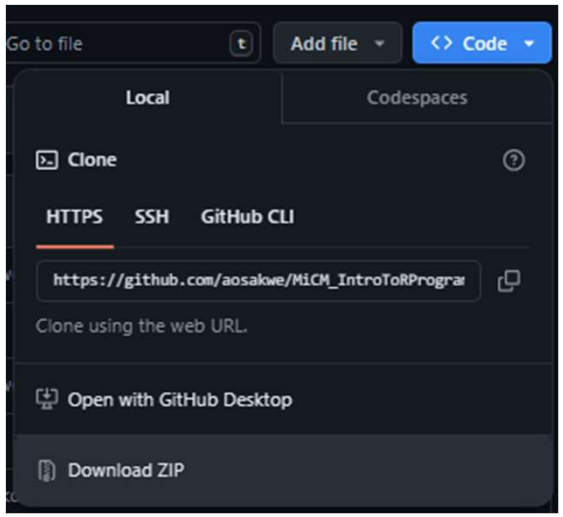
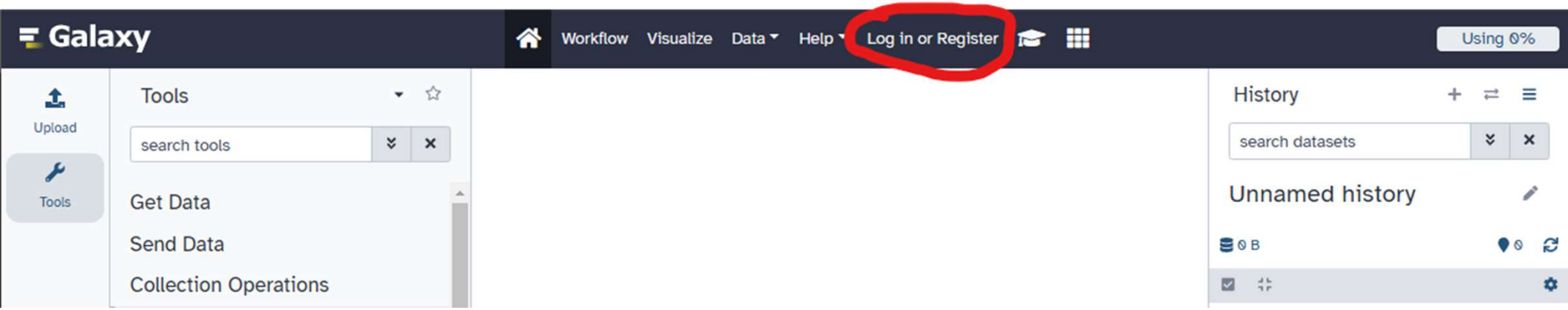
McGill initiative in Computational Medicine

1. Go to https://github.com/aosakwe/MiCM_IntroToRNA and download the workshop materials ZIP (under the blue code button)



2. Go to https://usegalaxy.org/ and create an account (don't forget to verify it!)



McGill initiative in Computational Medicine

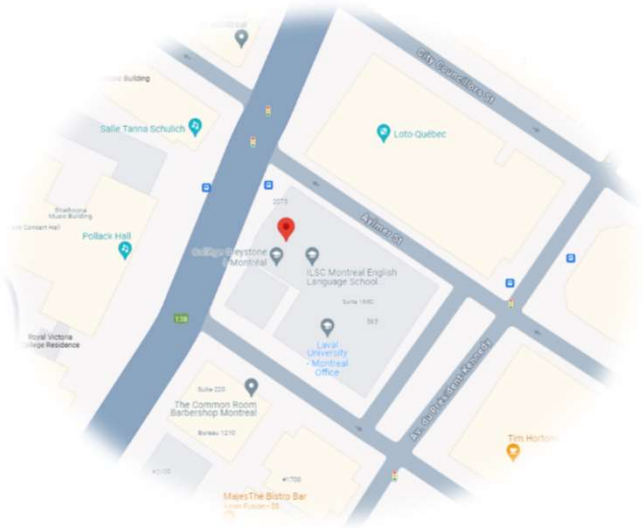
Introduction to RNA-seq

Workshop Lead: Adrien Osakwe
Facilitator: Jeffrey Yu
Date: June 27, 2024

McGill initiative in Computational Medicine

McGill initiative in Computational Medicine

Mission statement: deliver quality workshops designed to help biomedical researchers develop the skills they need to succeed.



Location: 550 Sherbrooke Street, Montreal, Quebec

Scan the QR code to sign up for our **mailing list**

Contact: workshop-micm@mcgill.ca

Summer 2024 Workshop Series

Workshop	Date	Lead/Facilitator	Location	Registration
How to think in Code	July 3 10AM-1PM	Thomas Zheng	Education Room 133	Open
Intro to UNIX and HPC	July 11 9AM-1pm	Georgi Mehri	Education Room 133	Open
Intro to Git & GitHub	July 12 1PM-5PM	Adrien Osakwe	Education Room 133	Open
Intro to Python (Part 1)	July 16 9AM-1PM	Benjamin Rudski	Education Room 133	Open
Intermediate Python (Part 2)	July 18 9AM-1PM	Benjamin Rudski	Education Room 133	Open
Fundamentals of Machine Learning	July 24 9AM-1PM	Tugce Gurbuz	Education Room 133	Open
Intro to Matlab	August 7 9AM-1PM	Meghana Munipalle	Education Room 133	TBA
Intro to R (Part 1)	August 12 9AM-1PM	TBA	Education Room 133	TBA
Intermediate R (Part 2)	August 14 1PM-5PM	Gerardo Martinez	Education Room 133	TBA
Intro to Bayesian Inference in R	August 16 1PM-5PM	Adrien Osakwe	Education Room 133	TBA
Proteogenomics	August 19 1PM-5PM	Thomas Zheng	Education Room 133	TBA

https://www.mcgill.ca/micm/training/workshops-series

Outline

1. **Introduction (~10 mins)**
2. **Experimental Design (~20 mins)**
3. **Processing & Quality Control (~30 mins)**
4. **Analysis (~1h30 mins)**
5. **Single Cell Data (~45 mins)**

Acknowledgements

Material

- Reinnier Padilla (HGEN)

Exercises

- Galaxy Tutorial Page
- DESeq2 Vignettes
- OSCA eBook

Data

- 10X Genomics
- Stephen Turner

Introduction

Limitations of Microarrays

- Microarrays are **insufficient for large-scale research**
- Can only study **known mRNA transcripts**
- Does not generate counts

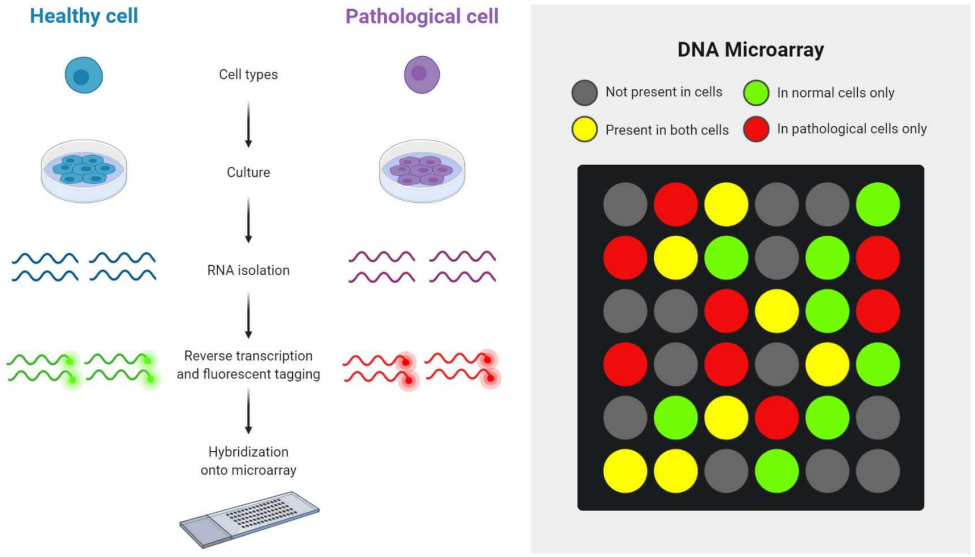


Image By Sagar Aryal, created using biorender.com

RNA Sequencing

- RNA-seq provides a **high-throughput** way to explore gene expression
- Can discover **novel transcripts & isoforms**
- Can perform more complicated analyses
  - Incorporate covariates (sex, age, etc.)

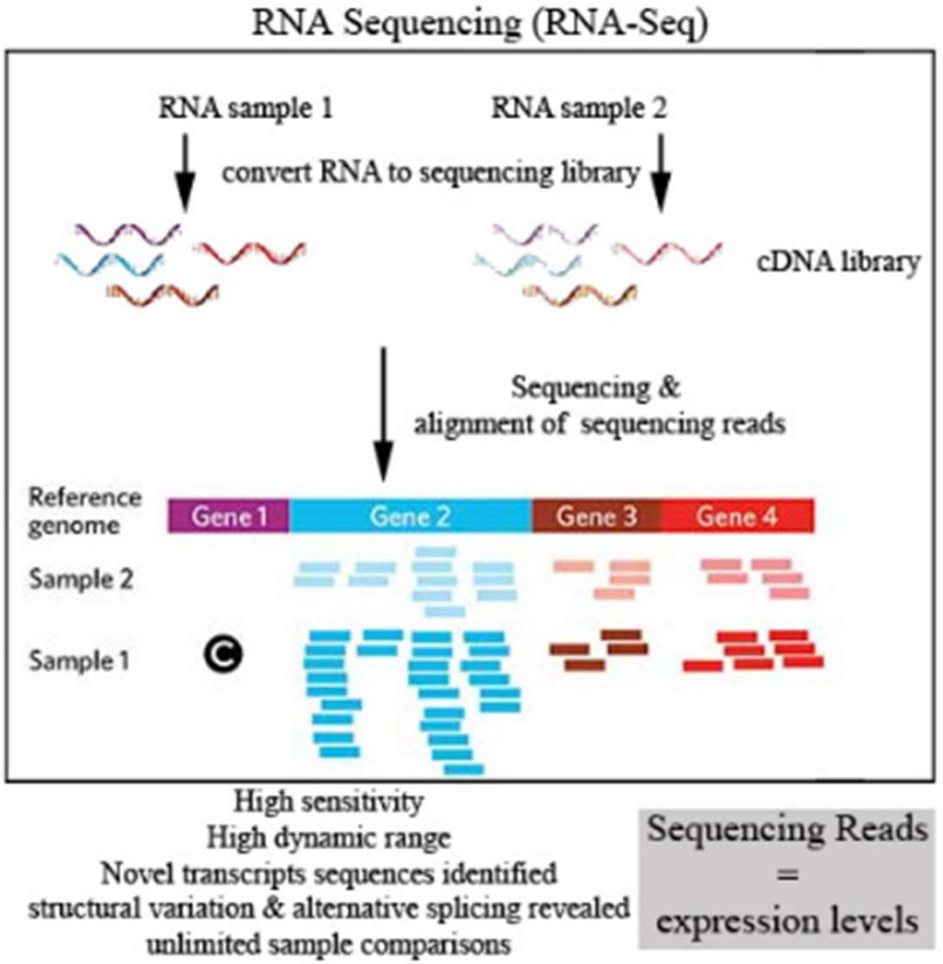


Image Source: otogenetics

Experimental Design

RNA-seq Protocol

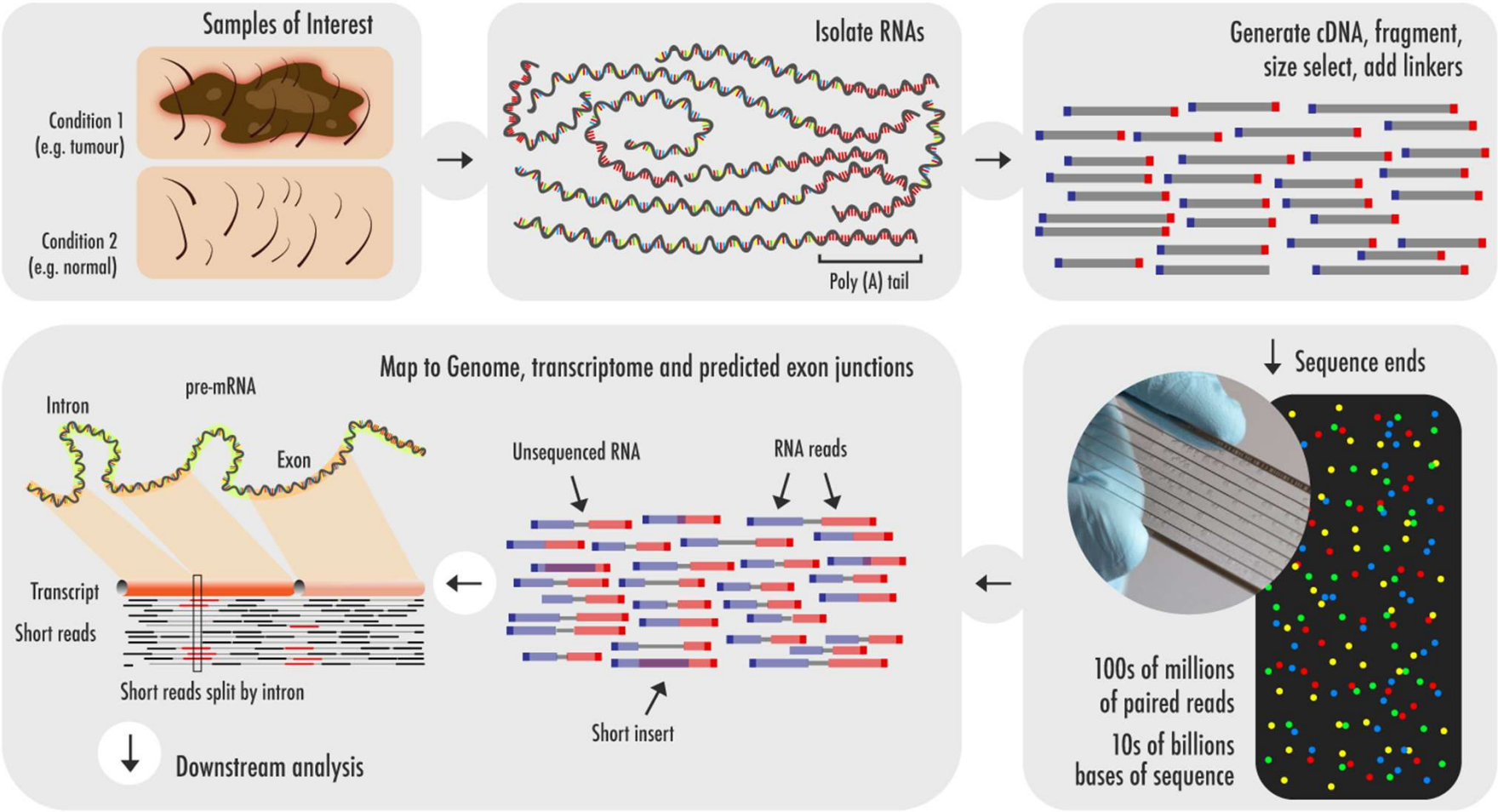


Image Source: Technology Networks – Genomics Research

Extraction Protocols

- Majority of RNA in sample may not be relevant
  - Ribosomal RNA (rRNA – 90% of all transcripts)

- Extraction protocols aim to enrich sample for mRNA

1. Poly(A) Selection
   - Requires a lot of mRNA and minimal degradation (RIN)
2. **rRNA Depletion**

Single-end vs. Paired-end Reads

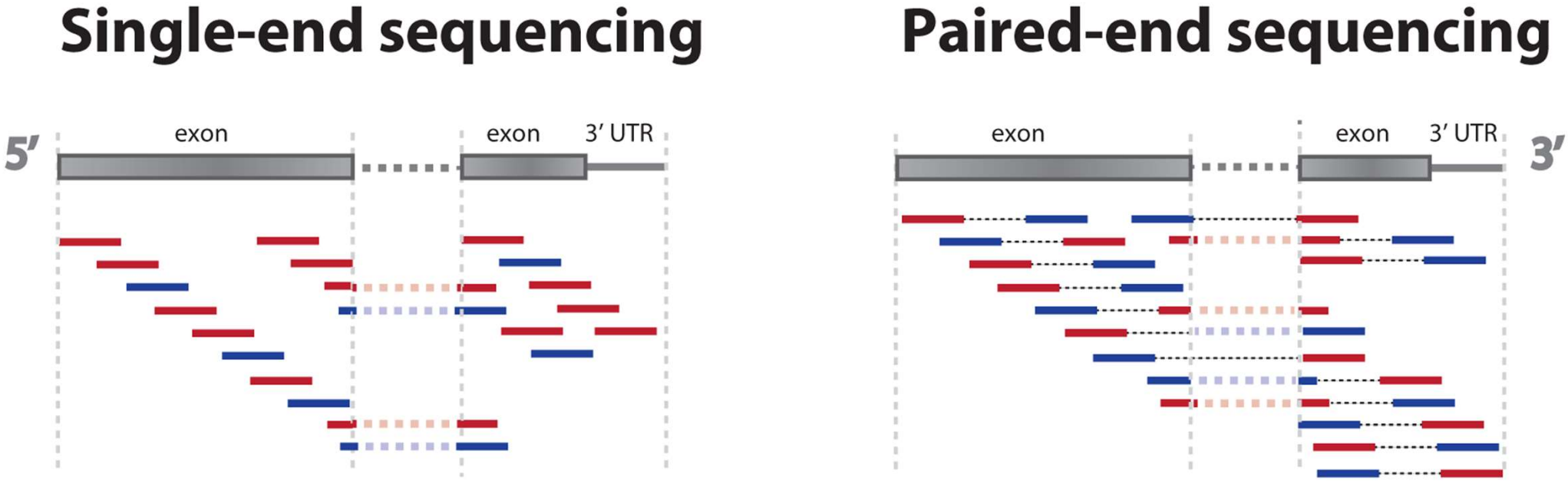


Image Source:
statomics.github.io

Single-end vs Paired-end Reads

Single-end Reads

- Cheaper
- Faster
- Sufficient for well-annotated genomes

Paired-end Reads

- Higher accuracy
- Better alignment
- Better for de-novo assembly

Read Length

- How many base pairs are read at a time
- 50, 100 or more base pairs

- Longer reads are better for alignment and assembly
  - Are also more expensive

- 50bp may be sufficient for differential expression
- 100bp> is more reliable for isoform studies

Library Size

- # of sequenced reads per sample
- Large libraries are better at finding rare transcripts
- BUT large libraries are at risk of **noise and contamination**
  - Consider using **saturation curves**

Library Size Saturation Curve

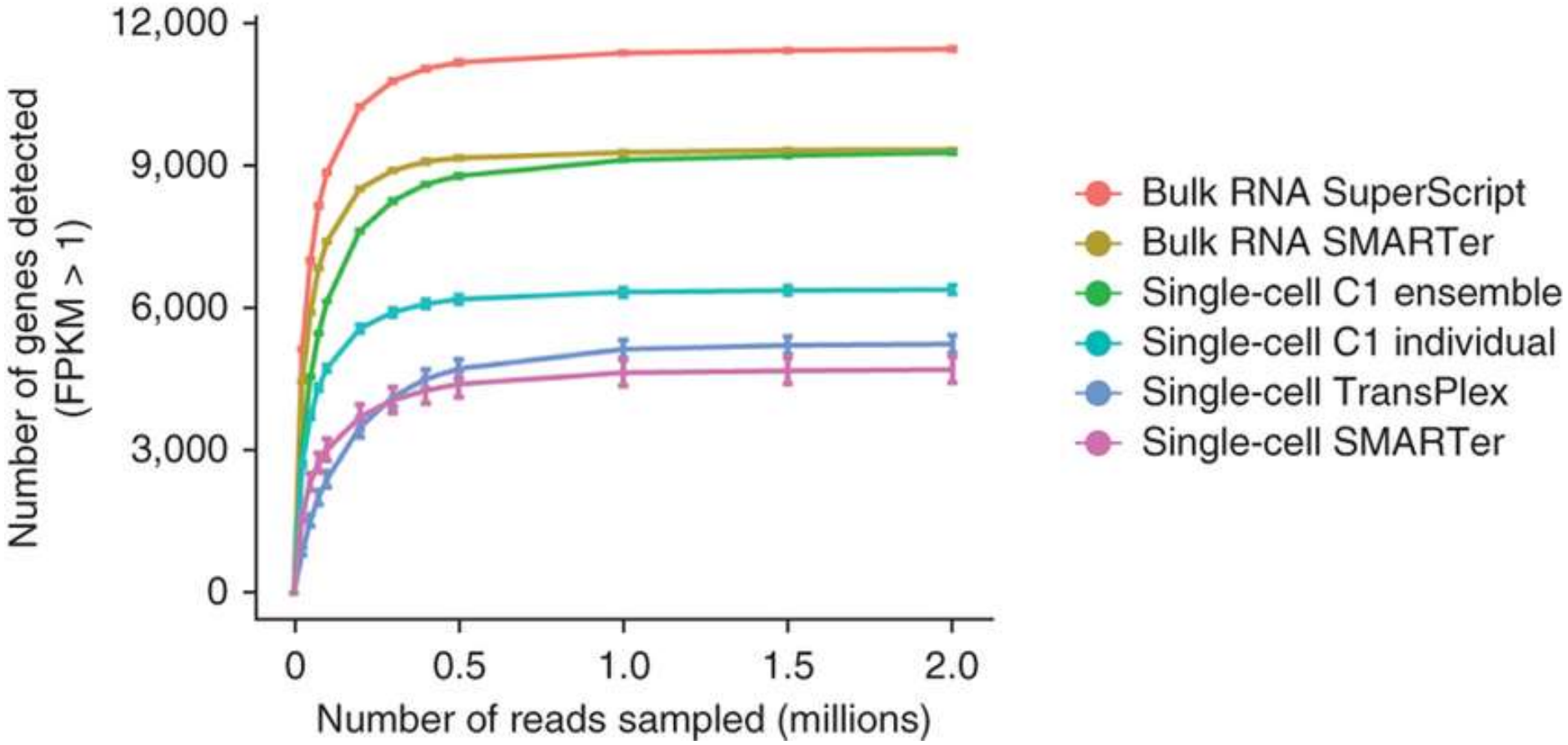
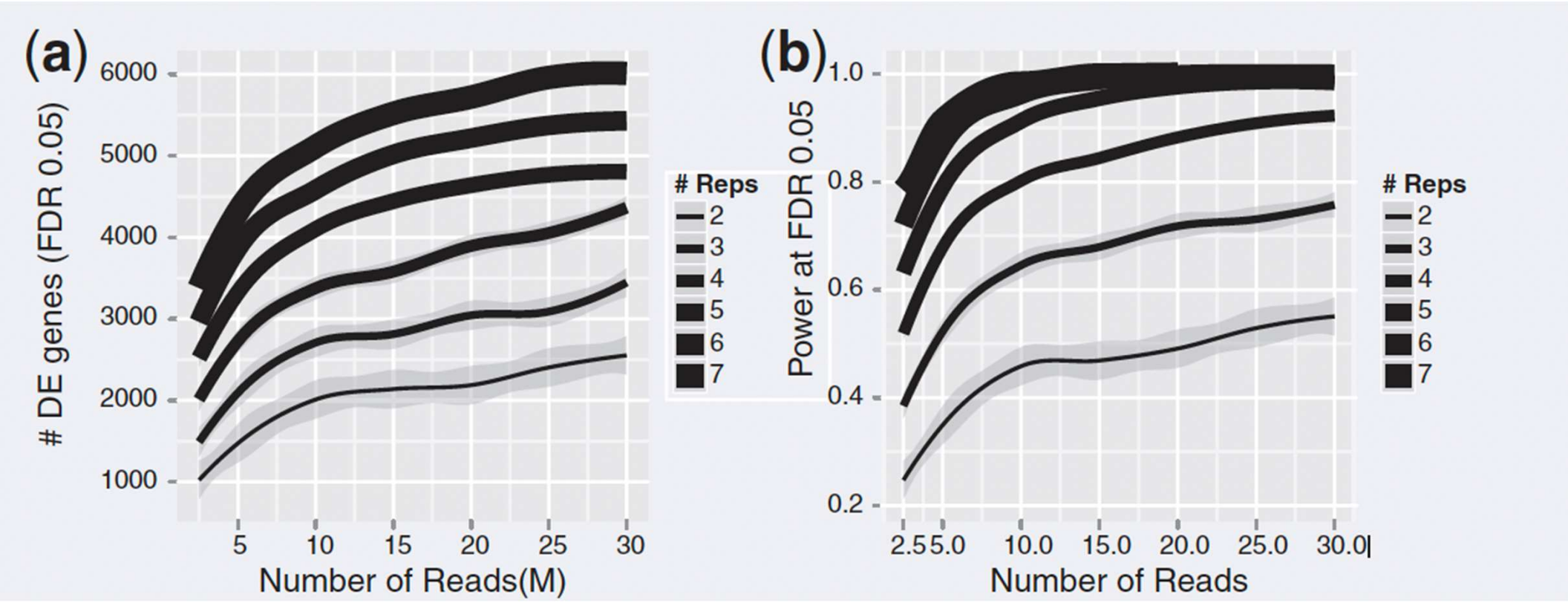


Image Source: Angela R Wu

Number of Replicates



Liu, Y., Zhou, J. & White, K. P. *Bioinformatics* **30**, 301-304, doi:10.1093/bioinformatics/btt688 (2014).

Other Considerations

- Need to consider technical and biological variability
  - What is our desired statistical power?
- Plan sequencing experiments
  - randomized sample processing
  - ensuring covariates are well mixed across batches

- Three replicates seen as **bare minimum** for inference
- Recommend running a **power analysis**
  - **Scotty** tool helps determine optimal # of replicates and library size

Scotty Tool

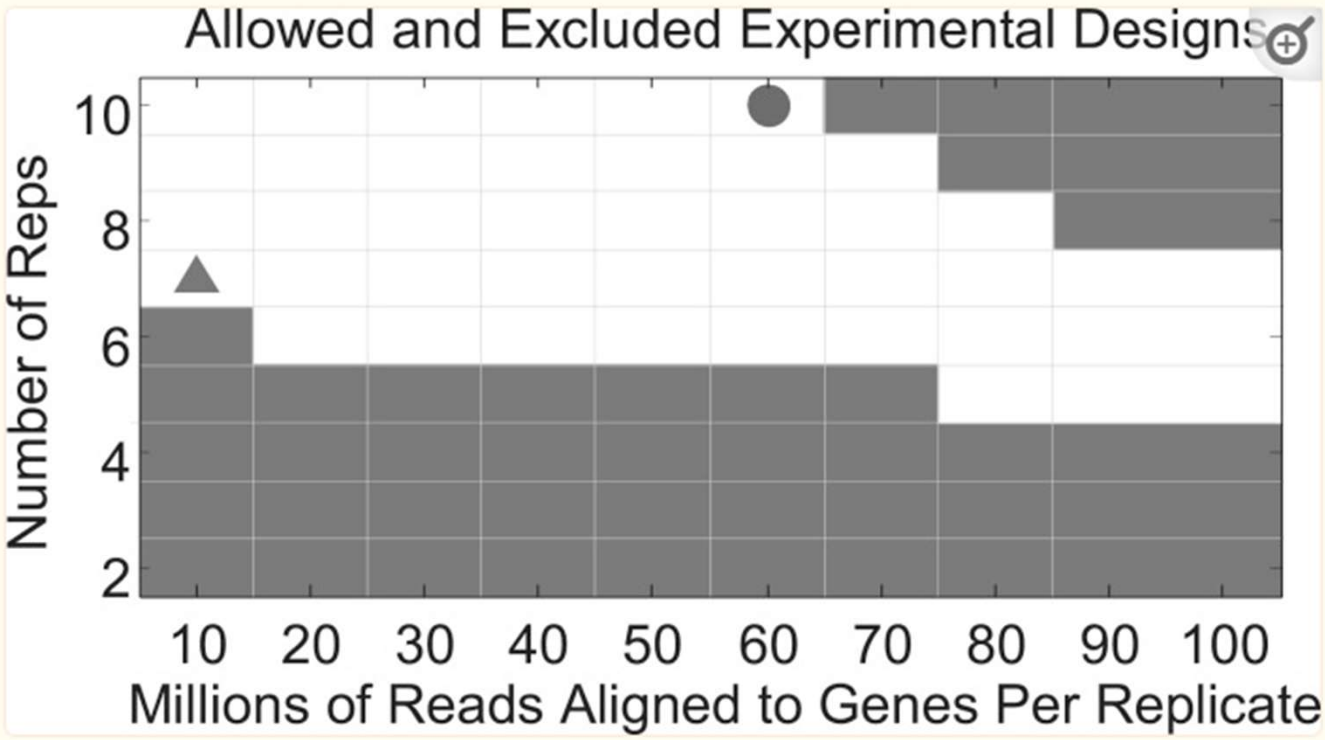


Fig. 1.

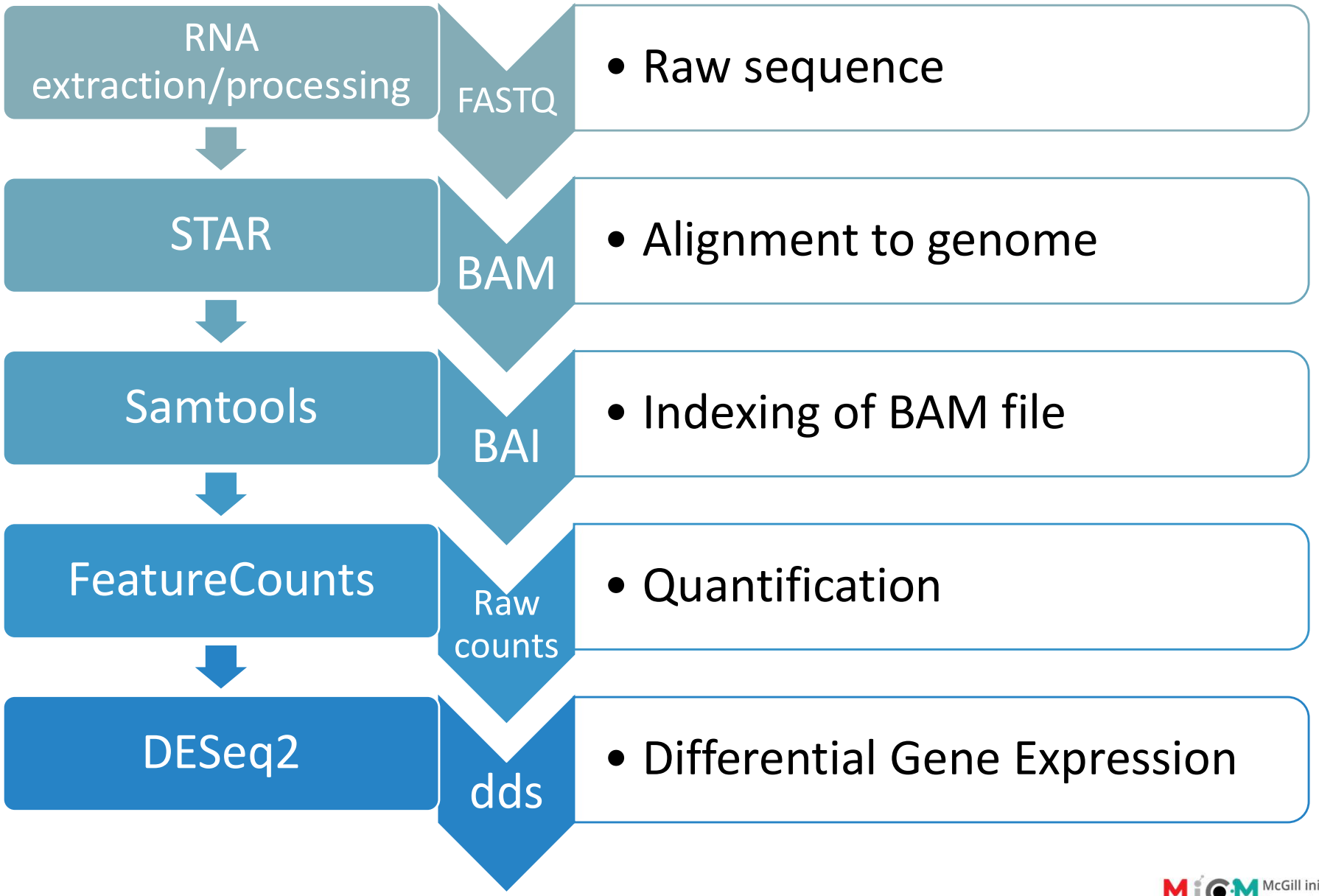
An example output from the Scotty application. This figure shows the user which of the tested experimental configurations do (white) and do not (shaded) conform to the user-defined constraints. Scotty then indicates the optimal configuration based on cost (filled triangle) and power (filled circle)

- http://scotty.genetics.utah.edu/

Processing & Quality Control (Galaxy)

Exercise: Galaxy Tool

RNA-seq Processing Protocol



Read Quality

- Quality control of raw reads involves the following metrics/indicators:
  - **sequence quality**
  - **GC content**
  - **k-mer overrepresentation**
  - **duplicated reads (sequencing errors)**
  - Adapter presence
  - PCR artifacts
  - sample/RNA contamination
- Satisfactory Duplication, k-mer and GC content levels will vary
  - expect samples to have similar estimates within an experiment.
  - Discard samples **over 30% disagreement.**
  - **FASTQC** for Illumina-based reads or NGSQC for any platform.
- Sequence quality decreases towards 3' end of the sequence
  - can trim 3' end to increase mappability (**Trimmomatic**)

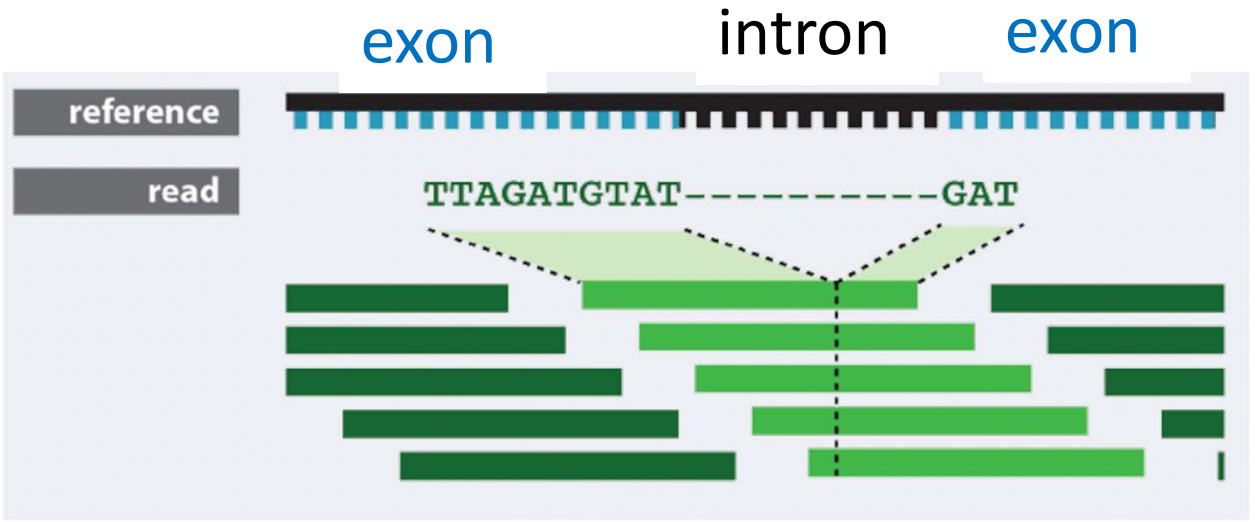
Exercise: FASTQC

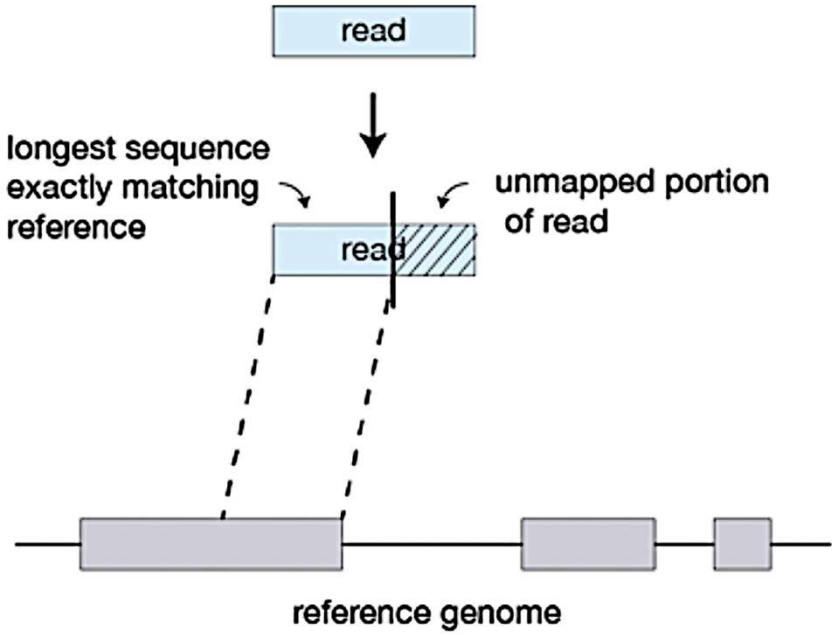
Alignment

- Two Approaches

1. Reference-based Alignment
   - Using either a reference genome or transcriptome
2. De-novo Assembly
   - Creates a new reference based on your sequenced reads

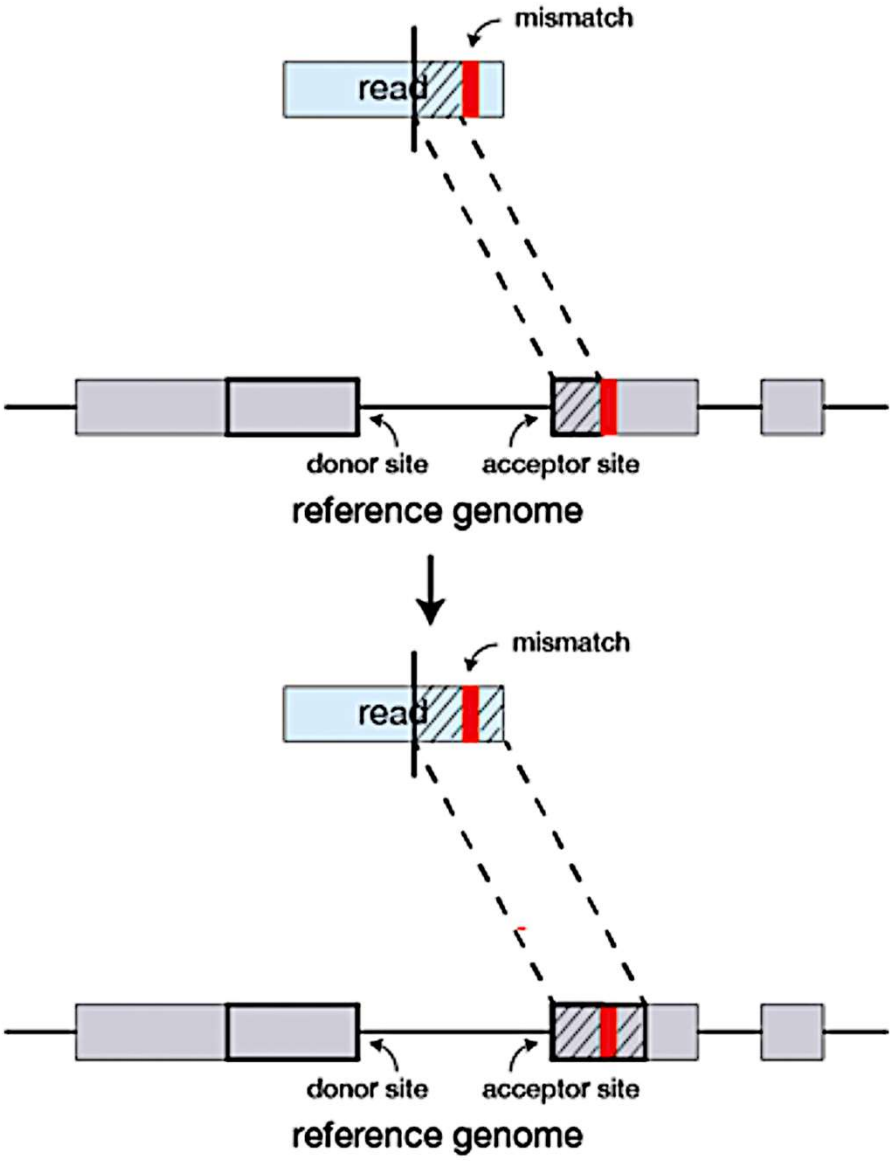
STAR Alignment





Maximal mappable prefixes (MMPs)

STAR Alignment



Transcript Alignment

- Crucial quality metric is the **percentage of mapped reads**.
  - estimate of sequencing accuracy and contamination.
  - expect **70-90% of reads** to map to the human genome
  - significant proportion of reads mapping to small number of identical regions (multi-mapping reads).

- Examine uniformity of read coverage on exons and the mapped strand.
- GC content of mapped reads can also reveal PCR biases
- Tools such as Picard, RSeQC and Qualimap can be used for alignment QC.

Transcript Identification

- With a reference genome, we can determine what transcripts RNA-seq reads represent.
  - Prevents the discovery of novel transcripts and emphasizes quantification.
- Without a reference, need to assemble reads into contigs that represent candidate transcripts.
- Read coverage can be used to determine expression level in both cases.

De novo alignment & Transcript Discovery

- Short reads struggle to accurately infer full-length transcripts.
  - Tools like GRIT can help improve this with CAGE/RAMPAGE data
  - Paired-end reads are also helpful

- An easier solution is to use **long read sequencing data**
  - **PacBio, Oxford-Nanopore**

- De novo assembly can be done by tools such as Trinity, SOAPdenovo-Trans, Oases and Trans-ABySS
- Important to pool reads in comparative analysis during De novo assembly

Transcript Quantification

- This step generates the **count matrix**
  - the number of reads mapping to a given transcript sequence.
  - HTSeq-count or **featureCounts** aggregate raw counts.
  - gene-level quantification uses a gene transfer format (GTF) file containing genome coordinates of exons and genes
- raw read counts are not ideal for analysis
  - Need normalized metrics for fair comparisons

Count Normalization

- **Reads per kilobase of exon model per million reads (RPKM)**
  - RPKM is a within-sample normalization method which removes transcript-length and library size effects
  - Designed for single-end reads
- **Fragments per kilobase of exen model per million mapped reads (FPKM)**
  - FPKM is an extension of RPKM designed for paired-end reads
  - Each pair of reads are treated as one fragment here (if they both were mapped). This avoids counting a fragment twice which cannot be done by RPKM.
  - Renders the same output as RPKM on single-end reads.
- **Transcripts per million (TPM)**
  - TPM is an extension of RPKM where we first normalize by transcript length and then normalize by sequencing depth.
  - The sum of all TPMs is the same for each sample, making it easier to compare the proportion of reads mapped to a gene across samples.
  - Can also convert FPKM counts into TPM

Analysis (R)

Exercise: Bioconductor

Exploratory vs. Inferential Analysis

Exploratory

- Hypothesis-generating
- More freedom in analysis pipeline
- Helps you **design** your analysis pipeline

Inferential Analysis

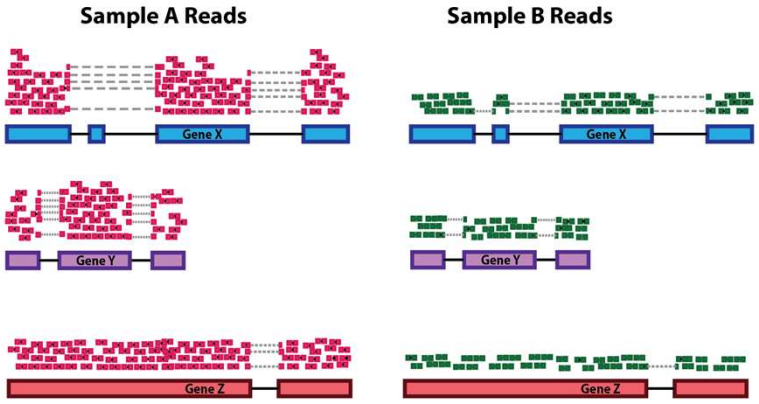
- Hypothesis testing
- Need to decide analysis pipeline **beforehand**
- Risk of **double-dipping**

Filtering Genes

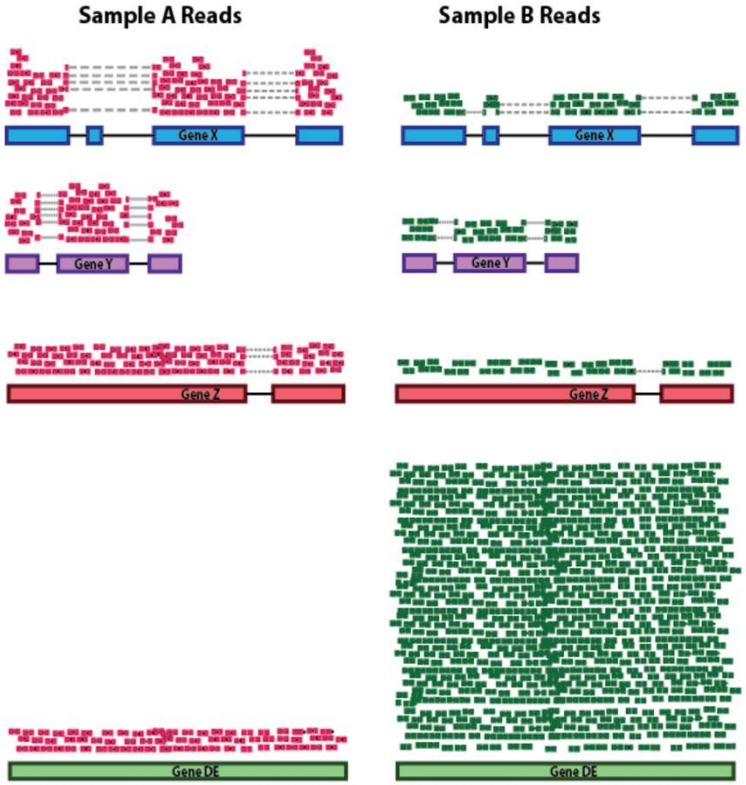
- Remove by functional category
  - Mitochondrial RNA
  - Ribosomal RNA
- Select for highly variable genes
  - Reduces computational cost
  - Potentially reduces noise
  - Risk losing novel genes

Normalization

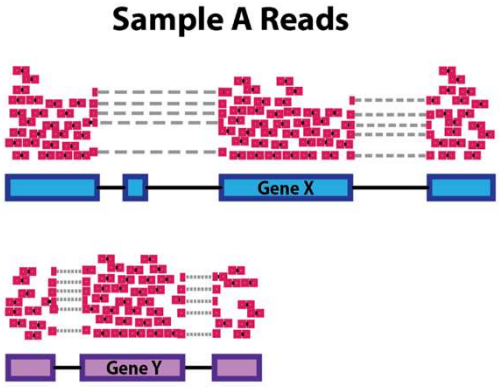
Sequencing Depth



RNA composition



Gene length



Normalization Limitations

- Assume we have two identical samples
  - Knockout expression of Gene D in sample 2
- Fixed Library size means remaining counts are redistributed over the remaining genes

Gene	Sample 1	Sample 2
A	30	235
B	24	188
C	0	0
D	563	0
E	5	39
F	13	102
Total	635	635

Normalization with DESeq2

Love *et al. Genome Biology* (2014) 15:550
DOI 10.1186/s13059-014-0550-8

METHOD Open Access

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love[1,2,3], Wolfgang Huber[2] and Simon Anders[2*]

Normalization with DESeq2

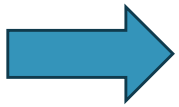
Gene	Sample 1	Sample 2	Sample 3
A	0	10	4
B	2	6	12
C	33	55	200

Gene	log(Sample 1)	log(Sample 2)	log(Sample 3)
A	-Inf	2.3	1.4
B	0.7	1.8	2.5
C	3.5	4.0	5.3

Normalization with DESeq2

Average each gene

Gene	Average of log values
A	-Inf
B	1.7
C	4.3



Filter out genes with infinite averages

Gene	Average of log values
B	1.7
C	4.3

Normalization with DESeq2

Subtract the average log value from the log counts:

Gene	log(Sample 1)	log(Sample 2)	log(Sample 3)
B	-1.0	0.1	0.8
C	-0.8	-0.3	1.0

$$\log(\text{counts for gene X}) - \text{average}(\log \text{ values for counts for gene X}) = \log\left(\frac{\text{counts for gene X}}{\text{average for gene X}}\right)$$

Calculate the median of the ratios for each sample:

Gene	log(Sample 1)	log(Sample 2)	log(Sample 3)
B	-1.0	0.1	0.8
C	-0.8	-0.3	1.0
Median	-0.9	-0.1	0.9

Normalization with DESeq2

Compute the scaling factor by taking the exponential of the medians:

Gene	Sample 1	Sample 2	Sample 3
Median	-0.9	-0.1	0.9
Scaling factors	0.4	0.9	2.5

Compute the normalized counts: divide the original counts by the scaling factors:

Gene	Sample 1	Sample 2	Sample 3
A	0	11.11	1.6
B	5	6.67	4.8
C	83	61.11	80

Exercise: Creating DESeq2 object

Sample Clustering (PCA)

- Principle component analysis compresses data into a smaller set of variables (**dimension reduction**)
  - From 20000 genes to 20 **principal components**

- Useful during exploratory analysis
  - Are there confounding variables in the data?
  - Sex-specific differences?

Exercise: PCA Plots

Differential Expression

Differential Gene Expression

- **Goal:** identify differentially expressed genes (DEGs)
  - **Determine how treatment and covariates affect expression**
  - **Infer underlying biological mechanisms**

- Common methods:
  - **DESeq2**
  - edgeR
  - limma

DESeq2

- Models Gene Expression using **Generalized Linear Models**
- Assumes counts are sampled from Negative-binomial distribution

The read count K_{ij} for gene i in sample j is described with a GLM of the negative binomial family with a logarithmic link:

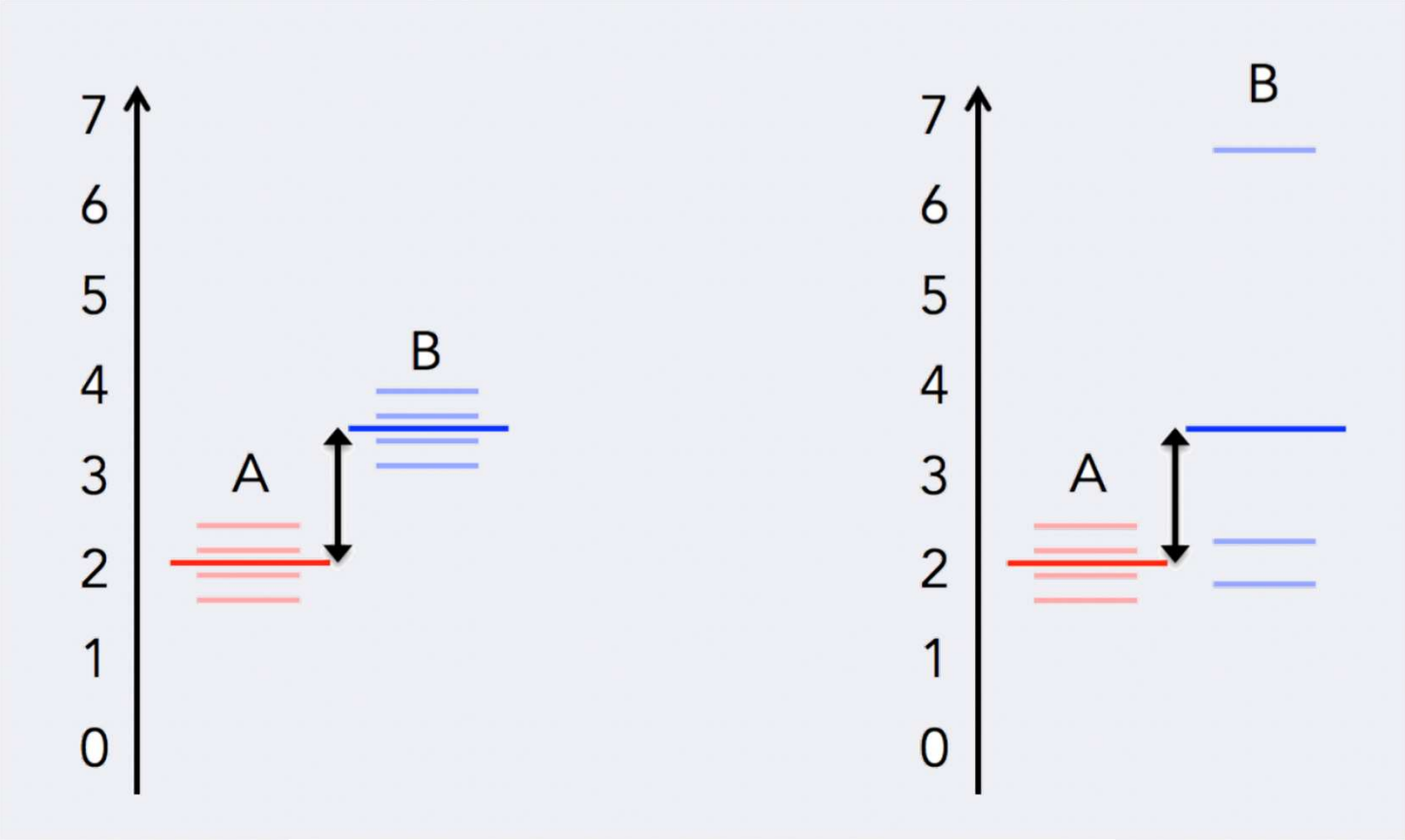
$$\begin{aligned} K_{ij} &\sim \mathrm{NB}\left(\mathrm{mean} = \mu_{ij}, \mathrm{dispersion} = \alpha_i\right) \\ \mu_{ij} &= s_{ij} q_{ij} \end{aligned} \tag{1}$$

$$\log q_{ij} = \sum_r x_{jr} \beta_{ir} . \tag{2}$$

- `baseMean` : mean of normalized counts for all samples
- `log2FoldChange` : log2 fold change
- `lfcSE` : standard error
- `stat` : Wald statistic
- `pvalue` : Wald test p-value
- `padj` : BH adjusted p-values

```
design = ~ sex + age + treatment + sex:treatment
```

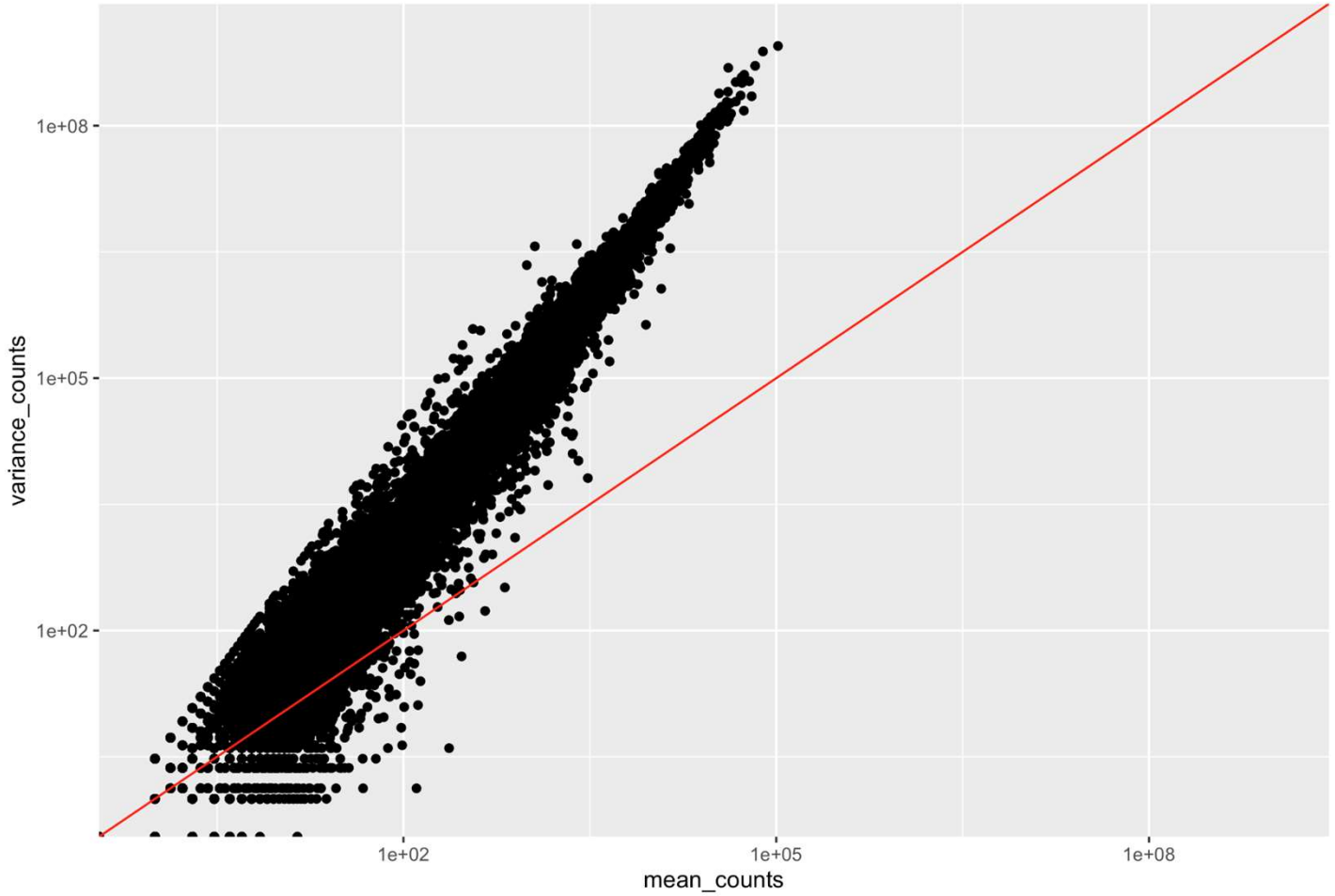
Uneven distribution of information



Replication introduces variance

Adapted from: https://hbctraining.github.io/DGE_workshop_salmon_online/schedule/links-to-lessons.html

Uneven distribution of information



Adapted from: https://hbctraining.github.io/DGE_workshop_salmon_online/schedule/links-to-lessons.html

What is Dispersion?

- A measure of spread or variability in the data

$$Var_{ij} = \mu_{ij} + \alpha_i \mu_{ij}^2$$

Rearranged:

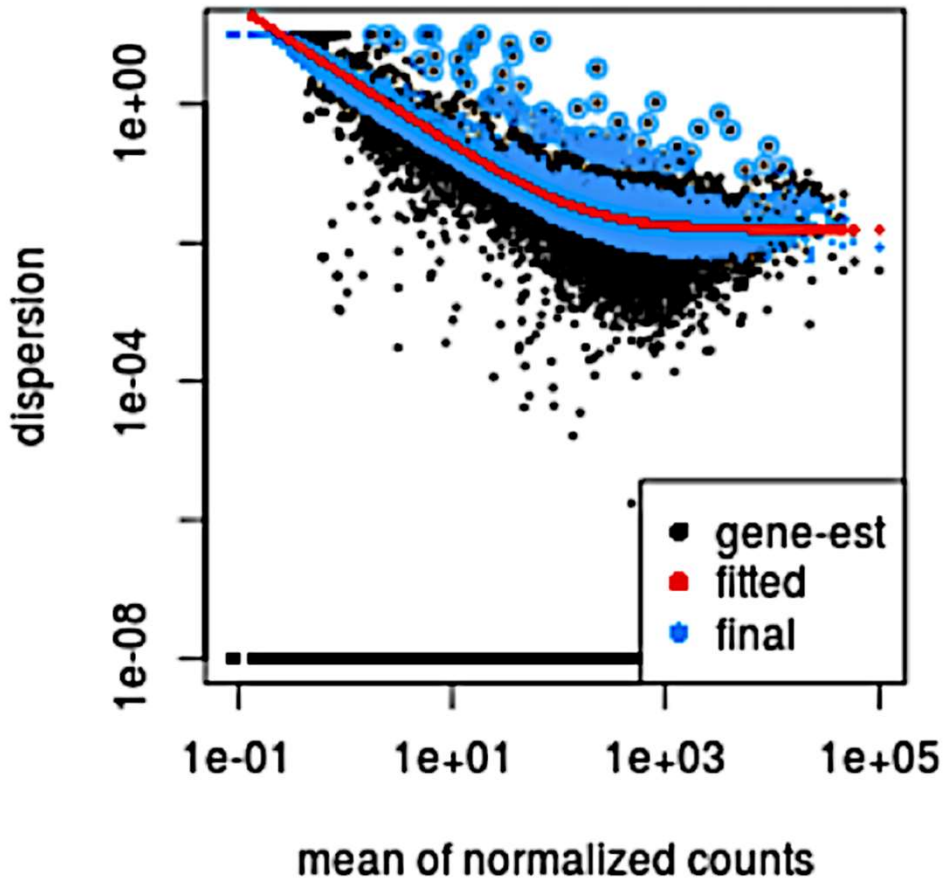
$$\sqrt{\alpha_i} = \frac{Var_{ij} - \mu_{ij}}{\mu_{ij}}$$

Which is also the same as:

$$\sqrt{\alpha_i} = \frac{Var_{ij}}{\mu_{ij}} - 1$$

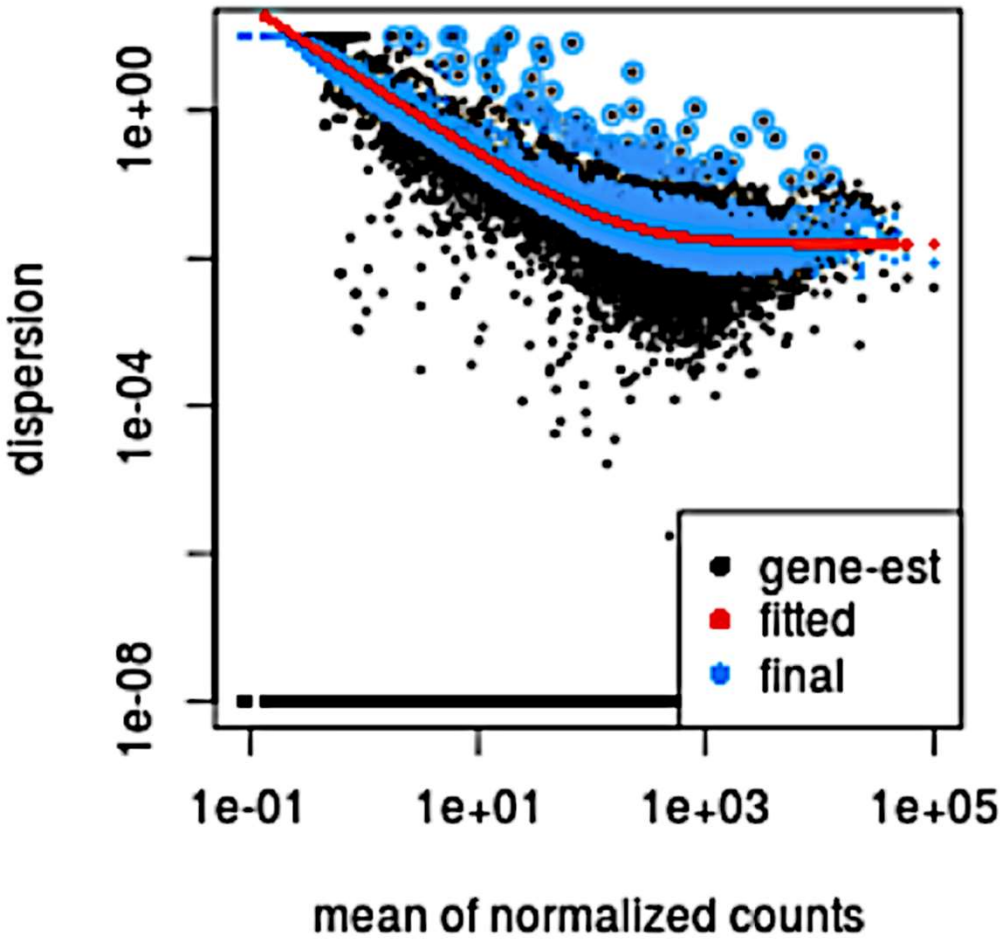
Adapted from: https://hbctraining.github.io/DGE_workshop_salmon_online/schedule/links-to-lessons.html

DESeq2 Dispersion



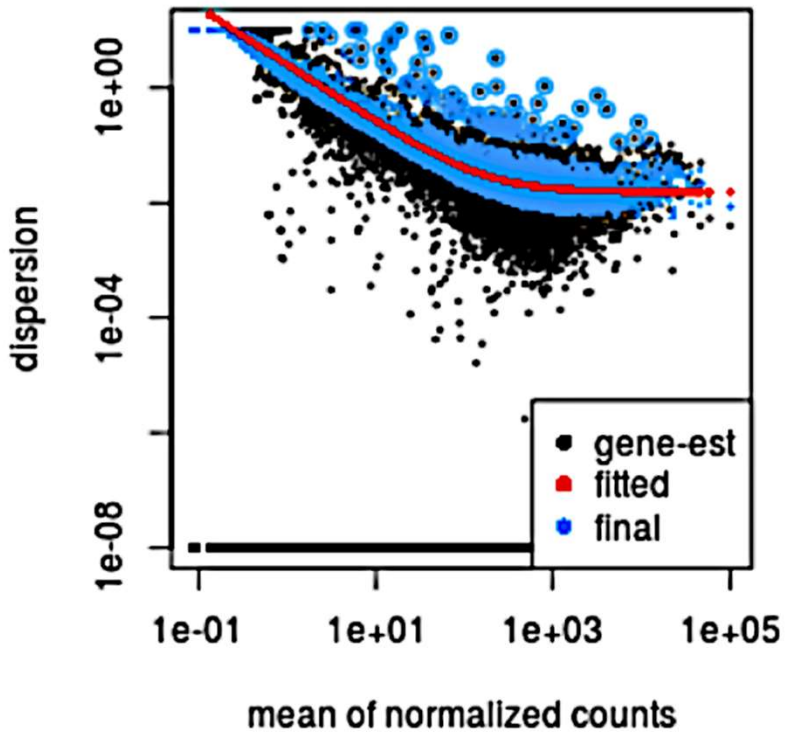
Adapted from: https://hbctraining.github.io/DGE_workshop_salmon_online/schedule/links-to-lessons.html

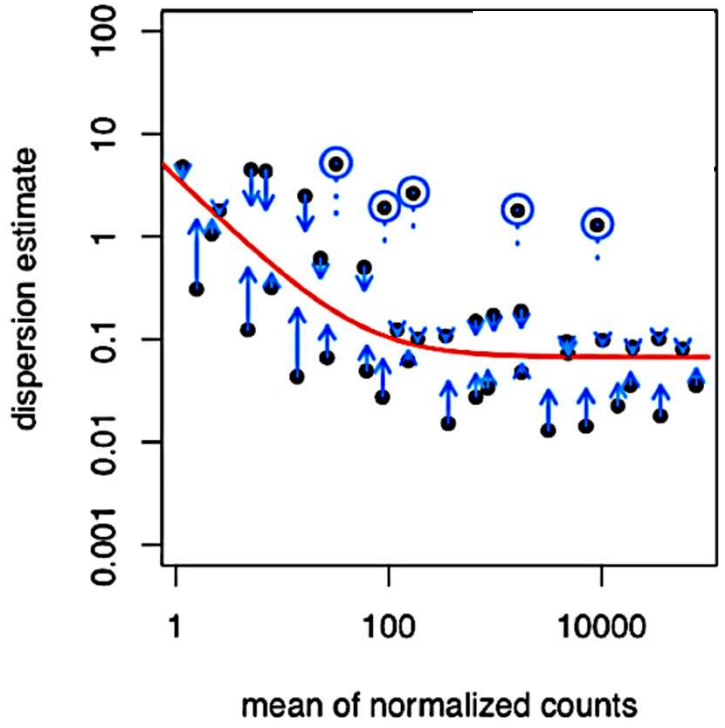
Fit curve to gene-wise estimates



Adapted from: https://hbctraining.github.io/DGE_workshop_salmon_online/schedule/links-to-lessons.html

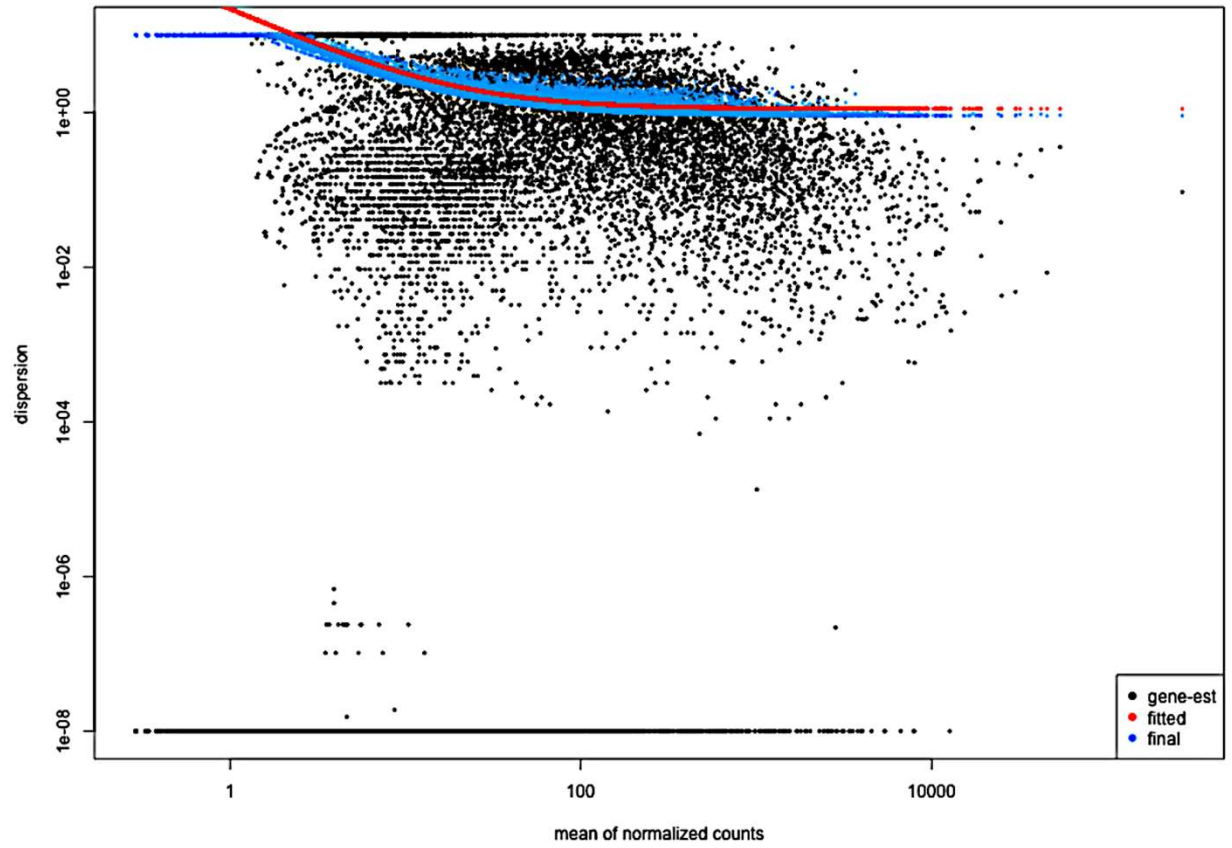
Dispersion shrinkage





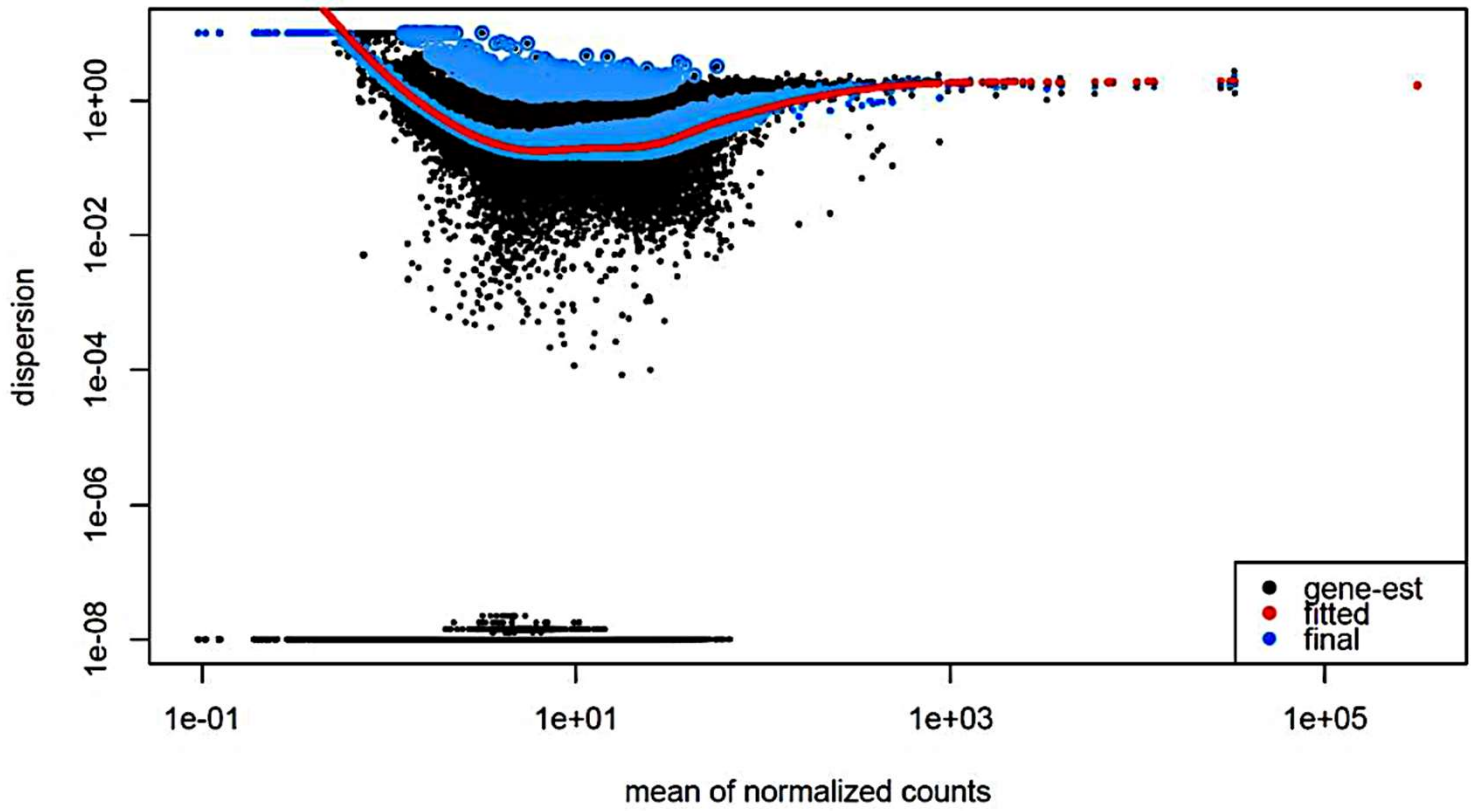
Adapted from: https://hbctraining.github.io/DGE_workshop_salmon_online/schedule/links-to-lessons.html

Concerning Plots



Adapted from: https://hbctraining.github.io/DGE_workshop_salmon_online/schedule/links-to-lessons.html

Concerning Plots

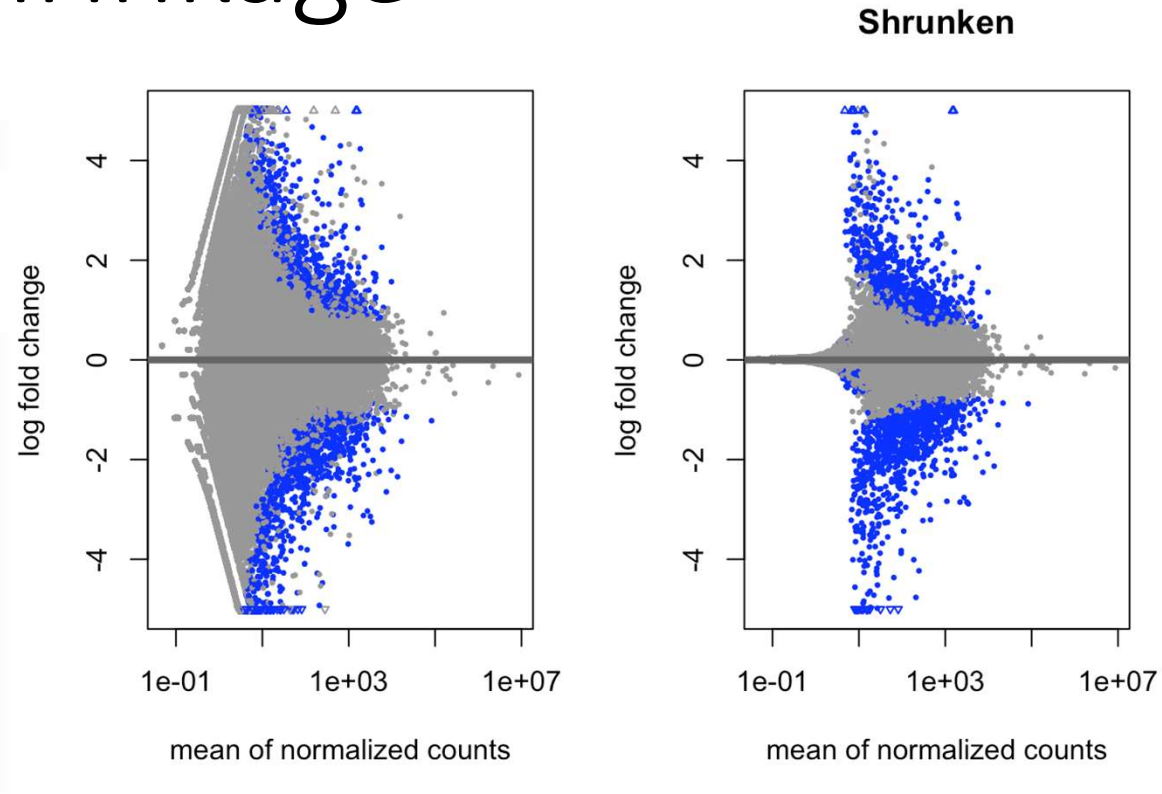


plotDispEsts(dds)

Adapted from: https://hbctraining.github.io/DGE_workshop_salmon_online/schedule/links-to-lessons.html

Exercise: Differential Expression

LFC Shrinkage

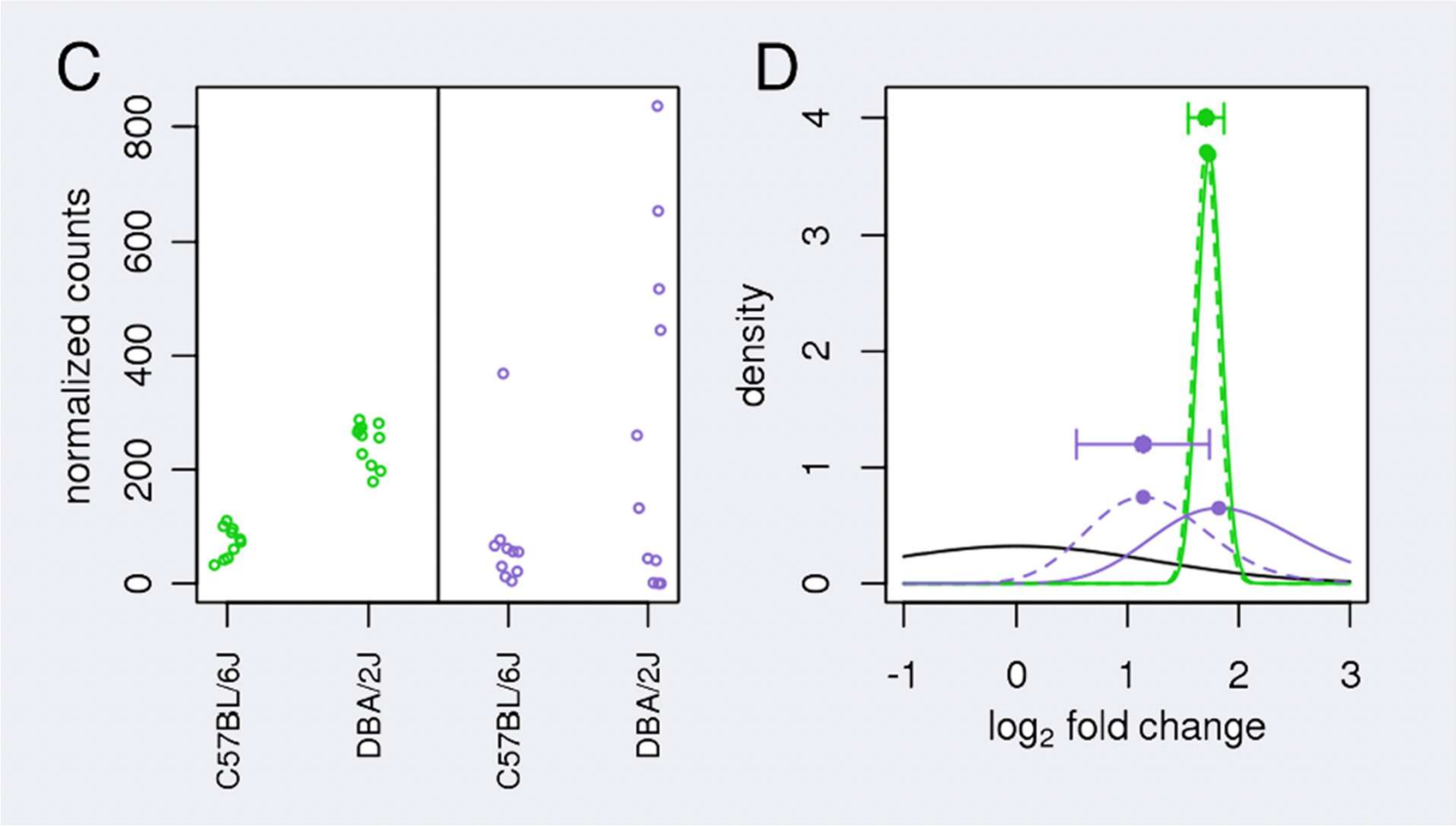


Reasons for shrinkage of LFC estimates of genes towards zero:

- Low counts
- High dispersion values

Love, M. I., Huber, W. & Anders, S. *Genome Biology* **15**, doi:10.1186/s13059-014-0550-8 (2014).

LFC Shrinkage Example



Love, M. I., Huber, W. & Anders, S. *Genome Biology* **15**, doi:10.1186/s13059-014-0550-8 (2014).

Exercise: DGE with LFC shrinkage

Visualizations

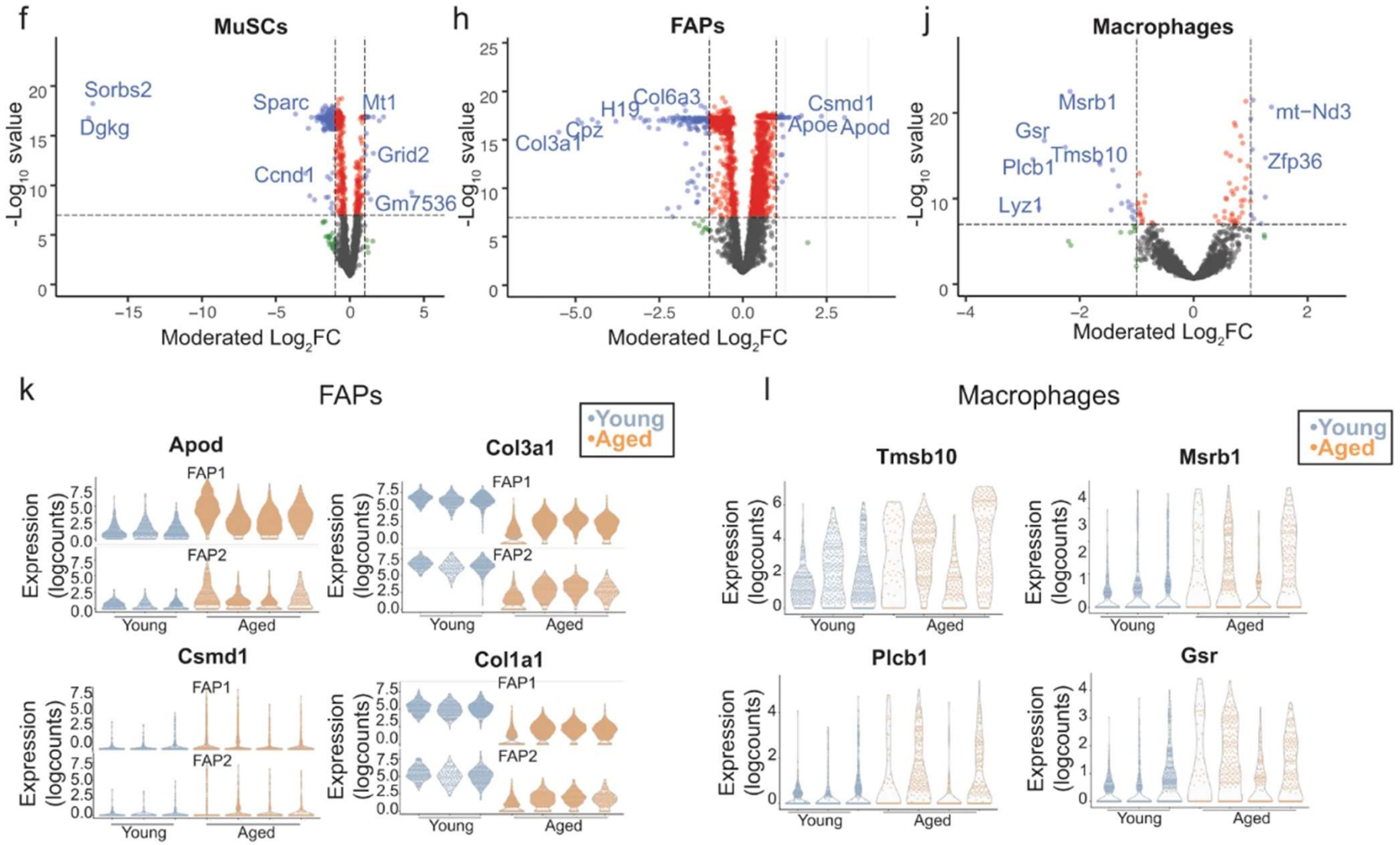


Image Source: Lazure et al., Nature Comms 2023

Exercise: Visualizations

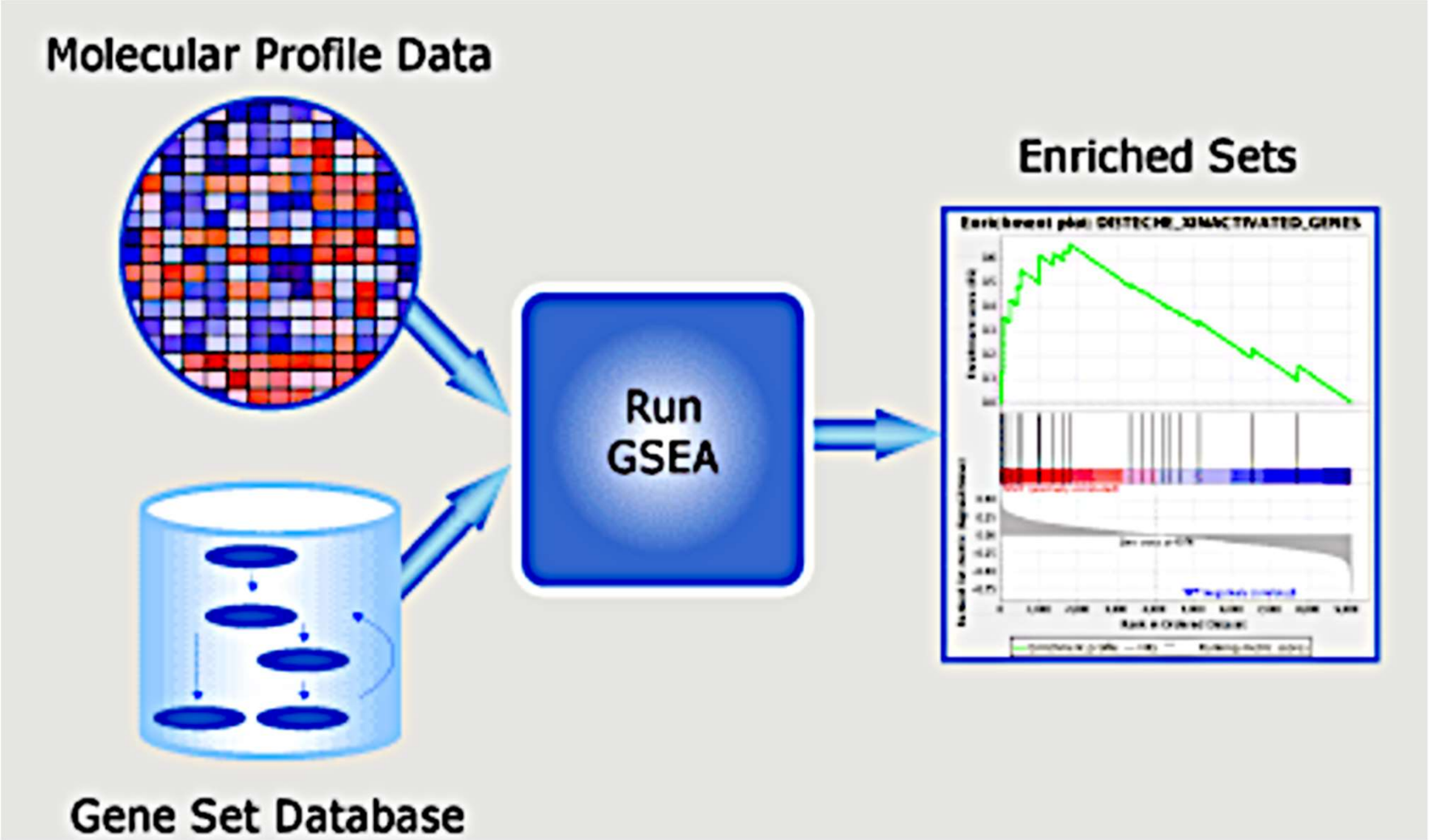
Functional Analysis

Functional Analyses

- List of DEGs may be insufficient to explain disease mechanisms
- Aim to determine what processes are represented by DEGs

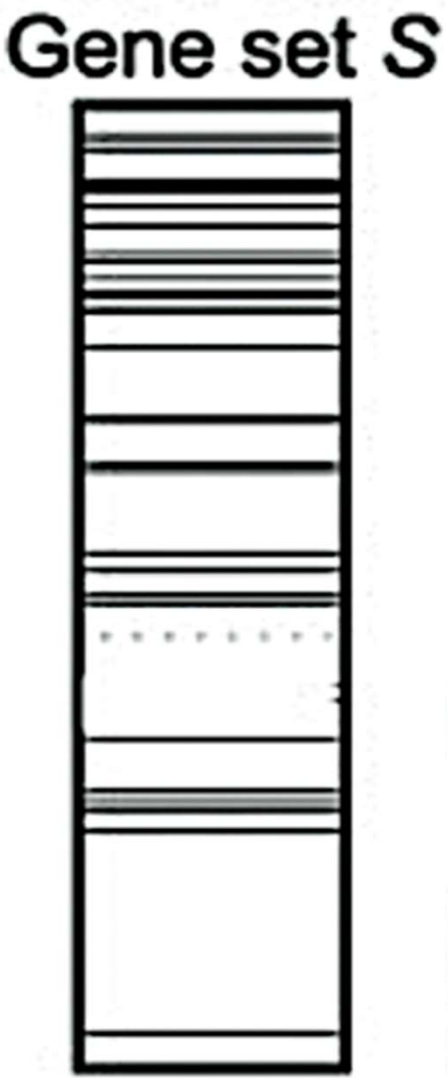
- Gene-set Enrichment Analysis (GSEA)
- Gene ontology

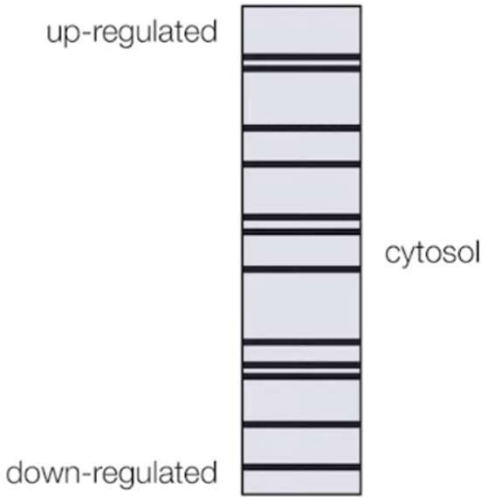
Gene set enrichment analysis (GSEA)

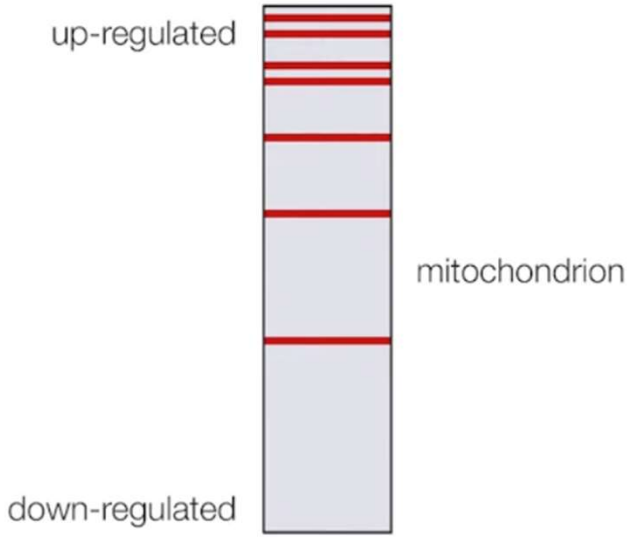


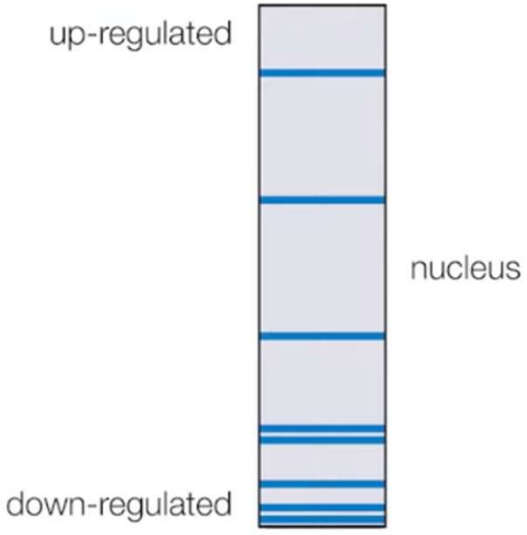
Subramanian, A. *et al. PNAS* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

Ranked List



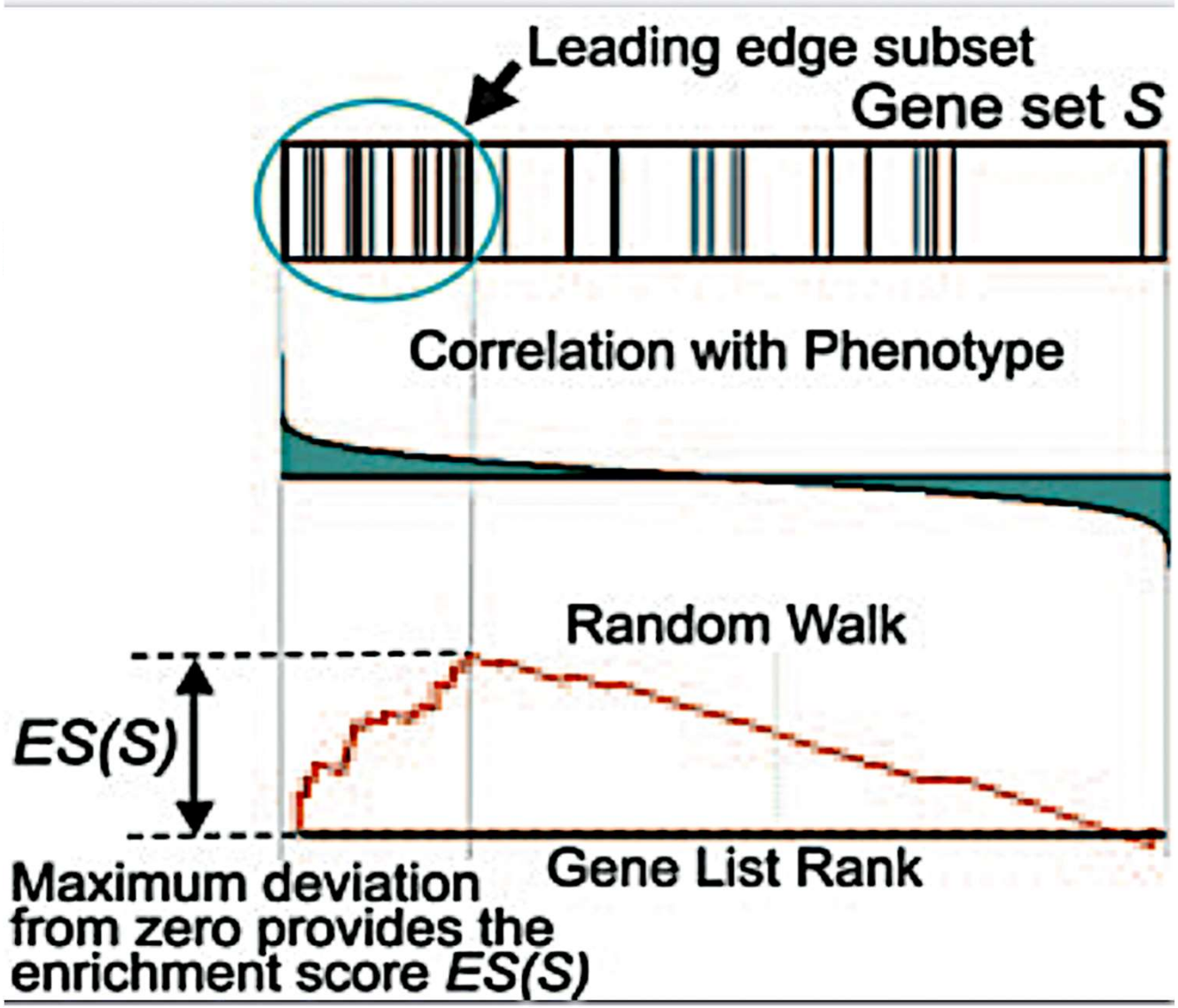






Subramanian, A. *et al. PNAS* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005). Eden, E., Navon, R.,

Enrichment Score Calculation

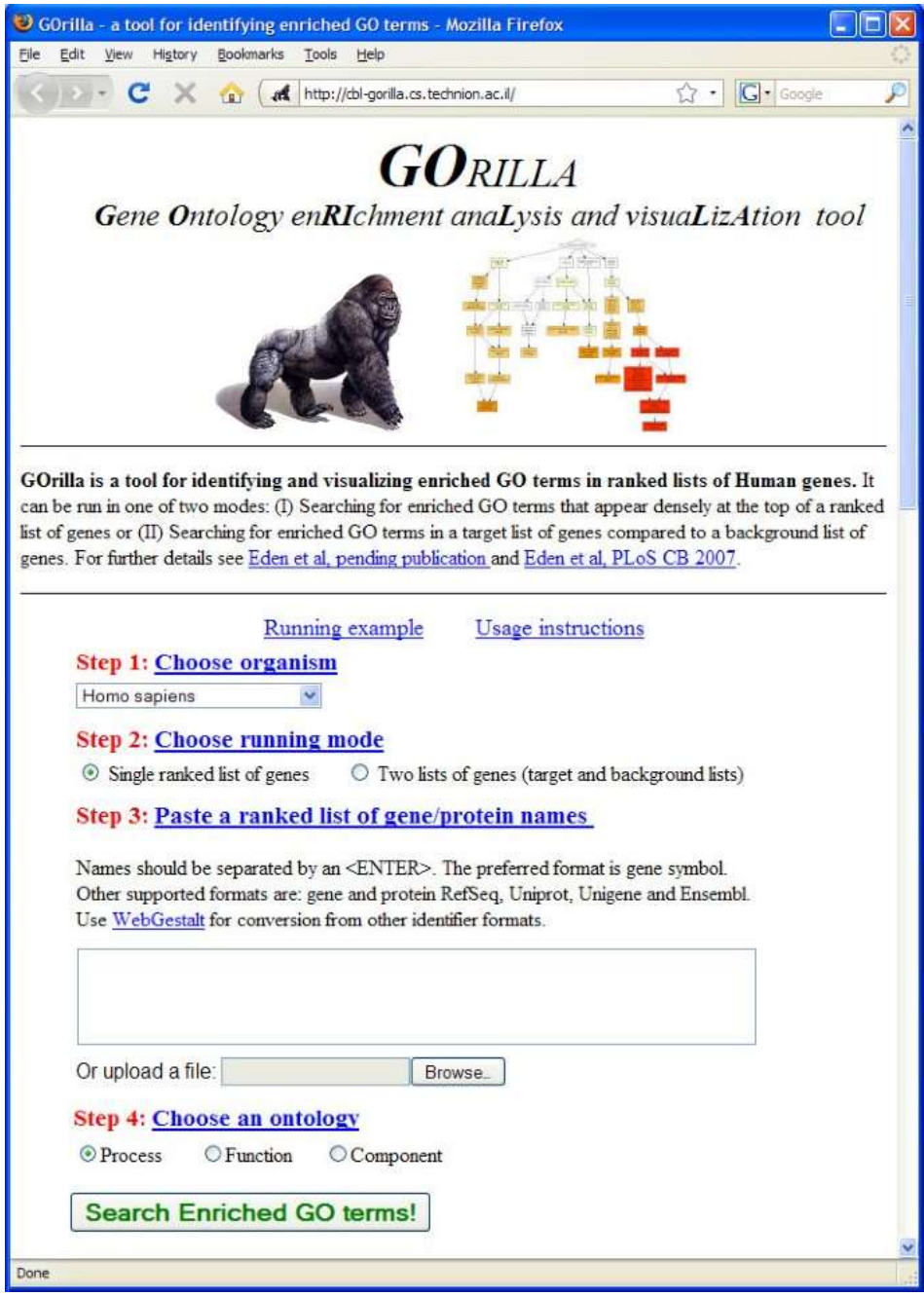


Exercise: fgsea Package

Exercise: MSigDB

Gene Ontology

- Gene Sets don't consider hierarchical structure
- **Gene Ontology** presents gene sets along a hierarchy
- Hypergeometric distribution

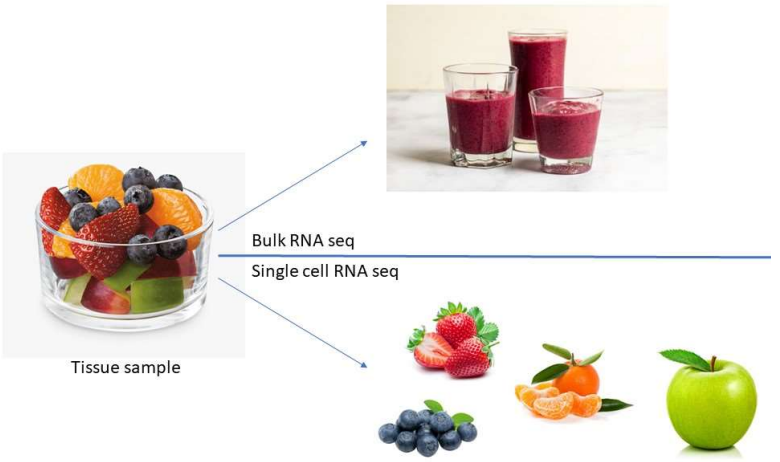


Steinfeld, I. *et al. GOrilla*: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009). https://doi.org/10.1186/1471-2105-10-48

Single Cell Data

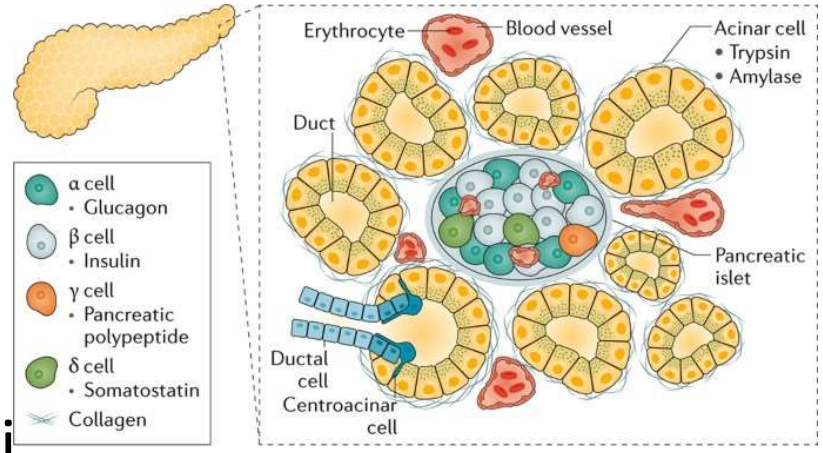
What is Single-Cell Sequencing

- Bulk Sequencing collects the **average** profile of a tissue
  - No information specific to a cell type
  - Can be biased by cell type proportions of sample
- Single-Cell sequencing collects **cell-specific** profiles of tissue
  - Have information for **individual** cells
  - Can get information for small cell populations



Why use it?

- Could collect cell type-specific bulk data by using FACS
  - Assumptions
    - Cell types of interest are known
    - Their markers are known
    - Conditions do not affect markers
- Advantages?
  - Can use the **entire** profile to identi
  - Can identify **new** cell types or **state**

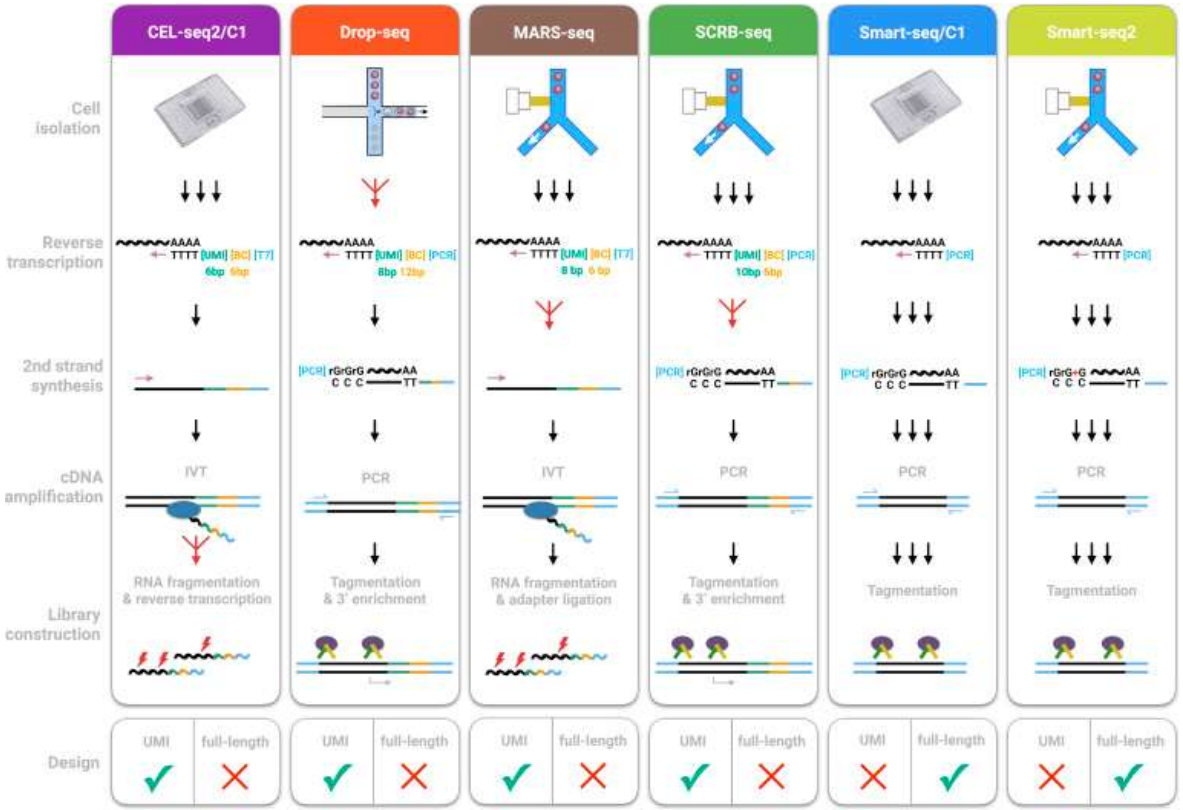


Nature Reviews | Gastroenterology & Hepatology

Ellis et al. 2017, Nature Reviews

Methods/Technology

- Cell Isolation
  - Plate-based
  - FACS-based
  - Droplet-based
- Reverse Transcription
- 2nd Strand Synthesis
- cDNA Amplification
- Library Construction

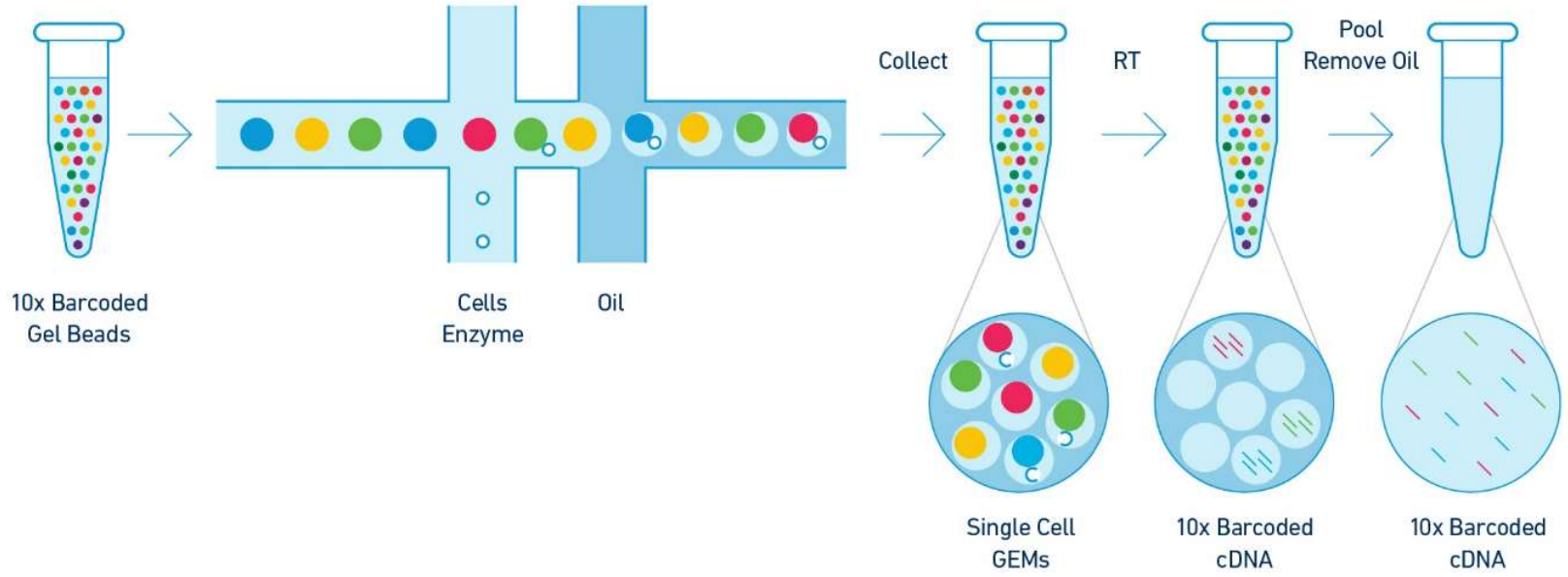


Ziegenhain et al., 2017, Molecular Cell

10X vs. SMART-seq2

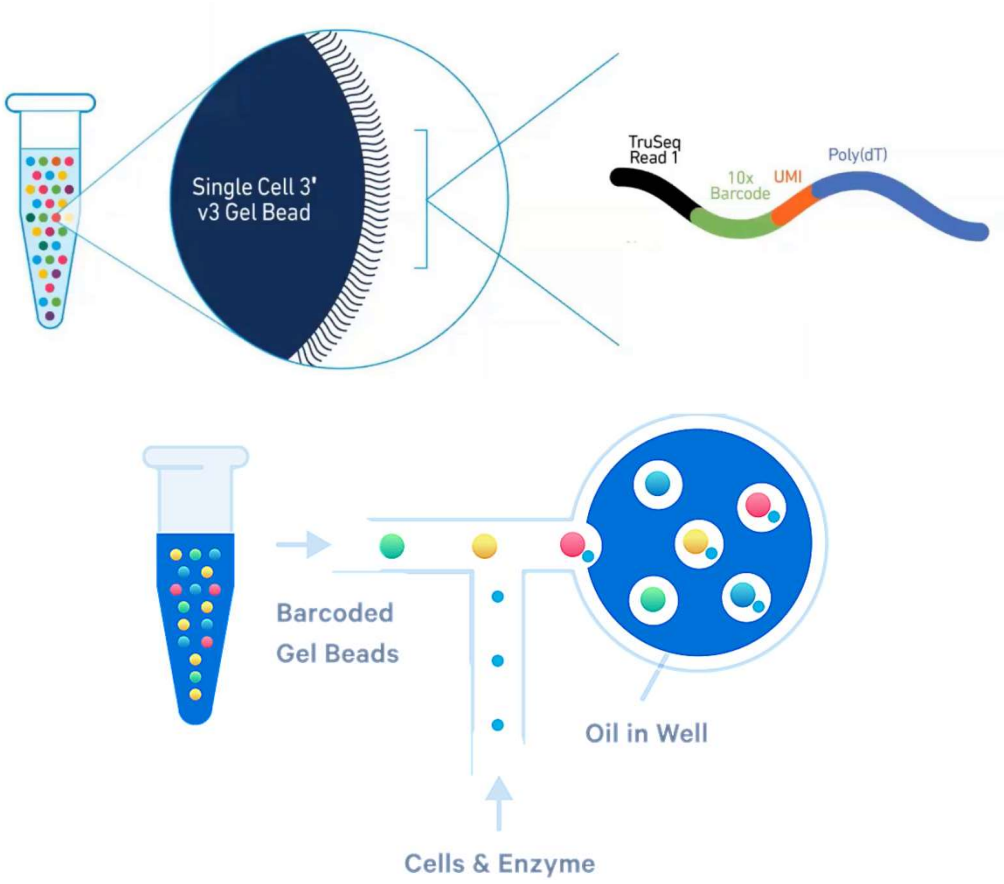
- 10X
  - Droplet-based – High throughput for cheap
  - UMI (Unique Molecular Identifier) – Can Multiplex cells and samples
  - High dropout rate
- SMART-seq2
  - Plate-based – lower throughput and more expensive
  - Full read design – More reliable gene detection and can assess splicing
  - Data is less noisy

10X Chromium

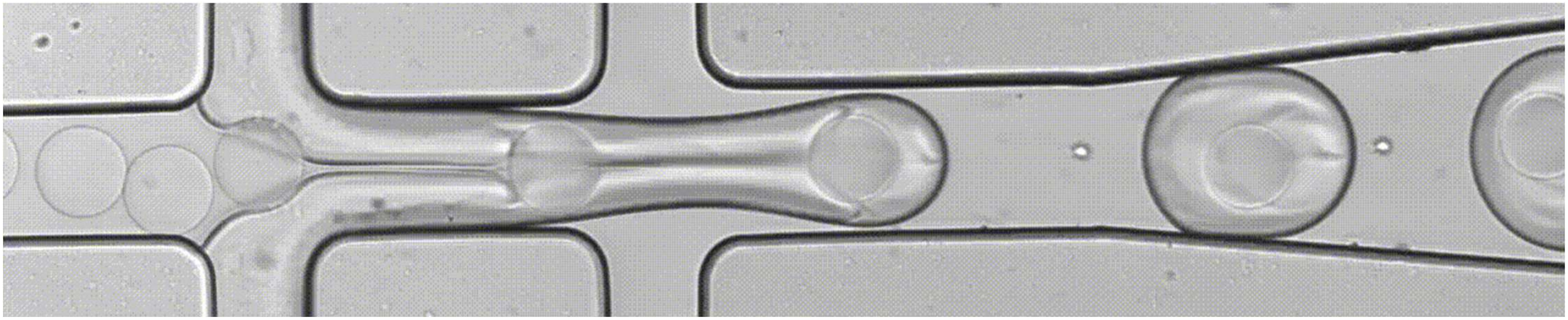


GEMs – Gel Bead Emulsion Droplets

- Gel Bead
  - UMI (Unique Molecular Id
  - Cell barcode
- Individual Cells
  - ONE cell per droplet
- Enzymes
  - Reverse Transcriptase



- Aerts lab (Katholieke Universiteit Leuven)



Sequencing

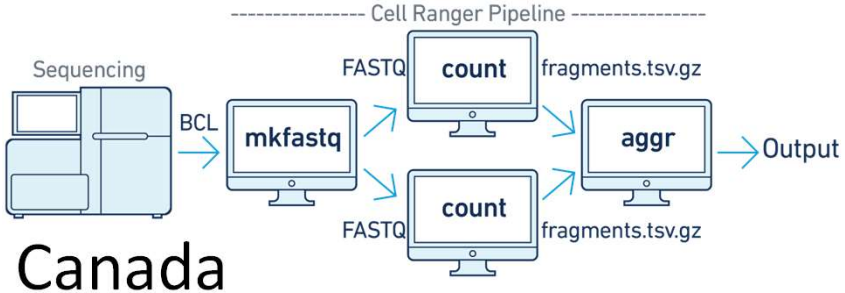
- Sequenced by standard short-read sequencing (e.g. Illumina).

- Paired-end
  - Read 1 – Barcode and UMI
  - Read 2 – mRNA transcript

- Output:
  - Base Call Files (BCL)

Processing

- Cellranger
  - UNIX-based software from 10X Genomics
  - Used to process single cell data
- Installation
  - 2.1.0 is pre-installed on Compute Canada
  - Newer versions available as .tar files on 10X website



Quality Control

- Cellranger calculates basic quality metrics for you
- Provides an html file as output which you can review
  - Read depth
  - Empty droplets etc.
- Also provides CLOUPE file with basic data exploration done
- Subsequent quality control can be done in your analysis software
  - Proportion of mt-genes
  - Sparsity etc.

Web Summary HTML

Summary | Analysis

5,527
Estimated Number of Cells

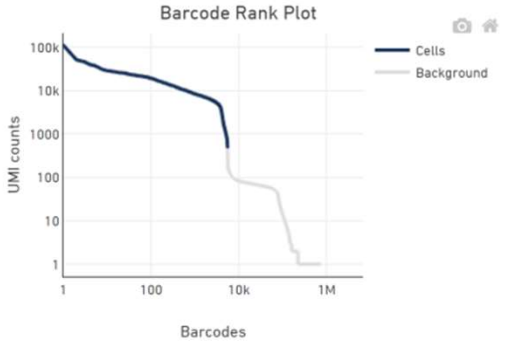
30,853
Mean Reads per Cell

1,836
Median Genes per Cell

Sequencing

Number of Reads	170,526,037
Valid Barcodes	97.7%
Valid UMIs	99.9%
Sequencing Saturation	54.8%
Q30 Bases in Barcode	95.8%
Q30 Bases in RNA Read	91.8%
Q30 Bases in Sample Index	92.0%
Q30 Bases in UMI	95.4%

Cells



Estimated Number of Cells	5,527
Fraction Reads in Cells	86.3%
Mean Reads per Cell	30,853
Median Genes per Cell	1,836
Total Genes Detected	21,421
Median UMI Counts per Cell	5,676

LOUPE Browser

SingleCellExperiment

SingleCellExperiment

- An R-based package that acts as an analysis framework for single cell data

- Also consider Seurat

Exercise: Loading 10X Data

Quality Control & Filtering

- Mitochondrial contamination
- # of Genes
- # of cells expressing gene X
- Doublet detection

Exercise: SCE Quality Control

UMAP/tSNE

- Non-linear dimension reduction techniques

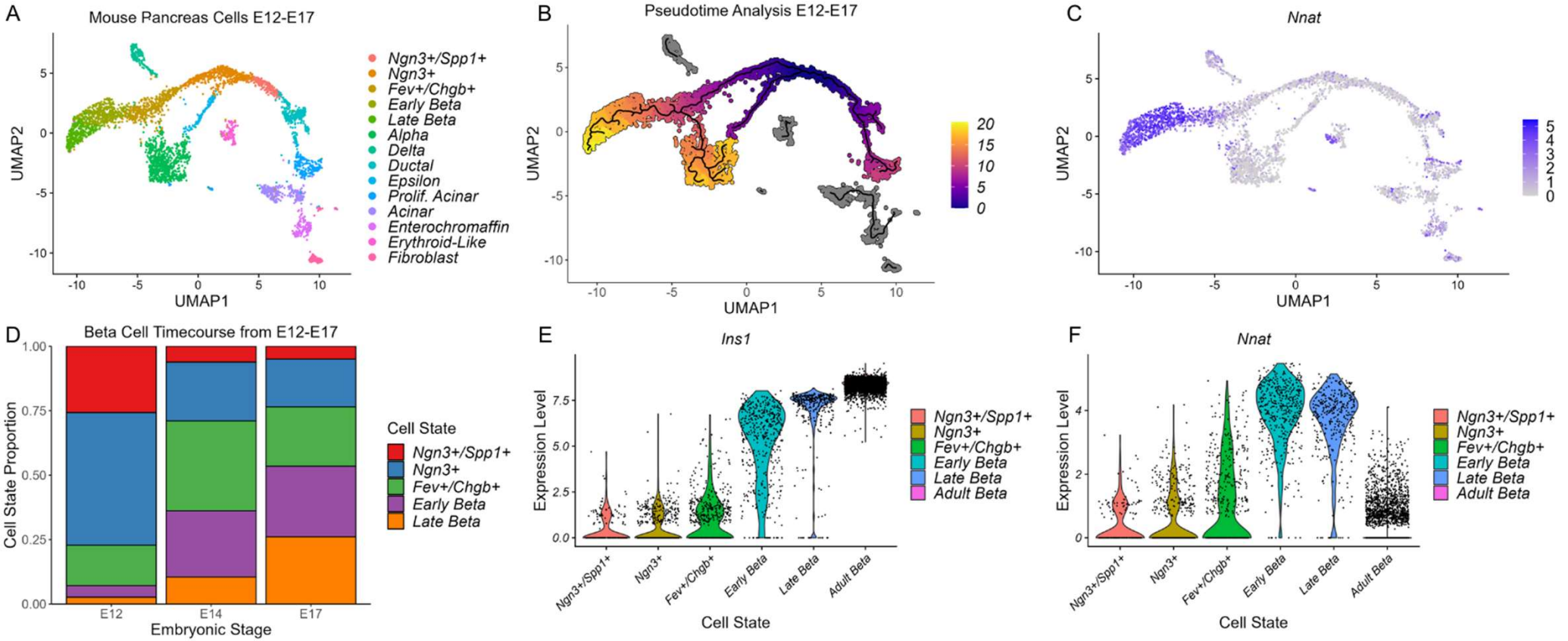
- More reliable than PCA for single cell data
  - Easier to discern cell type-specific clusters
  - Better stratification of cell states

Cell Type Labelling

- Marker Identification
- Violin plots
- Expression visualization with UMAP/tSNE

Exercise: UMAP and Cell Type Labelling

Single Cell Analyses



Closing Remarks

To summarize

- ✓ RNA-seq is a powerful tool for transcript identification AND quantification
- ✓ Experimental design decisions greatly depend on your research question
- ✓ Bulk & Single cell RNA-seq provide distinct pros and cons depending on your research question

Now you are ready to:

- Consider what experimental design you need for RNA-seq studies
- Run a standard DGE & GSEA analysis
- Use pipeline tools through the Galaxy framework
- Run basic analyses for single cell data

Acknowledgements

Material

- Reinnier Padilla (HGEN)

Exercises

- Galaxy Tutorial Page
- DESeq2 Vignettes
- OSCA eBook

Data

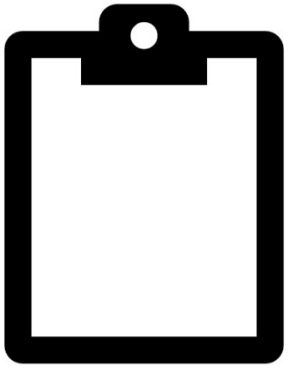
- 10X Genomics
- Stephen Turner

Thank you for attending!

1

Scan the QR code to confirm you attended today's workshop.

2



Fill out the feedback survey in the next 72h.

3

Get recognition for this workshop on your co-curricular record.