RNA-Seq Data Analysis

Helen Lockstone and Ben Wright
Bioinformatics Core, Wellcome Centre for Human Genetics

23rd November 2020

RNA-Seq: Workshop aims

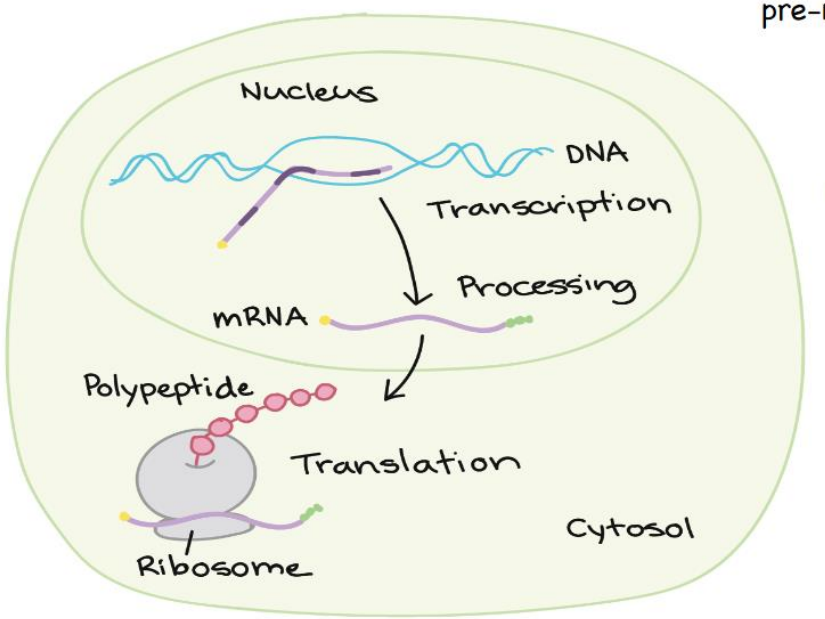
- Hands-on practice with the key steps in analysing RNA-Seq data

- Cover some key concepts, examples and use a real dataset to find differentially expressed genes

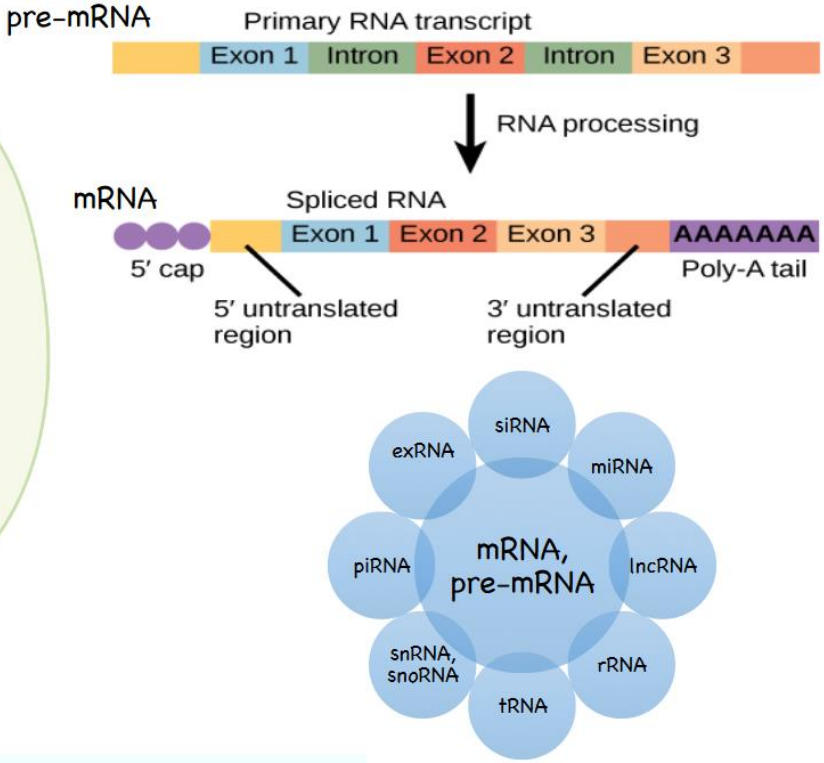
- Highlight crucial steps for performing analysis correctly

Course materials at
https://www.well.ox.ac.uk/bioinformatics/training/RNASeq_Nov2020/

Profiling the transcriptome



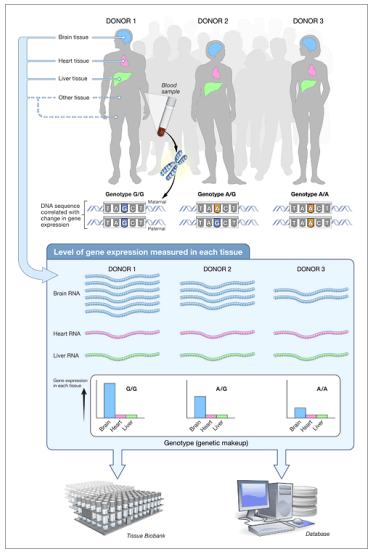
- Image credit Khan Academy Open Courses



Gene expression profiling techniques provide valuable insight into complex biological systems, albeit a snapshot – highly dynamic and tightly regulated process (splicing, gene methylation, RNA stability/degradation, miRNA regulation etc.)

Wide utility of gene expression profiling

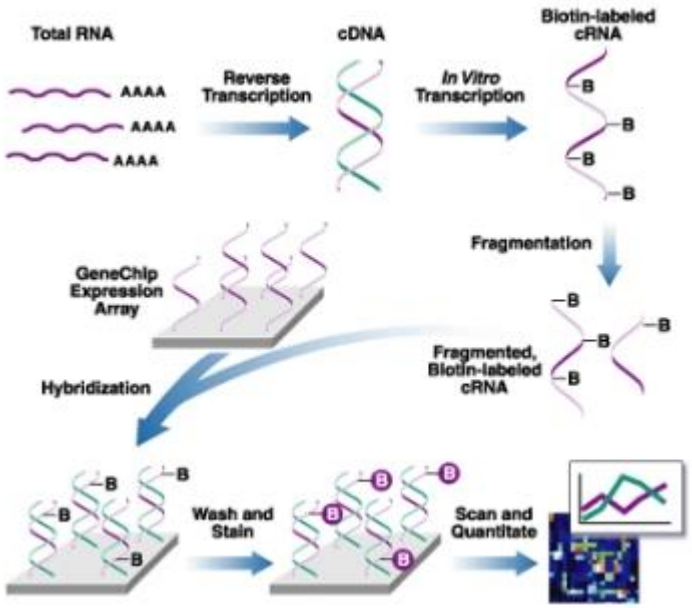
- ENCODE
- Allen brain atlas
- Genotype-Tissue Expression Project (GTEx)
- TCGA



- Public repositories containing tens of thousands of datasets
  - Gene Expression Omnibus (GEO) http://www.ncbi.nlm.nih.gov/geo/
  - ArrayExpress https://www.ebi.ac.uk/arrayexpress/
  - Sequence Read Archive (SRA) http://www.ncbi.nlm.nih.gov/sra

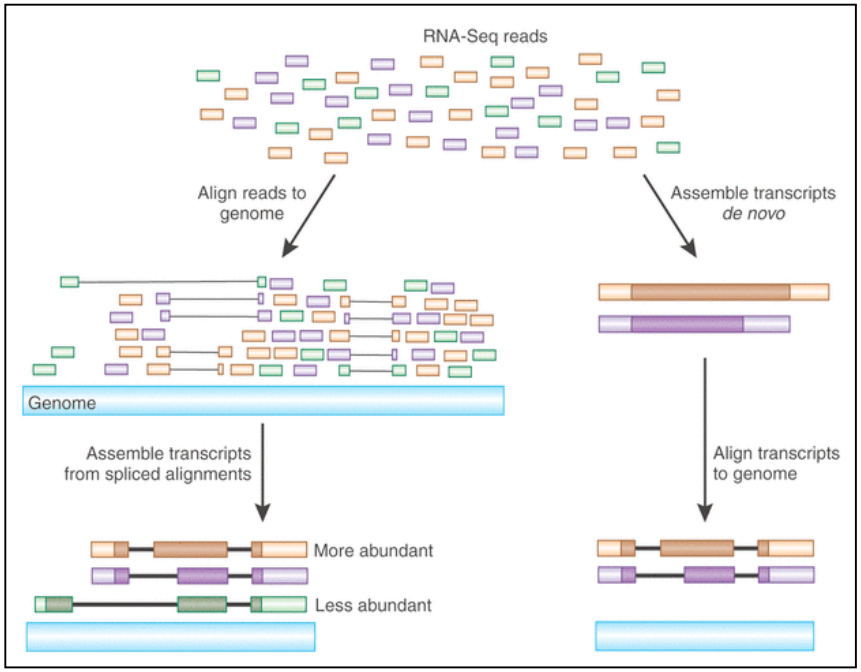
High throughput techniques for gene expression

Microarrays



Complementary hybridisation
early 1990s onwards

RNA-Seq



Next-generation sequencing
2007 onwards

RNA-Seq: Advantages and caveats

- No prior knowledge of gene sequences needed in RNA-Seq; array probes have to be designed
- Better suited for transcript discovery, isoform characterisation and refining existing annotations e.g. length of 5’ or 3’ UTR, uncharacterised exons
- Costs quite comparable between technologies now (RNA-Seq used to be more expensive)
- Even though RNA-Seq produces count data, these do not represent individual transcripts but many fragments of transcripts
- Short-read platforms like Illumina are not capable of sequencing entire transcripts with a single read, so isoform reconstruction and splicing characterisation are very challenging tasks
- Myth: RNA-Seq can detect low expressed genes better than arrays
  - In most experiments, up to half of all genes are not sequenced at all or generate just a handful of reads
  - Additional sequencing will still tend to be from more highly expressed genes, so lower end extremely hard to interrogate
- Caveat: what you sequence in an RNA-Seq library influences your data for all genes – very inter-dependent in a way that arrays are not

Gene expression results – what next?

- Whichever technology used, inherently limited to descriptive results no matter how well experiment performed or data analysed.
- Produce large amounts of information; subjective interpretation, can be mined in different ways requiring human decision-making to take the results further.
- What genes/pathways to focus follow-up experiments on? Different researchers could easily identify quite different themes from the same results.
- Best used as starting point for further work - following up hypotheses from gene expression data to uncover mechanistic/causal effects can produce elegant studies.

RNA-Seq: Library preparation

wellcome centre human genetics

RNA-Seq library preparation protocol	Pros/Strengths	Cons/Weaknesses	Requirements
PolyA enrichment	Selects mature mRNAs and full transcript covered	Requires good quality samples	>=200ng input
Ribo-depletion	Retains wider range of RNA species Full length of transcripts covered	Higher sequencing depth required, increasing cost	
3' mRNA	Very cost-effective; works well for variable quality/quantity input material	Can't characterise isoforms/splicing	
SMARTer low input	Good when only small amounts of RNA can be obtained	More expensive	10ng input
Small RNAs (miRNA)	Additional layer of informative data	Needs separate library prep so an extra cost	
Single cell	High resolution data, novel insights	Expensive, data analysis considerations	pg

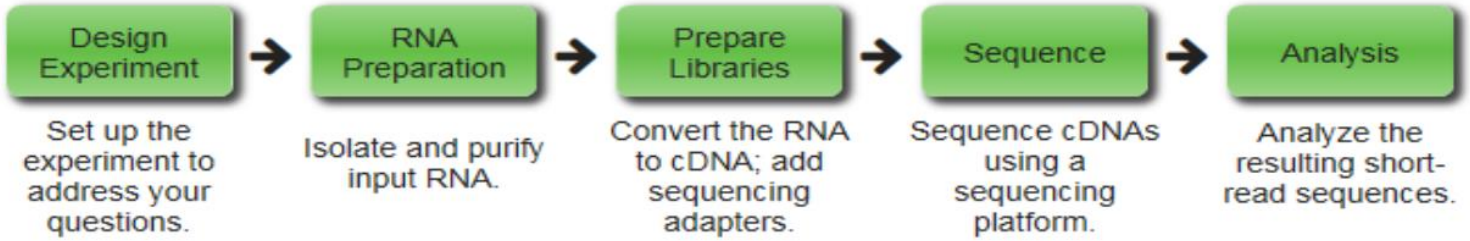
RNA-Seq: Sequencing Depth

- Number of reads required per sample depends on selected protocol and experimental questions

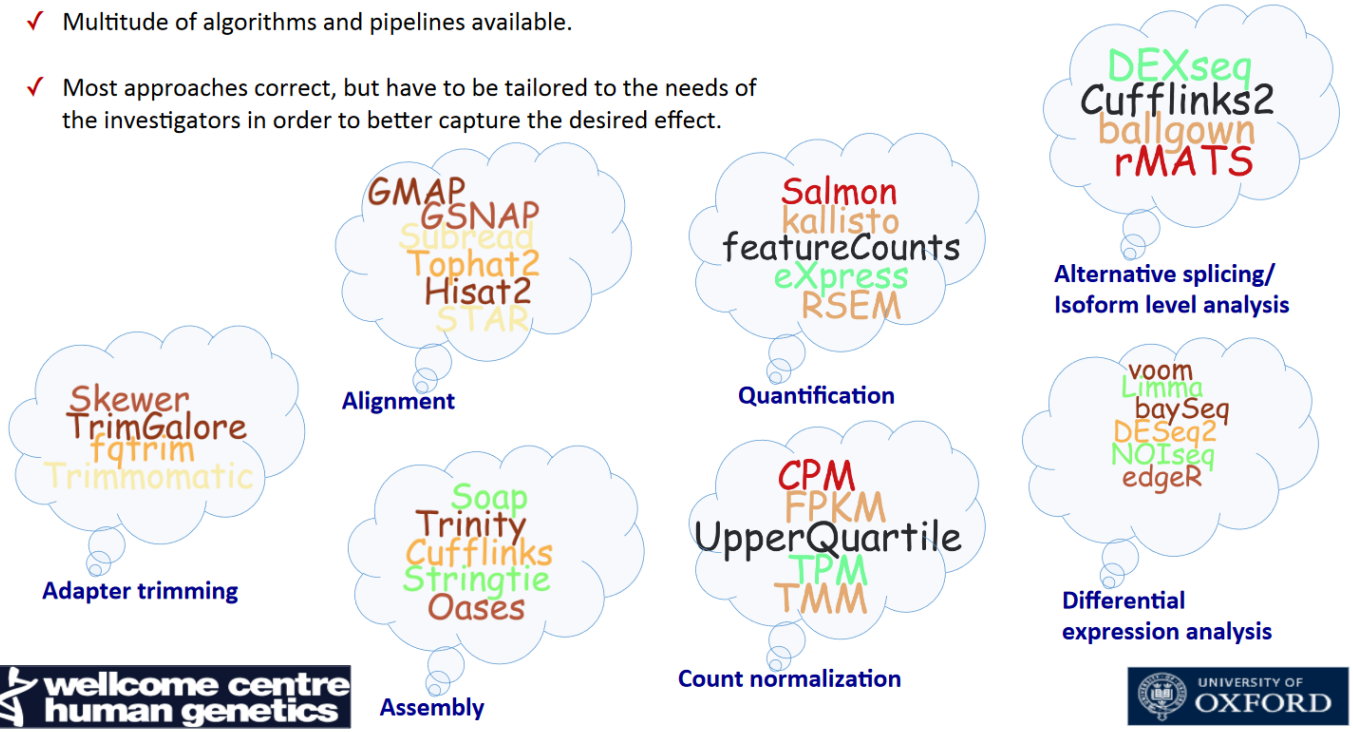
- HiSeq4000 – one lane = 250 million reads
- Multiplexing e.g. 10-plex human samples gives ~25m reads for each, plenty for quantifying gene expression (for those genes that are expressed)

- Higher depth required in some situations e.g. for splicing analysis, certain library prep methods (ribo-depletion)

RNA-Seq: Pipelines



- ✓ Multitude of algorithms and pipelines available.
- ✓ Most approaches correct, but have to be tailored to the needs of the investigators in order to better capture the desired effect.



Slide courtesy of Eshita Sharma

RNA-Seq: Choosing Tools

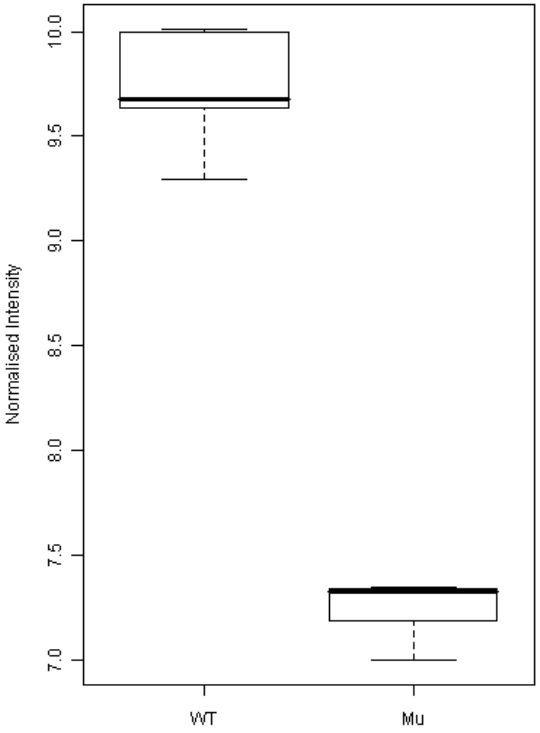
- Any sensible pipeline will produce reasonable results
- **DON'T** worry about trying to get to a definitive answer…..
- Instead, make sure appropriate tools are selected for the task and then carefully used:
  - Is the tool suitable for my question and data I have?
  - Have I understood how it works and how the parameter settings affect its behaviour?
  - Have I provided the right input and made sure the output is sensible?
  - Have I checked my R code for mistakes or unintended behaviour?

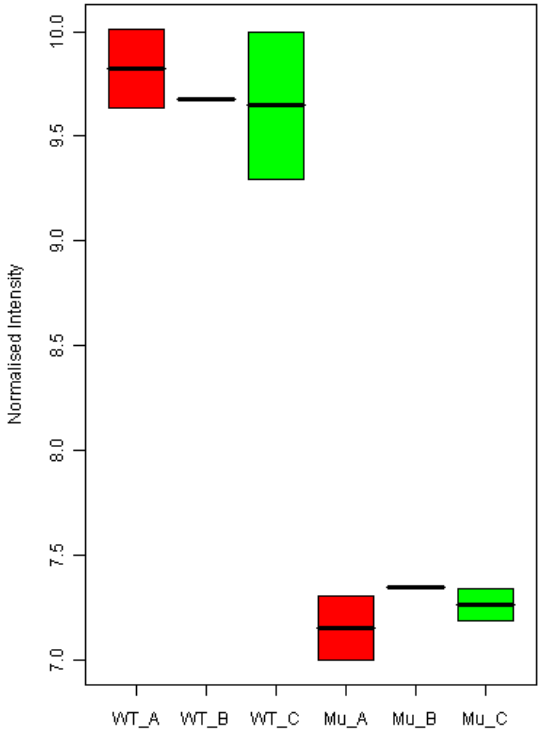
It is all too easy to for untoward things to happen somewhere along the line, and also surprisingly hard to spot them in high dimensional data like RNA-Seq

RNA-Seq: Influences on gene expression

- Gene expression data highly sensitive to many factors
  - Lab operator/conditions, day performed, sample collection methods, RNA extraction day and so on
  - Often influence the data to greater extent than any experimental effects
  - Any step where treated and control samples are handled differently could confound the experiment
  - If split into batches containing mix of treated/control samples, can account for potential effects in analysis
- Also be aware of potential effects from factors unrelated to the experiment on the data, which may need to be accounted for to optimise analysis

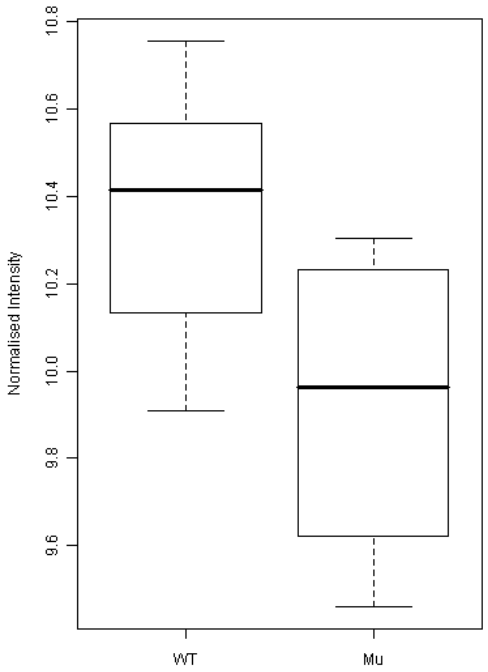
Example differentially expressed gene

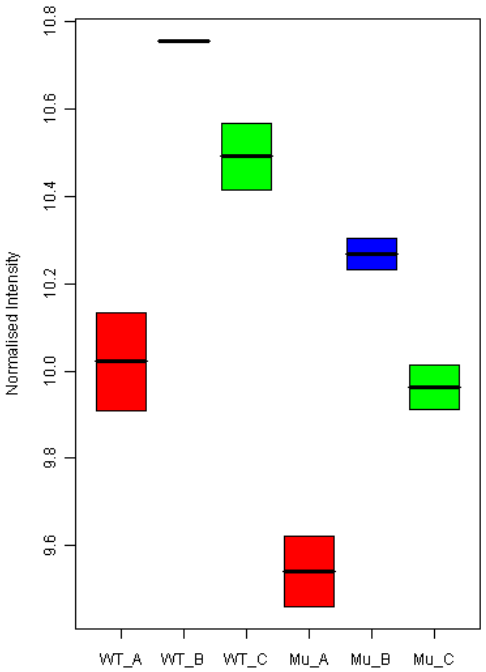




- Wt and Mut groups
- Three different litters (A, B, C)
- Top gene ~ 5x higher expression in Wt compared to Mut
- Similarly expressed across litters in both genotypes

Gene with strong litter effect





- Within litters, consistent pattern of higher expression in WT vs Mut
- Within genotypes, expression depends on litter: B > C > A
- Accounting for this source of variability increases power to detect changes of interest

wellcome centre human genetics

Practical Session I

Data Exploration and Quality Checks

Data exploration and QC practical

- Look at the tutorial at the link below:

https://www.well.ox.ac.uk/bioinformatics/training/RNASeq_Nov2020/practical_1/RNA_workshop_QC_practical.html

From the same location you can download the 'RNA_workshop_QC_practical.rmd' file, which can be opened in your R session for interactive running of the commands

If you have any problem getting started, please just let us know. You can spend approximately 20 minutes on this practical before we discuss the results.

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: 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 a big problem – producing false positives when the variance is underestimated
- The limma and edgeR packages for analysing gene expression use a clever strategy to limit this issue

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

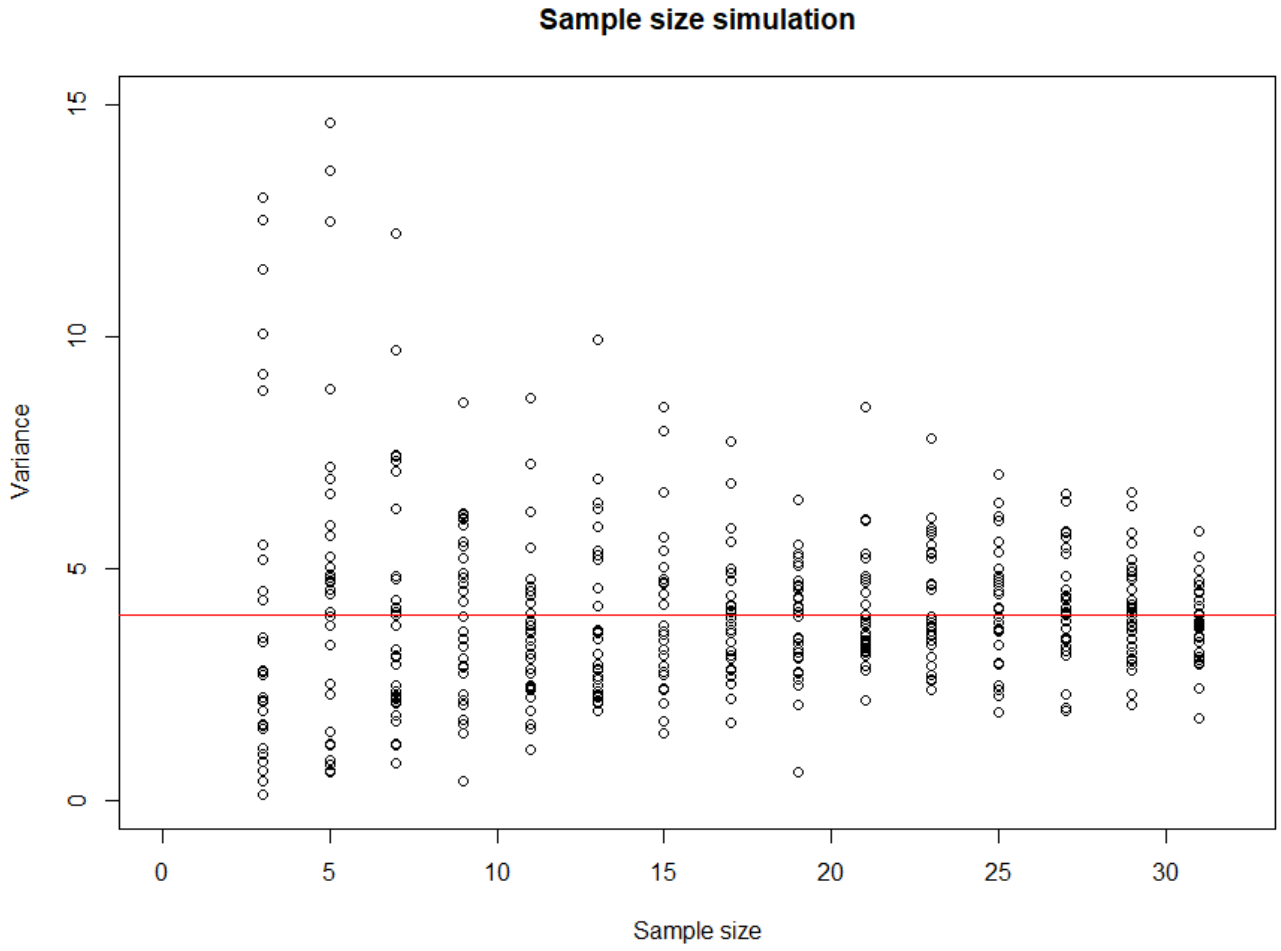
- We'll run a quick simulation first to understand this phenomenon better

- In R, enter the command

source("https://www.well.ox.ac.uk/bioinformatics/training/RNASeq_Nov2020/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

Data Preprocessing

RNA-Seq: Practical Session 2

- Now, we will start the analysis with some real gene expression data

- Use the tutorial document for practical 2 to work through the initial steps of reading data into R, formatting, and preprocessing steps of normalisation and filtering:

https://www.well.ox.ac.uk/bioinformatics/training/RNASeq_Nov2020/practical_2/RNA_workshop_preprocessing_practical.html

- Again there is an Rmarkdown version if you would like to use it
- Feel free to post questions or comments to the Teams chat

- We will re-start after lunch at 13:00

RNA-Seq: Analysis Models in R

wellcome centre human genetics

- R packages such as edgeR provide the user with a series of functions to perform different steps
- Often there is a lot of statistical computation happening (very efficiently) behind the scenes but we just need to run few lines of code
- One critical part for the user to set up correctly is the design matrix describing the experimental samples

RNA-Seq: Analysis Models in R

wellcome centre human genetics

- Assume we have 2 conditions, and 3 replicates per condition
- `conds <- factor(c(rep("WT", 3), rep("MU", 3)))`
- Inspecting this factor object in R will show the levels are ordered alphabetically by default, even though we ordered them WT, MU when creating the factor

- A design matrix can be created using the model.matrix function

- design <- model.matrix(~0+conds)

RNA-Seq: Analysis Models in R

- A design matrix has a row corresponding to each sample – the order absolutely **must** correspond to the sample order in the data object (holding our gene counts)
- So if your data object was ordered WT1 MU1, WT2, MU2, WT3, MU3 for example, the conds factor needs to capture this:
- conds2 <- factor(rep(c("WT", "MU"), 3)))
- Note again the levels are ordered alphabetically

wellcome centre human genetics

Practical Session III

Differential Expression Analysis