Single Cell Analysis "Under the Hood" Workshop
Session I

Organized by the NIH Single Cell Users Group

March 6th, 2019

NIH Single Cell Users Group

- This workshop is organized by

  Assaf Magen

  Stefan Cordes

  Mike Kelly

  Jamie Diemer

  Abdalla Abdelmaksoud

- Looking for volunteers for building an improved analysis pipeline

  Contact Assaf Magen or Stefan Cordes [assaf.magen@nih.gov; stefan.cordes@nih.gov]

Introduction

- Single-cell RNA sequencing (scRNAseq) revolution
- Require considerable computational analysis

- Broad objectives:
  - Beginners – how to make the first steps
  - Advanced – how to leverage the technology better
  - Consumers – what is being done and what to be careful of

Introduction

- Objectives
  - Discuss what can be done using scRNAseq
  - Building blocks of conventional analysis
  - Limitations of computational approaches

- This workshop will not:
  - Discuss or advocate specific pipelines
  - Make you a single-cell expert

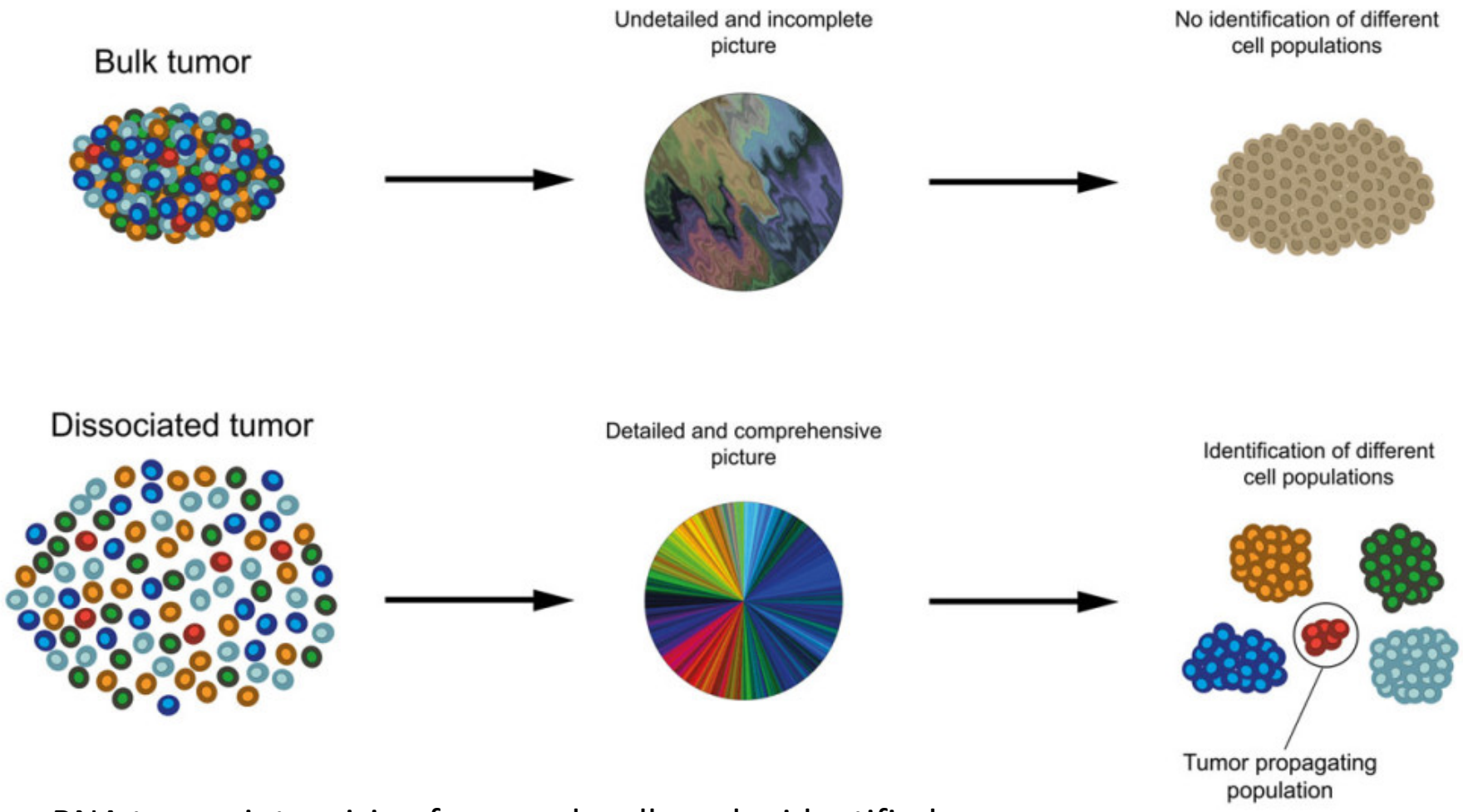
- Questions will be moderated by Jamie.
- Slides will be available online after the workshop.

Using Single Cell RNA-Seq to Study Heterogeneity

- What do we mean by cell heterogeneity
- Discrete versus continuous cell types and states
- Very brief concept of data generation

Mike Kelly

Why single cell RNA-Seq?

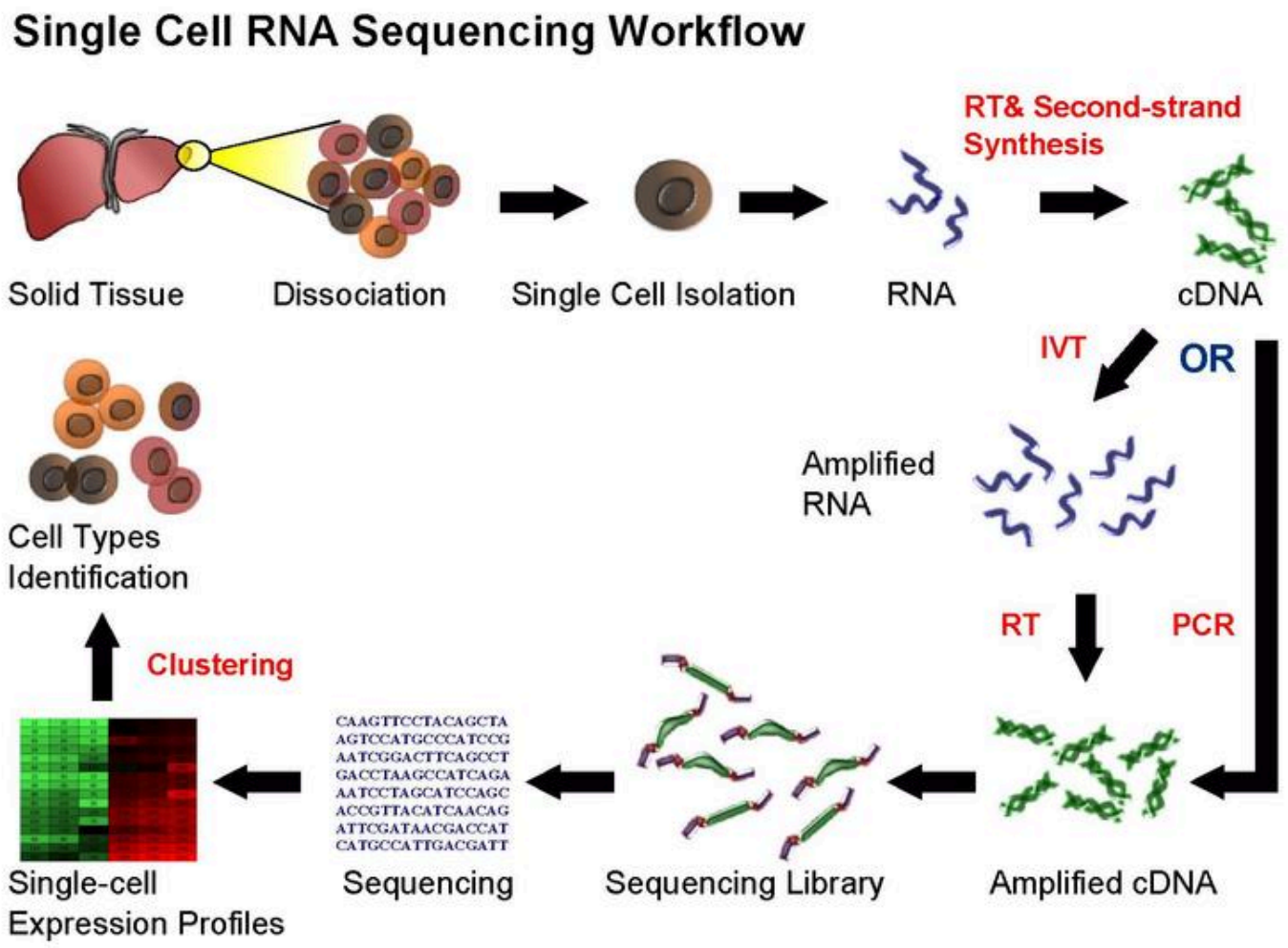


This is for discrete cell type classifications. Also need to add in use of single cell for capturing continuous state chances through snapshot(s) of asynchronous cells

mRNA transcripts arising from each cell can be identified by cell-specific barcodes that are added

Diagram from: http://fightdipg.org/research-projects/

Generalized workflow of generating single cell RNA-Seq data

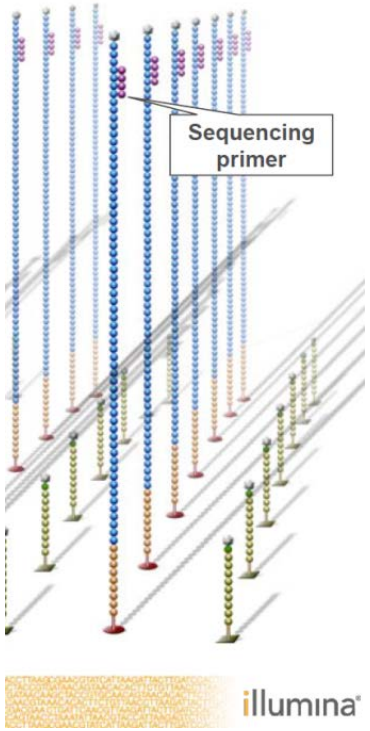


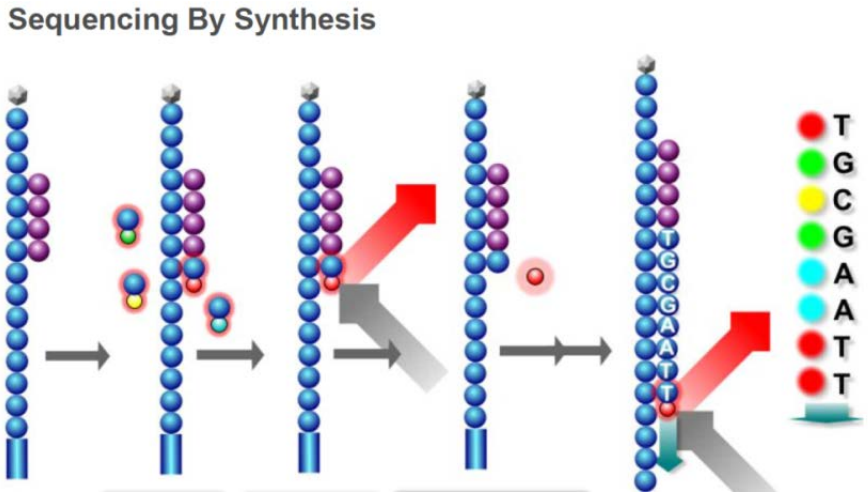
- Partition single cells
- Convert mRNA into cDNA
- Amplify cDNA
- Generate sequencing library
- Sequence
- Data analysis with identification of what transcripts are expressed by each cell profiled

https://hemberg-lab.github.io/scRNA.seq.course/

From cDNA library to millions of sequencing reads

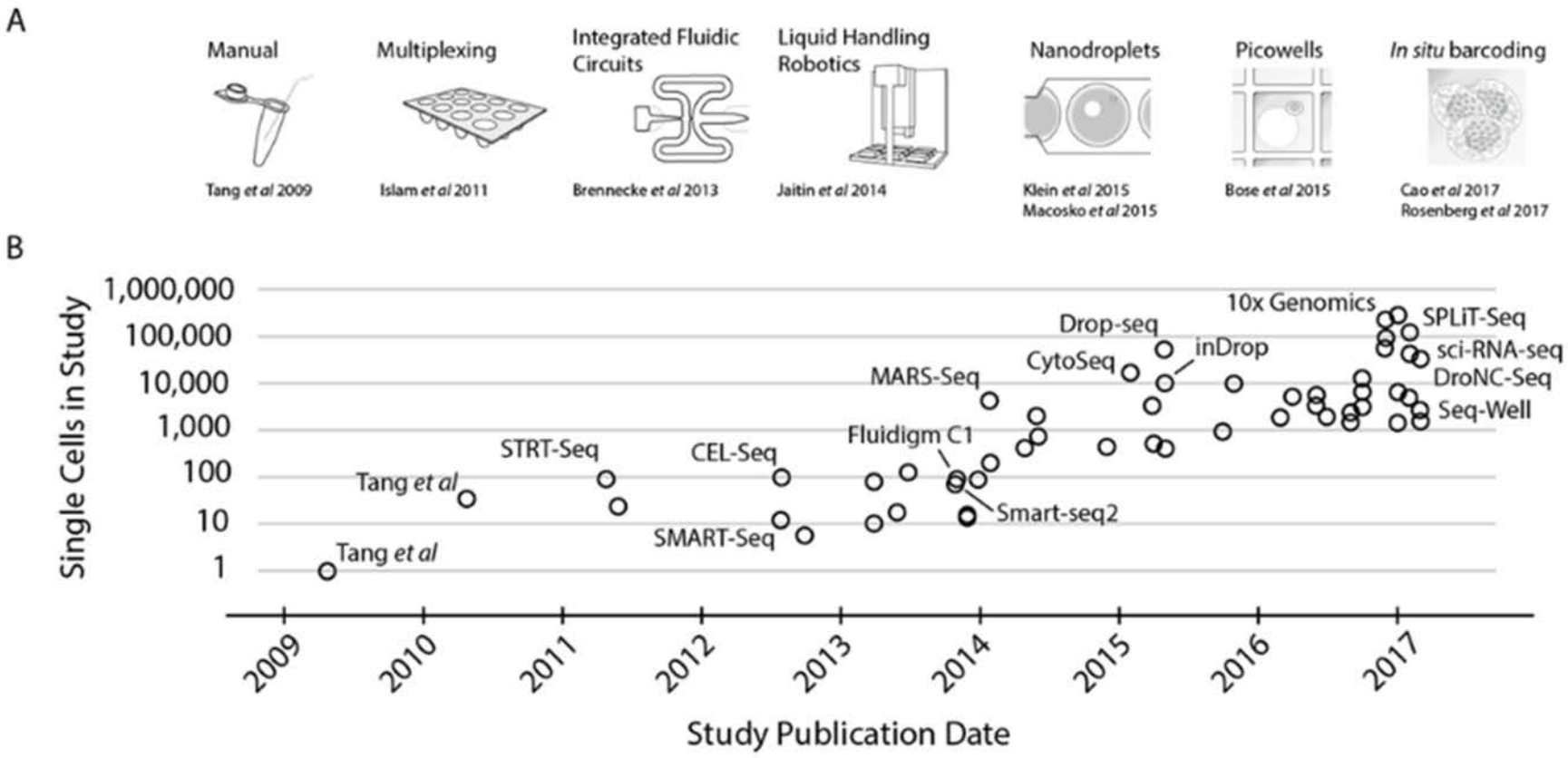
Illumina Next-Generation Sequencing (NGS) “Sequencing by Synthesis”





- Sequence read by fluorescent nucleotide incorporation during each “cycle”
- Each cluster dot will display a color associated with nucleotide (A, C, G, or T)
- Image processing -> conversion to Fastq output (sequence with quality score)

Single cell RNA-Seq has evolved quickly from lower throughput to higher throughput methods



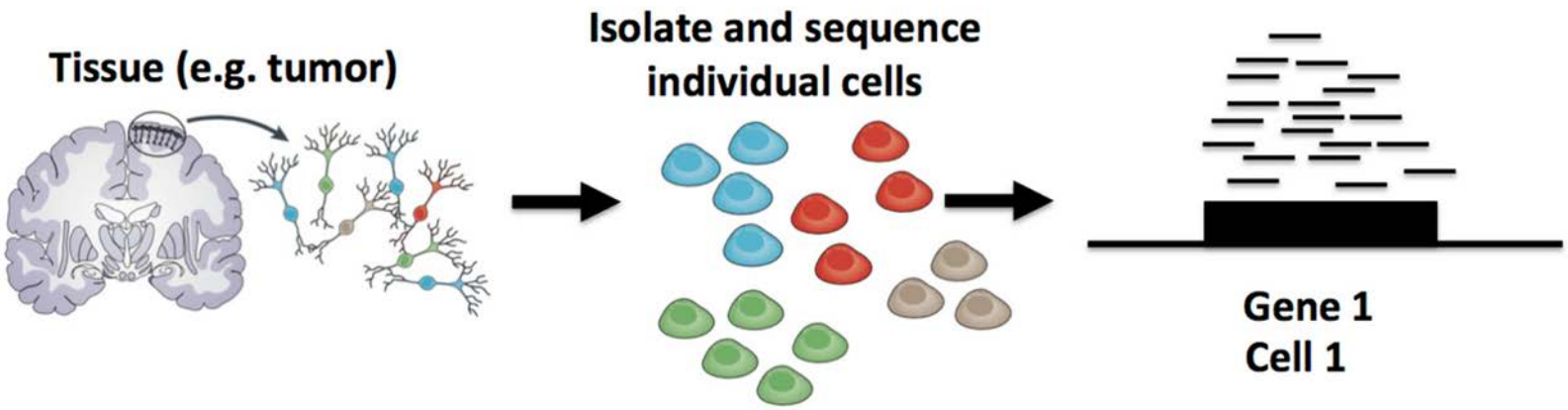
- First single cell whole transcriptome single cell RNA-Seq used manually picking of cells (2009)
- More widely adopted in 2012/2013 with Fluidigm C1 platform and SMARTer chemistry
- Huge increase in throughput with droplet based methods in 2015 (Drop-Seq / InDrops)
- Third generation of methods may see additional increase in throughput / decrease in cost (sciRNA-Seq / SPLiT-Seq / Seq-Well) ~2017/2018

Challenges in processing scRNA-Seq data

- What does this data look like
- Basic programming needed to interpret data
- How to get from highly multidimensional data to human interpretable format

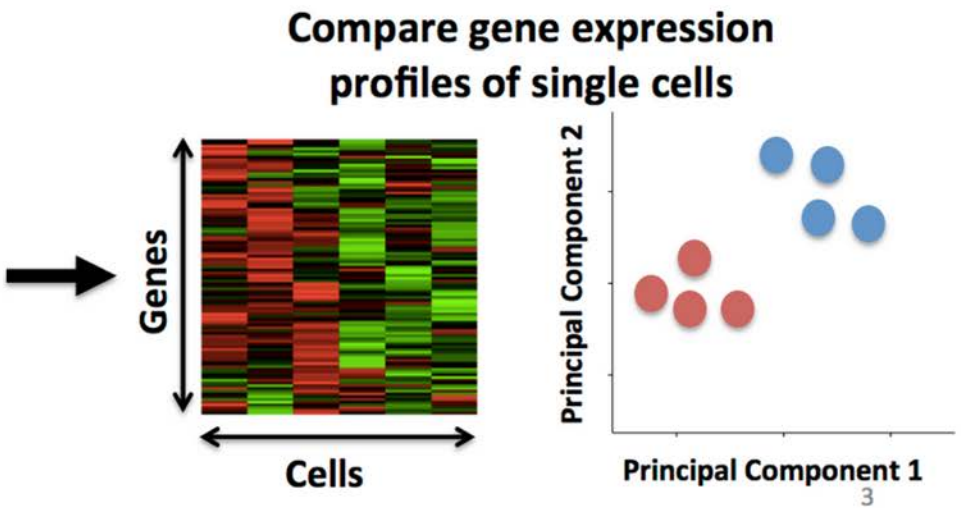
Jamie Diemer

Challenges in processing scRNA-Seq data



Read Counts

1	0	0	2
0	3	1	0
1	0	0	0
0	4	0	0



https://learn.gencore.bio.nyu.edu/single-cell-rnaseq/

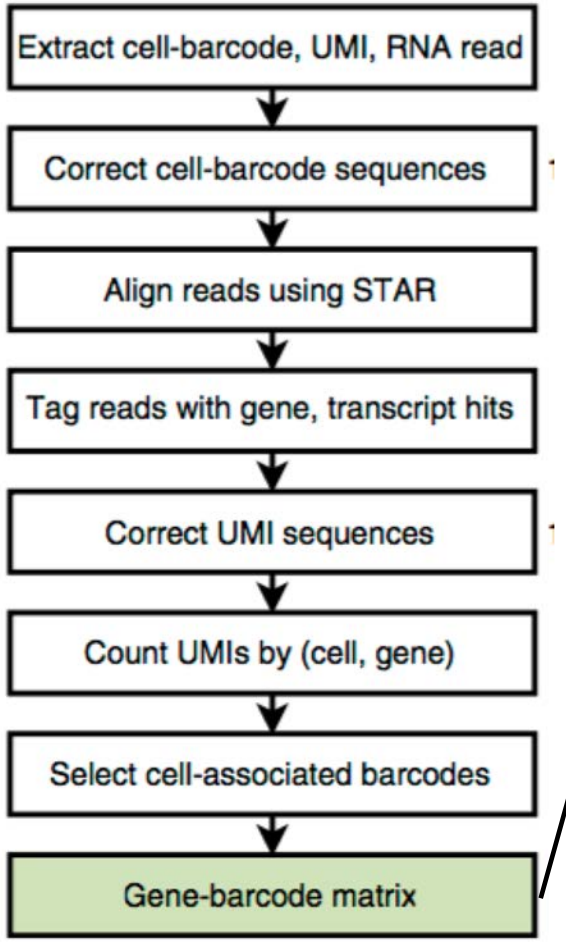
Challenges in processing scRNA-Seq data

Bench scientist

Bioinformatician

Analysis is very iterative

Cell Ranger (10X Genomics)



Seurat
Macosko et al., 2015

SC3
Kiselev et al., 2017

SINCERA
Guo et al., 2015

SNN-Cliq
C. Xu and Su 2015

<u>Many steps are the same</u>

- Filter
- Normalize
- Scale
- PCA/CCA analysis
- Clustering of cells/gene lists
- tSNE visualization

There are many R and R Studio resources

NIH Library

https://www.nihlibrary.nih.gov/training/calendar

Research Tools | Site Map - Site Guid... | AutoDraw | Submit a Service Re... | Cookbook for R | Gene - NCBI | NHGRI | Service Desk

Method

- In-Person Class
- Webinar

Location

- NIH Library Training Room
- NIH Library Media Room
- NIH Library Reading Room
- Masur Auditorium
- Lipsett Amphitheater
- Natcher Building (Building 45)
- Porter Neuroscience Center

Training Category

- Bibliometrics
- Bioinformatics
- BTRIS
- Data Services
- Databases & Searching
- Special Event
- Technology

March 2019 — CALENDAR | LIST

MON	TUE	WED	THU	FRI
25	26	27	28	1
4	5 Introduction to R and RStudio Intro to Systematic Reviews	6 BTRIS Basics	7 R Text Mining	8
11	12 Introduction to R Data Types Systematic Reviews: Types and Searching Publication Analysis	13 PubMed Basics	14 Openness and Reproducibility Workshop	15
18	19 Bibliometrics for Portfolio Analysis Screening & Managing Data for Systematic Reviews	20	21 Using EndNote for Systematic Reviews	22
25 Data Wrangling in R	26 Writing & Publishing Your Systematic Review	27	28 EndNote Desktop	29
1	2	3	4	5

https://www.rstudio.com/online-learning/#R

RStudio — Products | Resources | Pricing | About Us | Blogs

- R Programming
- Shiny
- R Markdown
- Data Science
- Books

Online learning

A wealth of tutorials, articles, and examples exist to help you learn R and its extensions. Scroll down or click a link below for a curated guide to learning R and its extensions.

R Programming — Read More ›

Shiny — Read More ›

R Markdown — Read More ›

Data Science — Read More ›


USA


USA

Safety first!

Be careful, re-run analyses to make sure they are reproducible, try different parameters, check with colleagues

There are MANY learning resources

https://hemberg-lab.github.io/scRNA.seq.course/index.html

Analysis of single cell RNA-seq data

Vladimir Kiselev (wikiselev), Tallulah Andrews, Jennifer Westoby (Jenni_Westoby), Davis McCarthy (davisjmcc), Maren Büttner (marenbuettner) and Martin Hemberg (m_hemberg)

2018-05-29

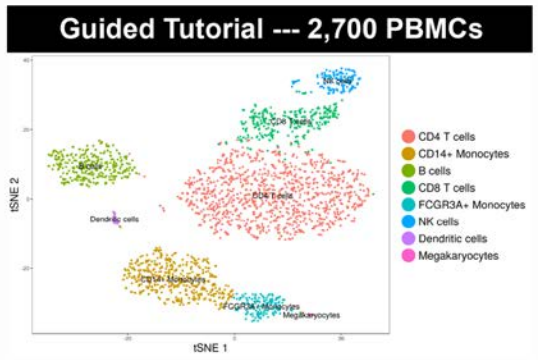
1 About the course

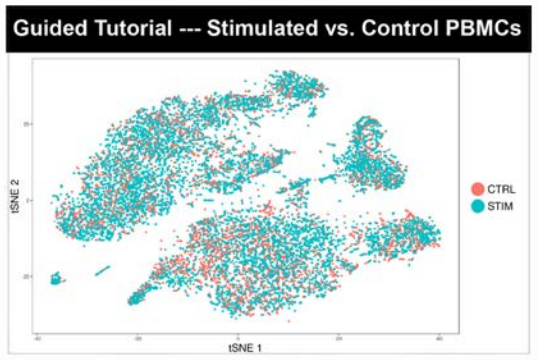
Today it is possible to obtain genome-wide transcriptome data from single cells using high-throughput sequencing (scRNA-seq). The main advantage of scRNA-seq is that the cellular resolution and the genome wide scope makes it possible to address issues that are intractable using other methods, e.g. bulk RNA-seq or single-cell RT-qPCR. However, to analyze scRNA-seq data, novel methods are required and some of the underlying assumptions for the methods developed for bulk RNA-seq experiments are no longer valid.

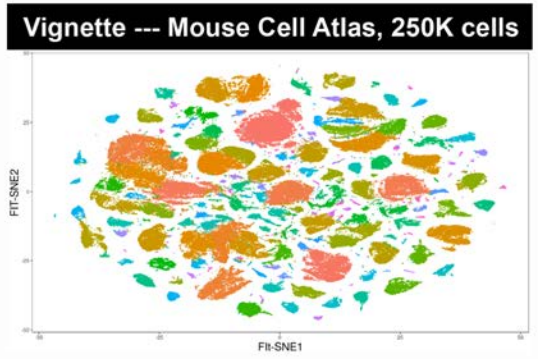
In this course we will discuss some of the questions that can be addressed using scRNA-seq as well as the available computational and statistical methods avialable. The course is taught through the University of Cambridge Bioinformatics training unit, but the material found on these pages is meant to be used for anyone interested in learning about computational analysis of scRNA-seq data. The course is taught twice per year and the material here is updated prior to each event.

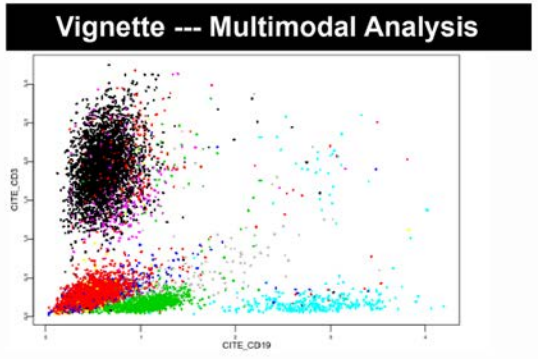
The number of computational tools is increasing rapidly and we are doing our best to keep up to date with what is available. One of the main constraints for this course is that we would like to use tools that

Seurat’s Guided tutorials

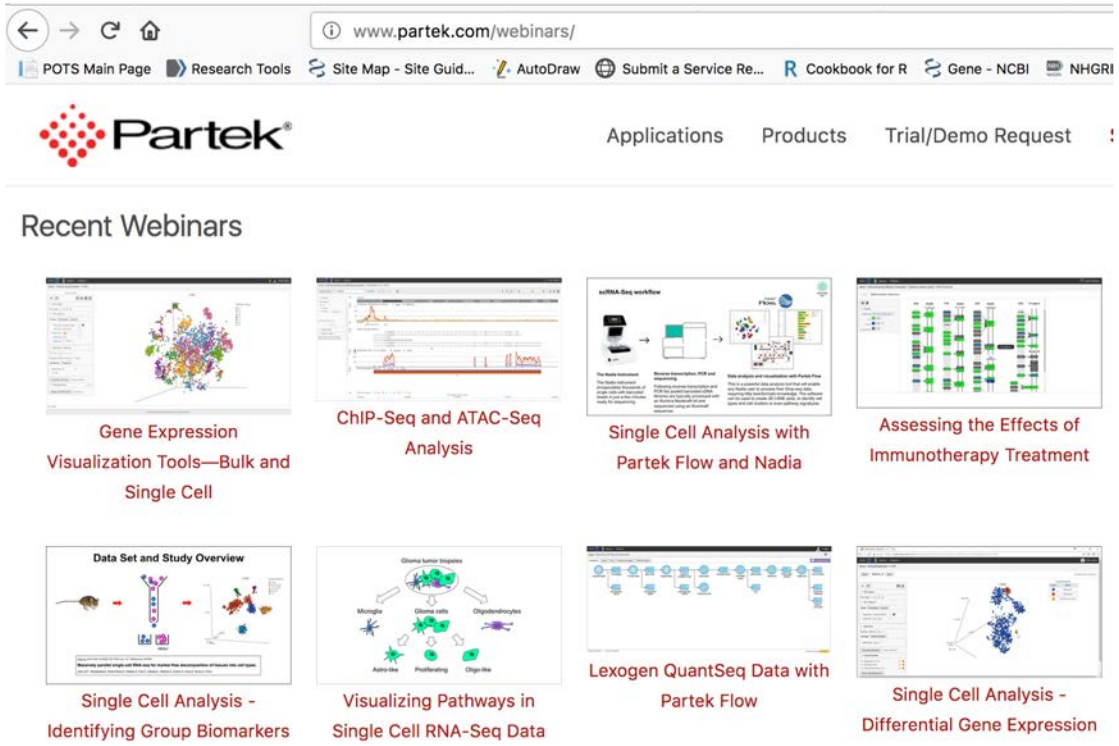








Partek’s webinar series



RStudio

WT2_PRE2NodePlaying.R | Ery_Subset_10ClustersTotal.R | WT2_PRE2.R

```r
library(Seurat)
library(Matrix)
library(dplyr)

setwd("~/Desktop/Liu Lab Science/All Experiments By Date_Not In Ketu/10XGenomics/04-19-18_ToBoostNumbers/Filtered Gene Matric
WT2.data <- Read10X (data.dir = "filtered_gene_bc_matrices/mm10")

setwd("~/Desktop/Liu Lab Science/All Experiments By Date_Not In Ketu/10XGenomics/04-19-18_ToBoostNumbers/Filtered Gene Matric
PRE2.data <- Read10X (data.dir = c("PRE2"= "filtered_gene_bc_matrices/mm10"))

#To prevent error from having exact cell barcodes in two condtions, add this
colnames(WT2.data) <- paste0("WT2_",colnames(WT2.data))
colnames(PRE2.data) <- paste0("PRE2_",colnames(PRE2.data))

# Set up object 1
WT2 <- CreateSeuratObject(raw.data = WT2.data, project = "WT2", min.cells = 5)
WT2@meta.data$WT2 <- "WT2"
WT2 <- FilterCells(WT2, subset.names = "nGene", low.thresholds = 500, high.thresholds = Inf)
WT2 <- NormalizeData(WT2)
WT2 <- ScaleData(WT2, display.progress = F)

# Set up object 2
PRE2 <- CreateSeuratObject(raw.data = PRE2.data, project = "PRE2", min.cells = 5)
PRE2@meta.data$WT2 <- "PRE2"
PRE2 <- FilterCells(PRE2, subset.names = "nGene", low.thresholds = 500, high.thresholds = Inf)
PRE2 <- NormalizeData(PRE2)
PRE2 <- ScaleData(PRE2, display.progress = F)
```

Console | Terminal

~/Desktop/Liu Lab Science/All Experiments By Date_Not In Ketu/10XGenomics/04-19-18_ToBoostNumbers/Filtered Gene Matrices/WT2_PRE2/AttemptsAtNodeCollapse

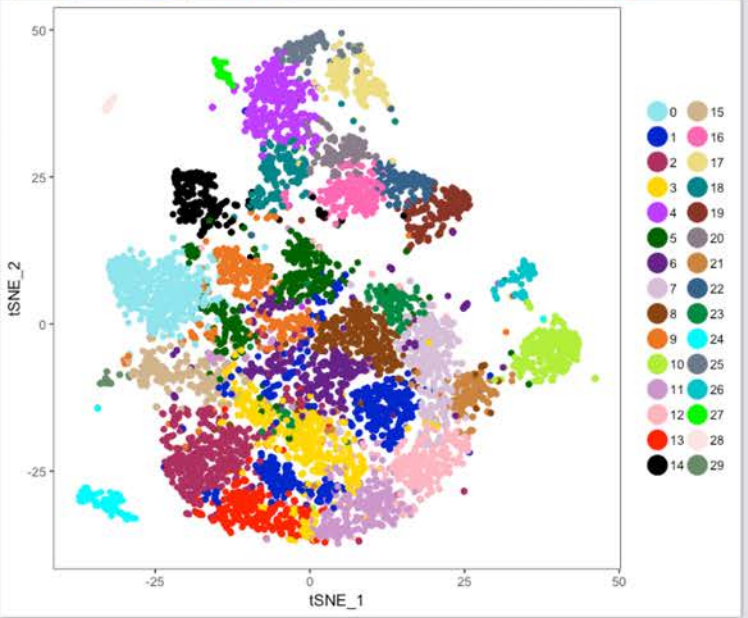
```
+                                                    "chocolate4",
+                                                    "chocolate2",
+                                                    "olivedrab2",
+                                                    "plum3",
+                                                    "lightpink",
+                                                    "red",
+                                                    "gray0",
+                                                    "tan",
+                                                    "hotpink",
+                                                    "lightgoldenrod",
+                                                    "turquoise4",
+                                                    "tomato4",
+                                                    "thistle4",
+                                                    "tan3",
+                                                    "steelblue4",
+                                                    "springgreen4",
+                                                    "cyan",
+                                                    "slategray4",
+                                                    "turquoise3",
+                                                    "green",
+                                                    "mistyrose",
+                                                    "darkseagreen4"))
> plot_grid(p1)
> plot_grid(p2)
>
```

Environment | History | Connections

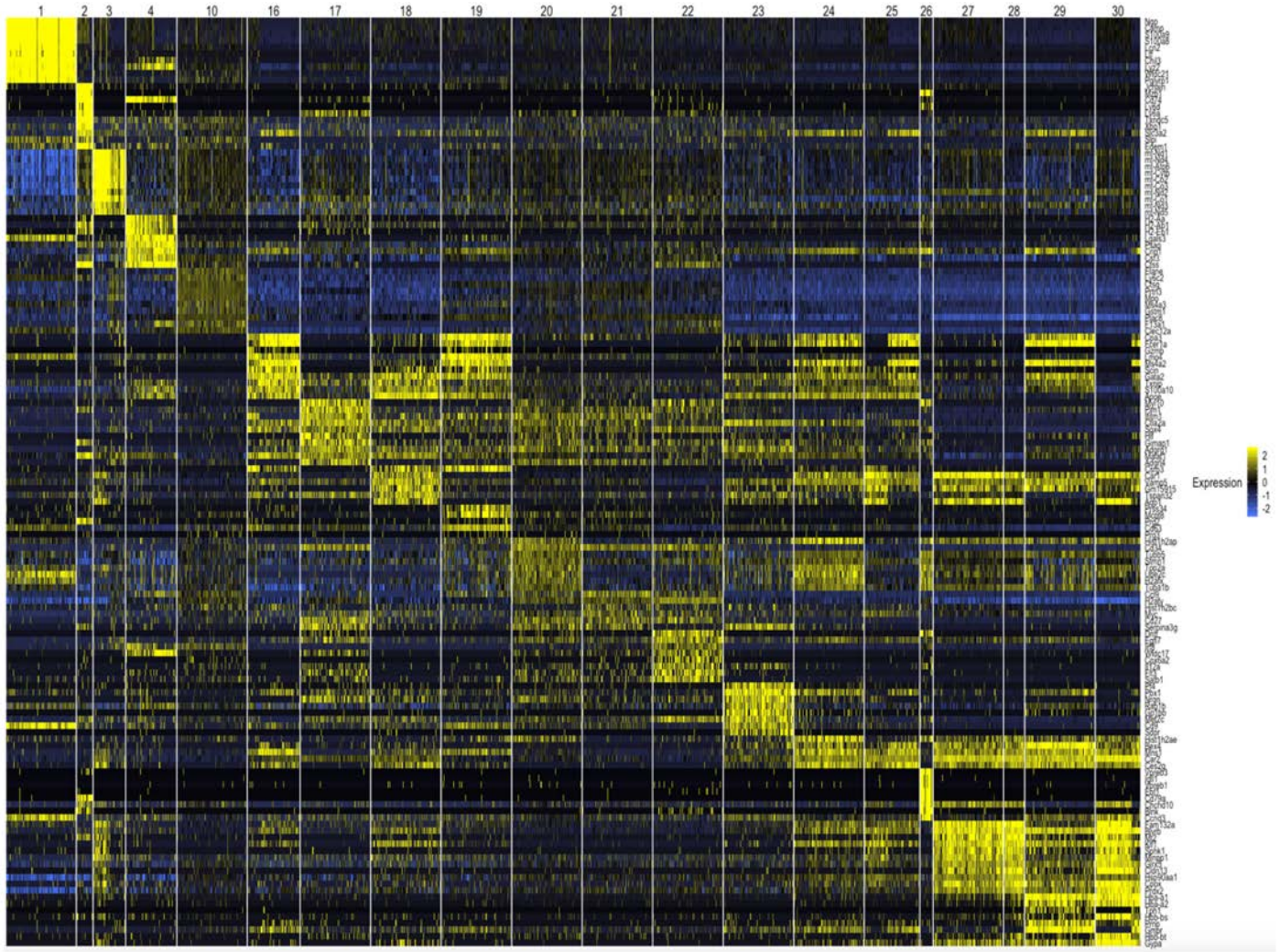
Global Environment

Data	
Ery.subset	Large seurat (559.4 Mb)
Merge.36	Large seurat (1.4 Gb)
p1	Large gg (10 elements, 1 Mb)
p2	Large gg (10 elements, 1 Mb)
WT2_PRE2.combined	Large seurat (1.4 Gb)
WT2_PRE2.markersNew	8946 obs. of 7 variables
Values	
features.plot	chr [1:200] "Ngp" "Camp" "S100a9" "S100a8" "Lcn2" "Ltf" "Chil3"…

Files | Plots | Packages | Help | Viewer



DoHeatmap (object = SubsetData(object = Merge.36, max.cells.per.ident = 200), genes.use = features.plot, slim.col.label = TRUE, group.label.rot = F, col.mid = "grey0", col.high = "yellow", col.low = "royalblue1", group.spacing = 0.10, group.label.loc = "top")



Why Is Preprocessing Important for Single Cell Analysis

- Clean out low quality information
- Separate biological data from artifacts
- Remove noise

Abdalla Abdelmaksoud

QC and Preprocessing

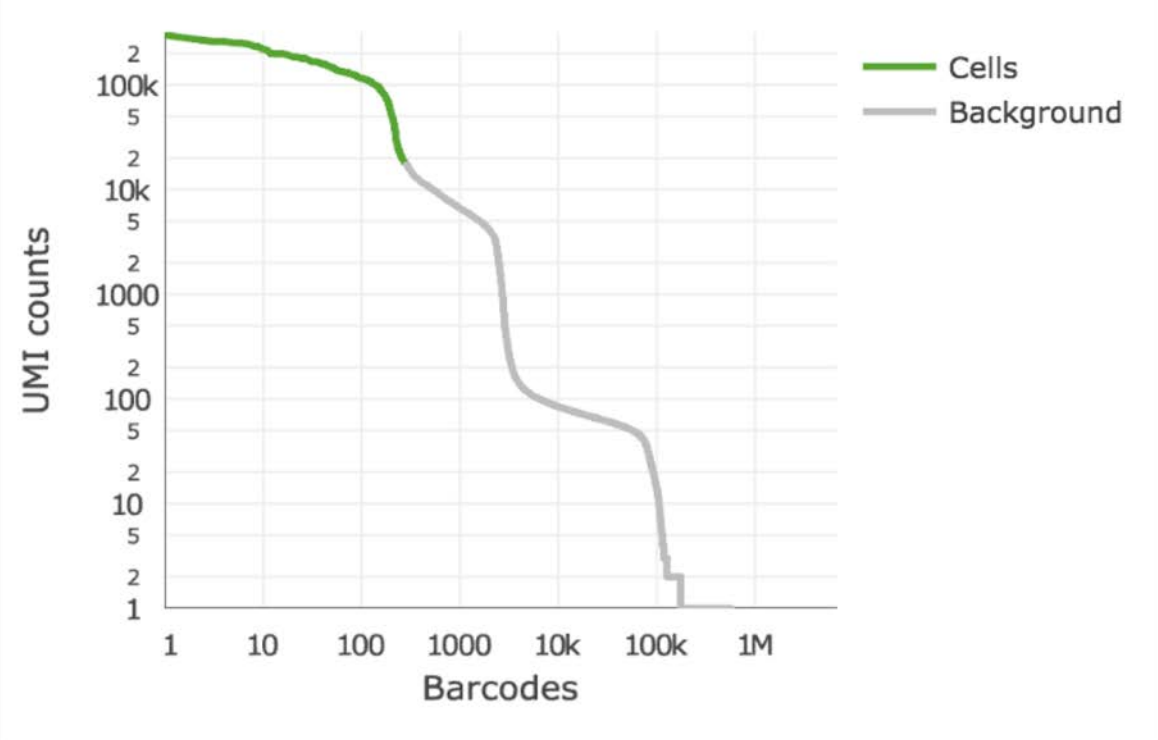
1. Empty Droplet
2. Low expressed genes
3. Cells expressing low number of genes
4. Cells with low reads
5. Dying cells
6. Doublets
7. Normalization
8. Imputation (denoise)

Empty Droplet

Cellranger takes care of it in 2 steps:

1. Maintain cells whose UMI counts/10 exceed UMI of 99th percentile

- Create background model to call remaining cells (DropletUtils)



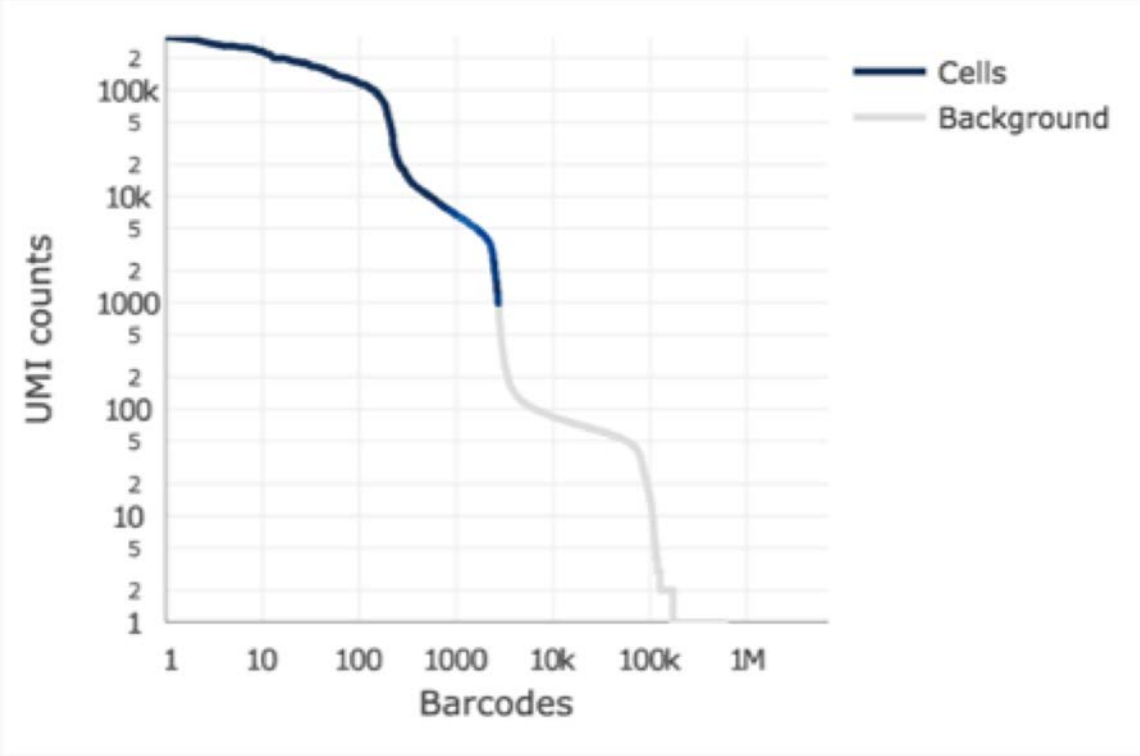
Cell Ranger 2.2

Empty Droplet

Cellranger takes care of it in 2 steps:

1.Maintain cells whose UMI counts/10 exceed UMI of 99th percentile

2.Create background model to call remaining cells (DropletUtils)



Cell Ranger 3.0

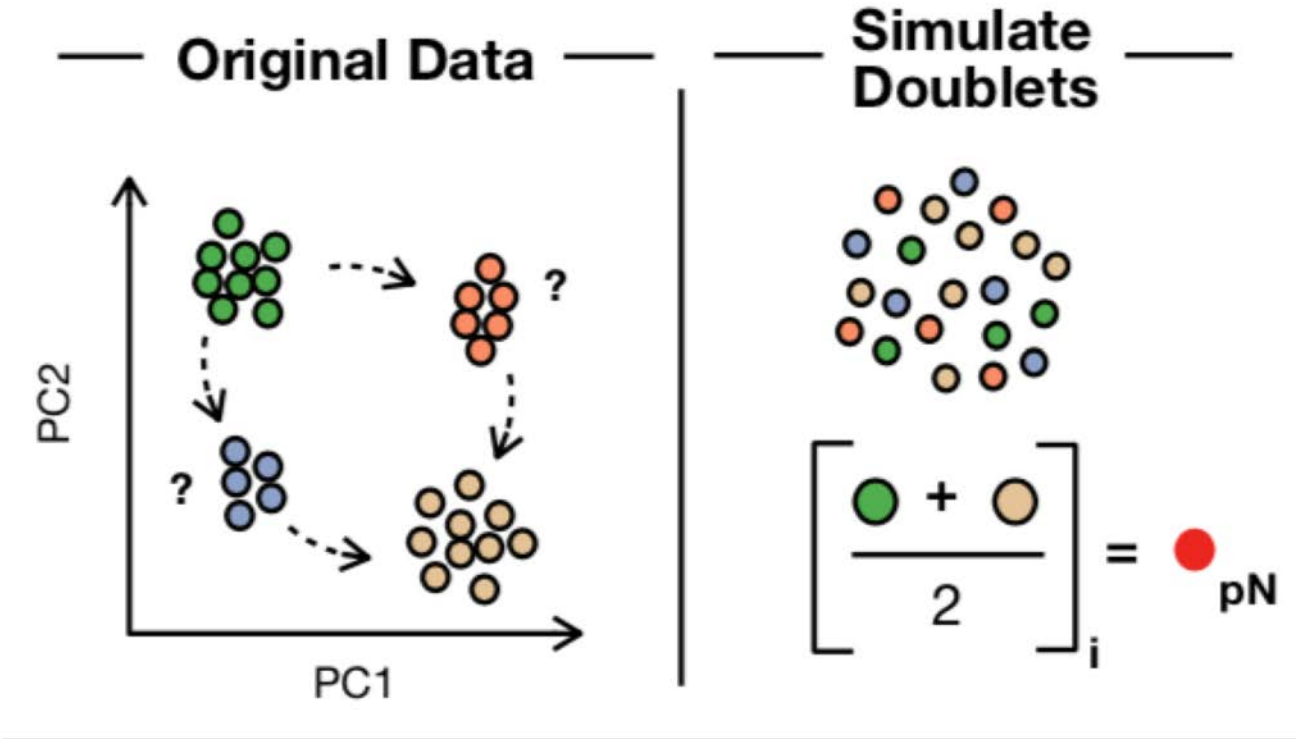
Filtering

1. **Genes filter:** keep genes that have expression in at least 0.1 percent of total number of cells
2. **Barcode Filters**: Based off of distribution of data (median +/- (3-5) deviations)
    - High percentage of mitochondria
    - Low number of genes per cell
    - Low number reads per cells

Doublets

DoubletFinder and Scrublet :

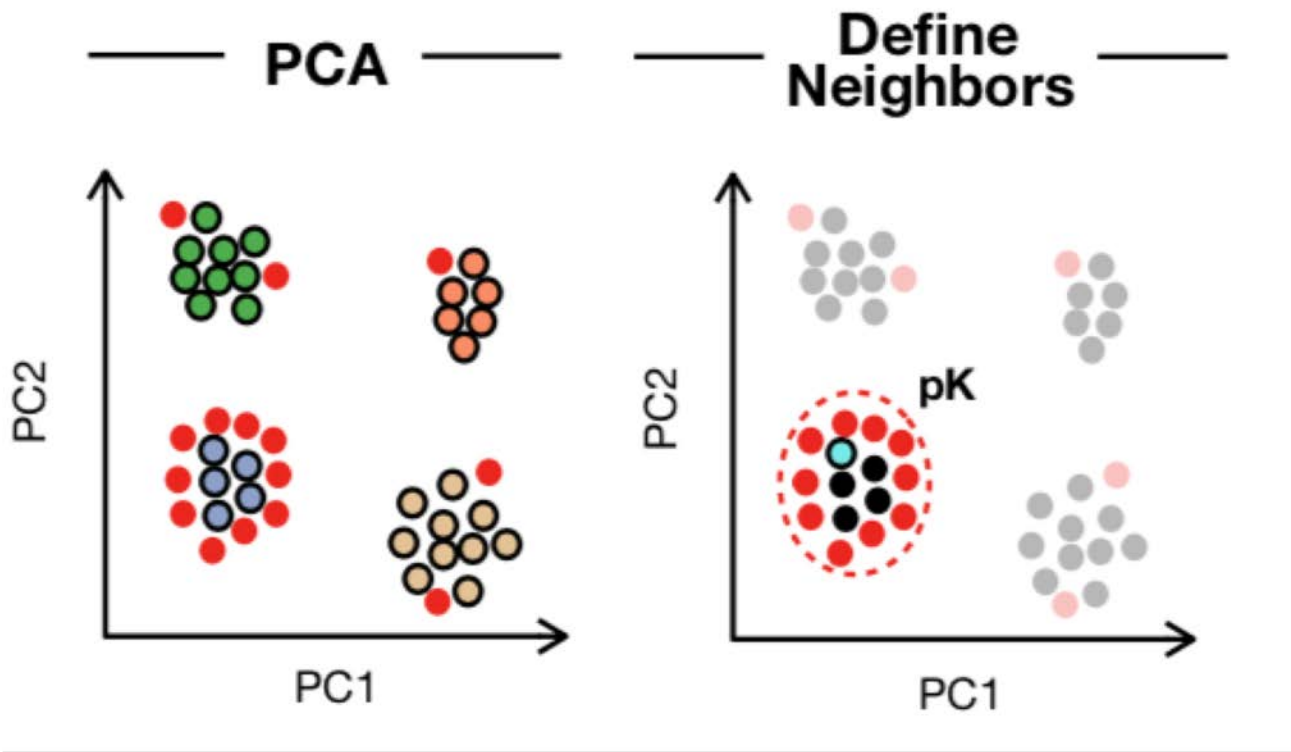
1. Generate artificial doublets from existing scRNA-seq data
2. Merge real-artificial data and find real cell's proportion of artificial k nearest neighbors
3. Rank order and threshold doublet values according to the expected number of doublets



Doublets

DoubletFinder and Scrublet :

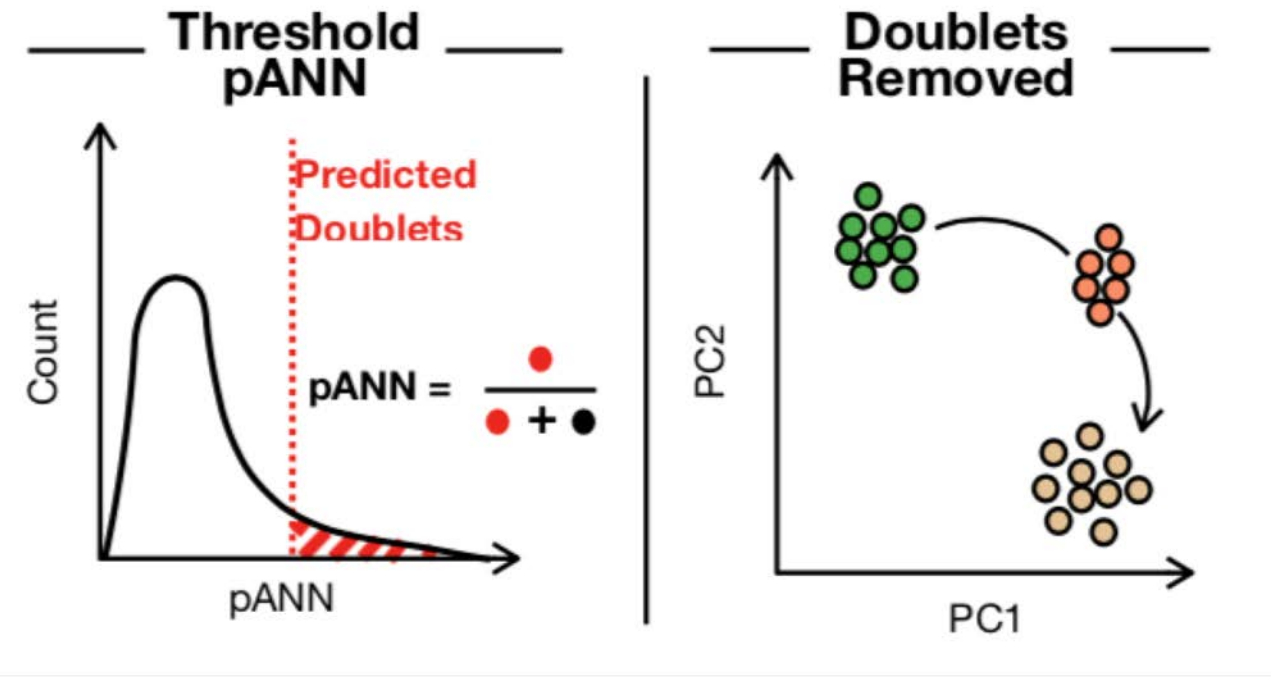
1. Generate artificial doublets from existing scRNA-seq data
2. Merge real-artificial data and find real cell's proportion of artificial k nearest neighbors
3. Rank order and threshold doublet values according to the expected number of doublets



Doublets

DoubletFinder and Scrublet :

1. Generate artificial doublets from existing scRNA-seq data
2. Merge real-artificial data and find real cell's proportion of artificial k nearest neighbors
3. Rank order and threshold doublet values according to the expected number of doublets



Normalization

Increases in sequencing typically lead to proportional increases in gene counts

Bulk RNA-seq

- Normalization methods estimate a scale factor per sample

scRNA-seq

- Data sequencing depth does not affect gene counts equally
- A lot more zeros

Normalization

TPM (Transcripts Per Million):

- This is the number of transcripts for each gene in each cell, divided by the total number of transcripts in that cell (in millions)

Scran:

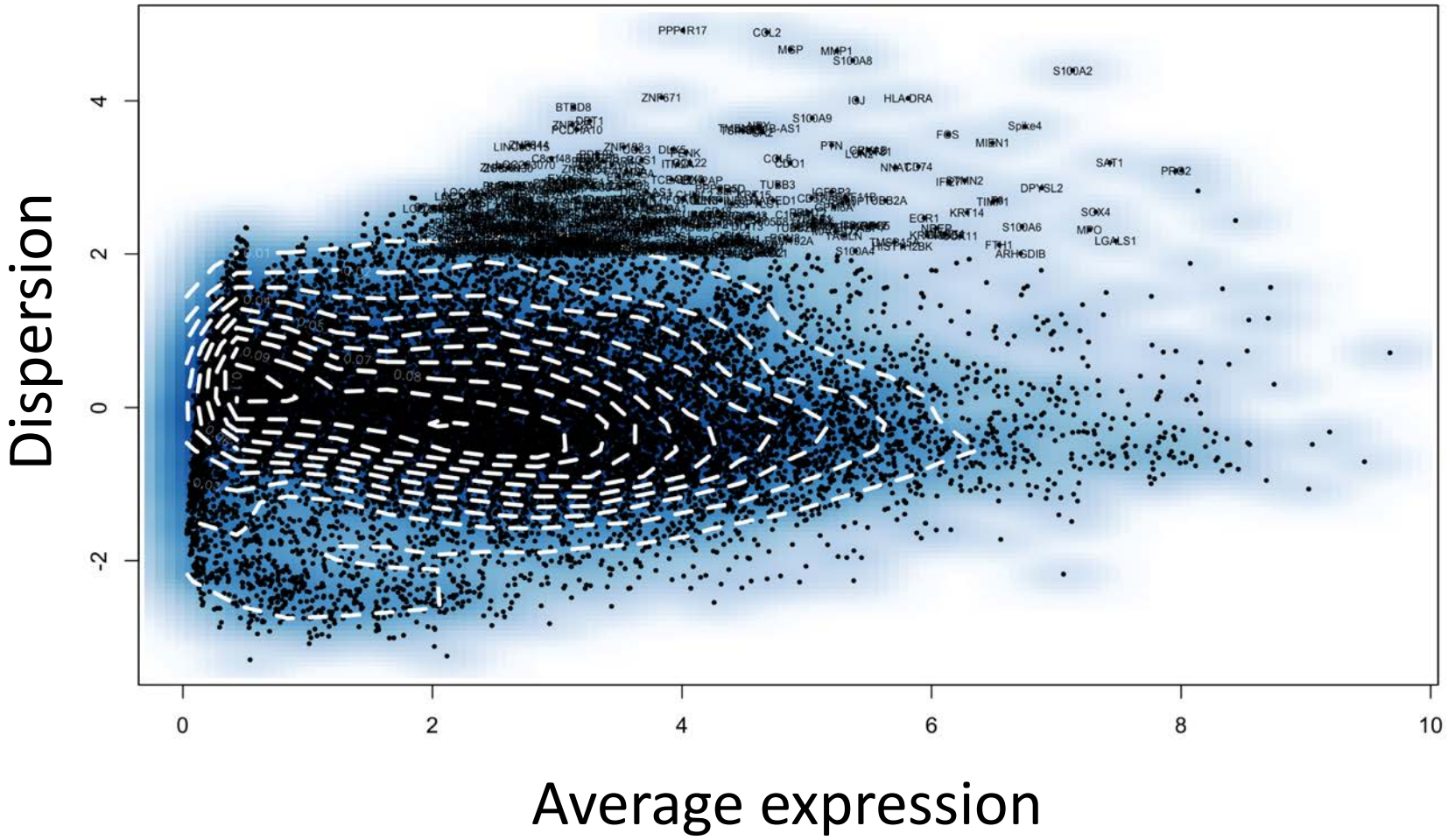
- Groups similar cells based on rank correlations in their gene expression profiles
- Normalizes across groups (CPM)
- Uses linear algebra to apply normalizations to cells

Scnorm:

- Groups genes based on their count-depth relationship
- Within each group applies a quantile regression to estimate scaling factors to remove the effect of sequencing depth from the counts

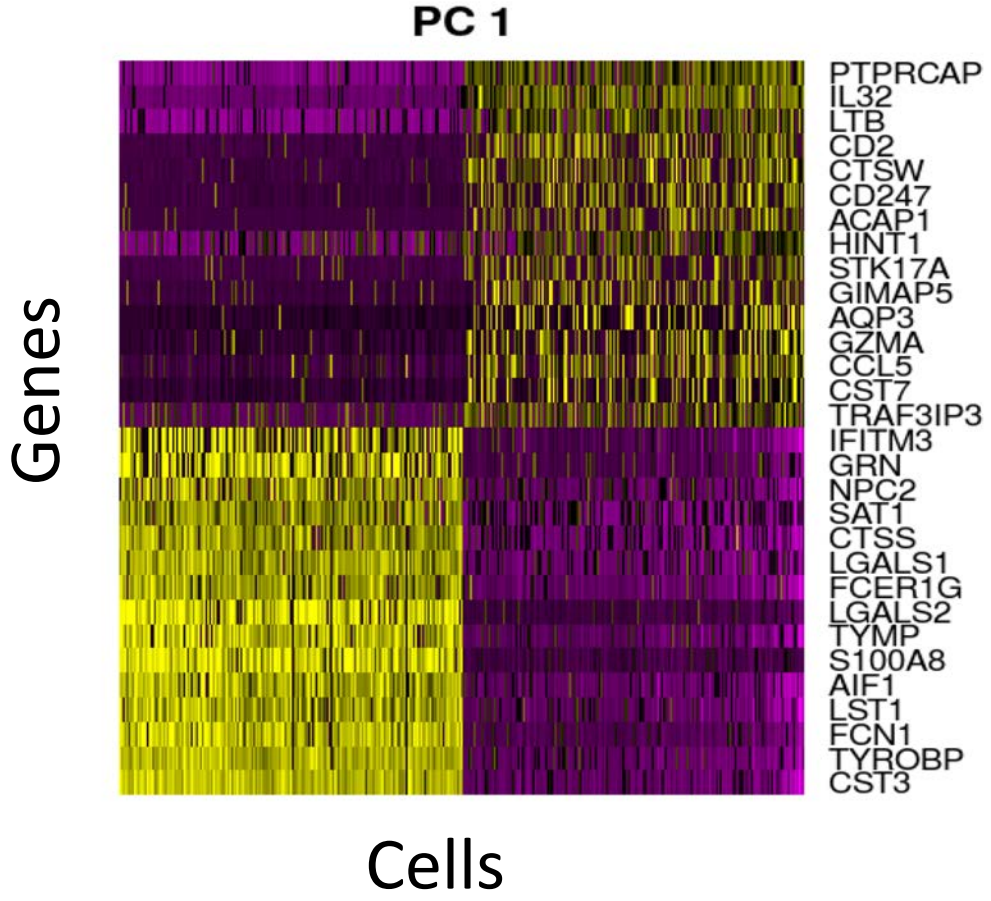
Select Interesting Genes

Selecting for genes with highest dispersion/variance values



Reducing Dimensions

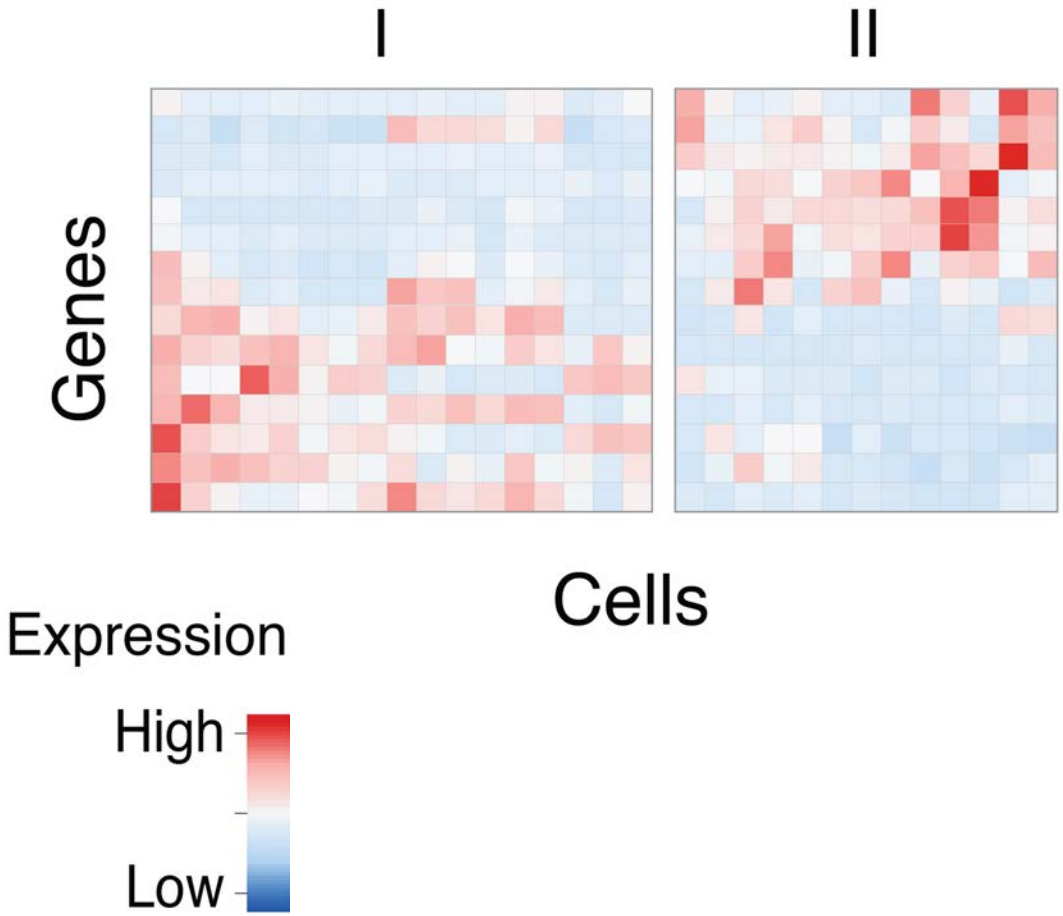
PCA reduces the dimensionality (the number of variables) of a data set by maintaining as much variance as possible.



primary sources of heterogeneity

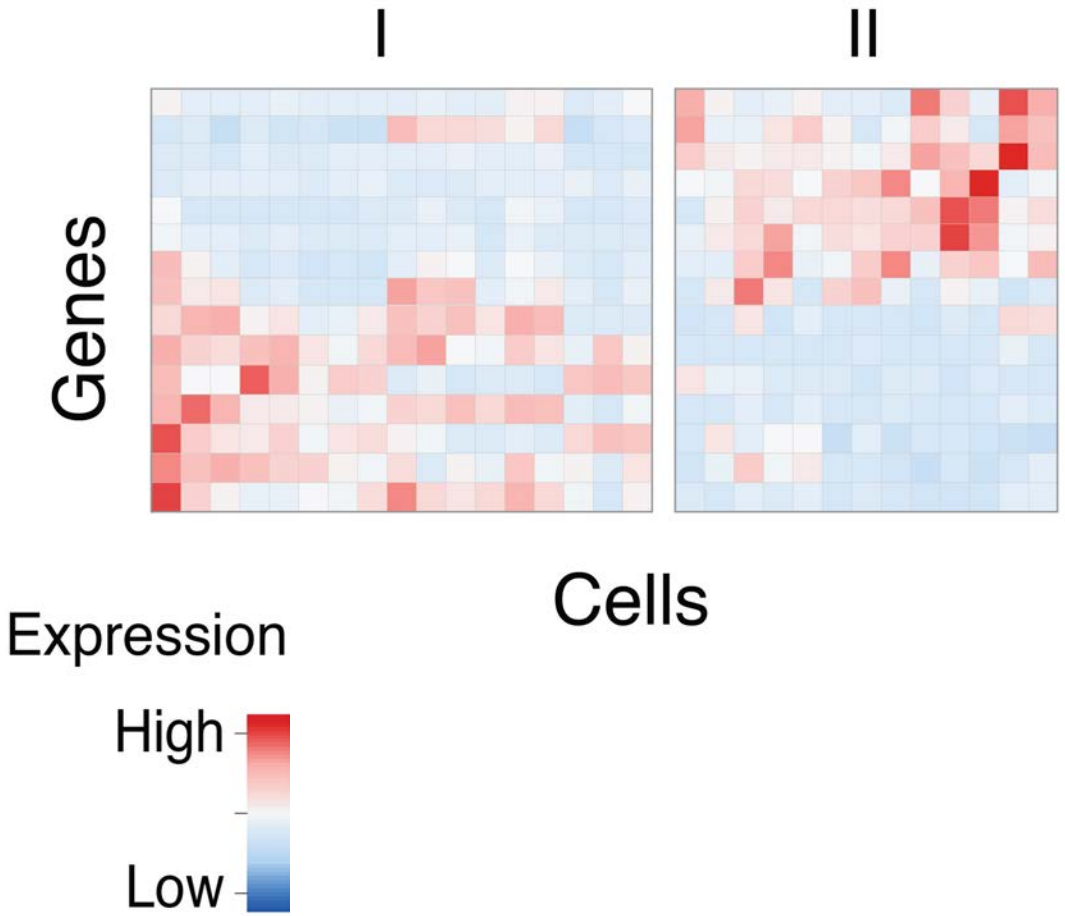
Characterizing cell types and states

- Basics of clustering analysis
- Adjusting cluster resolution
- Cluster robustness and reproducibility
- Identify high confidence populations
- Marker gene identification



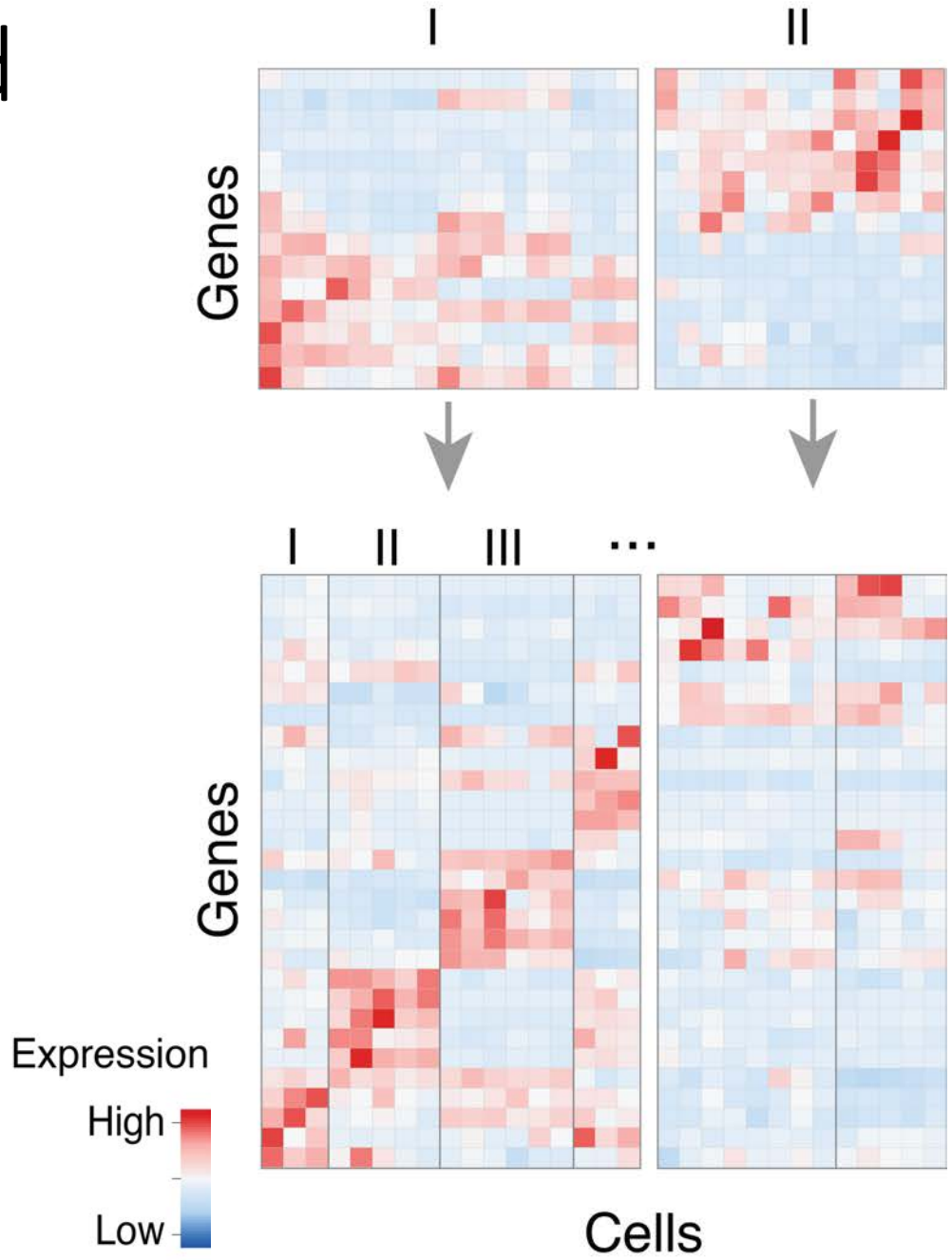
Characterizing cell types and states

- Basics of clustering analysis
  - Measure pairwise similarity
  - Identify highly similar groups
- Adjusting cluster resolution
- Cluster robustness and reproducibility
- Identify high confidence populations
- Marker gene identification



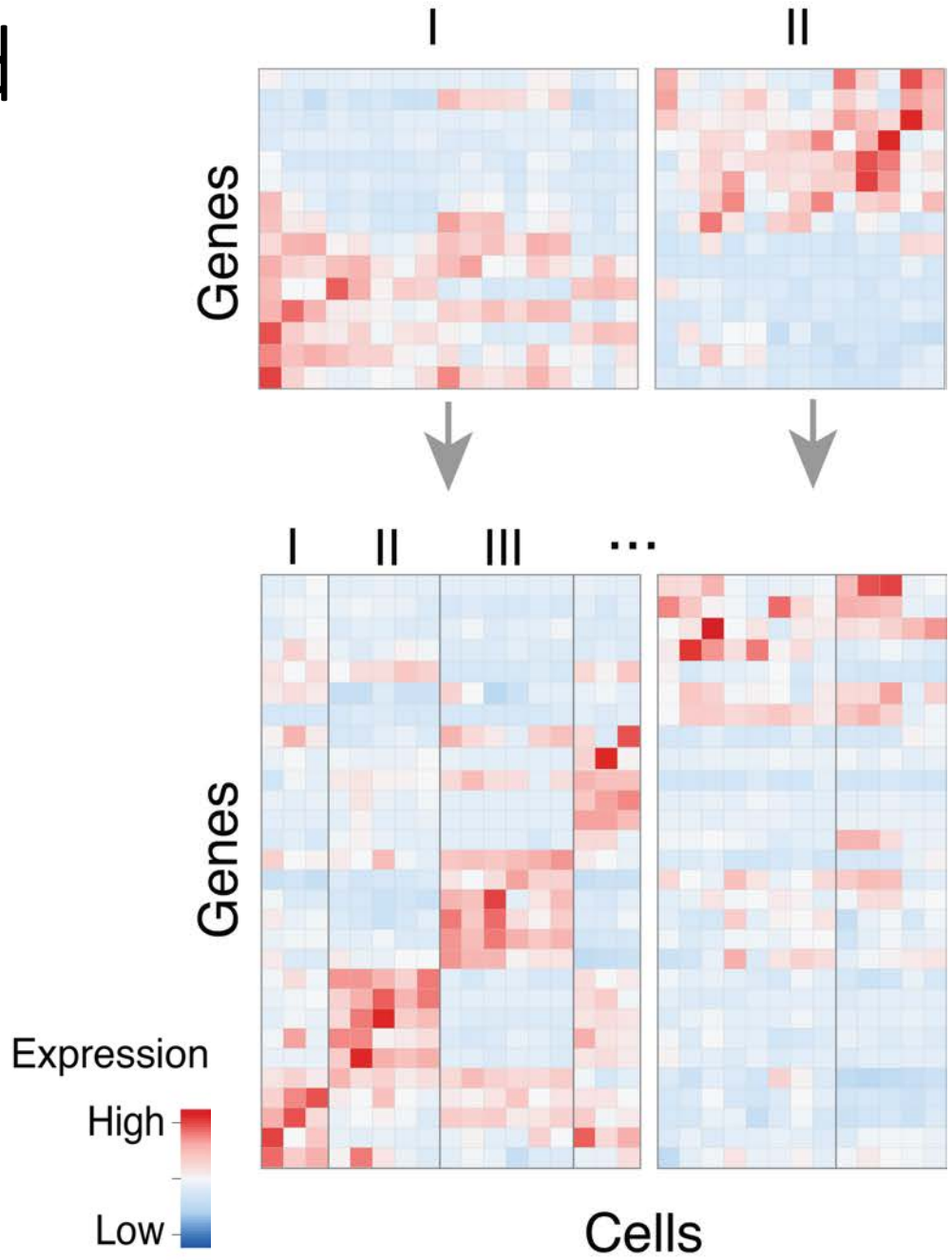
Characterizing cell types and states

- Basics of clustering analysis
- Adjusting cluster resolution
- Cluster robustness and reproducibility
- Identify high confidence populations
- Marker gene identification



Characterizing cell types and states

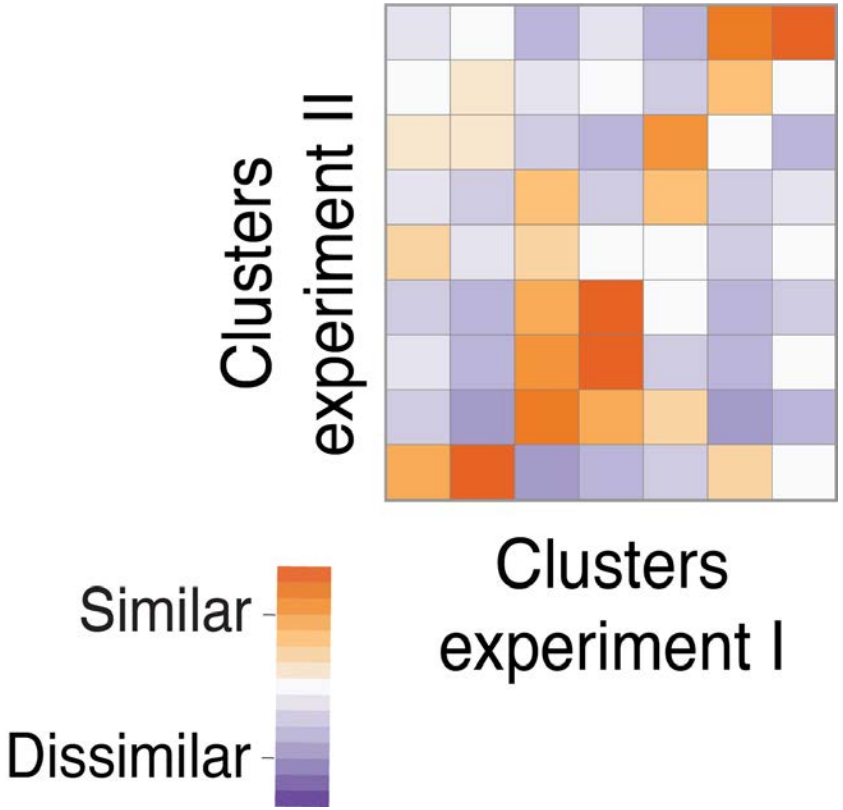
- Basics of clustering analysis
- Adjusting cluster resolution
  - Determining the number of populations
  - Manually – may be biased
  - Analytically – simulation analysis
- Cluster robustness and reproducibility
- Identify high confidence populations
- Marker gene identification



robustSingleCell @ GitHub

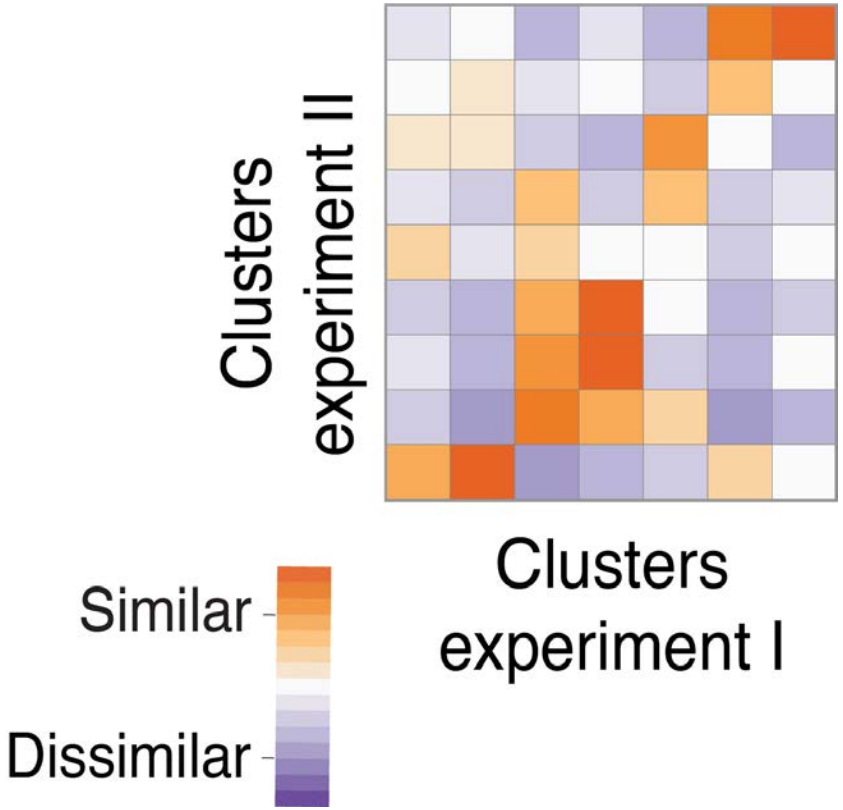
Characterizing cell types and states

- Basics of clustering analysis
- Adjusting cluster resolution
- Cluster robustness and reproducibility
- Identify high confidence populations
- Marker gene identification



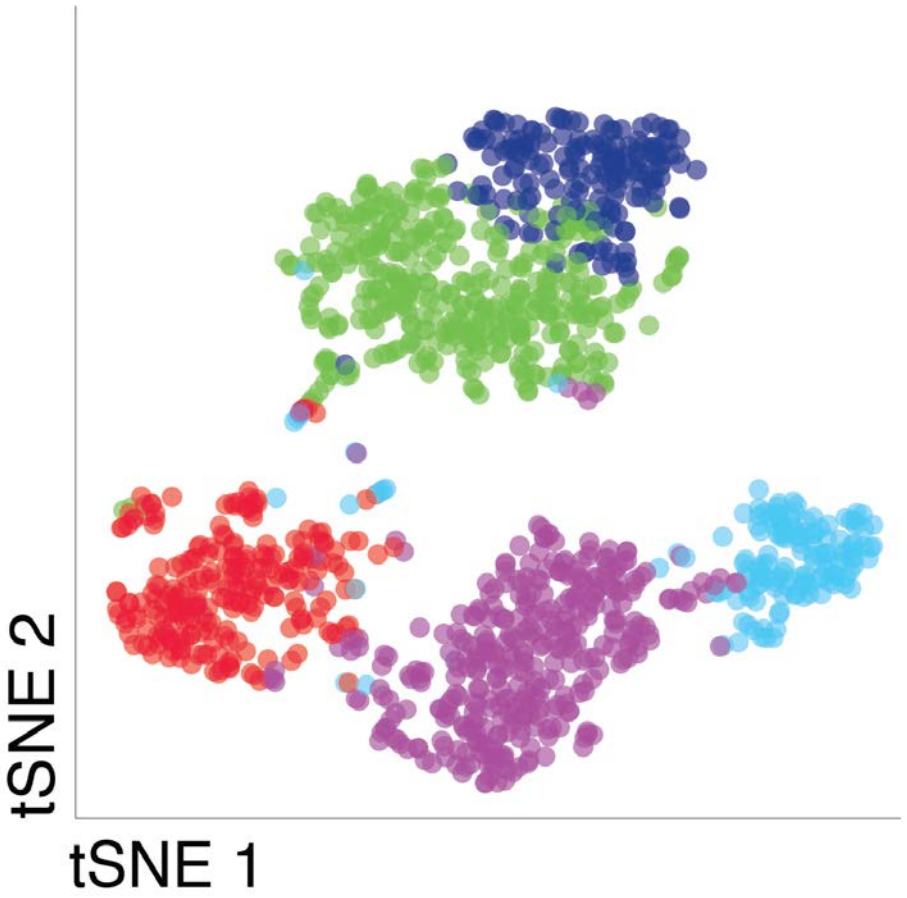
Characterizing cell types and states

- Basics of clustering analysis
- Adjusting cluster resolution
- Cluster robustness and reproducibility
  - Using sampling techniques
  - Using biological replicates
- Identify high confidence populations
- Marker gene identification



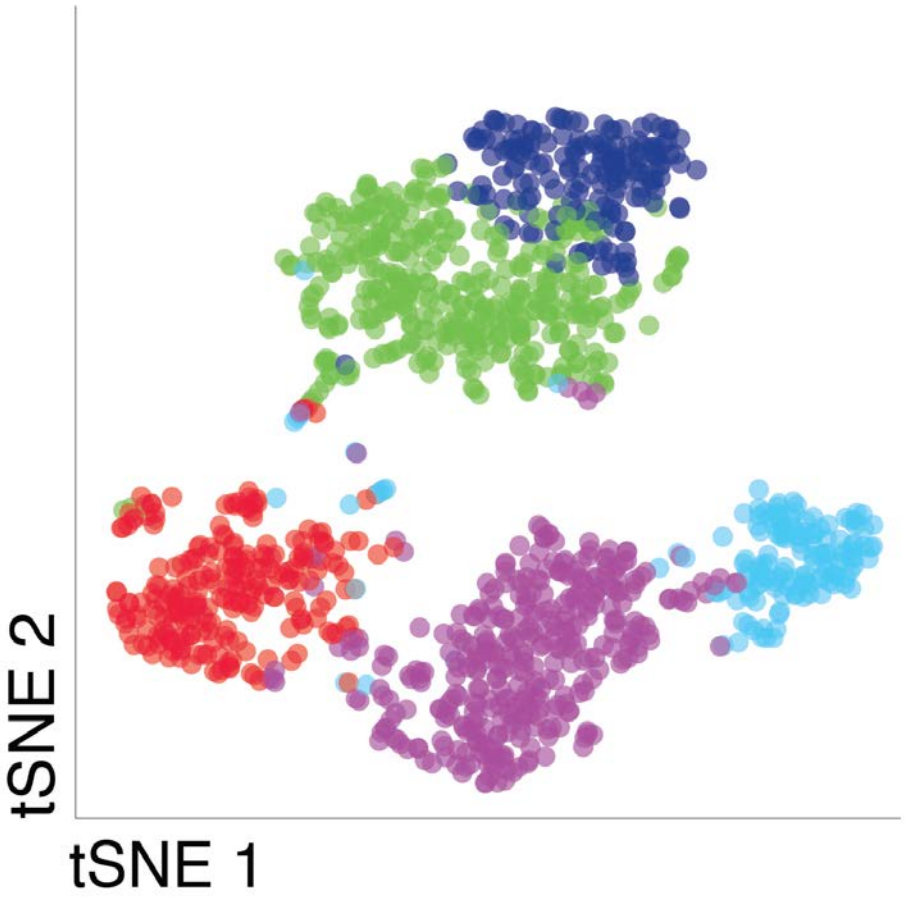
Characterizing cell types and states

- Basics of clustering analysis
- Adjusting cluster resolution
- Cluster robustness and reproducibility
- Identify high confidence populations
- Marker gene identification



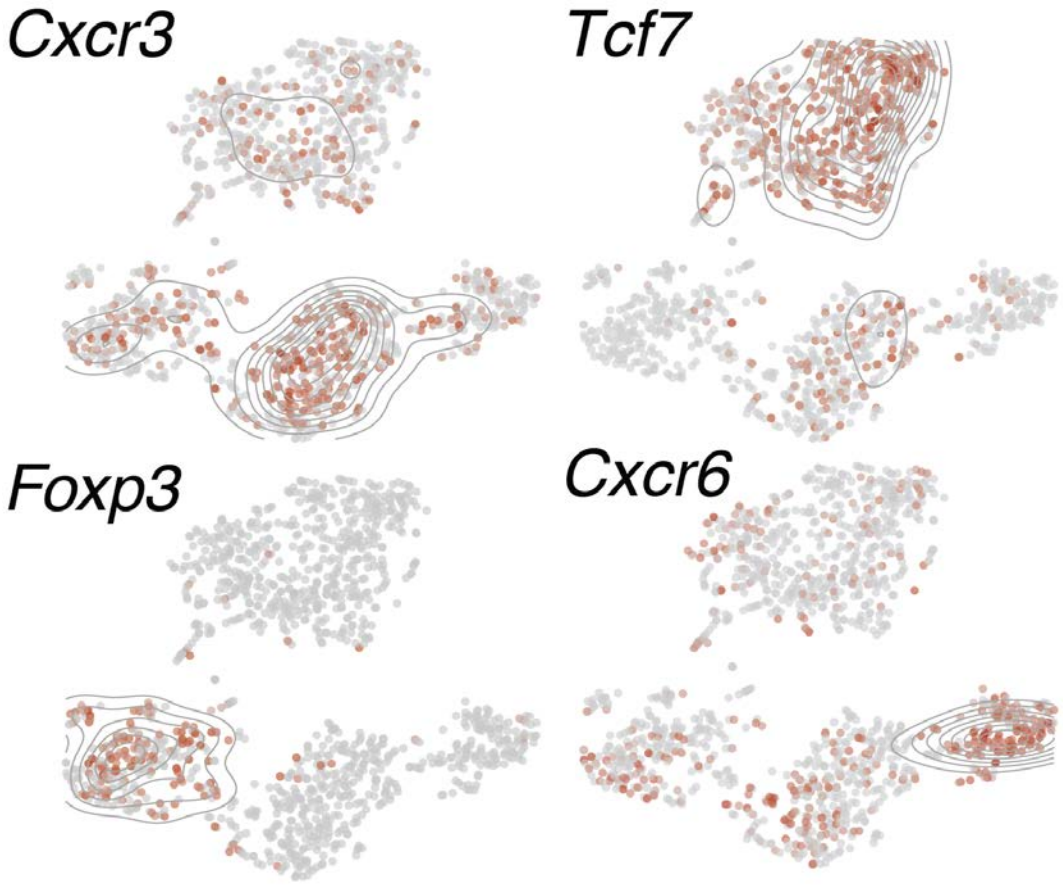
Characterizing cell types and states

- Basics of clustering analysis
- Adjusting cluster resolution
- Cluster robustness and reproducibility
- Identify high confidence populations
  - Determined by analytical criteria
  - Supported by biological knowledge
- Marker gene identification



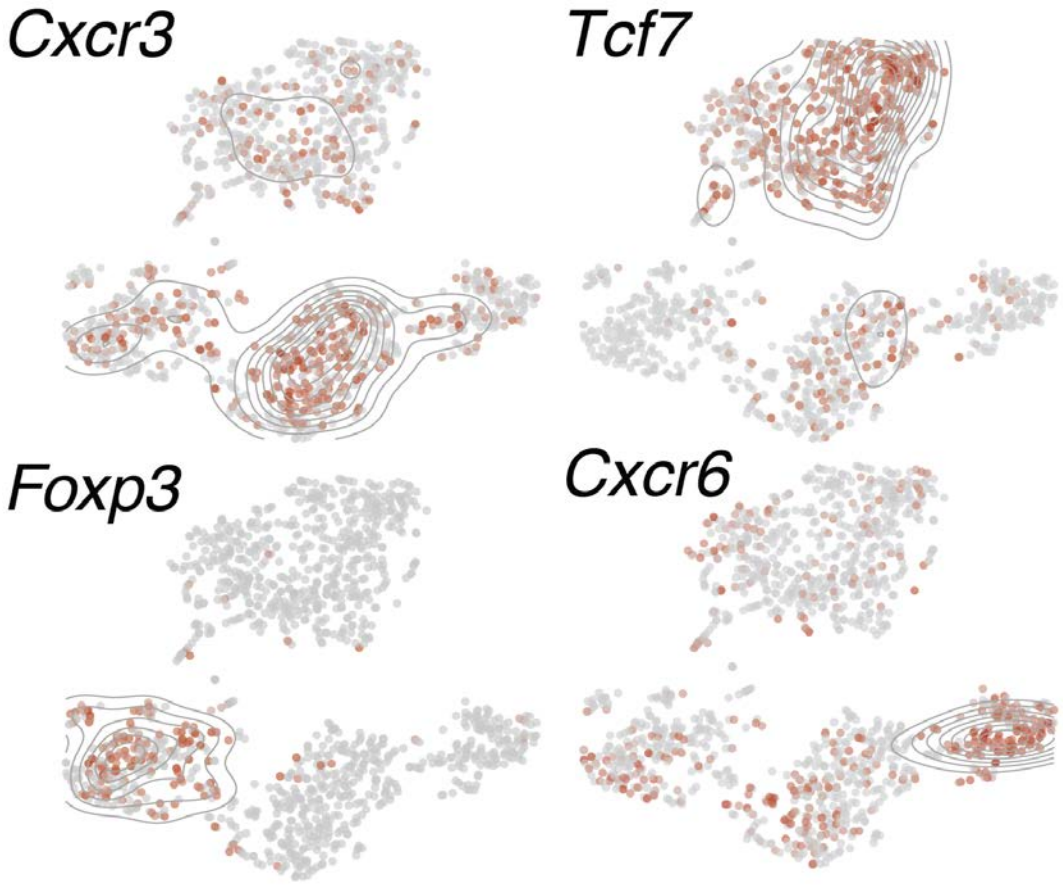
Characterizing cell types and states

- Basics of clustering analysis
- Adjusting cluster resolution
- Cluster robustness and reproducibility
- Identify high confidence populations
- Marker gene identification



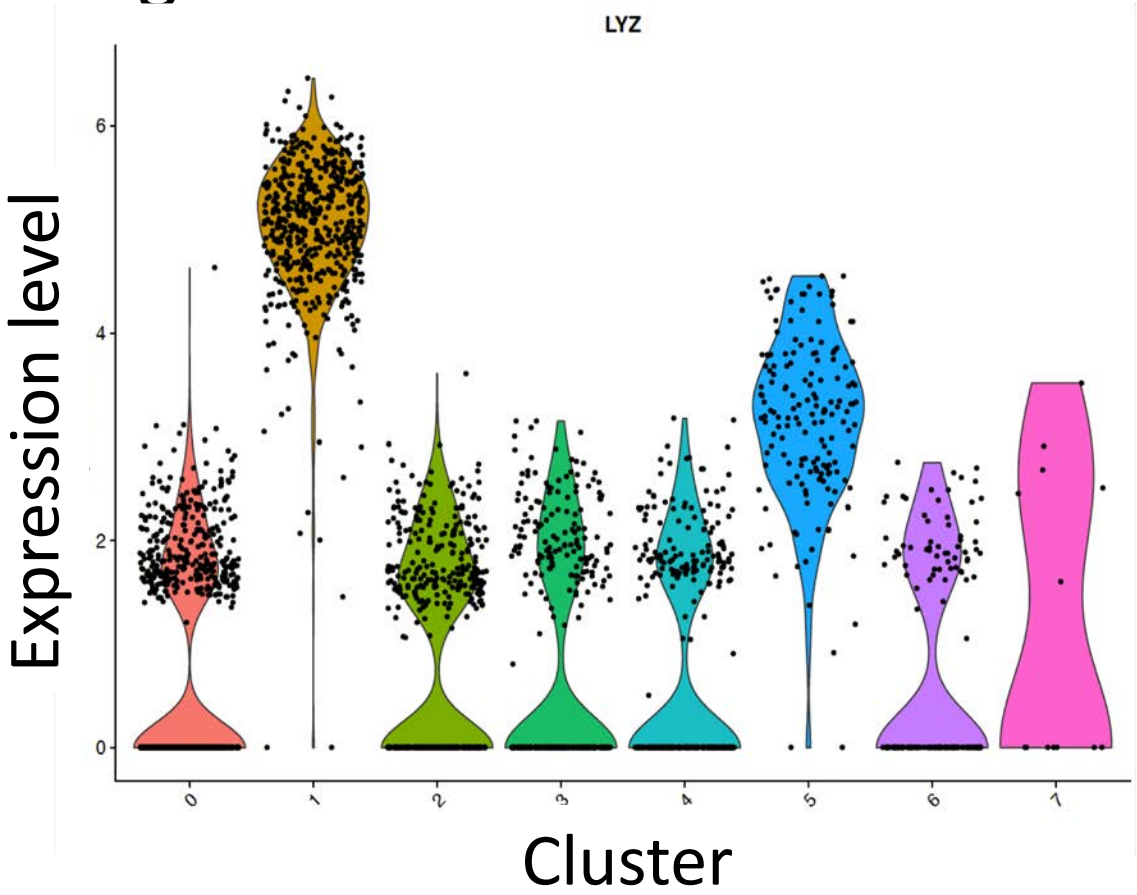
Characterizing cell types and states

- Basics of clustering analysis
- Adjusting cluster resolution
- Cluster robustness and reproducibility
- Identify high confidence populations
- Marker gene identification
  - Compare one cluster against the others
  - Rational biologically-driven comparisons



Marker Gene Identification

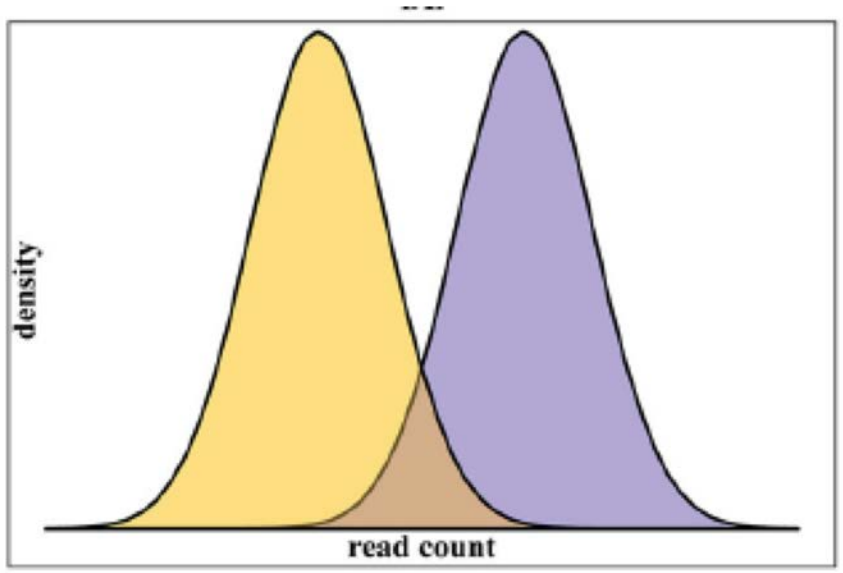
Having clustered the data, we'd like to understand the biological meaning of those clusters



Differential Gene Expression

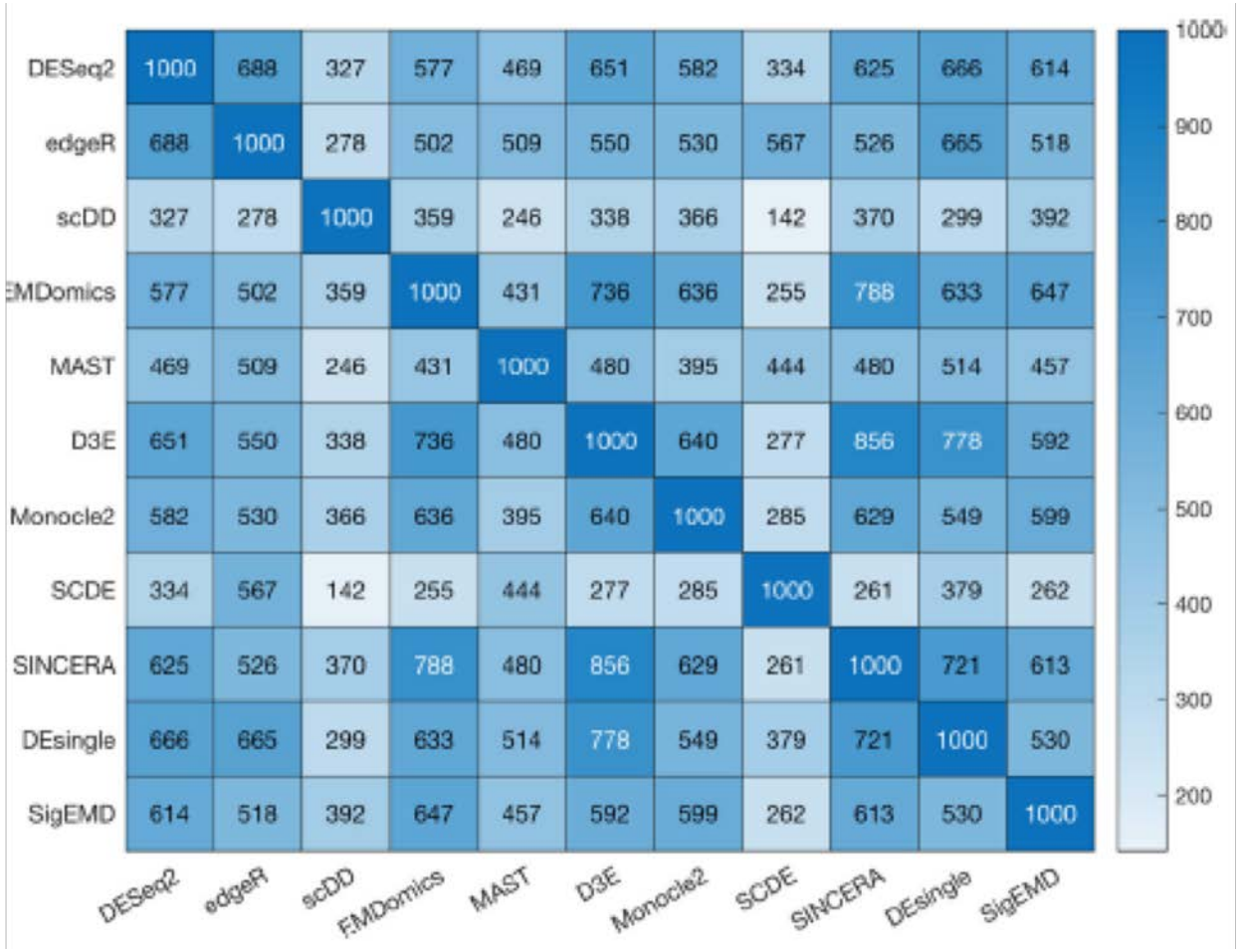
- Determine genes that are differentially expressed
  - To statistically significant degree (adjusted p-value)
  - In a biologically significant manner (log fold change)
- Complicating Factors:
  - scRNA-seq captures between 5 – 15% of the mRNA, resulting in an abundance of zero counts
  - The distribution of counts in a cluster is typically multimodal

Broad Approaches to Computing Differential Gene Expression



- **Non-parametric** approaches do not make assumptions about the distributions belonging to any particular family
- **Parametric** approaches make assumptions about the distributions
  - Negative Binomial
  - Normal

Agreement of Top 1000 genes detected by different different scDGE methods



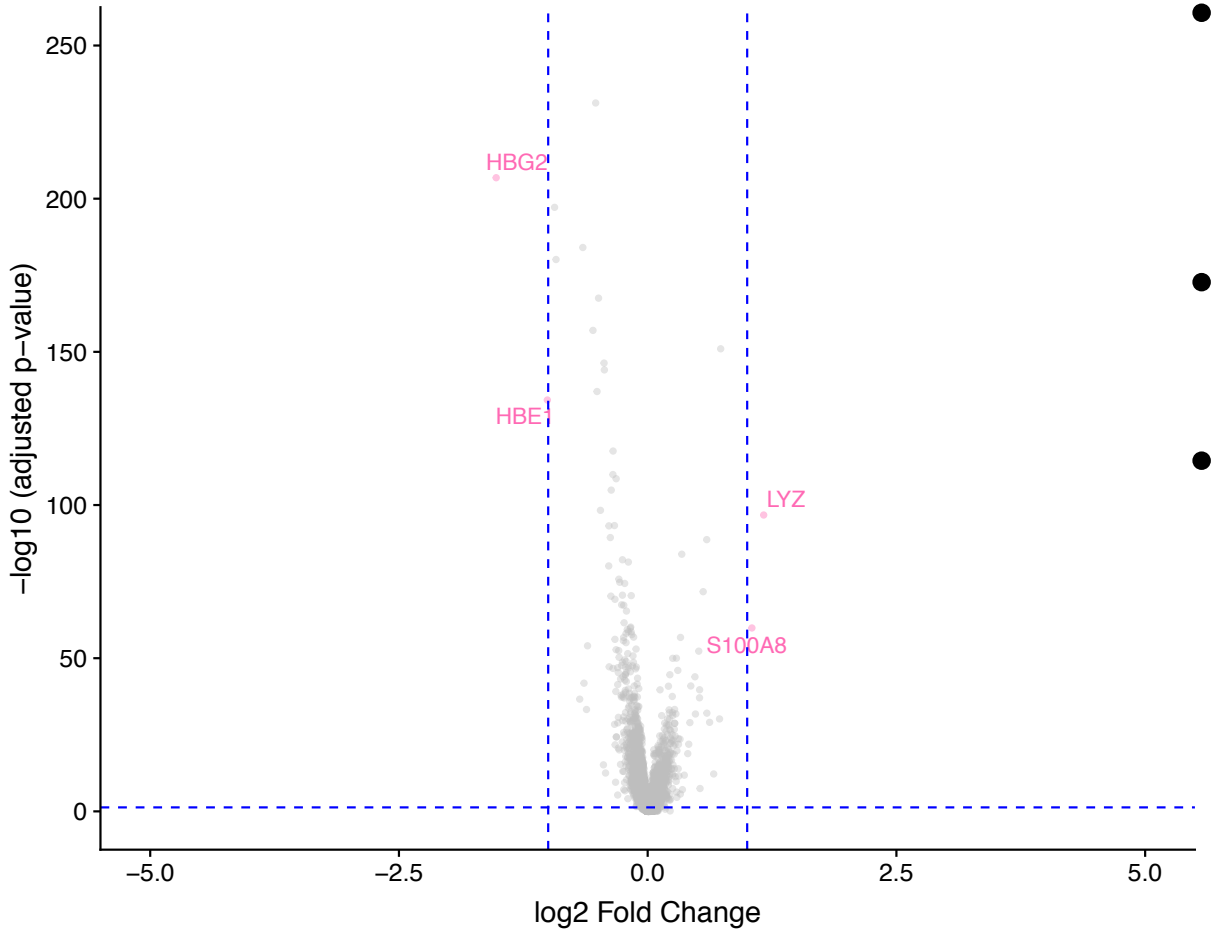
11 scDGE methods in wide spread usage. [Specific] real data.

T. Wang et al. *BMC Bioinformatics* **20**(2019):40

Differential Gene Expression – The current status

- Main conclusions:
  - Non-parametric methods handle multi-modality in data better
  - Parametric methods handle drop-outs better
  - **As a practical matter, agreement between the major methods is *adequate***
- Other considerations can be determining factor in what method to use
  - Computational speed (can vary by >2 orders of magnitude)
  - Robustness

Biologically and Statistically Significant Differential Gene Expression



- Statistical significance of differential gene expression is one component
- Log fold change can be a useful indicator of biological relevance
- Volcano plots can provide helpful visualizations of genes whose differential expression is both statistically and biologically significant

Cell Type Annotation

- Differential gene expression is often the first step to assigning cells in a cluster to a particular cell type
- Conventional approach is manual
  - Relies on association between marker genes and cell type
  - Are labor-intensive and increasingly becoming rate limiting
  - Annotations are not easily transferred to other data sets
- Supervised annotation
  - Supervised machine learning approaches
  - Marker genes are used to train classifiers
  - Classifier can then be used on new data sets
  - Effort (still in early stages) is underway to generate a repository of classifiers

H.A. Pliner et al. Supervised classification enables rapid annotation of cell atlases. bioRxiv preprint, Feb 4, 2019

Visualization

- The human sensory system is adapted to life in three dimensions
- Raw data from scRNA-seq experiments is very high dimensional (20,000 – 30,000 genes x 10s – 100s of thousands of cells)
- Fortunately much insight into the data can be obtained from a lower dimensional viewpoint

Dimensional Reduction

- Dimensional reduction is used at several points in the analysis of scRNA-seq data
- Some genes are ***uninteresting***
  - "Housekeeping" genes carry out the same functions and are essentially uniformly expressed across all cells
  - Other genes are expressed at such low levels in all cells that the "signal" from these genes is swamped by measurement "noise"
- Other genes are ***interesting***, but act in concerted fashion to carry out cellular processes such as differentiation, cell cycle, responses to environmental signals – reasonable to expect that their dynamics could be described by smaller number of degrees of freedom

Dimensional Reduction

- Dimensional reduction is used at several points in the analysis of scRNA-seq data
- ***Uninteresting*** genes were projected out during preprocessing: Helps machine learning algorithms to focus on differences that are biologically relevant
- Other genes are ***interesting***
  - Quietly used during clustering
  - Critical for visualization

Dimensional Reduction

- Broadly speaking there are two types of dimensional reductions:
  - Those that ***globally preserve distance*** (PCA, MDS and Sammon mapping)
  - Those that ***preserve distances only locally*** (t-SNE, diffusion maps, UMAP, etc.)

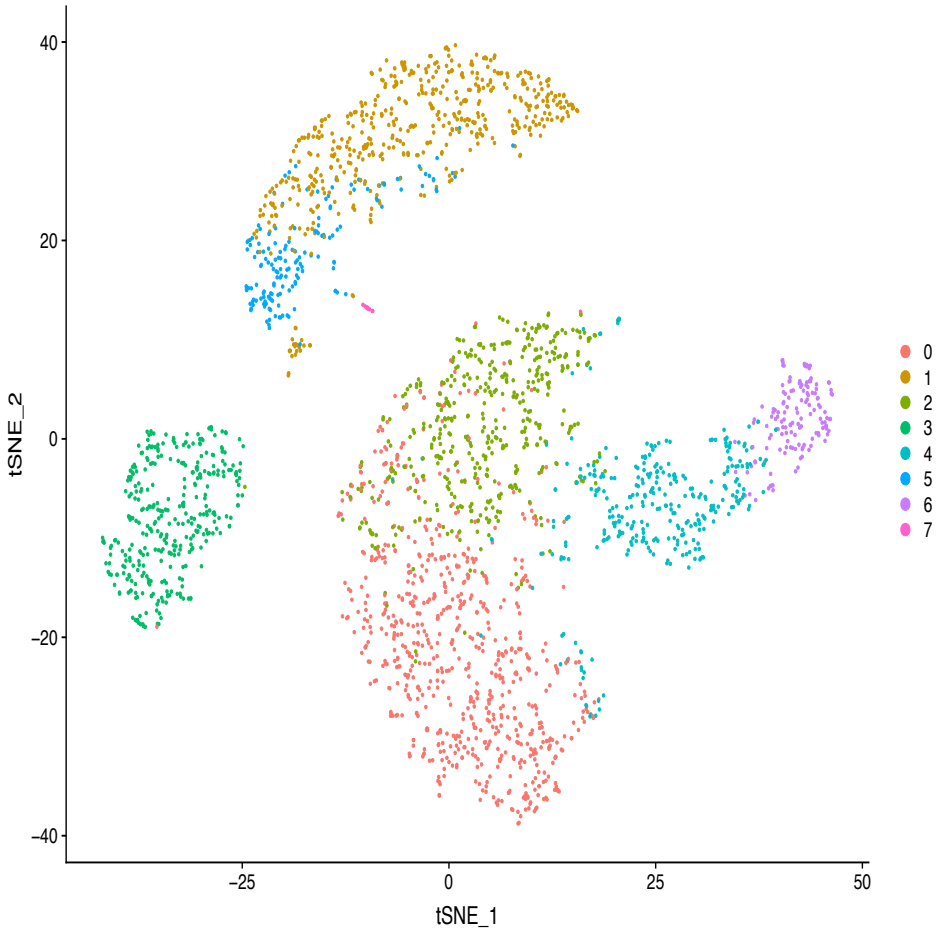
t-Distributed Stochastic Nearest (t-SNE) Neighbor Embedding

- One can the notion of *similarity* between two cells based on the probability that random walks in gene space connect them
- In dimensional reduction cells in a higher dimensional space are projected to a lower dimensional on
  - There is a notion of similarity in the higher dimensional space
  - There is another notion of similarity in the lower dimensional space
- t-SNE aims to learn a map from the higher dimensional space into a lower dimensional one, that preserves these notions of similarity as much as possible

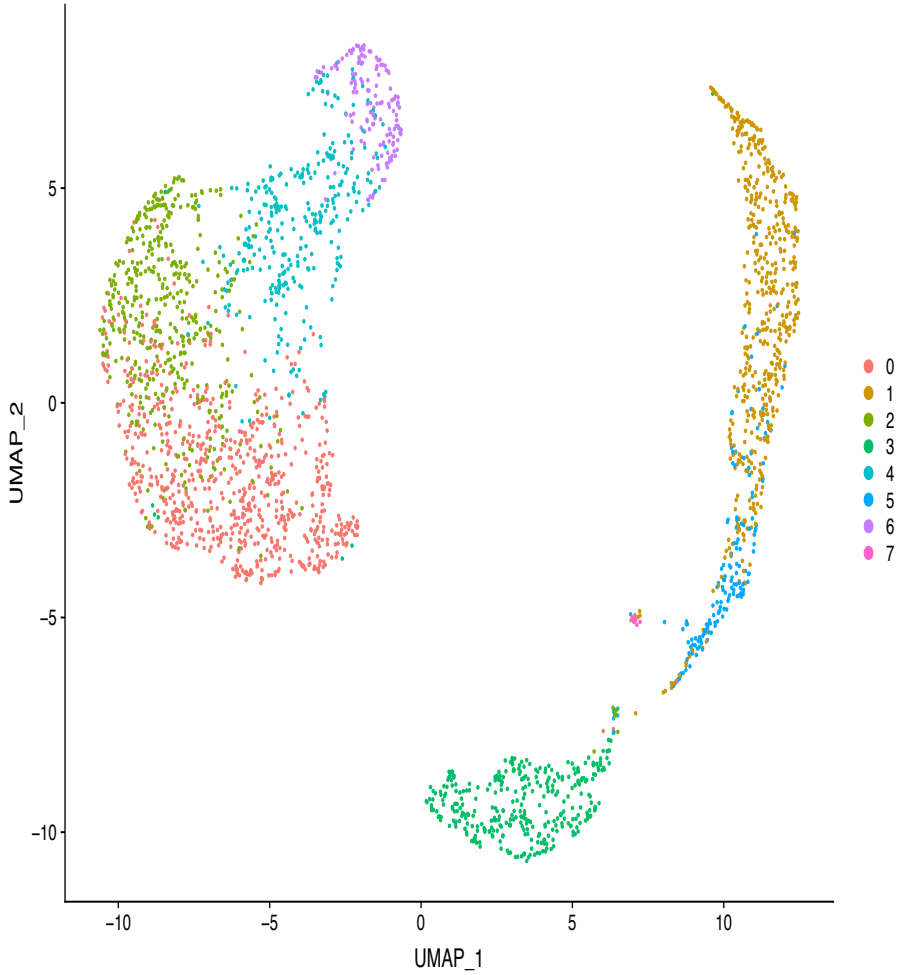
Uniform Manifold Approximation and Projection (UMAP)

- Under mild assumptions about the space in which the data intrinsically “lives”, one can model it by a “fuzzy topological structure”
- The lower dimensional representation of the data may similarly be modeled.
- Two fuzzy sets can be compared quantitatively via the “cross entropy”
- UMAP computes the lower dimensional representation that minimizes the cross entropy

Visualizing Data with t-SNE and UMAP



t-SNE



UMAP

Comparisons

- By design UMAP preserves more of the global structure
- UMAP preserves the continuity of cell subsets – which is critical for understanding cellular development
- UMAP has more hyperparameters which can be tuned to resolve subtle subsets of cells
  - Number of nearest neighbors in computing the local distance measure
  - Dimension of the space to which projecting
  - Desired separation between close points in the space to which we are projecting
  - Number of random lower dimensional representations from which we start the optimization

Choosing Hyperparameters and Final Words of Caution

- Rationally choosing and optimizing the hyperparameters is an open problem
- Current approaches address this problem iteratively in conjunction with clustering and require biological insight into how faithfully known cell types are separated
- UMAP should be used with caution on small data sets
- Both t-SNE and UMAP lack the interpretability of the reduced dimensional results of PCA

Recap and Discussion Questions

- High-throughput single-cell technologies are rich with information
- scRNAseq analysis can be challenging
- Data-driven analytical approaches make it possible
- Users should consider the limitations of analytical solutions
- **April 3rd Session:** Discuss dataset integration and comparison approaches
-

References and Useful Links

DropletUtils: Lun et al . “Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data.” *BioRxiv 2018*

DoubletFinder: McGinnis et al . "DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors." *BioRxiv,* 2018

Garnett: Pliner et al. “Supervised classification enables rapid annotation of cell atlases.” BioRxiv, 2019

Scrublet: Wolock et al. "Scrublet: computational identification of cell doublets in single-cell transcriptomic data." *BioRxiv* , 2018

SCnorm: Bacher et al. “SCnorm: Robust Normalization of Single-Cell RNA-Seq Data.” Nature Methods, 2017

Scran: Lun et al "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts." *Genome biology* , 2016

Scater: McCarthy et al. "Scater: pre-processing, quality control, normalisation and visualisation of single-cell RNA seq data in R." Bioinformatics, 2017.

robustSingleCell: A pipeline designed to identify robust cell subpopulations using scRNAseq data and compare population compositions across tissues and experimental models via similarity analysis. github.com/asmagen/robustSingleCell

tSNE: van der Maaten et al. “Visualizing High-Dimensional Data Using t-SNE.” *J. Mach. Learning Res*. 9(2008):2579.

UMAP: McInnes et al. “UMAP: Uniform Manifold Approximation and Projection.” arXiv, 2018.

Additional Links

- 10X University:

https://www.10xgenomics.com/10x-university/

- Sean Davis' List of Single Cell Analysis Tools: https://github.com/seandavi/awesome-single-cell

- Hemberg Lab Single Cell "Course":

https://hemberg-lab.github.io/scRNA.seq.course/