Gurdon scRNA-seq data analysis workshop – day 1

Adam Reid

Head of Bioinformatics

Gurdon Institute

21st March 2023

Course setup

Install R and RStudio

If you do not already have R and Rstudio installed, follow the instructions:

https://posit.co/download/rstudio-desktop/

Install R packages

Open R studio and, in the console, run:

install.packages('Seurat')

install.packages("sctransform")

install.packages("tidyr")

install.packages("dplyr")

Get dataset

You can download the data from DropBox (see course website)

Where these data are stored on your computer will affect some of the commands in the exercises!

Course website

https://adamjamesreid.github.io/gurdon-bioinformatics/docs/scrnaseq_workshop.html

Outline

Day 1

1. Background to scRNA-seq
2. CellRanger for mapping and read counts
3. QC – good and bad cells
4. Normalisation – why raw counts are misleading
5. Dimension reduction - high dimensional data are difficult to visualize
6. Batch effects - samples from different batches can have systematic differences that need removing
7. Clustering
8. The dataset
9. Cell Ranger demo
10. Exercises

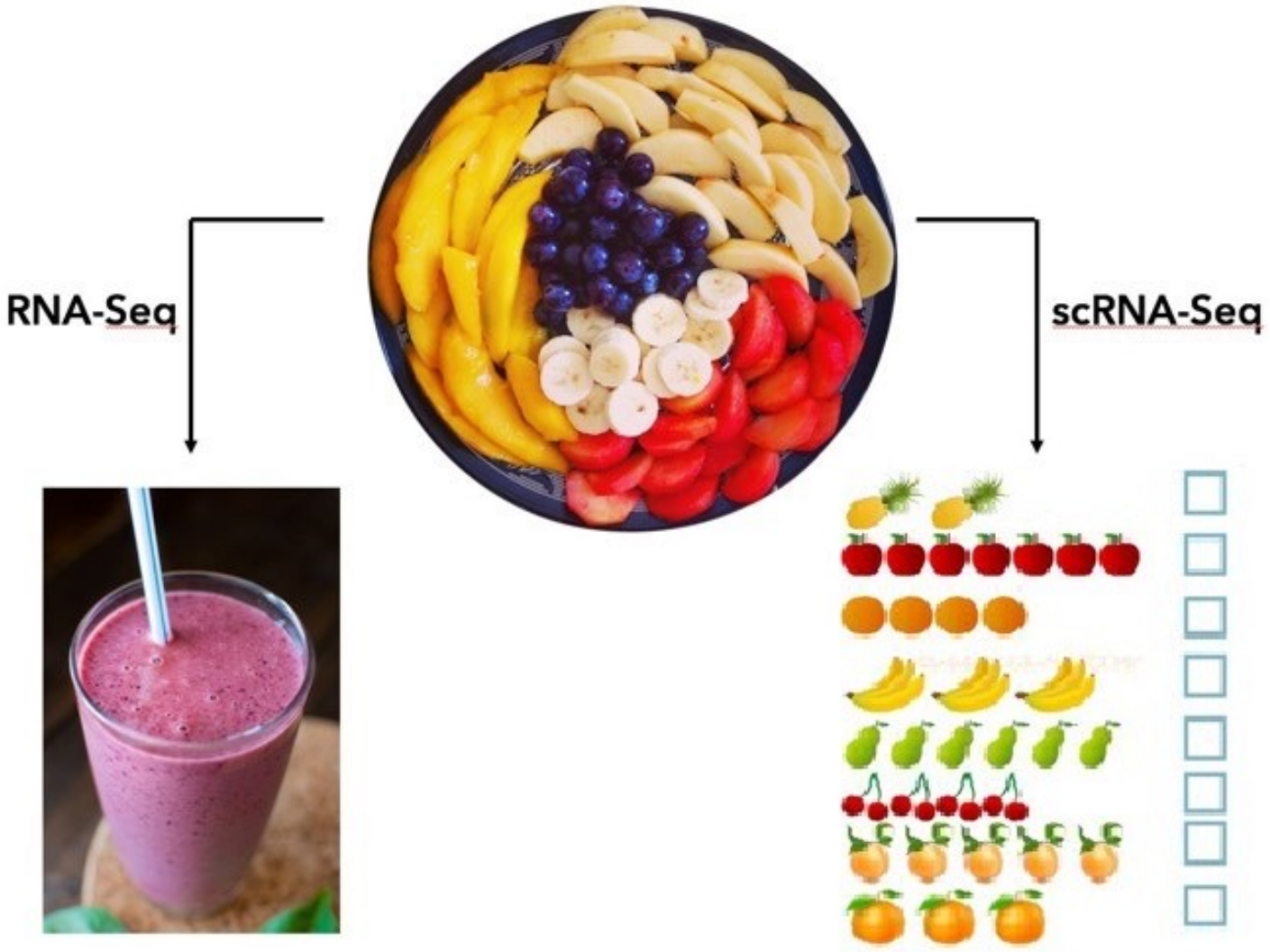
Outline

Day 2

1. Recap
2. Identifying cell types
3. Differentially expressed genes
4. Your own data

Bulk vs single cell RNA-seq

Average expression level
- Comparative transcriptomics
- Disease biomarker
- Homogenous systems



Separate populations
- Define heterogeneity
- Identify rare cell populations
- Cell population dynamics

BULK VS SINGLE CELL RNA-SEQ

1. mRNA: TruSeq RNA-Seq (Gold Standard)
   - ~20,000 transcripts
     - More when consider splice variants / isoforms
   - Observe 80-95% of transcripts depending on sequencing depth

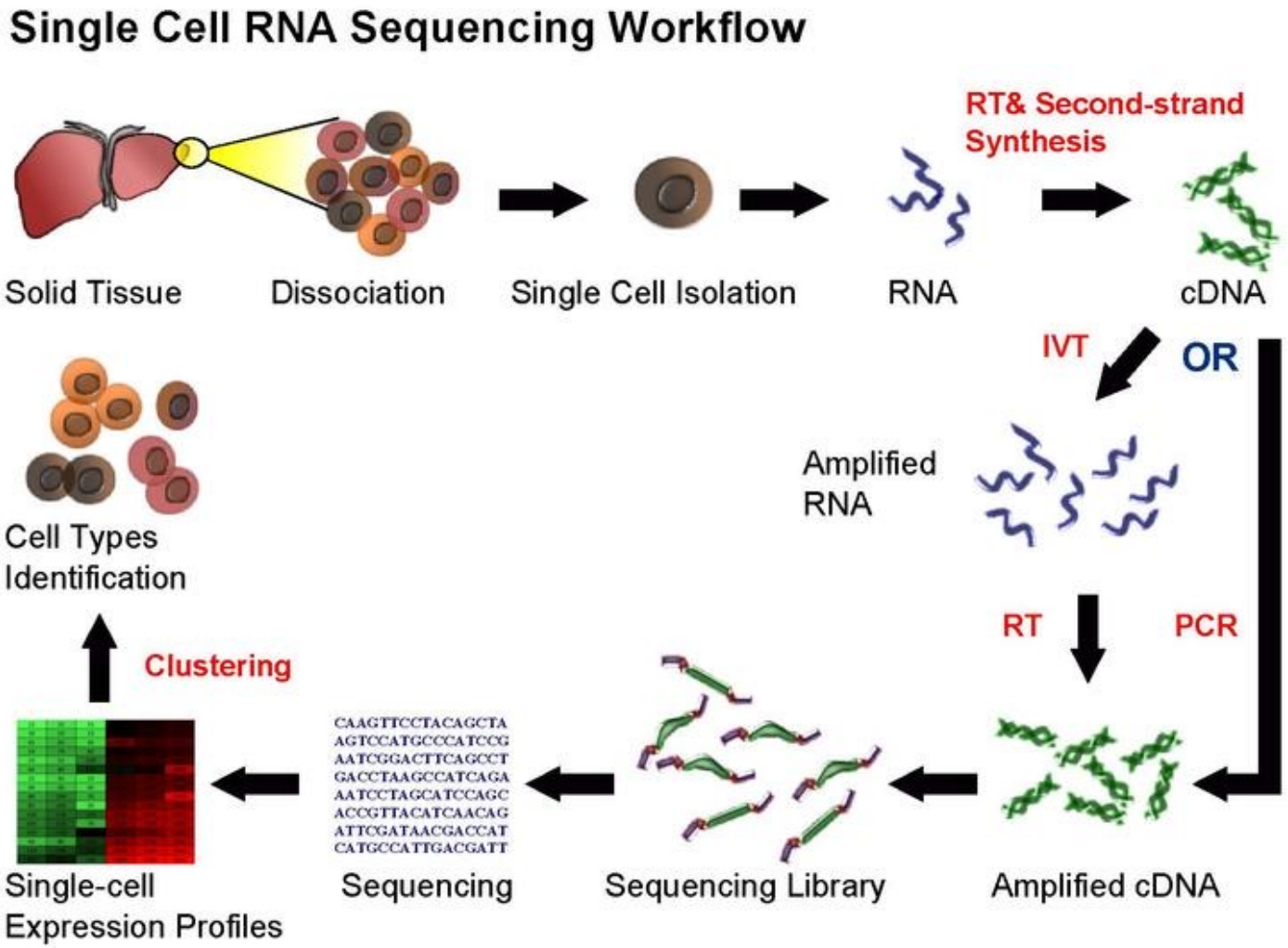
2. Single Cell Methods
   - 200 -10,000 transcripts per cell
   - Observe 10-50% of the transcriptome
   - Many transcripts will show up with zero counts in every cell. (even GAPDH)
   - If you only looked at transcripts observed in all cells numbers drop dramatically.

Disadvantages of scRNA-seq
- Dropouts and noisy data
- Lowly expressed genes might be undetected
- Samples will contain doublets
- Replicates without batch effect are unlikely
- Expensive

Source: Sarah Boswell, Harvard Medical School, September 2020

workflow

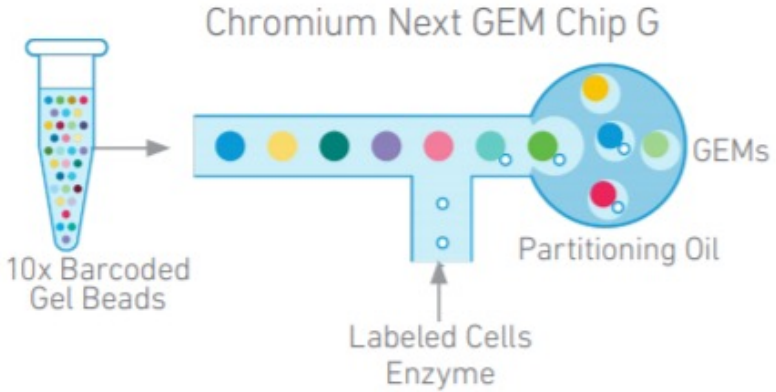
Good sample preparation is key to success!

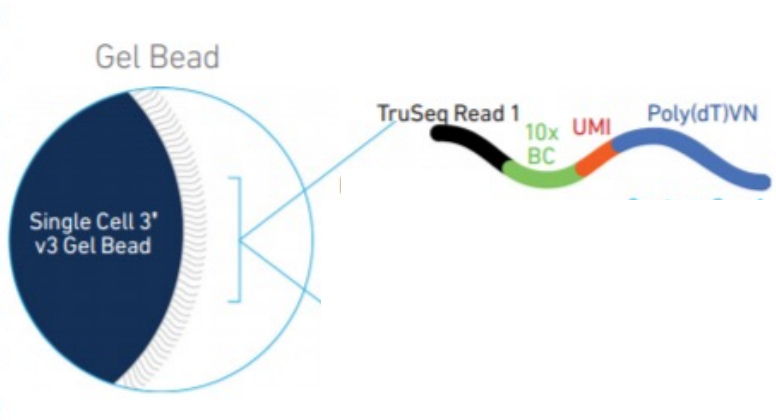


Source: https://en.wikipedia.org/wiki/Single_cell_sequencing

10x Genomics Chromium scRNA-seq platform overview

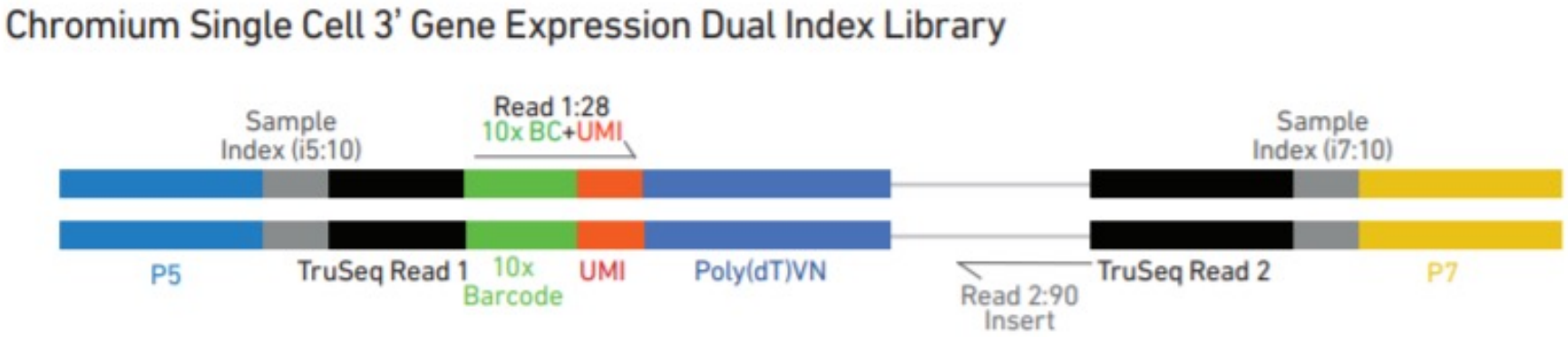
- Droplet-based capturing, either 3’ or 5’ mRNA
- Uses soft **gel beads** containing oligos. These enable “single Poisson loading” leading to capture of >60% of input cells.
- Standardized instrumentation and reagents (unhackable so no customisation or control)
- Very easy to use and less processing time
- More high-throughput scaling - 8 samples can be processed simultaneously with up to 10000 cells captured per sample
- The doublet rate increases with number of cells loaded
- CellRanger and CellLoupe software are available and user friendly



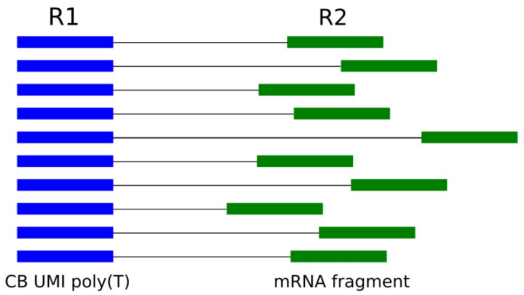


Source: 10x Genomics

10X Chromium libraries



Source: 10x Genomics



Sequencing Read	Description	Number of cycles
Read1	10x Barcode Read (Cell) + Randomer Read (UMI)	28bp
i7 index	Sample index read	10bp
i5 index	Sample index read	10bp
Read2	Insert Read (Transcript)	90bp

What platform Should I use?

Choose protocol based on:

- Throughput (number of cells per reaction)
- Sample of origin
- Cost / Labour / Time limitations
- Gene body coverage: 5'/ 3' biased or full-length?
- UMI vs no-UMI
- Sequencing depth per cell

Examples:

- If you sample is fairly homogeneous – bulk RNAseq
- If your sample is limited in cell number – plate-based method
- If you want re-annotate the transcriptome and discover new isoforms – full-length coverage (SMART-seq2, seqWell)
- If you are looking to classify all cell types in a diverse tissue - high throughput
- If you have only archival human samples – nuclei isolation

Experimental design

Firstly, what is the question?

- Classifying the different cell types in a sample? — Replication advised
- Rare/novel cell population? — Replication advised
- Gene expression differences between particular cell types in different individuals/conditions? — Replication essential
- Changes in cell number between conditions? — Replication essential
- Changes in expression through development? — Replication advised

Replication

- For some questions good replication is essential. For others it is advised.
- The best statistical approaches for DE in scRNA-seq are the same as for bulk, so ideally triplicates.
- The appropriate way to make replicates depends on your experiment and you should consult a bioinformatician to help with experimental design

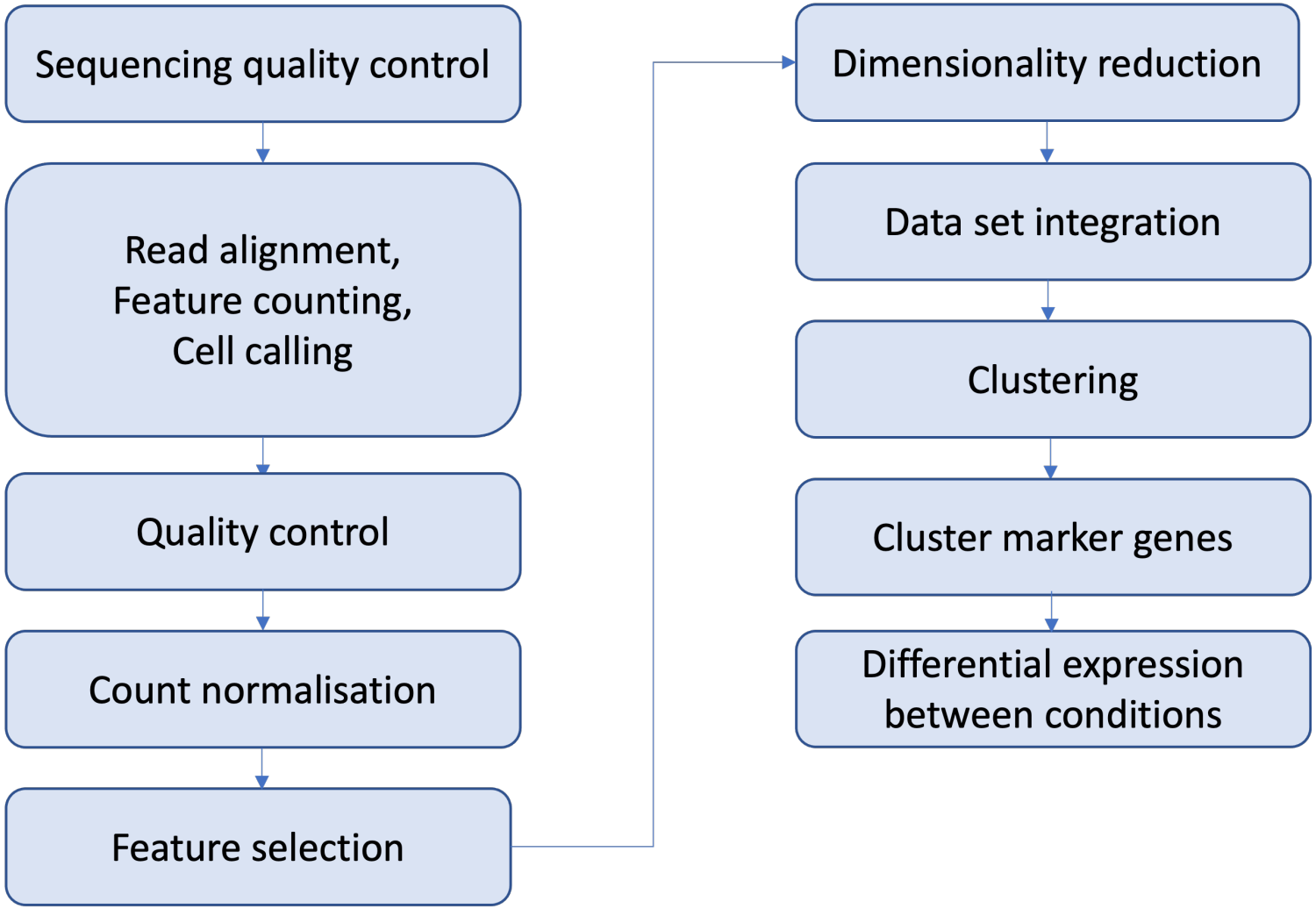
How many cells per samples?

- How rare is your rarest cell type of interest? How many cells might you need to compare cell numbers/gene expression levels in these cells?
- N.b. the number of cells going into the experiment is much more than you will get out!

Sequencing depth

- Aim for 100,000 reads per cell

Analysis workflow



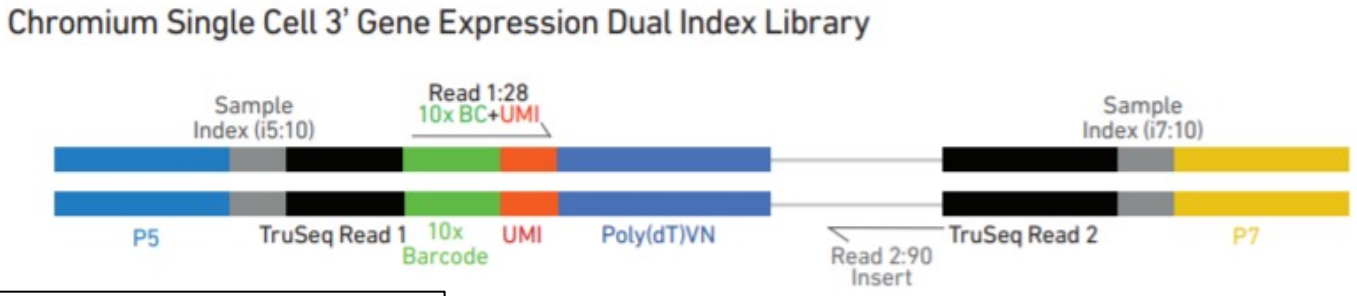
10X chromium library structure

The 10x library contains four pieces of information, in the form of DNA sequences:

- **sample index** - identifies the library, with one or two indexes per sample
- **10x barcode** - identifies the droplet in the library
- **UMI** - identifies the transcript molecule within a cell and gene
- **insert** - the transcript sequence

The sequences for any given fragment will generally be delivered in 3 or 4 files:

- **I1**: I7 sample index
- **I2**: I5 sample index if present (dual indexing only)
- **R1**: 10x barcode + UMI
- **R2**: insert sequence



<SampleName>_S<SampleNumber>_L00<Lane>_<Read>_001.fastq.gz
E.g.
SRR9264343_S0_L001_I1_001.fastq.gz
SRR9264343_S0_L001_R1_001.fastq.gz
SRR9264343_S0_L001_R2_001.fastq.gz

Mapping and counting reads

The reference sequence – fasta format

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

Reference annotation – gtf format

chr1	tool	gene	11218	15435	.	+	.	ID=gene1
chr1	tool	mRNA	11218	15435	.	+	.	ID=transcript1;Parent=gene1
chr1	tool	exon	11218	13000	.	+	.	ID=exon1;Parent=transcript1
chr1	tool	exon	13800	14002	.	+	.	ID=exon2;Parent=transcript1
chr1	tool	exon	15000	15360	.	+	.	ID=exon3;Parent=transcript1
chr1	tool	exon	15384	15435	.	+	.	ID=exon4;Parent=transcript1
chr1	tool	UTR5	11218	12000	.	+	.	ID=UTR5a;Parent=transcript1
chr1	tool	CDS	12801	13000	.	+	0	ID=CDS1;Parent=transcript1
chr1	tool	CDS	13800	14002	.	+	0	ID=exon1;Parent=transcript1
chr1	tool	CDS	15000	15234	.	+	0	ID=exon1;Parent=transcript1
chr1	tool	UTR3	15234	15360	.	+	.	ID=UTR3a;Parent=transcript1
chr1	tool	UTR3	15384	15435	.	+	.	ID=UTR3b;Parent=transcript1

Reads – fastq format

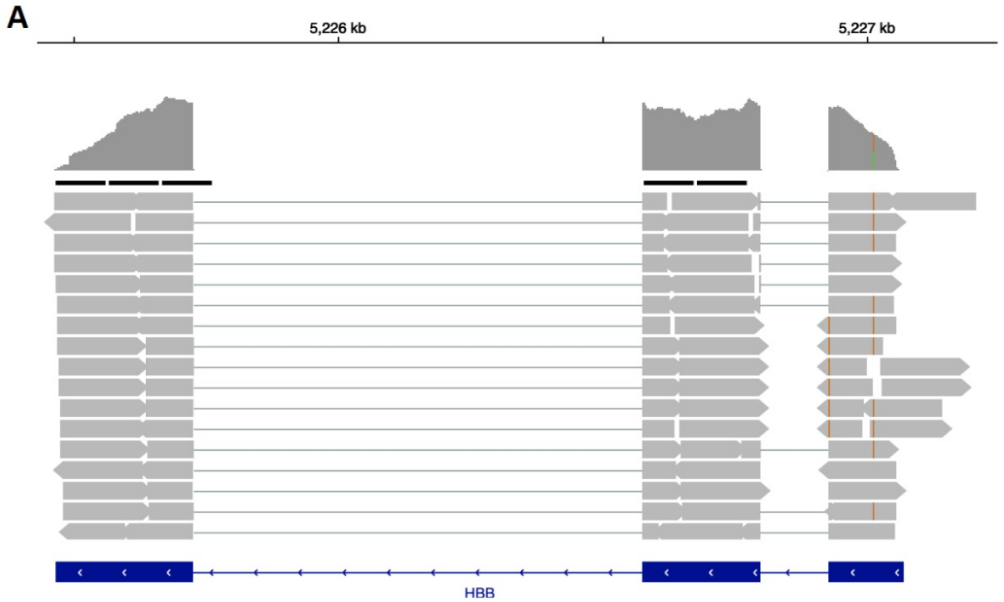
There are two popular sources of assembly files:
UCSC (hg19, hg38, mm10, etc)
GRC (GRCh37, GRCh38, GRCm38).

Major releases of UCSC and GRC assemblies are matched in main chromosomes (e.g. chr1 from hg38 = chr1 from GRCh38), but differ in additional contigs and so-called ALT loci, which change between minor releases

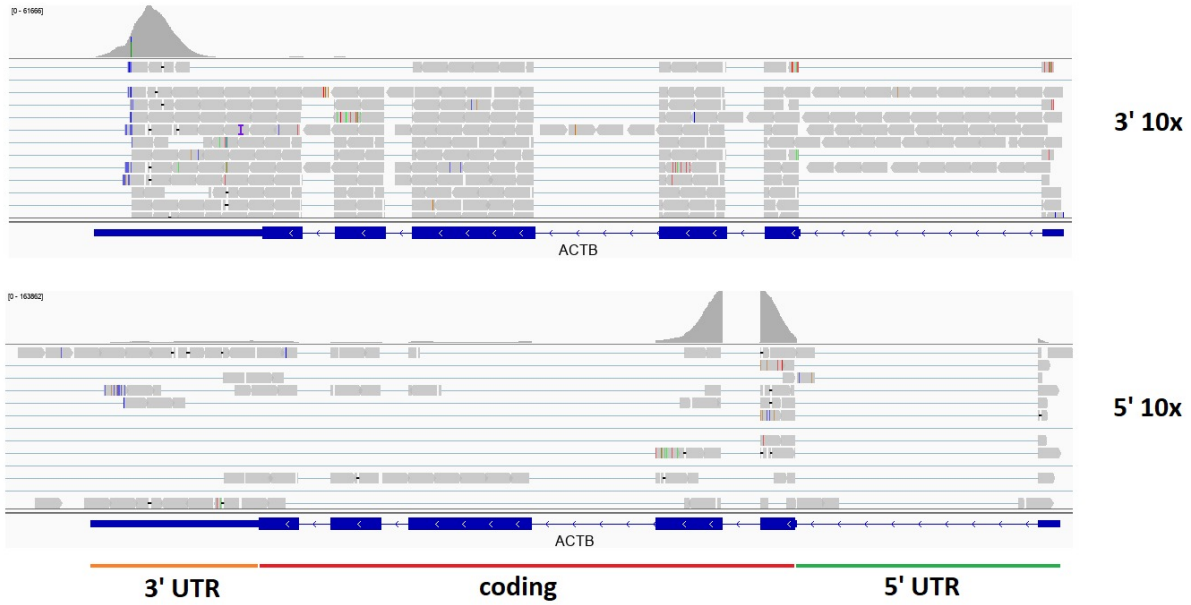
Popular sources of human and mouse genome annotation are RefSeq, ENSEMBL, and GENCODE. RefSeq is the most conservative of the three, and tends to have the fewest annotated transcripts per gene. RefSeq transcript IDs start with NM_ or NR, e.g. **NM_12345**.

ENSEMBL and GENCODE are very similar to each other and can be used interchangeably for our purposes. Gene names in these start with ENSG (for human) and ENSMUSG (for mouse); transcripts start with ENST and ENSMUST, respectively.

Mapping and counting reads



Bulk RNA-seq



3’ and 5’ 10X scRNA-seq

```
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
```

Mapped reads are stored in BAM files

Read counts are stored in a count matrix

Cell Ranger outputs

```
File  Edit  View  Search  Terminal  Help
_versions
%h%-$
%h%-$ ls SRR9264343/outs/
analysis
cloupe.cloupe
filtered_feature_bc_matrix
filtered_feature_bc_matrix.h5
metrics_summary.csv
molecule_info.h5
possorted_genome_bam.bam
possorted_genome_bam.bam.bai
raw_feature_bc_matrix
raw_feature_bc_matrix.h5
web_summary.html
%h%-$
```

10x GENOMICS Cell Ranger • count

SITTA6

Summary | Analysis

14,668
Estimated Number of Cells

20,065
Mean Reads per Cell

1,344
Median Genes per Cell

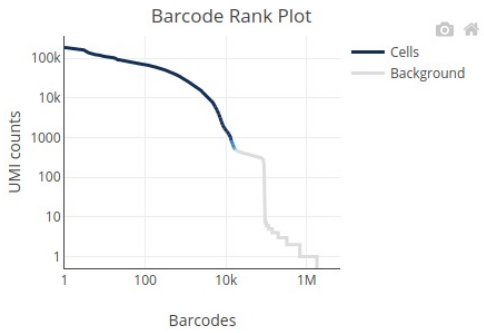
Sequencing

Number of Reads	294,310,066
Number of Short Reads Skipped	0
Valid Barcodes	97.7%
Valid UMIs	100.0%
Sequencing Saturation	18.6%
Q30 Bases in Barcode	96.1%
Q30 Bases in RNA Read	94.6%
Q30 Bases in UMI	95.7%

Mapping

Reads Mapped to Genome	93.6%
Reads Mapped Confidently to Genome	89.7%

Cells

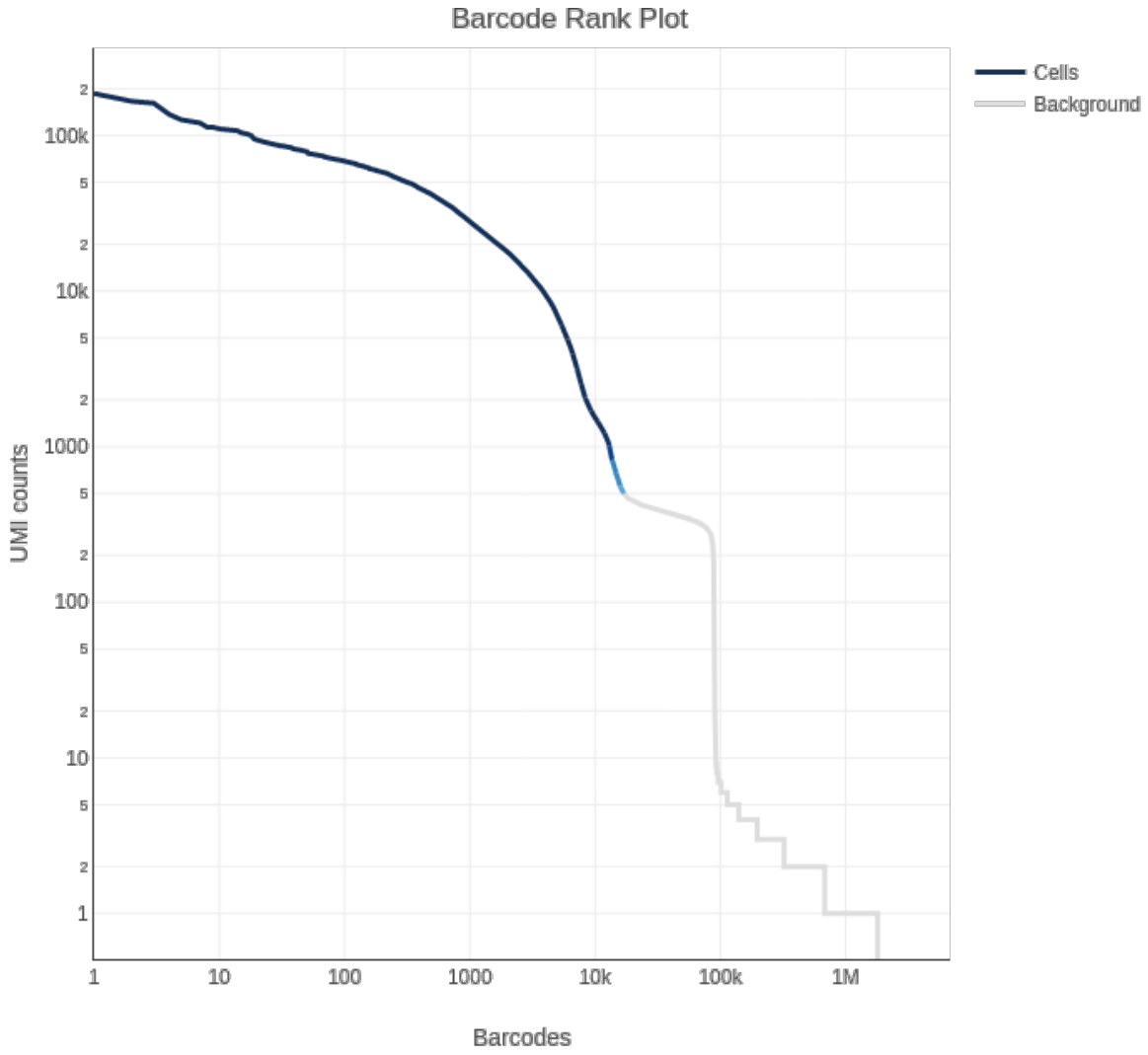


Estimated Number of Cells	14,668
Fraction Reads in Cells	80.8%
Mean Reads per Cell	20,065
Median Genes per Cell	1,344
Total Genes Detected	23,106
Median UMI Counts per Cell	2,928

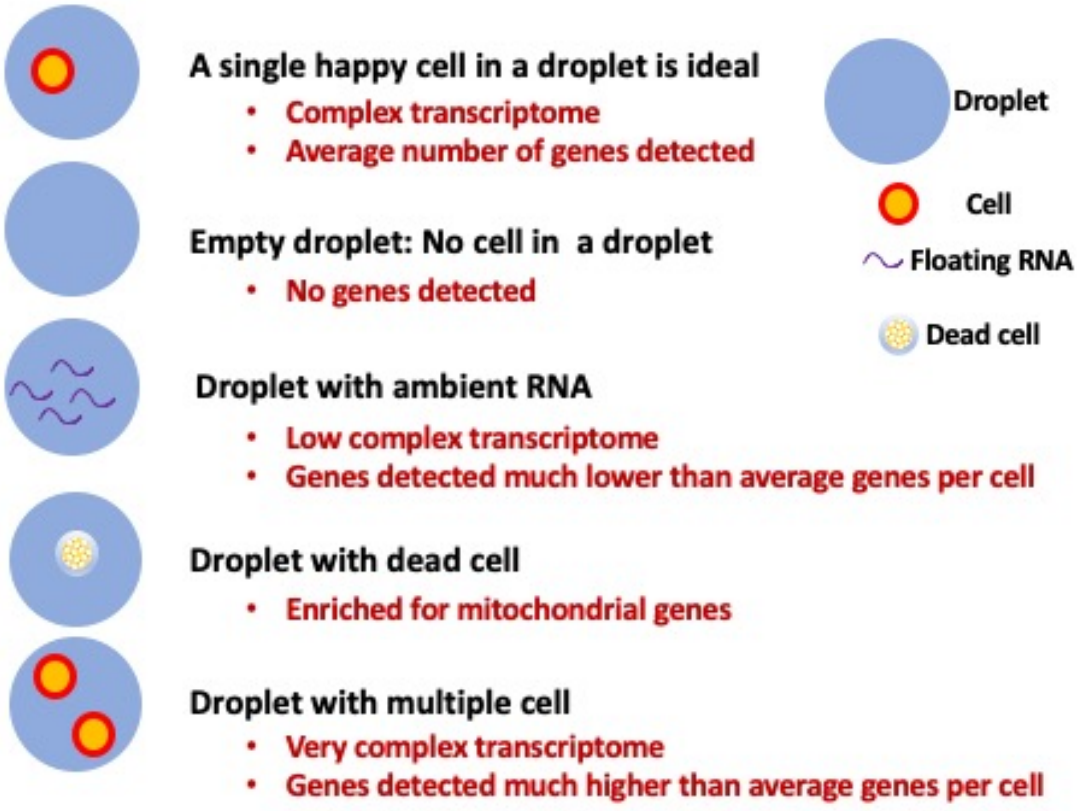
Sample

Sample ID	SITTA6
Sample Description	

Cell Ranger cell calling



Quality Control - QC



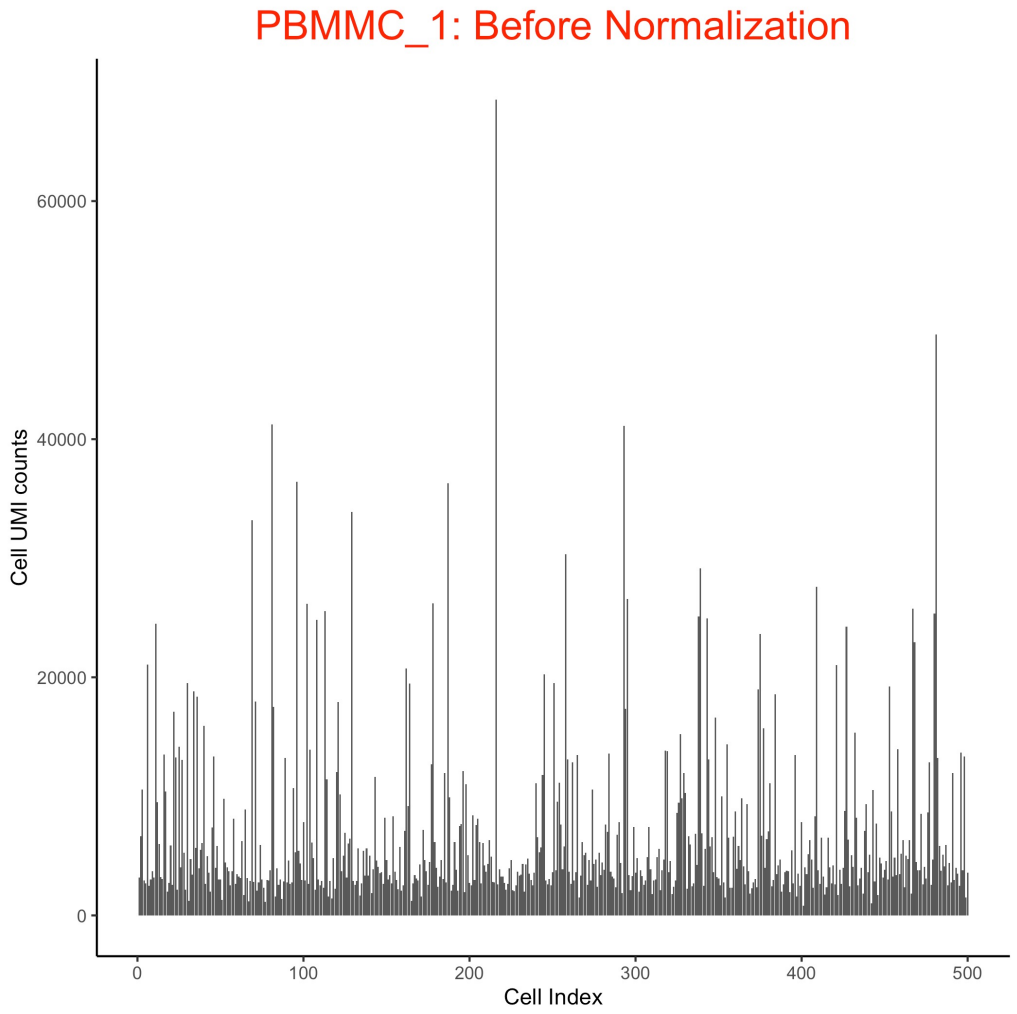
Aim of QC is:

- To remove undetected genes
- To remove empty droplets
- To remove droplets with dead cells
- To remove Doublet/multiplet cells
- Ultimately to filter the data to only include true cells that are of high quality

This is achieved by applying cut-offs on:

1. Number of genes detected per cell
2. Percent of mitochondrial genes per cell
3. Number of UMIs/transcripts detected per cell

Normalisation



We derive biological insights downstream by comparing cells against each other.

But the UMI count differences makes it harder to compare cells.

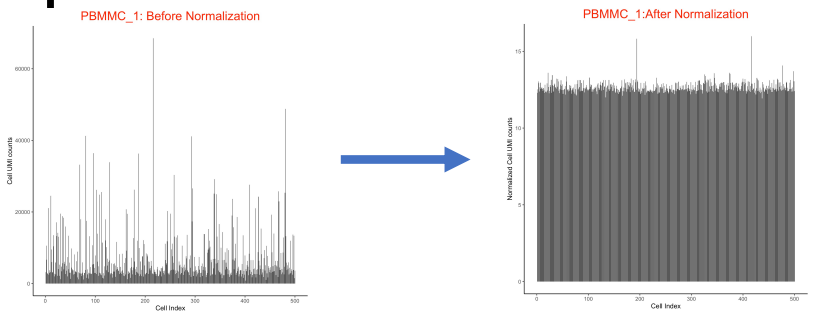
Why do total transcript molecules (UMI counts) differ between cells?

Biological:
- Cell type differences: size, transcriptional activity

Technical:
- scRNA data is inherently noisy
- Low mRNA content per cell
- Cell-to-cell differences in mRNA capture efficiency
- Variable sequencing depth
- PCR amplification efficiency

Normalization reduces technical differences to increase the relative strength of biological differences, allowing meaningful comparison of expression profiles between cells.

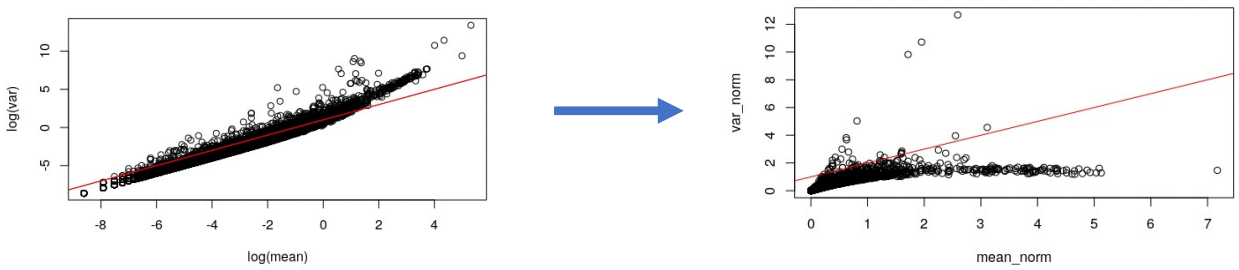
Normalisation – general principle



Normalization has two steps:

1. Scaling – correct depth bias (and composition bias)
- Calculate size factors or normalization factors that represents the relative depth bias in each cell
- Scale the counts for each gene in each cell by dividing the raw counts by a size factor

2. Transformation – correct mean-variance bias
- log2 (e.g. Deconvolution)
- Pearson residuals (eg. sctransform)



Normalisation approaches used for bulk RNA-seq are not appropriate for scRNA-seq. Here we use scTransform, which takes care of scaling and transformation for single-cell RNA-seq data

scTransform does both steps in one by regressing on library size and taking the Pearson residuals!

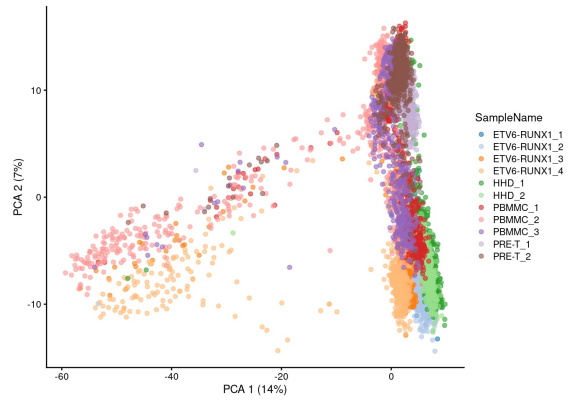
Dimension reduction

scRNA-seq data are high dimensional i.e. can be represented by a number of dimensions equal to the number of genes. **But I can't see in 30,000 dimensions!**

Not all 30,000 dimensions are useful – many have little or no information. Some contain the same information as others.

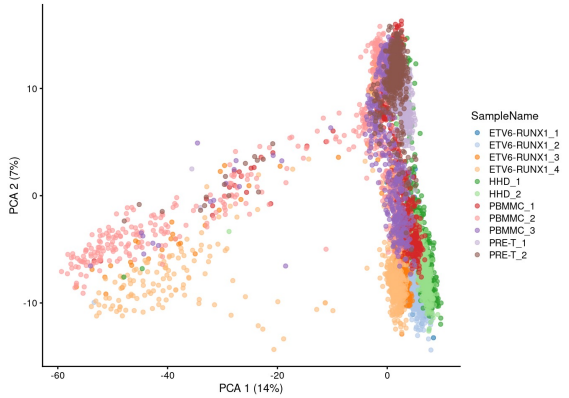
1. Identify the highly variable genes (HVGs) – those which are "doing something" in the data
2. Principal components analysis (PCA) – find a smaller number of linear paths through the n dimensions which capture as much variation in gene expression as possible, each across multiple genes. PC1 has the most variance of any PC, PC2 the next most…. We might select 10 or 20 PCs, instead of 30,000 genes.

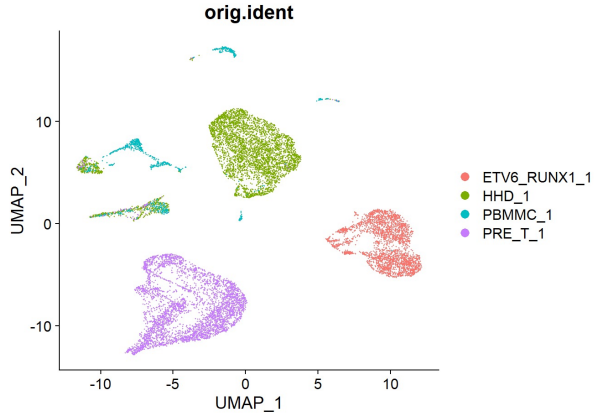
	Cell1	Cell2	…	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
…	.	.	.	.
…	.	.	.	.
…	.	.	.	.
GeneM	25	0	.	0



Dimension reduction - UMAP

- Non-linear graph-based dimension reduction method like t-SNE
- Better than PCA at showing overall similarities of cells
- Preserves the global structure better than t-SNE
- Newer & efficient = fast
- Based on the top Principal Components



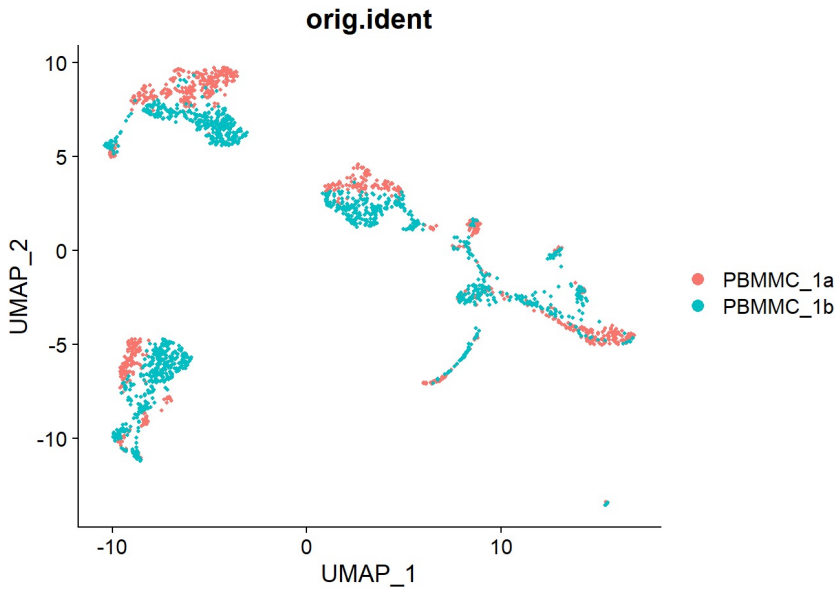


TIPS:

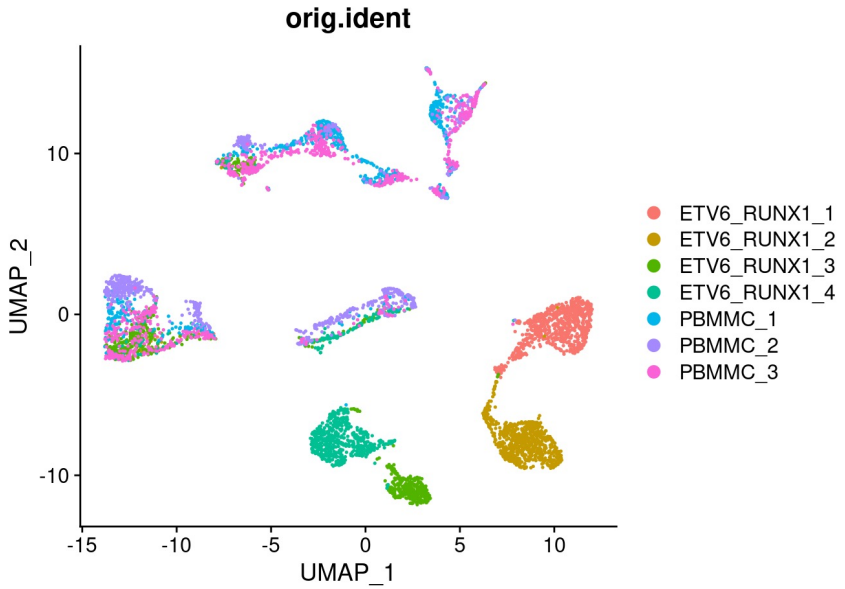
- Worth exploring the parameters e.g. n_neighbours
- Set a seed so you don't get different results each time!

Batch effects

Batch effect between technical replicates

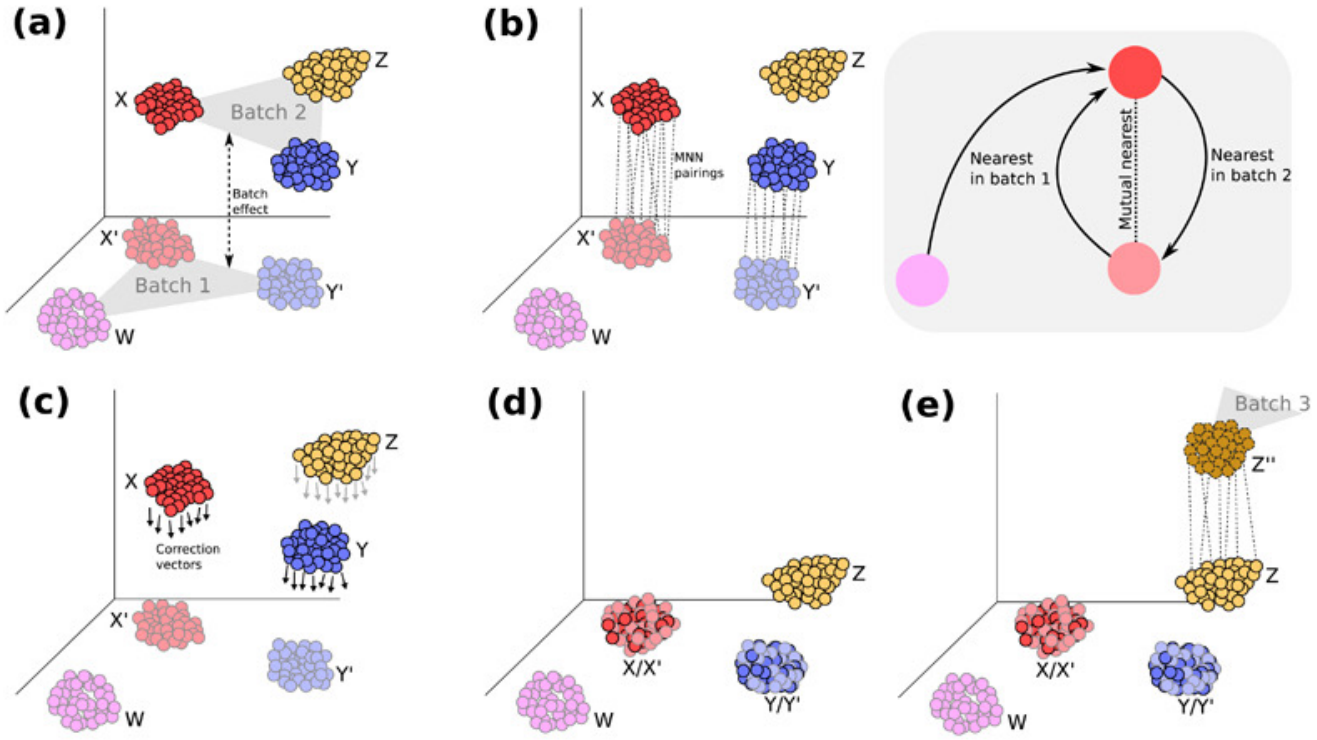


Batch effect between different samples



UMI counts, which have been corrected for batches cannot be used for gene-level analysis i.e. differential expression and marker identification. They are only suitable for cell-level analysis e.g. clustering.

Batch effects – mutual nearest neighbours



Here are the assumptions of this approach (taken from [Haghverdi et al 2018]):

1. There is at least one cell population that is present in both batches

2. the batch effect is not correlated to the biological effect

3. the batch-effect variation is much smaller than the biological-effect variation between different cell types

The “integration anchors” approach in Seurat is based on this Mutual Nearest Neighbours approach of Haghverdi et al.

Clustering

The data has been QC'd, normalized, and batch corrected.

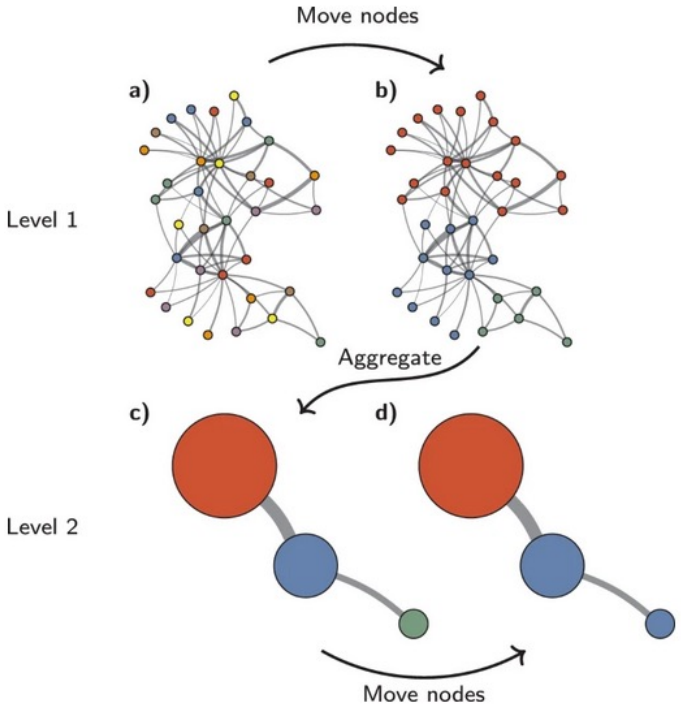
We can now start to understand the dataset by identifying cell types.

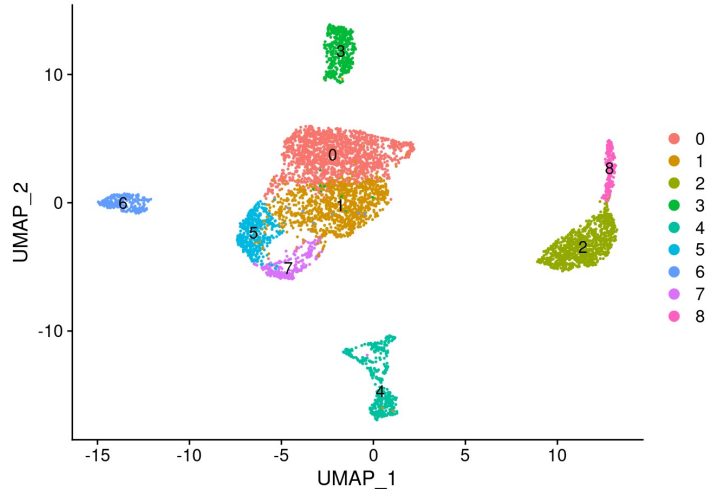
This involves two steps:

1. unsupervised clustering: identification of groups of cells based on the similarities of the transcriptomes without any prior knowledge of the labels usually using the PCA output

2. annotation of cell-types based on transcription profiles

TIPS:
It is worth trying different algorithms and parameters for each algorithm as it is difficult to know which will work best for your data





Is there a “correct” clustering?

Clustering, like a microscope, is a tool to explore the data.

We can zoom in and out by changing the resolution of the clustering parameters, and experiment with different clustering algorithms to obtain alternative perspectives on the data.

Asking for an unqualified “best” clustering is akin to asking for the best magnification on a microscope.

A more relevant question is “how well do the clusters approximate the cell types or states of interest?”. Do you want:

- resolution of the major cell types?
- Resolution of subtypes?
- Resolution of different states (e.g., metabolic activity, stress) within those subtypes?

Explore the data, use your biological knowledge!

The Dataset

Childhood acute lymphoblastic leukemia (cALL) is the most common pediatric cancer. The aim of the study was to characterise the heterogeneity of gene expression at the cell level, within and between patients.

Caron et al. 2020

- loaded thawed PBMMCs onto a 10X Genomics Chromium single cell platform (v2 chemistry).
- They aimed for 3,000 cells per sample and targeted 100,000 reads per cell by sequencing each sample on one lane of an Illumina HiSeq 4000 high-throughput sequencer (2x98 b.p. paired-end sequencing).
- They generated single cell gene expression data from 39,375 pediatric bone marrow mononuclear cells (PBMMCs) from eight cALL patients of common subtypes.
- Thus we have cells collected from four patients with ETV6/RUNX1 rearrangements, two HHD cases and two T-ALL cases. There are also PBMMCs from 3 healthy donors.

In the original paper they examined transcriptional variation within and between the cancers of different patients. Similarly, by the end of this workshop we will have:

- Performed QC, normalisation and batch correction on the data.
- Identified the cell types in the different samples and looked at DE genes.

Analysis software

Popular options

- Seurat - R
- Bioconductor (scater/scran) - R
- Scanpy – Python

They all do similar things but with different data structures and/or programming languages.

Sometimes the tool you want to use will be implemented in one or other “ecosystem”. You can usually convert your data into alternative format and proceed, but you might have to use a different language. Look for tutorials online!

The Seurat object

`@meta.data` slot - stores metadata for our droplets/cells (e.g. which batch of samples they belong to, total counts, total number of detected genes, etc.).

`@assays` slot - stores the matrix of raw counts (`$RNA`), as well as matrices of normalised (`$SCT`) and transformed data (`$integrated`).

`@reductions` slot – stores dimension reductions such as pca (`$pca`) and umap (`$umap`)

Cell Ranger demo

```
# run mkref
cellranger mkref \
        --fasta=Homo_sapiens.GRCh38.dna.chromosome.21.fa \
        --genes=gencode.v41.primary_assembly.annotation.chr21.gtf \
        --genome=cellranger_index_human \
        --nthreads=7

# run cellranger count (maximum CPUs 8; maximum RAM 24GB)
cellranger count \
        --id=ETV6_RUNX1_rep1 \
        --transcriptome=cellranger_index_human \
        --fastqs=/mnt/scratch/reid/ajr236/SRR9264343/ \
        --sample=SRR9264343 \
        --localcores=8 \
        --localmem=24
```

Acknowledgements

Lots of slides on scRNA-seq - Katarzyna Kania

Various course materials - Abigail Edwards, Ashley D Sawle, Chandra Chilamakuri, Kamal Kishore, Stephane Ballereau, Zeynep Kalendar Atak, Hugo Tavares, Jon Price, Roderik Kortlever, Adam Reid, Tom Smith