

Practical recommendations for controlling false discoveries in computational biology

Keegan Korthauer, PhD
Postdoctoral Research Fellow
 @keegankorthauer

European Bioconductor Meeting
Technical University of Munich, Germany
7 December 2018

Multiple comparisons in computational biology

20,000
genes

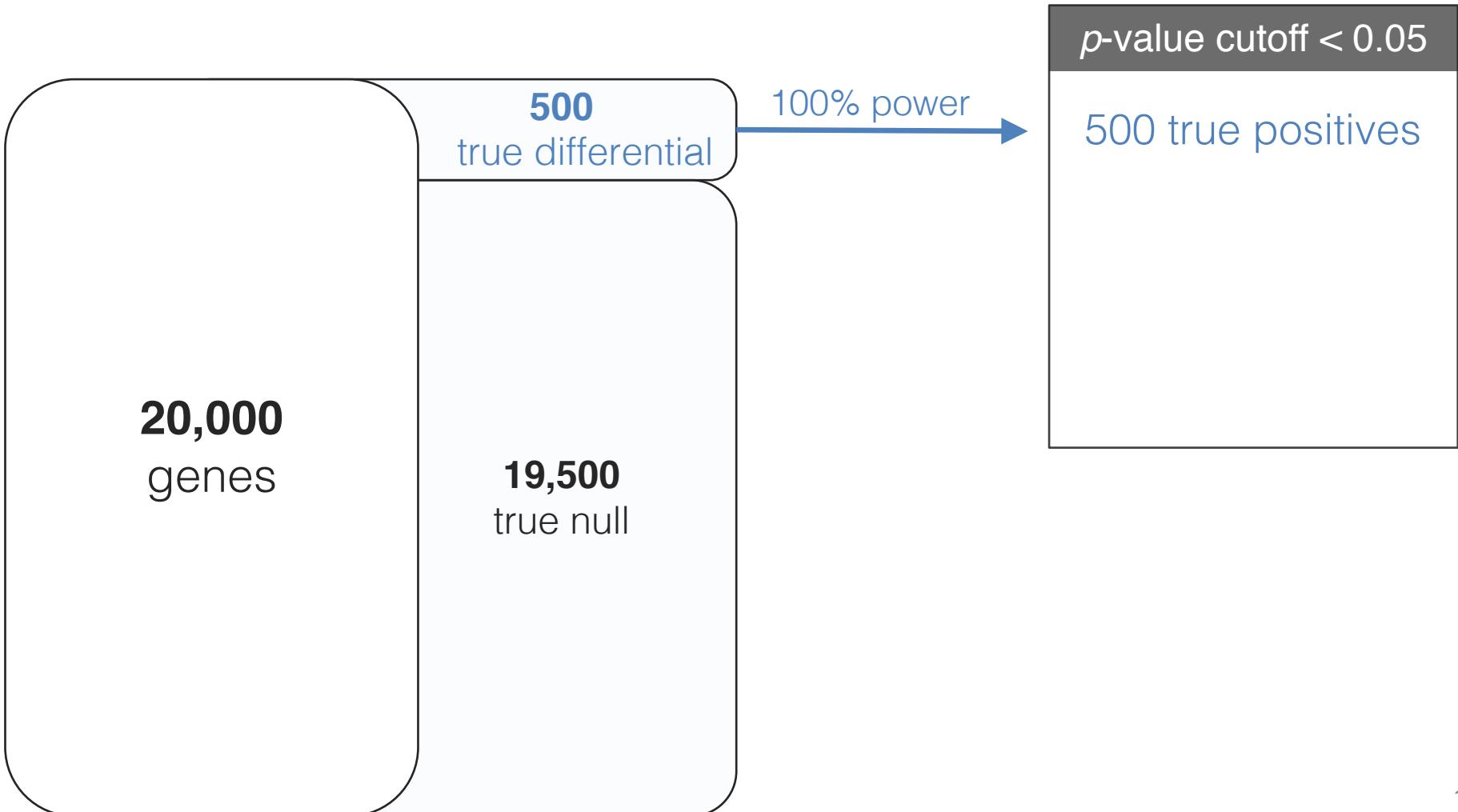
Multiple comparisons in computational biology

20,000
genes

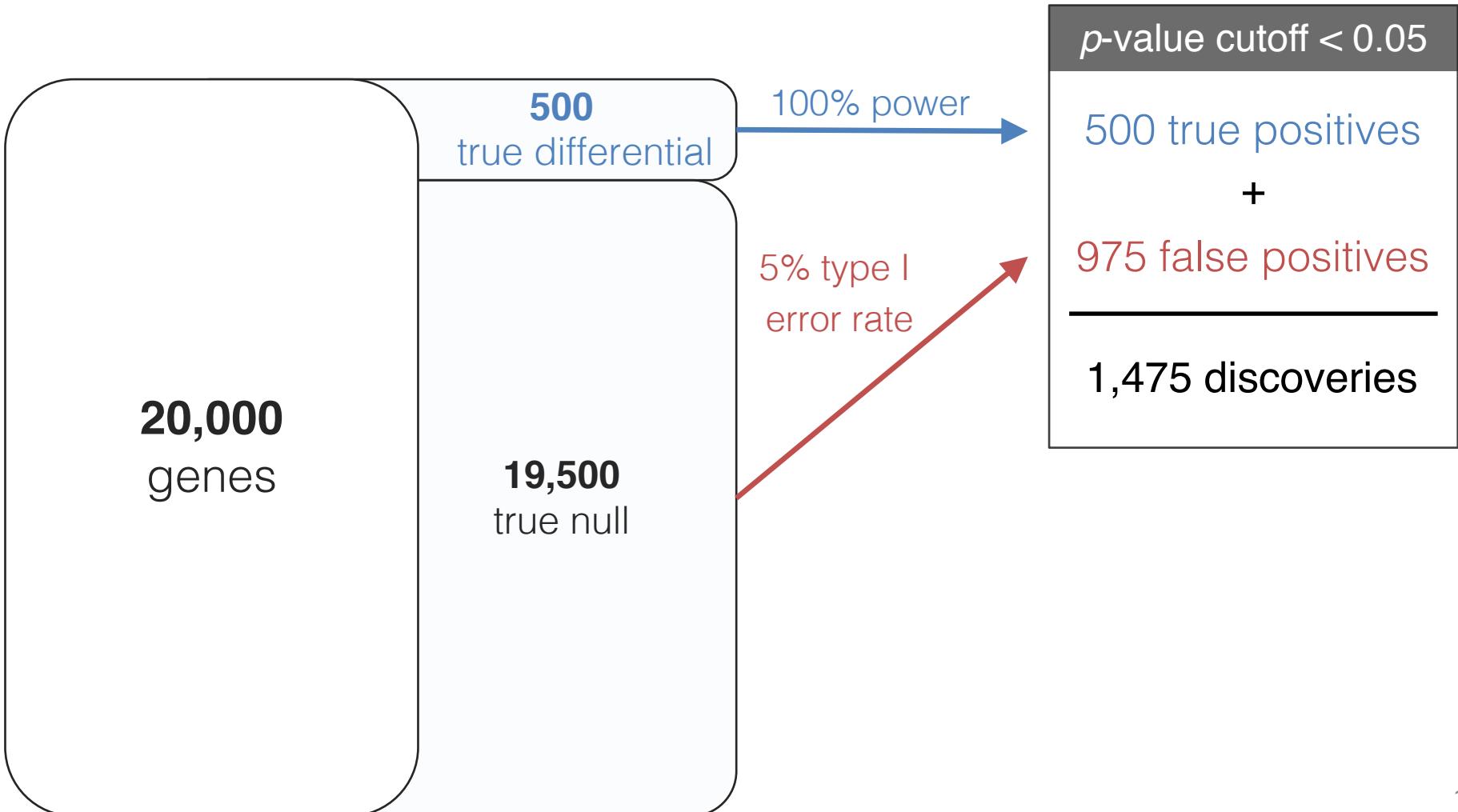
500
true differential

19,500
true null

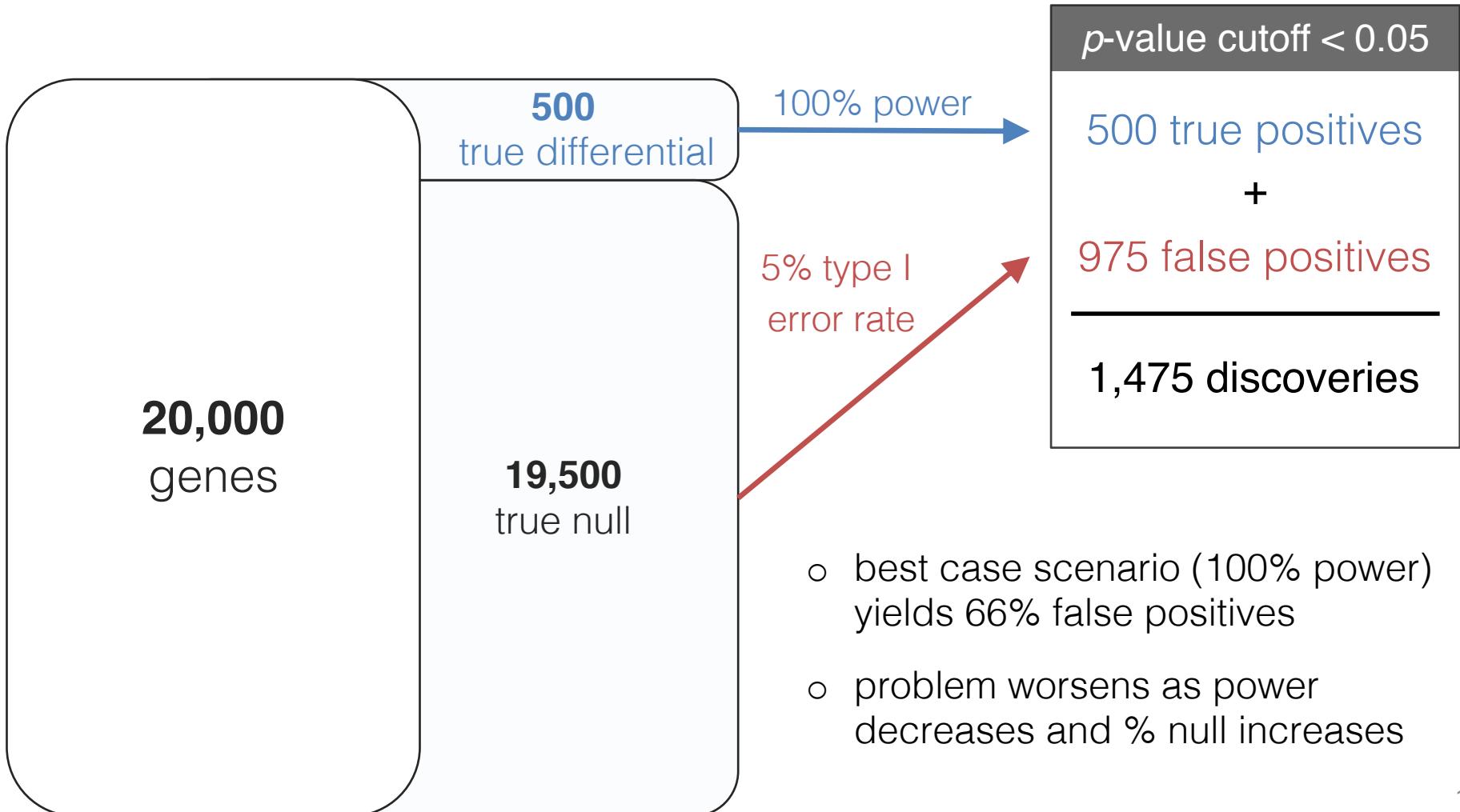
Multiple comparisons in computational biology



Multiple comparisons in computational biology



Multiple comparisons in computational biology



Controlling false positives

Family-wise Error Rate (FWER)

- Bonferroni correction

$$P(\text{at least one false positive}) < \alpha$$

Controlling false positives

Family-wise Error Rate (FWER)

- Bonferroni correction

$$P(\text{at least one false positive}) < \alpha$$

- lowers significance cutoff to $\frac{\alpha}{\# \text{ tests}}$
- e.g. for 20,000 tests and $\alpha = 0.05$, becomes 2.5×10^{-6}

Controlling false positives

Family-wise Error Rate (FWER)

- Bonferroni correction

$$P(\text{at least one false positive}) < \alpha$$

- lowers significance cutoff to $\frac{\alpha}{\# \text{ tests}}$
- e.g. for 20,000 tests and $\alpha = 0.05$, becomes 2.5×10^{-6}

False Discovery Rate (FDR)

- Benjamini-Hochberg (BH) adjustment
- Storey's q-value

$$E\left(\frac{\# \text{ false positives}}{\# \text{ total positives}}\right) < \alpha$$

Controlling false positives

Family-wise Error Rate (FWER)

- Bonferroni correction

$$P(\text{at least one false positive}) < \alpha$$

- lowers significance cutoff to $\frac{\alpha}{\# \text{ tests}}$
- e.g. for 20,000 tests and $\alpha = 0.05$, becomes 2.5×10^{-6}

False Discovery Rate (FDR)

- Benjamini-Hochberg (BH) adjustment
- Storey's q-value

$$E\left(\frac{\# \text{ false positives}}{\# \text{ total positives}}\right) < \alpha$$

- Most commonly used in high-throughput analyses

Moving beyond BH and Storey's q -value

BH and q -value

- all tests treated equal

Moving beyond BH and Storey's *q*-value

BH and *q*-value

- all tests treated equal

Reality

- all tests not equal
 - eQTL cis vs. trans
 - RNA-seq mean expression

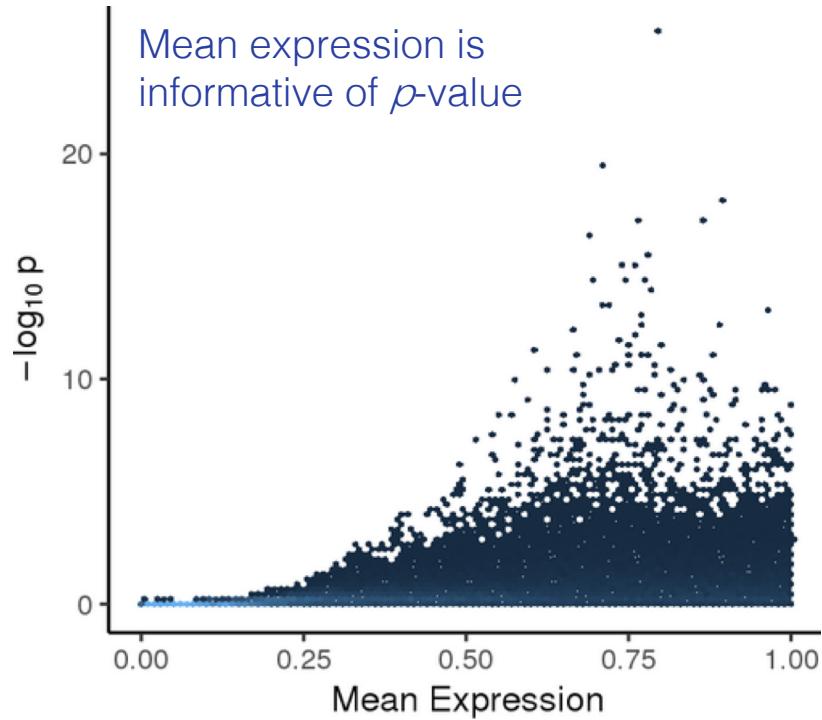
Moving beyond BH and Storey's q -value

BH and q -value

- all tests treated equal

Reality

- all tests not equal
 - eQTL cis vs. trans
 - RNA-seq mean expression



Moving beyond BH and Storey's q -value

BH and q -value

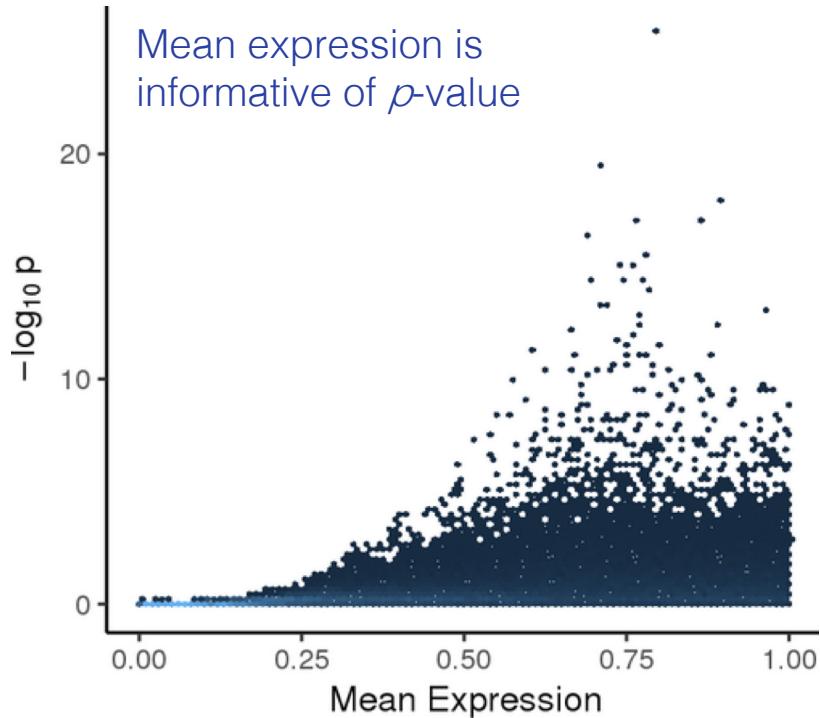
- all tests treated equal

Reality

- all tests not equal
 - eQTL cis vs. trans
 - RNA-seq mean expression

Covariate-aware methods

- model differences in tests via covariates
- recent explosion of methods



Timeline

1995	BH procedure
2001	Storey's q -value
2009	conditional local FDR (LFDR)
2015	FDR regression (FDRreg)
2016	Independent Hypothesis Weighting (IHW)
2017	Adaptive Shrinkage (ASH)
2018	Boca-Leek (BL)
	Adaptive p -value Thresholding (AdaPT)

Understanding covariate-aware methods for FDR control

consider the two-groups model:

classic methods

$$p_i \sim \pi_0 f_0 + (1 - \pi_0) f_1$$

probability of test being null

null distribution (uniform)

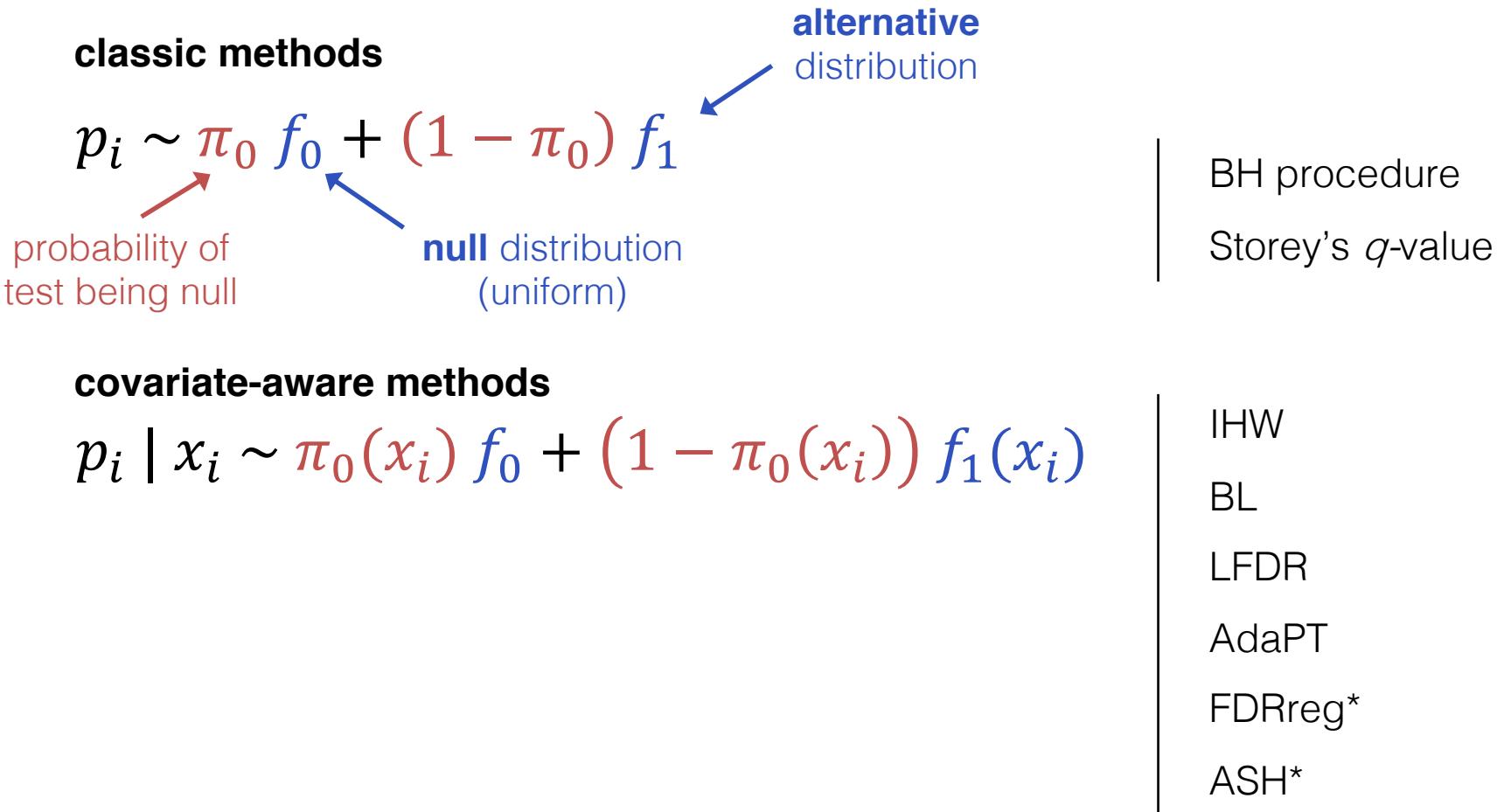
alternative distribution

BH procedure
Storey's q -value

The diagram illustrates the two-groups model for p-values. The equation $p_i \sim \pi_0 f_0 + (1 - \pi_0) f_1$ represents the distribution of p-values. A red arrow points from the term $\pi_0 f_0$ to the text "probability of test being null". A blue arrow points from the term $(1 - \pi_0) f_1$ to the text "null distribution (uniform)". Another blue arrow points from the term f_1 to the text "alternative distribution". To the right of the equation, there is a vertical line followed by two items: "BH procedure" and "Storey's q -value".

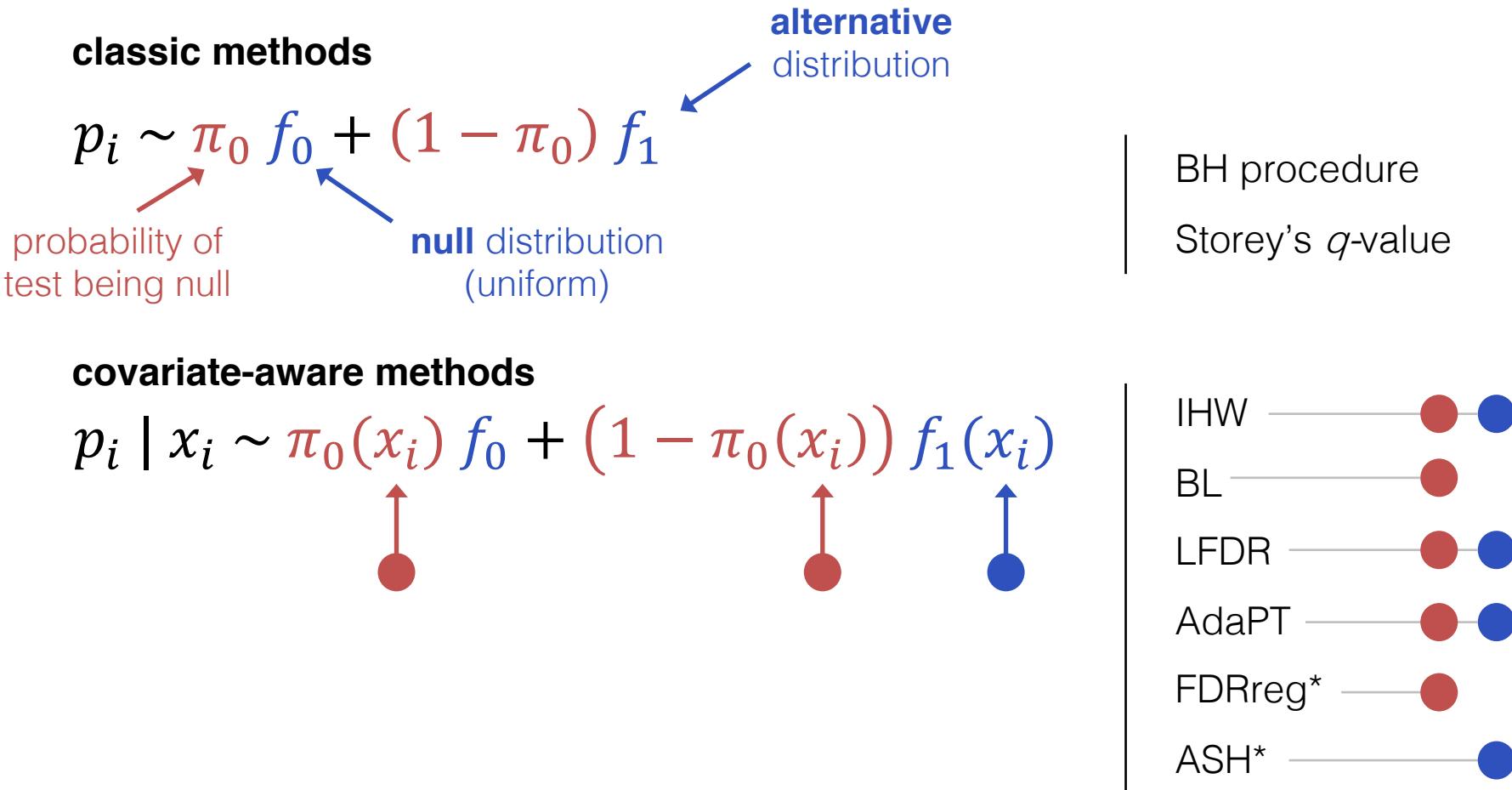
Understanding covariate-aware methods for FDR control

consider the two-groups model:



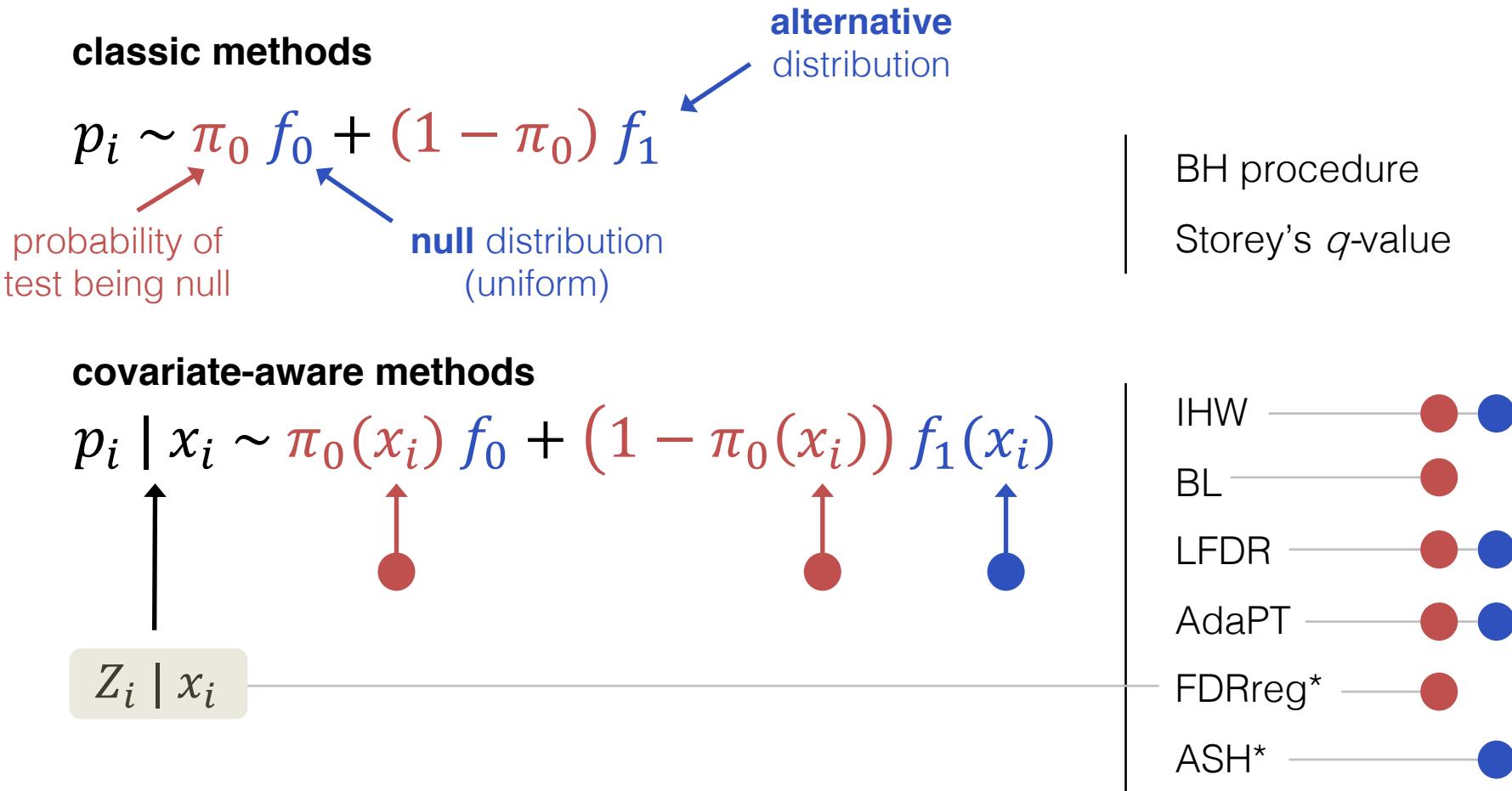
Understanding covariate-aware methods for FDR control

consider the two-groups model:



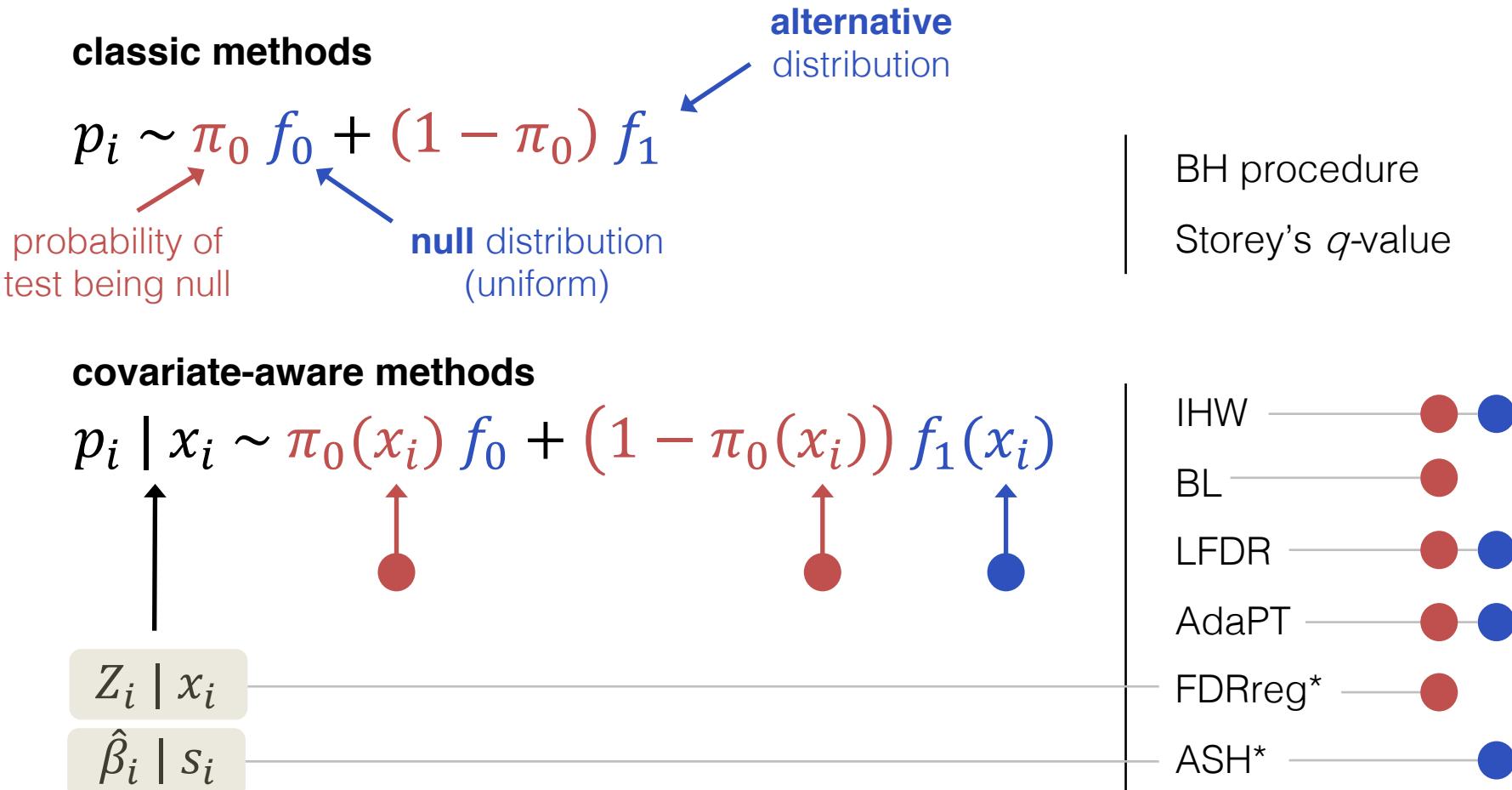
Understanding covariate-aware methods for FDR control

consider the two-groups model:



Understanding covariate-aware methods for FDR control

consider the two-groups model:



Inputs and outputs

	Input	Output	R package
BH	p -values	adjusted p -values	stats
IHW	(1) p -values (2) independent & informative covariate		ihw
q-value	p -values	q -values	qvalue
BL		adjusted p -values	swfdr
AdaPT	(1) p -values (2) independent & informative covariate	q -values	adaptMT
LFDR		adjusted p -values	none
FDRreg	(1) z-scores (2) independent & informative covariate	Bayesian FDRs	FDRreg
ASH	(1) effect sizes (2) standard errors of (1)	q -values	ash

Inputs and outputs

	Input	Output	R package	
BH	p -values	adjusted p -values	stats	
IHW	(1) p -values (2) independent & informative covariate		ihw	
q-value	p -values	q -values	qvalue	
BL		adjusted p -values	swfdr	
AdaPT	(1) p -values (2) independent & informative covariate	q -values	adaptMT	
LFDR		adjusted p -values	none	
FDRreg	(1) z -scores (2) independent & informative covariate	Bayesian FDRs	FDRreg	
ASH	(1) effect sizes (2) standard errors of (1)	q -values	ash	

Repository

 Bioconductor

 CRAN

 GitHub

Benchmarking for practical recommendations

Methods

Classic

BH procedure

Storey's q -value

Covariate-aware

IHW

BL

LFDR

AdaPT

FDRreg

ASH

Benchmarking for practical recommendations

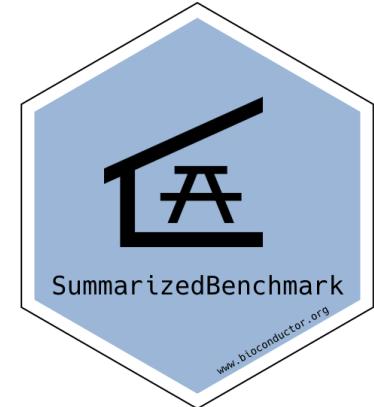
	Methods	Datasets
Classic	BH procedure Storey's q -value	<i>in silico</i> experiments pure simulations
Covariate-aware	IHW BL LFDR AdaPT FDRreg ASH	bulk RNA-seq DE scRNA-seq DE 16S microbiome DA ChIP-seq DB GWAS Gene Set Analyses

Benchmarking for practical recommendations

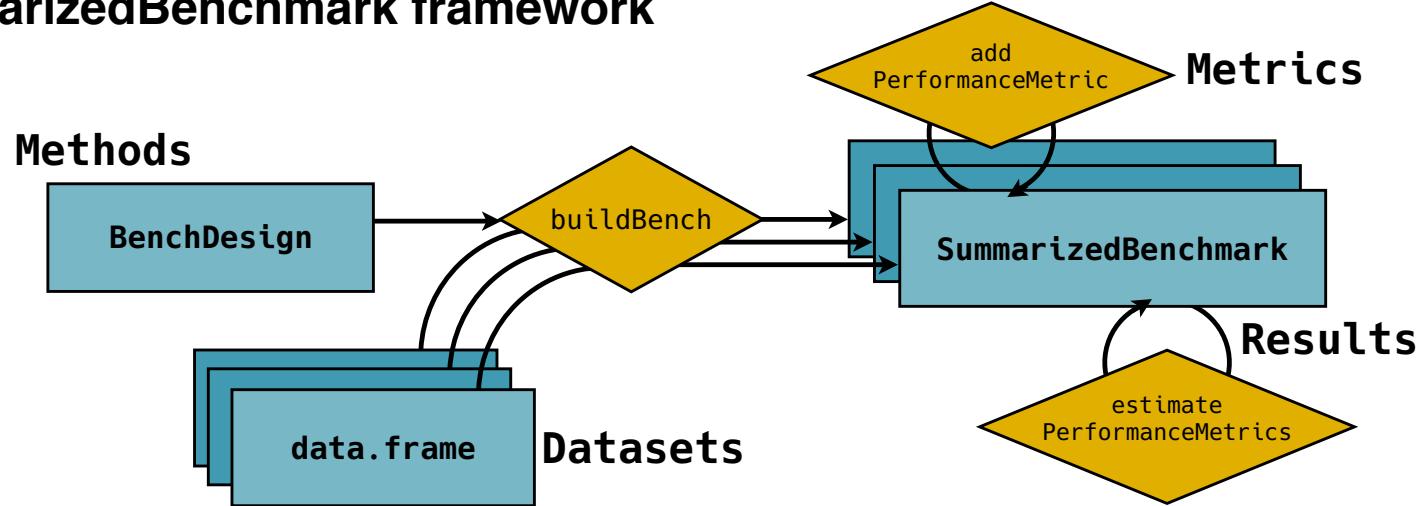
Methods	Datasets	Evaluations
Classic	BH procedure	FDR control
	Storey's q -value	
Covariate-aware	IHW	Power
	BL	
Case studies	LFDR	Applicability
	AdaPT	
	FDRreg	
	ASH	
	<i>in silico</i> experiments	
	pure simulations	
Case studies	bulk RNA-seq DE	Consistency
	scRNA-seq DE	
	16S microbiome DA	
	ChIP-seq DB	
	GWAS	
	Gene Set Analyses	

Software to facilitate benchmarking

Bioconductor package **SummarizedBenchmark**
enables reproducible comparisons across
methods + datasets

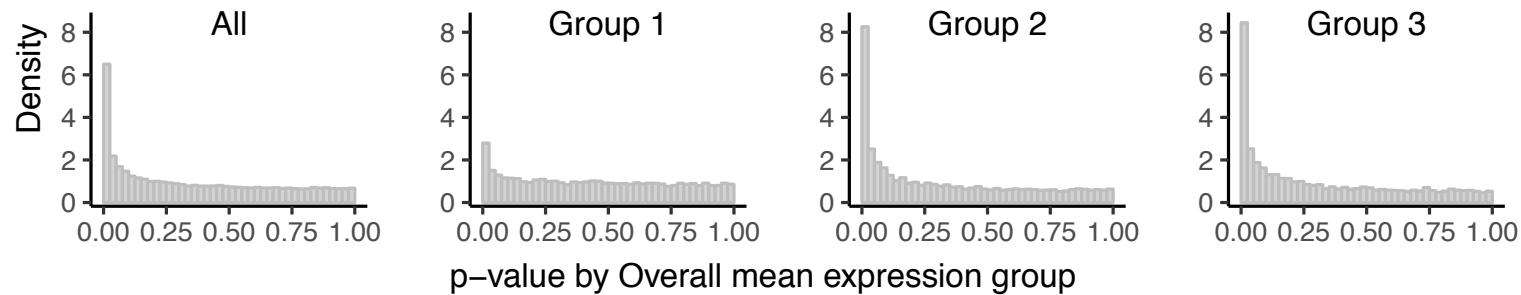


SummarizedBenchmark framework



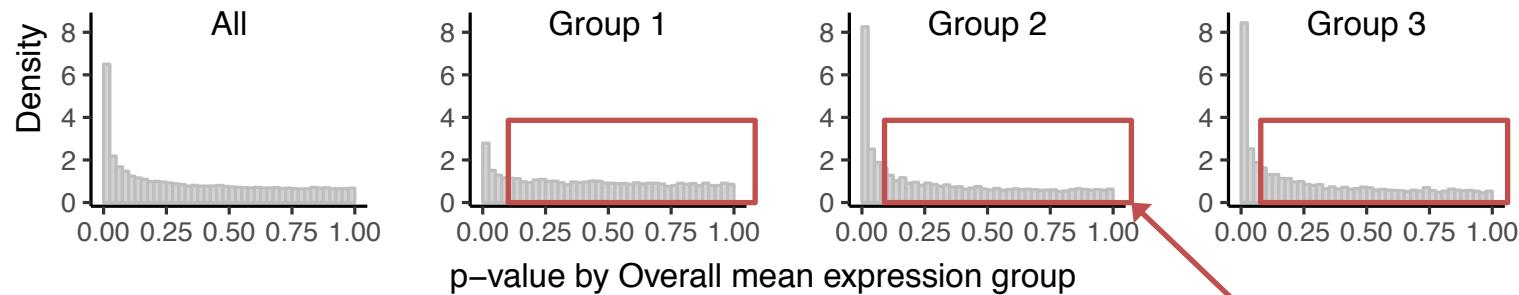
Independence and informativeness of covariates

RNA-seq case study: Brain



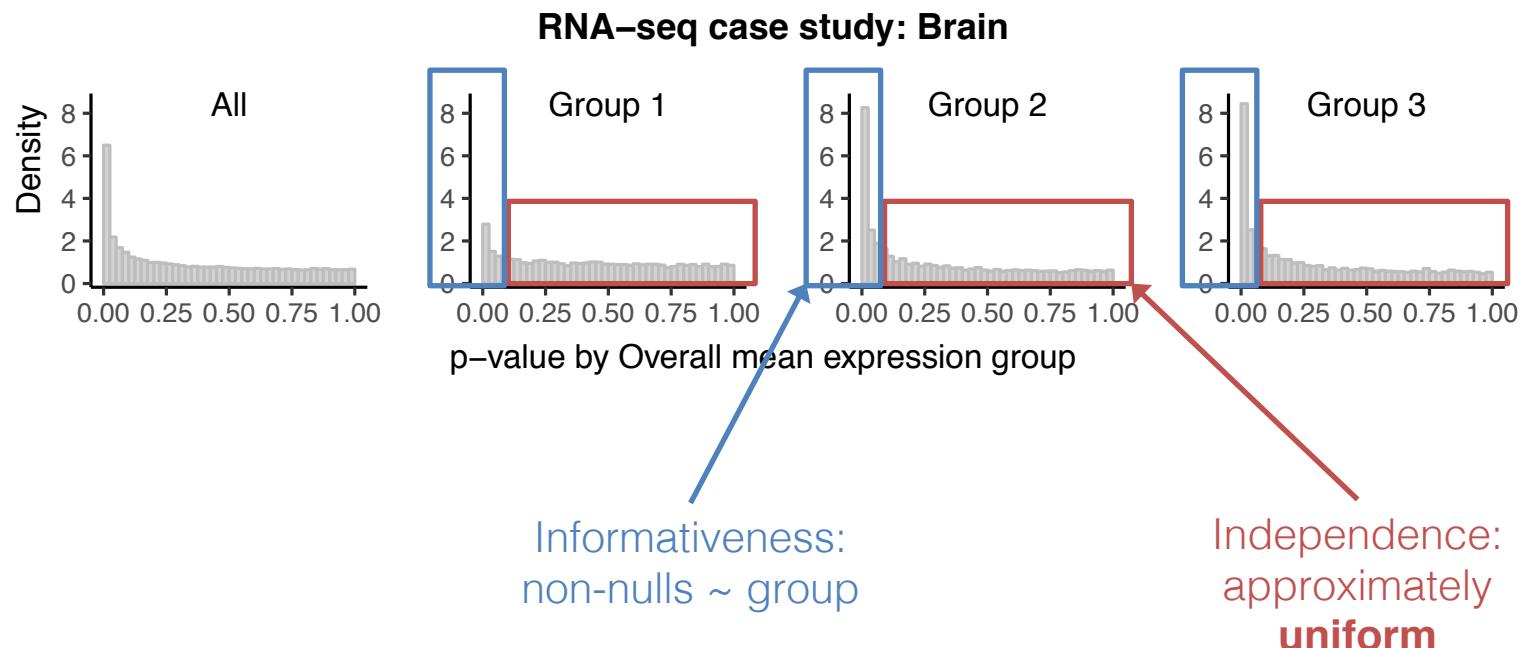
Independence and informativeness of covariates

RNA-seq case study: Brain



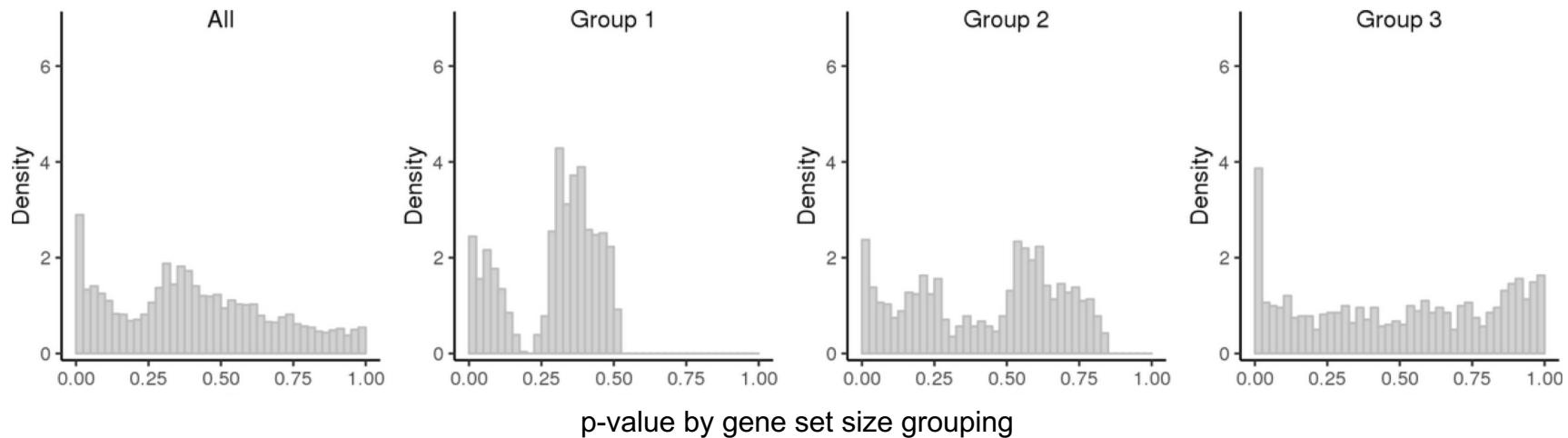
Independence:
approximately
uniform

Independence and informativeness of covariates



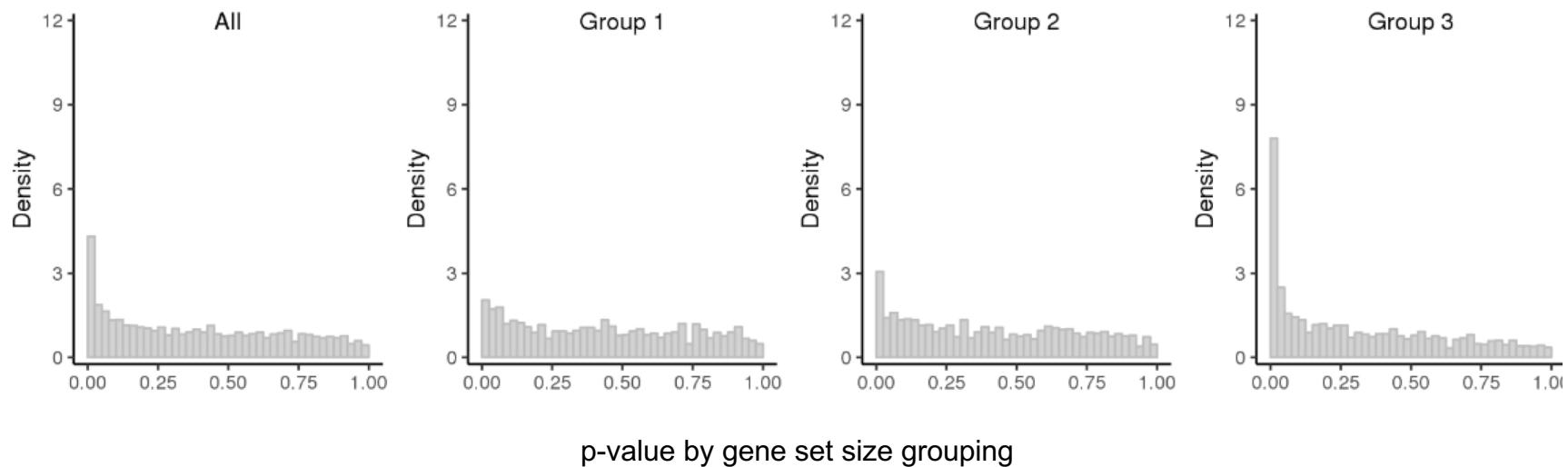
Gene set size is *not* independent for overrepresentation tests

`goseq` overrepresentation test p -value histograms



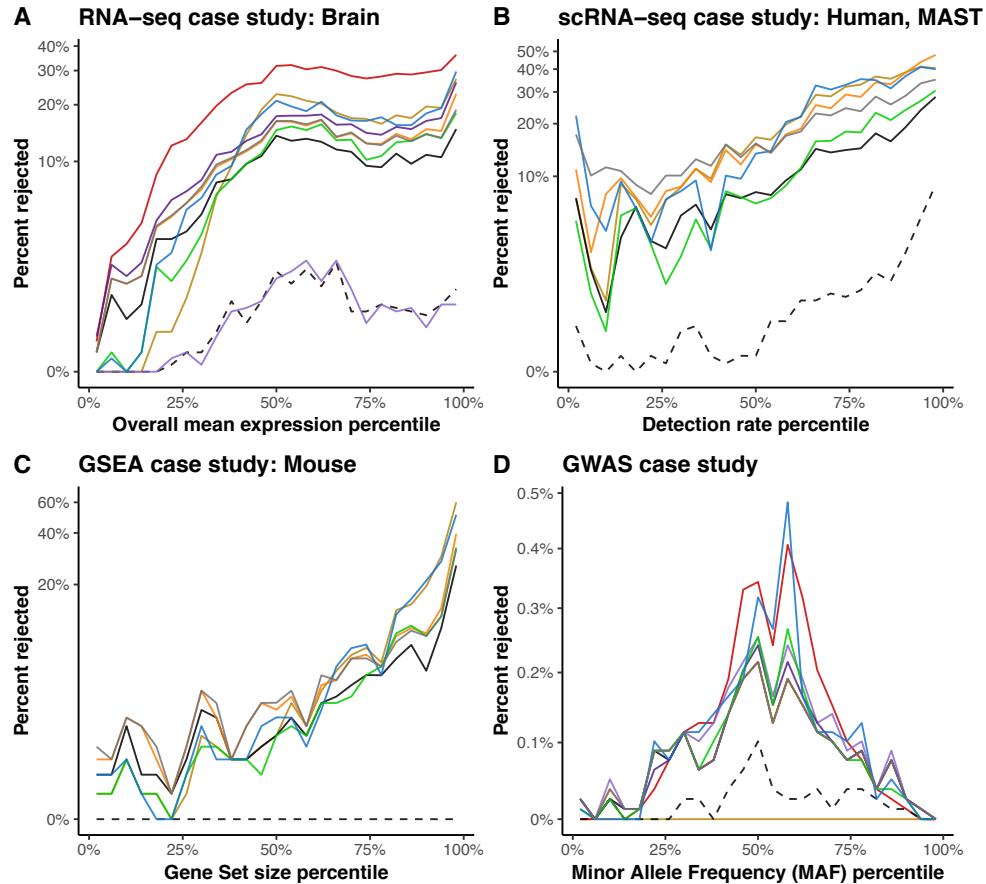
Same covariate is independent for GSEA

fgsea enrichment test p -value histograms



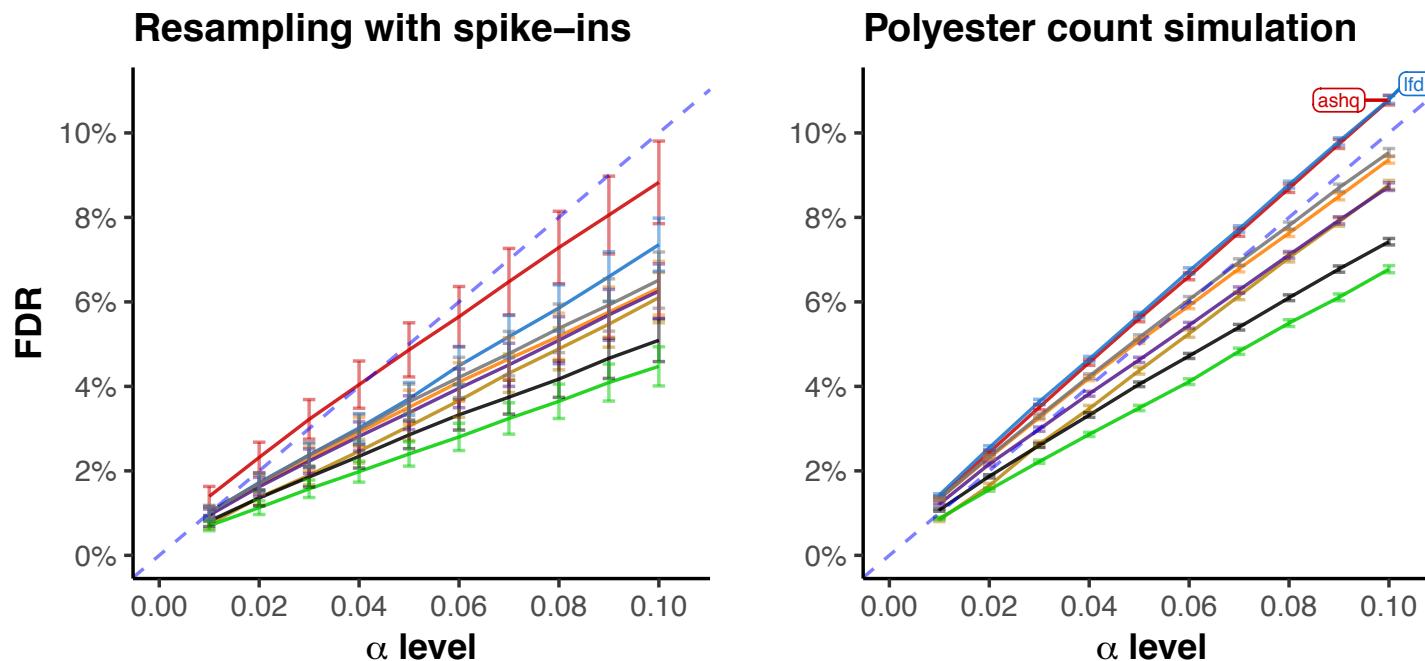
Informative covariates in case studies

Case Study	Covariate
Bulk RNA-seq	mean gene expression
Single Cell RNA-seq	mean non-zero gene expression, detection rate
Microbiome	mean non-zero abundance, ubiquity
ChIP-seq	mean read depth, window size
GWAS	minor allele frequency , sample size
Gene Set Analysis	gene set size



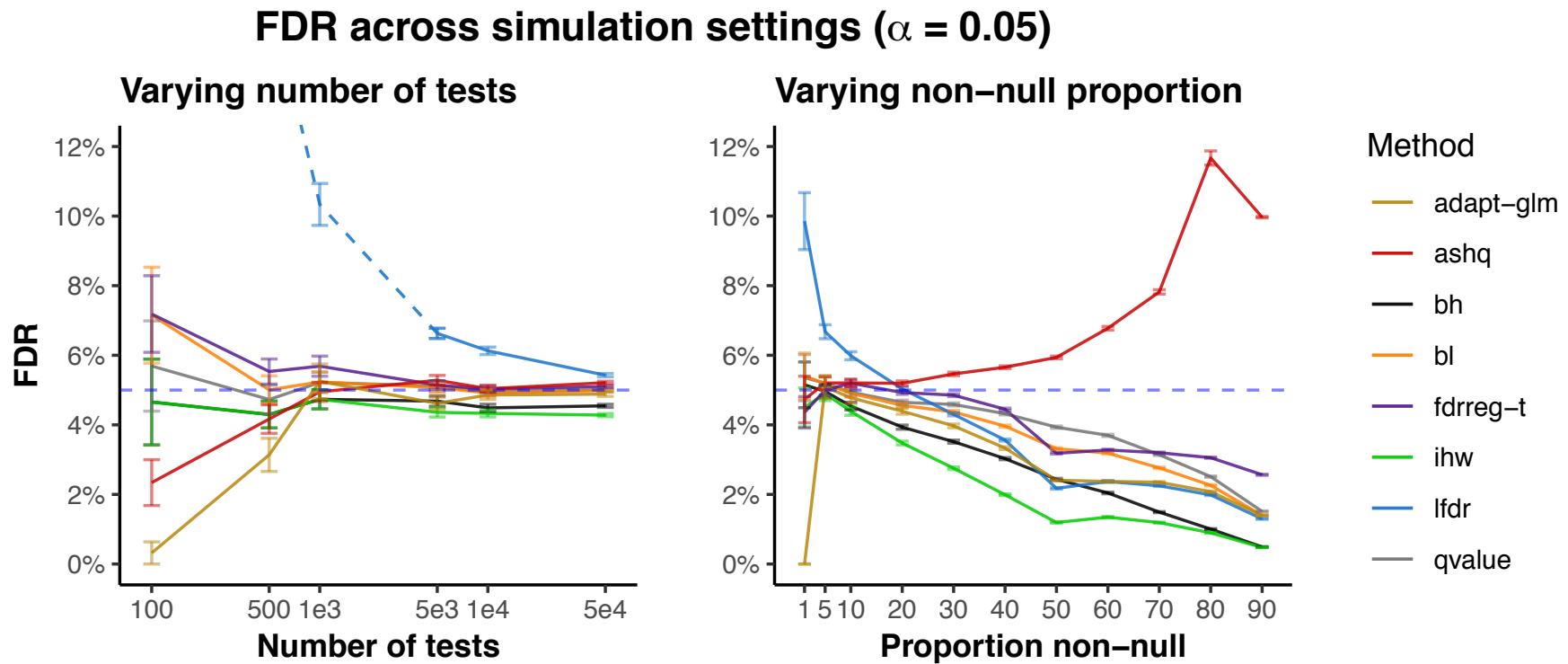
Most covariate-aware methods control FDR

FDR control in RNA-seq *in silico* experiments

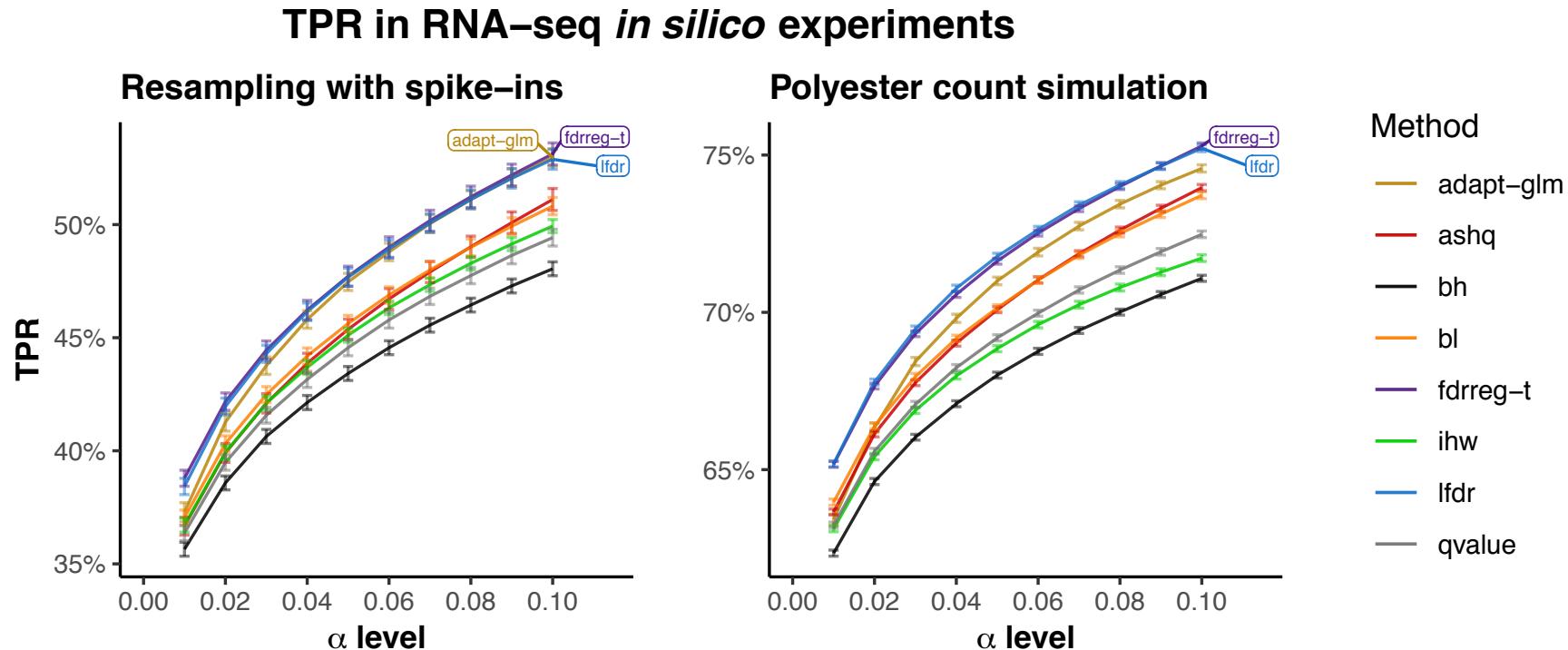


- Method
- adapt-glm
 - ashq
 - bh
 - bl
 - fdrreg-t
 - ihw
 - lfdr
 - qvalue

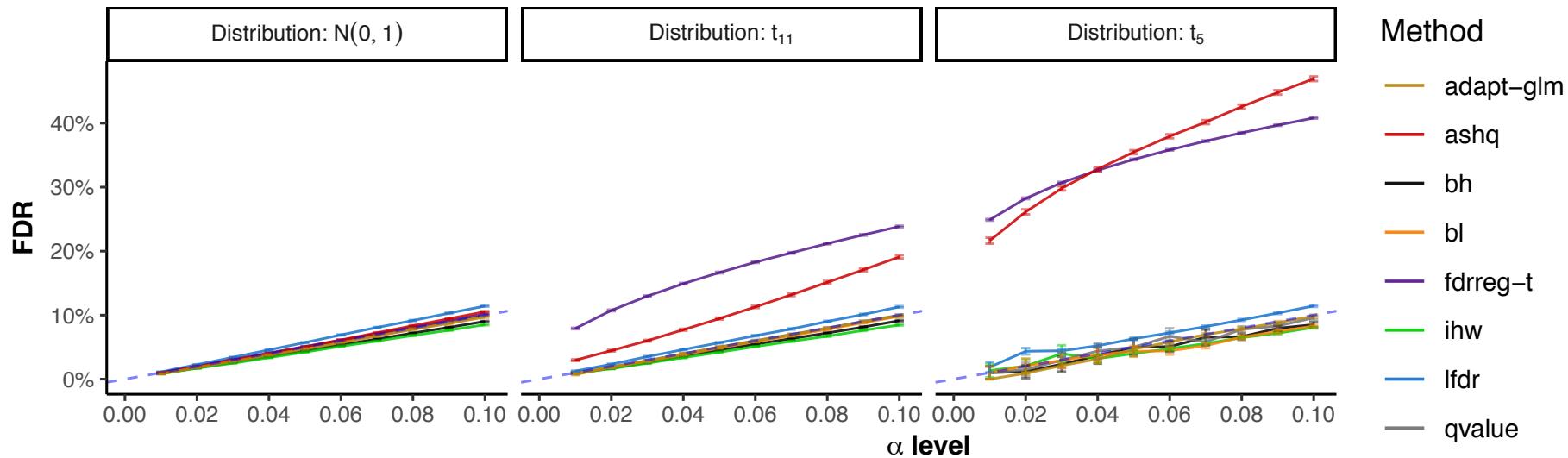
Some methods were sensitive to number of tests or null proportion



Covariate-aware methods were modestly more powerful

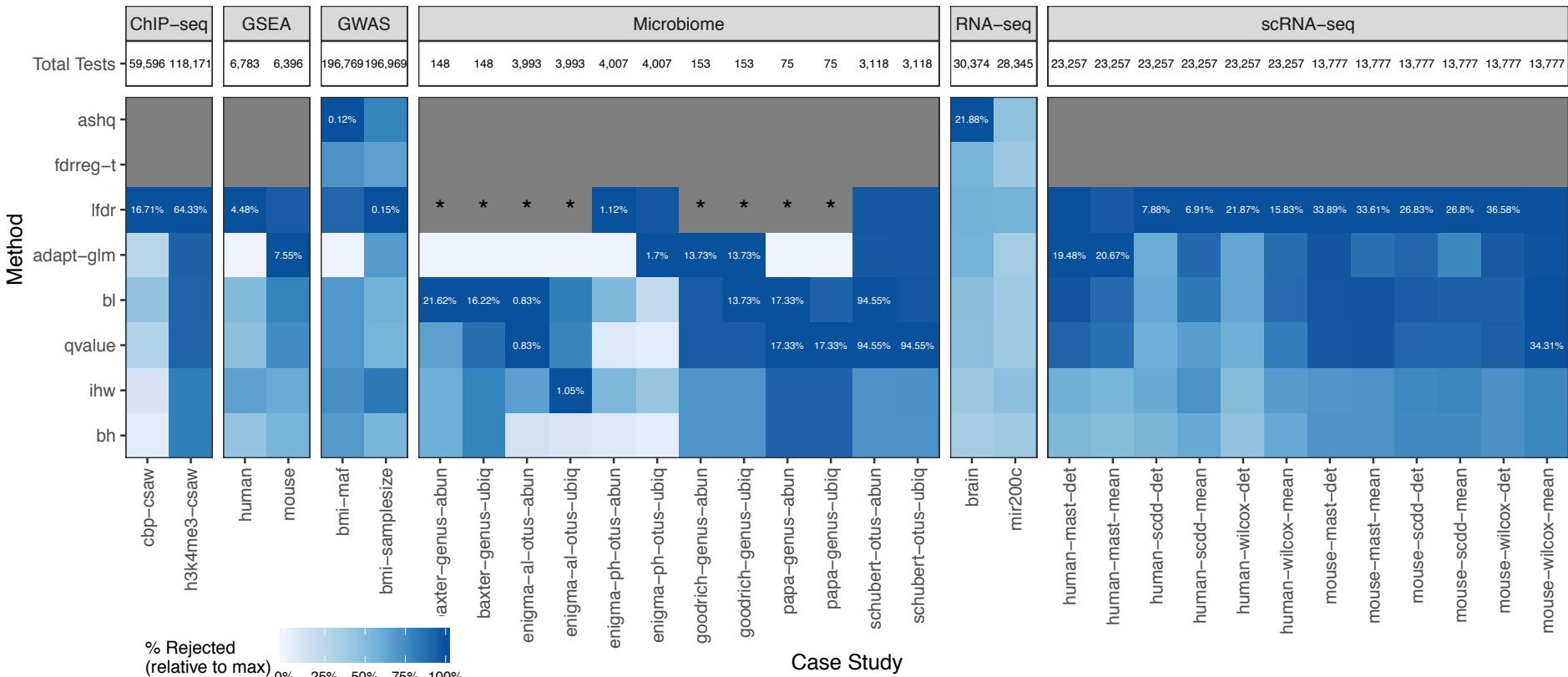


Some methods were sensitive to test statistic distribution

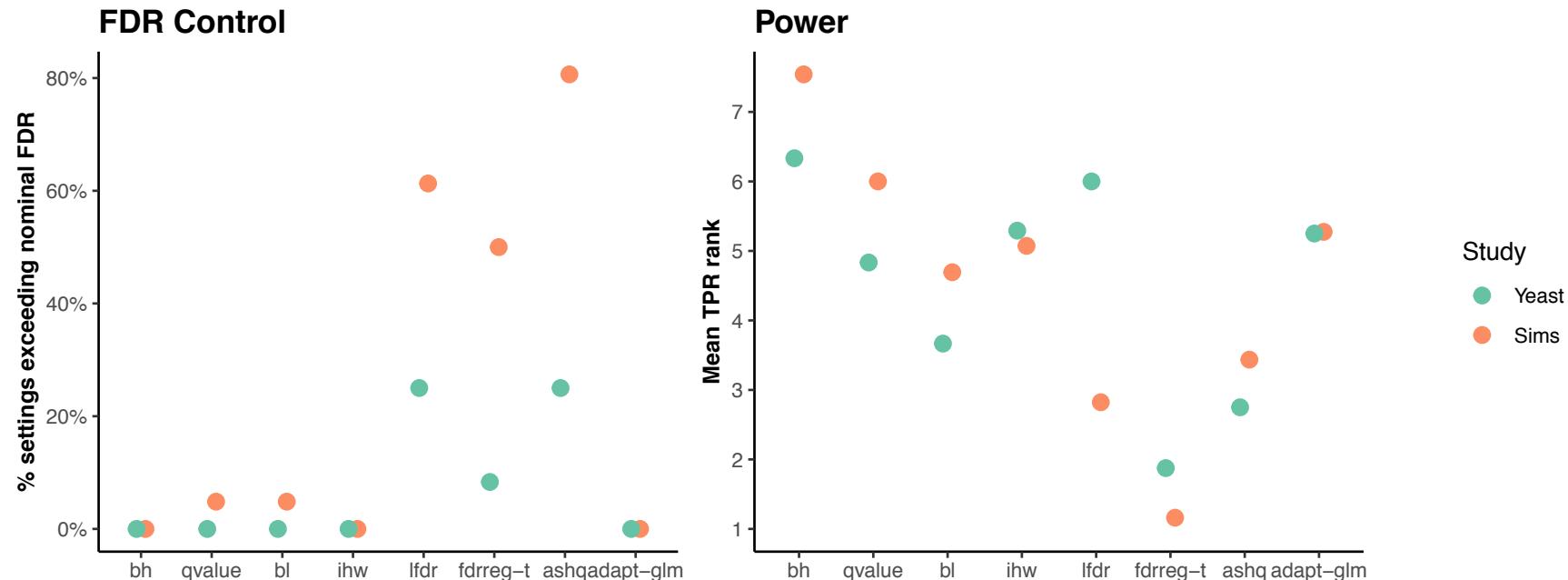


Not all methods could be applied to every case study

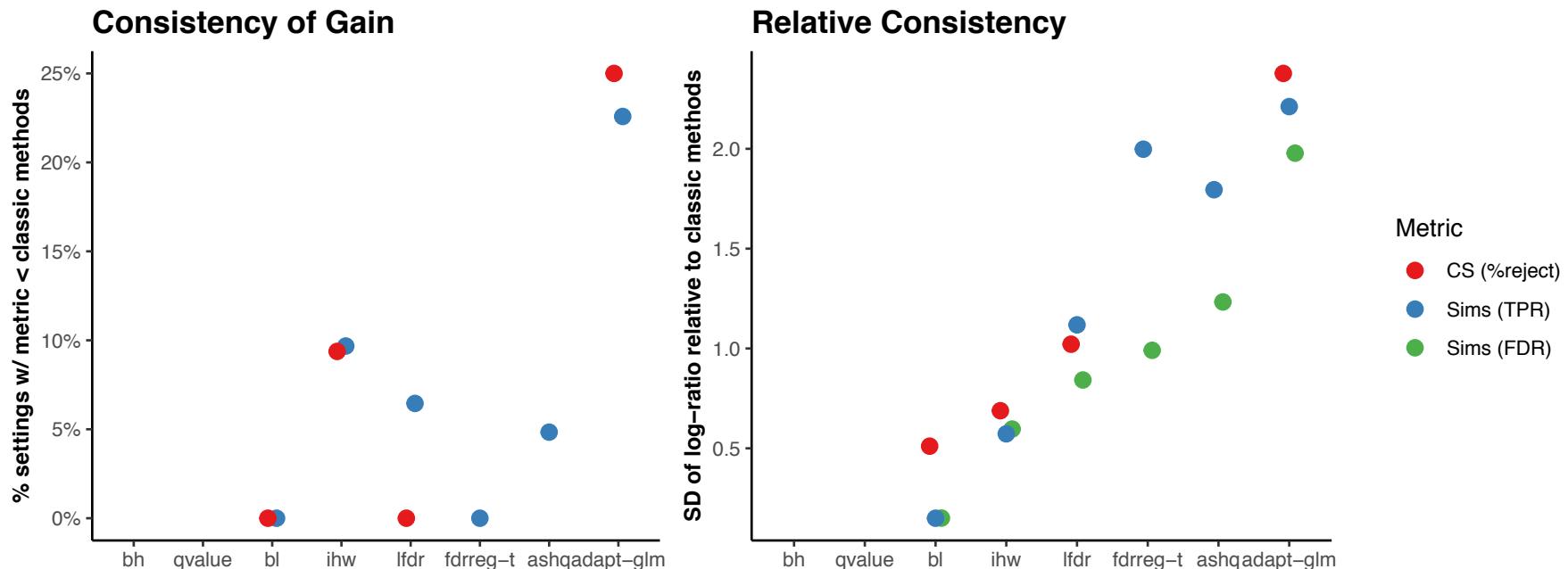
Number of rejections in case studies



Summary of FDR control and power across simulations

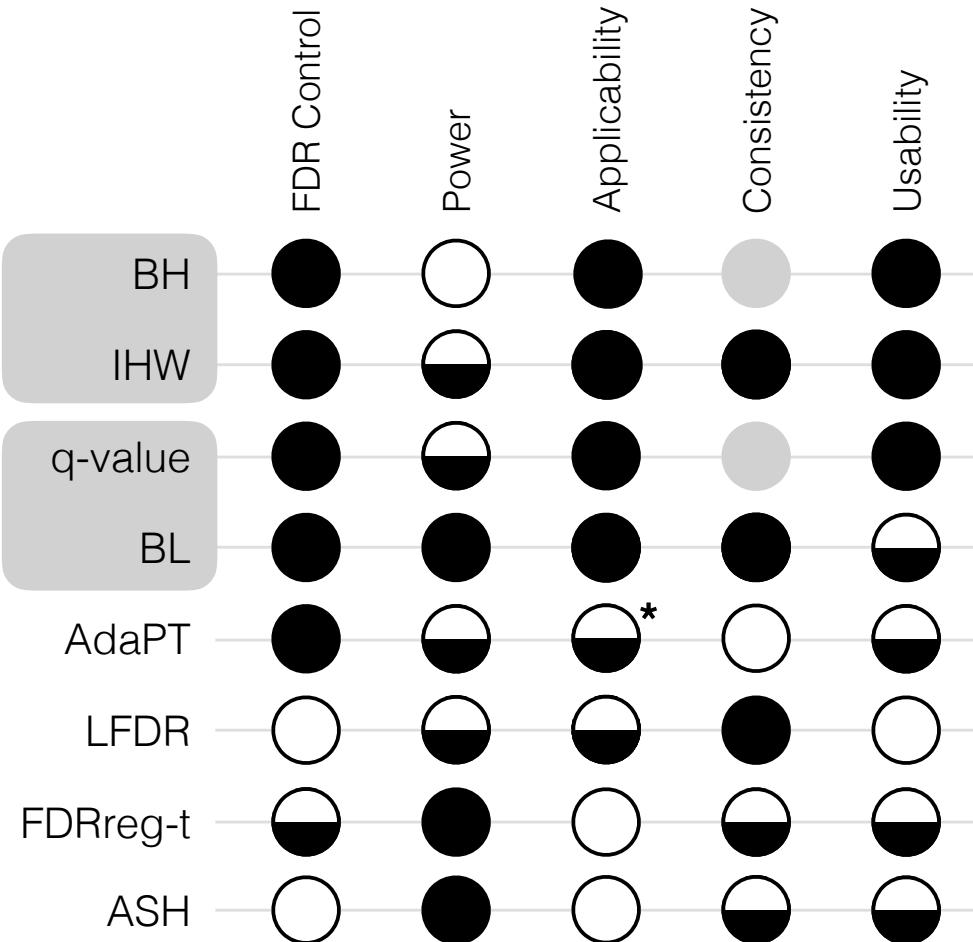


Gains relative to classic methods varied across methods



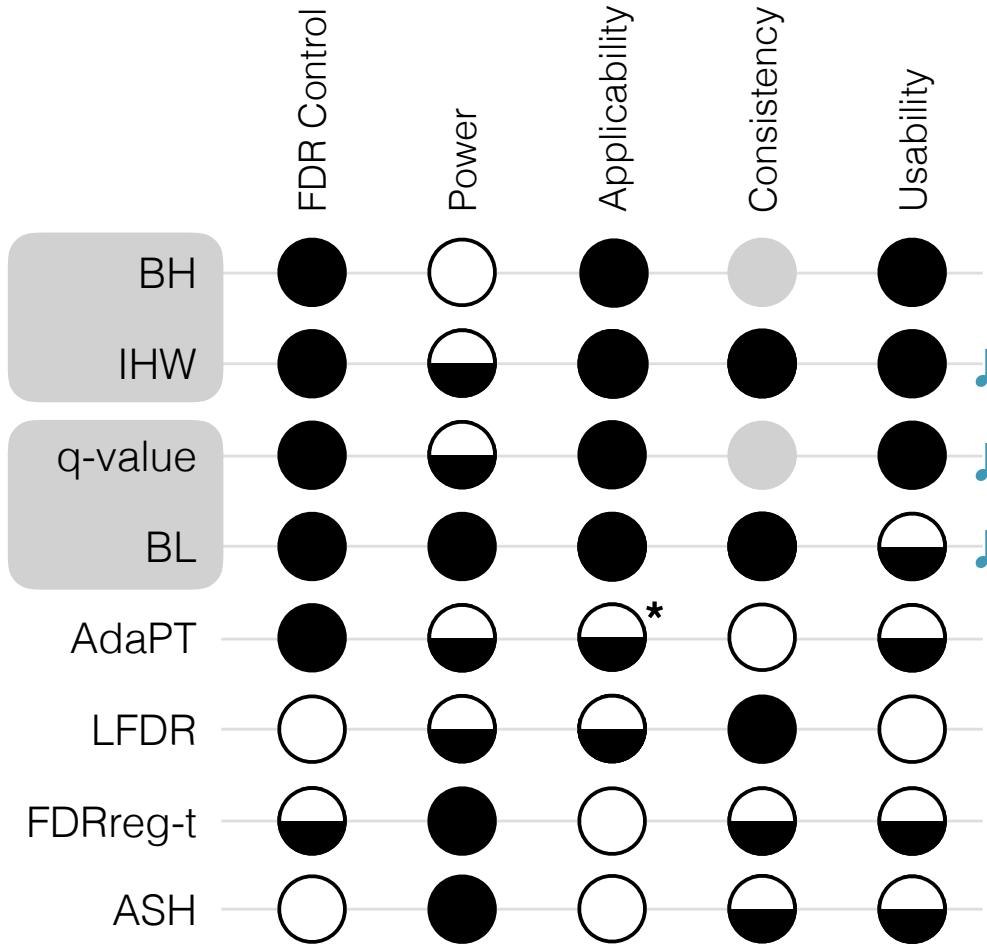
Recommendation summary

- Many covariate-aware methods provide consistent FDR control (**IHW**, **BL**, **AdaPT**)
- Gains in power achieved by covariate-aware methods are typically modest
- Not all methods could be applied to all simulations and case studies (**FDRreg**, **ASH**)
- Some methods showed highly variable performance across simulations and case studies (**AdaPT**)



Recommendation summary

- Many covariate-aware methods provide consistent FDR control (**IHW**, **BL**, **AdaPT**)
- Gains in power achieved by covariate-aware methods are typically modest
- Not all methods could be applied to all simulations and case studies (**FDRreg**, **ASH**)
- Some methods showed highly variable performance across simulations and case studies (**AdaPT**)
- **Some software implementations were more user-friendly than others**



Detailed case study & simulation reports

The screenshot shows a web browser window with the URL <https://www.pkimes.com/benchmark-fdr-html/>. The page title is "Additional files for *A practical guide to methods controlling false discoveries in computational biology*". Below the title, the authors are listed: Keegan Korthauer, Patrick K Kimes, Claire Duvallet, Alejandro Reyes, Ayshwarya Subramanian, Mingxiang Teng, Chinmay Shukla, Eric J Alm, and Stephanie C Hicks. A brief description follows: "This repository contains the knitted RMarkdown vignettes for simulations and case studies described in *A practical guide to methods controlling false discoveries in computational biology*." The main content is organized into sections for "Yeast *in silico* experiments" and "Polyester *in silico* experiments".

Yeast *in silico* experiments

- Additional file 2 - **Yeast *in silico* experiments I.**
 - Analysis and benchmarking results under the null, and using a unimodal alternative effect size distribution and large proportion (30%) of non-nulls using yeast RNA-seq data.
- Additional file 3 - **Yeast *in silico* experiments II.**
 - Analysis and benchmarking results using a unimodal alternative effect size distribution and small proportion (7.5%) of non-nulls using yeast RNA-seq data.
- Additional file 4 - **Yeast *in silico* experiments III.**
 - Analysis and benchmarking results using a bimodal alternative effect size distribution and large proportion (30%) of non-nulls using yeast RNA-seq data.
- Additional file 5 - **Yeast *in silico* experiments IV.**
 - Analysis and benchmarking results using a bimodal alternative effect size distribution and small proportion (7.5%) of non-nulls using yeast RNA-seq data.

Polyester *in silico* experiments

- Additional file 6 - **Polyester *in silico* experiments.**

Detailed case study & simulation reports

The screenshot shows a web browser with two tabs open. The left tab is titled "Additional files for A practical guide to methods" and contains a sidebar with sections for Yeast, Polymers, and other topics, each with a list of "Additional files". The right tab is titled "Case Study: Gene Set Enrichment (Mouse Data Set)" and displays the main content of the case study.

Case Study: Gene Set Enrichment Analysis (Mouse Data Set)

Alejandro Reyes and Keegan Korthauer
October 30, 2018

1 Summary

The objective of this vignette is to test different multiple testing methods in the context of Gene Set Enrichment Analysis (GSEA). To do this, we use data from the paper by Cabezas-Wallscheid et al. (Cell stem Cell, 2014). The data consist of RNA-seq data from mouse hematopoietic stem cells and multipotent progenitor lineages. The raw fastq data is available through the ArrayExpress database (<http://www.ebi.ac.uk/arrayexpress>) under accession number E-MTAB-2262. These data were mapped to the mouse reference genome GRCm38 (ENSEMBL release 69) using the Genomic Short-Read Nucleotide Alignment program (version 2012-07-20). We used htseq-count to count the number of reads overlapping with each gene and used the DESeq2 package to format the data as a DESeqDataSet R object.

Here we use the `fgsea` Bioconductor package to implement the GSEA method. This is a Functional Class Scoring approach, which does not require setting an arbitrary threshold for Differential Expression, but instead relies on the gene's rank (here we rank by DESeq2 test statistic).

2 Workspace Setup

```
library(dplyr)  
  
##  
## Attaching package: 'dplyr'
```

Acknowledgements



DANA-FARBER
CANCER INSTITUTE



HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Rafael Irizarry

Patrick Kimes*

Stephanie Hicks

Claire Duvallet

Alejandro Reyes

Ayshwarya Subramanian

Mingxiang Teng

Chinmay Shukla

Eric Alm



Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

keegan@jimmy.harvard.edu

@keegankorthauer

kkorthauer.org

