly, Laramie Duncan, Daniel MacArthur, Kathrin Samocha, Jordan Smoller.

Osaka University: Yoshinobu Kawahara.

University of Toronto: Recep Colak.

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