Single Cell RNASeq

Application and Analysis

Overview

Bulk RNASeq

- Library preparation
- Analysis methods
  - Normalization strategies
  - Differential gene expression
  - Linear models

Single cell RNASeq

- Library preparation
  - Demux and barcoding
- Analysis methods
  - Clustering
  - Cluster marker identification
  - Differential gene expression

Bulk vs. single cell RNASeq

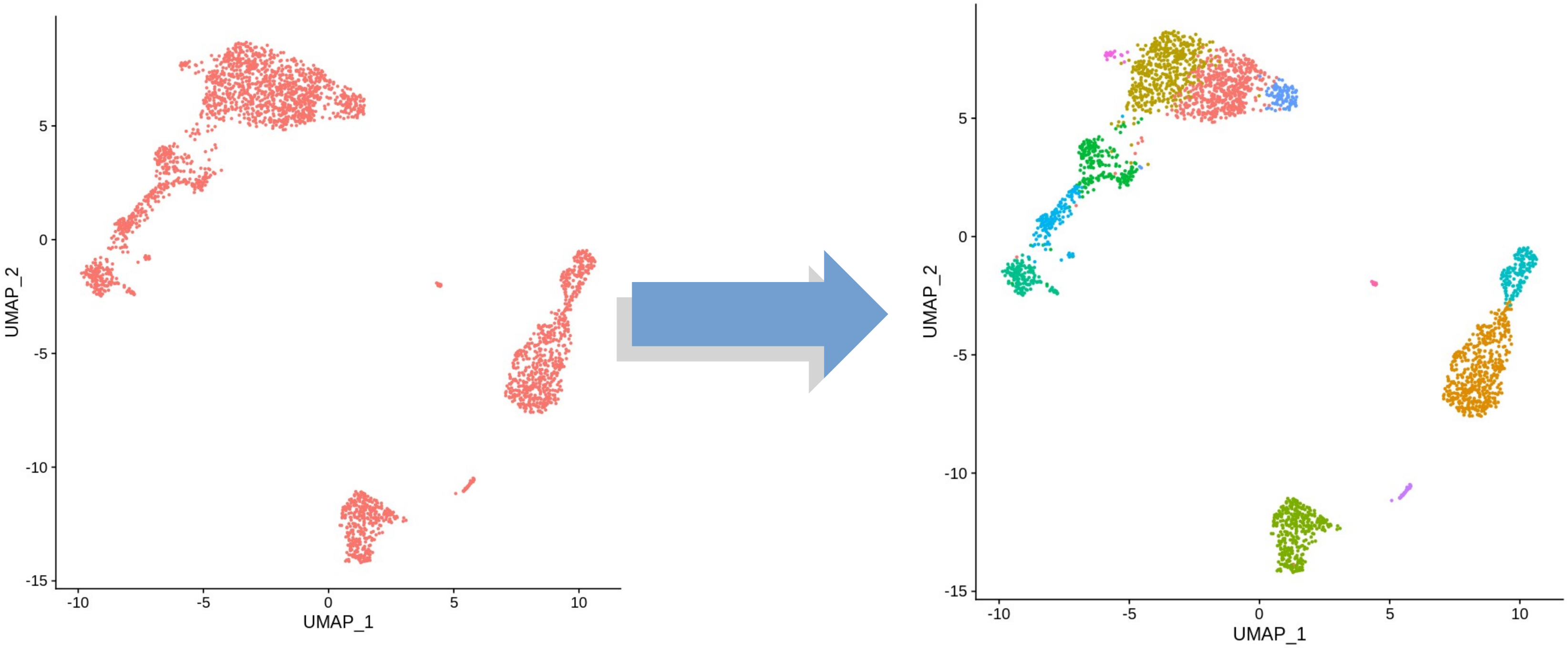
Bulk RNASeq

- Measures an average snapshot of the population of cells
- Well established methodology
  - Technology
  - Algorithms
- Requires extra work in cell sorting for cell type specific expression
  - Still does not have enough resolution

scRNASeq

- Addresses the inadequacies of bulk RNASeq as regards cell specific expression
- Shares much of the same tooling and methods as bulk RNASeq
  - Library preparation and sequencing
  - Alignment methods
  - Counting
- Introduces its own new problems
  - From its own chemistry
  - From the basic premise of what is asked

How do we define clusters?



Determining clusters of cell types

Identify by prior knowledge

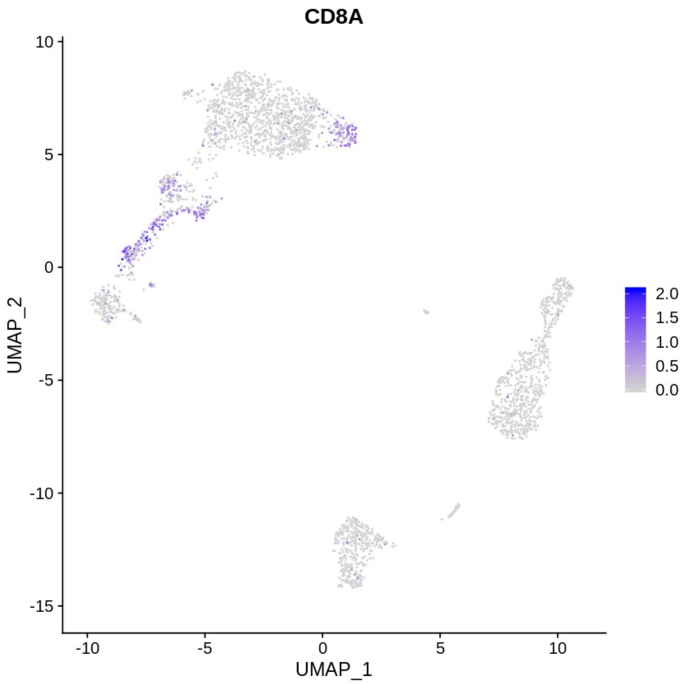
- Manually
  - Label with known biomarkers
  - Requires prior knowledge / expectations of cell populations
- Comparison to labeled single cell data sets
  - e.g. scMatch

***de novo* clustering**

- Methods
  - K-means
    - Parametric
    - Not recommended
  - K-nearest neighbors
    - Variants of KNN
    - Non-parametric
- Assumes all genes are worth the same importance

Using cell markers

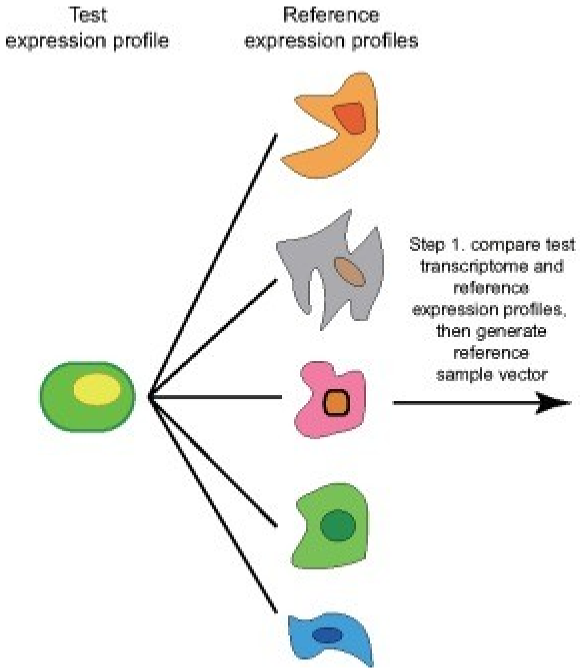
- Expression of known markers
- Investigator needs to know the genes prior
  - Not always as uniquely expressed as previously thought
  - Does not address subpopulations



Compare to other single cell databases

- Whole expression profile

- Requires a well curated database



Classification by reference

Neural nets and machine learning models

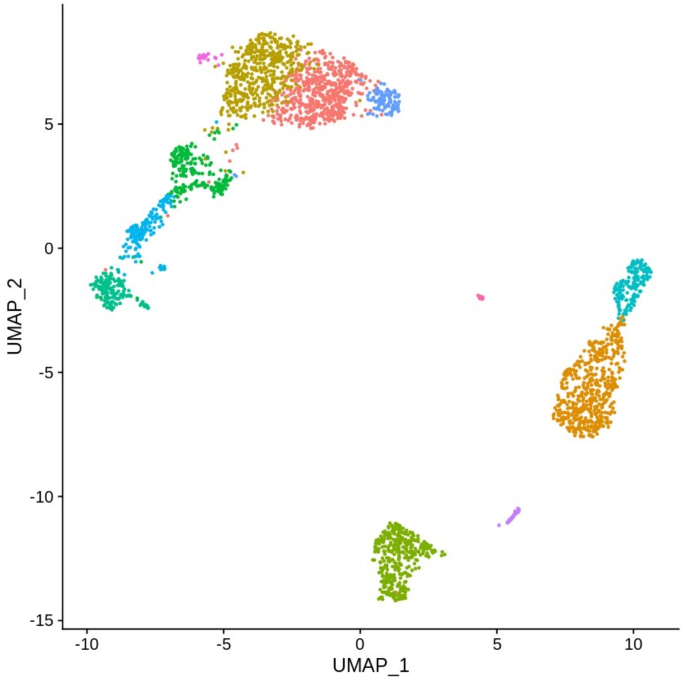
- Requires a data set
- Requires training
- Not portable
- Odd performance for a cell types not in the database

Correlation based

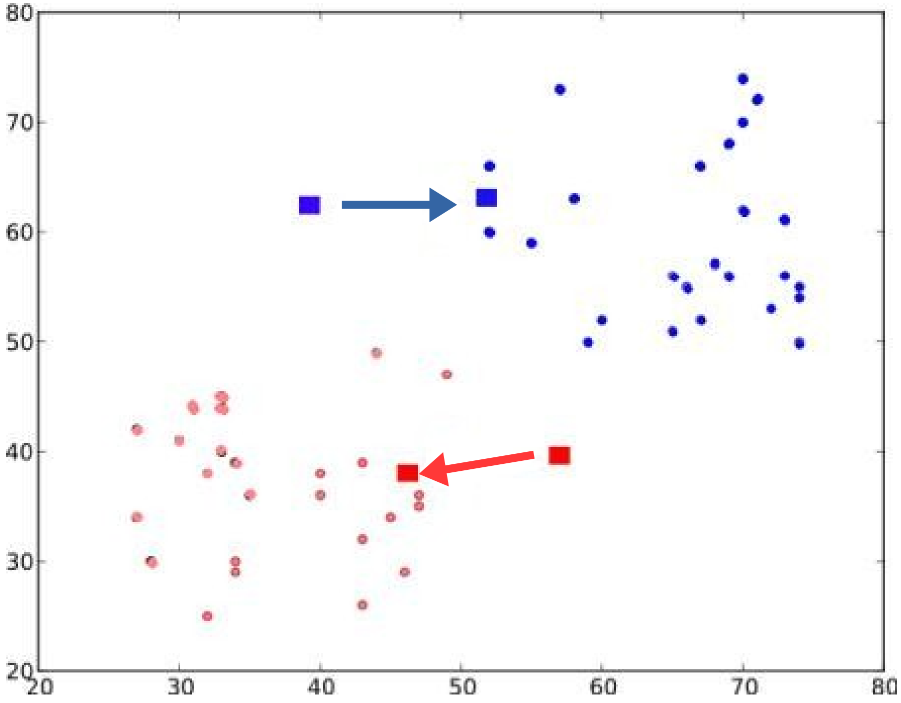
- e.g. scMatch
- Rank comparisons to curated database
  - Best correlation is the assignment

de novo clustering

- Clustering methods
  - K-means
  - KNN-based
  - Cell populations based on the data
    - Assume all genes are of same importance
      - Euclidean distance
- Identify markers
  - Differential gene expression
    - Against background (i.e. all other cells)



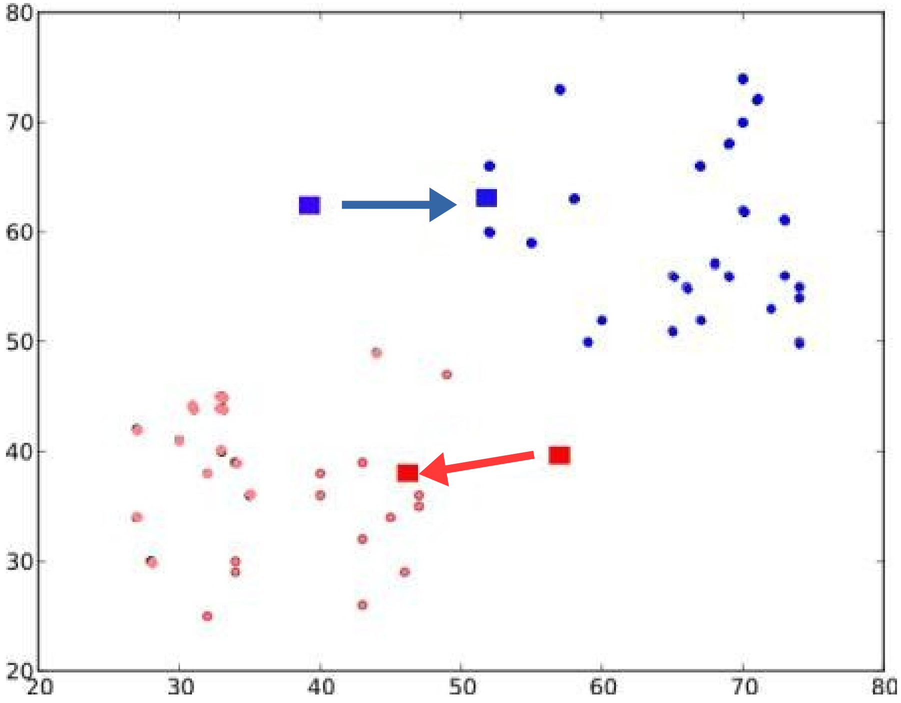
K-means



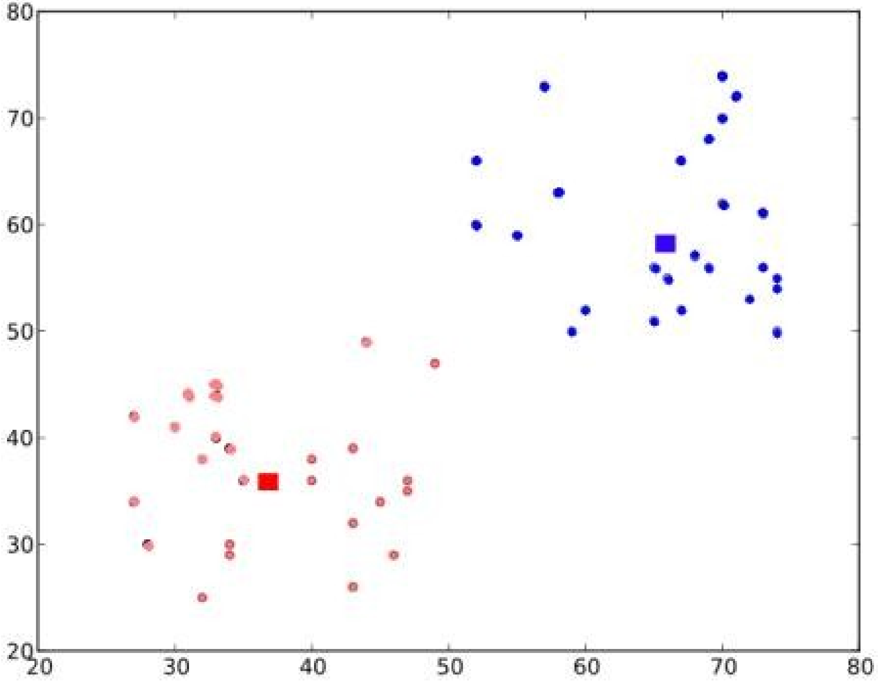
- Start at random points

- Keep updating to new points to find “center”

K-means



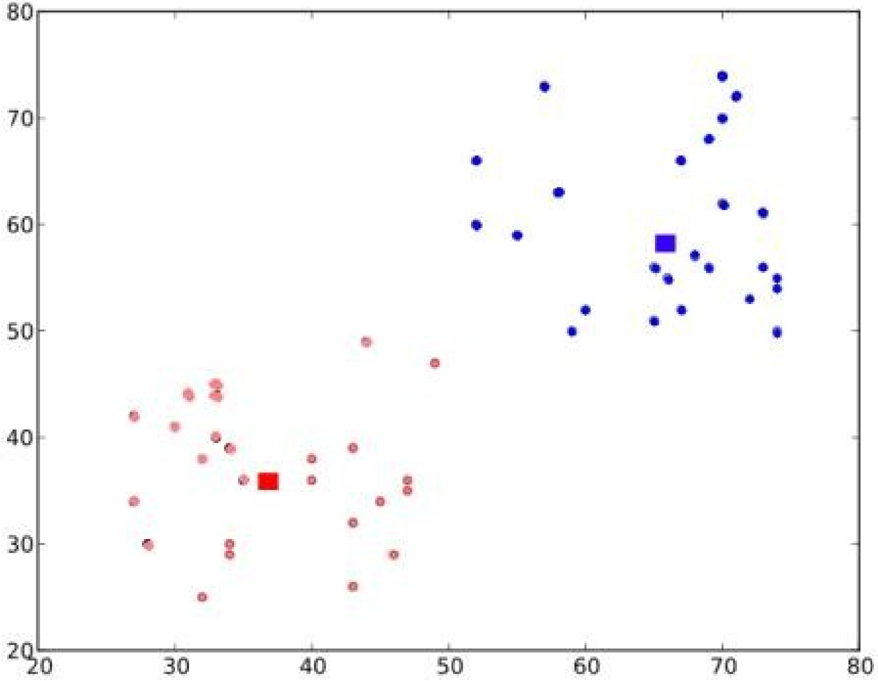
80
70
60
50
40
30
20
20
30
40
50
60
70
80



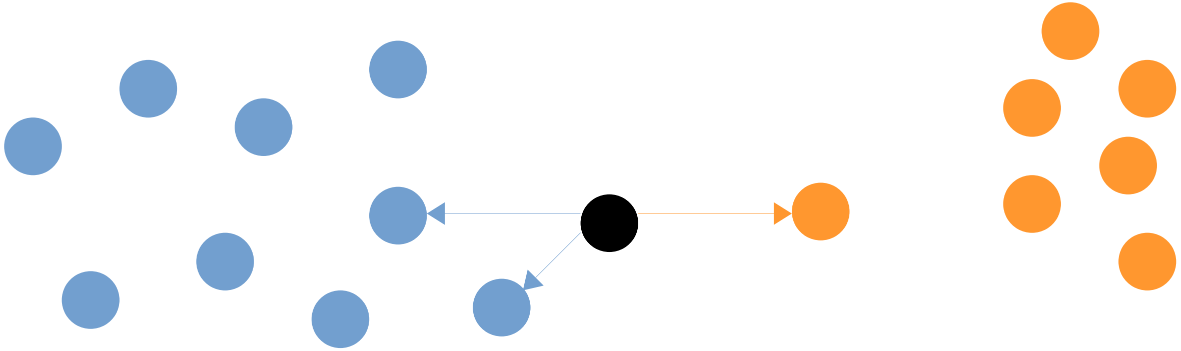
80
70
60
50
40
30
20
20
30
40
50
60
70
80

K-means

- Determines clusters by normal distribution of points
  - Clusters determined by minimizing variance from “centers”
- Assumes clusters are about same size
- Variance is same in all dimensions
  - Clusters are “spherical”



K-nearest neighbors



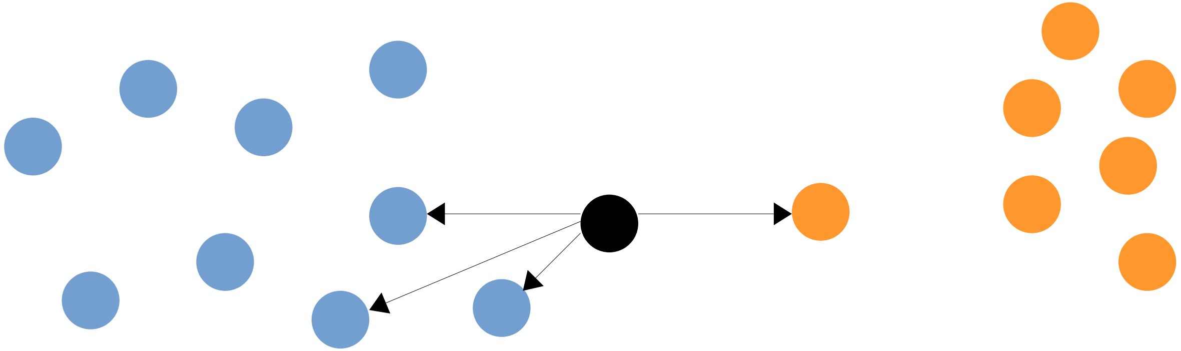
k = 3

2-1 majority vote

So assign black to blue

But KNN is a classification method, so how do we use it if we do not already know the classes.

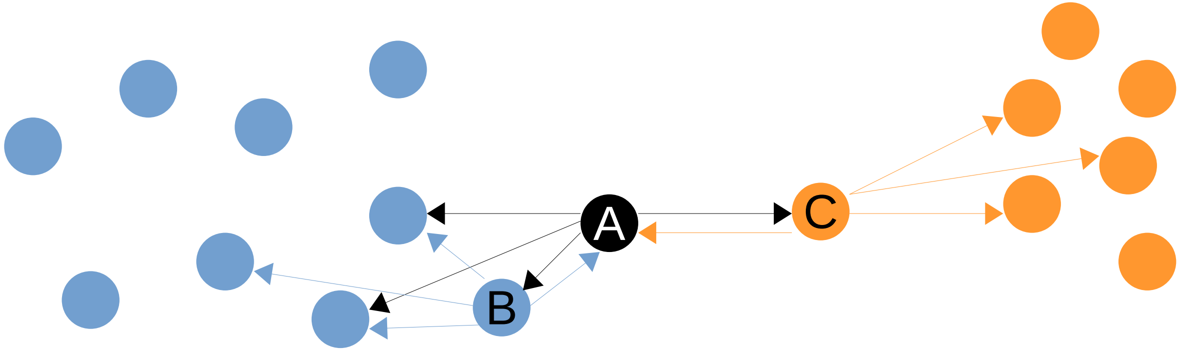
Shared nearest neighbors



Start with KNN.

Instead of Euclidean, compute “distance” as a number of shared neighbors with k_i neighbor.

Shared nearest neighbors



A shares 2 neighbors with B.

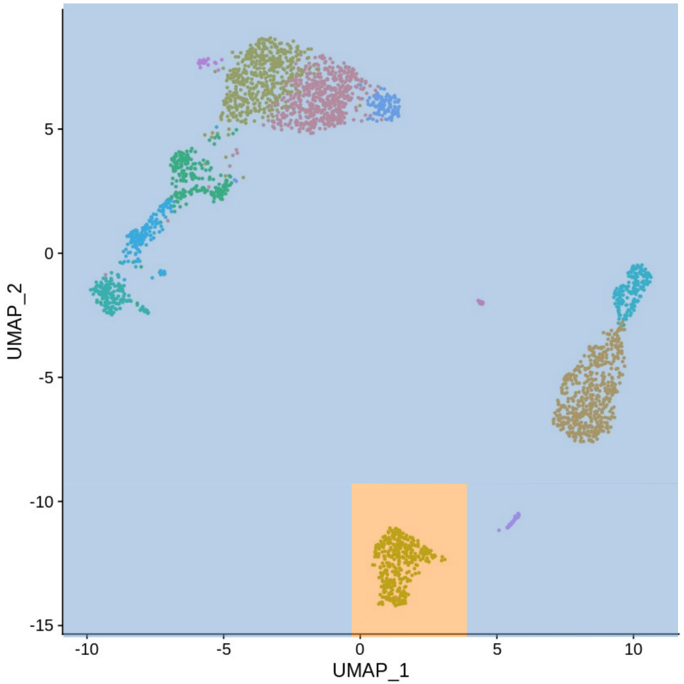
A shares 0 neighbors with C.

Set a threshold for minimum “sharing” to consider breaking cluster.
(Can be algorithmically derived.)

Differential gene expression analysis in scRNASeq

- Finding markers
  - Cluster against background

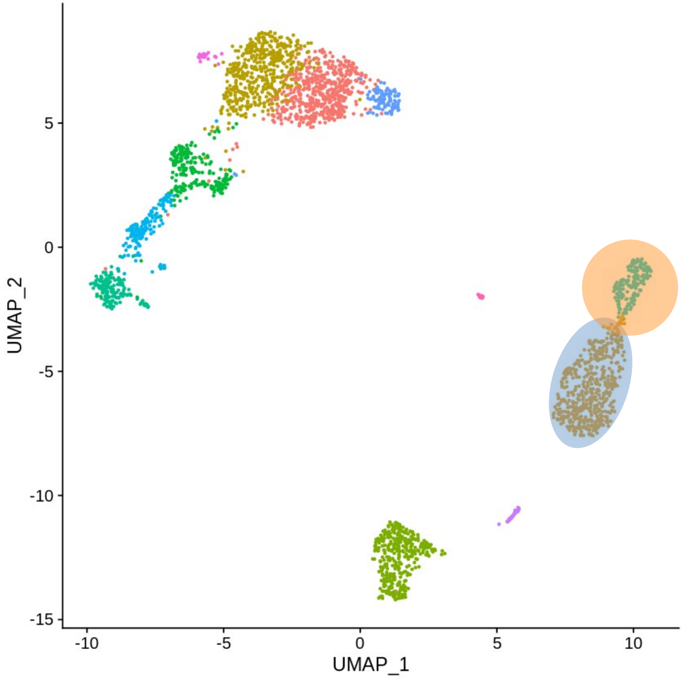
- Comparing clusters



Differential gene expression analysis in scRNASeq

- Finding markers
  - Cluster against background

- Comparing clusters



Differential gene expression analysis in scRNASeq

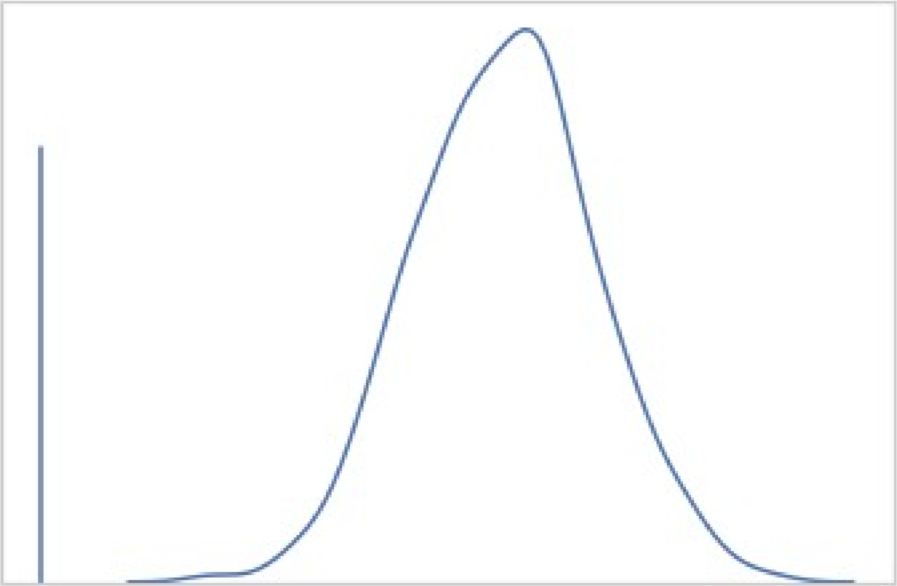
- Analogous to bulk RNASeq

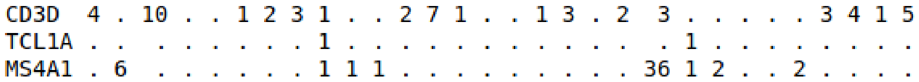
- Comparisons
  - Cluster choice
  - Conditional (e.g. KO vs WT)

Differential gene expression analysis in scRNASeq

- Handling sparsity
- Total RNA per cell
  - Is itself a feature
  - i.e. not every cell intrinsically has the same amount of RNA
- Normalize the data
  - Especially when combining data sets
- Differential gene expression
  - Pick clusters
  - MAST

Those zeroes are still a problem





```
CD3D  4 . 10 . . 1 2 3 1 . . 2 7 1 . . 1 3 . 2  3 . . . . . 3 4 1 5
TCL1A . .  . . . . . . 1 . . . . . . . . . . .  . 1 . . . . . . . .
MS4A1 . 6  . . . . . . 1 1 1 . . . . . . . . . 36 1 2 . . 2 . . . .
```

Cannot just use DESeq2, edgeR, voom, *etc.*

MAST

- Accounts for sparsity in the expression data
  - Simultaneously accounts for:
    - Rate of expression (*i.e.* does it express)
    - Extent of positive expression
- Preserves each cell’s sequencing depth information
  - As a covariate in the model
- Includes gene set enrichment analysis
  - GO terms

MAST: models depth as a factor

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Gene expression

Sequencing depth

Condition

Model more effects in experiment

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

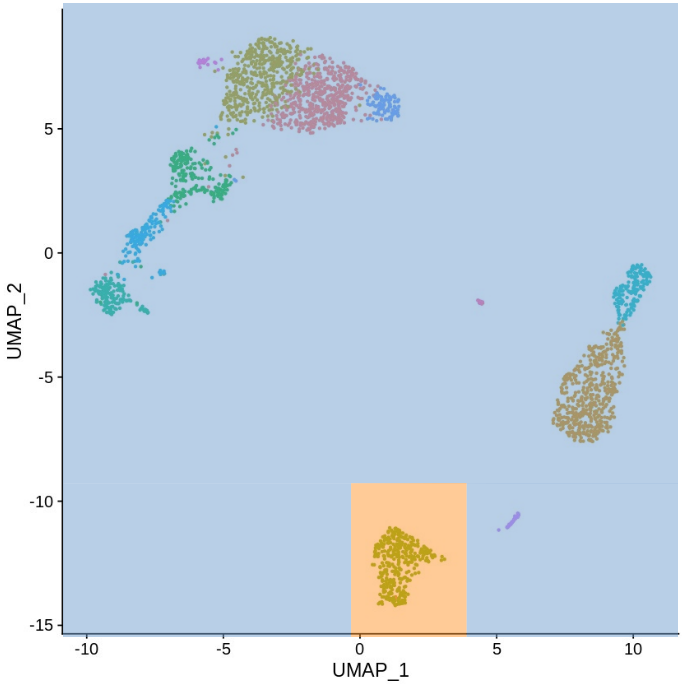
Gene expression

Sequencing depth

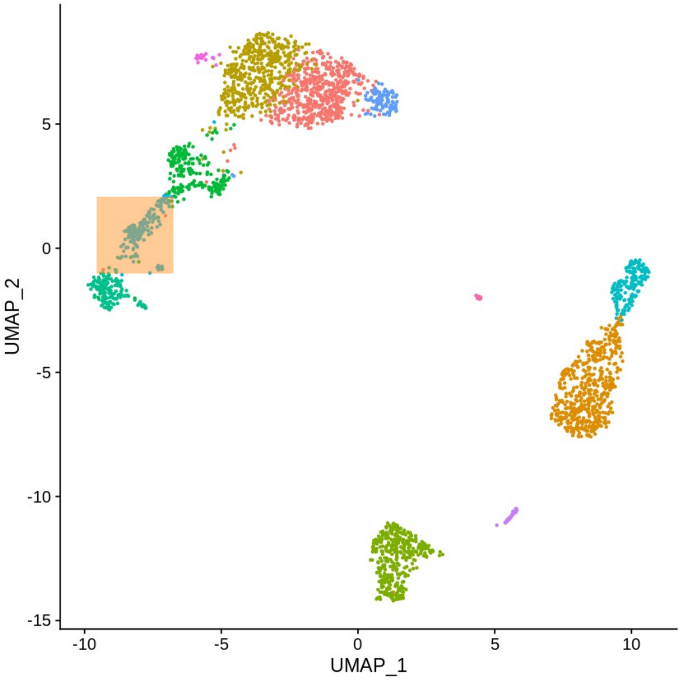
Condition

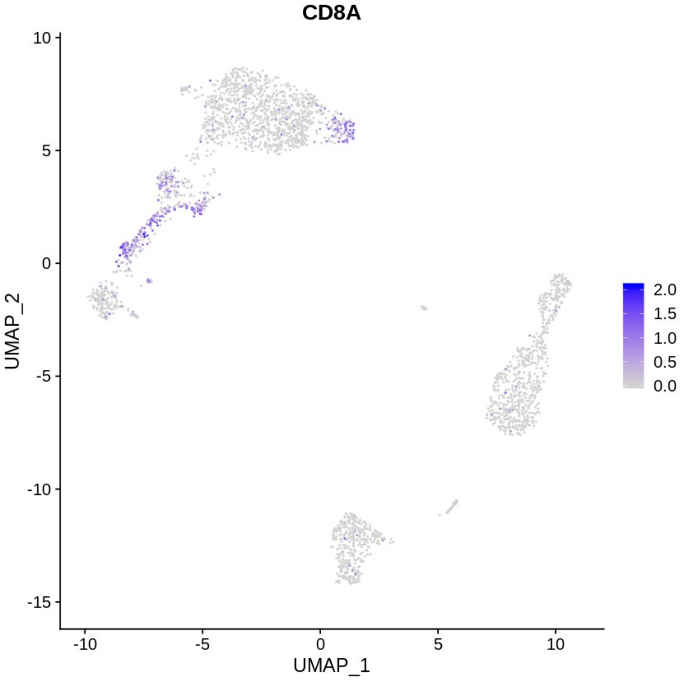
Identify marker genes

- Identify cluster of interest
- Differential gene expression
  - Cluster vs. pool of other cells

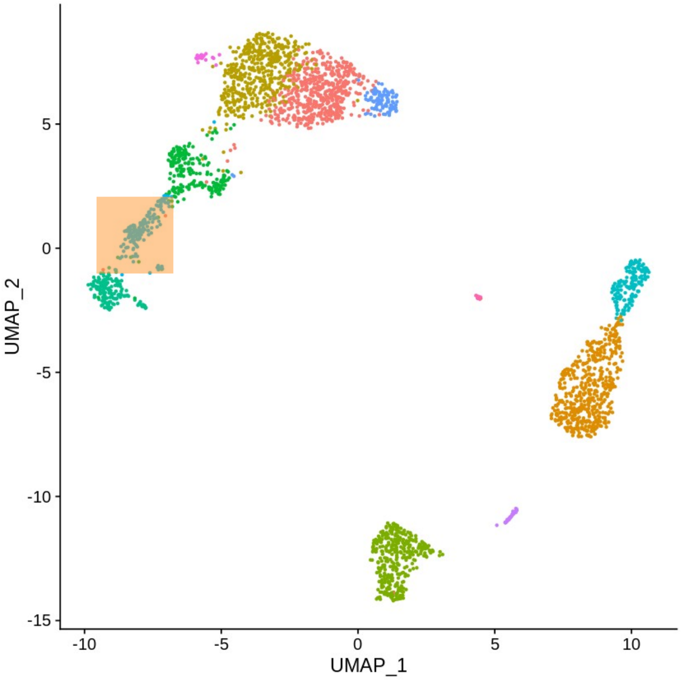


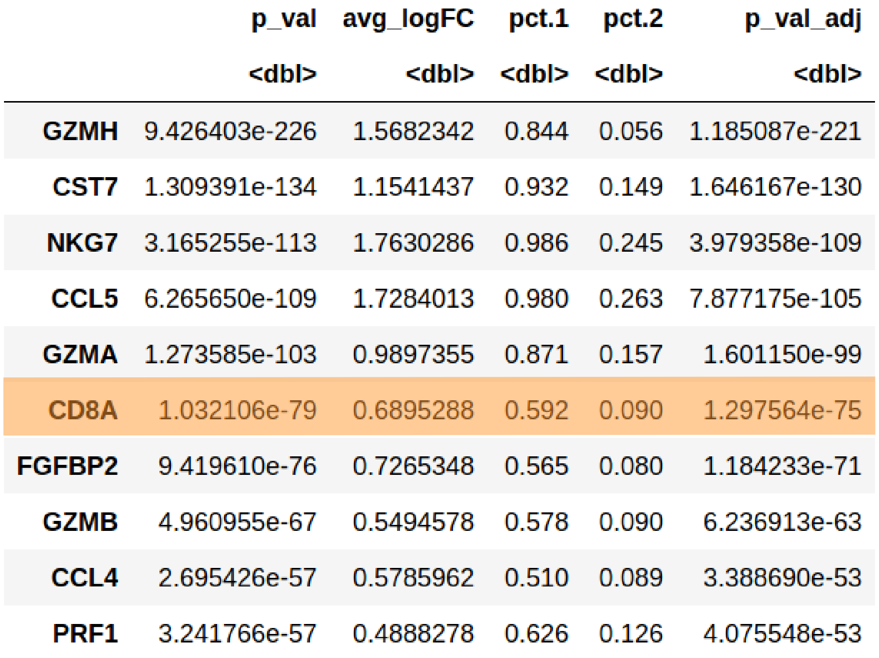
Cluster gene markers





Cluster gene markers



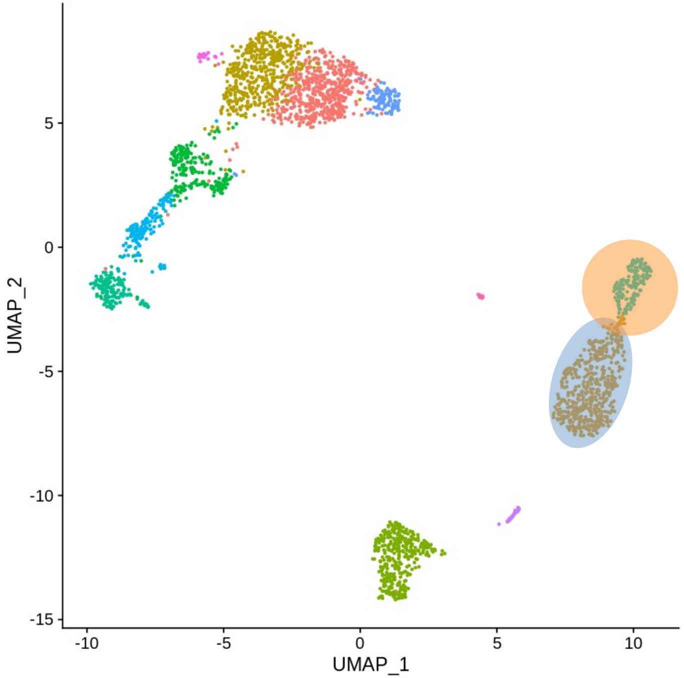


	p_val	avg_logFC	pct.1	pct.2	p_val_adj
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
GZMH	9.426403e-226	1.5682342	0.844	0.056	1.185087e-221
CST7	1.309391e-134	1.1541437	0.932	0.149	1.646167e-130
NKG7	3.165255e-113	1.7630286	0.986	0.245	3.979358e-109
CCL5	6.265650e-109	1.7284013	0.980	0.263	7.877175e-105
GZMA	1.273585e-103	0.9897355	0.871	0.157	1.601150e-99
CD8A	1.032106e-79	0.6895288	0.592	0.090	1.297564e-75
FGFBP2	9.419610e-76	0.7265348	0.565	0.080	1.184233e-71
GZMB	4.960955e-67	0.5494578	0.578	0.090	6.236913e-63
CCL4	2.695426e-57	0.5785962	0.510	0.089	3.388690e-53
PRF1	3.241766e-57	0.4888278	0.626	0.126	4.075548e-53

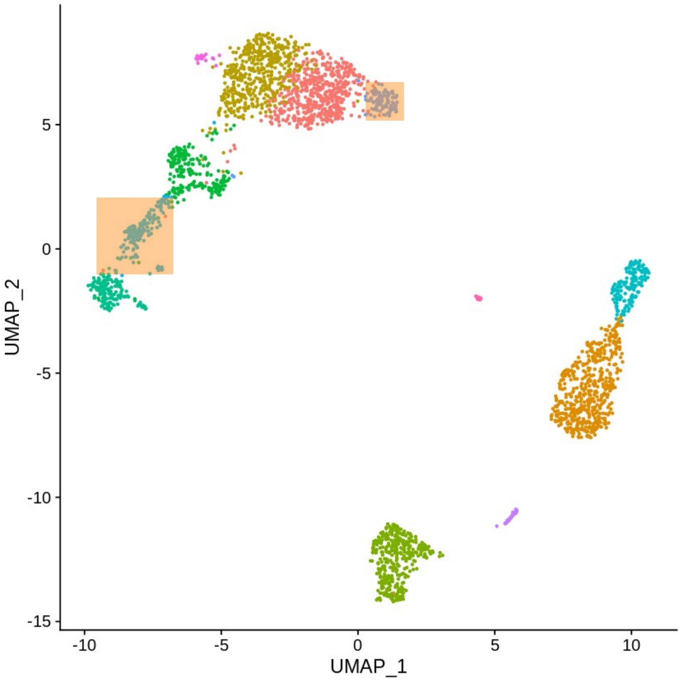
Differential expression

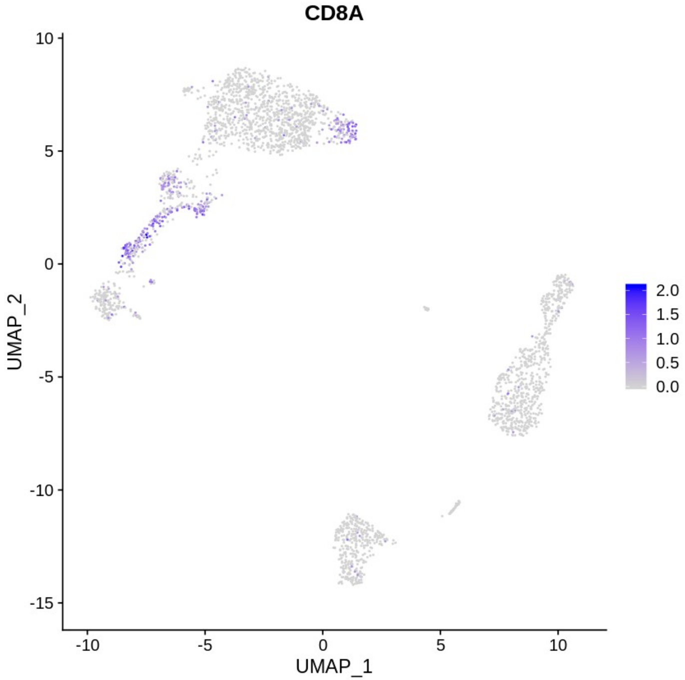
- Between different clusters in same sample
  - Analysis of sub-populations

- Between two samples
  - Within a select cluster
  - Throughout all cells

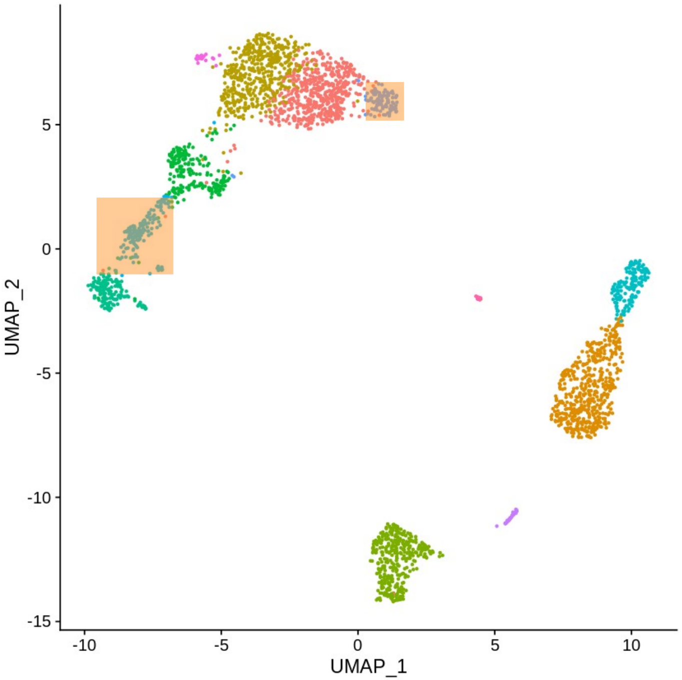


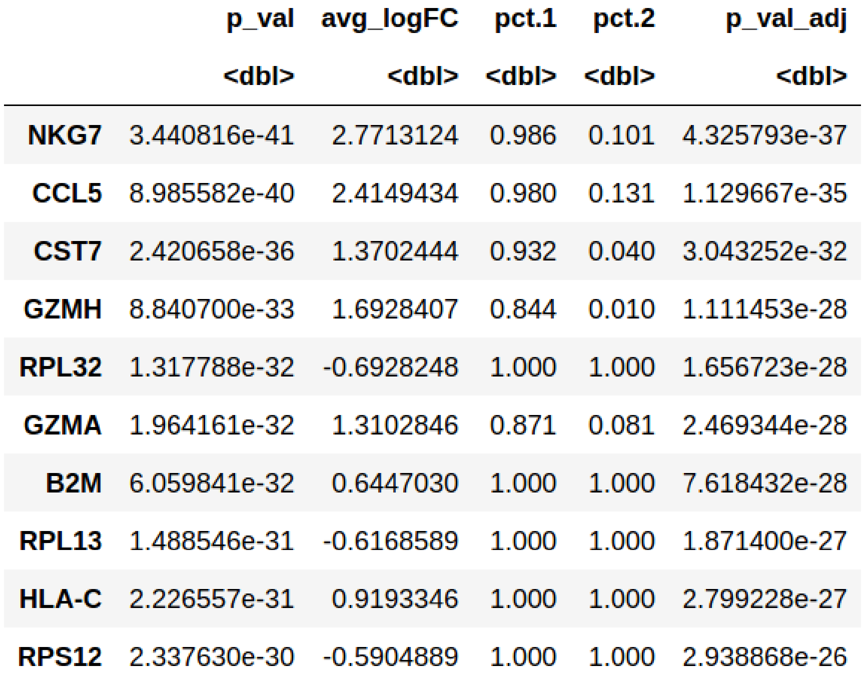
What is the difference between these two clusters expressing CD8?





What is the difference between these two clusters expressing CD8?



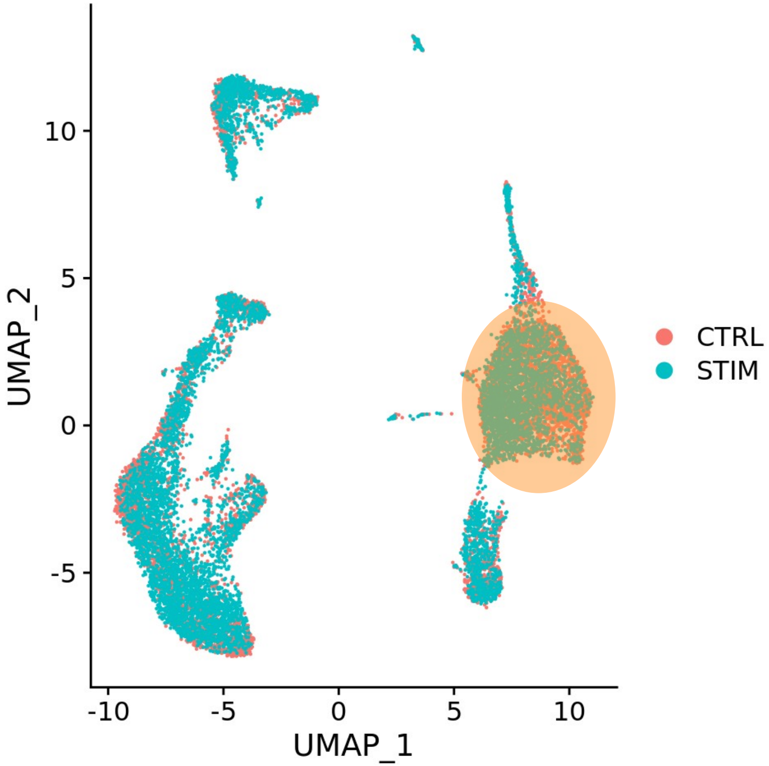


	p_val	avg_logFC	pct.1	pct.2	p_val_adj
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
NKG7	3.440816e-41	2.7713124	0.986	0.101	4.325793e-37
CCL5	8.985582e-40	2.4149434	0.980	0.131	1.129667e-35
CST7	2.420658e-36	1.3702444	0.932	0.040	3.043252e-32
GZMH	8.840700e-33	1.6928407	0.844	0.010	1.111453e-28
RPL32	1.317788e-32	-0.6928248	1.000	1.000	1.656723e-28
GZMA	1.964161e-32	1.3102846	0.871	0.081	2.469344e-28
B2M	6.059841e-32	0.6447030	1.000	1.000	7.618432e-28
RPL13	1.488546e-31	-0.6168589	1.000	1.000	1.871400e-27
HLA-C	2.226557e-31	0.9193346	1.000	1.000	2.799228e-27
RPS12	2.337630e-30	-0.5904889	1.000	1.000	2.938868e-26

Differential expression

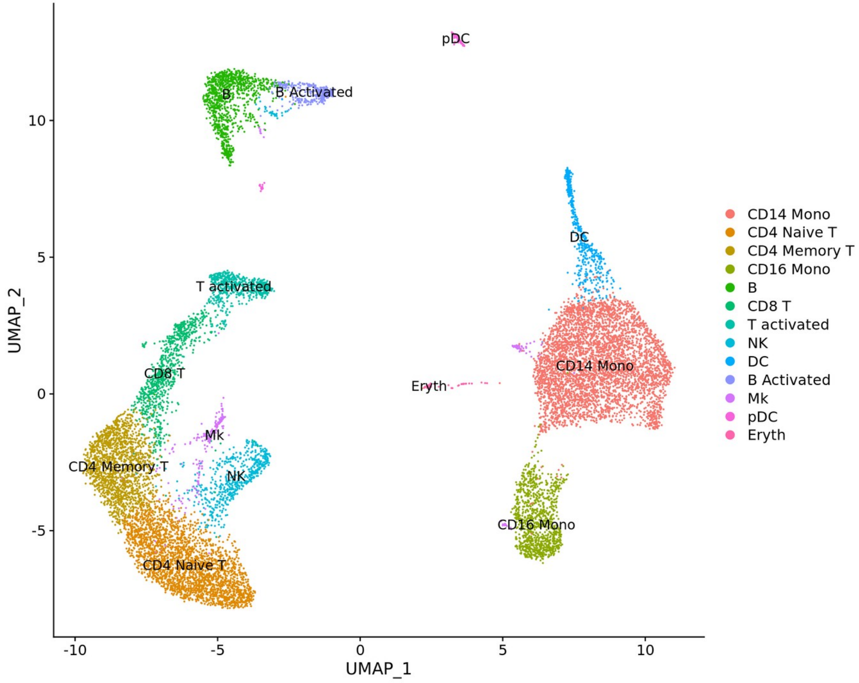
- Between different clusters in same sample
  - Analysis of sub-populations

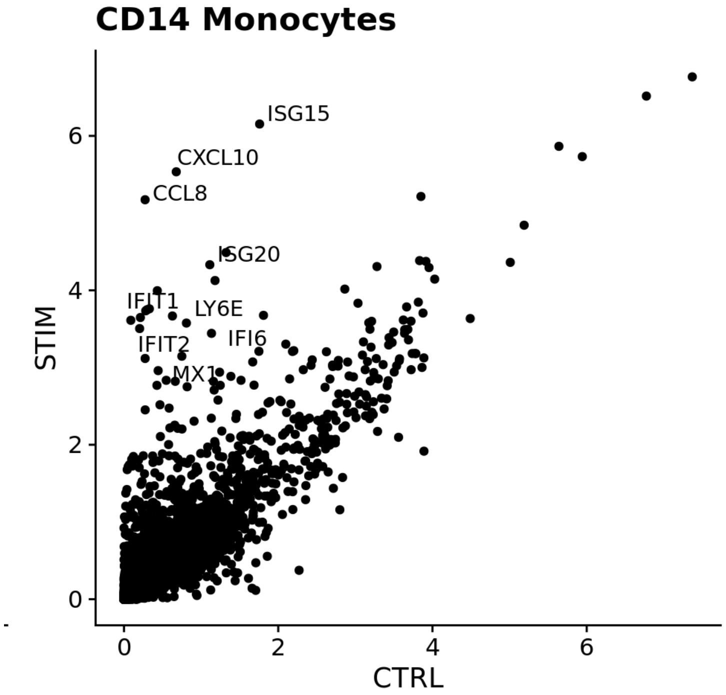
- Between two samples
  - Within a select cluster
  - Throughout all cells



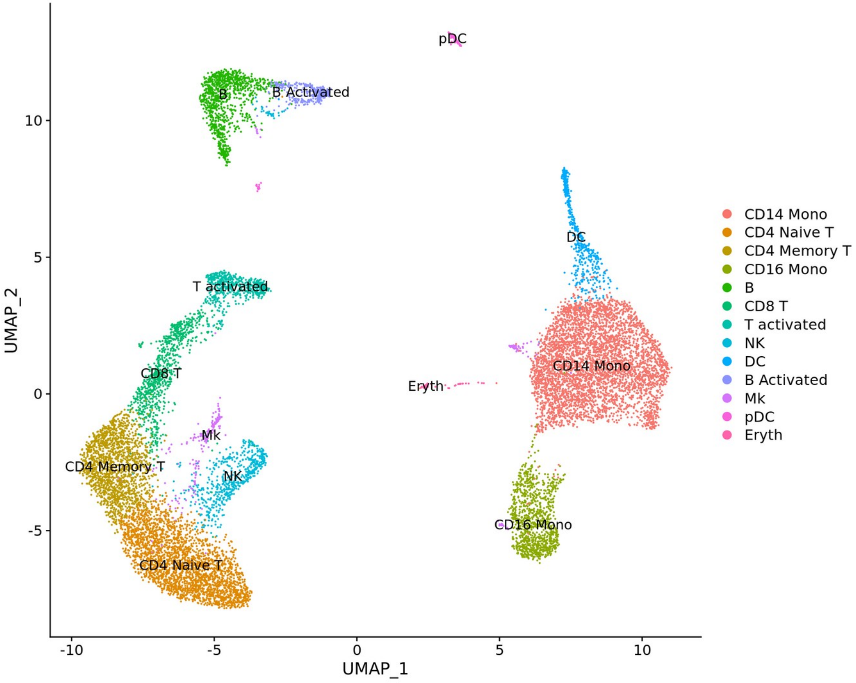
Do we see a difference in expression between CTRL and STIM in selected cluster?

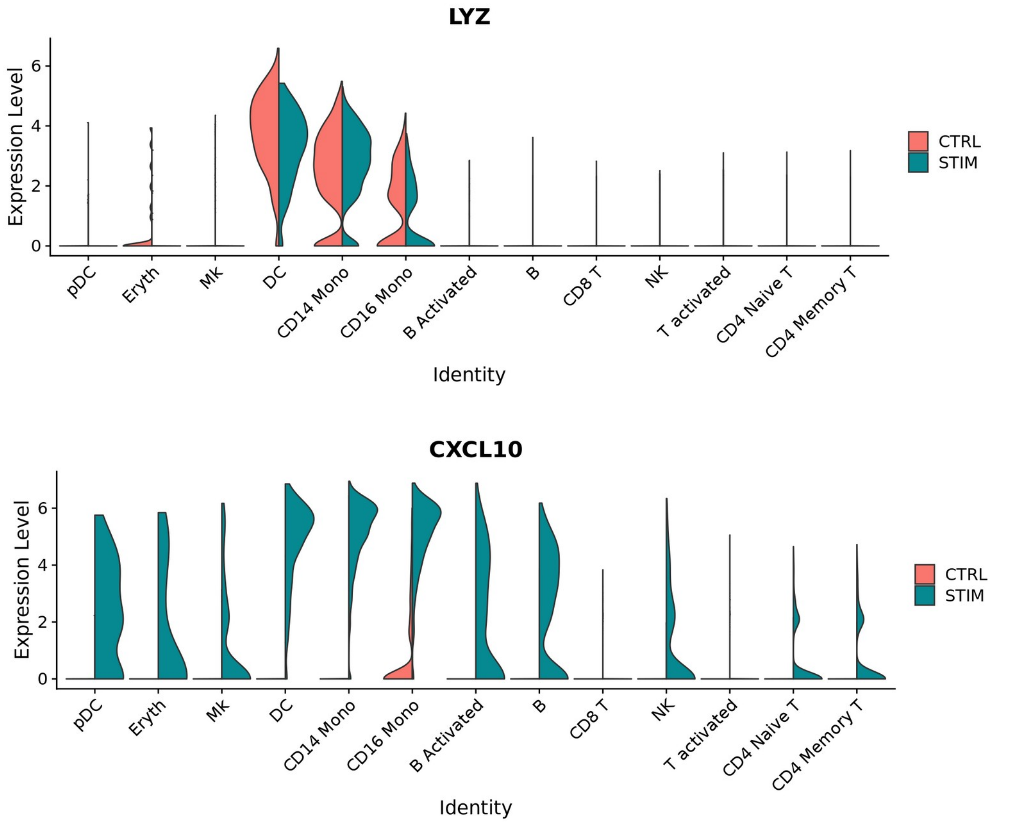
Differential expression




CD14 Monocytes

Differential expression

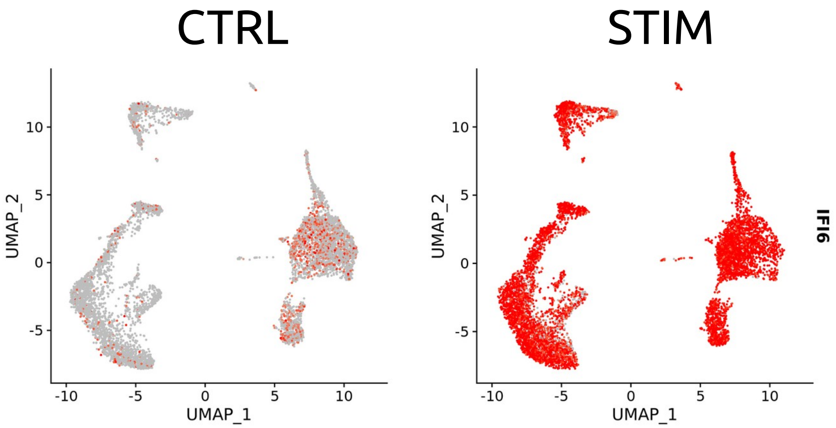




Differential expression

- Between different clusters in same sample
  - Analysis of sub-populations

- Between two samples
  - Within a select cluster
  - Throughout all cells



Lineage / Differentiation

Identify the expression changes involved in:

- Cell cycle changes
- Cell type differentiation
- Cellular activation

Analysis of lineage

Manually

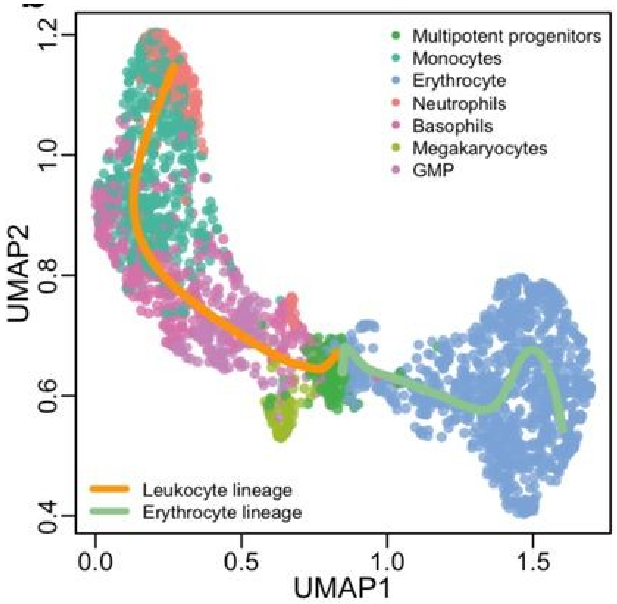
- Pick out clusters along presumed lineage
- Differential expression analysis between all the clusters

Algorithmically

- Auto detects related cells
- Predicts direction of lineage
- Test for the pattern of differential expression

Algorithmic lineage analysis

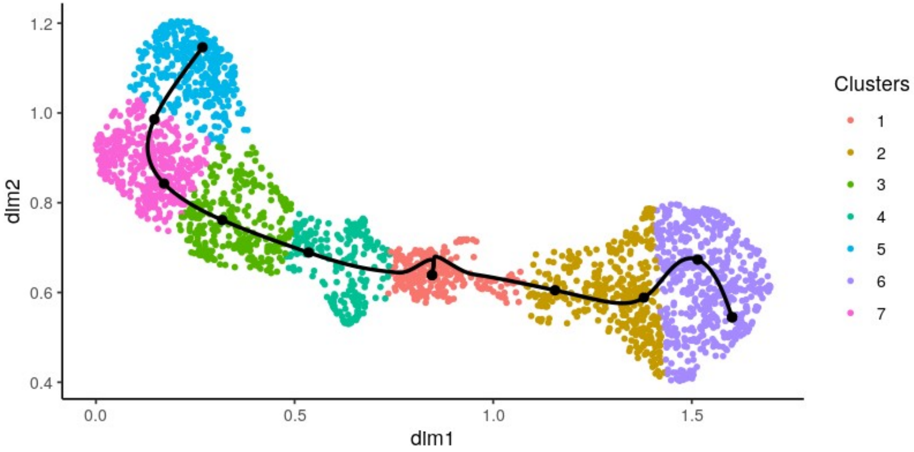
- Automatically identifies lineage progression
  - Some can predict multiple branching points
- Identifies “clusters” as pseduo-time along lineage
  - Tracks gene expression changes along axis of lineage
- Differential expression tests



Van den Berge K, Nature Communications, 2020

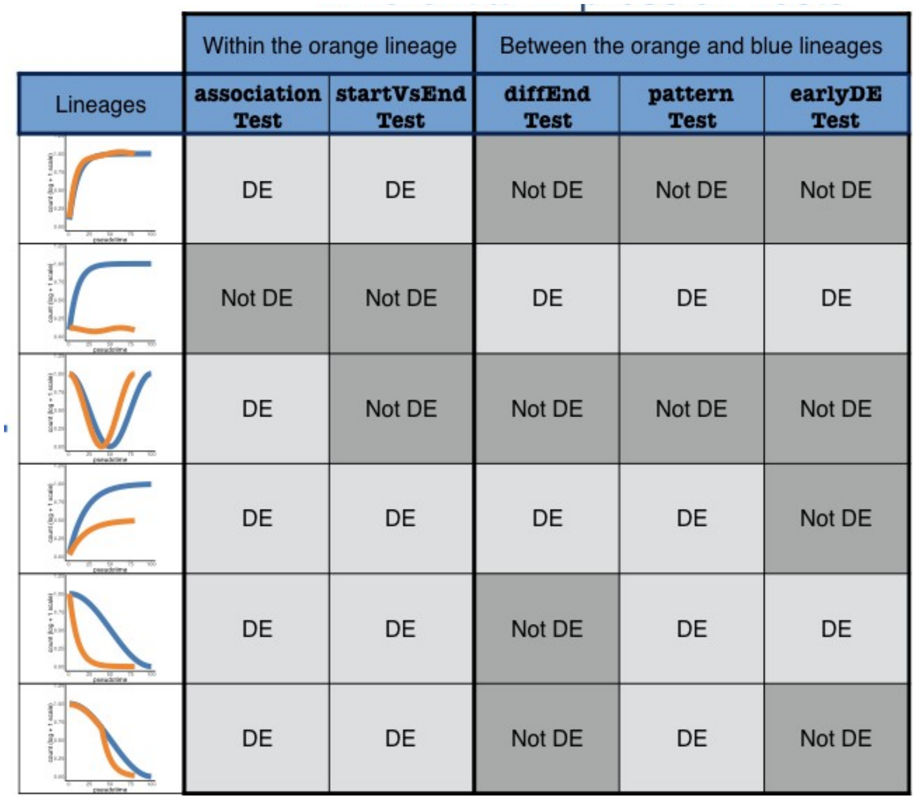
Algorithmic lineage analysis

- Automatically identifies lineage progression
  - Some can predict multiple branching points
- Identifies "clusters" as pseduo-time along lineage
  - Tracks gene expression changes along axis of lineage
- Differential expression tests



Algorithmic lineage analysis

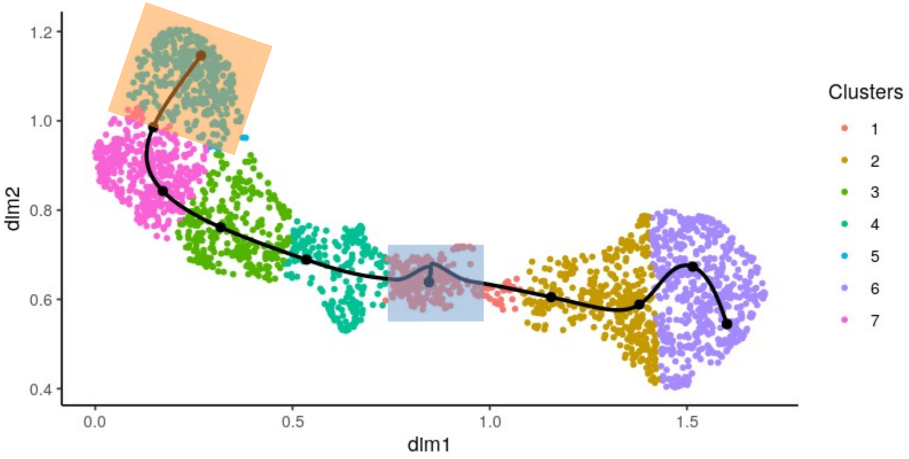
- Automatically identifies lineage progression
  - Some can predict multiple branching points
- Identifies “clusters” as pseduo-time along lineage
  - Tracks gene expression changes along axis of lineage
- Differential expression tests

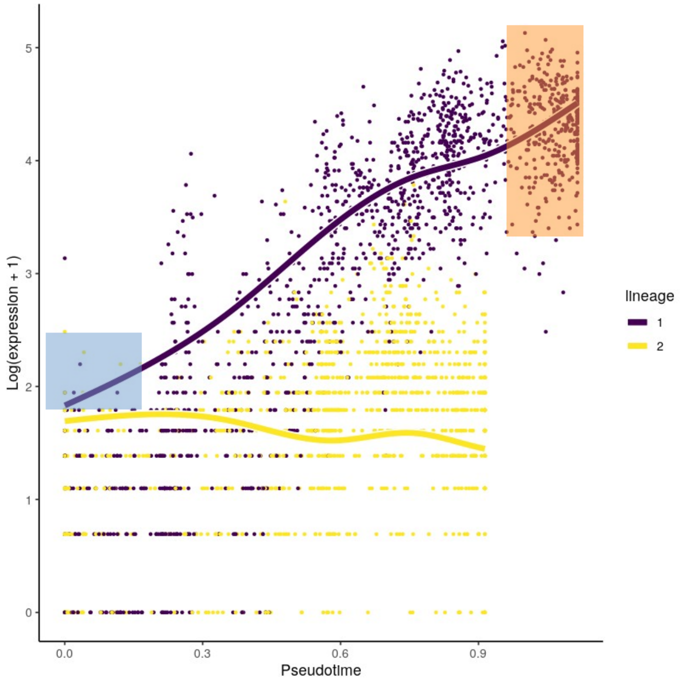


	Within the orange lineage		Between the orange and blue lineages		
Lineages	**associationTest**	**startVsEndTest**	**diffEndTest**	**patternTest**	**earlyDETest**
	DE	DE	Not DE	Not DE	Not DE
	Not DE	Not DE	DE	DE	DE
	DE	Not DE	Not DE	Not DE	Not DE
	DE	DE	DE	DE	Not DE
	DE	DE	Not DE	DE	DE
	DE	DE	Not DE	DE	Not DE

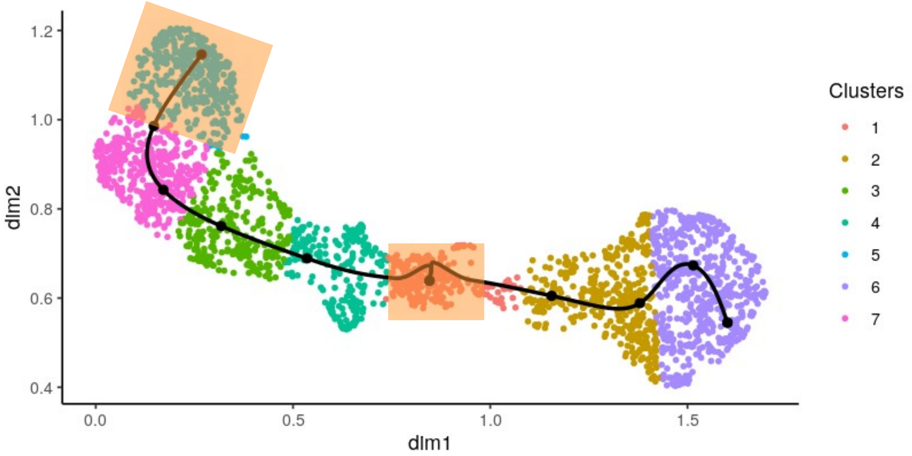
Van den Berge K, Nature Communications, 2020

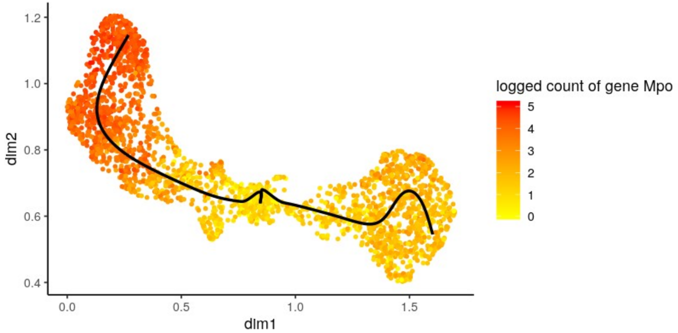
Markers for lineage progression



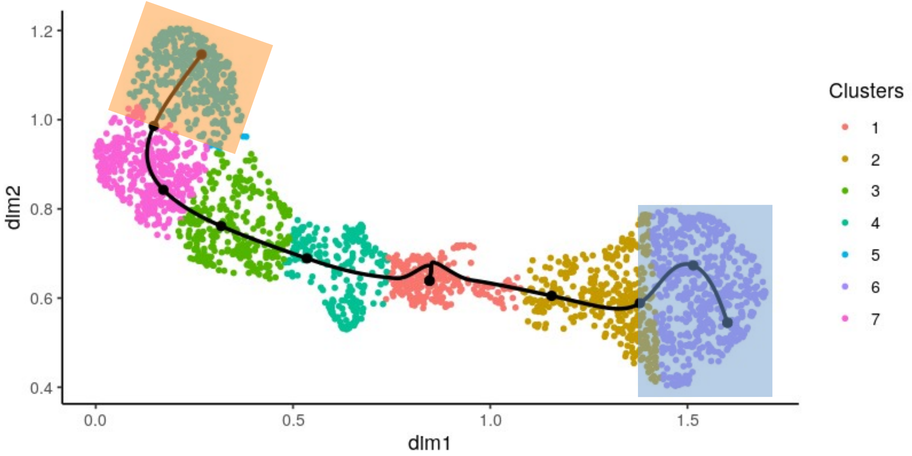


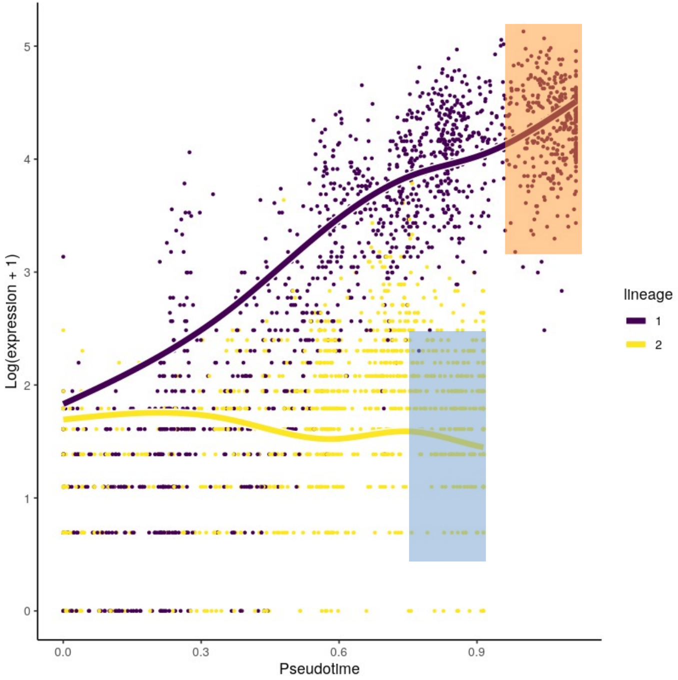
Markers for lineage progression



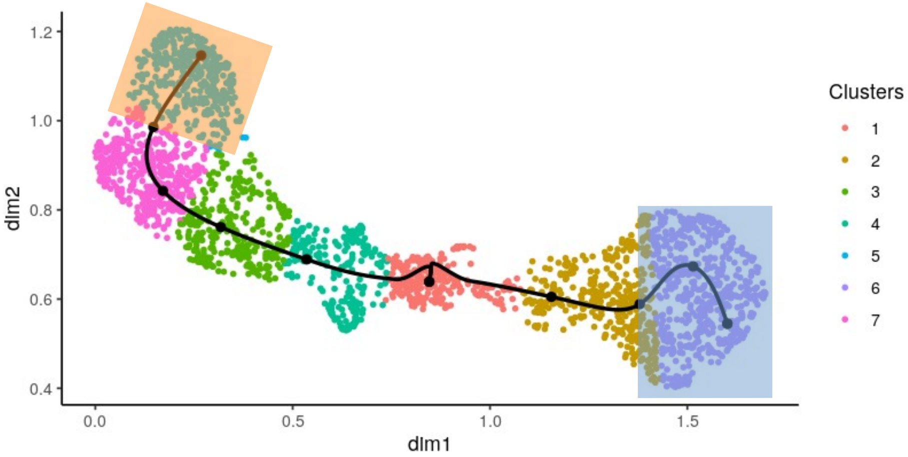


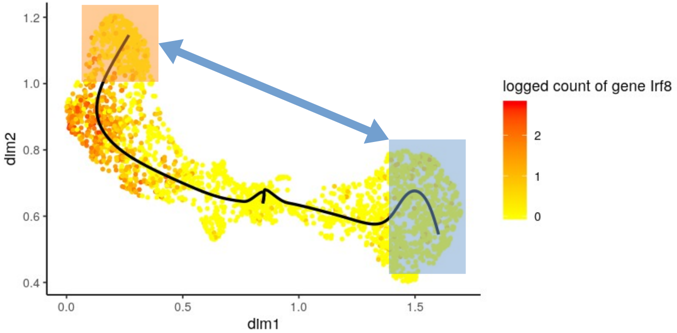
Markers between lineages





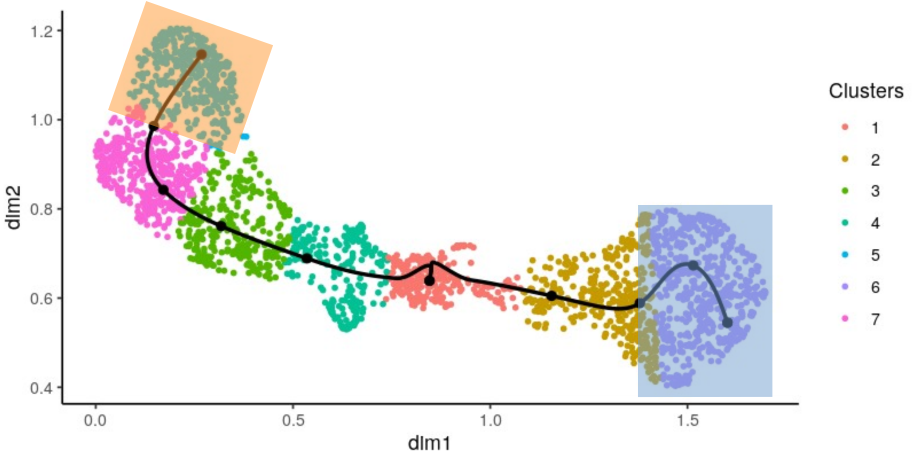
Marker as a pattern for lineage progression

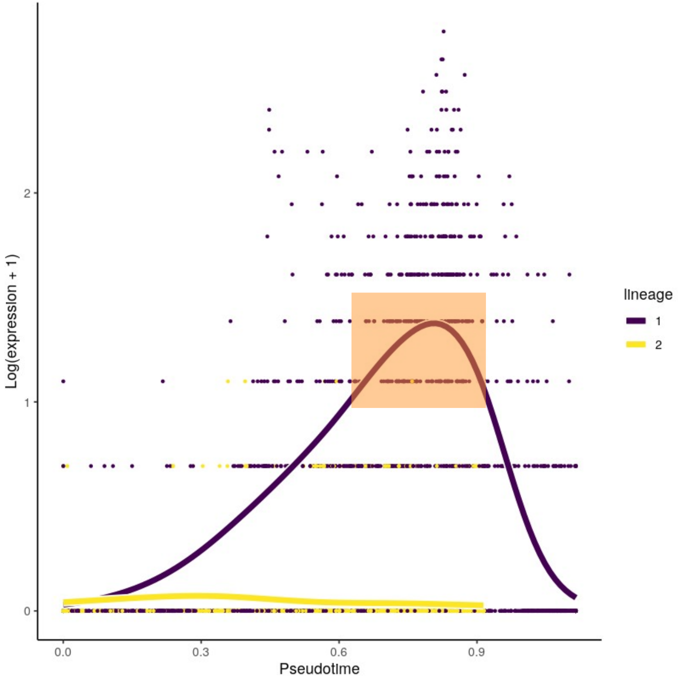




Expression at end of differentiation is not statistically significant.

Marker as a pattern for lineage progression





While end result is the same, there is a transient expression profile change to differentiation.

Summary

- Normalization
  - Concerns in scRNASeq not present in bulk RNASeq
- Visualization
  - Careful not to read into “clusters” too deeply
- Identifying populations of cell types
  - Clustering
  - Identification of cluster’s cell type
    - With *a priori* biological knowledge
    - Automatically with curated databases
  - Determination of cell type markers

- Differential expression
  - Between cell populations
  - Between conditions
    - Stimulated versus control
    - Time
  - For lineage / differentiation
    - Identification of lineage
    - Differential gene expression between states
    - Identification of transient expression profile changes