Differential Expression Analysis using RNA-seq

Mudassar Iqbal

WT ICD Informatics Bootcamp

9th – 13th November, 2020

Accompanying R notebook, covering:

- Example data
- Expression quantification (data preparation for DE analysis using *DESeq2*)
- Working with *DESeqDataSet* object
- Data exploration and visualization
- Differential expression analysis
- Exploring the DE output

Experimental data --- see R notebook

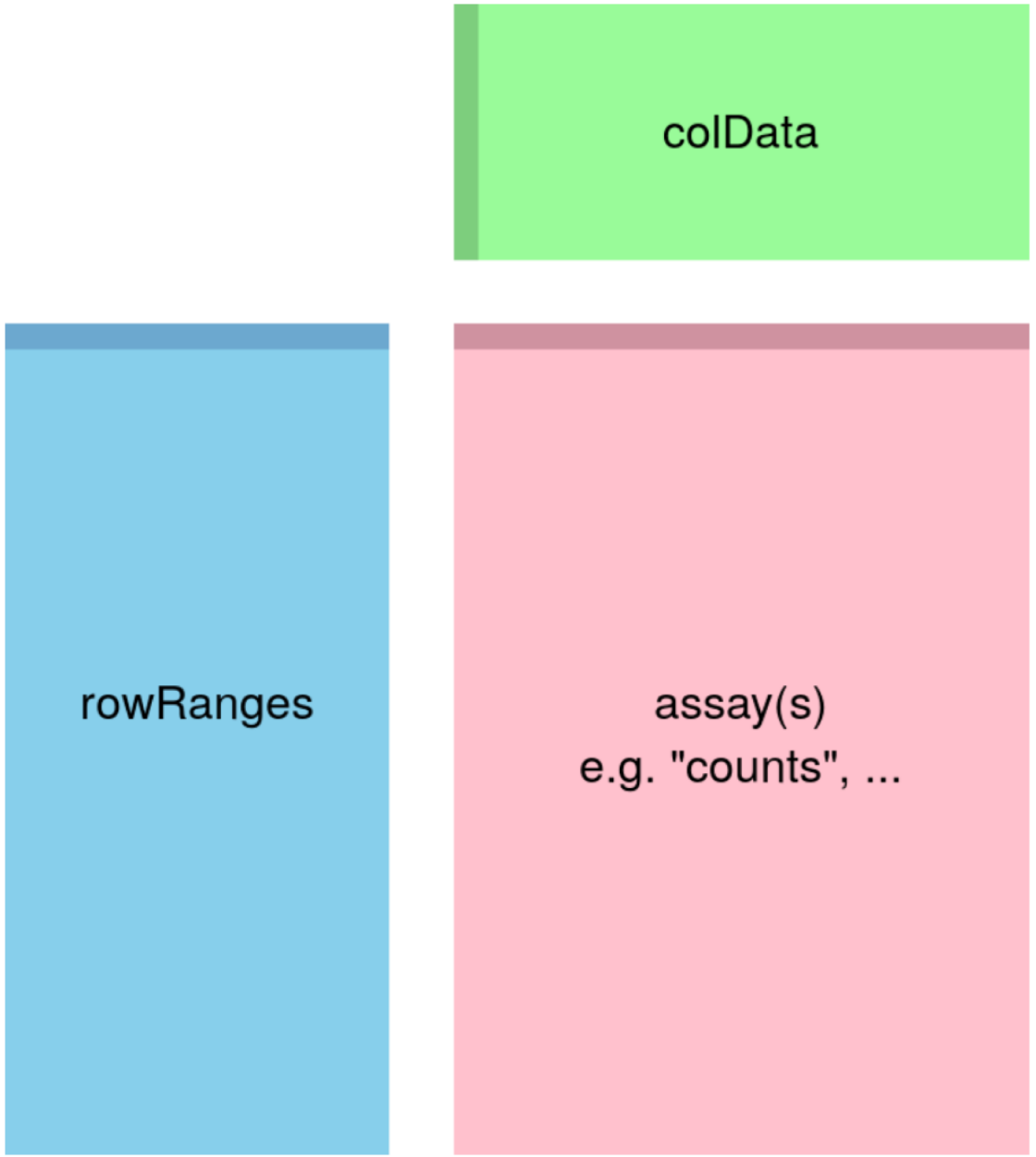
PLoS One. 2014 Jun 13;9(6):e99625. doi: 10.1371/journal.pone.0099625. eCollection 2014.

RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells

```
##  [1] "GSE52778_series_matrix.txt"       "Homo_sapiens.GRCh37.75_subset.gtf"
##  [3] "SRR1039508_subset.bam"            "SRR1039509_subset.bam"
##  [5] "SRR1039512_subset.bam"            "SRR1039513_subset.bam"
##  [7] "SRR1039516_subset.bam"            "SRR1039517_subset.bam"
##  [9] "SRR1039520_subset.bam"            "SRR1039521_subset.bam"
## [11] "SraRunInfo_SRP033351.csv"         "quants"
## [13] "sample_table.csv"
```

```
## DataFrame with 8 rows and 5 columns
##                  names    donor     condition     cell      dex
##              <factor> <factor>      <factor> <factor> <factor>
## SRR1039508 SRR1039508  N61311  Untreated      N61311     untrt
## SRR1039509 SRR1039509  N61311  Dexamethasone  N61311     trt
## SRR1039512 SRR1039512  N052611 Untreated      N052611    untrt
## SRR1039513 SRR1039513  N052611 Dexamethasone  N052611    trt
## SRR1039516 SRR1039516  N080611 Untreated      N080611    untrt
## SRR1039517 SRR1039517  N080611 Dexamethasone  N080611    trt
## SRR1039520 SRR1039520  N061011 Untreated      N061011    untrt
## SRR1039521 SRR1039521  N061011 Dexamethasone  N061011    trt
```

SummarizedExperiment



Starting with count matrix

```
##                    SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516
## ENSG00000000003.14        708        468        901        424       1188
## ENSG00000000005.5           0          0          0          0          0
## ENSG00000000419.12        455        510        604        352        583
##                    SRR1039517 SRR1039520 SRR1039521
## ENSG00000000003.14       1091        806        599
## ENSG00000000005.5           0          0          0
## ENSG00000000419.12        774        410        499
```

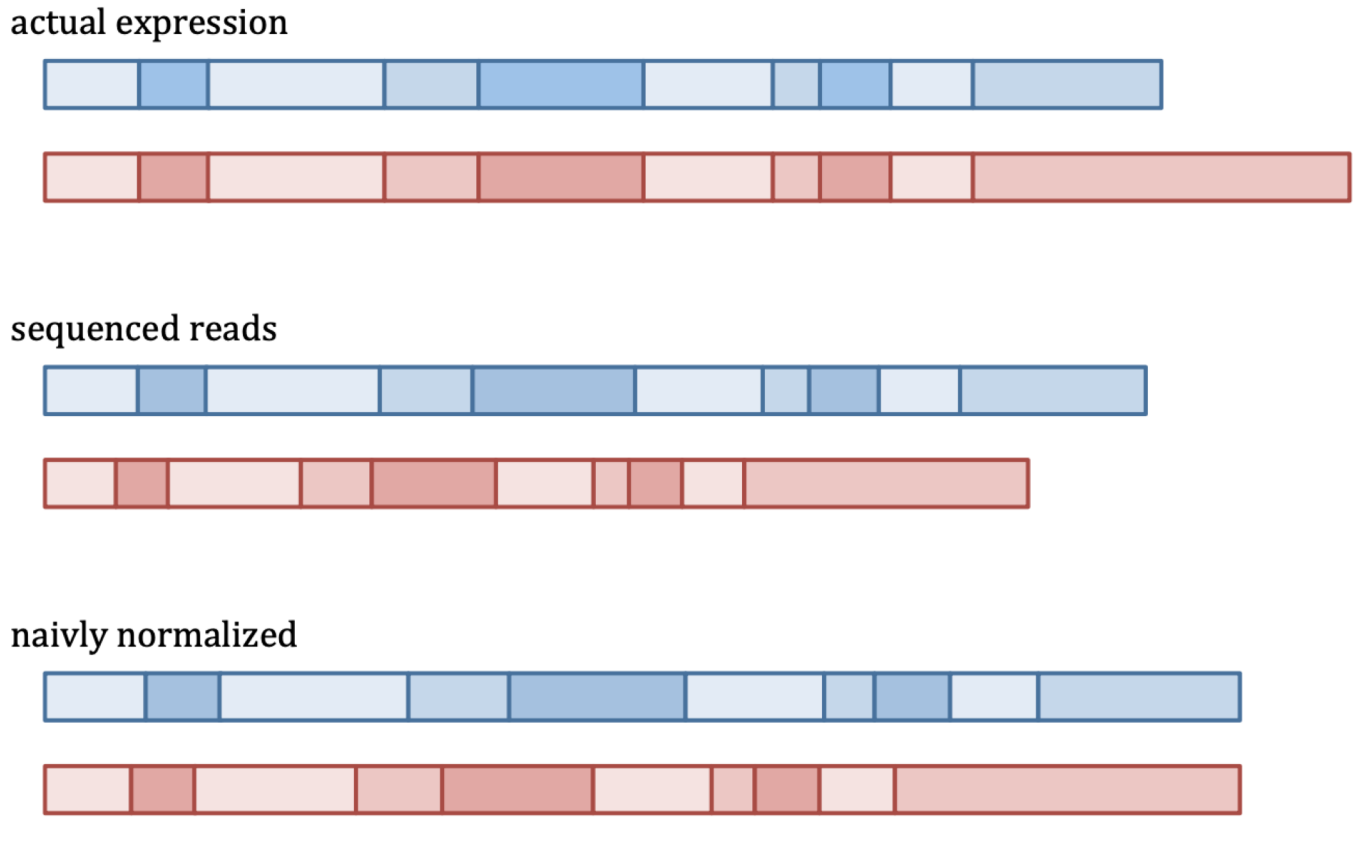
Key Challenges:

- Large dynamic range
- Data are non-negative integers, with non-symmetric distribution (normal or log-normal distributions are not suitable!)
- Sampling biases – normalization (total sequencing depth of an experiment)
- And more!

Normalization for library size

Normalization for library size

- If sample A has been sampled deeper than sample B, we expect counts to be higher.
- Naive approach: Divide by the total number of reads per sample
- Problem: Genes that are strongly and differentially expressed may distort the ratio of total reads.



Normalization for library size

To compare more than two samples:

- Form a “virtual reference sample” by taking, for each gene, the geometric mean of counts over all samples
- Normalize each sample to this reference, to get one scaling factor (“size factor”) per sample.

Anders and Huber, 2010

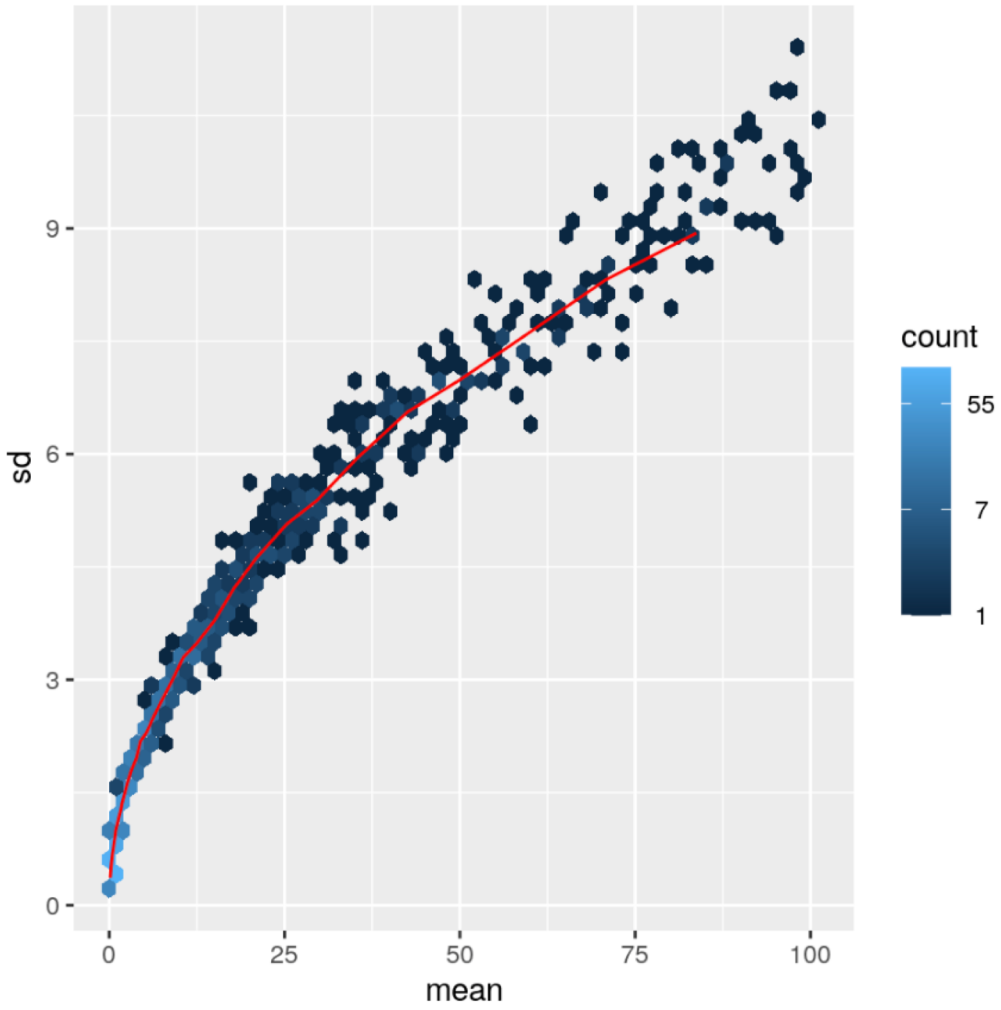
similar approach: Robinson and Oshlack, 2010

Variance stabilizing and rlog transformations:

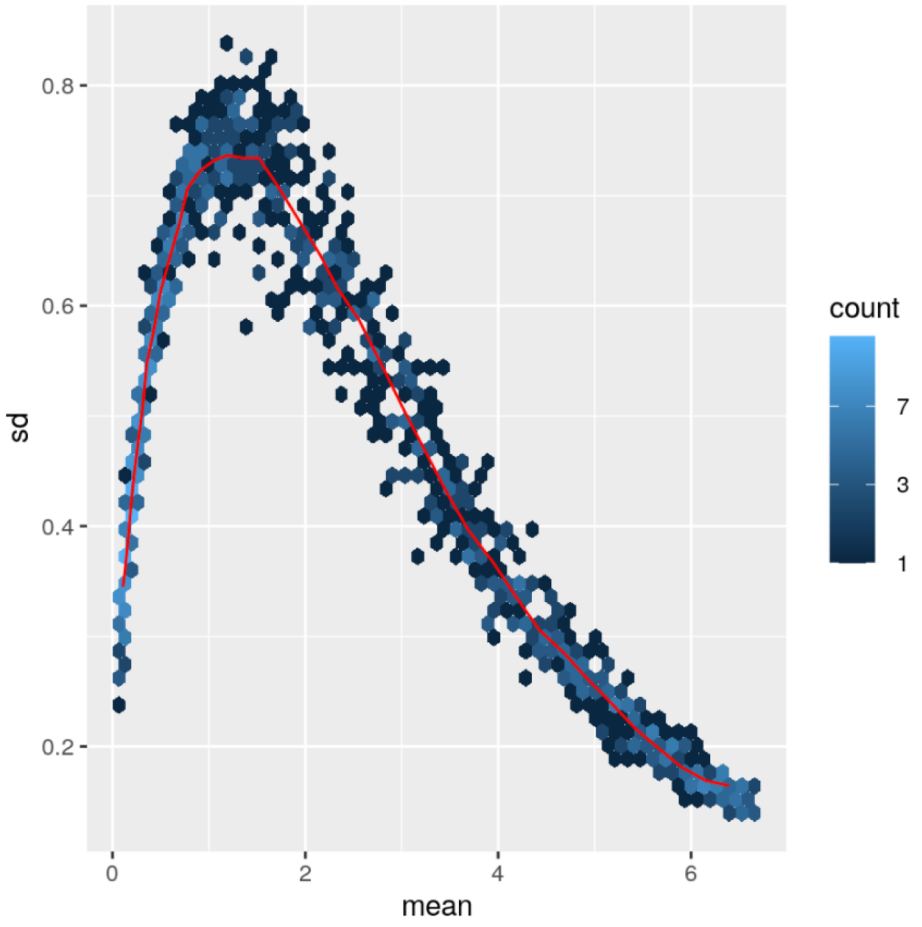
Why we need them!

Simulated data

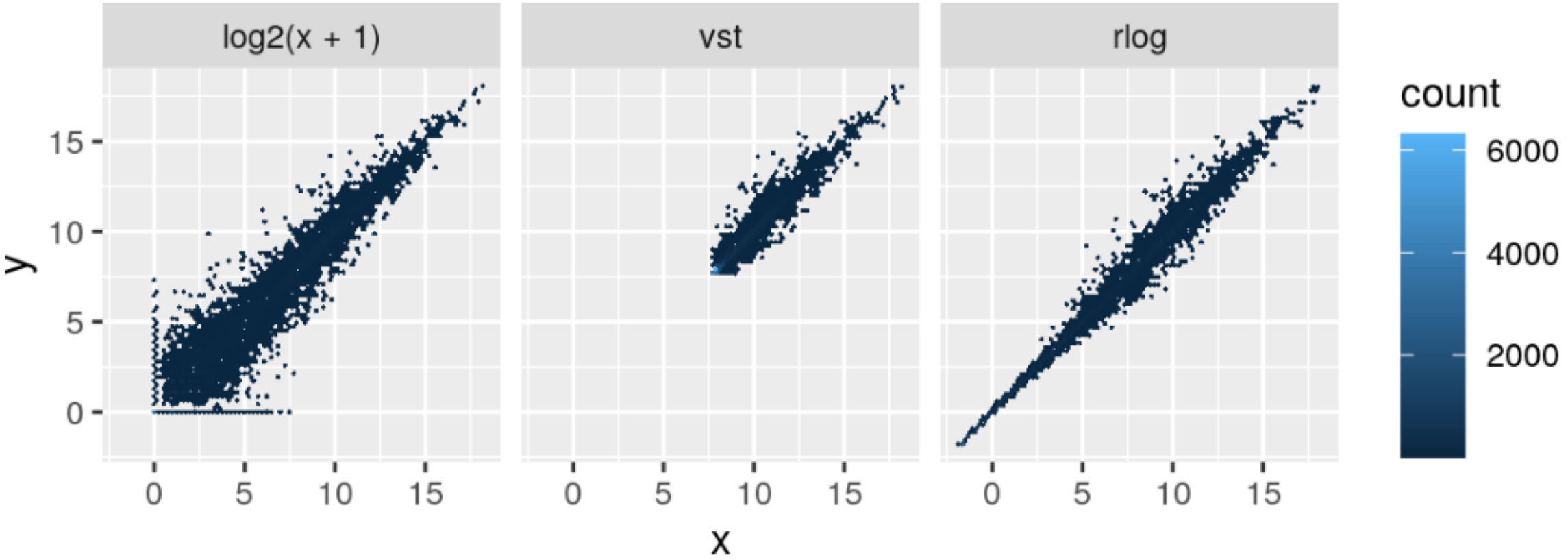
Poisson counts



logarithm-transformed counts

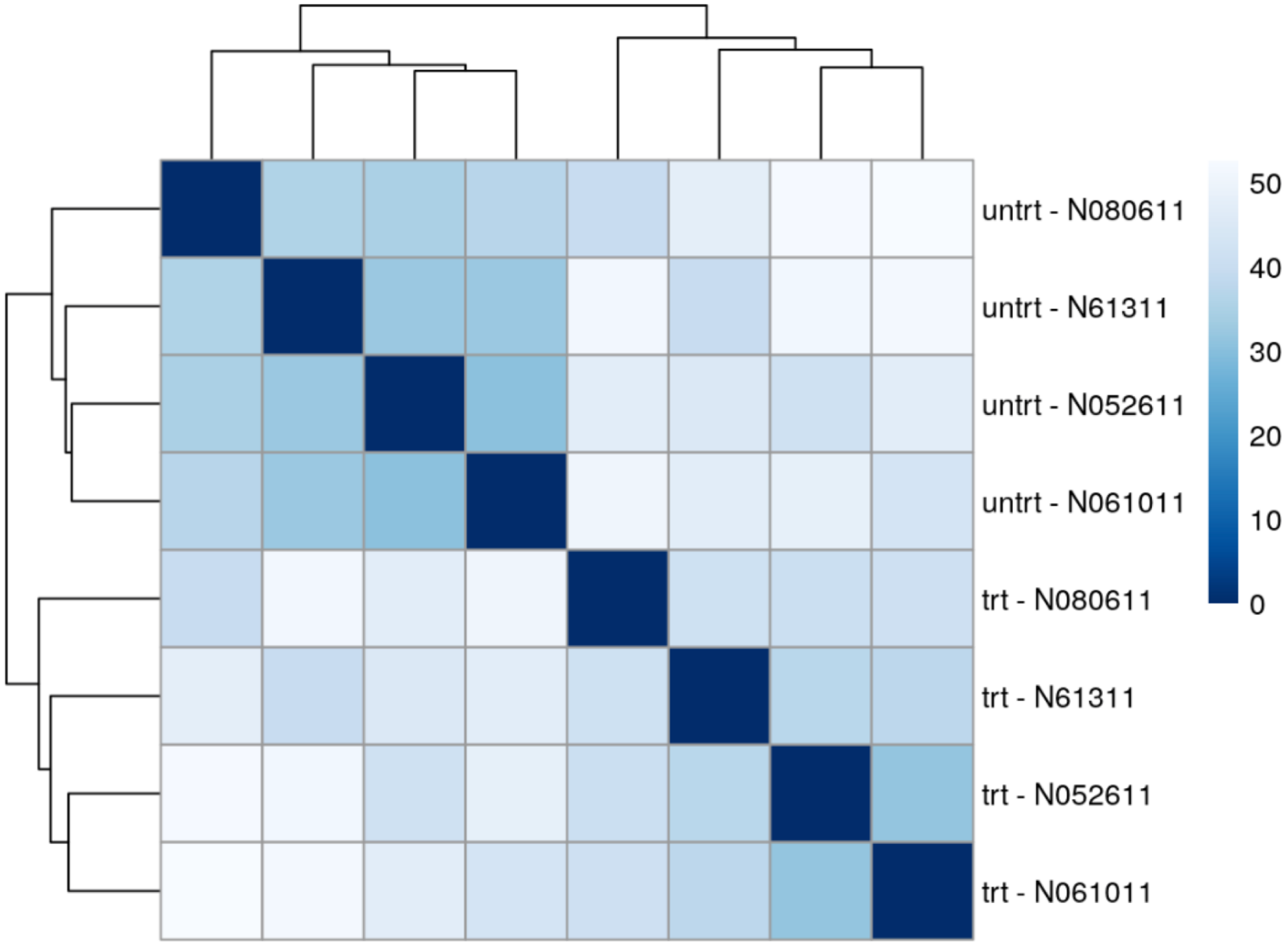


First two samples from ‘airway’ data

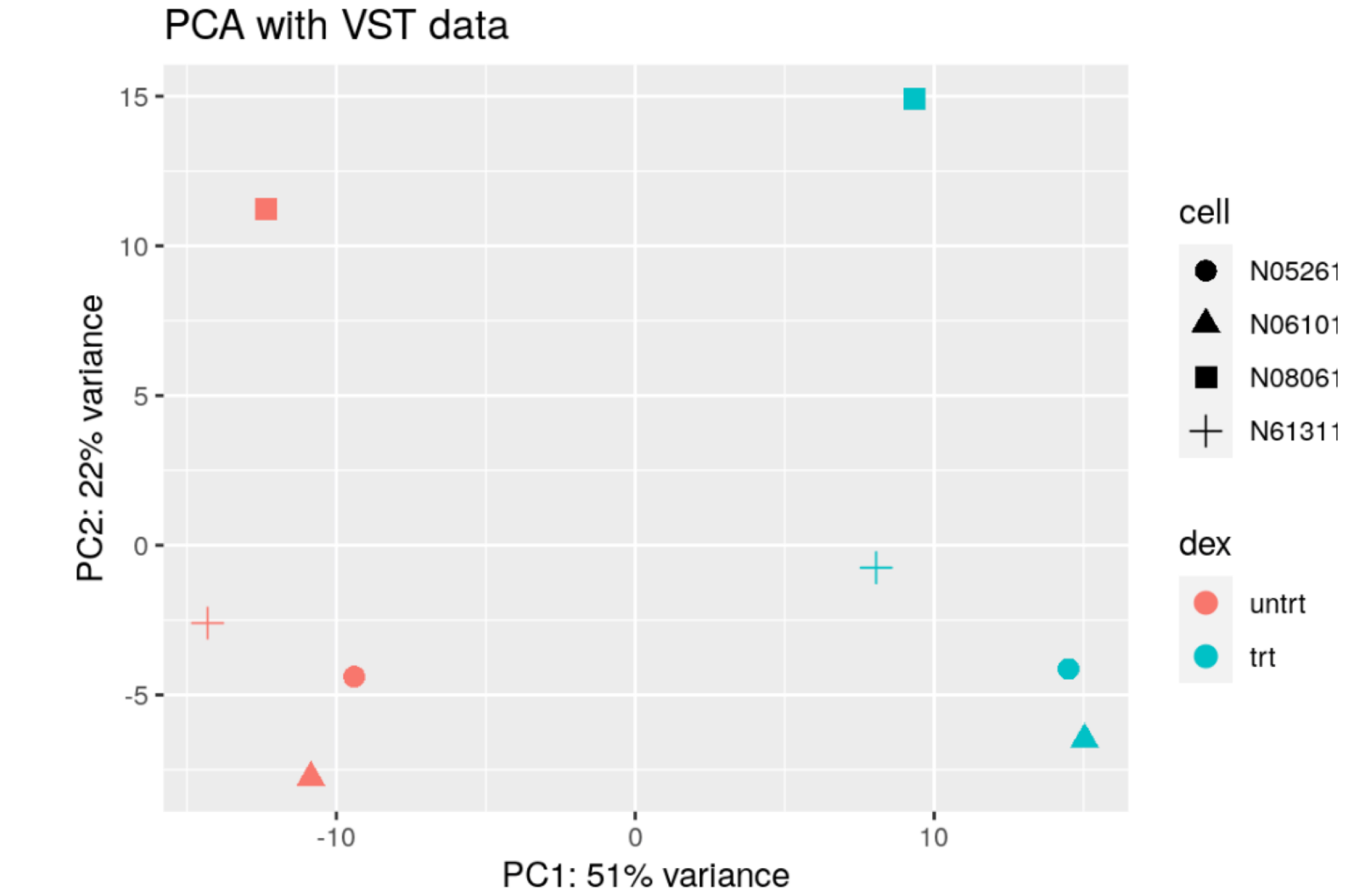


Transform the counts x with a function g(x) such that variability of values of x is not related to their mean.

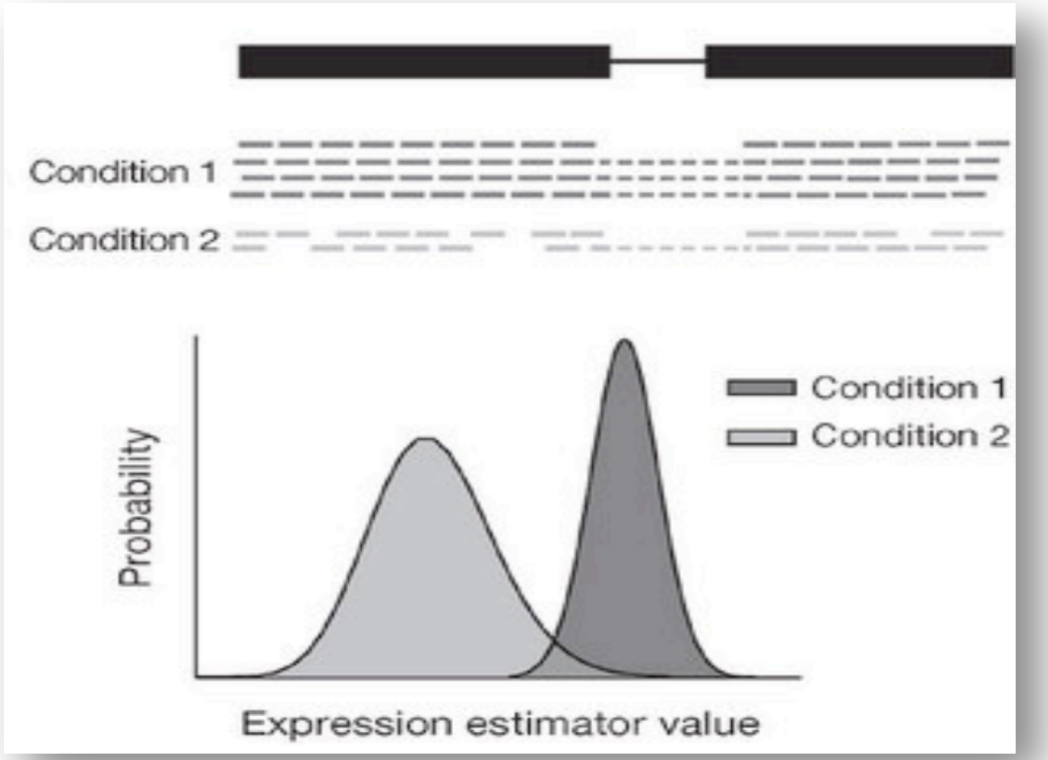
Data Exploration: Sample distances



Data Exploration: PCA



DE basics



1 test per gene!

1. Estimate **magnitude** of DE taking into account differences in sequencing depth, technical, and biological read count variability.

logFC

2. Estimate the **significance** of the difference accounting for performing thousands of tests.

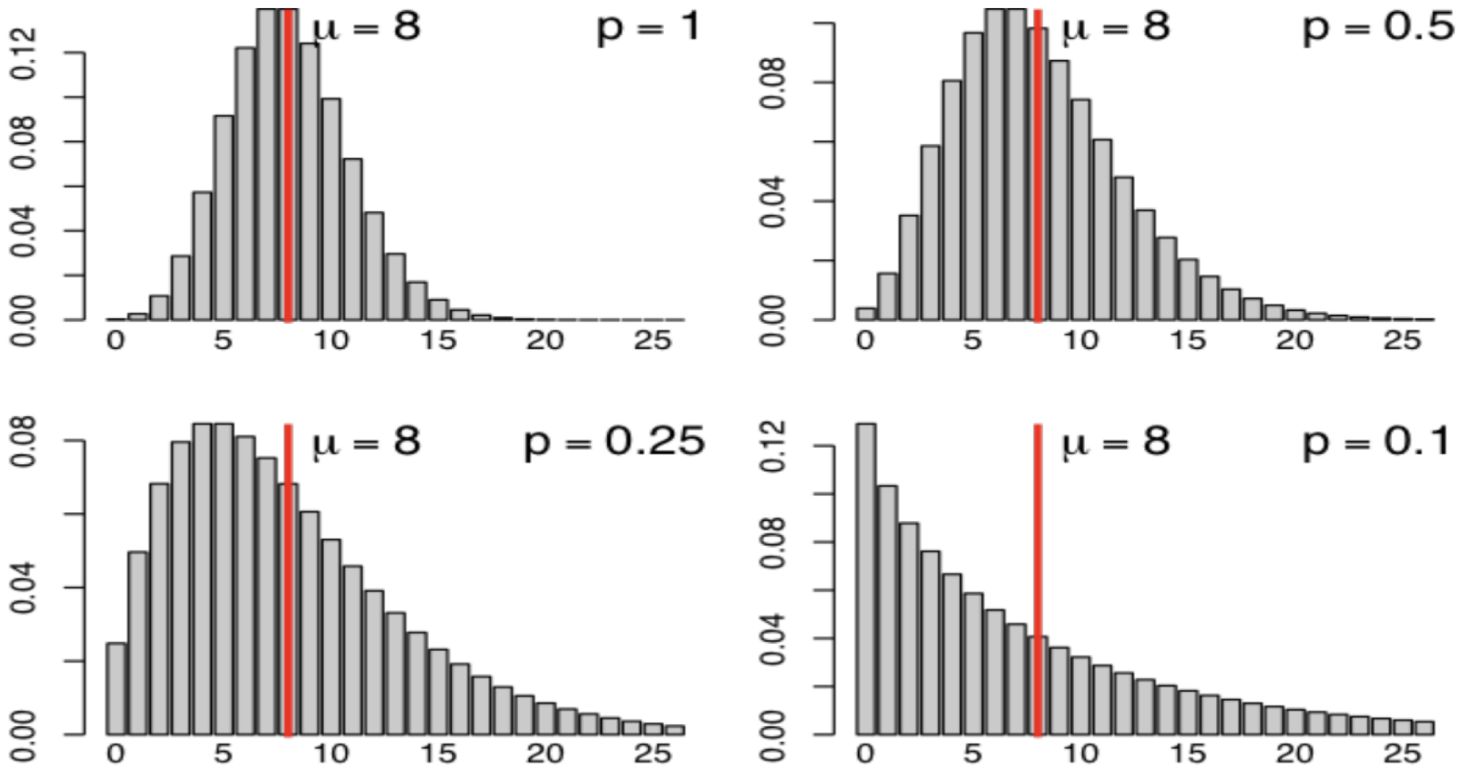
(adjusted) p-value

H0*: no difference in the read distribution between two conditions*

Garber et al. (2011) Nature Methods, 8(6), 469–477. doi:10.1038/nmeth.1613

Models for RNA-seq data: Negative Binomial distribution

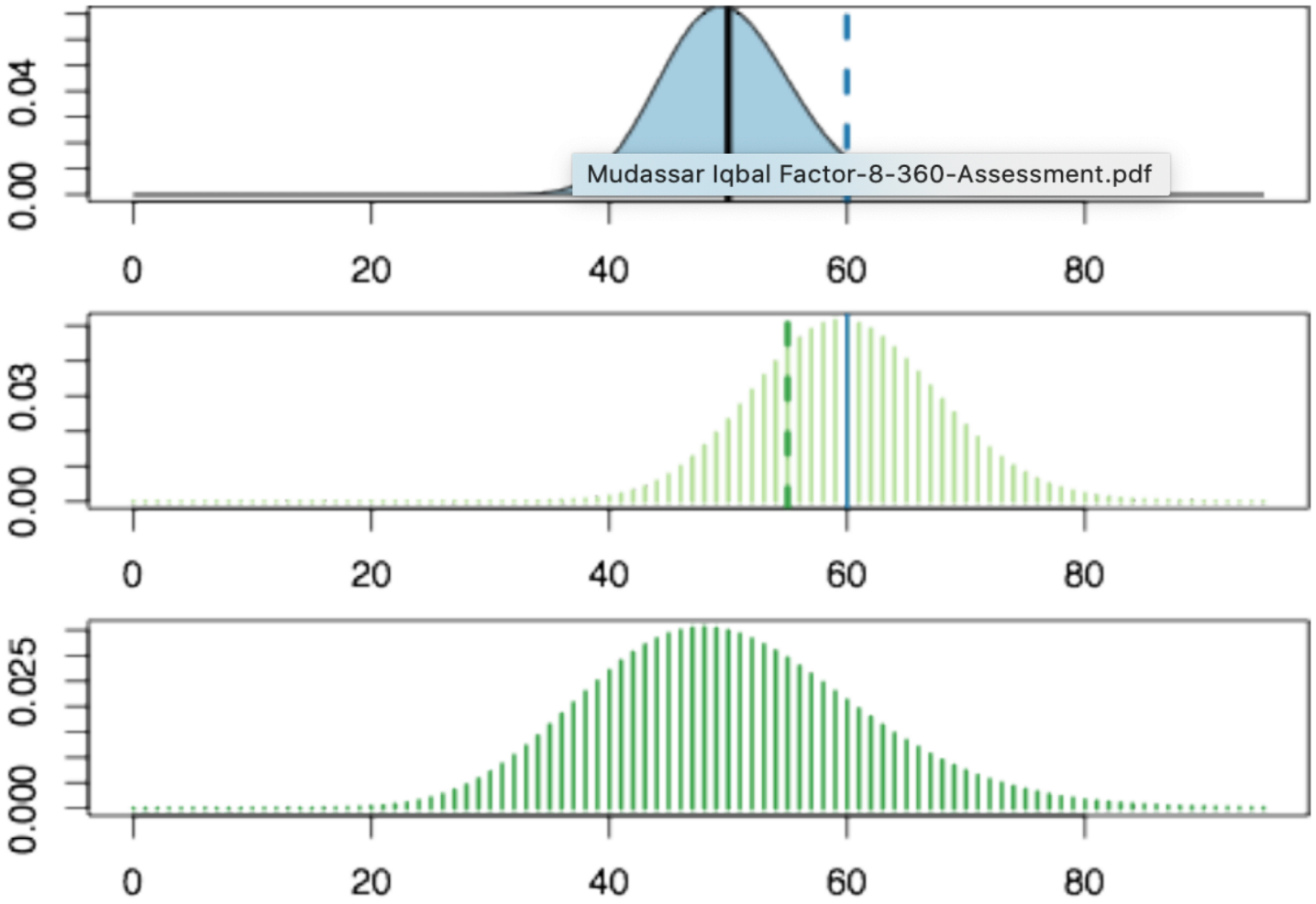
A commonly used generalization of the Poisson distribution with *two* parameters



$$\Pr(Y = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k \quad \text{for } k = 0, 1, 2, \ldots$$

http://www.bioconductor.org/help/course-materials/2015/CSAMA2015/lect/L05-deseq2-anders.pdf

The NB from a hierarchical model



Biological sample with mean μ and variance *v*

Poisson distribution with mean *q* and variance *q*.

Negative binomial with mean *μ* and variance *q*+*v*.

Fig4.22, https://www.huber.embl.de/msmb/Chap-Mixtures.html

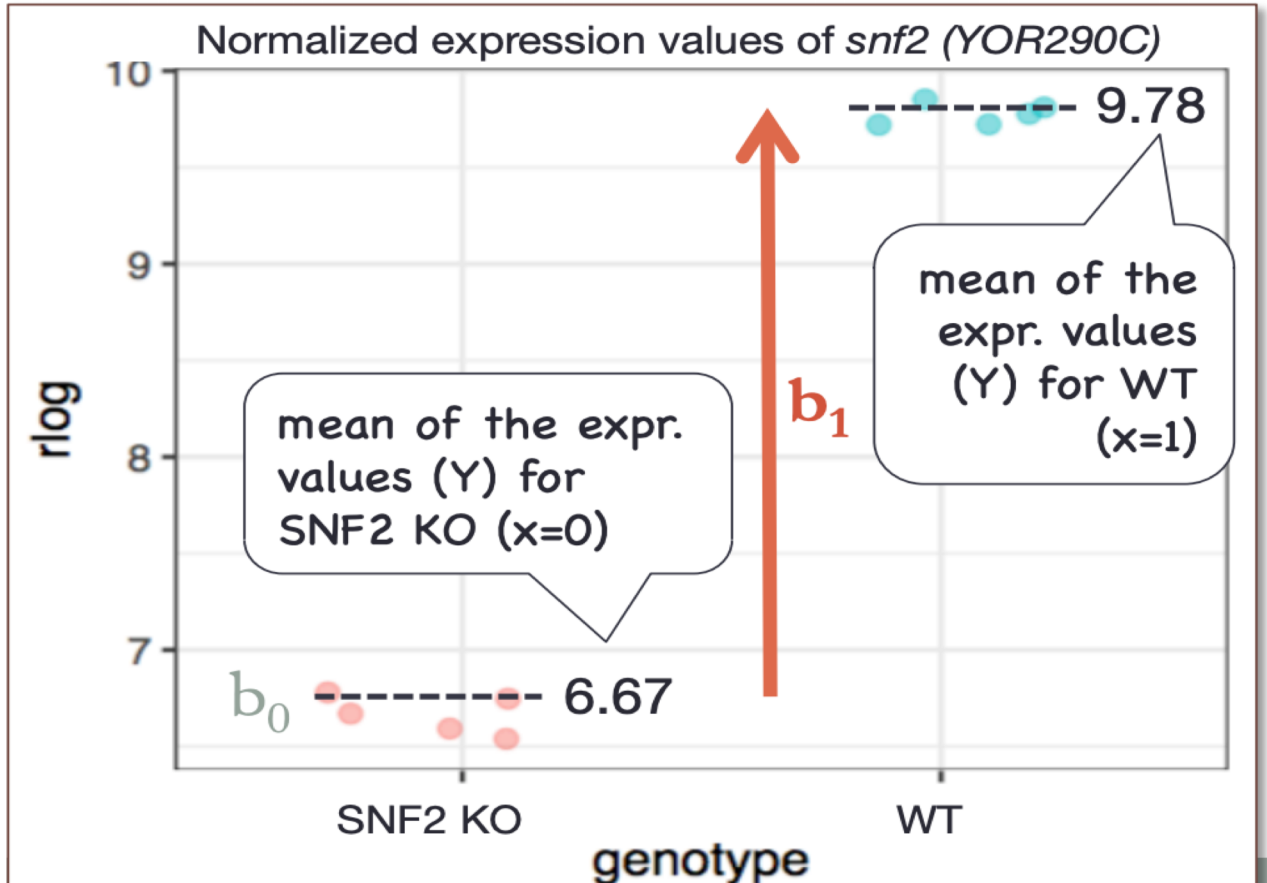
Estimating the difference with regression models

Example: Modeling normalized gene expression values using a linear model

describing all normalized expression values of one example gene using a simple linear model of the following form:

$$\underset{\text{expr. values}}{Y} = \underset{\text{intercept}}{b_0} + \underset{\text{delta}}{b_1} * \underset{\text{genotype}}{x} + e$$

b_0: **intercept**, i.e. average of the baseline group
b_1: **difference** between baseline & non-reference group
x : 0 if genotype == "SNF2", 1 if genotype == "WT"




```
# 1. FIT the model
> lmfit <- lm(rlog.norm ~ genotype)
# 2. ESTIMATE the coefficients
> coef(lmfit)
(Intercept)   genotypeWT
      6.666        3.111
```

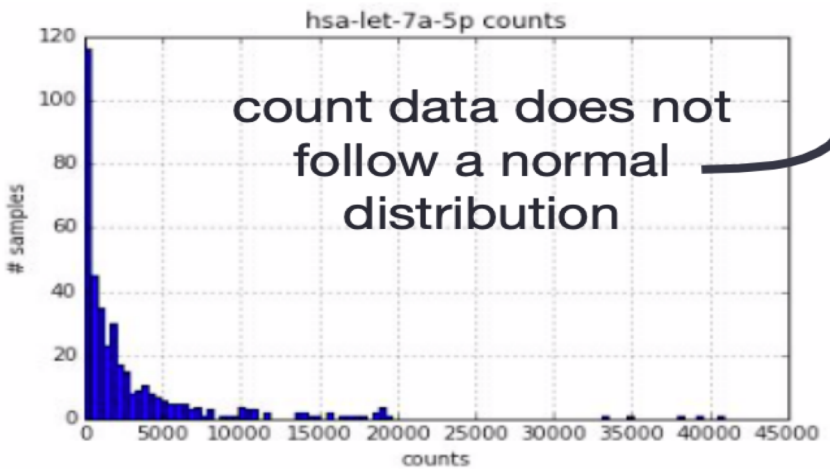
b_1

b_0 b_1

both beta values are **estimates**!
(they're spot-on because the data is so clear for this example and the model is so simple)

DE analysis: dealing with raw read counts

1. **Fitting** a sophisticated model (not a basic linear model) to get a grip on the read counts (done per gene; includes normalization)
   - library size factor
   - dispersion estimate using information across multiple genes
   - assuming a neg. binomial distribution of read counts



negative binomial (NB) model

gene-specific dispersion parameter (fitted towards the average dispersion)

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_i)$$

read counts for gene i and sample j

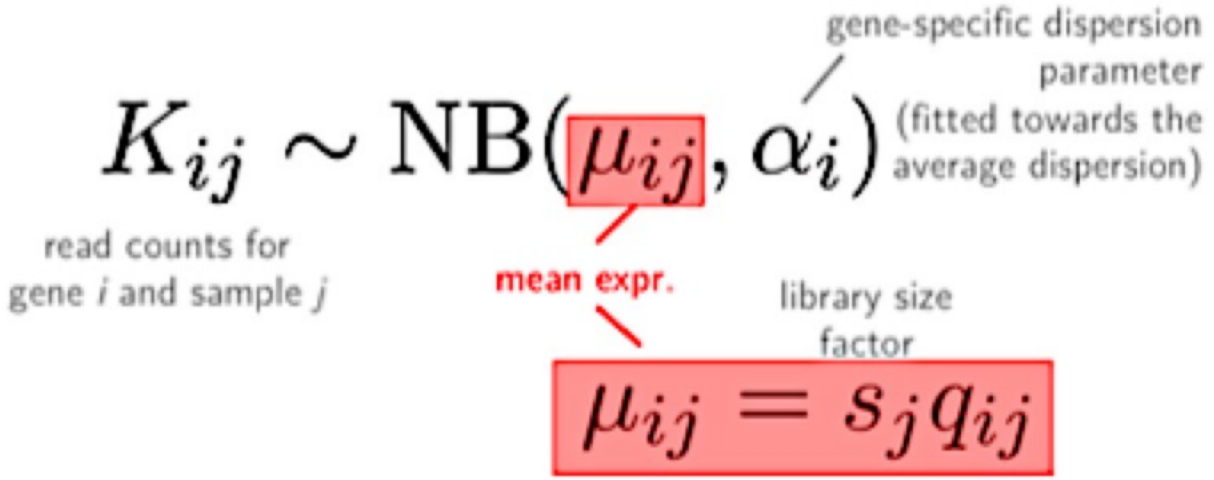
mean expr.

library size factor

$$\mu_{ij} = s_j q_{ij}$$

DE analysis

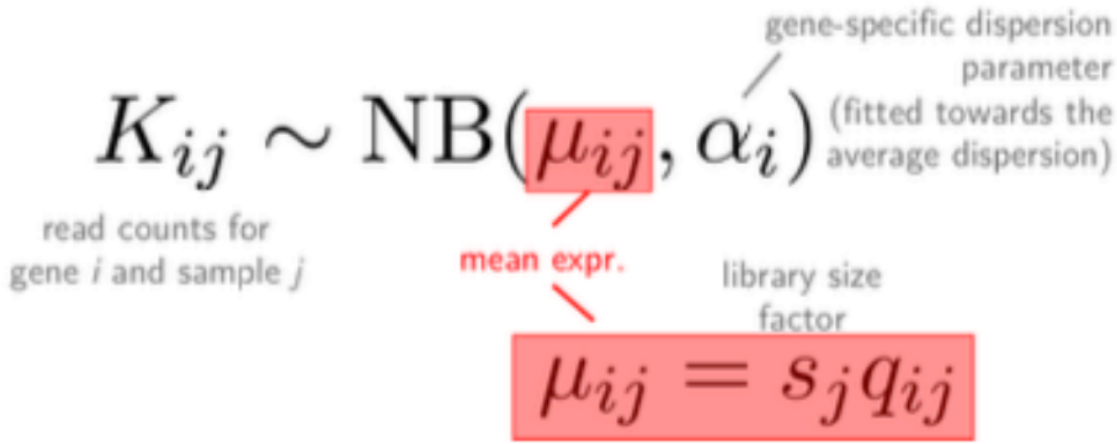
1. **Fitting** a sophisticated model to get a grip on the read counts (done per gene; includes normalization)



2. Estimating **coefficients** of the model to obtain the difference between the estimated mean expression of the different groups (log2FC)
   - define the **contrast of interest**, e.g. ~ batchEffect + condition
   - always put **the factor of interest last**
   - order of the factor levels determines the direction of log2FC

DE analysis

1. **Fitting** a sophisticated model to get a grip on the read counts (done per gene; includes normalization)



2. Estimating **coefficients** of the model to obtain the difference between the estimated mean expression of the different groups (log2FC)

3. **Test** whether the log2FC is “far away” from 0
   - log-likelihood test or Wald test are used by DESeq2
   - multiple hypothesis test correction

Modeling read counts and estimating the log2-fold-change (DESeq2)

fitted mean

gene-specific dispersion parameter (fitted towards the average dispersion)

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_i)$$

read counts for gene *i* and sample *j*

$$\mu_{ij} = s_j q_{ij}$$

library size factor

expression value estimate

$$\log_2(q_{ij}) = x_j . \beta_i$$

moderated log-fold change for gene *i*

model matrix column for sample *j*

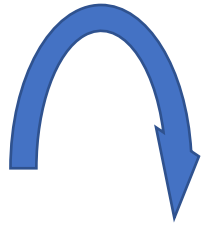
Once the coefficients are estimated, the significance tests need to test how far away from zero they are since zero would mean "no difference".

H0: *no difference in the read distribution between two conditions*

Let's do this!

From raw (read) counts to DE

```
##                    SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516
## ENSG00000000003.14        708        468        901        424       1188
## ENSG00000000005.5           0          0          0          0          0
## ENSG00000000419.12        455        510        604        352        583
##                    SRR1039517 SRR1039520 SRR1039521
## ENSG00000000003.14       1091        806        599
## ENSG00000000005.5           0          0          0
## ENSG00000000419.12        774        410        499
```



dds <- DESeq(dds)

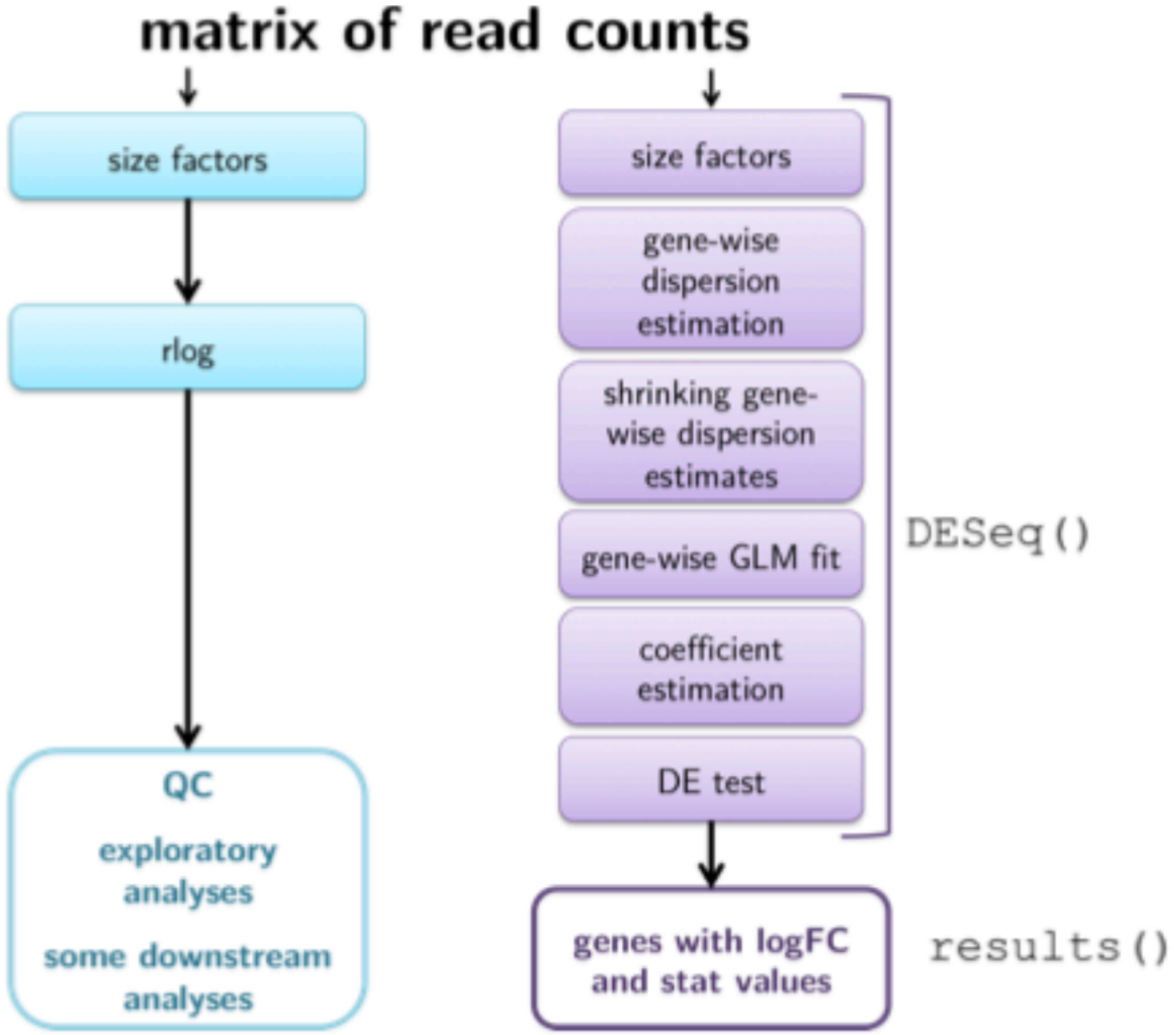
```
log2 fold change (MLE): dex trt vs untrt
Wald test p-value: dex trt vs untrt
DataFrame with 6 rows and 6 columns
                     baseMean    log2FoldChange             lfcSE                stat                pvalue
             padj
                    <numeric>         <numeric>         <numeric>         <numeric>             <numeric>
        <numeric>
ENSG00000216490.3 42.3007106248874 -5.72482604043858  1.47565155349927 -3.87952428665362 0.000104660951388532 0.0
00987852735620848
ENSG00000267339.5 30.5206063861615 -5.39781111590036 0.773017198758288 -6.98278269173179 2.89389866967638e-12  9.
4586266383349e-11
ENSG00000257542.5 10.0398869404764 -5.25991351125893  1.28200100388257 -4.10289344183752 4.08015195589968e-05 0.0
00430645894758231
ENSG00000146006.7 61.6448430808005 -4.49504194549369 0.663821044679586 -6.77146646904417 1.27483509915645e-11 3.8
2631678606496e-10
ENSG00000108700.4 14.6323552674909 -4.09068745839218 0.941842256424086 -4.34328299722226 1.40369125072207e-05  0.
00016668710298455
ENSG00000213240.8  12.096158074824 -3.87312575409988  1.27413298748988 -3.03981279201488  0.00236725244353312
0.0144986704569854
```

Exploratory vs. DE analysis workflow

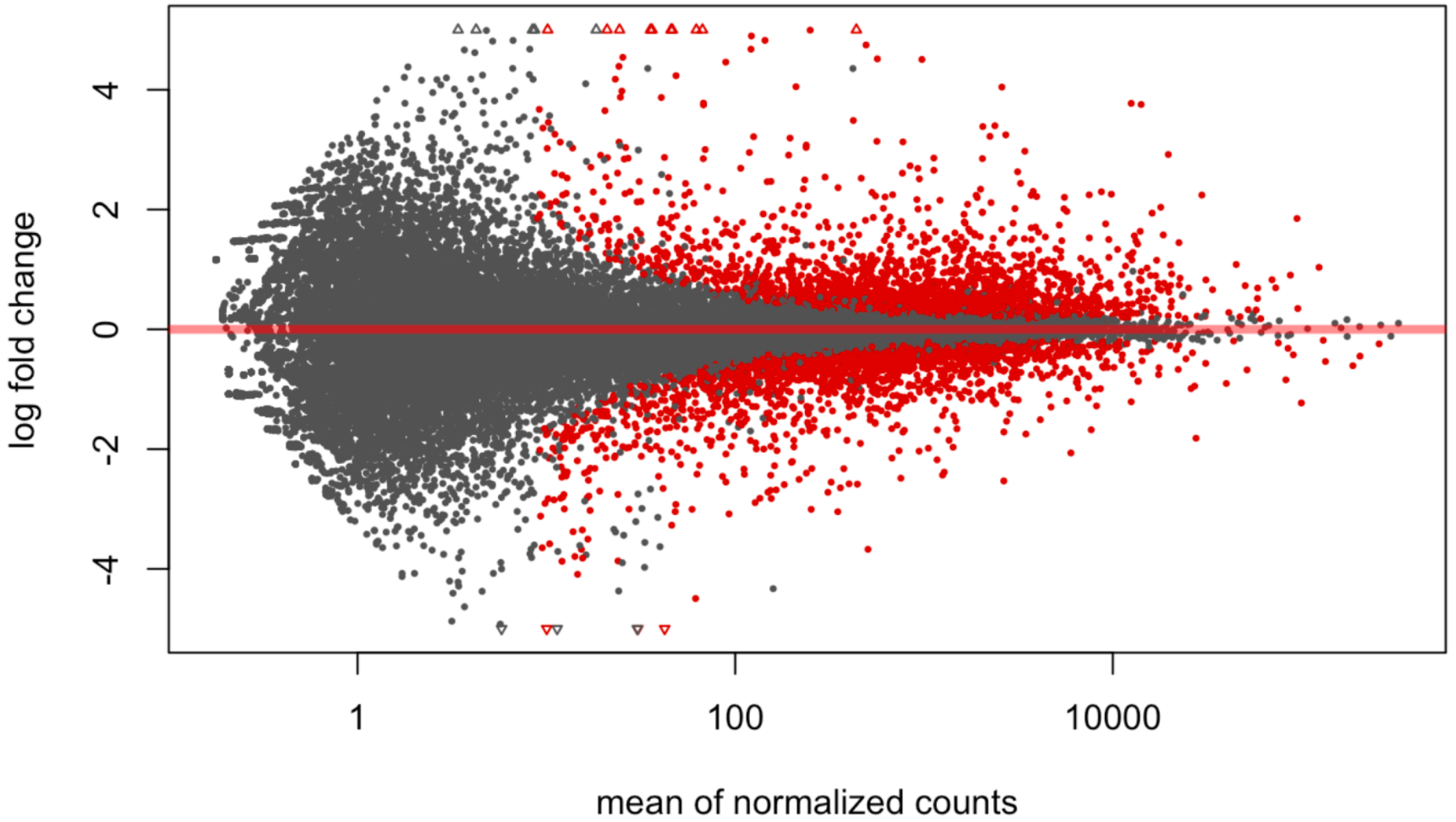
raw reads

NORMALIZATION

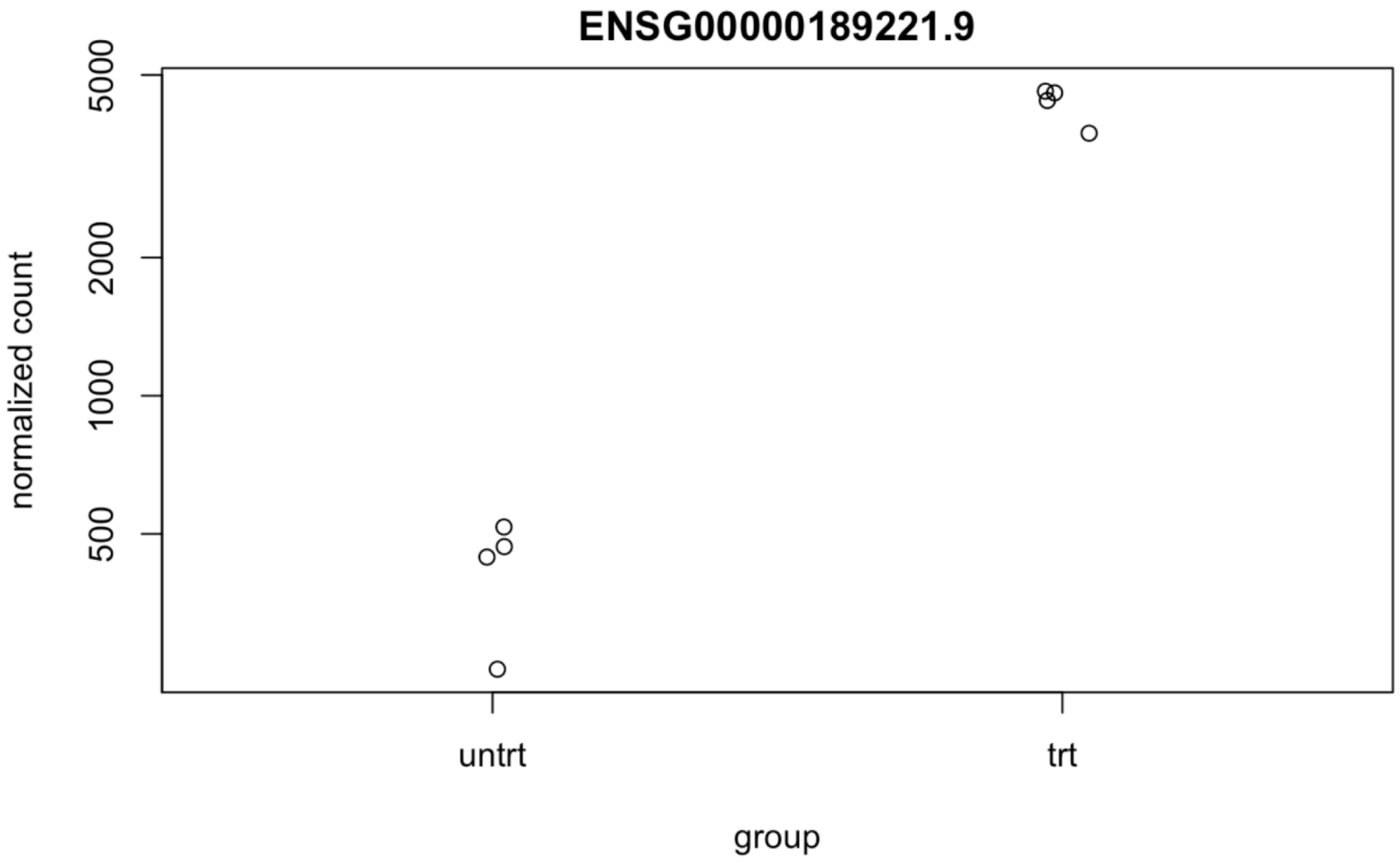
- lib sizes
- variance
- log-transformation

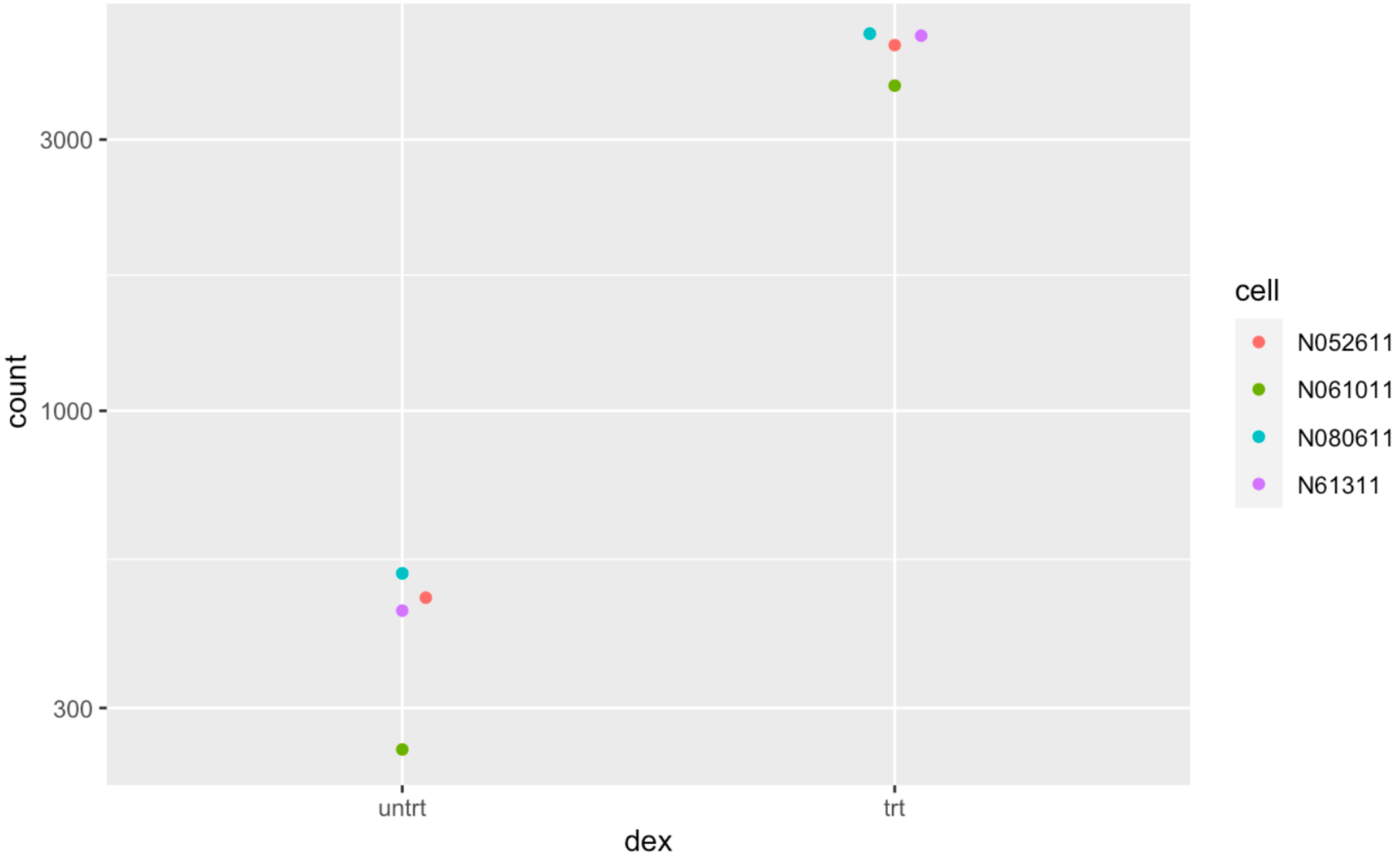


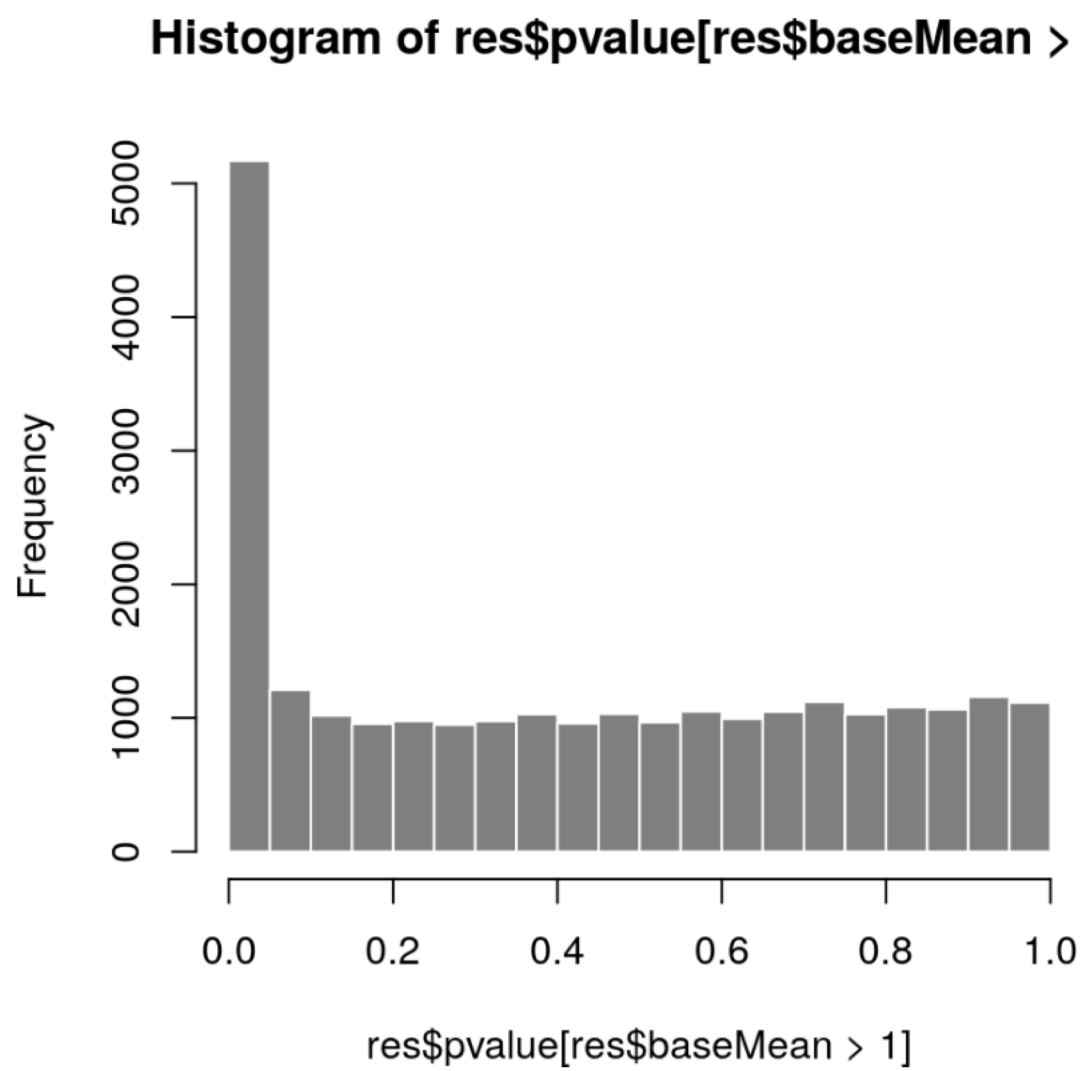
The MA plot



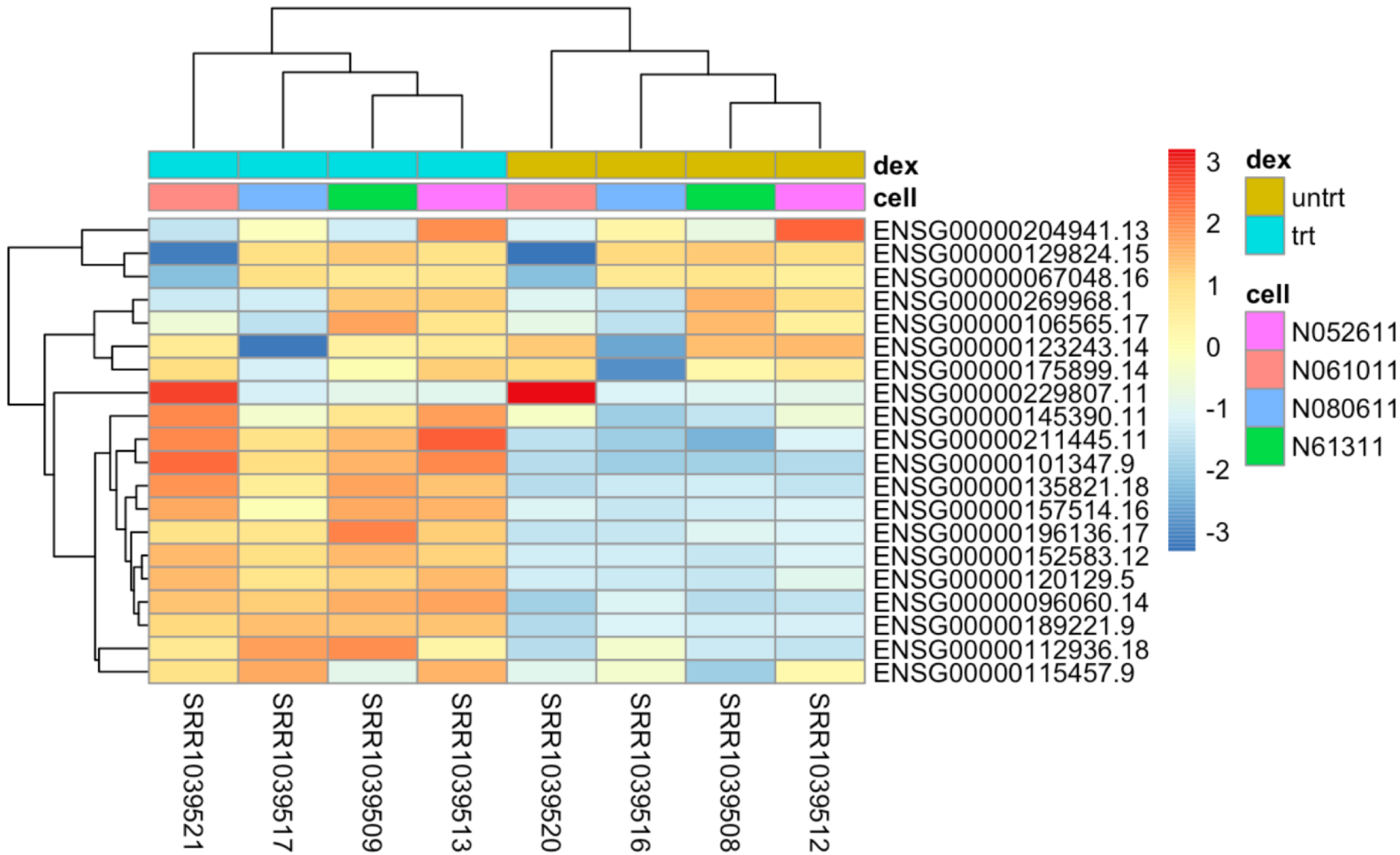
Plotting results – gene wise comparisons




3000
1000
300
count
untrt
trt
dex
cell
N052611
N061011
N080611
N61311

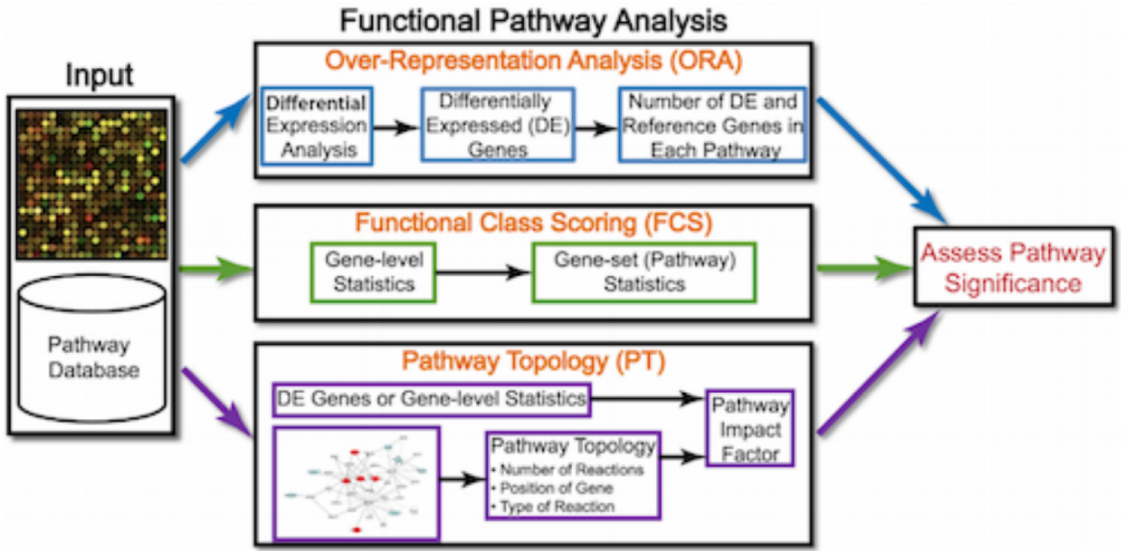

Histogram of res$pvalue[res$baseMean > 1]
Frequency
0
1000
2000
3000
4000
5000
0.0
0.2
0.4
0.6
0.8
1.0
res$pvalue[res$baseMean > 1]

Gene Clustering



and much more!

- Annotating, exporting results
- Integrating with other assays
- GO terms enrichments
- Network and Pathway analysis
- Experimental follow-up



https://hbctraining.github.io/GCC-BOSC-2018/slides/functional_analysis_mp.pdf

Resources used

- Love *et al.* (2015) RNA-Seq workflow: gene-level exploratory analysis and differential expression, F1000R 4 – 1070
- Modern Statistics for Modern Biology

  Susan Holmes, Wolfgang Huber

  https://www.huber.embl.de/msmb/
- https://chagall.med.cornell.edu/RNASEQcourse/
- https://www.bioconductor.org/help/course-materials/2014/CSAMA2014/2_Tuesday/lectures/DESeq2-Anders.pdf

Further reading (recommended)

Modern Statistics for Modern Biology

Susan Holmes, Wolfgang Huber

https://www.huber.embl.de/msmb/