RNA-Seq Workshop

December 2021

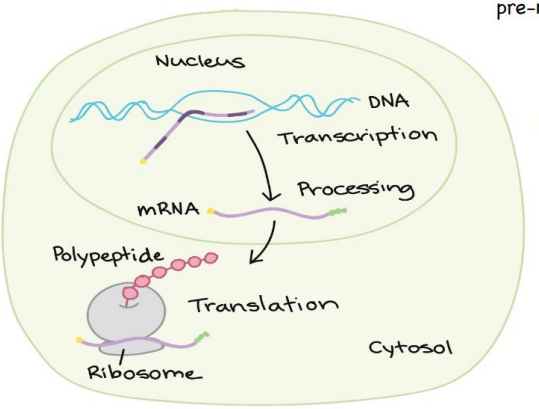
Helen Lockstone & Ben Wright
Bioinformatics Core, WHG

Schedule

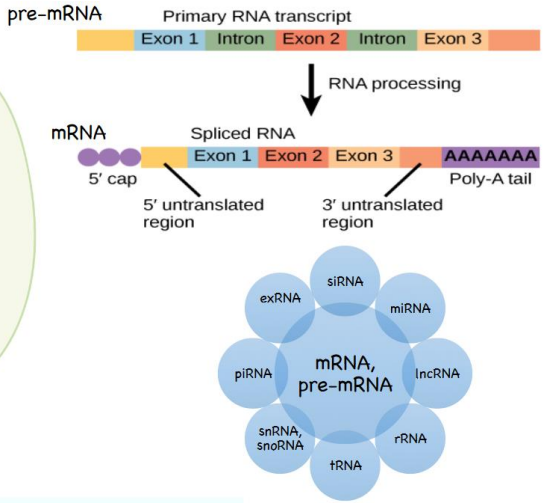
10:00-10:15	Introduction to RNA-sequencing
10:15-11:00	Overview of data generation and processing steps; quality control
11:00-11:15	*Break*
11:15-12:00	Data analysis practical in R part 1
12:00-13:00	*Lunch*
13:00-~15:00	Data analysis practical in R parts 2 & 3

Profiling the transcriptome

Gene transcription is a complex, highly dynamic and tightly regulated process… but even a snapshot gives valuable insight into biological systems and disease



- Image credit Khan Academy Open Courses

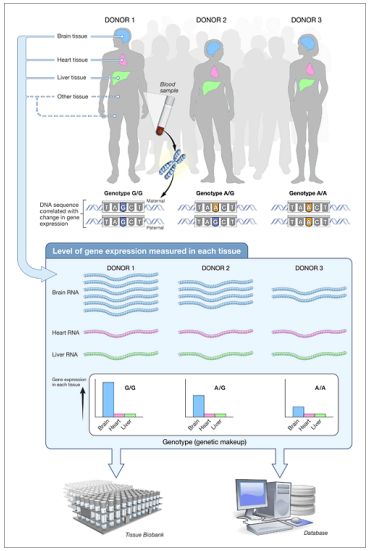


- Interrogating gene expression - accessible and applicable to wide variety of situations and research questions
- Important link between genotype and phenotype

Utility of gene expression profiling

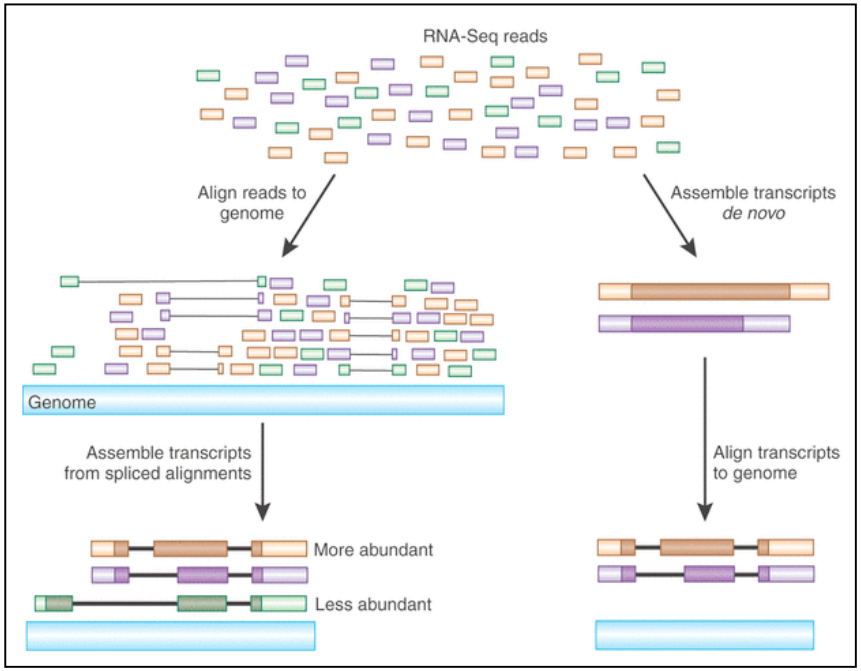
- ENCODE
- Allen brain atlas
- Genotype-Tissue Expression Project (GTEx)
- TCGA

ALLEN BRAIN ATLAS
DATA PORTAL



- 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

RNA-Sequencing



- High throughput - costs reduced and throughput increased over time
- RNA fragmented, converted to cDNA library and sequenced
- No prior knowledge of transcript sequence or structure required

Next-generation sequencing (since ~2007)

RNA-Seq: Library preparation

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	picograms

RNA-Seq: Sequencing Depth

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

  HiSeq4000/NovaSeq – one lane produces ~250-350 million reads

  Multiplex samples as only require 10-20million reads per sample for gene quantification

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

RNA-Seq: Sequencing Depth

- Typically, up to half of all genes are not sequenced at all or generate just a handful of reads – either unexpressed or expressed at very low levels

- Other genes can have many thousands of sequenced reads mapping to them (highly expressed, well-represented in the library)

Do you think that sequencing more deeply would help pick up lower expressed genes?

RNA-Seq: Sequencing Depth

Do you think that sequencing more deeply would help pick up lower expressed genes?

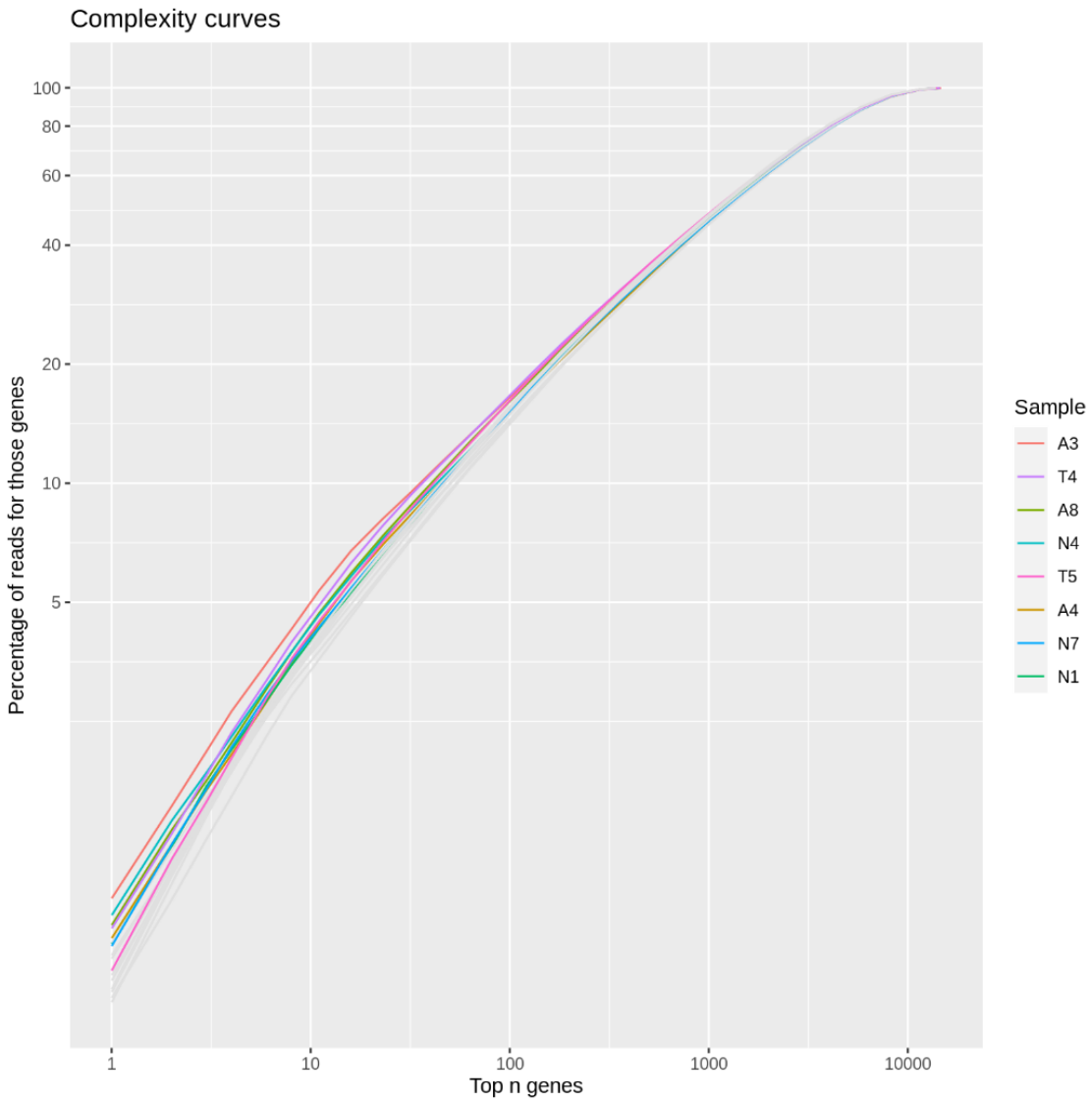
It may seem like it should help but in fact it doesn't because another lane (or lanes) of data would give you plenty more reads for the genes you already have enough data for, and very unlikely to increase the counts for low-expressed genes sufficiently to analyse them

RNA-Seq: Library composition

- Input RNA amount affects library composition and complexity (the range of transcripts that are represented in the library)
- Typically, low input amounts of RNA leads to the most highly-expressed genes dominating the sequencing data

- For this reason, normalising input to that of the lowest sample may not be recommended
  - E.g. if 8 samples had >250ng and 2 samples had <100ng, it would be preferable to use 250ng input where available and have a couple with lower RNA input

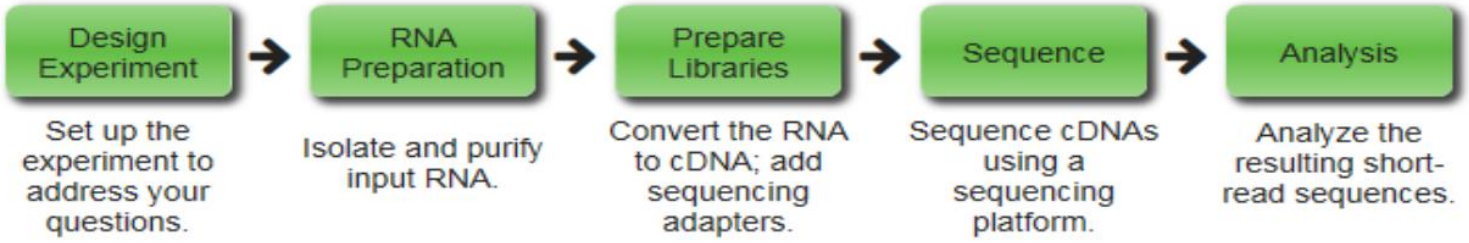
RNA-Seq: Library complexity



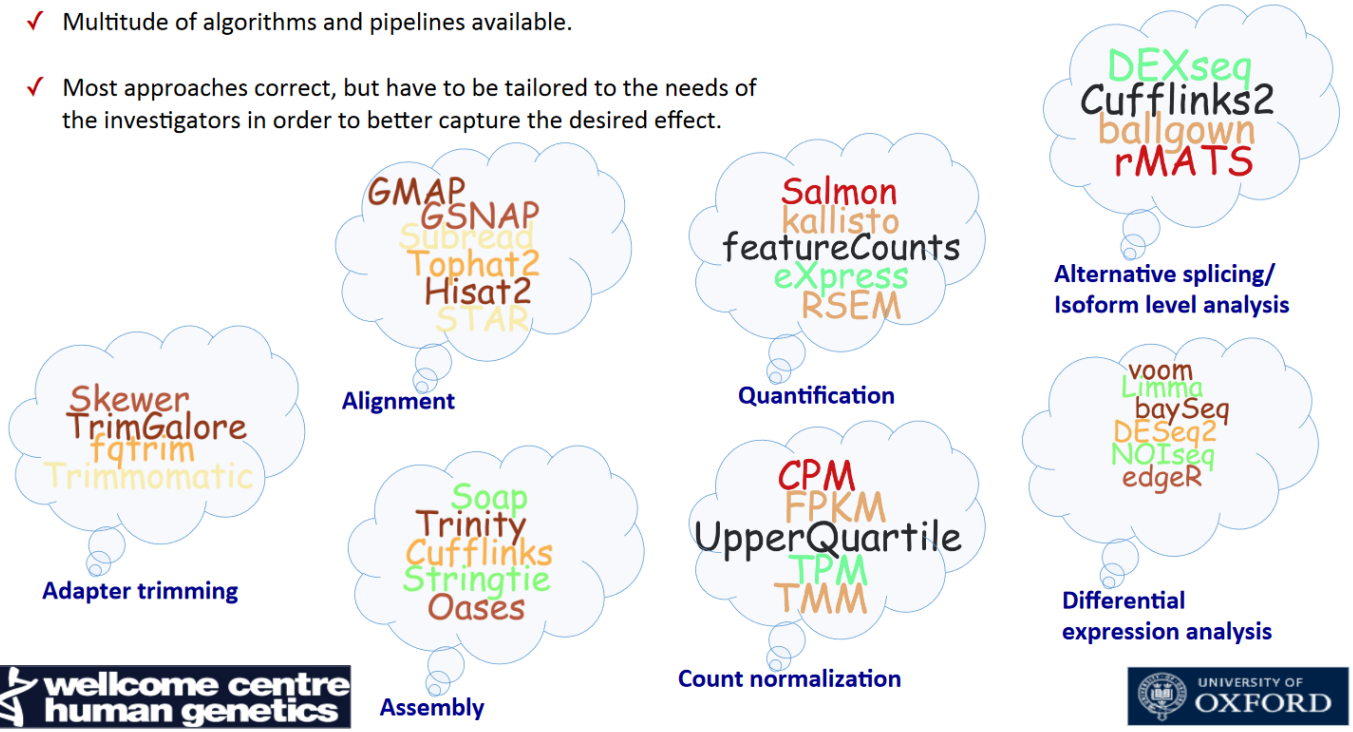
A useful metric is the % of reads arising from the top100 most highly expressed genes in each sample.

For bulk RNA-Seq, this might be around 20-40%. For low input and single cell data up to 70-90% of all sequenced reads can come from the top 100 genes.

RNA-Seq: Pipelines and tools



- ✓ 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.



wellcome centre human genetics

UNIVERSITY OF OXFORD

RNA-Seq: General advice

- 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 the research question and data?
  - Have I understood how it works and how the parameter settings might 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