Gurdon scRNA-seq data analysis workshop – day 2

Adam Reid

Head of Bioinformatics

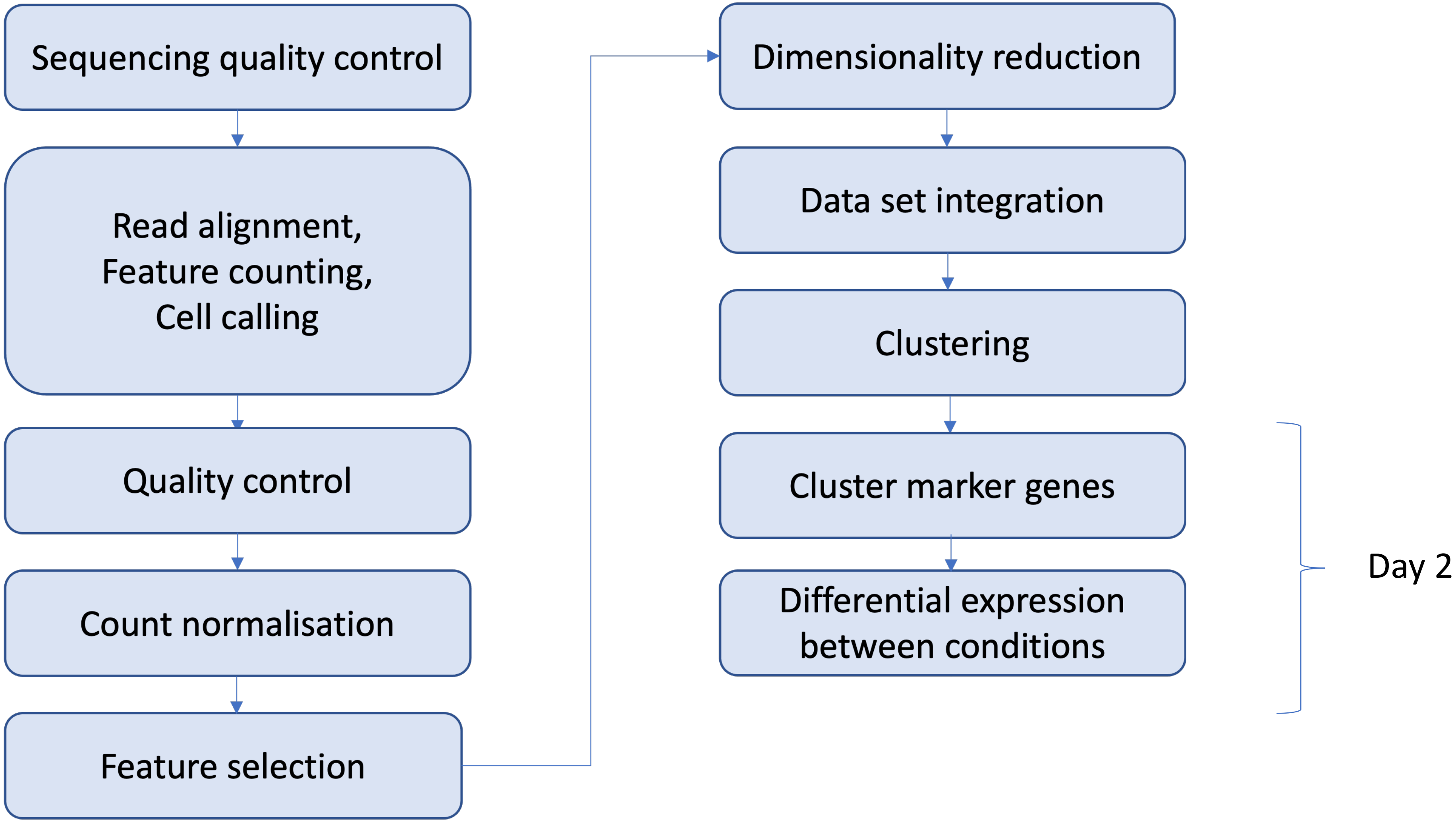
Gurdon Institute

22nd March 2023

Outline

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

Analysis workflow



Sequencing quality control

Read alignment,
Feature counting,
Cell calling

Quality control

Count normalisation

Feature selection

Dimensionality reduction

Data set integration

Clustering

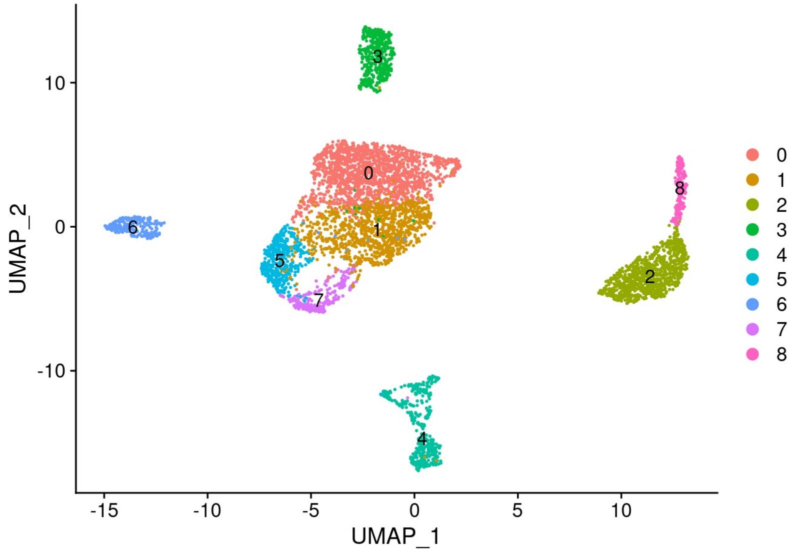
Cluster marker genes

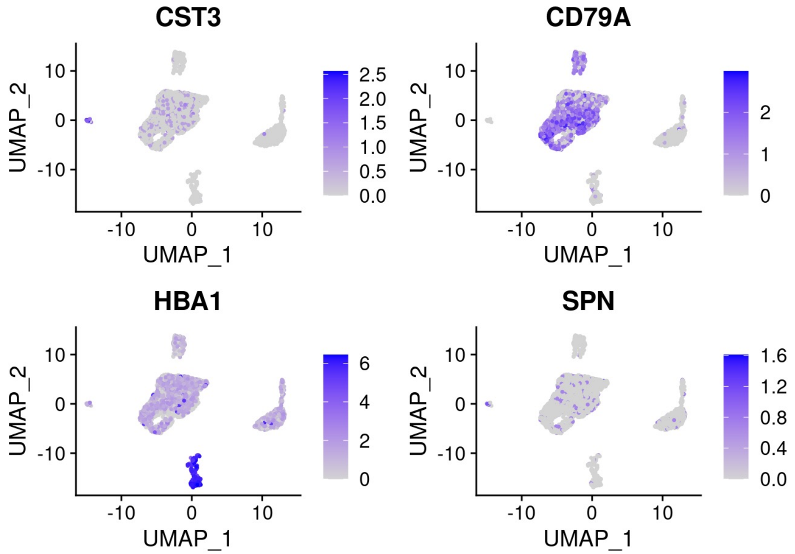
Differential expression between conditions

Day 2

Identifying cell types

Using known markers for cell types you expect



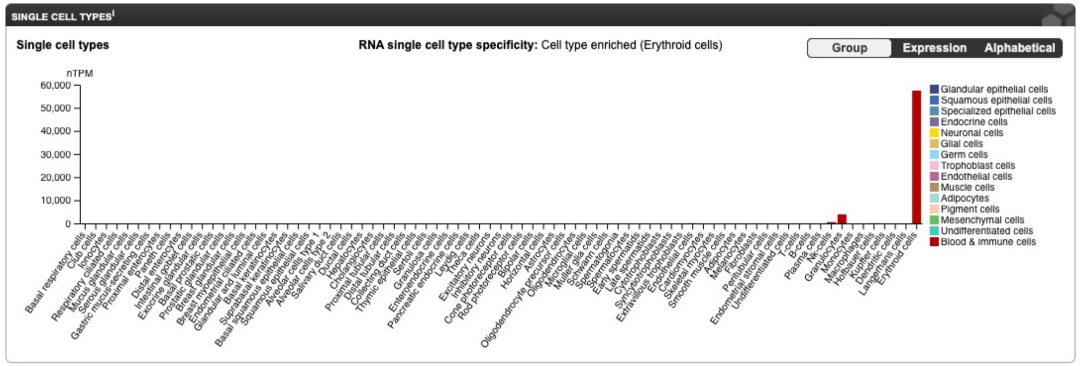


Identifying markers *de novo*

- Identify genes which are more highly expressed in one cell cluster than in others
- In Seurat use 'FindMarkers'

```
        p_val   avg_log2FC pct.1 pct.2 p_val_adj
GNLY    0       2.648152   0.476 0.009 0
CTSW    0       1.112248   0.601 0.022 0
KLRD1   0       1.084170   0.519 0.008 0
CCL5    0       2.858279   0.841 0.020 0
NKG7    0       2.532095   0.841 0.042 0
CST7    0       1.298939   0.692 0.008 0
```

Look for where the marker genes are known to be expressed e.g. using Human Protein Atlas in



SINGLE CELL TYPES

Single cell types

RNA single cell type specificity: Cell type enriched (Erythroid cells)

Automated typing tools are available, e.g. CellTypist

Differentially expressed genes

- Clusters and/or cell types have been identified, we now want to compare sample groups:
  - Differential expression – E.g. differences in gene expression between sample groups within a particular cell type.
  - Differential abundance - Differences in cell numbers between sample groups for a particular cell type.

Differential expression

There are a variety of different approaches. These were assessed for accuracy in Squair et al (2021)

The best methods, account properly for variability between replicates and in essence work just as for bulk RNA-seq

Replicates should be samples, not cells:

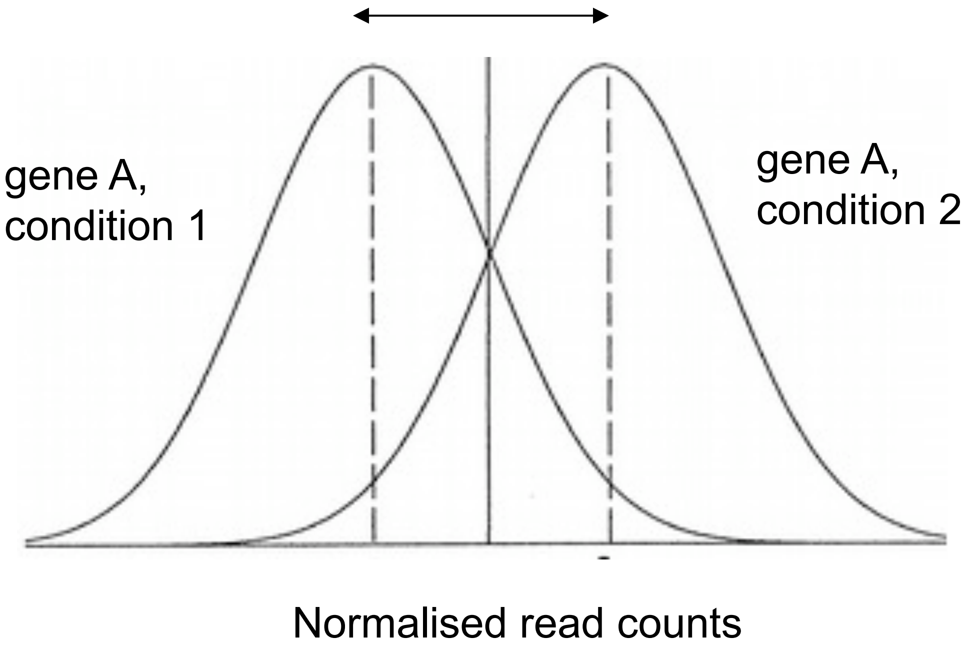
- it is important to treat each sample and not each cell as the experimental unit of interest.
- cells from one sample are likely more similar to each other than cells taken from different samples.
- ignoring variability in gene expression with a condition, and treating cells (as opposed to samples) as independent biological replicates can lead to false discoveries (Squair et al. 2021, Zimmerman et al. 2021).

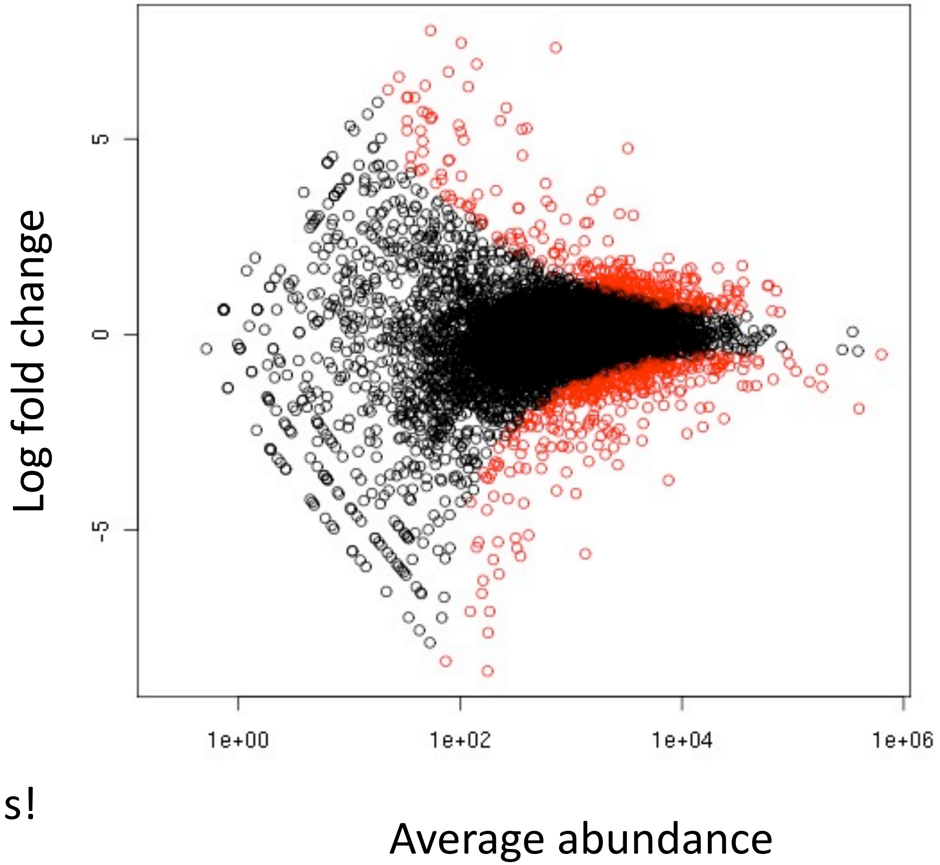
Pseudo-bulk:

- gene expression levels for each cluster in each sample are obtained by summing across cells

Determining differential expression

- We normally don't have enough replicates to do traditional tests of significance for RNA-seq data
- Instead most methods look for outliers in the relationship between average abundance and fold change, assuming most genes are not differentially expressed





Single-cell RNA-seq data is even more noisy than bulk methods!

A big caveat

- We are going to use a simple but poorly performing approach (Wilcoxon Rank Sum test)
- It will not give you the best results, because it doesn’t treat replicates properly and does not account for some of the properties of scRNA-seq data
- The current best practice approach is to use pseudobulking and DESeq2, but this is currently relatively code-heavy in Seurat (although quite nicely implemented in scran).