Significance Calibration with Control Genes

David Gerard

Department of Human Genetics
University of Chicago
dcgerard@uchicago.edu
Boss: Matthew Stephens

June, 2016

Regression Model

Usual assumed model:

$$\mathbf{Y}_{n\times p} = \mathbf{X}_{n\times k}\beta_{k\times p} + \mathbf{E}_{n\times p}$$

Regression Model

Usual assumed model:

$$\mathbf{Y}_{n\times p} = \mathbf{X}_{n\times k}\boldsymbol{\beta}_{k\times p} + \mathbf{E}_{n\times p}$$

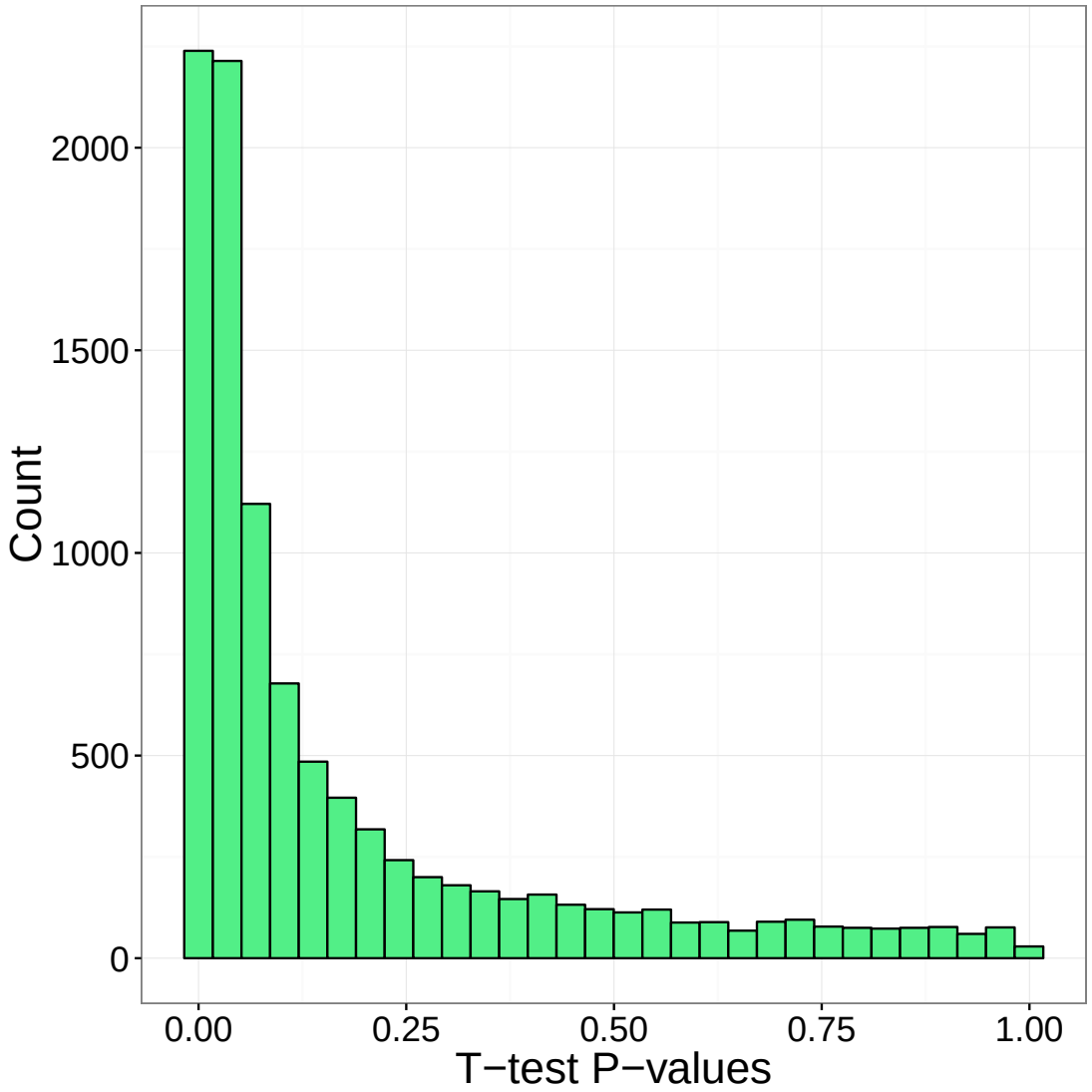
Actual model:

$$\mathbf{Y}_{n\times p} = \mathbf{X}_{n\times k}\boldsymbol{\beta}_{k\times p} + \mathbf{Z}_{n\times q}\boldsymbol{\alpha}_{q\times p} + \mathbf{E}_{n\times p}$$

Why We Need to Adjust for **Z**

- I took a 20 (samples) by 10,000 (genes) matrix of log-transformed RNAseq data from the Genotype-Tissue Expression (GTEx) project [Lonsdale et al., 2013].
- Randomly assigned half a "treatment" label and the other half a "control" label.
- Calculated p-values from two-sample t-tests.
- Since assignment to treatment group was random, all genes are theoretically null and any signal comes from hidden confounding.
- Ideally, want p-values to look uniform.

t-test p-values



Lots of Methods to Deal with Unobserved Confounding

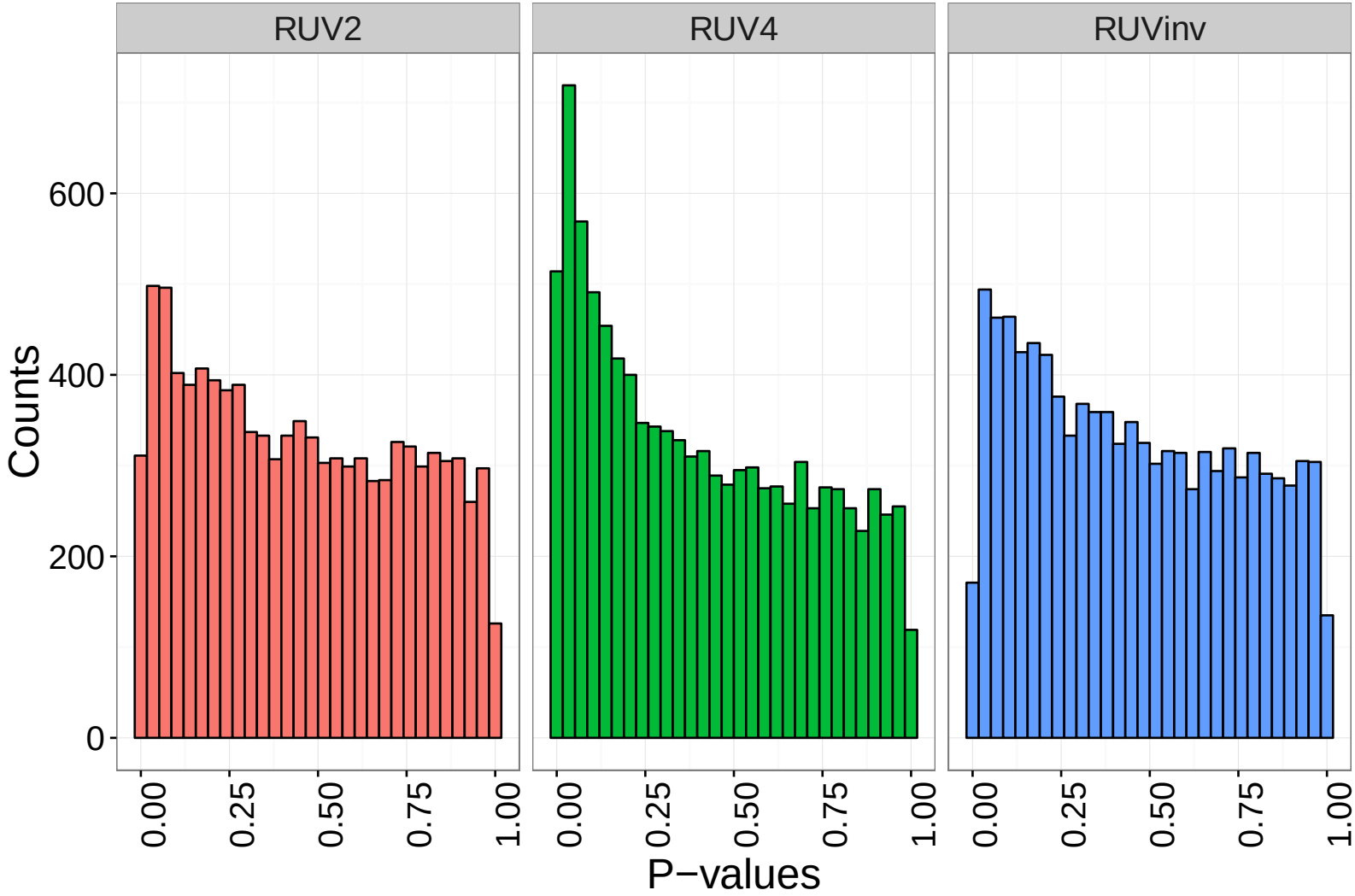
- **S**urrogate **V**ariable **A**nalysis (SVA),
- **R**emoving **U**nwanted **V**ariation (RUV) — multiple versions,
- **L**atent **E**ffect **A**djustment after **P**rimary **P**rojection (LEAPP),
- **C**onfounder **A**djusted **T**esting and **E**stimation (CATE) — multiple versions.

Lots of Methods to Deal with Unobserved Confounding

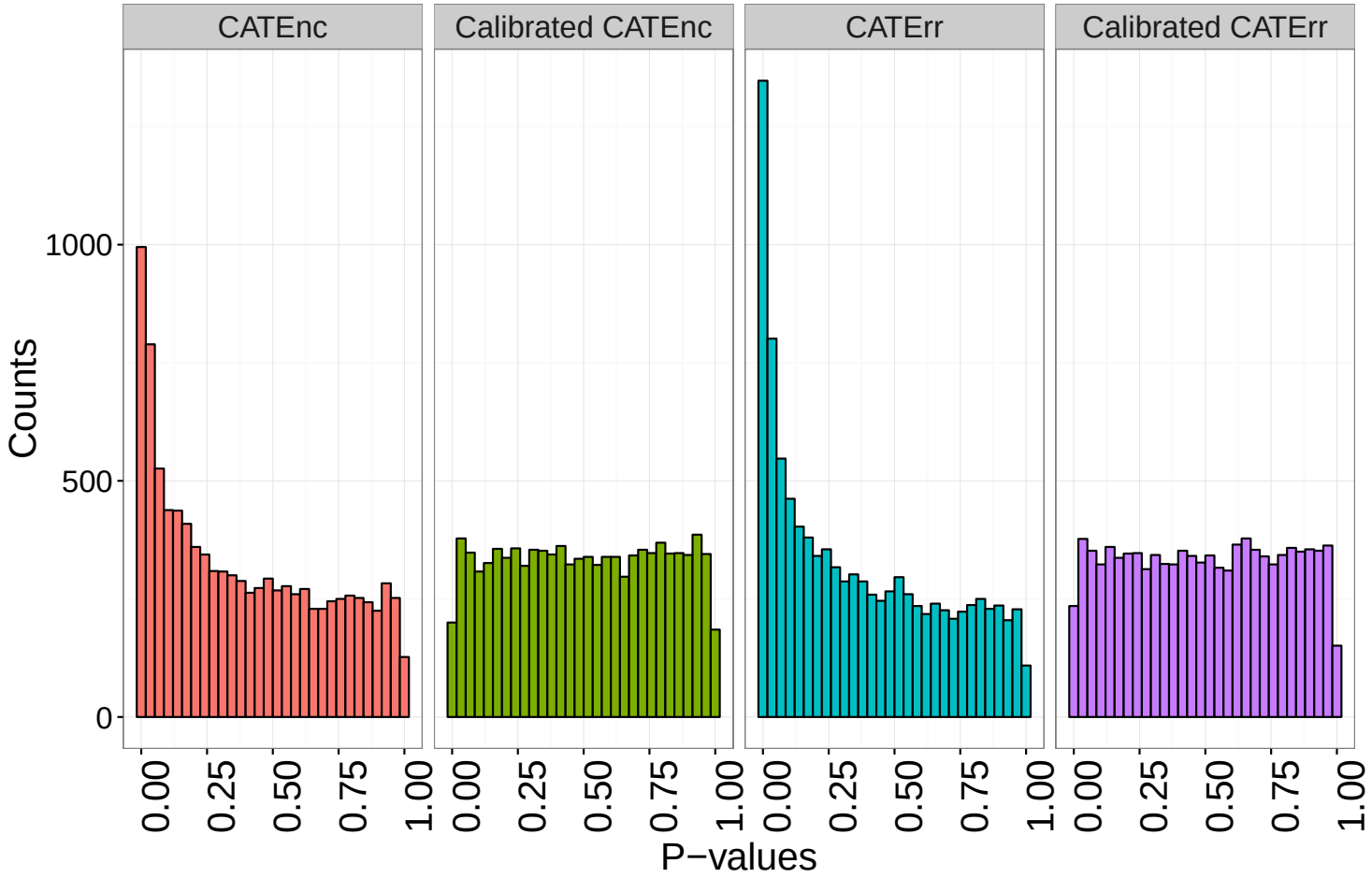
- **S**urrogate **V**ariable **A**nalysis (SVA),
- **R**emoving **U**nwanted **V**ariation (RUV) — multiple versions,
- **L**atent **E**ffect **A**djustment after **P**rimary **P**rojection (LEAPP),
- **C**onfounder **A**djusted **T**esting and **E**stimation (CATE) — multiple versions.

How do they all do?

RUV


RUV2
RUV4
RUVinv
Counts
600
400
200
0
0.00
0.25
0.50
0.75
1.00
P−values

CATE

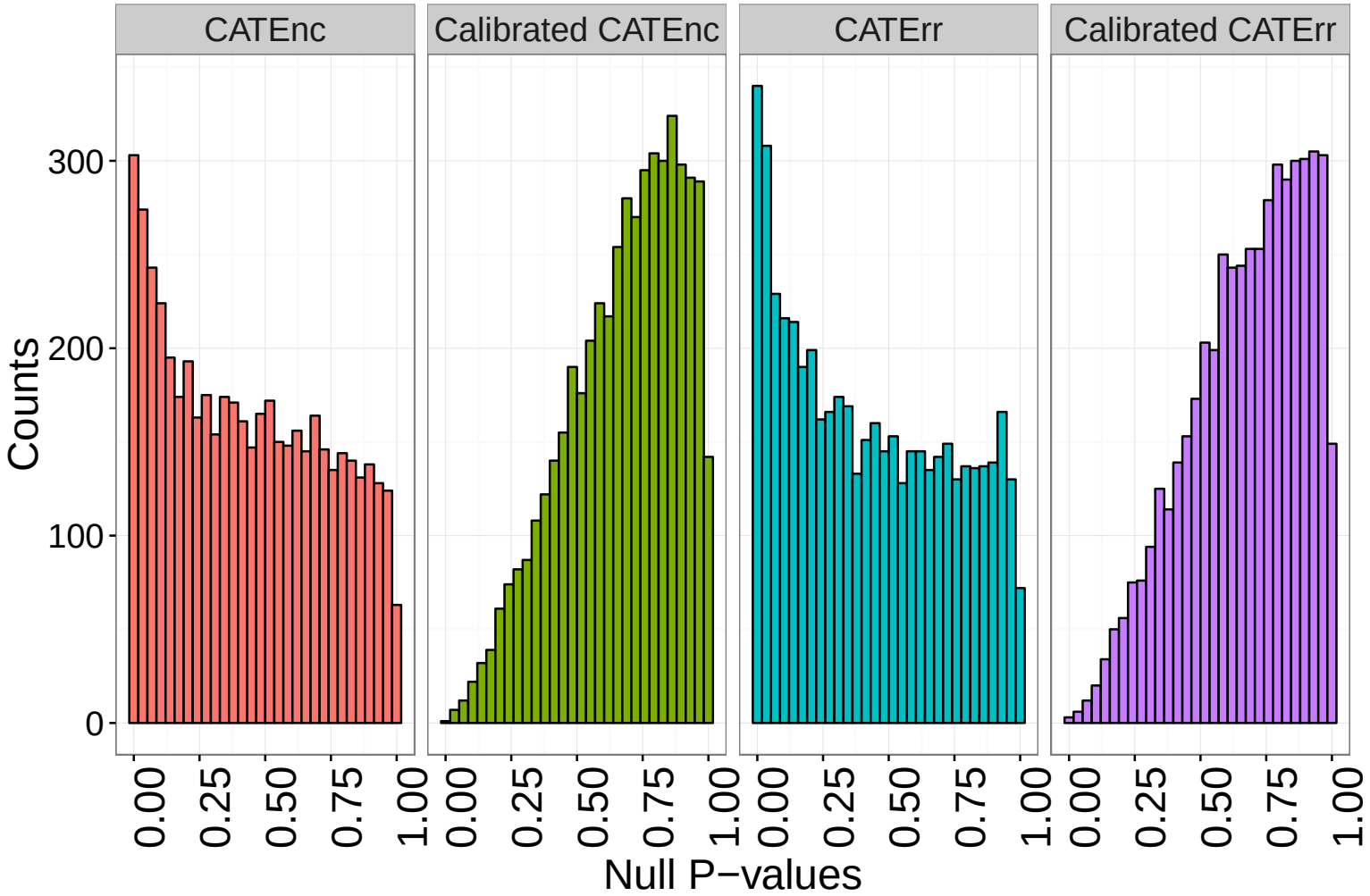

CATEnc
Calibrated CATEnc
CATErr
Calibrated CATErr
Counts
1000
500
0
0.00
0.25
0.50
0.75
1.00
P−values

Added Signal

- Calibrated CATE works really well in all-null setting.
- Calibration procedure assumes that most genes are null.
- What about when a large proportion of genes are non-null?
- Added N(0, 1) signal to half of the genes. Details

Calibrated CATE Overshrinks When Signal is Added

Looking only at null p-values:



Our Solution: Use Control Genes for Calibration

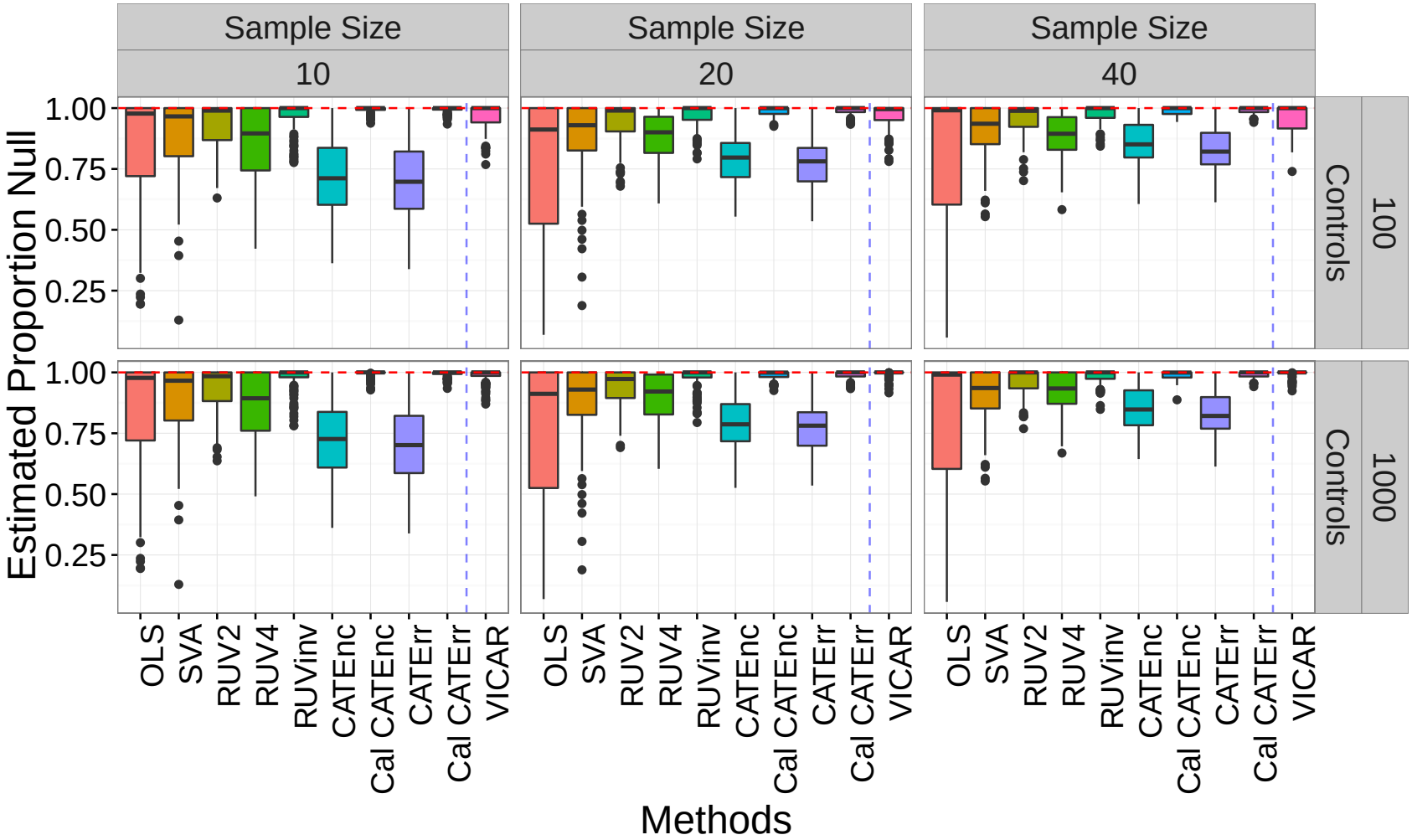
- Control Genes: Genes that are known to be unassociated with a covariate.
- Examples: Housekeeping genes, spike-in controls.
- RUV and CATEnc use controls to estimate confounders.
- We additionally use controls to calibrate variance estimates (which calibrates test statistics).
- Method: **V**ariance **I**nflation for **C**onfounder **A**djustment in **R**egression (VICAR). Details

Simulations

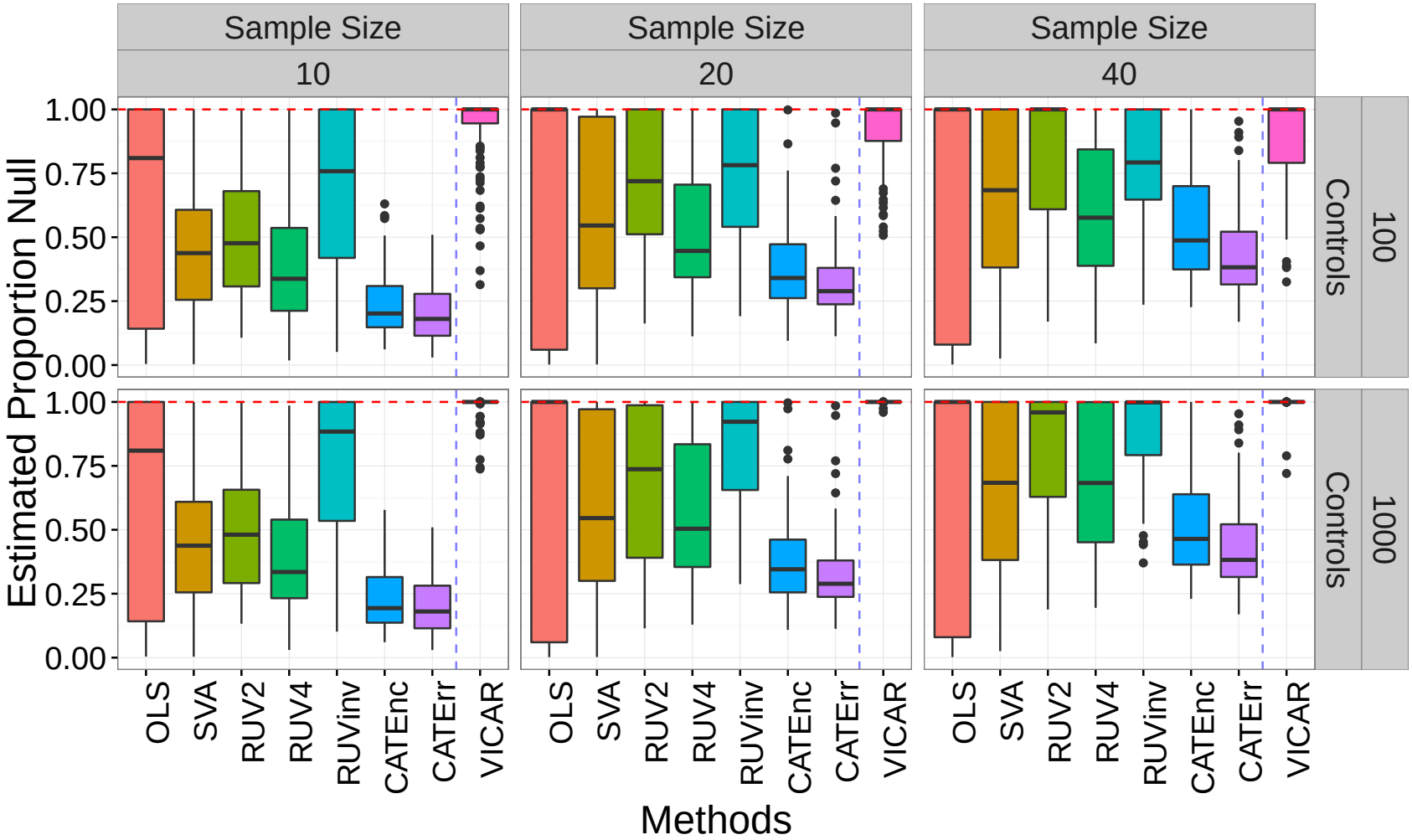
Summary Statistics Methods

- Common to use summary statistics methods to, e.g., estimate the proportion of null genes (π_0).
- How well does each method work using these summary statistics method?
- `qvalue` [Storey, 2003] take p-values.
- `ashr` [Stephens, 2016] takes estimates of the effects and the standard errors.

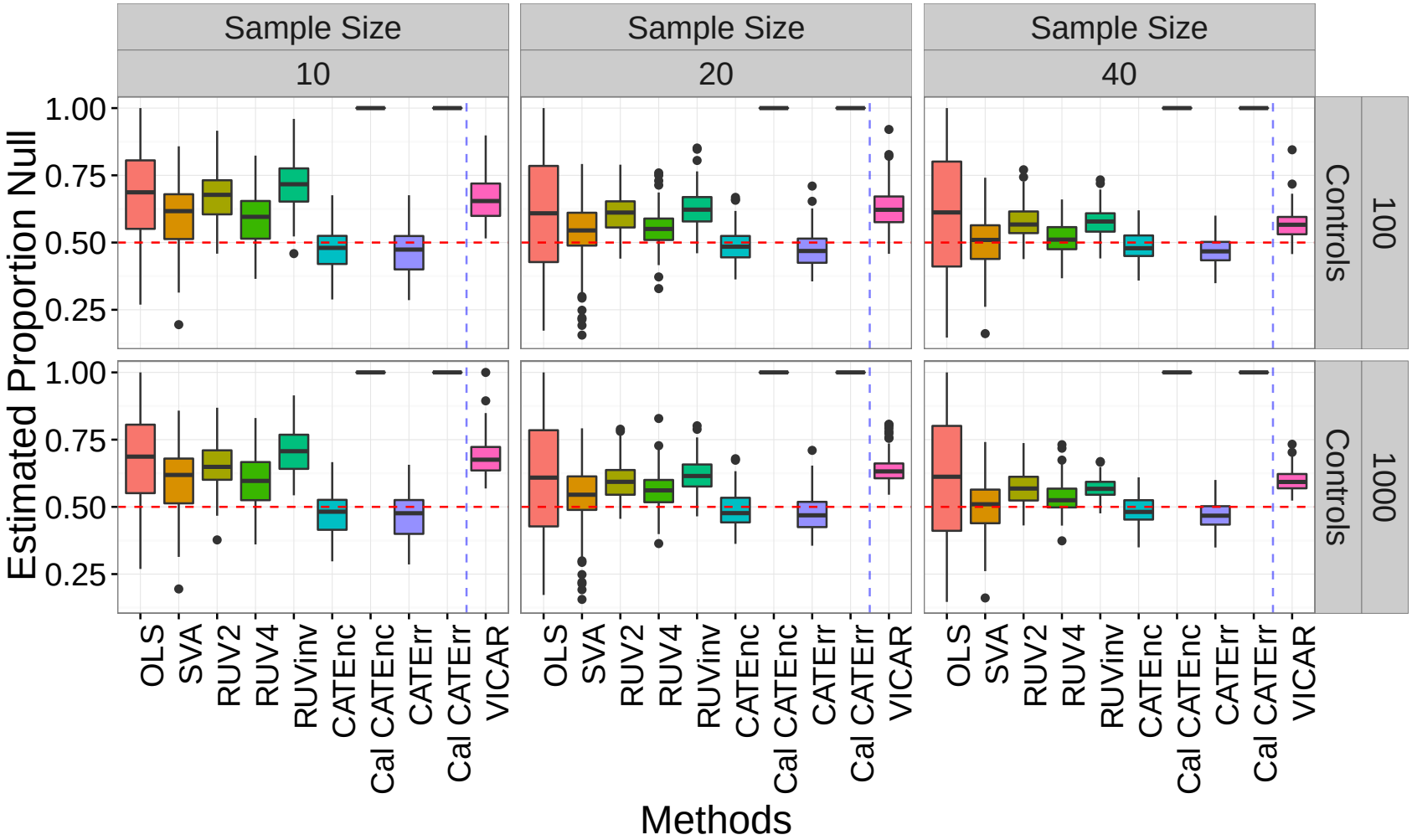
All Null Simulation Results: qvalue



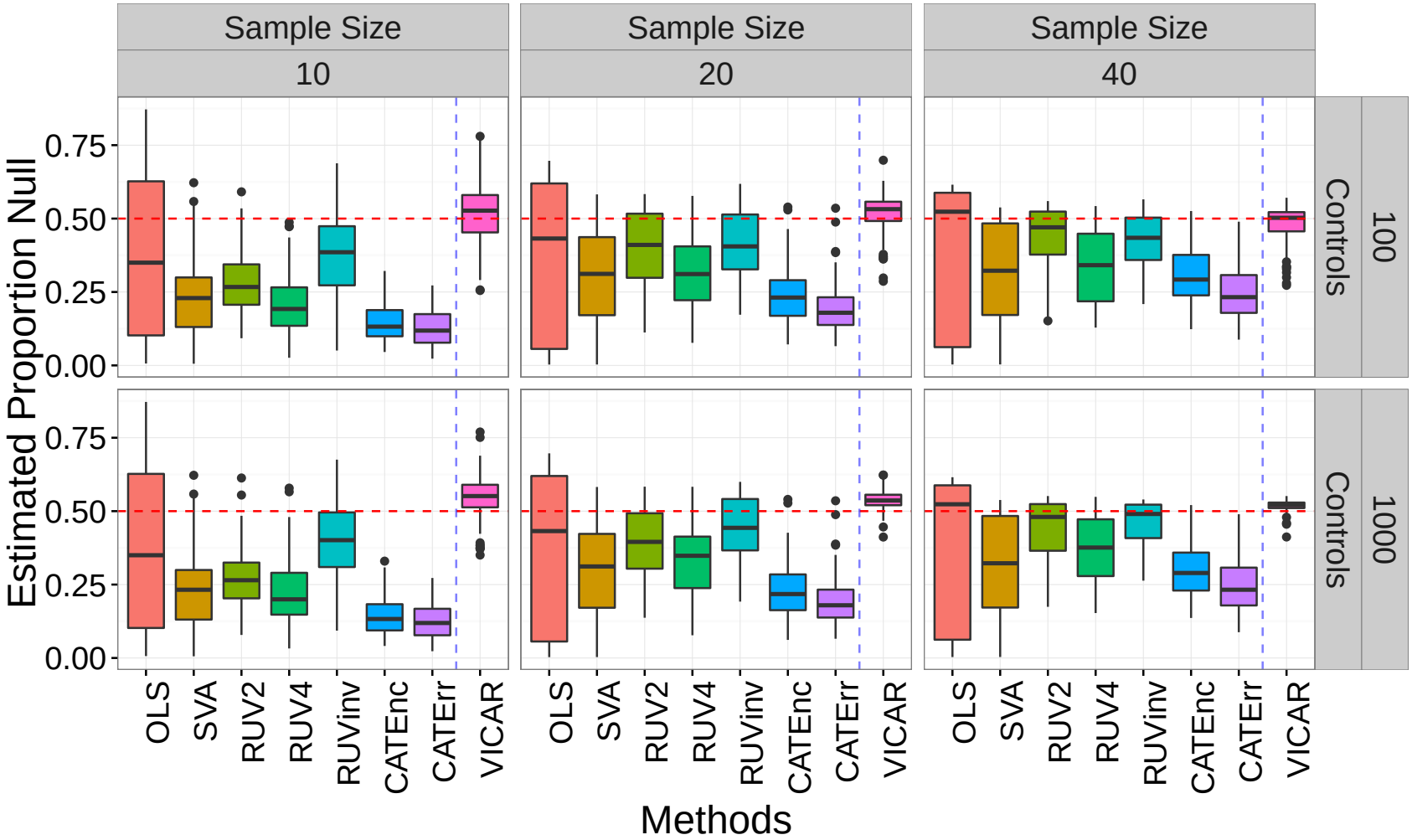
All Null Simulation Results: ashr



Added Signal Simulations Results: qvalue



Added Signal Simulations Results: `ashr`

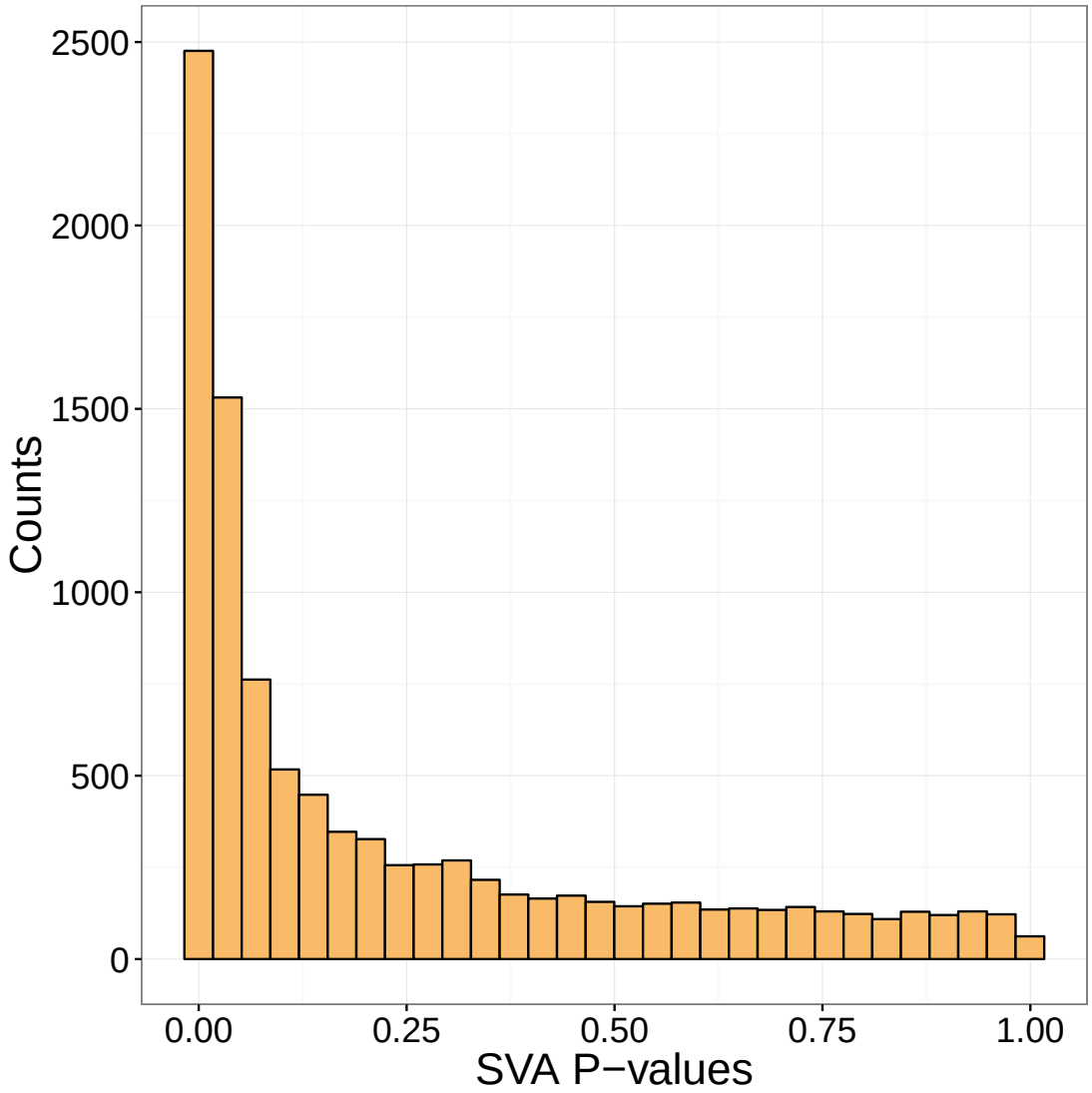


Thank You

`vicar` R package: https://github.com/dcgerard/vicar

Appendix

SVA All-null P-values



Details of Adding Signal

Draw

$$a_{i_1}, \ldots, a_{i_{\pi_0 p}} \sim N(0, 1)$$

where $i_\ell \in \Omega \subseteq \{1, \ldots, p\}$, the set of non-null genes. Let

$$z_{ij}|y_{ij} = \begin{cases} Binom(y_{ij}, 2^{a_j x_i}) & \text{if } a_j < 0 \text{ and } j \in \Omega \\ Binom(y_{ij}, 2^{-a_j(1-x_i)}) & \text{if } a_j > 0 \text{ and } j \in \Omega \\ y_{ij} & \text{if } j \notin \Omega. \end{cases}$$

We use $\mathbf{Z}$ as our new gene-expression data set.

Justification for This Approach

Suppose $y_{ij} \sim Poisson(\lambda_j)$, and let x_i be the indicator of treatment versus control for sample i. Then

$$z_{ij}|a_j, a_j < 0, j \in \Omega \sim Poisson(2^{a_j x_i}\lambda_j)$$
$$z_{ij}|a_j, a_j > 0, j \in \Omega \sim Poisson(2^{-a_j(1-x_i)}\lambda_j),$$

and

$$E[\log_2(z_{ij}) - \log_2(z_{kj})|a_j, a_j < 0, j \in \Omega] \approx a_j x_i - a_j x_k, \text{ and}$$
$$E[\log_2(z_{ij}) - \log_2(z_{kj})|a_j, a_j > 0, j \in \Omega] \approx -a_j(1 - x_i) + a_j(1 - x_k).$$

So a_j is the $\log_2$-fold signal between treatment and control. Go Back

Details of VICAR: Setup

$$\mathbf{Y}_{n\times p} = \mathbf{X}_{n\times k}\boldsymbol{\beta}_{k\times p} + \mathbf{Z}_{n\times q}\boldsymbol{\alpha}_{q\times p} + \mathbf{E}_{n\times p},$$
$$\mathbf{E} \sim N_{n\times p}(0, \boldsymbol{\Sigma} \otimes \mathbf{I}_n)$$
$$\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2).$$

Apply rotation from [Wang et al., 2015] to obtain three independent models:

$$\mathbf{Y}_1 = \mathbf{R}_{11}\boldsymbol{\beta}_1 + \mathbf{R}_{12}\boldsymbol{\beta}_2 + \mathbf{Z}_1\boldsymbol{\alpha} + \mathbf{E}_1 \quad (1)$$
$$\mathbf{Y}_2 = \mathbf{R}_{22}\boldsymbol{\beta}_2 + \mathbf{Z}_2\boldsymbol{\alpha} + \mathbf{E}_2 \quad (2)$$
$$\mathbf{Y}_3 = \mathbf{Z}_3\boldsymbol{\alpha} + \mathbf{E}_3, \quad (3)$$

where $\boldsymbol{\beta}_2$ are the coefficients of interest.

Details of VICAR: What CATEnc and RUV4 do

1. Estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ using (3).
2. Estimate $\mathbf{Z}_2$ using (2) and the control genes assuming $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ are known from step 1.
3. Estimate β_2 by $\mathbf{R}_{22}^{-1}(\mathbf{Y}_2 - \hat{\mathbf{Z}}_2\hat{\boldsymbol{\alpha}})$.
4. Standard errors are same as using $(\mathbf{X}, \hat{\mathbf{Z}})$ as your covariates and $\hat{\boldsymbol{\Sigma}}$ as your variance estimates.

Details of VICAR: What VICAR does

1. Estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$ using (3).
2. Estimate $\mathbf{Z}_2$ *and a variance inflation parameter* λ using (2) and the control genes assuming $\boldsymbol{\alpha}$ is known and the variances are $\lambda\boldsymbol{\Sigma}$ with $\boldsymbol{\Sigma}$ known from step 1.
3. Estimate β_2 by $\mathbf{R}_{22}^{-1}(\mathbf{Y}_2 - \hat{\mathbf{Z}}_2\hat{\boldsymbol{\alpha}})$.
4. Standard errors are same as using $(\mathbf{X}, \hat{\mathbf{Z}})$ as your covariates and $\hat{\lambda}\hat{\boldsymbol{\Sigma}}$ as your variance estimates.

Go Back

References I

Gagnon-Bartsch, J., Jacob, L., and Speed, T. (2013).

Removing unwanted variation from high dimensional data with negative controls.

Technical report, Technical Report 820, Department of Statistics, University of California, Berkeley.

Gagnon-Bartsch, J. A. and Speed, T. P. (2012).

Using control genes to correct for unwanted variation in microarray data.

Biostatistics, 13(3):539–552.

Leek, J. T. and Storey, J. D. (2008).

A general framework for multiple testing dependence.

Proceedings of the National Academy of Sciences, 105(48):18718–18723.

References II

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013).

The genotype-tissue expression (gtex) project.

Nature genetics, 45(6):580–585.

Stephens, M. (2016).

False discovery rates: A new deal.

bioRxiv, page 038216.

Storey, J. D. (2003).

The positive false discovery rate: a bayesian interpretation and the q-value.

Annals of statistics, pages 2013–2035.

References III

Sun, Y., Zhang, N. R., Owen, A. B., et al. (2012).

Multiple hypothesis testing adjusted for latent variables, with an application to the agemap gene expression data.

The Annals of Applied Statistics, 6(4):1664–1688.

Wang, J., Zhao, Q., Hastie, T., and Owen, A. B. (2015).

Confounder adjustment in multiple hypotheses testing.

arXiv preprint arXiv:1508.04178.