Differential gene expression

Jan Kubovčiak
with acknowledgements to Deepak Tanwar and SIB

Goals

- Identify cell cluster/group markers
- Identify differentially expressed genes between groups of cells from biological replicates

Goals

- Identify cell cluster/group markers
  - Identify genes that are differentially expressed between cell populations
  - e.g., cell types, clusters, or conditions while accounting for the single-cell nature of the data (e.g., sparsity, dropout events)
- Identify differentially expressed genes between groups of cells from biological replicates
  - Aggregate single-cell data into pseudo-bulk profiles to perform DGE analysis using bulk RNA-seq methods, reducing noise and leveraging biological replicates.

Marker identification

Use Cases:

- Annotating cell types (e.g., identifying CD3 as a marker for T cells)
- Discovering novel cell states or subpopulations

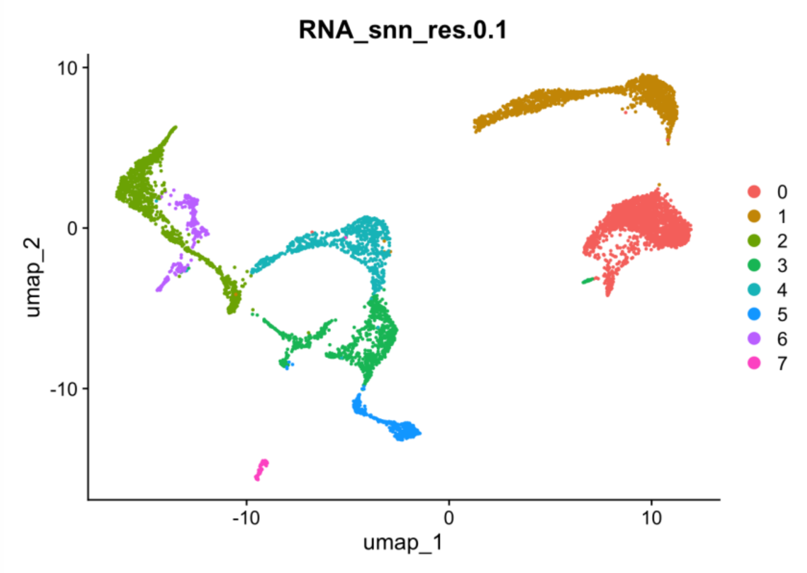
Methods:

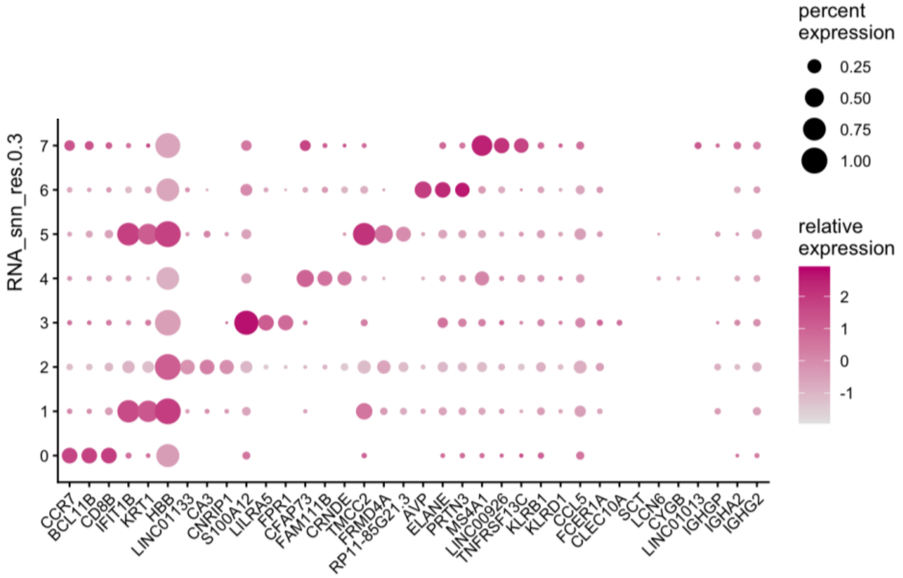
- Log-Fold Change (LFC) Analysis: Calculate the log-fold change in expression between a target cluster and all other cells.
- Wilcoxon Rank-Sum Test: Test for genes with significantly higher expression in one group compared to others (e.g., in Seurat).

Challenges:

- Marker genes may not be unique to a single cell type, requiring careful validation
- Dropout events can obscure marker gene detection

Marker identification

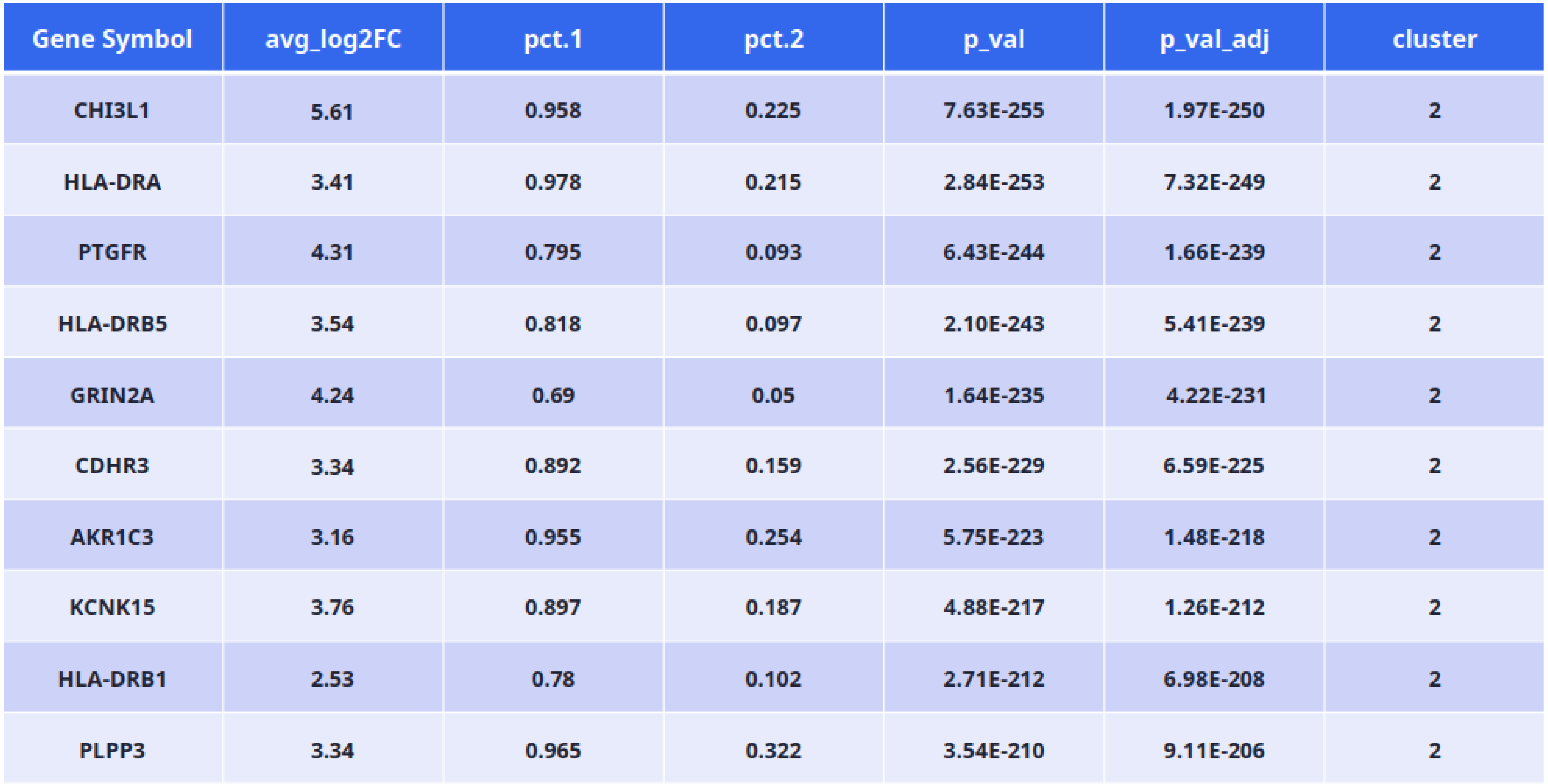




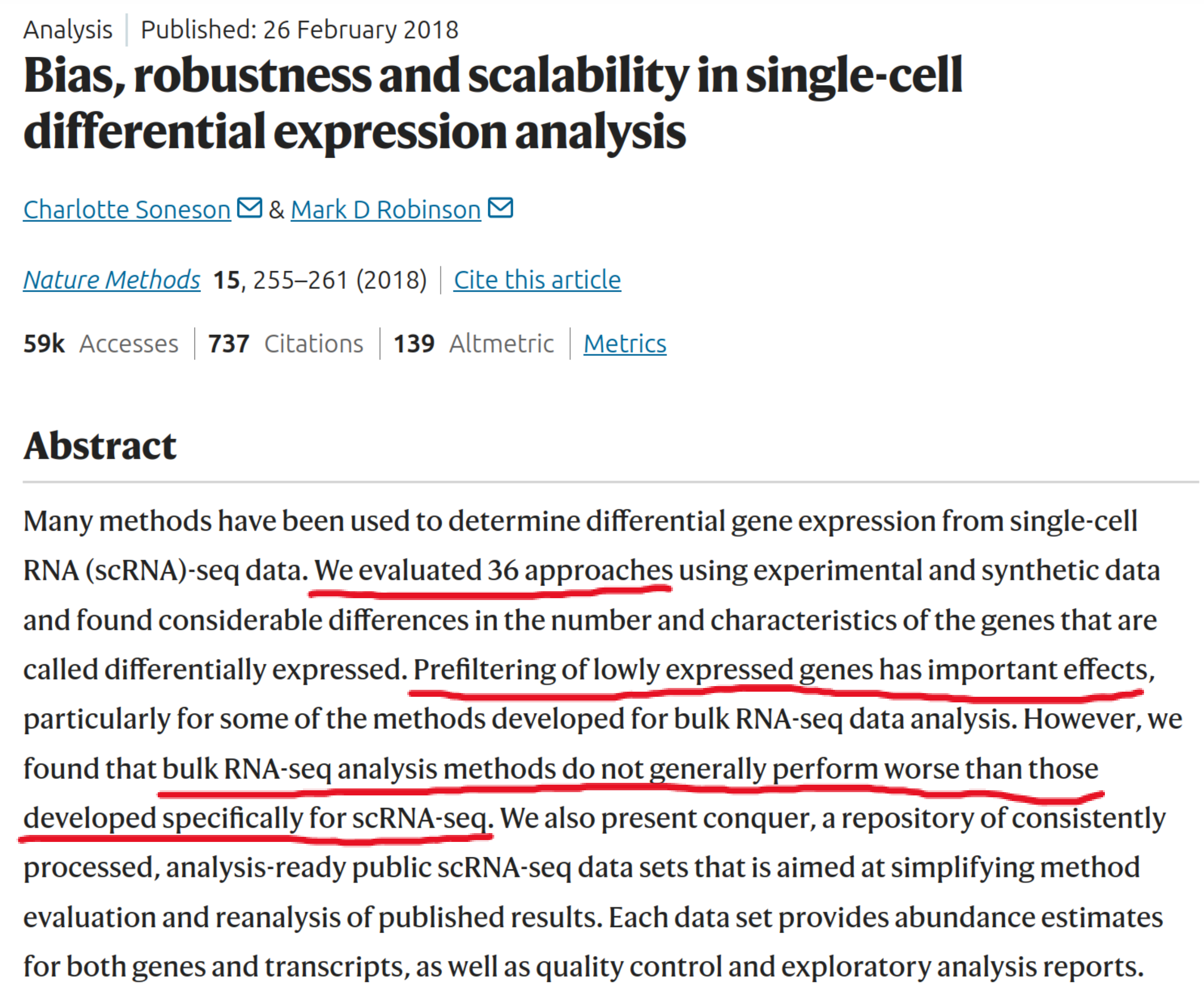
Seurat::FindAllMarkers()
Seurat::FindMarkers()

What is a marker?

Typical output



Gene Symbol	avg_log2FC	pct.1	pct.2	p_val	p_val_adj	cluster
CHI3L1	5.61	0.958	0.225	7.63E-255	1.97E-250	2
HLA-DRA	3.41	0.978	0.215	2.84E-253	7.32E-249	2
PTGFR	4.31	0.795	0.093	6.43E-244	1.66E-239	2
HLA-DRB5	3.54	0.818	0.097	2.10E-243	5.41E-239	2
GRIN2A	4.24	0.69	0.05	1.64E-235	4.22E-231	2
CDHR3	3.34	0.892	0.159	2.56E-229	6.59E-225	2
AKR1C3	3.16	0.955	0.254	5.75E-223	1.48E-218	2
KCNK15	3.76	0.897	0.187	4.88E-217	1.26E-212	2
HLA-DRB1	2.53	0.78	0.102	2.71E-212	6.98E-208	2
PLPP3	3.34	0.965	0.322	3.54E-210	9.11E-206	2



Analysis | Published: 26 February 2018

Bias, robustness and scalability in single-cell differential expression analysis

Charlotte Soneson & Mark D Robinson

Nature Methods **15**, 255–261 (2018) | Cite this article

59k Accesses | **737** Citations | **139** Altmetric | Metrics

Abstract

Many methods have been used to determine differential gene expression from single-cell RNA (scRNA)-seq data. We evaluated 36 approaches using experimental and synthetic data and found considerable differences in the number and characteristics of the genes that are called differentially expressed. Prefiltering of lowly expressed genes has important effects, particularly for some of the methods developed for bulk RNA-seq data analysis. However, we found that bulk RNA-seq analysis methods do not generally perform worse than those developed specifically for scRNA-seq. We also present conquer, a repository of consistently processed, analysis-ready public scRNA-seq data sets that is aimed at simplifying method evaluation and reanalysis of published results. Each data set provides abundance estimates for both genes and transcripts, as well as quality control and exploratory analysis reports.

https://www.nature.com/articles/nmeth.4612

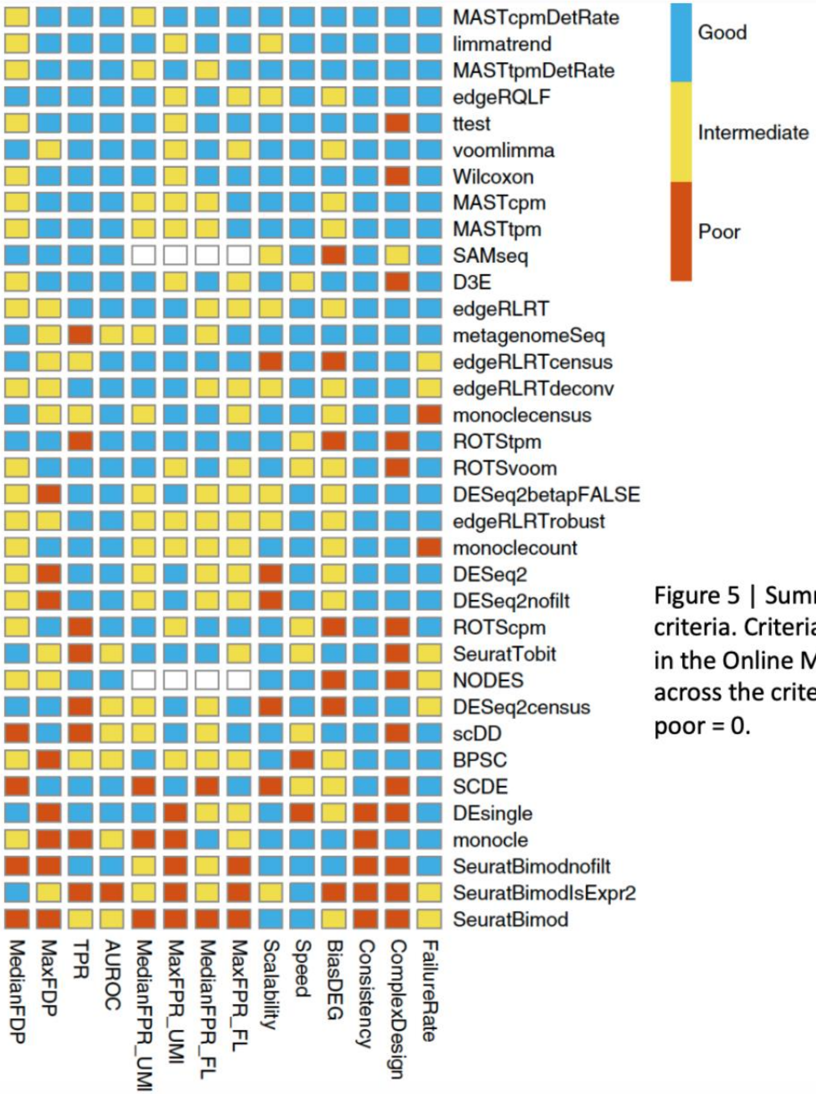


Figure 5 | Summary of DE method performance across all major evaluation criteria. Criteria and cutoff values for performance categories are available in the Online Methods. Methods are ranked by their average performance across the criteria, with the numerical encoding good = 2, intermediate = 1, poor = 0.

https://github.com/csoneson/conquer_comparison

Pseudo-bulk DEA

Use cases:

- Comparing **sample groups**, e.g. treated vs untreated, controls vs patients
- Complex designs

Methods:

- Aggregate gene expression counts within groups (e.g., by cell type and sample) and use bulk RNA-seq tools like DESeq2, edgeR, or limma
- Tools like Muscat in Bioconductor are specifically designed for pseudo-bulk DGE analysis in scRNA-seq

Advantages:

- Reduces noise and dropout effects
- Leverages well-validated bulk RNA-seq tools

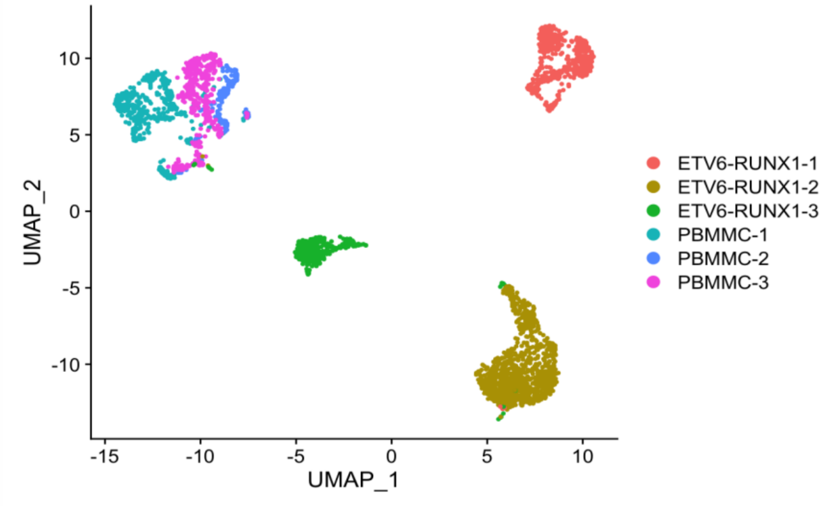
Limitations:

- Loses single-cell resolution and cannot detect cell-to-cell variability

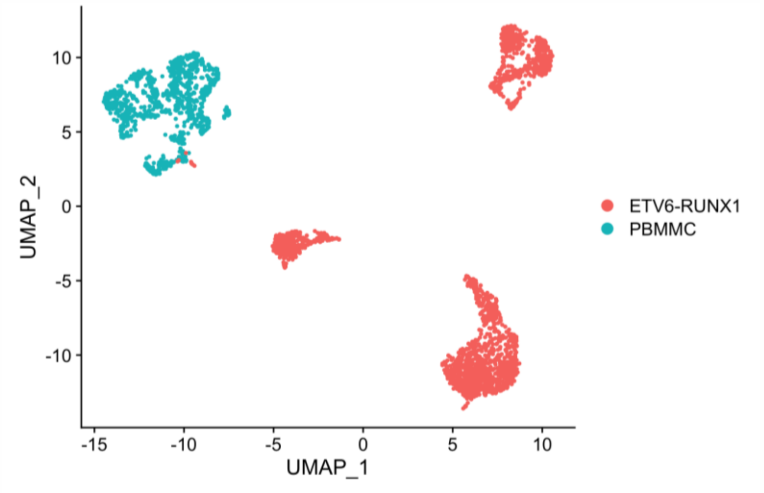
Pseudo-bulk DEA

pro B cells subset of PBMMCs
from multiple samples

By sample



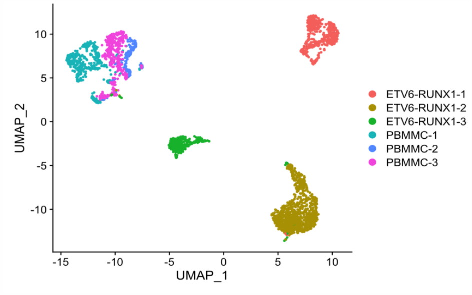
By condition

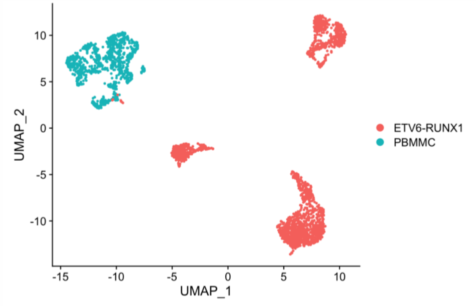


Identify DE genes between healthy and leukemic pro B cells

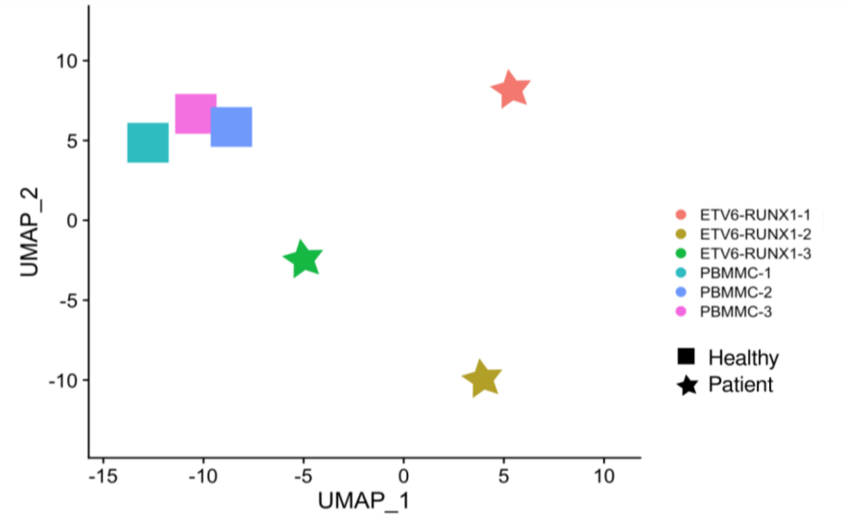
Pseudo-bulk DEA

pro B cells pseudobulk

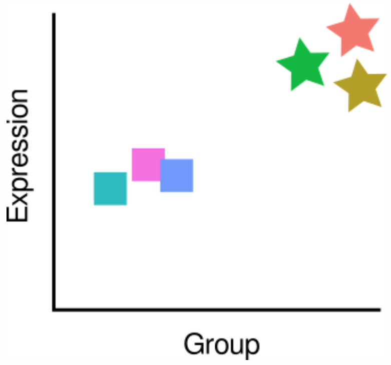




By sample & condition

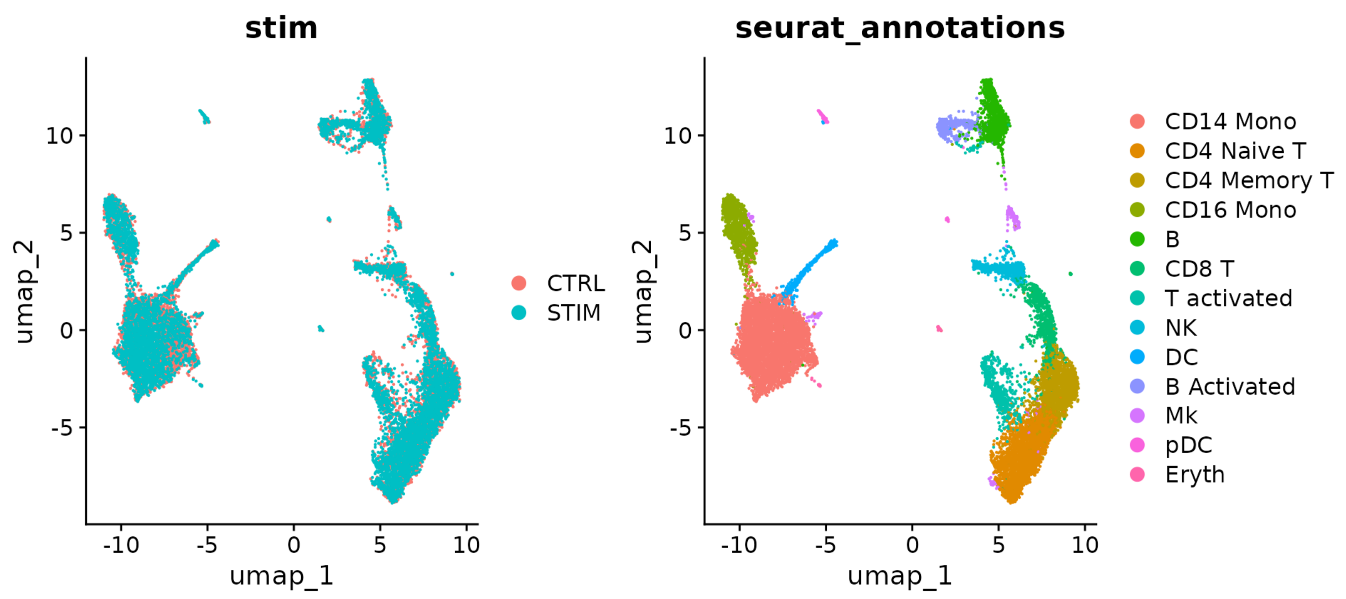


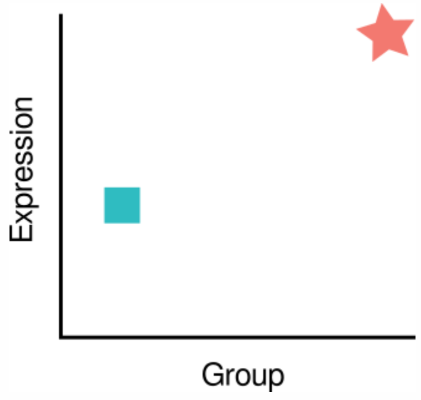
Gene XY



Beware of pseudo-replication

interferon-stimulated and control PBMCs, one sample per condition





Thanks for your attention!