Data Obesity

Inflated Estimation and Irreproducible Results with High Throughput Data

Naomi Altman
Penn State University

Dec 2013

Feature Selection

The Measurement Process in a Nutshell

- Do the bench experiment.
- Preprocess the data.
- **Select the interesting features.**
- Downstream analysis of interesting features.
- Interpretation of interesting features.
- Design the next bench experiment.

Feature Selection

The Measurement Process in a Nutshell

- Do the bench experiment.
- Preprocess the data.
- **Select the interesting features.**
- Downstream analysis of interesting features.
- Interpretation of interesting features.
- Design the next bench experiment.

But often we cannot reproduce the results.

Today's Talk

1. Feature selection using False Discovery Rates
   - Continuous test statistics
   - Discrete test statistics (joint with I. Dialsingh)
2. Irreproducible feature selection.
3. Data "obesity"
   - effect size in selected t-tests
   - effect size in selected correlations
4. A slimming diet?

False Discovery Rate

Typically in high throughput studies we do a statistical test for each feature.

If we do 10 thousand tests and reject when $P \leq 0.05$ with totally random differences between treatments we expect 500 rejections.

Multiple comparisons adjustments account for this by rejecting at a more stringent p-value.

Adjustments can be made less conservative by estimating π_0 the percentage of tests which are truly null.

We start this section by discussing adaptive False Discovery Rate (FDR) as a feature selection method.

False Discovery Rate

Benjamini & Hochberg (1995) realized that when testing 1000's of hypotheses a few errors could be tolerated.

False Discovery Rate (FDR) is the expected percentage of null hypotheses among the statistically significant tests.

False Discovery Rate

Benjamini & Hochberg (1995) realized that when testing 1000's of hypotheses a few errors could be tolerated.

False Discovery Rate (FDR) is the expected percentage of null hypotheses among the statistically significant tests.

Table : Outcomes of m tests.

	Not Significant	Significant	Total
True Null	U	V	m_0
False Null	T	S	m_1
Total	W	R	m

$$FDR = E\left(\frac{V}{R}|R>0\right)P(R>0)$$

R: number of rejections
V: number of false rejections

False Discovery Rate

Benjamini & Hochberg (1995) realized that when testing 1000's of hypotheses a few errors could be tolerated.

False Discovery Rate (FDR) is the expected percentage of null hypotheses among the statistically significant tests.

Table : Outcomes of m tests.

	Not Significant	Significant	Total
True Null	U	V	m_0
False Null	T	S	m_1
Total	W	R	m

$$FDR = E\left(\frac{V}{R}|R>0\right)P(R>0)$$

R: number of rejections
V: number of false rejections

Since we are trying to control *false discoveries* we do not need to control for the truly non-null tests.
Adaptive FDR methods use an estimate of $\pi_0 = m_0/m$ to improve the power of the multiple comparisons adjustments.

Implementing FDR procedures

- Compute a test statistic for each hypothesis.
  - We might use the p-value as the test statistic.

Implementing FDR procedures

- Compute a test statistic for each hypothesis.
  - We might use the p-value as the test statistic.
- Order the hypotheses from most to least significant, so that H_{0k} has the k^{th} significant test statistic.
- Estimate FDR(k) if we reject $H_{01} \cdots H_{0k}$.

Implementing FDR procedures

- Compute a test statistic for each hypothesis.
  - We might use the p-value as the test statistic.
- Order the hypotheses from most to least significant, so that H_{0k} has the k^{th} significant test statistic.
- Estimate FDR(k) if we reject $H_{01} \cdots H_{0k}$.
- Either
  - Pick a level q and reject $H_{01} \cdots H_{0k}$ if FDR(k)$< q$ OR
  - Pick a p-value α and reject H_{0i} if its p-value is less than α. Then estimate FDR(k) for the rejected list of k features.

Implementing FDR procedures

Popular methods for estimating FDR or pFDR=E(V/R |R>0).

Benjamini and Hochberg (1995,2000)- FDR

- Find the maximal i such that
$$p_{(i)} \leq \frac{i \times q}{m_0}$$
- When the test statistics are continuous and independent, then this algorithm controls the FDR at level at least q

Implementing FDR procedures

Popular methods for estimating FDR or pFDR=E(V/R |R>0).

Benjamini and Hochberg (1995,2000)- FDR

- Find the maximal i such that
$$p_{(i)} \leq \frac{i \times q}{m_0}$$
- When the test statistics are continuous and independent, then this algorithm controls the FDR at level at least q

Storey (2001) - pFDR

- Let
$$\hat{q}_i = \frac{p_{(i)} m_0}{i}$$
- $q_1 = \hat{q}_1$.
- If $\hat{q}_i \geq q_{i-1}$ then $q_i = \hat{q}_i$. Otherwise $q_i = q_{i-1}$.
- Find the maximal i such that $q_i \leq q$.
- When the test statistics are continuous and independent, then this algorithm controls the pFDR at level at least q

Estimating FDR for Discrete Data

Discrete test statistics

arise from binary and count data such as

- read counts in RNA-seq and ChIP-seq
- SNP studies
- thresholding (above/below)
- multiple 2-way tables (e.g. surveys)

Estimating FDR for Discrete Data

Discrete test statistics

arise from binary and count data such as

- read counts in RNA-seq and ChIP-seq
- SNP studies
- thresholding (above/below)
- multiple 2-way tables (e.g. surveys)

Methods developed for continuous data are conservative for discrete data.

- due to the non-uniformity of the Null distribution of p-values
- and the lack of power for small counts.

Why does discreteness matter?

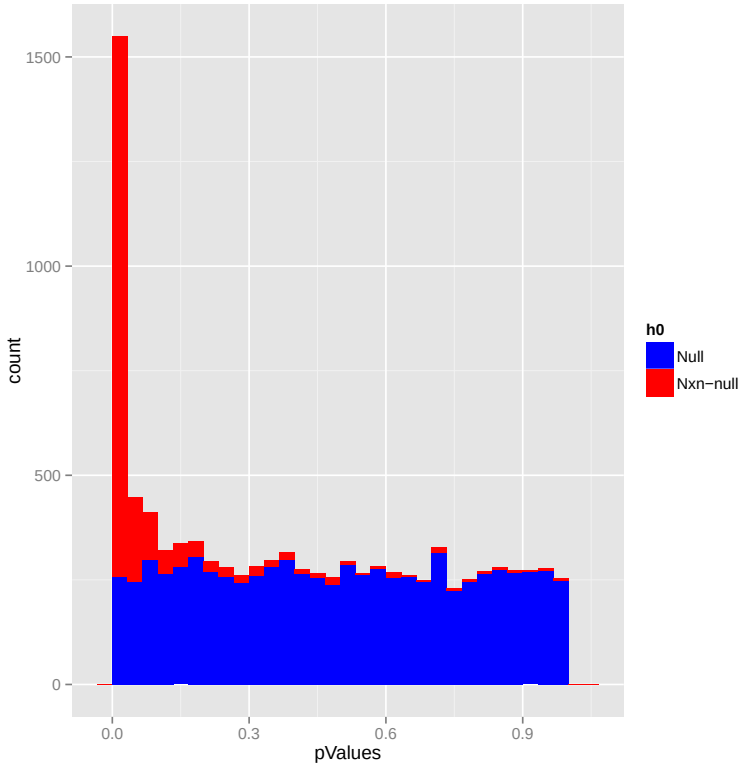


Figure : Continuous p-values $\pi_0 = 0.8$

P-values from 10000 t-tests.

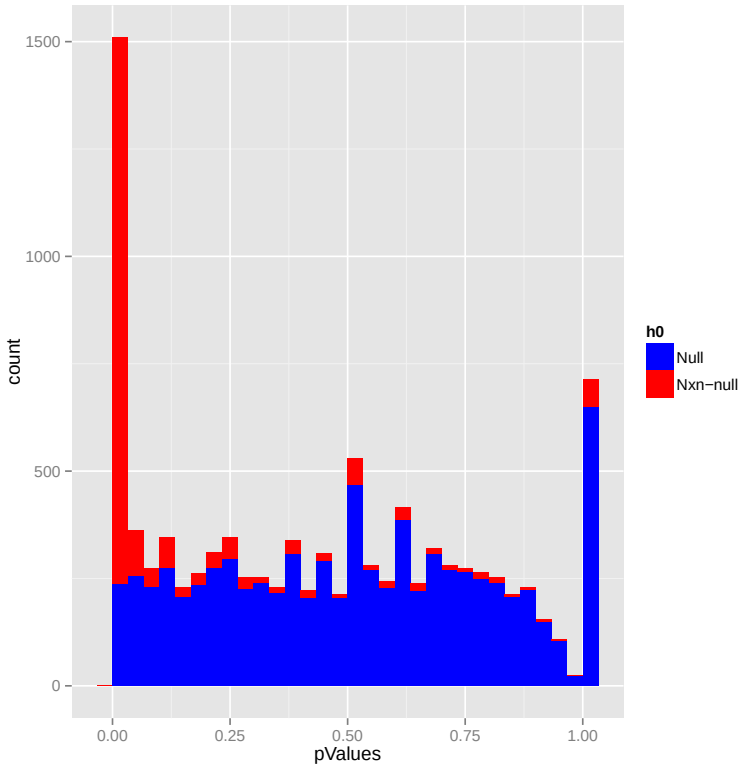
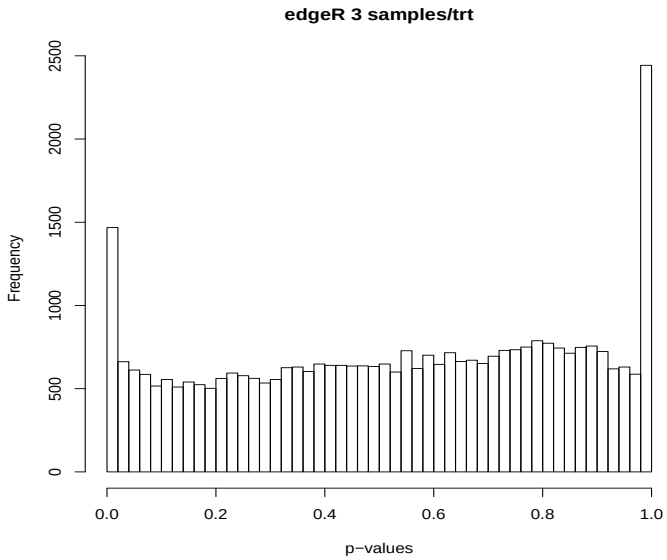


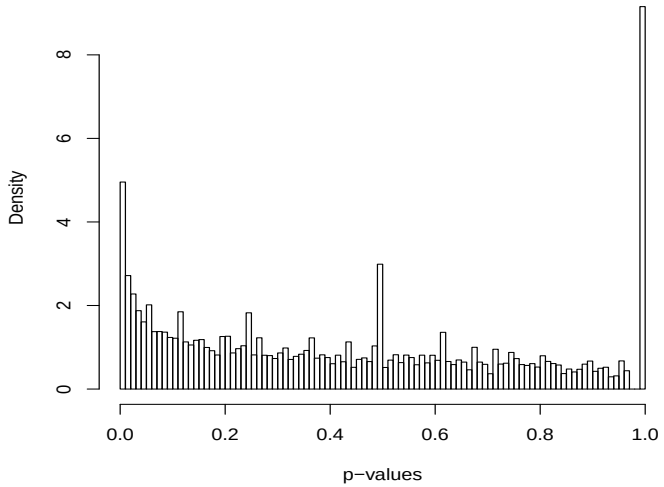
Figure : Discrete p-values $\pi_0 = 0.8$

P-values from 10000 Fisher exact tests.

Why does discreteness matter?



P-values from an RNA-seq study of 2 maize genotypes with 3 biological replicates.



P-values from a SNP study in cattle.

Why does discreteness matter?

We will use the same heuristics for discrete and continuous tests. BUT

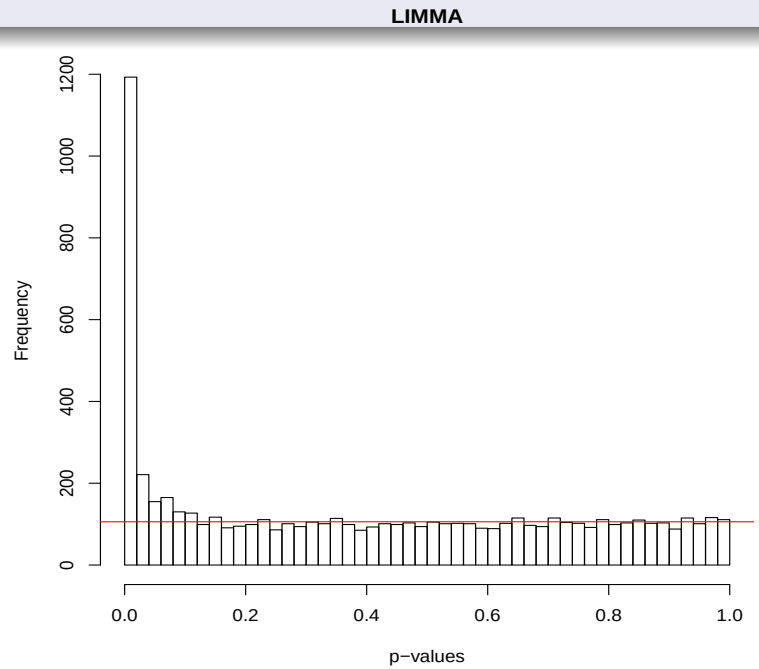
Table : Distribution of p-values.

	Continuous	Discrete
null p-value distribution	uniform	depends on an ancillary
Prob(p=1)	0	>0
percent of support points with positive probability	0%	100%
minimum achievable p-value	0	>0

Estimating π_0

Estimating π_0 from continuous p-values

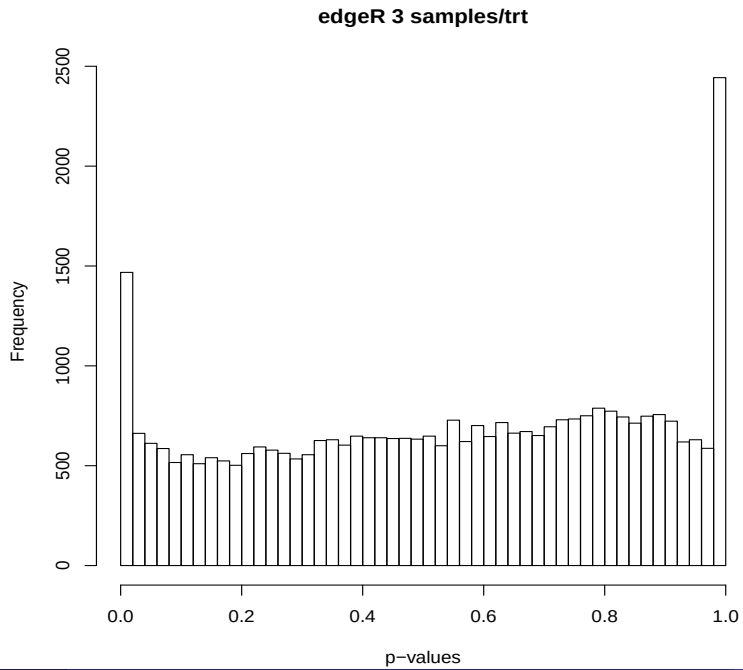
- Storey (2002) estimates height of flat part of histogram.
- Nettleton et al (2006) estimate the heights of the bins in excess of expected given $\hat{\pi}_0$.
- Pounds and Cheng (2004) assume all true non-nulls have p=0, so $2 * \bar{p} \approx \pi_0$.



Estimating π_0

Estimating π_0 from discrete p-values

- These methods seem less plausible since low power non-null tests may have p-values far from 0.
- As well, both null and non-null tests have p-values with mass at 1, leading to a peak at p=1.
- We add 2 new methods.



Estimating π_0 from expected frequencies

Estimating π_0 from discrete p-values

- There is often an ancillary statistic which determines the distribution of the test statistic. e.g. row totals.
- If there are many tests with the same value of the ancillary, $\hat{f}(p)$ the empirical distribution of the p-values can be estimated by the observed frequencies.
- Regression method - regress empirical frequencies of p-values against expected frequencies under H_0.
- The slope is approximately π_0.

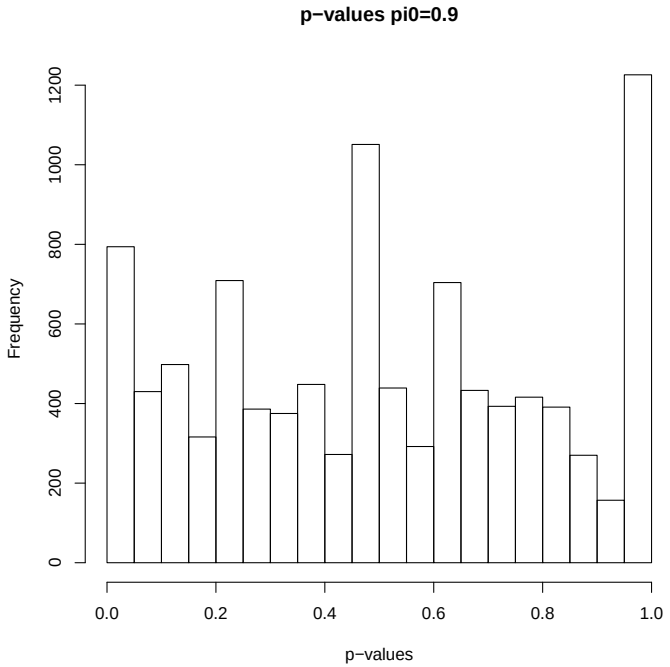
Estimating π_0 by removing zero-power tests

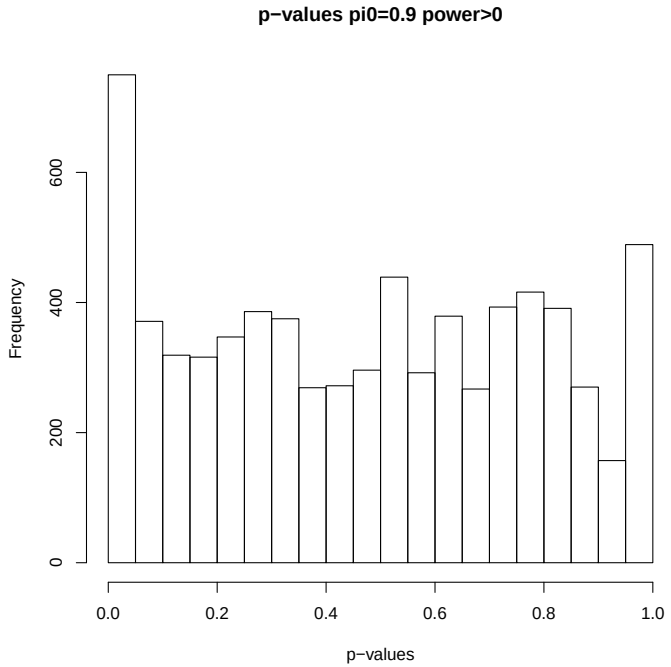
Minimum achievable p-value

- For a given test statistic, there is some set $\psi_1 < \cdots < \psi_k = 1$ of achievable p-values.
- ψ_1 is the minimal achievable p-value for the test.
- Select a level α the maximum p-value at which to reject the null hypothesis.
- If $\psi_1 > \alpha$ then the test has zero power.

Estimating π_0 by removing zero-power tests

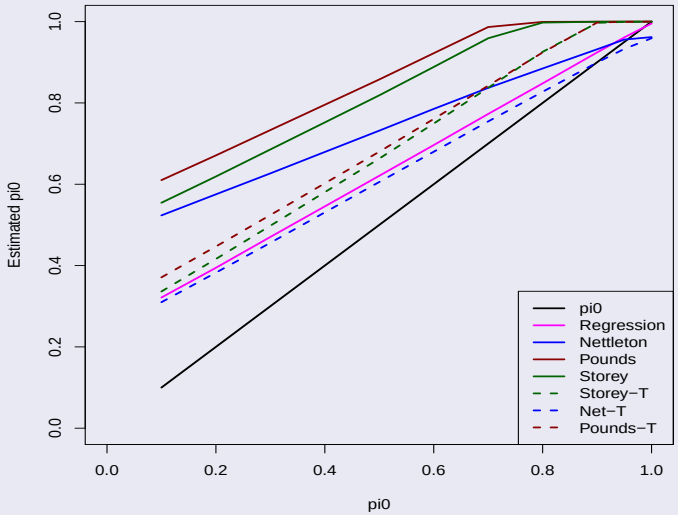
- Tarone (1990) suggested removing tests with zero power.
- Our "T" methods remove the zero-power tests and then proceed with a method for continuous data.
- "T" methods remove some of the excess mass at $p = 1$ and make the histogram of p-values more uniform.
- In this talk, we use the Storey-T method.
- We remove the tests with zero power at $\alpha = 0.01$ and then use Storey's method on the remaining tests (right plot).



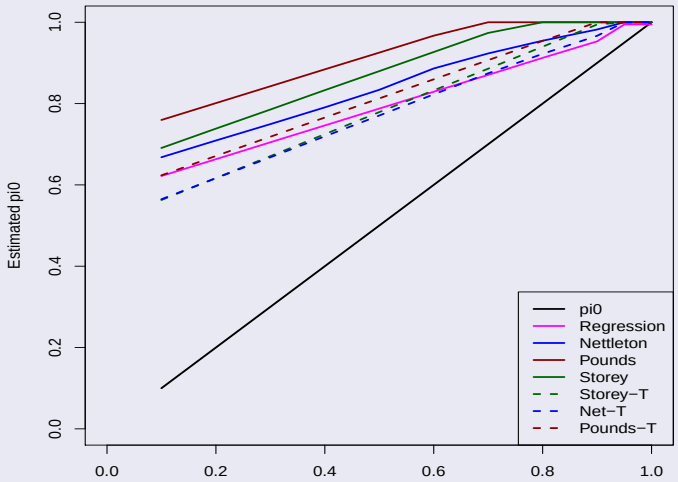


Estimating π_0

Simulation Results



25^{th} percentile of $\hat{\pi}_0$ using simulated RNA-seq data.



25^{th} percentile of $\hat{\pi}_0$ using simulated SNP data.

Estimating FDR

3 Adaptive Methods

- The adaptive Benjamini and Hochberg method uses BH control of FDR with $\hat{m}_0 = \hat{\pi}_0 m$.
- Gilbert (2005) filters zero power tests then applies the BH method to the remaining m_F tests.
- We suggest an adaptive Gilbert method that uses an estimate of π_0 with Gilbert's method.

Simulation Results

Using the same simulation scenario as before we implemented

- Benjamini and Hochberg's 1995 method
- Gilbert's (2005) method using $\alpha = 0.01$
- Adaptive versions of BH and Gilbert using

Simulation Results

Using the same simulation scenario as before we implemented

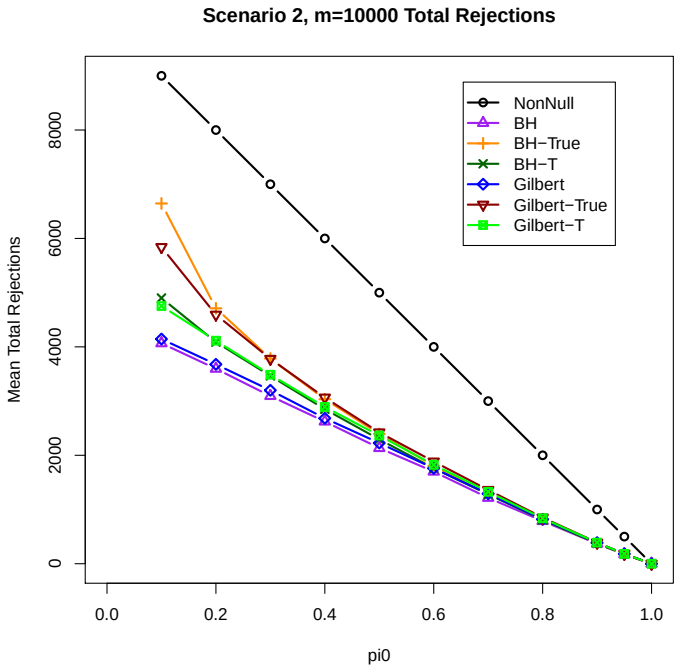
- Benjamini and Hochberg's 1995 method
- Gilbert's (2005) method using $\alpha = 0.01$
- Adaptive versions of BH and Gilbert using
  - true π_0
  - estimated π_0 using the Storey-T method

Simulation Results

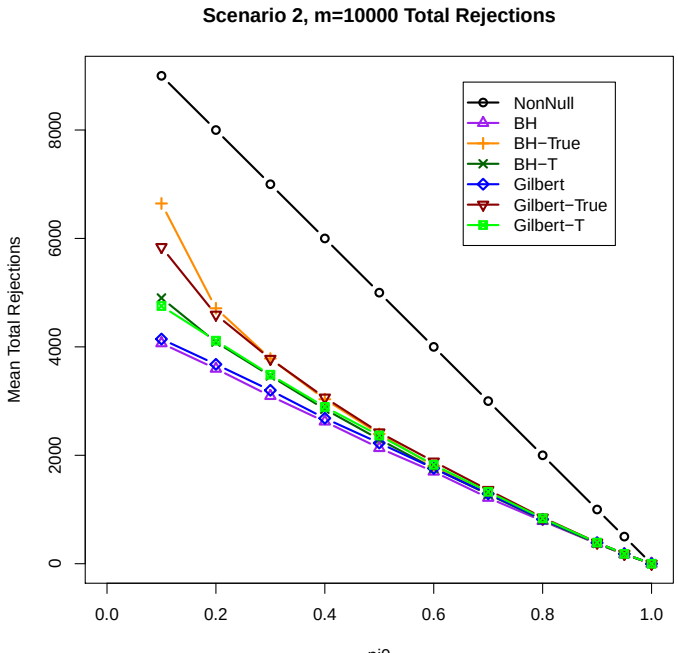
Using the same simulation scenario as before we implemented

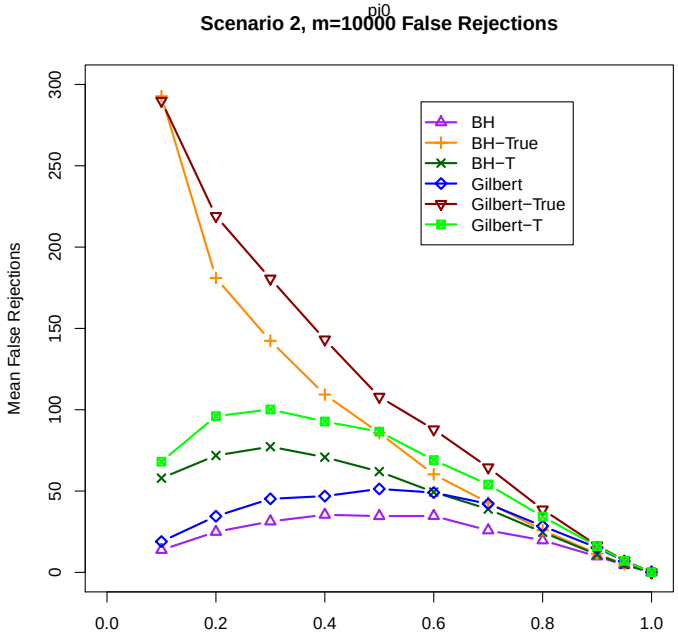
- Benjamini and Hochberg's 1995 method
- Gilbert's (2005) method using $\alpha = 0.01$
- Adaptive versions of BH and Gilbert using
  - true π_0
  - estimated π_0 using the Storey-T method
- We considered error rates for
  - false detection
  - false nondetection (including nondetection due to zero power)
  - total errors

Results m=10,000 few small margins

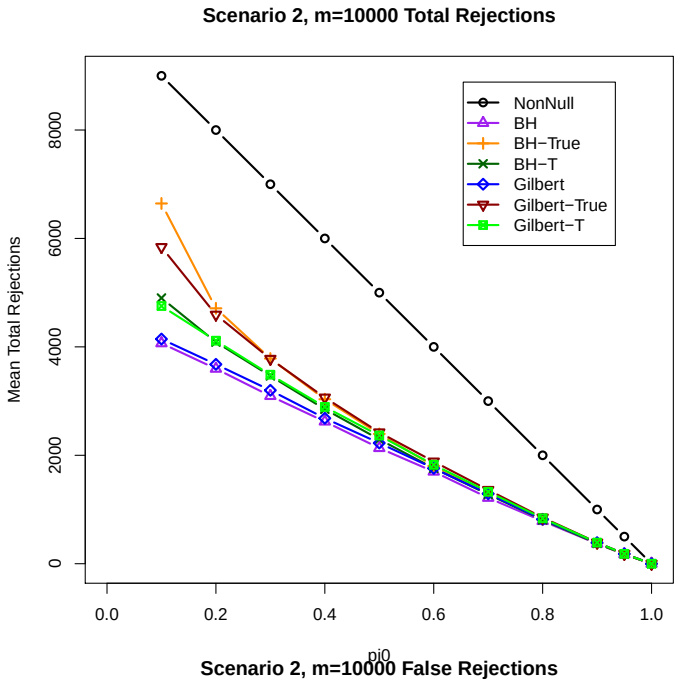


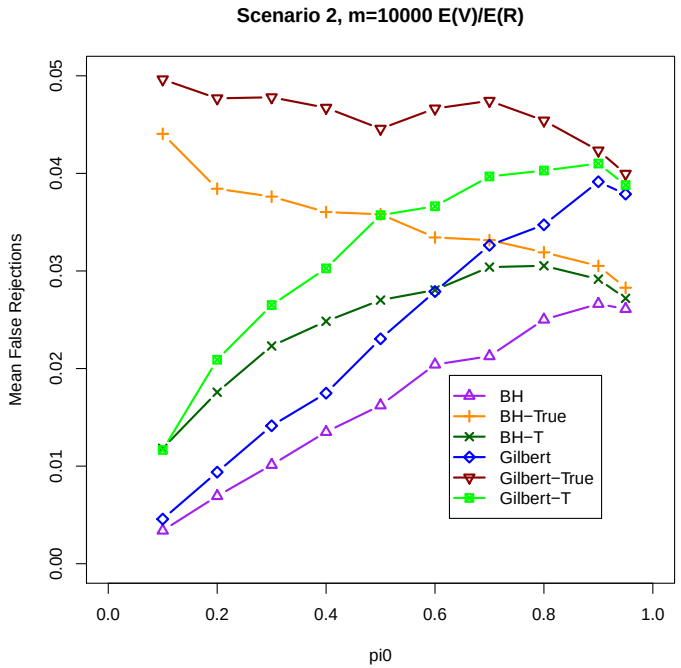
Results m=10,000 few small margins

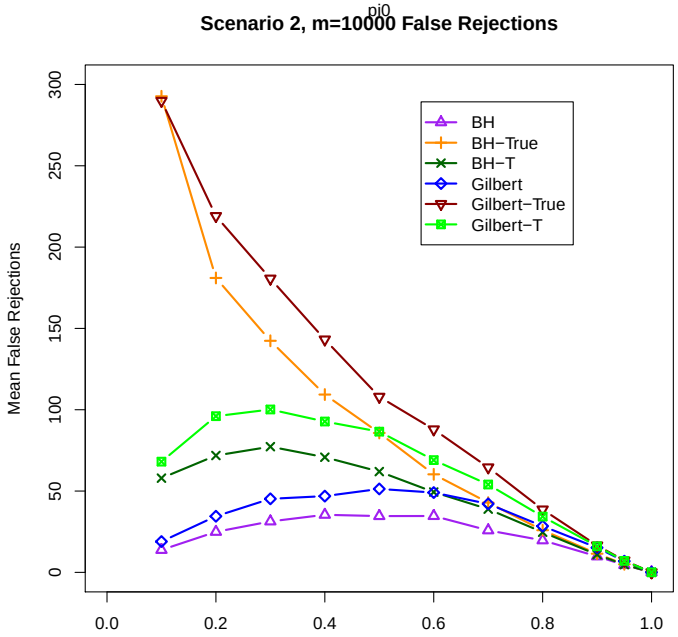




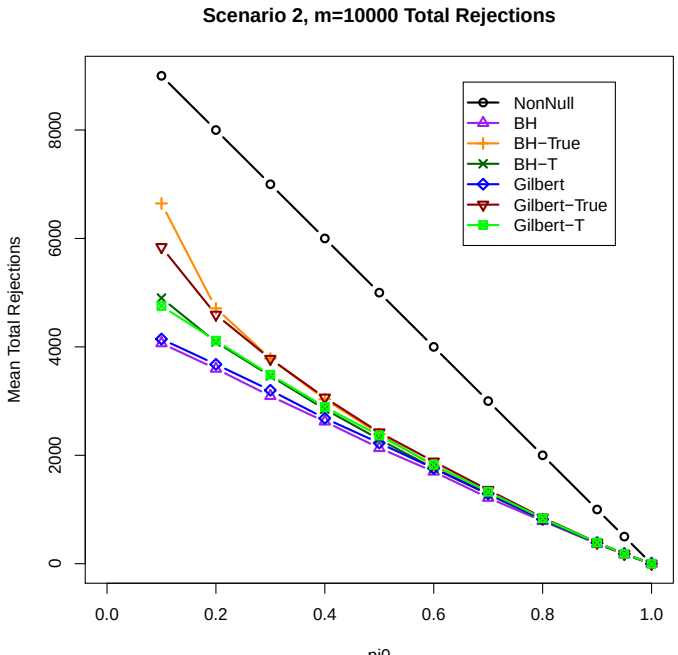
Results m=10,000 few small margins

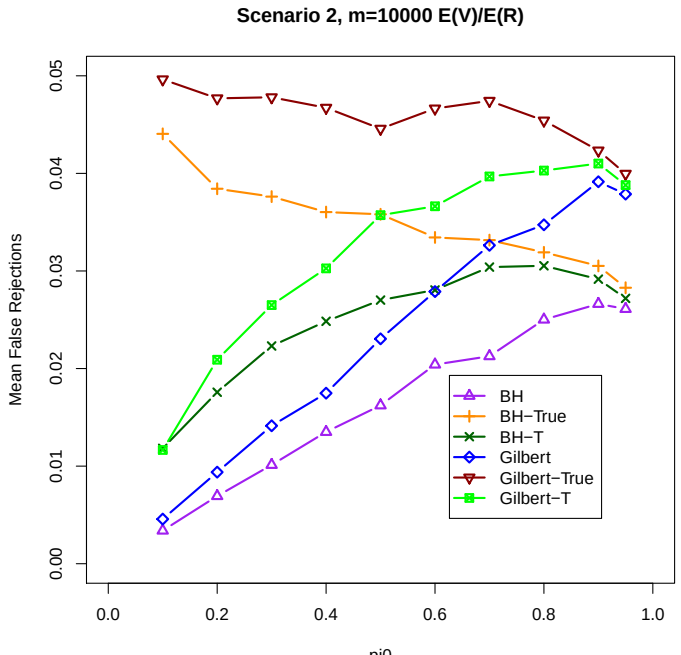


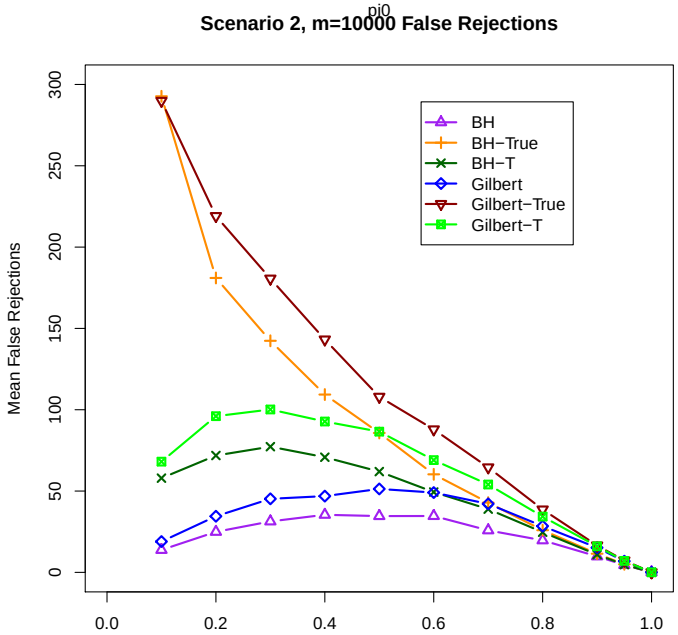


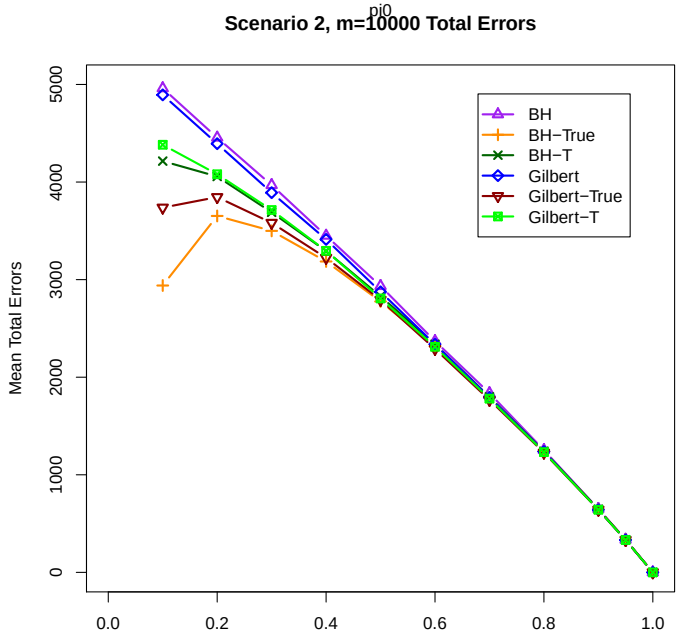


Results m=10,000 few small margins

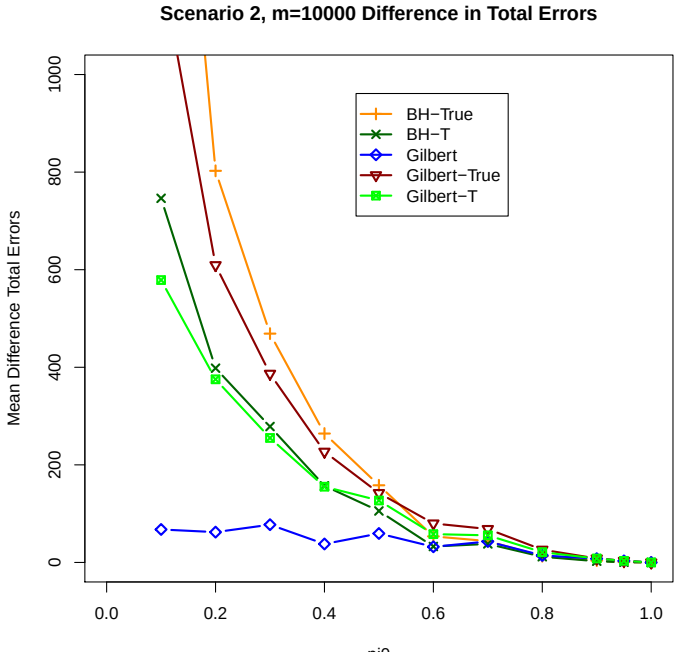


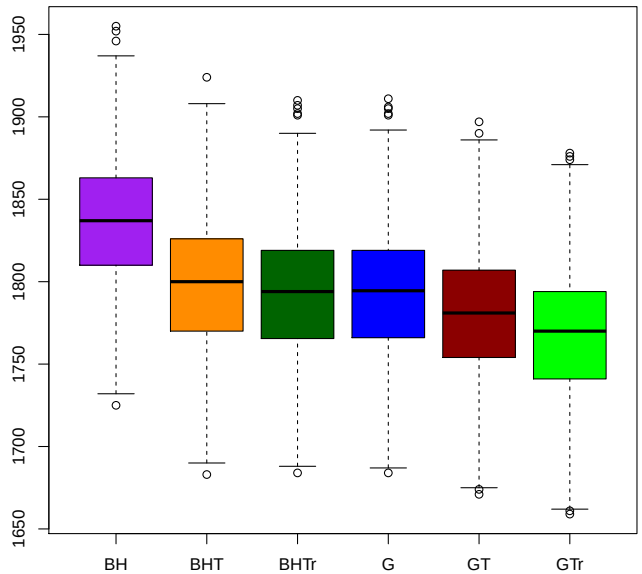






Does it matter?





Blekhman Primate Liver Data

Blekhman, et al, (2010) used RNA-seq to interrogate liver samples in male and female human, chimpanzee and rhesus monkey.

- There were 20689 features but only 16979 features had at least 2 reads.
- There were 3 biological samples for each species by gender combination.

We look at 2 comparisons:

Comparison	Test
2 lanes same human	Fisher's exact test
male human versus chimpanzee	moderated Negative Binomial test

Blekhman Primate Liver Data

Blekhman, et al, (2010) used RNA-seq to interrogate liver samples in male and female human, chimpanzee and rhesus monkey.

- There were 20689 features but only 16979 features had at least 2 reads.
- There were 3 biological samples for each species by gender combination.

We look at 2 comparisons:

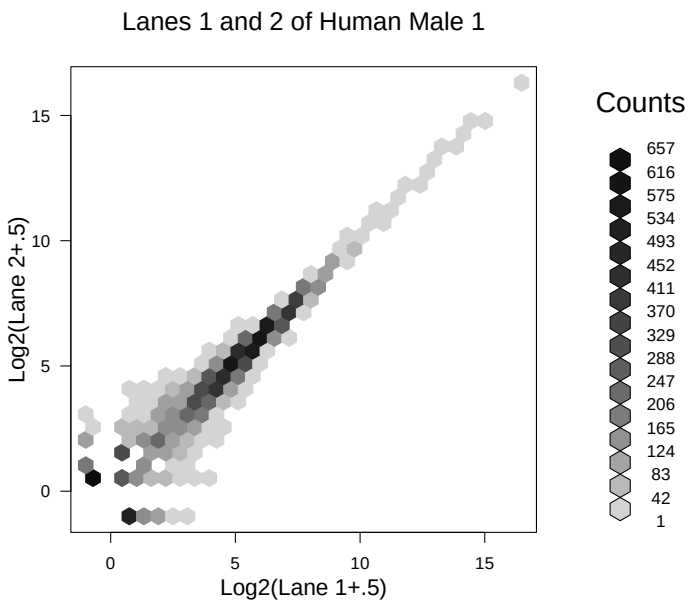
Comparison	Test
2 lanes same human	Fisher's exact test
male human versus chimpanzee	moderated Negative Binomial test

- It is difficult to compute expected counts for the moderated Negative Binomial test, so we use the T-method.
- We use Fisher's exact test to estimate the minimal achievable p-value. It is conservative.
- Analysis is done using `edgeR` in `Bioconductor`.

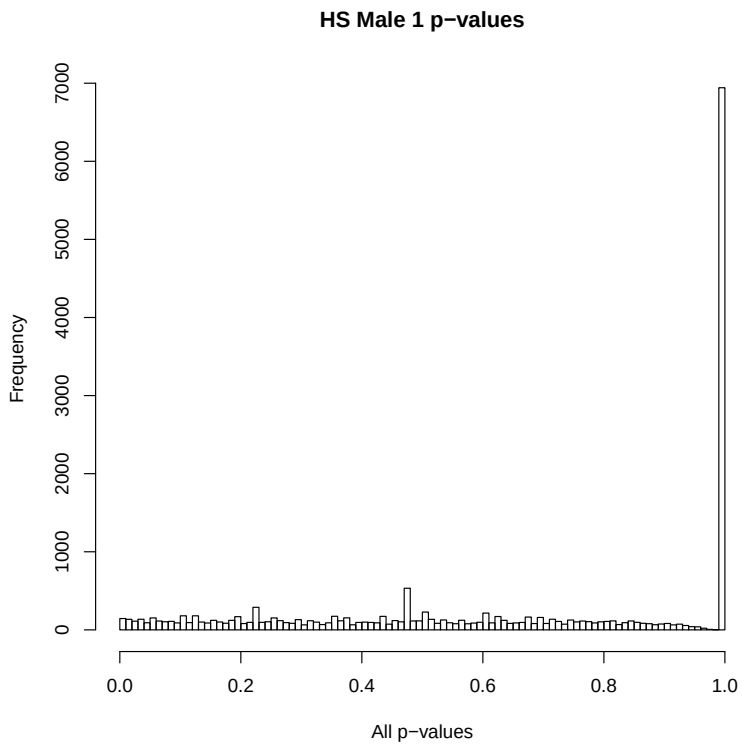
Human Male 1

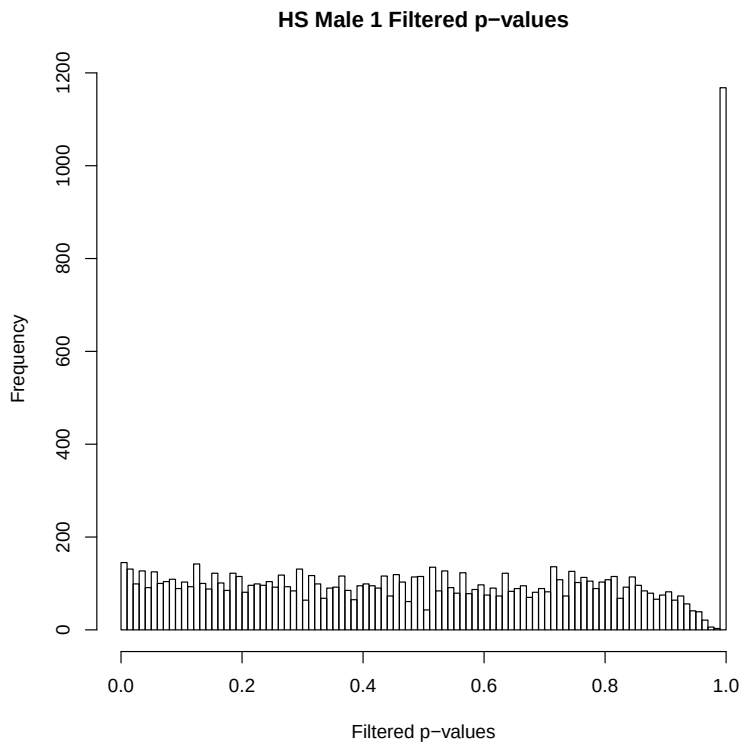
We compared the two lanes of sequencing data for Human male 1.

- 13553 features were detected by at least 1 read.
- 10359 features were detected by at least 7 reads (giving minimum achievable p-value>0.01.)
- We do not expect any differences between the two lanes.



Human Male 1



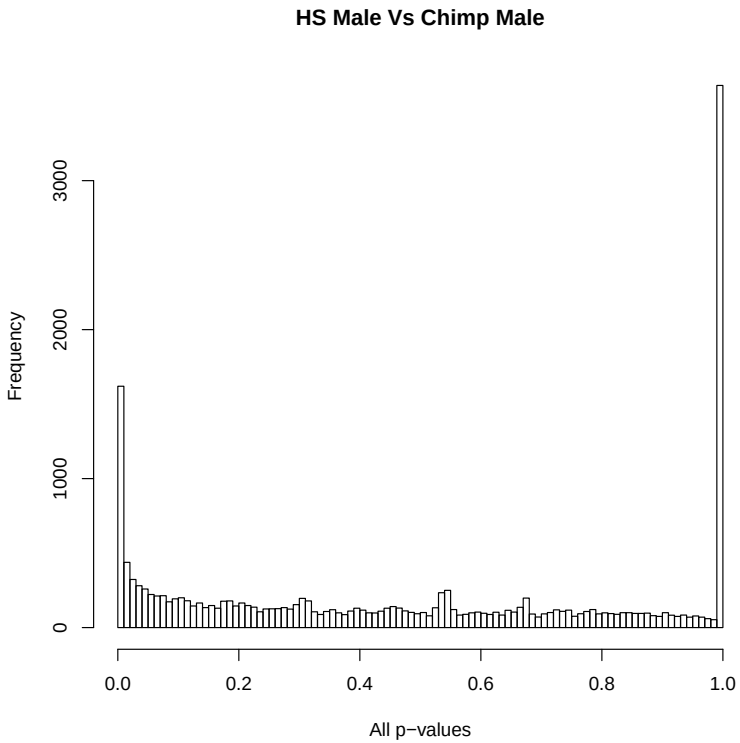


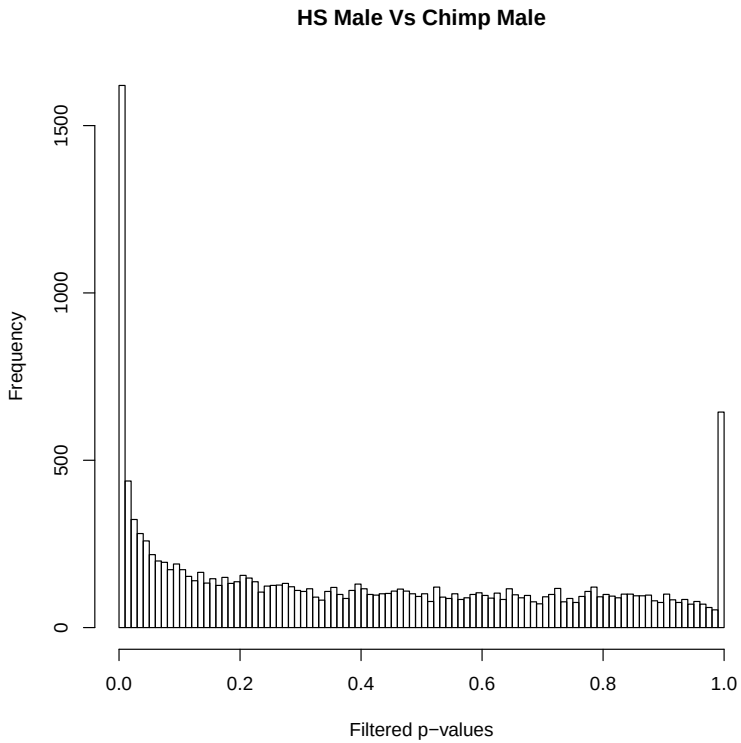
$\hat{\pi}_0 = 1.0$ using both Storey's method and the Storey-T method.

Method	Number Significant FDR<0.05
Benjamini & Hochberg	3
Gilbert	5

Human Males versus Chimpanzee Males

- The data were normalized, and the dispersion shrinkage factors were computed.
- There are 3 biological replicates of each.
- 16375 features were detected with at least 1 read.
- 13809 features were detected with at least 7 reads.





Human Males versus Chimpanzee Males

$\hat{\pi}_0 = 1.0$ using Storey's method and 0.87 using the Storey-T method.

Method	Number Significant Non-adaptive	Number Significant Adaptive
Benjamini & Hochberg	1166	1239
Gilbert	1251	1325

Note: $1166/16375 = 7.1\%$ so $\pi_0 = 1$ is not reasonable.

FDR for Discrete Tests

What did we learn?

For tabular count data (e.g. RNA-seq, SNP)

- It is important to have an estimate of π_0.
  - When most counts are big, remove features with small margins and use methods for continuous data.
  - When many counts are small, use the regression method.

FDR for Discrete Tests

What did we learn?

For tabular count data (e.g. RNA-seq, SNP)

- It is important to have an estimate of π_0.
  - When most counts are big, remove features with small margins and use methods for continuous data.
  - When many counts are small, use the regression method.
- Adaptive methods
  - do not help much for $\pi_0 > 0.9$.
  - can significantly reduce total errors when $\pi_0 < 0.5$

FDR for Discrete Tests

What did we learn?

For tabular count data (e.g. RNA-seq, SNP)

- It is important to have an estimate of π_0.
  - When most counts are big, remove features with small margins and use methods for continuous data.
  - When many counts are small, use the regression method.
- Adaptive methods
  - do not help much for $\pi_0 > 0.9$.
  - can significantly reduce total errors when $\pi_0 < 0.5$
- Gilbert's method is preferable to vanilla BH.

FDR for Discrete Tests

What did we learn?

For tabular count data (e.g. RNA-seq, SNP)

- It is important to have an estimate of π_0.
  - When most counts are big, remove features with small margins and use methods for continuous data.
  - When many counts are small, use the regression method.
- Adaptive methods
  - do not help much for $\pi_0 > 0.9$.
  - can significantly reduce total errors when $\pi_0 < 0.5$
- Gilbert's method is preferable to vanilla BH.
  - Happily Gilbert's method is equivalent to removing features with small margins and then using BH.

FDR for Discrete Tests

What did we learn?

For tabular count data (e.g. RNA-seq, SNP)

- It is important to have an estimate of π_0.
  - When most counts are big, remove features with small margins and use methods for continuous data.
  - When many counts are small, use the regression method.
- Adaptive methods
  - do not help much for $\pi_0 > 0.9$.
  - can significantly reduce total errors when $\pi_0 < 0.5$
- Gilbert's method is preferable to vanilla BH.
  - Happily Gilbert's method is equivalent to removing features with small margins and then using BH.

Summary

For RNA-seq and SNP data, remove features with small margins and proceed as if p-values were continuous.

Irreproducible feature selection

Why don't match our results match yours?

Lets assume there are 10 thousand features.

- 2 careful labs replicate the same experiment.
- Both labs pick a sample size to achieve 80% power testing at level $\alpha = 0.05$.
- Both labs use the best possible methods and use FDR=0.05 to reject.

Shouldn't they get the same results?

Irreproducible feature selection

Matching results?

Suppose $\pi_0 = 0.90$.

Table : Results of testing

	$\alpha = 0.05$	FDR=0.05
Significance level	0.05	0.0015
Power	0.80	0.27
FDR	0.56	0.05
Total Discoveries	1250	283
Reproducible False Positives	2	0
Reproducible True Positives	640	73
% Reproducible	52%	26%

Irreproducible feature selection

Matching results?

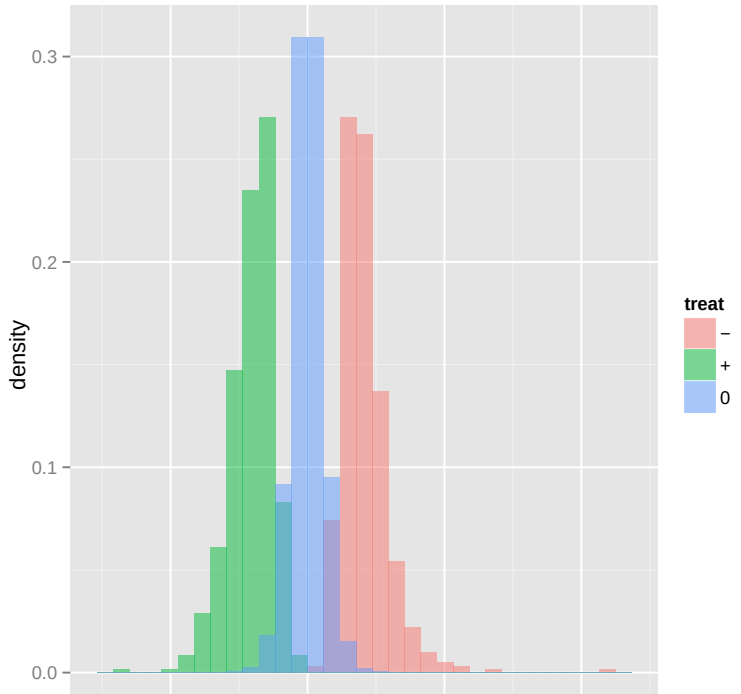
Of course, things are really more complex:

- Some of the features are correlated.
- Most studies are underpowered.
- Each feature has a different variance, so the power differs among the tests.
- There are systematic errors in the measurements so that some false positives and false negatives are more likely to reoccur.

The Screening Problem

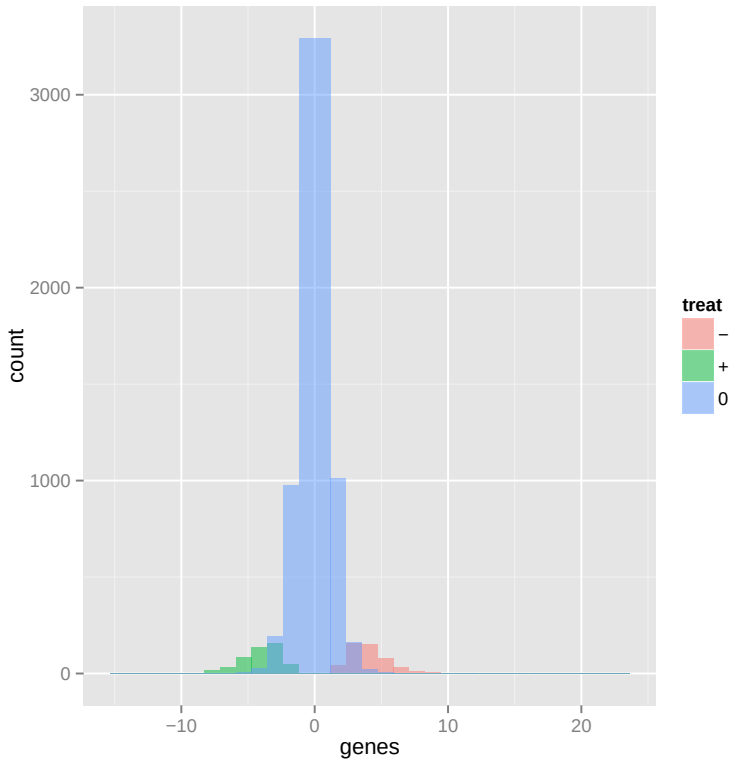
Irreproducibility

Most of us are aware of the FDR (and also the False Negative Rate). But that is not the only problem. To see the problem, lets consider some simulated data with $\pi_0 = 0.9$.
I set 5% of the data to be 2.35-fold up and 5% 2.35-fold down (90% power with sample size 5).



Irreproducibility

But it is not so clear when the histogram is plotted proportional to the number of features.



The Screening Problem

Screening with p-values

But that is not all, if we test at $\alpha = 0.05$:

Number Significant $p < 0.05$	1303
False Discoveries	409
False Nondiscoveries	106
FDP	31%

The Screening Problem

Screening with p-values

But that is not all, if we test at $\alpha = 0.05$:

Number Significant $p < 0.05$	1303
False Discoveries	409
False Nondiscoveries	106
FDP	31%

Screening with FDR

Using the Benjamini and Hochberg method to obtain FDR=0.05:

Number Significant $FDR < 0.05$	120
False Discoveries	9
False Nondiscoveries	889
FDP	0.075%

The estimated mean "effect" is 2.91 although the actual effect size was 2.35.

The Screening Problem

Screening with p-values

But that is not all, if we test at $\alpha = 0.05$:

Number Significant $p < 0.05$	1303
False Discoveries	409
False Nondiscoveries	106
FDP	31%

Screening with FDR

Using the Benjamini and Hochberg method to obtain FDR=0.05:

Number Significant $FDR < 0.05$	120
False Discoveries	9
False Nondiscoveries	889
FDP	0.075%

The estimated mean "effect" is 2.91 although the actual effect size was 2.35.

Correlation and Clustering?

Bloated Correlation

We often cluster selected features based on pairwise correlation.
For example:

- Select a set of differentially (expressed, bound, methylated) genes.
- Compute a pairwise (absolute) correlation.
- Cluster.