Statistics for Genomics (140.688)

Instructor: Jeff Leek

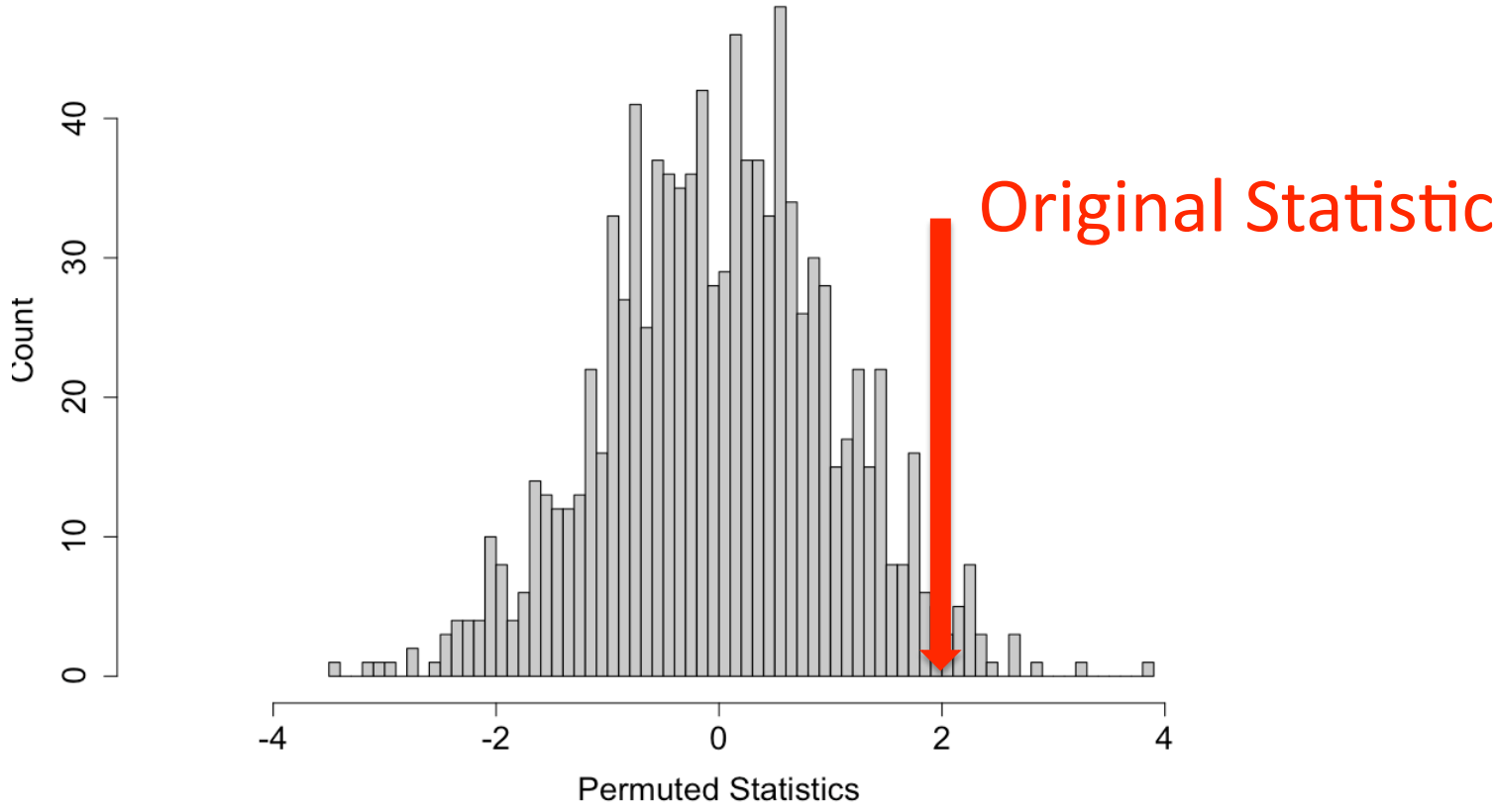
Slide Credits: Rafael Irizarry, John Storey

No announcements today.

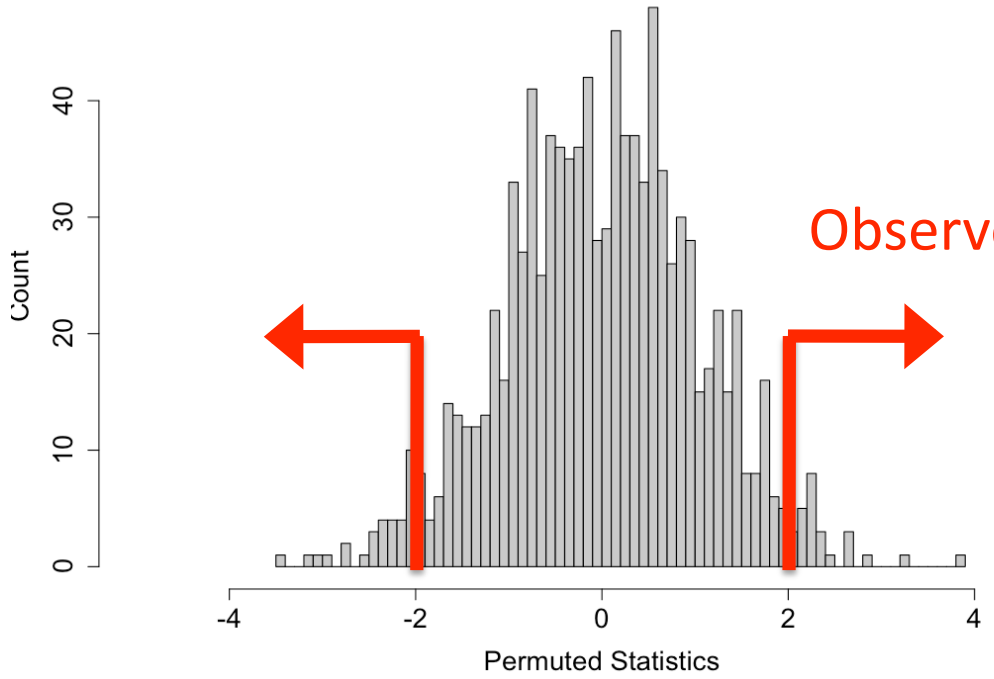
Hypothesis testing

- Once you have a given score for each gene, how do you decide on a cut-off?
- p-values are popular.
- But how do we decide on a cut-off?
- Are 0.05 and 0.01 appropriate?
- Are the p-values correct?

Recalculate the Statistic And Compare

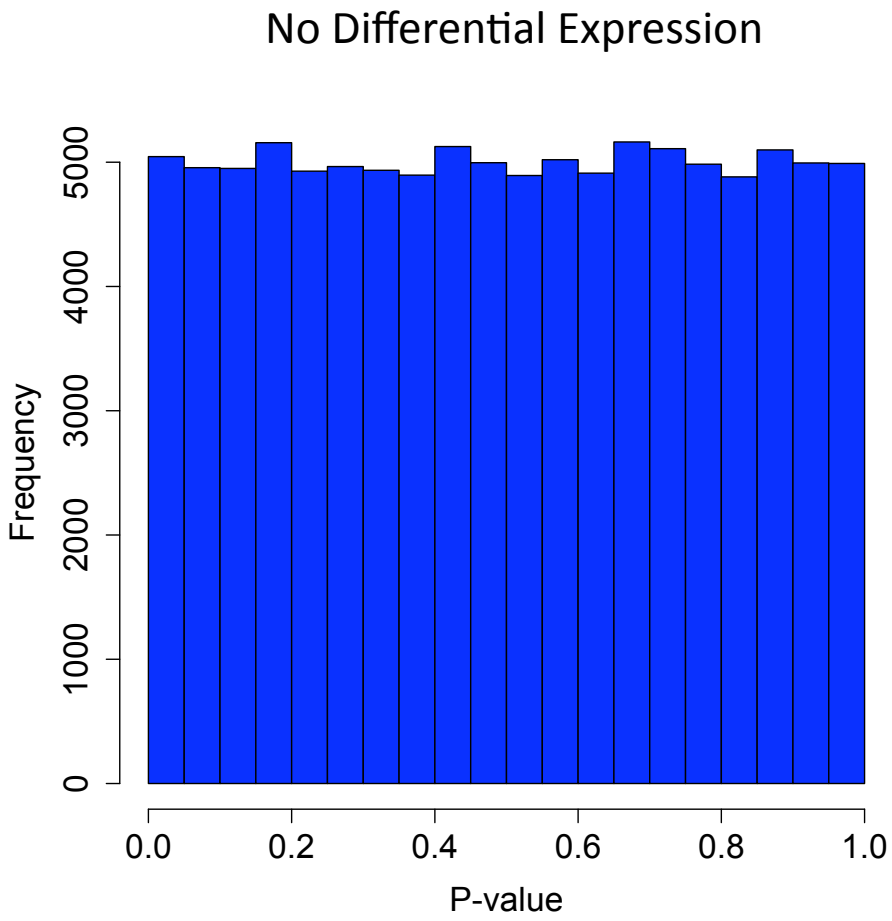


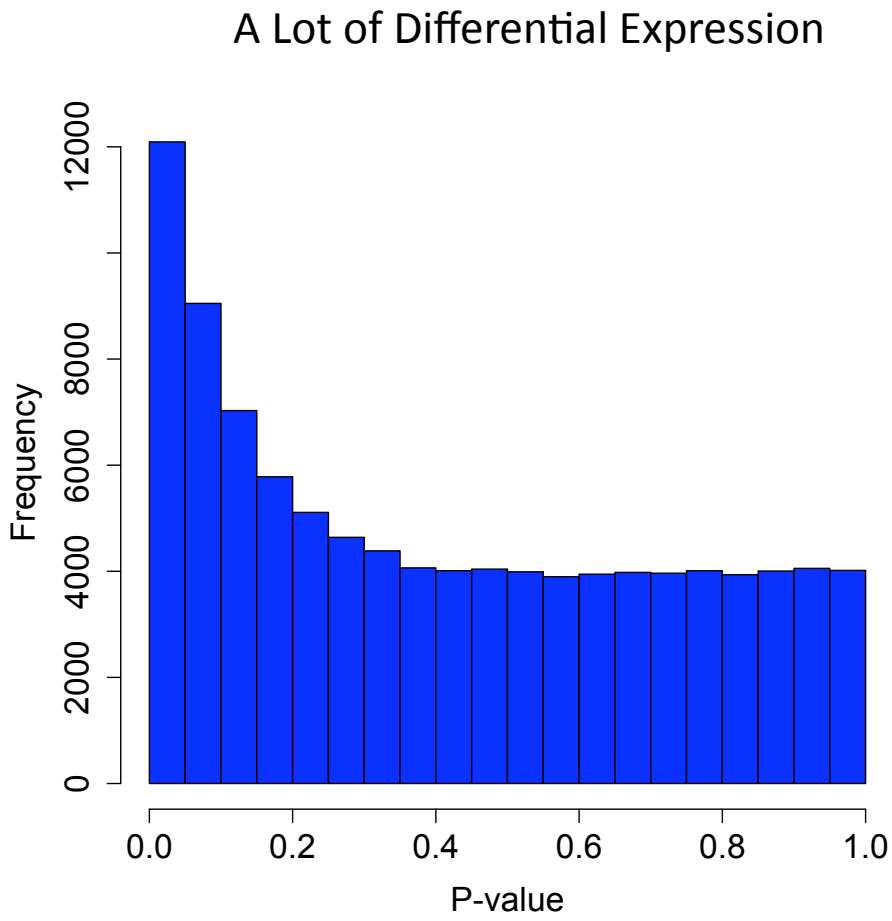
Calculating a P-value



$$\text{P-value} = \frac{\{\#\ |S^{perm}| \geq |S^{obs}|\}}{\#\ \text{of Permutations}}$$

P-values


No Differential Expression
Frequency
0
1000
2000
3000
4000
5000
0.0
0.2
0.4
0.6
0.8
1.0
P-value


A Lot of Differential Expression
Frequency
0
2000
4000
6000
8000
12000
0.0
0.2
0.4
0.6
0.8
1.0
P-value

Multiple Comparison Problem

- If we do have useful approximations of our p-values, we still face the multiple comparison problem

- When performing many independent tests p-values no longer have the same interpretation

Hypothesis Testing

- Test for each gene null hypothesis: no differential expression.

- Two types of errors can be committed
  - Type I error or false positive (say that a gene is differentially expressed when it is not, i.e., reject a true null hypothesis).
  - Type II error or false negative (fail to identify a truly differentially expressed gene, i.e.,fail to reject a false null hypothesis)

Hypothetical Example

- Microarray with 10,000 genes
- Calculate 10,000 p-values
- Call genes “significant” if p-value < 0.05
- Expected Number of False Positives:

$$10{,}000 \times 0.05 = 500 \text{ False Positives}$$

Multiple Hypothesis Testing

- What happens if we call all genes significant with p-values ≤ 0.05, for example?

	Called Significant	Not Called Significant	Total
Null True	*V*	$m_0 - V$	m_0
Altern.True	*S*	$m_1 - S$	m_1
Total	*R*	$m - R$	m

Error Rates

- Per comparison error rate (PCER): the expected value of the number of Type I errors over the number of hypotheses

  $PCER = E(V)/m$

- Per family error rate (PFER): the expected number of Type I errors

  $PFER = E(V)$

- Family-wise error rate: the probability of at least one Type I error

  $FEWR = Pr(V \geq 1)$

- False discovery rate (FDR) rate that false discoveries occur

  $FDR = E(V/R; R>0) = E(V/R \mid R>0)Pr(R>0)$

- Positive false discovery rate (pFDR): rate that discoveries are false

  $pFDR = E(V/R \mid R>0)$.

Multiple Comparison Error Rates

- Family wise error rate:

$$\Pr(\#\ \text{False Positives} \geq 1)$$

- False discovery rate:

$$\mathrm{E}\left[\frac{\#\,\text{False Positives}}{\#\ \text{Of Discoveries}}\right]$$

False Discovery Rate

- The "false discovery rate" measures the proportion of false positives among all genes called significant:

$$\frac{\#\,\text{false positives}}{\#\,\text{called significant}} = \frac{V}{V+S} = \frac{V}{R}$$

- This is usually appropriate because one wants to find as many truly differentially expressed genes as possible with relatively few false positives
- The false discovery rate gives the rate at which further biological verification will result in dead-ends

False Positive Rate versus False Discovery Rate

- False positive rate is the rate at which truly null genes are called significant

$$\text{FPR} \approx \frac{\#\,\text{false positives}}{\#\,\text{truly null}} = \frac{V}{m_0}$$

- False discovery rate is the rate at which significant genes are truly null

$$\text{FDR} \approx \frac{\#\,\text{false positives}}{\#\,\text{called significant}} = \frac{V}{R}$$

Difference in Interpretation

Suppose 550 out of 10,000 genes are significant at 0.05 level

P-value < 0.05

Expect 0.05*10,000 = 500 false positives

False Discovery Rate < 0.05

Expect 0.05*550 = 27.5 false positives

Family Wise Error Rate < 0.05

The probability of at least 1 false positive ≤ 0.05

Controlling Error Rates

Corrections when doing m tests:

Bonferroni Correction (FWER)

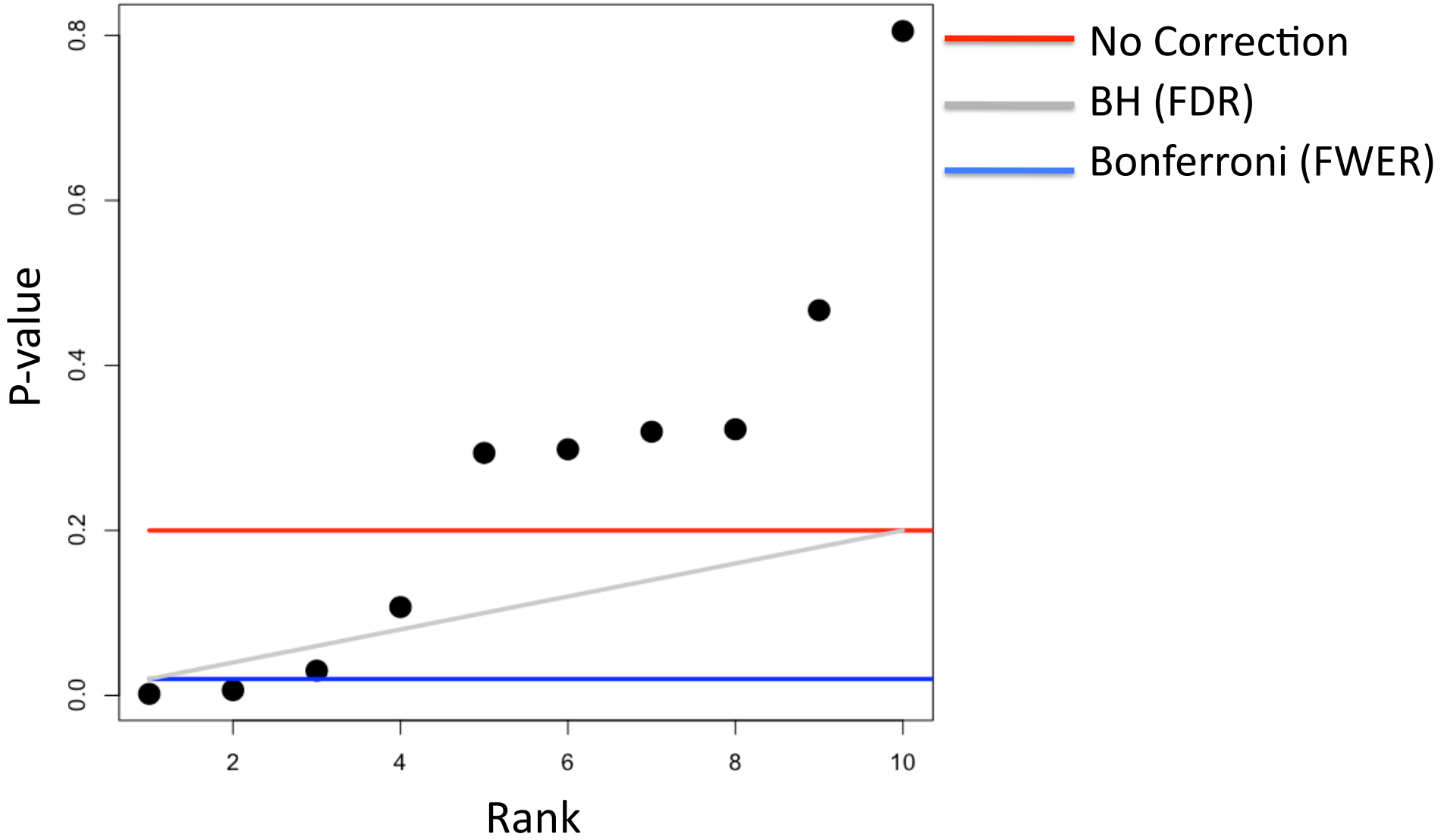
P-values less than α/m are significant

Benjamini-Hochberg Correction (FDR)

Order the p-values: $p_{(1)},\ldots,p_{(m)}$

If $p_{(i)} \leq \alpha \times i/m$ then it is significant

Example With 10 P-values



False Positive Rate and P-values

- The *p-value* is a measure of significance in terms of the false positive rate (aka Type I error rate)
- P-value is defined to be the minimum false positive rate at which the statistic can be called significant
- Can be described as the probability a truly null statistic is “as or more extreme” than the observed one

False Discovery Rate and Q-values

- The *q-value* is a measure of significance in terms of the false discovery rate
- Q-value is defined to be the minimum false discovery rate at which the statistic can be called significant
- Can be described as the probability a statistic "as or more extreme" is truly null

Estimate of FDR

- We begin by estimating FDR when calling all genes significant with p-values ≤ *t*
- *Heuristic* motivation:

$$\mathrm{FDR}(t) \approx \frac{\mathrm{E}[V(t)]}{\mathrm{E}[R(t)]} = \frac{\mathrm{E}[\#\{\text{null } p_i \le t\}]}{\mathrm{E}[\#\{p_i \le t\}]}$$

$=m_0 t$

$$\hat{\mathrm{FDR}}(t) = \frac{\hat{m}_0 \cdot t}{\#\{p_i \le t\}}$$

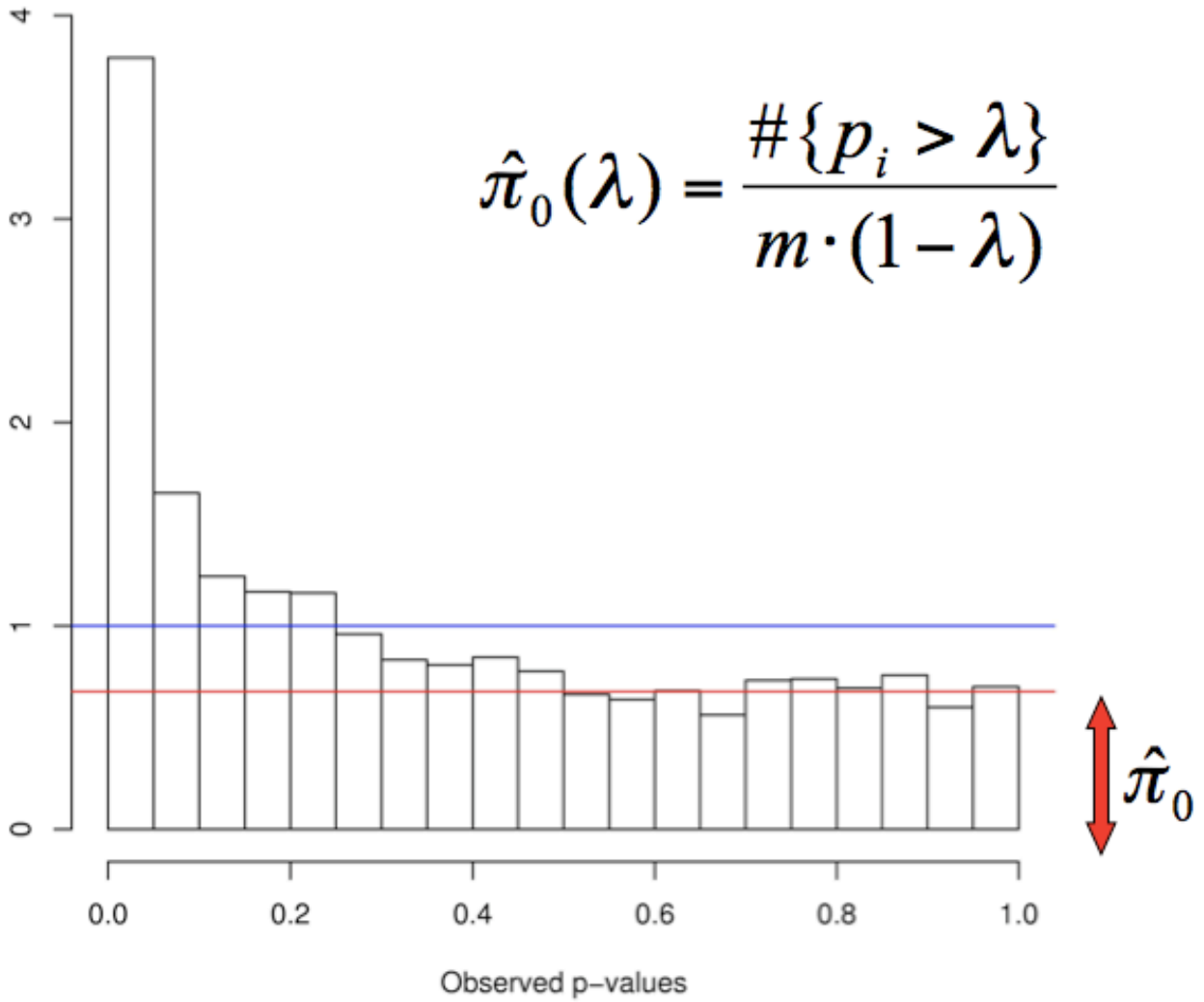
Estimate of π_0

- We first estimate the more easily interpreted $\pi_0 = m_0/m$, the proportion of truly null (non-differentially expressed) genes:

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

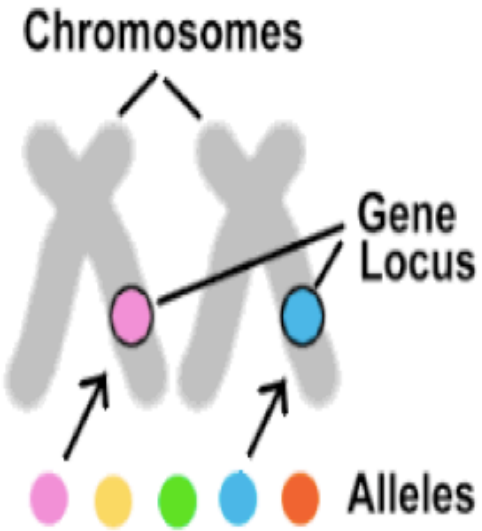
- Then clearly

$$\hat{m}_0 = \hat{\pi}_0 \cdot m$$


$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{m \cdot (1-\lambda)}$
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
Observed p−values
$\hat{\pi}_0$

Sources of Heterogeneity

External Factors (like environment)

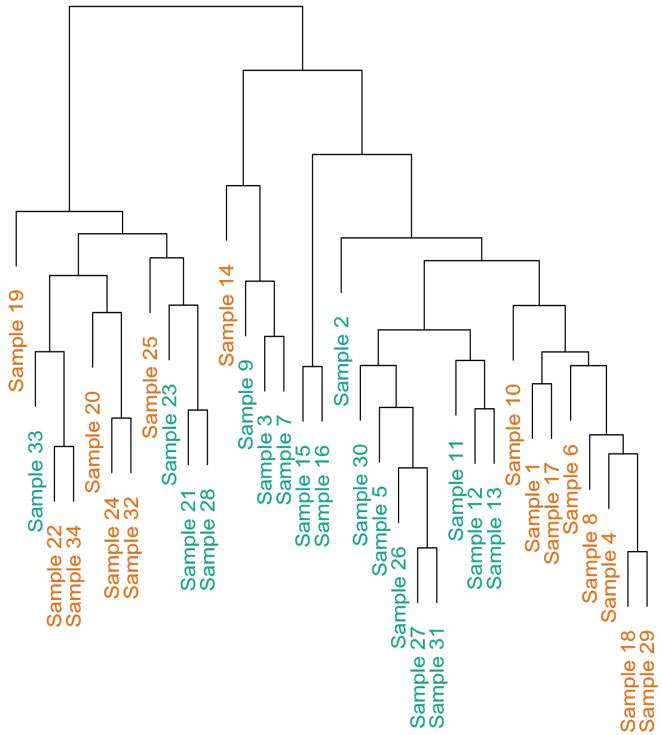


Genetics/Epigenetics

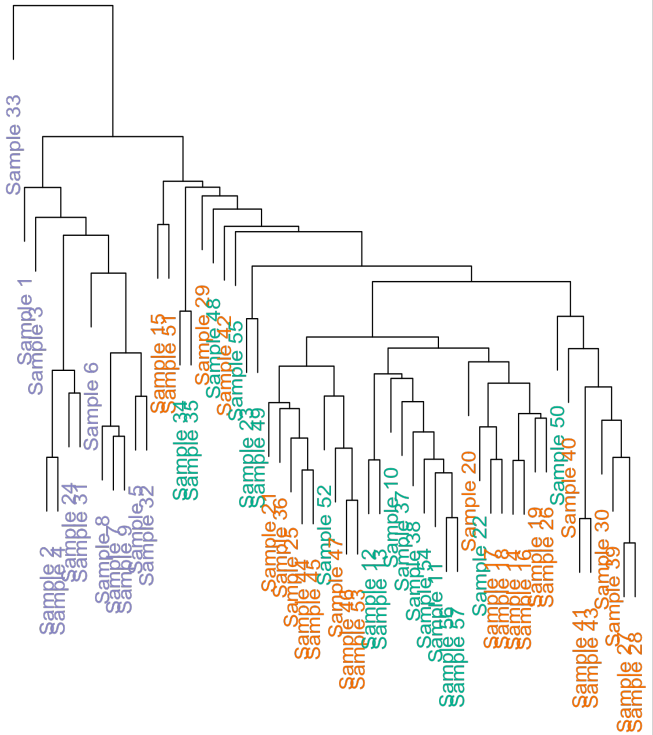
Technical Factors

The Effect of Heterogeneity

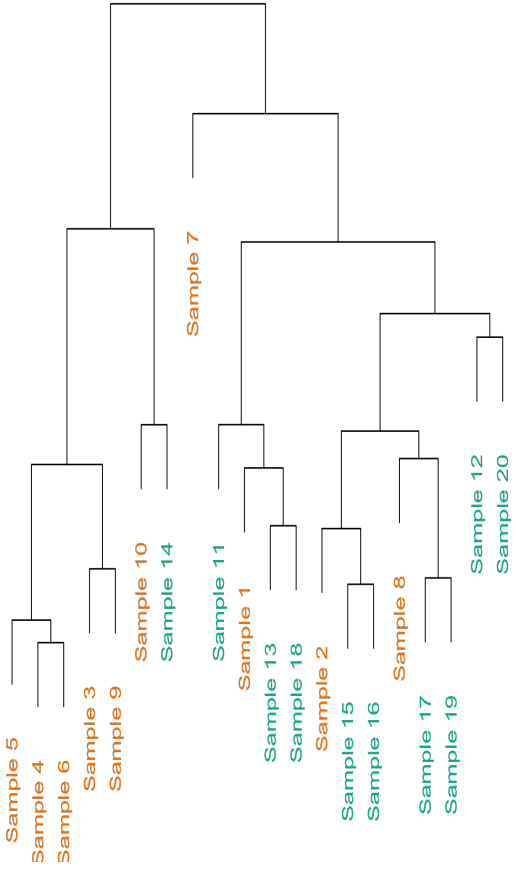
Color = Environment
(Idaghdour et al. 2008)



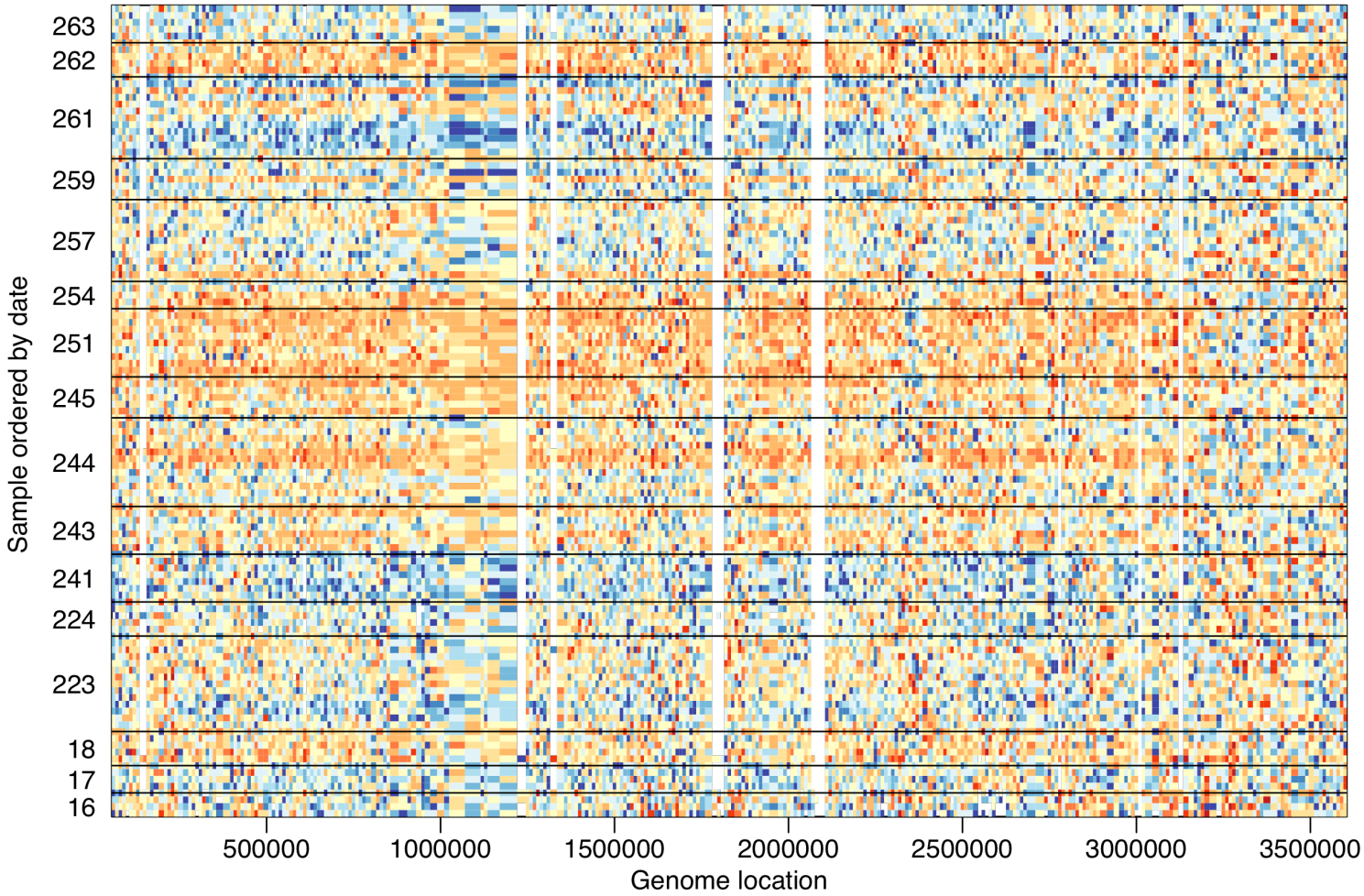
Color = Processing Year
(Cheung et al. 2008)



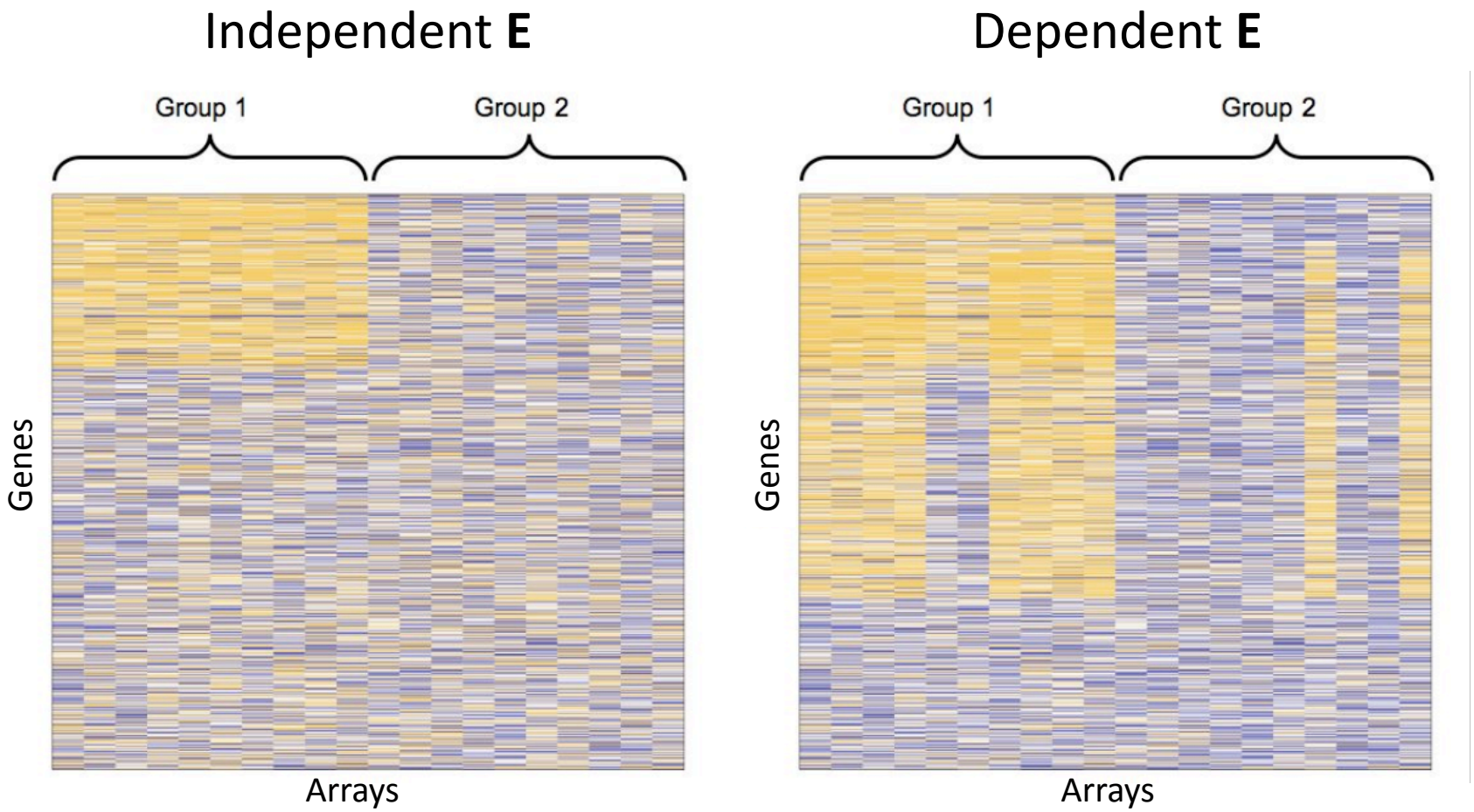
Color = Allele
(Brem et al. 2005)



Batch Effects in Sequencing



A Simple Simulated Example



Gene by Gene Model

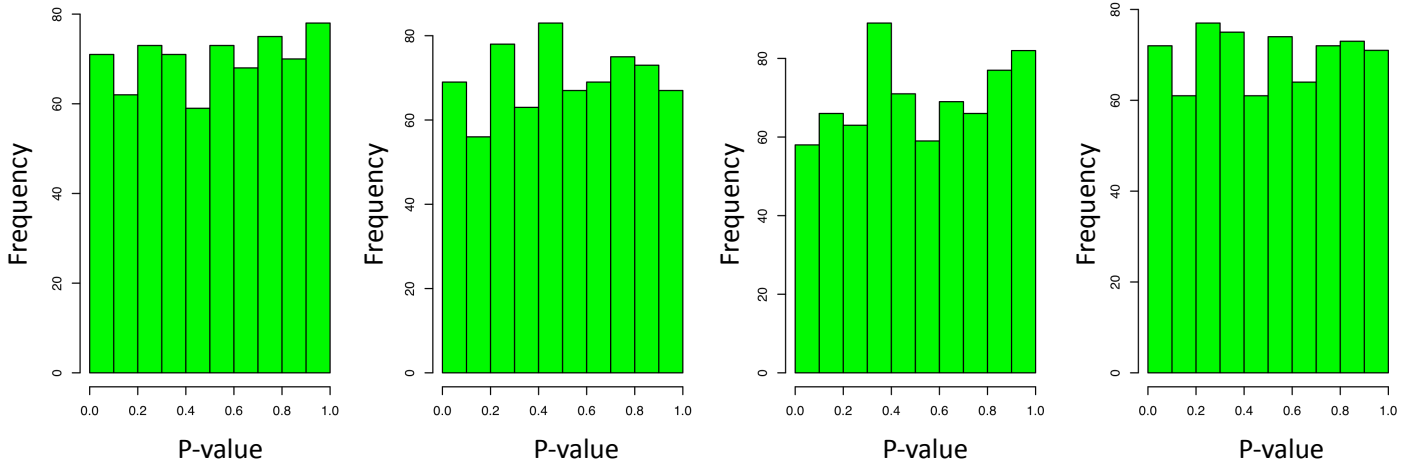
$$\text{expression} = b_0 + b_1 \times \text{group} + \text{noise}$$

Test whether $b_1 = 0$ <=> T-test for gene I

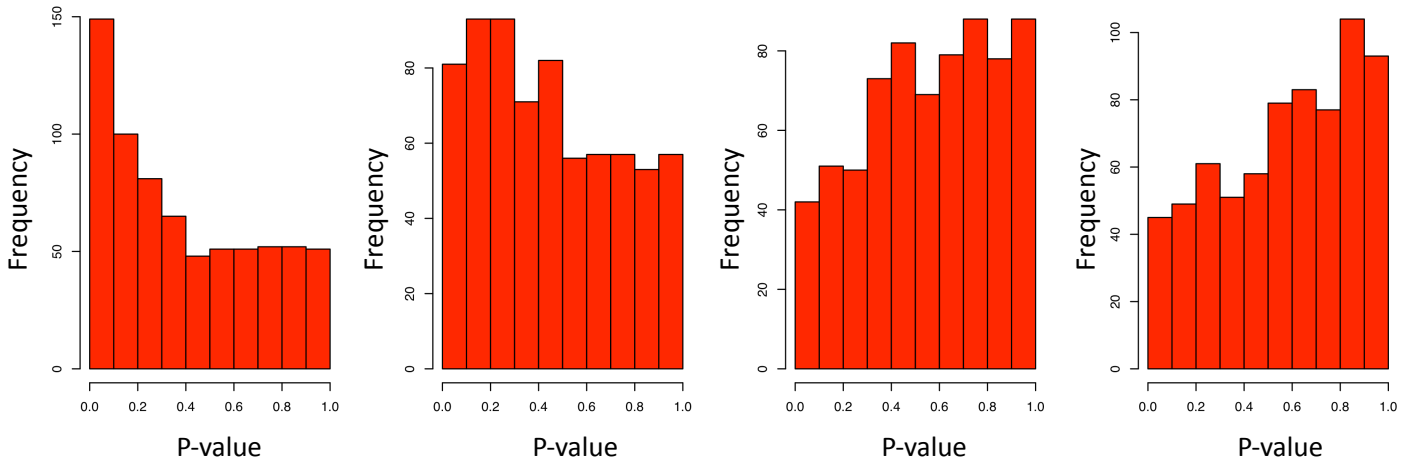
Calculate a P-value

Null P-Value Distributions

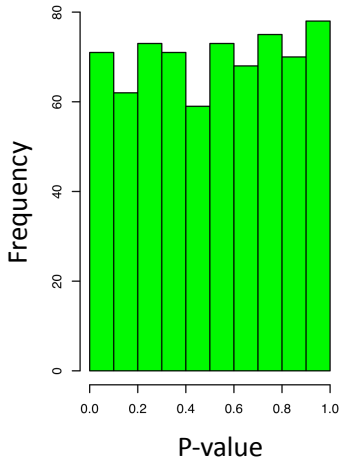
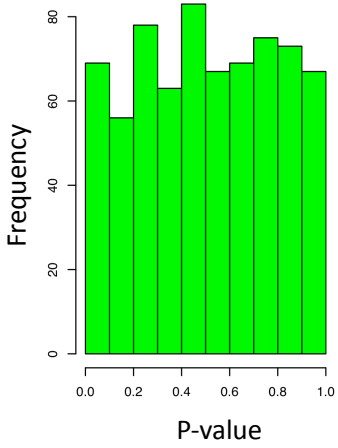
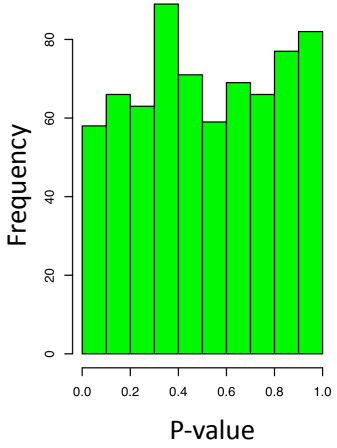
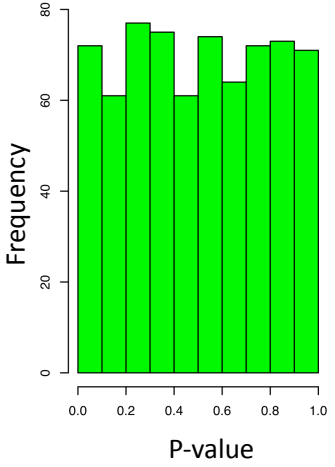
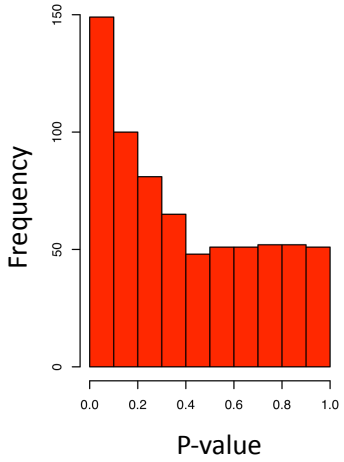
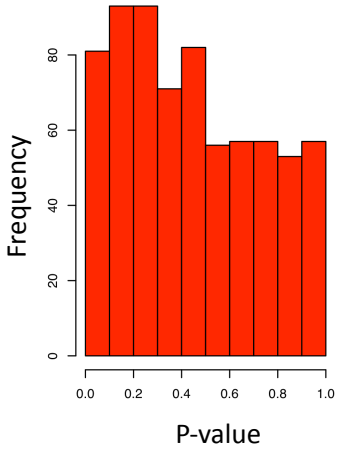
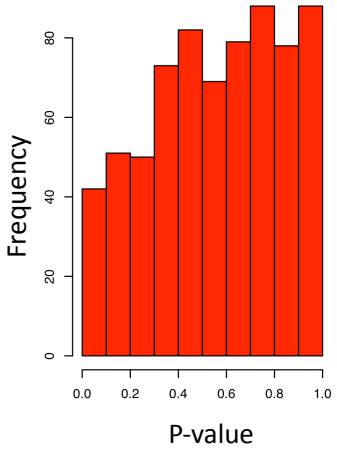
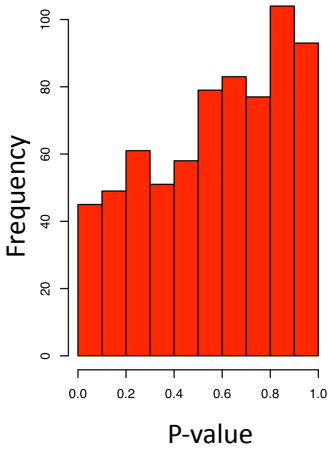
Independent **E**



Dependent **E**

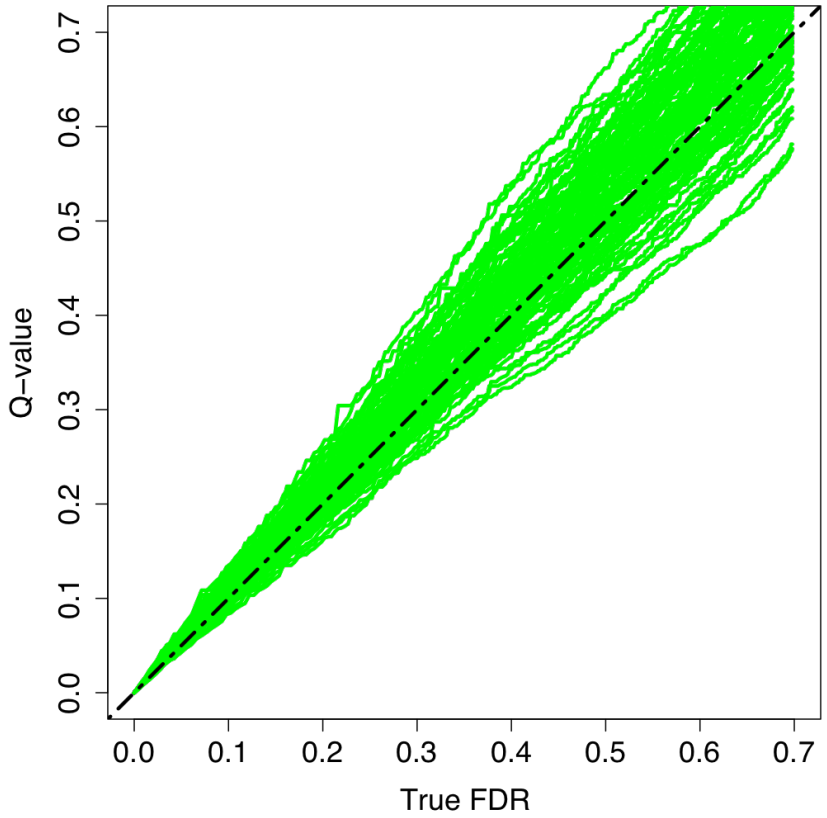


Null P-Value Distributions

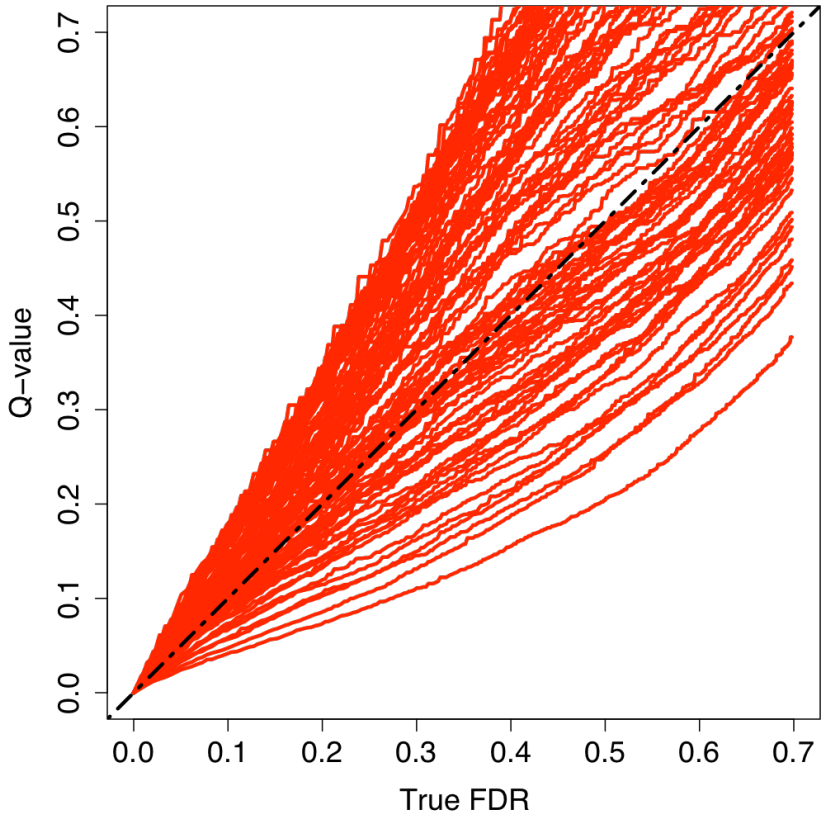
Correlation	$\lvert\rho\rvert = 0.40$	$\lvert\rho\rvert = 0.31$	$\lvert\rho\rvert = 0.10$	$\lvert\rho\rvert = 0.00$
Independent **E**				
Dependent **E**				

False Discovery Rate Estimates

Independent **E**

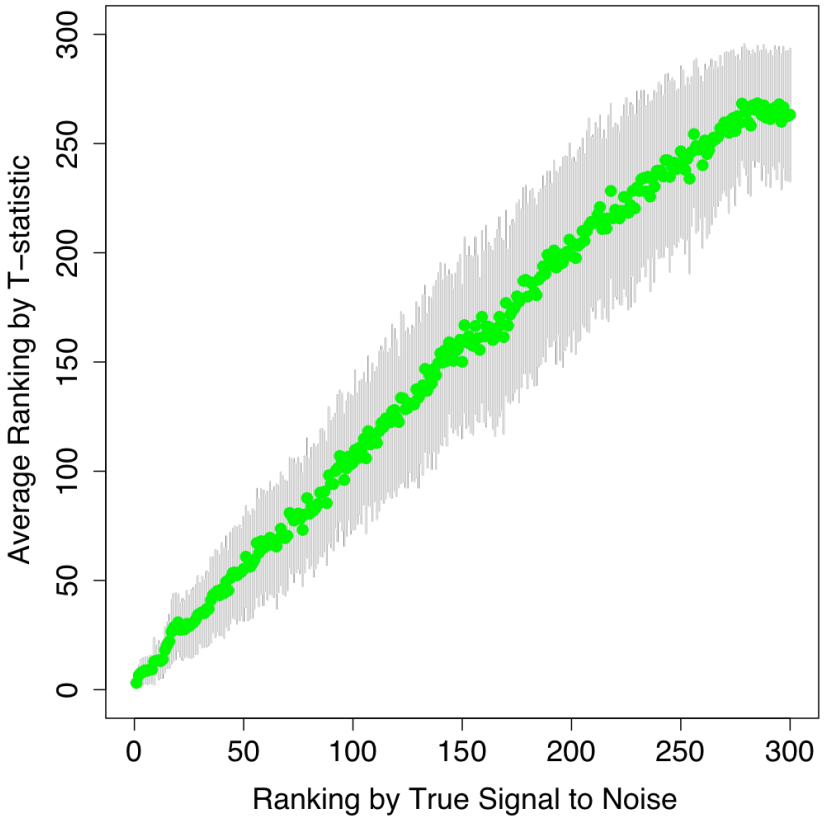


Dependent **E**

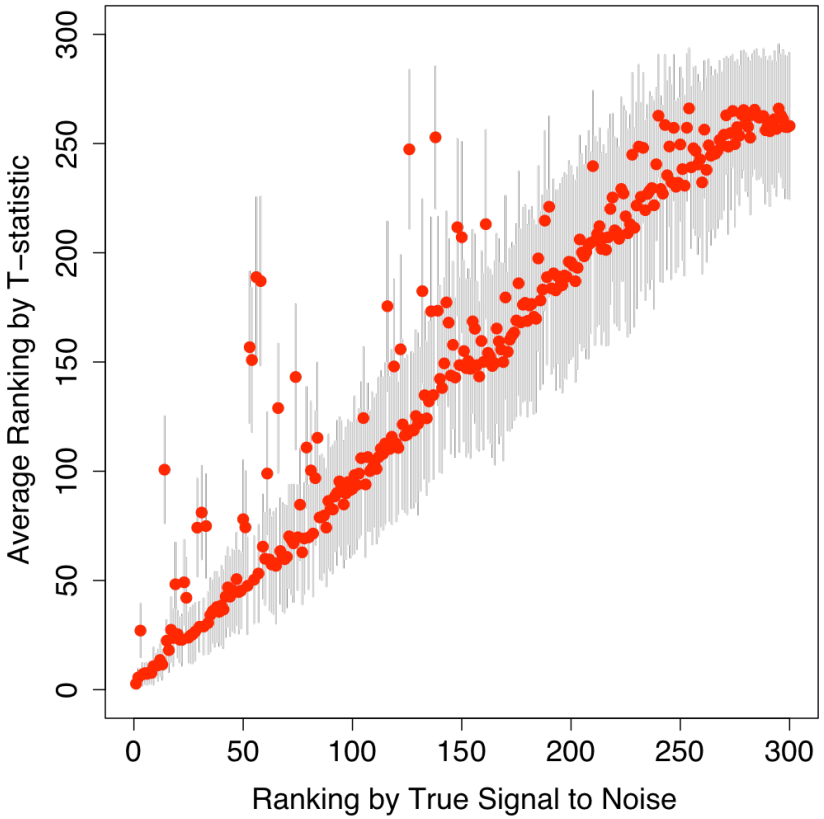


Ranking Estimates

Independent **E**



Dependent **E**



Batch and rankings

The Model

$$\text{expression} = b_0 + b_1 \times \text{group} + b_2 \text{ batch} + \text{noise}$$

Gene 1

$$\text{expression} = b_0 + 3 \times \text{group} + 10 \text{ batch} + \text{noise}$$

Gene 2

$$\text{expression} = b_0 + 1 \times \text{group} + 1 \text{ batch} + \text{noise}$$

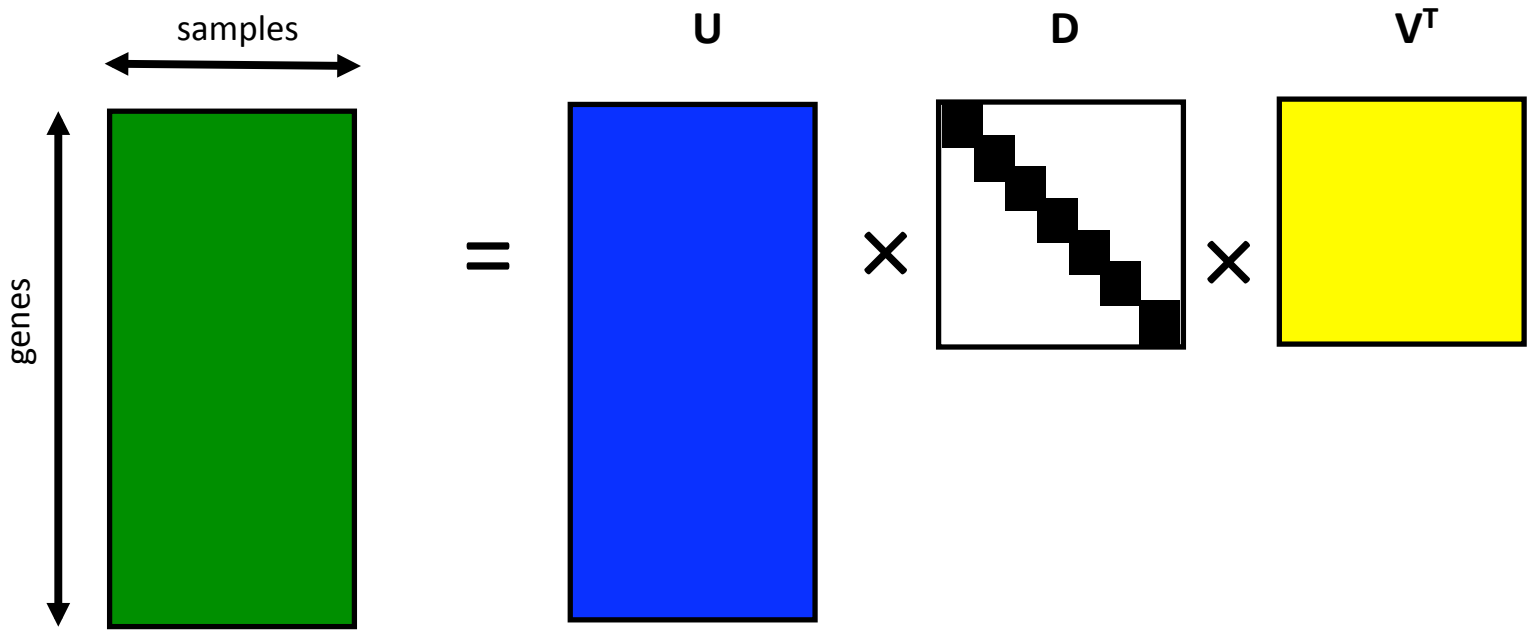
Principal Components Analysis / Singular Value Decomposition

- A method to identify patterns in the data that explain a large percentage of the variation

- PCA and SVD have different mathematical goals but end up estimating the same thing

- First proposed for genomics by Alter et al. (2000) PNAS

Singular Value Decomposition



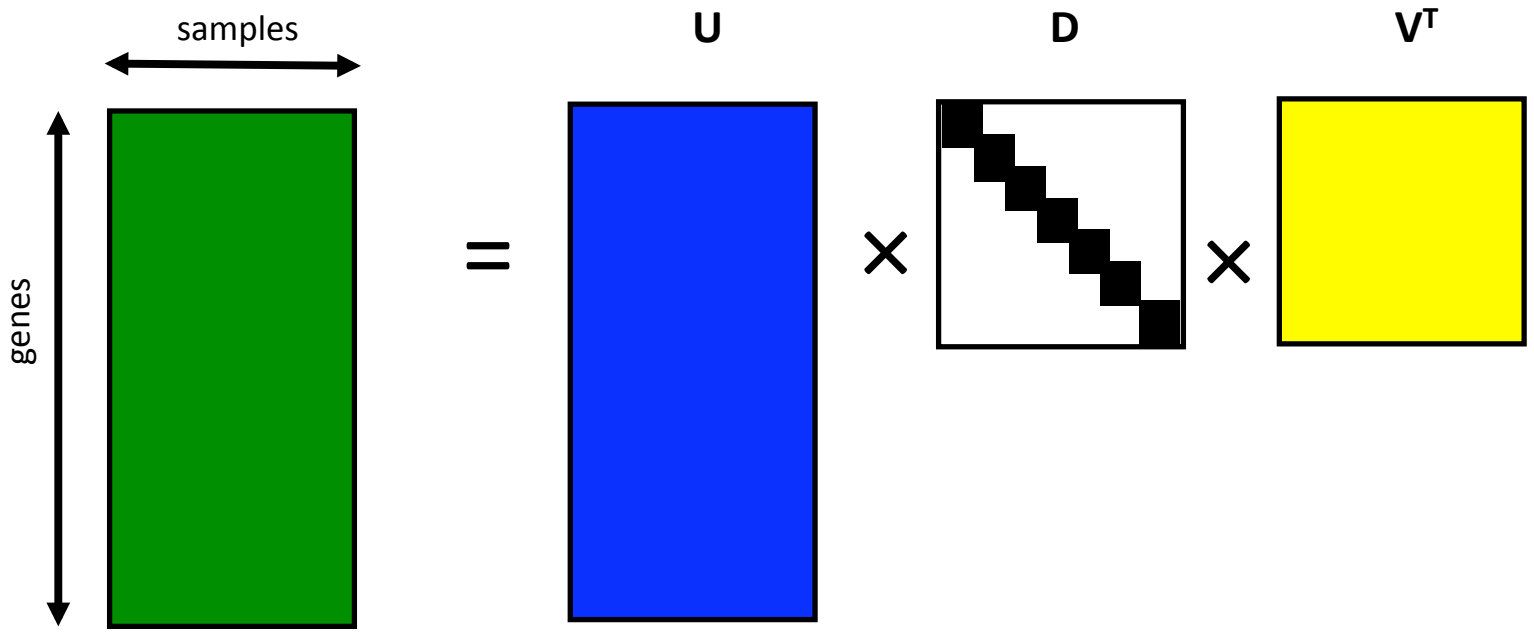
Data

eigenarrays/
left singular vectors/
loadings

singular
values

eigengenes/
right singular vectors/
principal components

Properties of SVD



Columns of V^T/rows of U are orthogonal and calculated one at a time

Columns of V^T describe patterns across genes

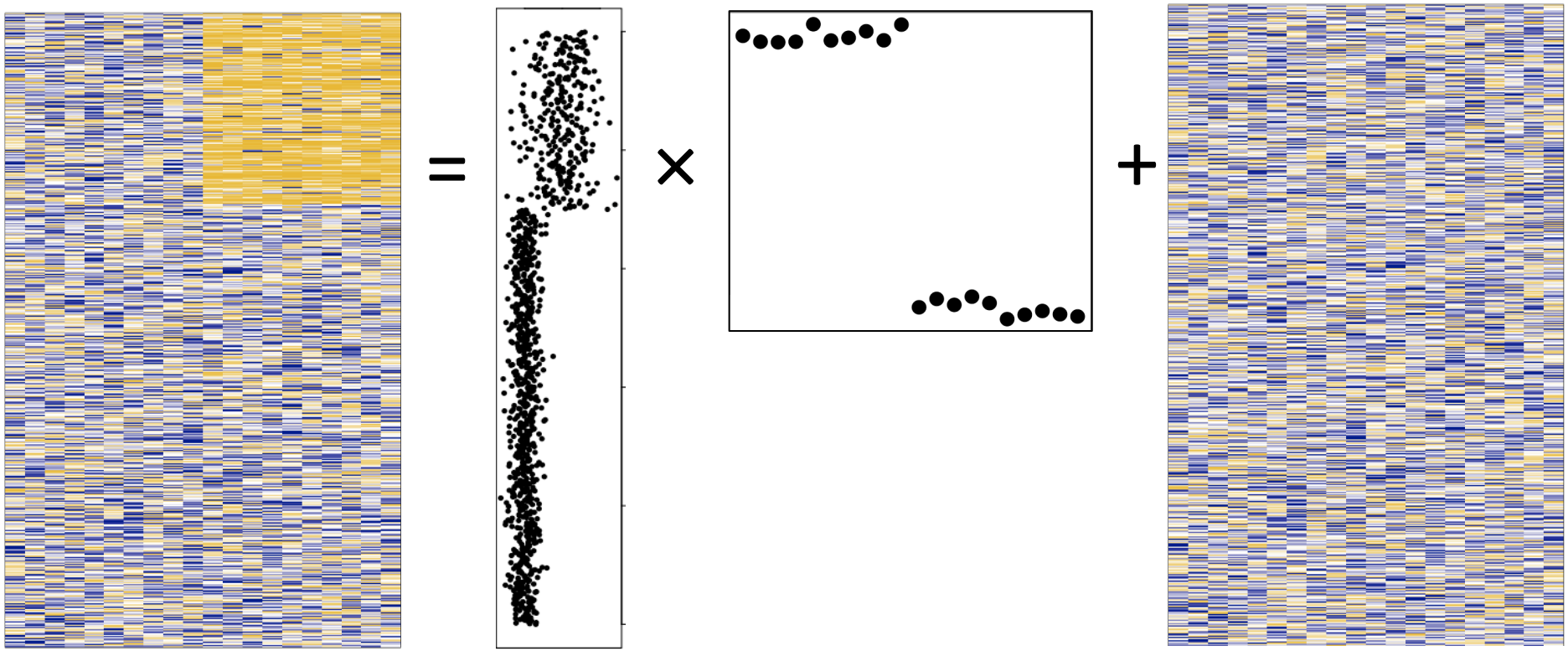
Columns of U describe patterns across arrays

$d_i^2 / \sum_{i=1}^{n} d_i^2$ is the percent of variation explained by the ith column of V

1 Pattern 1st SV

1st Column of U

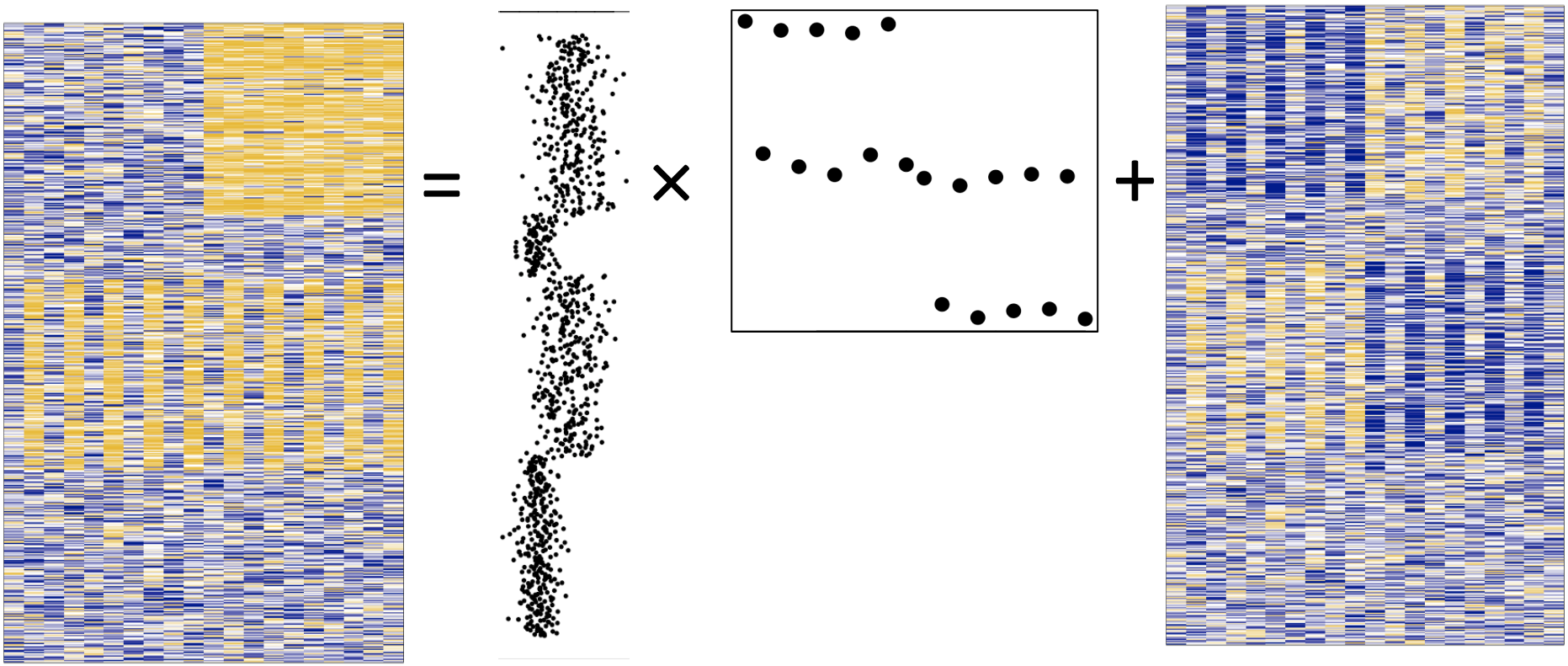
1st Column of V^T



2 Patterns, 1st SV

1st Column of U

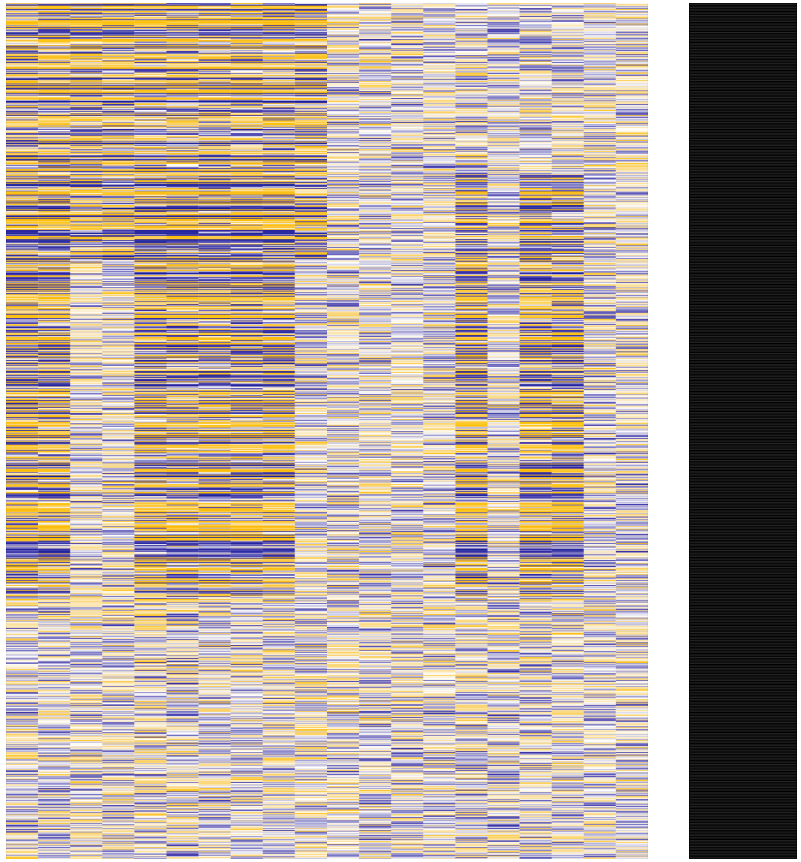
1st Column of V^T



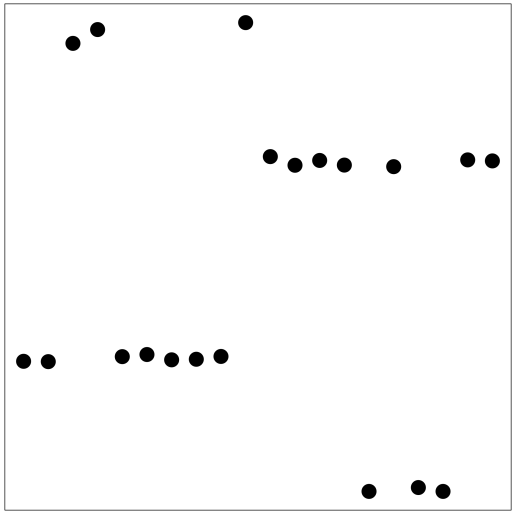
Surrogate Variable Analysis

The Data

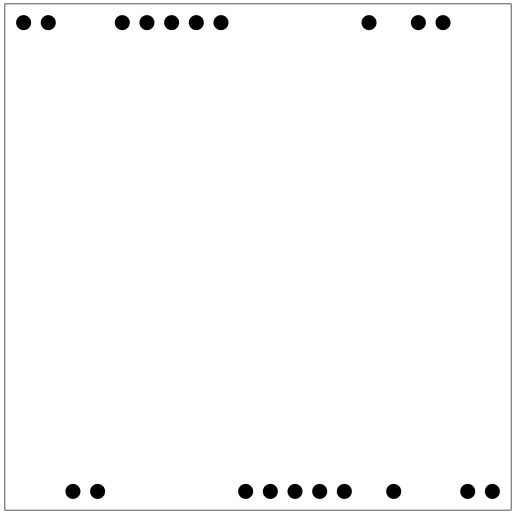
Pr(!**Group** & **Batch**)



Estimate of **Batch**



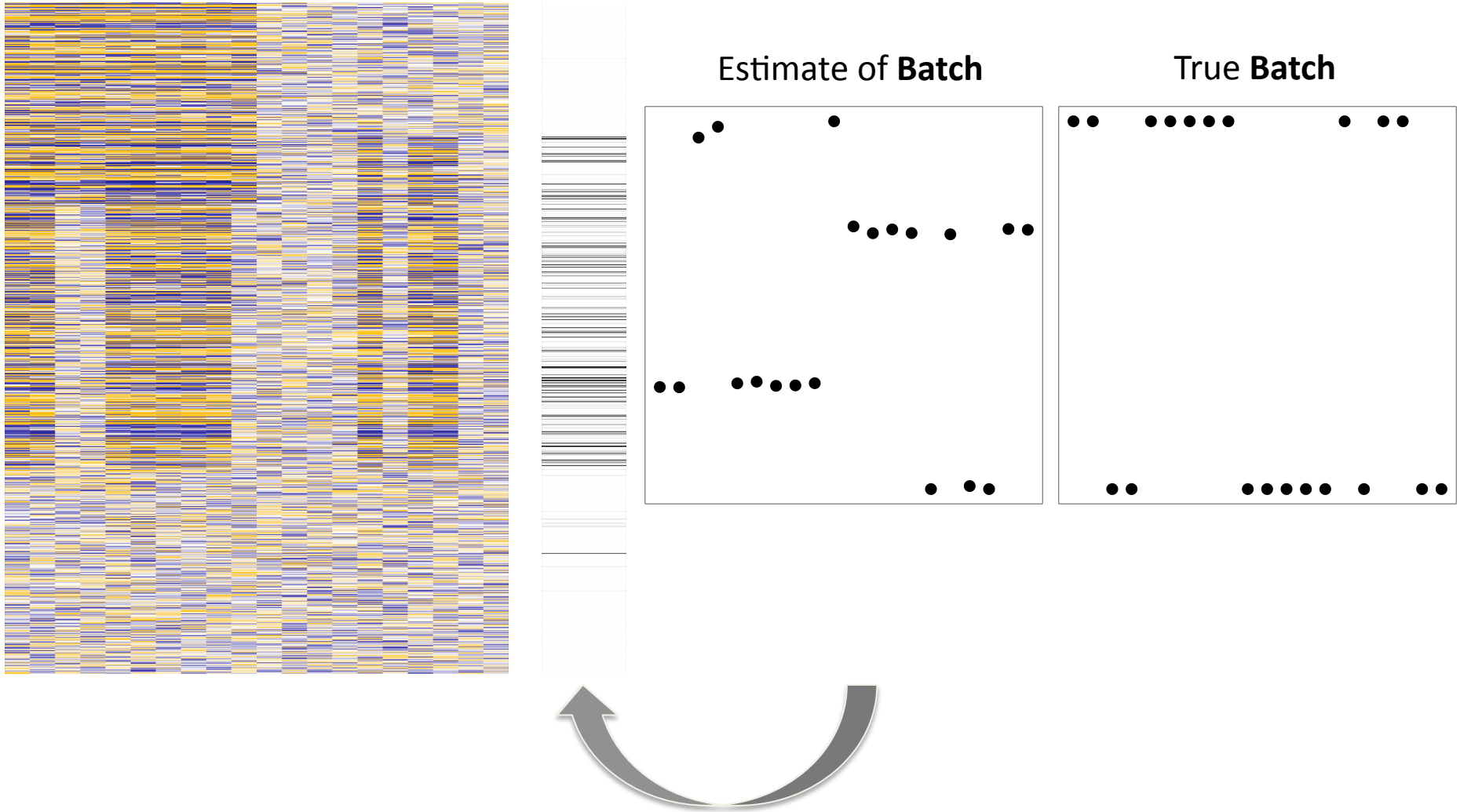
True **Batch**



Surrogate Variable Analysis

The Data

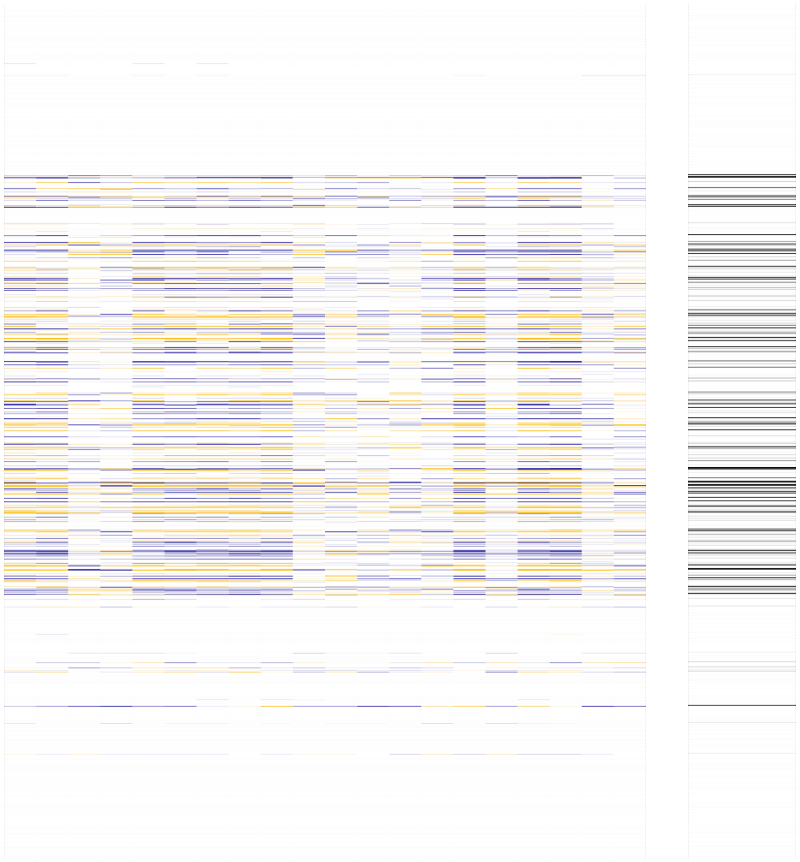
Pr(!**Group** & **Batch**)



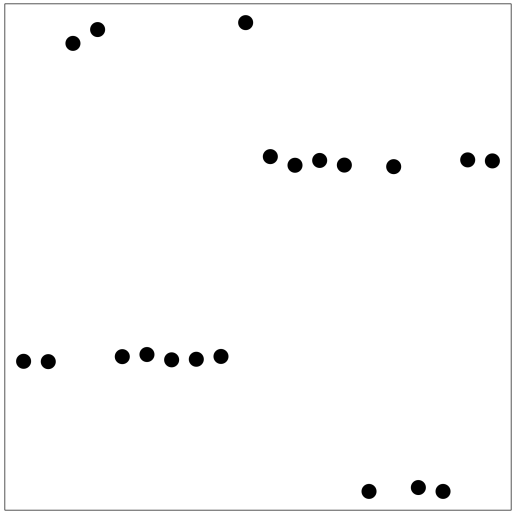
Surrogate Variable Analysis

The Data

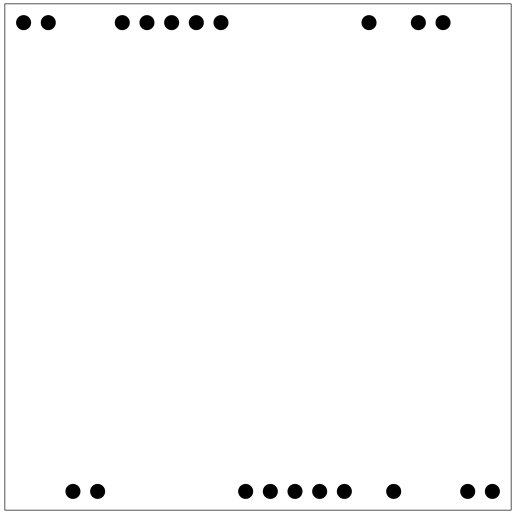
Pr(!**Group** & **Batch**)



Estimate of **Batch**



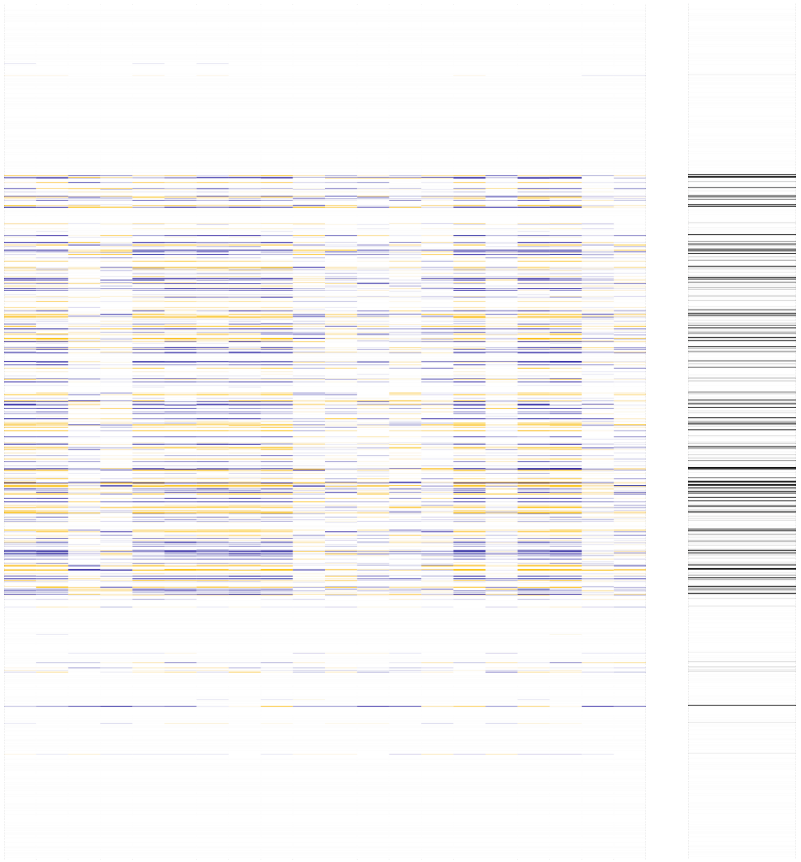
True **Batch**



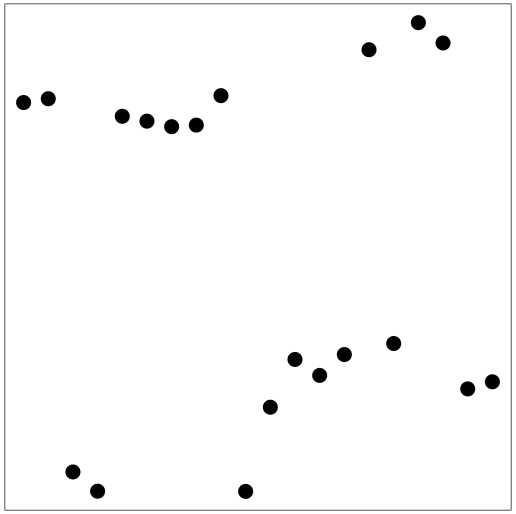
Surrogate Variable Analysis

The Data

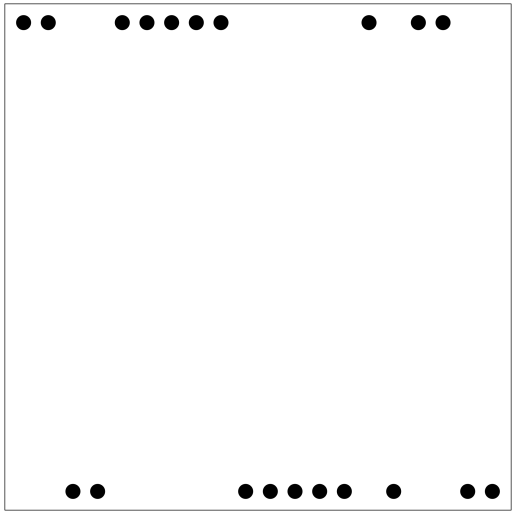
Pr(!**Group** & **Batch**)



Estimate of **Batch**



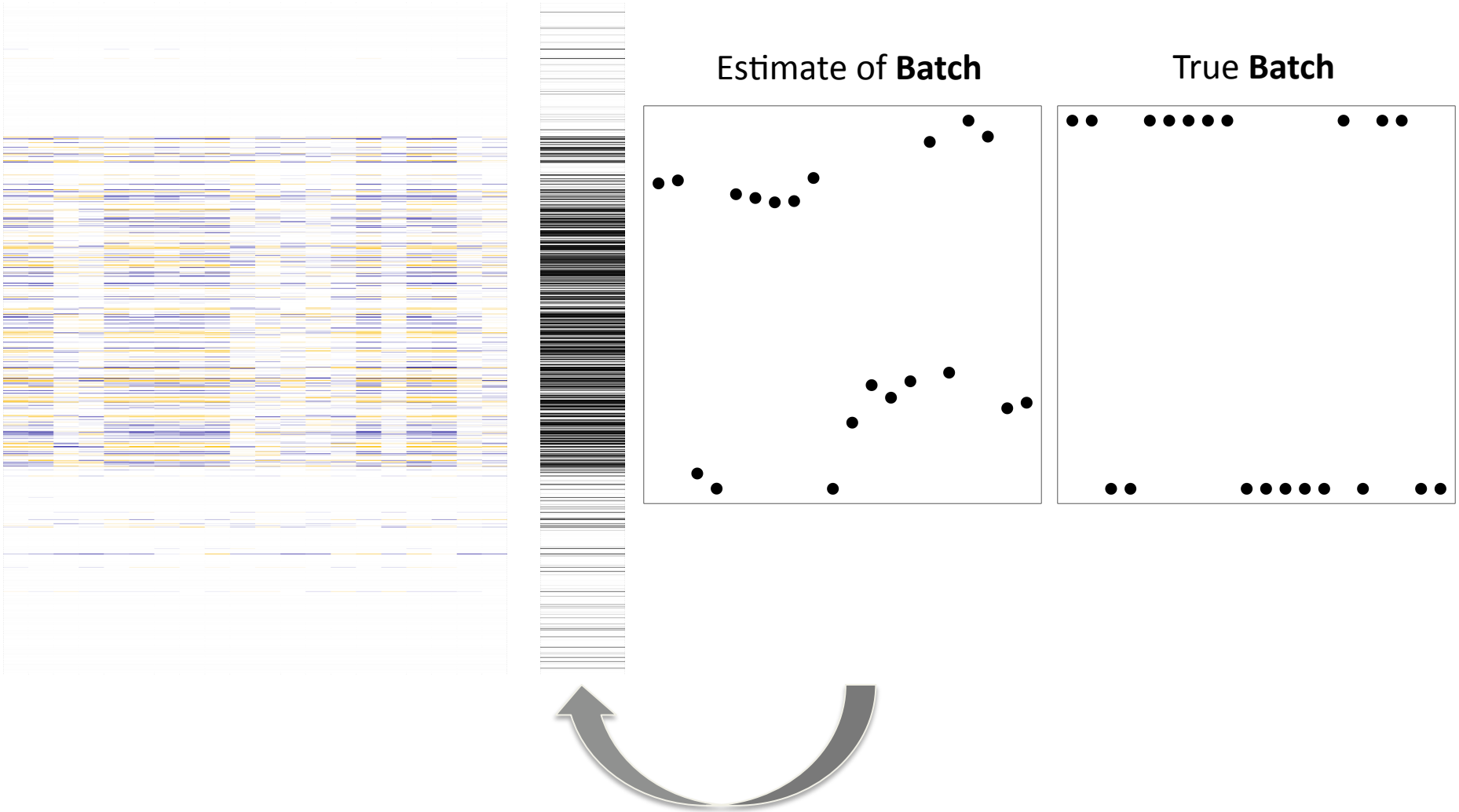
True **Batch**



Surrogate Variable Analysis

The Data

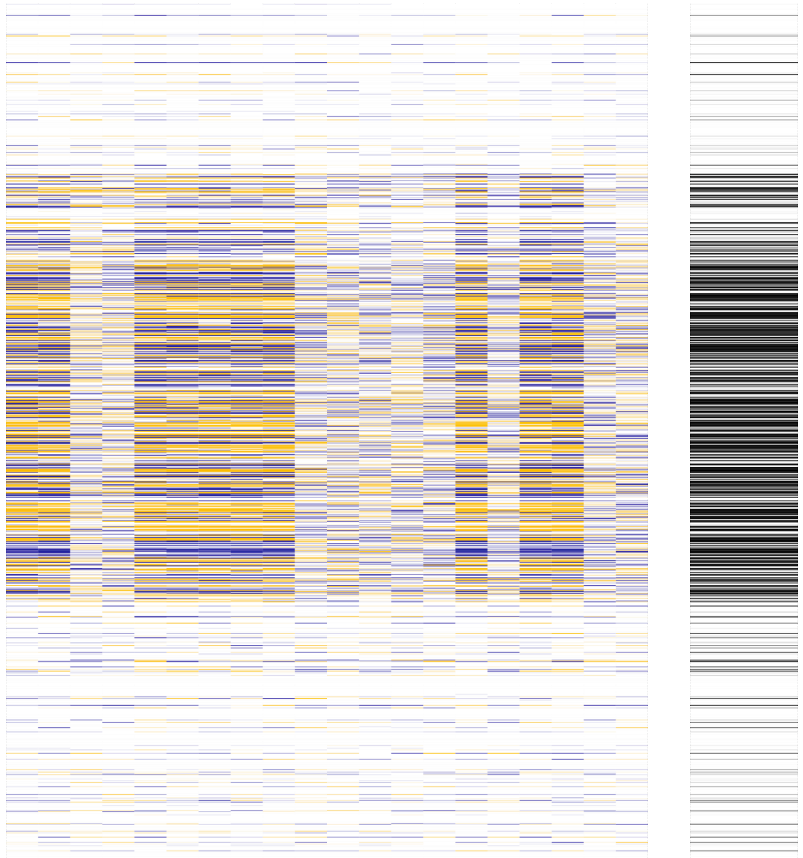
Pr(!**Group** & **Batch**)



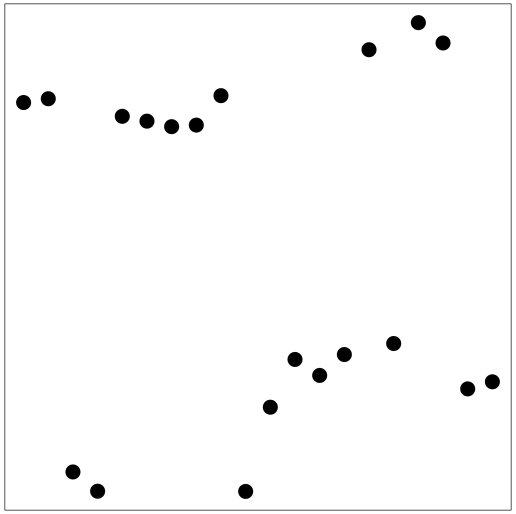
Surrogate Variable Analysis

The Data

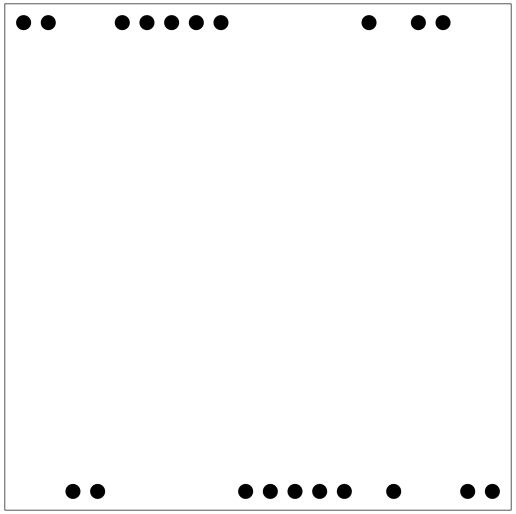
Pr(!**Group** & **Batch**)



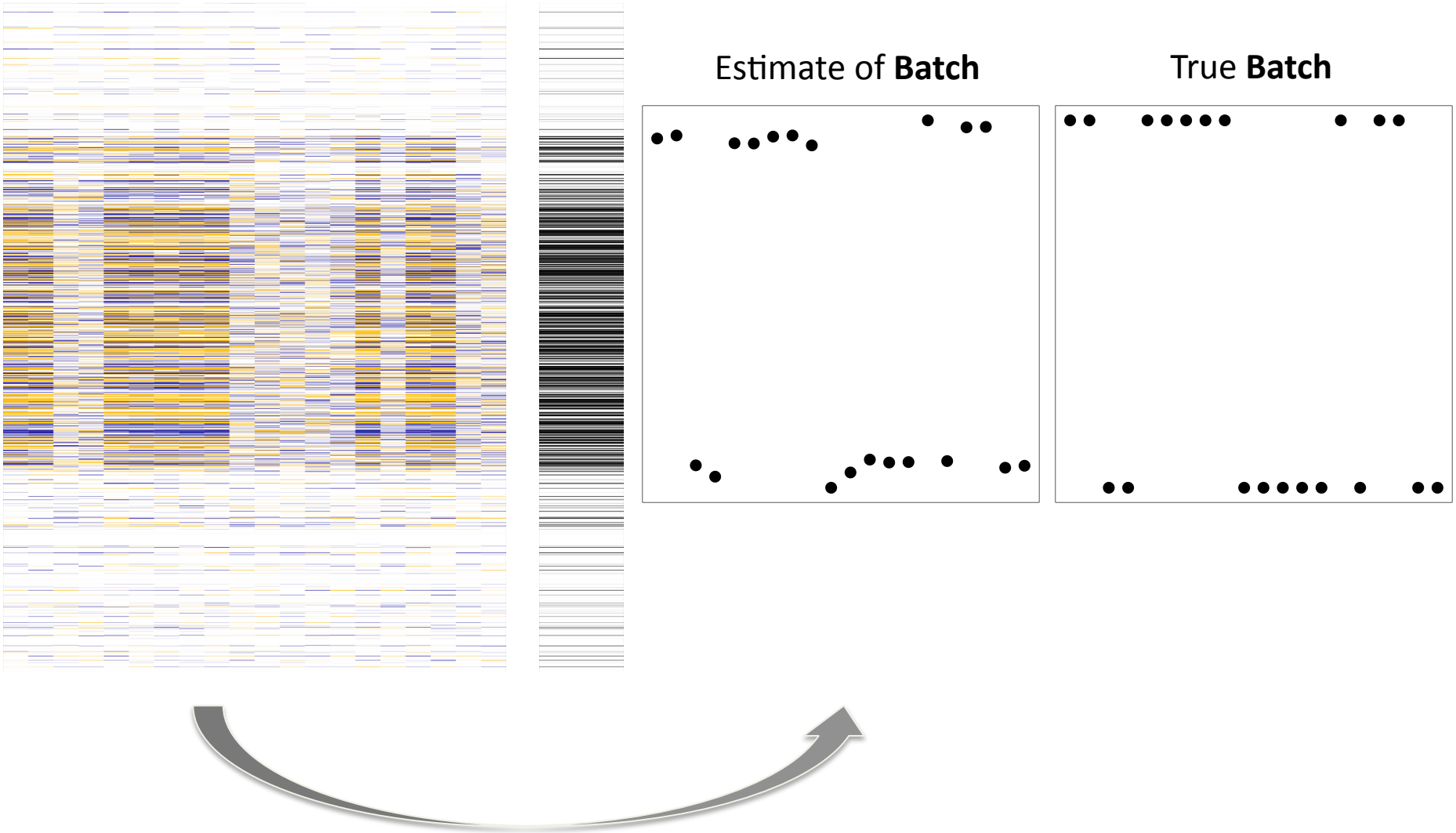
Estimate of **Batch**



True **Batch**



Surrogate Variable Analysis



SVA Adjusted Gene by Gene Model

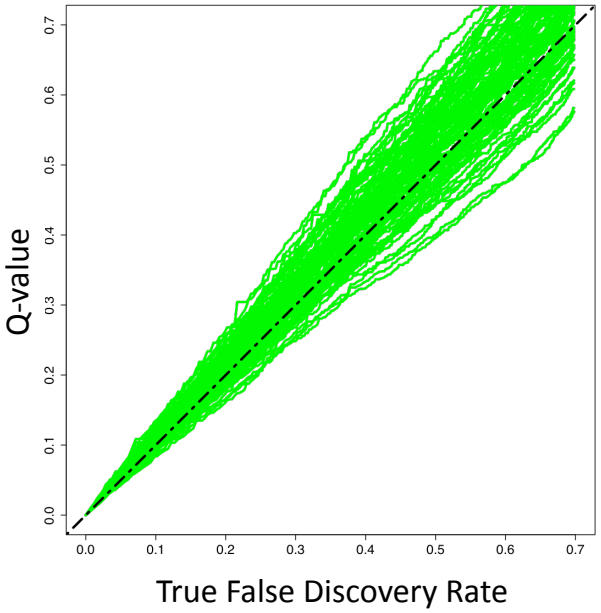
$$\text{expression} = b_0 + b_1 \times \text{group} + \text{surrogates} + \text{noise}$$

Test whether $b_1 = 0$

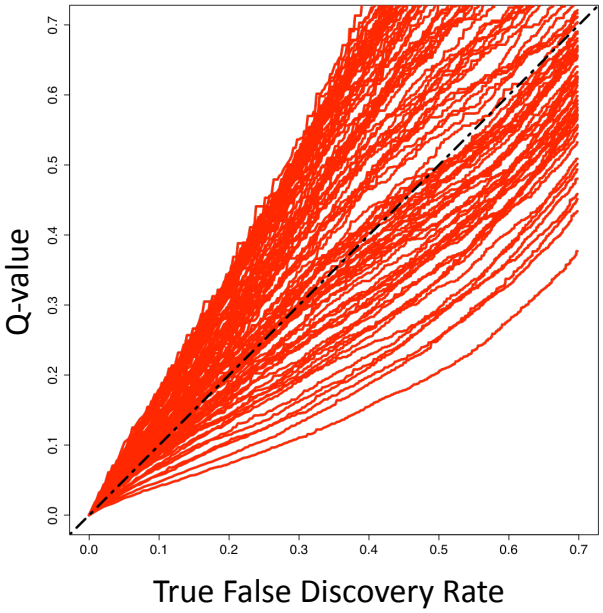
Calculate a P-value

False Discovery Rate Estimates

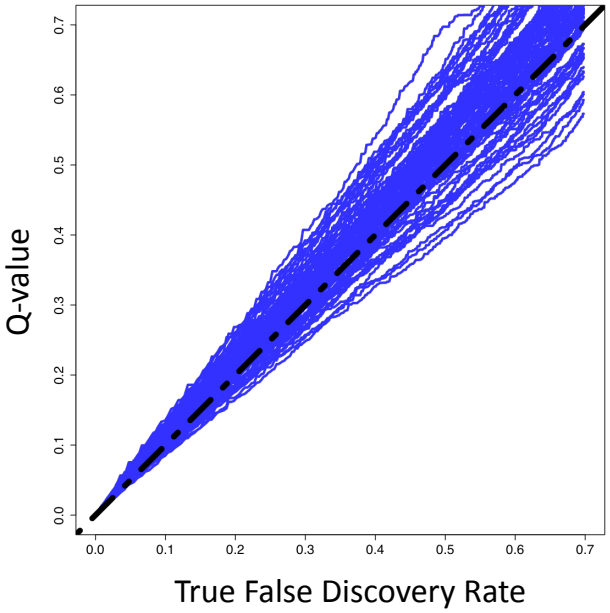
Independent **E**



Dependent **E**

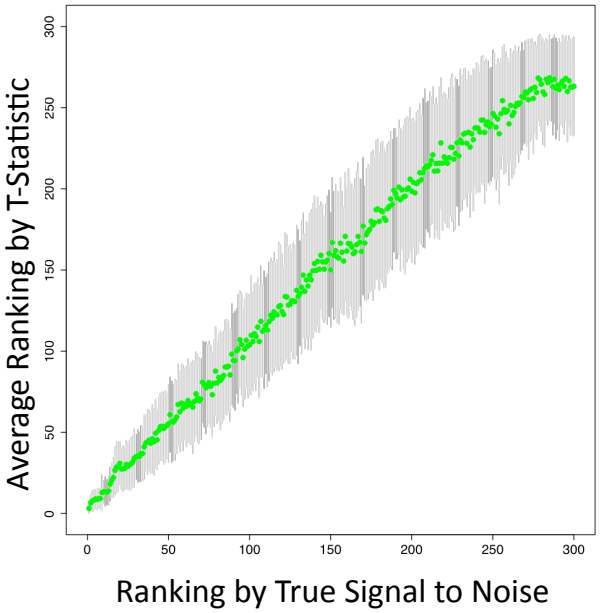


Dependent **E**
+ IRW-SVA

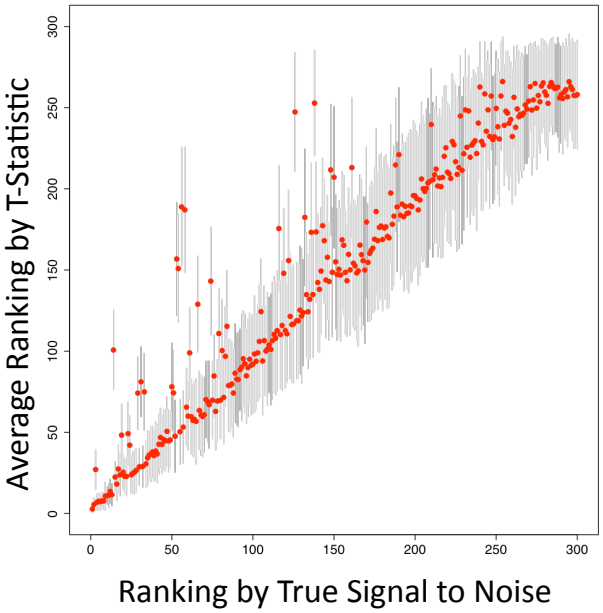


Ranking Estimates

Independent **E**



Dependent **E**



Dependent **E**
+ IRW-SVA