B@G 2018

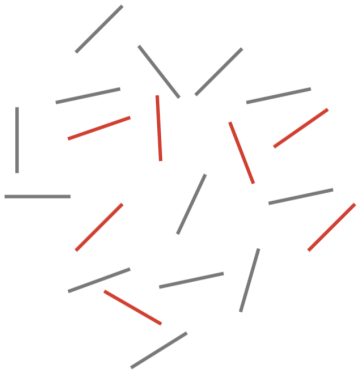
Theory and methods of RNA-seq studies

Part II
normalization, plots and
gene level analyses: DGE & eQTL

Simone Tiberi, University of Zurich

16/02/2018

Data analysis



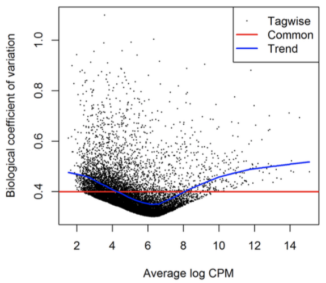
ALIGNMENT

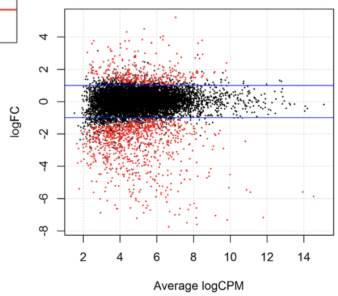
COUNTING

Gene A
Gene B
.
.
.
Gene X

$$\begin{pmatrix} 7 & \cdot & \cdot & \cdot \\ 13 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

ANALYSIS





Charlotte Soneson, UZH

1. Normalizations
2. Exploratory plots
3. Models for gene expression data
4. Incorporating transcript-level information in DGE

References

Comparability

- When comparing the expression of different genes/transcripts, keep in mind that genes/transcripts have different lengths: if two genes/transcripts are equally expressed, but one is twice as long, it will have about twice as many reads mapping to it.
- When comparing one gene/transcript across different samples, beware of:
  - possible batch effects;
  - sequencing biases (Salmon tries to correct for it via the option *seqBias*);
  - library size, i.e. the total number of reads in the sample, across all genes.
- We can calculate normalization factors to make counts comparable between different experiments. Remember that, except a few cases, the counts are not scaled themselves: we compute normalization factors to include in the model.

Normalization

How to calculate normalization factors?

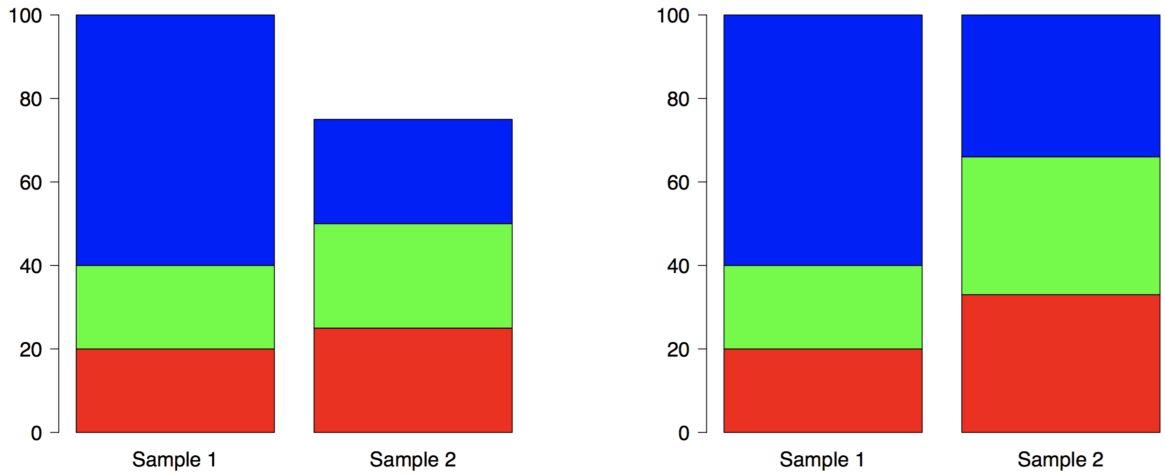
- Attempt 1: **total count** (library size)
  - Define a reference sample (one of the observed samples or a "pseudo-sample") - gives a "target library size"
  - Normalization factor for sample *j* is defined by

$$\frac{\text{total count in sample } j}{\text{total count in reference sample}}$$

Normalization

The influence of RNA composition

- Observed counts are relative
- High counts for some genes are “compensated” by low counts for other genes



Charlotte Soneson, UZH

Normalization

How to calculate normalization factors?

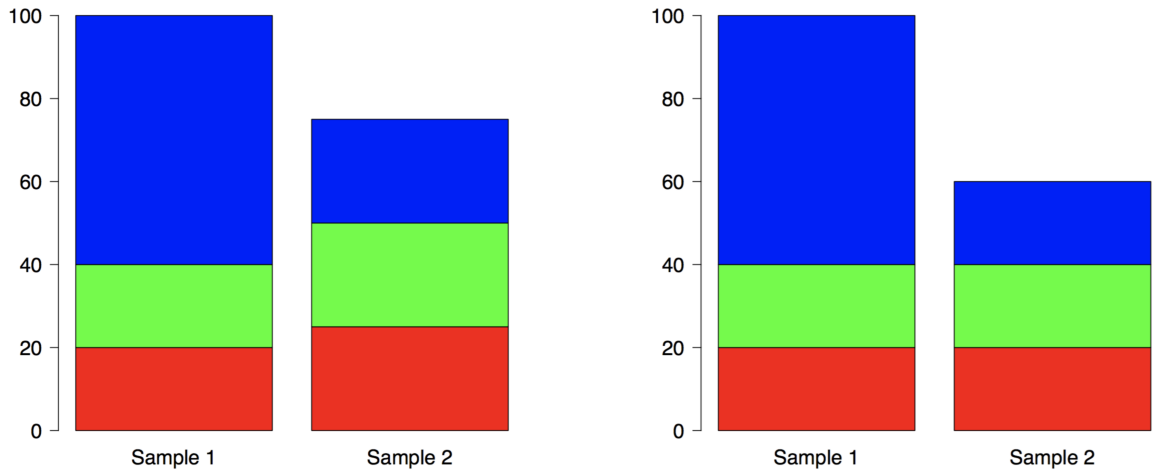
- Attempt 2: total count (library size) * compensation for differences in composition
- Idea: use only non-differentially expressed genes to compute the normalization factor
- Implemented by both edgeR (TMM) and DESeq2 (median count ratio)
- Both these methods assume that most genes are not differentially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

Charlotte Soneson, UZH

Normalization

How to calculate normalization factors?

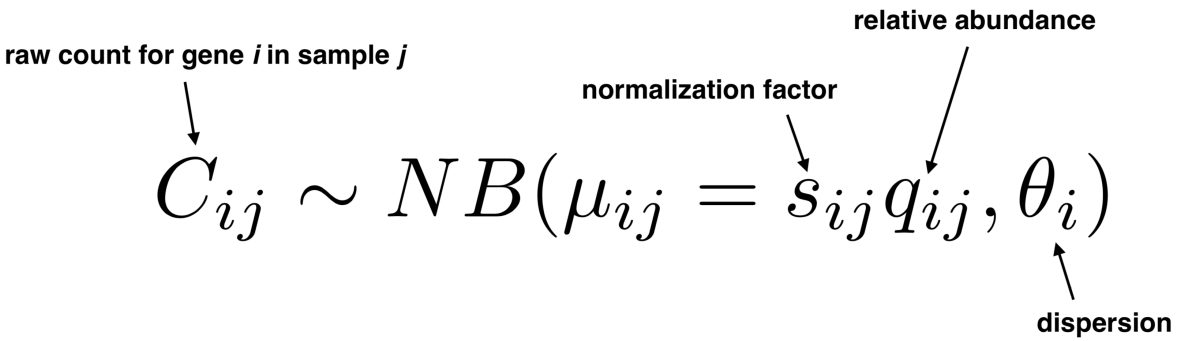
- Attempt 2: total count (library size) * compensation for differences in composition



Charlotte Soneson, UZH

Normalization

Normalization



- 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)

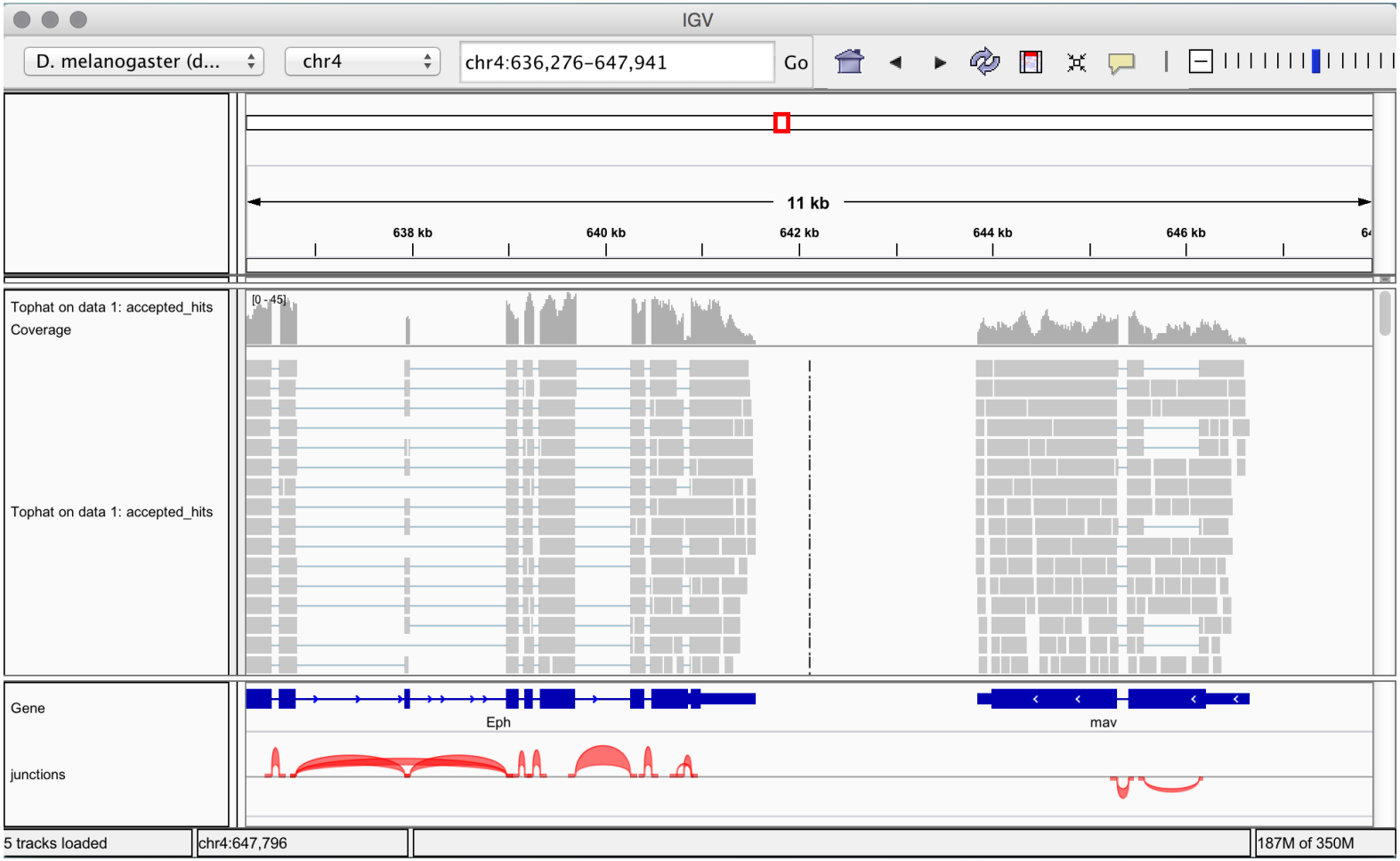
Definition of expression levels

- We introduce the following elements:
  - X_t = number of reads arising from transcript t;
  - N = total number of reads sequenced from all genes, i.e. the library size;
  - l_t = effective length of the transcript t.
    We do not consider the actual length of transcripts, but the effective transcript length: the part of the transcript I can map my read on: l_t = length_transcript - length_read + 1.
  - $Z = \sum_t \pi_t \, l_t$, where π_t represents the relative abundance of transcript t. Z is a normalization factor representing the mean effective length of transcripts, weighted by transcript relative abundance.
- CPM $= \frac{X_t}{N} \times 10^6$, counts per million; how many reads map to transcript t every 10^6 reads. CPM can also be expressed at the gene level by replacing the counts mapping to gene g, X_g, instead of X_t. CPM normalizes for the library size.

Definition of expression levels

- RPKM $= \frac{X_t}{l_t\, N} \times 10^9$, reads per kilobase per million of mapped reads. It normalizes counts for transcript length (or gene length) and for sequencing depth, i.e. library size. It is the rate of reads/fragments mapping to transcript t per base multiplied by $\frac{10^9}{N}$ to make it more convenient. It is also called fragment per kilobase per million of mapped reads, FPKM: RPKM and FPKM are the same, one name refers to reads and one to fragments (paired-end reads).
- TPM $= \frac{X_t}{l_t\, N} \times Z \times 10^6$, transcripts per million. TPM is the number of transcripts of type t you expect when sequencing 10^6 transcripts. TPM is usually preferred over RPKM/TPKM. It normalizes for transcript length and for library size.
- Remember, these normalized counts are NOT used for differential expression analyses; however, they are useful for plotting the data.

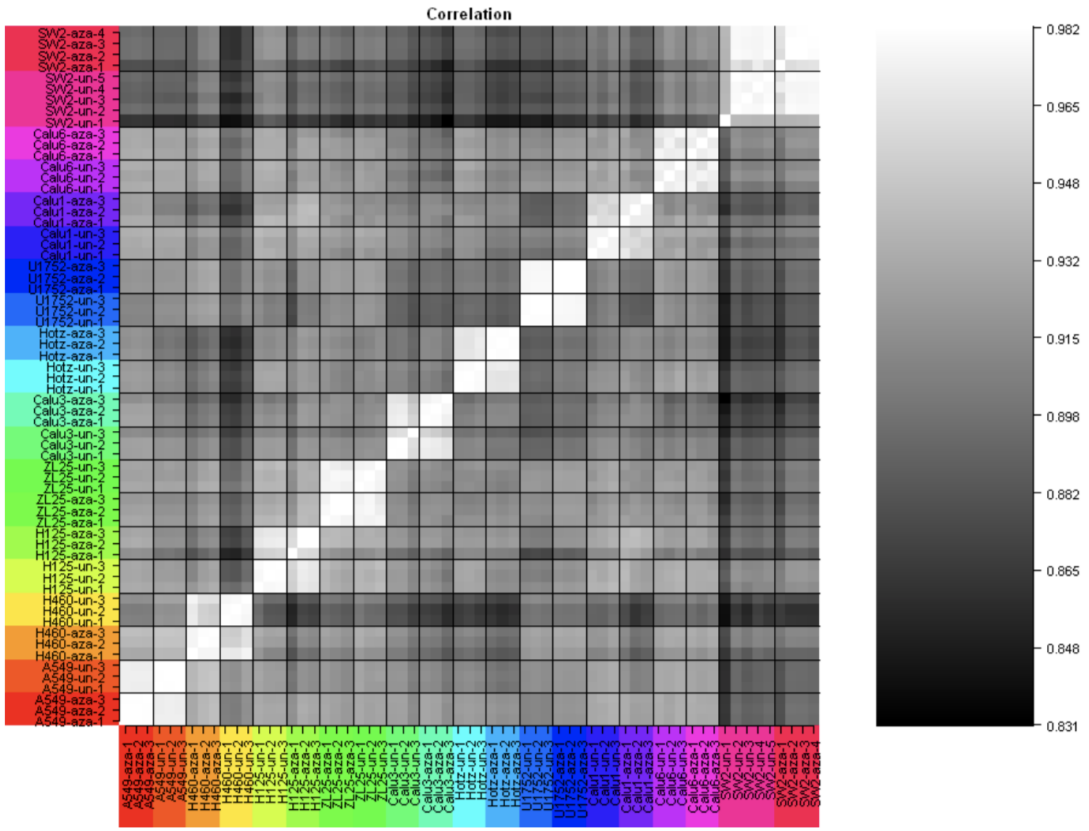
Visualization via IGV



http://gvlproject.github.io/game2017_docs/tutorials/rna_seq_dge_basic/rna_seq_basic_tutorial/

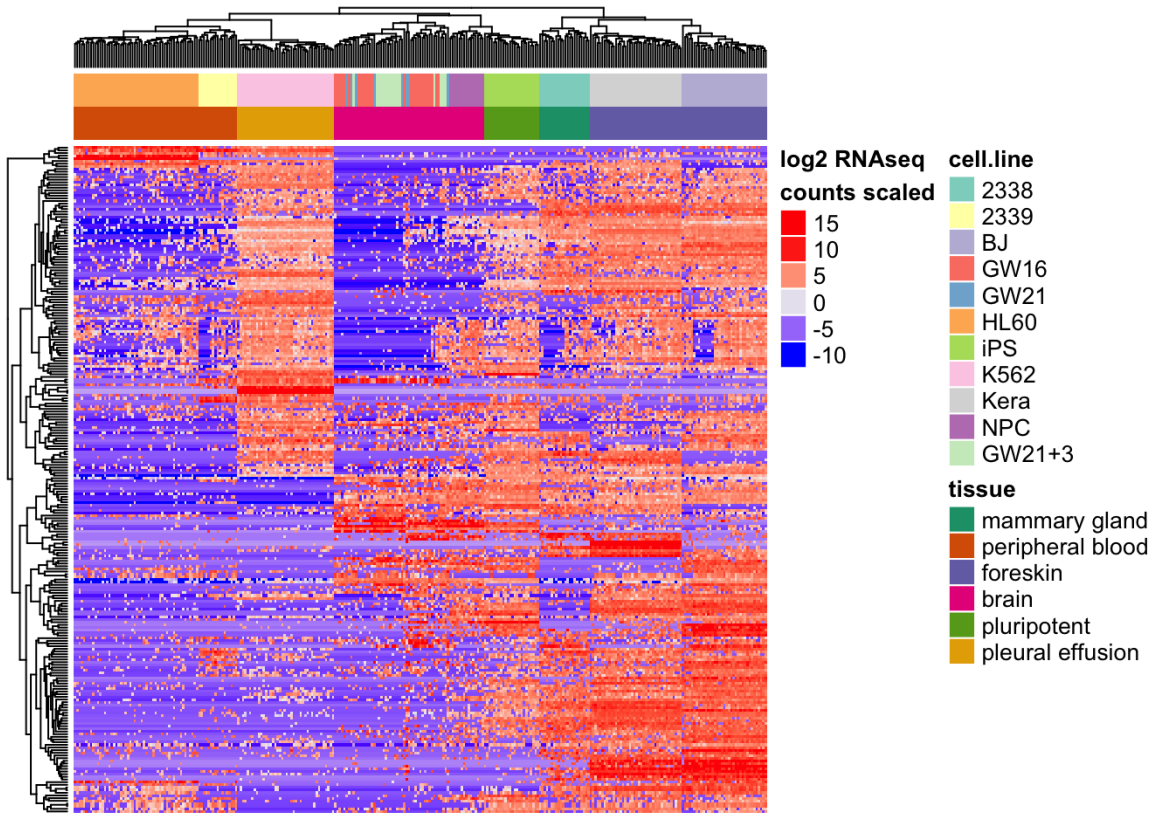
Correlation plot

- Correlation between sample pairs.
- Useful to see what samples are similar: it can indicate potentially interesting biological similarities or possible batch effects, if samples of the same batch have very high correlation.



Hierarchical clustering

- Joint hierarchical clustering of samples (columns) and genes (rows).
- Genes clustering together often have similar functions or are involved in the same pathways.
- Useful to see how samples cluster together and if there are possible batch or experimental effects.
- The length of the branches is proportional to the distance between the groups.

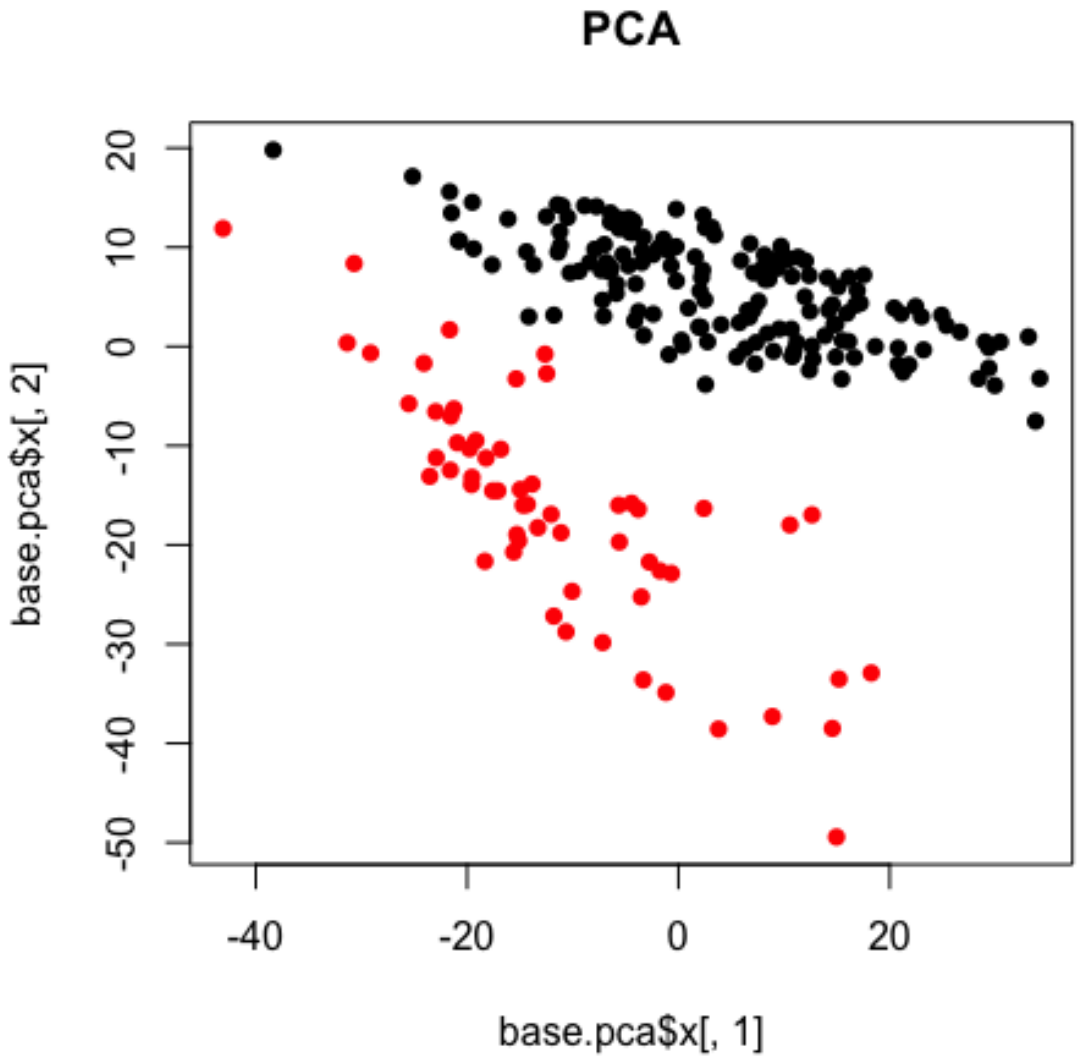


https://rpubs.com/crazyhottommy/heatmap_demystified

Dimensionality reduction

- In the exploratory stages of the analysis, it is common to perform a dimensionality reduction on the genes and plot the samples.
- Three common dimensionality reduction techniques are:
  - PCA, principal component analysis, which applies a linear transformation of the covariates (gene expression) to maximise the explained variance;
  - MDS, multidimensional scaling (more popular than PCA in RNA-seq studies);
  - tSNE, t-distributed stochastic neighbor embedding, that applies a non-linear dimensionality reduction; also very popular in the field.
- In all cases the output is a 2 or 3 dimensional plot of the samples, which we can colour by grouping to informally study how samples cluster together. If covariates are available, we can also colour by covariates to see if they also induce clustering: useful to study if a covariate should be included in the analysis or to investigate possible batch effects.

Dimensionality reduction



https://hms-dbmi.github.io/scw/heterogeneity.html

All models are wrong

George Box:

"Essentially, all models are wrong, but some are useful."

"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."

David Cox:

"The very word model implies simplification and idealization. The idea that complex physical, biological or sociological systems can be exactly described by a few formulae is patently absurd. The construction of idealized representations that capture important stable aspects of such systems is, however, a vital part of general scientific analysis and statistical models."

Andrew Gelman:

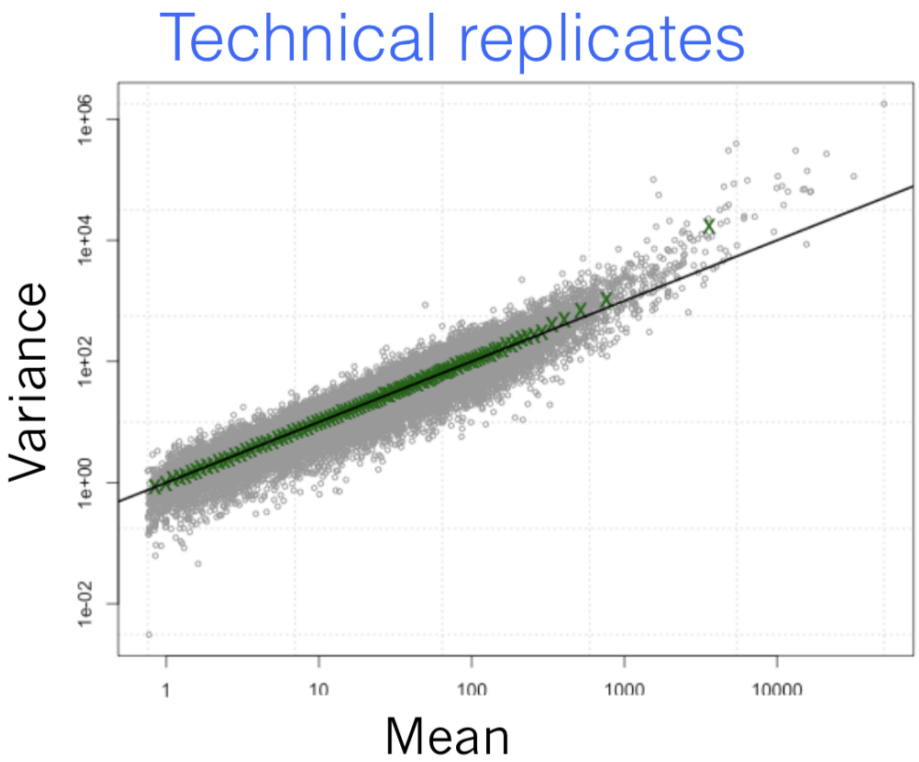
"The saying, "all models are wrong", is helpful because it is not completely obvious...This is a simple point...But, the trouble is, many people don't realize that all models are wrong."

Modelling counts

- Most RNA-seq methods input raw count data, not transformed counts (log, TPM, CPM, FPKM, etc...): they take into account for normalization factors.
- An exception: **voom**, which log-transforms the data to a suitable scale and assumes the output is normally distributed.
- Standard model for count data: the Poisson distribution.
- In the Poisson mean and variance are identical: $X \sim Po(\mu)$, $E(X) = Var(X) = \mu$.
- As $\mu \to \infty$, $X \dot{\sim} \mathcal{N}(\mu, \mu)$; the normal approximation is only valid for a big enough μ: models for discrete data are preferable for RNA-seq counts.

Poisson counts

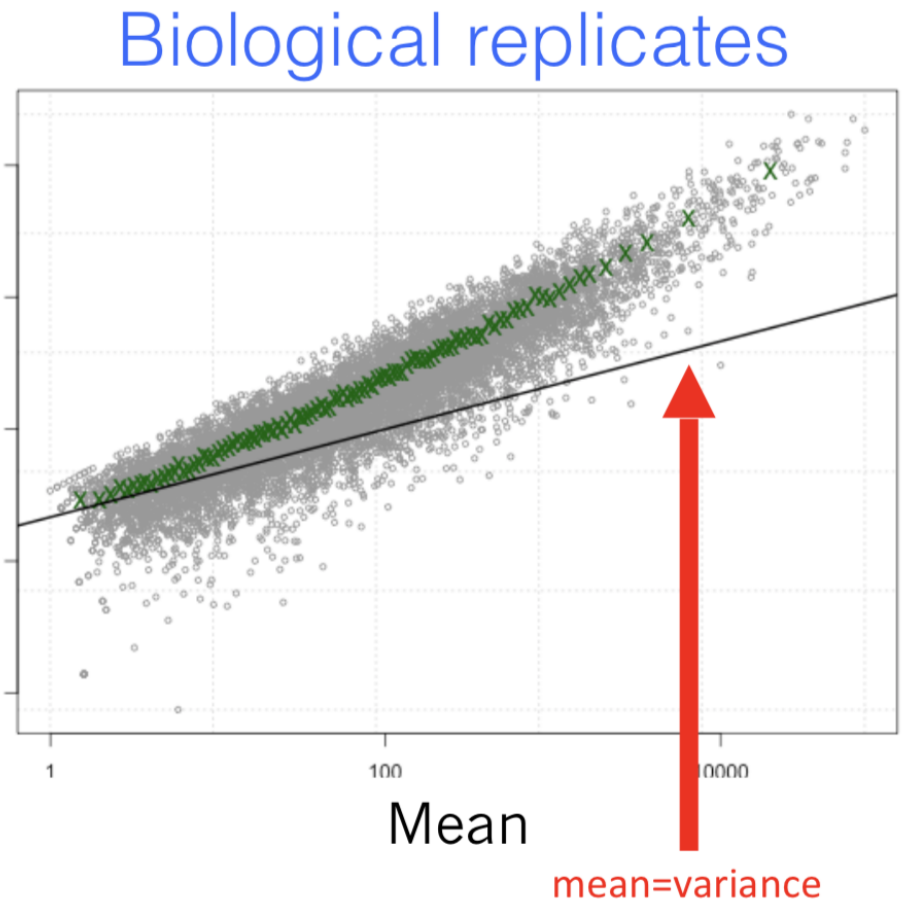
- We compute the mean and variance of the expression of genes across experimental replicates (same biological sample).
- The Poisson distribution represents well technical replicates: mean $\simeq$ variance.



Marioni et al. (2008), Genome Research

Over-dispersed counts

- For different biological samples, the variance tends to exceed the mean: the counts are over-dispersed w.r.t. the Poisson distribution.
- We consider the Negative-Binomial (NB) distribution, which generalizes the Poisson and allows for extra variability: $Var(X) = \mu + \theta\mu^2$, where θ is the dispersion parameter.



Parikh et al. (2010), Genome Biology.

Normalization

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)

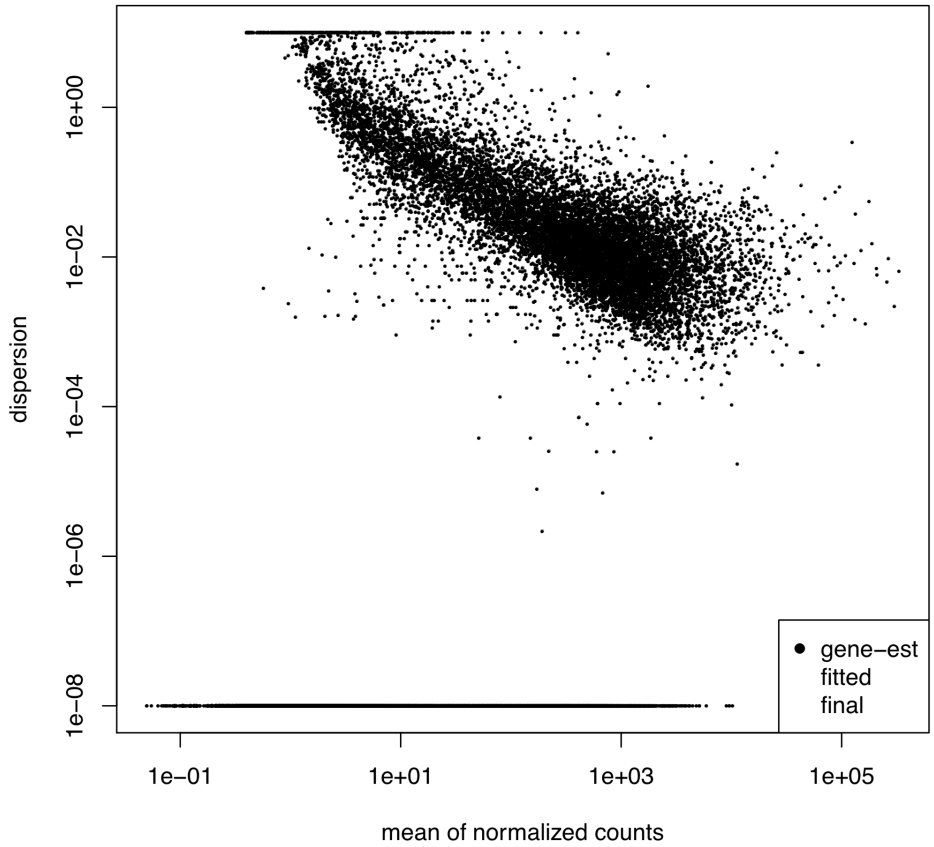
Charlotte Soneson, UZH

The negative binomial

- Once keeping the dispersion fixed, the NB is a generalized linear model (GLM).
- It is the most common model for RNA-seq data (**edgeR**, **DESeq**, **DESeq2**).
- The three methods follow a similar approach:
  - first, they estimate the dispersion parameter, θ, for every gene (a crucial step);
  - then, they fit the NB model to gene level counts, keeping θ fixed to the estimated value (*plug-in* approach);
  - finally, they test if expression levels, after accounting for normalization factors, vary between conditions: edgeR and DESeq use a likelihood ratio test, DESeq2 uses a Wald test.
- The three methods mentioned (edgeR, DESeq, DESeq2) are very similar (and indeed perform similarly!).
- To infer the gene-wise dispersion parameter, the three methods take advantage of the information from the other genes.

Dispersion estimation

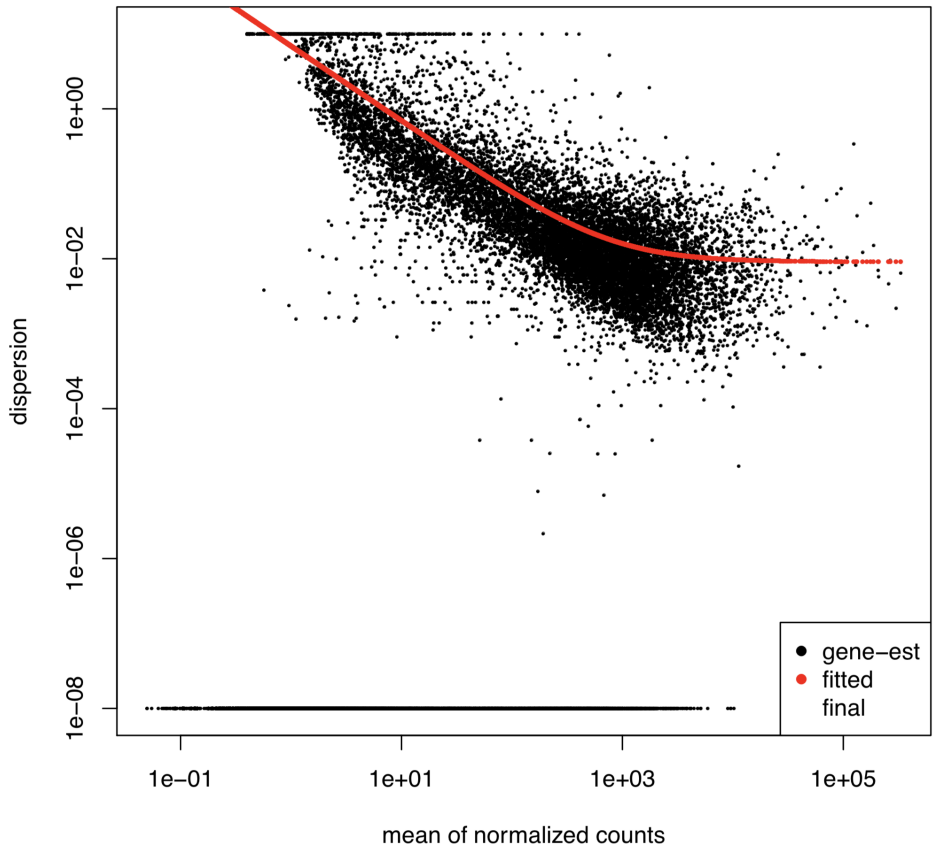
- first, the individual gene-wise dispersions are estimated (in black);



Charlotte Soneson, UZH

Dispersion estimation

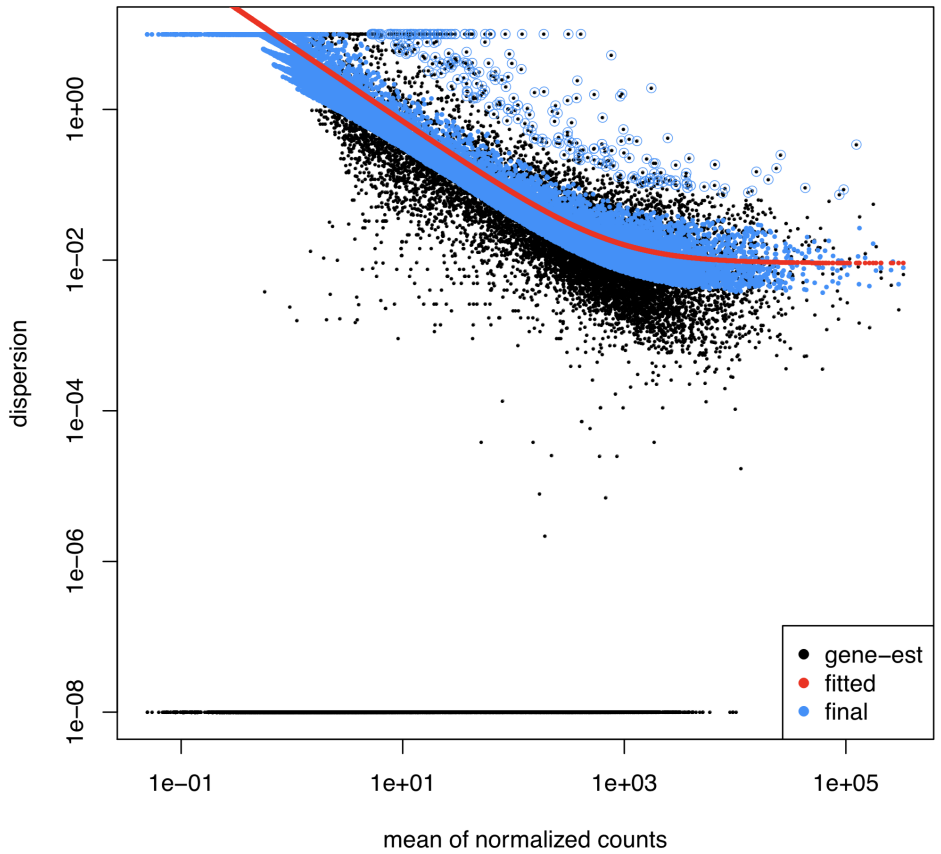
- first, the individual gene-wise dispersions are estimated (in black);
- then, a common trend is computed (in red);



Charlotte Soneson, UZH

Dispersion estimation

- first, the individual gene-wise dispersions are estimated (in black);
- then, a common trend is computed (in red);
- finally, the individual estimates are shrunk towards the common trend (in blue).



Charlotte Soneson, UZH

Differential gene expression (DGE)

- Once we fit our model to data, we typically test if the expression levels (i.e. the counts) of a gene differ between two groups, often healthy vs disease.
- In practice, DESeq2 substituted DESeq: very similar model but more sophisticated moderation of the dispersion; it works slightly better.
- If covariates are available, most DGE methods allow the user to include them in the model.
- If the experimental design requires it, they also allow to include fixed effects in the model.
- Random effects are not allowed in the three methods we saw though: use **ShrinkBayes** if random effects are needed.

How many replicates?

- Mathematically, each group must have at the very least 2 vs 2 samples: comparing 1 vs 1, we cannot disentangle the biological variability between identical samples (within group variance), from the variability between conditions, which we are interested in (between groups variance).
- Most studies perform 3 vs 3 comparisons, however a recent paper highlights that at least 6 samples per condition should be used (i.e. 6 vs 6): Schurch (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?, RNA.
- When allocating money, should we sequence deeper (more reads per sample) or sequence more samples?
  (Almost always) sequence more samples.

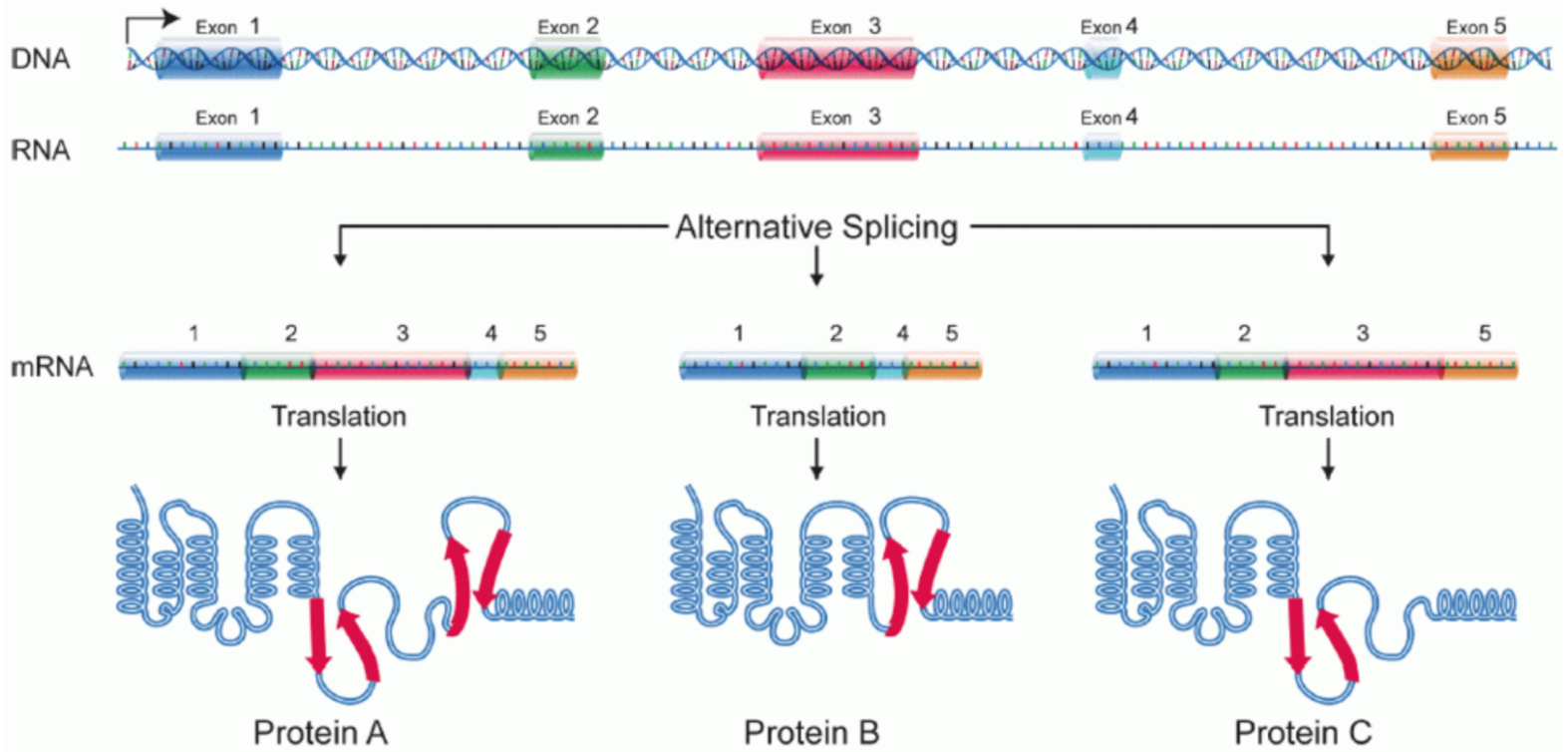
Expression quantitative trait loci (eQTL)

- In DGE, we test if a gene has different expression levels between groups.
- If we have information about the phenotype, represented by single nucleotide polymorphisms (SNPs), we can test if a gene has different expression levels between phenotypes.
- This kind of analysis is called expression quantitative trait loci (eQTL).
- The analysis is performed in the same way as for DGE where the separation in groups is defined by phenotypes.
- For eQTL, we perform many more tests than for DGE: for every gene we need to test many SNPs; we typically only test the SNPs in the neighbourhood of the gene we are considering.

Alternative splicing affects DGE

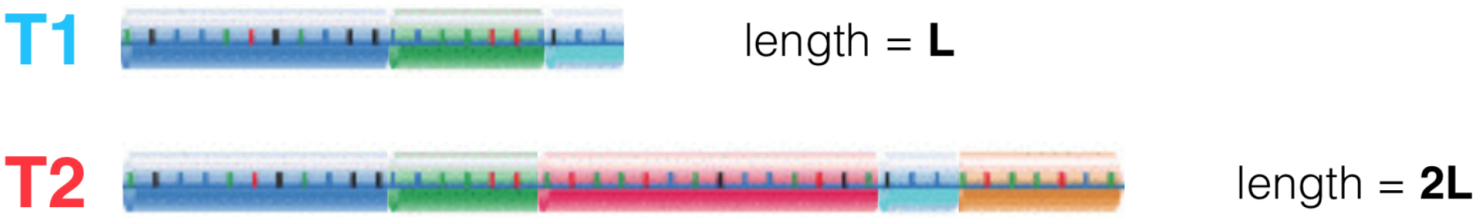
- Since transcripts have different lengths, alternative splicing can affect the detection of differentially expressed genes.
- Problem well described in Soneson et al. (2015), where the authors provide an R package (**tximport**) which allows to load easily transcript level estimates, counts and effective transcript lengths estimated from Salmon, kallisto, etc...
- Very simple idea, easy to implement into other methods (edgeR, DESeq, DESeq2, etc...) and extremely important.

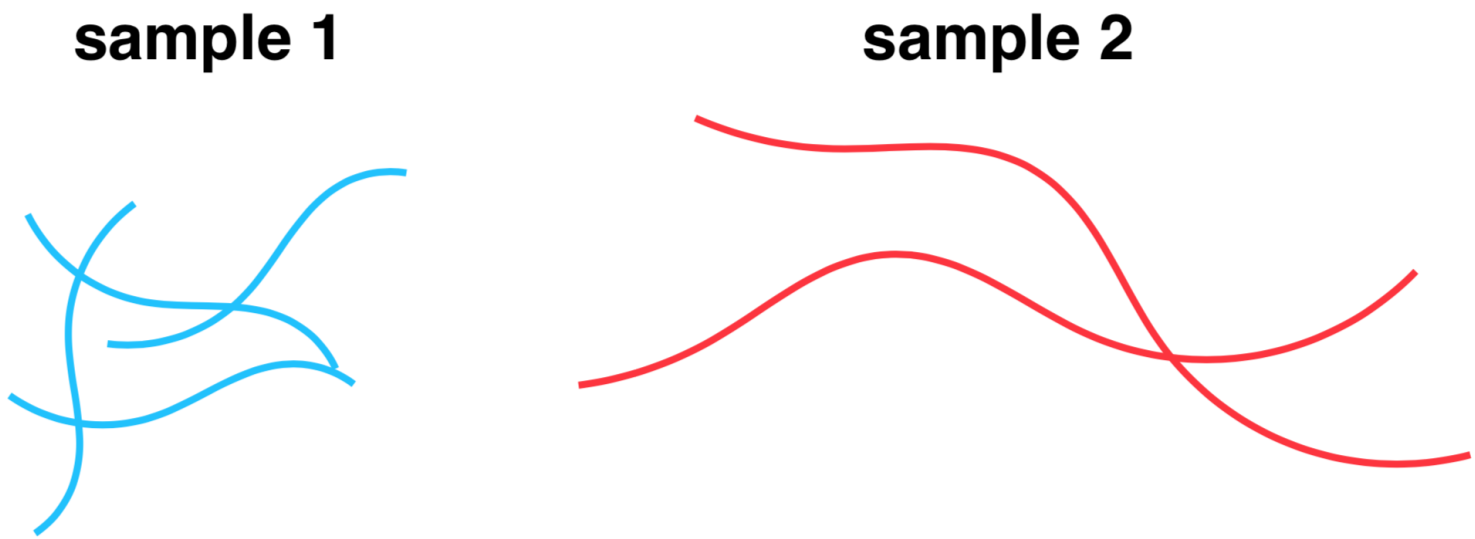
Alternative splicing affects DGE



Wikipedia

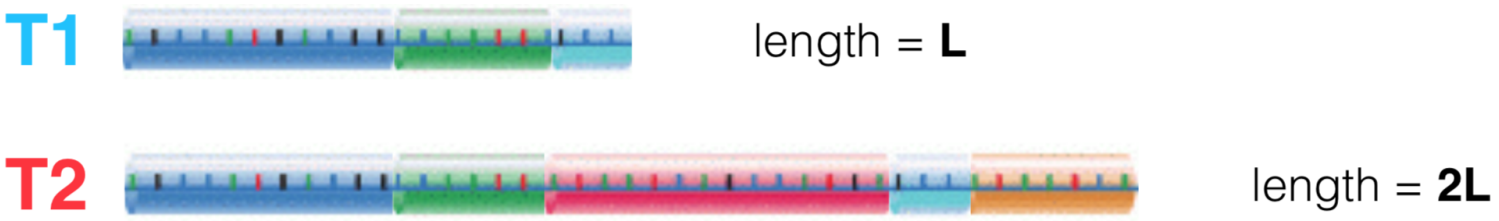
Alternative splicing affects DGE

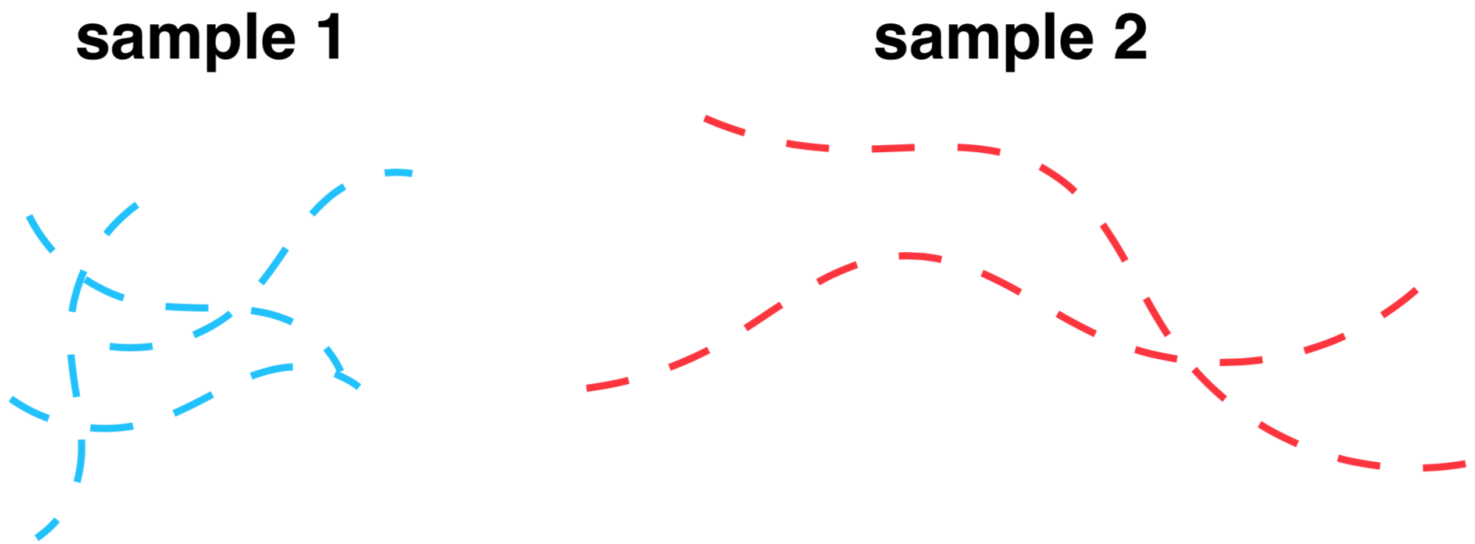




Charlotte Soneson, UZH

Alternative splicing affects DGE





Charlotte Soneson, UZH

Alternative splicing affects DGE

Gene	S1	S2
Count	150	150

Charlotte Soneson, UZH

Average transcript length (ATL)

$$ATL_{g1} = 1 \cdot L + 0 \cdot 2L = L$$

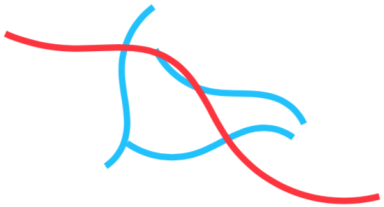
$$ATL_{g2} = 0 \cdot L + 1 \cdot 2L = 2L$$

Charlotte Soneson, UZH

Average transcript length (ATL)

length = **L**

length = **2L**



$$ATL_{g1} = 0.75 \cdot L + 0.25 \cdot 2L = 1.25L$$

$$ATL_{g2} = 0.5 \cdot L + 0.5 \cdot 2L = 1.5L$$

Include ATL in DGE analysis

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

scaling factor

relative abundance

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

dispersion

- Extend scaling factor for given sample and gene to include the average length of the transcripts from the gene that are present in the sample

Charlotte Soneson, UZH

Include ATL in DGE analysis

- Similar to correction factors for library size, but sample-**and** gene-specific
- Transcript abundance levels (TPMs) can be obtained from (e.g.) Salmon or kallisto
- Average transcript length for gene *g* in sample *s*:

$$ATL_{gs} = \sum_{i \in g} \theta_{is} \bar{\ell}_{is}, \qquad \sum_{i \in g} \theta_{is} = 1$$

$\bar{\ell}_{is}$ = effective length of isoform i (in sample s)
θ_{is} = relative abundance of isoform i in sample s

Include ATL in DGE analysis

- The tximport vignette, in the "Use with downstream Bioconductor differential expression packages" section, explains how to incorporate the ATL in the DGE analyses of edgeR, DESeq2 and limma-voom.
- For more info, read:
  `https://bioconductor.org/packages/release/bioc/vignettes/tximport/inst/doc/tximport.html`

References

References

- General RNA-seq workflow: Love et al. (2015). RNA-Seq workflow: gene-level exploratory analysis and differential expression, F1000Research.
- **tximport**: Soneson et al. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences, F1000Research.
- **edgeR**: Robinson et al. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics.
- **DESeq**: Anders et al. (2010). Differential expression analysis for sequence count data, Genome Biology.
- **DESeq2**: Love et al. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biology.
- **ShrinkBayes**: Van de Wiel et al. (2014). ShrinkBayes: a versatile R-package for analysis of count-based sequencing data in complex study designs, BMC Bioinformatics.
- Number of replicates: Schurch (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?, RNA.

Questions?