Start Setting up for the Workshop

Workshop webpage: https://go.tufts.edu/discworkshop

Log on with Tufts Credentials to On Demand on Tufts Cluster https://ondemand.pax.tufts.edu/

Click on `Interactive Apps > RStudio Pax` and you will see a form to fill out to request compute resources to use RStudio on the Tufts HPC cluster. We will fill out the form with the following entries:

- `Number of hours` : `5`
- `Number of cores` : `1`
- `Amount of memory` : `16GB`
- `R version` : `4.0.0`
- `Reservation for class, training, workshop` : `Bioinformatics Workshops`

Single-cell RNA-Seq Bioinformatics

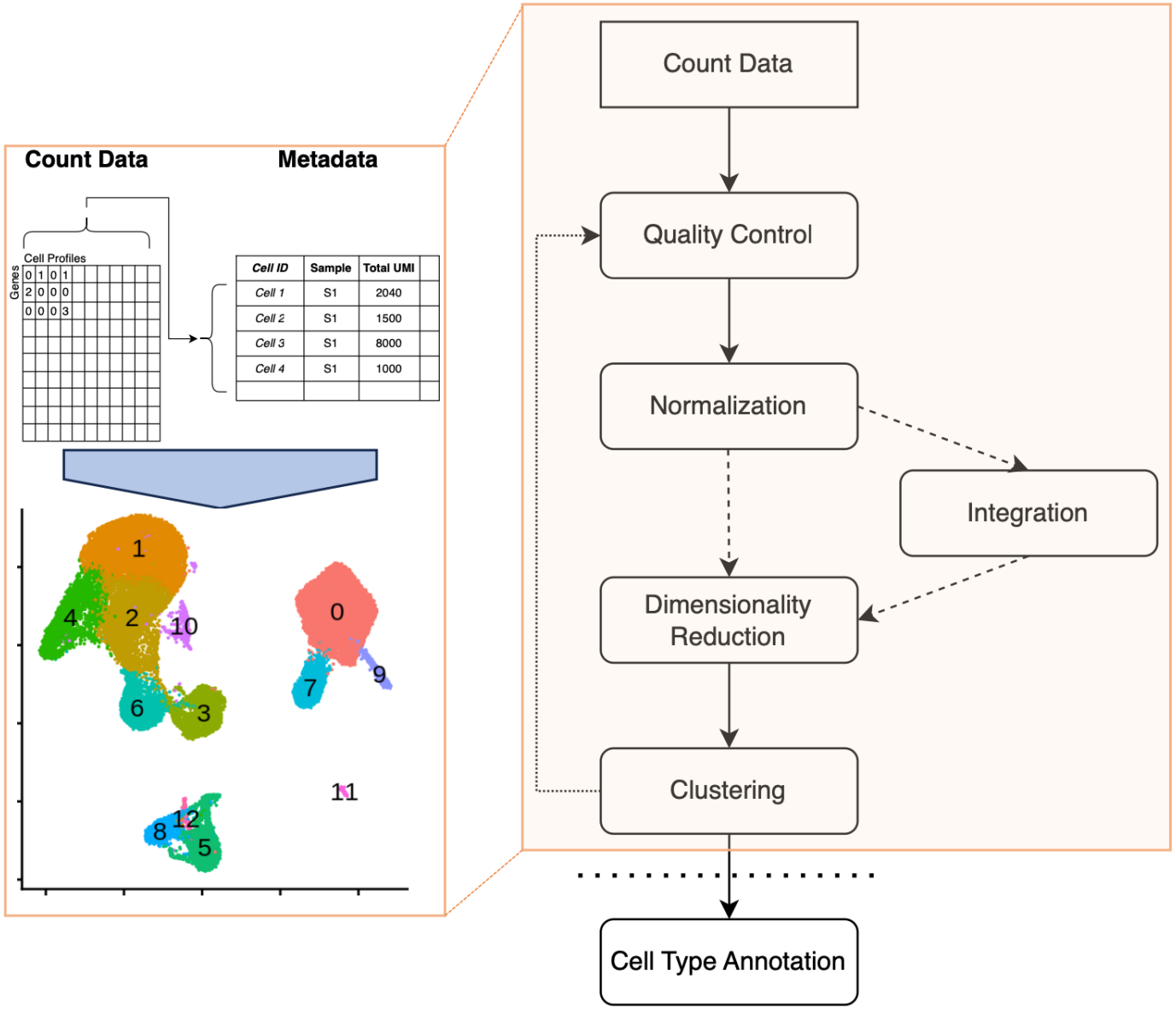
Day 2: Cell-type classification and Differential Expression

Data Intensive Studies Center (DISC) Single-cell series Oct-Nov 2023

Rebecca Batorsky, Data Scientist, Rebecca.Batorsky@tufts.edu

Eric Reed, Data Scientist, Eric.Reed@tufts.edu

Workshop part I recap



Case Study

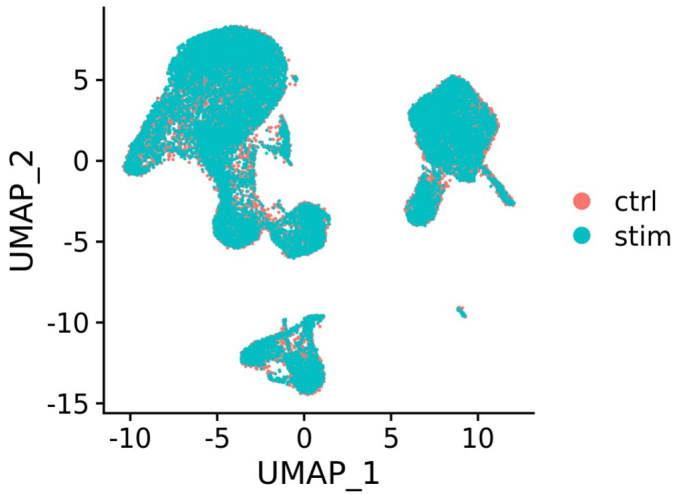
scRNAseq Data

Peripheral Blood Mononuclear Cells (PBMCs)

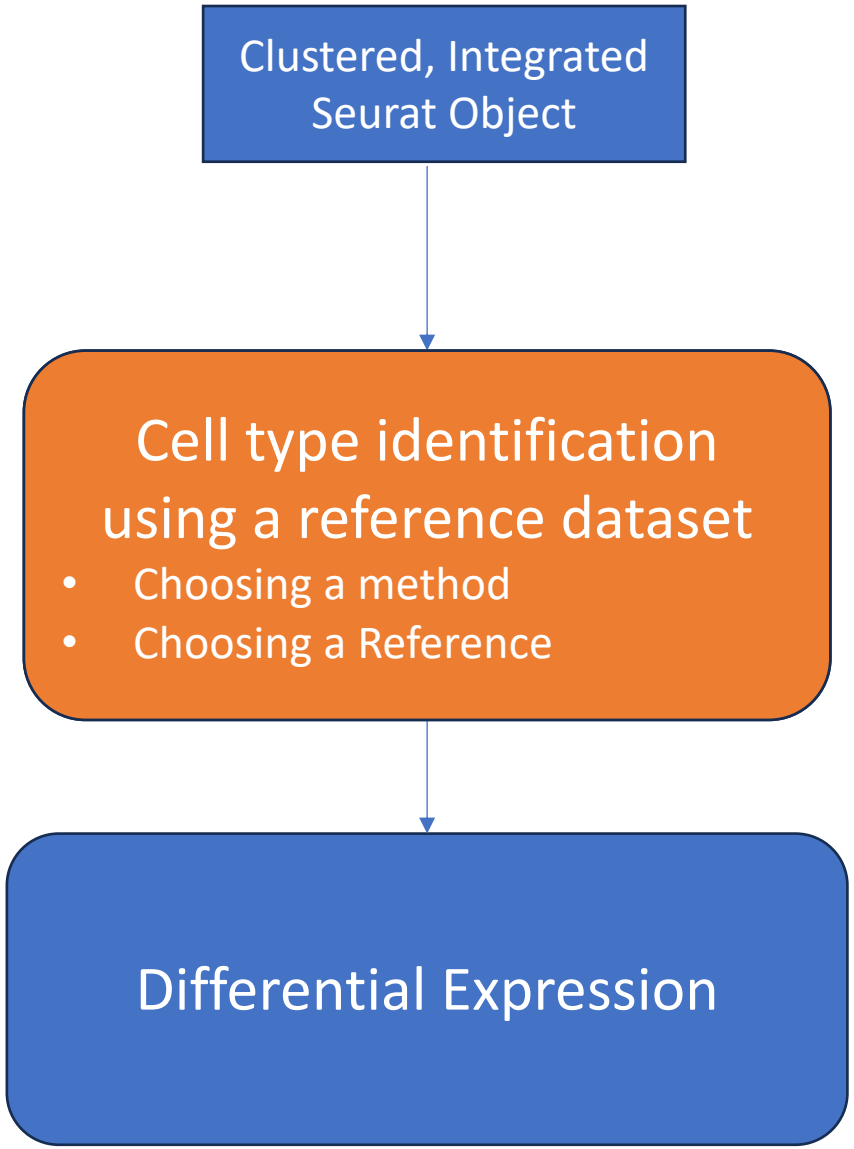
Includes:
- B cells
- T cells
- Natural killer cells
- Monocytes
- Macrophages

Doesn't include:
- Neutrophils
- Platelets
- Red blood cells

Two samples
- Control: "ctrl"
- Interferon beta-treated (stimulated): "stim"



Workshop day 2



Our Seurat Object

@metadata

	Sample	nCount_RNA	nFeature_RNA	…	integrated_snn_res.0.4
Cell-1	ctrl	2344	874		1
Cell-2	ctrl	3125	896		2
Cell-3	stim	2578	725		3

@reductions

…

@assays

RNA

counts

	Cell-1	Cell-2	Cell-3
gene1	0	3	0
gene2	1	1	0

data
- Norm
- log1p

	Cell-1	Cell-2	Cell-3
gene1	0	2.02	0
gene2	2.42	1.15	0

SCT

counts
- SCT norm for sequencing depth

	Cell-1	Cell-2	Cell-3
gene1	0	2	0
gene2	1	0	0

data
- log1p

	Cell-1	Cell-2	Cell-3
gene1	0	1.09	0
gene2	0.69	0	0

- Scale.data
- SCTModel

Integrated

data
- integrated

	Cell-1	Cell-2	Cell-3
gene1	-0.55	6.29	0.35
gene2	1.7	0.01	-0.36

- Scale.data
- Var.features

Adding Cell-type labels

Query dataset

query@metadata

	Sample	nCount_RNA	nFeature_RNA	…	integrated_snn_res.0.4	Cell-type
Cell-1	ctrl	2344	874		1	?
Cell-2	ctrl	3125	896		2	?
Cell-3	stim	2578	725		3	?

query@assays$RNA@data

	Cell-1	Cell-2	Cell-3
gene1	0	2.02	0
gene2	2.42	1.15	0

Reference dataset

Reference Metadata

	…	Cell-type
ref-cell-1		T-cell
ref-cell-2		B-cell
ref-cell-3		Macrophage

Normalized data

	ref-cell-1	ref-cell-2	ref-cell-3
gene1	0	5.5	0
gene2	5.3	20.1	0

SingleR Correlation Method

Using a reference database of cell-type gene expression, calculate correlation of gene expression for each cell

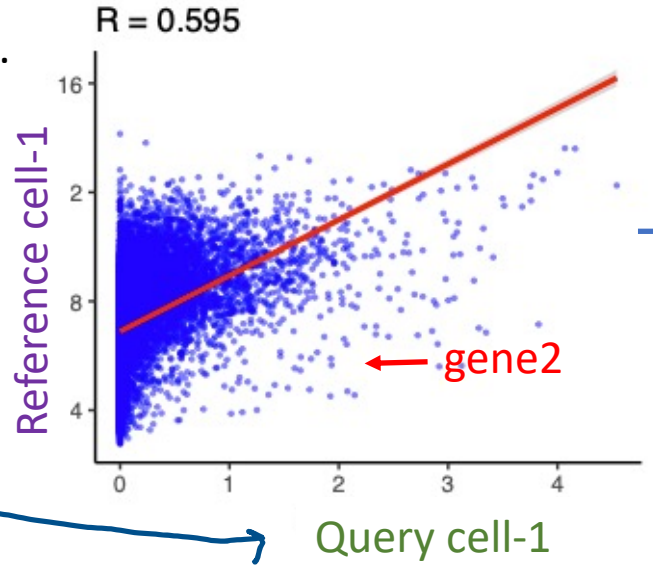
Reference

	ref-cell-1	ref-cell-2	ref-cell-3	…
gene1	0	5.5	0	
gene2	5.3	20.1	0	
…				

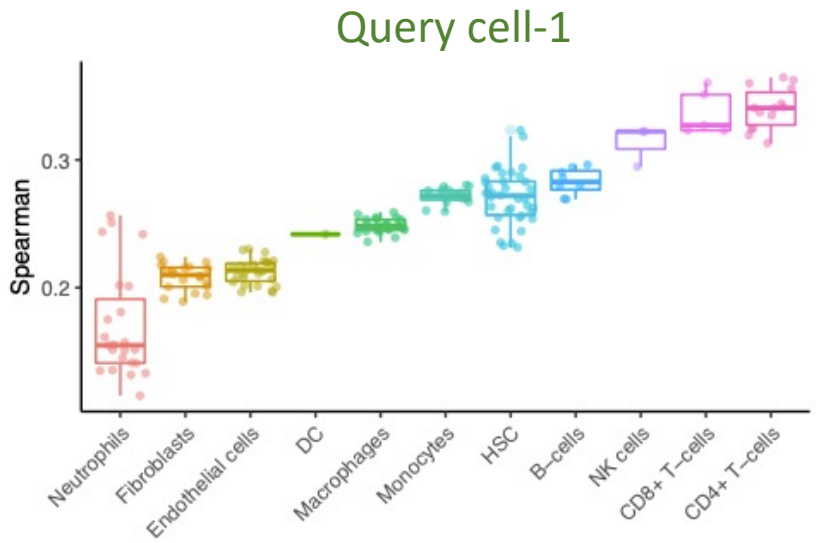
Query

	Cell-1	Cell-2	Cell-3	…
gene1	0	2.02	0	
gene2	2.42	1.15	0	
…				

Step 1: Spearman Correlation



Step 2: Aggregation



Step3: Fine tuning

Step 4: Labeling

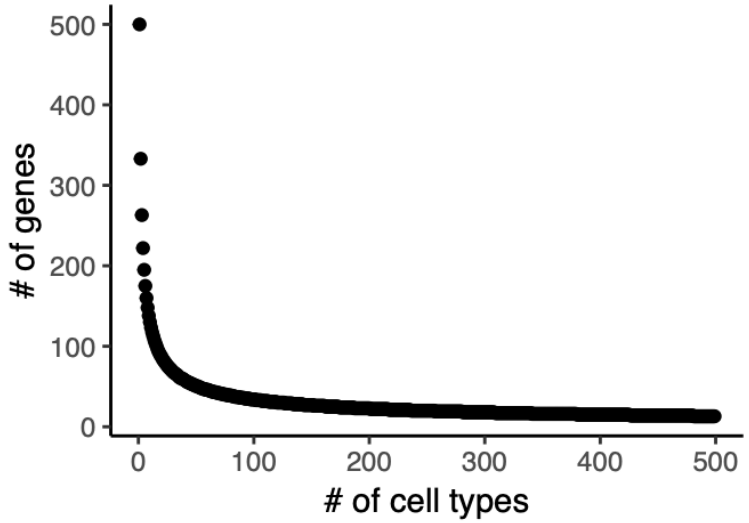
	Sample	…	Cell-type
Cell-1	ctrl		CD4+T-

Aran et al “Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage.” *Nat. Immunol.* 20 (2): 163–72.

SingleR method

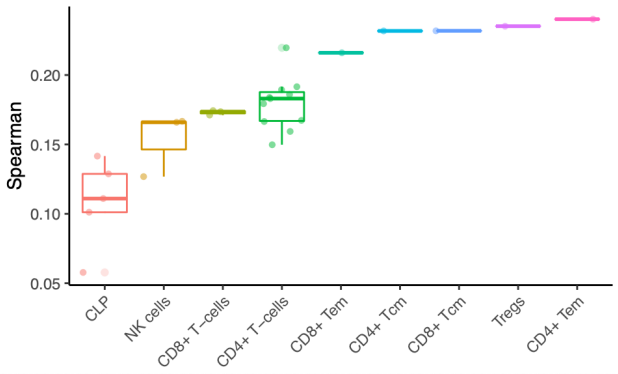
Step 3: Fine tuning

- Assignment is performed in multiple iterations.
- At each iterations, only the top N variable genes between cell-types are considered
- After each iteration, only best scoring cell types are kept



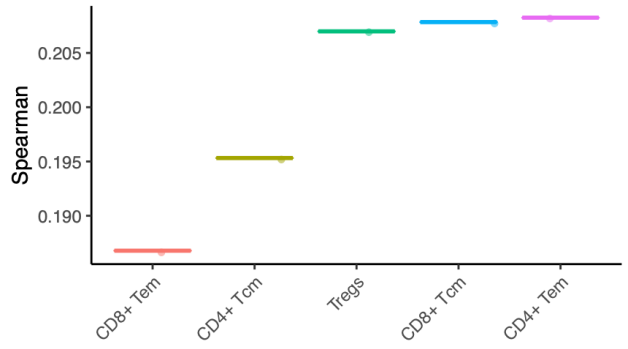

```
## [1] "Number of DE genes:1819"
```

Query cell-1



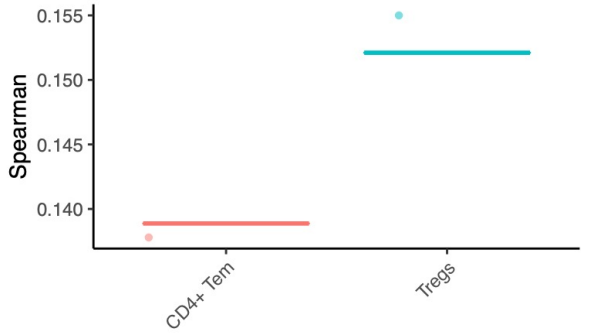

```
## [1] "Number of DE genes:1287"
```

Query cell-1




```
## [1] "Number of DE genes:666"
```

Query cell-1



Aran et al “Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage.” *Nat. Immunol.* 20 (2): 163–72.

SingleR with celldex database of pure cell types

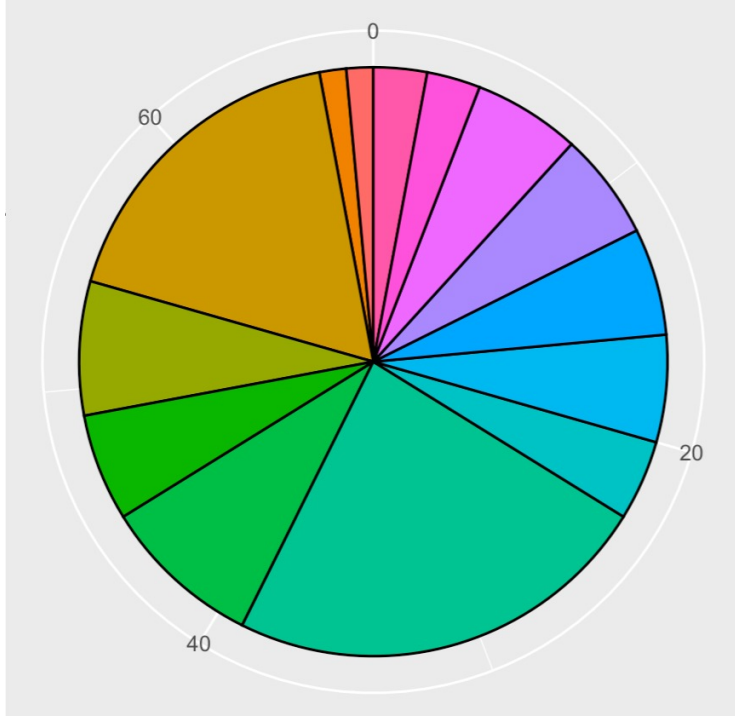
Celldex R library is a general purpose reference, giving convenient access to expression profiles for many cell types

Some examples	Description
Human primary cell atlas (HPCA)	Microarray datasets derived from human primary cells
Blueprint/ENCODE	Bulk RNA-seq data for pure stroma and immune cells
Immunological Genome Project (ImmGen)	Microarray profiles of pure mouse immune cells

Labels are given at two resolutions, e.g. for HPCA **label.main. = T cells**

label.fine =

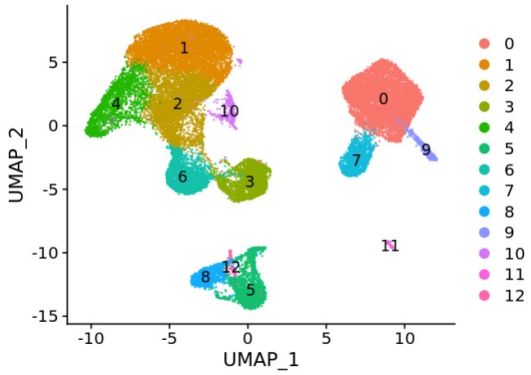
- T_cell:CCR10-CLA+1,25(OH)2_vit_D3/IL-12
- T_cell:CCR10+CLA+1,25(OH)2_vit_D3/IL-12
- T_cell:CD4+
- T_cell:CD4+_central_memory
- T_cell:CD4+_effector_memory
- T_cell:CD4+_Naive
- T_cell:CD8+
- T_cell:CD8+_Central_memory
- T_cell:CD8+_effector_memory
- T_cell:CD8+_effector_memory_RA
- T_cell:CD8+_naive
- T_cell:effector
- T_cell:gamma-delta
- T_cell:Treg:Naive



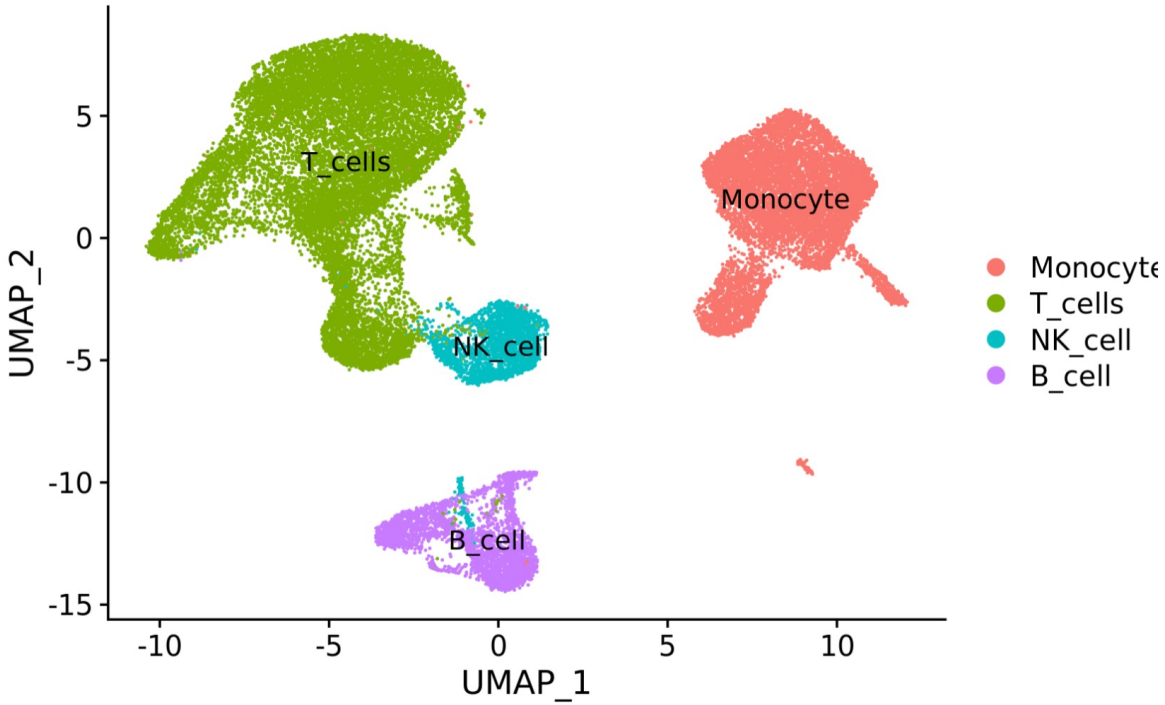
https://bioconductor.org/packages/devel/data/experiment/vignettes/celldex/inst/doc/userguide.html

SingleR Results

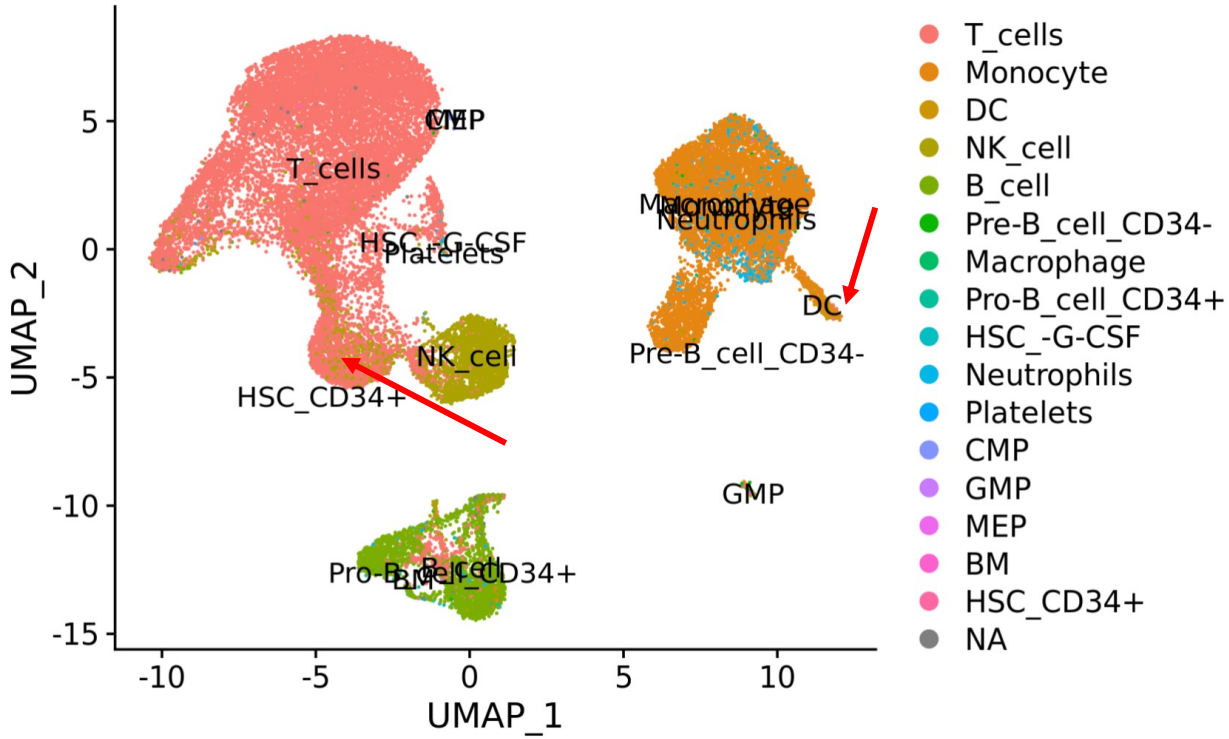
Query



Cluster level, HPCA, label.main



Cell level, HPCA, label.main



Seurat Integration mapping method

10X GENOMICS Products Research

Support › Single Cell Gene Expression › Datasets

3k PBMCs from a Healthy Donor

Single Cell Gene Expression Dataset by Cell Ranger 1.1.0

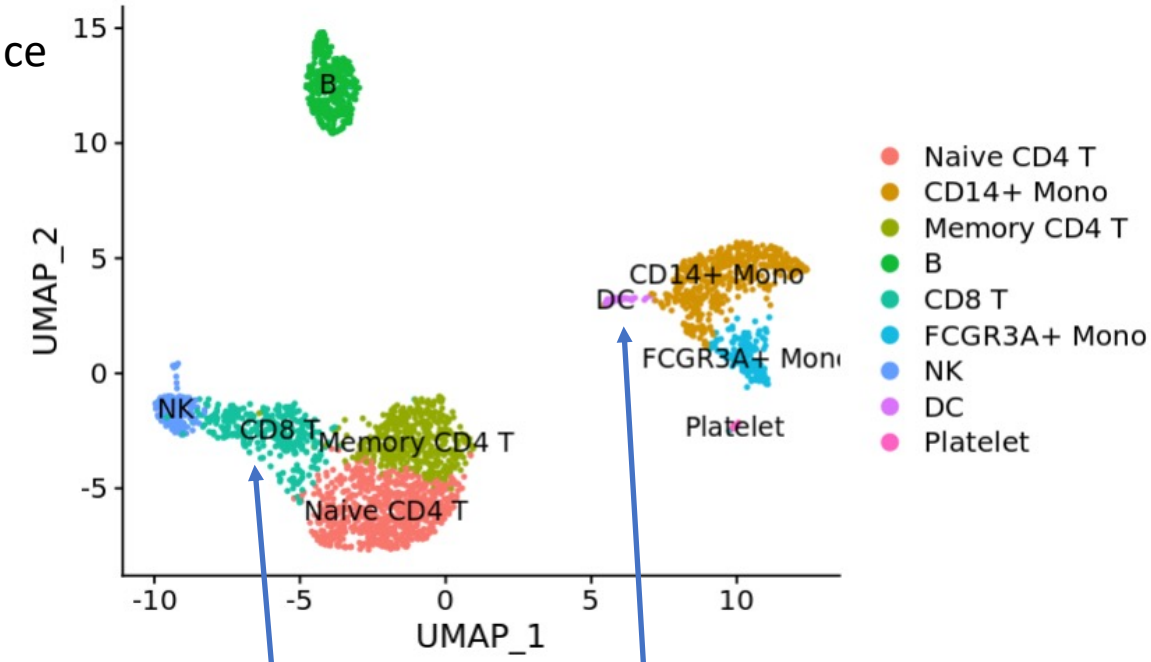
Peripheral blood mononuclear cells (PBMCs) from a healthy donor (same donor as pbmc6k).

PBMCs are primary cells with relatively small amounts of RNA (~1pg RNA/cell).

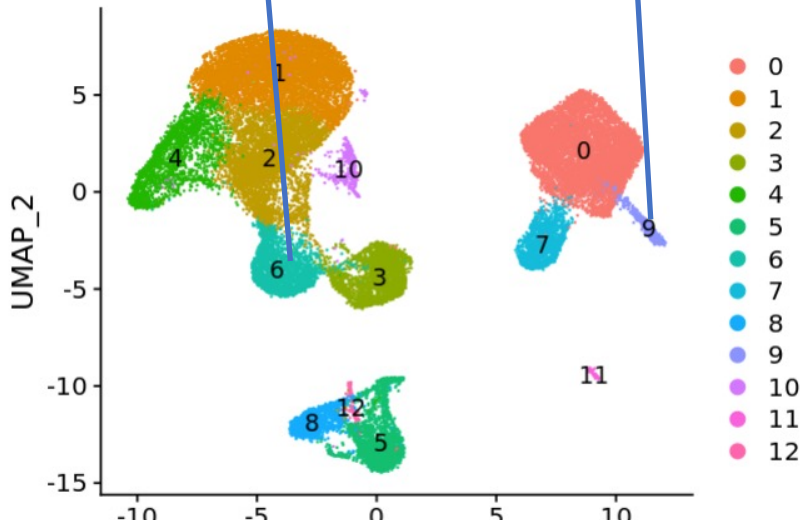
- 2,700 cells detected
- Sequenced on Illumina NextSeq 500 with ~69,000 reads per cell
- 98bp read1 (transcript), 8bp I5 sample barcode, 14bp I7 GemCode barcode and 10bp read2 (UMI)
- Analysis run with --cells=3000

Published on May 26, 2016

Reference

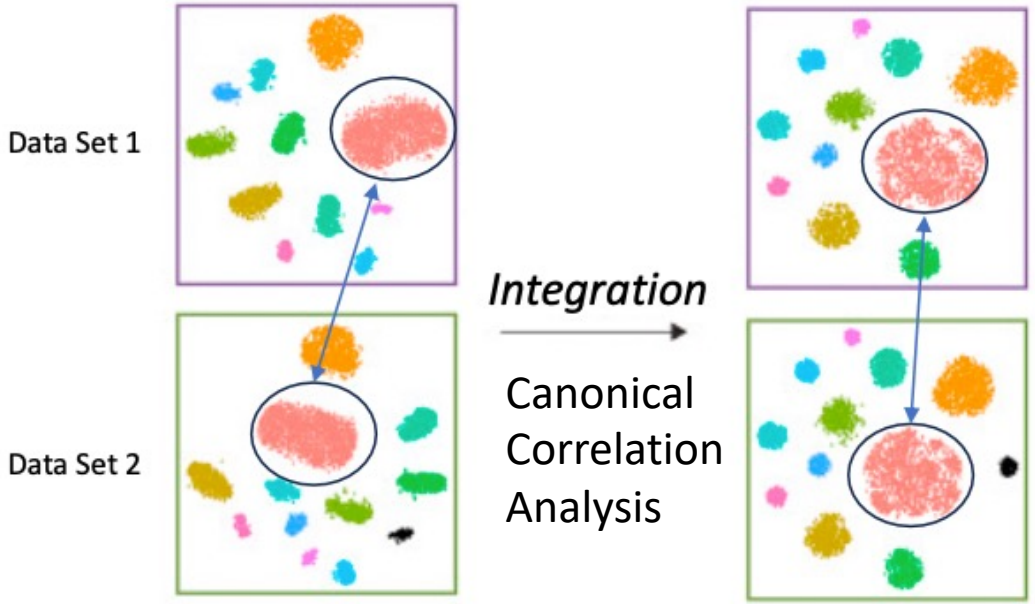


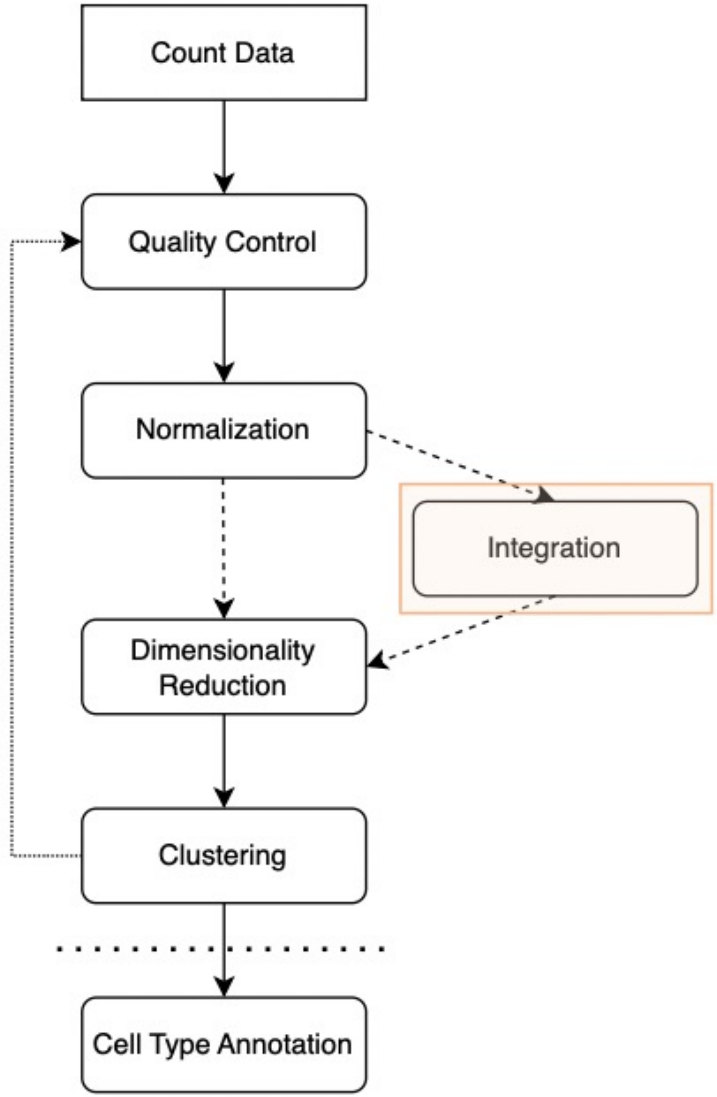
Query



Integration (recap from workshop 1)

Transform our data such that the profiles of the same cell types are more similar across data sets

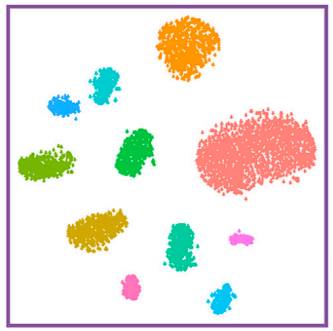




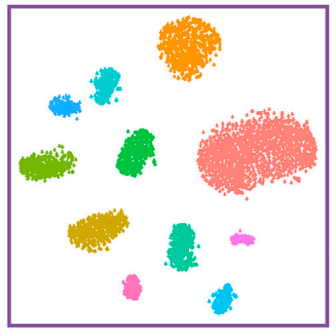
Seurat Integration mapping method

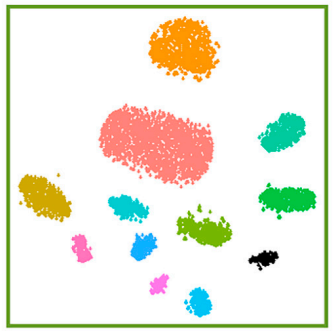
Reference

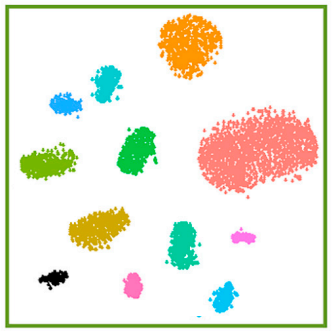
Query



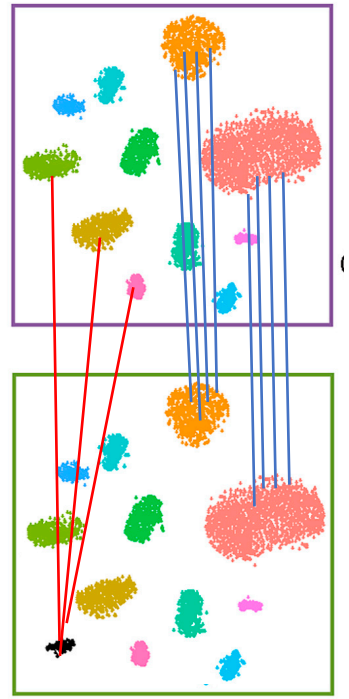
Project Query using reference PCA structure







Identify Mutual Nearest Neighbors (“Anchors”)

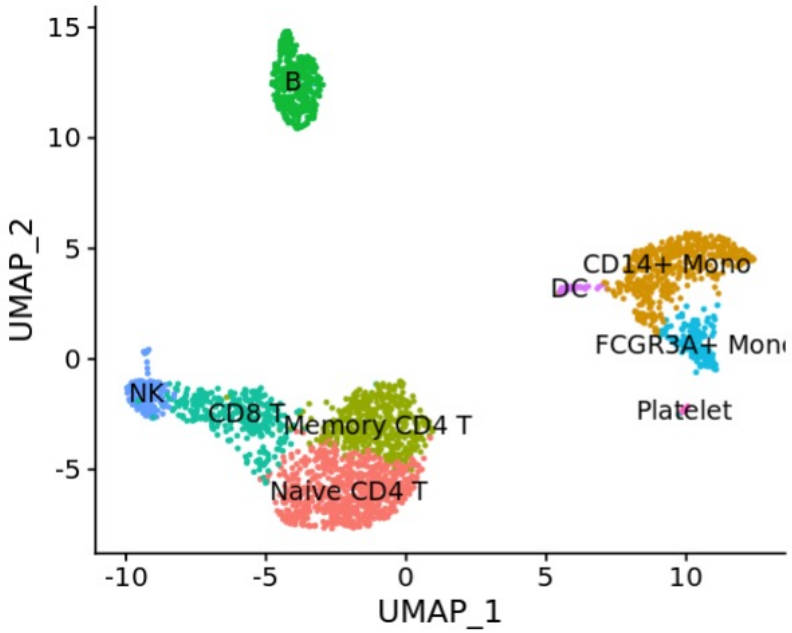


High scoring correspondence => Transfer labels

Low scoring correspondence => Unique population

Where is the reference PCA?

Reference PBMC



@metadata

	...	Cell-type
ref-cell-1		T-cell
ref-cell-2		B-cell
ref-cell-3		Macrophage

@assays

RNA, SCT

	ref-cell-1	ref-cell-2	ref-cell-3
gene1	0	1.09	0
gene2	5.04	0	0

@reductions

PCA

Feature loadings

	PC_1	PC_2	PC_3	...
PPBP	-1.172699e-02	1.523301e-02	-1.475722e-01	
S100A9	-1.186967e-01	2.094446e-02	-2.730169e-02	
IGLL5	8.678533e-03	5.451734e-02	5.480080e-02	
LYZ	-1.195951e-01	1.637529e-02	-1.491376e-02	
...				

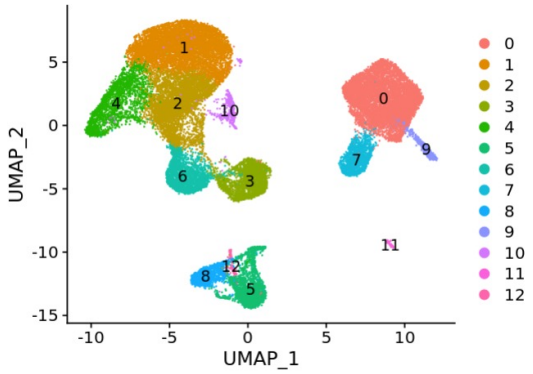
Cell embeddings

	PC_1	PC_2	PC_3	...
AAACATACAACCAC-1	4.60604661	-0.603719506	-0.605242902	
AAACATTGAGCTAC-1	0.16708090	4.544217123	6.451886683	
...				

UMAP

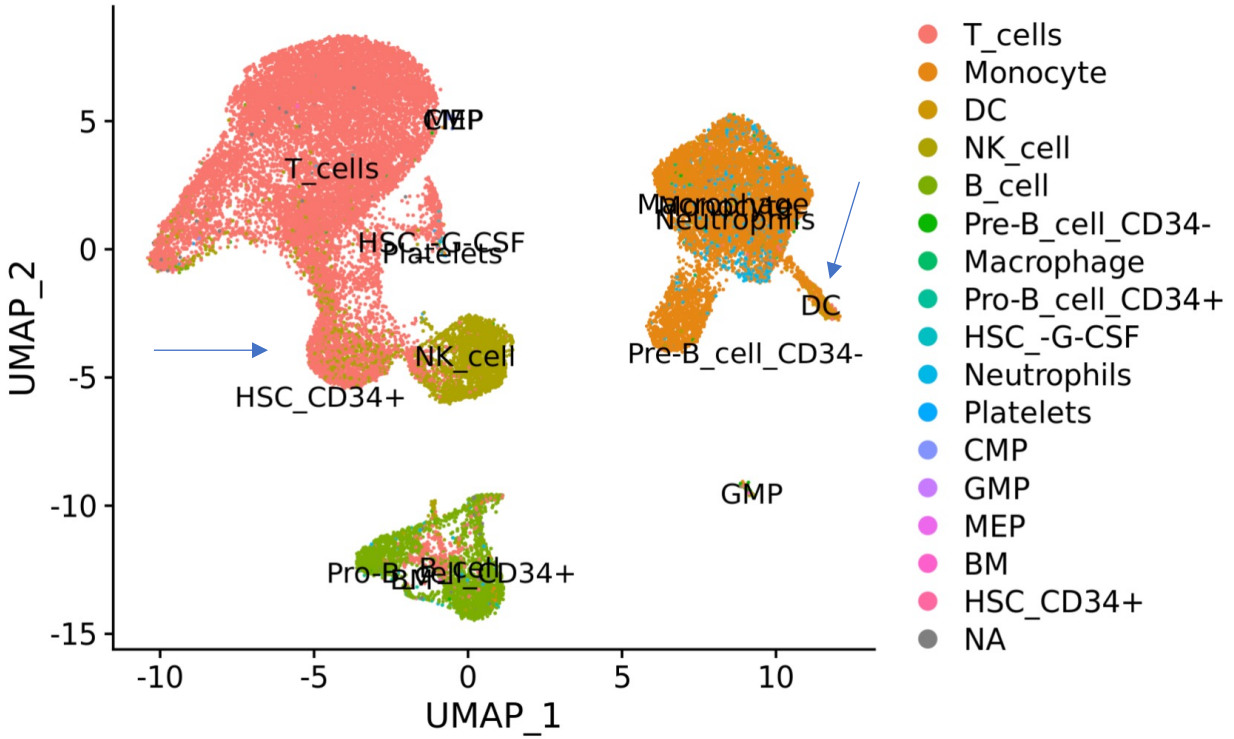
	UMAP_1	UMAP_2
AAACATACAACCAC-1	-3.198993068	-4.242983556
AAACATTGAGCTAC-1	-4.163467985	10.733455443
...		

Query



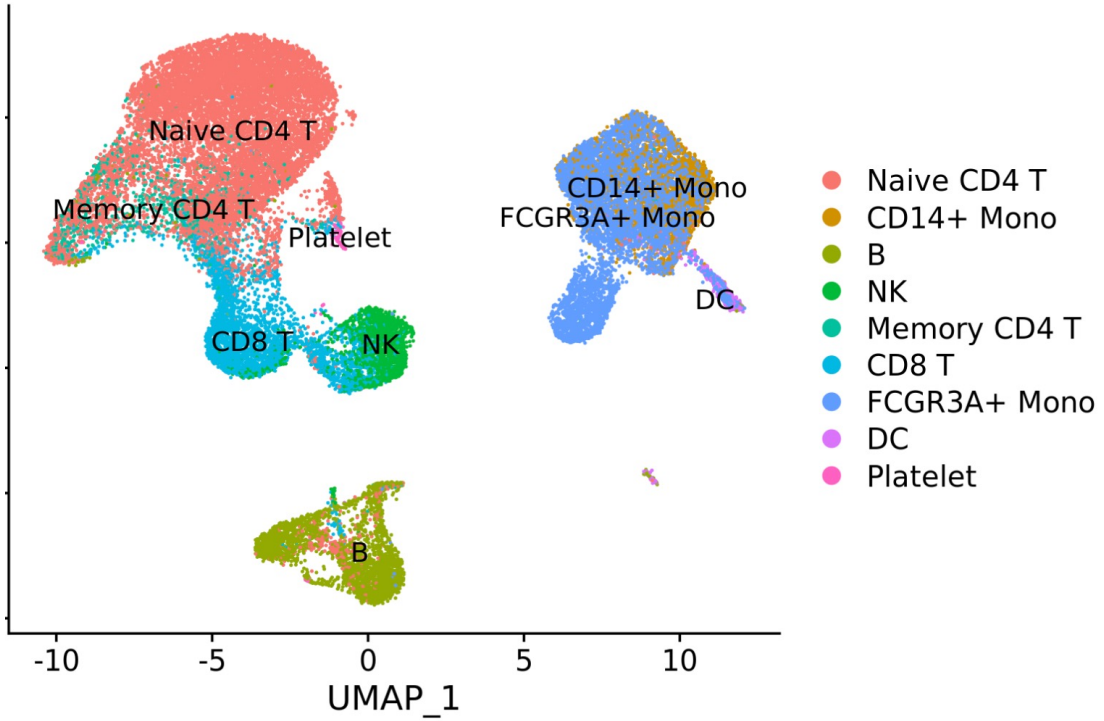
SingleR Results

Cell level, HPCA, label.main



Seurat Results

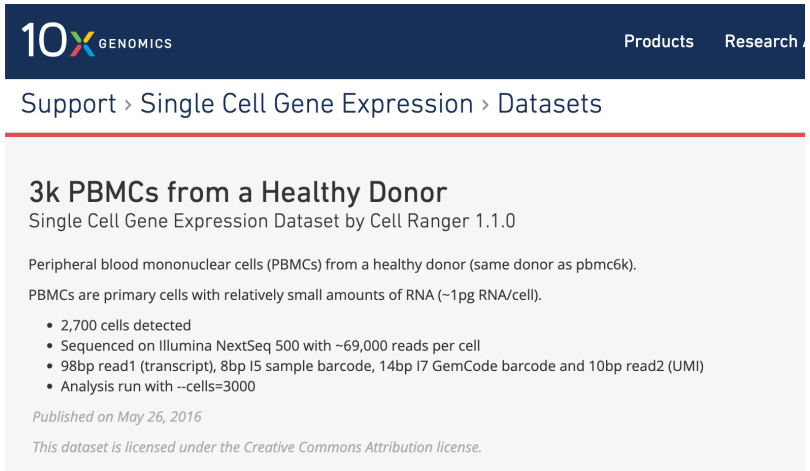
Human PBMC



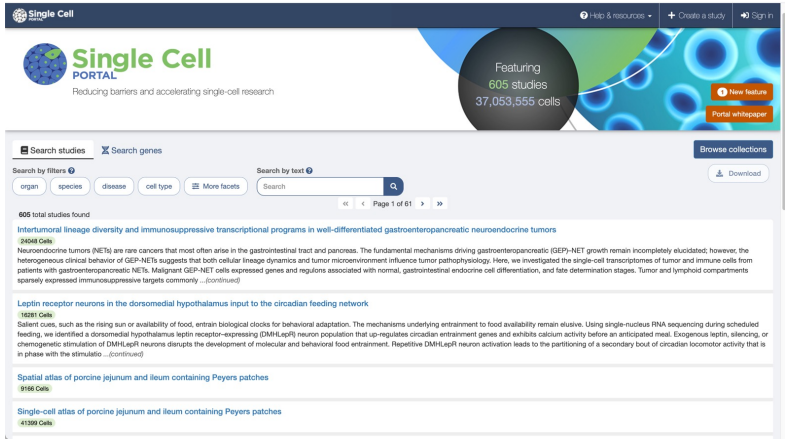
Using single-cell datasets as a reference

Typically contains cells present in a single organism, tissue, condition and should be chosen to match your experiment as closely as possible.

10x Genomics reference database



Broad Single Cell Portal



Literature search

https://tabula-sapiens-portal.ds.czbiohub.org/

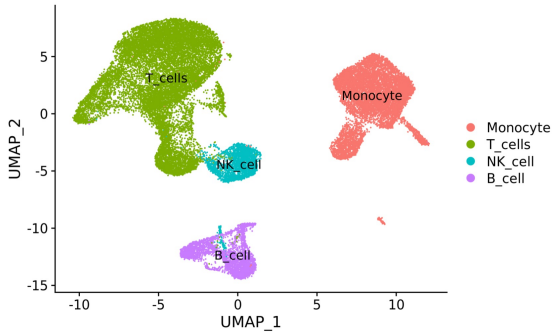
https://chanzuckerberg.github.io/cellxgene-census/

Pro: can match well

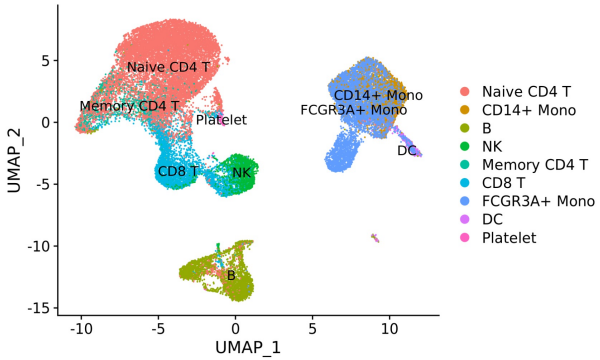
Con: may not be available, and takes a lot of pre-processing

Example workflow

1. Rough cell type mapping using label.main in Celldex.



2. Fine cell type mapping using a well-matched single-cell dataset.

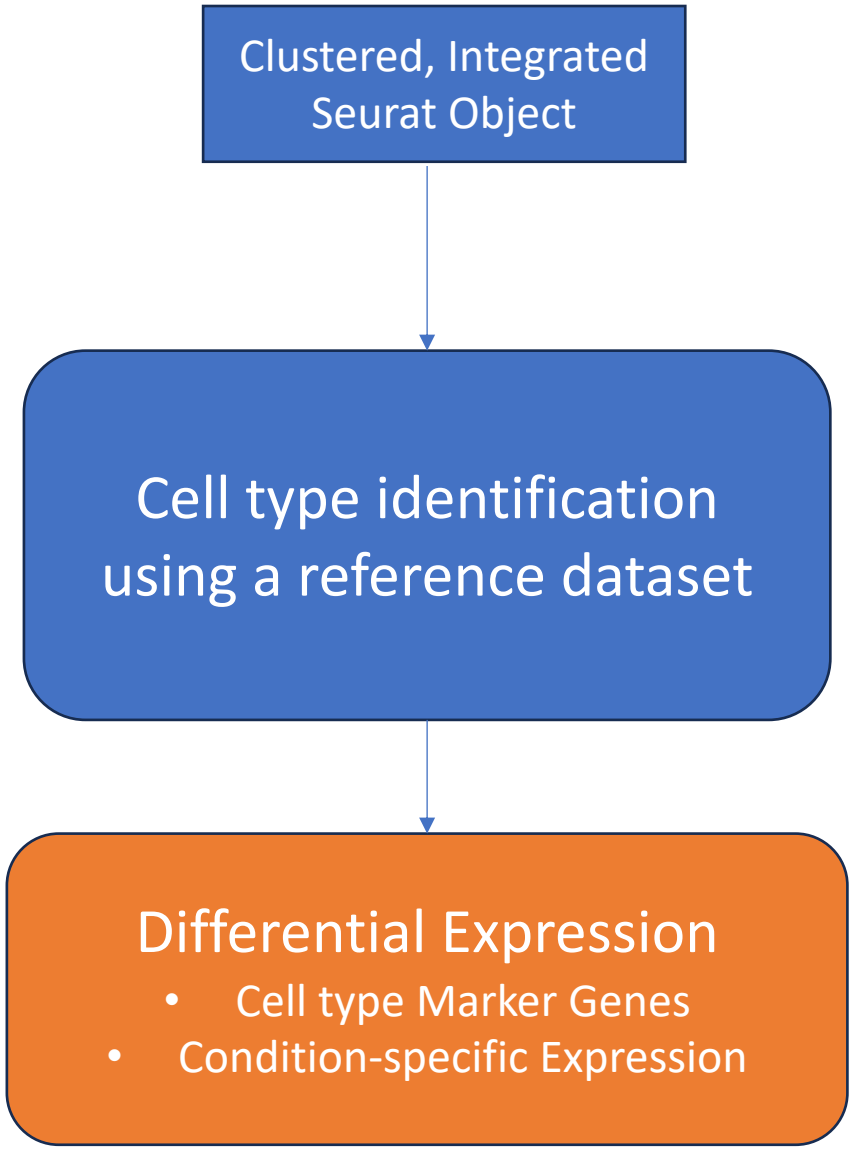


3. Known and calculated marker genes (next section)

What if there are no good references for your cell type?

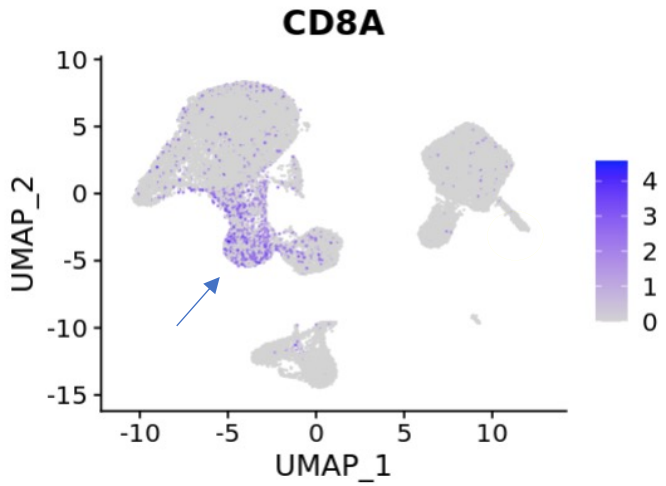
Questions?

Workshop day 2



Differential Expression

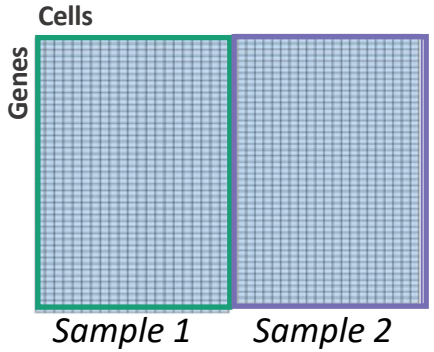
1) Genes that are overexpressed in on cell-type compared to all other cell-types



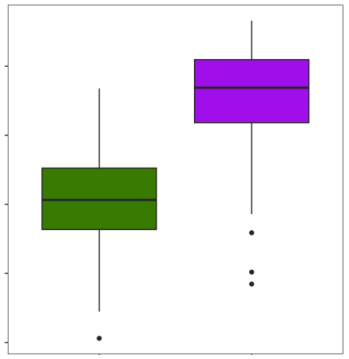
"Marker Genes"
- can be computed within a single sample

2) Gene that are statistically differently expressed between phenotypes or conditions, usually within a cell-type

Control PBMC Stimulated PBMC



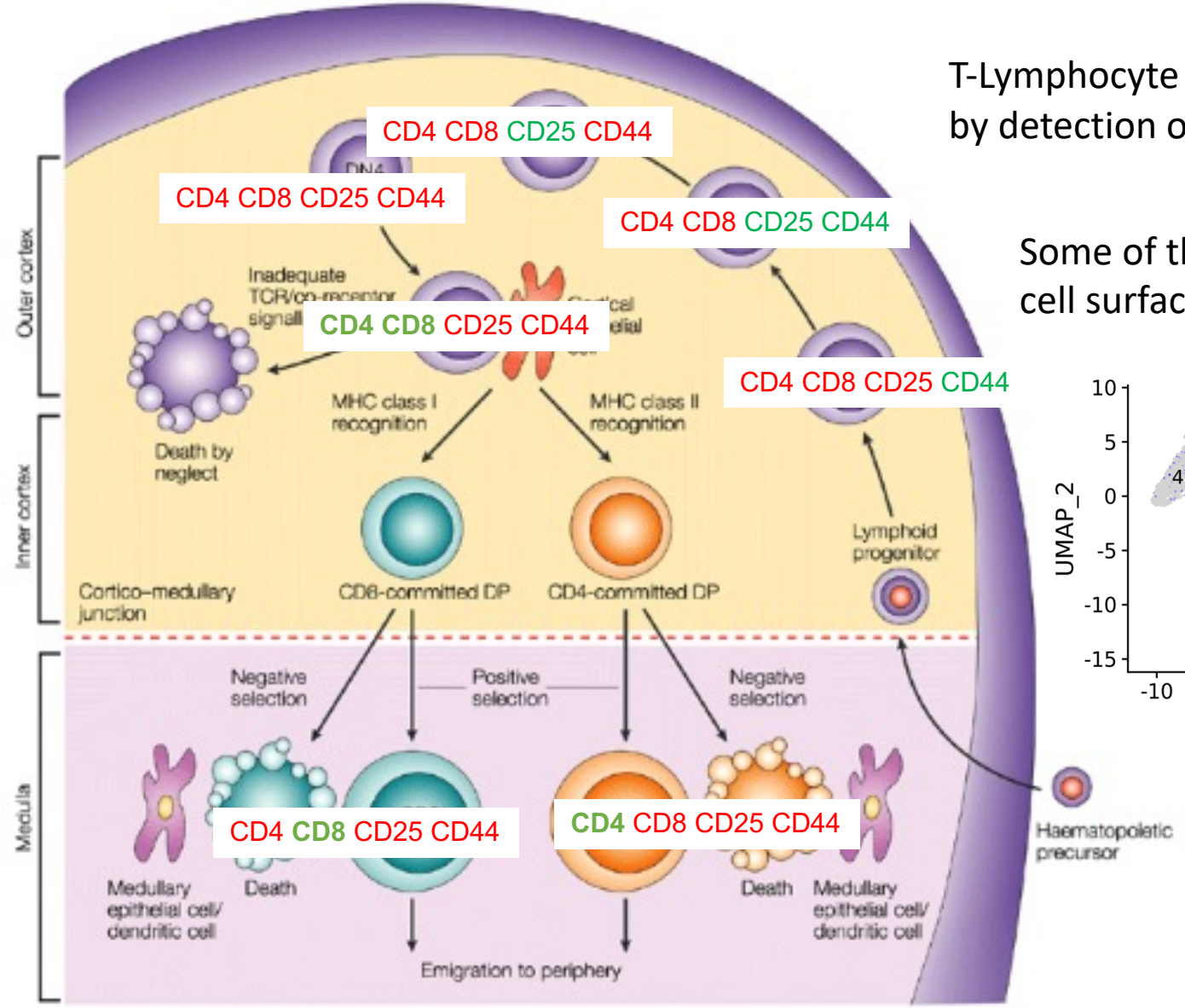
Gene1 Expression



Control Stimulated

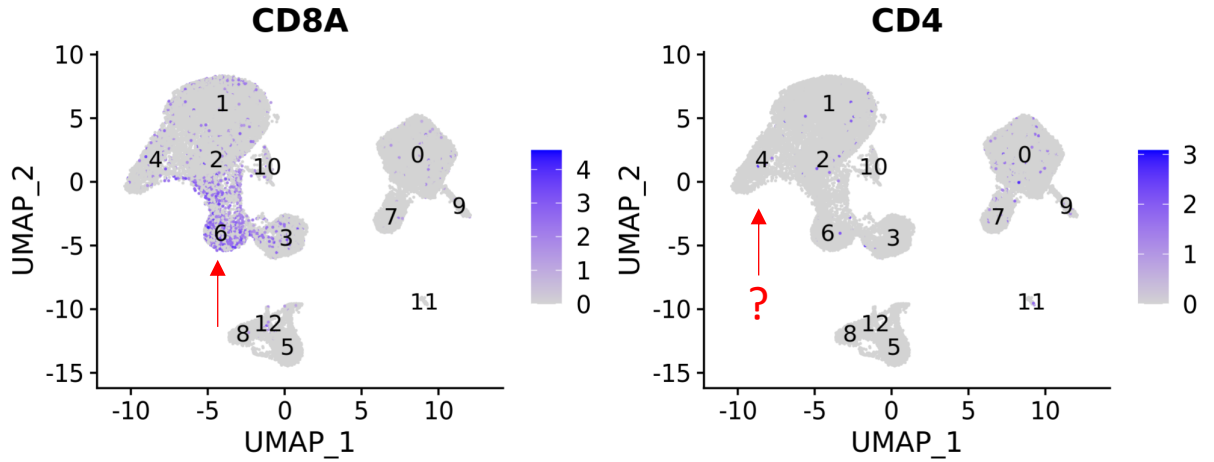
"Differentially Expressed Genes"
- typically requires biological replicates and modeling of inter-replicate variability

Cell-type Marker Genes



T-Lymphocyte development in the Thymus was historically defined by detection of 4 cell surface proteins CD4 CD8 CD25 CD44.

Some of these markers must be redefined for scRNAseq because cell surface protein quantity != gene transcript quantity

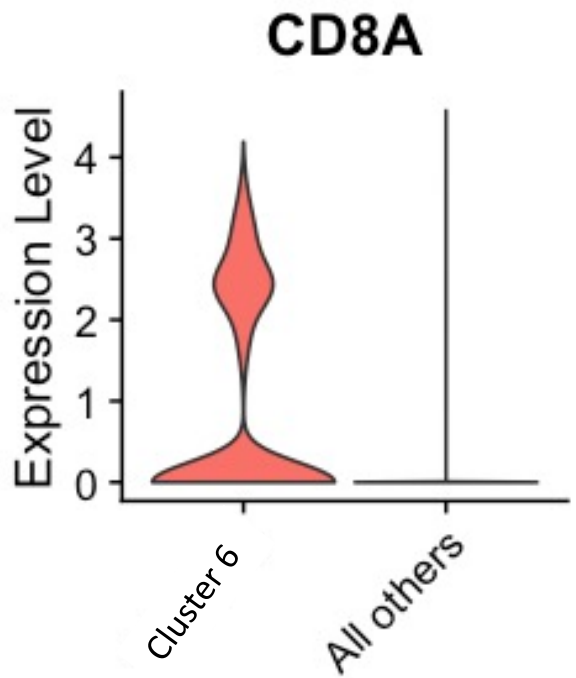


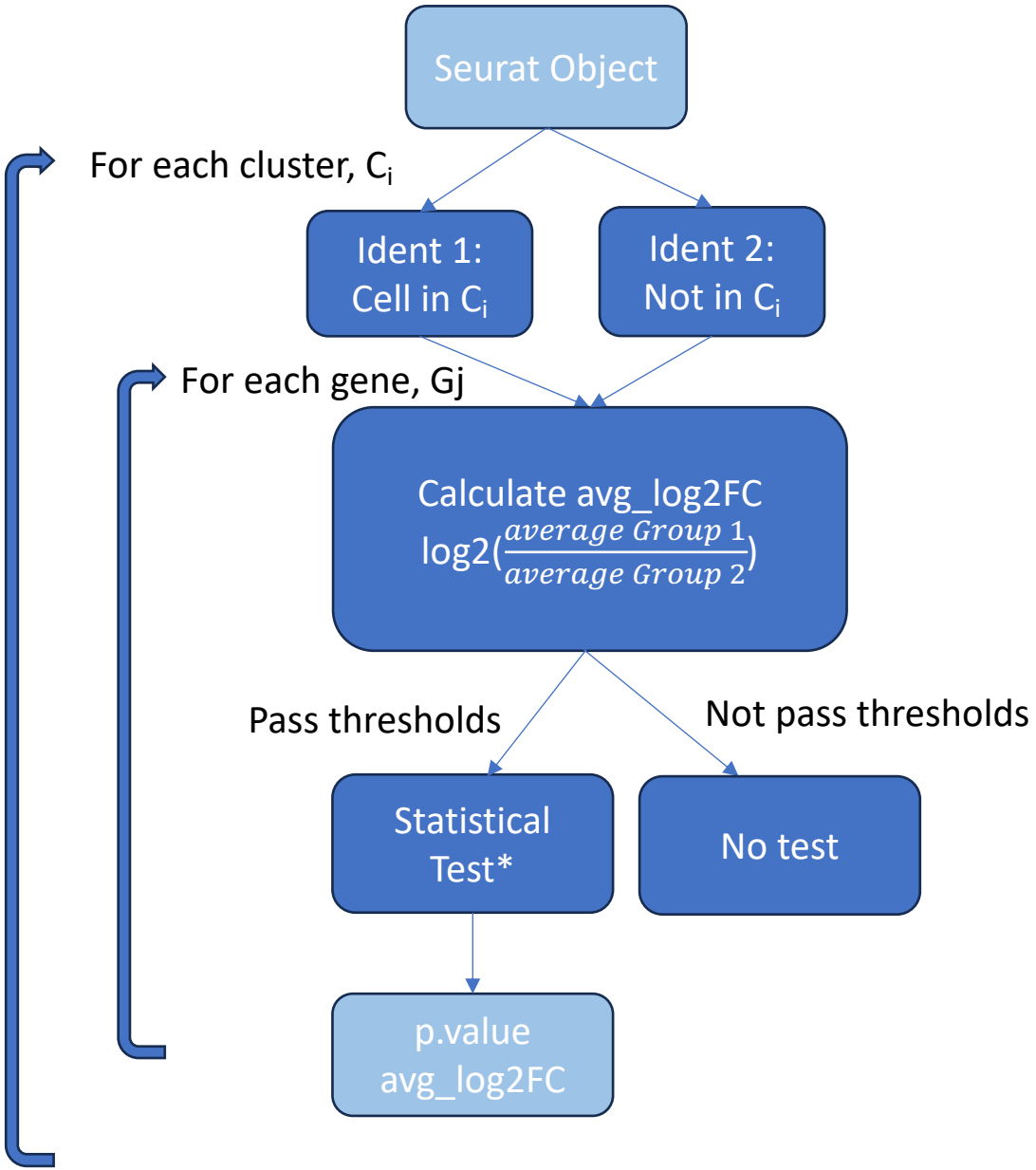
https://www.labome.com/method/T-Cell-Markers-and-B-Cell-Markers.html

Find All Markers

Genes that are more highly expressed by cells in one cluster compared to all other cells

```
FindAllMarkers(seurat_object,
               test.use = "wilcox",
               logfc.threshold = 0.25,
               min.pct = 0.1,
               only.pos = TRUE,
               …)
```





- Adjust p-values for the number of tests performed

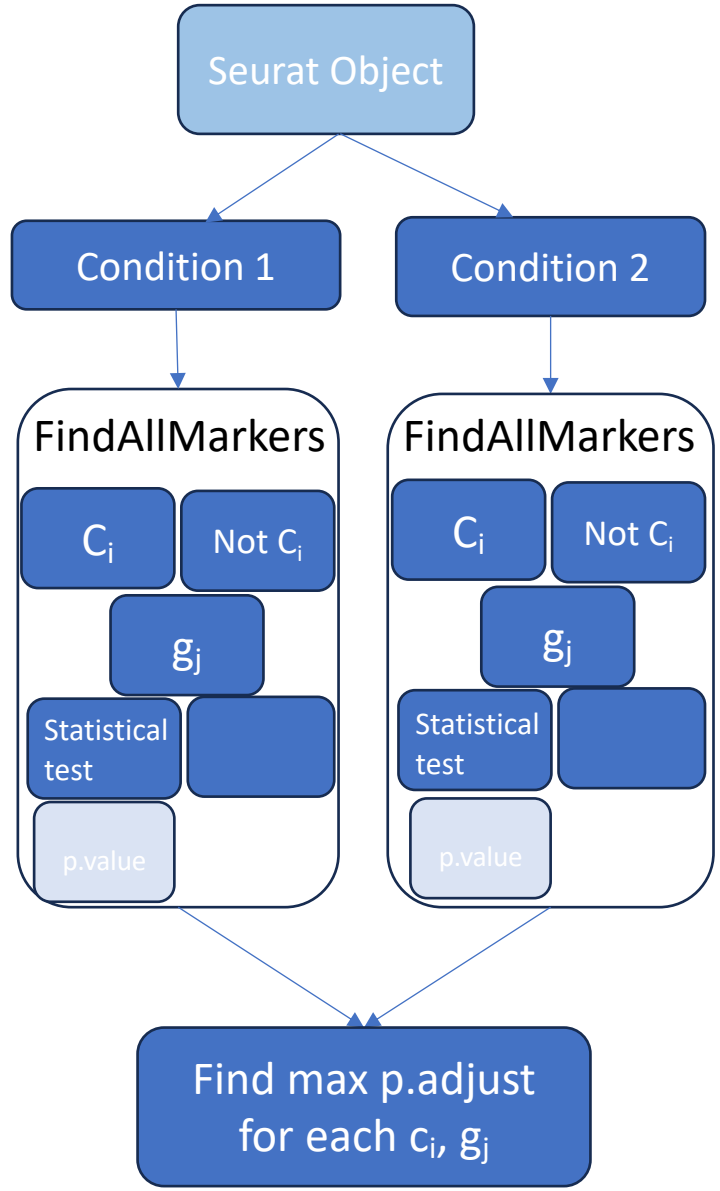
*more on this soon

Find Conserved Markers

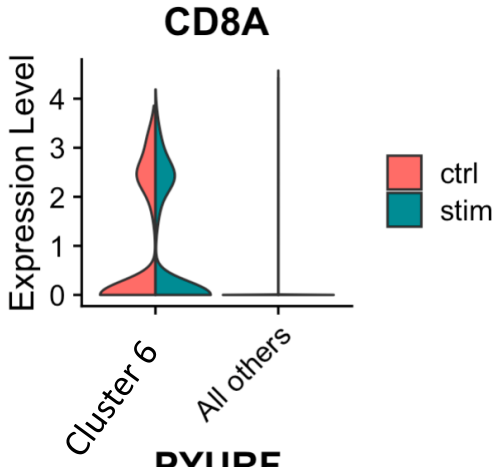
- Find Genes that are markers in two groups

```
FindConservedMarkers(seurat_object,
                     ident.1 = 6,
                     grouping.var = "sample",
                     …)
```

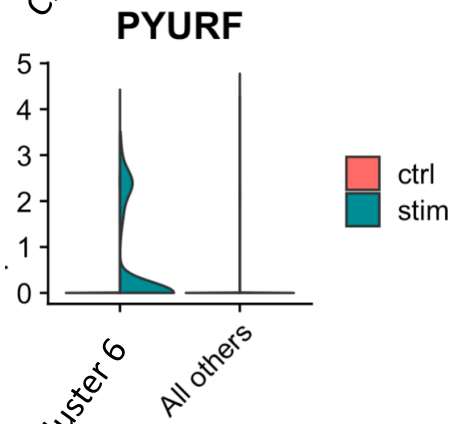
- Split cells by grouping.var



Cluster 6 marker in **both conditions**



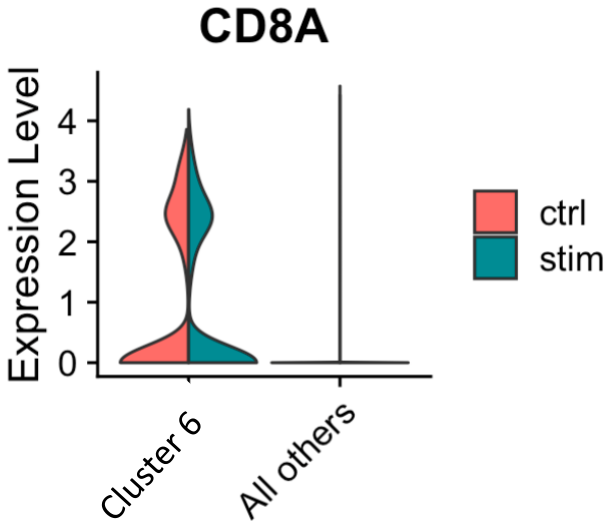
Cluster 6 marker in **stim only**



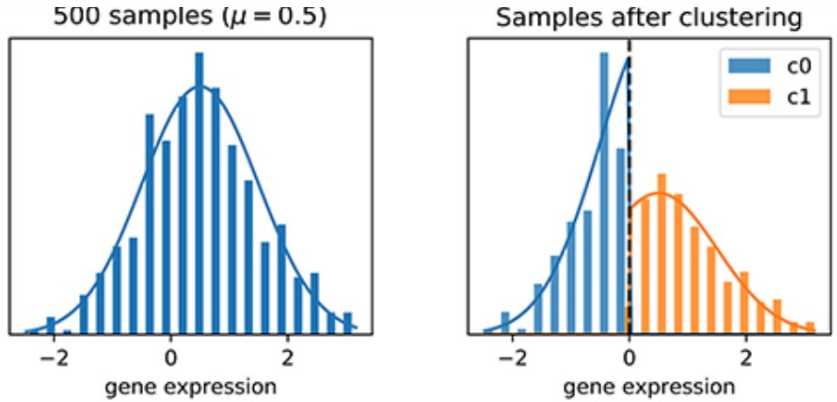
Markers: Which statistical test?

```
FindConservedMarkers(seurat_object,
                     ident.1 = 6,
                     grouping.var = "sample",
                     test.use = “wilcox”,
                     …)
```

There are 9 options, we'll discuss two.



Interpretation of p-values



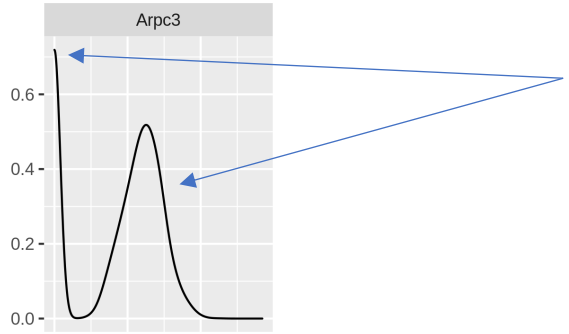
Most statistical tests are designed for hypotheses to be defined before experiments are carried out. In finding cluster marker genes, testing is performed **on the same data that is used to extract the clusters in the first place**. Thus, there will be some genes that are differentially expressed between clusters by construction.

Valid post-clustering differential analysis for single-cell RNA-Seq. Zhang et al Cell. Syst. 2020

Markers: Which statistical test?

- Wilcoxon Rank Sum test - default
  - Assumes independence of samples (cells)
  - no assumption of statistical distribution

- MAST (Model-based Analysis of Single-cell Transcriptomics):
  - Assumes independence of samples (cells)
  - Assumes counts are zero-inflated, *usually* true for single-cell data (though see ref)



Bimodal expression pattern due to “drop out”

  - Allows adjusting for additional covariates such as batch, sex, sample, through the latent.vars parameter

Markers: Results

```
cluster_6_markers = FindConservedMarkers(seurat_object,
                                         ident.1 = 6,
                                         grouping.var = "sample",
                                         only.pos = TRUE)
```

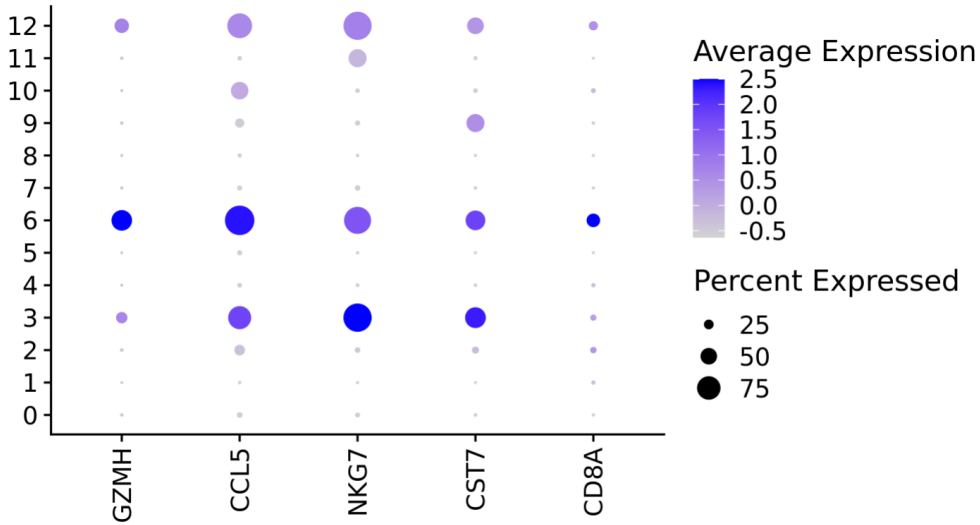
	stim_p_val	stim_avg_log2FC	stim_pct.1	stim_pct.2	stim_p_val_adj	ctrl_p_val	ctrl_avg_log2FC	ctrl_pct.1	ctrl_pct.2	ctrl_p_val_adj	max_pval	minimump_p_val
GNLY	0.000000e+00	1.2250346	0.666	0.127	0.000000e+00	0.000000e+00	1.1372563	0.603	0.111	0.000000e+00	0.000000e+00	0.000000e+00
CD8A	0.000000e+00	2.0559922	0.393	0.055	0.000000e+00	0.000000e+00	2.1068908	0.387	0.058	0.000000e+00	0.000000e+00	0.000000e+00
PRF1	0.000000e+00	1.6820287	0.578	0.110	0.000000e+00	1.916110e-165	1.3894355	0.260	0.046	2.695008e-161	1.916110e-165	0.000000e+00

Multiple p-values!

- `stim_p_val`, `ctrl_p_val` - the raw p-value for each group
- `stim_p_val_adj`, `ctrl_p_val_adj` - the FDR adjusted p-value for each group
- `max_pval` - largest p-value of the two groups
- `minimump_p_val` - combined p-value, using package `metap`

Fold changes

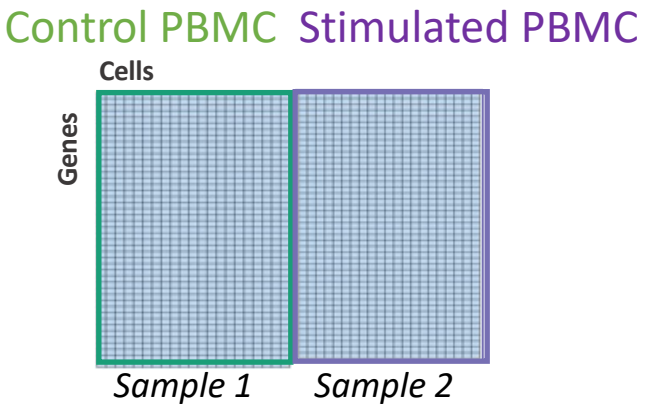
- `stim_avg_log2FC` and `ctrl_avg_log2FC`: log fold-change of the average expression between the two cluster 6 and all other clusters. Positive values indicate that the feature is more highly expressed in the cluster 6.

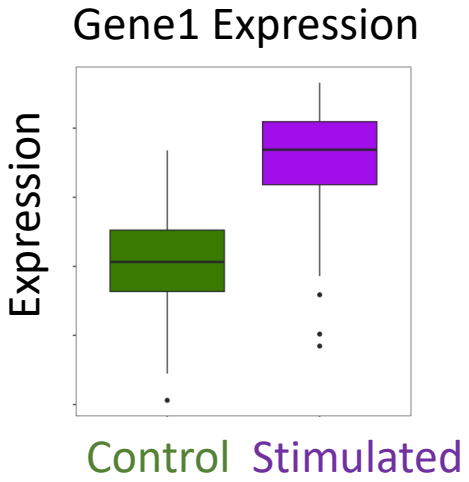


Differential Expression in scRNAseq

1) Genes that are overexpressed in on cell-type compared to all other cell-types

2) Gene that are statistically differently expressed between phenotypes or conditions, usually within a cell-type



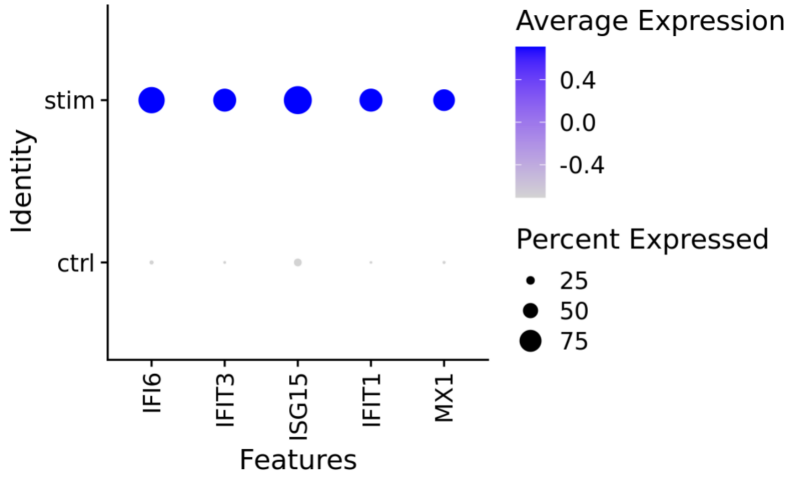


"Differentially Expressed Genes"
- typically requires biological replicates and modeling of inter-replicate variability

Differential Expression: Which statistical test?

With no replicates, naïve method: Seurat FindMarkers with test.use = ”wilcox” or “MAST”

```
deg = FindMarkers(cluster_6,
                  ident.1 = "stim",
                  ident.2 = "ctrl",
                  only.pos = FALSE)
```

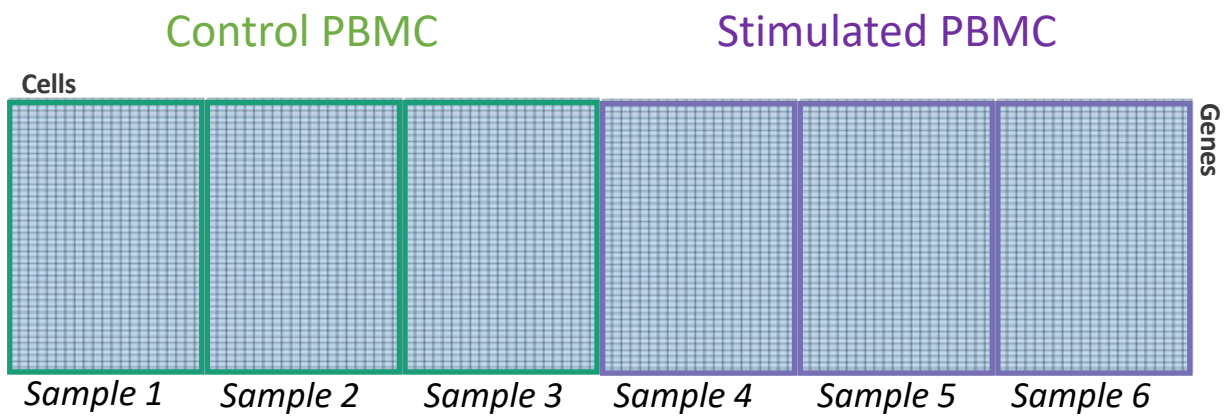


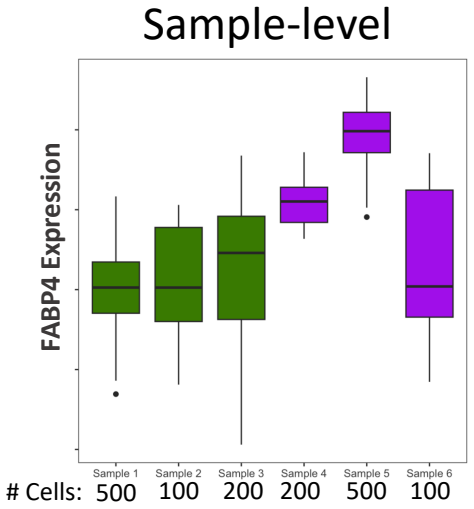
Pseudoreplication bias

“Cells from the same individual share common genetic and environmental backgrounds and are not statistically independent; therefore, they are **subsamples or pseudoreplicates**. Thus, single-cell data have a hierarchical structure that many current single-cell methods do not address, leading to biased inference, highly inflated type 1 error rates, and reduced robustness and reproducibility.”

Benchmarking methods for detecting differential states between conditions from multi-subject scRNAseq data. Juntilla et al Briefings in Bioinformatics 2022

A better study design





(Thanks Eric Reed, for the image)

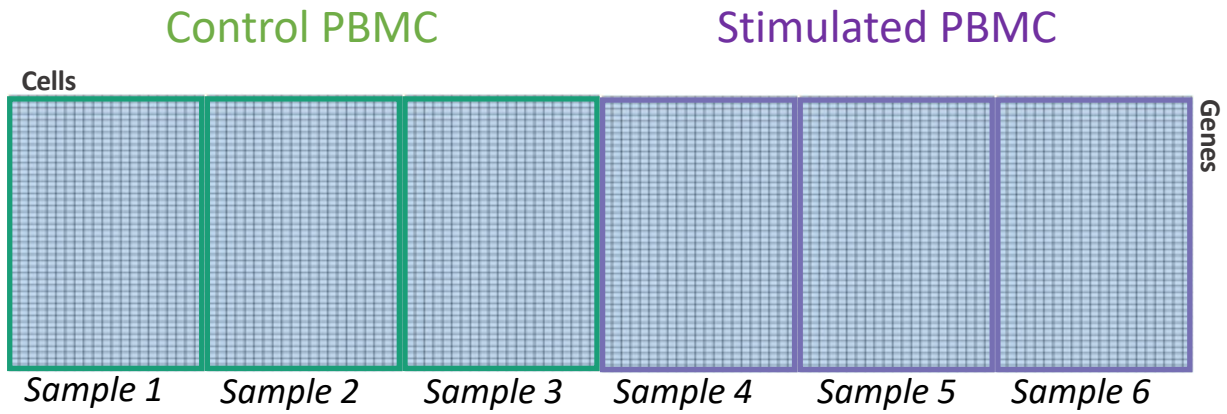
With replicates, there are options to model inter-individual variability:

- **Pseudobulk**: Transform data into bulk-RNAseq-like data by aggregating gene counts within each sample
- **Latent Variable**: Test whether the difference in gene expression between the groups can be explained by the difference in one or multiple latent variables (e.g. batch)
- **Mixed-model**: modeled as random effect

Helpful to talk to statistician to choose

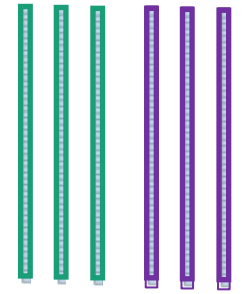
Benchmarking methods for detecting differential states between conditions from multi-subject scRNAseq data. Juntilla et al Briefings in Bioinformatics 2022

Pseudobulk approach



Average gene expression in each sample

(or sample + cell-type)



Bulk-RNAseq differential expression tools can then be used, e.g. linear regression with covariates. For each gene, condition i:

Expression Values

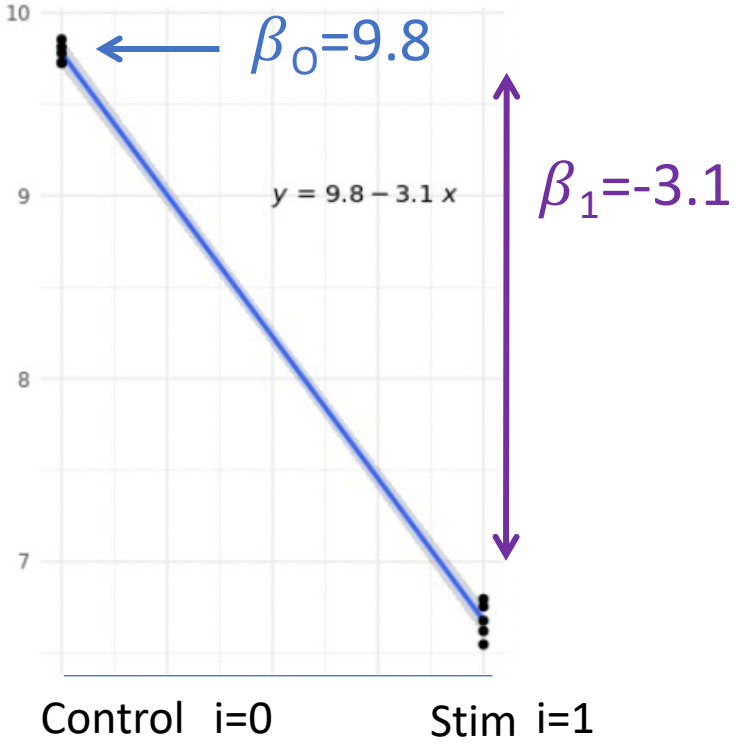
$$y \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \beta_o + \beta_1 X_i$$

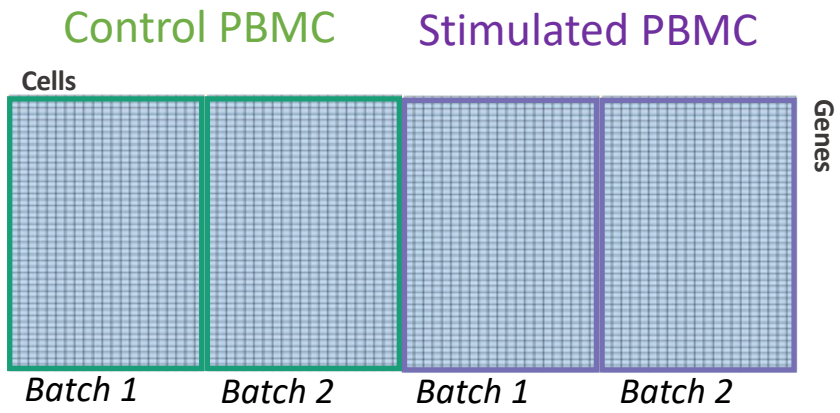
Intercept

Slope: difference between conditions

Condition: 0-control 1-stim



Latent Variables for batch effects



In this case, our samples were grouped into batches.
e.g. subject tested in both conditions.

Can difference in gene expression between the conditions be explained by batch?
For each gene, the expression in cell c:

Expression Values $y_c \sim N(\mu_{ij}, \sigma^2)$

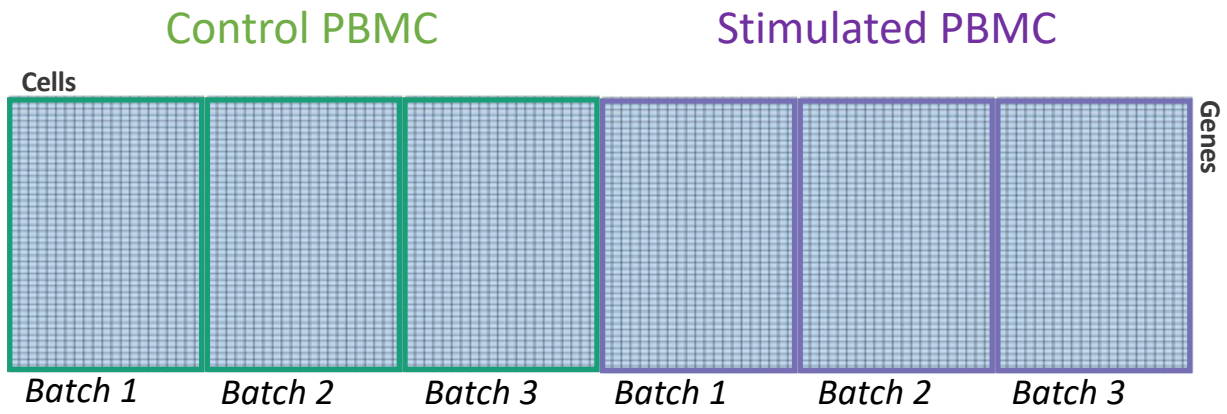
$$\mu_{ij} = \beta_o + \beta_1 X_i + \beta_2 X_j$$

- β_o: Intercept
- $\beta_1 X_i$: Condition effect — Condition: 0-control, 1- stim
- $\beta_2 X_j$: Batch effect — Batch: 0 - Batch 1, 1 - Batch 2

Seurat can do this:

```
deg = FindMarkers(cluster_6,
                  ident.1 = "stim",
                  ident.2 = "ctrl",
                  test.use="MAST",
                  latent.vars = "batch"
                  only.pos = FALSE)
```

Linear mixed-effect models are flexible



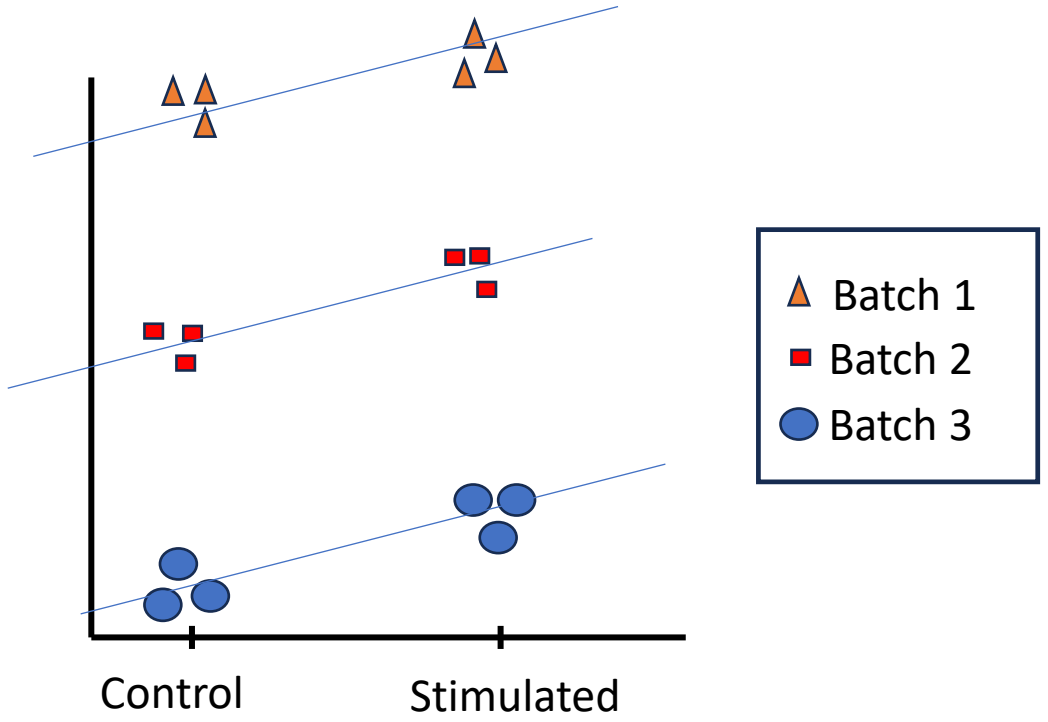
Mixed-effect models allow us to model correlations within multiple batches and between conditions.

$$y_c \sim N(\mu_{ij}, \sigma^2),$$
$$\boldsymbol{\gamma_j} \sim N(0, \varphi^2)$$

$$\mu_{ij} = \beta_1 X_i + \boldsymbol{\gamma_j} \boldsymbol{W_j}$$

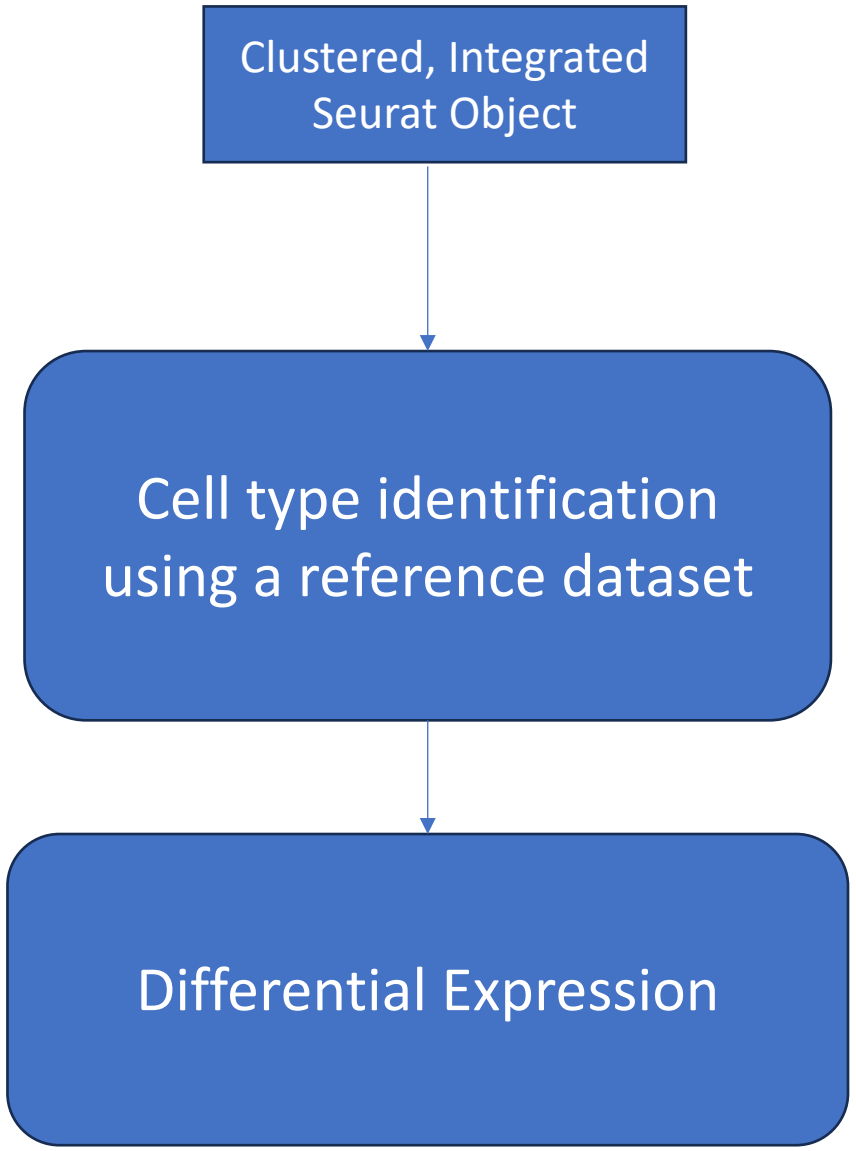
Random coefficient for batch j

Design matrix for random effects of batch j, condition i



https://stats.oarc.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models/

Wrap Up



- Choosing a method and reference
  - SingleR correlation with bulk database
  - Integration mapping with matching scRNAseq

- Cell type Marker Genes
- Condition-specific Expression

References for statistical methods

- Zhang et al Cell. Syst. 2020 Valid post-clustering differential analysis for single-cell RNA-Seq.
- Zimmerman 2021 **A practical solution to pseudoreplication bias in single-cell studies**
- Juntilla et al Briefings in Bioinformatics 2022 Benchmarking methods for detecting differential states between conditions from multi-subject scRNAseq data.
- Das 2021 Statistical methods for analysis of single-cell RNA-sequencing data

Bayesian mixed effects regression example:

Karagiannis et al 2023 Multi-modal profiling of peripheral blood cells across the human lifespan reveals distinct immune cell signatures of aging and longevity

UCLA Stats: Mixed-effects model tutorial