scRNA-seq Integration and Differential Expression Workshop

Working with treatment versus control data

Learning Outcomes

- Understand and get comfortable using various integration strategies
- Understand all DE functions offered by Seurat and when to use them: FindMarkers(), FindConservedMarkers(), and FindAllMarkers()
- Learn how to use DE tools meant for bulk data (e.g. DESeq2 and limma) for single cell ‘pseudobulk’ data, and understand why you might choose this approach
- Learn different ways to visualise DEGs 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)

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
  - Conduct differential expression analysis using an alternative ‘pseudobulk’ approach

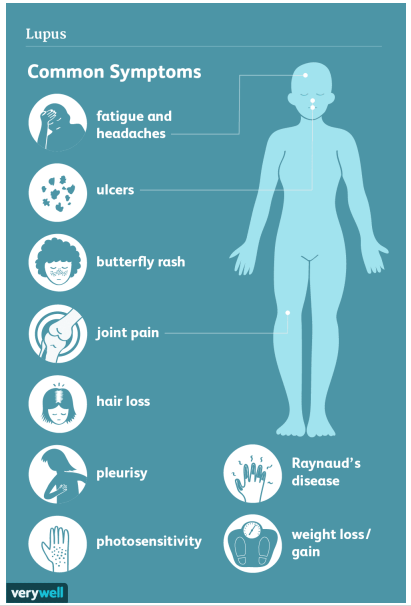
Article | Published: 11 December 2017

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

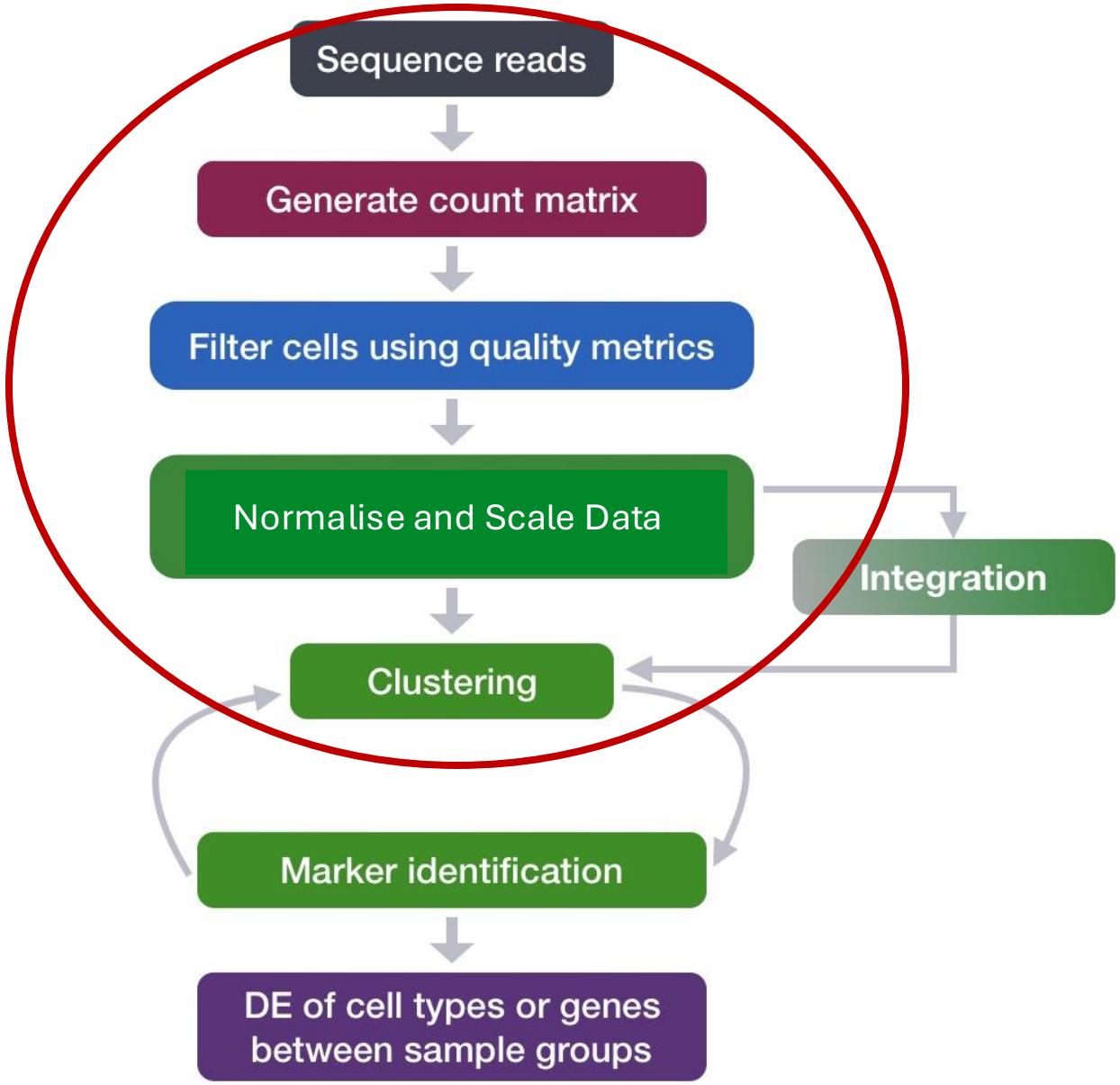
Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, Rachel E Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A Criswell & Chun Jimmie Ye

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

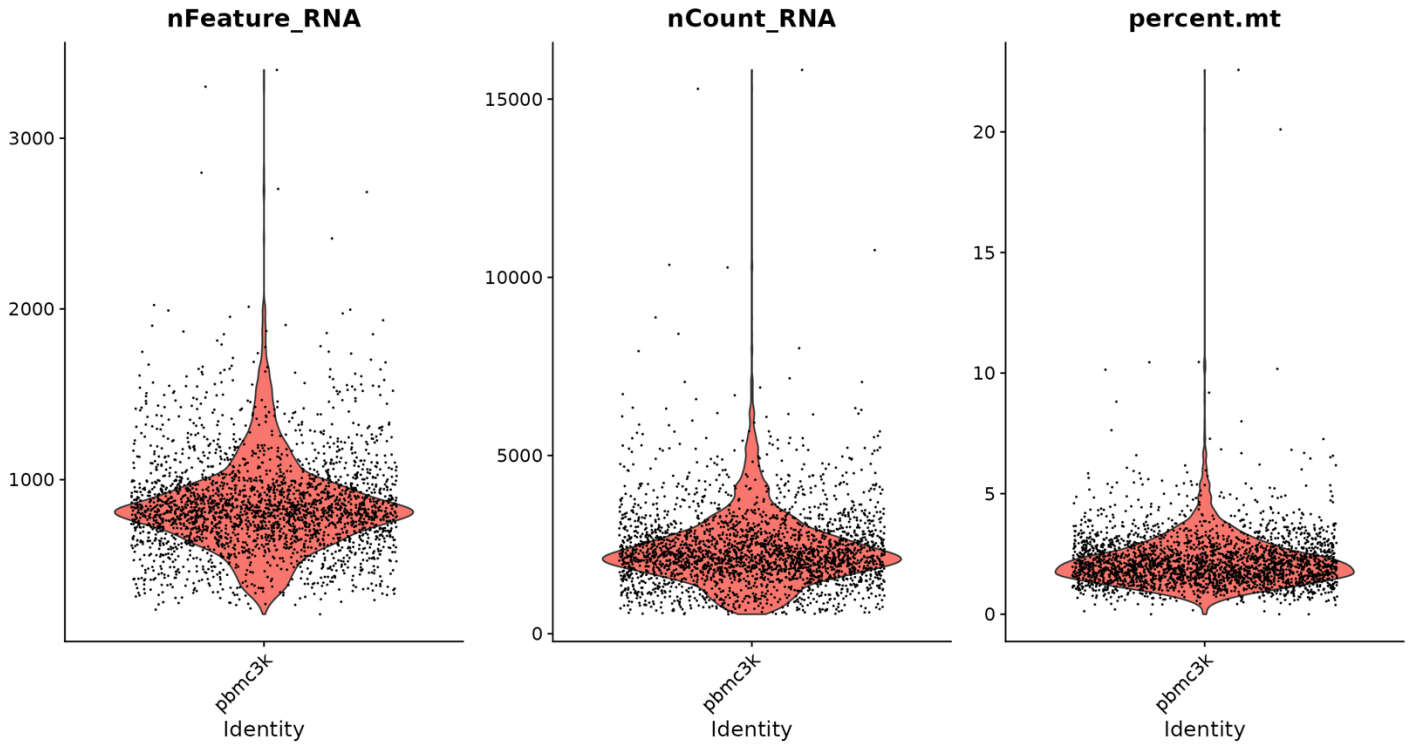
68k Accesses | **481** Citations | **177** Altmetric | Metrics



Refresher: General scRNA-seq Workflow



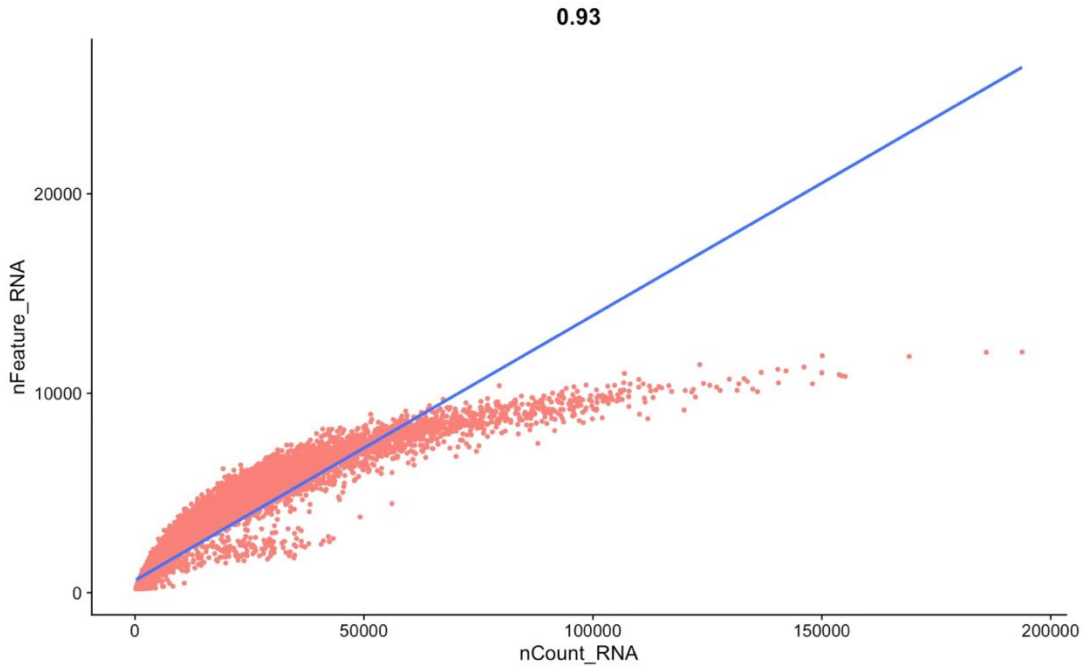
Guidelines for removing low quality cells



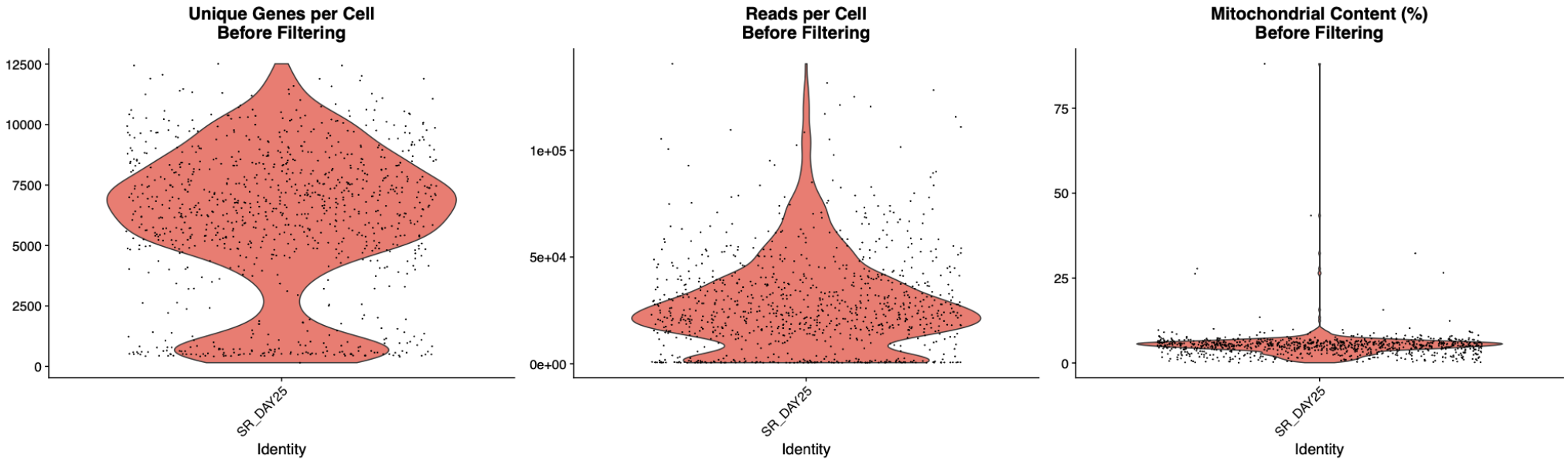
- Low quality cells or empty droplets will have fewer genes and fewer counts
- Cell doublets (>1 cell assigned to a single barcode) will have significantly more genes and counts
- Dying cells will have higher mitochondrial contamination
  - (<=5% is a good guideline)
- We can use violin plots to determine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

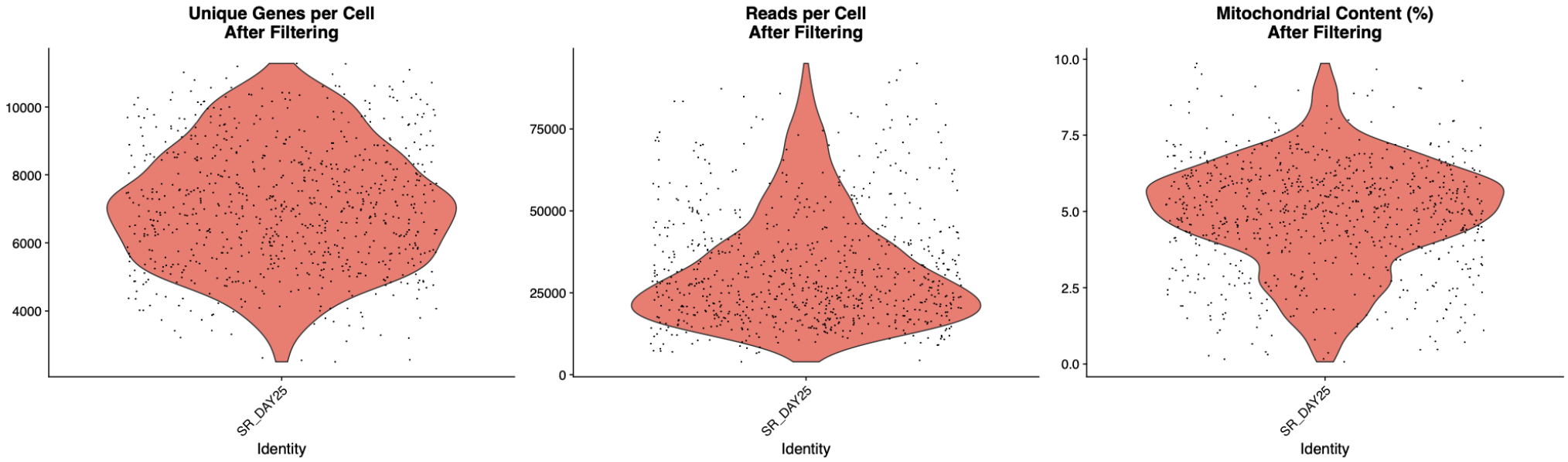
Consider Metrics Together: Feature and Molecule Association Plots

- 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

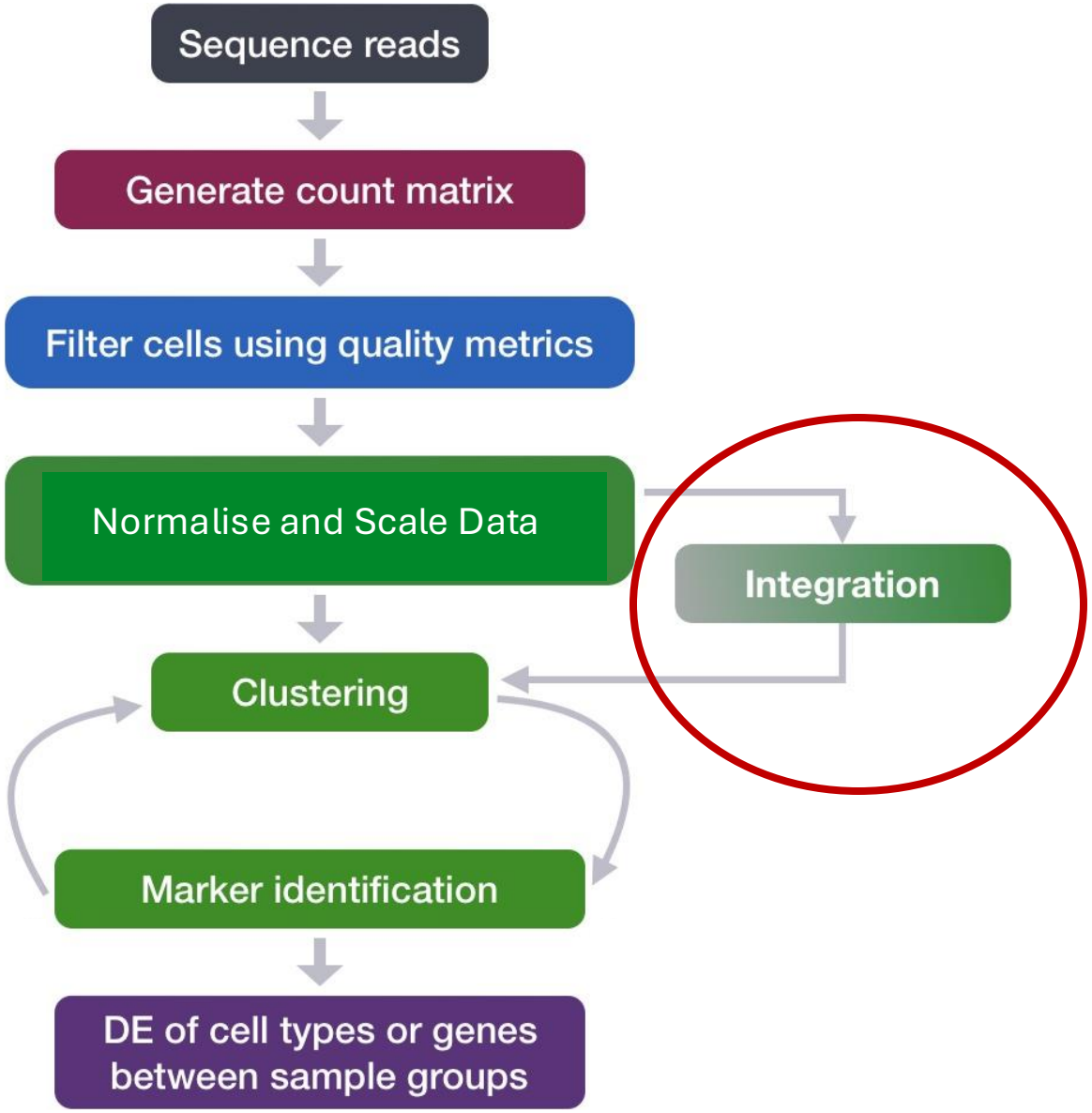


Example Before and After QC Plots





Integration – What, When, Why?



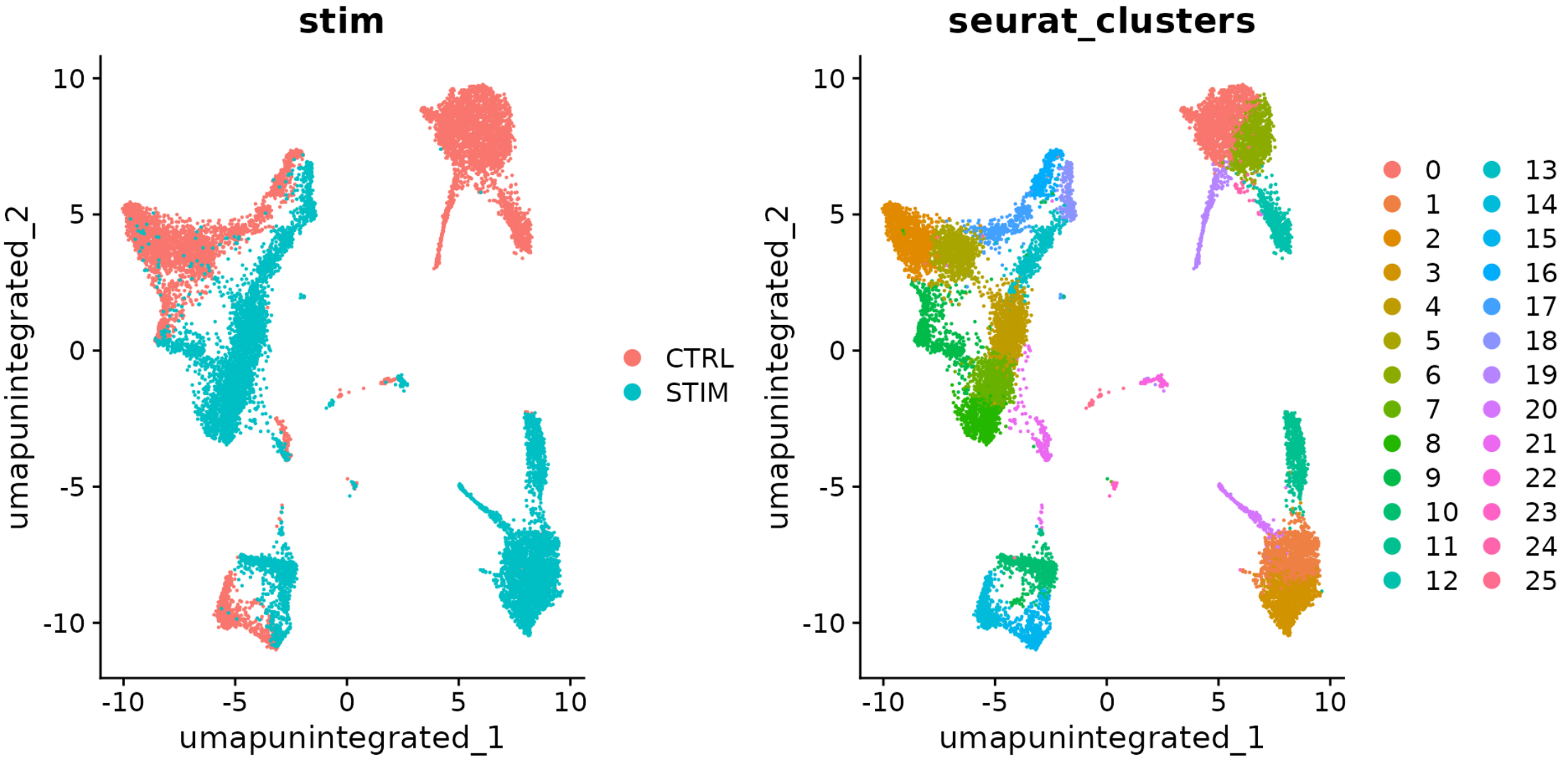
Integration – What, When, Why?

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 stimulation/condition.

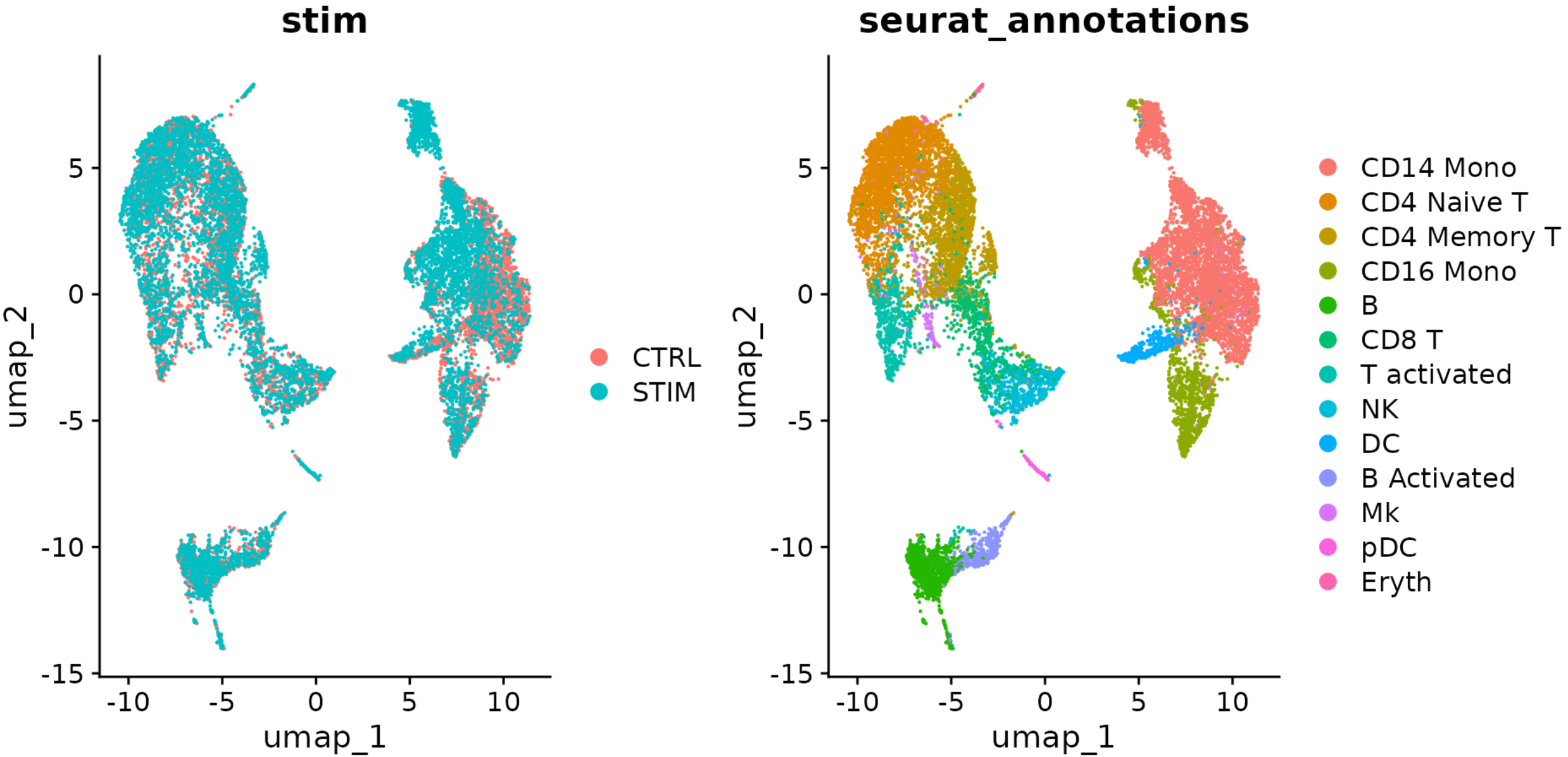
These steps rely on **integration**—a process that aligns shared cell states across datasets, enhancing statistical power and enabling these comparative analyses across multiple scRNA-seq datasets.

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



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, modalities, or batches

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

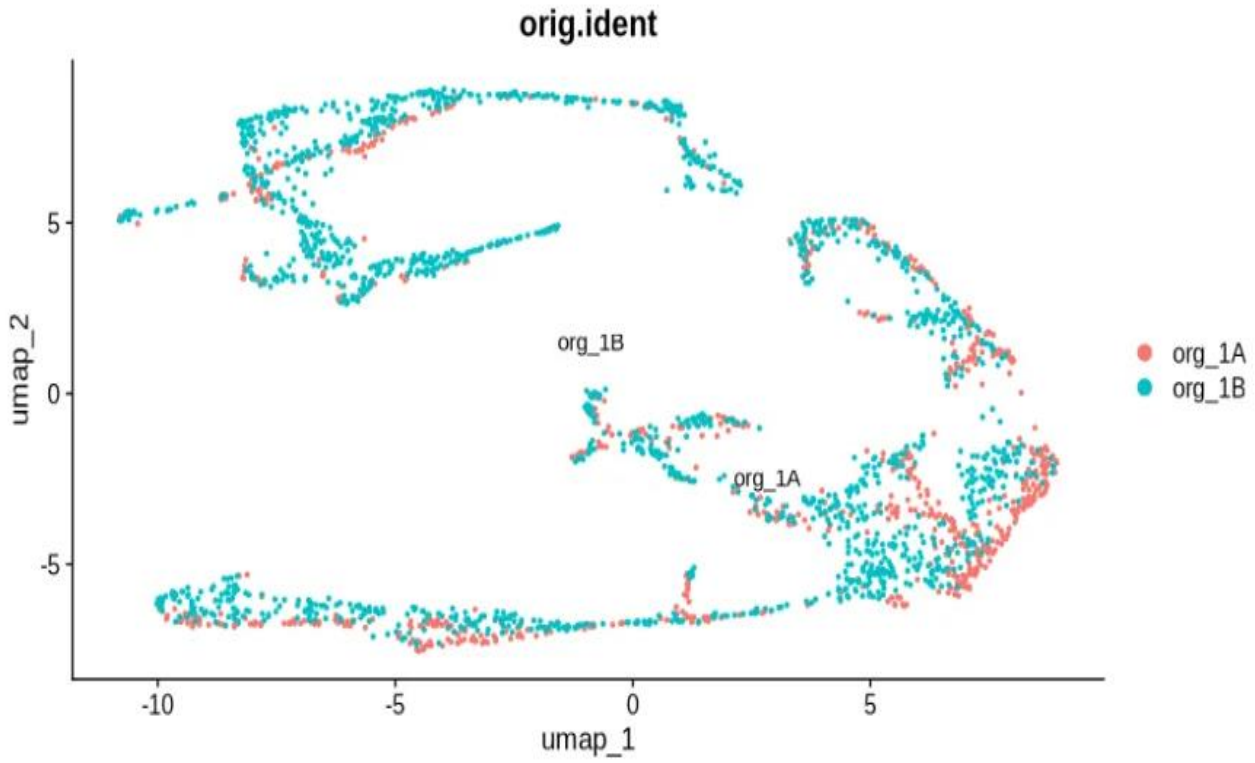
  (see next slide)

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.

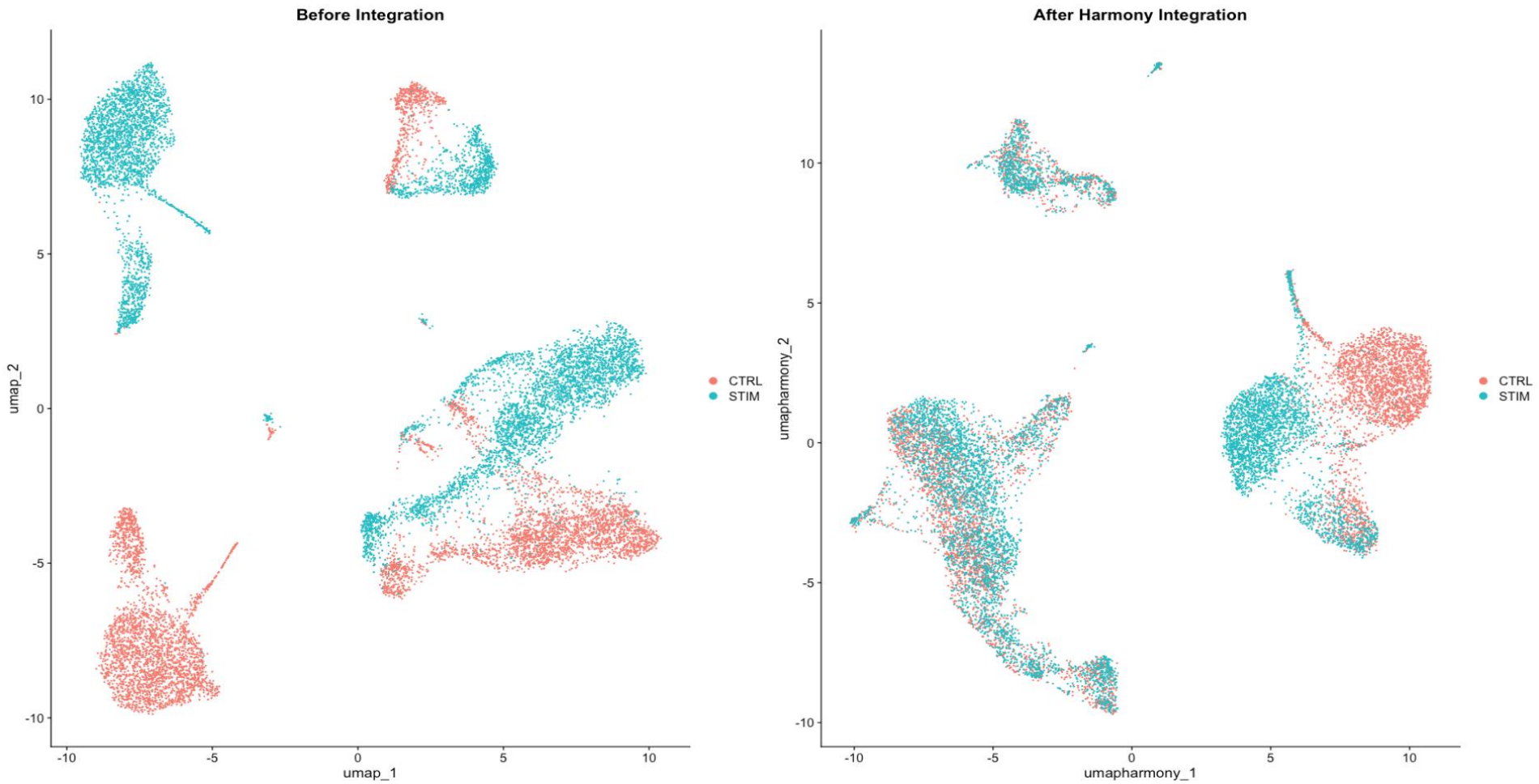
Example:

- 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

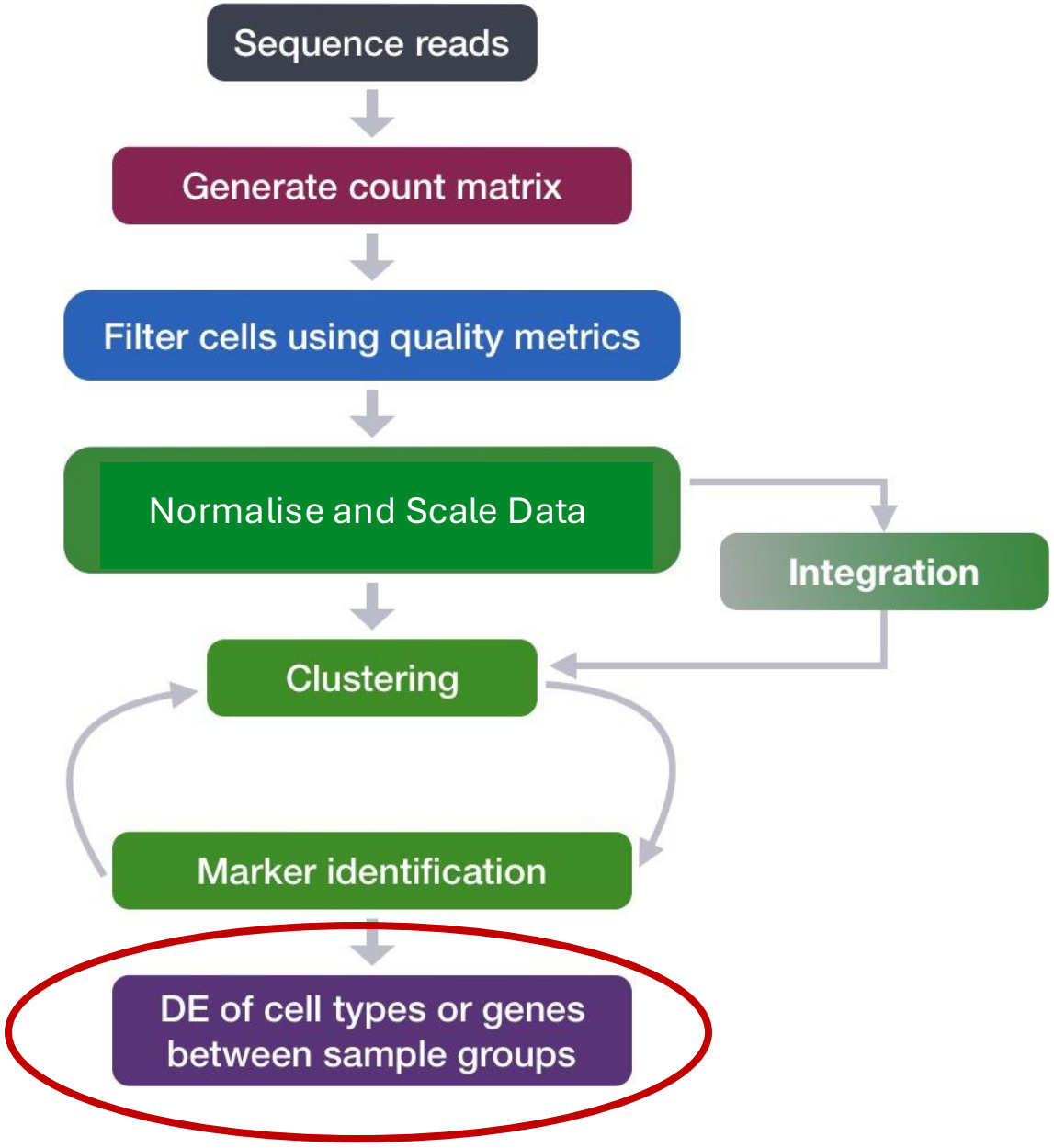
Benchmarking atlas-level data integration in single-cell genomics

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

135k Accesses | **368** Altmetric | Metrics

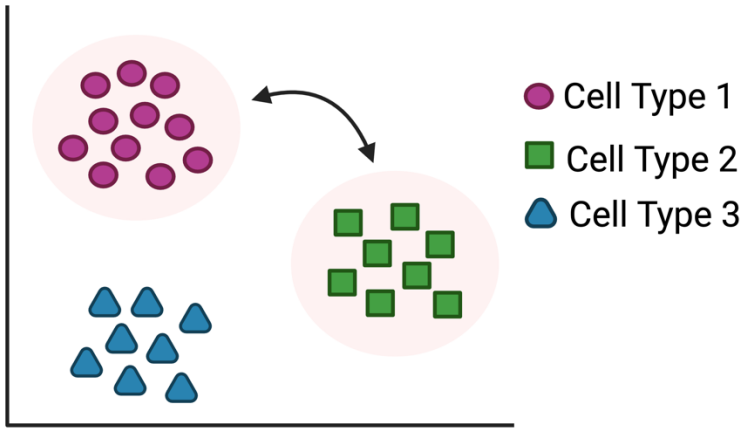
Differential Expression Analyses in Seurat



In-built Seurat Functions for DE Analysis

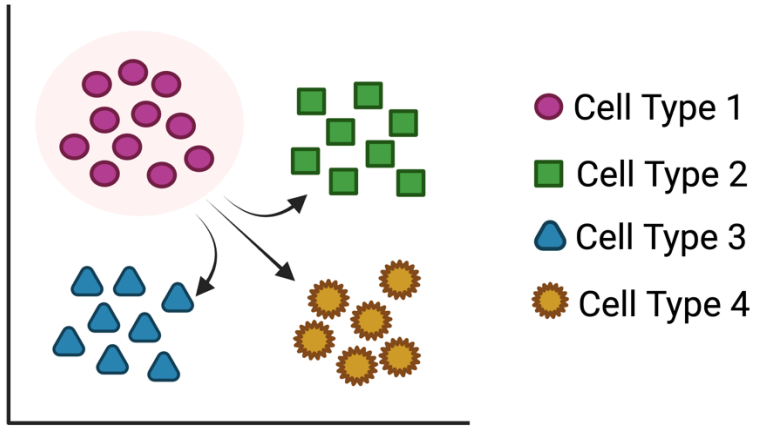
findMarkers()

Find DEGs between two clusters



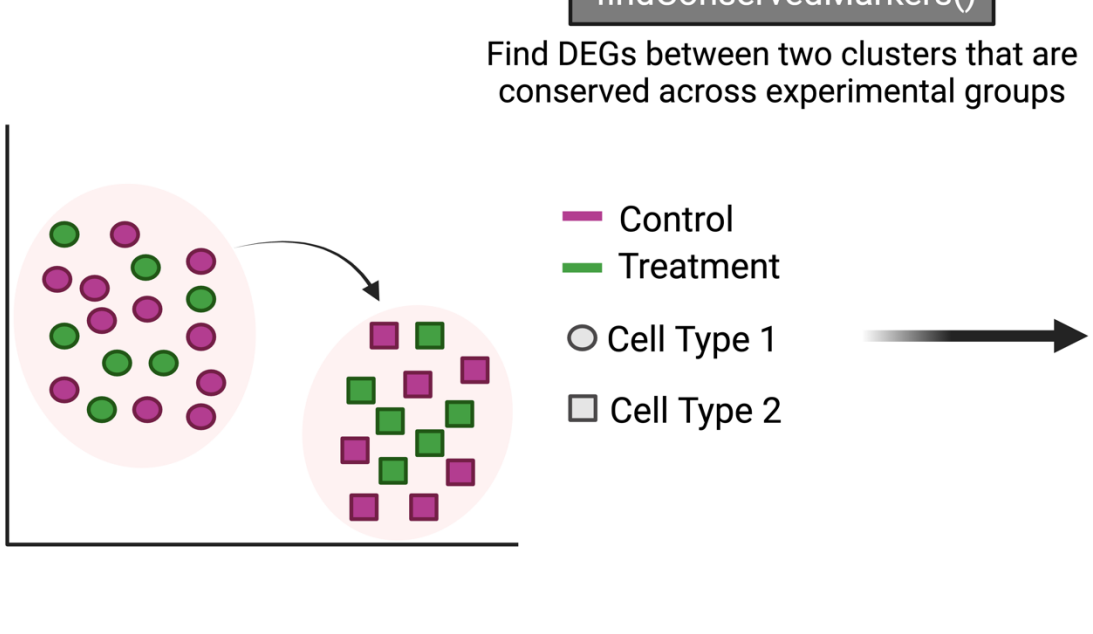
findAllMarkers()

Find DEGs in a cluster compared to all clusters



findConservedMarkers()

Find DEGs between two clusters that are conserved across experimental groups

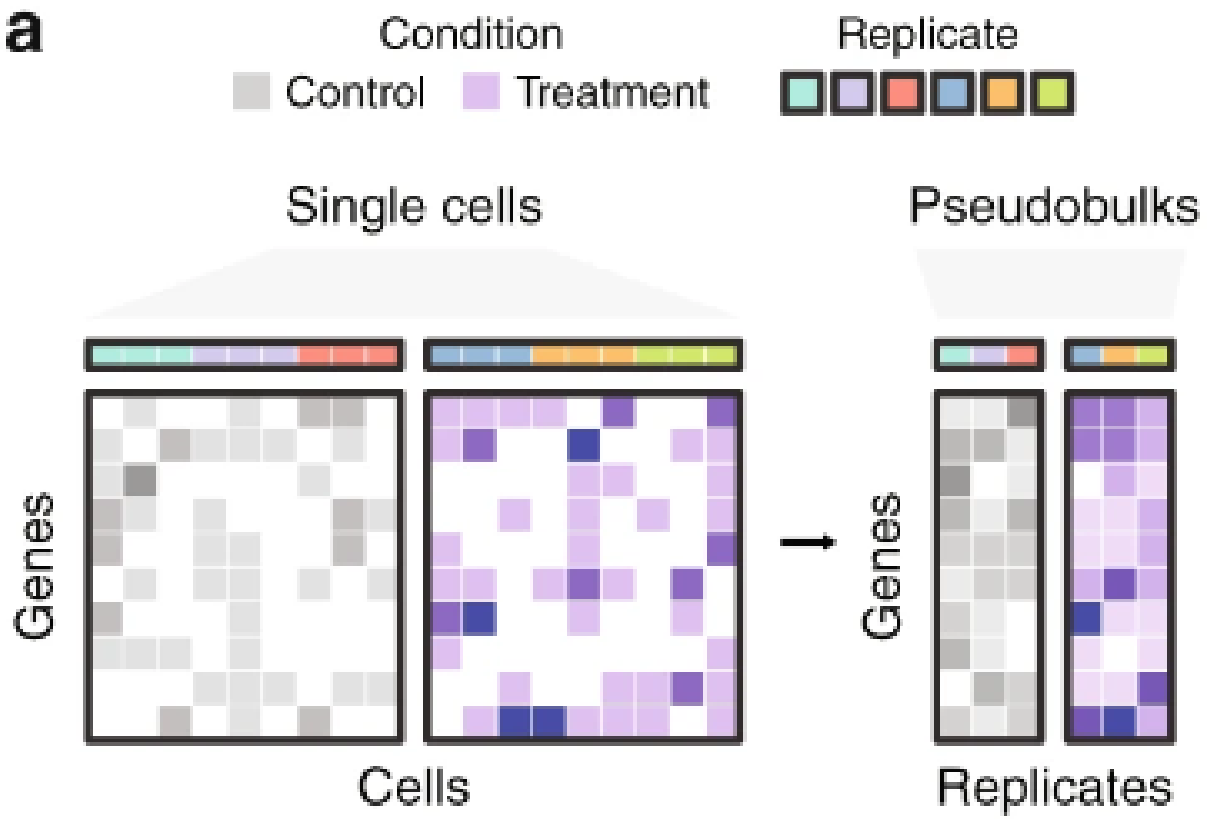


Pseudobulk Analyses – An alternative DE approach

- 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.



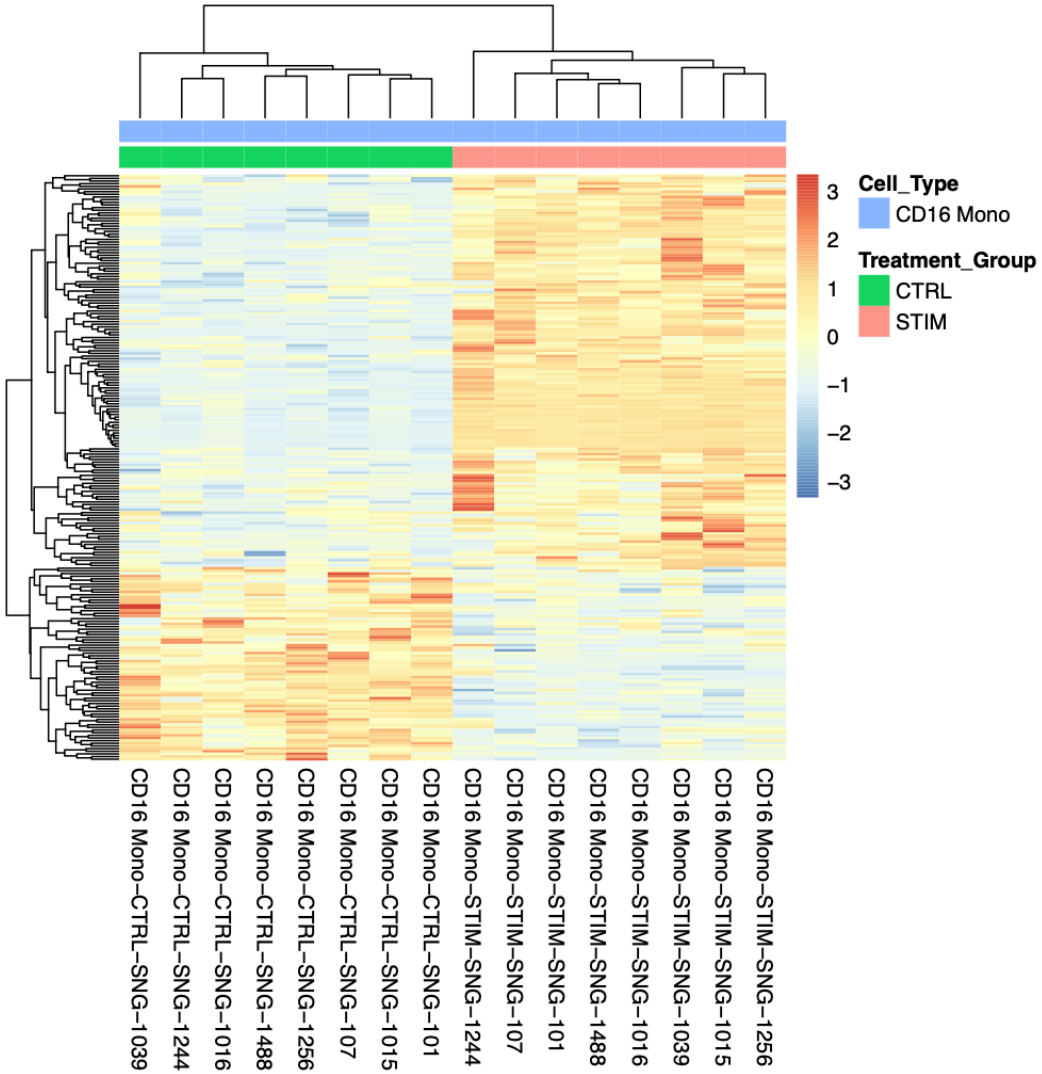
https://www.nature.com/articles/s41467-021-25960-2

Why use a pseudobulk approach?

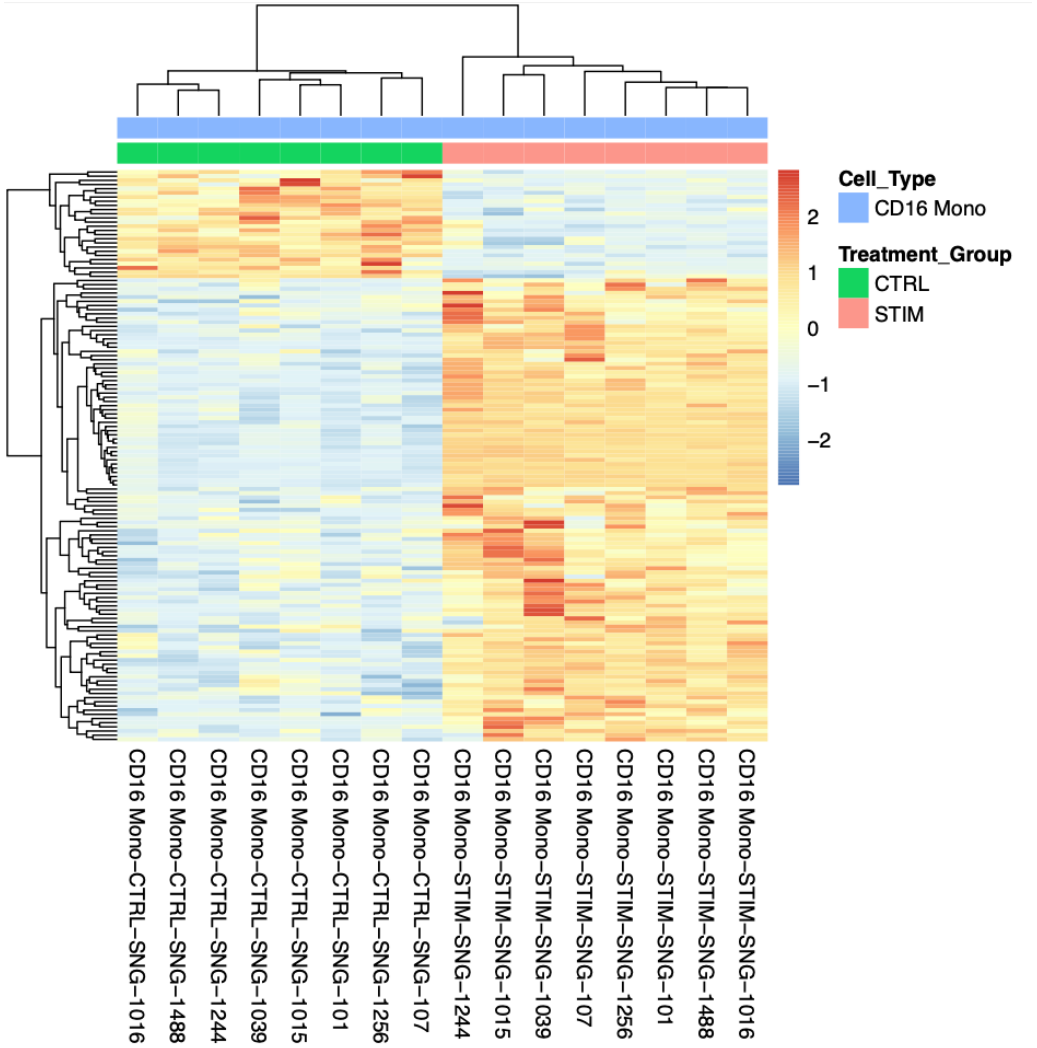
- scRNA-seq data is notoriously sparse, with a complicated distribution and substantial heterogeneity across and within cell populations.
- Single-cell DE methods often struggle to identify low-expression DEGs and overemphasize highly expressed genes.
- They also inflate p-values by treating individual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

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?

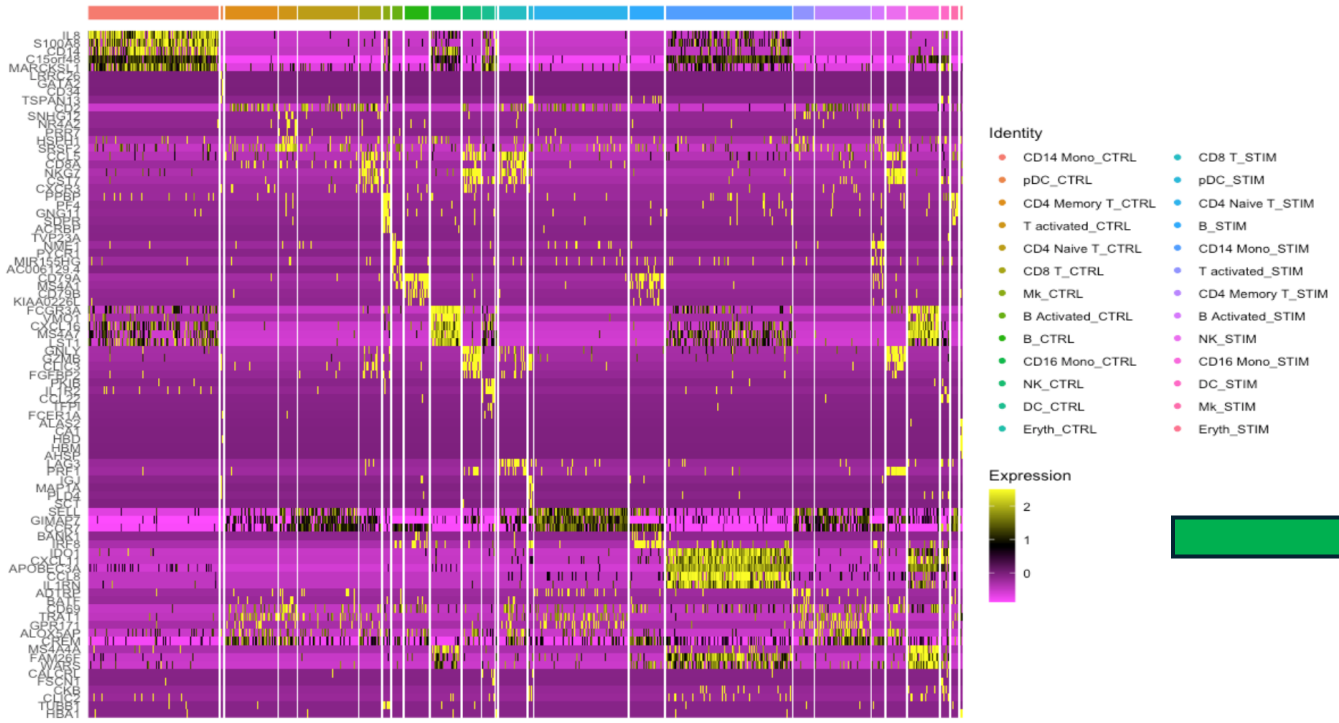


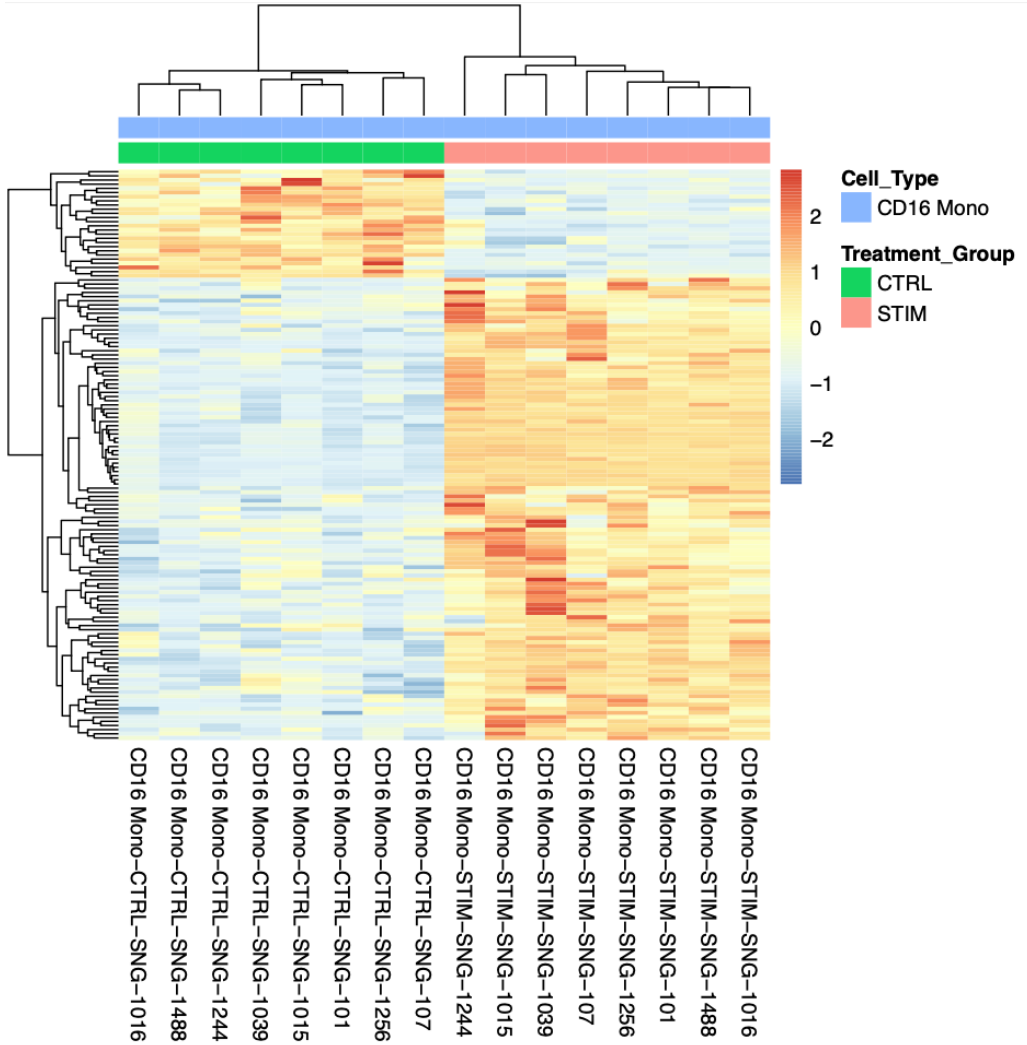
DEGs found by Seurat single-cell method



DEGs found by DESeq2 pseudo-bulk method

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





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 or **DAVID** can help you analyse and visualize these functional categories.

2. Pathway Analysis

- Use tools such as **KEGG**, **Reactome**, or **Ingenuity Pathway Analysis (IPA)** 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 such as **GTEx**, **TCGA**, or publicly available single-cell RNA-seq datasets to validate your findings or explore how they relate to known disease states, tissues, or conditions.