Properties of machine learning and FDRs for discovery in large scale data

Megan Hollister Murray

Department of Biostatistics
Vanderbilt University

October 22nd, 2020

Table of Contents

Introduction

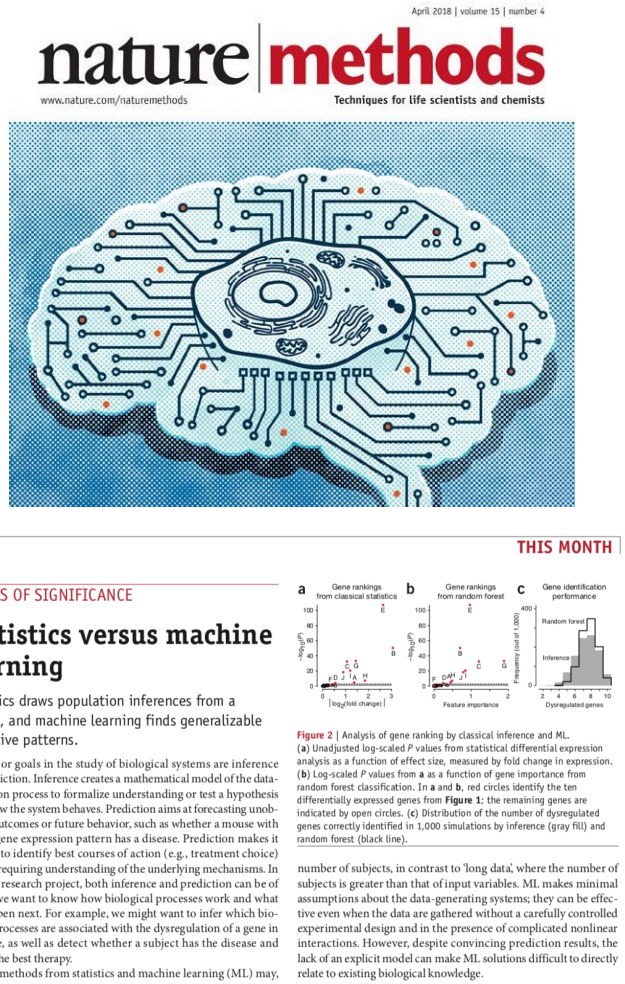
Two projects are included in this presentation:

- Machine Learning and Multiple Testing
  - Presented at ENAR 2019
  - Machine Learning
  - Traditional p-value methods
  - Second-generation p-values
  - Discovery in large-scale data
- False discovery rates
  - Estimation vs. control
  - Limitations with `stats::p.adjust`
  - R package `FDRestimation`

Table of Contents

Machine Learning and Multiple Testing

Background



- A paper in April 2018 *Nature Methods* on statistical discovery in large-scale data
- Concluded random forests outperformed Benjamini-Hochberg p-value based approaches
- Based on simulations of dysregulated genes in expression data
- Not all approaches were given the same a priori information

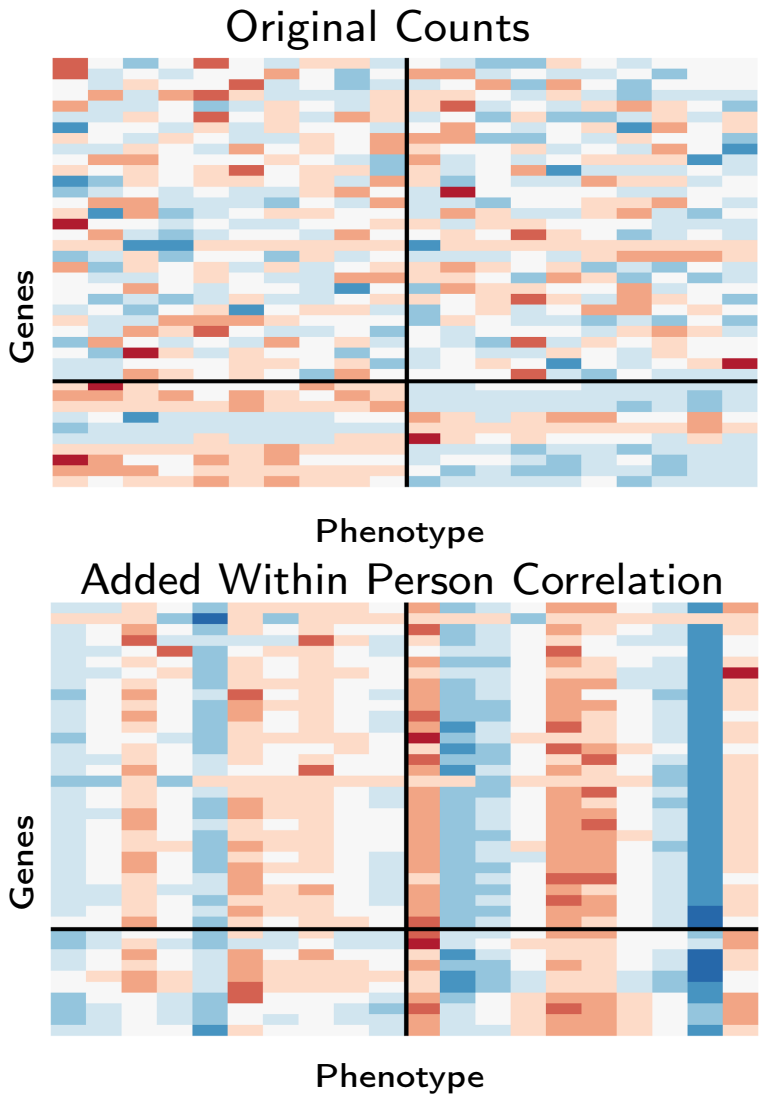
Machine Learning and Multiple Testing
Goals

$\rightarrow$ Paper received much press and substantial twitter discussion

Objectives:

1. Examine claims using unbiased and fair comparisons
2. Estimate accuracy of machine learning and "traditional" methods
3. Identify methods with the best performance characteristics

Simulated Gene Expression Data



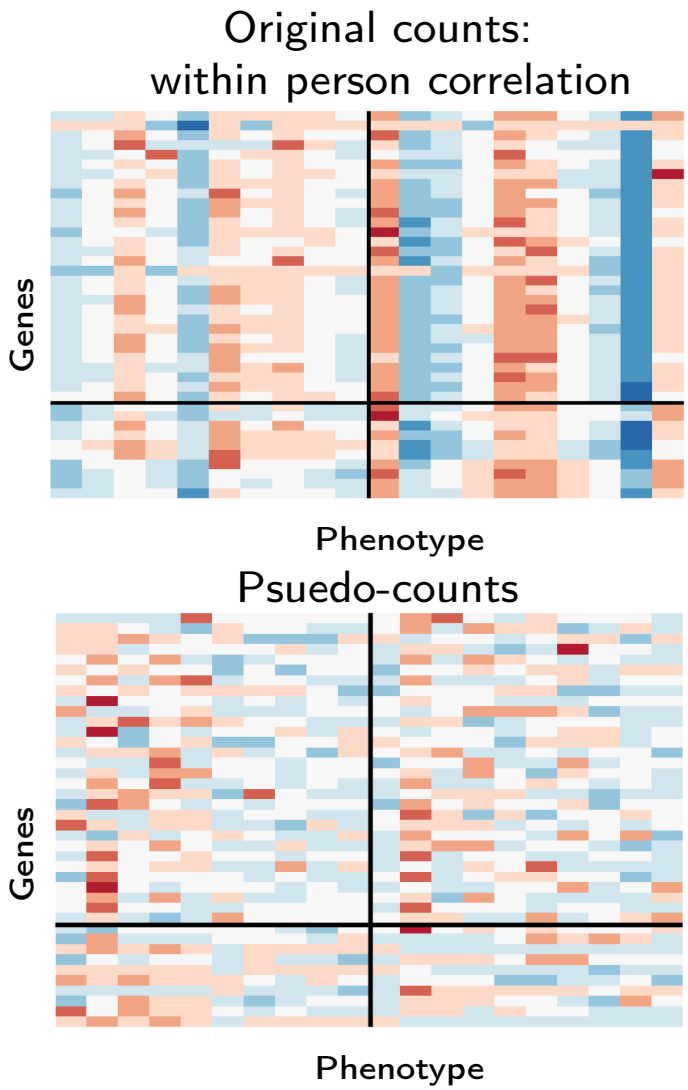
- 40 genes ; 20 people
- 10 phenotype positive ; 10 negative
- 25% (10) of genes are "dysregulated" across phenotype
- Computed pseudo-counts

- *Allowed within person correlation across genes (new)*

Algorithm 1:

Result: Simulated RNA-seq counts

1. Generate the observed counts for each gene by sampling from a Poisson distribution. Counts $\sim Pois(\lambda)$
2. Compute the mean gene expression $\lambda = \exp(\alpha_i + \mathrm{I}_{\textbf{positive}}\beta_i + \epsilon_{ij} + \gamma_j)$
    1. For all 40 genes simulate log mean expression levels from $\alpha_i \sim N(4, 2)$
        - For the positive(+) phenotype include the addition of a standard normal to each mean expression $\beta_i \sim N(0, 1)$
    2. For each gene and person simulate the genetic variation $\epsilon_{ij} \sim N(0, 0.15)$
    3. OPTIONAL: For each person simulate the within-person correlation $\gamma_j \sim N(0, 1)$



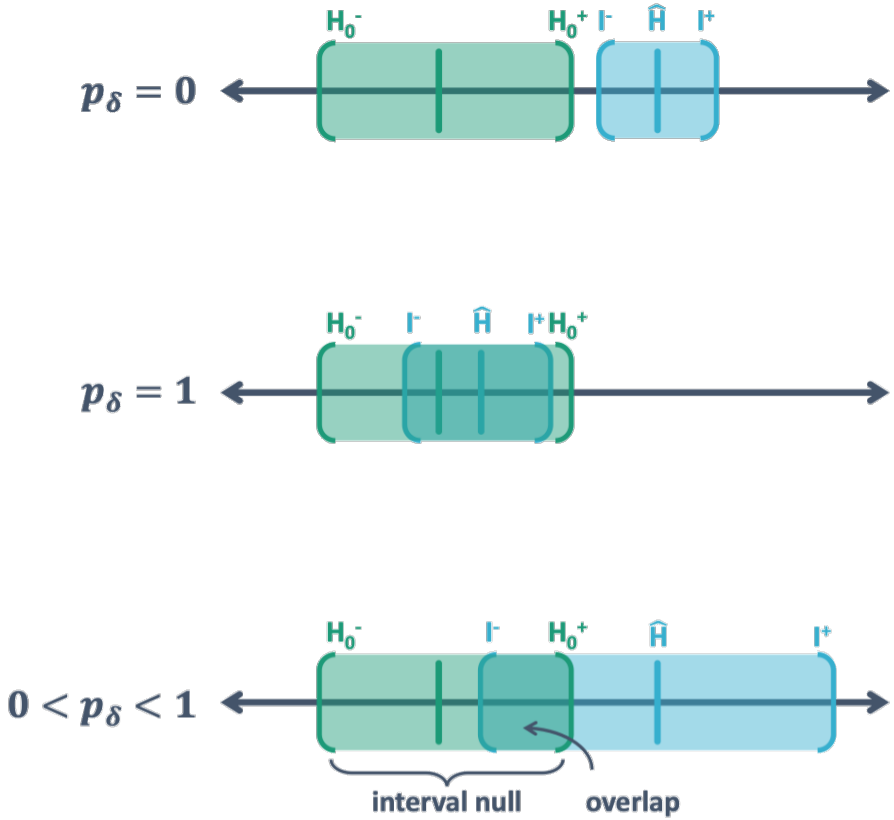
Pseudo-counts: "normalized counts"

- From edgeR package
  - Method of Robinson and Smyth (2008)
  - Poisson distribution is used to model RNA-seq counts
  - Accounts for overdispersion
  - Preserves differences between genes and variability within each gene

Traditional	**Machine Learning**
Nominal *p*-values	Random Forest importance levels
Bonferroni adjusted *p*-values	Neural Net prediction weights
Benjamini-Hochberg Emp FDRs	
Second-generation *p*-values	

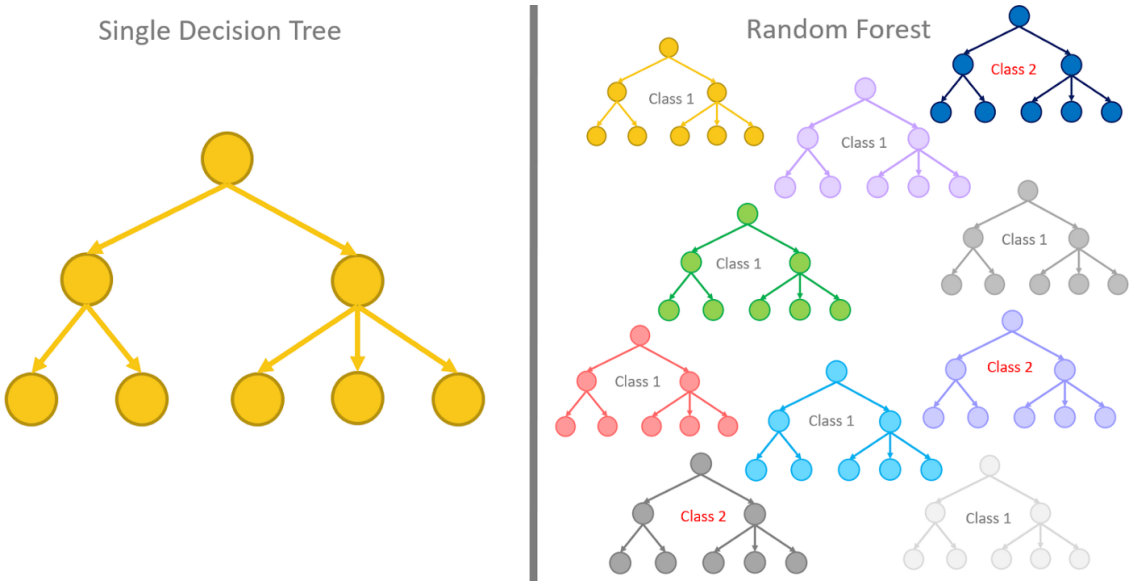
1. 5% significance level / FWER / FDR
2. Top 10 ranked genes by ML criteria
3. *Top 10 ranked genes by Traditional criteria (new)*

- SGPV is denoted by p_δ
- δ: interval null hypothesis
- The fraction of data-supported effect sizes that are null
- **Cases:**
  1. $p_\delta = 0$ when data incompatible with null region
  2. $p_\delta = 1$ when data compatible with null region
  3. $0 < p_\delta < 1$ when data are inconclusive



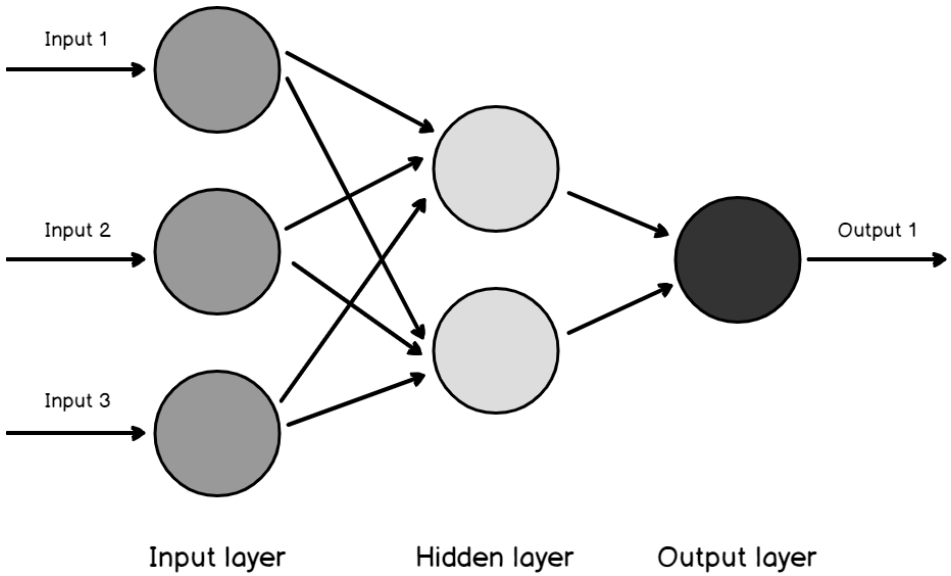
Random Forest importance levels

- Classification for phenotype with 100 trees
- Mean decrease in Gini index
- Quantifies a gene's contribution to the average classification when the tree is split



Neural Net prediction weights

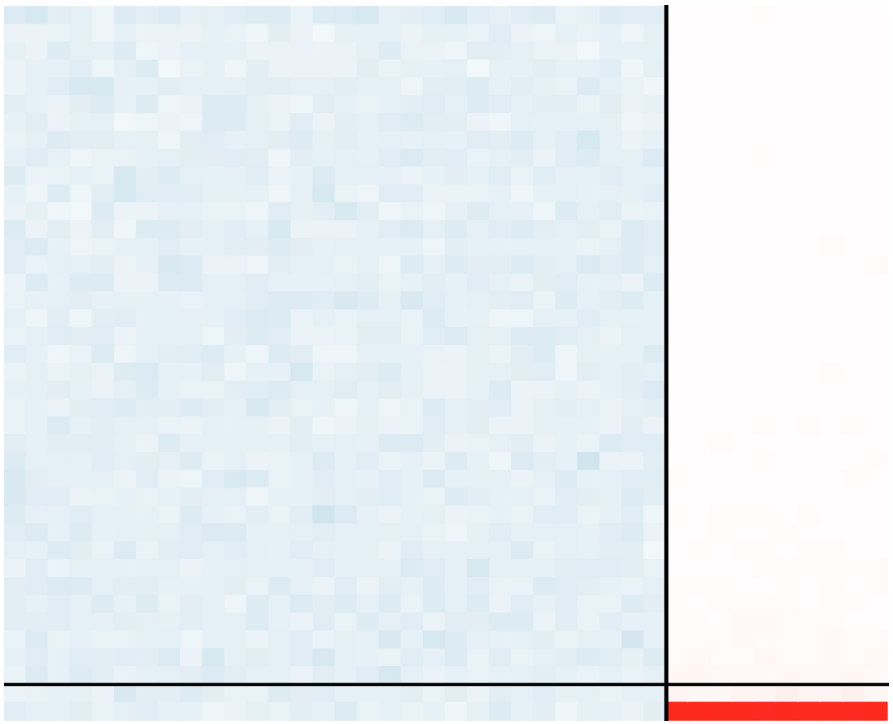
- Predict phenotype for each person using the 40 genes as predictors
- Method proposed by Garson 1991 identifies the relative importance of explanatory variables for response in a supervised neural network by deconstructing the model weights
- Used `gar.fun` function created by Marcus Beck



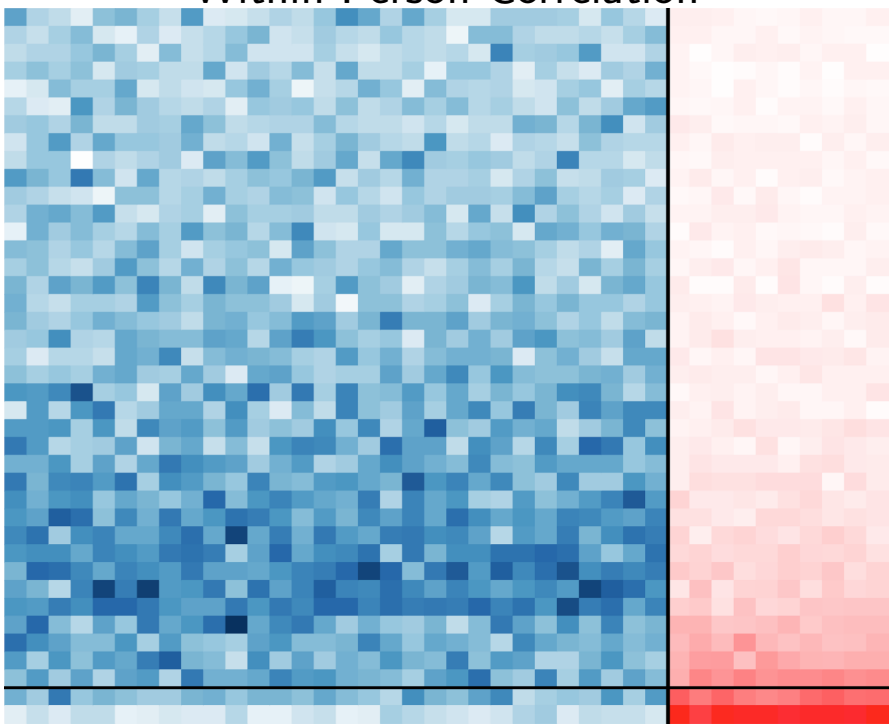
Results

- Heatmap of discovery for nominal p-values
- Values below horizontal line less than 0.05

Nominal p-values of Original Counts

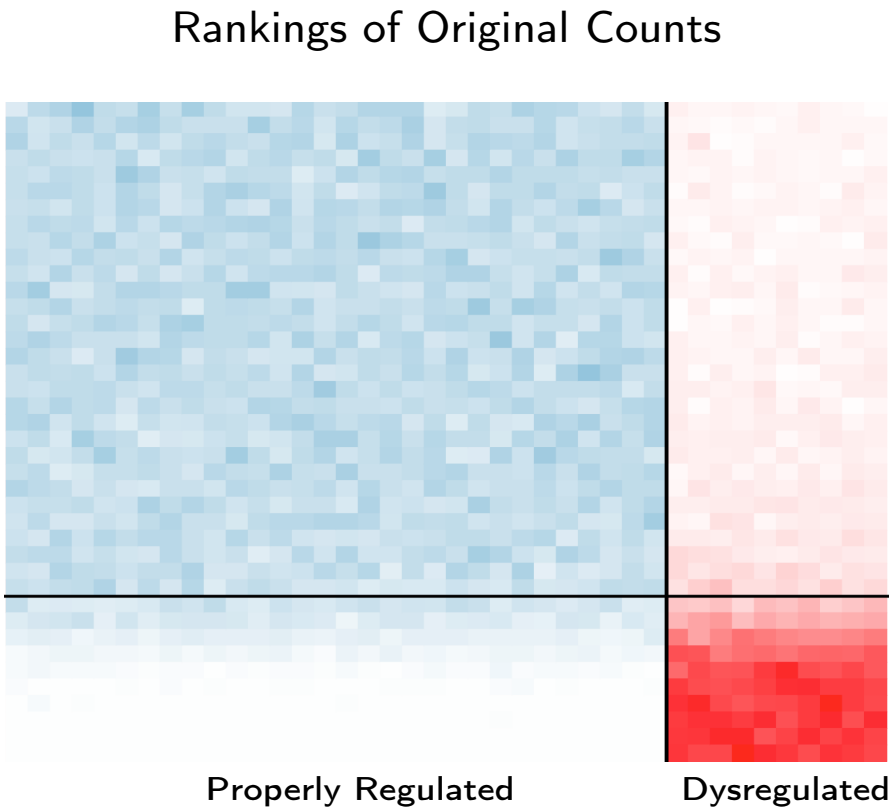


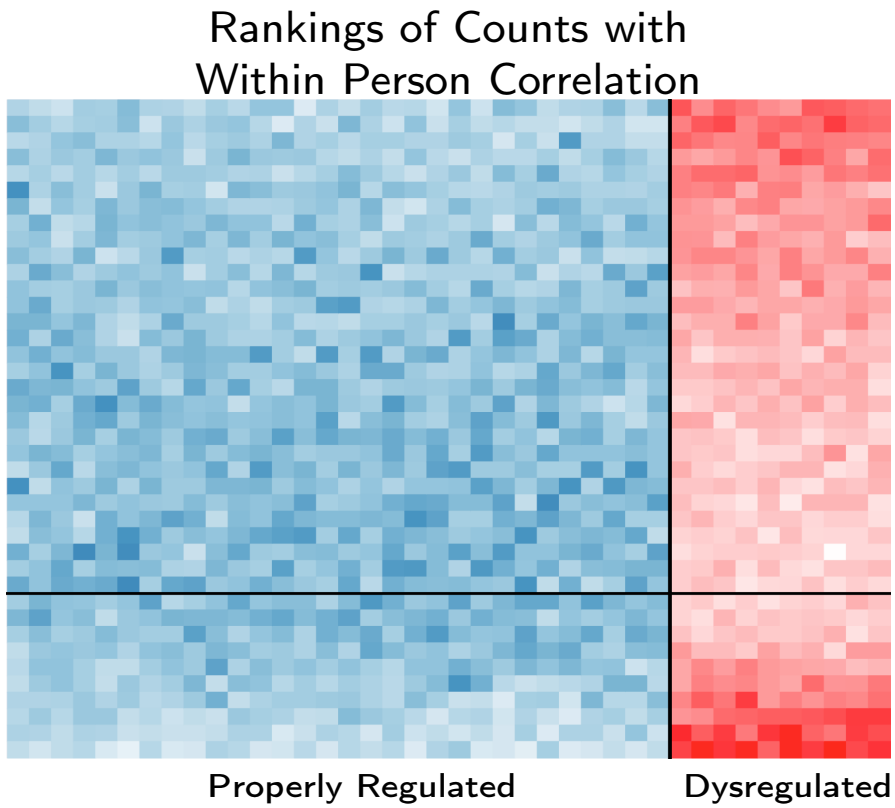
Nominal p-values of Counts with Within Person Correlation

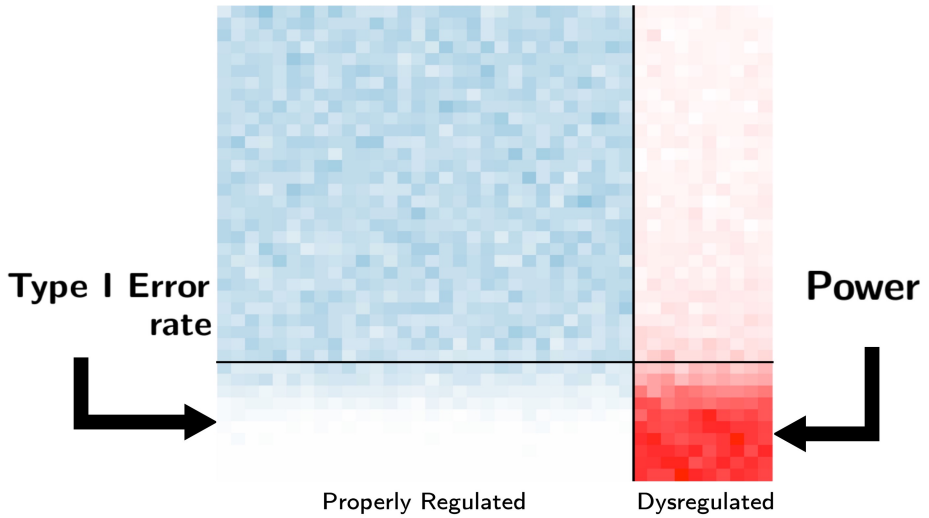


Heatmaps of Rankings

- Heatmap of gene rankings by FDR (Benjamini-Hochberg)
- Top 10 rankings below horizontal line





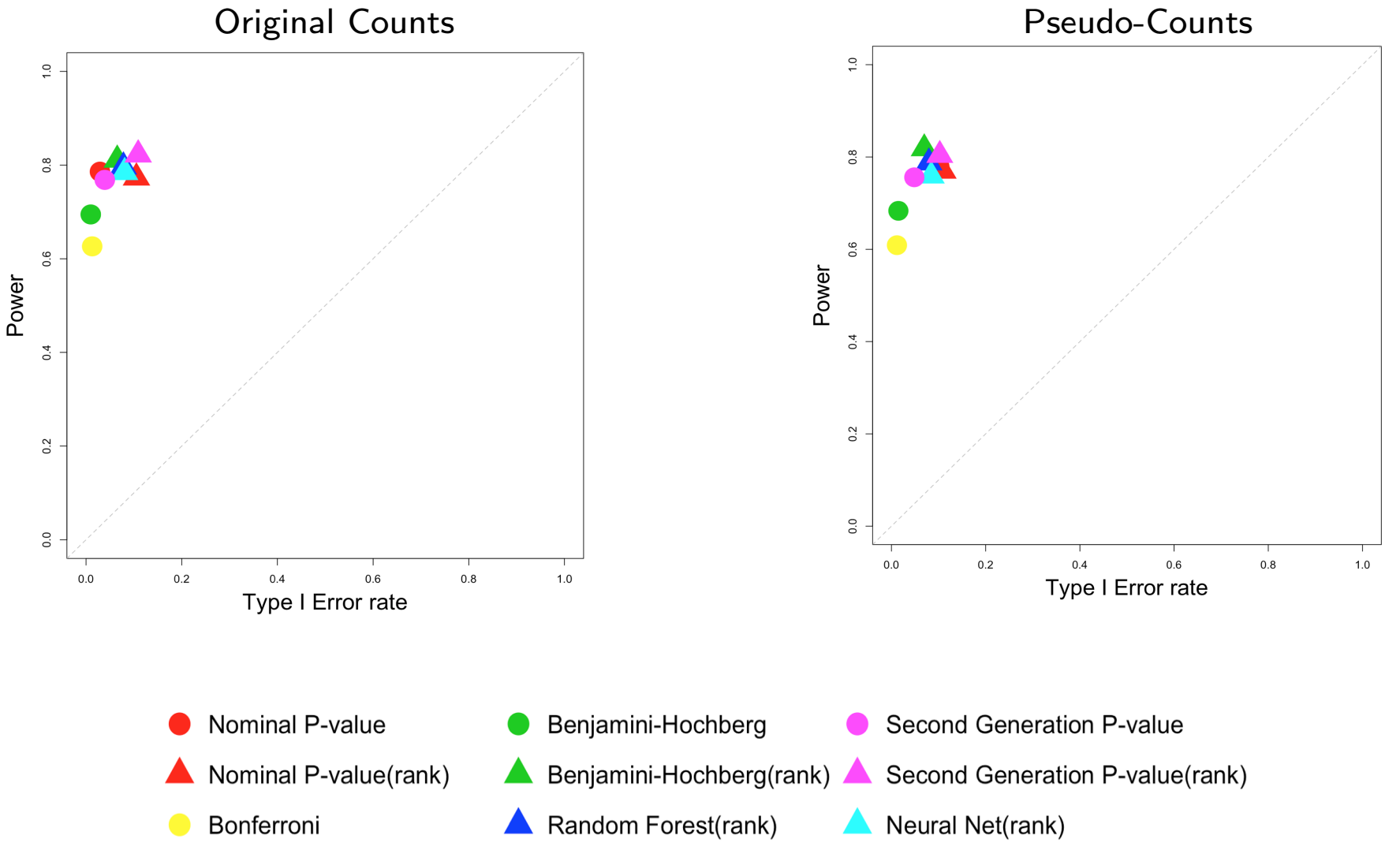


Accuracy statistics:

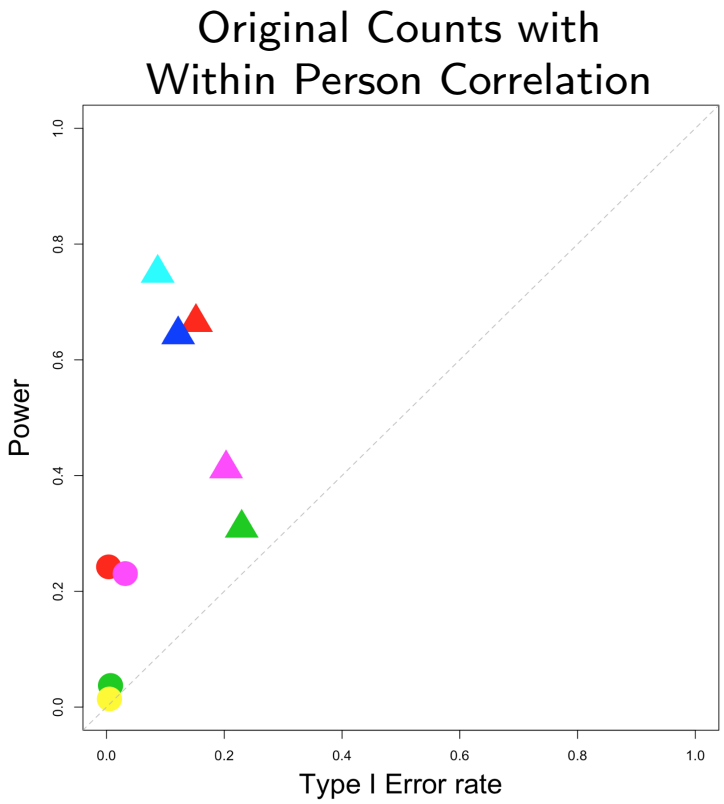
- **Power**
  - → Proportion of "dysregulated" genes identified as "dysregulated"
- **Type I Error rate**
  - → Proportion of "properly regulated" genes identified as "dysregulated"

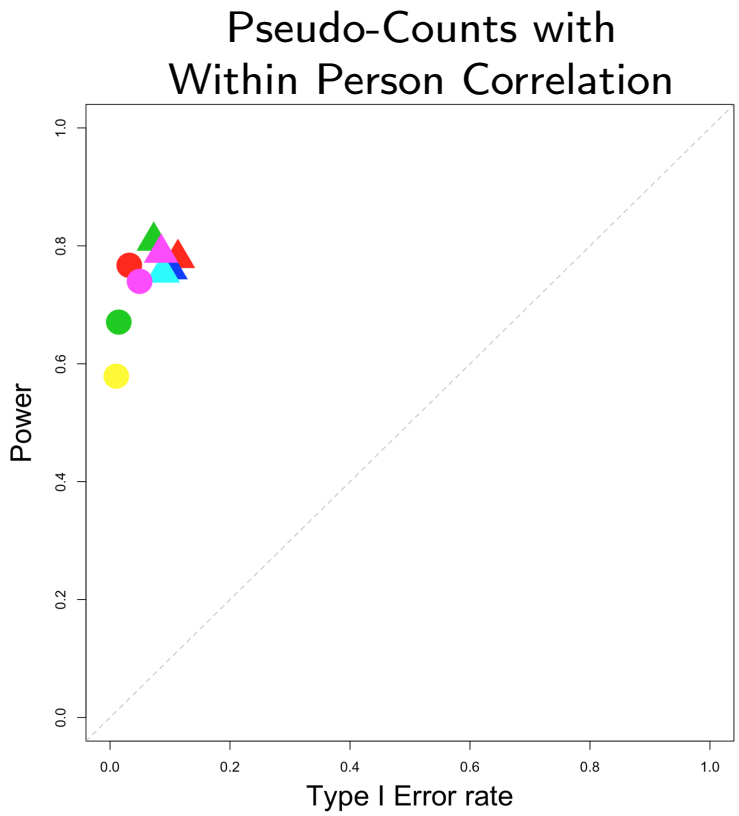
Machine Learning and Multiple Testing

Results



Results





Nominal P-value
Nominal P-value(rank)
Bonferroni
Benjamini-Hochberg
Benjamini-Hochberg(rank)
Random Forest(rank)
Second Generation P-value
Second Generation P-value(rank)
Neural Net(rank)

- Normalizing step is critical for some methods
- Methods perform identically when properly compared (by rankings)
- Comparing ranking vs threshold discovery gives *false* impression of differential statistical accuracy (ie. *Nature Methods*)

	Traditional Methods	**Machine Learning**
Pros	• Significance level criterion • Can be ranked • Interpretable coefficients	• Handles complexity with ease • Variety of flexible algorithms
Cons	• Complexity poses challenges • Significance criterion not universal • Models can be simplistic	• Must pre-specify number of findings • No threshold criterion • Coefficients hard to interpret

Table of Contents

False Discovery Rates

The performance of the ranked BH empirical FDRs and the use of `stats::p.adjust` motivated us to create our own package.

False discovery rates (FDRs)

- The propensity for an observed result to be mistaken
- Should accompany observed results
- Not always monotonic in p-value space
- Can control error rate (BH adjusted p-values)

Benjamini-Hochberg (BH) procedure:

Find the largest index, k, such that Equation (1) holds. Then all features with $p_{(1)}, ..., p_{(k)}$ are deemed interesting at the FDR γ threshold and considered "findings".

$$p_{(i)} \leq \gamma \frac{i}{m} \text{ for } i \in \{1, 2, ..., m\} \tag{1}$$

BH adjusted p-value:

$$\tilde{p}_{(i)} := \min_{j \geq i} \left(\frac{p_{(j)} m}{j} \right) \leq \gamma \tag{2}$$

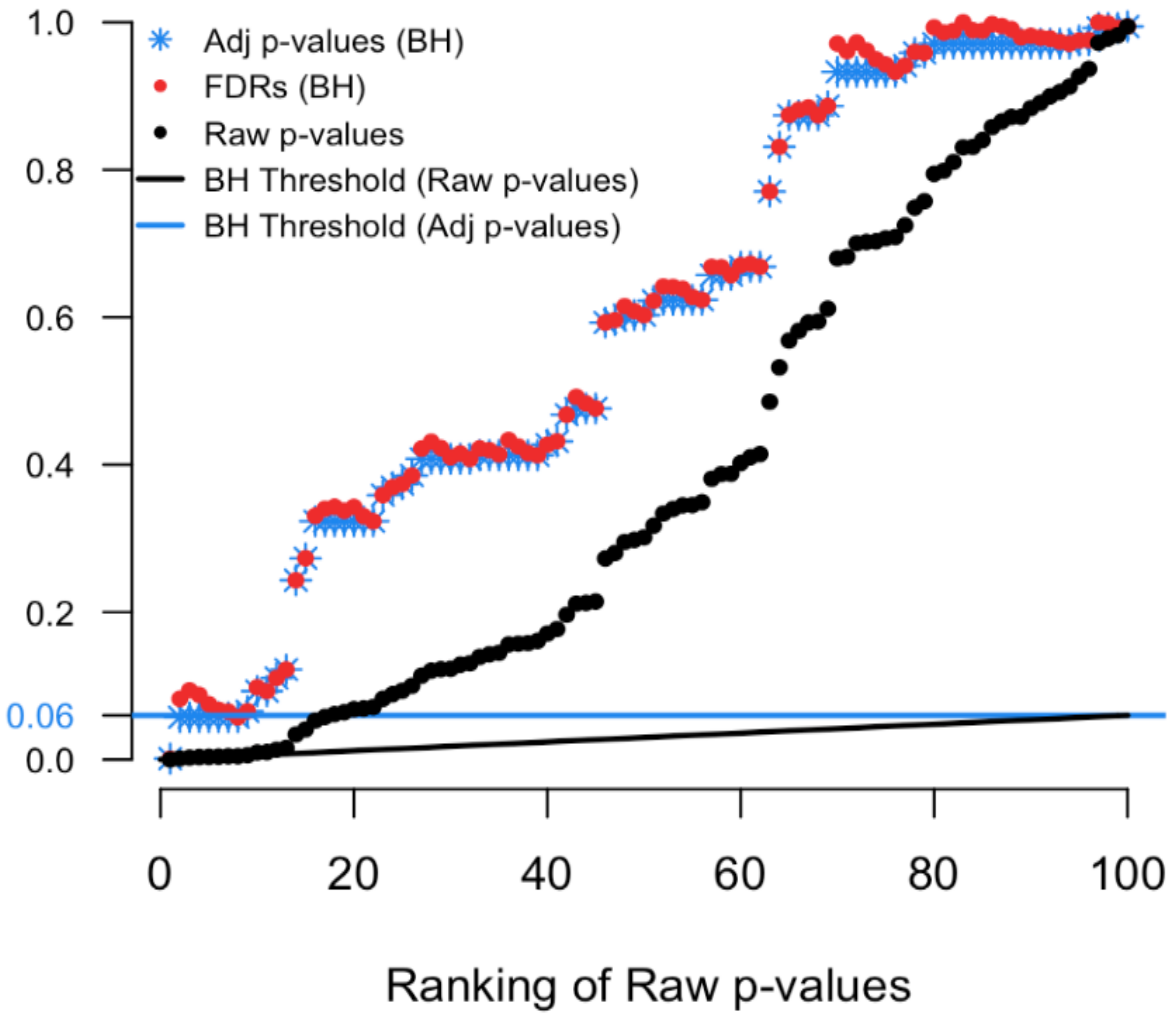
BH FDR:

$$FDR_i := \frac{p_i m}{\text{rank}(p_i)} \cdot \hat{\pi}_0 \tag{3}$$

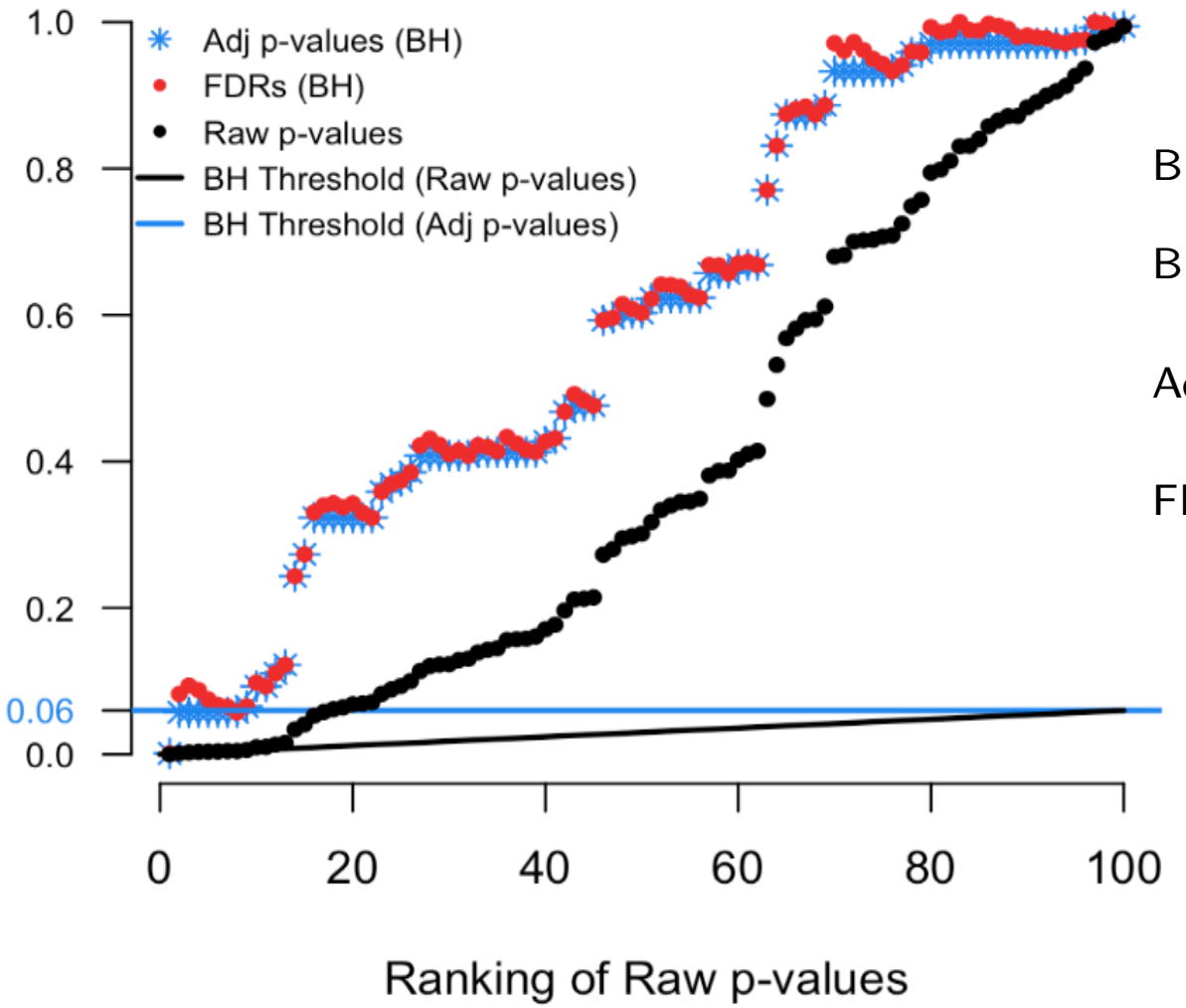
False Discovery Rates

p-value Based Methods

Simulated example



Simulated example



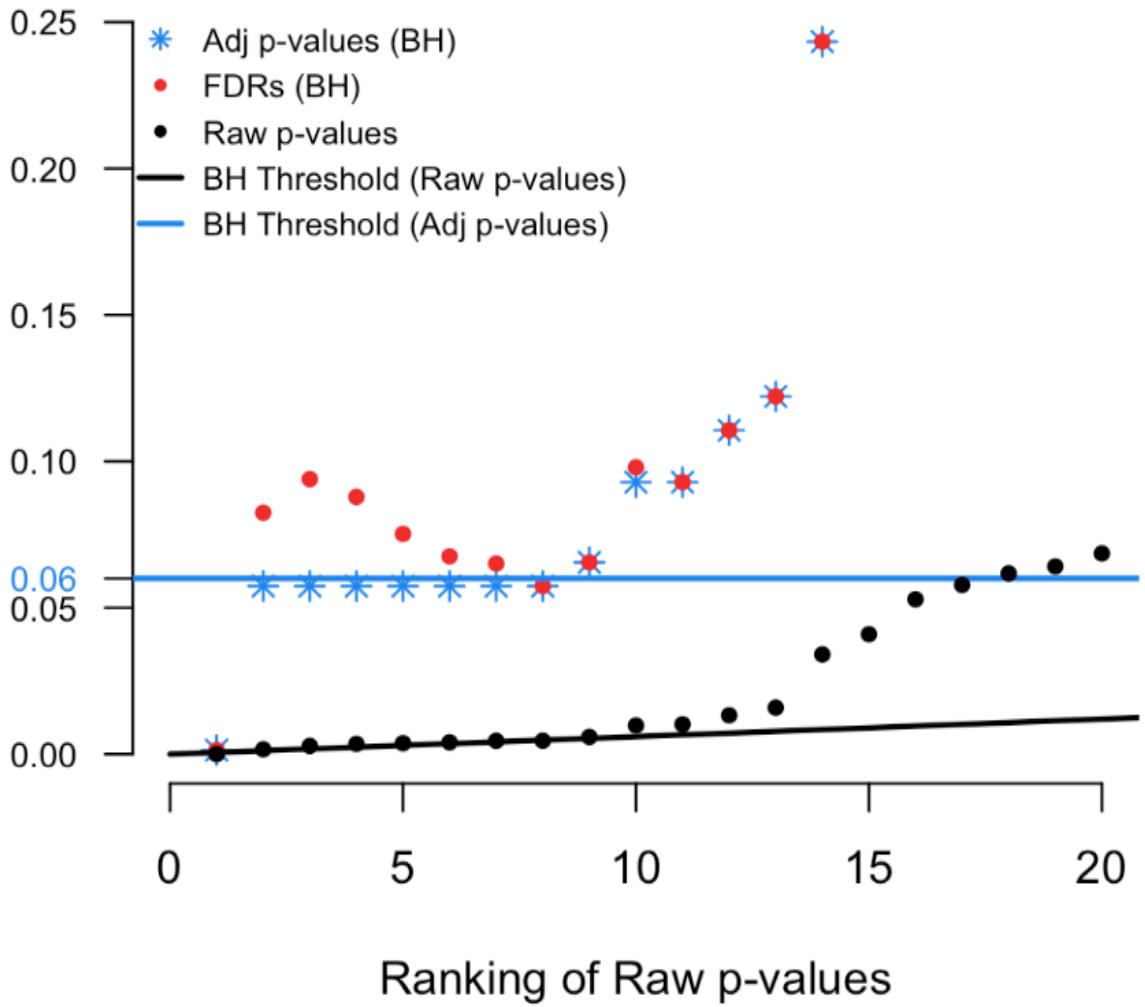
Blue threshold line $= \gamma$

Black threshold line $= \gamma \frac{i}{m}$

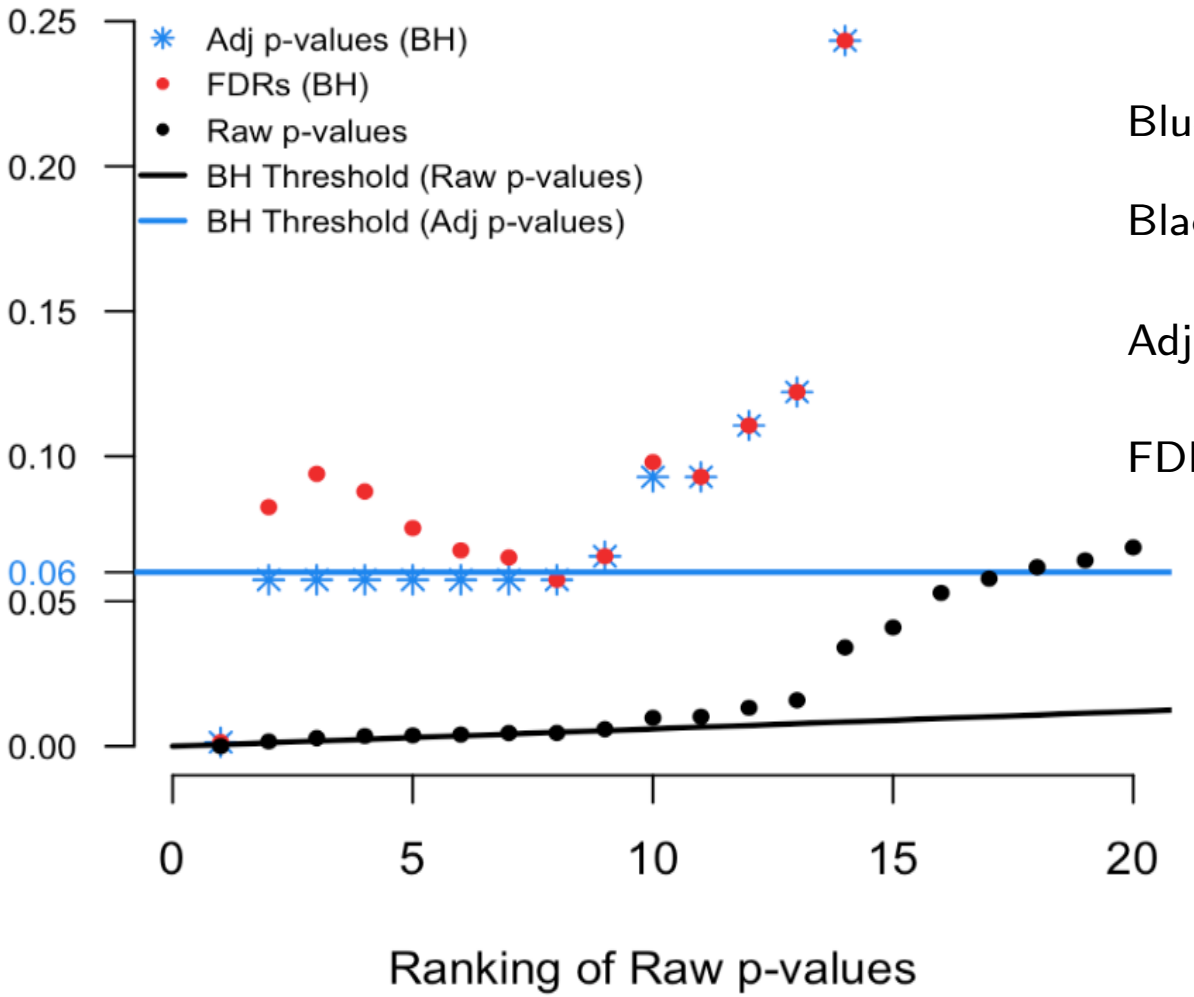
Adj p-values $= \min_{j \geq i} \left(\frac{p_{(j)} m}{j} \right)$

FDRs $= \frac{p_i m}{\text{rank}(p_i)} \cdot \hat{\pi}_0$

Zoomed in simulated example



Zoomed in simulated example



Blue threshold line $= \gamma$

Black threshold line $= \gamma \frac{i}{m}$

Adj p-values $= \min_{j \geq i} \left(\frac{p_{(j)} m}{j} \right)$

FDRs $= \frac{p_i m}{\text{rank}(p_i)} \cdot \hat{\pi}_0$

False Discovery Rates

Derivation from p-value space to $\mathcal{Z}$-value space:

$$p_{(i)} \leq \frac{i}{m}\gamma$$

$$p_{(i)}\frac{m}{i} \leq \gamma$$

$$F_0(\mathcal{Z}_{(i)})\frac{m}{i} \leq \gamma$$

$$\frac{\pi_0 F_0(\mathcal{Z}_{(i)})}{F(\mathcal{Z}_{(i)})} \leq \pi_0\gamma$$

False Discovery Rates

Derivation from p-value space to $\mathcal{Z}$-value space:

$$p_{(i)} \leq \frac{i}{m}\gamma$$

$$p_{(i)}\frac{m}{i} \leq \gamma$$

$$F_0(\mathcal{Z}_{(i)})\frac{m}{i} \leq \gamma$$

$$FDR_i = \frac{\pi_0 F_0(\mathcal{Z}_{(i)})}{F(\mathcal{Z}_{(i)})} \leq \pi_0\gamma$$

□

Null and alternative distributions: $F_0(\mathcal{Z}) = \int_{\mathcal{Z}} f_0(z)dz$ *and* $F_1(\mathcal{Z}) = \int_{\mathcal{Z}} f_1(z)dz$

Mixing distribution function:

$$F(\mathcal{Z}) = \pi_0 F_0(\mathcal{Z}) + \pi_1 F_1(\mathcal{Z}) \tag{4}$$

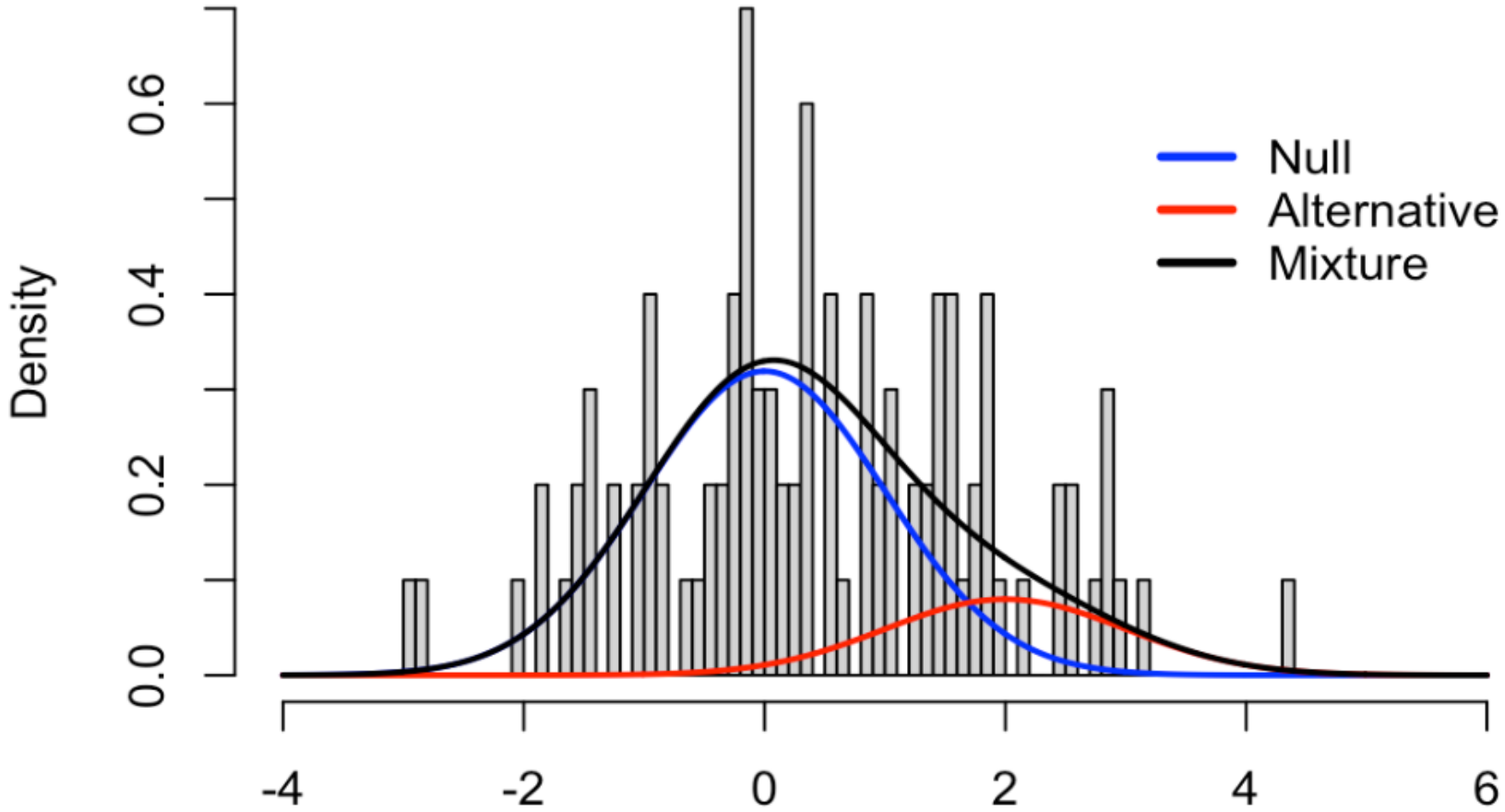
The global FDR:

$$FDR(\mathcal{Z}) := Pr\{null|z \in \mathcal{Z}\} = \frac{\pi_0 F_0(\mathcal{Z})}{F(\mathcal{Z})} \tag{5}$$

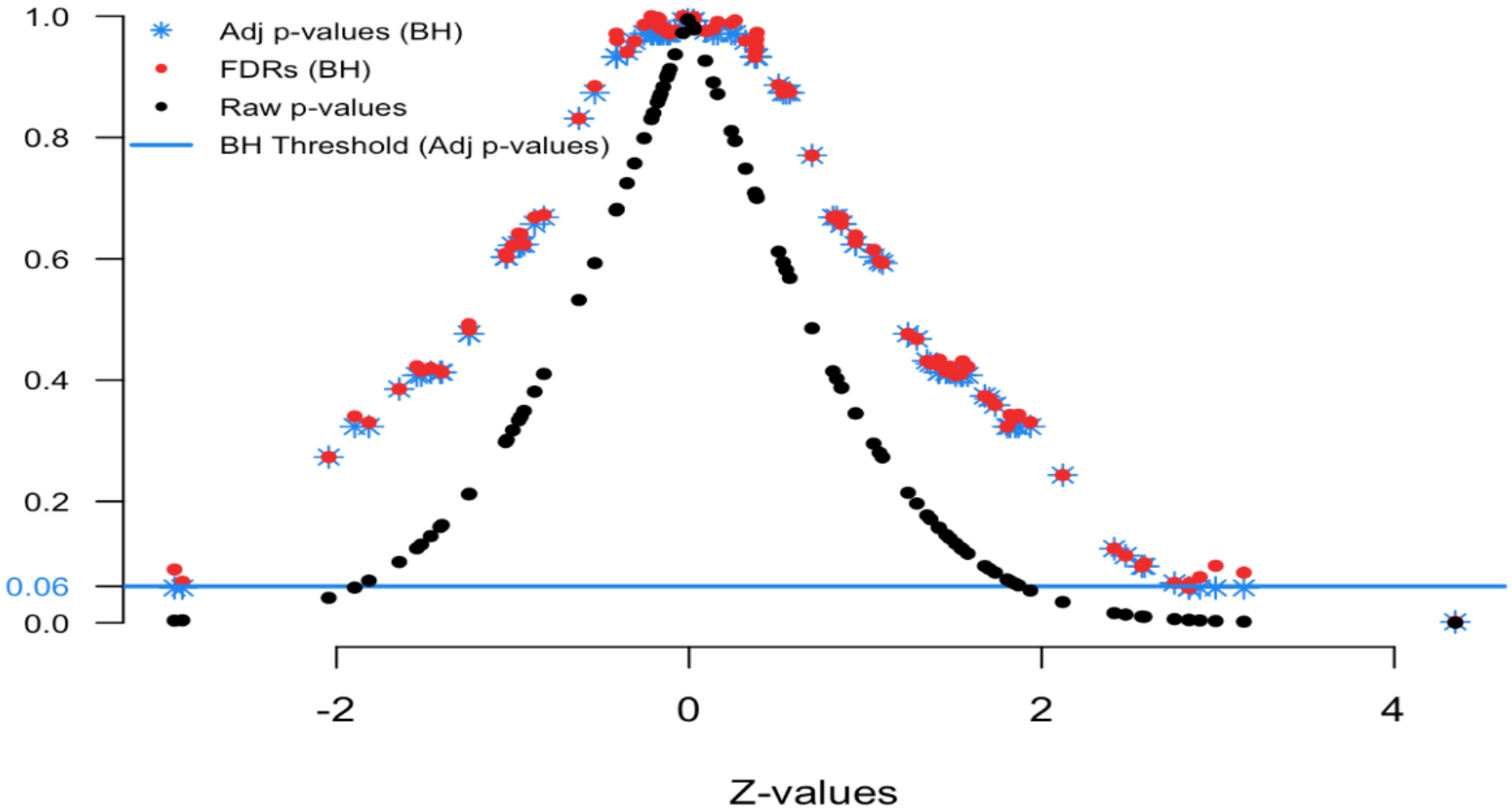
Empirical Bayes estimate of the global FDR:

$$\frac{\hat{\pi}_0 \hat{F}_0(\mathcal{Z})}{\hat{F}(\mathcal{Z})} \tag{6}$$

Simulated Example Density Histogram



FDR Z-values plot



False Discovery Rates

Null Proportion (π_0) Estimation

- The proportion of truly null features (π_0) in a mixture distribution
- Important component of the FDR estimates
- Conservative approach is to set $\pi_0 = 1$
- In our package users are able to specify an estimation routine
  - Storey
  - Meinshausen
  - Jiang
  - Nettleton
  - Pounds
  - New method **“Last Histogram Height"**

Algorithm 2: Last Histogram Height Method

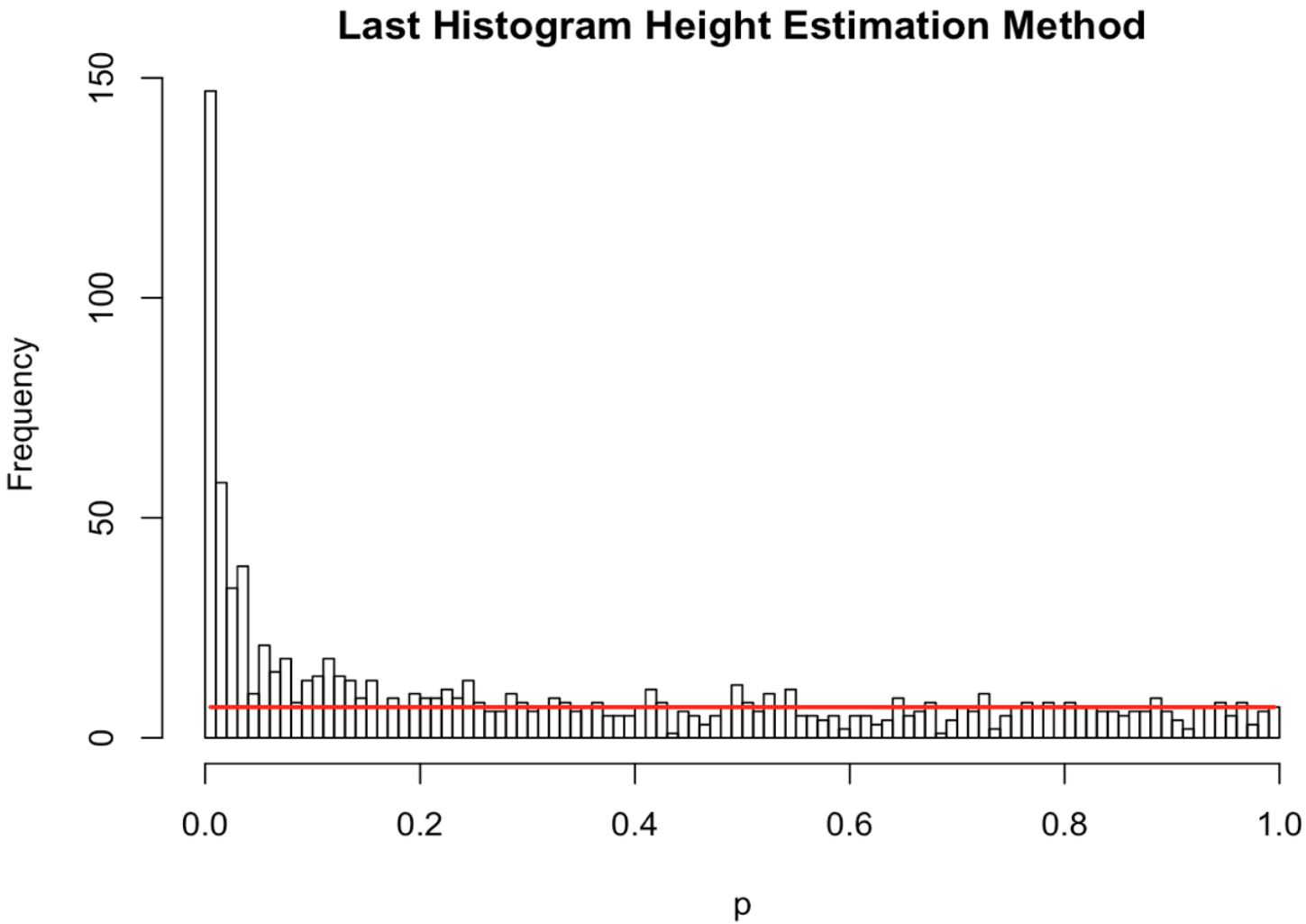
Result: Null proportion estimate

1. Plot a histogram of the raw p-values, $p_1, p_2, ...p_m$, with B number of bins, where $B < m$
   - The most stable bin method is `scott`, according to our simulations
2. Store the histogram bin heights H_b for each bin $b = 1, 2, ..., B$
3. Call the height of last bin H_B the "null height"
4. Set the estimate of π_0 to be

$$\hat{\pi}_0 = \frac{H_B B}{m}$$

False Discovery Rates

Null Proportion (π_0) Estimation



Algorithm 3: Storey's Method

Result: Null proportion estimate

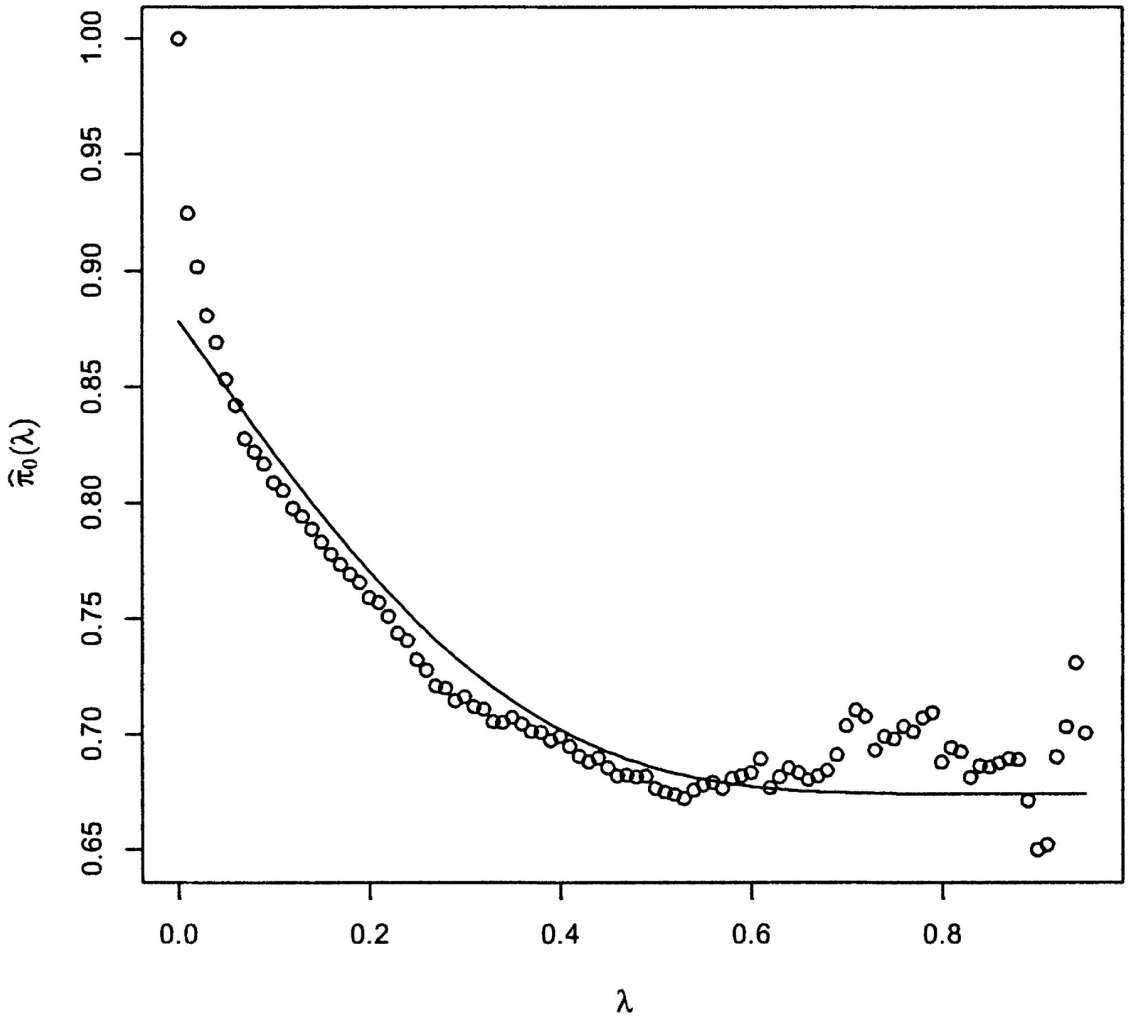
1. Let $p_{(1)}, p_{(2)}, ...p_{(m)}$ be the ordered p-values.
2. For a range of λ, say $\lambda = 0, 0.05, 0.10, ..., 0.95$, and $i = 1, ..., m$, calculate

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{m(1-\lambda)}$$

3. Let $\hat{h}(\cdot)$ be the natural cubic spline with 3 df of $\hat{\pi}_0(\lambda)$ on λ
4. Set the estimate of π_0 to be when $\lambda = 1$:

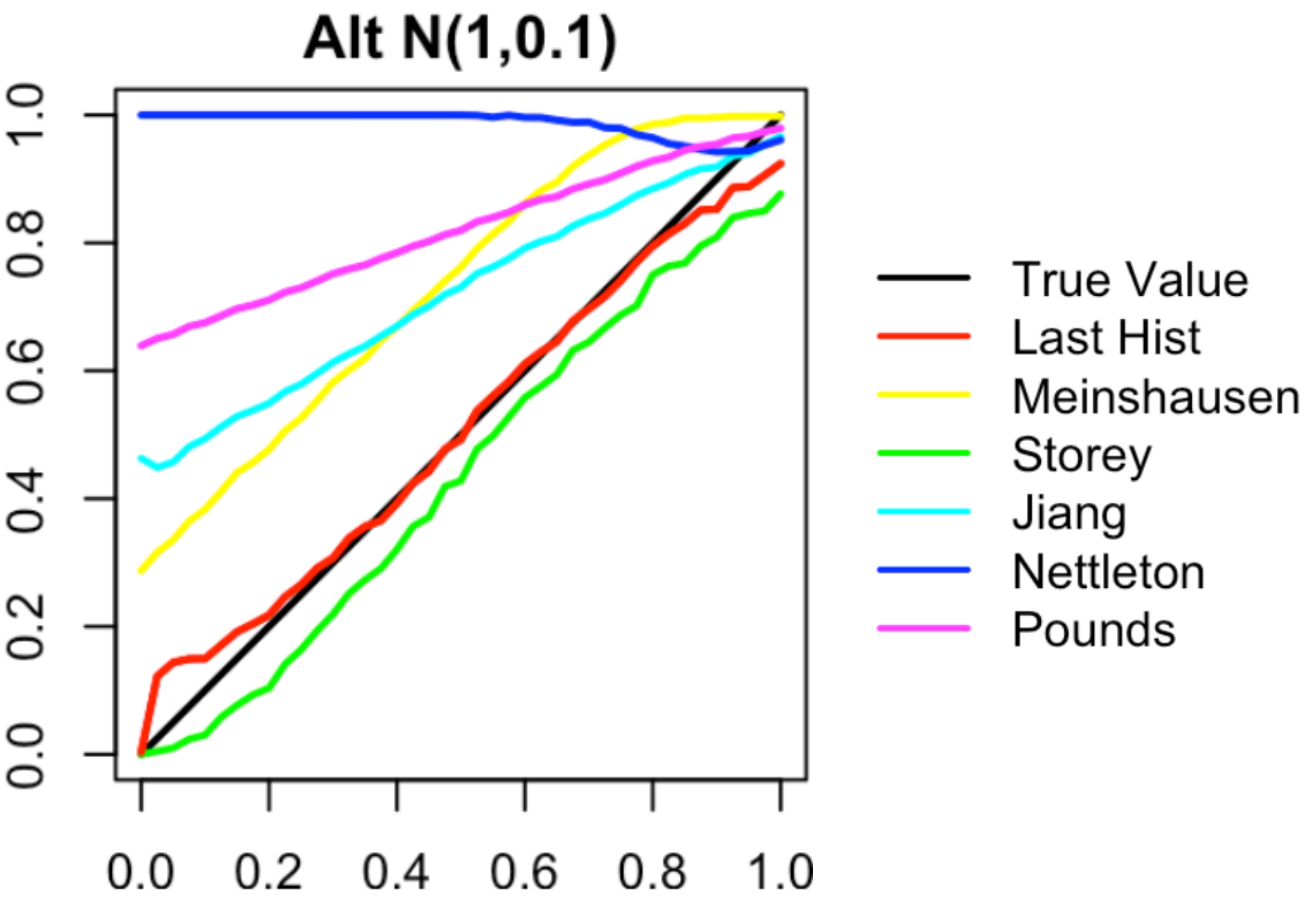
$$\hat{\pi}_0 = \hat{h}(1)$$

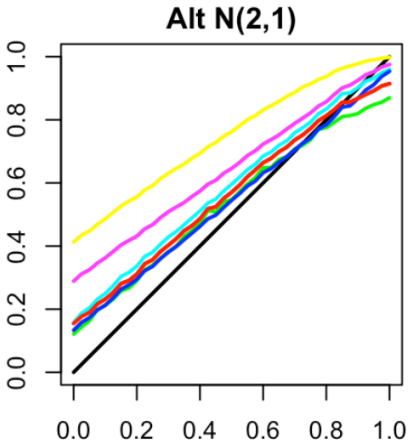
Natural cubic spline fit to the $\hat{\pi}_0(\lambda)$ outputs

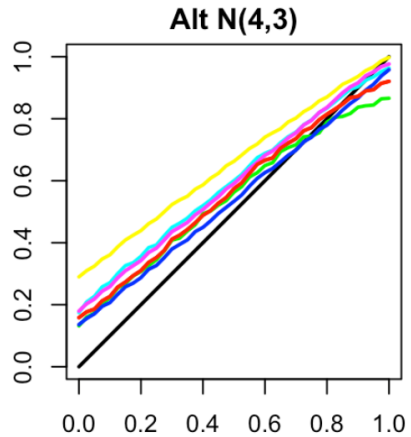


False Discovery Rates

Null Proportion (π_0) Estimation

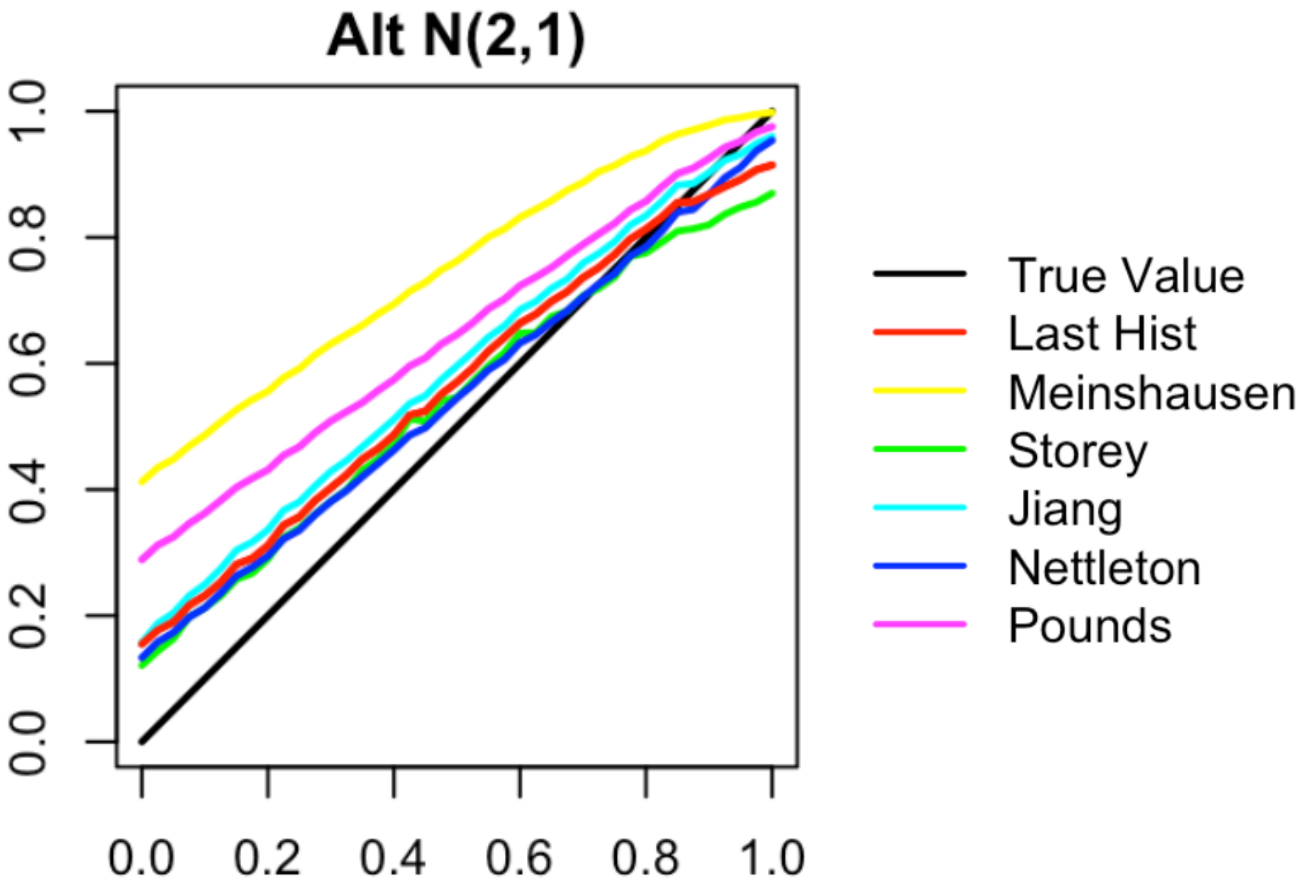


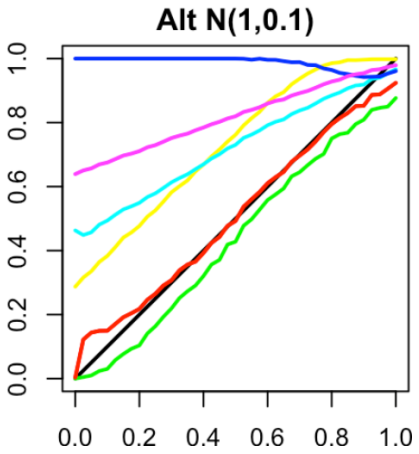


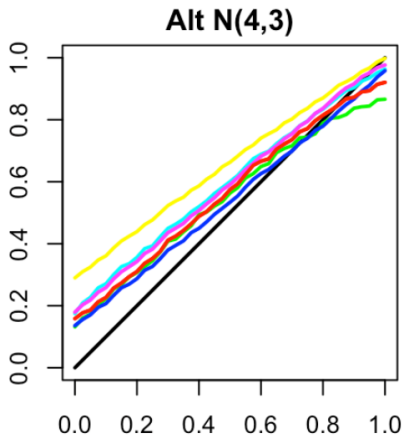


False Discovery Rates

Null Proportion (π_0) Estimation

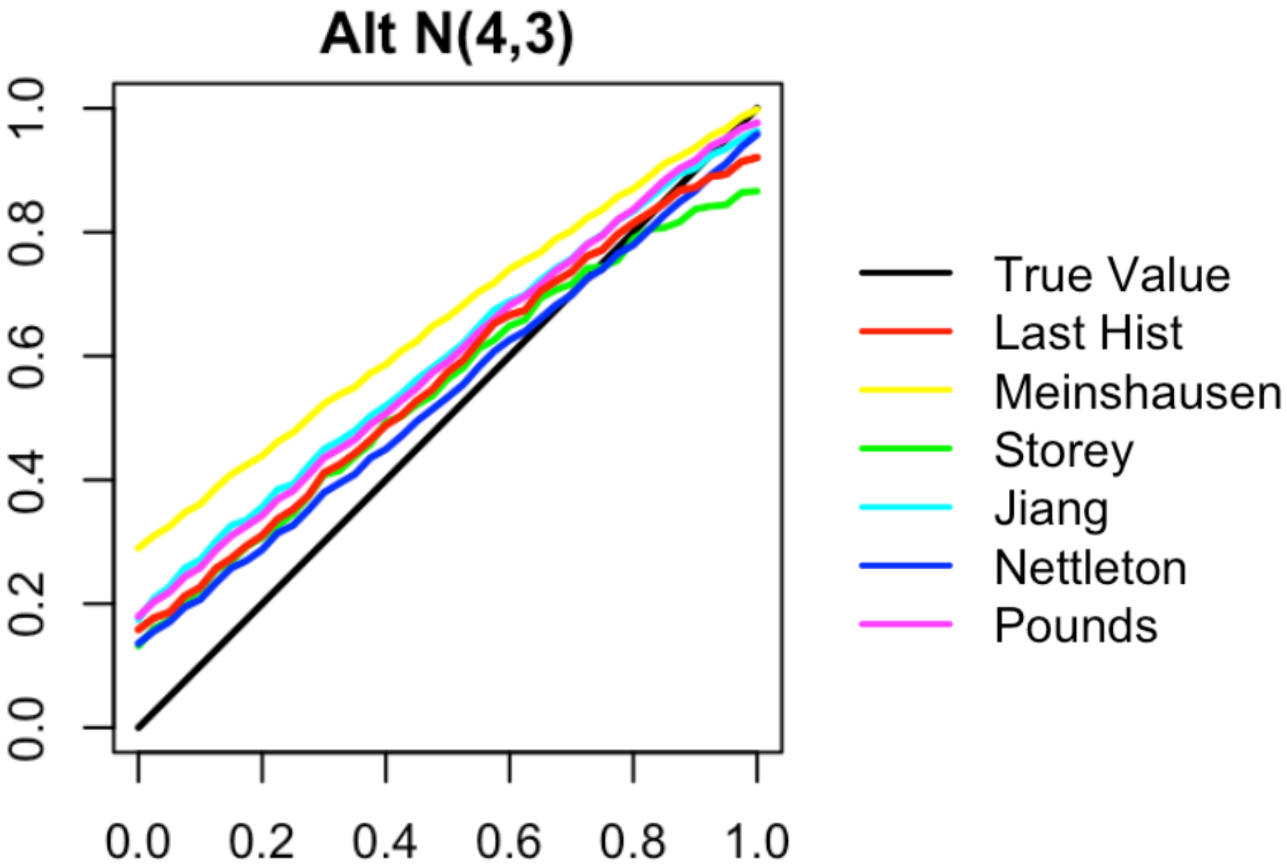


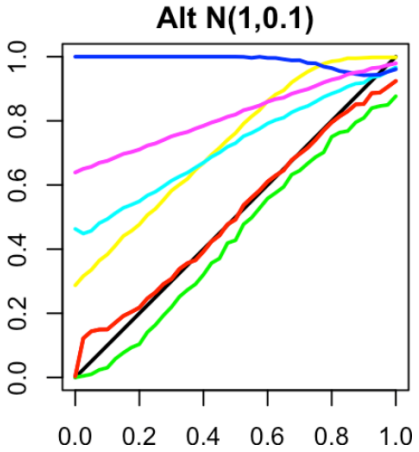




False Discovery Rates

Null Proportion (π_0) Estimation





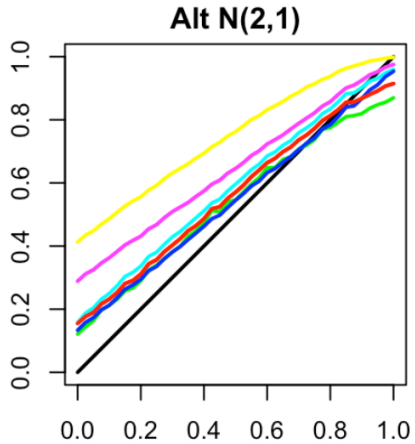


Table of Contents

R Package `FDRestimation`

`FDRestimation`

- A user-friendly R package
- Outputs false discovery rates
- Inputs are p-values or $\mathcal{Z}$-values and a variety of assumptions

FDRestimation :: p.fdr

`stats::p.adjust` is a popular multiple comparisons R function

The problems:

- Returns the BH adjusted p-value labeled as the FDR estimate
- Removing NAs
- Certain key assumptions are not adjustable

FDRestimation :: p.fdr

Adjustment Methods:

- Benjamini-Hochberg
- Benjamini-Yeukateili (with both positive and negative correlation)
- Bonferroni
- Holm
- Hochberg
- Sidak

All FDR estimates can be adjusted for pi0.

FDRestimation :: p.fdr

Other inputs:

- Threshold for important findings
- The assumed pi_0 value
- The desired pi_0 estimation method
- Whether to sort the results
- Whether to remove NAs in the imputed raw p-value vector count

FDRestimation :: p.fdr

The function will return a list object of the `p.fdr` class.

- **fdrs**
- **Results Matrix**
- **Reject Vector**
- **pi0**
- **threshold**
- **Adjustment Method**

Summary:

```
Call:
p.fdr(pvalues = sim.data.p)

Number of tests: 100
Raw p-value Range: [8e-04, 0.9941]

Adjustment Method: BH
False Discovery Rate Range: [0.04094, 1]
Findings at 0.05 level:
   Significant (Reject): 20
   Inconclusive (Fail to Reject): 80
---
Estimated/Assumed Null Proportion (pi0): 1
```

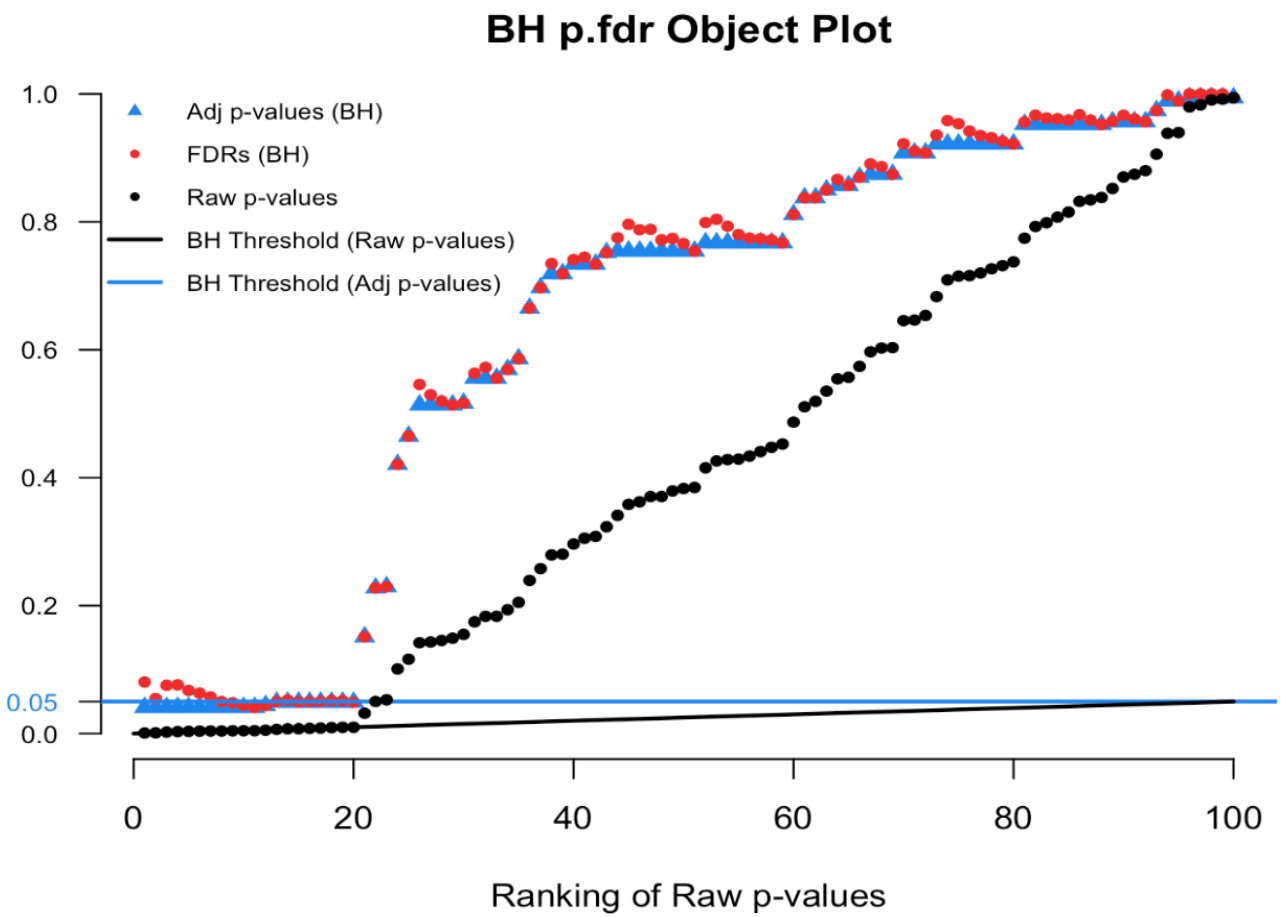
FDRestimation :: plot.p.fdr

- Plots the results from `p.fdr`
- By default:
  - the adjusted FDRs
  - adjusted p-values
  - raw p-values are plotted
  - threshold line for raw p-values
  - threshold line for adjusted p-values
- Other inputs:
  - axis limits
  - location of the legend
  - title of the plot
  - plotting symbols
  - colors of points and lines

FDRestimation :: plot.p.fdr

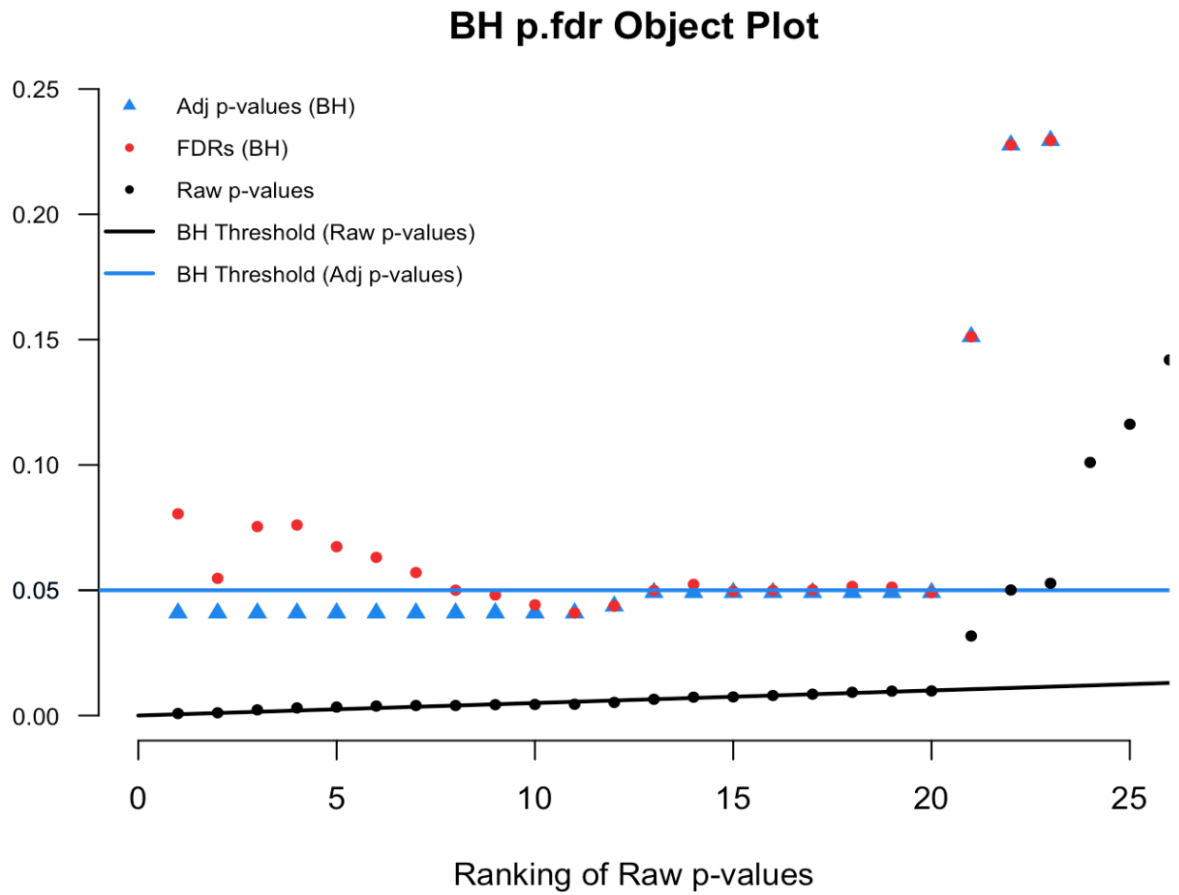
```
set.seed(88888)
sim.data.p=c(runif(80), runif(20, min=0, max=0.01))

plot(p.fdr(p=sim.data.p))
```



FDRestimation :: plot.p.fdr

```
plot(p.fdr(p=sim.data.p), xlim=c(0,25), ylim=c(0,0.25))
```



FDRestimation :: get.pi0

- Estimates the null proportion from the raw p-values
- 6 different methods:
  - Last Histogram Height
  - Storey
  - Meinshausen
  - Jiang
  - Nettleton
  - Pounds
- Other inputs:
  - Histogram breaks method
  - Threshold of importance
  - Z-values

FDRestimation :: get.pi0

```
set.seed(88888)

get.pi0(sim.data.p, estim.method="set.pi0", set.pi0=0.8)

[1] 0.8

get.pi0(sim.data.p, estim.method="last.hist")

[1] 0.85

get.pi0(sim.data.p, estim.method="storey")

[1] 0.8867
```

Table of Contents

Final Thoughts

- Encourage the use of FDR methods
- Make clear that p-value adjustments are not interchangeable with estimated FDRs

- Provide a useful and easy tool for computing false discovery rates
- Flexible function that allows the user to specify all assumptions

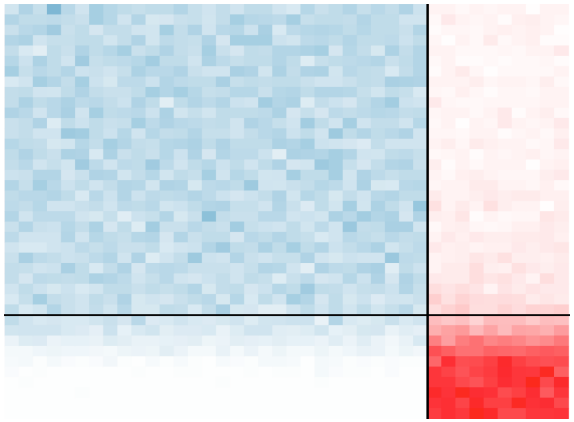
Final Thoughts

Questions?

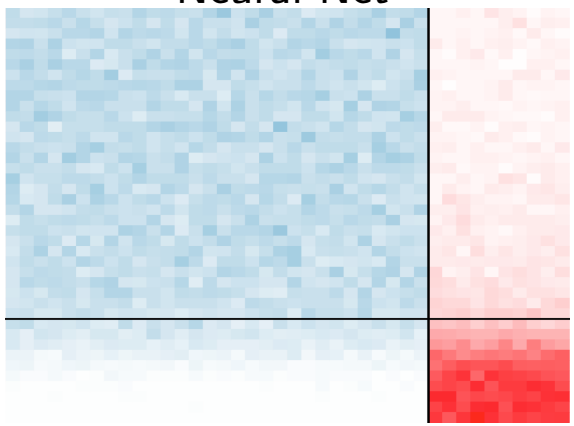
Heatmaps of Rankings

- Heatmaps of rankings of the original gene expression counts

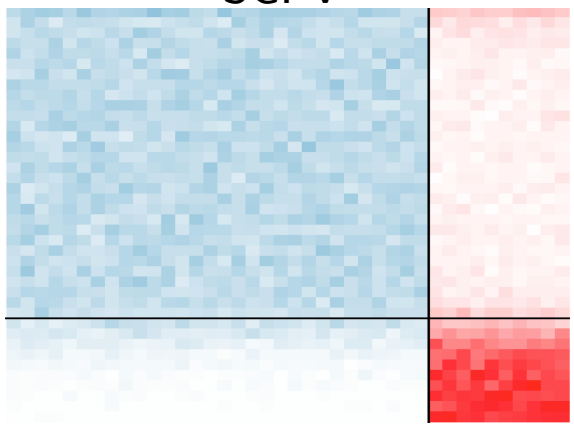
Nominal p-values



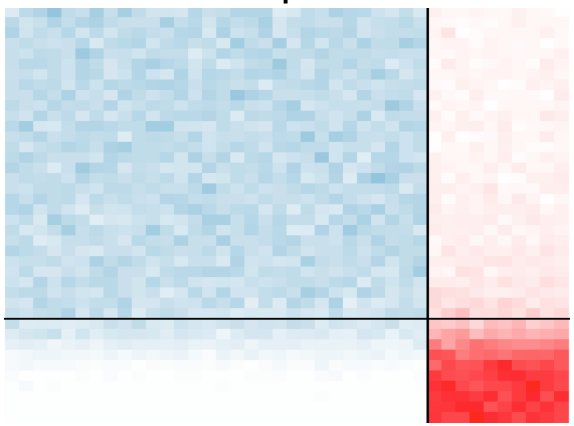
Neural Net



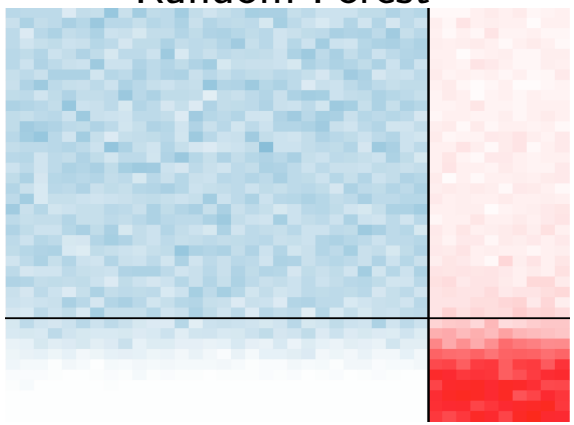
SGPV



BH Emp FDRs



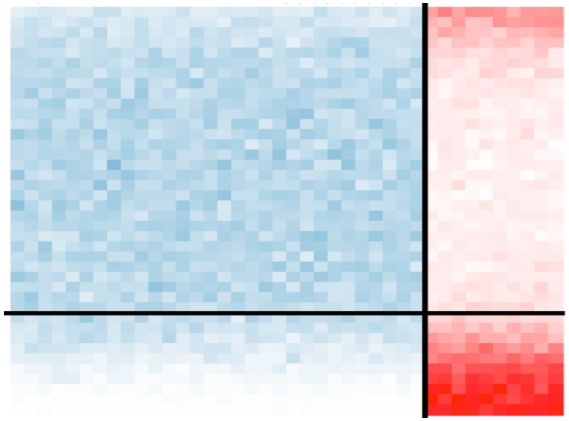
Random Forest



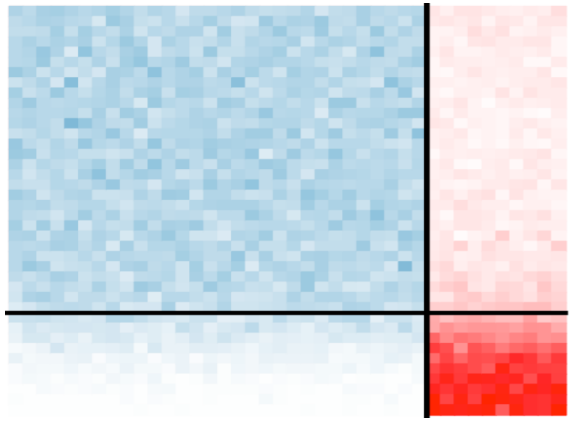
Heatmaps

- Heatmaps of rankings of the counts **with added within person correlation**

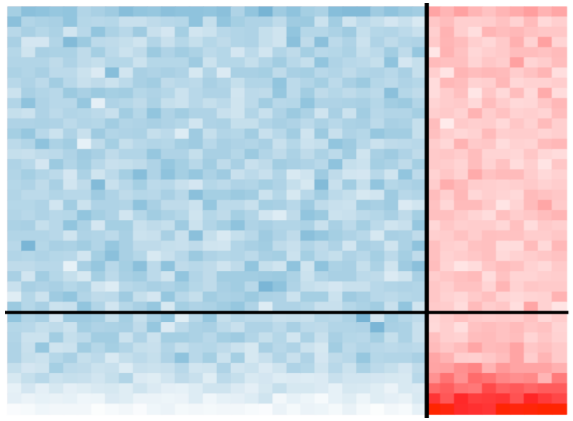
Nominal *p*-values



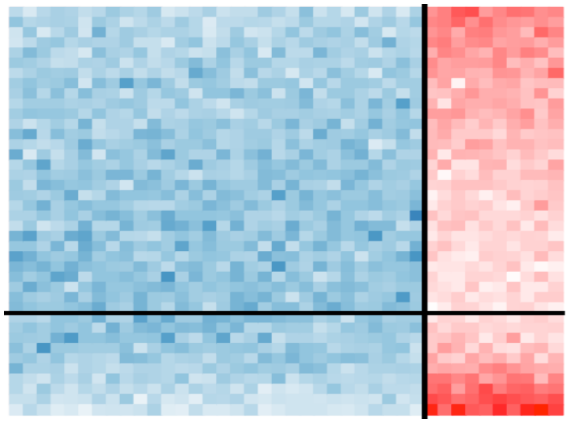
Neural Net



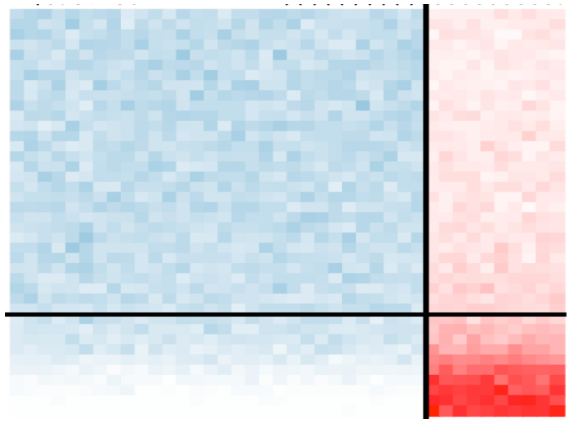
SGPV



BH Emp FDRs



Random Forest



Final Thoughts

Simple Motivating Example

Feature	Raw p-value	Z-value	Adjusted p-value	FDR	Lower Bound FDR
Feature 1	0.005	2.807	0.025	0.025	0.019
Feature 2	0.049	1.969	0.064	0.122	0.126
Feature 3	0.050	1.960	0.064	0.083	0.128
Feature 4	0.051	1.951	0.064	0.064	0.130
Feature 5	0.700	0.385	0.700	0.700	0.481

Table 1: Example with 5 features using the Benjamini-Hochberg adjustment and assuming a two-sided normal distribution.

FDRestimation :: p.fdr

Simulate 100 features with a true null proportion of 80%.

Input:

```
set.seed(88888)

sim.data.p= c(runif(80),
              runif(20,
                 min=0,
                 max=0.01))

p.fdr(p=sim.data.p[1:5],
         threshold=0.05,
         adjust.method="BH")
```

Output:

```
$fdrs
[1] 1.000 0.957 0.968 0.834 0.698

$'Results Matrix'
      BH FDRs     Adjusted p-values     Raw p-values
1       1.000                 0.698            0.239
2       0.957                 0.834            0.574
3       0.968                 0.834            0.774
4       0.834                 0.834            0.834
5       0.698                 0.698            0.279

$'Reject Vector'
[1]  "FTR.H0" "FTR.H0" "FTR.H0" "FTR.H0" "FTR.H0"

$pi0
[1] 1

$threshold
[1]  0.05

$'Adjustment Method'
[1] "BH"

$Call
p.fdr(p=sim.data.p[1:5], threshold=0.05, adjust.method="BH")
```

FDRestimation::p.fdr

Summary of p.fdr

```
Call:
p.fdr(pvalues = sim.data.p)

Number of tests: 100
Raw p-value Range: [8e-04, 0.9941]

Adjustment Method: BH
False Discovery Rate Range: [0.04094, 1]
Findings at 0.05 level:
   Significant (Reject): 20
   Inconclusive (Fail to Reject): 80
---
Estimated/Assumed Null Proportion (pi0): 1
```

References I

Y. Benjamini and Y. Hochberg.

Controlling the false discovery rate: A practical and powerful approach to multiple testing.

Journal of the Royal Statistical Society, 57(1):289–300, 1995.

Y. Benjamini and D. Yekutieli.

The control of the false discovery rate in multiple testing under dependency.

Annals of statistics, pages 1165–1188, 2001.

C. Bonferroni.

Teoria statistica delle classi e calcolo delle probabilita.

Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze, 8: 3–62, 1936.

Y. Hochberg.

A sharper bonferroni procedure for multiple tests of significance.

Biometrika, 75(4):800–802, 1988.

References II

S. Holm.

A simple sequentially rejective multiple test procedure.

Scandinavian journal of statistics, pages 65–70, 1979.

H. Jiang and R. Doerge.

Estimating the proportion of true null hypotheses for multiple comparisons.

Cancer informatics, 6:117693510800600001, 2008.

N. Meinshausen, J. Rice, et al.

Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses.

The Annals of Statistics, 34(1):373–393, 2006.

D. Nettleton, J. G. Hwang, R. A. Caldo, and R. P. Wise.

Estimating the number of true null hypotheses from a histogram of p values.

Journal of agricultural, biological, and environmental statistics, 11(3):337, 2006.

References III

S. Pounds and S. W. Morris.

Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values.

Bioinformatics, 19(10):1236–1242, 2003.

Z. Šidák.

Rectangular confidence regions for the means of multivariate normal distributions.

Journal of the American Statistical Association, 62(318):626–633, 1967.

J. D. Storey and R. Tibshirani.

Statistical significance for genomewide studies.

Proceedings of the National Academy of Sciences, 100(16):9440–9445, 2003.