From samples to counts' table with 4SeqGUI: output files

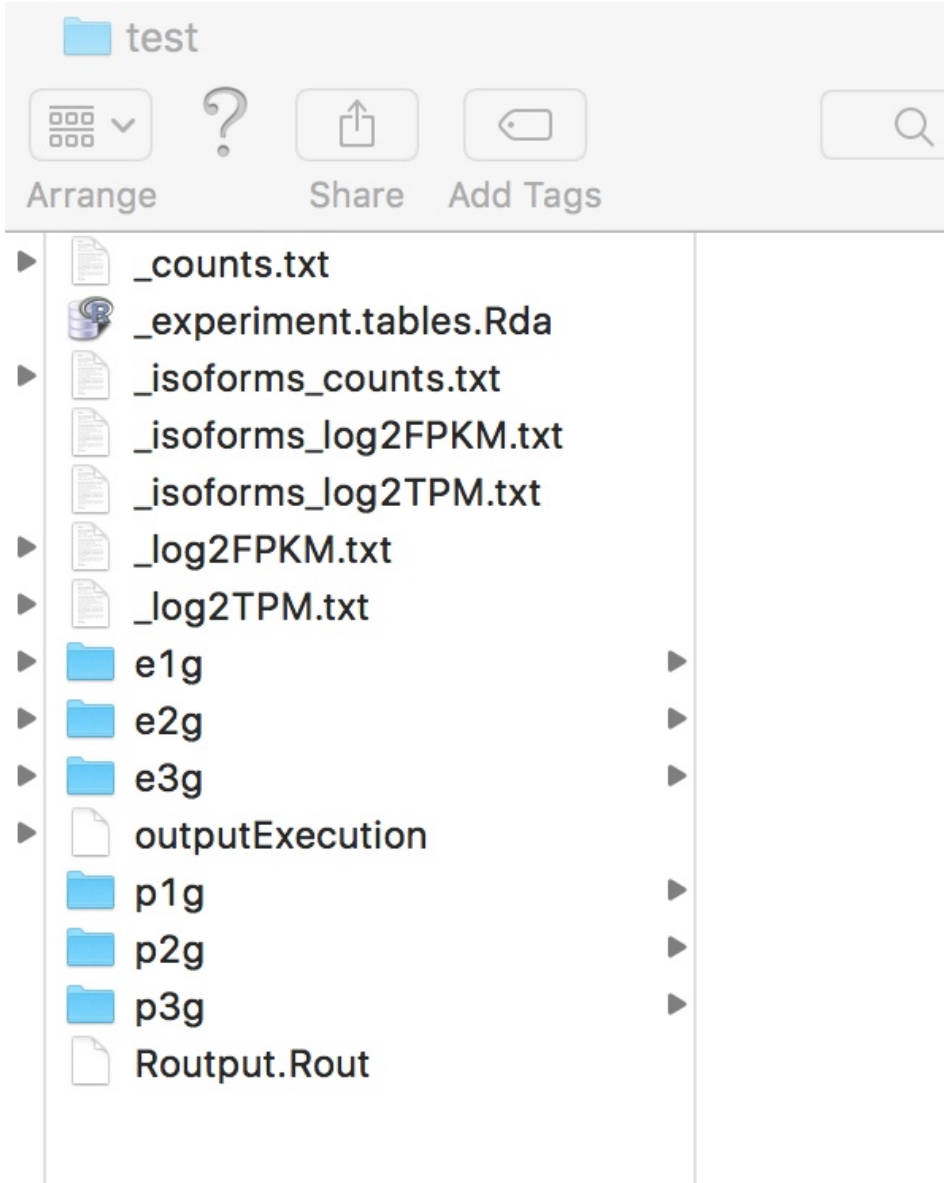


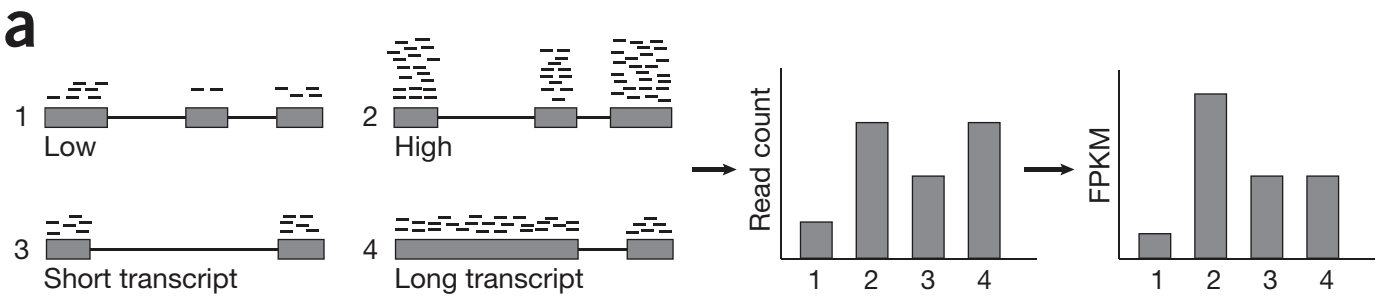
Figure 16: From samples to experiment tables

Differential Expressed Genes – Normalization

RNA-seq to estimate gene expression, read counts need to be properly normalized to extract meaningful expression estimates

There are two main sources of systematic variability that require normalization.

1. RNA fragmentation during library construction causes longer transcripts to generate more reads compared to shorter transcripts present at the same abundance in the sample.
2. The variability in the number of reads produced for each run causes fluctuations in the number of fragments mapped across samples



To account for these issues, the reads per kilobase of transcript per million mapped reads (RPKM) metric normalizes a transcript's read count by both its length and the total number of mapped reads in the sample When data originate from paired-end sequencing, the analogous fragments per kilo- base of transcript per million mapped reads (FPKM)

There's a new RNA-seq metric on the block…

- We used to report RPKM (Reads Per Kilobase Million) or FPKM (Fragments Per Kilobase Million)
  - These normalized read counts for:
    - 1) The sequencing depth (that's the "Million" part)
      - Sequencing runs with more depth will have more reads mapping to each gene.
    - 2) The length of the gene (that's the "Kilobase" part)
      - Longer genes will have more reads mapping to them.
- Now they want us to use TPM – Transcripts per million

RPKM Summary

BEFORE

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

Read counts were…

1) Normalized for differences in sequencing depth.
2) Normalized for gene size.

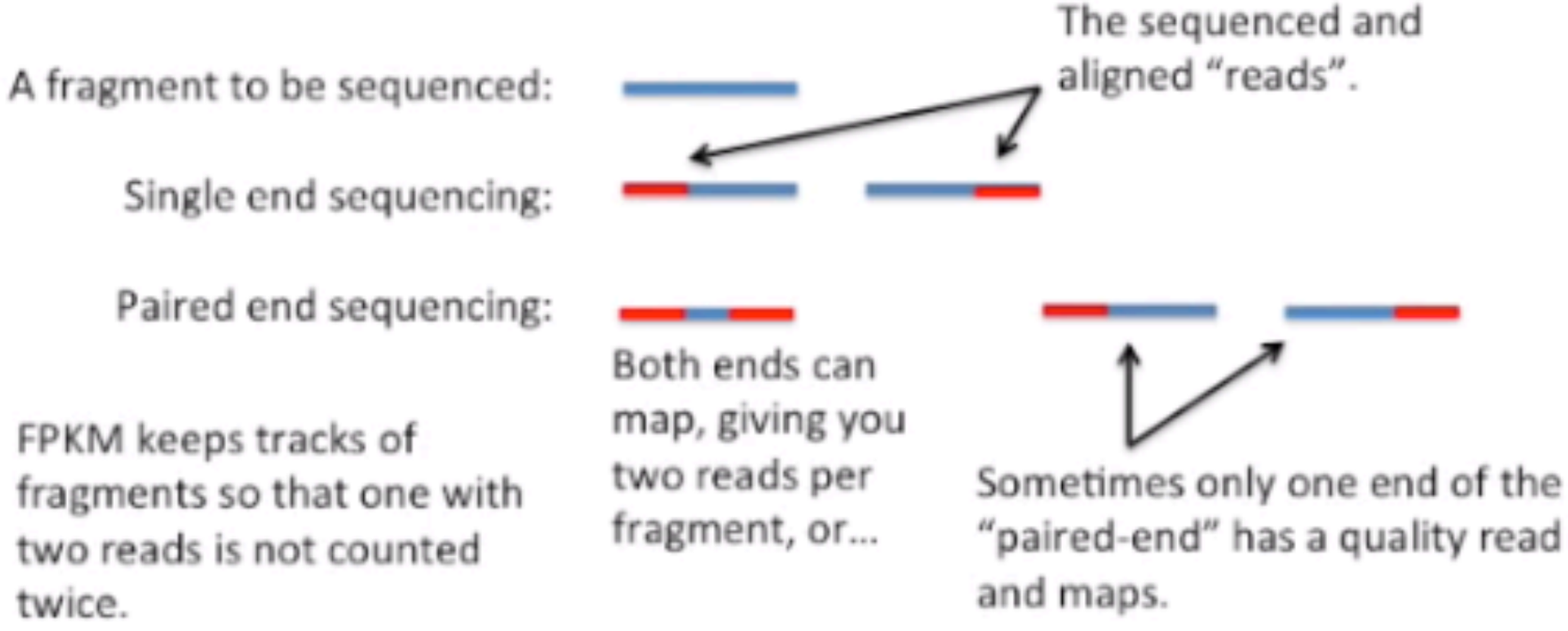
AFTER

Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009

RPKM and FPKM – two very closely related terms…

RPKM = Reads Per Kilobase Million
FPKM = Fragments per Kilobase Million

RPKM is for single end RNA-seq.
FPKM is very similar to RPKM, but for paired end RNA-seq.



FPKM keeps tracks of fragments so that one with two reads is not counted twice.

TPM Summary

BEFORE

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

Read counts were…

1) Normalized for gene size.

2) Normalized for differences in sequencing depth.

AFTER

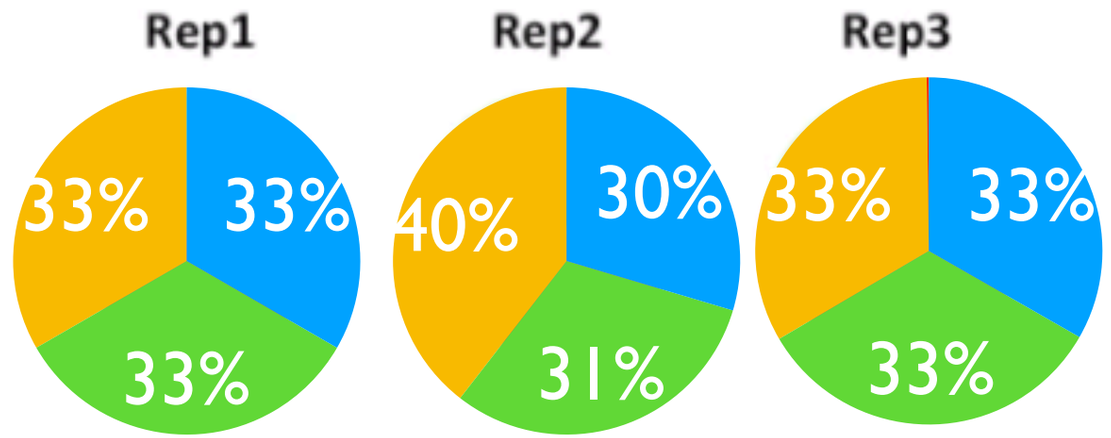
Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02

RPKM vs TPM

Consider 3 pies, each the same size (10).

A 3.33 sized slice is the same in each pie, and is always larger than 3.32.

TPM makes it clear that in Rep1, more of its total reads mapped to gene A than in Rep3.



TPM

Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02
Total:	10	10	10

RPKM vs TPM

RPKM

Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009
Total:	4.29	4.5	4.25

With RPKM, it is harder to compare the proportion of total reads because each replicate has different total (each pie has a different size)

A 1.43 size slice represents a different proportion of reads in in different pies.

Samples QC by PCA

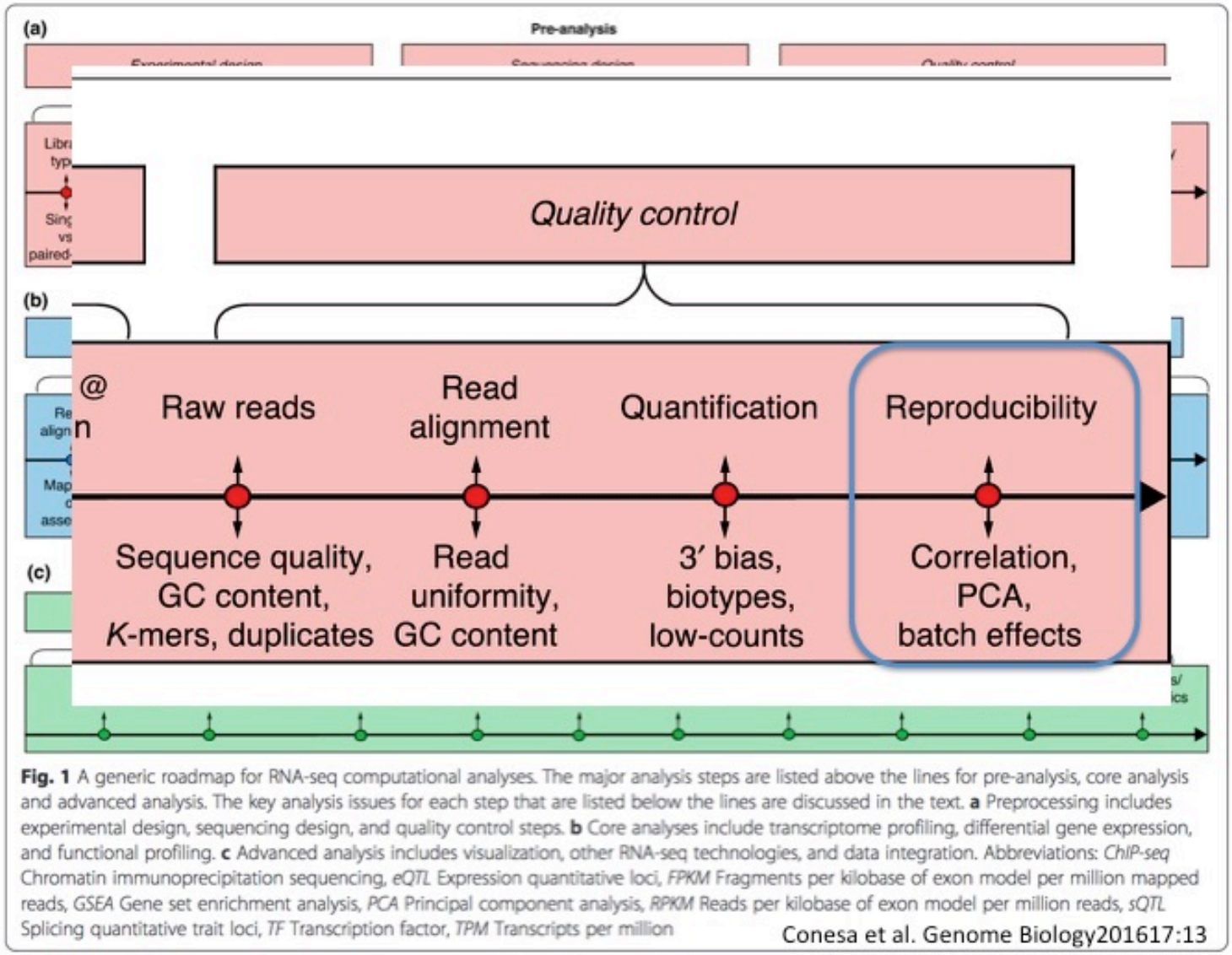


Fig. 1 A generic roadmap for RNA-seq computational analyses. The major analysis steps are listed above the lines for pre-analysis, core analysis and advanced analysis. The key analysis issues for each step that are listed below the lines are discussed in the text. **a** Preprocessing includes experimental design, sequencing design, and quality control steps. **b** Core analyses include transcriptome profiling, differential gene expression, and functional profiling. **c** Advanced analysis includes visualization, other RNA-seq technologies, and data integration. Abbreviations: *ChIP-seq* Chromatin immunoprecipitation sequencing, *eQTL* Expression quantitative loci, *FPKM* Fragments per kilobase of exon model per million mapped reads, *GSEA* Gene set enrichment analysis, *PCA* Principal component analysis, *RPKM* Reads per kilobase of exon model per million reads, *sQTL* Splicing quantitative trait loci, *TF* Transcription factor, *TPM* Transcripts per million

Conesa et al. Genome Biology201617:13

Figure 19: RNAseq workflow

PCA: Principal Component Analysis

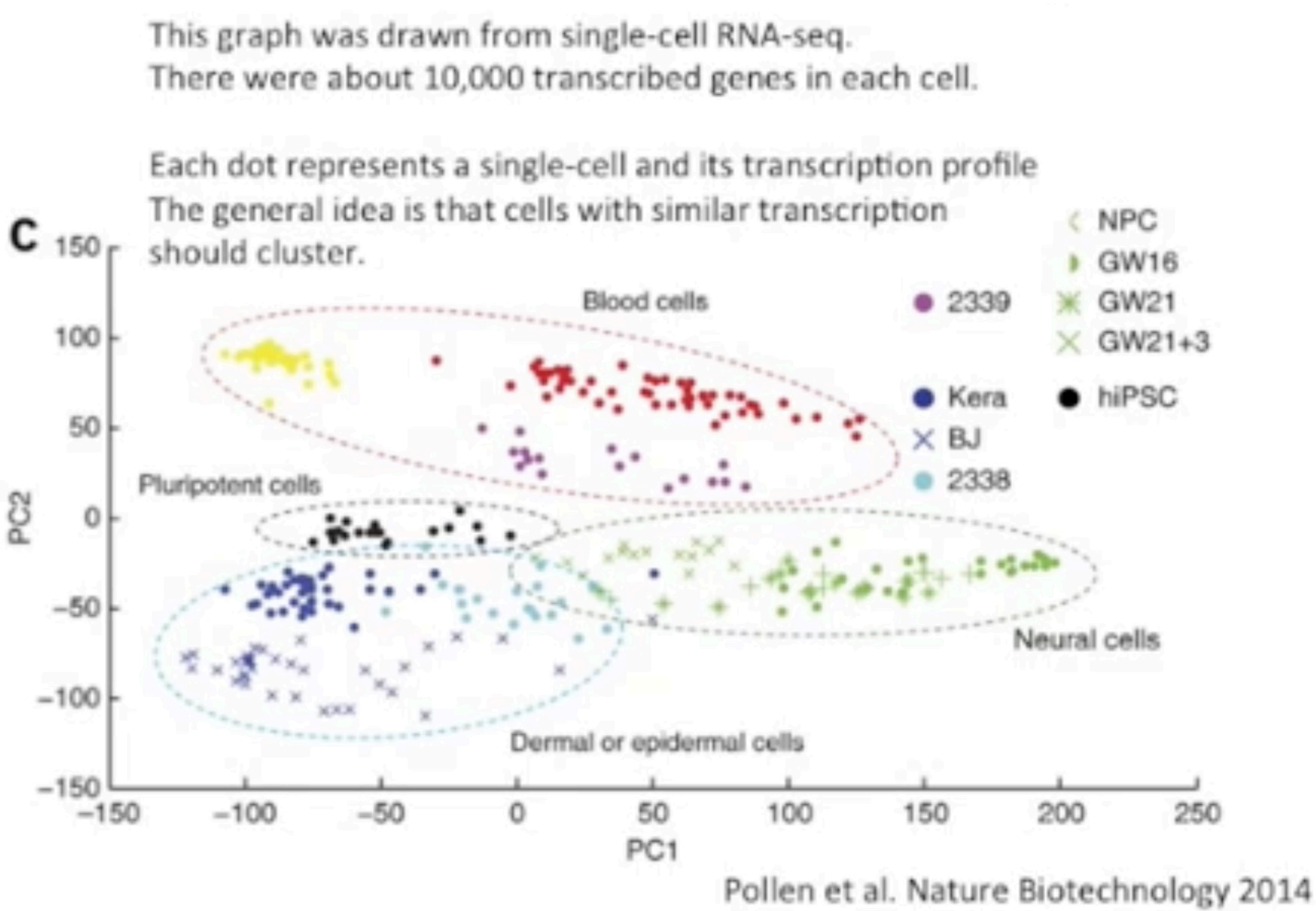


Figure 20:

Differential expression

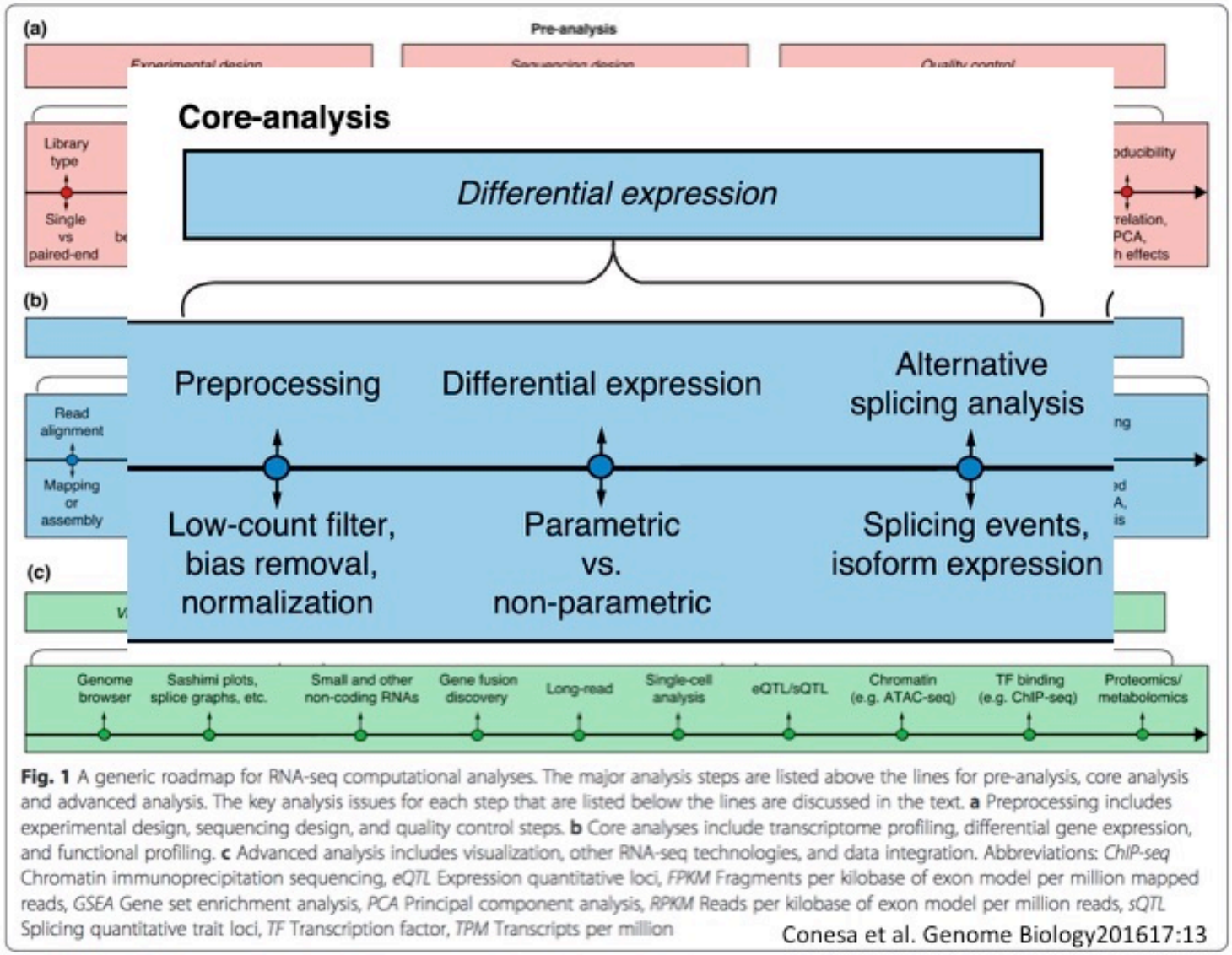


Fig. 1 A generic roadmap for RNA-seq computational analyses. The major analysis steps are listed above the lines for pre-analysis, core analysis and advanced analysis. The key analysis issues for each step that are listed below the lines are discussed in the text. **a** Preprocessing includes experimental design, sequencing design, and quality control steps. **b** Core analyses include transcriptome profiling, differential gene expression, and functional profiling. **c** Advanced analysis includes visualization, other RNA-seq technologies, and data integration. Abbreviations: *ChIP-seq* Chromatin immunoprecipitation sequencing, *eQTL* Expression quantitative loci, *FPKM* Fragments per kilobase of exon model per million mapped reads, *GSEA* Gene set enrichment analysis, *PCA* Principal component analysis, *RPKM* Reads per kilobase of exon model per million reads, *sQTL* Splicing quantitative trait loci, *TF* Transcription factor, *TPM* Transcripts per million

Conesa et al. Genome Biology201617:13

Figure 60: Differential expression

Differential Expressed Genes

Differential expression (DE) analysis refers to the identification of genes (or other types of genomic features, such as, transcripts or exons) that are expressed in significantly different quantities in distinct groups of samples, be it biological conditions (drug-treated vs. controls), diseased vs. healthy individuals, different tissues, different stages of development, or something else.

Although genes (if we focus on those for a while) are of course not expressed independent of each other, differential expression analysis is typically done on one gene at a time (although information is sometimes borrowed across genes, as we will see below) in a ***univariate* way.**

WHY? the number of *examples* is much smaller than the number of *features*, which makes it harder to fit a statistical model that considers all genes as a whole.

Multivariate dimension reduction methods such as principal component analysis (PCA) can be used to construct **low-dimensional representations** of the expression profiles that retain some of the properties of the complete data set and are thus often useful for visualization

Differential Expressed Genes – Replicates

The purpose of **replication** is to be able to estimate the variability between and among groups, which is important for, for example, hypothesis testing. Technical replication is used to estimate the variability of the measurement technique, for example, RNA-seq. **Biological replication is used to find out the variability within a biological group**. Roughly speaking, a change observed in gene expression between two groups can only be called significant if the difference between the groups is large compared to the variability within the group, while taking the sample size into consideration.

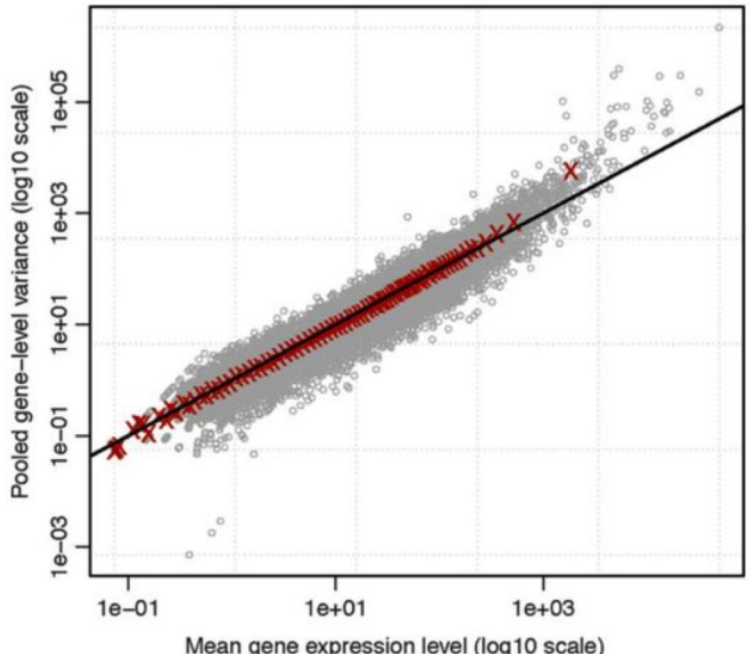
How many replicates should you use? This depends on the specifics of the experiment. The biological homogeneity of the different samples, the purpose of the experiment and the desired level of statistical power, among other things, will affect the number of replicates needed.

Many sequencing core facilities require or suggest using at least three or four replicates per group to be compared; two is almost always too few. With three, there is the risk that at least one sample will fail in library preparation or sequencing and you still end up with only two replicates in one of the groups.
Human blood and some tissue samples used for clinical case–control transcriptomics studies seem to exhibit considerable variation between individuals. Particularly for complex diseases, very large numbers of replicates (perhaps hundreds or thousands) may be needed to observe differential expression between cases and controls. For cell lines or samples from distinct tissues, only a few replicates may be needed.

Differential Expressed Genes – Statistical Distribution

For RNA-seq experiments, where one might assume that sequences are sampled at random from the sequencing library, the raw read counts would be expected to be **Poisson-distributed**.

You would expect to get slightly different counts even for the same library in an idealized scenario where it was sequenced twice under the same conditions. This inevitable noise which arises from the sampling process is called ***shot noise***, and often the variability between technical replicates in RNA-seq can be described quite well by this type of Poisson noise



Mean–variance plot for Marioni et al. dataset (Marioni et al. 2008). The variability in technically replicated RNA-seq data can be adequately captured using a Poisson model. The grey points in this plot shows the mean and pooled variance for each gene, scaled to account for differences in library size between samples. The black line displays the theoretical variance under the Poisson model where the variance is equal to the mean. The red crosses show binned variance, where genes are grouped by mean level.

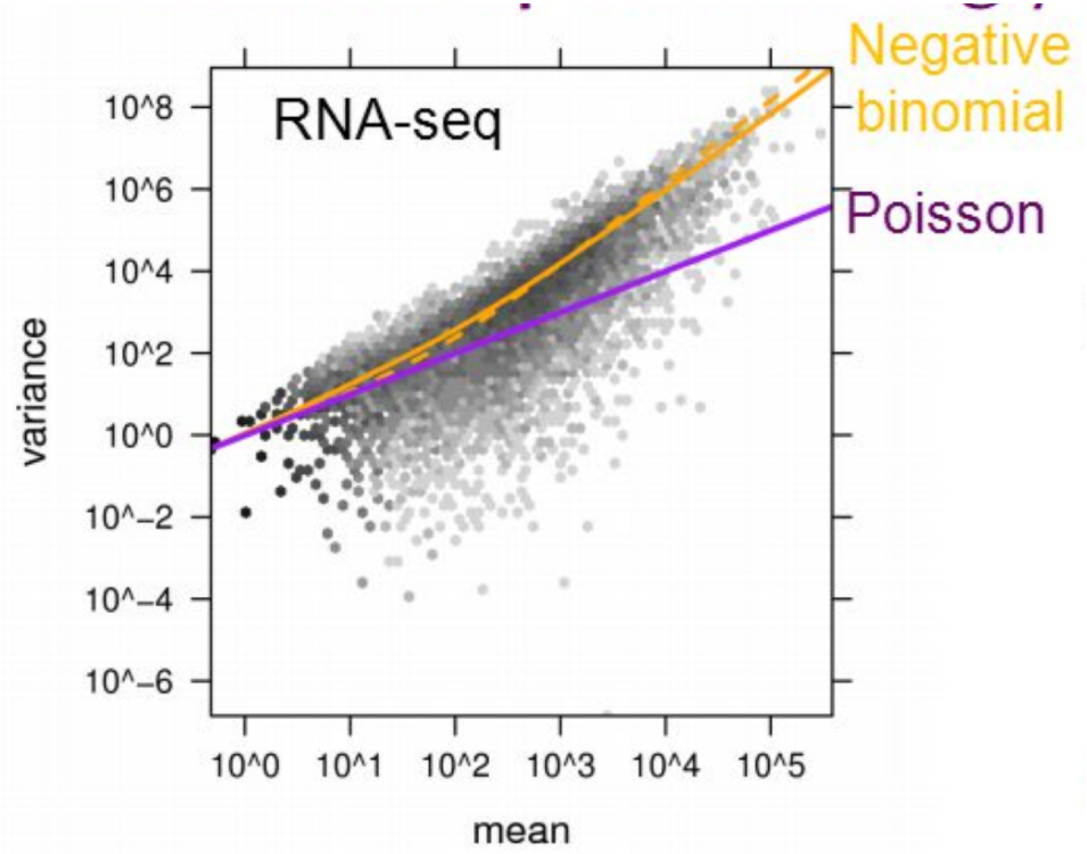
Differential Expressed Genes – Noise

We distinguish:

- Shot noise
  - unavoidable, appears even with perfect replication
  - dominant noise for weakly expressed genes
- Technical noise
  - from sample preparation and sequencing
  - negligible (if all goes well)
- Biological noise
  - unaccounted-for differenced between samples
  - Dominant noise for strongly expressed genes

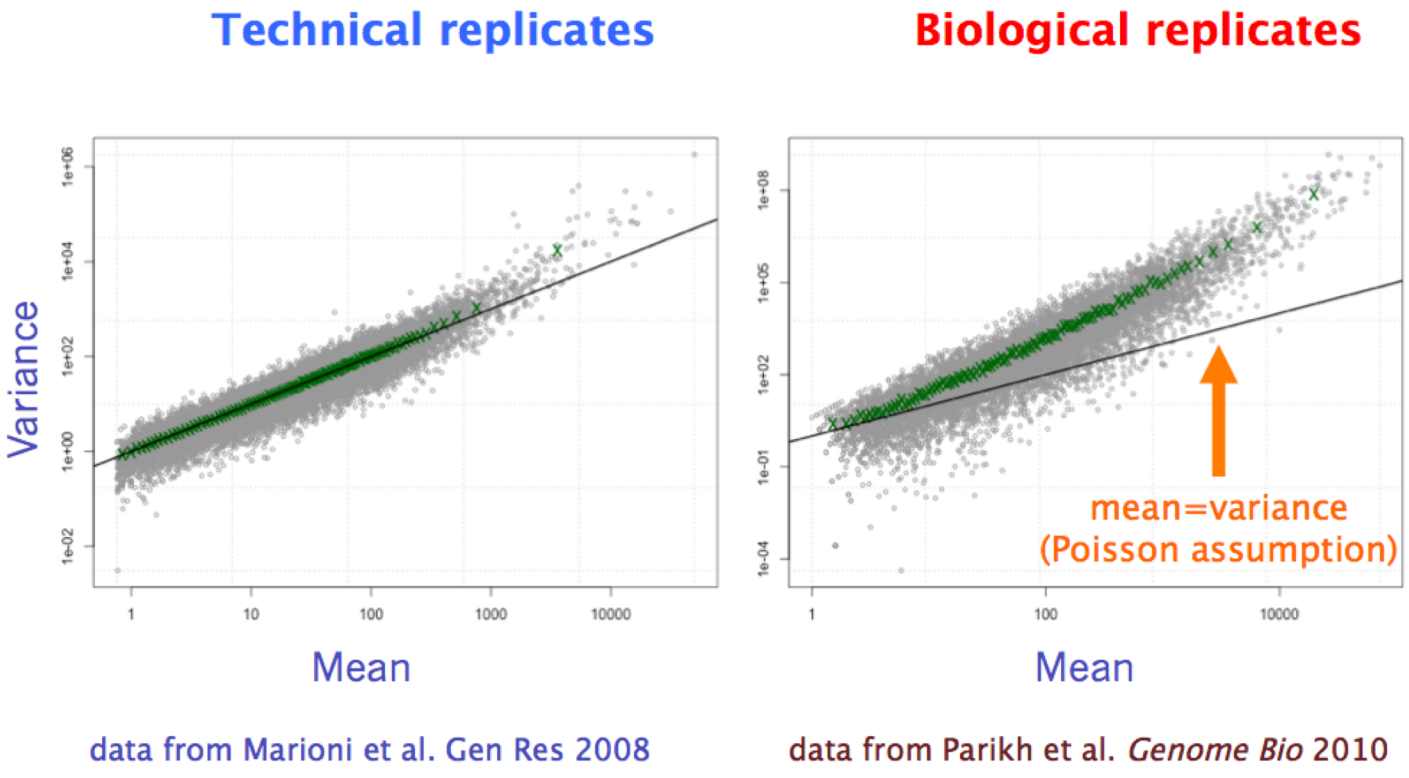
Differential Expressed Genes – Statistical Distribution

When samples are taken from biologically distinct sources, such as different individuals, the variability between them has often been modeled by a ***negative binomial* distribution** (sometimes called gamma-Poisson distribution). This distribution can be described as an *overdispersed* Poisson distribution



In RNAseq genes with high mean counts, because they are long or highly expressed, tend to show more variance between sample than genes with low mean counts. Thus this data fits a Negative Binomial Distribution.

Differential Expressed Genes – Statistical Distribution



Counts for the same gene from different **technical replicates** have variance equal to the mean (Poisson)

Counts for the same gene from different **biological replicates** have a variance exceeding the mean (overdispersion)

Differential Expressed Genes – Normalization, DESeq2

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**.

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.

Differential Expressed Genes – Generalized linear models

Two sample groups, treatment and control.

Assumption:

Count value for a gene in sample j is generated by Negative Binomial distribution with mean μj and dispersion α.

Null hypothesis:

All samples have the same μj .

Alternative hypothesis:

Mean is the same only within groups:

$$\log \mu_j = \beta_C + x_j \beta_T$$

where $x_j = 0$ if j is control sample

$x_j = 1$ if j is treatment sample

Linear Models

– The observed value of Y is a linear combination of the effects of the independent variables

Arbitrary number of independent variables

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Polynomials are valid

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \ldots + \beta_p X_1^p$$

$$E(Y) = \beta_0 + \beta_1 \log(X_1) + \beta_2 f(X_2) + \ldots + \beta_k X_k$$

We can use functions of the variables if the effects are linear

Smooth functions: not exactly the same as the so-called ***additive models***

– If we include categorical variables the model is called **General Linear Model**

In DeSeq2

RNA-seq raw count data follows a negative binomial distribution, as reported in the previous slide.

The DESeq2 authors model the data i.e. imply that for each gene is built a regression model of the data such that it is possibile to make statistical inferences from the data.

The normalised counts, are used to compute a logistic regression model fro each gene **with the negative binomial distribution.**

Once modelled each gene, the way to derive a P value for each model coefficient is by the Wald Test.

In DeSeq2

RNA-seq raw count data follows a negative binomial distribution, as reported in the previous slide.

The DESeq2 authors model the data i.e. imply that for each gene is built a regression model of the data such that it is possibile to make statistical inferences from the data.

The normalised counts, are used to compute a logistic regression model fro each gene **with the negative binomial distribution.**

Once modelled each gene, the way to derive a P value for each model coefficient is by the Wald Test.

The likelihood ratio (LTR) test

We are working with models, therefore we would like to do hypothesis tests on coefficients or contrasts of those models:

- We fit two models M1 without the coefficient to test and M2 with the coefficient.

- We compute the likelihoods of the two models (L1 and L2) and obtain LRT=-2log(L1 /L2) that has a known distribution under the null hypothesis that the two models are equivalent. This is also known as model selection

```
ddsLRT = DESeq(dds, test="LRT", full=~sex+age+smoke+disease, reduced=~sex+age+smoke)
```

The LRT It tests whether the increase in the log likelihood from the additional coefficients would be expected if those coefficients were equal to zero. It doesn't mean the reduced model is a good model or a good fit.

The adjusted p-value computed stay for: if it is small, then for the set of genes with those small adjusted p-values, the additional coefficient in full and not in reduced increased the log likelihood more than would be expected if their true value was zero.

```
ddsLRT = DESeq(dds, test="LRT", full=~sex+age+smoke+geneA+disease, reduced=~sex+age+smoke+disease)
```

Differential Expressed Genes – FDR

```
## log2 fold change (MAP): 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>
## ENSG00000179593  67.24305       4.880507 0.3308119   14.75312
2.937594e-49  9.418996e-47
## ENSG00000109906 385.07103       4.860877 0.3321627   14.63403
1.704000e-48  5.181040e-46
## ENSG00000152583 997.43977       4.315374 0.1723805   25.03400
2.608143e-138 4.599460e-134
## ENSG00000250978  56.31819       4.090157 0.3288246   12.43872
1.610666e-35  2.679631e-33
## ENSG00000163884 561.10717       4.078073 0.2103212   19.38974
9.421379e-84  1.038413e-80
## ENSG00000168309 159.52692       3.991146 0.2547755   15.66534
2.610147e-55  1.180255e-52
```

Why we need a p-value?

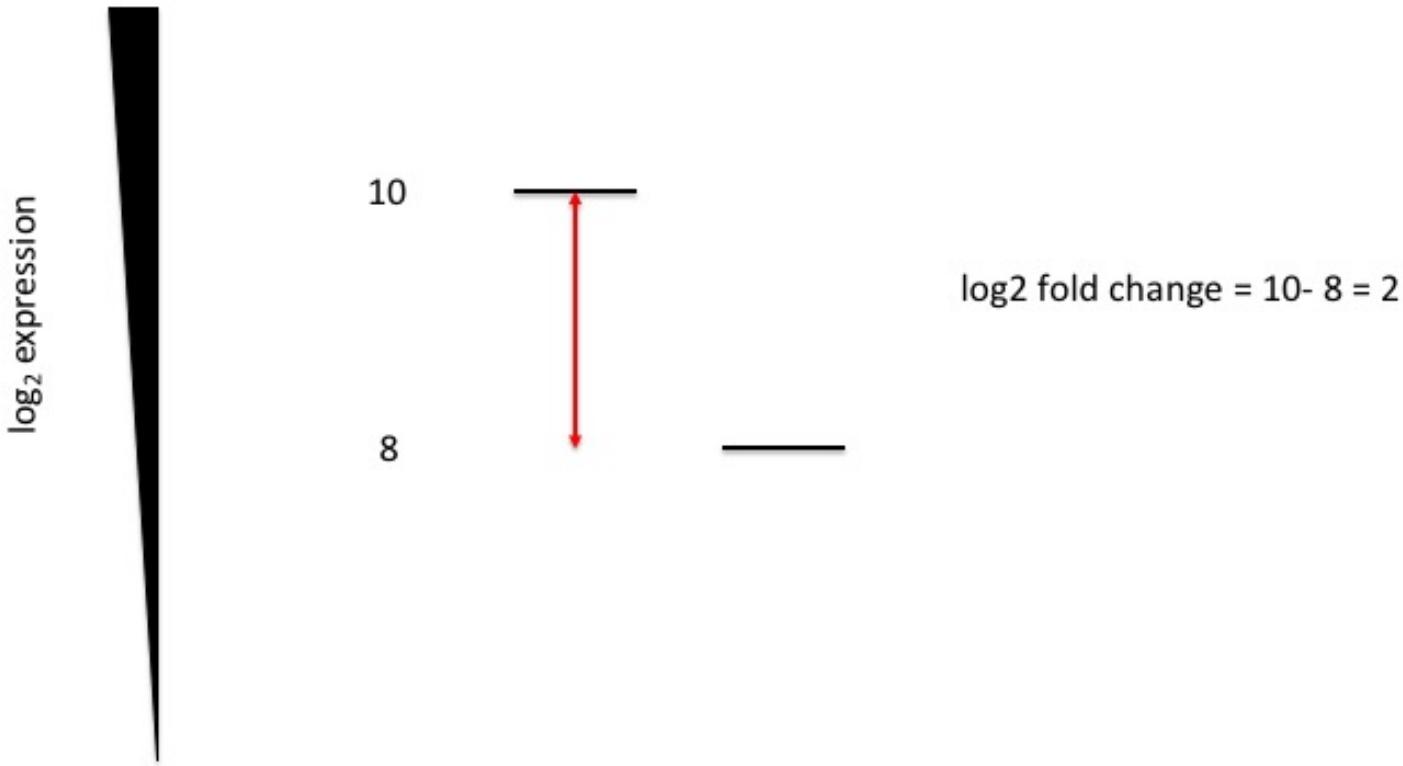


Figure 62: log2FC

Why we need a p-value?

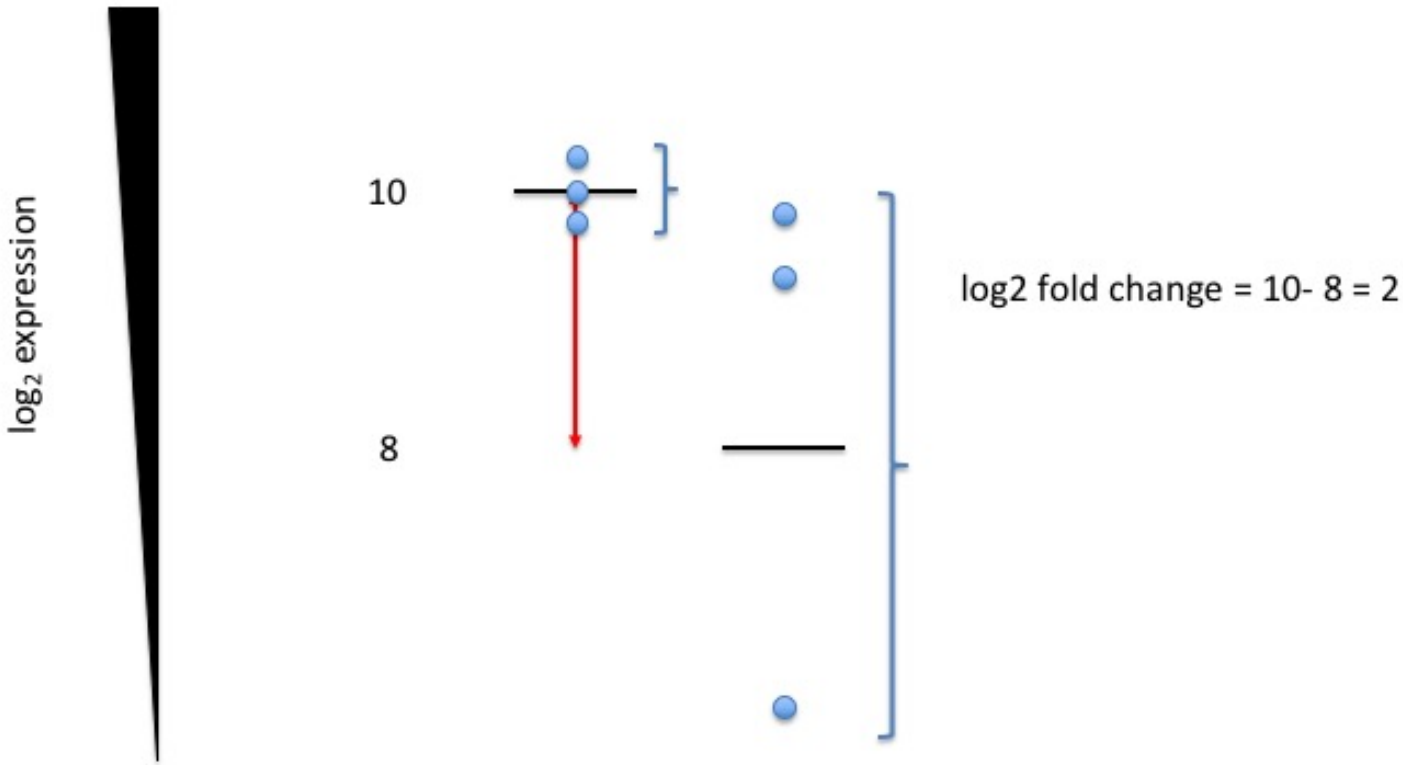


Figure 63: log2FC

Why we need a p-value?

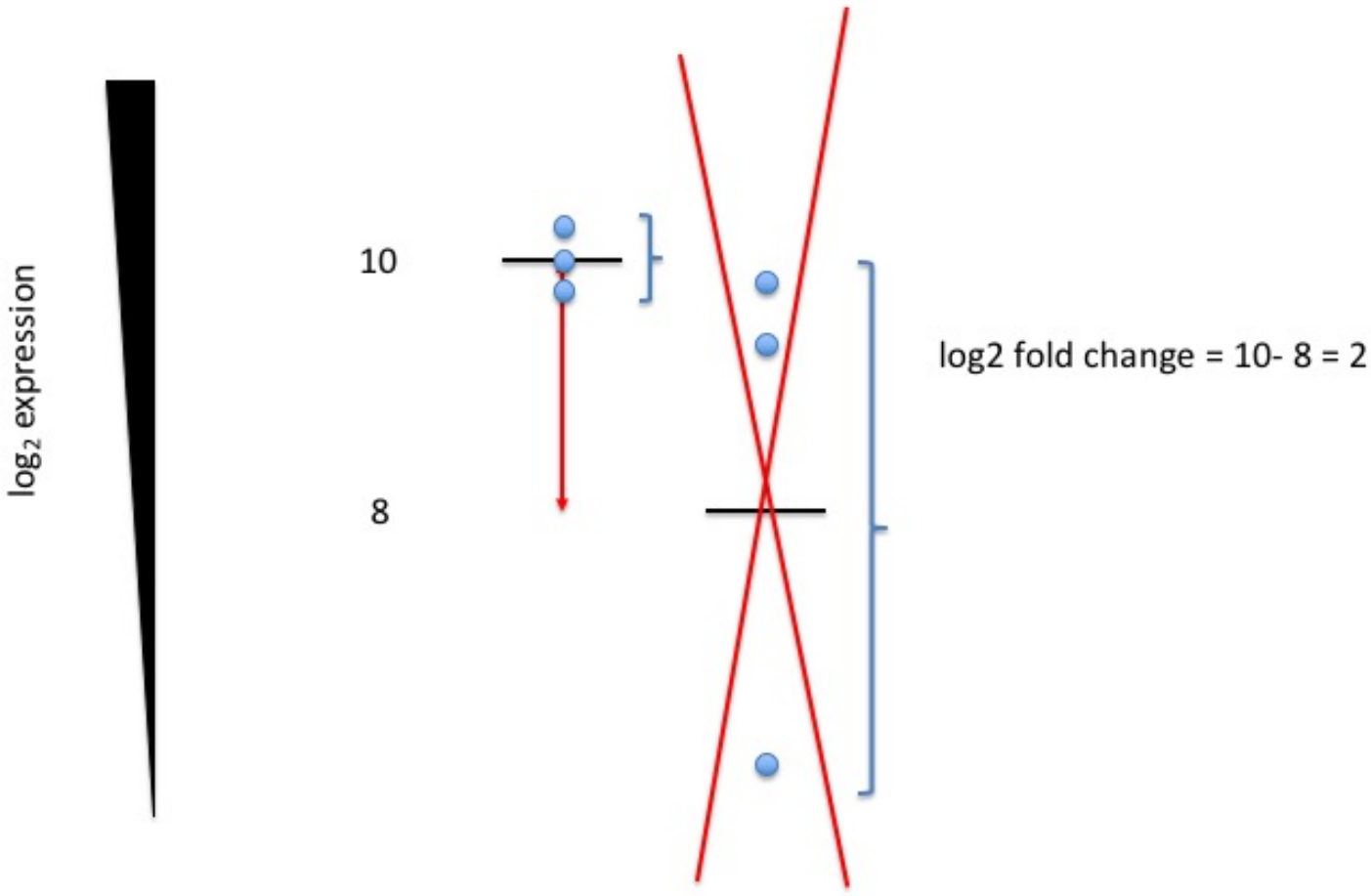
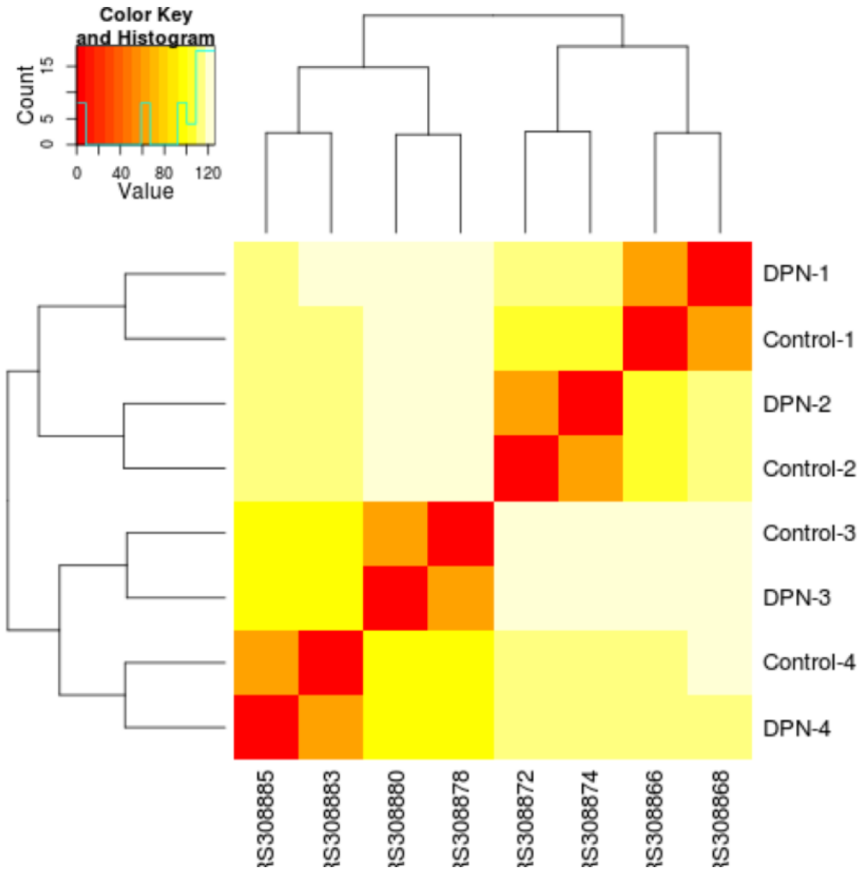
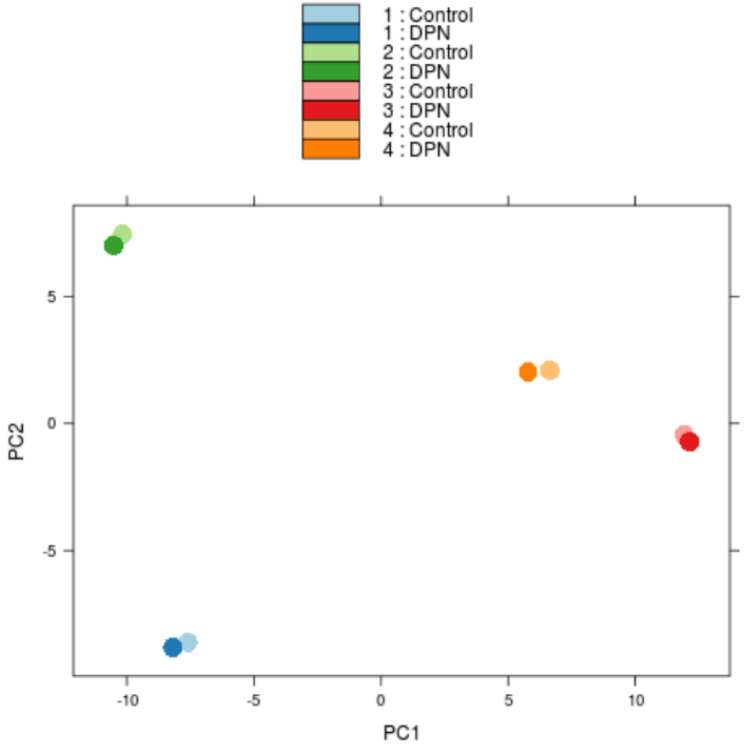


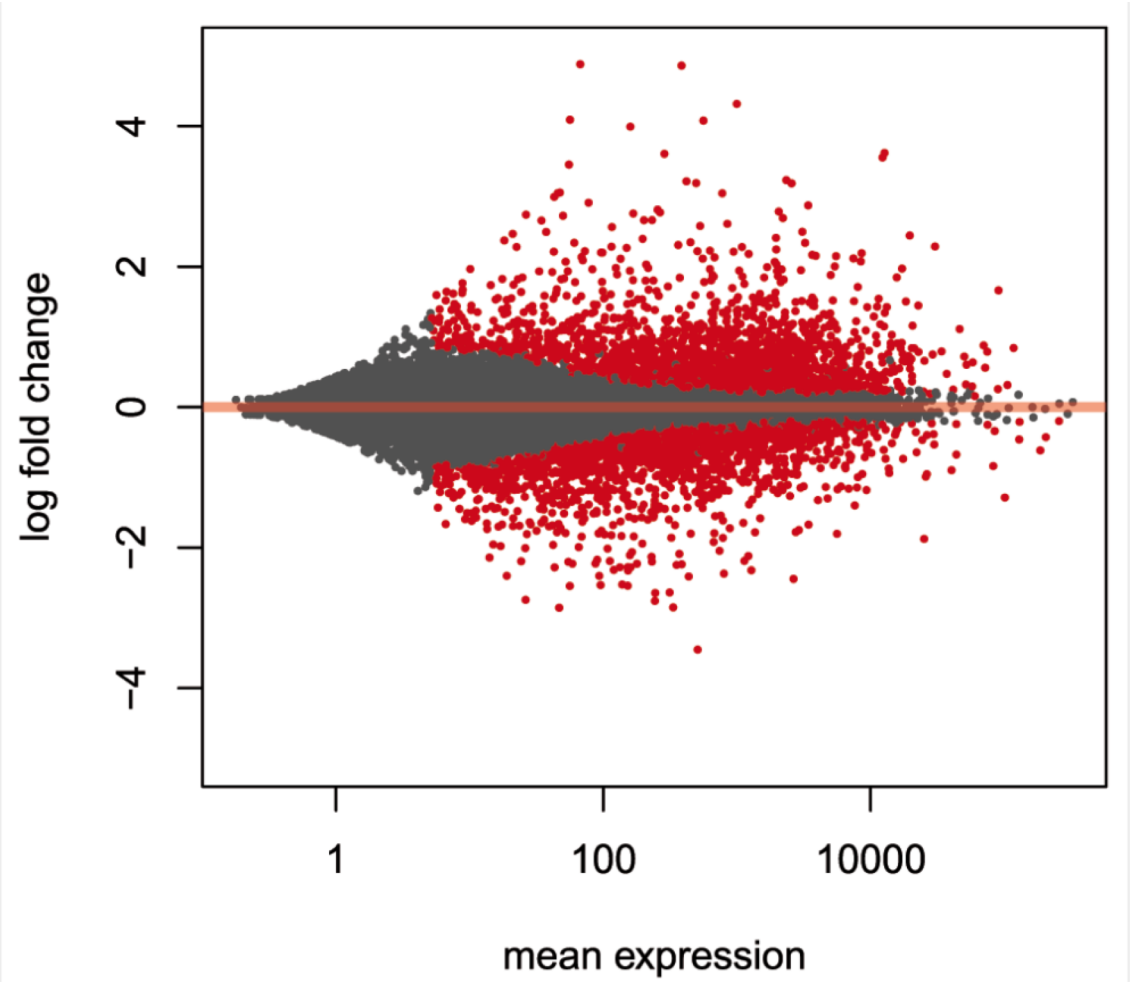
Figure 64: log2FC

Differential Expressed Genes – Visualization





Differential Expressed Genes – Visualization

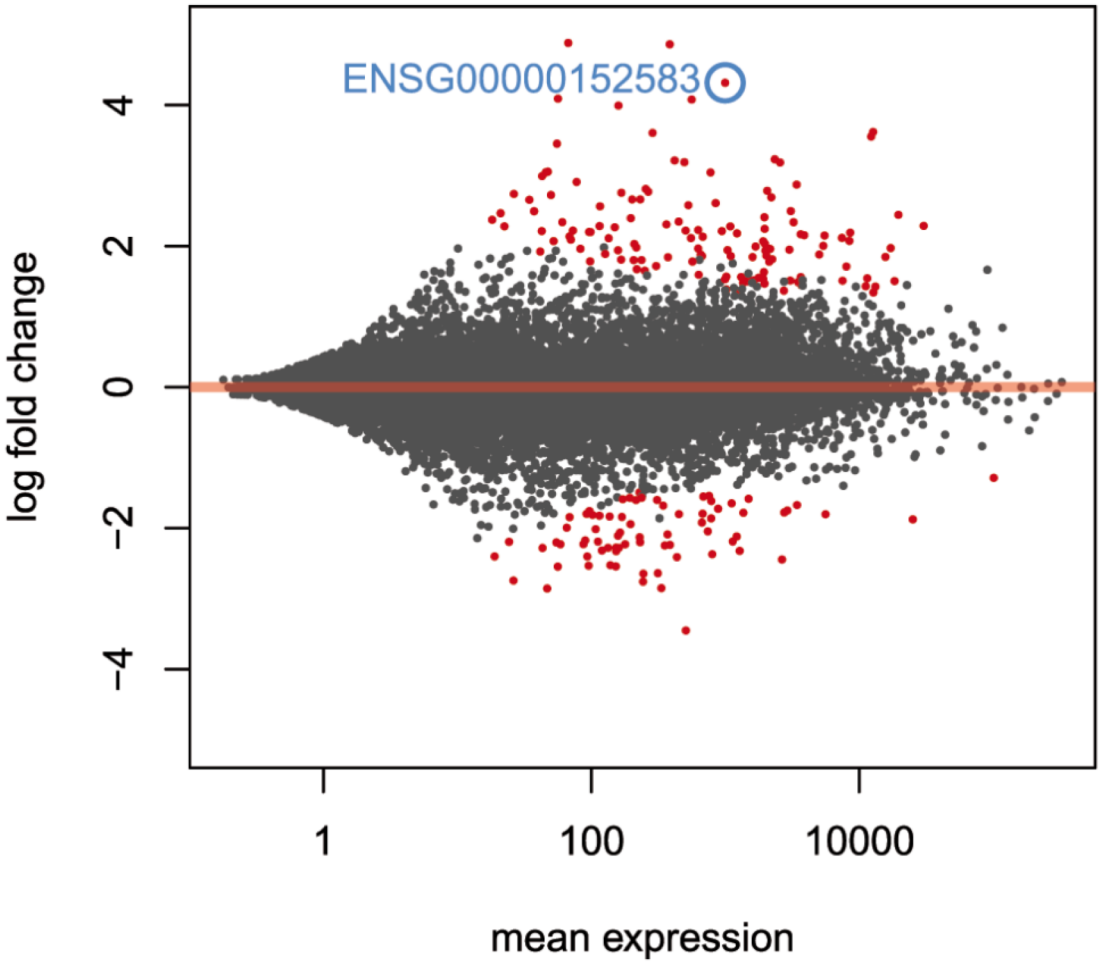


MA-plot of changes induced by treatment.

The log2 fold change for a particular comparison is plotted on the y-axis and the average of the counts normalized by size factor is shown on the x-axis (“M” for minus, because a log ratio is equal to log minus log, and “A” for average). Each gene is represented with a dot.

Genes with an adjusted *p* value below a threshold (here 0.1, the default) are shown in red.

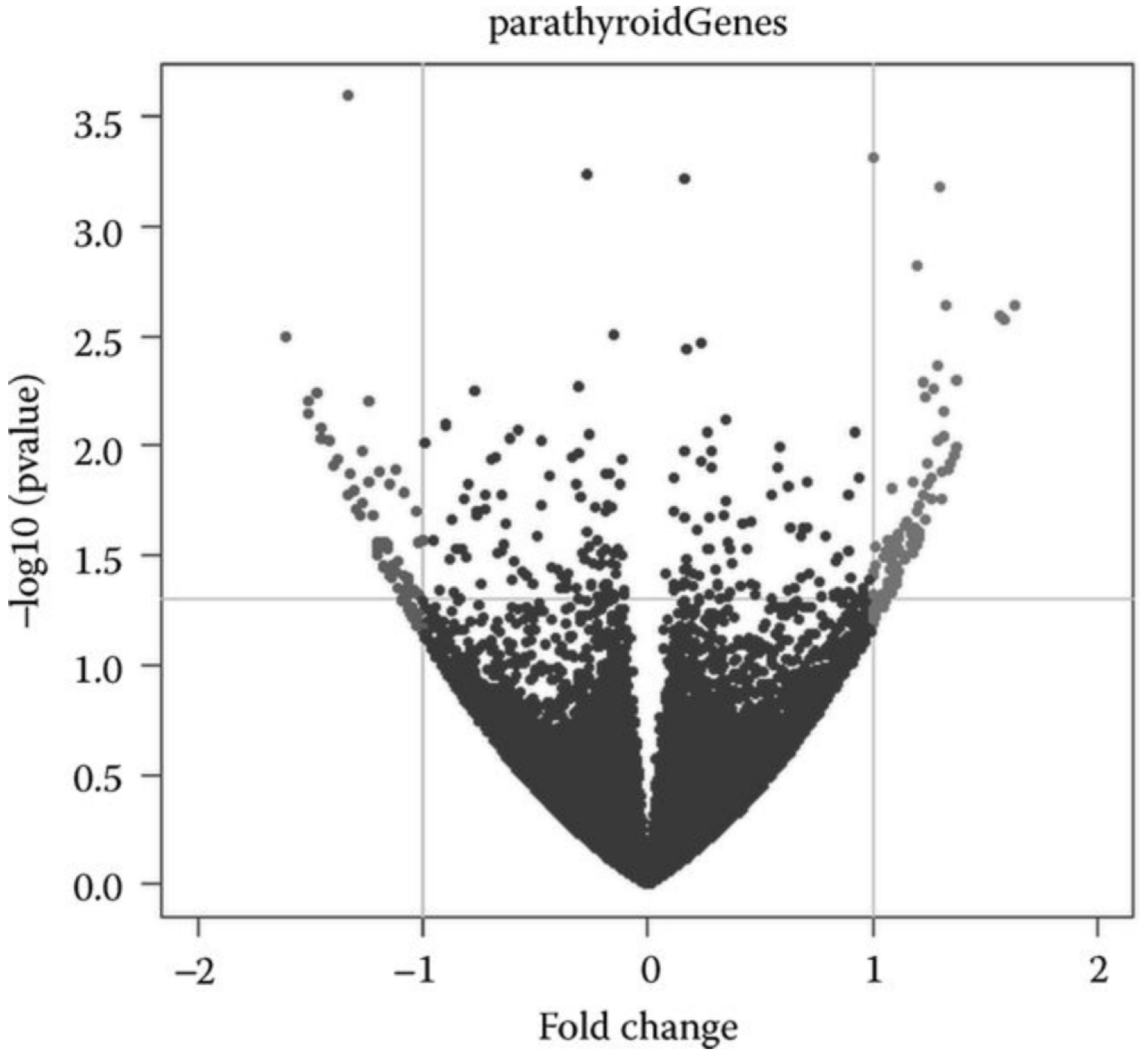
Differential Expressed Genes – Visualization



The red points indicate genes for which the log2 fold change was significantly higher than 1 or less than -1 (treatment resulting in more than doubling or less than halving of the normalized counts) with adjusted *p* value less than 0.1.

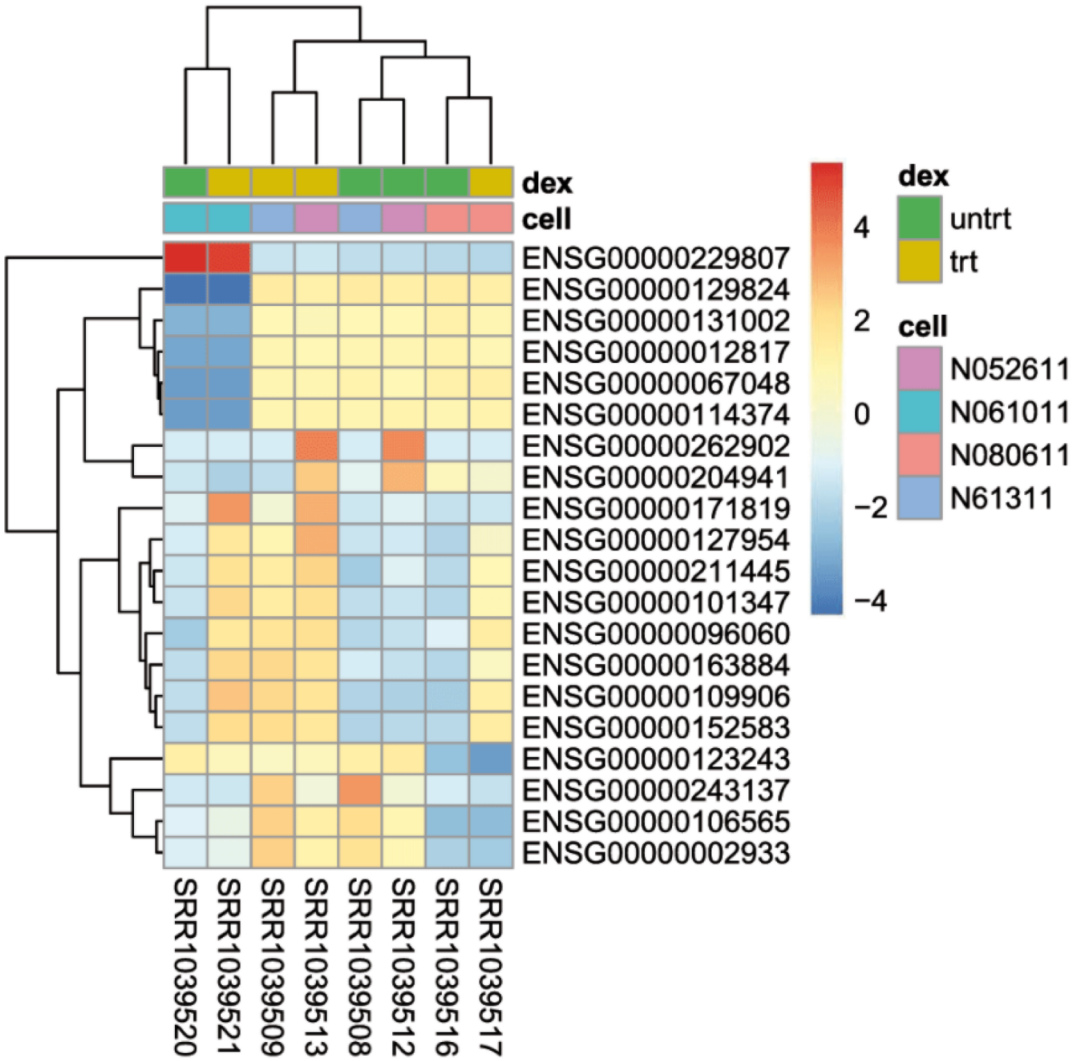
The point circled in blue indicates the gene with the lowest adjusted *p* value.

Differential Expressed Genes – Visualization



A **Volcano plot** is simply a scatterplot that has the fold change values for all features on the horizontal (*x*) axis, and the −log 10-transformed *p*-value on the vertical (*y*) axis.

Differential Expressed Genes – Visualization

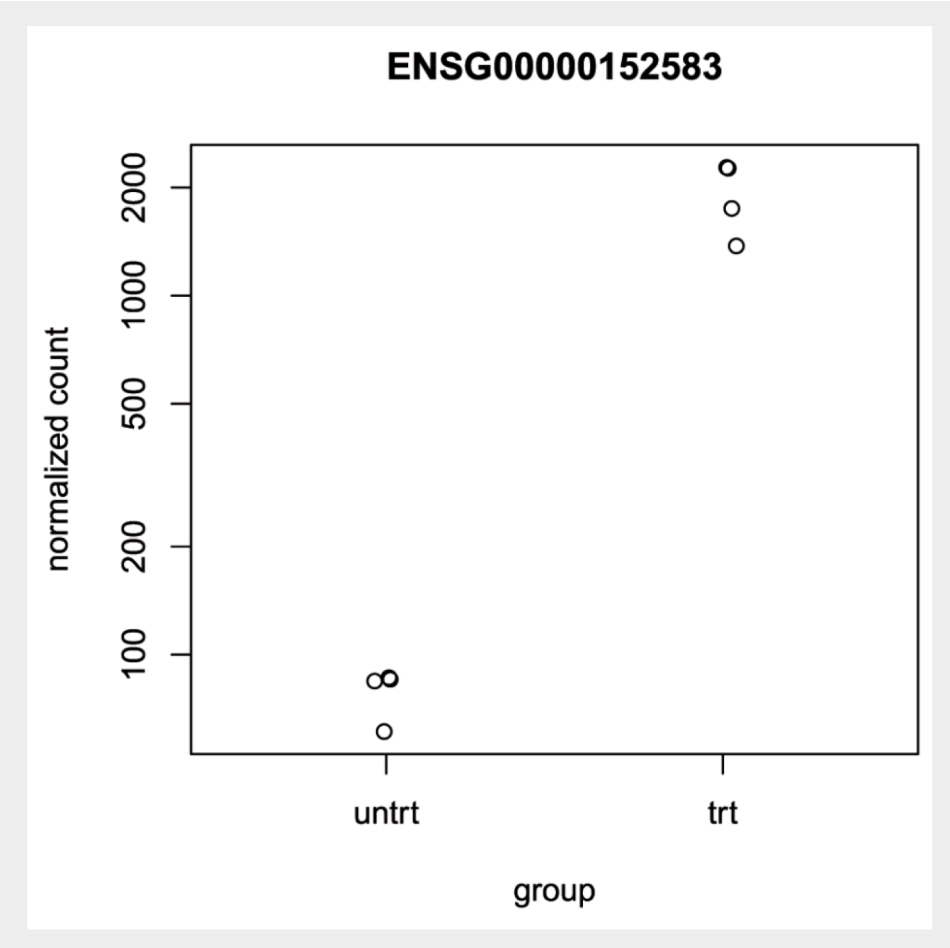


Heatmap of relative rlog-transformed values across samples.

Treatment status and cell line information are shown with colored bars at the top of the heatmap.

Note that a set of genes at the top of the heatmap are separating the N061011 cell line from the others. In the center of the heatmap, we see a set of genes for which the dexamethasone treated samples have higher gene expression.

Differential Expressed Genes – Visualization



Normalized counts for a single gene over treatment group.