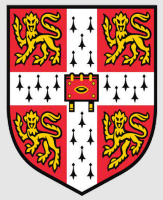
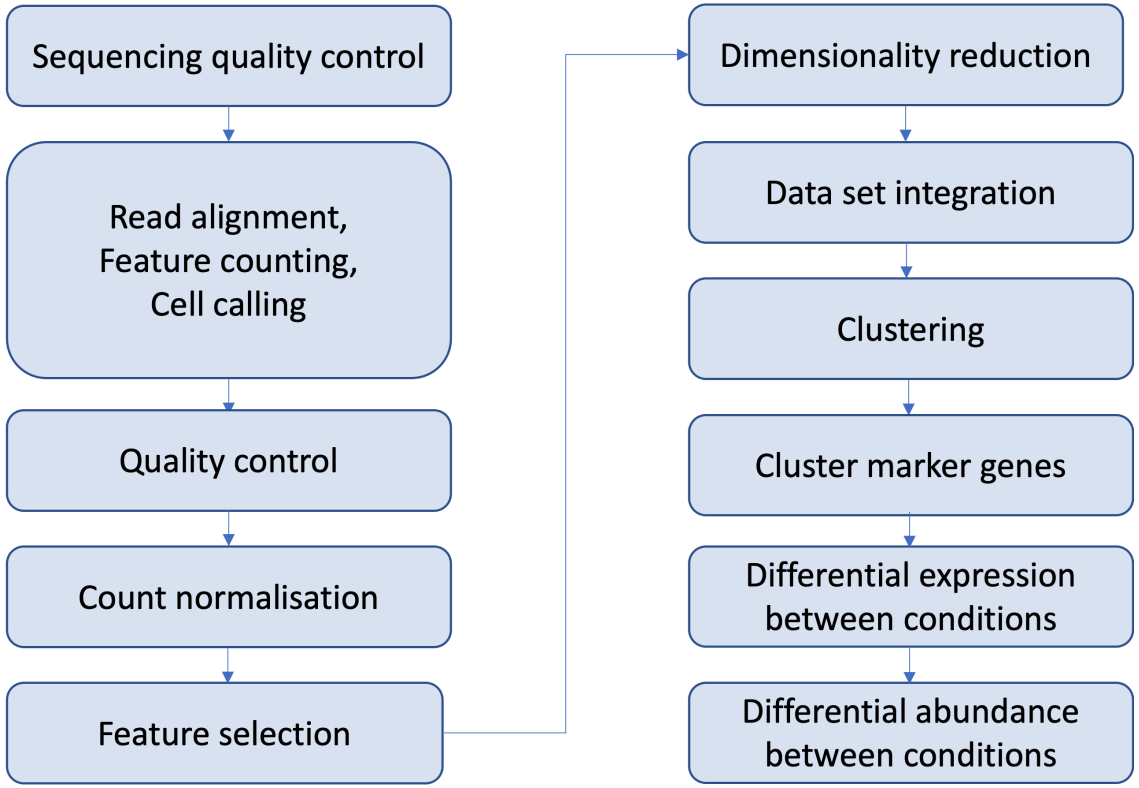


UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | Cambridge Institute

Day 1 Recap

Workflow


Sequencing quality control
Read alignment,
Feature counting,
Cell calling
Quality control
Count normalisation
Feature selection
Dimensionality reduction
Data set integration
Clustering
Cluster marker genes
Differential expression
between conditions
Differential abundance
between conditions

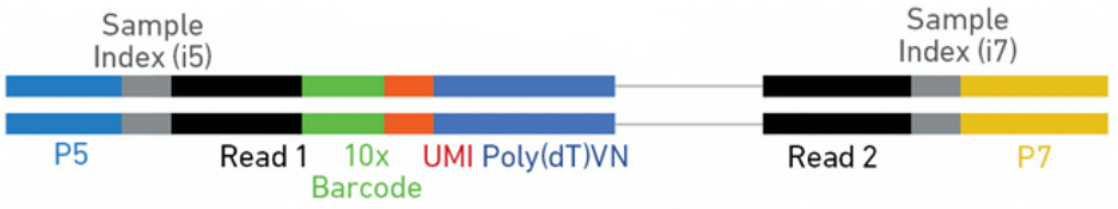
Data set

- Data set: CaronBourque2020: Data from Childhood acute lymphoblastic leukemia (cALL)
- cells: Bone Marrow Mononuclear cells (BMMCs)
  - 12 samples
  - 4 Sample groups
    - HHD: The high hyper diploid cases (51–67 chromosomes).
      - Two replicates.
    - PBMMC: healthy pediatric BMMC.
      - Four replicates.
      - There are two PBMMC_1 samples. These are two libraries from the same sample material.
    - ETV6-RUNX1: ETV6/RUNX1 rearrangement
      - Four replicates
    - Pre-T: Pre-T ALL
      - Two replicates
- Aim: characterize the heterogeneity of gene expression at the cell level, within and between patients

10x library file structure

The 10x library contains four pieces of information, in the form of DNA sequences, for each “read”.

- **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 molecule



Cell Ranger

- 10x Cell Ranger - This not only carries out the alignment and feature counting, but will also:
  - Call cells
  - Generates counts matrix
  - Generate a summary report in html format
  - Generate a “cloupe” file

Cell Ranger references

cellranger mkref
–fasta={GENOME FASTA}
–genes={ANNOTATION GTF}
–genome={OUTPUT FOLDER FOR INDEX}
–nthreads={CPUS}

Running cellranger count

```
cellranger count –id={OUTPUT_SAMPLE_NAME}
–transcriptome={DIRECTORY_WITH_REFERENCE}
–fastqs={DIRECTORY_WITH_FASTQ_FILES}
–sample={NAME_OF_SAMPLE_IN_FASTQ_FILES}
–localcores={NUMBER_OF_CPUS}
–localmem={RAM_MEMORY}
```

Cell Ranger outputs

The contents of the `outs` directory are:

```
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%-$
```

Not every droplet is useble

A single happy cell in a droplet is ideal

- Complex transcriptome
- Average number of genes detected

Empty droplet: No cell in a droplet

- No genes detected

Droplet with ambient RNA

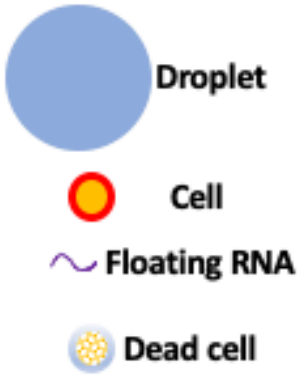
- Low complex transcriptome
- Genes detected much lower than average genes per cell

Droplet with dead cell

- Enriched for mitochondrial genes

Droplet with multiple cell

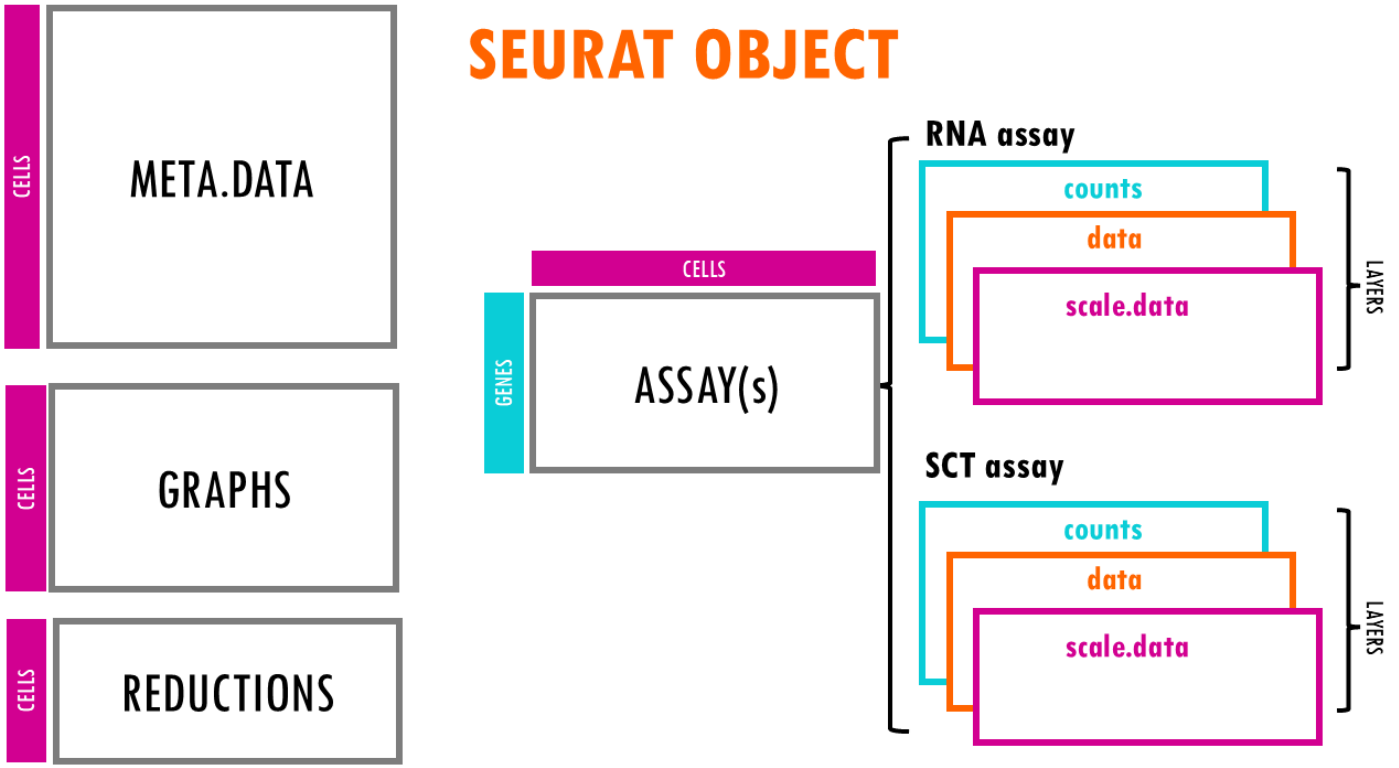
- Very complex transcriptome
- Genes detected much higher than average genes per cell



Quality Control overview

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

The *Seurat* object



Useful commands in QC

- access counts:

```
seurat_object[["RNA"]]$counts
```

- access cell metadata:

```
seurat_object[[]]
```

QC parameters

- The library size
- Number of expressed genes in each cell
- proportion of UMIs mapped to genes in the mitochondrial genome