# Differential expression analysis

Charlotte Soneson

Friedrich Miescher Institute for Biomedical Research &
SIB Swiss Institute of Bioinformatics

Hinxton, April 9, 2019

# Differential analysis types for RNA-seq

- Does the total output of a gene change between conditions? **DGE**
- Does the expression of individual transcripts change? **DTE**
- Does *any* isoform of a given gene change? **DTE+G**
- Does the isoform composition for a given gene change? **DTU/DIU/DEU**
- (Does *anything* change? GDE*)

- need **different** abundance quantification of transcriptomic features (genes, transcripts, exons)

*https://liorpachter.wordpress.com/2018/02/15/gde%C2%B2-dge%C2%B2-dtu%C2%B2-dte%E2%82%81%C2%B2-dte%E2%82%82%C2%B2/

# Challenges for RNA-seq data

- Choice of statistical distribution
- Normalization between samples
- Few samples -> difficult to estimate parameters (e.g., variance)
- High dimensionality (many genes) -> many tests

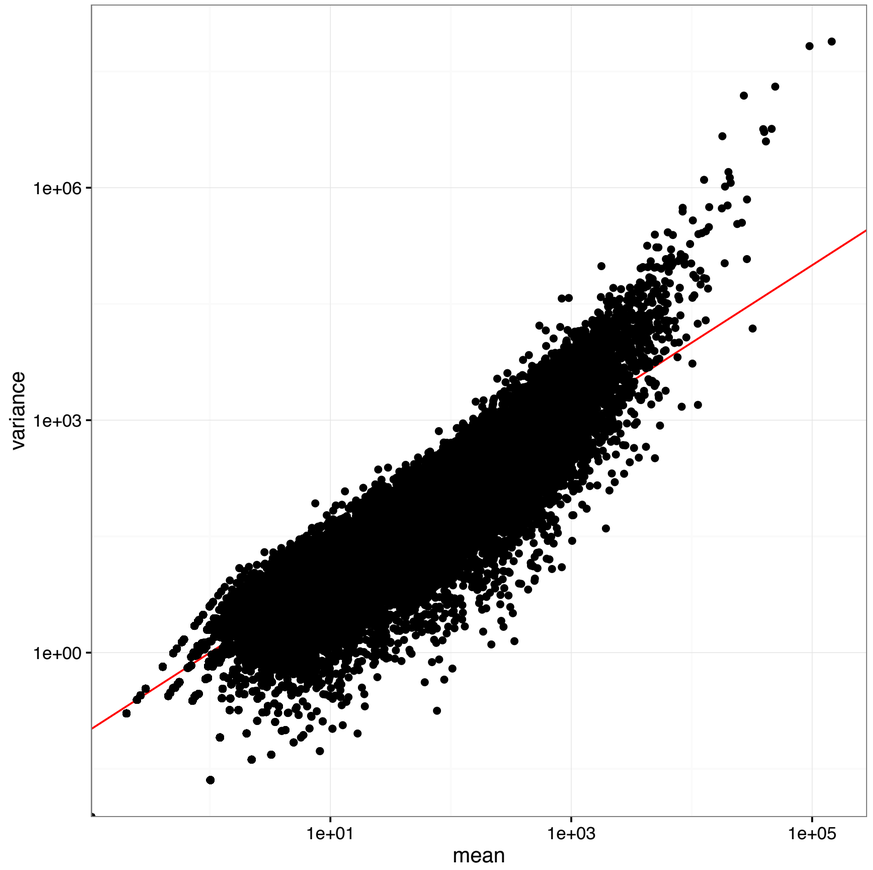
# Challenges for RNA-seq data

- **Choice of statistical distribution**
- Normalization between samples
- Few samples -> difficult to estimate parameters (e.g., variance)
- High dimensionality (many genes) -> many tests

# Modeling counts

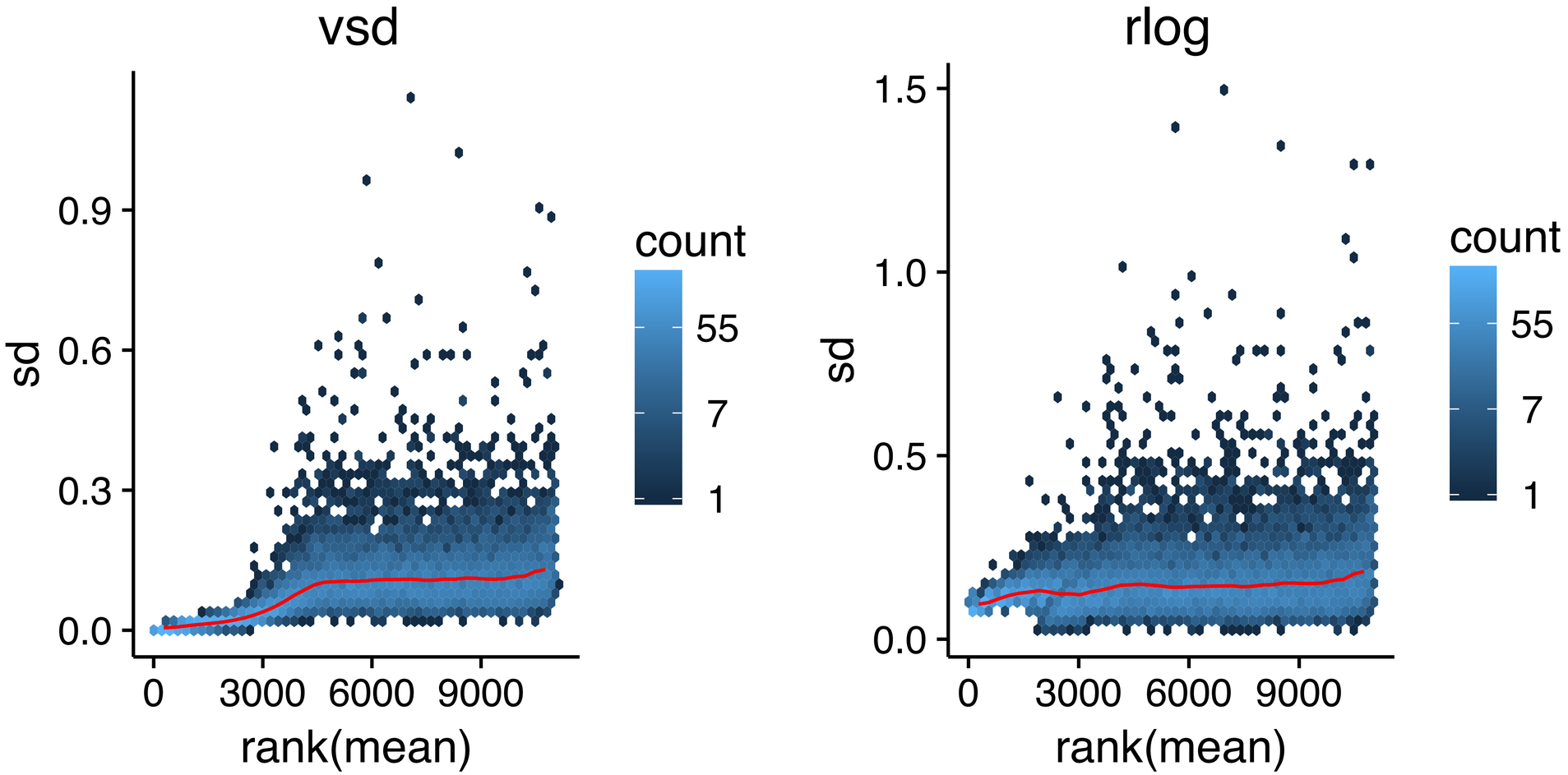
- **Negative binomial distribution**
  - $var(X) = \mu + \theta\mu^2$
  - $\theta$ = dispersion
  - $\sqrt{\theta}$ = "biological coefficent of variation”
  - Allows mRNA proportions to vary across samples
  - Captures variability across biological replicates better

**Example from SEQC data, replicates of the same RNA mix**



# Data transformations - DESeq2

- Two approaches: rlog, variance stabilizing transformation
- Aim: remove dependence of variance on mean after transformation



# Challenges for RNA-seq data

- Choice of statistical distribution
- **Normalization between samples**
- Few samples -> difficult to estimate parameters (e.g., variance)
- High dimensionality (many genes) -> many tests

# Normalization

- Observed counts depend on:
  - abundance
  - gene length
  - sequencing depth
  - sequencing biases
  - …
- “As-is”, not directly comparable across samples

# Normalization

raw count for gene *i* in sample *j*

relative abundance

normalization factor

$$C_{ij} \sim NB(\mu_{ij} = s_{ij} q_{ij}, \theta_i)$$

dispersion

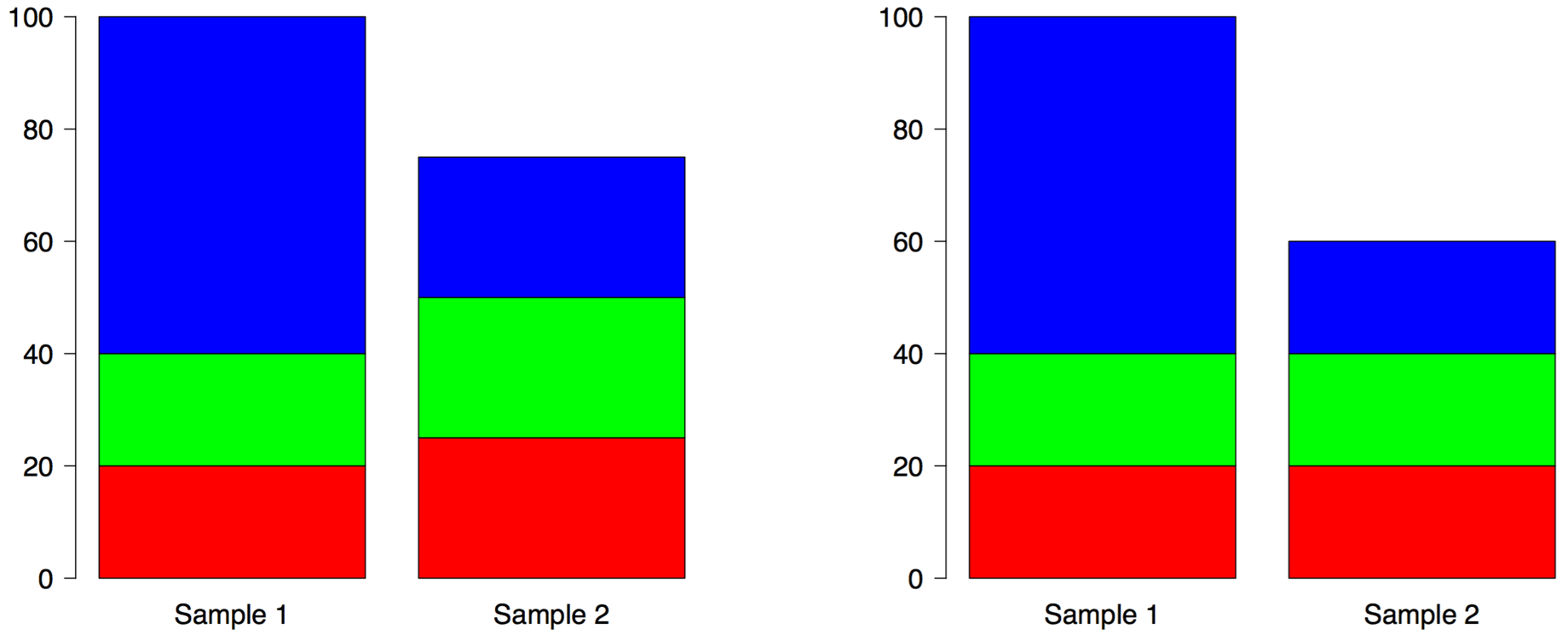
- $s_{ij}$ is a normalization factor (or *offset*) in the model
- counts are not explicitly scaled
  - important exception: voom/limma (followed by explicit modeling of mean-variance association)

# How to calculate normalization factors?

- Attempt 2: total count (library size) * compensation for differences in composition
- Idea: use only non-differentially expressed genes to compute the normalization factor
- Implemented by both edgeR (TMM) and DESeq2 (median count ratio)
- Both these methods assume that most genes are not differentially expressed

# How to calculate normalization factors?

- Attempt 2: total count (library size) * compensation for differences in composition

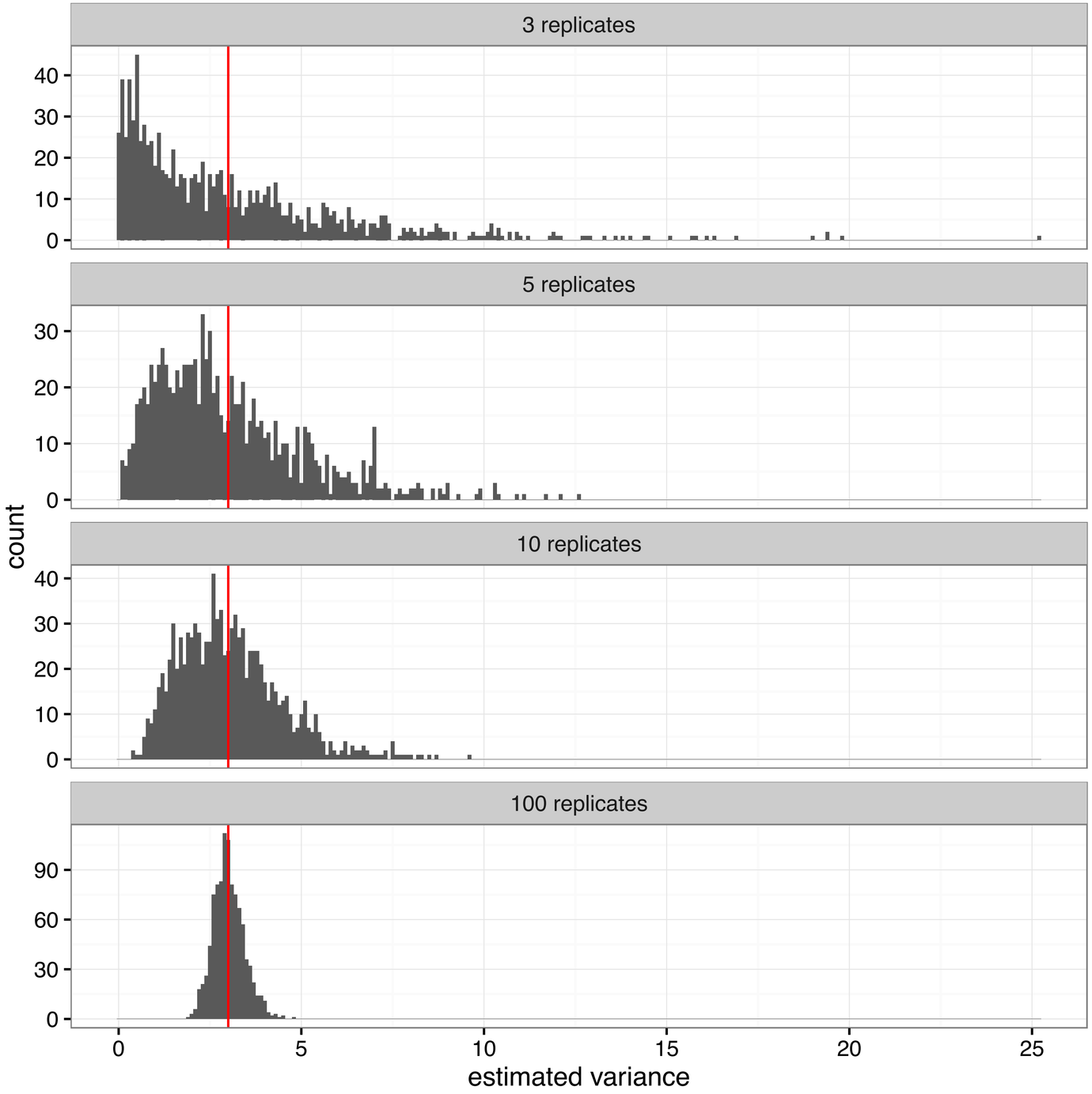


# Challenges for RNA-seq data

- Choice of statistical distribution
- Normalization between samples
- **Few samples -> difficult to estimate parameters (e.g., variance)**
- High dimensionality (many genes) -> many tests

**Example:**
estimate variance of normally distributed variable

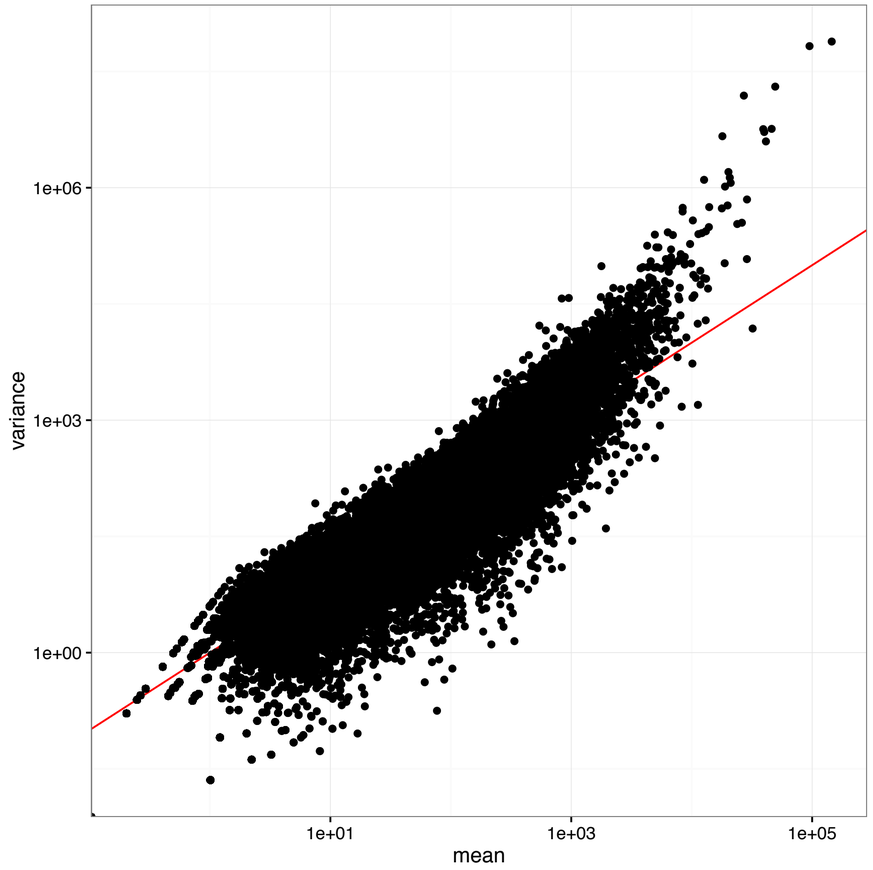
True value = 3



# Modeling counts

- **Negative binomial distribution**
  - $var(X) = \mu + \theta\mu^2$
  - $\theta$ = dispersion
  - $\sqrt{\theta}$ = "biological coefficent of variation”
  - Allows mRNA proportions to vary across samples
  - Captures variability across biological replicates better

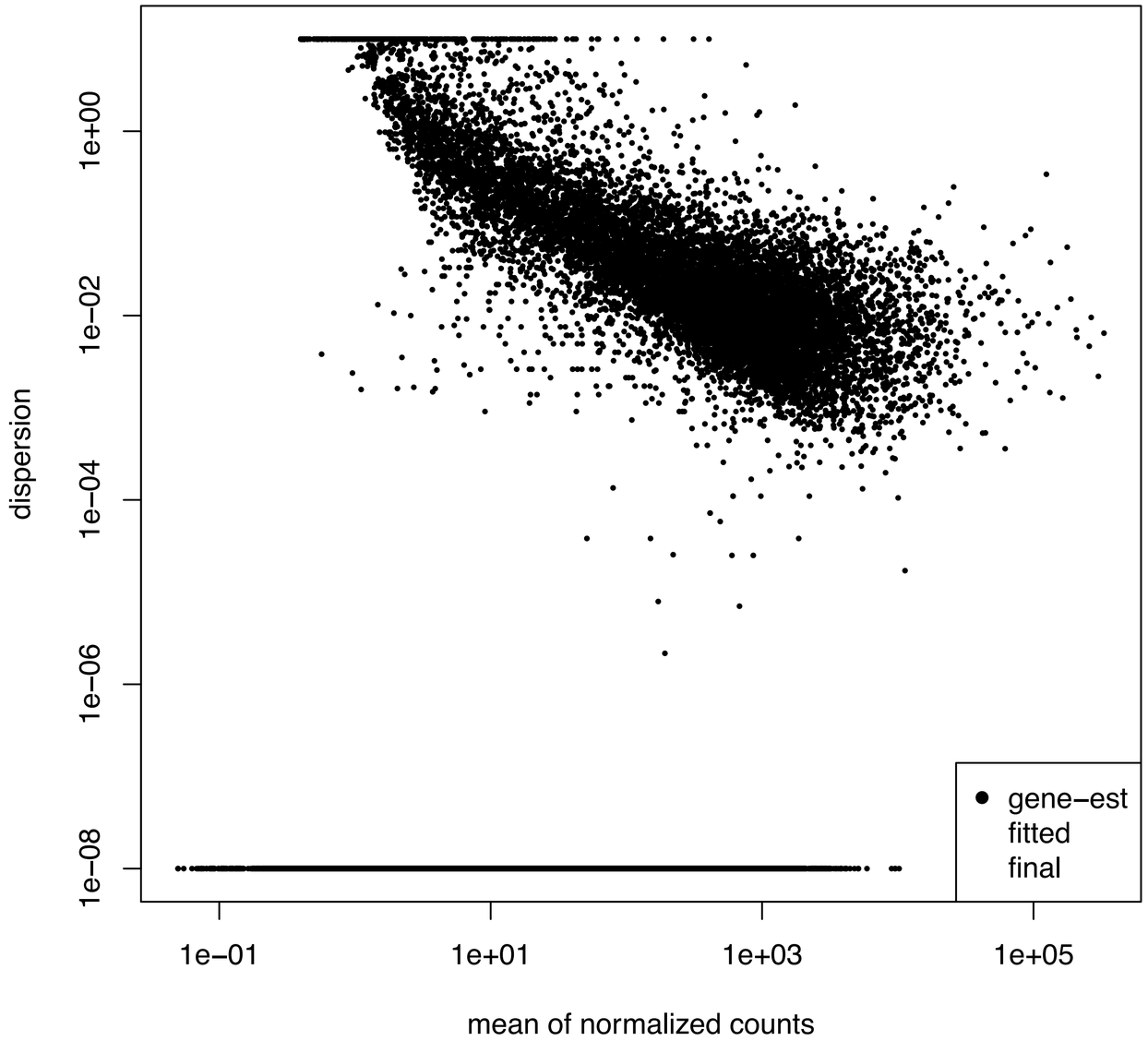
**Example from SEQC data, replicates of the same RNA mix**



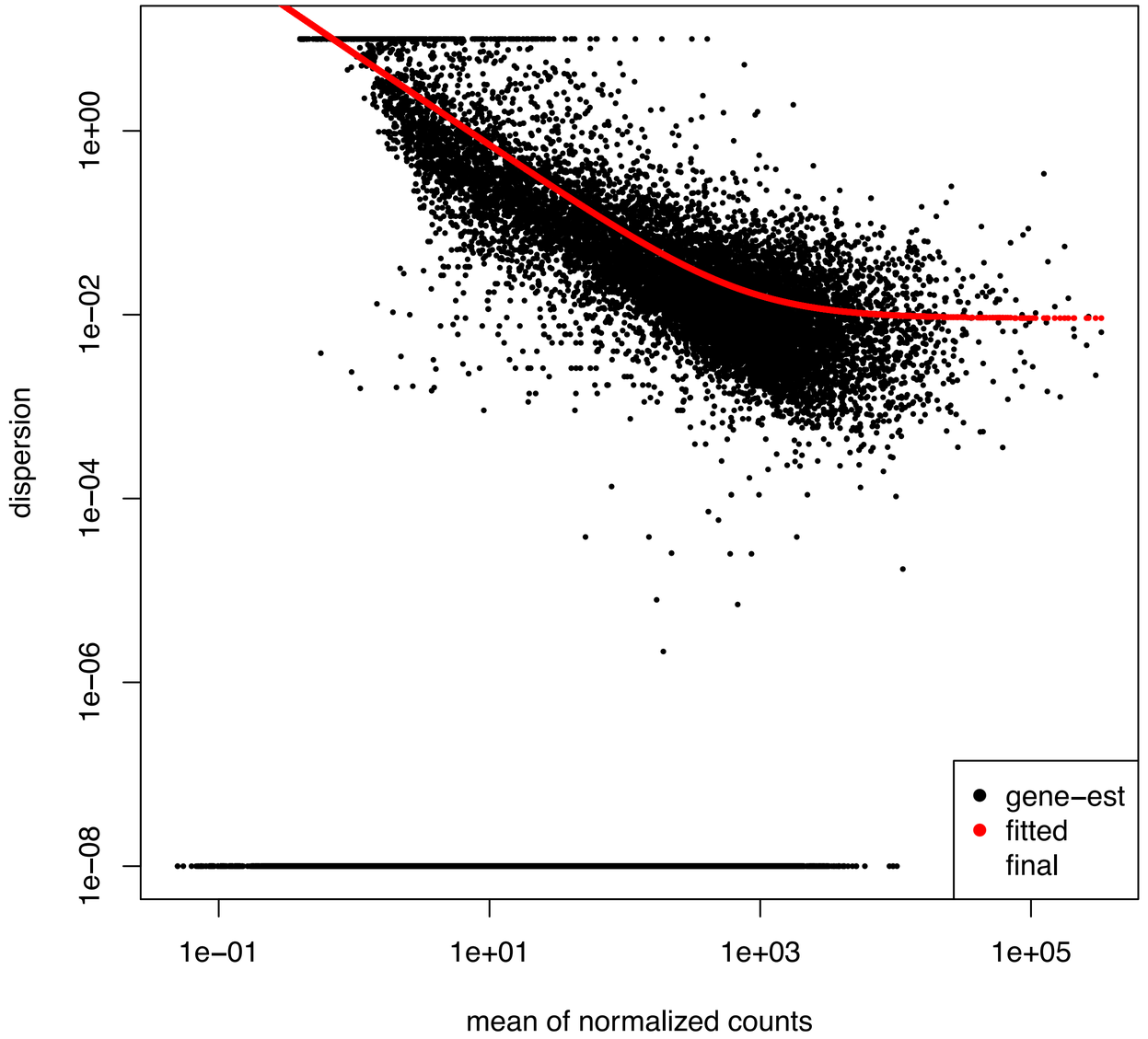
# Shrinkage dispersion estimation

- Take advantage of the large number of genes
- Shrink the gene-wise estimates towards a center value defined by the observed distribution of dispersions across
  - all genes (“common” dispersion estimate)
  - genes with similar expression (“trended” dispersion estimate)

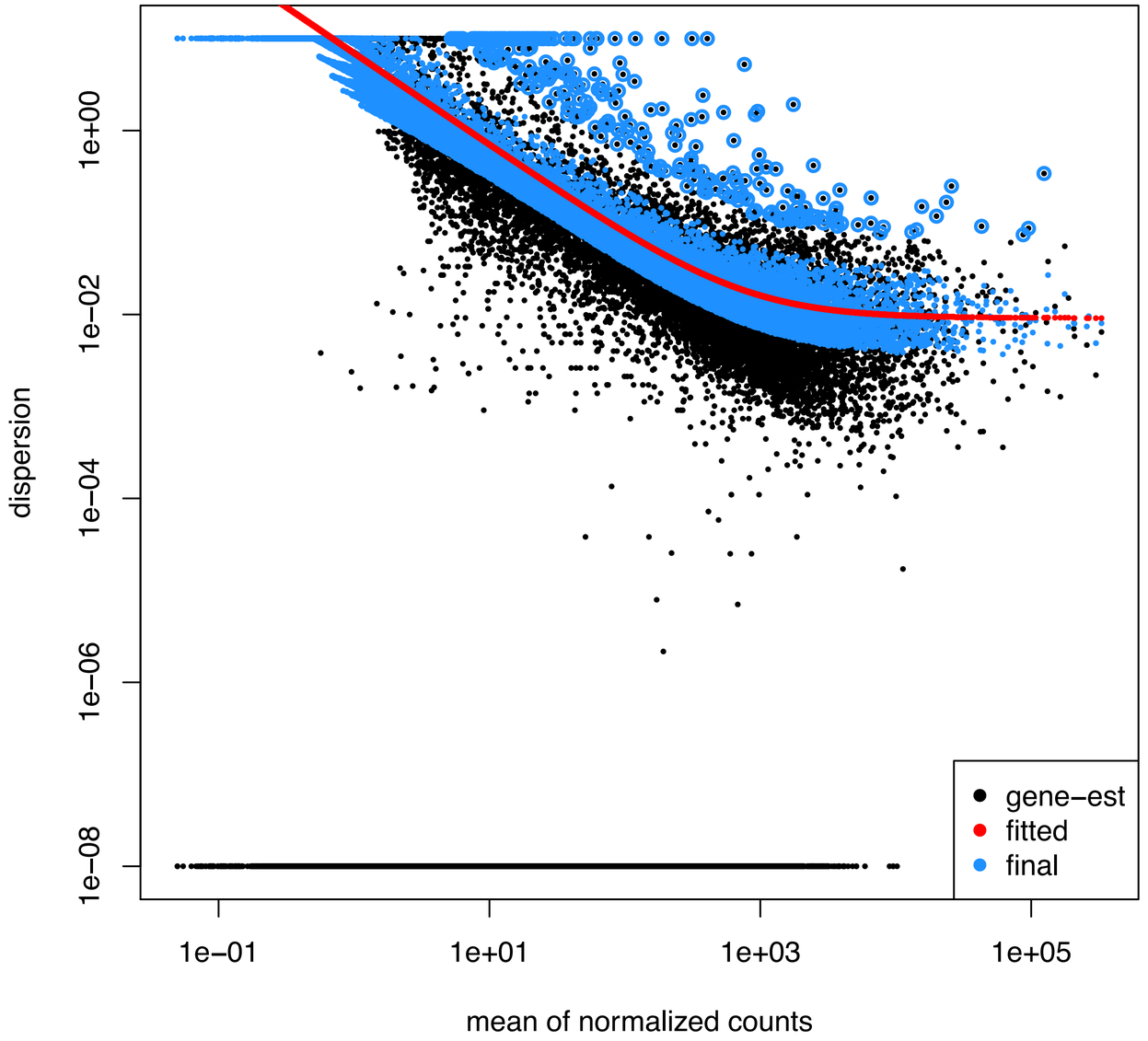
# Shrinkage dispersion estimation



# Shrinkage dispersion estimation



# Shrinkage dispersion estimation



# Challenges for RNA-seq data

- Choice of statistical distribution
- Normalization between samples
- Few samples -> difficult to estimate parameters (e.g., variance)
- **High dimensionality (many genes) -> many tests**

# What is a p-value?

- The p-value is the probability of obtaining a test statistic *at least as extreme* as the one observed, *if the null hypothesis is true* (i.e., if there is no true signal in the data)

- Hence, if we get a p-value of **0.05**, it means that there is a **5%** chance of getting that extreme results even in the absence of real signal!

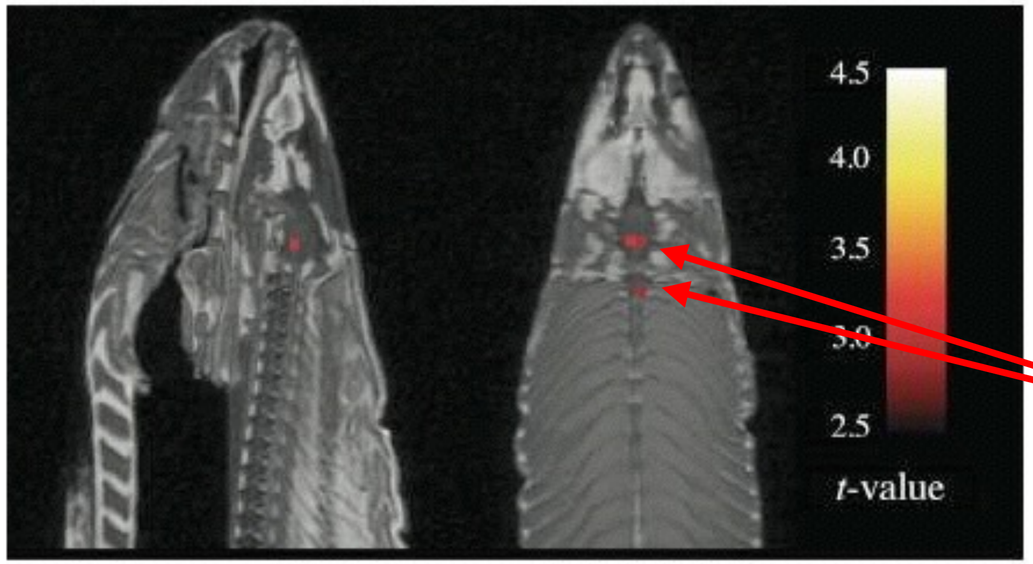
# What does this mean for high-throughput studies?

- Assume that we perform 10,000 tests (one for each gene)…
- … and that there is no true signal at all in the data
- Then we would expect to get around 500 p-values below 0.05
- Relying solely on p-values would be misleading!

# So, what can happen if we don't pay attention?

**NEUROSCIENCE PRIZE**: Craig Bennett, Abigail Baird, Michael Miller, and George Wolford [USA], for demonstrating that brain researchers, by using complicated instruments and simple statistics, can see meaningful brain activity anywhere — even in a dead salmon.

IMPROBABLE RESEARCH




## METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

Craig M. Bennett, Abigail A. Baird, Michael B. Miller, and George L. Wolford, Journal of Serendipitous and Unexpected Results, vol. 1, no. 1, 2010, pp. 1-5

# We need to change perspective

- Instead of limiting the false positive probability for *each individual test*, try to limit
  - the probability of obtaining *any* false positives (FWER)
  - the fraction of false positives among the significant genes (FDR)

# Benjamini-Hochberg correction - controlling the FDR

- Assume we are performing N tests
- Intuition:
  - for each threshold α, we can estimate the expected number of false discoveries by αN
  - Compare this to the actual number of discoveries at that threshold ($N_\alpha$)
  - Choose a so that $\alpha N / N_\alpha \leq 0.05$ (or another desired threshold)

# Interpreting the FDR

- The FDR is a measure for a *set* of genes
- In a set of genes with FDR = 0.05, approximately 5% can be expected to be false discoveries
- However, we don’t know *which ones*! It could be the most significant!
- *q-values* are gene-wise significance measures (“adjusted p-values”) - the smallest FDR we have to accept in order to call the gene significant

# Independent filtering

- Idea: filter out genes that have little chance of showing significance (**without** looking at the test results!!!)
- Improves detection power for remaining genes (fewer tests - less strict correction for multiple testing)
- For RNA-seq, typically filter based on expression

# Independent filtering

- DESeq2:
  - filters based on the average normalized counts, using an optimized threshold.
  - p-values for excluded genes are set to NA in results
- edgeR:
  - manual filtering before applying test
  - all remaining genes are tested, and get a p-value

# Testing against a threshold

- By default, the null hypothesis is that the log-fold change ($\beta$) between conditions is 0
- Both edgeR and DESeq2 can test more general null hypothesis, e.g. $|\beta| \leq 1$
- Useful if very small fold changes are not of interest
- Note that this is **not** the same as setting both a p-value and a fold change threshold on the regular test results!

# Which method to choose?

- edgeR
- DESeq2
- voom/limma
- sleuth
- NOISeq
- DSS
- ShrinkBayes
- EBSeq
- baySeq
- SAMseq

RESEARCH ARTICLE

# A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data

Zong Hong Zhang, Dhanisha J. Jhaveri, Vikki M. Marshall, Denis C. Bauer, Janette Edson, Ramesh K. Narayanan, Gregory J. Robinson, Andreas E. Lundberg, Perry F. Bartlett, Naomi R. Wray, Qiong-Yi Zhao

# Comparison of software packages for detecting differential expression in RNA-seq studies

*Fatemeh Seyednasrollah, Asta Laiho and Laura L. Elo*

**Research article** Highly accessed Open Access

# A comparison of methods for differential expression analysis of RNA-seq data

**Charlotte Soneson[1]* and Mauro Delorenzi[1,2]**

# How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?

NICHOLAS J. SCHURCH,[1,6] PIETÀ SCHOFIELD,[1,2,6] MAREK GIERLIŃSKI,[1,2,6] CHRISTIAN COLE,[1,6] ALEXANDER SHERSTNEV,[1,6] VIJENDER SINGH,[2] NICOLA WROBEL,[3] KARIM GHARBI,[3] GORDON G. SIMPSON,[4] TOM OWEN-HUGHES,[2] MARK BLAXTER,[3] and GEOFFREY J. BARTON[1,2,5]

**Method** Highly accessed Open Access

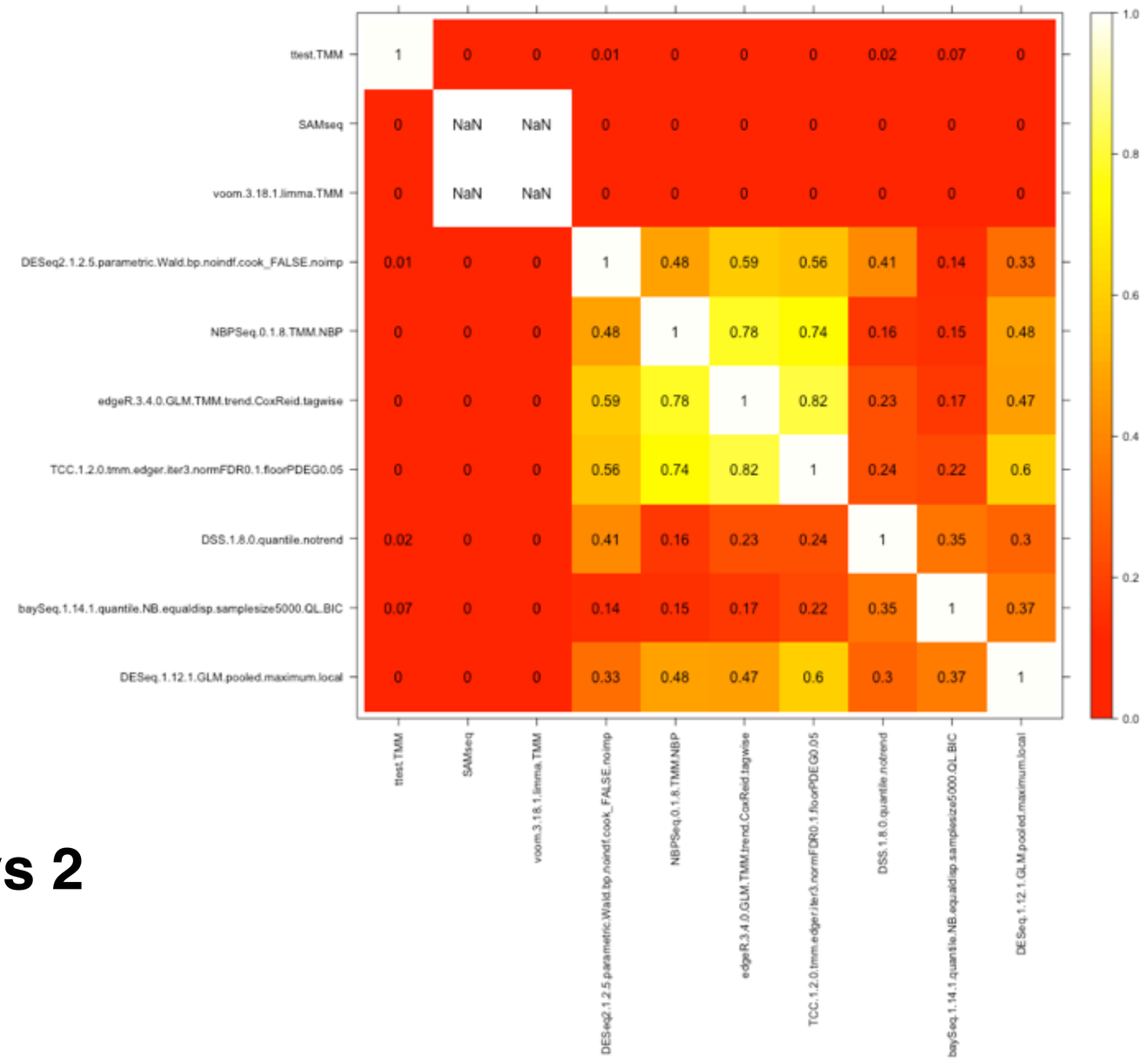
# Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data

**Franck Rapaport[1], Raya Khanin[1], Yupu Liang[1], Mono Pirun[1], Azra Krek[1], Paul Zumbo[2,3], Christopher E Mason[2,3], Nicholas D Socci[1] and Doron Betel[3,4]***

# Sometimes, there is not much choice…

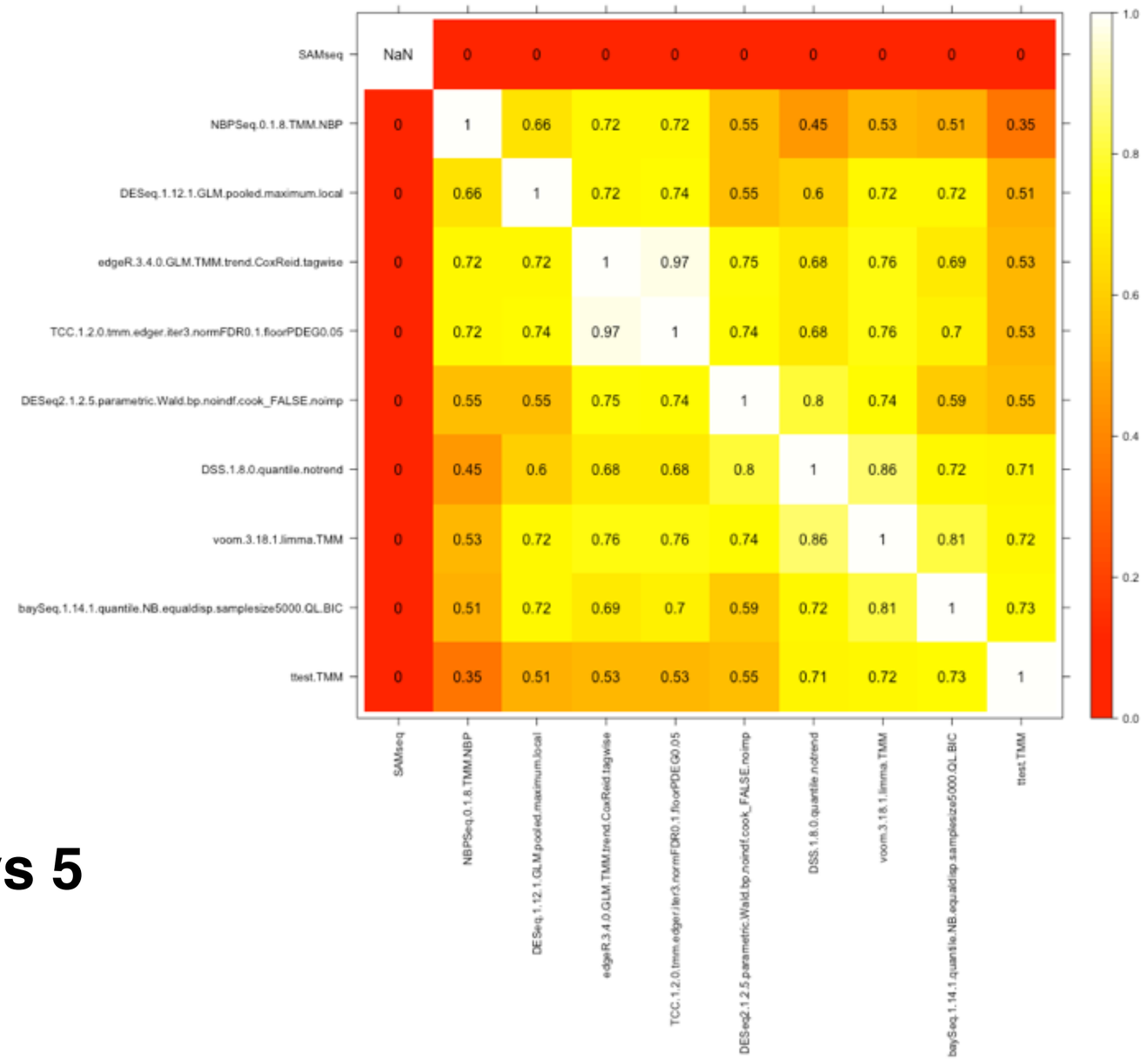
- Random effects - ShrinkBayes (also allows zero-inflation)
- Incorporate assignment uncertainty - sleuth

# As sample size increases, methods perform more similarly



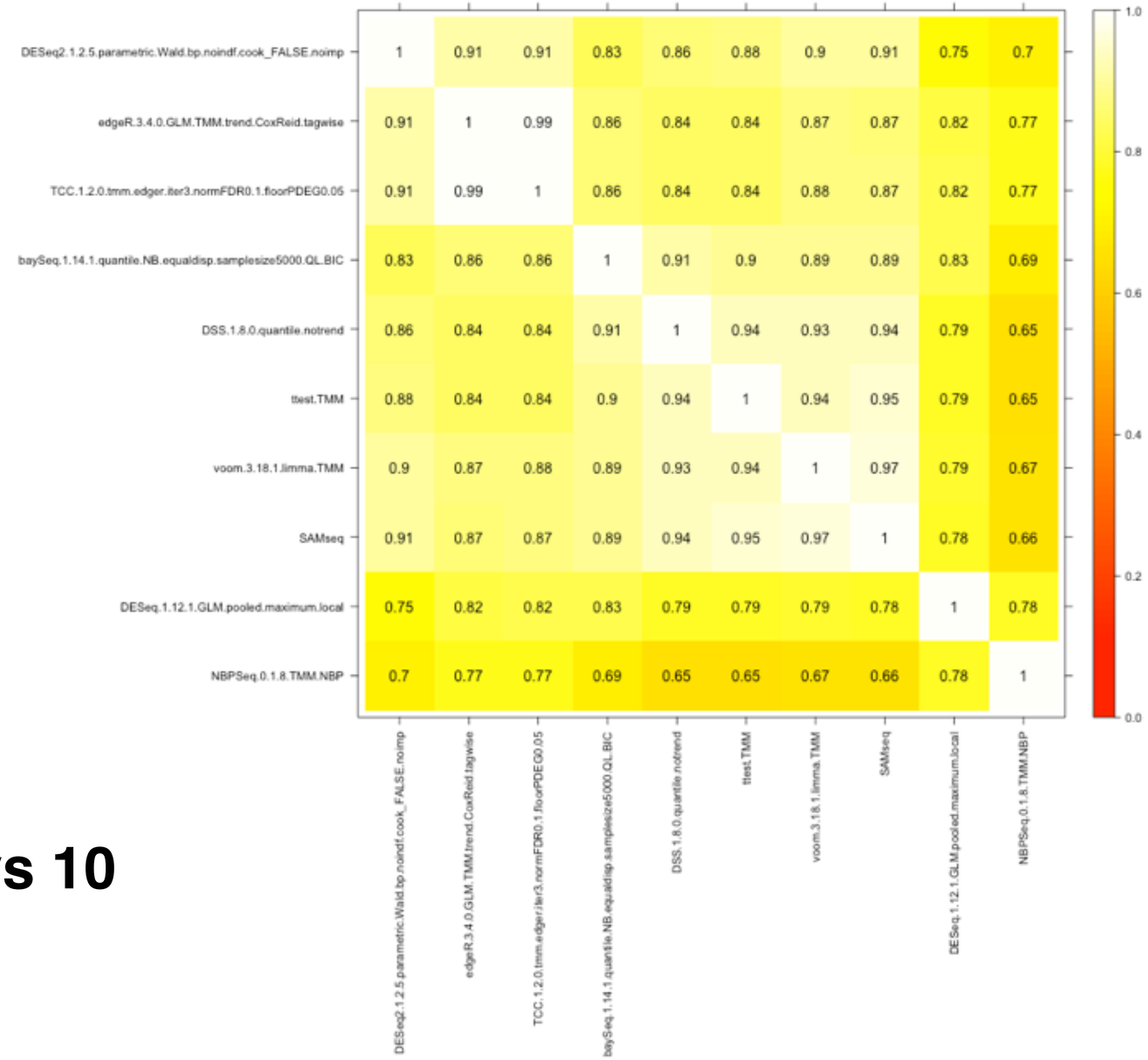

## 2 vs 2

# As sample size increases, methods perform more similarly



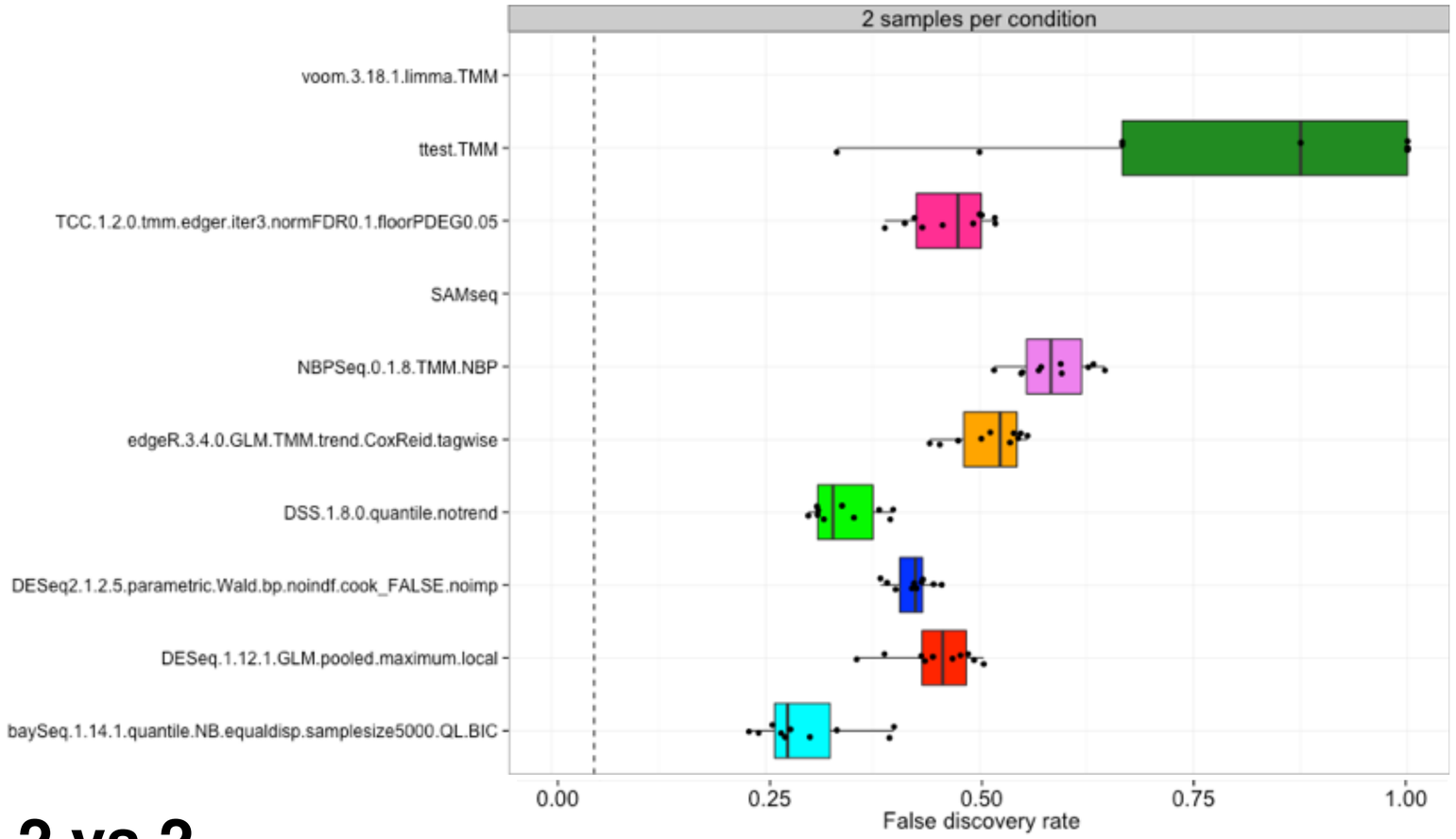

5 vs 5

# As sample size increases, methods perform more similarly



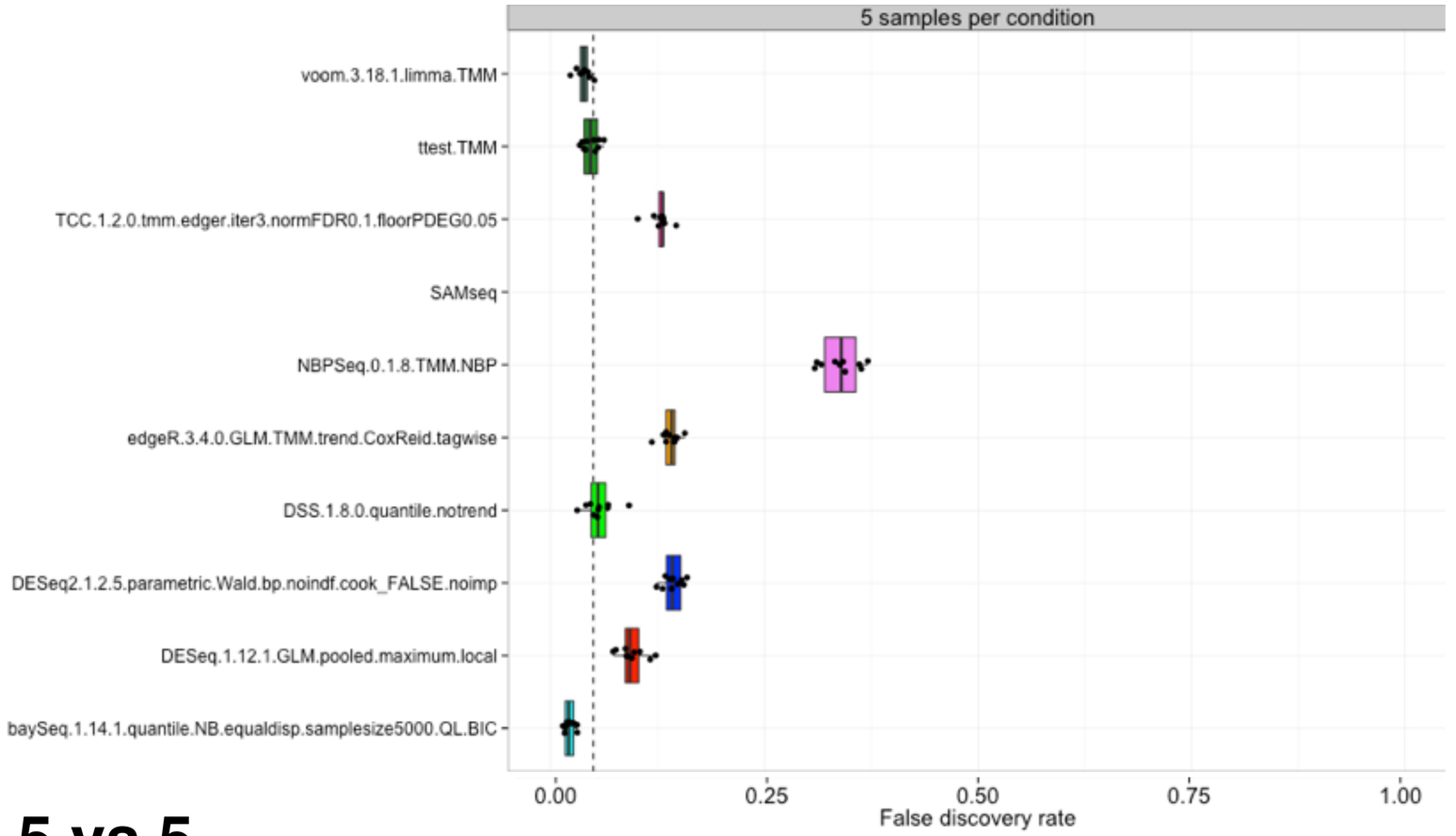

## 10 vs 10

# As sample size increases, methods perform better



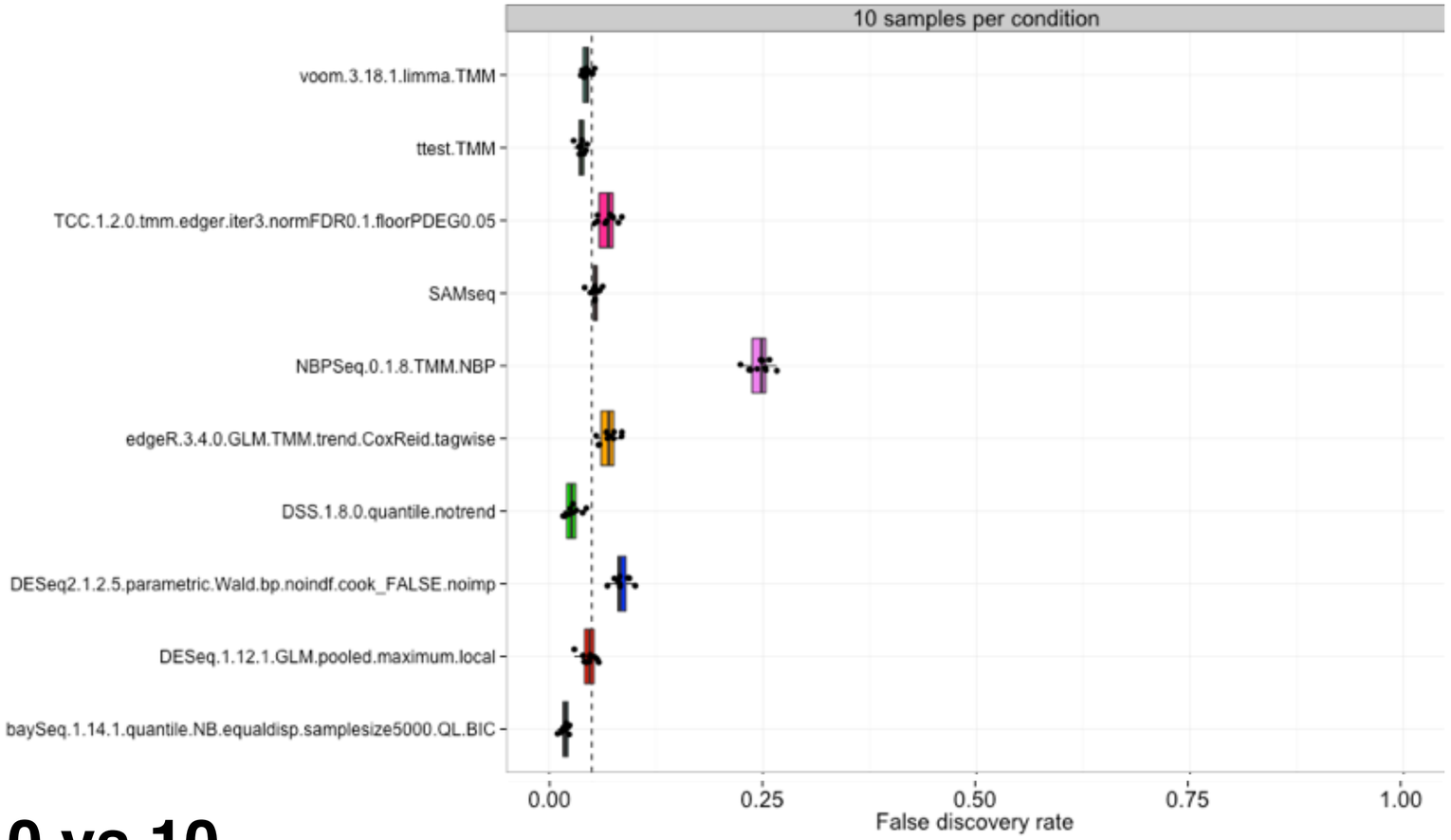

## 2 vs 2

# As sample size increases, methods perform better



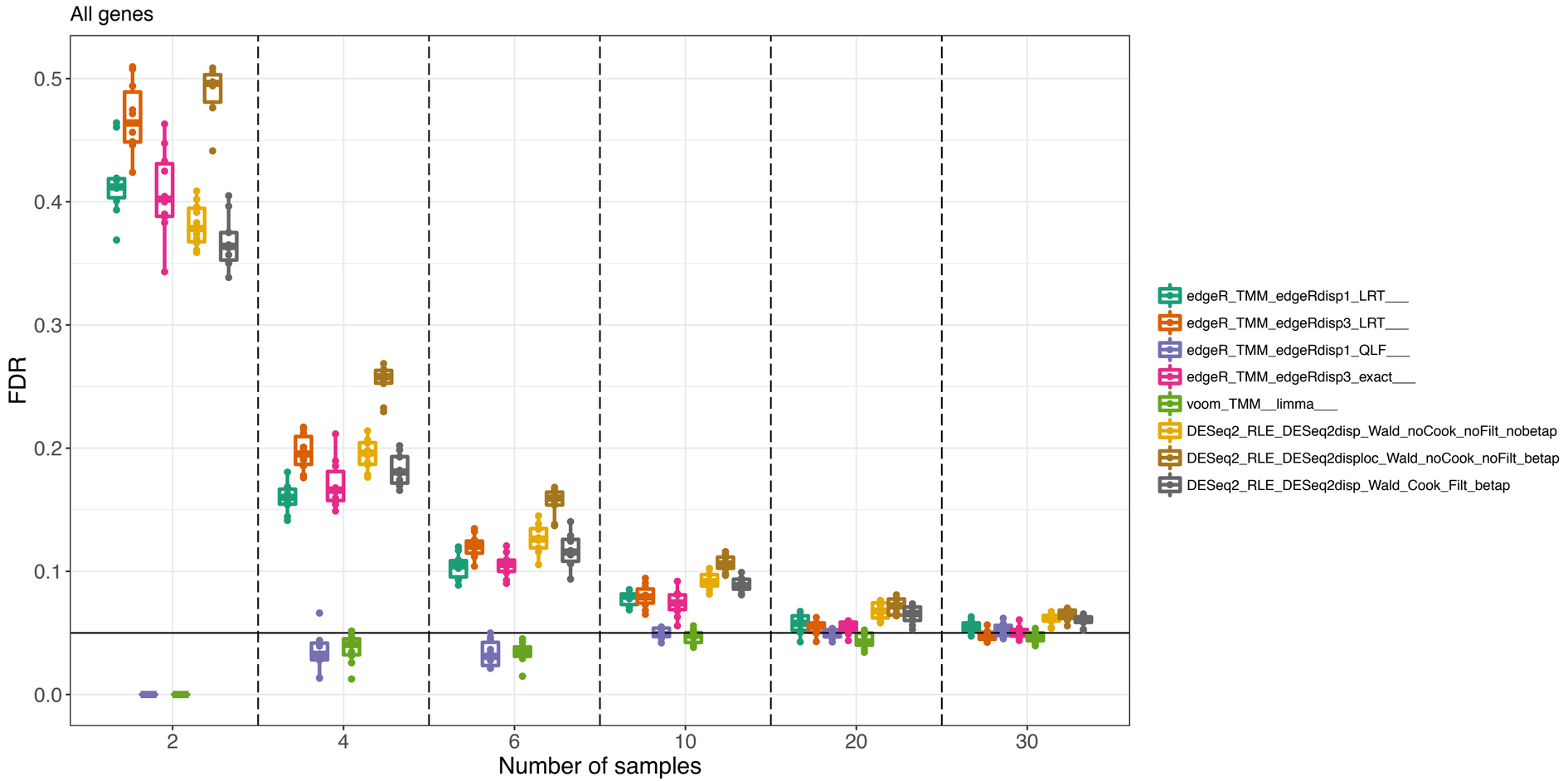

5 vs 5

# As sample size increases, methods perform better

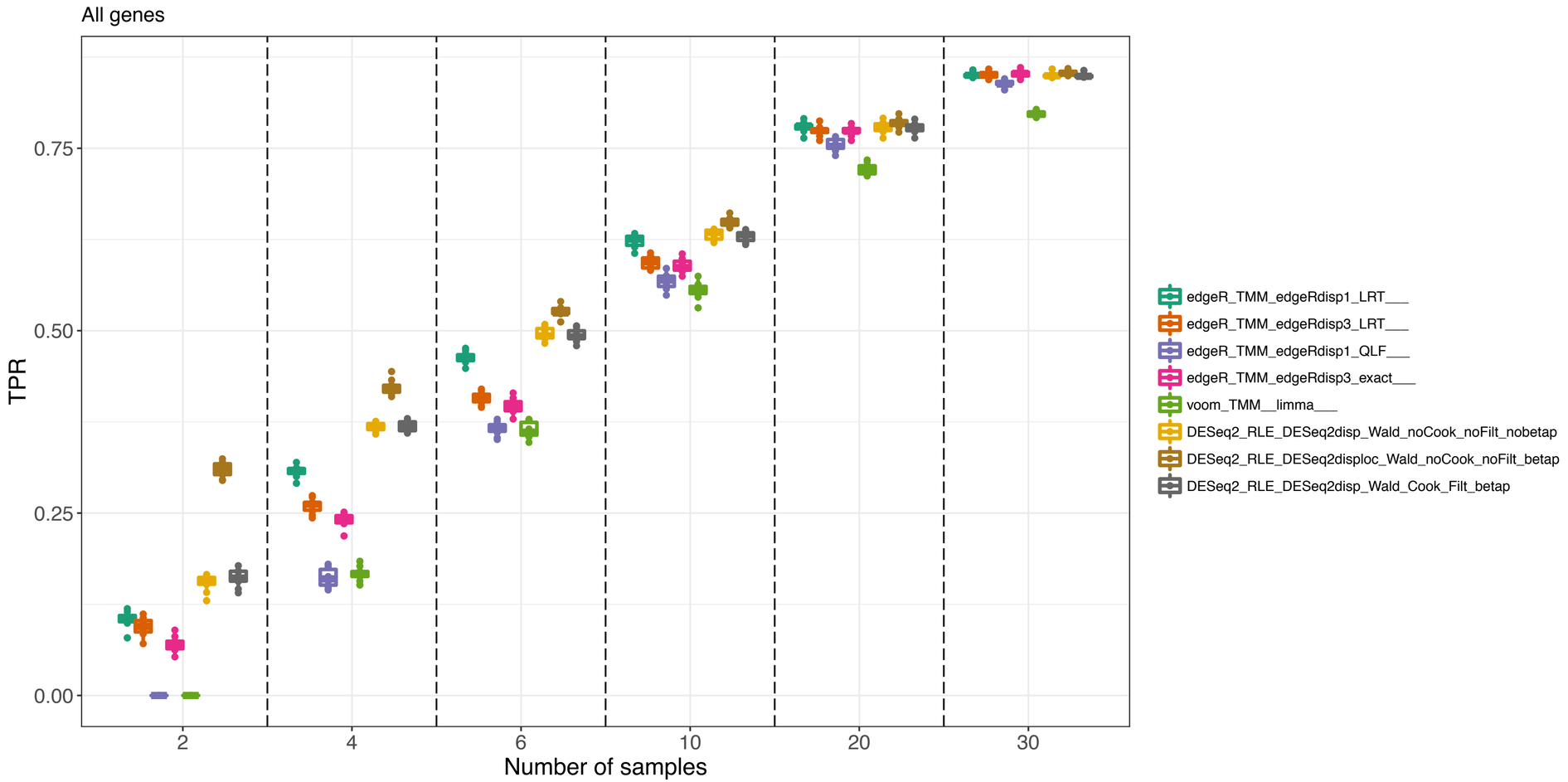



10 vs 10

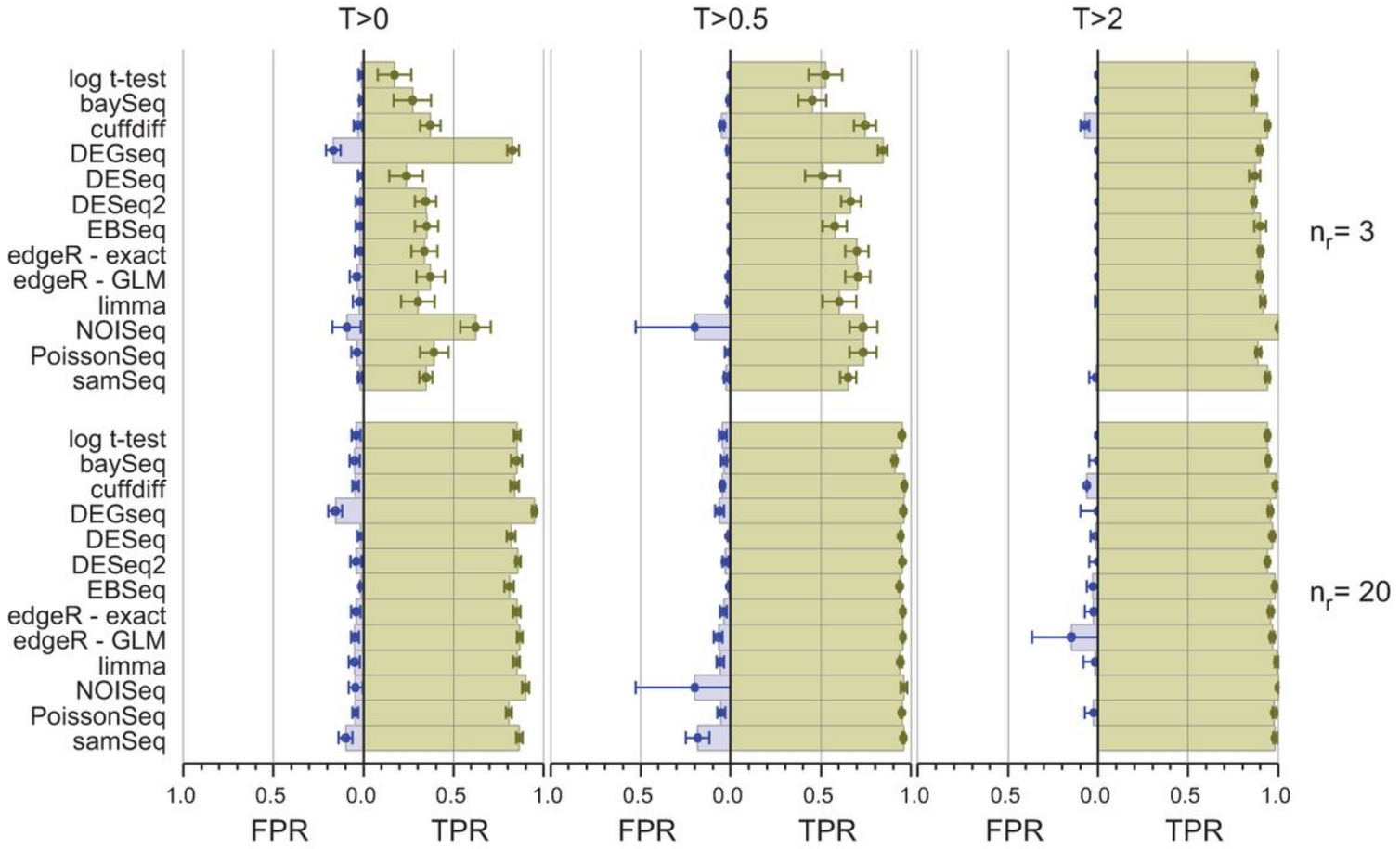
# As sample size increases, methods perform better



# As sample size increases, methods perform better



# Strong signals can be detected with few samples




Schurch et al., RNA (2016)

# edgeR or DESeq2?

- Based on similar statistical models
- Implement similar basic functionality
- Both are very much “alive” and actively developed
- Beware of the different default settings!

# edgeR vs DESeq2 - differences

- Type of test (Wald/LRT in DESeq2, LRT/QLF in edgeR)
- Dealing with outliers (default in DESeq2, *estimateGLMRobustDisp*() in edgeR)
- Independent filtering (included by default in DESeq2, manually preceding test in edgeR)
- exploratory analysis - data transformation (variance-stabilizing transformation in DESeq2, logCPM in edgeR)

# Model formulas and design matrices

- Testing is done separately for each gene
- We must tell the packages **which model** to fit (e.g. which predictors to use)
- The design does *not* follow “automatically” from having the sample annotation table - many different designs are often possible
- Model formulas in R:

$$\text{response variable} \sim \text{predictors}$$

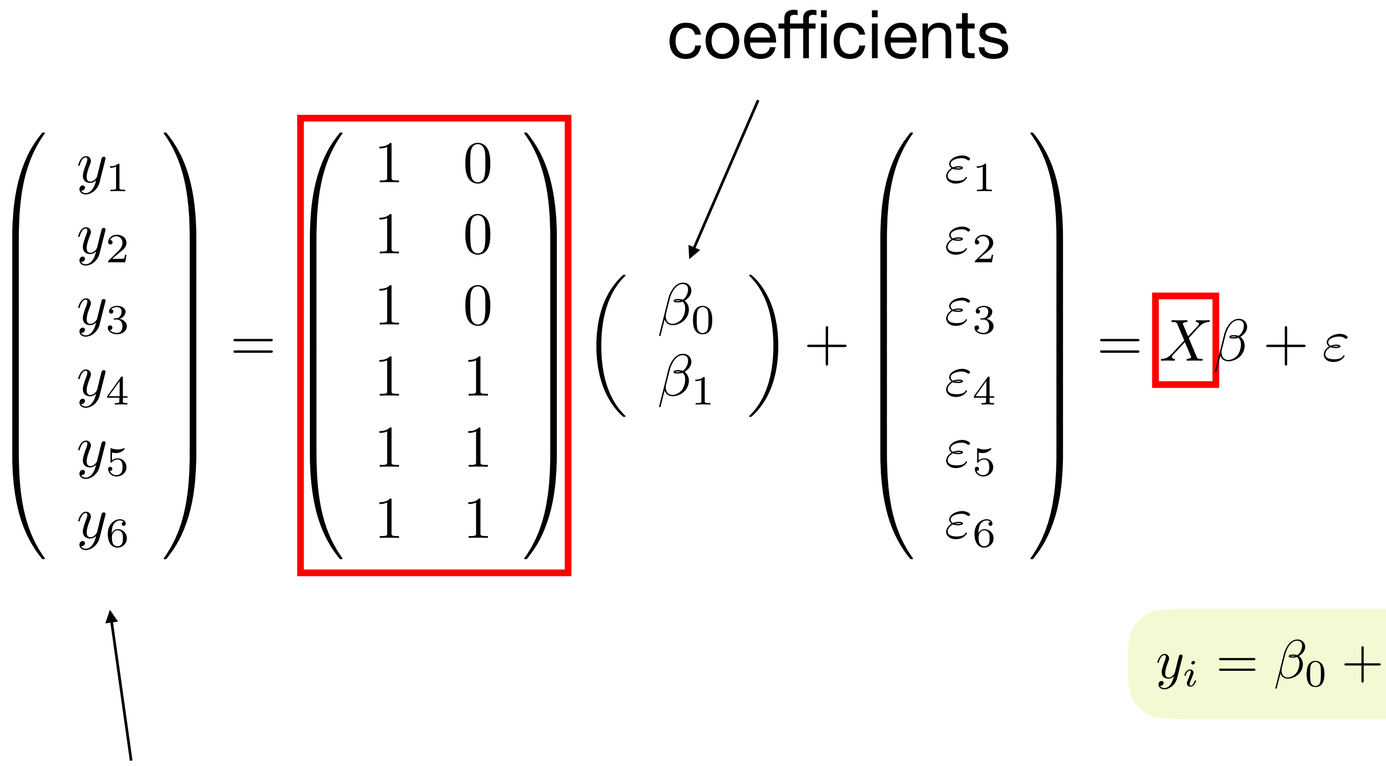
- Fit a separate model for each gene - response variable changes. Specify only predictors

# Testing and contrasts

- After fitting the model(s), we must decide *which* coefficient (or combination thereof) we want to apply a hypothesis test for.
- Combinations of coefficients are called *contrasts*.
- Design matrices can often be defined in many equivalent ways - important that the contrast is defined accordingly!

# Model formulas and design matrices

- A design matrix contains the values of the predictor variables for each sample



# Model formulas and design matrices - example 1
# One predictor, two levels (without intercept)

## Sample table:

| | sample | treatment |
|---|---|---|
| 1 | s1 | control |
| 2 | s2 | control |
| 3 | s3 | control |
| 4 | s4 | treated |
| 5 | s5 | treated |
| 6 | s6 | treated |

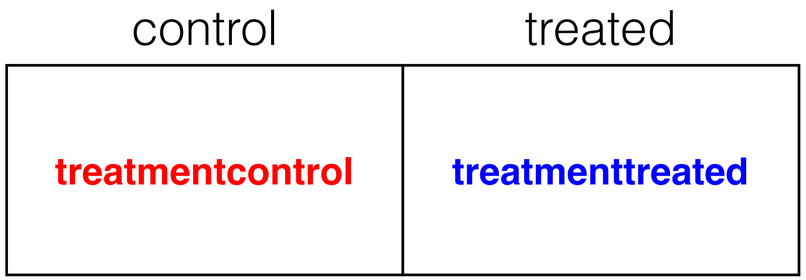
## Design matrix:

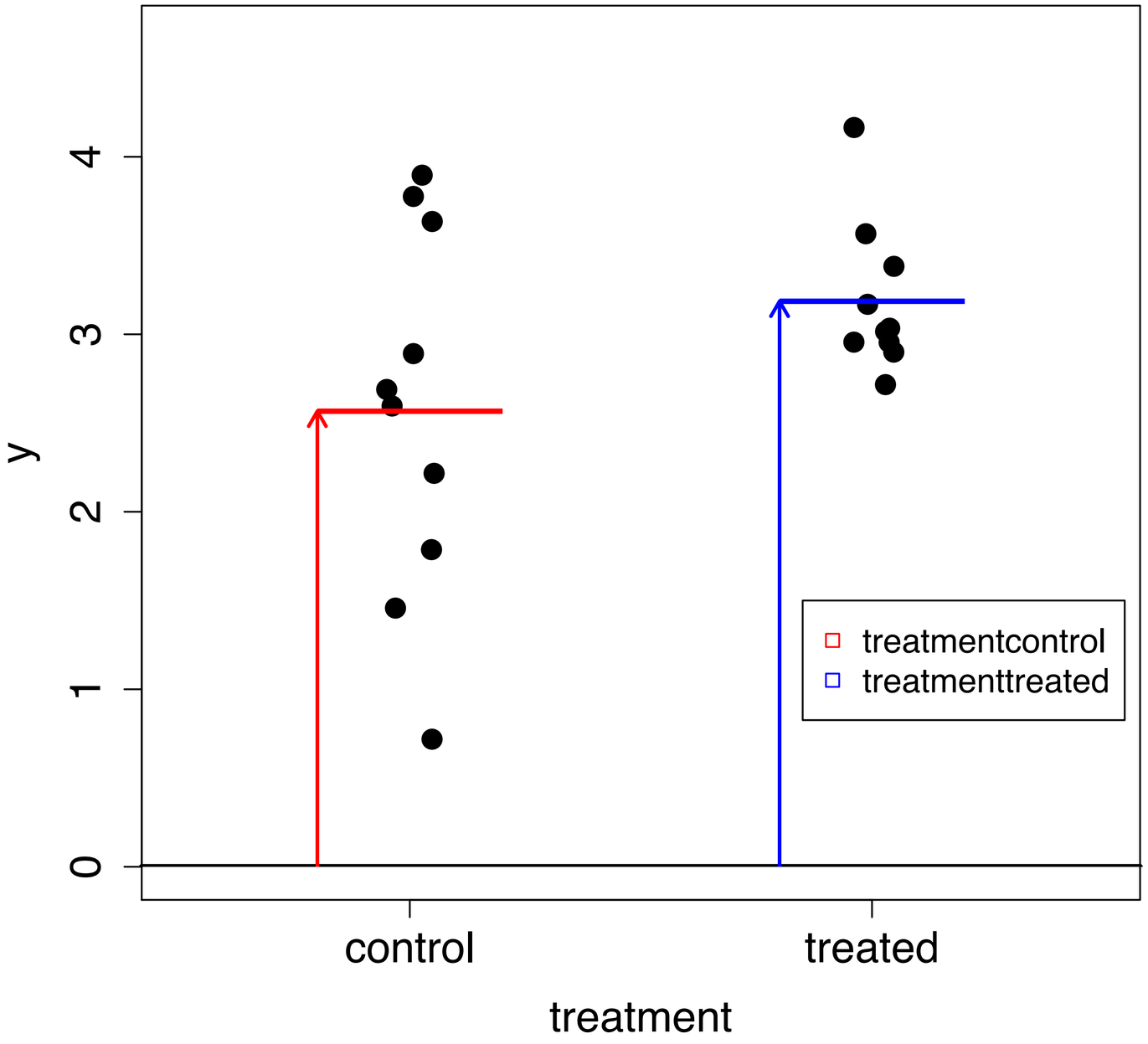
| | treatmentcontrol | treatmenttreated |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 1 | 0 |
| 3 | 1 | 0 |
| 4 | 0 | 1 |
| 5 | 0 | 1 |
| 6 | 0 | 1 |

## Formula:

$$\sim 0 + \text{treatment}$$

## Modeled values:




y
0
1
2
3
4
control
treated
treatment
treatmentcontrol
treatmenttreated

# Model formulas and design matrices - example 1
# One predictor, two levels (with intercept)

## Sample table:

| | sample | treatment |
|---|---|---|
| 1 | s1 | control |
| 2 | s2 | control |
| 3 | s3 | control |
| 4 | s4 | treated |
| 5 | s5 | treated |
| 6 | s6 | treated |

## Design matrix:

| | (Intercept) | treatmenttreated |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 1 | 0 |
| 3 | 1 | 0 |
| 4 | 1 | 1 |
| 5 | 1 | 1 |
| 6 | 1 | 1 |

## Formula:

$$\sim \text{treatment}$$

## Modeled values:

| control | treated |
|---|---|
| 1 * Intercept + 0 * treatmenttreated | 1 * Intercept + 1 * treatmenttreated |

# Model formulas and design matrices - example 1
# One predictor, two levels (with intercept)

## Sample table:

| | sample | treatment |
|---|---|---|
| 1 | s1 | control |
| 2 | s2 | control |
| 3 | s3 | control |
| 4 | s4 | treated |
| 5 | s5 | treated |
| 6 | s6 | treated |

## Design matrix:

| | (Intercept) | treatmenttreated |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 1 | 0 |
| 3 | 1 | 0 |
| 4 | 1 | 1 |
| 5 | 1 | 1 |
| 6 | 1 | 1 |

## Formula:

$$\sim \text{treatment}$$

## Modeled values:

| control | treated |
|---|---|
| 1 * Intercept + 0 * treatmenttreated | 1 * Intercept + 1 * treatmenttreated |

# Model formulas and design matrices - example 1
# One predictor, two levels (with intercept)

## Sample table:

| | sample | treatment |
|---|---|---|
| 1 | s1 | control |
| 2 | s2 | control |
| 3 | s3 | control |
| 4 | s4 | treated |
| 5 | s5 | treated |
| 6 | s6 | treated |

## Design matrix:

| | (Intercept) | treatmenttreated |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 1 | 0 |
| 3 | 1 | 0 |
| 4 | 1 | 1 |
| 5 | 1 | 1 |
| 6 | 1 | 1 |

## Formula:

$$\sim \text{treatment}$$

## Modeled values:

| control | treated |
|---|---|
| 1 * Intercept + 0 * treatmenttreated | 1 * Intercept + 1 * treatmenttreated |

# Model formulas and design matrices - example 1
# One predictor, two levels (with intercept)

## Sample table:

| | sample | treatment |
|---|---|---|
| 1 | s1 | control |
| 2 | s2 | control |
| 3 | s3 | control |
| 4 | s4 | treated |
| 5 | s5 | treated |
| 6 | s6 | treated |

## Design matrix:

| | (Intercept) | treatmenttreated |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 1 | 0 |
| 3 | 1 | 0 |
| 4 | 1 | 1 |
| 5 | 1 | 1 |
| 6 | 1 | 1 |

## Formula:

$$\sim \text{treatment}$$

## Modeled values:

| control | treated |
|---|---|
| 1 * Intercept + 0 * treatmenttreated | 1 * Intercept + 1 * treatmenttreated |

# Model formulas and design matrices - example 1
# One predictor, two levels (with intercept)

## Sample table:

| | sample | treatment |
|---|---|---|
| 1 | s1 | control |
| 2 | s2 | control |
| 3 | s3 | control |
| 4 | s4 | treated |
| 5 | s5 | treated |
| 6 | s6 | treated |

## Design matrix:

| | (Intercept) | treatmenttreated |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 1 | 0 |
| 3 | 1 | 0 |
| 4 | 1 | 1 |
| 5 | 1 | 1 |
| 6 | 1 | 1 |

## Formula:

$$\sim \text{treatment}$$

## Modeled values:

| control | treated |
|---|---|
| 1 * Intercept + 0 * treatmenttreated | 1 * Intercept + 1 * treatmenttreated |

# Model formulas and design matrices - example 1
# One predictor, two levels (with intercept)

## Sample table:

| | sample | treatment |
|---|---|---|
| 1 | s1 | control |
| 2 | s2 | control |
| 3 | s3 | control |
| 4 | s4 | treated |
| 5 | s5 | treated |
| 6 | s6 | treated |

## Design matrix:

| | (Intercept) | treatmenttreated |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 1 | 0 |
| 3 | 1 | 0 |
| 4 | 1 | 1 |
| 5 | 1 | 1 |
| 6 | 1 | 1 |

## Formula:

$$\sim \text{treatment}$$

## Modeled values:

| control | treated |
|---|---|
| 1 * Intercept + 0 * treatmenttreated | 1 * Intercept + 1 * treatmenttreated |

# Model formulas and design matrices - example 1
# One predictor, two levels (with intercept)

## Sample table:

| | sample | treatment |
|---|---|---|
| 1 | s1 | control |
| 2 | s2 | control |
| 3 | s3 | control |
| 4 | s4 | treated |
| 5 | s5 | treated |
| 6 | s6 | treated |

## Design matrix:

| | (Intercept) | treatmenttreated |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 1 | 0 |
| 3 | 1 | 0 |
| 4 | 1 | 1 |
| 5 | 1 | 1 |
| 6 | 1 | 1 |

## Formula:

$$\sim \text{treatment}$$

## Modeled values:

| control | treated |
|---|---|
| 1 * Intercept + 0 * treatmenttreated | 1 * Intercept + 1 * treatmenttreated |

# Model formulas and design matrices - example 1
# One predictor, two levels (with intercept)

## Sample table:

| | sample | treatment |
|---|---|---|
| 1 | s1 | control |
| 2 | s2 | control |
| 3 | s3 | control |
| 4 | s4 | treated |
| 5 | s5 | treated |
| 6 | s6 | treated |

## Design matrix:

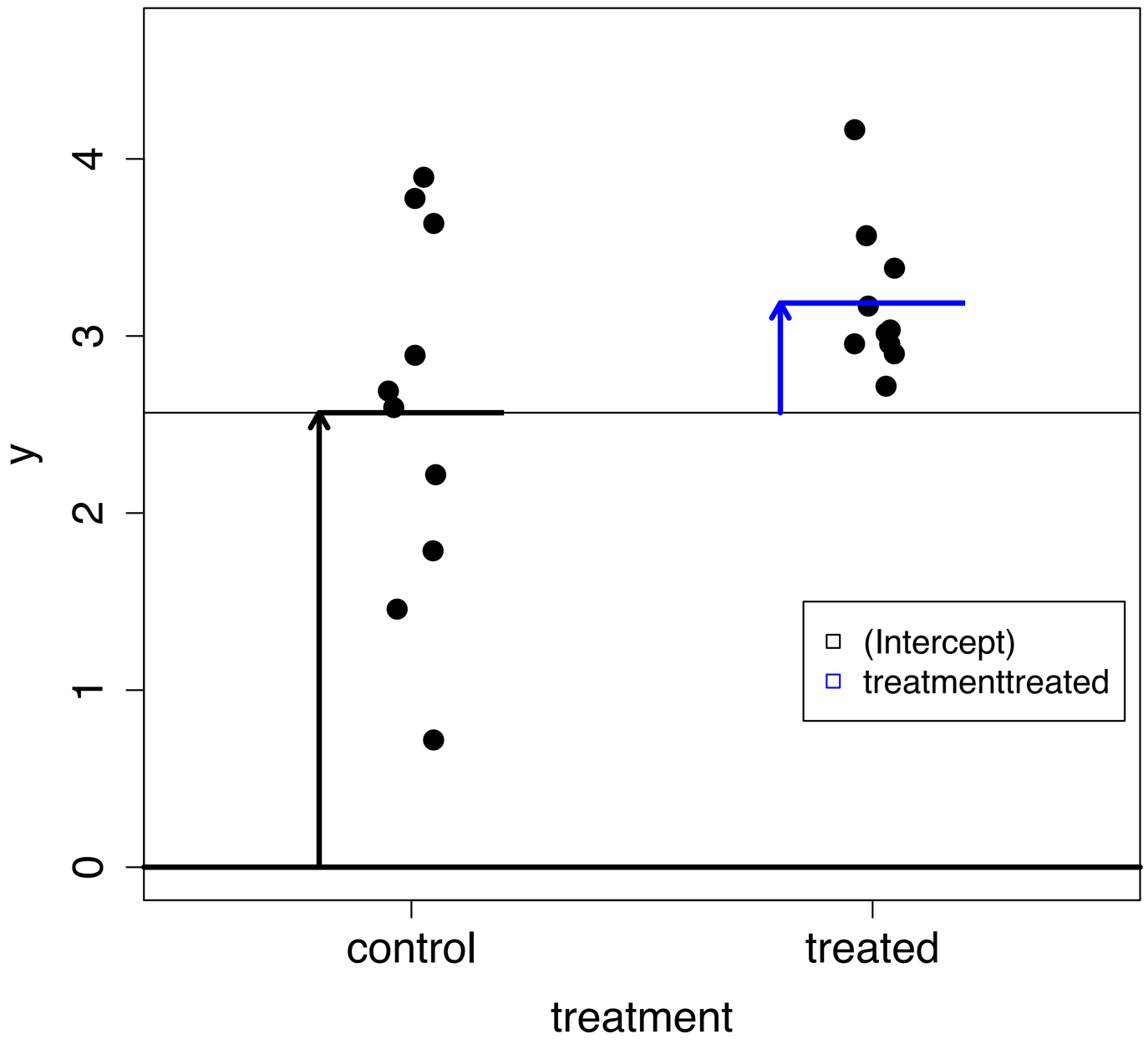
| | (Intercept) | treatmenttreated |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 1 | 0 |
| 3 | 1 | 0 |
| 4 | 1 | 1 |
| 5 | 1 | 1 |
| 6 | 1 | 1 |

## Formula:

$$\sim \text{treatment}$$

## Modeled values:

| control | treated |
|---|---|
| **Intercept** | **Intercept + treatmenttreated** |


y
0
1
2
3
4
control
treated
treatment
(Intercept)
treatmenttreated

# Model formulas and design matrices - example 2
# One continuous predictor

## Sample table:

| | sample | age |
|---|---|---|
| 1 | s1 | 21 |
| 2 | s2 | 12 |
| 3 | s3 | 64 |
| 4 | s4 | 44 |
| 5 | s5 | 19 |
| 6 | s6 | 26 |

## Design matrix:

| | (Intercept) | age |
|---|---|---|
| 1 | 1 | 21 |
| 2 | 1 | 12 |
| 3 | 1 | 64 |
| 4 | 1 | 44 |
| 5 | 1 | 19 |
| 6 | 1 | 26 |

## Formula:

$$\sim \text{age}$$

## Modeled values:

| s1 | s2 | s3 | s4 | s5 | s6 |
|---|---|---|---|---|---|
| Intercept + 21 * age | Intercept + 12 * age | Intercept + 64 * age | Intercept + 44 * age | Intercept + 19 * age | Intercept + 26 * age |

# Model formulas and design matrices - example 3
# One predictor, three levels

## Sample table:

| | sample | treatment |
|---|---|---|
| 1 | s1 | control |
| 2 | s2 | control |
| 3 | s3 | treatA |
| 4 | s4 | treatA |
| 5 | s5 | treatB |
| 6 | s6 | treatB |

## Design matrix:

| | (Intercept) | treatmenttreatA | treatmenttreatB |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 2 | 1 | 0 | 0 |
| 3 | 1 | 1 | 0 |
| 4 | 1 | 1 | 0 |
| 5 | 1 | 0 | 1 |
| 6 | 1 | 0 | 1 |

## Formula:

$$\sim \text{treatment}$$

## Modeled values:

| control | treatA | treatB |
|---|---|---|
| **Intercept** | **Intercept + treatmenttreatA** | **Intercept + treatmenttreatB** |

# Model formulas and design matrices - example 4
# One predictor, paired data (or two predictors)

## Sample table:

| | sample | treatment |
|---|---|---|
| 1 | s1 | control |
| 2 | s1 | treated |
| 3 | s2 | control |
| 4 | s2 | treated |
| 5 | s3 | control |
| 6 | s3 | treated |

## Design matrix:

| | (Intercept) | samples2 | samples3 | treatmenttreated |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 1 |
| 3 | 1 | 1 | 0 | 0 |
| 4 | 1 | 1 | 0 | 1 |
| 5 | 1 | 0 | 1 | 0 |
| 6 | 1 | 0 | 1 | 1 |

## Formula:

$$\sim \text{sample} + \text{treatment}$$

## Modeled values:

| | s1 | s2 | s3 |
|---|---|---|---|
| control | **Intercept** | **Intercept + samples2** | **Intercept + samples3** |
| treated | **Intercept + treatmenttreated** | **Intercept + samples2 + treatmenttreated** | **Intercept + samples3 + treatmenttreated** |

# Model formulas and design matrices - example 4
# One predictor, paired data (or two predictors)

## Sample table:

| | genotype | treatment |
|---|---|---|
| 1 | A | control |
| 2 | A | control |
| 3 | A | treated |
| 4 | A | treated |
| 5 | B | control |
| 6 | B | control |
| 7 | B | treated |
| 8 | B | treated |

## Design matrix:

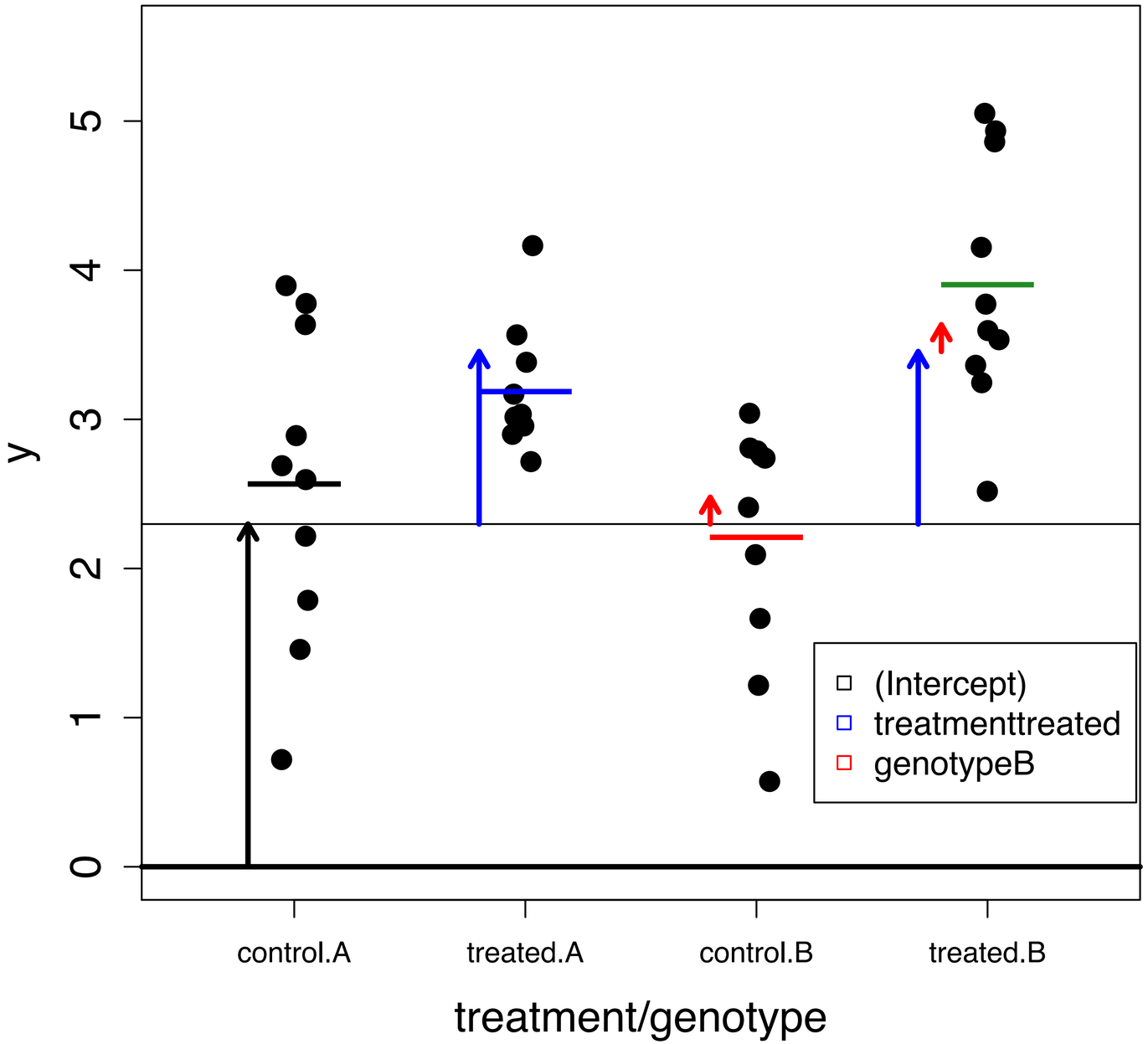
| | (Intercept) | genotypeB | treatmenttreated |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 2 | 1 | 0 | 0 |
| 3 | 1 | 0 | 1 |
| 4 | 1 | 0 | 1 |
| 5 | 1 | 1 | 0 |
| 6 | 1 | 1 | 0 |
| 7 | 1 | 1 | 1 |
| 8 | 1 | 1 | 1 |

## Formula:

$$\sim \text{genotype} + \text{treatment}$$

## Modeled values:

| | genotype A | genotype B |
|---|---|---|
| control | **Intercept** | **Intercept + genotypeB** |
| treated | **Intercept + treatmenttreated** | **Intercept + genotypeB + treatmenttreated** |


y
0
1
2
3
4
5
control.A
treated.A
control.B
treated.B
treatment/genotype
(Intercept)
treatmenttreated
genotypeB

# Model formulas and design matrices - example 5
# Two predictors, with interaction

## Sample table:

| | genotype | treatment |
|---|---|---|
| 1 | A | control |
| 2 | A | control |
| 3 | A | treated |
| 4 | A | treated |
| 5 | B | control |
| 6 | B | control |
| 7 | B | treated |
| 8 | B | treated |

## Design matrix:

| | (Intercept) | genotypeB | treatmenttreated | genotypeB:treatmenttreated |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 0 |
| 3 | 1 | 0 | 1 | 0 |
| 4 | 1 | 0 | 1 | 0 |
| 5 | 1 | 1 | 0 | 0 |
| 6 | 1 | 1 | 0 | 0 |
| 7 | 1 | 1 | 1 | 1 |
| 8 | 1 | 1 | 1 | 1 |

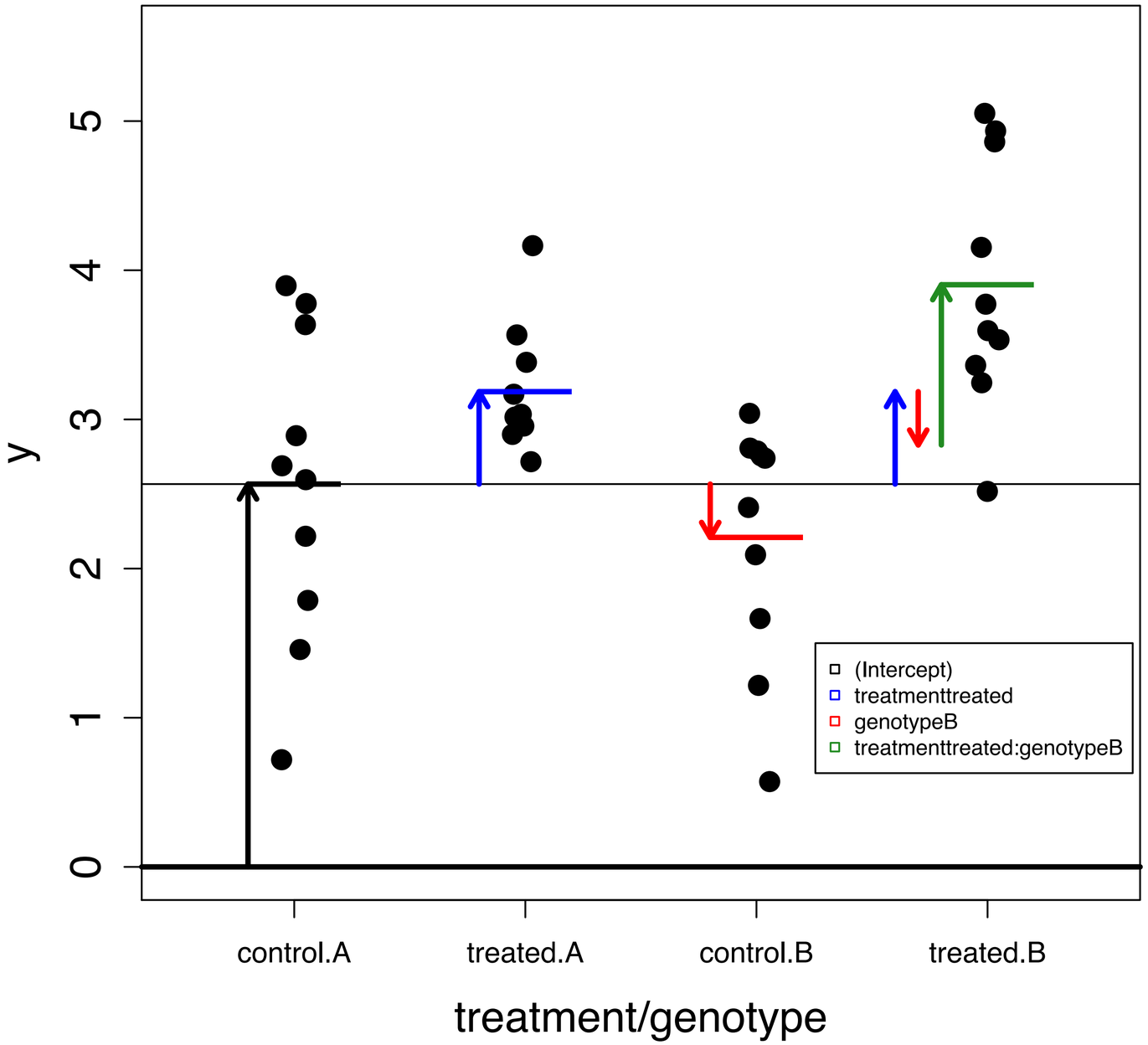
## Formula:

$\sim$ genotype * treatment

$\sim$ genotype + treatment + genotype:treatment

## Modeled values:

| | genotype A | genotype B |
|---|---|---|
| control | **Intercept** | **Intercept + genotypeB** |
| treated | **Intercept + treatmenttreated** | **Intercept + genotypeB + treatmenttreated + genotypeB:treatmenttreated** |


y
0
1
2
3
4
5
control.A
treated.A
control.B
treated.B
treatment/genotype
(Intercept)
treatmenttreated
genotypeB
treatmenttreated:genotypeB

# Model formulas and design matrices - example 6
# Two predictors, with interaction

## Sample table:

| | treat.gt |
|---|---|
| 1 | control.A |
| 2 | control.A |
| 3 | treated.A |
| 4 | treated.A |
| 5 | control.B |
| 6 | control.B |
| 7 | treated.B |
| 8 | treated.B |

## Design matrix:

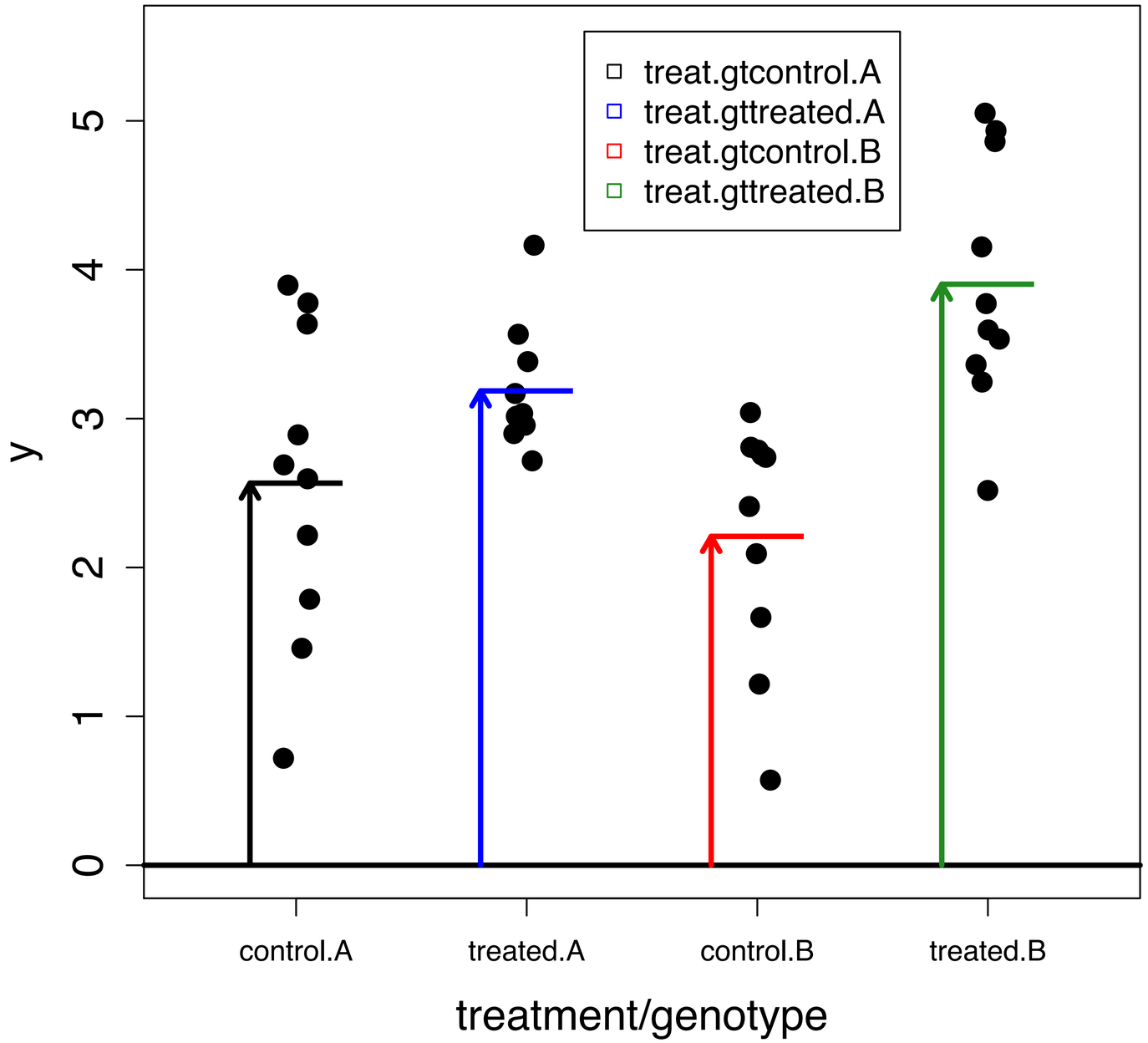
| | treat.gtcontrol.A | treat.gttreated.A | treat.gtcontrol.B | treat.gttreated.B |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 0 |
| 3 | 0 | 1 | 0 | 0 |
| 4 | 0 | 1 | 0 | 0 |
| 5 | 0 | 0 | 1 | 0 |
| 6 | 0 | 0 | 1 | 0 |
| 7 | 0 | 0 | 0 | 1 |
| 8 | 0 | 0 | 0 | 1 |

## Formula:

$$\sim 0 + \text{treat.gt}$$

## Modeled values:

| | genotype A | genotype B |
|---|---|---|
| control | **treat.gtcontrol.A** | **treat.gtcontrol.B** |
| treated | **treat.gttreated.A** | **treat.gttreated.B** |


treat.gtcontrol.A
treat.gttreated.A
treat.gtcontrol.B
treat.gttreated.B
y
0
1
2
3
4
5
control.A
treated.A
control.B
treated.B
treatment/genotype

# Using contrasts vs subsetting data set

- Fitting model to full data set and using contrasts gives more samples to estimate parameters (generally recommended)
- Also assumes that dispersion is similar in all groups (estimates one dispersion parameter per gene)
- In some situations, subsetting to only groups of interest is advantageous:

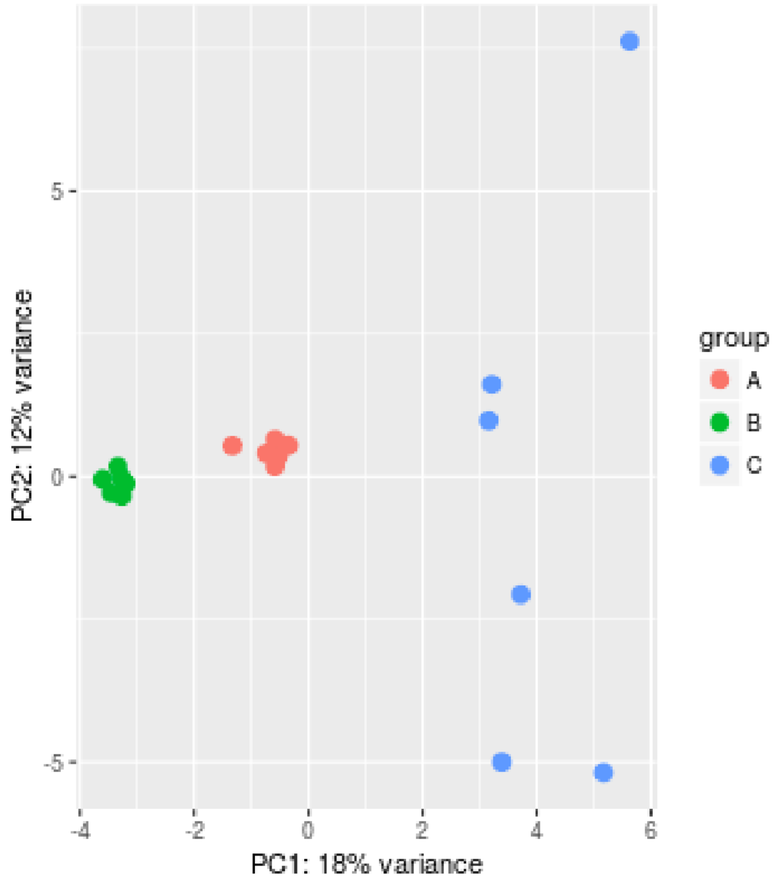
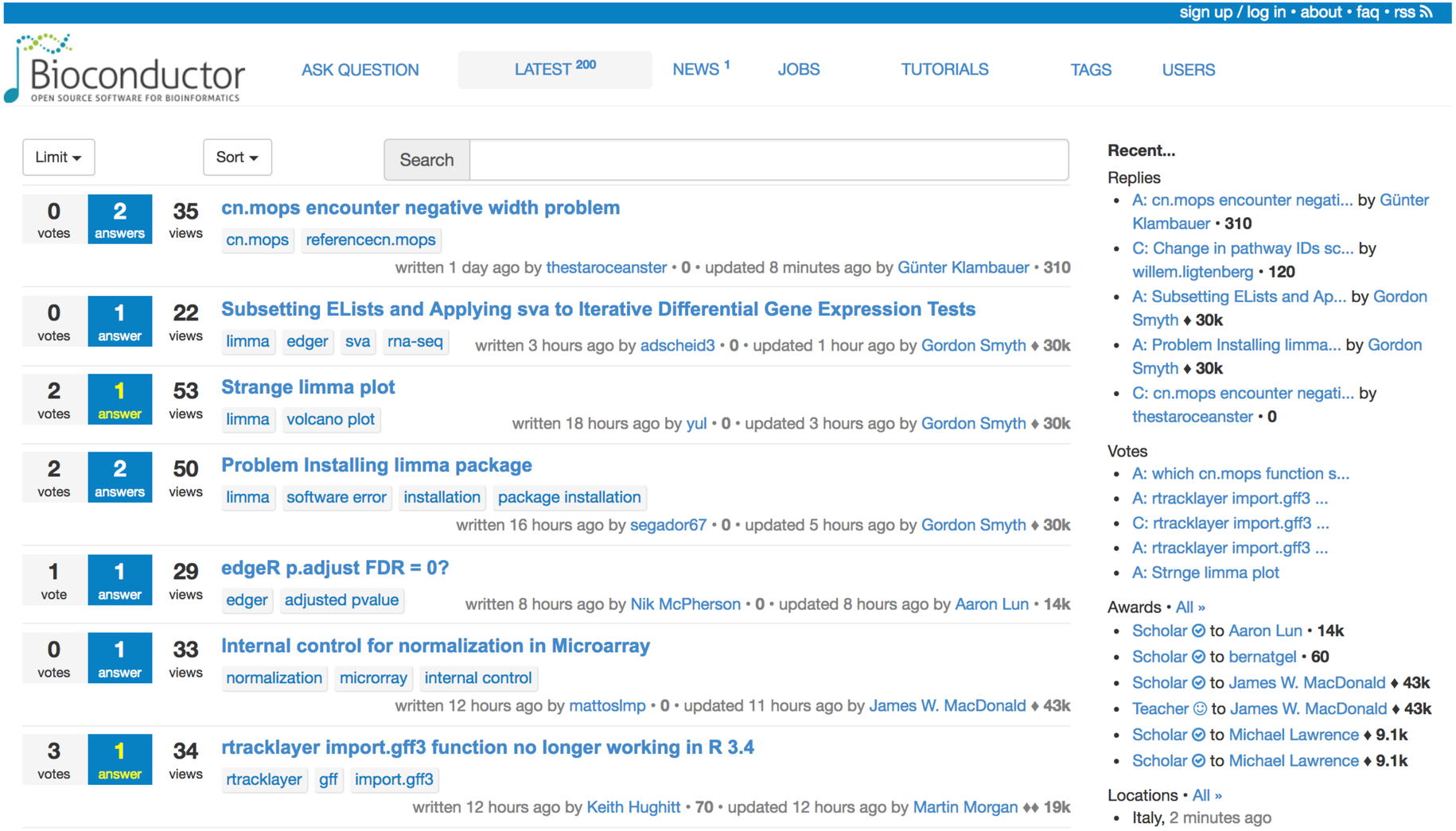



Illustration from DESeq2 vignette

# https://support.bioconductor.org/



# DESeq2

platforms all | downloads top 5% | posts 213 / 1 / 3 / 29 | in Bioc 4 years
build ok | commits 9.00 | test coverage 95%

## Differential gene expression analysis based on the negative binomial distribution

Bioconductor version: Release (3.4)

Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution.

Author: Michael Love, Simon Anders, Wolfgang Huber

Maintainer: Michael Love <michaelisaiahlove at gmail.com>

Citation (from within R, enter `citation("DESeq2")`):

> Love MI, Huber W and Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology*, **15**, pp. 550. doi: 10.1186/s13059-014-0550-8.

### Installation

To install this package, start R and enter:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("DESeq2")
```

### Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("DESeq2")
```

| | | |
|---|---|---|
| PDF | R Script | Analyzing RNA-seq data with the "DESeq2" package |
| PDF | | Reference Manual |
| Text | | NEWS |

## Documentation »

*Bioconductor*

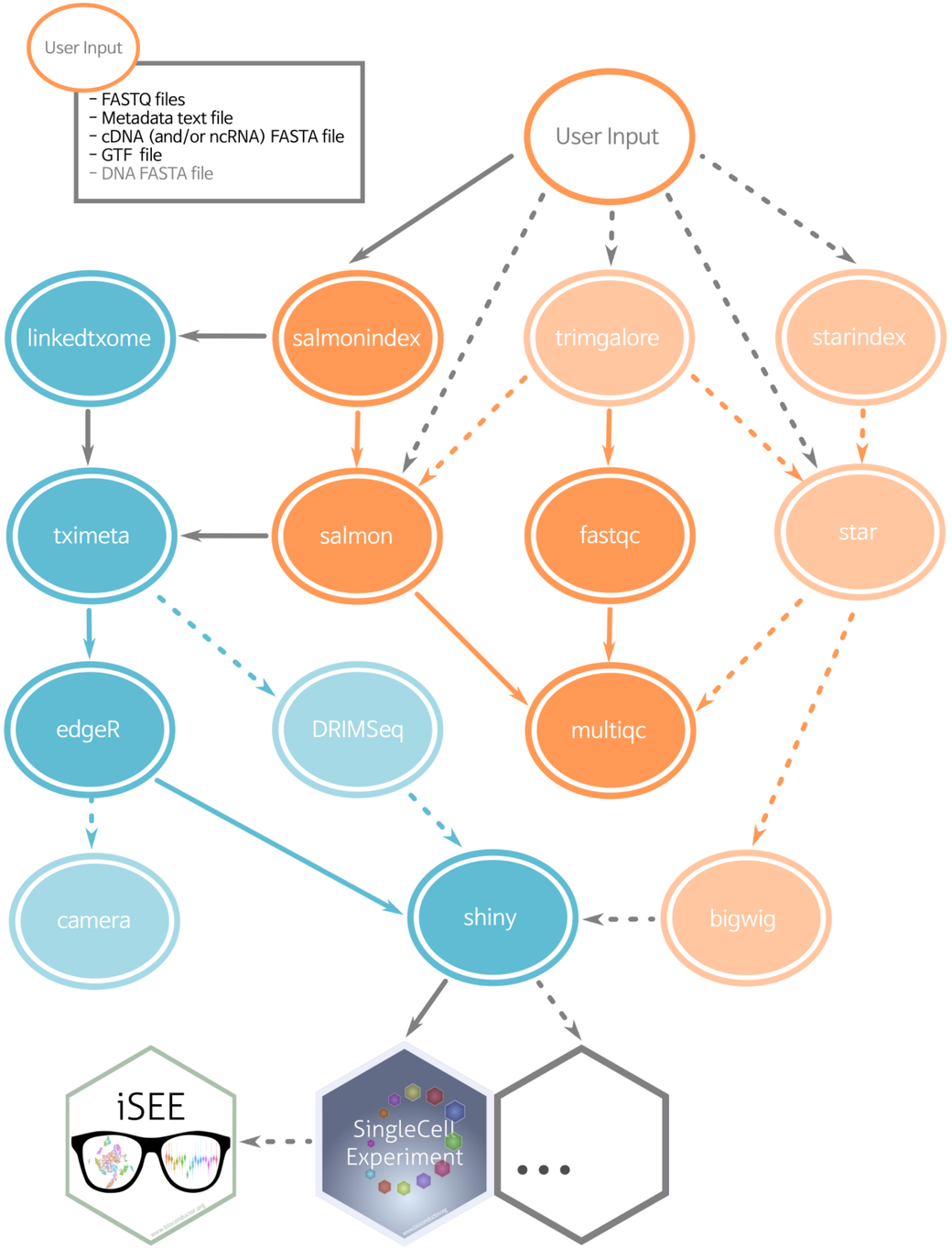
- Package vignettes and manuals.
- Workflows for learning and use.
- Course and conference material.
- Videos.
- Community resources and tutorials.

*R* / CRAN packages and documentation

## Support »

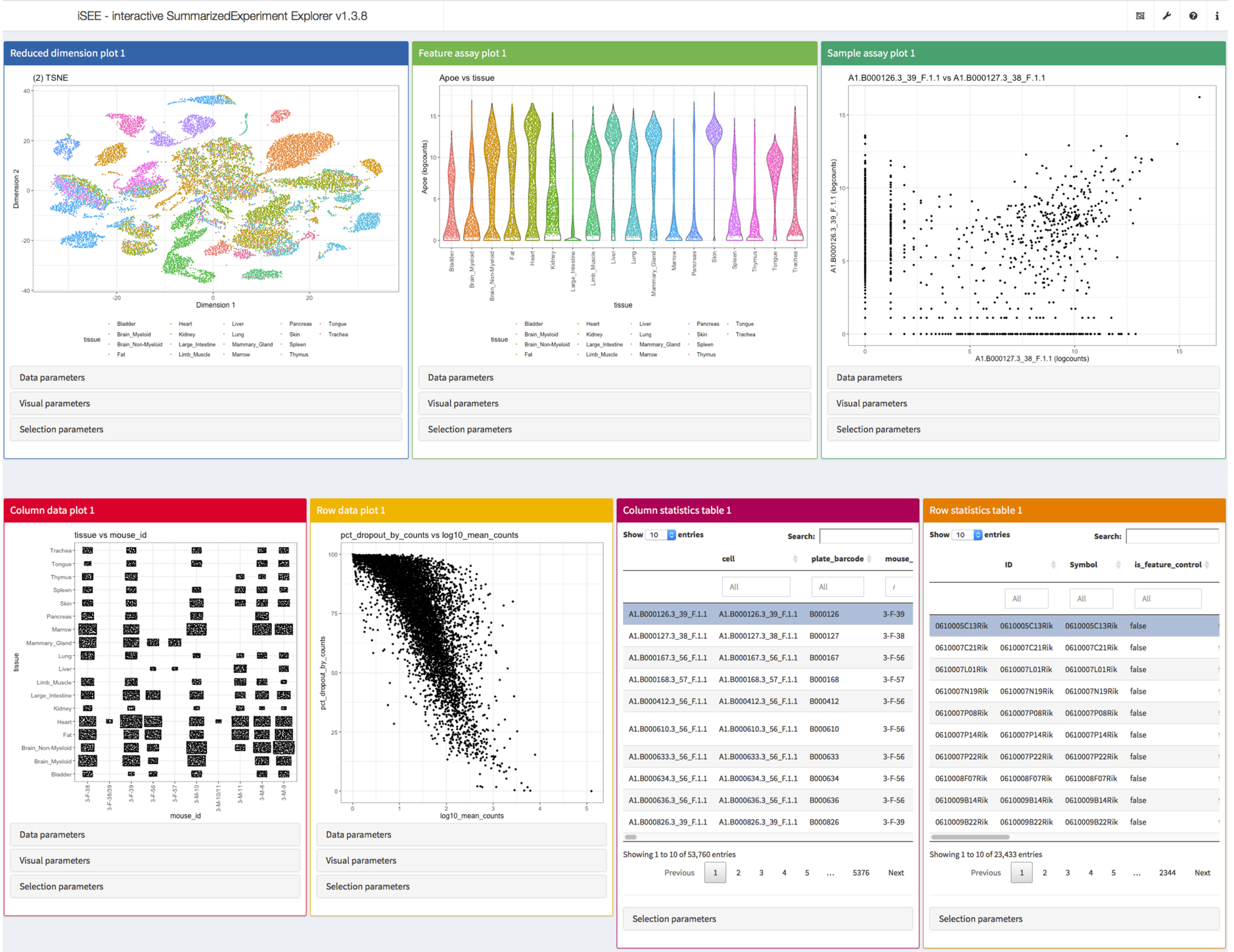
Please read the posting guide. Post questions about Bioconductor to one of the following locations:

- Support site - for questions about Bioconductor packages
- Bioc-devel mailing list - for package developers




- Automated, reproducible, modular RNA-seq workflow
- Covers a standard “end-to-end” RNA-seq analysis
- Based on Snakemake and conda environments
- https://github.com/csoneson/ARMOR


- Interactive exploration tool for any ‘rectangular’ numeric data
- Available from Bioconductor



# References


- van den Berge, Hembach, Soneson, Tiberi et al: RNA sequencing data: hitchhiker's guide to expression analysis. PeerJ Preprints 6:e27283v2 (2018) - **review of RNA-seq**
- Orjuela, Huang, Hembach et al: ARMOR: an Automated Reproducible MOdular workflow for preprocessing and differential analysis of RNA-seq data. bioRxiv doi:10.1101/575951 (2019) - **RNA-seq workflow**
- Rue-Albrecht, Marini, Soneson & Lun: iSEE: Interactive SummarizedExperiment Explorer. F1000Research 7:741 (2018) - **iSEE**
- Love, Soneson & Patro: Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. F1000Research 7:952 (2018) - **DTU workflow**
- Robinson et al.: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139-140 (2010) - **edgeR**
- Love et al.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15:550 (2014) - **DESeq2**
- Law et al.: voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology 15:R29 (2014) - **voom**
- Patro et al.: Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods 14:4170419 (2017) - **Salmon**
- Bray et al.: Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology 34(5):525-527 (2016) - **kallisto**
- Patro et al.: Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nature Biotechnology 32:462-464 (2014) - **Sailfish**
- Pimentel et al.: Differential analysis of RNA-Seq incorporating quantification uncertainty. bioRxiv http://dx.doi.org/10.1101/058164 (2016) - **sleuth**
- Wagner et al.: Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences 131:281-285 (2012) - **TPM vs FPKM**
- Soneson et al.: Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research 4:1521 (2016) - **ATL offsets (tximport package)**
- Li et al.: RNA-seq gene expression estimation with read mapping uncertainty. Bioinformatics 26(4):493-500 (2010) - **TPM, RSEM**
- Soneson, Matthes et al.: Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. Genome Biology 17:12 (2016)
- Schurch et al.: How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 22:839-851 (2016)
- Dillies et al.: A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics 14(6):671-683 (2013)
- Soneson & Delorenzi: A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 14:91 (2013)
- Anders et al.: Detecting differential usage of exons from RNA-seq data. Genome Research 22(10):2008-2017 (2012) - **DEXSeq**
- Love et al.: RNA-Seq workflow: gene-level exploratory analysis and differential expression. F1000 Research 4:1070 (2016) - **RNA-seq workflow**
- Law et al.: RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000 Research 5:1408 (2016) - **RNA-seq workflow**
- Chen et al: From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000 Research 5:1438 (2016) - **RNA-seq workflow**