NORTH TEXAS
C{}DE-A-TH{}N

All in one scRNA-seq Pipeline:

Data downloading to analysis

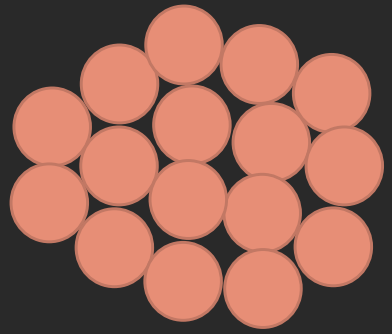
Kaitlyn Saunders, Alexa Salsbury, Yan Fang, and Edmund Miller

Overview

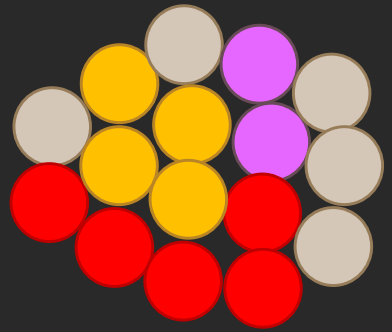
Background: scRNA seq is powerful tool to get highly dimensional data which bulk seq cannot provide

Problem: Multiple analysis tools required and data format not compatible; Intensive coding required; Biologist unfriendly

Solution: Build up an all-in-one automatic scRNA seq analysis pipeline, from data downloading to analysis visualization



Bulk RNA-seq detects the mRNA content across all cells in the sample.

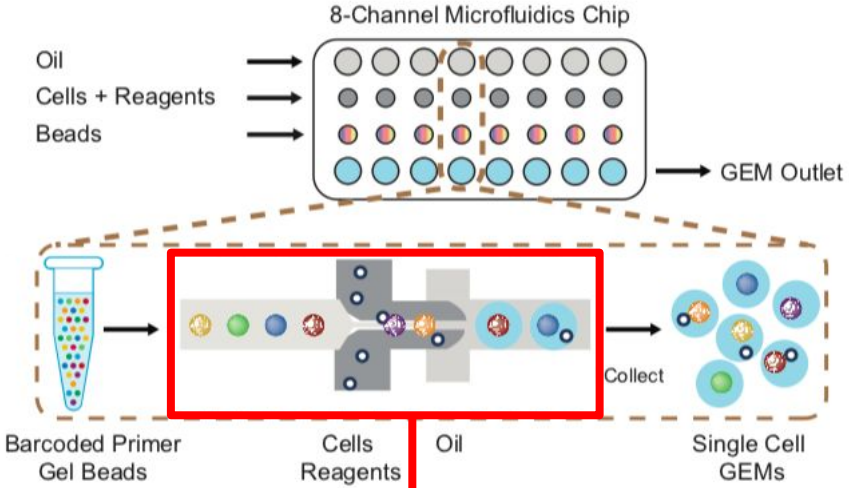


sc-RNA seq detects the mRNA content of each individual cell in the sample

Background on scRNA-seq

- Single-cell RNA sequencing (scRNA-seq) detects and quantifies individual cell mRNA content
- Looking at the whole tissue without accounting for cellular heterogeneity hide important cell-specific differences that can affect cell type and cell function.

Drop seq Process



What data looks like

Cell barcode – sample

	AAACCTGCAGTATCTG-1_healthy1_cd45+	AAACCTGTCGGCGCTA-1_healthy1_cd45+
RP11.34P13.3	0	0
FAM138A	0	0
OR4F5	0	0
RP11.34P13.7	0	0
RP11.34P13.8	0	0
RP11.34P13.14	0	0
RP11.34P13.9	0	0

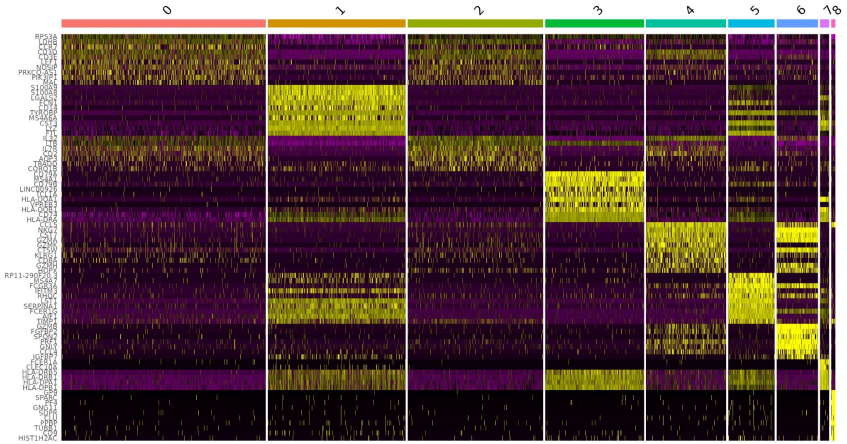
Gene names

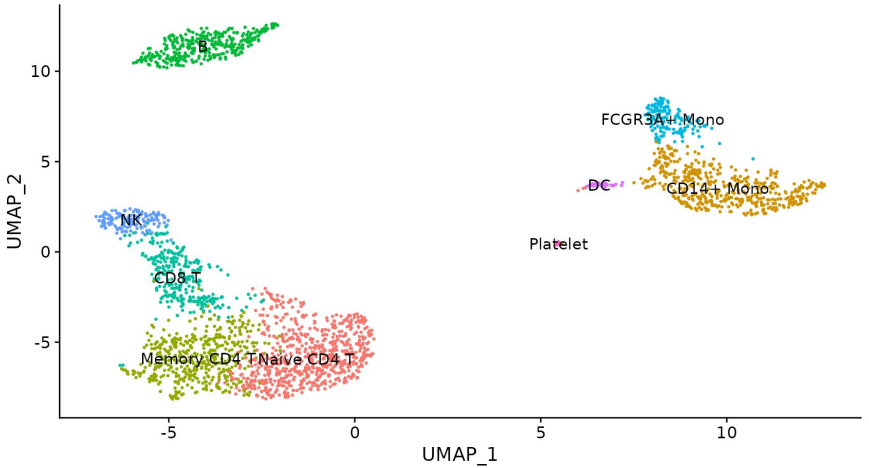
Number of RNA molecules for each gene

What information we can get from scRNAseq dataset?

- Identify (new/rare) cell types
- Find differential expressed genes after certain treatment
- Cell fate and differential direction

Clustering and scSorter

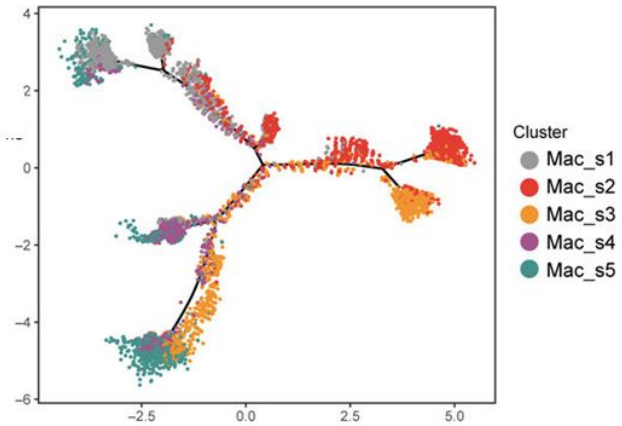




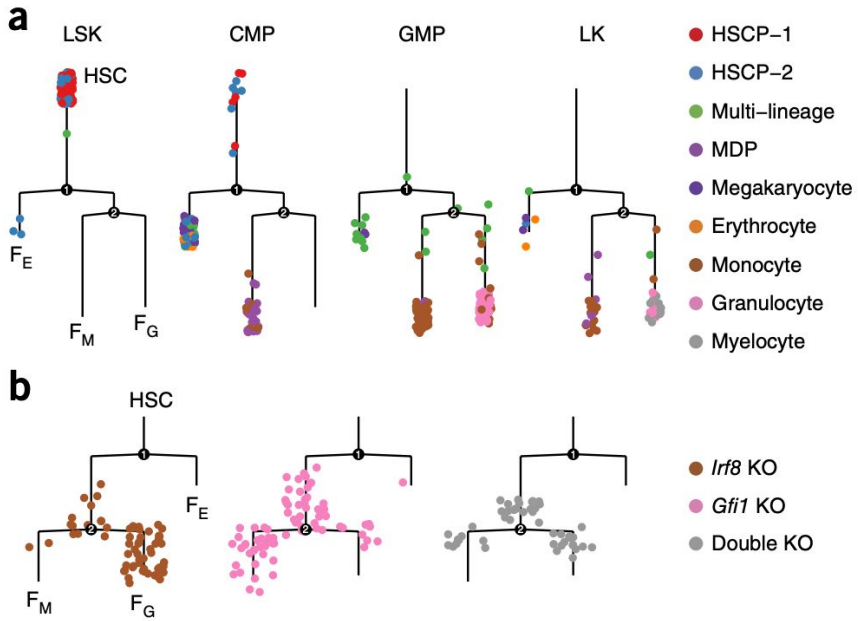
https://satijalab.org/seurat/

Pseudotime and Monocle3

- Machine learning
- Learn the sequence of gene expression changes
- Buildup overall "trajectory" of gene expression changes
- Setup a root and assign pseudotime



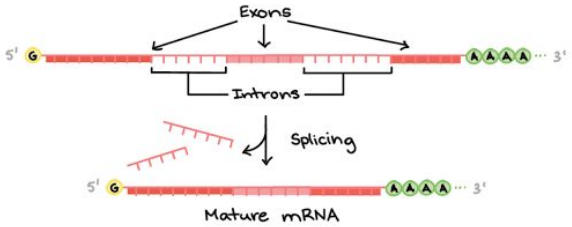
Differentiating blood cells



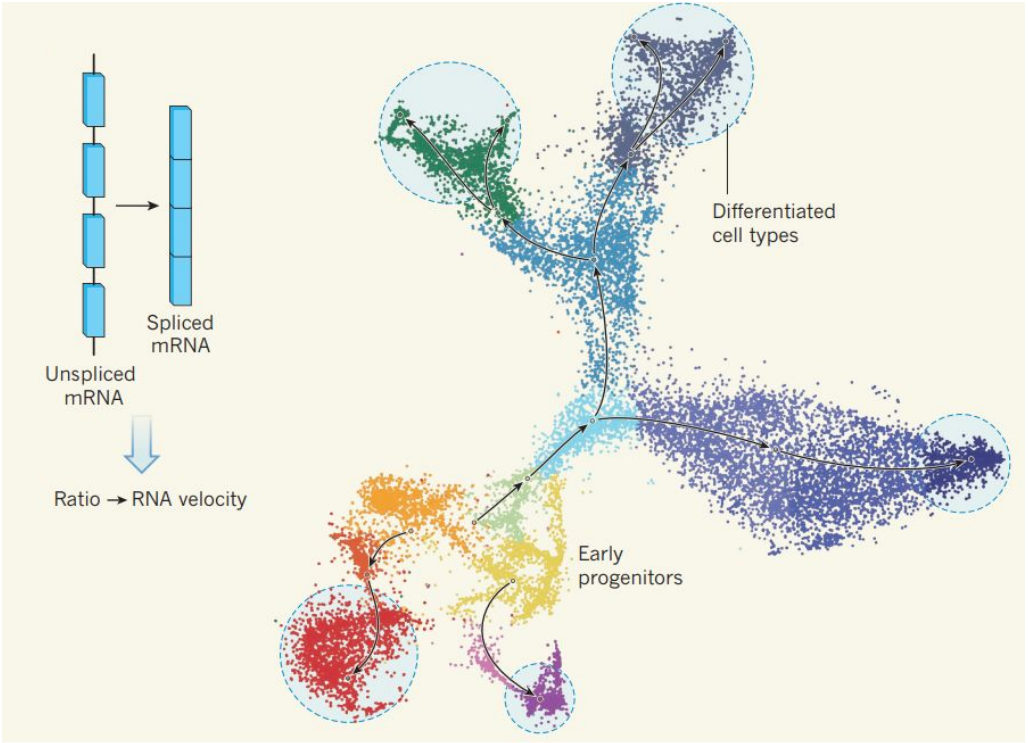
Qiu, X., *et al*. *Nat Methods* 14, 979–982 (2017).

RNA velocity

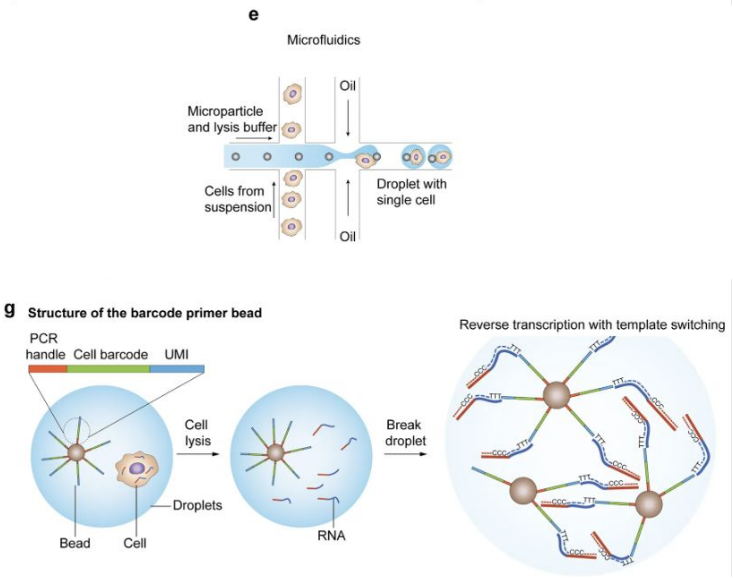
- Uses the ratio of unspliced to spliced mRNA transcripts to predict which cells other cells will become similar to in the future



- Unspliced mRNA decays quickly.
- Stably expressed genes will always have a small fraction of unspliced mRNA as it will continuously produce the mature spliced mRNA, and by extension, the unspliced form as well.



Multiple analysis tools required and data format not compatible

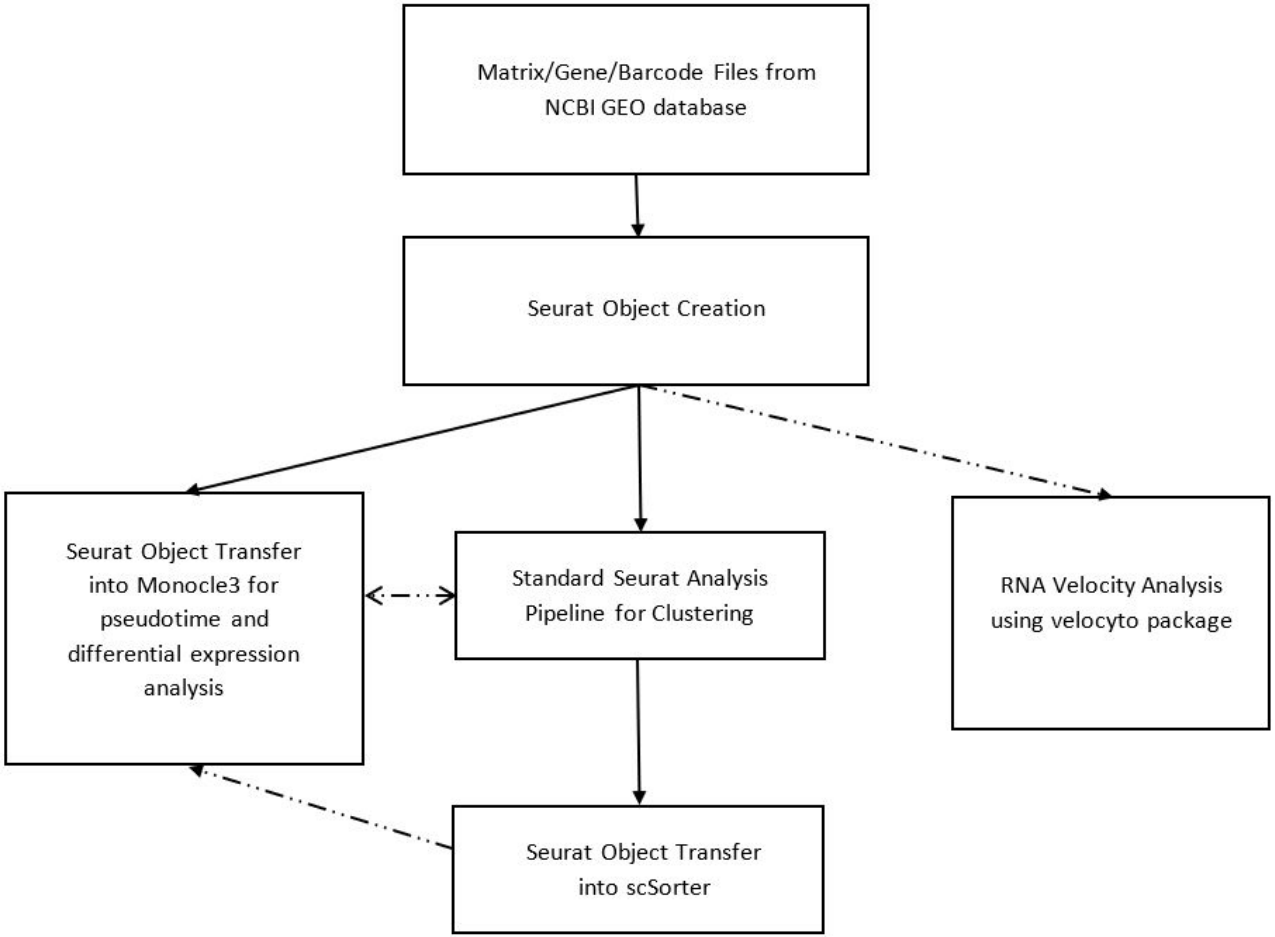


Hwang, B. *Experimental & Molecular Medicine* 2018

1. Data format clean up
   a. For the SRA data, we need to start with reads alignment with cellranger
   b. For the GEO dataset, we need to merge samples into a single seurat dataset.

   → (Cell ranger), GEOquery, R, Seurat
2. QC
3. Data normalization and Scale
4. Dimension reduction
5. Clustering
6. Differential expressed genes in each cell cluster

   → Seurat
7. Annotate each cell types → scSorter
8. Cell differentiation direction with trajectory analysis → Monocle3
9. RNA velocity → velocyto.R


Matrix/Gene/Barcode Files from NCBI GEO database
Seurat Object Creation
Seurat Object Transfer into Monocle3 for pseudotime and differential expression analysis
Standard Seurat Analysis Pipeline for Clustering
RNA Velocity Analysis using velocyto package
Seurat Object Transfer into scSorter

Demo

Real data: Data Downloading and Clustering

NCBI

Gene Expression Omnibus

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

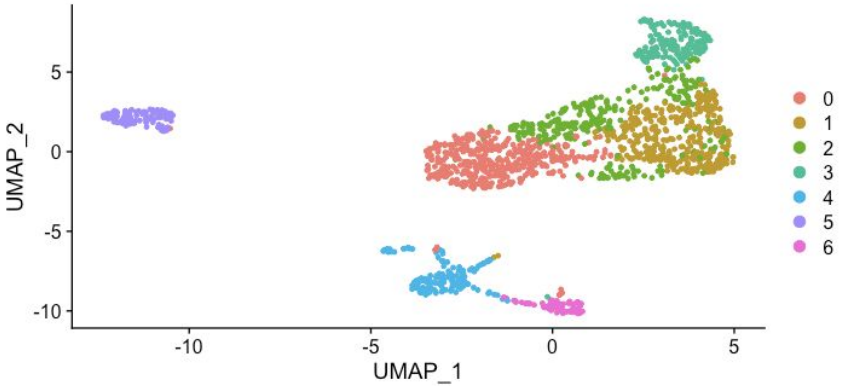
NCBI > GEO > **Accession Display** Not logged in | Login

GSE92332_RAW.tar	322.8 Mb	(http) (custom)	TAR (of MTX, TSV)

Series GSE92332 Query DataSets for GSE92332

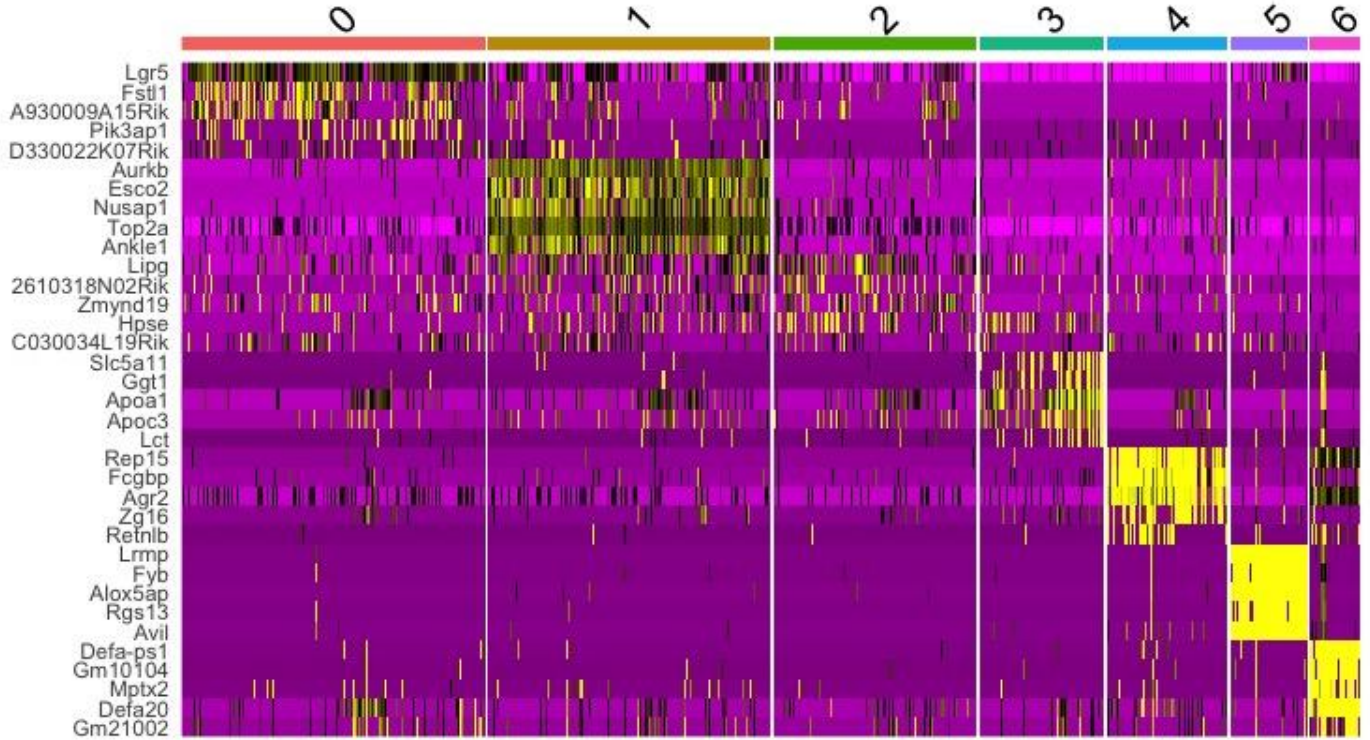
Status	Public on Nov 06, 2017
Title	A single-cell survey of the small intestinal epithelium
Organism	Mus musculus
Experiment type	Expression profiling by high throughput sequencing
Summary	To understand the diversity of cellular states within mouse intestinal epithelial tissue, we obtained whole intestines from wild type mice, dissagregated the samples, sorted into single cells and profiled them by single-cell RNA-seq.
Overall design	To understand normal tissue homeostasis, untreated cells were profiled, while to investiage the impact of enteric pathogens on epithelial cells, mice infected with both Salmonella Enterica and the parasitic worm H. polygyrus were profiled using both 3'-droplet-based and full length plate-based single-cell RNAseq. RANKL-mediated differentiation of organoid cultures were used to sequence the rare intestinal M cell, these were then compared to M cells obtained from the in vivo follicle associated epithelium (FAE). Please note that the FAE_UMIcounts.txt file has been updated on May 7th 2020.
Contributor(s)	Haber AL, Biton M, Rogel N, Herbst RH, Shekhar K, Smillie C, Burgin G, Delorey TM, Howitt MR, Katz Y, Tirosh I, Beyaz S, Dionne D, Zhang M, Raychowdhury R, Garret WS, Shi HN, Rozenblatt-Rosen O, Yilmaz O, Xavier R, Regev A

Data Loading and Clustering

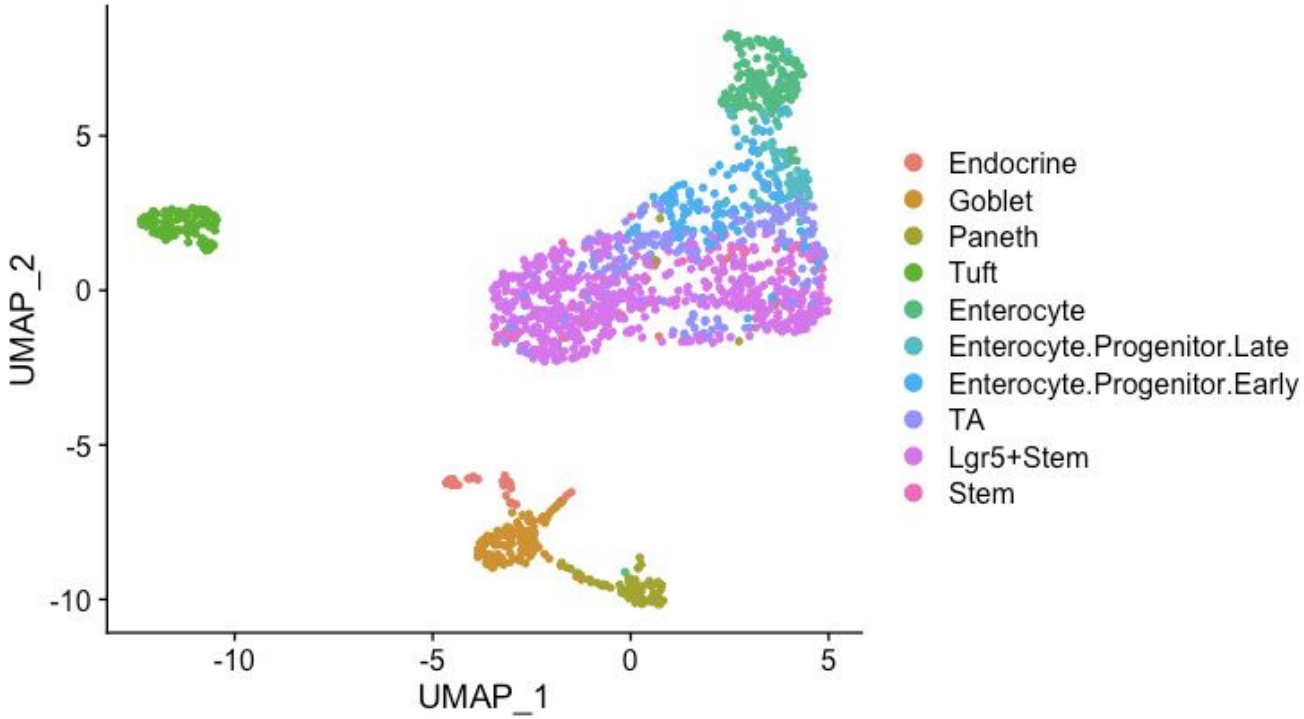


1. Data format clean up
2. Data normalization and Scale
3. Dimension reduction
4. Clustering

Real data: differentially expressed genes

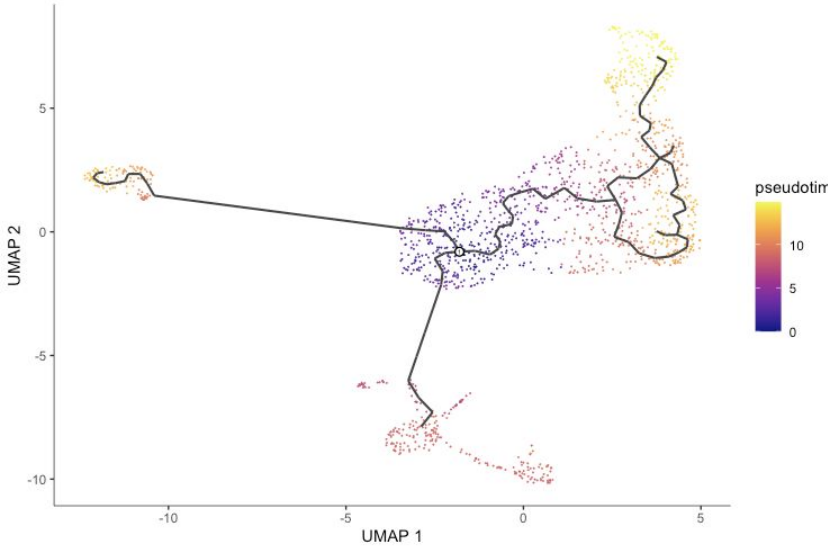


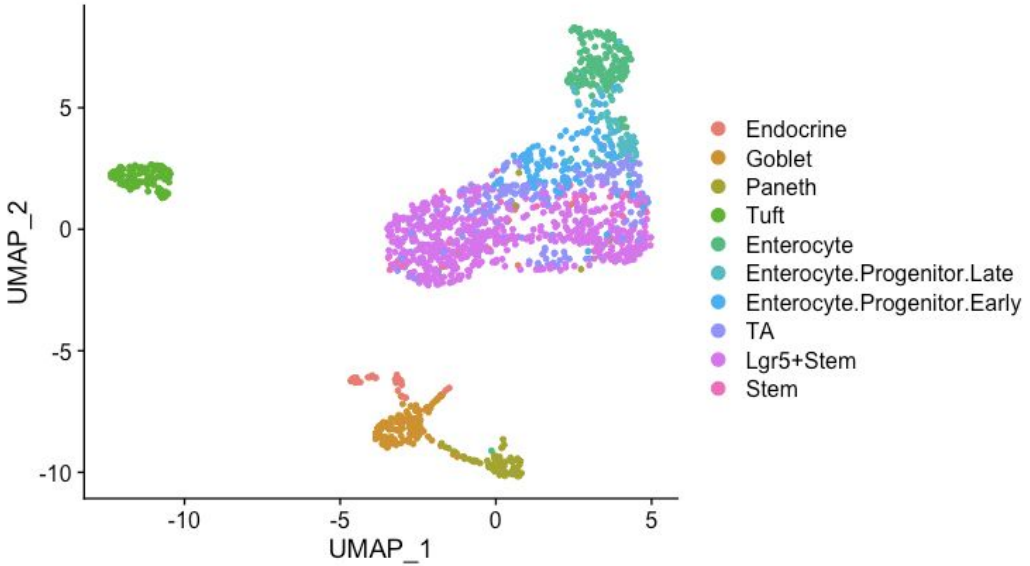
Real data: Annotate each cell types



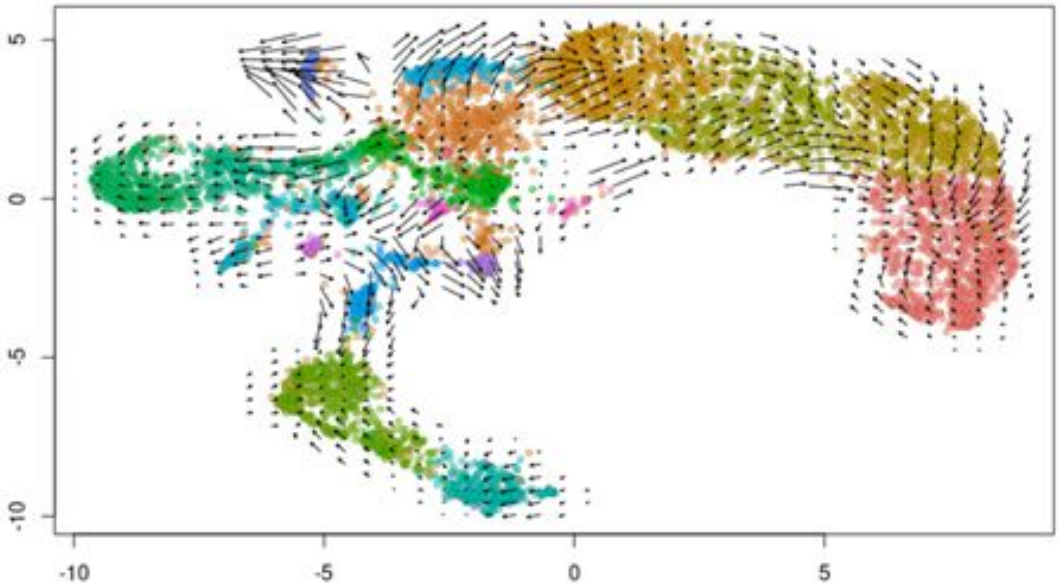
Real data: Pseudotime

Monocle3: Pseudotime assign





Real data: RNA velocity



Future Directions

- Automate rest of code using nf-core template
- Convert outputs of Monocle3 to scSorter
- Interconvert outputs of Monocle3 and Seurat, such that the pseudotime plot can be overlaid on top of the Seurat UMAP, and the like
- Convert outputs of Monocle3 and Seurat to the RNA velocity pipeline

Thanks :D