RNA-Seq Data Analysis

14th & 15th June 2021

Helen Lockstone

Head of Bioinformatics and Training

Bioinformatics Core, Wellcome Centre for Human Genetics

Assistant Instructors: Irina Chelysheva and Srinivasa Rao

RNA-Seq: Data Normalisation

wellcome centre human genetics

Raw gene counts are influenced by:

- Sequencing depth - varies by sample
- Gene length
- Amplification/sequencing biases – GC content
- Positional biases – preferential locations for RNA fragmentation
- Also by sample composition, which is less obvious

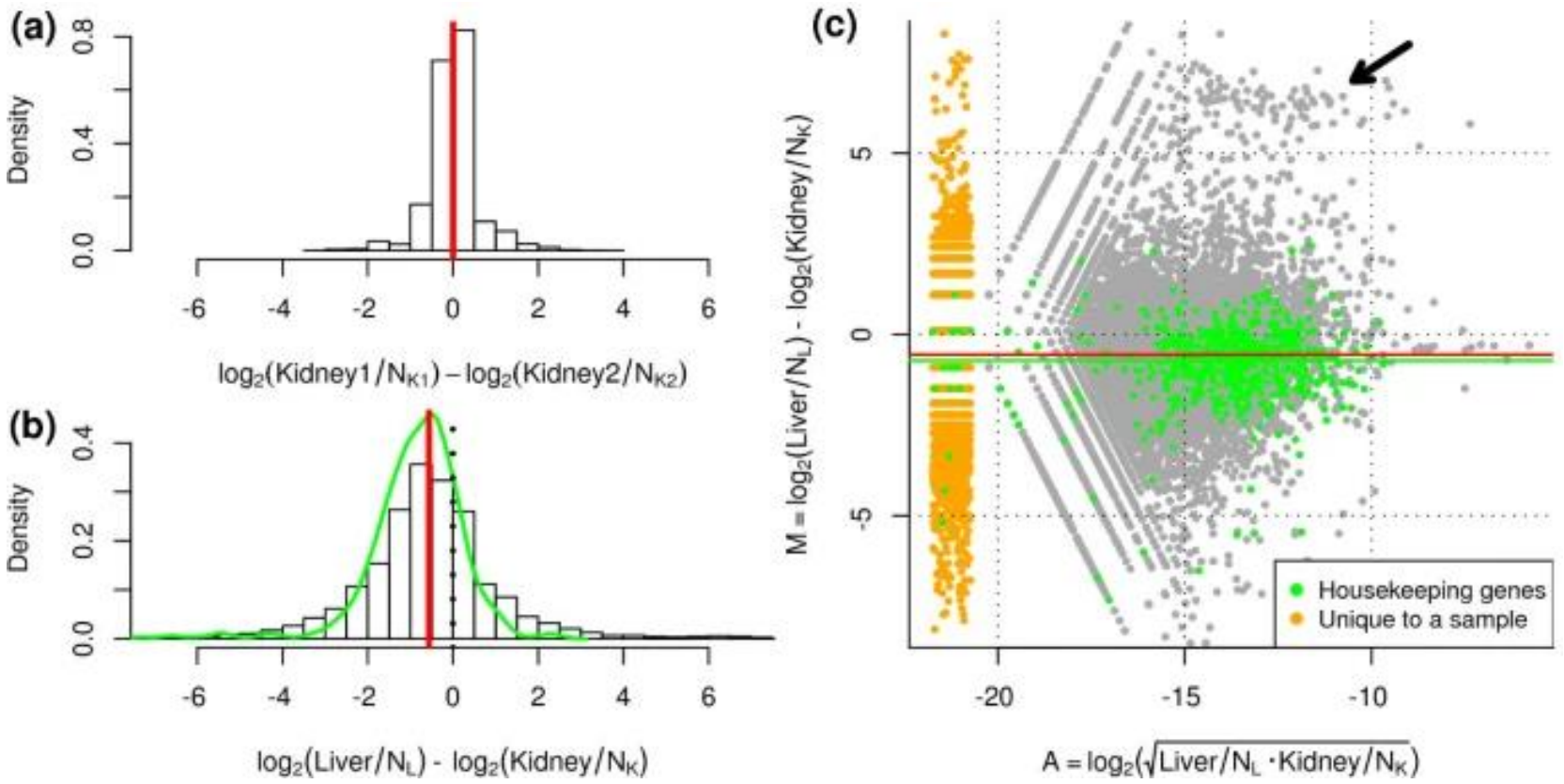
Count summaries accounting for various factors

- RPKM (reads per kilobase per million reads
- FPKM (fragments per kilobase per million reads)
- CPM (counts per million reads)
- TPM (transcripts per million reads)

RNA-Seq: Data Normalisation

- edgeR's normalisation procedure adjusts for library size (sequencing depth variations) and sample composition differences
- Other factors like gene length and GC content are shared between genes, although some sample-specific effect has been detected, particularly for GC-content
- TMM or trimmed mean of M-values (fold changes or FCs)
- Finds a set of scaling factors that minimizes the logFCs between the samples for most genes
- Assumes majority of genes are not likely to be differentially expressed, which is reasonable for most studies

RNA-Seq: Normalisation



RNA-Seq: Small sample sizes [1/4]

- A typical RNA-Seq experiment has very small sample sizes, maybe 3-6 replicates per condition
- Small samples generally make it harder to infer differential expression, because there is more uncertainty in the estimation of the mean and variance, which are in turn used to compute the test statistic and p-value.
- Unreliable estimates of the sample variance due to small sample size can be problematic – producing false positives when the variance is underestimated and reducing power if overestimated
- The limma and edgeR packages for analysing gene expression use a clever strategy to limit this issue

RNA-Seq: Small sample sizes [2/4]

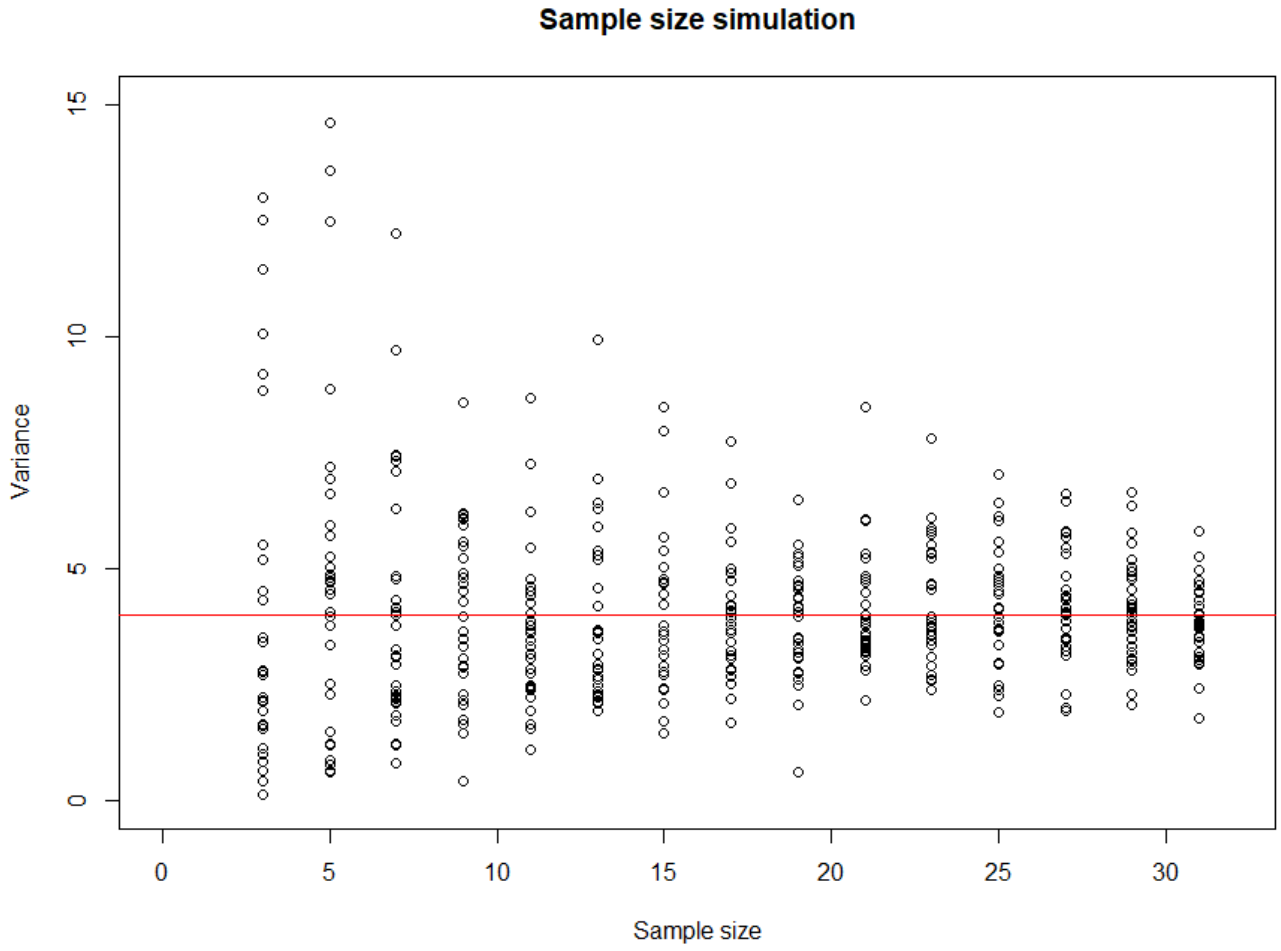
- We'll run a quick simulation first to understand this phenomenon better (taken from Statistics: An Introduction using R by Michael J Crawley)

- In R, enter the command

source("https://www.well.ox.ac.uk/bioinformatics/training/RNASeq_materials/scripts/sample_size_exercise.R")

- This should produce a plot similar to the one on the next slide

RNA-Seq: Small sample sizes [3/4]



RNA-Seq: Small sample sizes [4/4]

- Our simulation clearly shows how the variance estimates are poorest when n<10
- Built into limma and edgeR is a procedure to ‘borrow information between genes’ and improve the robustness of the variance estimates
- Individual gene variance estimates are ‘squeezed’ towards a common value according to how other genes with similar expression levels behave
- A very low variance estimate will be inflated, reducing the chance that a very small difference in expression incorrectly generates a significant p-value

wellcome centre human genetics

Practical Session II
(to finish)

Data Preprocessing

Practical Tutorial: Differential Expression

We'll finish the last few steps of the second tutorial (note you may need to quickly run the code chunks again as we did not save files yesterday (if you happened to save your R session, you should be able to re-open at the point you left off).

https://www.well.ox.ac.uk/bioinformatics/training/RNASeq_June2021/Practical_tutorial_2/

wellcome centre human genetics

Practical Session III

Differential Expression Analysis

Practical Tutorial: Differential Expression

The final tutorial for the workshop can be found here:

https://www.well.ox.ac.uk/bioinformatics/training/RNASeq_June2021/Practical_tutorial_3/

Further resources

Some useful resources on RNA-Seq data analysis are collated here:

https://www.well.ox.ac.uk/bioinformatics/training/RNASeq_materials/resources

Thank you for attending, we hope you found it useful!