Symbiose Seminar

Feature selection in high-dimensional genomics data

Chloé-Agathe Azencott

Centre for Computational Biology (CBIO)
Mines Paris – PSL, Institut Curie & INSERM U1331
PR[AI]RIE-PSAI

March 26, 2026

http://cazencott.info chloe-agathe.azencott@minesparis.psl.eu @cazencott@lipn.info

MINES PARIS | PSL

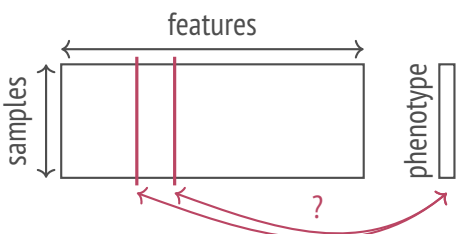
PR[AI]RIE Paris School

institut Curie

Inserm

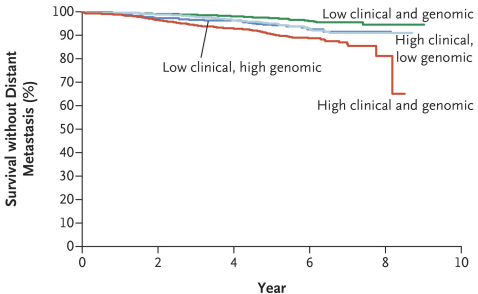
Selecting features in high-dimensional genomic data

- **Data**: For each sample,
  - **genomic features** measured along the entire genome
  - **phenotype** (observed trait of interest)
- **Goal**: identify which genomic features are **linked** with the phenotype



Selecting features in high-dimensional genomic data

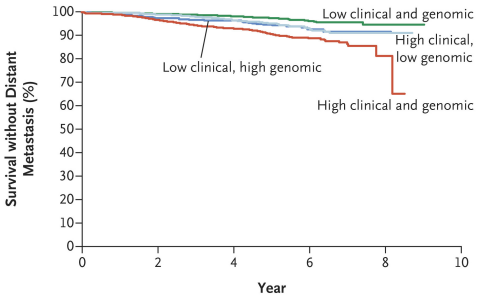
- **Motivation:** Key to providing **improved care** or **better agricultural yields**
  - **biomarker** or **signature** discovery



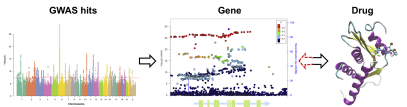
F. Cardoso et al., **70-gene signature as an aid to treatment decisions in early-stage breast cancer** NEJM 2016

Selecting features in high-dimensional genomic data

- **Motivation:** Key to providing **improved care** or **better agricultural yields**
  - **biomarker** or **signature** discovery
  - hypotheses on the underlying **biological mechanisms**



F. Cardoso et al., **70-gene signature as an aid to treatment decisions in early-stage breast cancer** NEJM 2016



Trait	Gene with GWAS hits	Known or candidate drug
Type 2 Diabetes	*SLC30A8/KCNJ11*	ZnT-8 antagonists/Glyburide
Rheumatoid Arthritis	*PADI4/IL6R*	BB-CI-amidine/Tocilizumab
Ankylosing Spondylitis(AS)	*TNFR1/PTGER4/TYK2*	TNF-inhibitors/NSAIDs/fostamatinib
Psoriasis(Ps)	*IL23A*	Risankizumab
Osteoporosis	*RANKL/ESR1*	Denosumab/Raloxifene and HRT
Schizophrenia	*DRD2*	Anti-psychotics
LDL cholesterol	*HMGCR*	Pravastatin
AS, Ps, Psoriatic Arthritis	*IL12B*	Ustekinumab

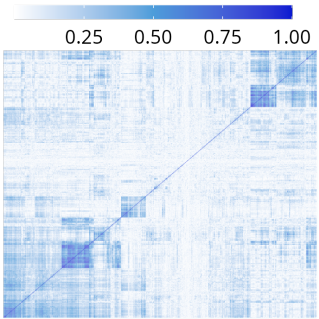
P. Visscher et al., **10 years of GWAS discovery: biology, function and translation** AJHG 2017

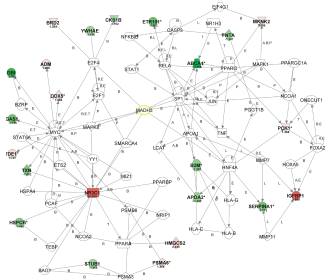
Selecting features in high-dimensional genomic data

- **Challenges:**
  - Orders of magnitude **more features than samples** ($p \gg n$)
    - 20 000 **transcripts** or 500 000 **Single Nucleotide Poymorphisms** for a few hundreds/thousands samples

Selecting features in high-dimensional genomic data

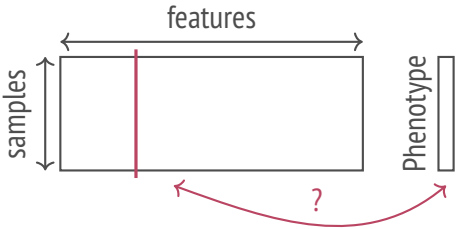
- **Challenges:**
  - Orders of magnitude **more features than samples** ($p \gg n$)
    - 20 000 **transcripts** or 500 000 **Single Nucleotide Poymorphisms** for a few hundreds/thousands samples
  - **Correlations** between features

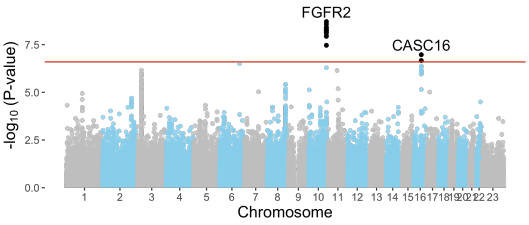


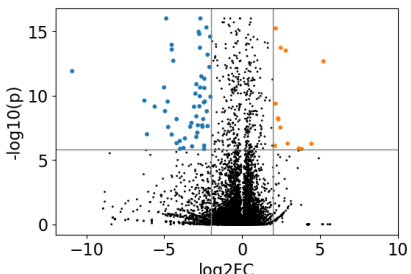


P. Lee et al., **Glucocorticoid Receptor-Dependent Gene Regulatory Networks** AJHG 2005

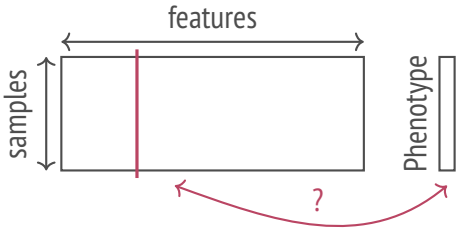
Classical statistical testing

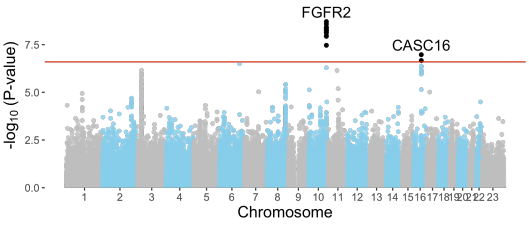




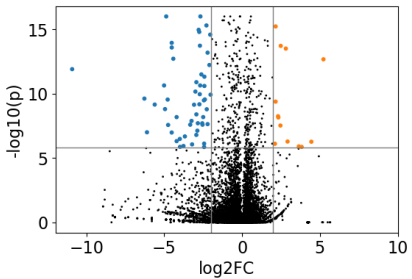


Classical statistical testing





- **Issues:**
  - Lack of **power**
  - Does not account for **other features** (joint/conditional effects)
  - Potentially high **false discovery rate** [Hej+24]



1. Structured sparsity

Regularized empirical risk minimization

- Idea: impose **a priori constraints** on the solution of the empirical risk minimization problem

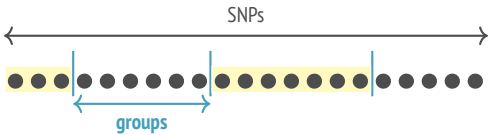
  In particular: **sparsity** that respects **pre-defined structures** (groups, graphs, trees)

- **Parametric** models: $\mathcal{F} = \{f_{\vec{w}}; \vec{w} \in \mathbb{R}^d\}$

$$\min_{\vec{w} \in \mathbb{R}^d} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f_{\vec{w}}(\vec{x}_i))}_{\text{loss}} + \lambda \underbrace{\Omega(\vec{w})}_{\text{regularizer}}$$

SMuGLasso for GWAS in diverse populations

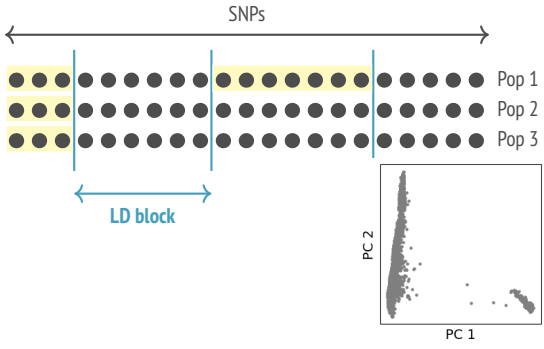
- **Group:** Group SNPs by **linkage disequilibrium blocks** [DAN15]



A. Nouira & C.-A. Azencott. **Sparse Multitask group Lasso for genome-wide association studies**. PLoS Comp Bio 2025

SMuGLasso for GWAS in diverse populations

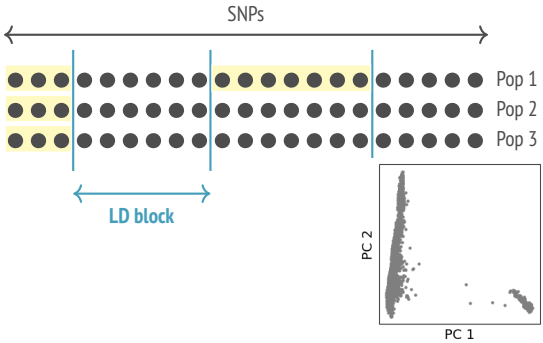
- **Group:** Group SNPs by **linkage disequilibrium blocks** [DAN15]
- Cluster samples into **genetically homogeneous populations** (PCA + clustering) $\longrightarrow$ **tasks**
- **Multitask:** blocks selected across tasks [OTJ09]
- **Sparse:** some blocks are task-specific



A. Nouira & C.-A. Azencott. **Sparse Multitask group Lasso for genome-wide association studies**. PLoS Comp Bio 2025

SMuGLasso for GWAS in diverse populations

- **Group:** Group SNPs by **linkage disequilibrium blocks** [DAN15]
- Cluster samples into **genetically homogeneous populations** (PCA + clustering) $\rightarrow$ **tasks**
- **Multitask:** blocks selected across tasks [OTJ09]
- **Sparse:** some blocks are task-specific



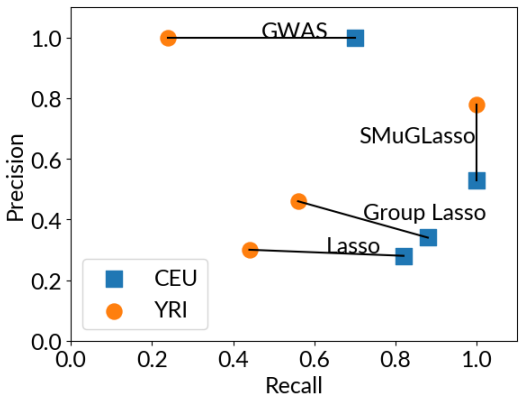
$$\min_{\vec{w}\in\mathbb{R}^p} \sum_{t=1}^{T} \underbrace{\frac{1}{n_t}\sum_{i=1}^{n_t}\mathcal{L}(y_i^{(t)}, \langle \vec{w}^{(t)}, \vec{x}_i^{(t)}\rangle)}_{\text{loss}} + \lambda \underbrace{\sum_{g=1}^{G}\sqrt{p_g}\sum_{j\in\mathcal{G}_g}\sum_{t=1}^{T}\left(w_j^{(t)}\right)^2}_{\text{mutltitask group-level sparsity}} + \lambda_2 \underbrace{\sum_{g=1}^{G}\sqrt{p_g}\sum_{j\in\mathcal{G}_g}\sum_{t=1}^{T}\left|w_j^{(t)}\right|}_{\text{task-level sparsity}}$$

A. Nouira & C.-A. Azencott. **Sparse Multitask group Lasso for genome-wide association studies**. PLoS Comp Bio 2025

SMuGLasso has better recall than other methods

Simulation with GWAsimulator [LL07]

- 2 populations from HapMap3:
  - **CEU** (1300 cases, 1700 controls)
  - **YRI** (400 cases, 600 controls)
- 50 000 SNPs
  - 200 disease-causing SNPs
  - 50 **CEU-specific** SNPs
  - 50 **YRI-specific** SNPs

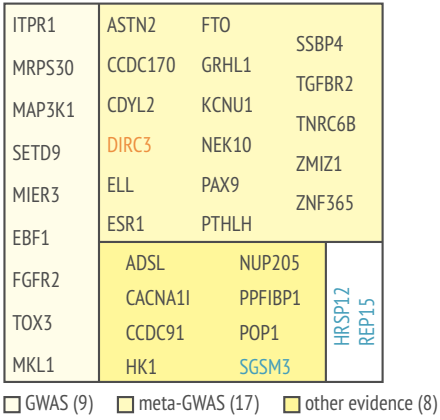


A. Nouira & C.-A. Azencott. **Sparse Multitask group Lasso for genome-wide association studies**. PLoS Comp Bio 2025

SMuGLasso identifies disease genes

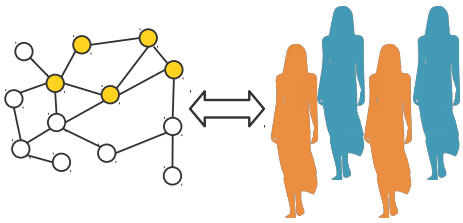
DRIVE dataset [Hun+10]

- 13 846 breast cancer cases, 14 435 controls
- 312 237 SNPs after quality control
- PCA + kmeans $\rightarrow$ 2 populations:
  - **Pop1** (USA, Australia, Denmark)
  - **Pop2** (USA, Cameroon, Nigeria, Uganda)



A. Nouira & C.-A. Azencott. **Sparse Multitask group Lasso for genome-wide association studies**. PLoS Comp Bio 2025

Graph-based regularization



Variants of the **Lasso** encourage the sparsity pattern to **respect the structure of a given graph**: features that are connected on the provided graph will tend to be selected together

- Network-constrained Lasso [LL08]
- Graph Lasso [JOV09]
- Graph-guided fused Lasso [KSX09]

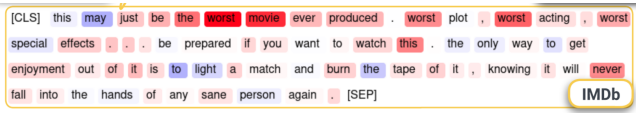
H. Climente-González et al. **A network-guided protocol to discover susceptibility genes in genome-wide association studies**. STAR Prot 2023
H. Climente-González et al. **Boosting GWAS using biological networks: A study on susceptibility to familial breast cancer**. PLoS Comp Bio 2021
C.-A. Azencott. **Network-guided biomarker discovery**. Lecture Notes in Computer Science 2016

Explainable AI

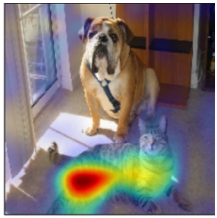
- **XAI:** methods that help understand how ML algorithms arrived at a given result [ISO20; Mol25]
- **coefficients** in a linear model; **mean decrease in impurity** in a random forest

Explainable AI

- **XAI:** methods that help understand how ML algorithms arrived at a given result [ISO20; Mol25]
- **coefficients** in a linear model; **mean decrease in impurity** in a random forest
- permutation importance, LINE, SHAP, gradient-based attribution



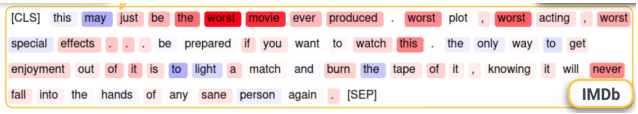
Integrated gradients



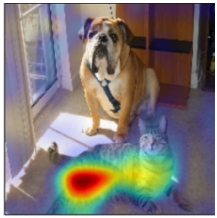
Grad-CAM

Explainable AI

- **XAI:** methods that help understand how ML algorithms arrived at a given result [ISO20; Mol25]
- **coefficients** in a linear model; **mean decrease in impurity** in a random forest
- permutation importance, LINE, SHAP, gradient-based attribution



Integrated gradients



Grad-CAM

- How **reliable** are those?

Stability

- **Stability:** Selecting the same features on subsets of the data
- **Gene signatures** are notoriously unstable [VDD11]

Stability

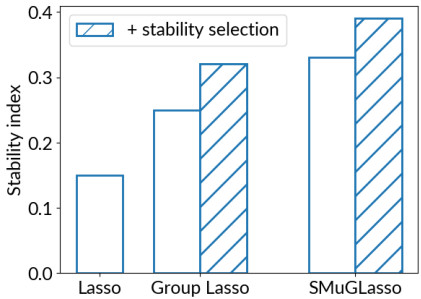
- **Stability:** Selecting the same features on subsets of the data
- **Gene signatures** are notoriously unstable [VDD11]
- **Measuring stability:** [NB16]
  - Apply the feature selection procedure on K subsamples of the data
  - Compare solutions using Pearson's correlation

Enforcing stability

- Bolasso and **stability selection** [Bac08; MB10; SS13]
  - Repeat on **multiple subsamples** of the data.
  - Only keep the features that are **selected often.**

Enforcing stability

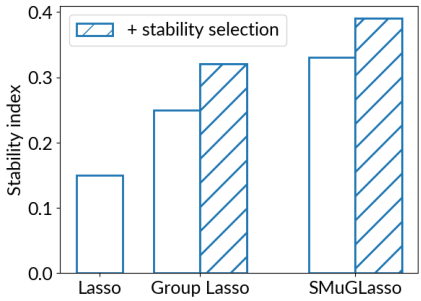
- Bolasso and **stability selection** [Bac08; MB10; SS13]
  - Repeat on **multiple subsamples** of the data.
  - Only keep the features that are **selected often.**
- **SMuGLasso** uses a stability selection procedure



A. Nouira & C.-A. Azencott. **Sparse Multitask group Lasso for genome-wide association studies**. PLoS Comp Bio 2025

Enforcing stability

- Bolasso and **stability selection** [Bac08; MB10; SS13]
  - Repeat on **multiple subsamples** of the data.
  - Only keep the features that are **selected often.**
- **SMuGLasso** uses a stability selection procedure



- What about **statistical guarantees** (e.g. confidence intervals, p-values, FDR control)?

A. Nouira & C.-A. Azencott. **Sparse Multitask group Lasso for genome-wide association studies**. PLoS Comp Bio 2025

2. Post-selection inference

Inference in linear models

- **Hypothesis testing** in a linear model $x \mapsto \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
  - Is β_j significantly different from 0?
  - **Wald test**

Inference in linear models

- **Hypothesis testing** in a linear model $x \mapsto \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
  - Is β_j significantly different from 0?
  - **Wald test**
- If the linear model was learned **under sparsity constraints:**
  - naturally biased towards non-zero β_j!
  - **Post-Selection Inference:** perform inference that **accounts for the selection event** [Lee+16]

Inference in linear models

- **Hypothesis testing** in a linear model $x \mapsto \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
  - Is β_j significantly different from 0?
  - **Wald test** relies on the hypothesis of normally-distributed residuals $Y \sim \mathcal{N}(\mu(X), \sigma^2)$
- If the linear model was learned **under sparsity constraints:**
  - naturally biased towards non-zero β_j!
  - **Post-Selection Inference:** perform inference that **accounts for the selection event** [Lee+16]

Inference in linear models

- **Hypothesis testing** in a linear model $x \mapsto \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
  - Is β_j significantly different from 0?
  - **Wald test** relies on the hypothesis of normally-distributed residuals $Y \sim \mathcal{N}(\mu(X), \sigma^2)$

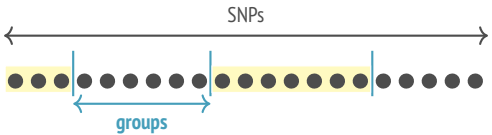
- If the linear model was learned **under sparsity constraints:**
  - naturally biased towards non-zero β_j!

    Y must be compatible with the observed set of selected features
  - **Post-Selection Inference:** perform inference that **accounts for the selection event** [Lee+16]

    i.e. characterize $Y | \widehat{S}(Y) = S$

Group-based feature selection with kernels

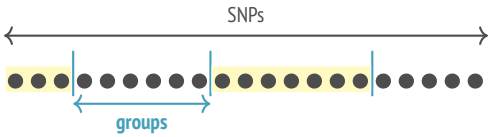
- Is group g associated with the phenotype?
- E.g. **SKAT** [Wu+11] test statistic $T = \vec{y}^\top K_g \vec{y}$



- **kernel** K_g quantifies (possibly nonlinear) similarities between samples based on SNPs in g

Group-based feature selection with kernels

- Is group g associated with the phenotype?
- E.g. **SKAT** [Wu+11] test statistic $T = \vec{y}^\top K_g \vec{y}$

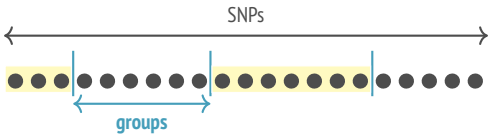


- **kernel** K_g quantifies (possibly nonlinear) similarities between samples based on SNPs in g
- **Quadratic Kernel Association Test** $\vec{y}^\top q(K_g)\vec{y}$
  - E.g. **HSIC** estimators [Gre+05]

L. Slim et al. **kernelPSI: a post-selection inference framework for nonlinear variable selection**, ICML 2019

Group-based feature selection with kernels

- Is group g associated with the phenotype?
- E.g. **SKAT** [Wu+11] test statistic $T = \vec{y}^\top K_g \vec{y}$

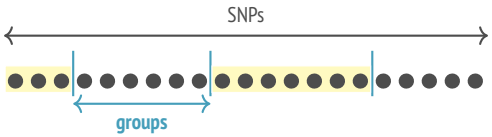


- **kernel** K_g quantifies (possibly nonlinear) similarities between samples based on SNPs in g
- **Quadratic Kernel Association Test** $\vec{y}^\top q(K_g)\vec{y}$
  - E.g. **HSIC** estimators [Gre+05]
- **Group selection procedure:** select the kernels with the highest scores

L. Slim et al. **kernelPSI: a post-selection inference framework for nonlinear variable selection**, ICML 2019

Group-based feature selection with kernels

- Is group g associated with the phenotype?
- E.g. **SKAT** [Wu+11] test statistic $T = \vec{y}^\top K_g \vec{y}$



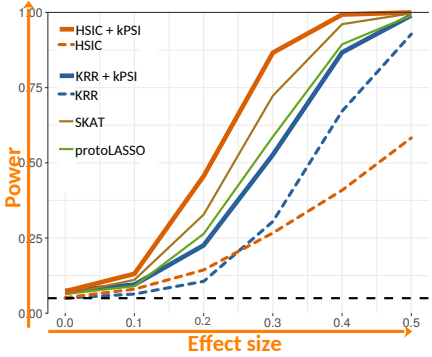
- **kernel** K_g quantifies (possibly nonlinear) similarities between samples based on SNPs in g
- **Quadratic Kernel Association Test** $\vec{y}^\top q(K_g)\vec{y}$
  - E.g. **HSIC** estimators [Gre+05]
- **Group selection procedure:** select the kernels with the highest scores

  Post-selection inference needed

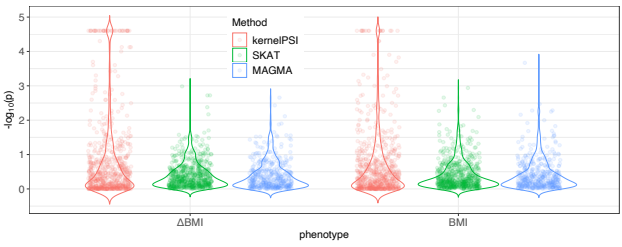
L. Slim et al. **kernelPSI: a post-selection inference framework for nonlinear variable selection**, ICML 2019

kernelPSI

– **kernelPSI:** characterize the selection event and sample efficiently from the resulting distribution



Simulation

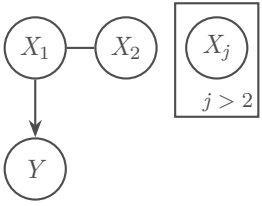


On UKBiobank data

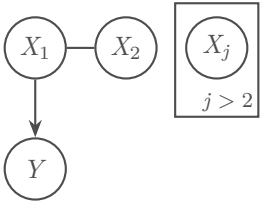
L. Slim et al. **kernelPSI: a post-selection inference framework for nonlinear variable selection**, ICML 2019

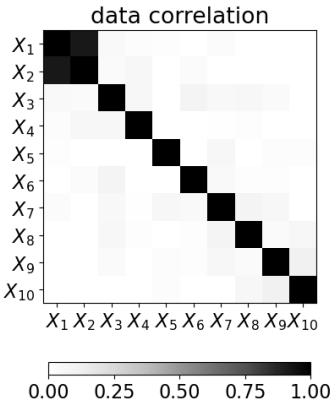
3. False Discovery Rate control with Knockoffs

Challenge: correlation between variables

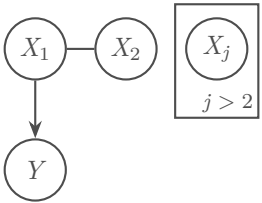


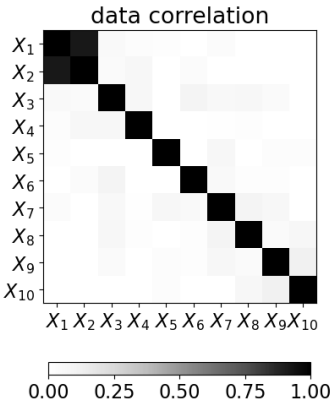
Challenge: correlation between variables

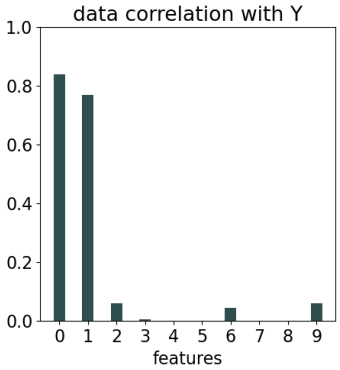




Challenge: correlation between variables

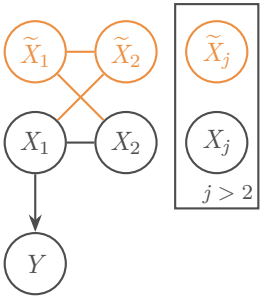


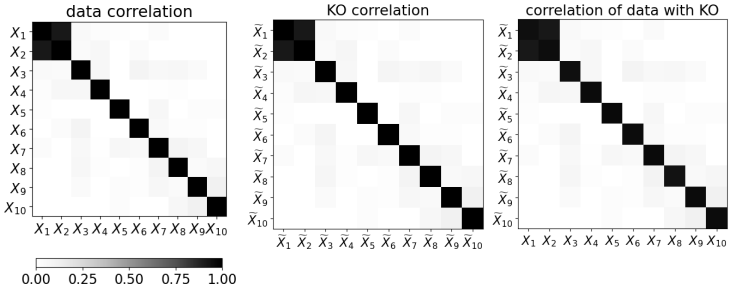




Features that are **correlated** with features associated with the phenotype **appear associated** with the phenotype

Knockoffs

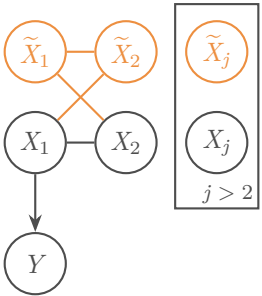


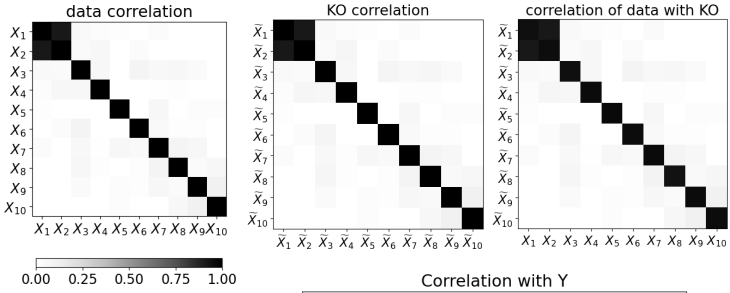


- $\widetilde{X}$ built from X only: $\widetilde{X} \perp Y | X$
- $\widetilde{X}$ has the **same covariance structure** as X
- $\widetilde{X}$ as **different** from X as possible

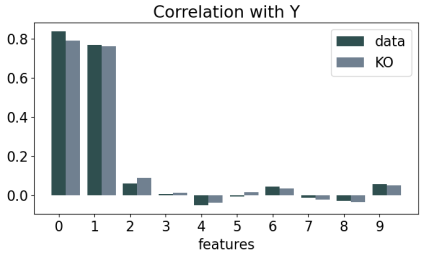
R. Barber & E. Candès **Controlling the false discovery rate via knockoffs**
Ann. Stat. 2015

Knockoffs



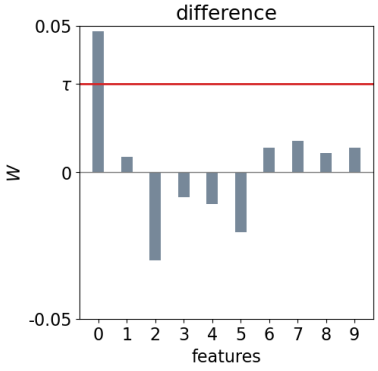


- $\widetilde{X}$ built from X only: $\widetilde{X} \perp Y | X$
- $\widetilde{X}$ has the **same covariance structure** as X
- $\widetilde{X}$ as **different** from X as possible



R. Barber & E. Candès **Controlling the false discovery rate via knockoffs**
Ann. Stat. 2015

Knockoff Statistics

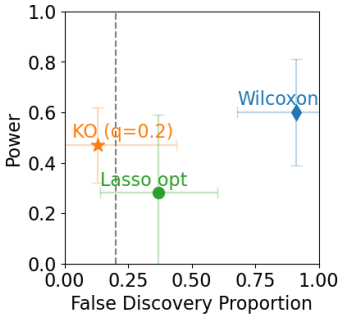


- $W_j = Q(X_j, Y) - Q(\widetilde{X}_j, Y)$
- Example: **Lasso Coefficient-Difference** [Can+18]
- $\tau = \min\left\{t > 0 \colon \frac{|\{j \colon W_j \leq -t\}|}{|\{j \colon W_j \geq t\}|} \leq q\right\}$ guarantees **control of the target FDR** q

R. Barber & E. Candès et al. **Controlling the false discovery rate via knockoffs** Ann. Stat. 2015
E. Candès et al. **Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection** J. R. Stat. Soc. B 2018

Applying knockoffs to transcriptomics data

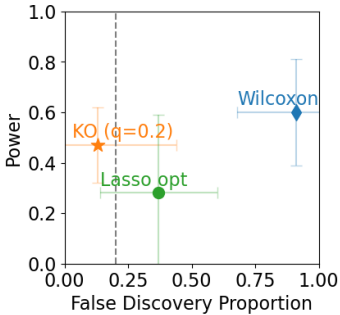
- **Simulation framework:** real transcriptomics data, simulated case/control outcomes
  - ☺ KO **control FDR well**, and better than Wilcoxon rank-summed test or the Lasso



J. Cartier et al. **Statistical knockoffs improve biomarker discovery from transcriptomic data** BioRxiv 2025

Applying knockoffs to transcriptomics data

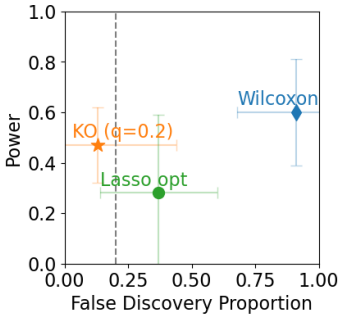
- **Simulation framework:** real transcriptomics data, simulated case/control outcomes
  - ☺ KO **control FDR well**, and better than Wilcoxon rank-summed test or the Lasso
  - ☹ The KO framework fails if the relationship between phenotype and features is **non-linear**



J. Cartier et al. **Statistical knockoffs improve biomarker discovery from transcriptomic data** BioRxiv 2025

Applying knockoffs to transcriptomics data

- **Simulation framework:** real transcriptomics data, simulated case/control outcomes
  - ☺ KO **control FDR well**, and better than Wilcoxon rank-summed test or the Lasso
  - ☹ The KO framework fails if the relationship between phenotype and features is **non-linear**



- On **real** outcomes:
  - ☹ The KO framework is **overly conservative**

cohort	CRUKPAP	AEGIS	BC HER2	BC ER	BC PgR
discoveries (q=0.5)	3	7	13	1	0

J. Cartier et al. **Statistical knockoffs improve biomarker discovery from transcriptomic data** BioRxiv 2025

Take home messages

- **Feature selection** in high-dimensional, correlated data is hard
- **Structured sparsity** allows us to leverage biological knowledge and prior hypotheses

  but **XAI** comes with no guarantees and **inference** remains difficult
- New(ish) tools:
  - **Post-selection inference** for p-values / confidence intervals on model coefficients
  - **Statistical knockoffs** to control for the rate of false discoveries
  - (not discussed) **Conditional feature importance** [Str+08; RLNT26]

Thanks

CBIO: Youmna Ayadi, **Julie Cartier**, Adeline Fermanian (now Califrais), Florian Massip, **Asma Nouira** (now ASNR), **Lotfi Slim** (now NVIDIA), Jean-Philippe Vert (now Owkins/Bioptimus).

Sanofi: Clément Chatelain.

Funding: Agence Nationale pour la Recherche (SCAPHE, PRAIRIE, Label Carnot), Association Nationale Recherche Technologie, Sanofi R&D.

References I

[Aze+13] Chloé-Agathe Azencott et al. "Efficient network-guided multi-locus association mapping with graph cuts". In: *Bioinformatics* 29.13 (2013), pp. i171i179. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt238.

[Aze16] Chloé-Agathe Azencott. "Network-guided biomarker discovery". In: *Machine Learning for Health Informatics*. Springer International Publishing, 2016, pp. 319336. DOI: 10.1007/978-3-319-50478-0_16.

[Bac08] Francis R Bach. "Bolasso: model consistent lasso estimation through the bootstrap". In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 33–40.

[Ber+87] H. C. P. Berbee et al. "Hit-and-run algorithms for the identification of nonredundant linear inequalities". In: *Mathematical Programming* 37.2 (1987), pp. 184207. ISSN: 1436-4646. DOI: 10.1007/bf02591694. URL: http://dx.doi.org/10.1007/BF02591694.

[Bla+23] Alexandre Blain et al. "False Discovery Proportion control for aggregated Knockoffs". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 78193–78204. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/f6712d5191d2501dfc7024389f7bfcdd-Paper-Conference.pdf.

[Bla+25] Alexandre Blain et al. *When Knockoffs fail: diagnosing and fixing non-exchangeability of Knockoffs*. 2025. arXiv: 2407.06892 [stat.ME]. URL: https://arxiv.org/abs/2407.06892.

References II

[Can+18] Emmanuel Candès et al. "Panning for Gold: Model-X Knockoffs for High Dimensional Controlled Variable Selection". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80.3 (2018), pp. 551577. ISSN: 1467-9868. DOI: 10.1111/rssb.12265.

[Car+25] Julie Cartier et al. "Statistical knockoffs improve biomarker discovery from transcriptomic data". In: (2025). DOI: 10.1101/2025.07.04.663147.

[CG+21] Héctor Climente-González et al. "Boosting GWAS using biological networks: A study on susceptibility to familial breast cancer". In: *PLOS Computational Biology* 17.3 (2021), e1008819. DOI: 10.1371/journal.pcbi.1008819.

[CGAY23] Héctor Climente-González, Chloé-Agathe Azencott, and Makoto Yamada. "A network-guided protocol to discover susceptibility genes in genome-wide association studies using stability selection". In: *STAR Protocols* 4.1 (2023). DOI: 10.1016/j.xpro.2022.101998.

[DAN15] Alia Dehman, Christophe Ambroise, and Pierre Neuvial. "Performance of a blockwise approach in variable selection using linkage disequilibrium information". In: *BMC Bioinformatics* 16.1 (2015). DOI: 10.1186/s12859-015-0556-6.

[Gre+05] Arthur Gretton et al. "Measuring Statistical Dependence with Hilbert-Schmidt Norms". In: *Algorithmic Learning Theory*. Springer Berlin Heidelberg, 2005, pp. 6377. ISBN: 9783540316961. DOI: 10.1007/11564089_7. URL: http://dx.doi.org/10.1007/11564089_7.

References III

[Hej+24] Boris P. Hejblum et al. "Neglecting the impact of normalization in semi-synthetic RNA-seq data simulations generates artificial false positives". In: *Genome Biology* 25.1 (2024). DOI: 10.1186/s13059-024-03231-9.

[Hun+10] David J. Hunter et al. *Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) Oncoarray Genotypes.* 2010. URL: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001265.v1.p1.

[ISO20] ISO/IEC JTC 1/SC 7. *TR 29119-11:2020 Part 11: Guidelines on the testing of AI-based systems.* International Organization for Standardization and International Electrotechnical Commission. 2020.

[JOV09] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. "Group lasso with overlap and graph lasso". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML 09. ACM, 2009, pp. 433440. DOI: 10.1145/1553374.1553431.

[KSX09] Seyoung Kim, Kyung-Ah Sohn, and Eric P. Xing. "A multivariate regression approach to association analysis of a quantitative trait network". In: *Bioinformatics* 25.12 (2009), pp. i204i212. DOI: 10.1093/bioinformatics/btp218.

[Lee+16] Jason D. Lee et al. "Exact post-selection inference, with application to the lasso". In: *The Annals of Statistics* 44.3 (2016). ISSN: 0090-5364. DOI: 10.1214/15-aos1371. URL: http://dx.doi.org/10.1214/15-AOS1371.

[LL07] Chun Li and Mingyao Li. "GWAsimulator: a rapid whole-genome simulation program". In: *Bioinformatics* 24.1 (2007), pp. 140142. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm549.

References IV

[LL08] Caiyan Li and Hongzhe Li. "Network-constrained regularization and variable selection for analysis of genomic data". In: *Bioinformatics* 24.9 (2008), pp. 11751182. DOI: 10.1093/bioinformatics/btn081.

[LT15] Joshua R. Loftus and Jonathan E. Taylor. *Selective inference in regression models with groups of variables.* 2015. arXiv: 1511.01478 [stat.ME]. URL: https://arxiv.org/abs/1511.01478.

[MB10] Nicolai Meinshausen and Peter Bühlmann. "Stability selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4 (2010). ISSN: 1467-9868.

[Mol25] Christoph Molnar. *Interpretable machine learning: A guide for making black box models explainable.* 2025.

[NA] Asma Nouira and Chloé-Agathe Azencott. "Multitask group Lasso for Genome Wide association Studies in diverse populations". In: *Pacific Symposium on Biocomputing 2022*, pp. 163–174. DOI: 10.1142/9789811250477_0016.

[NA25] Asma Nouira and Chloé-Agathe Azencott. "Sparse multitask group Lasso for genome-wide association studies". In: *PLOS Computational Biology* 21.9 (2025), e1012734. DOI: 10.1371/journal.pcbi.1012734.

[NB16] Sarah Nogueira and Gavin Brown. "Measuring the stability of feature selection". In: *Machine Learning and Knowledge Discovery in Databases*. Lecture Notes in Computer Science 9852. Springer International Publishing, 2016, pp. 442–457.

References V

[OTJ09] Guillaume Obozinski, Ben Taskar, and Michael I. Jordan. "Joint covariate selection and joint subspace selection for multiple classification problems". In: *Statistics and Computing* 20.2 (2009), pp. 231252. DOI: 10.1007/s11222-008-9111-x.

[RLNT26] Angel Reyero-Lobo, Pierre Neuvial, and Bertrand Thirion. *Conditional Feature Importance revisited: Double Robustness, Efficiency and Inference.* 2026. arXiv: 2501.17520 [stat.ME]. URL: https://arxiv.org/abs/2501.17520.

[RTT17] Stephen Reid, Jonathan Taylor, and Robert Tibshirani. "A General Framework for Estimation and Inference From Clusters of Features". In: *Journal of the American Statistical Association* 113.521 (2017), pp. 280293. ISSN: 1537-274X. DOI: 10.1080/01621459.2016.1246368. URL: http://dx.doi.org/10.1080/01621459.2016.1246368.

[SS13] Rajen D. Shah and Richard J. Samworth. "Variable selection with error control: another look at stability selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75.1 (2013), pp. 55–80.

[Str+08] Carolin Strobl et al. "Conditional variable importance for random forests". In: *BMC Bioinformatics* 9.1 (2008), p. 307. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-307.

[Sug+14] Mahito Sugiyama et al. "Multi-task feature selection on multiple networks via maximum flows". In: *Proceedings of the 2014 SIAM International Conference on Data Mining*. 2014, pp. 199207. DOI: 10.1137/1.9781611973440.23.

[Tib96] Robert Tibshirani. "Regression Shrinkage and Selection Via the Lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267288. ISSN: 1467-9868. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

References VI

[Toz+22] Veronica Tozzo et al. "Where do we stand in regularization for life science studies?" In: *Journal of Computational Biology* 29.3 (2022), pp. 213232. DOI: 10.1089/cmb.2019.0371.

[VDD11] David Venet, Jacques E. Dumont, and Vincent Detours. "Most random gene expression signatures are significantly associated with breast cancer outcome". In: *PLoS Computational Biology* 7.10 (2011), e1002240.

[Wu+11] MichaelăC. Wu et al. "Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test". In: *The American Journal of Human Genetics* 89.1 (2011), pp. 8293. ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2011.05.029. URL: http://dx.doi.org/10.1016/j.ajhg.2011.05.029.

[YL05] Ming Yuan and Yi Lin. "Model Selection and Estimation in Regression with Grouped Variables". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68.1 (2005), pp. 4967. DOI: 10.1111/j.1467-9868.2005.00532.x.