Quantitative RNAseq analysis

Irina Mohorianu (CSCI)

Overview

- Pre-processing of count matrices [Aim: data clean-up]
- Normalisation [Aim: comparable expression distributions]
- Differential expression analyses [Aim: bio-ok results]
- Examples of other types of analyses
- Brief introduction to small non-coding RNAs

mRNAseq. Preprocessing of count matrices

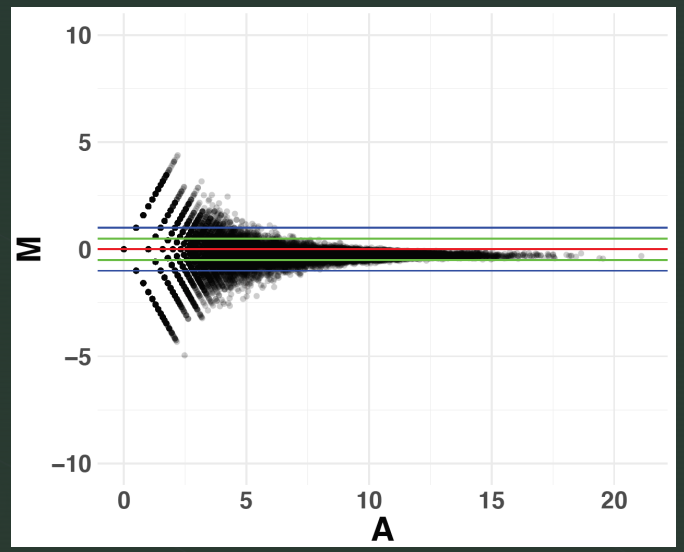
[1] **MA plots**

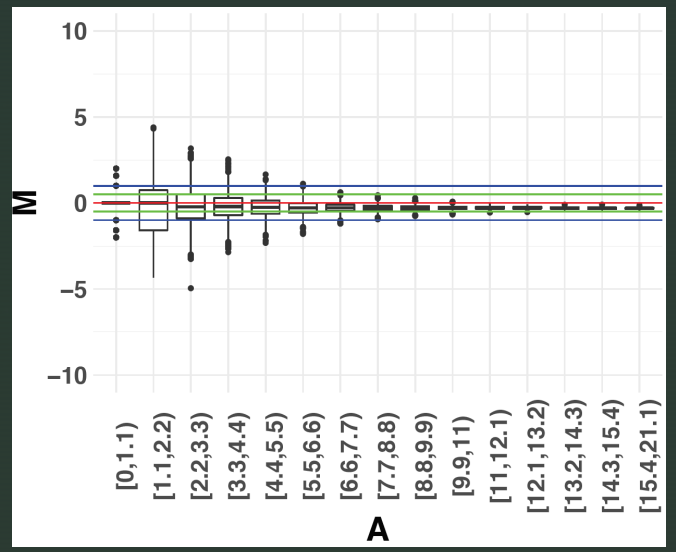
[2] density plots

[3] PCAs

[4] Jaccard Similarity Plots

[5] Noise detection





[a] wide variation for low abundance entries
[b] funneling shape for higher abundance

Summarization is possible for these samples.

Data processed by I Moutsoupoulos

mRNAseq.
Preprocessing of count matrices

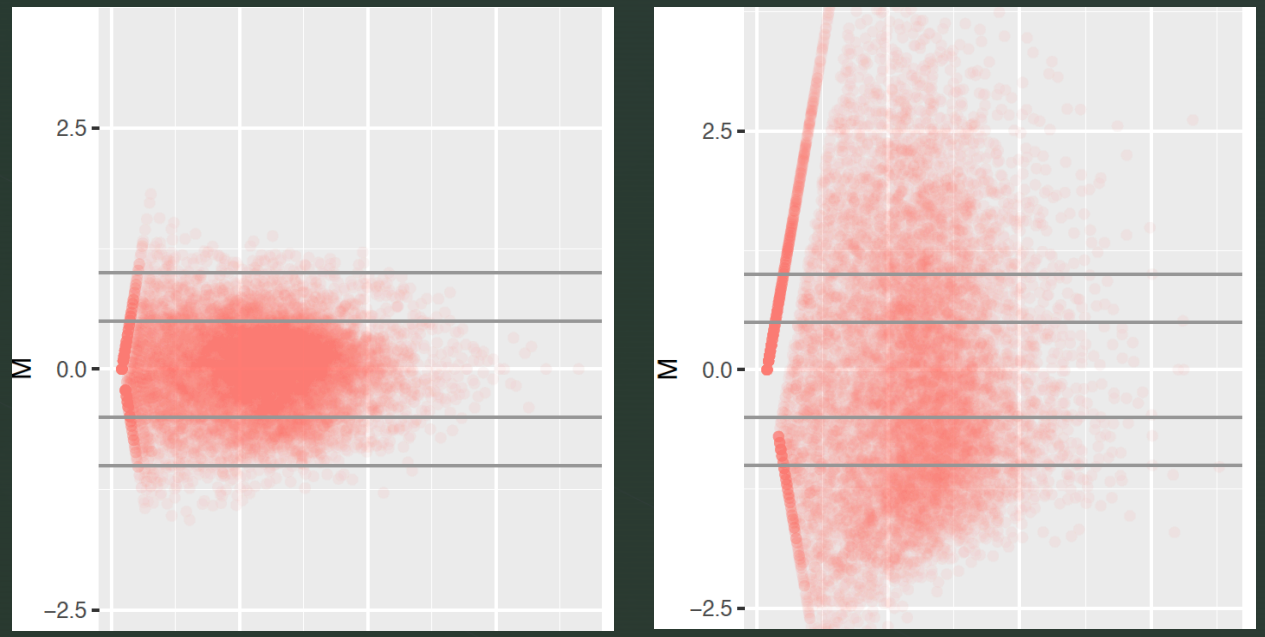
[1] **MA plots**

[2] density plots

[3] PCAs

[4] Jaccard Similarity Plots

[5] Noise detection



MA plots can also highlight issues with the data
[a] illustrates a random relationship between replicates
[b] illustrates degradation
The results on the count matrix usually align with previous feedback e.g. from library quantification

Data processed by E Williams

mRNAseq.
Preprocessing of count matrices

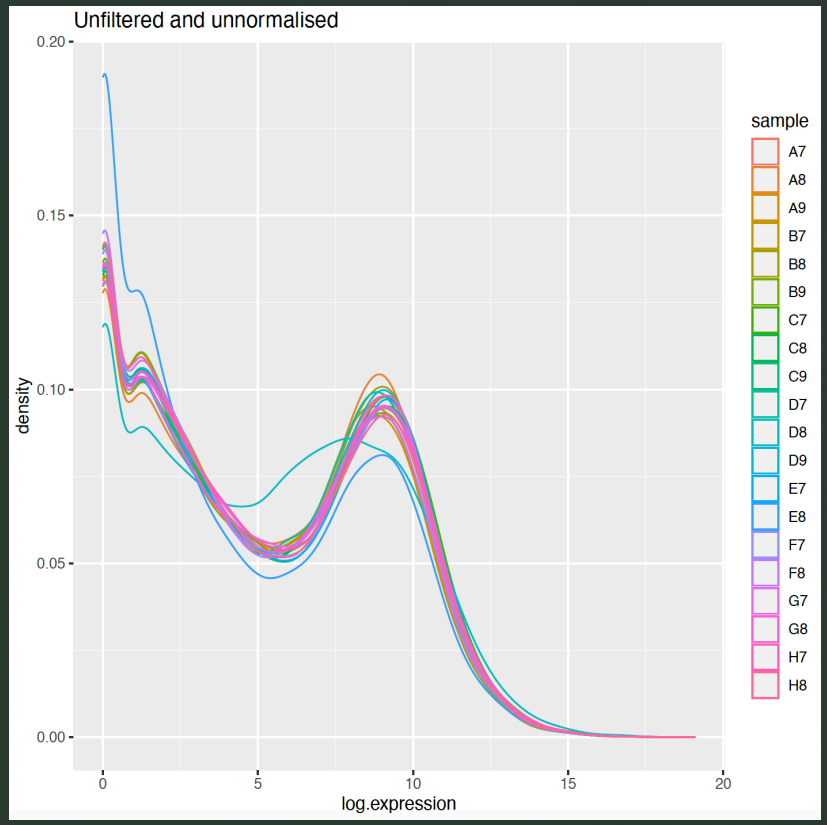
[1] MA plots

[2] **density plots**

[3] PCAs

[4] Jaccard Similarity Plots

[5] Noise detection



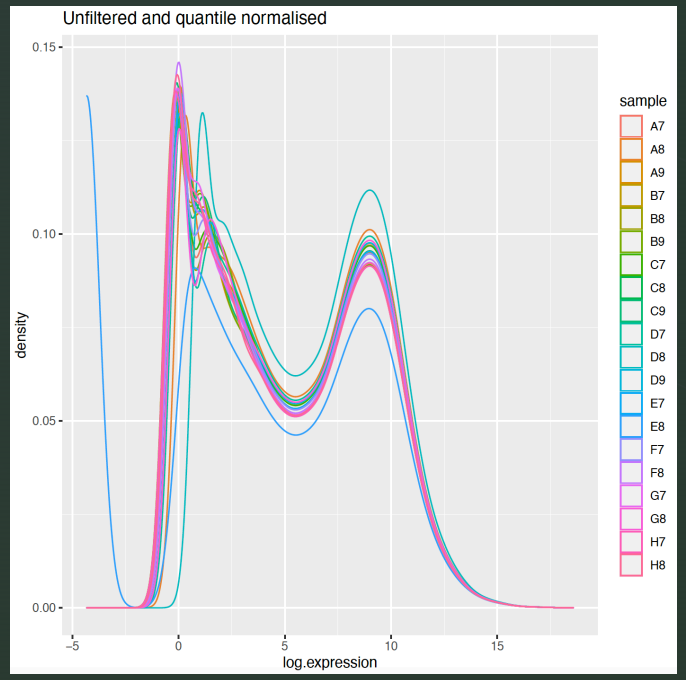
Important elements:
[a] shape of the distribution
[b] variation between samples
[c] outliers

Tasks:
[a] exclusion of noise
[b] normalisation

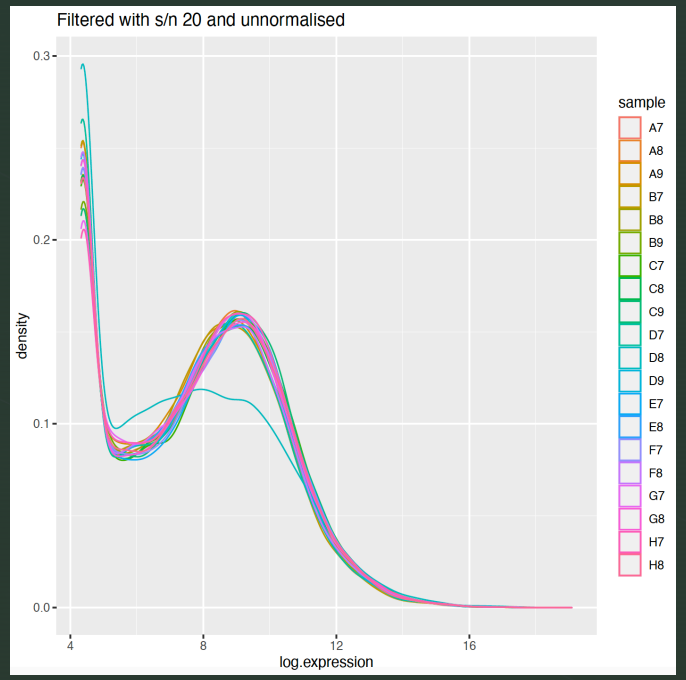
Data processed by E Williams

mRNAseq.
Preprocessing of count matrices

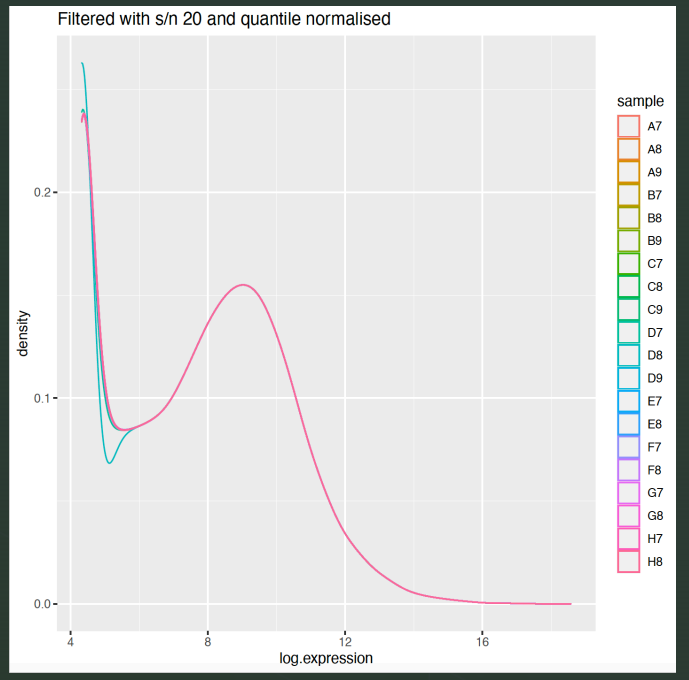
Tasks:
[a] exclusion of noise
[b] normalisation



No noise filtering
Quantile normalisation



Noise filtering
No normalisation



Noise filtering
Quantile normalisation

Data processed by E Williams

mRNAseq.
Preprocessing of count matrices

[1] MA plots

[2] density plots

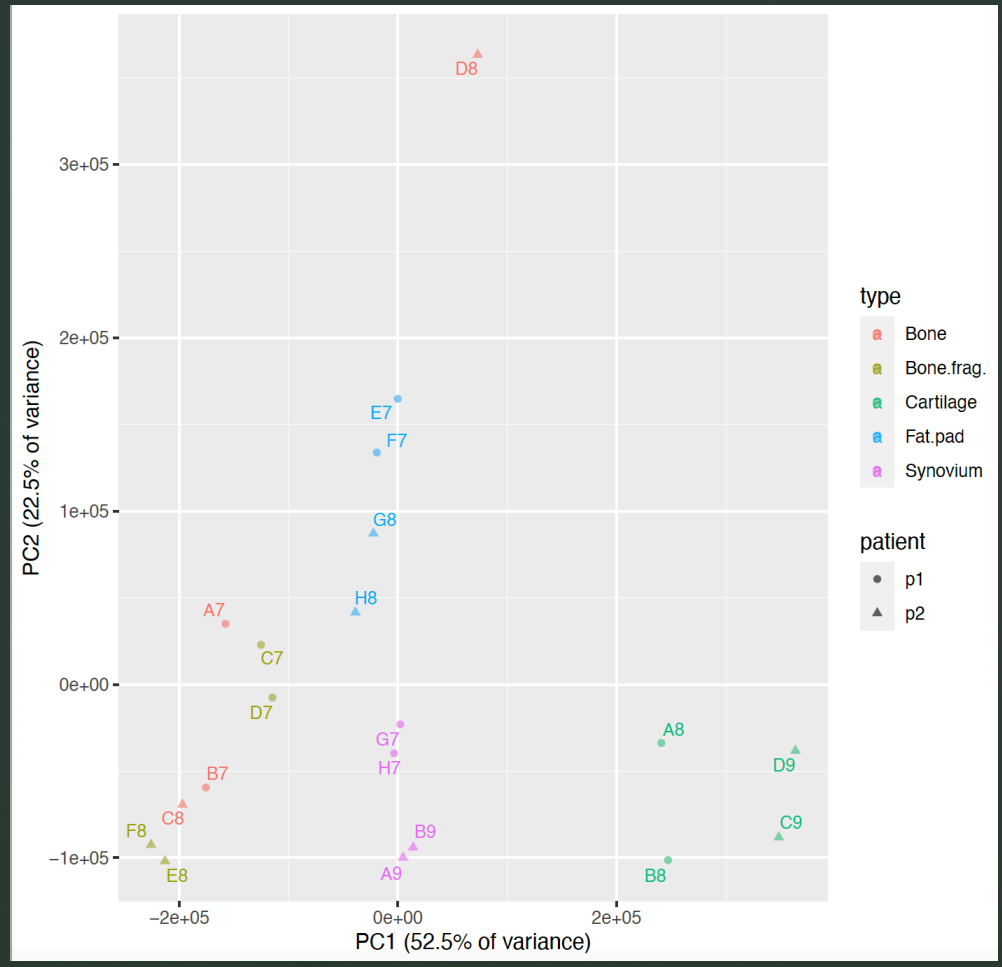
[3] **PCAs**

[4] Jaccard Similarity Plots

[5] Noise detection

PCAs are influenced by variations in expression [either at distribution level or noise]
> Should be performed incrementally

Aim: consistency of replicates
Good separation between conditions



Data processed by E Williams

mRNAseq.
Preprocessing of count matrices

[1] MA plots

[2] density plots

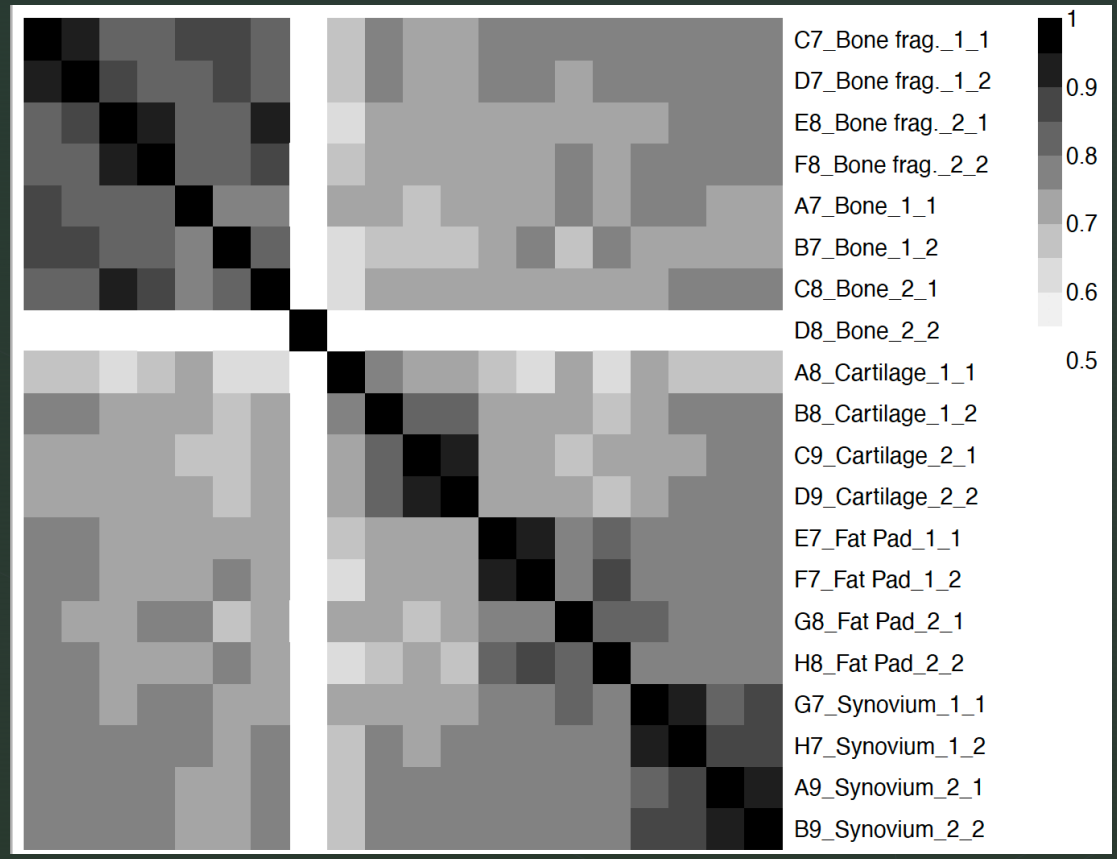
[3] PCAs

[4] **Jaccard Similarity Plots**

[5] Noise detection

JSI = ratio between the genes found in common vs the genes found in at least one sample.

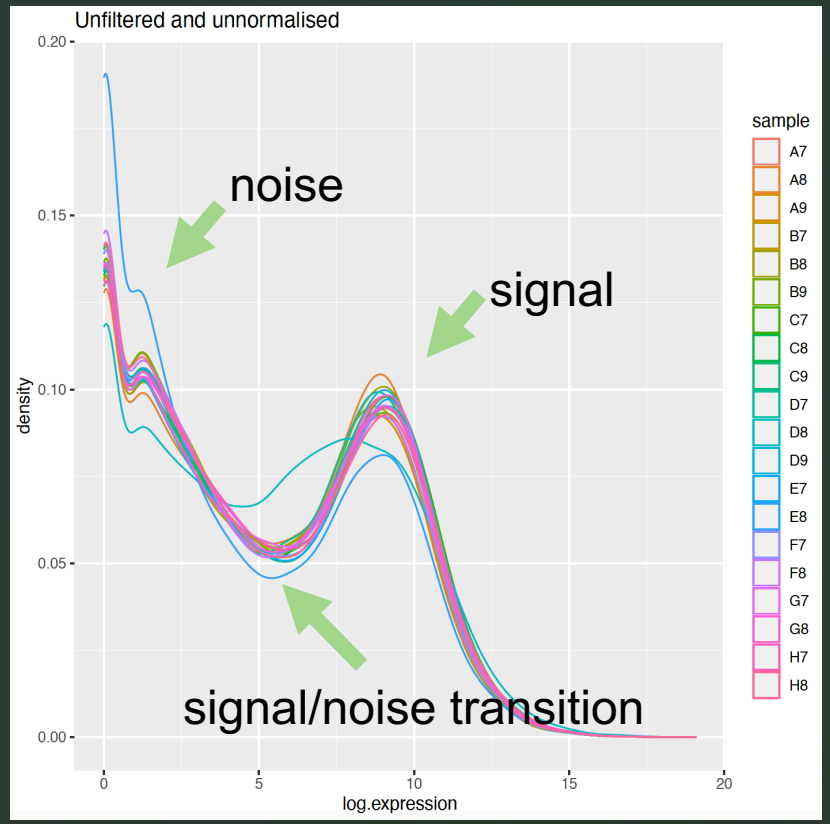
> Should be performed incrementally

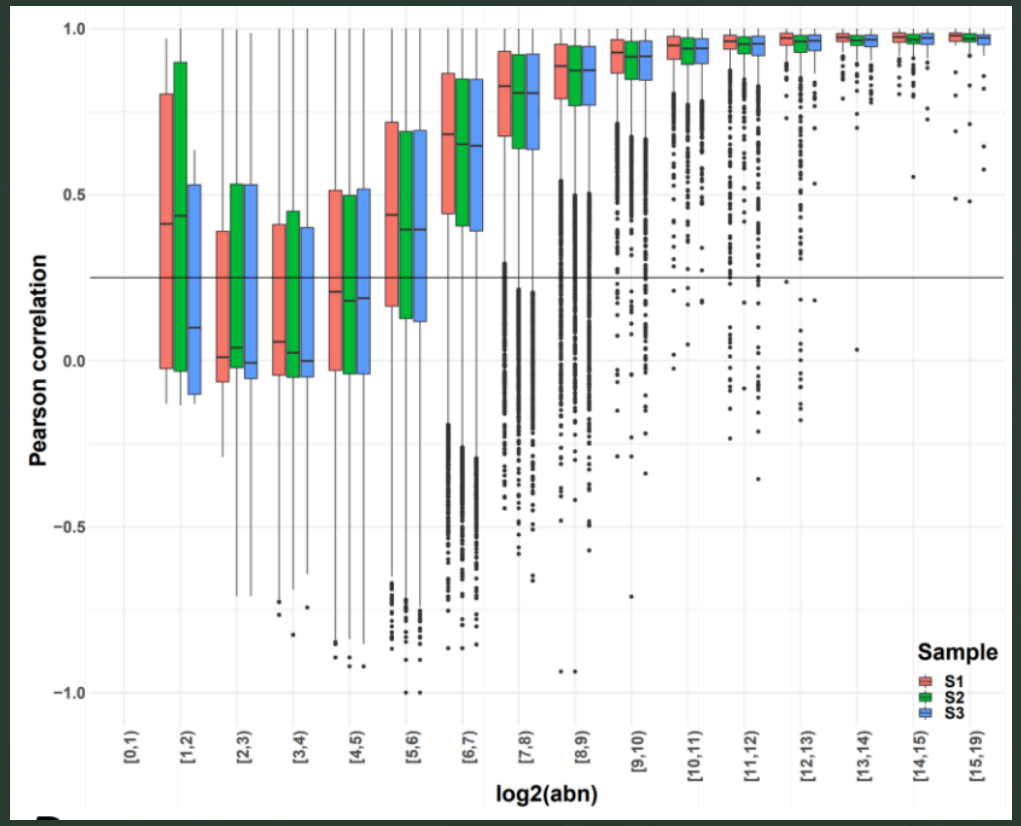


Data processed by E Williams

mRNAseq.
Preprocessing of count matrices

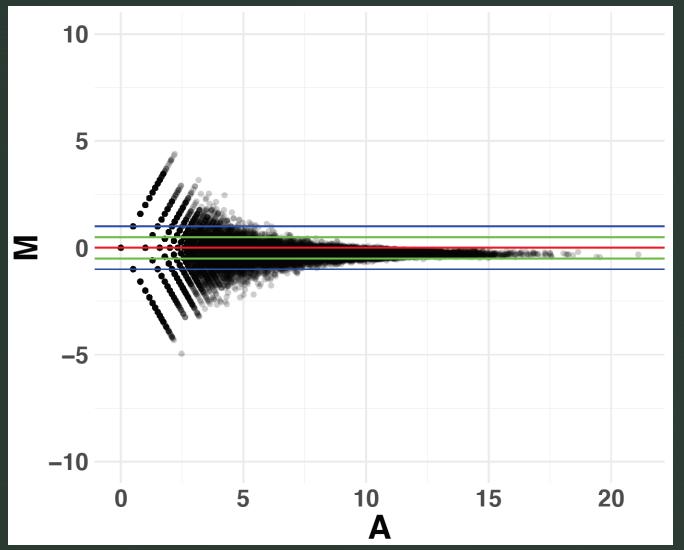
[5] **Noise detection**





mRNAseq.
Preprocessing of count matrices

[5] **Noise detection**



The variability observed at low abundances is a side effect of noise and can propagate into differential expression analysis.

The filtering of noisy entries increases the convergence of DE analyses

Moutsoupoulos et al 2021, https://www.biorxiv.org/content/10.1101/2021.01.17.427026v2

mRNAseq. Normalisation approaches

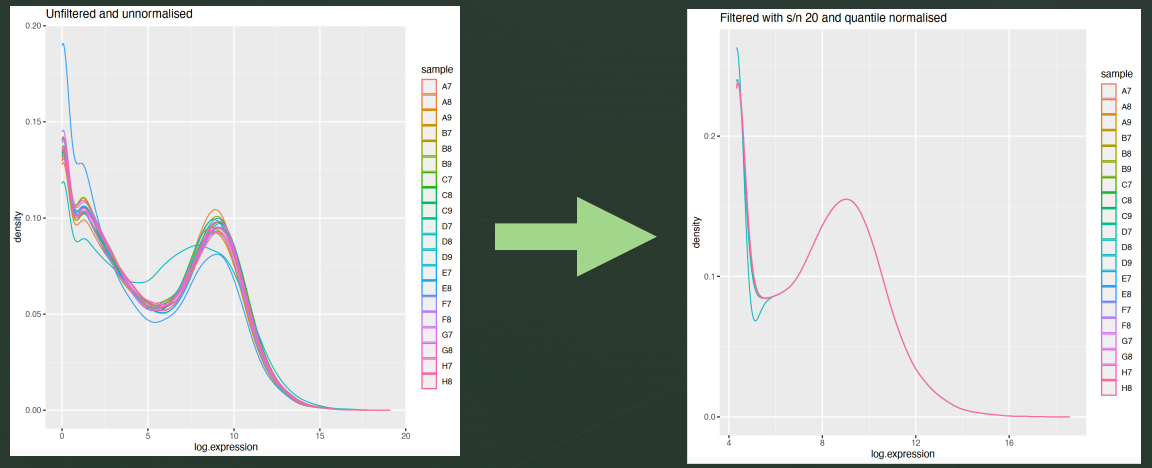
Parametric normalizations
RPM TPM
RPKM FPKM

DESeq
TMM

Non-parametric normalisations
Quantile Normalisation

Single cell normalisations
SCnorm

Through normalisation we obtain a digital measure of the abundance of transcripts. Normalization methods are necessary to remove technical biases in sequenced data such as depth of sequencing (higher sequencing depth produces higher read counts for gene expressed at same level) and gene length (differences in gene length generate unequal reads count for genes expressed at the same level).



mRNAseq.
Normalisation approaches

Parametric normalization
[within sample]
RPM = reads per million
CPM = counts per million

$$\text{RPM or CPM} = \frac{\text{Number of reads mapped to gene} \times 10^6}{\text{Total number of mapped reads}}$$

The nomalisation is performed on the sequencing depth.
The actual distribution of counts is not important

Cons:
[1] only one scaling factor is applied across abundances
as shown, low abundance genes are more variable that high abundance ones

[2] the transcript length is not taken into account.
i.e. genes are comparable across samples, but not within sample.

Based on information from: https://www.reneshbedre.com/blog/expression_units.html

mRNAseq. Normalisation approaches

Parametric normalization
[within sample]

RPKM = reads per kilo base per million mapped reads [single-end sequencing]
FPKM = fragments per kilo base per million mapped reads [paired-end sequencing]

$$\mathrm{RPKM} = \frac{\text{Number of reads mapped to gene} \times 10^3 \times 10^6}{\text{Total number of mapped reads} \times \text{gene length in bp}}$$

The normalization is performed per sequencing depth and gene length.

Median normalisation

The scaling total is no longer the sequencing depth per sample, but the median.
An averaged median across all samples is use as baseline.

Based on information from: https://www.reneshbedre.com/blog/expression_units.html
And Dilies et al

mRNAseq. Normalisation approaches

Parametric normalization – TMM [trimmed mean of means]

[a] TMM is a between-sample normalization method
[b] TMM assumes that most of the genes are not differentially expressed
[c] TMM does not consider gene length or library size for normalization

Steps for TMM:
[a] normalise the gene counts to library size
[b] select one sample as control; calculate the M and A values
[c] trim the M and A values (e.g. on IQRs)
[d] calculate the weighted mean of M
and derive the per-sample normalisation factor

$$\mathrm{M} = log_2 \frac{\text{treated sample count}}{\text{control sample count}}$$

$$\mathrm{A} = \frac{log_2(\text{treated sample count}) + log_2(\text{control sample count})}{2}$$

Smid et al., 2018 proposed a GeTMM (Gene length corrected TMM)
We calculate RPK for each gene from raw read count data which is then corrected by TMM normalization factor

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data
Mark D. Robinson, Davis J. McCarthy,and Gordon K. Smyth; Bioinformatics. 2010 Jan 1; 26(1): 139–140.

Based on information from: https://www.reneshbedre.com/blog/expression_units.html

mRNAseq. Normalisation approaches

Parametric normalization – DeSeq2

DESeq2 normalization method is proposed by Anders and Huber, 2010 and is similar to TMM

DESeq assumes that most of the genes are not differentially expressed

DESeq normalization uses the median of the ratios of observed counts to calculate size factors. The size factor is calculated by first dividing the observed counts for each sample by its geometric mean. Next the median of these ratios for each sample are calculated.

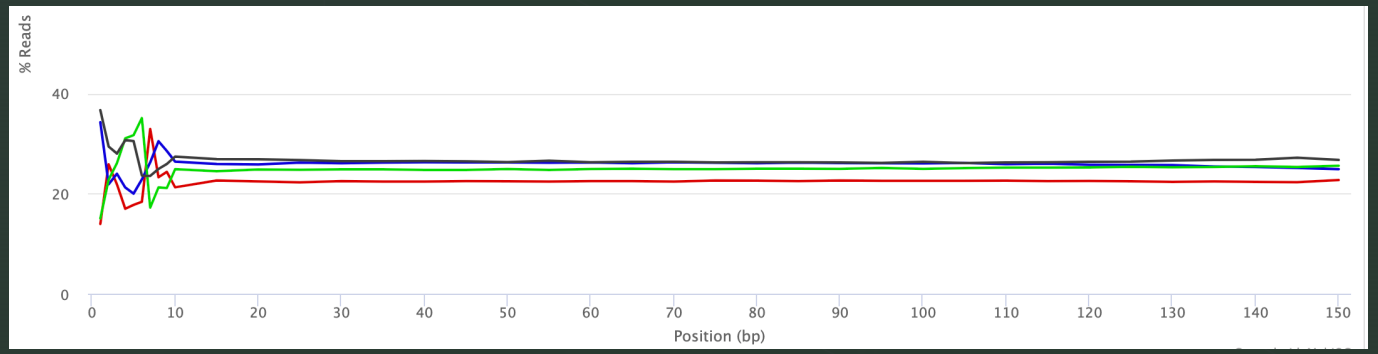
DESeq does not consider gene length for normalization
DESeq performs better for between-samples comparisons

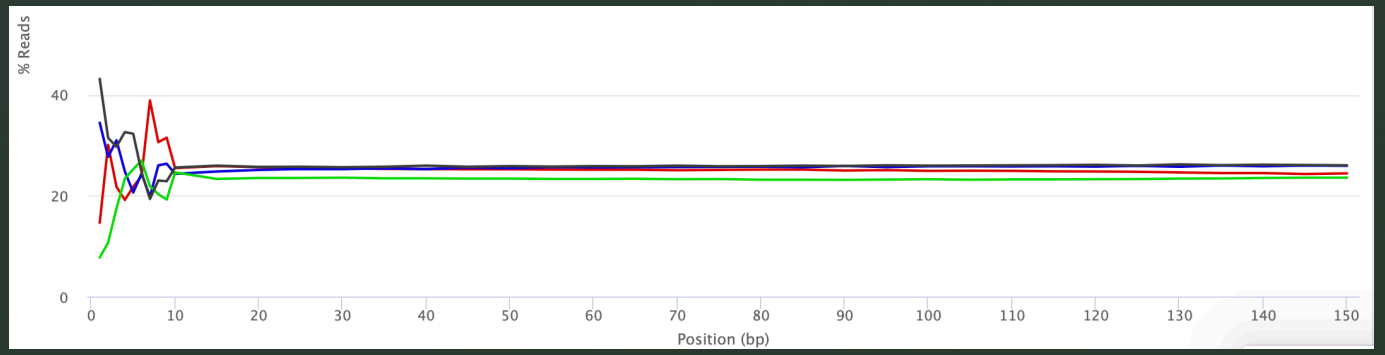
Differential expression analysis for sequence count data
Simon Anders & Wolfgang Huber
Genome Biology volume 11, Article number: R106 (2010)

Based on information from: https://www.reneshbedre.com/blog/expression_units.html

mRNAseq. Normalisation approaches

Is there a drawback to scaling normalisations?





Also, the abundance distribution has the properties of a binomial distribution for medium-high abundances and of a Poisson distribution for low abundances.

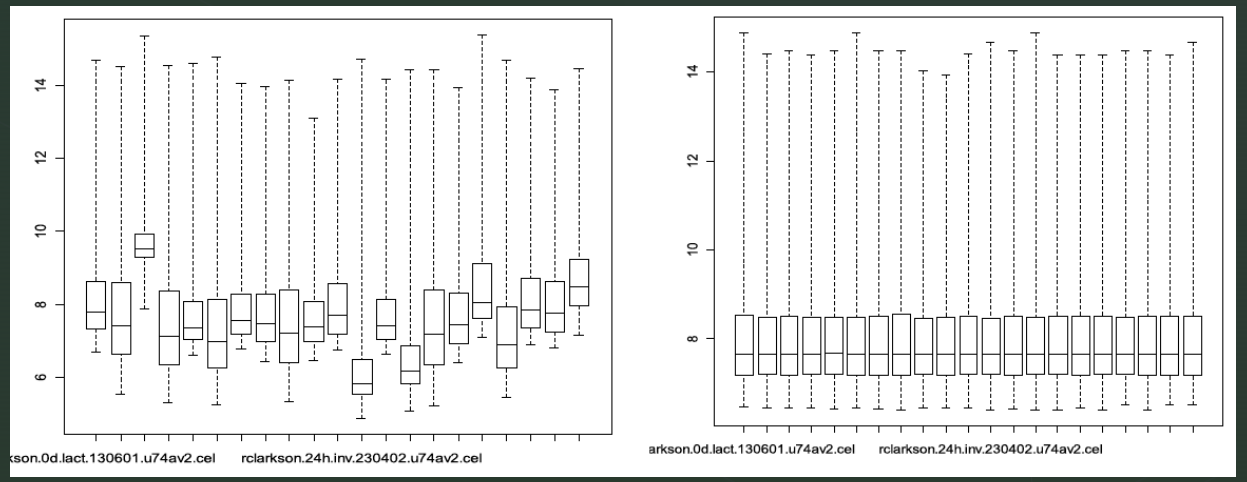
mRNAseq.
Normalisation approaches

Non-Parametric normalisation

Quantile normalisation

First introduced for microarray assays [Bolstad2003]
It is based on balancing the rank distributions across samples.

Upper Quartile
The normalisation is focused
on the upper Quartile only

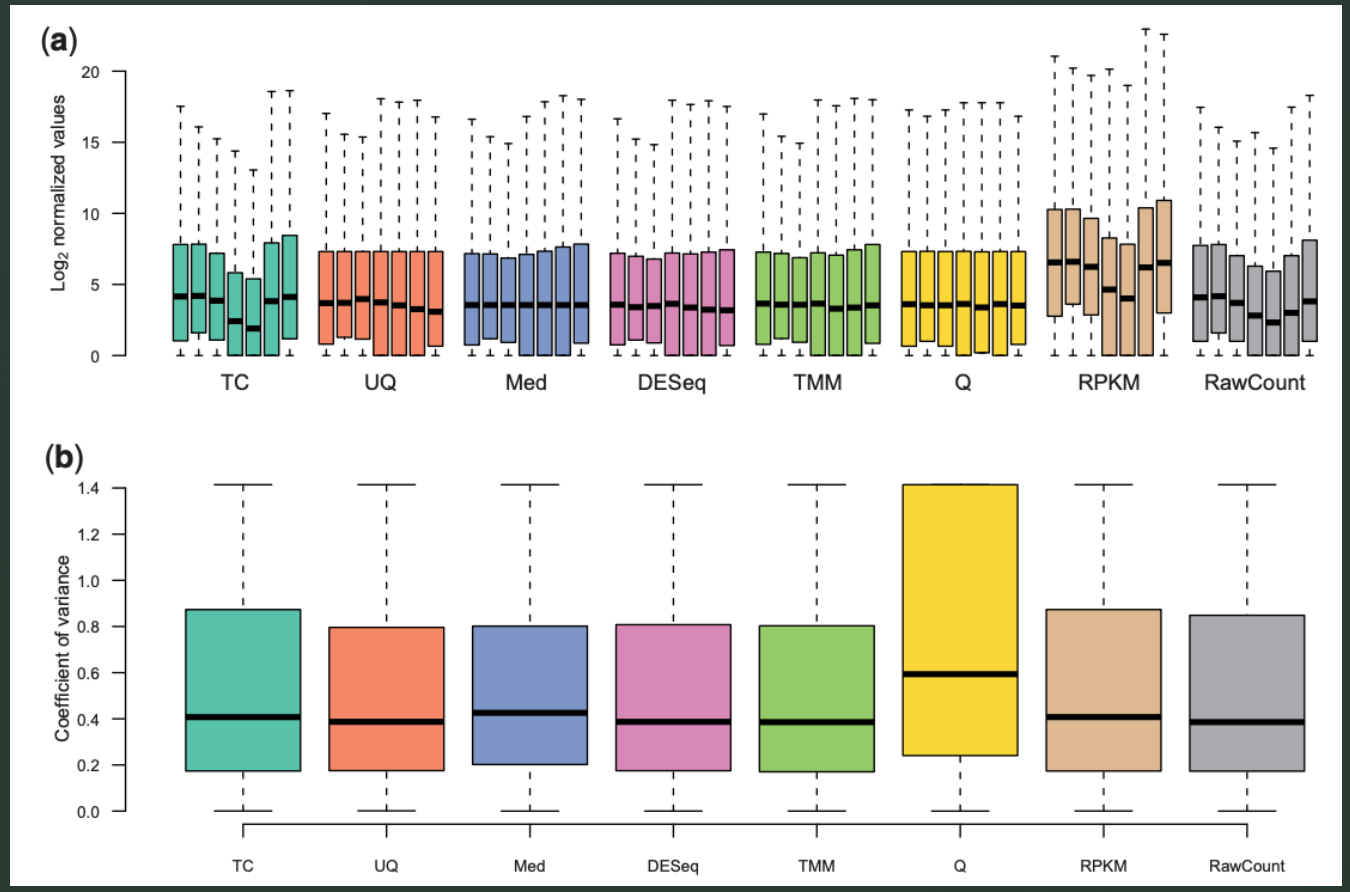


Raw.

Quantile normalised

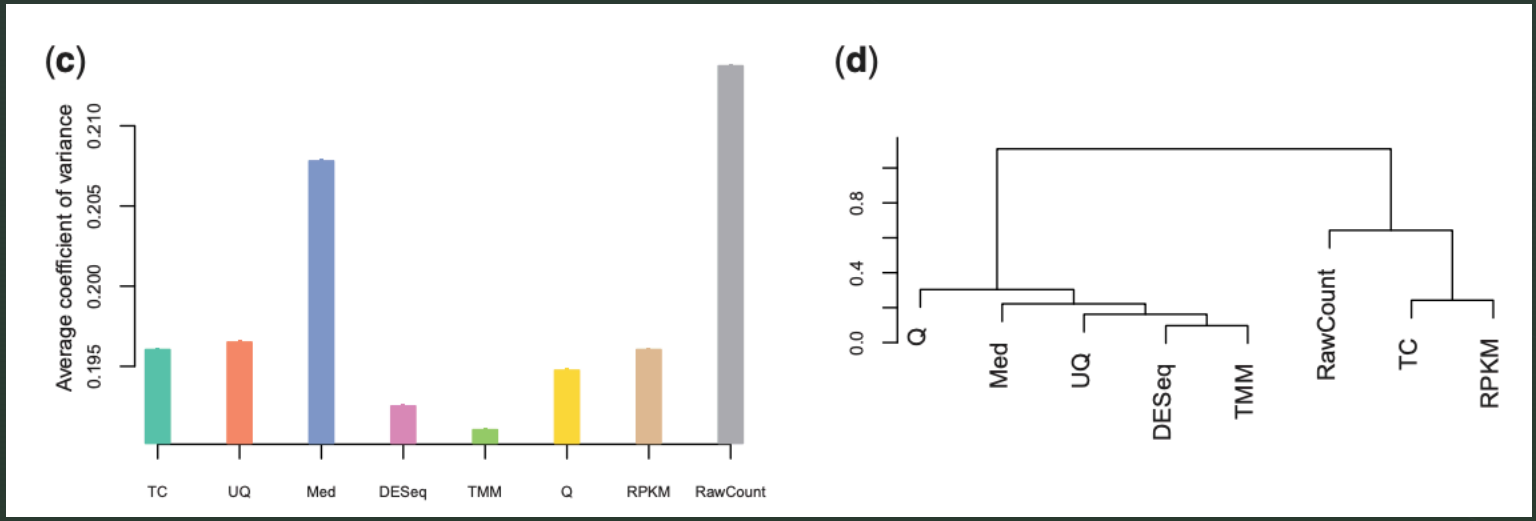
Bolstad, B. M.; Irizarry, R. A.; Astrand, M.; Speed, T. P. (2003); Bioinformatics. 19 (2): 185–193.
"A comparison of normalization methods for high density oligonucleotide array data based on variance and bias".

mRNAseq. Normalisation approaches



Dilies et al 2013 “A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis”

mRNAseq. Normalisation approaches



Dilies et al 2013 “A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis”

mRNAseq. Normalisation approaches

Single cell normalisation

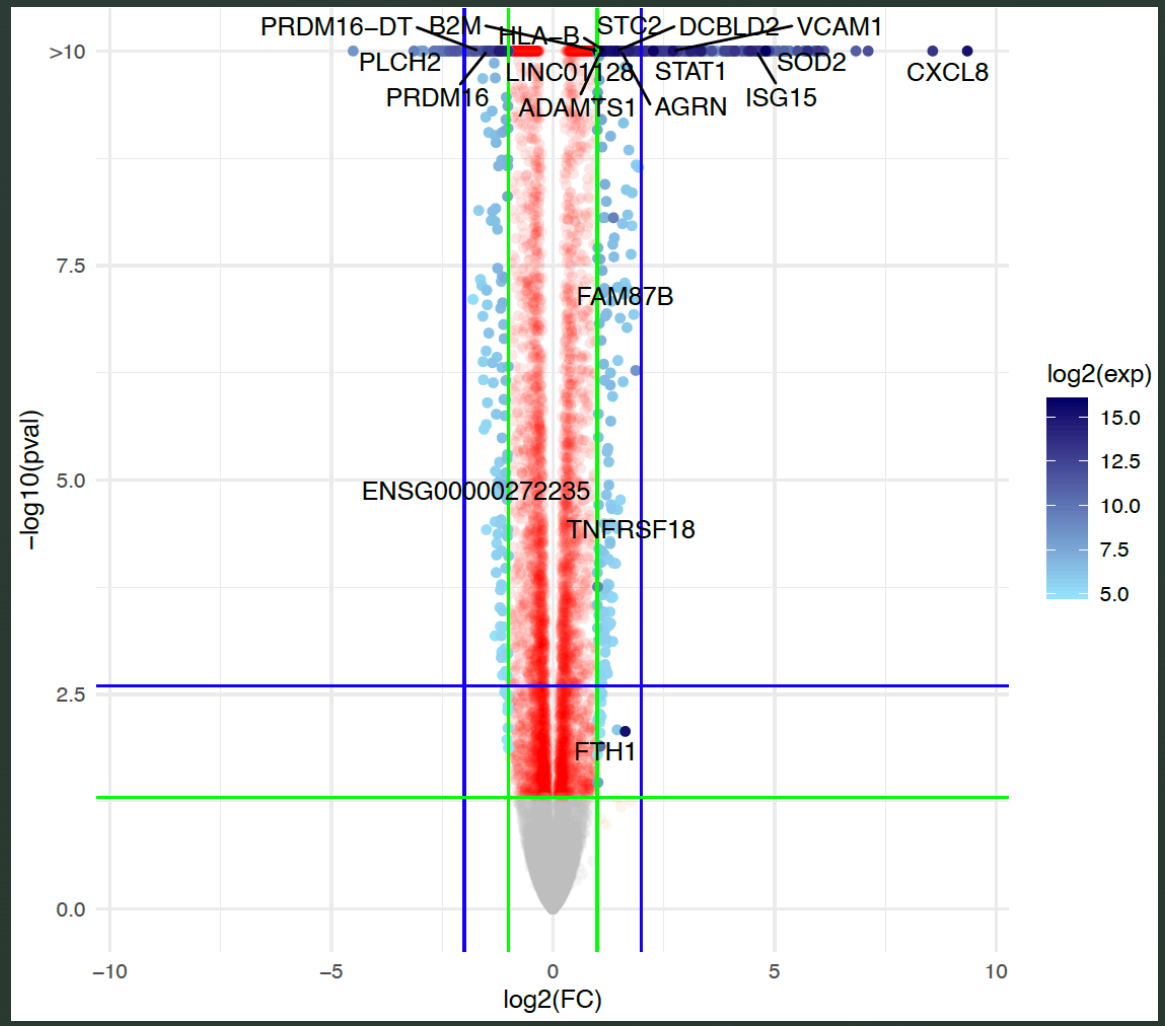
Bulk normalization are biased for scRNA-seq usage due to abundance of non-zero expression counts,& variable count-depth relationship (dependence of gene expression on seq depth)

Scnorm = robust and accurate between-sample normalization for scRNA-seq

Input: estimates of expression counts
[1] Genes with low expression counts are filtered out
[2] estimate the count-depth relationship using quantile regression
[3] Cluster genes into groups with similar count-depth relationship
[4] A scale factor is calculated for each group and used estimating normalized expression

Bacher R, Chu LF, et al. SCnorm: robust normalization of single-cell RNA-seq data. Nature methods. 2017 Jun;14(6):584.

mRNAseq.
Differential expression analysis



The DE calls are summarised in volcano plots; on the x-axis we represent the FC, on the y-axis –log10(adj p-value).

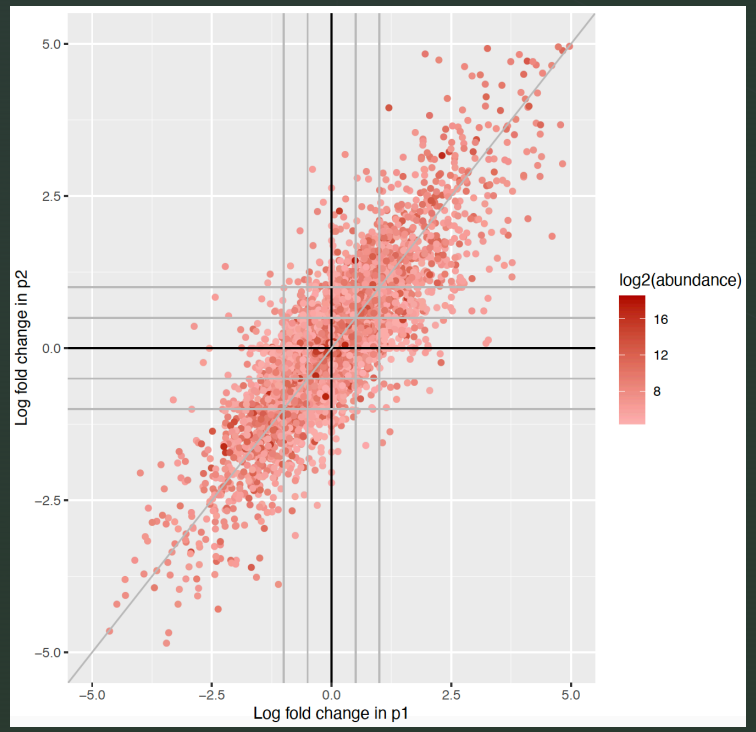
Questions that need to be addressed:

Stability and Robustness:
Do different methods produce similar/ convergent results?

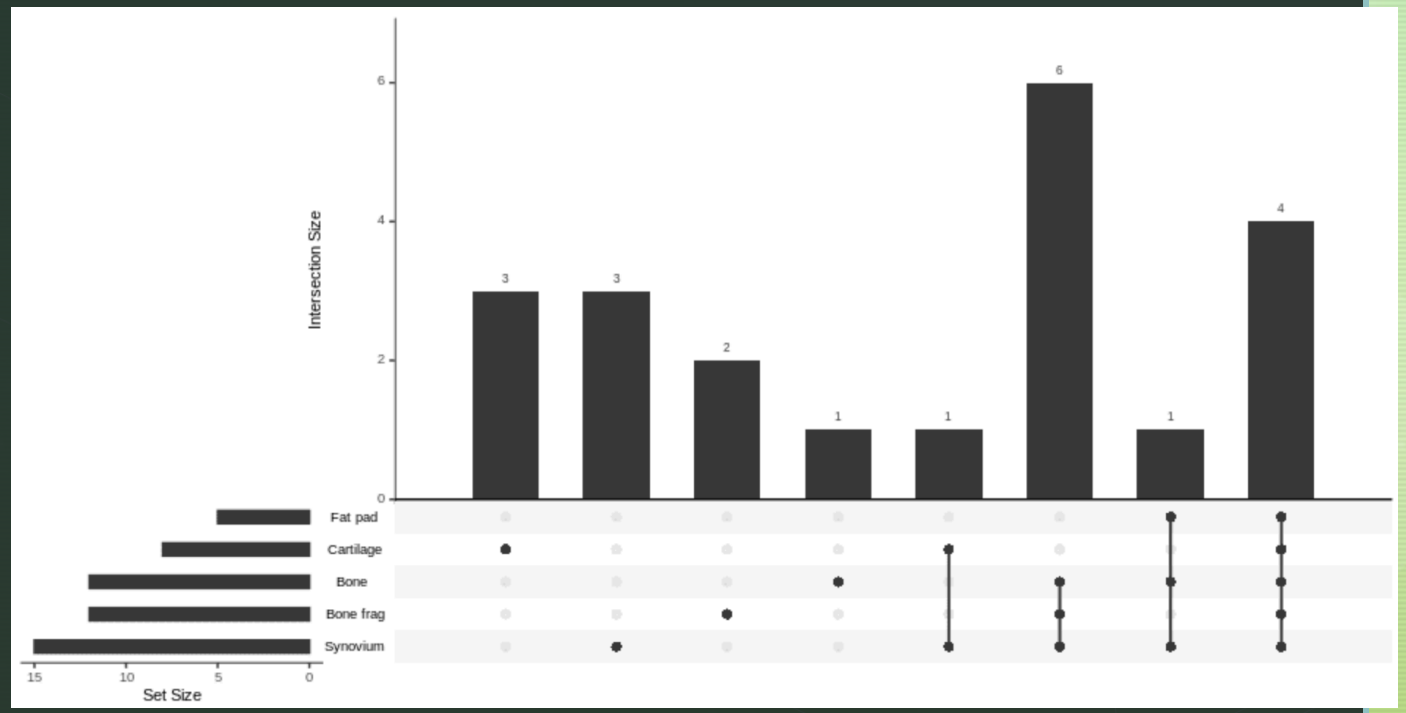
What do the results mean from a biological perspective?
[Enrichment analysis lecture]

mRNAseq.
Differential expression analysis

To summarise differences across multiple comparisons:

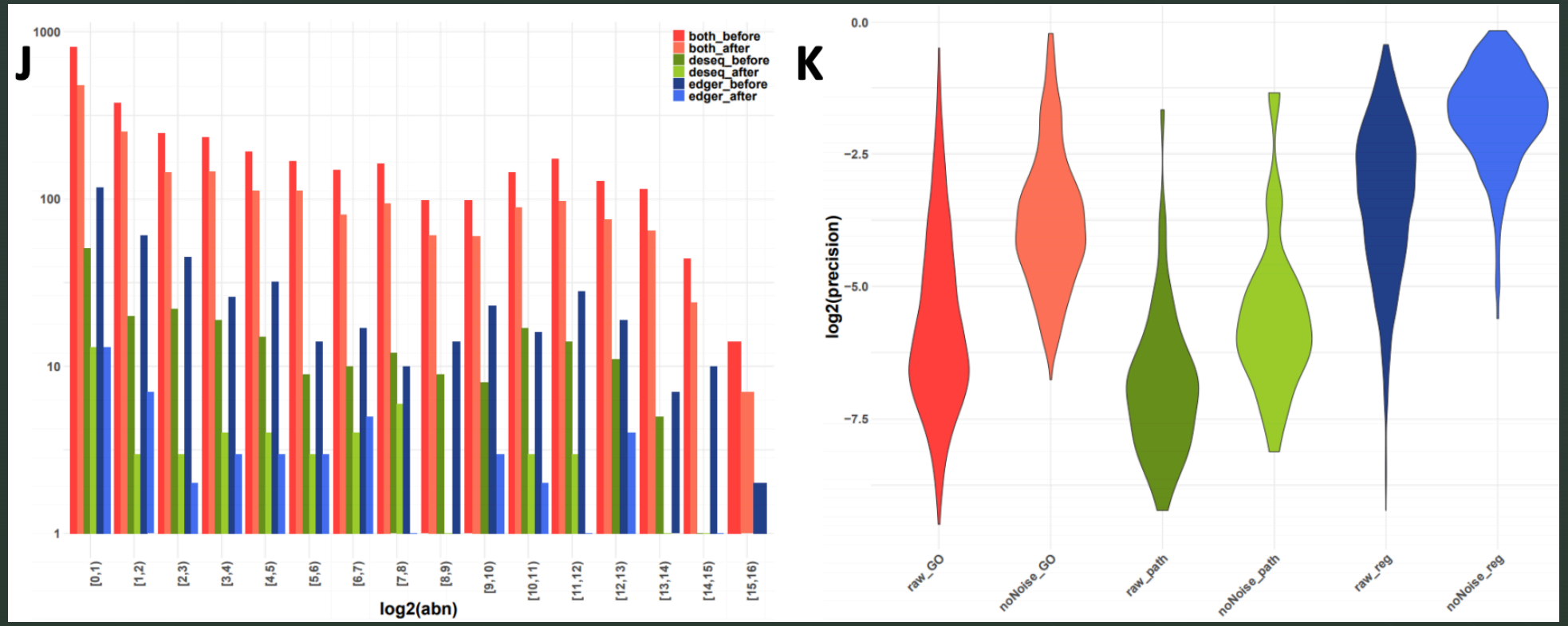


Cross-plots: on each axis we represent the FC in one comparison



UpSet plots – quantitative multiple Venn diagrams

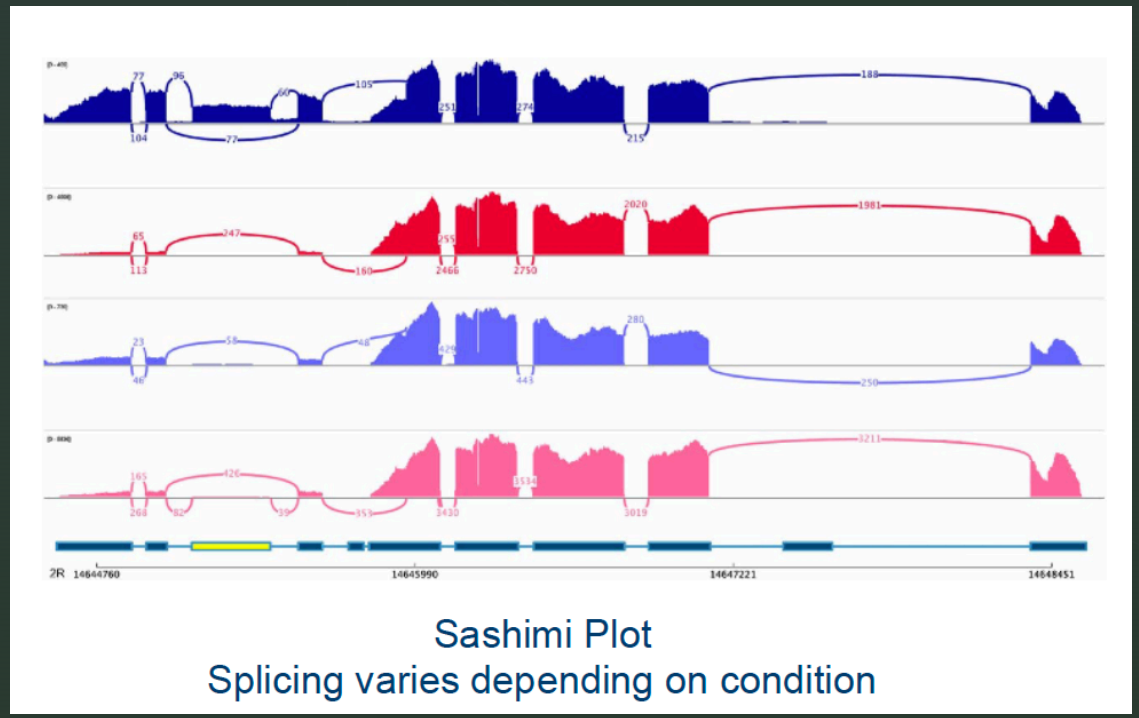
mRNAseq.
Differential expression analysis



Post noise-filtering the DE genes called using EdgeR and DEseq2 converge. (left panel) The precision (intersection size divided by the query size) for the results of the enrichment analysis performed on the DE genes also suggests convergence (right panel)

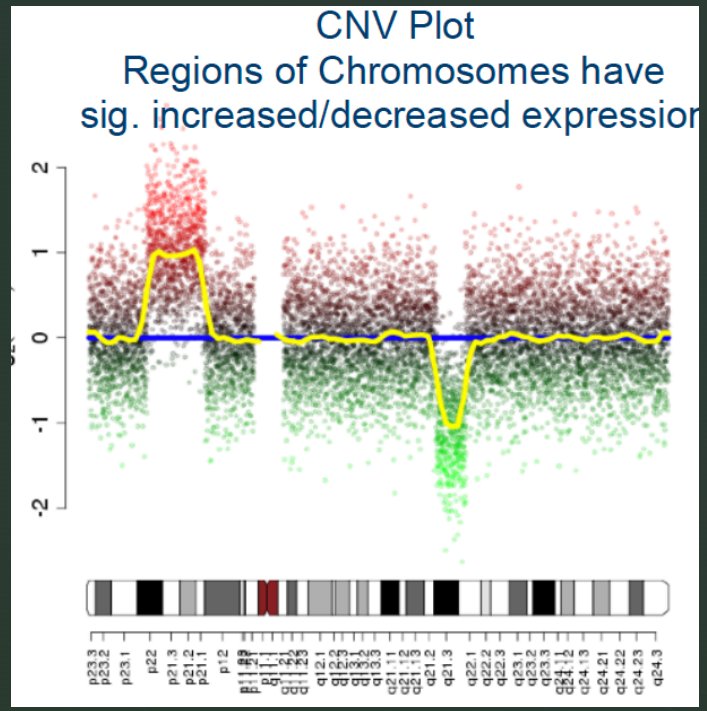
Examples of other quantitative analyses

An aim of RNAseq is the detection of differentially expressed genes.



Sashimi Plot
Splicing varies depending on condition

Genes affected by alternative splicing can also be found.



Detection of regions with copy number variation (CNV)

Slide by dr A Enright

Examples of other quantitative analyses

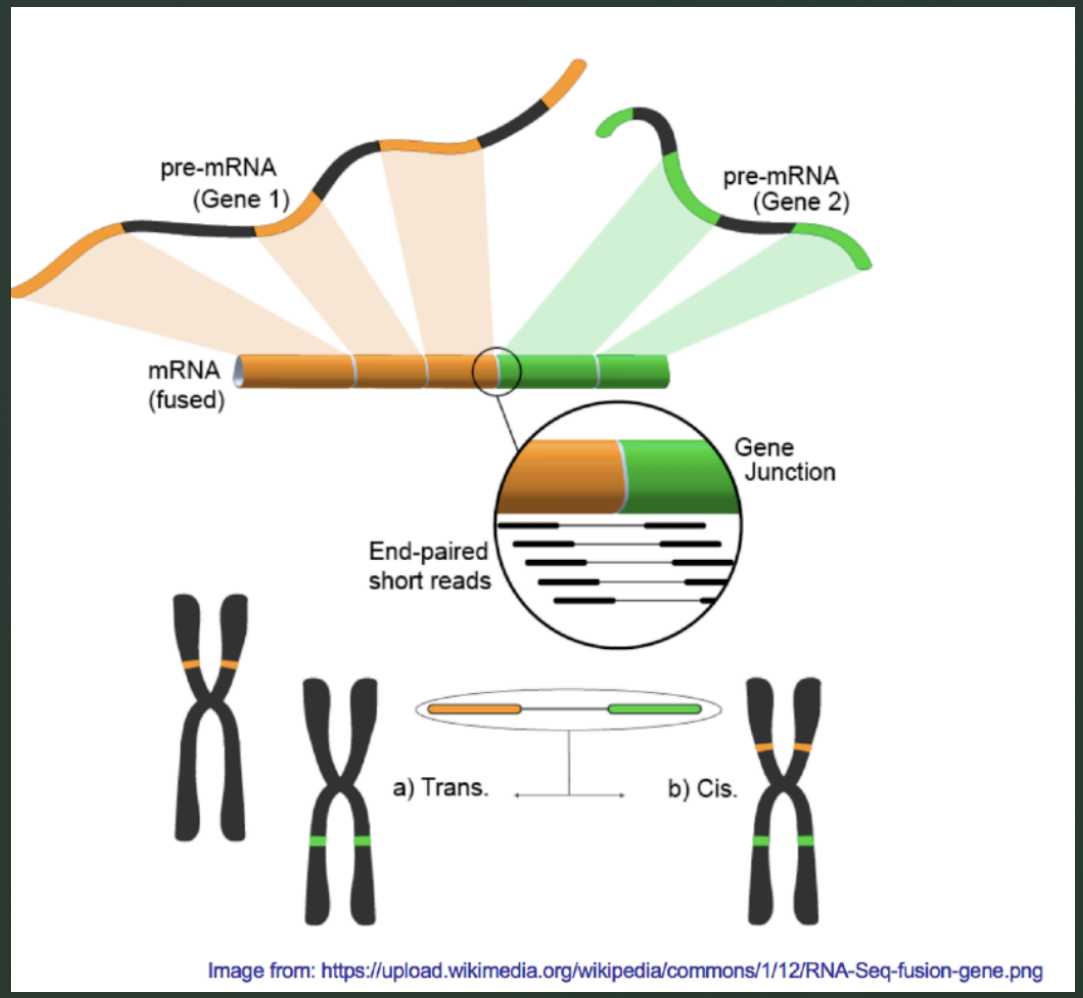
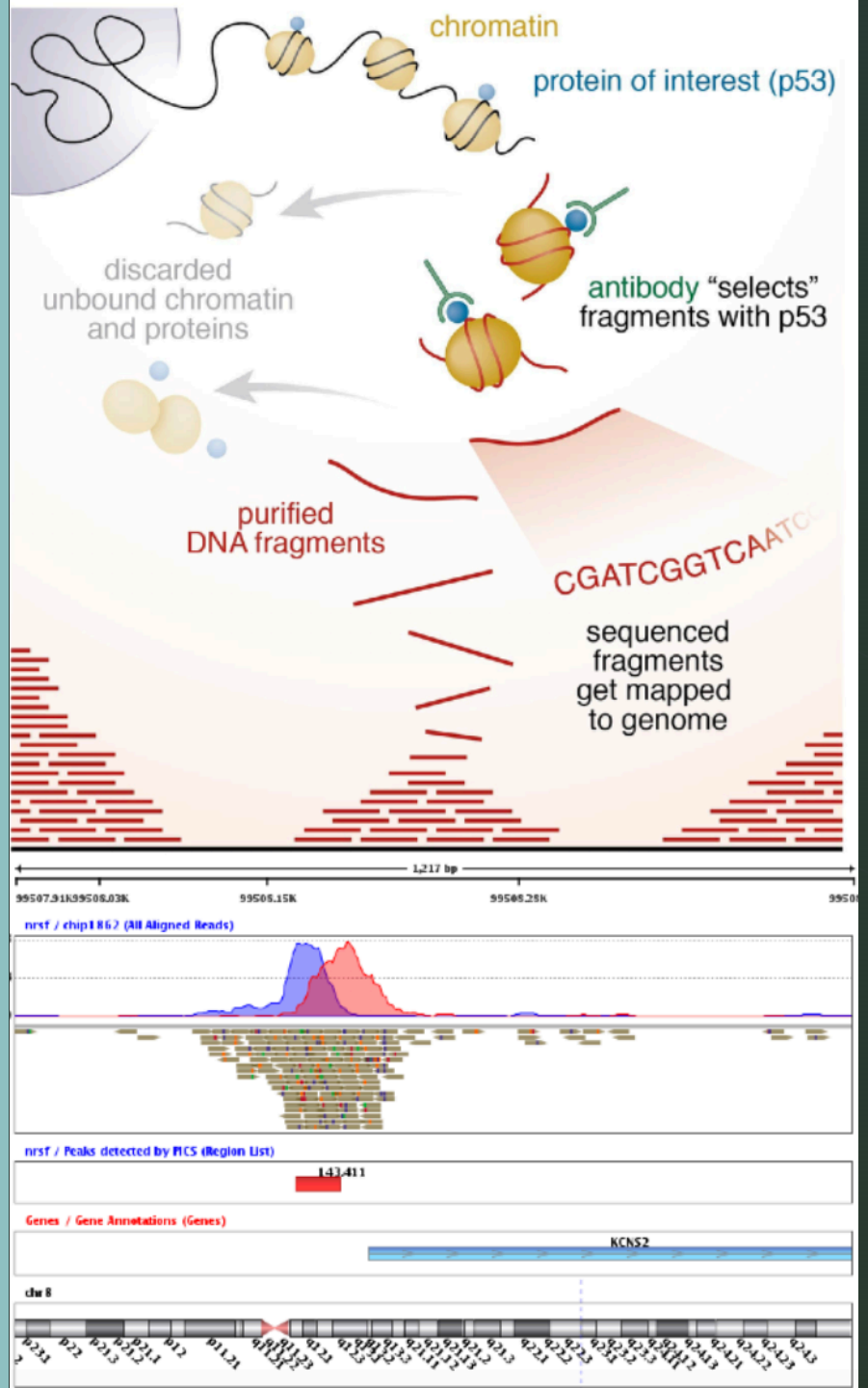


Image from: https://upload.wikimedia.org/wikipedia/commons/1/12/RNA-Seq-fusion-gene.png

Example of gene fusion.

Slide by dr A Enright



Examples of other quantitative analyses

Other high throughput assays include:
[1] using antibodies to captures genes of interest

[2] crosslink antibodies to DNA/ RNA

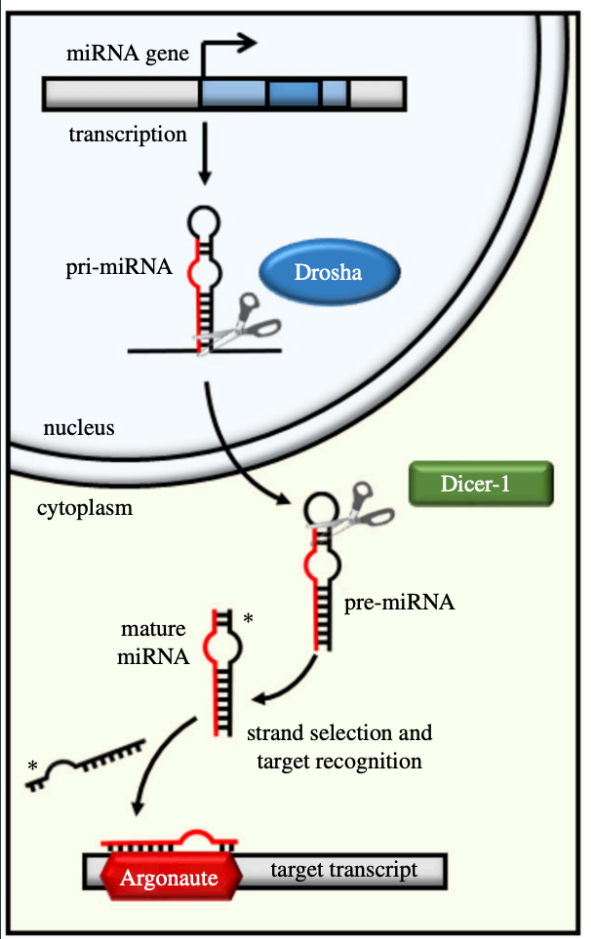
[3] wash and sequence approaches
- ChIPseq – chromatin bound proteins
- CLIPseq – protein:RNA
- RIPseq – RNA binding proteins
- meRIPseq - methylation sites

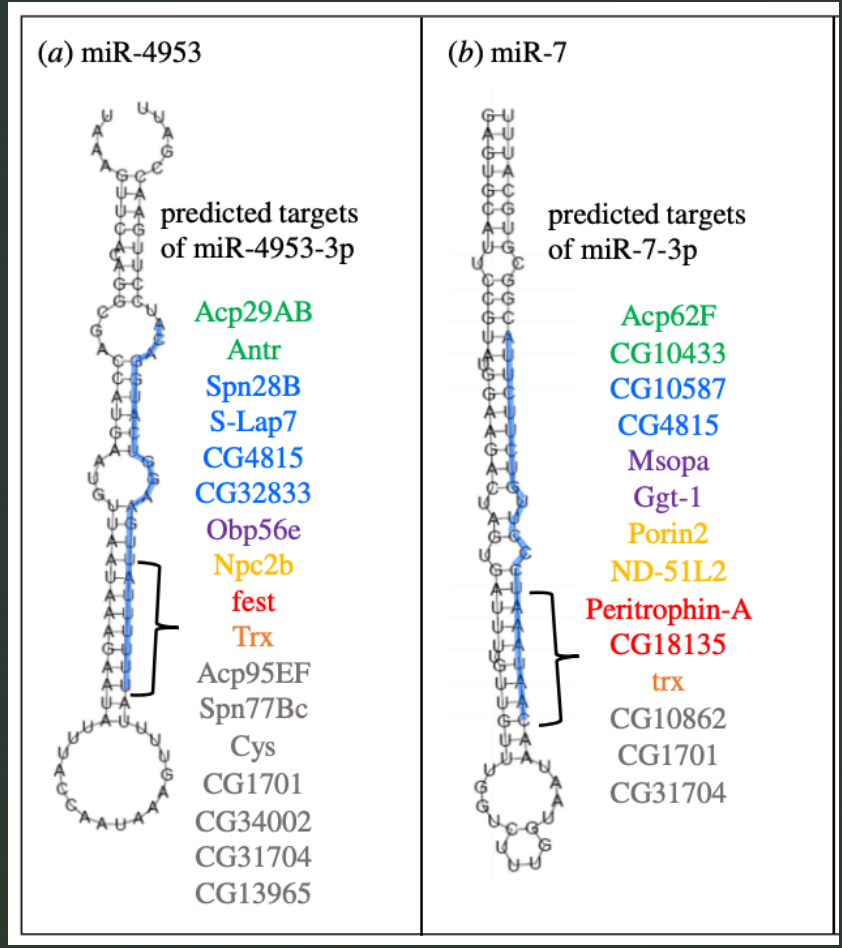
[4] to study DNA
- BSseq – bisulphite sequening
- ATACseq – chromatin accessibility

[5] multi-omics assays at single cell level

Slide by dr A Enright

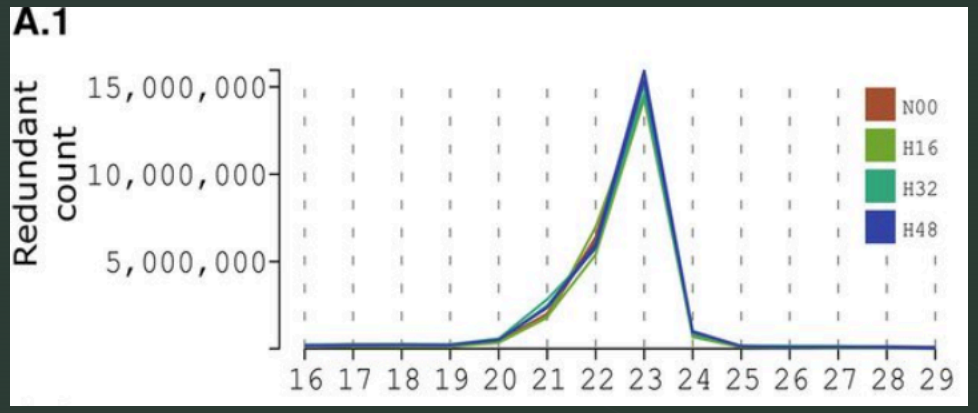
Non-coding RNAs.
Small Non-coding RNAseq



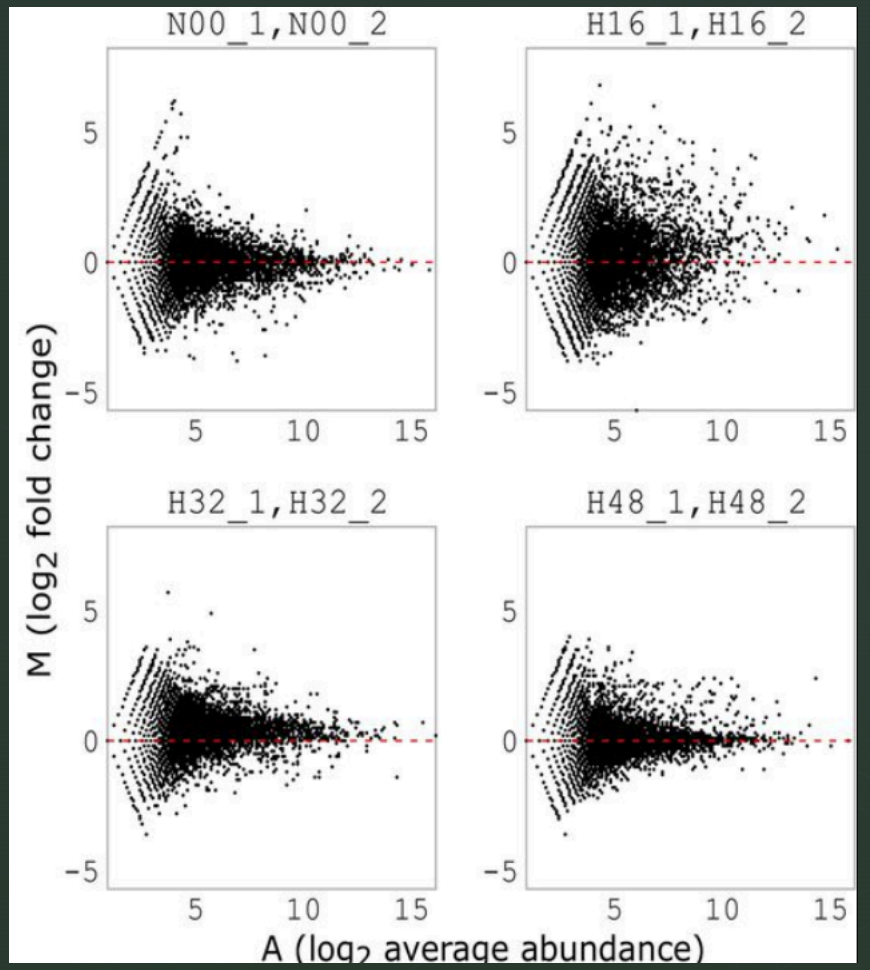


Control of seminal fluid protein expression via regulatory hubs in Drosophila melanogaster
Mohorianu et al 2018, Proc Royal Soc B.;https://doi.org/10.1098/rspb.2018.1681

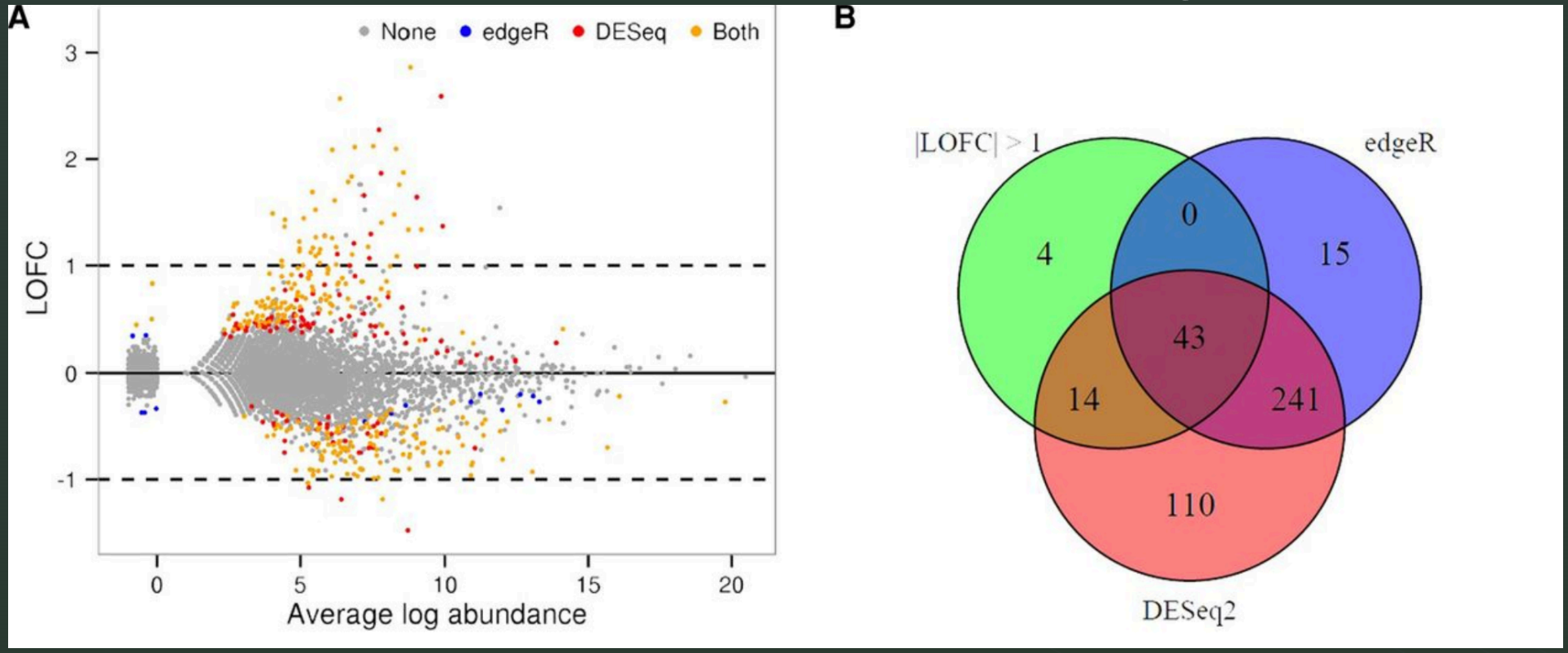
Non-coding RNAs. Small Non-coding RNAseq



The main difference between sRNAseq and mRNAseq is that the reads (small RNAs) are not summarised to generate the expression of a transcript, but considered individually.



Non-coding RNAs.
Small Non-coding RNAseq



Similar questions, related to differentially expressed sRNAs are asked.

Comprehensive processing of high-throughput small RNA sequencing data including quality checking, normalization, and differential expression analysis using the UEA sRNA Workbench
M Beckers et al RNA 23 (6), 823-835