Selection-adjusted estimation of effect sizes

with an application in eQTL studies

Snigdha Panigrahi

19 October, 2017

Stanford University

Selective inference - introduction

Selective inference

- Statistical inference - target of inference, generative model

  Confirmatory paradigm: target and model known before looking at data

  Exploratory paradigm : target and model chosen after *data snooping*

- Selective inference bridges the gap between the two paradigms

  Selection - look for interesting inferential questions

  decide/change existing beliefs on a generative model

  Inference - retain some guarantees post data exploration

Selective guarantees

- Simultaneous coverage - Berk et al. 2013
- Conditional coverage - Lee et al. 2013; Fithian et al. 2014; Loftus and Taylor 2015
- Asymptotic coverage and consistency - Tian and Taylor 2015
- FDR control - Barber and Candes 2014; Candes et al. 2016; Li and Barber 2016; Lei and Fithian 2016

Reproducibility in GWAS

Nature editorial: *Over the past year, Nature has published a string of articles that highlight failures in the reliability and reproducibility of published research.*

- Analyst starts with an exhaustive pool of explanatory variables that affect an outcome/ disease/ phenotype
- She reports the strongest associations suggested by the data
- FDR is a well-accepted global error criterion over reported discoveries

  Prior works -Peterson et al. 2016; Stell and Sabatti 2016; Brzyski et al. 2017

How to measure the strengths of interesting effects post a genome-wide search over the space of predictors/models?

Conditional framework of inference

Inferential guarantees

- Selective Type-I error
- Consistency of estimates conditional on selection

Data-splitting (e.g. Cox 1975) is common wisdom

- Conditional regime is a generalization of data-splitting
- More powerful than data-splitting: uses left-over information from selection
- Small sample settings: data splitting suffers from low screening and inferential powers

Talk today

- Selection and inference in one-go : describe methods that allow geneticist to identify interesting effects and infer about them using a single data-set
- Higher inferential power : than an ad-hoc hold-out method
- New selective inference tools : optimization-based methods that approximate an intractable reference measure in selection-adjusted law
- Estimates : provide a useful point estimate that is consistent and selection-adjusted intervals that empirically have target coverage

GTEx - a multi-phenotype study

A cis-eQTL study

eQTL (Expression quantitative trait locus) studies: identify regions of the genome that influence whether and how much a gene is expressed

Cis-eQTL mapping - associations between

- gene expression - the response
- local genetic variations (SNPs in this data) - predictors

  restricted within 1MB of the transcription start site of each gene

Overview of GTEx data set

https://www.gtexportal.org/home/

44 Human Tissues, **449** Donors

GTEx data - Gene expression data for genes in various tissues
Genotypic information for donors.

Tissues - **Liver** (**97** samples), Brain (**72** samples), Stomach (**170** samples)

Multi-phenotype analysis

Gene expression data (RNA-seq) for G genes that correspond to Liver

Each gene g has its own predictor matrix of local variants.

$$(y^{(1)}, X^{(1)}),\ (y^{(2)}, X^{(2)}),\ \cdots,\ (y^{(G)}, X^{(G)})$$

$$y^{(g)} \in \mathbb{R}^{97\times 1}, X^{(g)} \in \mathbb{R}^{97\times V_g}$$

$G \approx 22K,\ V_g$ varies between $3000 - 25000$

Selection- eGenes and eVariants

A hierarchical selection

Selection of regulatory variants is modeled as a 2-stage selection

- **eGenes** - select genes with $>= 1$ signal from G genes

  Consortium et al. 2015; Ongen et al. 2015; Carithers et al. 2015 identify eQTLs (eGenes)

- **eVariants**: From V_g variants, select regulatory variants

  Aguet et al. 2016 conduct secondary analysis for eGenes to further identify multiple regulatory variants

Randomized screenings

For data (y, X), a convex constrained (randomized) query:

$$\text{minimize}_\beta \; \ell(y, X; \beta) + \mathcal{P}_\lambda(\beta)$$

Randomized screenings

For data (y, X), a convex constrained (randomized) query:

$$\text{minimize}_\beta \; \ell(y, X; \beta) + \mathcal{P}_\lambda(\beta) - \omega^T \beta, \;\; \omega \sim \mathcal{N}(0, \tau^2 I)$$

- Randomization : preserves more left-over information for inference
- Randomized queries introduced in Tian et al. 2016
- Computational advantage with randomization : simplify support of selective density through a change of measure

Marginal screening of eGenes

eGene screening: for every gene

- Compute p-values $\{p_1, p_2, \cdots, p_{V_g}\}$ based on marginal T-stats

$$T_j^{(g)} = \sqrt{n-2} \cdot \rho\left(X_j^{(g)}, y^{(g)}\right) \Big/ \sqrt{1-\rho^2\left(X_j^{(g)}, y^{(g)}\right)}$$

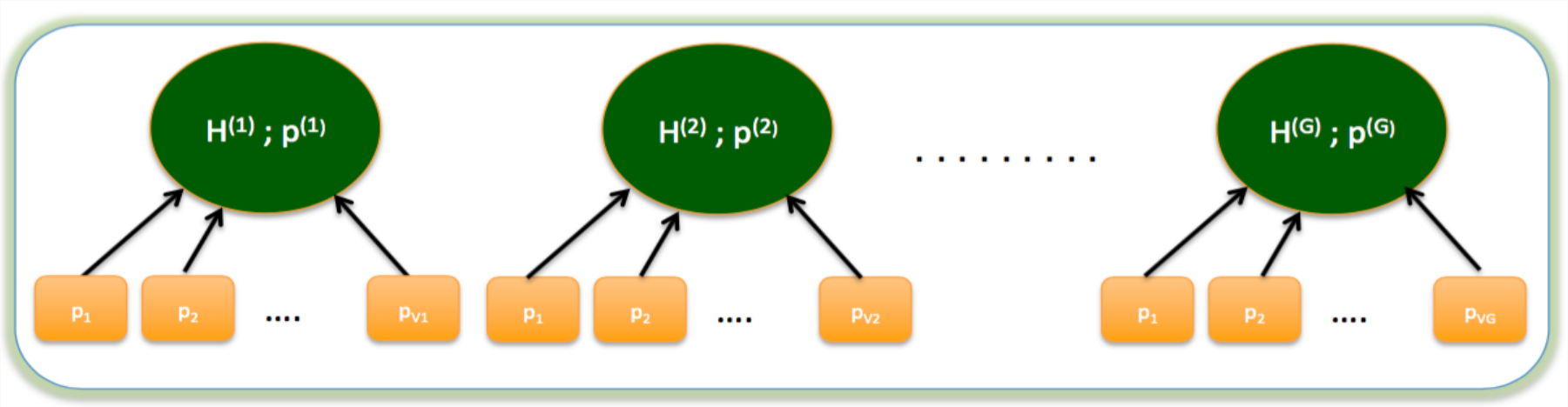
Marginal screening of eGenes

eGene screening: for every gene

- Compute p-values $\{p_1, p_2, \cdots, p_{V_g}\}$ based on marginal T-stats

$$T_j^{(g)} = \sqrt{n-2} \cdot \frac{\rho\left(X_j^{(g)}, y^{(g)}\right)}{\sqrt{1-\rho^2\left(X_j^{(g)}, y^{(g)}\right)}} + \omega_j,\ \omega_j \sim \mathcal{N}(0, \gamma^2)$$

- **Bonferroni**-adjusted $V_g \min_i p_i$ or **Simes**-adjusted - $\min_i \frac{V_g}{i} p_{(i)}$
- Pass all G Simes p-values through **BH**-q sieve



eVariant screening

eVariant screening: for each eGene, identify promising variants

- A multivariate regression to explain polygenic nature of the phenotype
- Aguet et al. 2016: A forward-backward with an adaptive threshold
- Employ a (randomized) LASSO

$$\text{minimize}_\beta \ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

eVariant screening

eVariant screening: for each eGene, identify promising variants

- A multivariate regression to explain polygenic nature of the phenotype
- Aguet et al. 2016: A forward-backward with an adaptive threshold
- Employ a (randomized) LASSO

$$\text{minimize}_\beta \ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 - \zeta^T\beta, \ \zeta \sim \mathcal{N}(0, \tau^2 I_p)$$

Adaptive target & GWAS model

Target post selection

Post selection of eGene g and eVariants $\boldsymbol{E}$

$$\text{Adaptive target} \ : b_{\boldsymbol{E}} = \arg\min_{b} \mathbb{E}\| Y^{(g)} - X_{\boldsymbol{E}}^{(g)} b\|_2^2.$$

Unadjusted inference for $e_j^T b_{\boldsymbol{E}}$ is based on

- Point estimate - $Z = e_j^T (X_{\boldsymbol{E}}^{(g)} X_{\boldsymbol{E}}^{(g)})^{-1} X_{\boldsymbol{E}}^{(g)^T} Y^{(g)}$
- Intervals - $[Z - z_{1-\alpha/2} \cdot \sigma (X_{\boldsymbol{E}}^T X_{\boldsymbol{E}})^{-1}_{j,\,j}, Z + z_{1-\alpha/2} \cdot \sigma (X_{\boldsymbol{E}}^T X_{\boldsymbol{E}})^{-1}_{j,\,j}]$

No guarantees on unadjusted inference!

GWAS generative model

- Saturated model framework - Lee et al. 2013

$$Y^{(g)} = \mu + \epsilon^{(g)},\ \epsilon^{(g)} \overset{\text{IND}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

- Selected model framework - Fithian et al. 2014

$$Y^{(g)} = X_{\boldsymbol{E}}^{(g)} \beta_{\boldsymbol{E}} + \epsilon^{(g)},\ \epsilon^{(g)} \overset{\text{IND}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

- Other selection guided model - $\bar{\boldsymbol{E}}$ function of $\boldsymbol{E}$

$$Y^{(g)} = X_{\bar{\boldsymbol{E}}}^{(g)} \beta_{\bar{\boldsymbol{E}}} + \epsilon^{(g)},\ \epsilon^{(g)} \overset{\text{IND}}{\sim} \mathcal{N}(0, \sigma^2 I)$$

Like in classical regime, there is always the possibility of mis-specification!

Conditional density

Selection modifies a generative model - conditioning on the selection event.

Conditional density relies on an explicit description of selection event

Selection applied is a function of outcome and randomization $\{(y^{(g)}, \omega^{(g)})\}$

$$\{\hat{E}^{(g)} = E^{(g)} \text{ for } g \in G\}, \quad G = \text{set of egenes}$$

Recall : Target statistic $Z^{(g)} = e_j^T (X_E^{(g)} X_E^{(g)})^{-1} X_E^{(g)\,T} Y^{(g)}$

Recall : Target parameter $b = e_j^T b_E$

Joint density $(Z^{(g)}, \Omega^{(g)})$ at $\{(z^{(g)}, \omega^{(g)})\}$; $\Omega^{(g)} \sim G$ with density g

$$\prod_g \exp(-(z^{(g)} - b)^2 / 2\sigma_j^2) \times g(\omega^{(g)}) \times 1_{\{\hat{E}^{(g)} = E^{(g)} \text{ for } g \in G\}}$$

Separable and convex selection region

Selection probability decouples into eGenes with a convex region $\mathcal{K}_g$

$$P(\hat{E}^{(g)} = E^{(g)} \text{ for } g \in G) = \Pi_{g \in G} P((y^{(g)}, \omega^{(g)}) \in \mathcal{K}_g) \times P_0.$$

- independent GWAS structure across genes
- some additional conditioning beyond selected eGenes and eVariants

Joint density $(Z^{(g)}, \Omega^{(g)})$ at $\{(z^{(g)}, \omega^{(g)})\}$; $\Omega^{(g)} \sim G$ with density g

$$\prod_g \exp(-(z^{(g)} - b)^2/2\sigma_j^2) \times g(\omega^{(g)}) \times 1_{\{(z^{(g)}, \omega^{(g)}) \in \mathcal{K}_g\}}$$

Decoupled selection-adjusted law across all eGenes - focus on 1 eGene.

Surrogate - (soft) truncated likelihood

Randomized selective pivot

- Joint density (Z, ω) at (z, w)

$$\exp(-(z-b)^2/2\sigma^2) \times g(w) \times 1_{(z,w)\in\mathcal{K}}$$

- Selective density of $Z|\ \hat{E}(Z,\omega) = E$

$$\exp(-(z-b)^2/2\sigma^2) \times P(\omega \in \mathcal{K}_0 | Z = z)$$

- Selective pivot - $\dfrac{\int_{Z^{\text{obs}}}^{\infty} \exp(-(z-b)^2/2\sigma^2) P(\omega \in \mathcal{K}_0 | Z = z) dz}{\int_{-\infty}^{\infty} \exp(-(z-b)^2/2\sigma^2) P(\omega \in \mathcal{K}_0 | Z = z) dz}$

Approximate pivot

$$\frac{\int_{Z^{\text{obs}}}^{\infty} \exp(-(z-b)^2/2\sigma^2) P(\omega \in \mathcal{K}_0 | Z = z) dz}{\int_{-\infty}^{\infty} \exp(-(z-b)^2/2\sigma^2) P(\omega \in \mathcal{K}_0 | Z = z) dz}$$

Approximate pivot

$$\frac{\sum_{z \in G : z \geq Z^{\text{obs}}} \exp(-(z-b)^2/2\sigma^2)\hat{P}(\omega \in \mathcal{K}_0 | Z = z)}{\sum_{z \in G} \exp(-(z-b)^2/2\sigma^2)\hat{P}(\omega \in \mathcal{K}_0 | Z = z)}$$

Theorem (Panigrahi et al. 2017)

With inversion map from K.K.T. conditions as

$$\omega = A_E Z + B_E O + c_E,$$

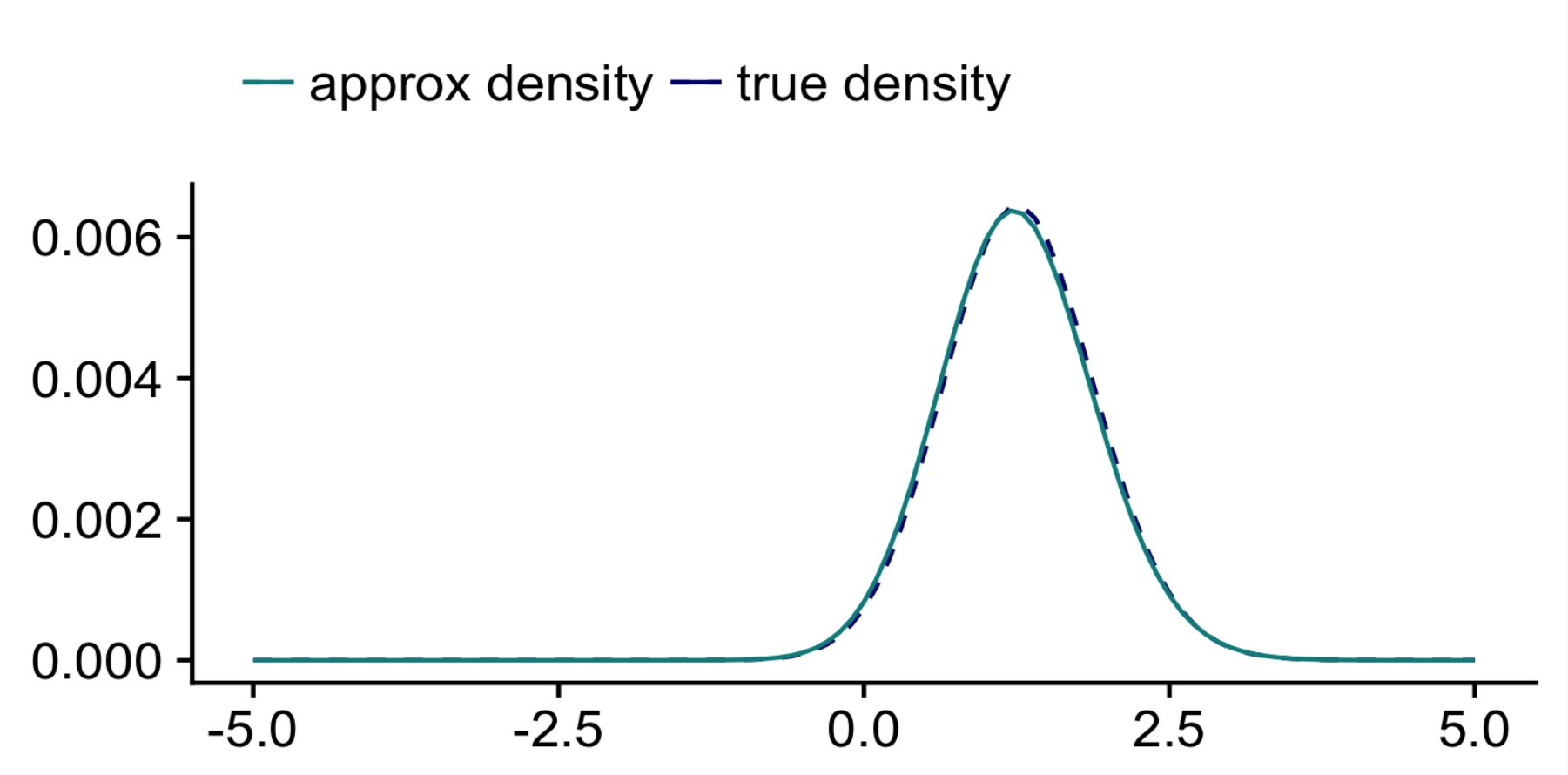
$$\log \hat{P}(\omega \in \mathcal{K}_0 | Z = z) = -\inf_{\mathbb{R}^{|E|} \ni o > 0} \{\Lambda_G^*(A_E z + B_E o + c_E) - \log H(z)\}$$

for a convex, compact $\mathcal{K}_0$

$$\frac{1}{n}\left\{\log \hat{P}(\omega \in \mathcal{K}_0 | Z = z) - \log P(\omega \in \mathcal{K}_0 | Z = z)\right\} \to 0 \text{ as } n \to \infty.$$

Selection-adjusted density

$$Z \sim \mathcal{N}(-1, 1),\ \omega \sim \mathcal{N}(0, 0.5);\ \text{Selection: } Z + w > 2$$



Selective MLE

- Log-partition function

$$b^2/2\sigma^2 + \log \int_{-\infty}^{\infty} \exp(-(z-b)^2/2\sigma^2) P(\omega \in \mathcal{K}_0 | Z = z) dz$$

Selective MLE

- Log-partition function

$$b^2/2\sigma^2 + \log \sum_{z \in G} \exp(-(z-b)^2/2\sigma^2) \hat{P}(\omega \in \mathcal{K}_0 | Z = z)$$

- The approximate selective MLE objective is convex

Theorem (Panigrahi et al. 2016)

For a convex selection region and Gaussian data, randomization, the selective MLE is consistent for non-local sequences of parameters.

Simple thresholding example: $w(x) = \dfrac{x^2}{2(1+\tau^2)} + b(x)$

$$\hat{b} = (1+\tau^2)\nabla(w(1+\tau^2)z - \tau^2 \hat{b})$$

Adjusted inference in eQTL

A GWAS generative model

$$Y^{(g)} = \sum_{k \in \mathcal{S}^{(g)}} X_k^{(g)} \beta_k + \epsilon^{(g)}, \ \epsilon^{(g)} \sim \mathcal{N}(0, \sigma^2 I).$$

$$|S^{(g)}| \in \{0, 1, 2, \cdots, 9\}, \ S^{(g)} \subset \{1, 2, \cdots, V_g\}.$$

Signal amplitude $|\beta_k| = 2.5 \approx \sqrt{2 \log p}$

Probability distribution of signals is as below:

0	1	2	3	4	5	6	7	8	9
60%	8%	7%	6%	5.5%	4.5%	3.5%	2.5%	2%	1%

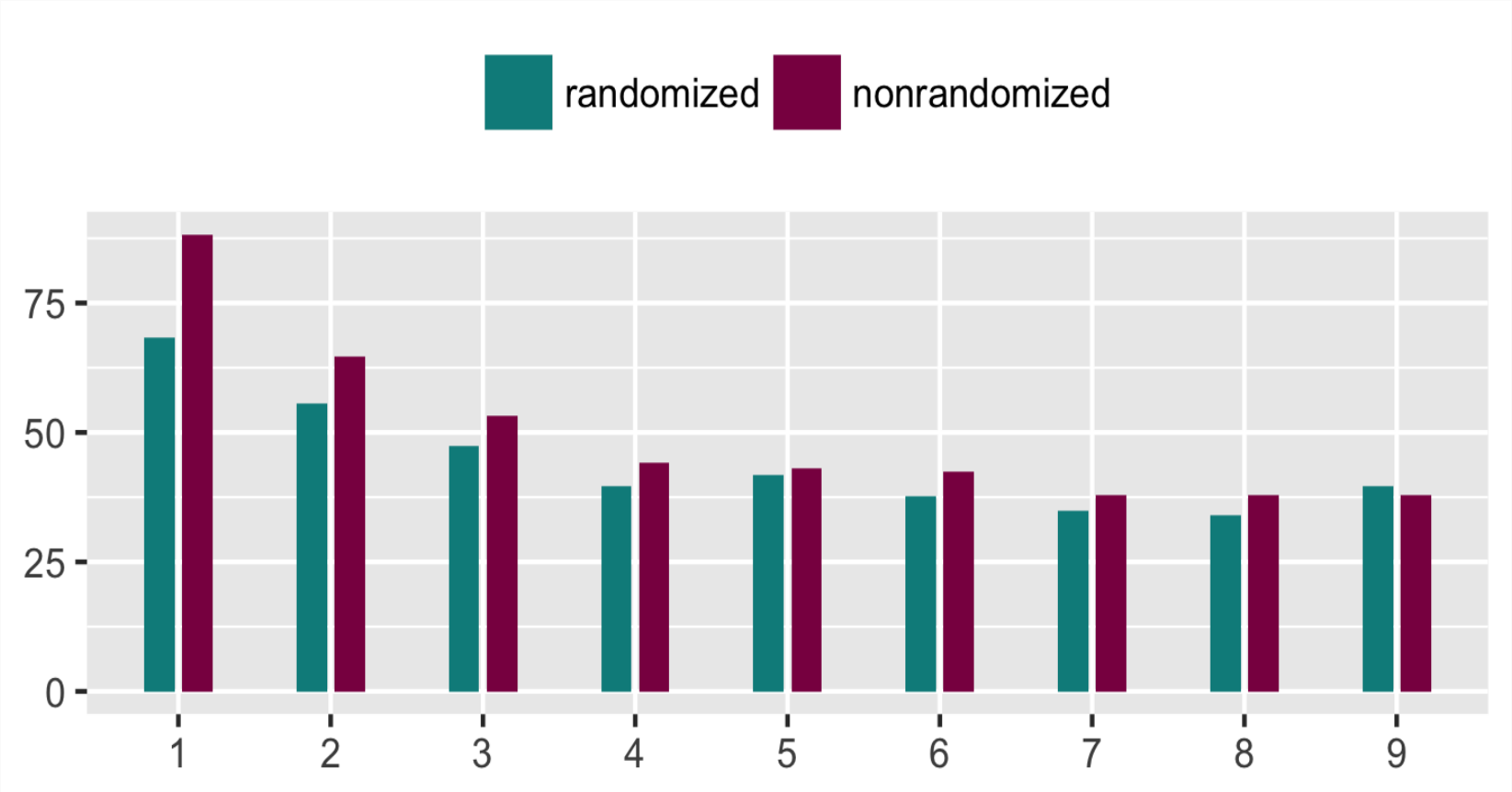
Screening powers of randomized versus non-randomized

Randomized screening - 1770 eGenes; non-randomized - 795 eGenes.

Comparison of power and FDR for eGene discoveries

Method	FDR	Power
Randomized screening	11.24	21.33
Non-randomized screening	1.65	10.51

Screening quality - randomized LASSO



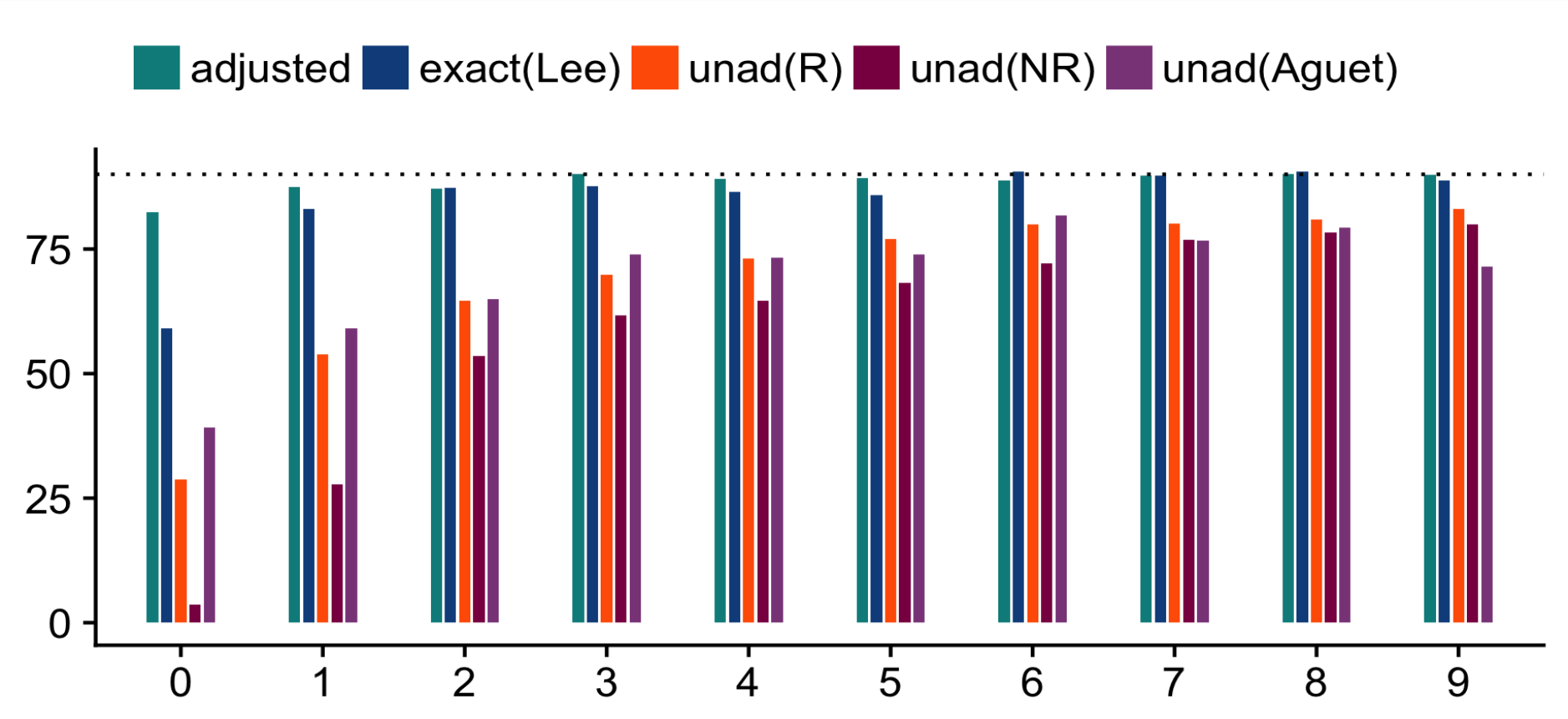
Overall screening: randomized LASSO - 39%, non-randomized LASSO - 42%.

Effect sizes of eVariants

Compare coverages and lengths of intervals, risks of the point estimates

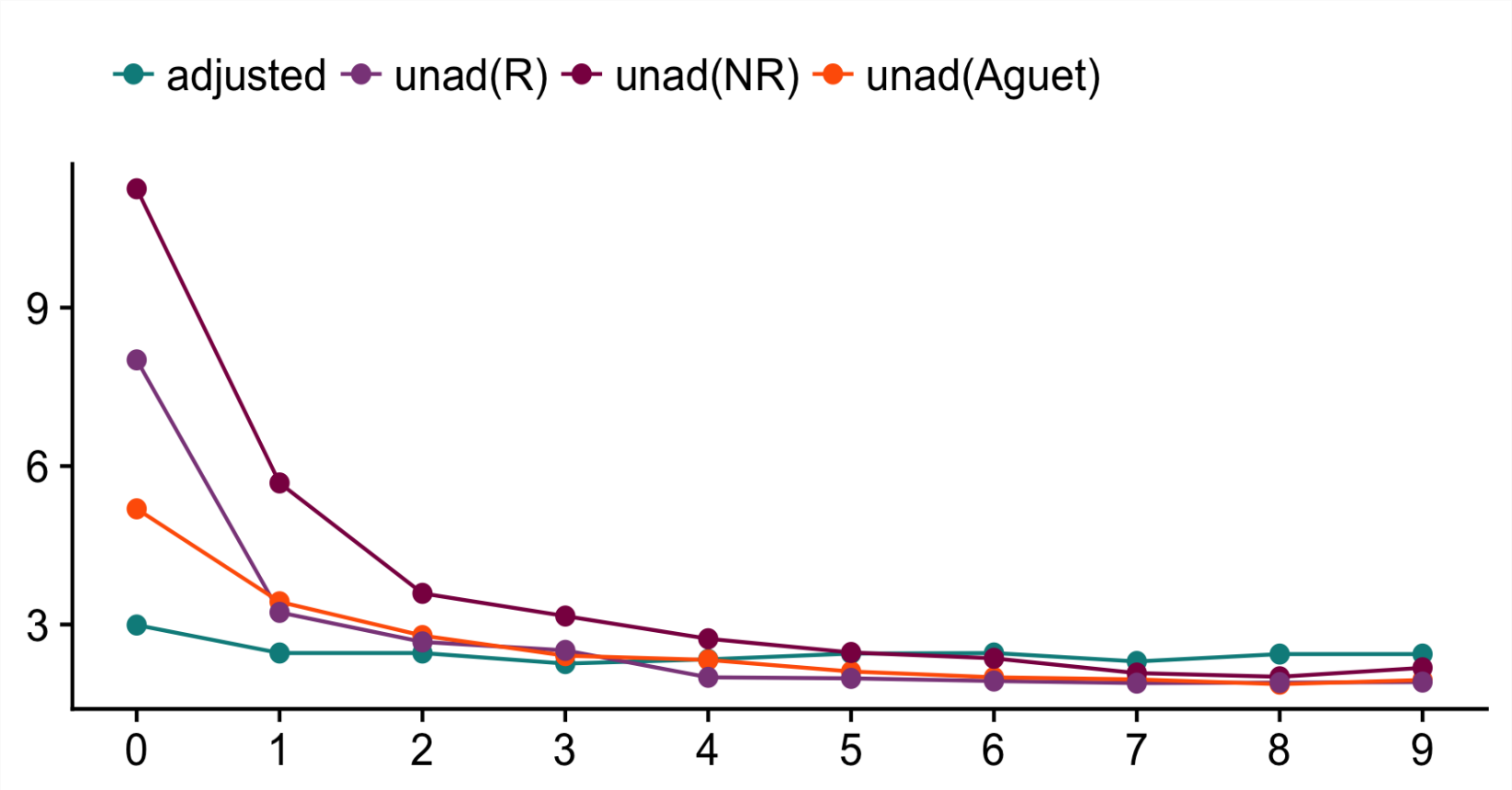
- Adjusted estimates based on our "approximation" methods
- Unadjusted estimates post randomized screenings
- Unadjusted estimates post nonrandomized screenings
- Unadjusted estimates post Aguet et al. 2016
- Lee et al. 2013 estimates adjusted for LASSO

Comparison of coverages



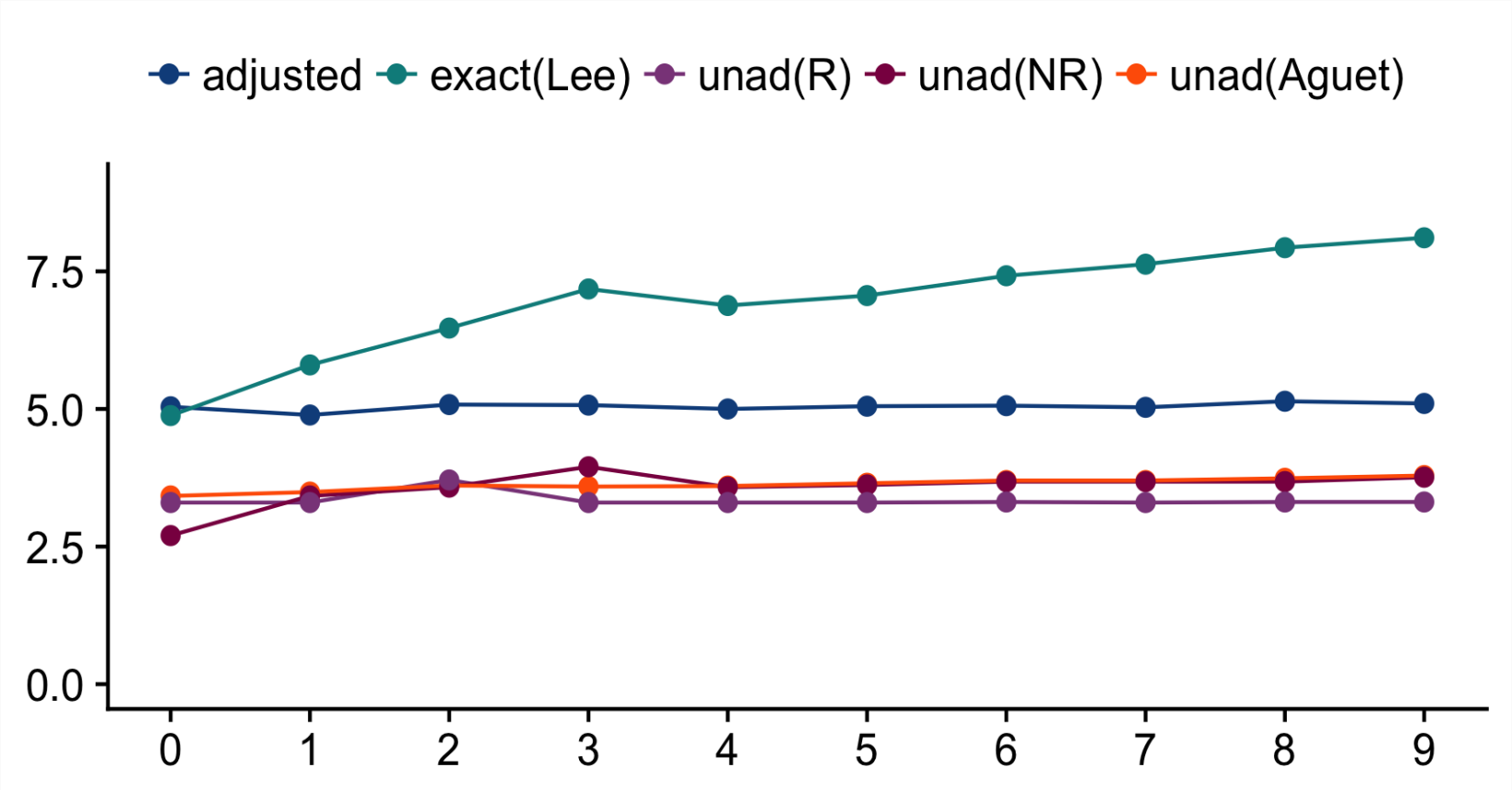
Coverage of adjusted intervals is close to target 90%, others fall short of target

Comparison of risks



Adjusted MLE strictly dominates in risk when number of signals less than 5, risks become comparable in other regimes

Comparison of lengths



Selection-adjusted intervals pay a price in length; though are only 1.5σ longer with randomization in comparison to unadjusted intervals

Effect size estimates in GTEx

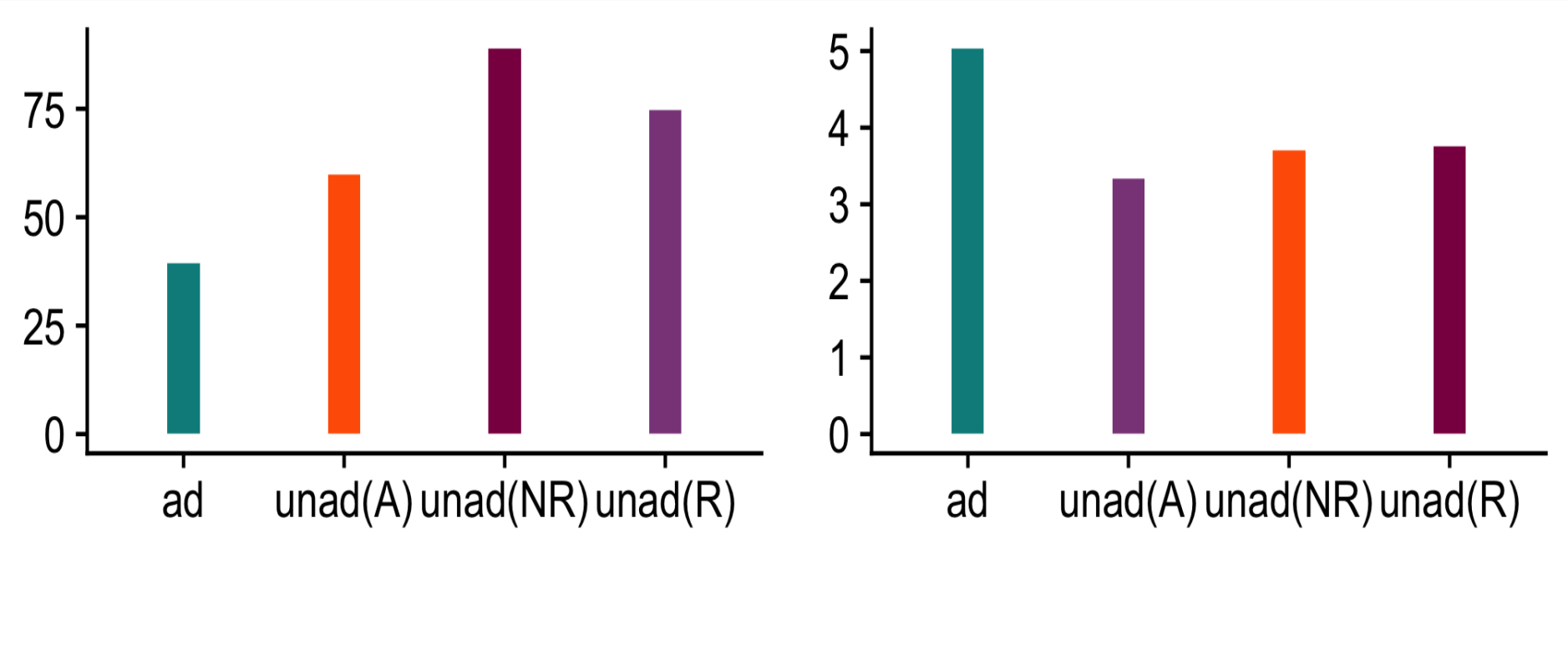
Summary statistics

- Screen **2218** eGenes with an average of **5500** eVariants
- Screen for each eGene **11** eVariants on an average

Comparison with GTEx paper:

Aguet et al. 2016 reports **5** eVariants on an average for **1500** eGenes.

Comparison of inference



Proportion of effects that are significant and comparison of lengths of intervals

Concluding remarks

- Inference post (randomized) selection: choose target and model
- Power boost in inferential stage due to randomized selection
- Scope of applying methods to other problems: effect sizes in RVA
- Use prior information to give Bayesian inference: Panigrahi et al. 2016

References

Francois Aguet et al. “Local genetic effects on gene expression across 44 human tissues”. In: *BiorXiv* (2016), p. 074450.

Rina Barber and Emmanuel Candes. “Controlling the false discovery rate via knockoffs”. In: *arXiv preprint arXiv:1404.5609* (2014).

Rina Foygel Barber and Aaditya Ramdas. “The p-filter: multi-layer FDR control for grouped hypotheses”. In: *arXiv preprint arXiv:1512.03397* (2015).

Richard Berk et al. “Valid post-selection inference”. In: *The Annals of Statistics* 41.2 (2013), pp. 802–837.

Damian Brzyski et al. “Controlling the rate of GWAS false discoveries”. In: *Genetics* 205.1 (2017), pp. 61–75.

Emmanuel Candes et al. “Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection”. In: *arXiv preprint arXiv:1610.02351* (2016).

Latarsha J Carithers et al. “A novel approach to high-quality postmortem tissue procurement: the GTEx project”. In: *Biopreservation and biobanking* 13.5 (2015), pp. 311–319.

GTEx Consortium et al. “The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans”. In: *Science* 348.6235 (2015), pp. 648–660.


DR Cox. "A note on data-splitting for the evaluation of significance levels". In: *Biometrika* 62.2 (1975), pp. 441–444.


William Fithian et al. "Optimal inference after model selection". In: *arXiv preprint arXiv:1410.2597* (2014).


Jason D Lee et al. "Exact post-selection inference with the lasso". In: *arXiv preprint arXiv:1311.6238* (2013).


Lihua Lei and William Fithian. "AdaPT: An interactive procedure for multiple testing with side information". In: *arXiv preprint arXiv:1609.06035* (2016).


Ang Li and Rina Foygel Barber. "Multiple testing with the structure adaptive Benjamini-Hochberg algorithm". In: *arXiv preprint arXiv:1606.07926* (2016).


Joshua R Loftus and Jonathan E Taylor. "Selective inference in regression models with groups of variables". In: *arXiv preprint arXiv:1511.01478* (2015).


Halit Ongen et al. "Fast and efficient QTL mapper for thousands of molecular phenotypes". In: *Bioinformatics* 32.10 (2015), pp. 1479–1485.


Snigdha Panigrahi et al. "An MCMC free approach to post-selective inference". In: *arXiv preprint arXiv:1703.06154* (2017).


Snigdha Panigrahi et al. "Bayesian Post-Selection Inference in the Linear Model". In: *arXiv preprint arXiv:1605.08824* (2016).


Christine B Peterson et al. "Many phenotypes without many false discoveries: error controlling strategies for multitrait association studies". In: *Genetic epidemiology* 40.1 (2016), pp. 45–56.


Laurel Stell and Chiara Sabatti. "Genetic variant selection: learning across traits and sites". In: *Genetics* 202.2 (2016), pp. 439–455.

Xiaoying Tian and Jonathan E Taylor. "Selective inference with a randomized response". In: *arXiv preprint arXiv:1507.06739* (2015).

Xiaoying Tian et al. "Selective sampling after solving a convex problem". In: *arXiv preprint arXiv:1609.05609* (2016).