Melbourne Bioinformatics
BIOINFORMATICS + DATA SERVICES + INFRASTRUCTURE, FOR LIFE SCIENCES TODAY

THE UNIVERSITY OF MELBOURNE

Workshop: scRNA-seq Differential Expression

Trainers: Manveer Chauhan
Xiaochen Zhang
Steven Morgan

Organiser: Vicky Perreau

Melbourne Bioinformatics
BIOINFORMATICS + DATA SERVICES + INFRASTRUCTURE, FOR LIFE SCIENCES TODAY

THE UNIVERSITY OF MELBOURNE

Acknowledgement of Country

The University of Melbourne acknowledges the Traditional Owners of the unceded land on which we work, learn and live: the Wurundjeri Woi-wurrung and Bunurong peoples.

We pay respect to Elders past, present and future, and acknowledge the importance of Indigenous knowledge in the Academy

Melbourne Bioinformatics
BIOINFORMATICS + DATA SERVICES + INFRASTRUCTURE, FOR LIFE SCIENCES TODAY

THE UNIVERSITY OF MELBOURNE

House keeping:

- WiFi
- SLACK
- Refreshments
- Building access
- Emergency

Melbourne Bioinformatics
BIOINFORMATICS + DATA SERVICES + INFRASTRUCTURE, FOR LIFE SCIENCES TODAY

THE UNIVERSITY OF MELBOURNE

Providing bioinformatics support for all researchers and students in Melbourne's biomedical and biosciences precinct.

Melbourne Bioinformatics
BIOINFORMATICS + DATA SERVICES + INFRASTRUCTURE, FOR LIFE SCIENCES TODAY

THE UNIVERSITY OF MELBOURNE

Capabilities

Training

Access our suite of self-directed tutorials or attend our hands-on training workshops (free for UoM & Affiliates researchers and students)

Free expert advice (for UoM & Affiliates)

Consult with some of our 30 bioinformaticians and high-end computing experts about grants, research design, data analysis/management, software and hardware

Subscriptions

Form a deep collaboration with our world-leading academic staff who can rapidly progress your research project and contribute to publications, usually via a subscription.

Online training repository

https://www.melbournebioinformatics.org.au/tutorials/

Melbourne Bioinformatics
BIOINFORMATICS + DATA SERVICES + INFRASTRUCTURE, FOR LIFE SCIENCES TODAY

THE UNIVERSITY OF MELBOURNE

@MelBioInf

www.melbournebioinformatics.org.au

sign up for eNews

bioinformatics-training@unimelb.edu.au

scRNA-seq Integration and Differential Expression Workshop

Working with treatment versus control data

Manveer Chauhan

Study Design

- Peripheral Mononuclear Blood Cells (PBMCs) were sequenced using scRNA-seq from 8 lupus patients. Patients were randomly split into a treatment and control group. The treatment group received interferon beta.

- Goals of our analysis:
  - Integrate data, so that batch effects are removed and similar cell types across both conditions are grouped together.
  - Identify upregulated genes in cell-types in a treatment versus control experiment.
  - Identify and visualise genes that are differentially expressed between conditions in a particular cell type
  - Perform differential expression analysis using an alternative ‘pseudobulk’ approach

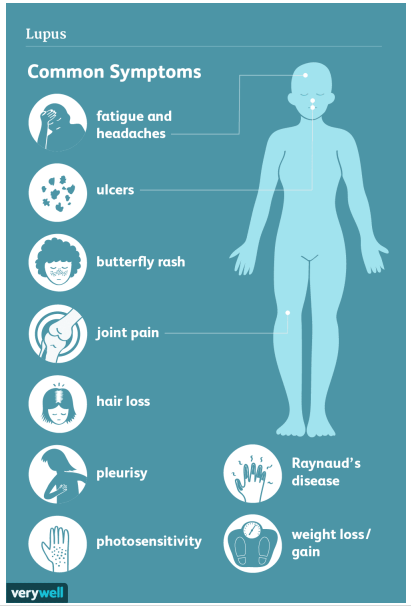
Article | Published: 11 December 2017

Multiplexed droplet single-cell RNA-sequencing using natural genetic variation

Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, Rachel E Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A Criswell & Chun Jimmie Ye

Nature Biotechnology **36**, 89–94 (2018) | Cite this article

68k Accesses | **481** Citations | **177** Altmetric | Metrics



Learning Outcomes

- Understand and get comfortable using various integration strategies
- Understand all differential expression functions offered by Seurat and when to use them
- Learn how to use differential expression tools meant for bulk data (e.g. DESeq2) on ‘pseudobulk’ data, and understand why you might choose this approach
- Learn different ways to visualize differentially expressed genes using both in-built Seurat functions and external packages (pheatmap)

Software and Package Requirements

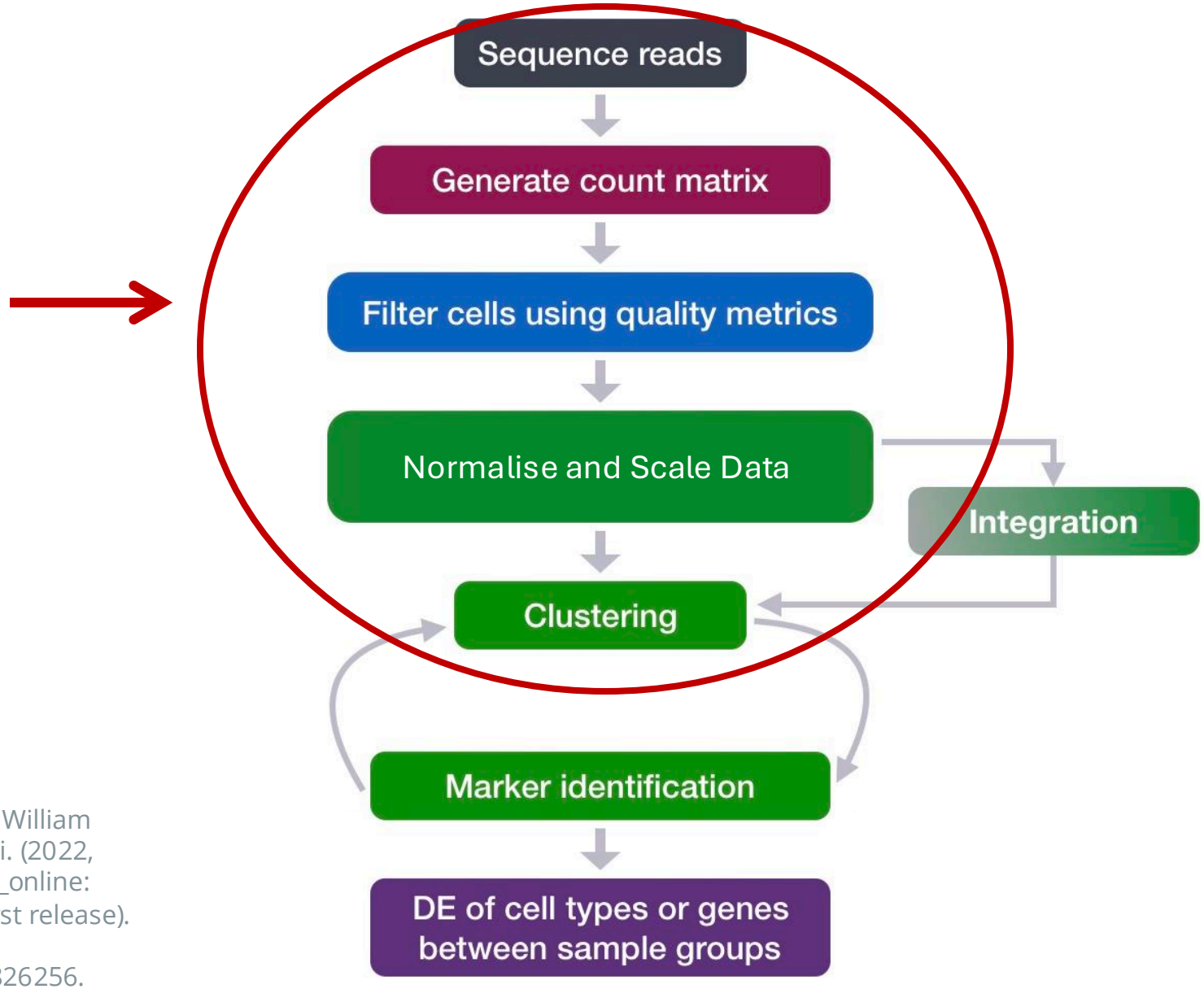
- R (v4.3.0)
- RStudio

R packages:

- Seurat (v5.0.1)
- DESeq2 (v1.42.1)
- tidyverse (v2.0.0)
- SeuratData (v0.2.2.9001)
- pheatmap (v1.0.12)
- grid (v4.0.3)
- metap (v1.11)

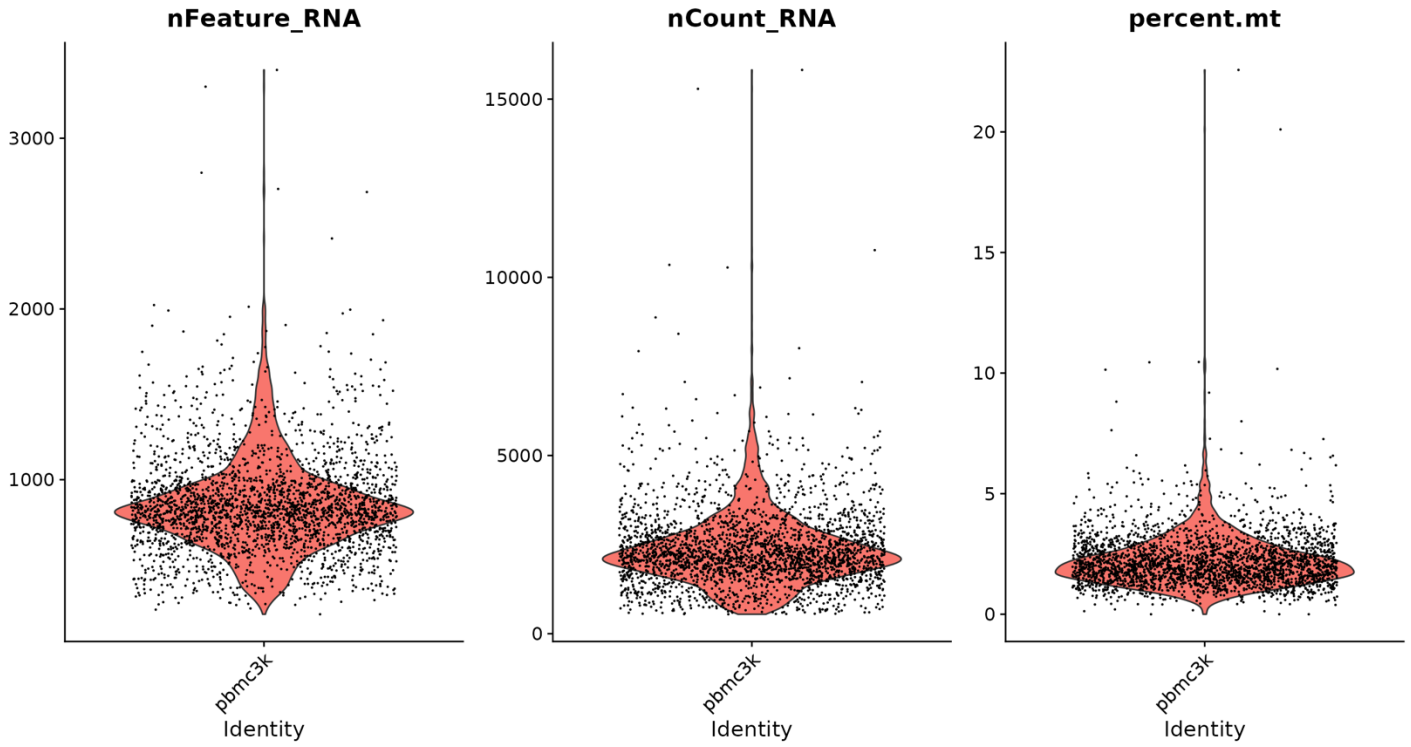
General scRNA-seq Workflow

CreateSeuratObject()
NormalizeData()
FindVariableFeatures()
ScaleData()
RunPCA()
FindNeighbors()
FindClusters()
RunUMAP()



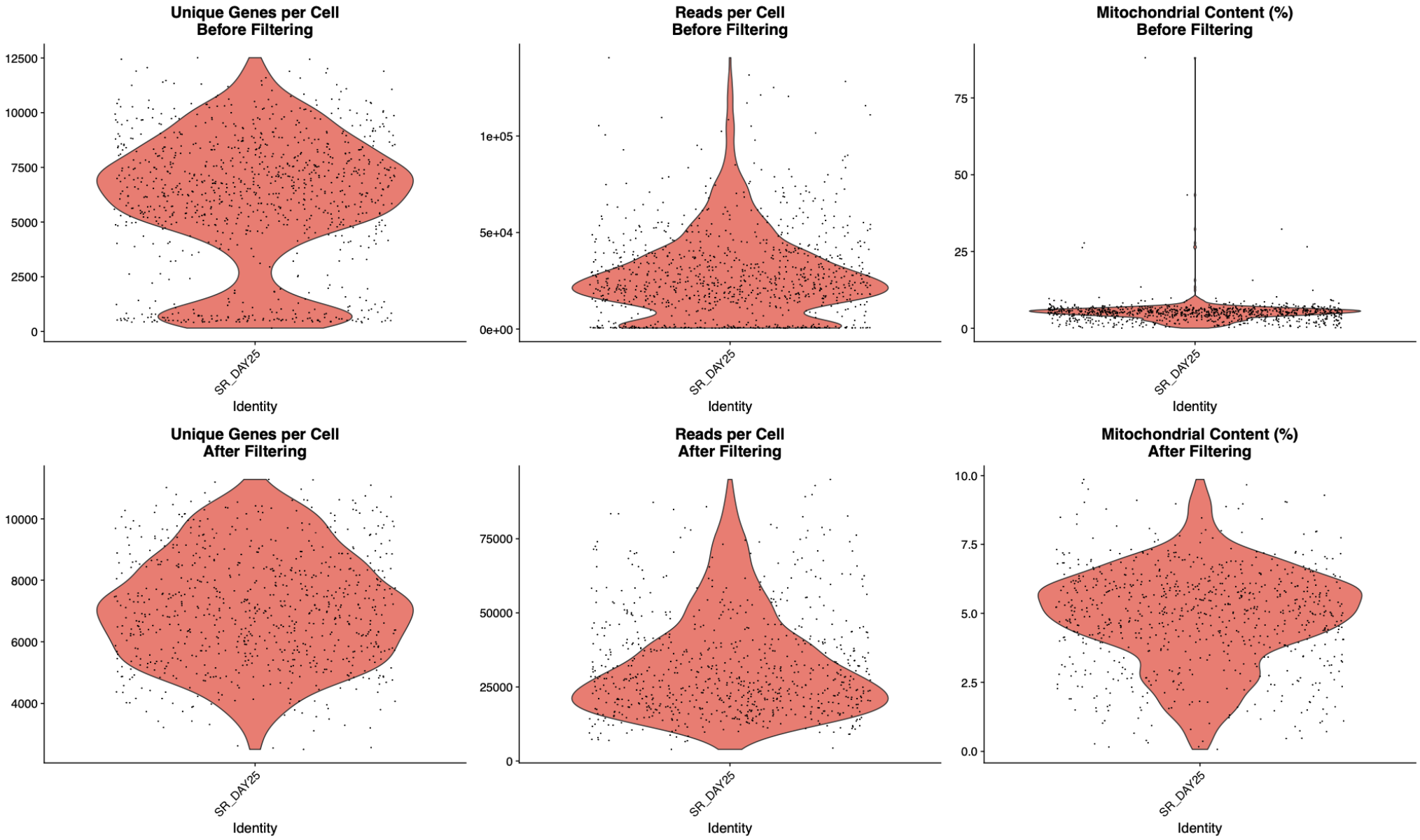
Mary Piper, Meeta Mistry, Jihe Liu, William Gammerdinger, & Radhika Khetani. (2022, January 6). hbctraining/scRNA-seq_online: scRNA-seq Lessons from HCBC (first release). Zenodo. https://doi.org/10.5281/zenodo.5826256.

Guidelines for removing low quality cells



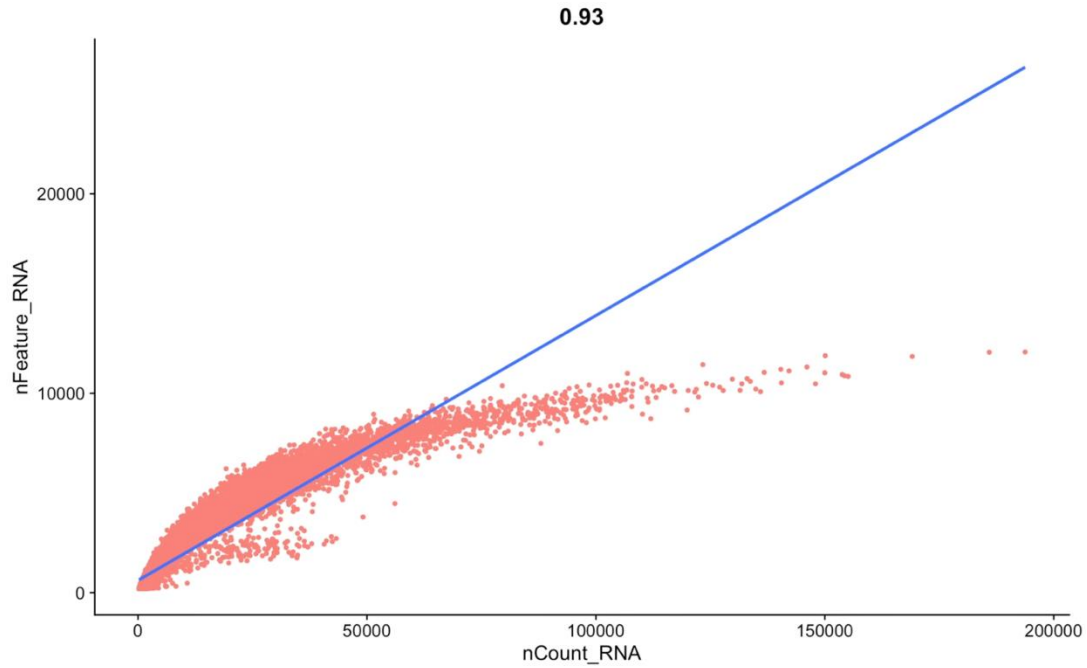
- Low quality cells or empty droplets will have fewer genes and fewer counts
- Cell doublets (>1 cell assigned to a single barcode) will have significantly more genes and counts
- Dying cells will have higher mitochondrial contamination
  - (<=5% or 10% is a good guideline)
- We can use violin plots to determine thresholds for filtering based on these metrics

Example Before and After QC Plots



Consider Metrics Together: Gene and UMI Association Plots

- X axis = number of transcripts/counts per cell
- Y axis = number of unique genes per cell
- Generally, for good quality data, we expect a strong positive correlation between the number of counts and unique genes.
- Using the line as a guide, we can figure out cells that are potentially lower quality
  - Cells in the bottom right quadrant indicates you've captured a few number of genes that are being sequenced over and over again
  - Cells in the top left quadrant indicates you're capturing many genes but not sequenced deep enough



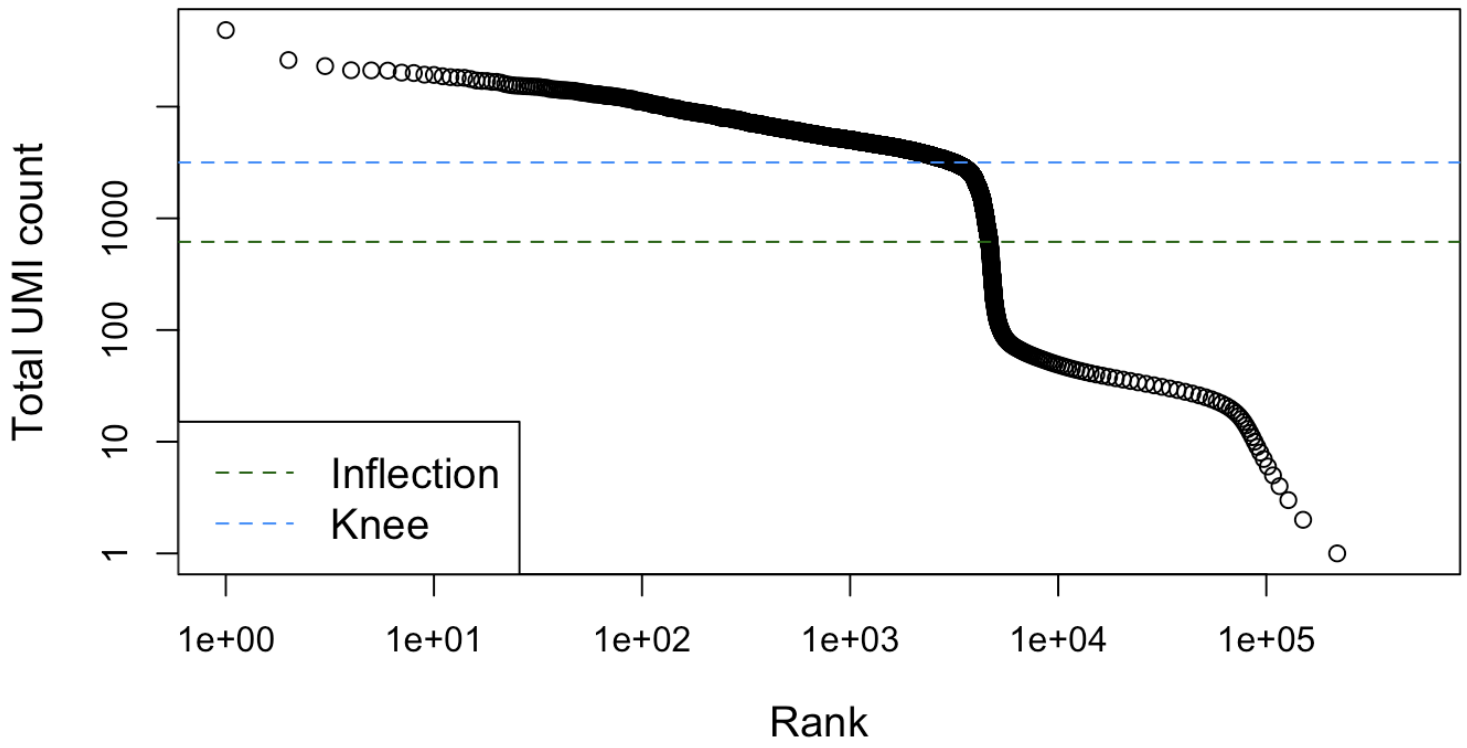
QC - Empty Droplet Filtering

- Using a threshold to keep only barcodes with total counts above X
  - Empty droplets can have relatively high counts (if they captured lots of ambient RNA).
  - Some **real cells** can have low counts (if they're small or naturally low in RNA).

Alternative algorithms

- The emptyDrops() function (Lun et al. 2019) helps by comparing the expression profile of each droplet against the ambient RNA background:
  - If a droplet looks *too similar* to background → it's probably empty.
  - If it looks *different enough* → it's probably a real cell.

QC - Empty Droplet Filtering



Dataset = Unfiltered human PBMCs (10X Genomics)

- Barcodes sorted by their total UMI counts
- Sharp drop (“knee” or “inflection”) separates high-count barcodes (likely cells) from low-count barcodes (likely empties)
- Keep everything above the knee

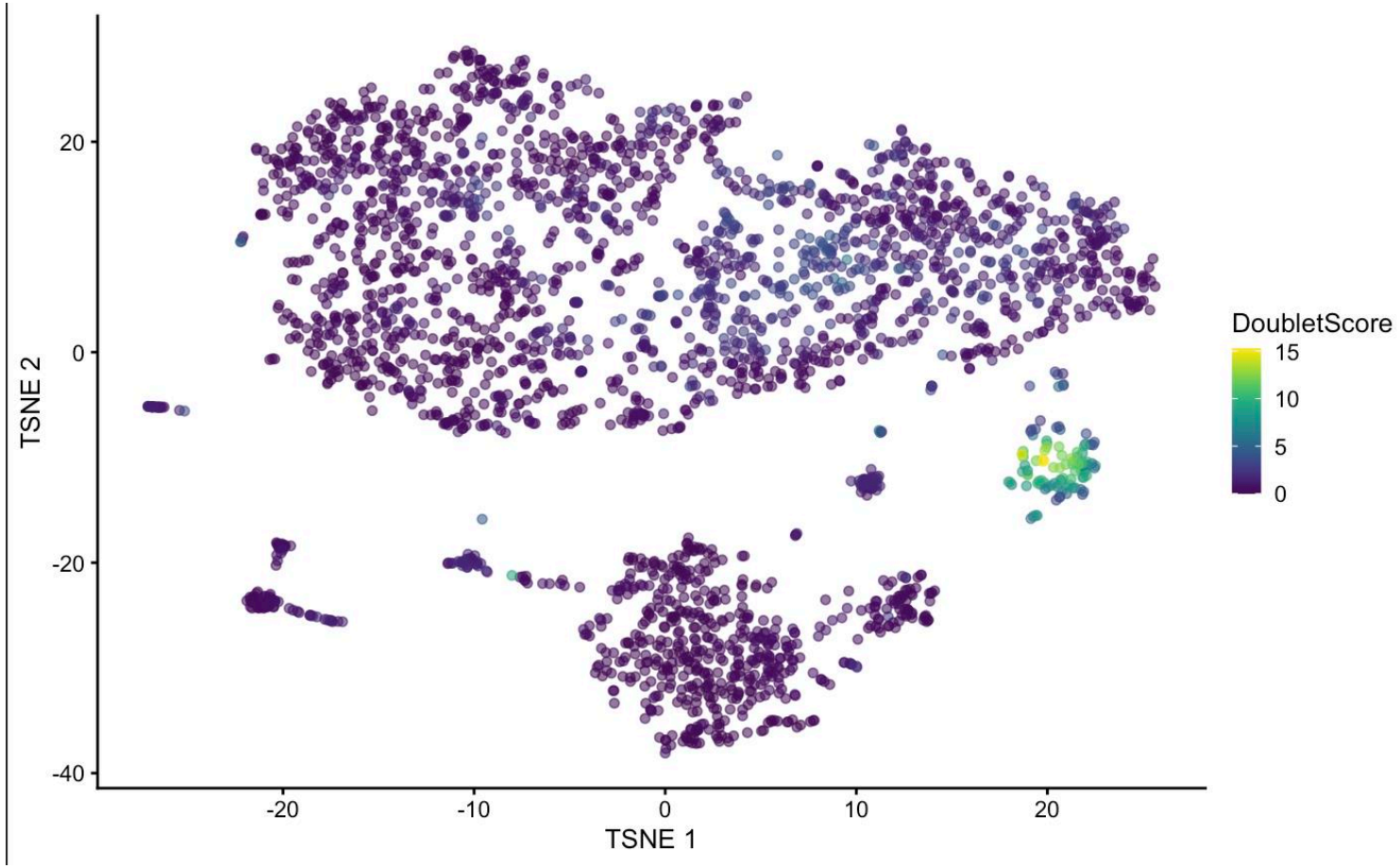
https://bioconductor.org/books/3.14/OSCA.advanced/droplet-processing.html#fig:rankplot

QC - Doublet Filtering

2 commonly used approaches (scDblFinder)

- Cluster-based method: findDoubletClusters()
  - Looks for clusters that sit *between* two other clusters in expression space.
- Simulation-based: computeDoubletDensity()
  - Create **synthetic doublets** by adding real cell profiles together.
  - For each cell, check whether its neighbors are more similar to real cells or simulated doublets.
  - Returns a **doublet score** for each cell -> High scores = cells likely to be doublets.

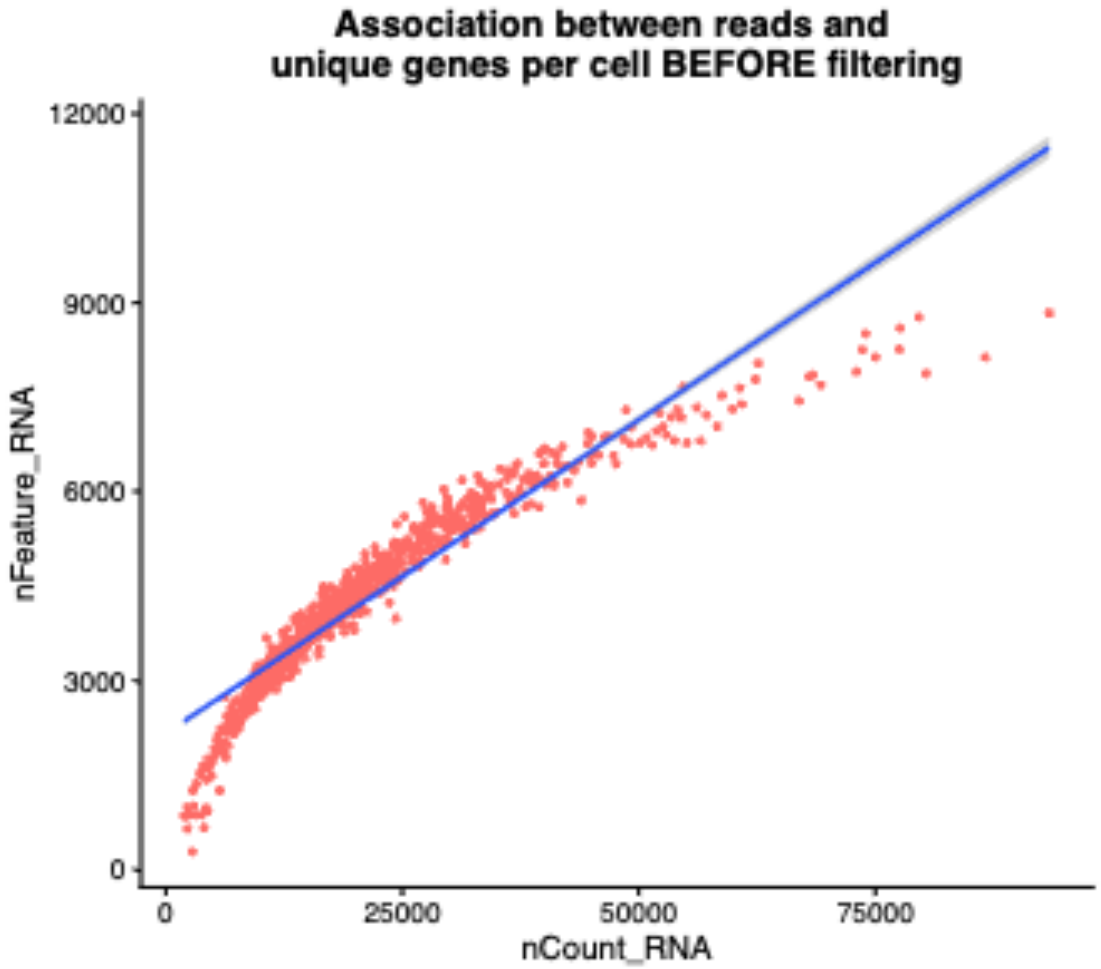
QC - Doublet Filtering

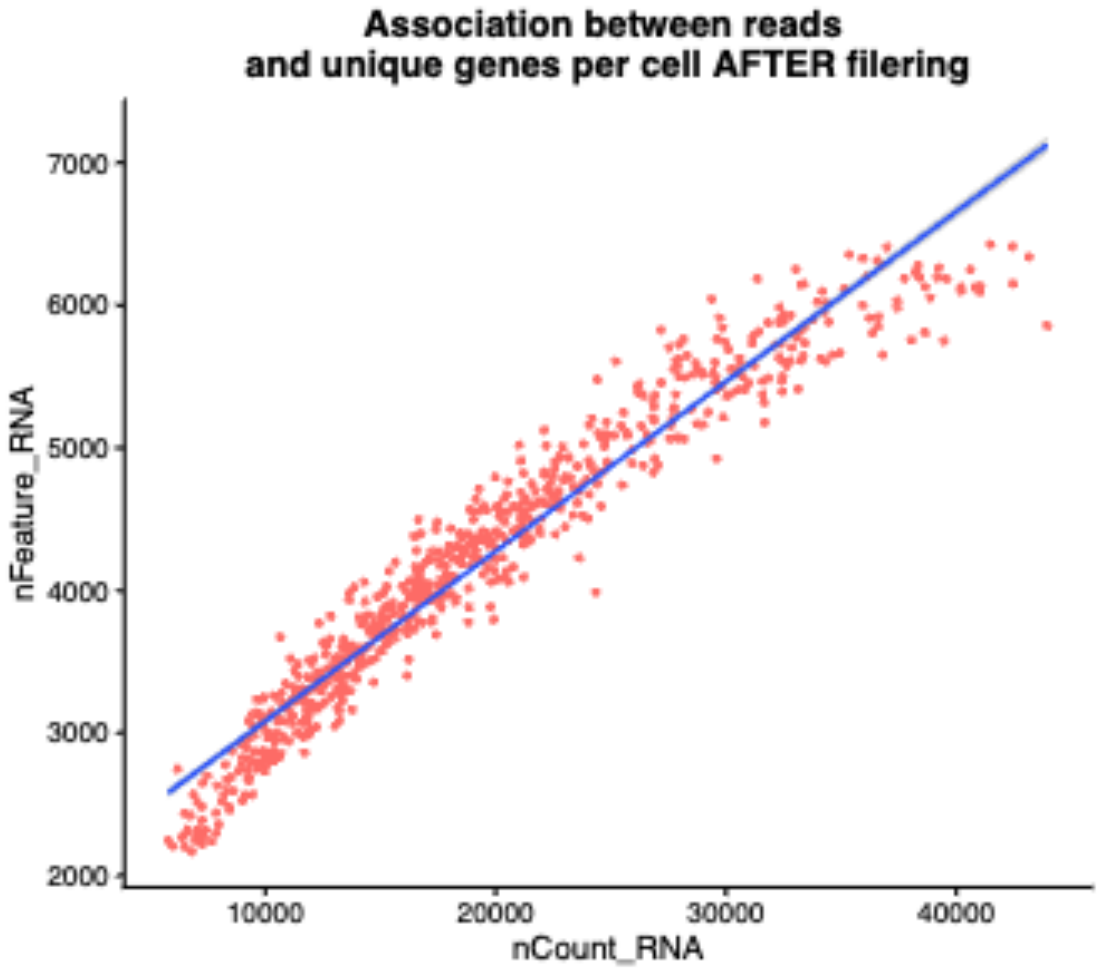


Dataset: Mouse mammary gland (10X Genomics data)

- Each point is a cell coloured according to its doublet density.
- The highest doublet scores are concentrated in a single cluster of cells in the centre

Example Association Plots Before and After Filtering





Cell-cycle scoring

What is cell-cycle scoring:

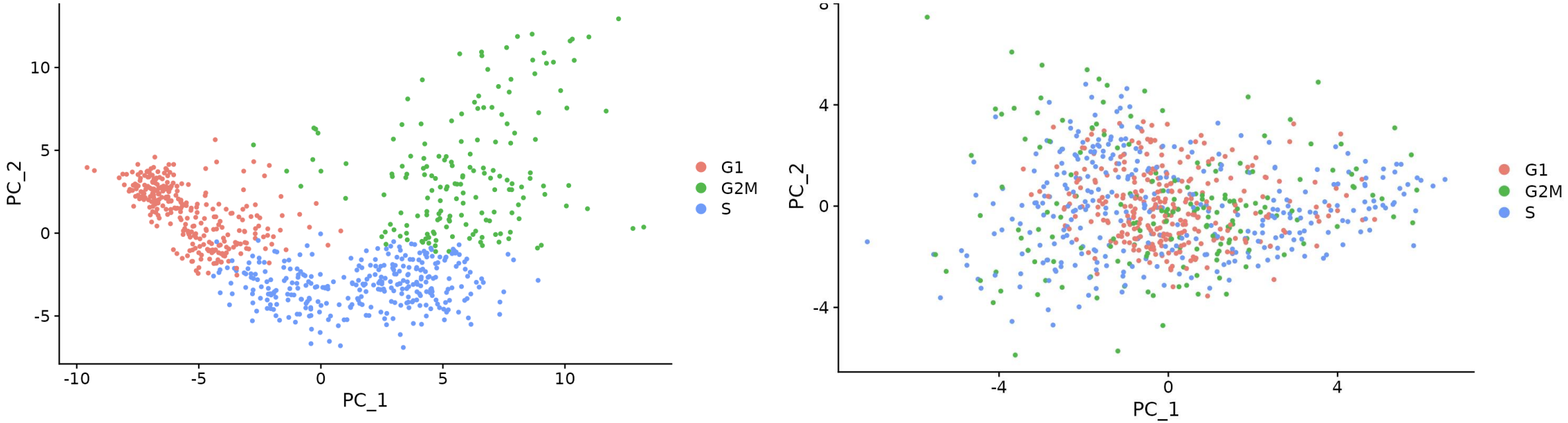
- A quick way to estimate each cell's cell-cycle phase (G1, S, G2/M) from its gene expression.
- Seurat does this with curated S-phase and G2/M marker lists, computing two numeric scores per cell:
  - S.Score – how strongly the cell expresses S-phase genes
  - G2M.Score – how strongly it expresses G2/M genes

Why it matters:

- Proliferation (S / G2M) can drive PCs/UMAP, splitting clusters by phase instead of biology.
- Scoring reveals whether cell cycle is a confounder for clustering or DE.
- If dominant, regress S/G2M scores during scaling.

Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science [Internet]. 2016;352:189–96. Available from: https://science.sciencemag.org/content/352/6282/189.full

Before and After Cell-cycle Score Regression



https://satijalab.org/seurat/articles/cell_cycle_vignette.html#regress-out-cell-cycle-scores-during-data-scaling

Good resource for further reading

Mol Syst Biol. 2019 Jun 19;15(6):e8746. doi: 10.15252/msb.20188746

Current best practices in single-cell RNA-seq analysis: a tutorial

Malte D Luecken [1], Fabian J Theis [1,2,✉]

Author information Article notes Copyright and License information

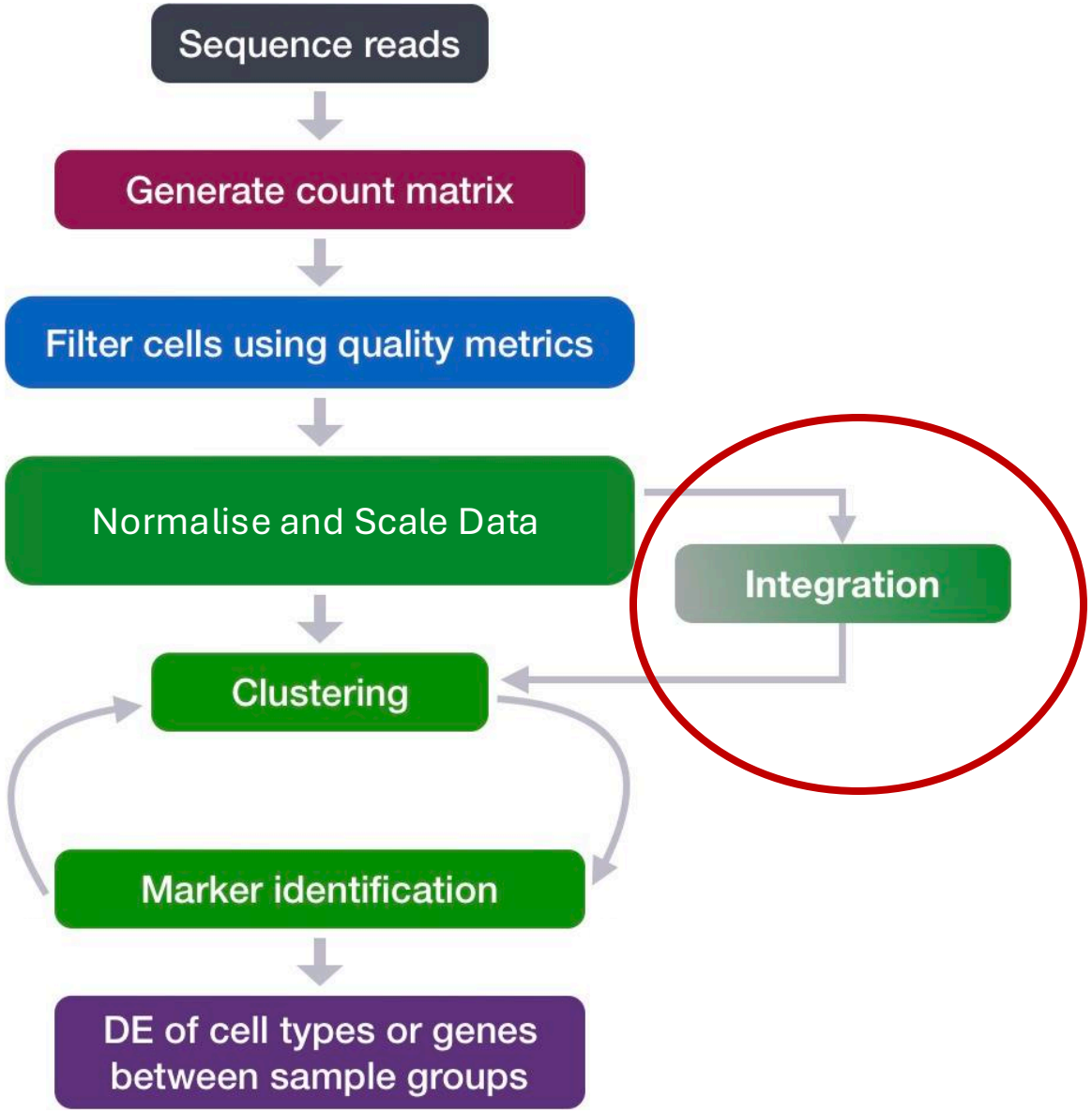
PMCID: PMC6582955 PMID: 31217225

https://pmc.ncbi.nlm.nih.gov/articles/PMC6582955/

Luecken and Theis (2019)

Training Material
Section 1 – Steps 1 and 2

Integration – What, When, Why?



Mary Piper, Meeta Mistry, Jihe Liu, William Gammerdinger, & Radhika Khetani. (2022, January 6). hbctraining/scRNA-seq_online: scRNA-seq Lessons from HCBC (first release). Zenodo. https://doi.org/10.5281/zenodo.5826256.

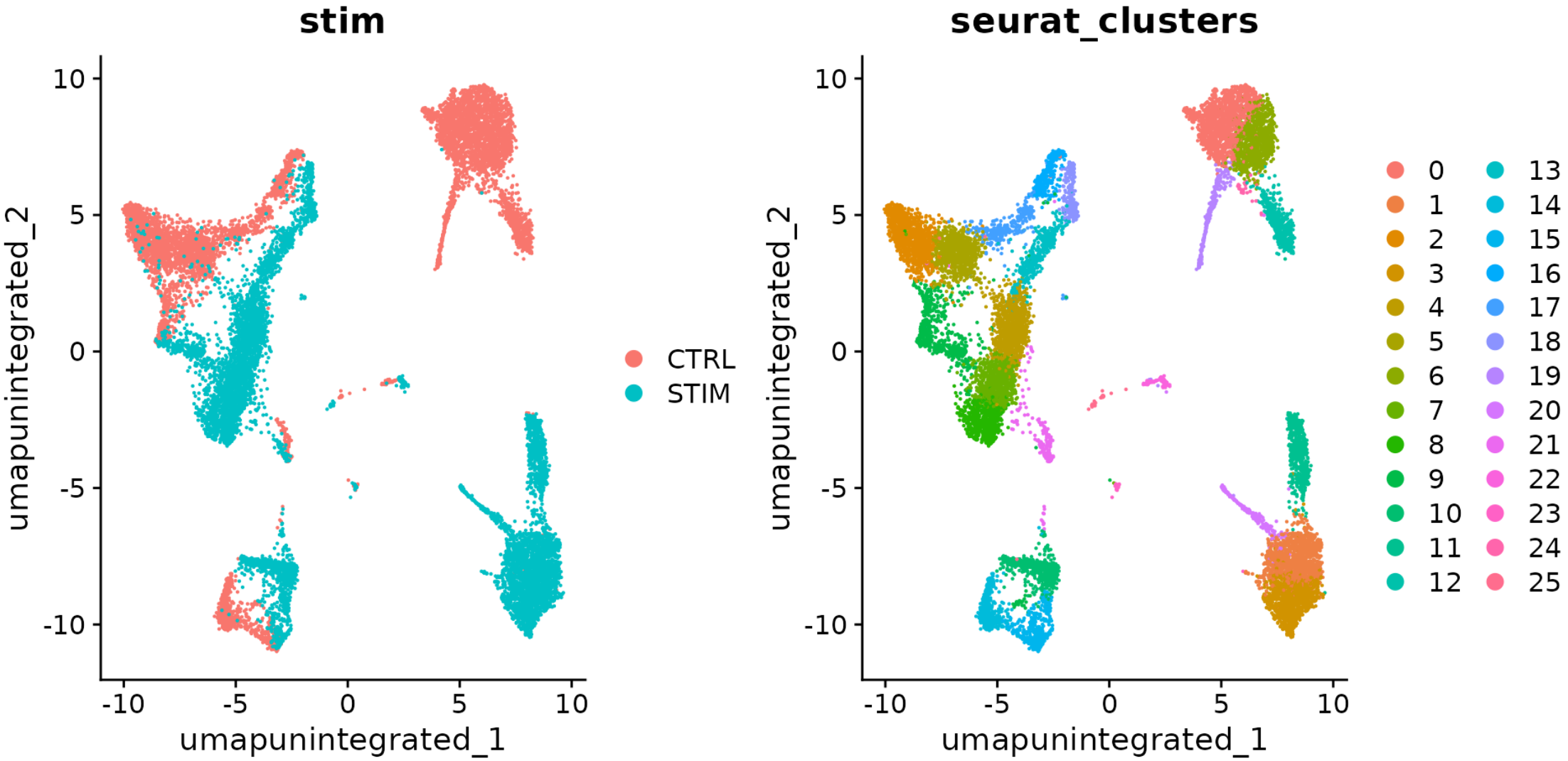
Integration – What, When, Why?

When comparing 2 Experimental Groups (e.g., Treatment/Control, KO/WT), we want to:

1. Identify shared cell subpopulations across both datasets.
2. Obtain conserved cell-type markers in both control and stimulated cells.
3. Compare datasets to reveal cell-type specific responses to treatment/condition.

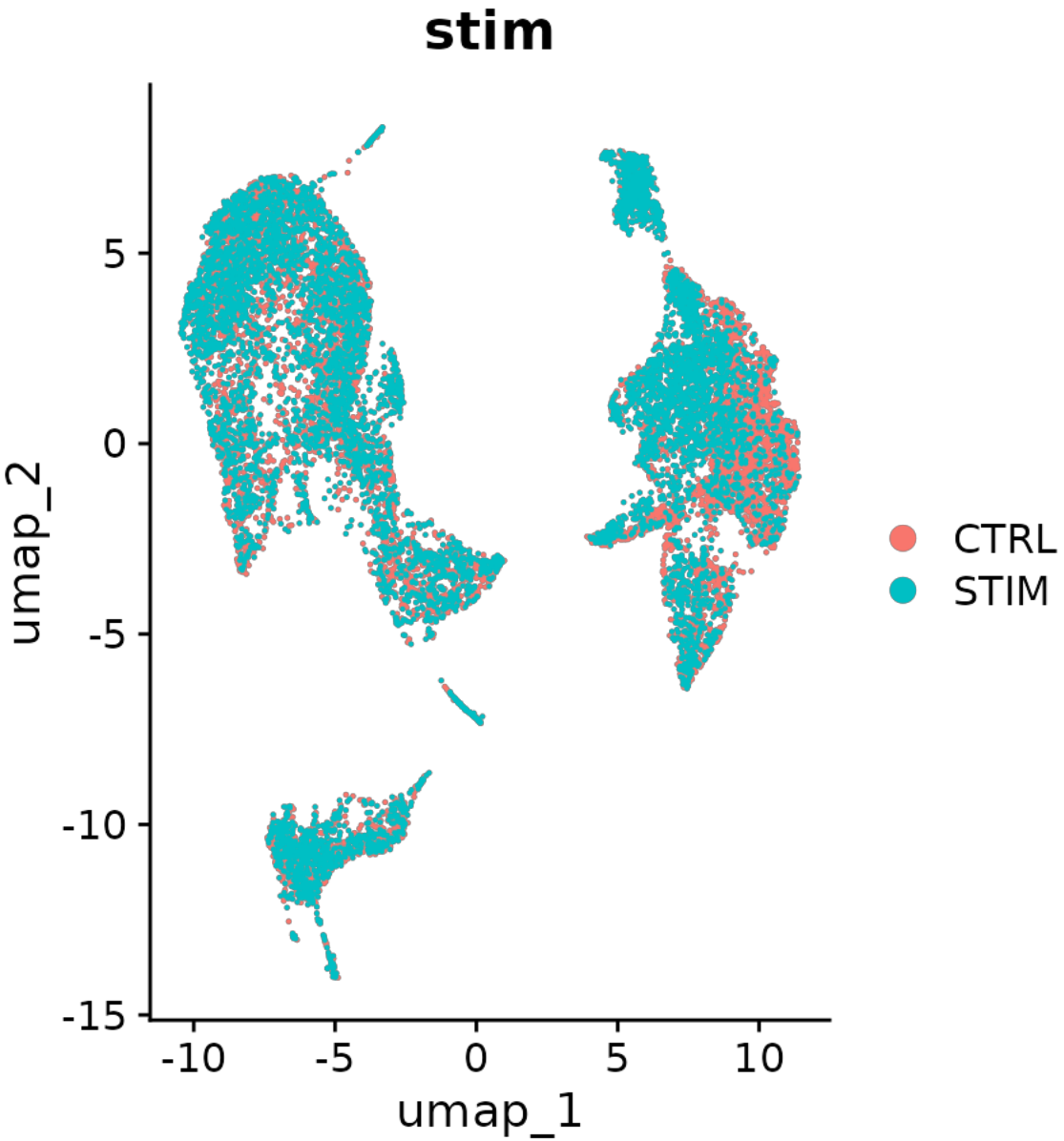
These steps rely on **integration**—a process that aligns shared cell states across datasets, enhancing statistical power and enabling these comparative analyses across multiple scRNA-seq datasets.

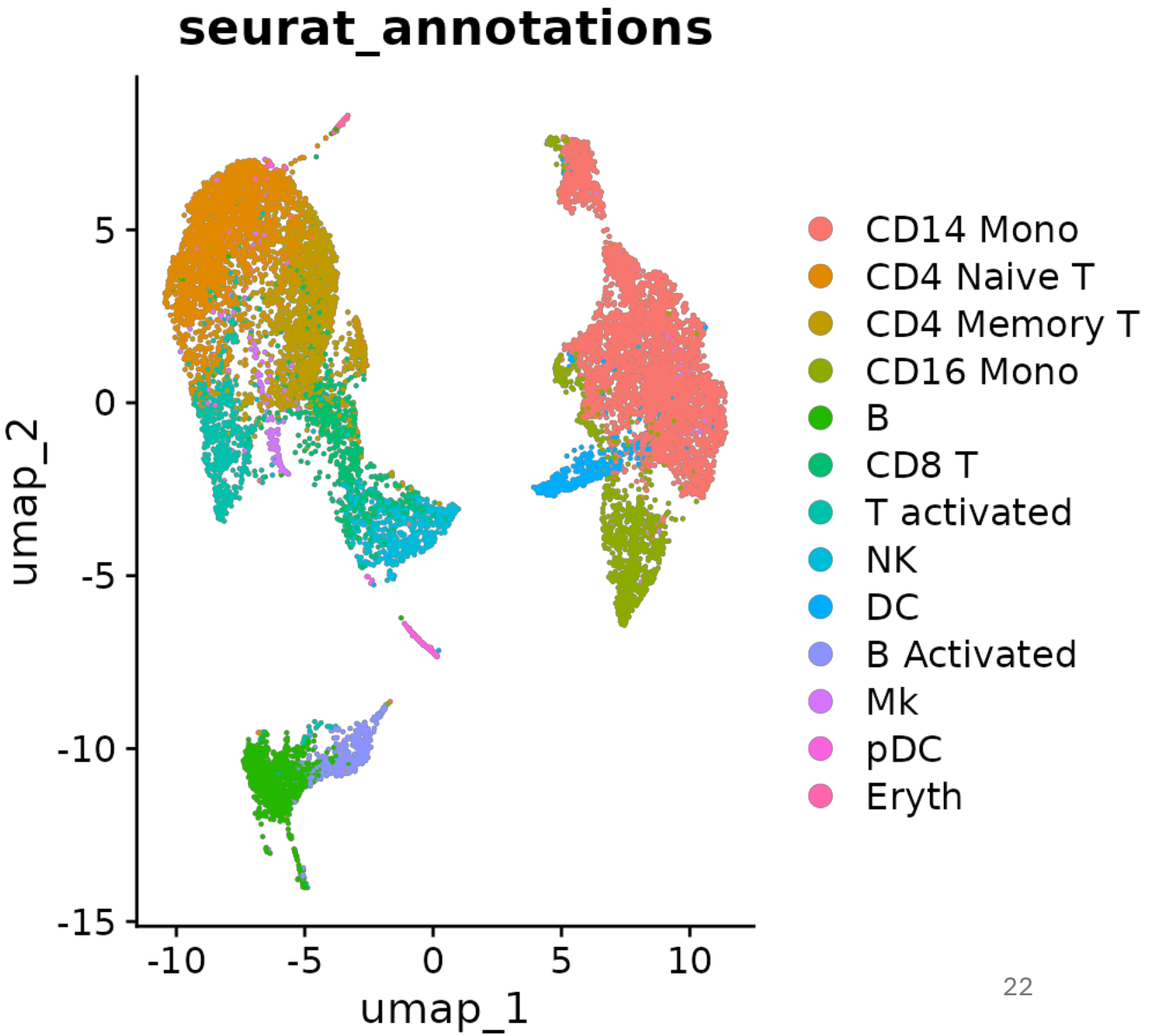
Unsupervised Clustering Without Integration



Clusters are defined by both cell-types and experimental group, complicating downstream analyses

With integration – we can group cells by their shared biology, making cell type annotation and DE analysis easier





K-MEANS Clustering

🔷 What is K-means?

- A fast-clustering algorithm that partitions data into k groups.
- Minimizes within-cluster squared error (SSE).

$$\mathrm{SSE}(X, C) = \sum_{x \in X} \min_{c \in C} \|x - c\|^2 .$$

- X: the dataset (all points)
- C: the cluster centers (centroids)
- $\| x - c \|^2$: squared Euclidean distance between a point and its closest cluster center

🔷 Challenge: Choosing the right k:

- Commonly, people use the “Elbow method”, but it is unreliable.
- Elbow plots always look similar (even for random/unclustered data).
- No theoretical support; results vary by scaling & chosen k range.

elbow could be misleading → because SSE always decreases with k, even for random data.

🔷 Better Alternatives:

- Variance-based: Variance Ratio Criterion (VRC).
- Information-theoretic: Bayesian Information Criterion (BIC).

Schubert E. Stop using the elbow criterion for k-means and how to choose the number of clusters instead. 2023;25:36–42.

Integration Summary

- **Goal:** To align same cell types across conditions.

- **Challenge:** Aligning cells of similar cell types so that we do not have clustering downstream due to differences between samples, conditions, or batches

- **Recommendation:** Go through the analysis without integration first to determine whether integration is necessary!

  (we'll talk a bit more about this later)

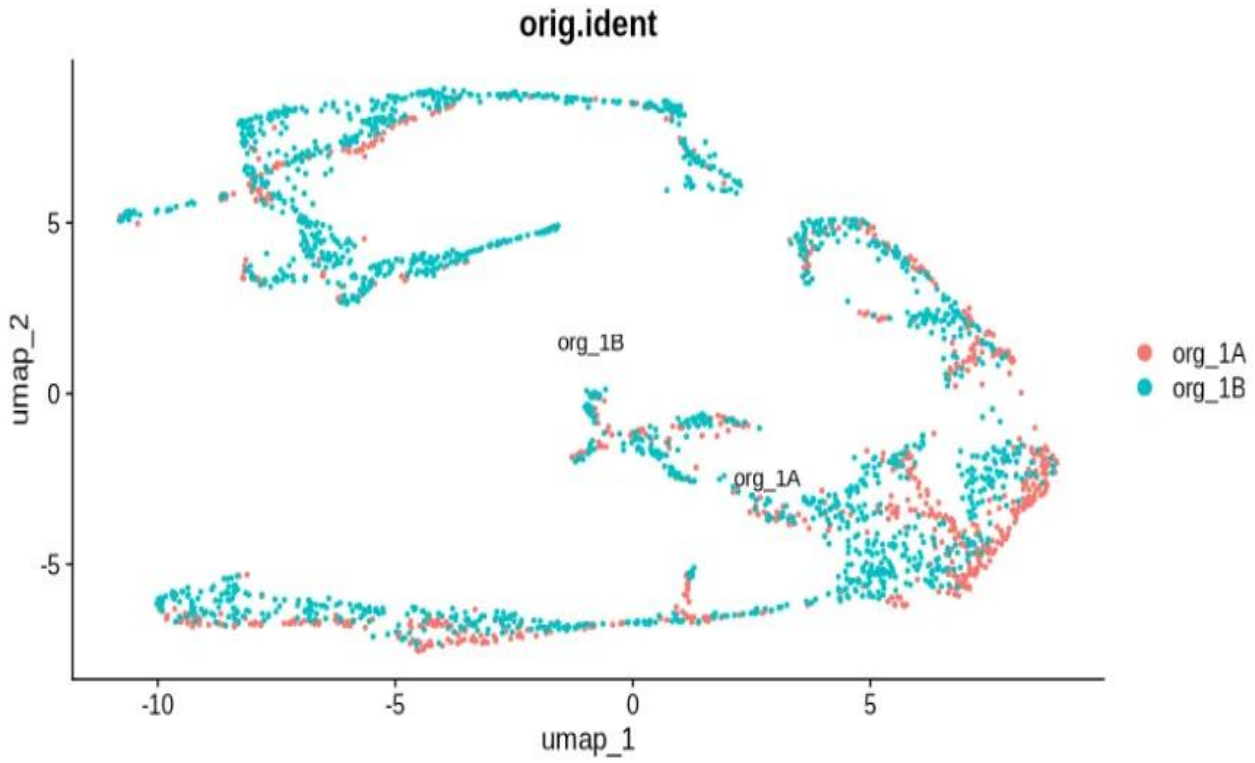
Training Material
Section 1 – Steps 3 to 5

Integration Caveats – Decide first whether its needed

- Integration can sometimes remove biologically relevant signals to artificially force cells to align.
- However, it's not always needed and can be avoided with thoughtful experimental design.

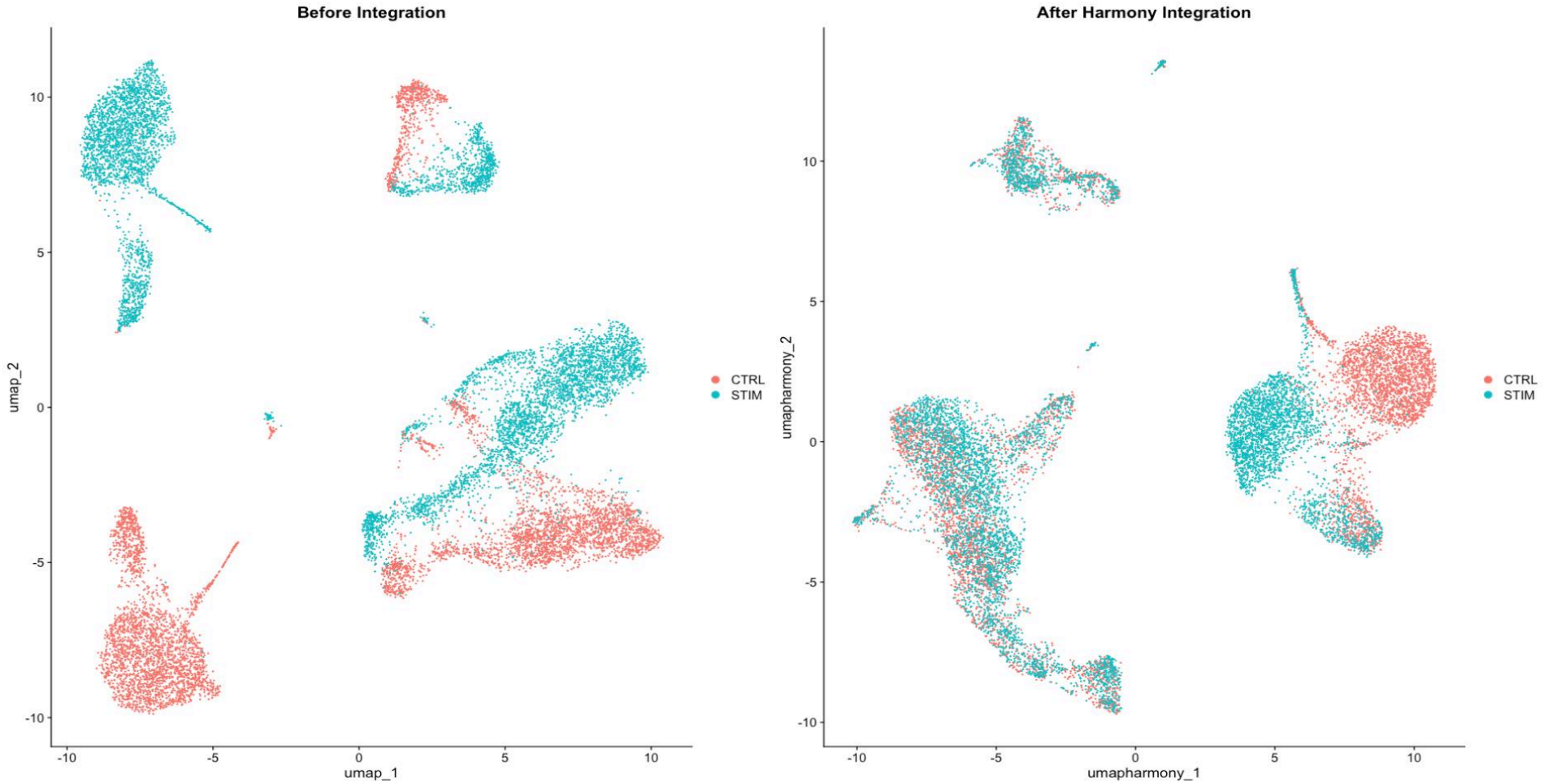
Example:

- The UMAP on the right shows two organoid samples at the same differentiation stage, processed and sequenced together.
- In this case, integration would likely result in the loss of meaningful data, with little to no benefit.



(Unpublished data)

Discussion



How can we determine whether the integration method (shown on the right) has failed due to genuine cell-type differences between the two datasets?

How do you decide on the integration tool to use?

- The optimal integration method depends on the complexity of the integration task and dataset you are working with
- Luecken et al. found that Harmony is good for simple integration tasks
- For more complex data scenarios other integration methods may be better such as Seurat CCA

Analysis | Open access | Published: 23 December 2021

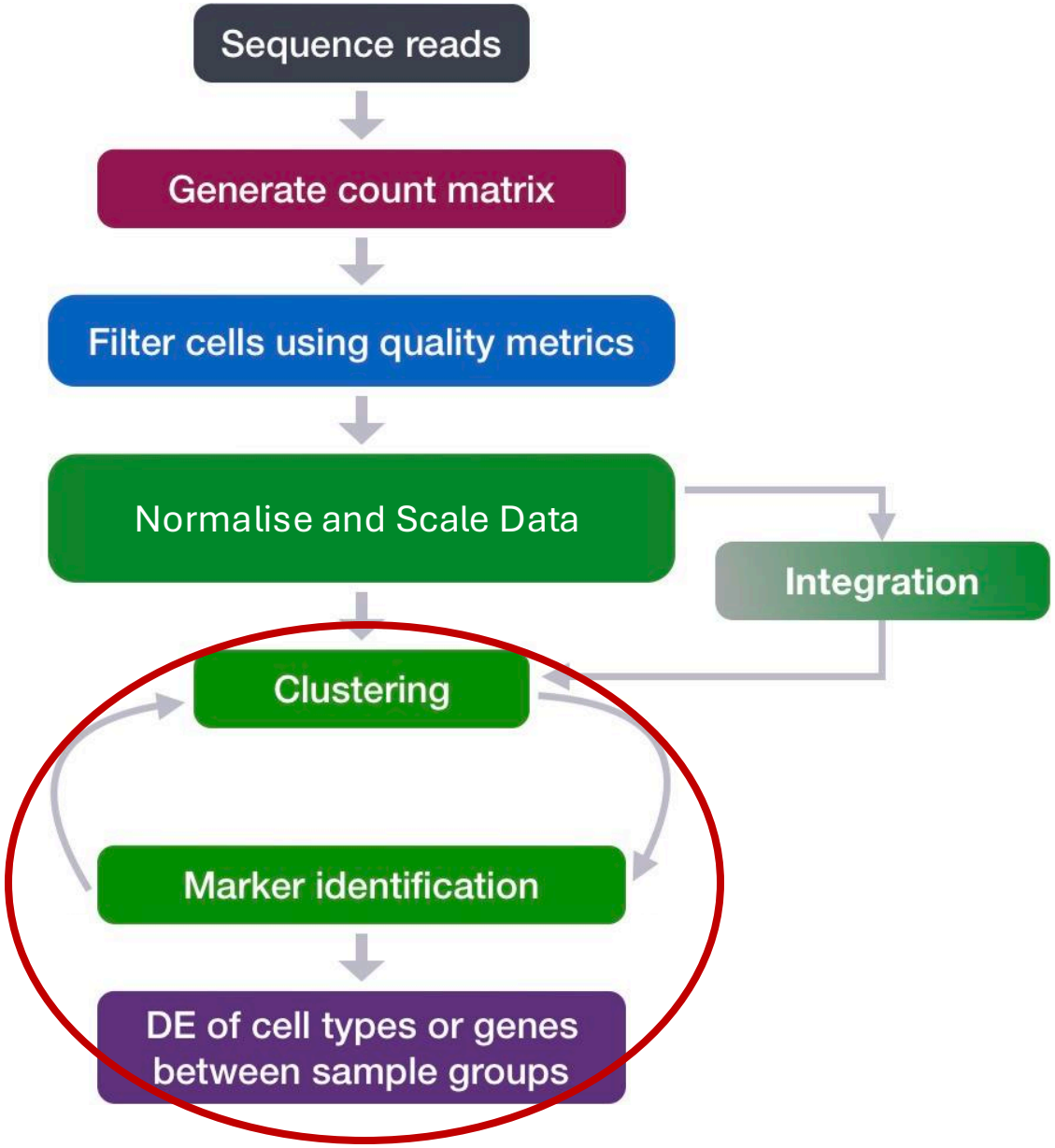
Benchmarking atlas-level data integration in single-cell genomics

Malte D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché & Fabian J. Theis

Nature Methods **19**, 41–50 (2022) | Cite this article

Break

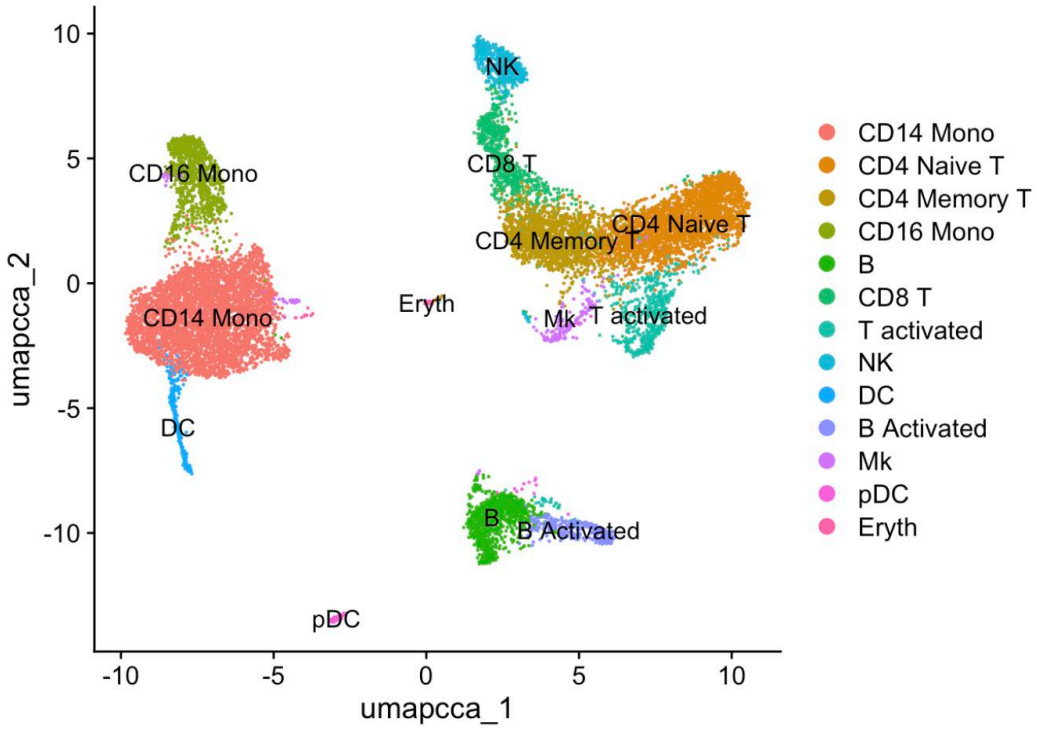
Differential Expression Analyses in Seurat



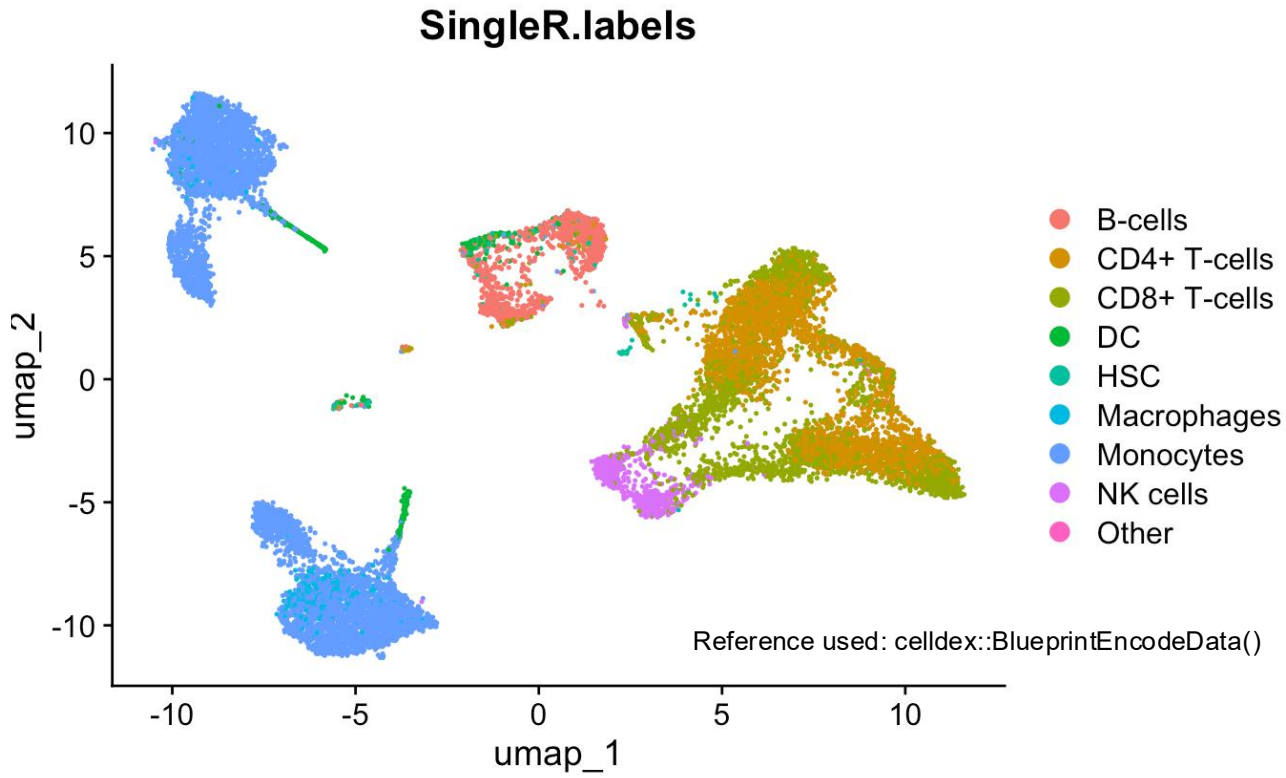
Mary Piper, Meeta Mistry, Jihe Liu, William Gammerdinger, & Radhika Khetani. (2022, January 6). hbctraining/scRNA-seq_online: scRNA-seq Lessons from HCBC (first release). Zenodo. https://doi.org/10.5281/zenodo.5826256.

Cell type labelling

Marker-based (manual) annotation after clustering



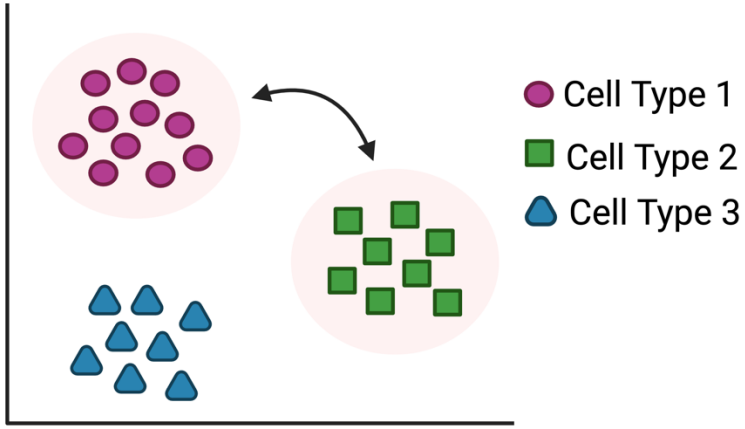
Automated labelling: Celldex and SingleR



In-built Seurat Functions for DE Analysis

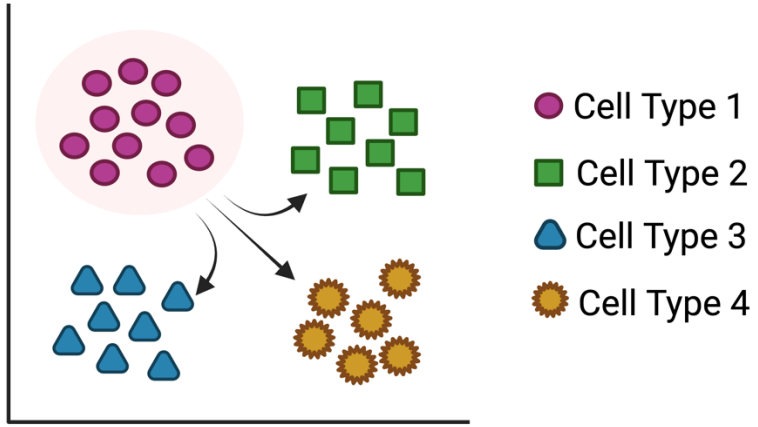
findMarkers()

Find DEGs between two clusters



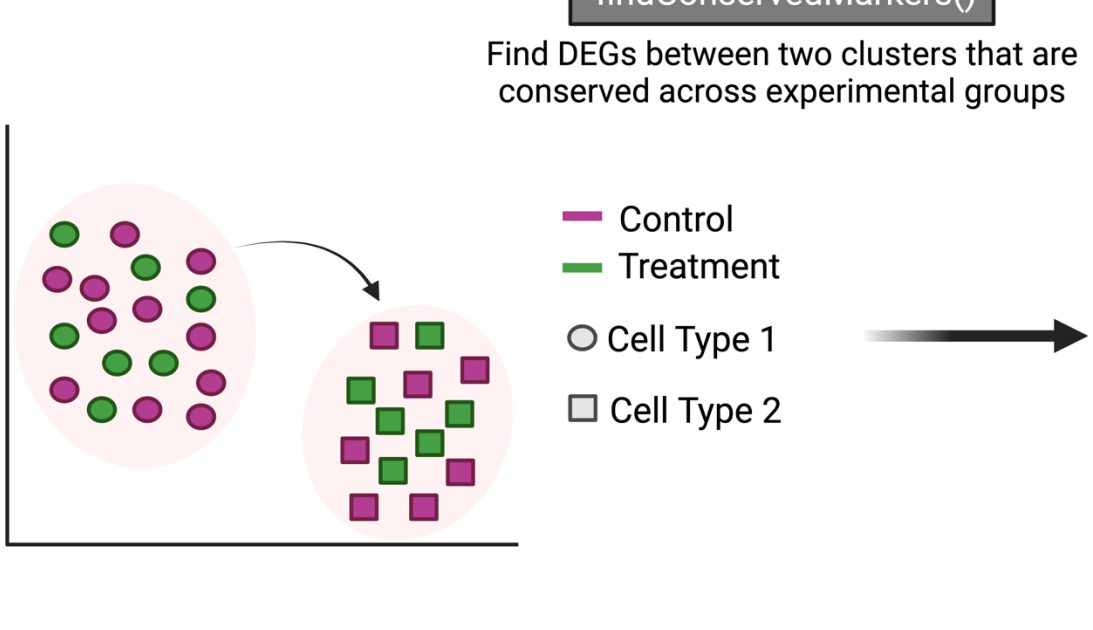
findAllMarkers()

Find DEGs in a cluster compared to all clusters



findConservedMarkers()

Find DEGs between two clusters that are conserved across experimental groups



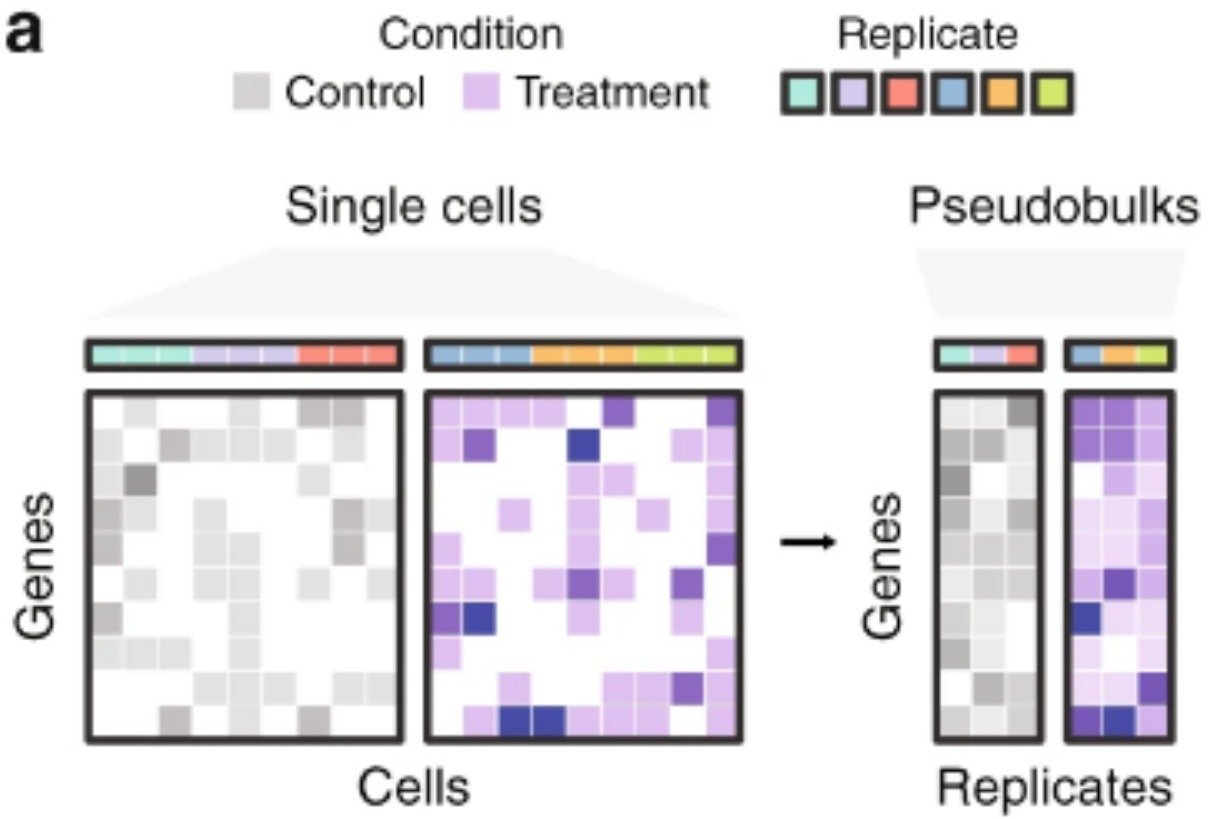
Training Material
Section 2 – Steps 1 to 6

Pseudobulk Analyses – An alternative DE approach

- Combines single-cell counts and metadata into 'bulk' count matrices at the sample or replicate level.

Advantages:

- Uses well-established bulk RNA-seq tools (DESeq2, edgeR, limma).
- Enhances statistical robustness by averaging out single-cell variability and reducing sparsity.
- Facilitates straightforward DE analysis with familiar methods.



https://www.nature.com/articles/s41467-021-25960-2

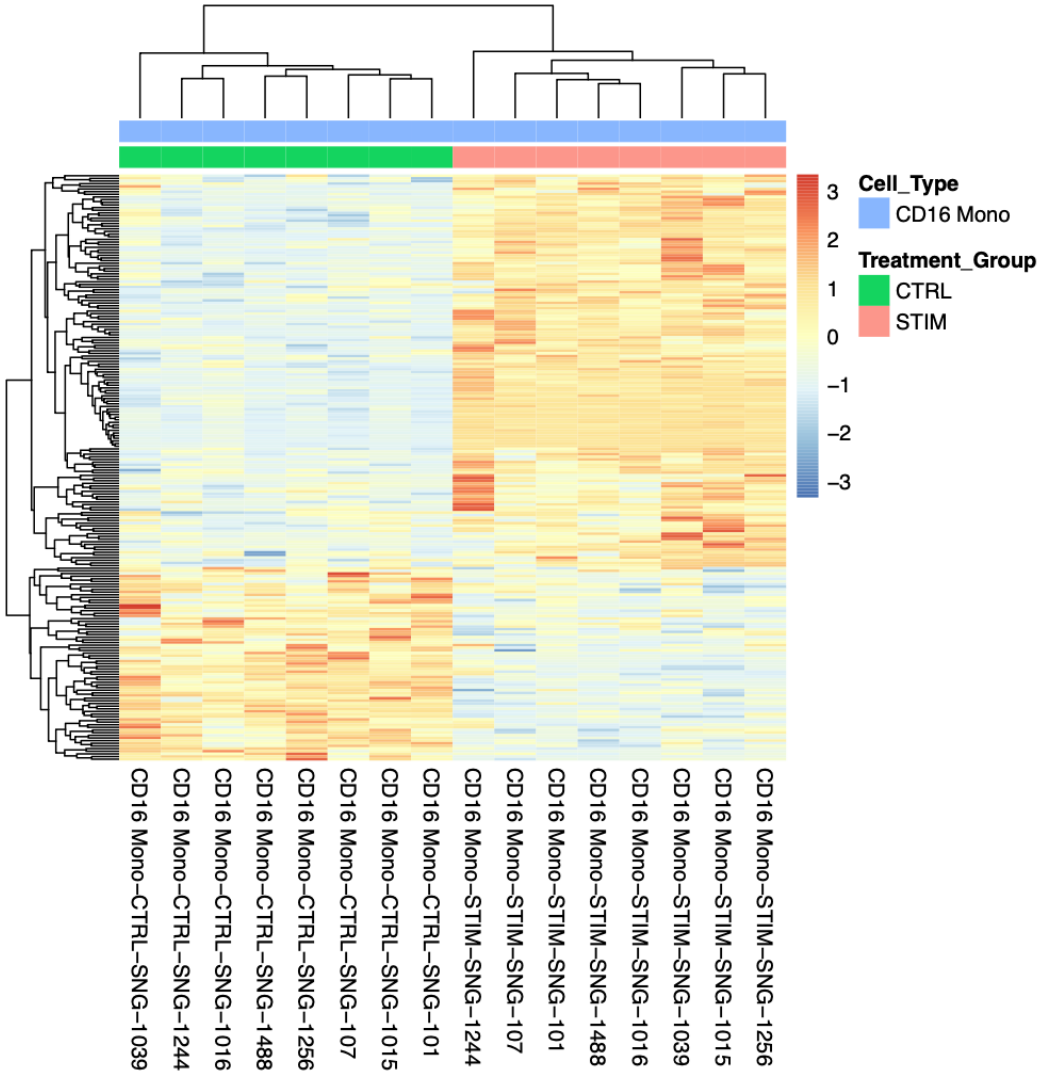
Why use a pseudobulk approach?

- scRNA-seq data is notoriously sparse, with a complicated distribution and heterogeneity across and within cell populations.
- Single-cell DE methods often struggle to identify low-expression DEGs and overemphasize highly expressed genes.
- **Inflates p-values by treating individual cells as separate samples, reducing statistical reliability.**
- Pseudobulk analysis aggregates cells by sample, preserving cell-type resolution while allowing for statistical testing using bulk RNA-seq tools
    - This leads to more accurate and robust differential expression findings.

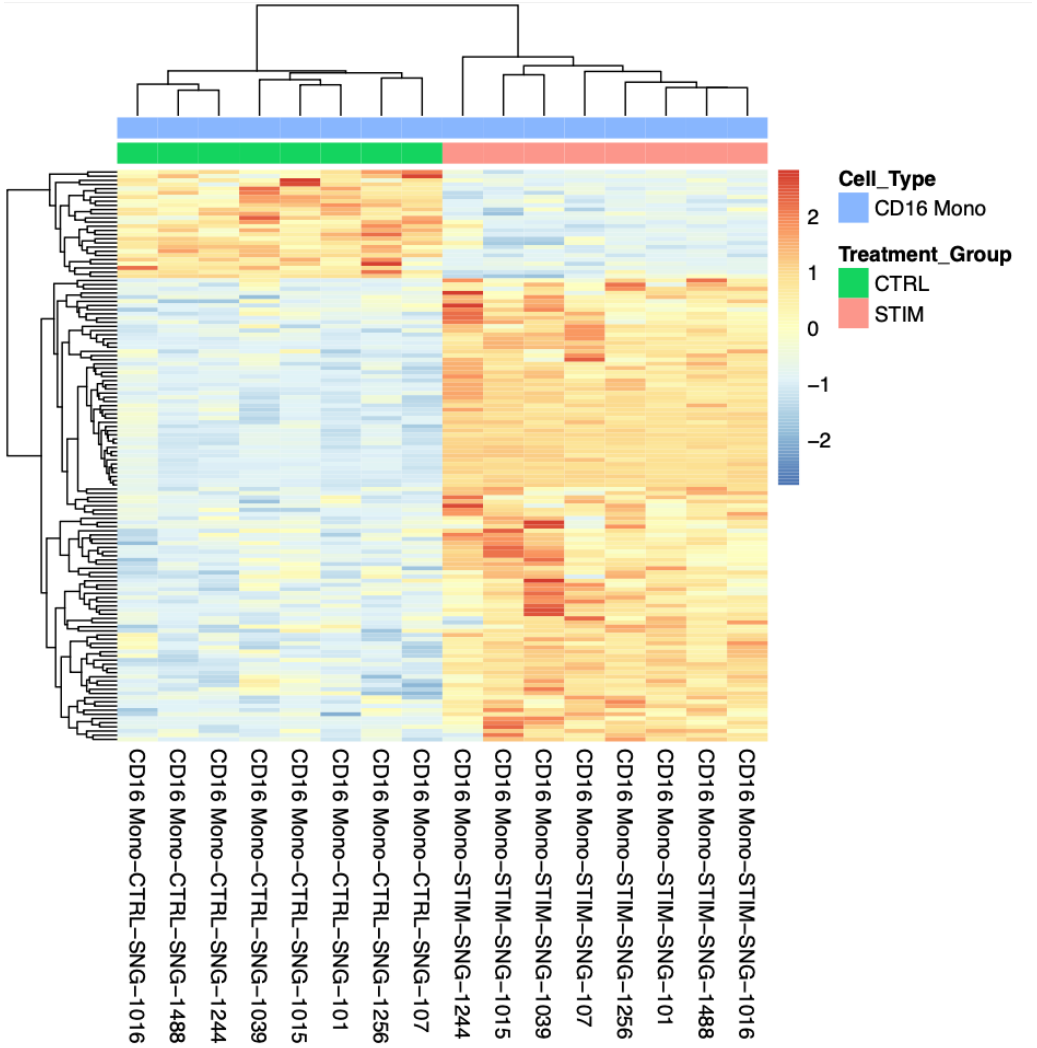
Training Material
Section 3 – Steps 1 to 5

Discussion: Compare single-cell versus pseudo-bulk DE approaches

These heatmaps display the expression of differentially expressed genes (DEGs) along the y-axis, with cells grouped by patient replicates on the x-axis. Can you spot the differences?

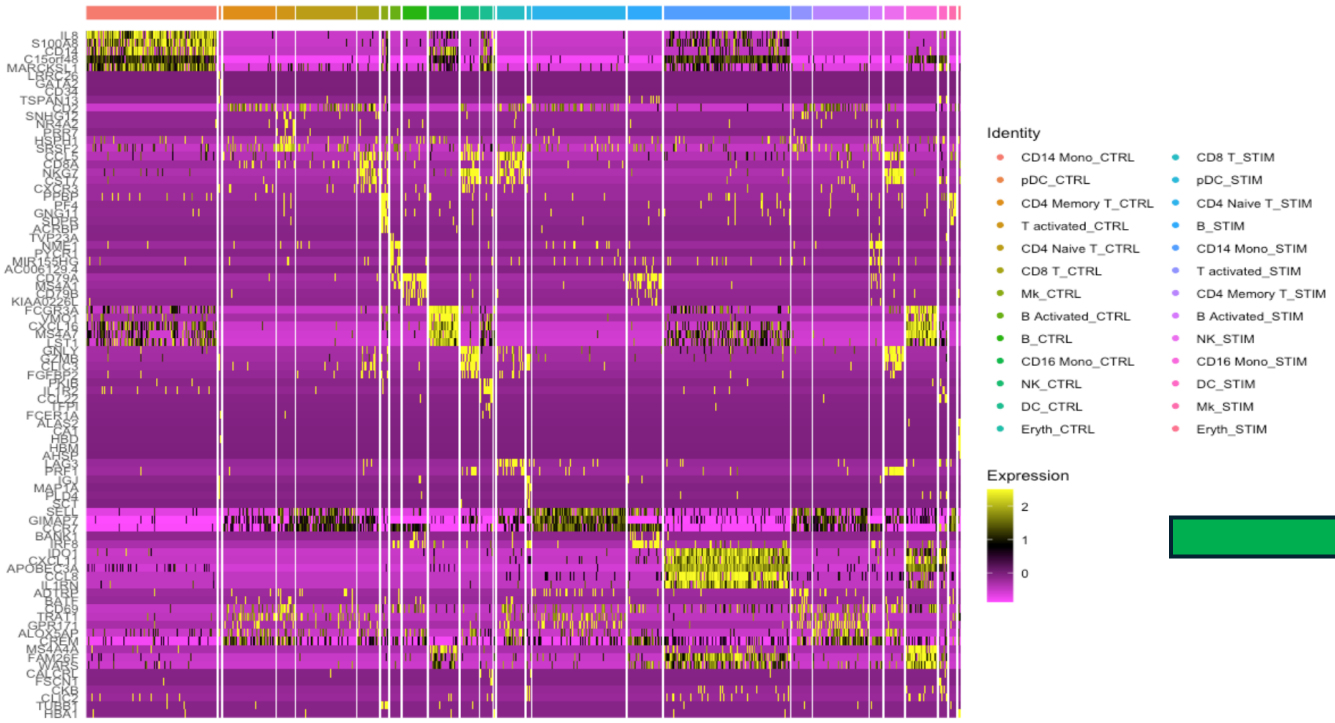


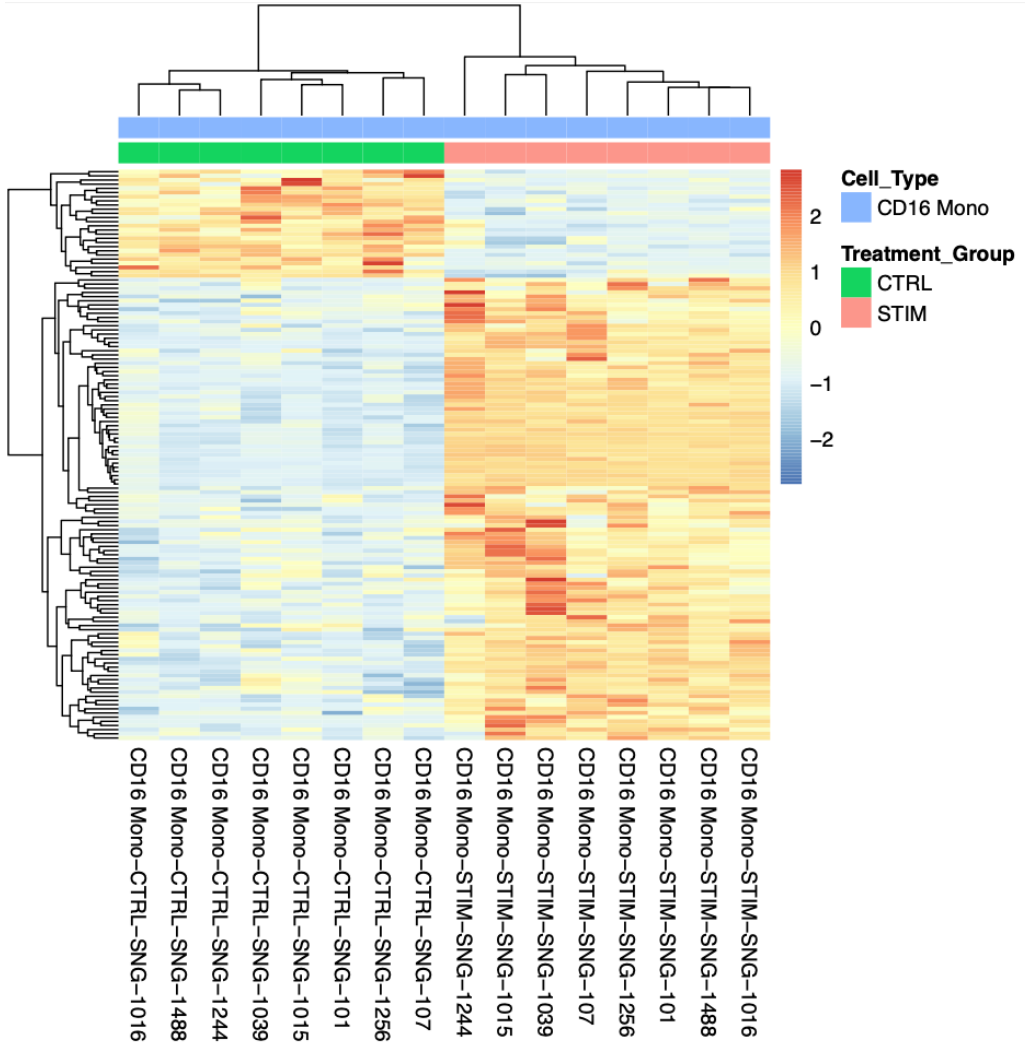
DEGs found by Seurat single-cell method



DEGs found by DESeq2 pseudo-bulk method

Walk Through: Extracting DEG data from Seurat to make custom visualisations with other packages (pheatmap)





Training Material
Section 3 – Step 6

What comes next?

1. Gene Ontology (GO) Enrichment Analysis

- Perform GO enrichment analysis to identify biological processes, molecular functions, or cellular components that are significantly enriched in your DEG list.
- Tools like **clusterProfiler** in R can help you analyse and visualize these functional categories.

2. Pathway Analysis

- Use tools such as **KEGG and Reactome** to map your DEGs onto known biological pathways. This helps in understanding the broader biological context of gene expression changes.
- **GSEA (Gene Set Enrichment Analysis)** can also be used to assess whether specific gene sets (e.g., pathways) are significantly enriched in your data.

3. Validation with External Datasets

- Compare your DEGs with external datasets or publicly available single-cell RNA-seq datasets to validate your findings or explore how they relate to known disease states, tissues, or conditions.

Melbourne Bioinformatics
BIOINFORMATICS + DATA SERVICES + INFRASTRUCTURE, FOR LIFE SCIENCES TODAY

THE UNIVERSITY OF MELBOURNE

Please complete our survey before you leave today.