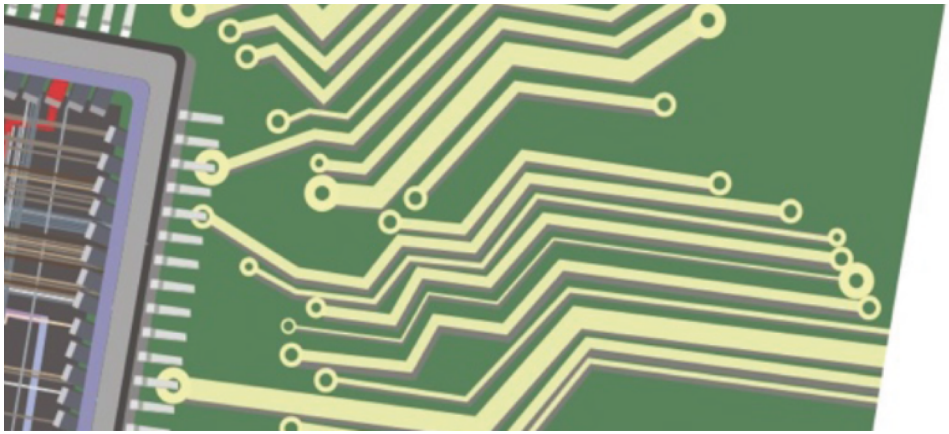


Computational Genomics Workshop

September 10 & 11, 2018

Introduction to single-cell RNA-seq analysis

Orr Ashenberg
Caroline Porter
Ayshwarya Subramanian

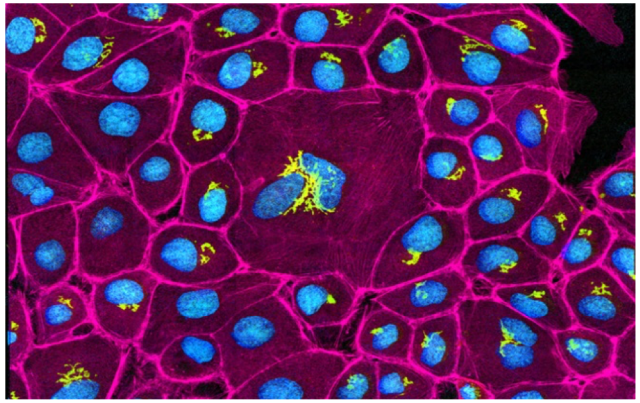
BROAD INSTITUTE

Goals for today

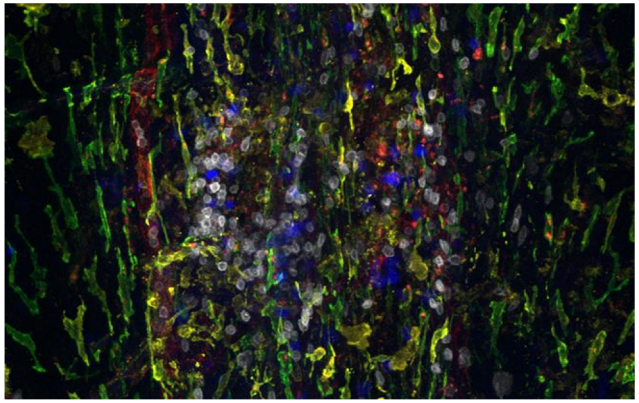
- Give an overview of single-cell RNA-seq data analysis
- Provide hands on experience stepping through an analysis pipeline, including performing quality control and identifying cell type subsets
- Introduce you to the Seurat analysis package

Incredible diversity in cell types, states, & interactions across human tissues

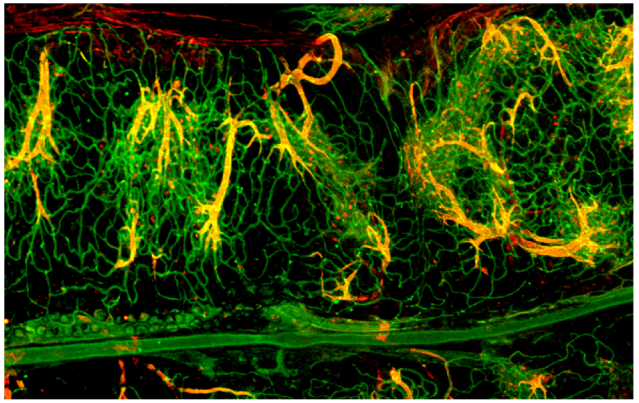
Skin epithelium



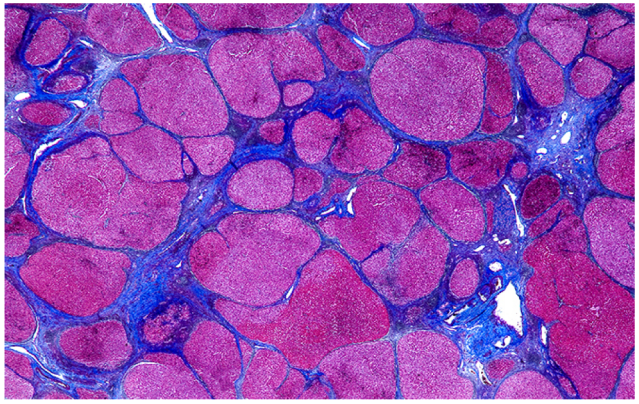
Brain meninges



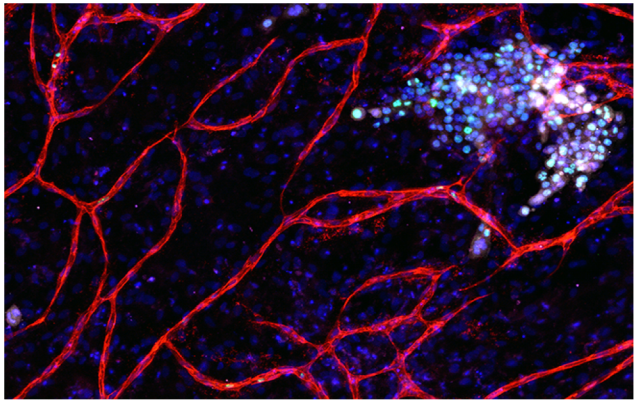
Blood vessels



Small intestine



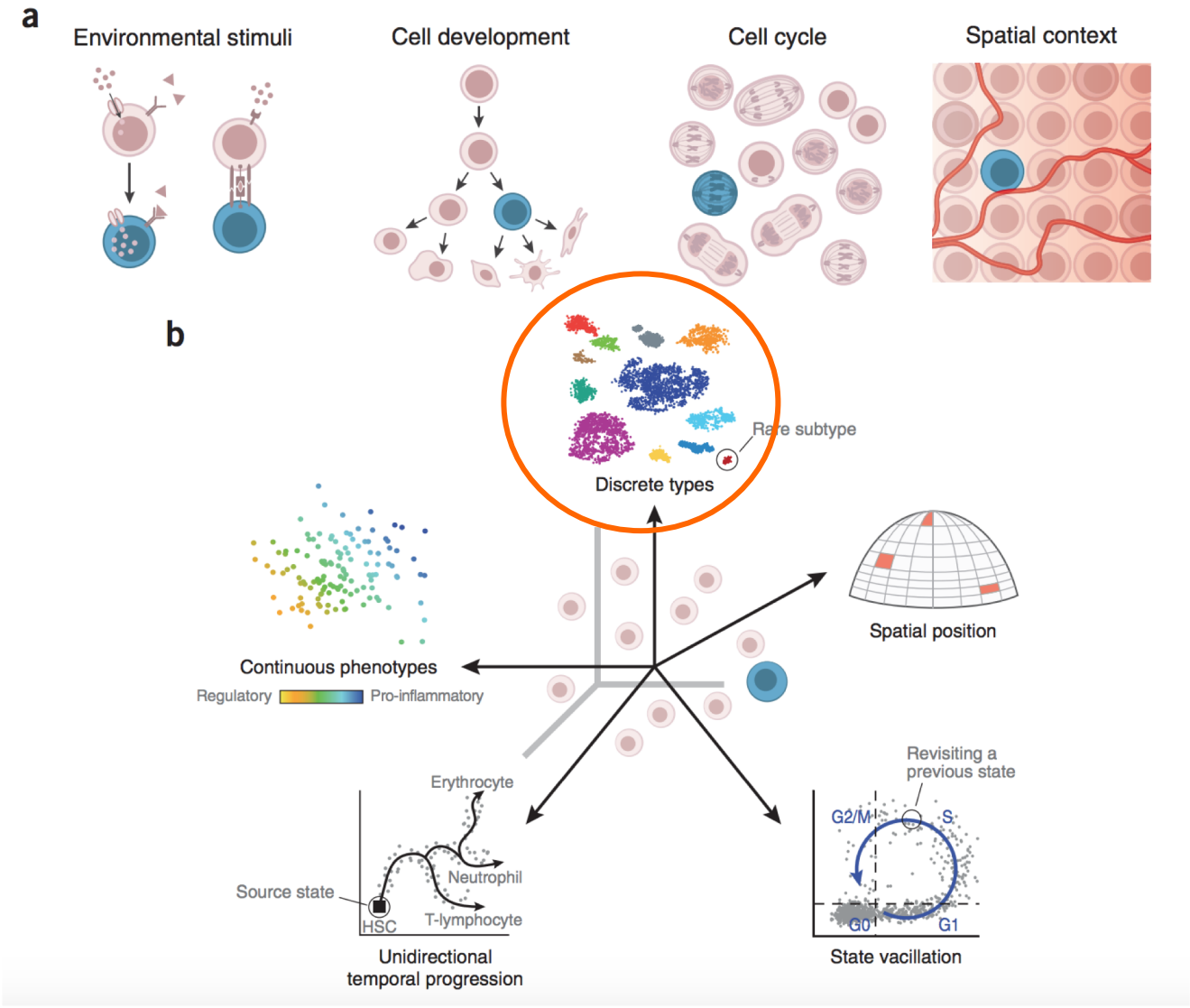
Liver cirrhosis



Breast cancer

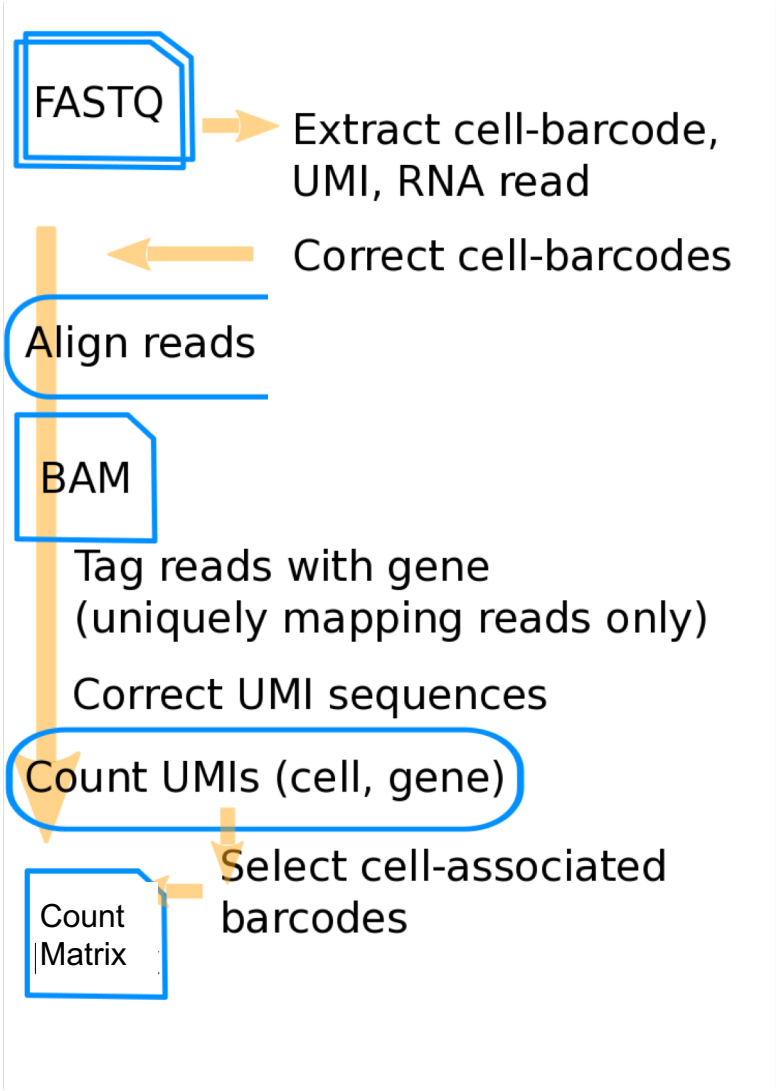
http://www.cell.com/pictureshow/skin | https://library.med.utah.edu/WebPath/webpath.html

The identity and fate of a cell are shaped by many features

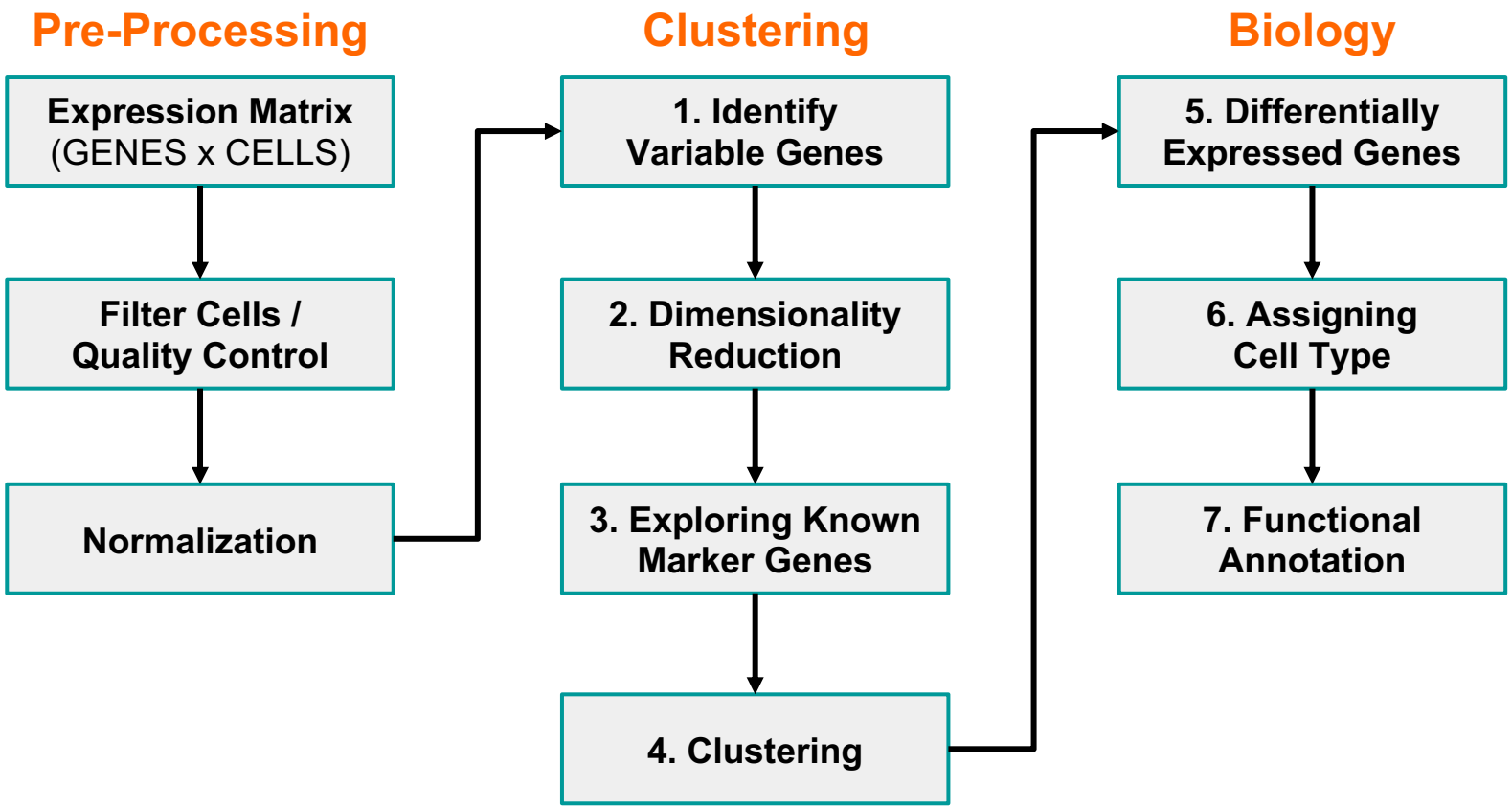


Wagner, Regev and Yosef. Nat Biotech 2016

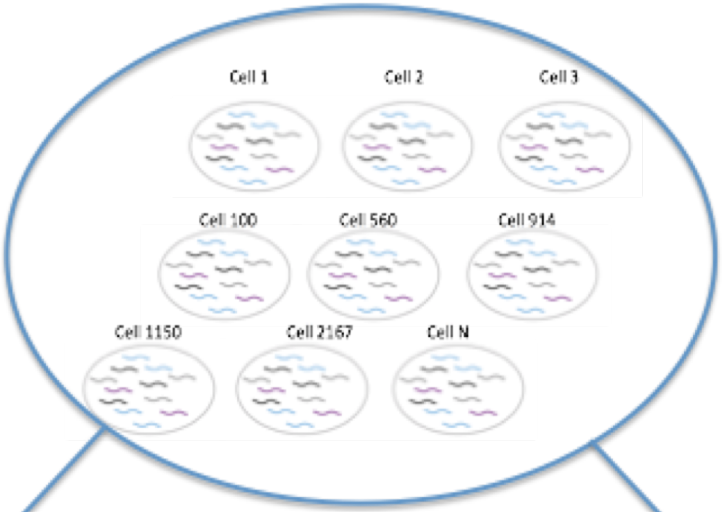
Single-cell RNA-seq analysis pipeline: Generating the count matrix



Single-cell RNA-seq analysis pipeline: Analyzing the expression data



Single-cell and bulk gene expression distributions are very different



Expression Matrix

	Exp
Gene 1	100
Gene 2	3.5
Gene 3	10
...	
Gene K	95
...	
Gene 25K	0.4

	Cell 1	Cell 2	Cell 3	....	Cell 5K
Gene 1	3	0	1		2
Gene 2	0	2	0		1
Gene 3	1	0	3		5
...					
Gene K	14	7	1		0
...					
Gene 25K	0	13	1		0

Population Average

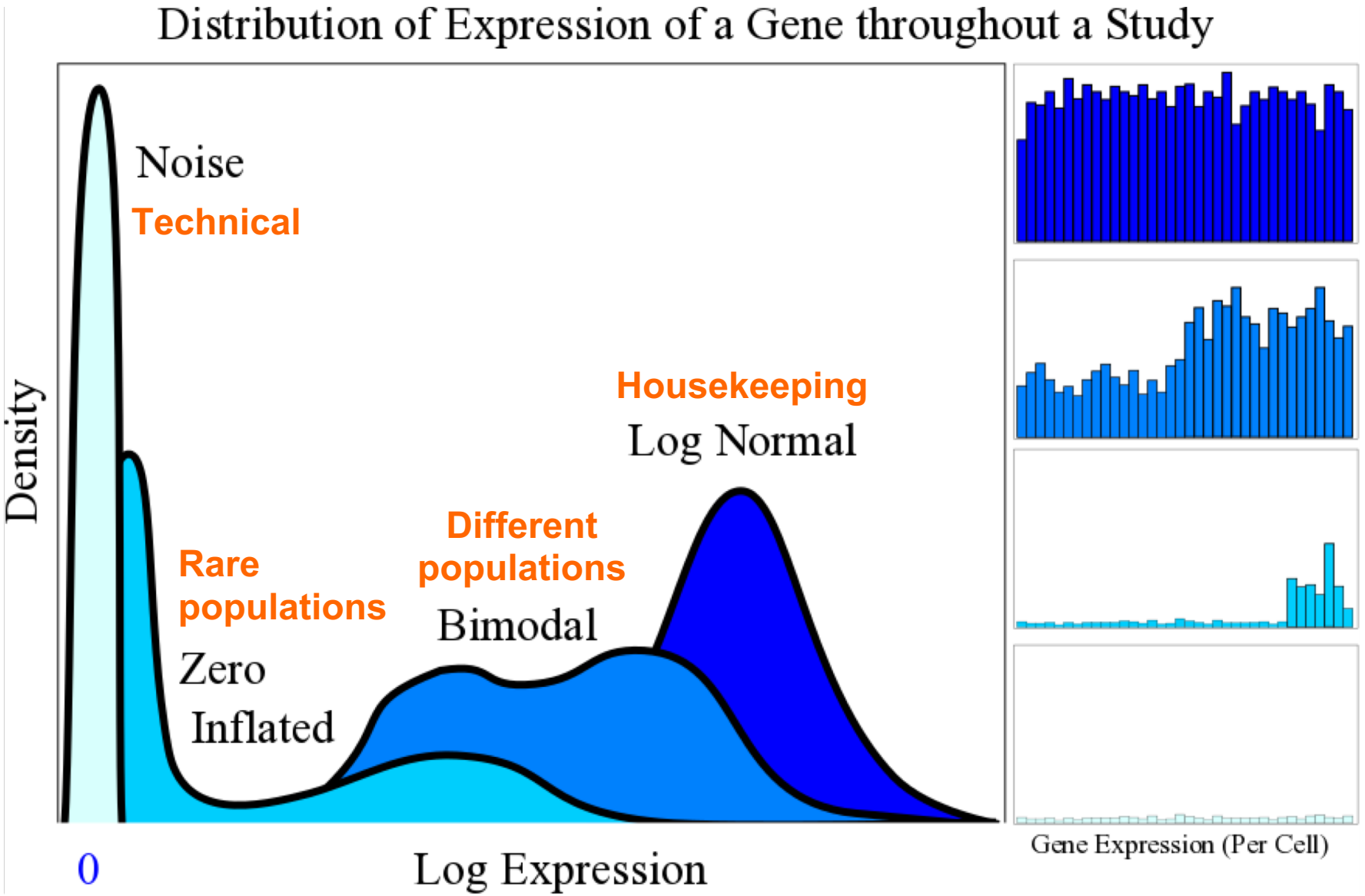
Bulk

Cellular resolution

Single cell

Slide courtesy of Karthik Shekhar

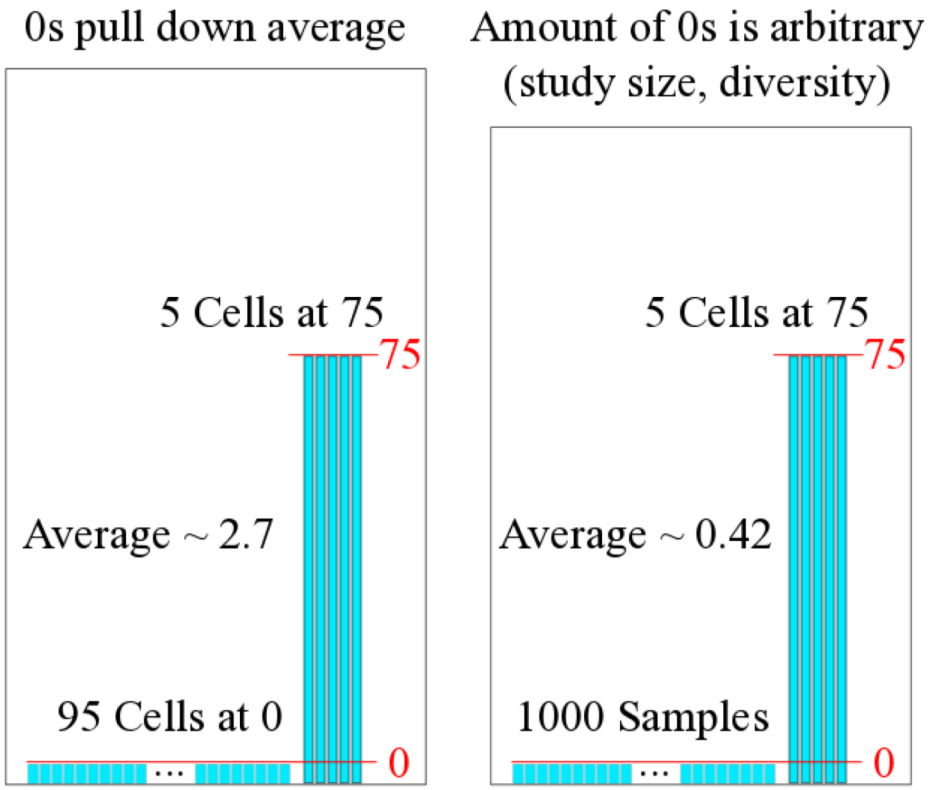
Genes have different distributions



Some single-cell RNA-seq data challenges to remember

- **Drop out**: data has an excessive amount of zeros due to limiting mRNA

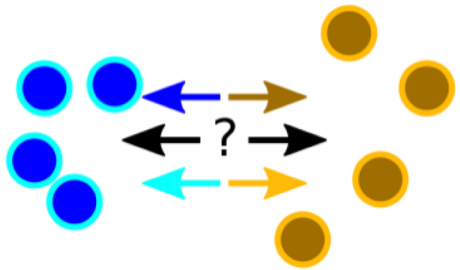
  Zero expression doesn't mean the gene isn't on



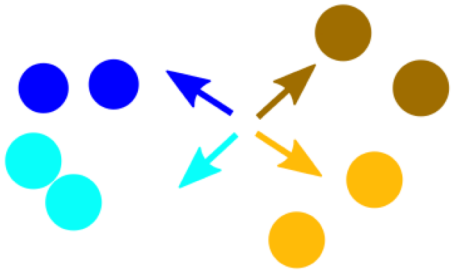
Some single-cell RNA-seq data challenges to remember

- **Confounding**: quality control metrics have the potential to be confounded with biology

Cell	Site	Treatment
1	Main	A
2	Main	A
3	Main	A
4	Main	A
5	Remote	B
6	Remote	B
7	Remote	B
8	Remote	B

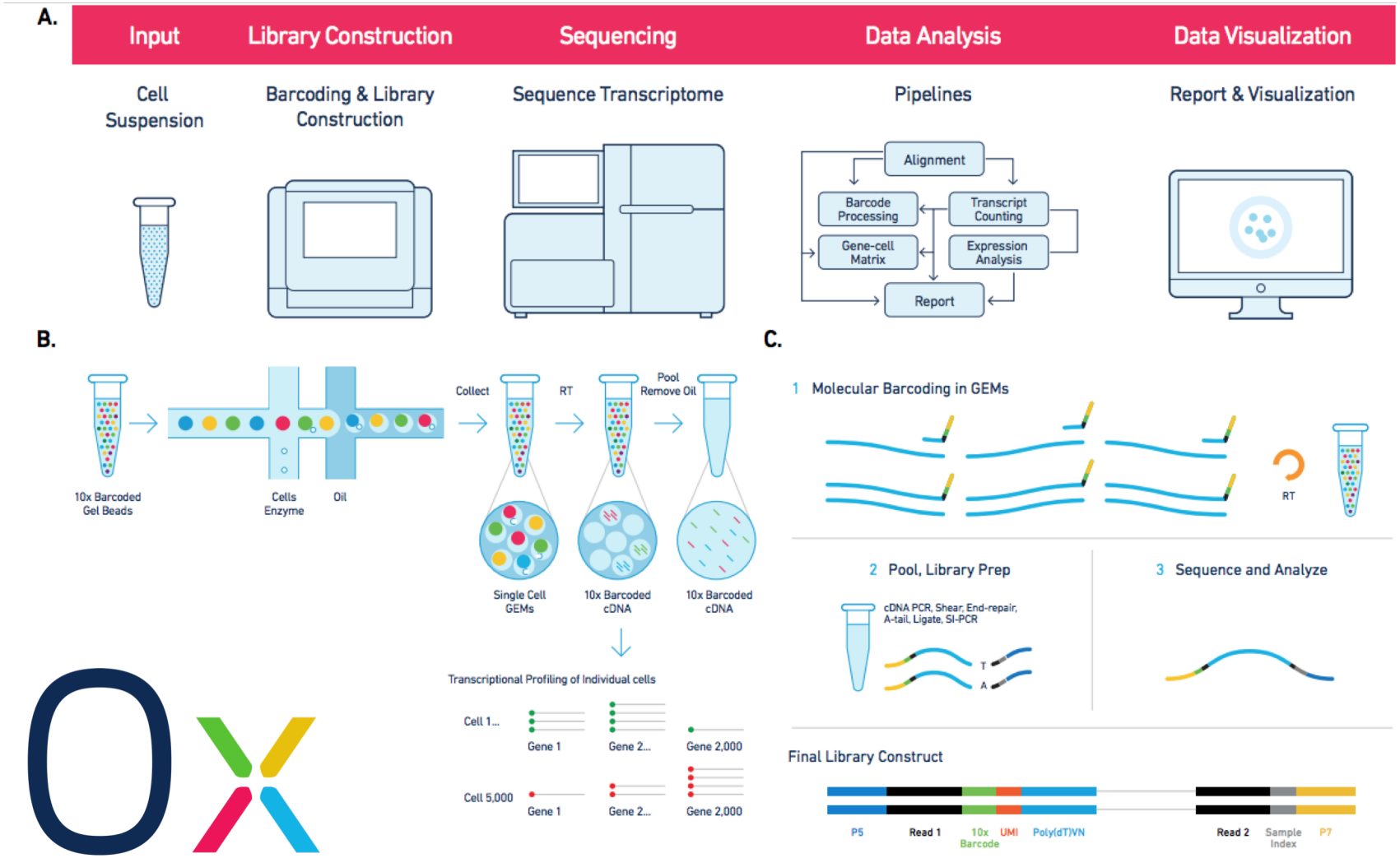


Cell	Site	Treatment
1	Main	A
2	Main	A
3	Main	B
4	Main	B
5	Remote	A
6	Remote	A
7	Remote	B
8	Remote	B



Batch effects can be removed from the data if the batch effect isn't completely confounded with biology

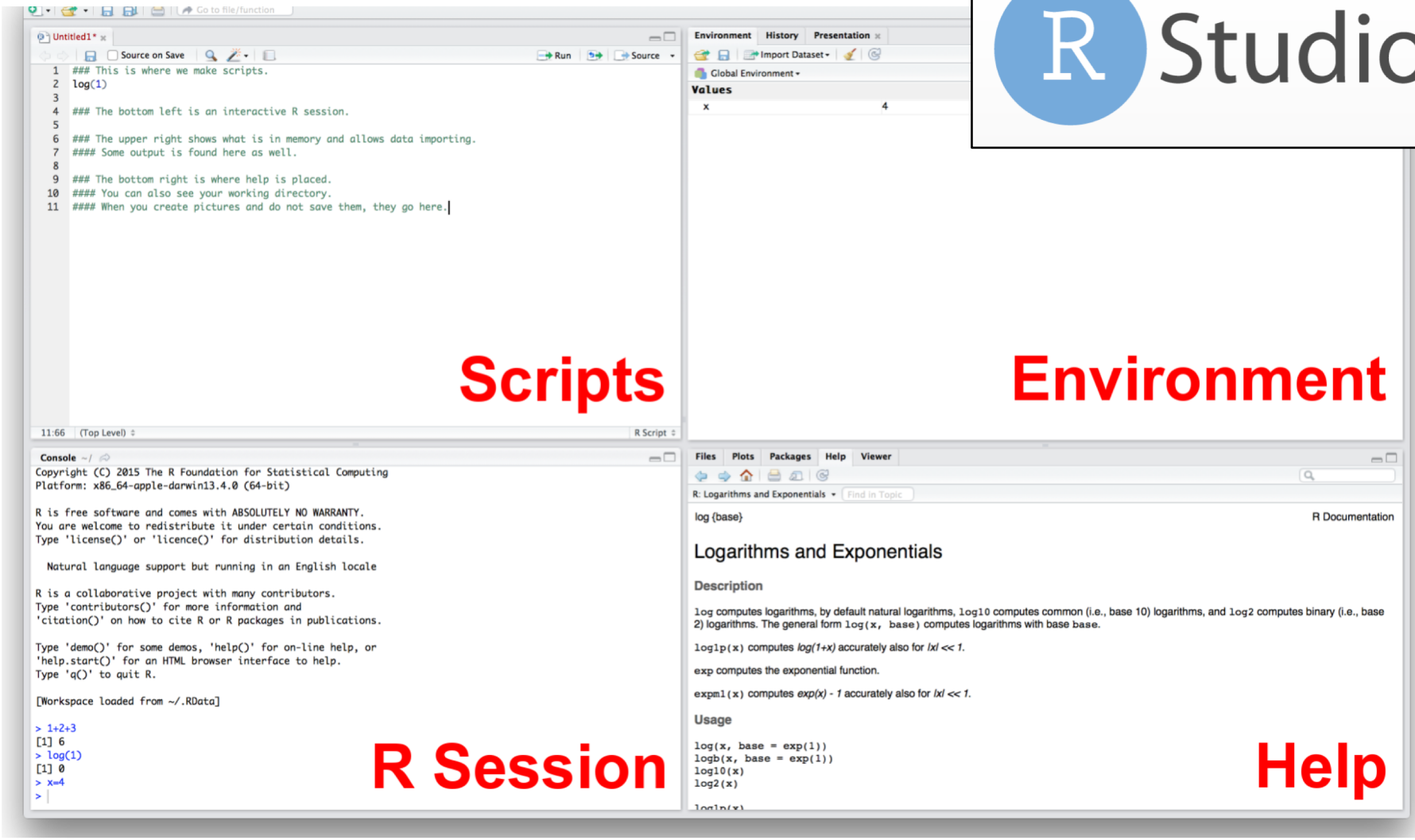
10x Genomics: The Chromium Single Cell Gene Expression Solution



10x GENOMICS®

http://go.10xgenomics.com/l/172142/2016-08-10/3svkf/172142/8090/LIT00002_RevB_Chromium_Single_Cell_3_Solution_Application_Note_Digital.pdf

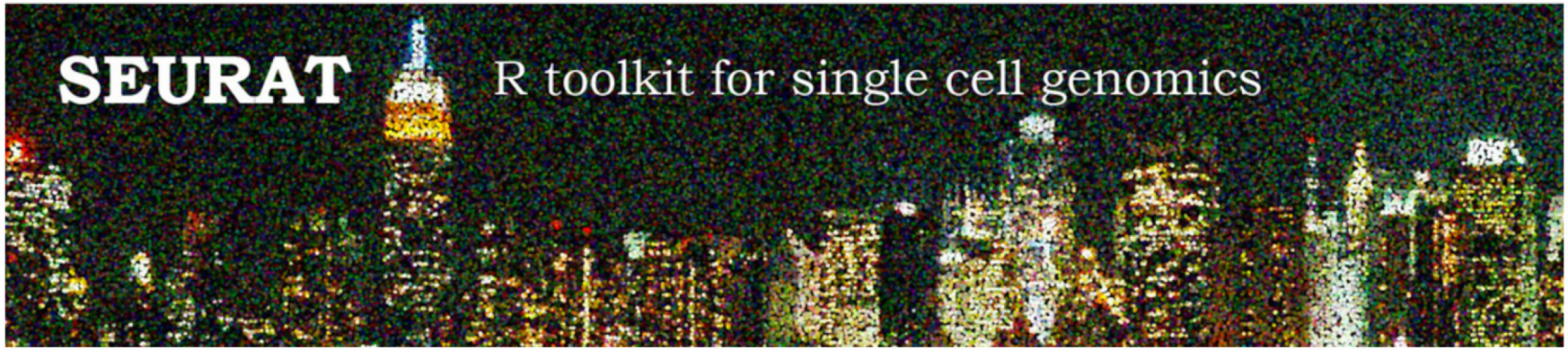
RStudio:
integrated development environment for R



Seurat
https://satijalab.org/seurat/

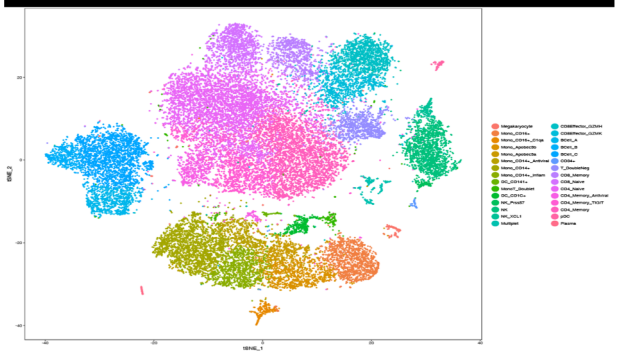
SATIJA LAB

HOME NEWS PEOPLE RESEARCH PUBLICATIONS SEURAT JOIN/CONTACT

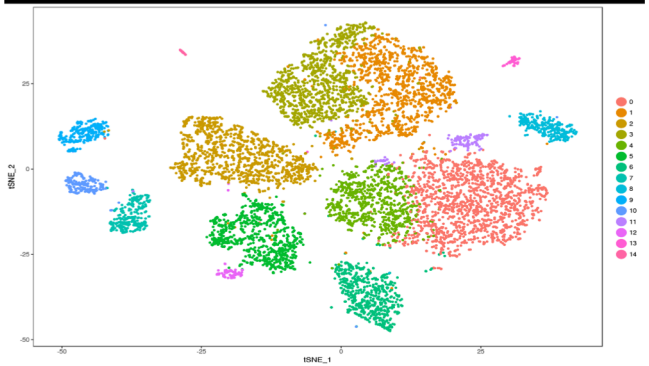


About | Install | Get Started | Frequently Asked Questions | Frequently Requested Vignettes | Contact

Command List --- 33,000 PBMCs



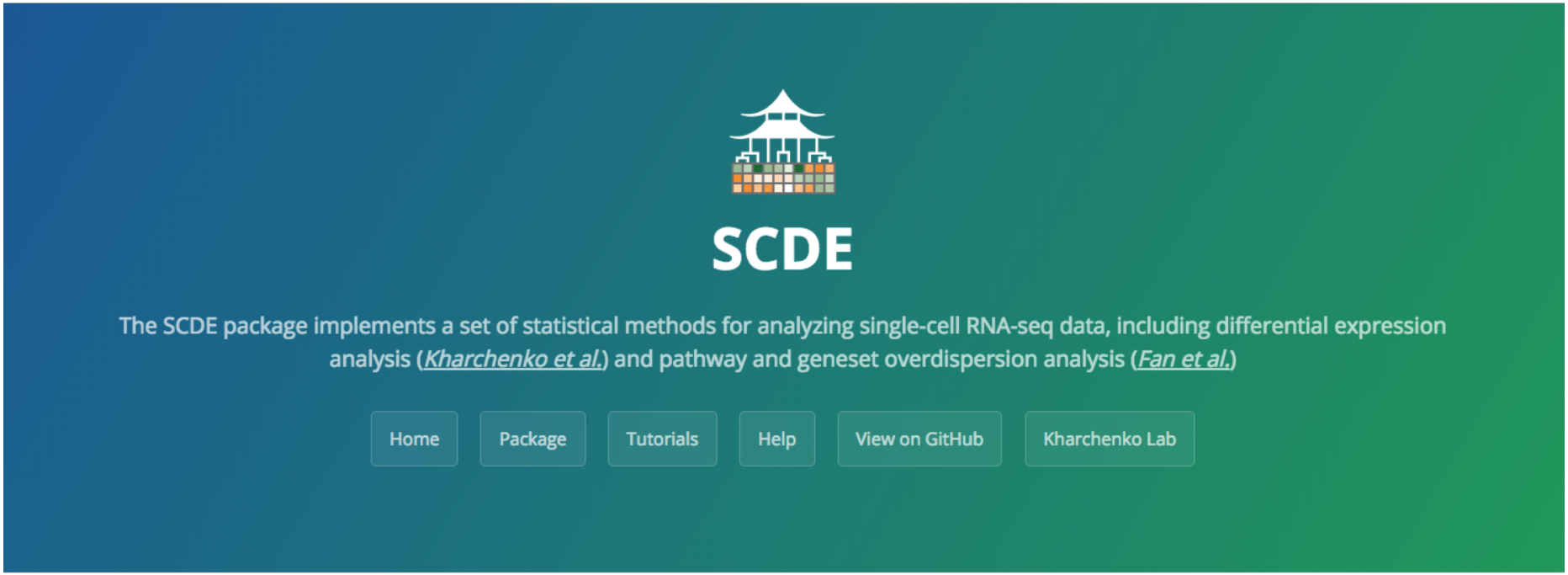
Command List --- 8,500 Pancreas Cells



Rahul Satija

D. Phil, Statistics
Oxford University
rsatija@nygenome.org

PAGODA/SCDE
http://hms-dbmi.github.io/scde/pagoda.html



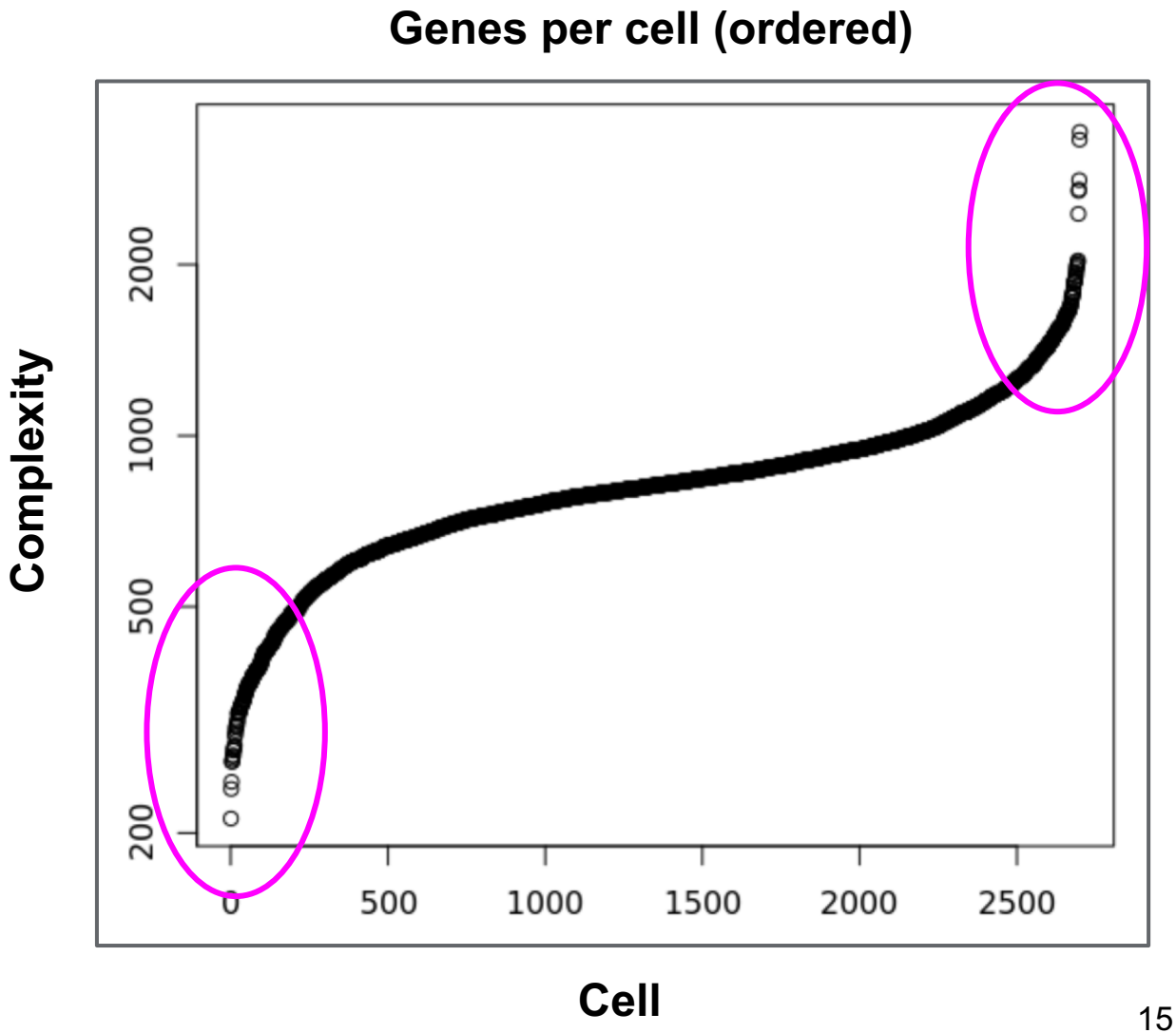
Peter Kharchenko (PI)

Assistant Professor of Biomedical Informatics | DBMI | HSCI
phone: 617-432-7377
e-mail: peter.kharchenko@post.harvard.edu
office: Countway room 312, 10 Shattuck St., Boston, MA 02115

Peter received a PhD in Biophysics at Harvard University, studying gene regulation and metabolic networks under the advisement of George Church. He then completed a four-year postdoctoral fellowship in computational biology and genomics in the laboratory of Peter Park.

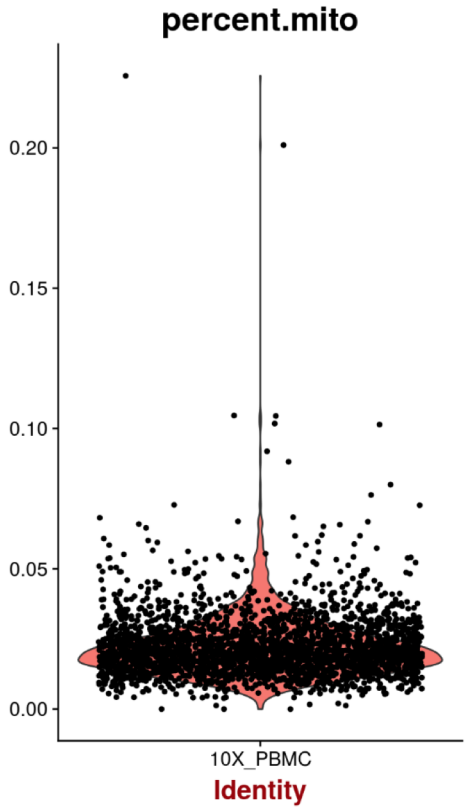
Filtering and quality control:
Number of genes per cell

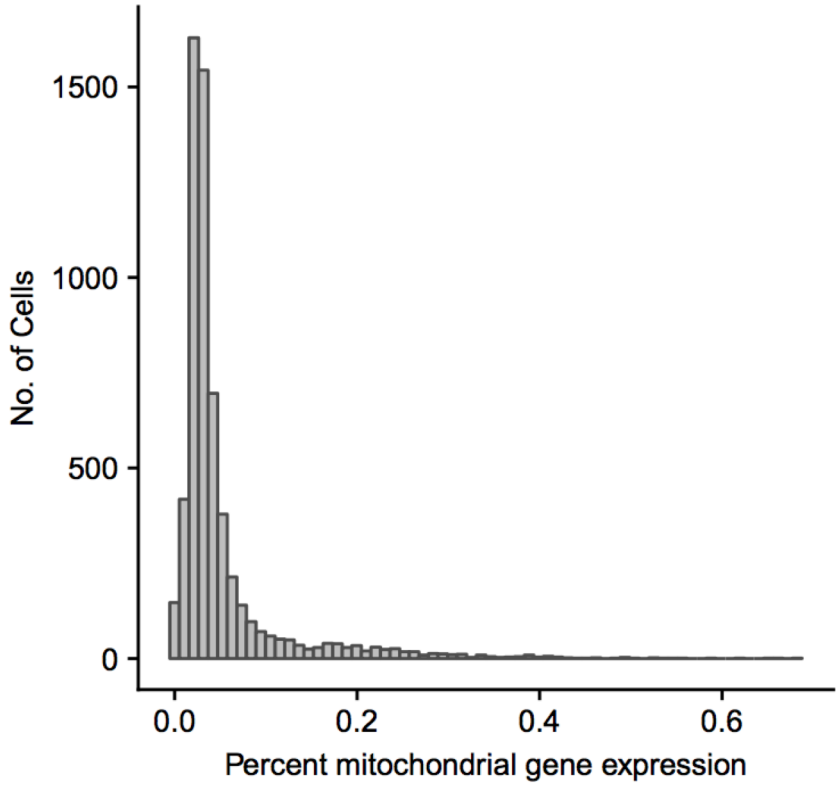
complexity =
number of genes
detected in a cell



Filtering and quality control:
Mitochondrial gene expression

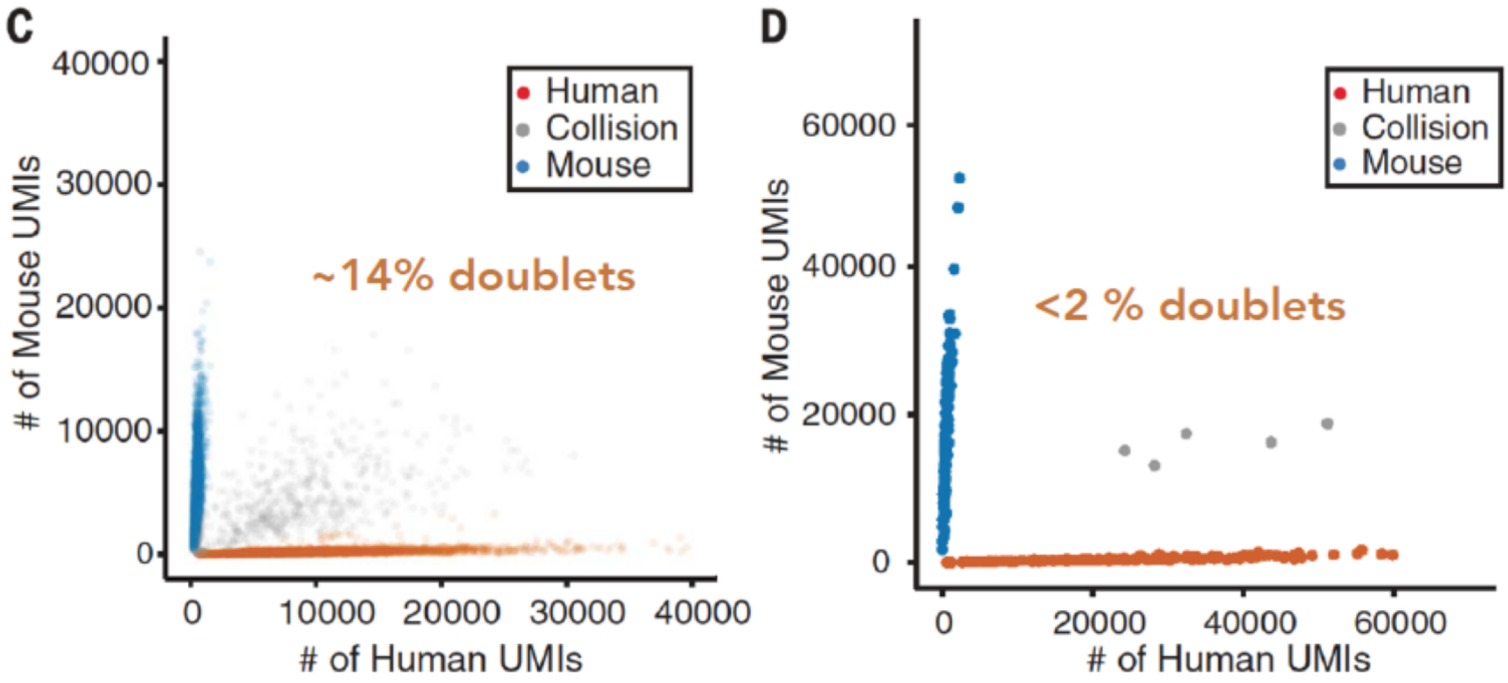
Percent of reads in a cell coming from mitochondrial genes is a good measure of cell quality - high mitochondrial gene expression indicates stressed cells (e..g, from damage during tissue dissociation)





Filtering and quality control:
Doublets - what are they?

Because of the setup, it is possible that two or more cells can enter the same droplet.
Studies estimate doublet frequency through a "mixed-species" experiment



The doublet frequency is +vely correlated with throughput

Slide courtesy of Karthik Shekhar

Filtering and quality control:
Doublets - resources for identifying them

Most simple way to filter for doublets is to choose an upper threshold on the number of genes or counts per cell in your data - a doublet (which is two cells viewed as one) should in theory have a lot more genes and counts than other cells

More sophisticated way to remove doublets is to use a package for identifying doublets, such as:

https://github.com/JonathanShor/DoubletDetection
https://github.com/AllonKleinLab/scrublet
https://www.biorxiv.org/content/early/2018/06/20/352484

DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors

Christopher S McGinnis, Lyndsay M Murrow, Zev J Gartner

doi: https://doi.org/10.1101/352484

Data normalization and scaling

Typically, we:

- Normalize gene expression for each cell by total expression and multiply by a scale factor

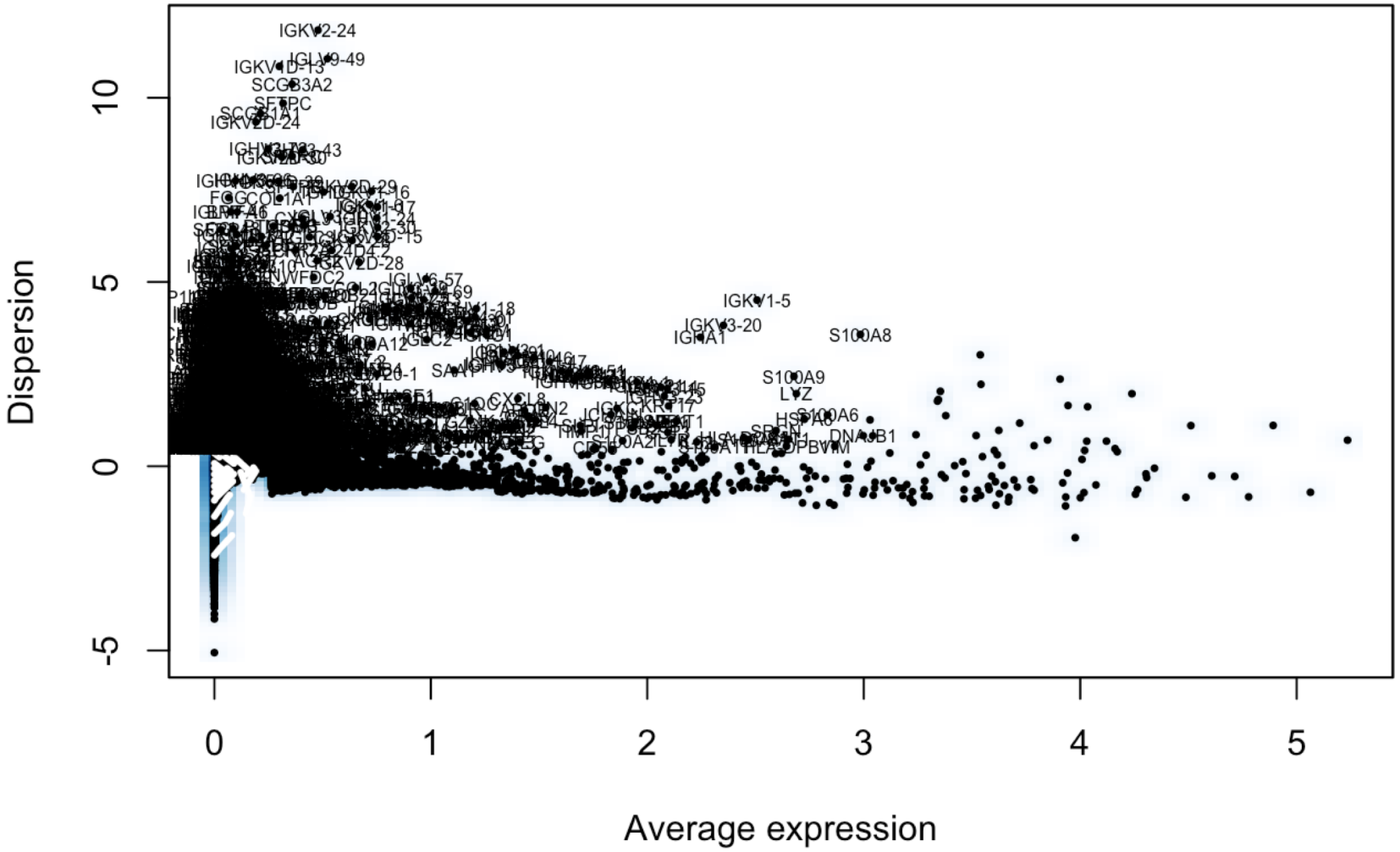
  Objective is to have relative gene expression to eliminate technical factors that impact the variation in the number of molecules per cell

  As a caution, there are biological factors that can impact this variation too

- Log transform the resulting normalized expression

  Helps get rid of extreme values in the data

Determining cell type, state, and/or function:
1. Identifying highly variable genes



Determining cell type, state, and/or function:
2. Dimensionality reduction

Cells are in 20,000 dimensional space

- many genes are lowly detected / noisy measurements
- genes are not independent of one another! rather they operate in coregulatory modules

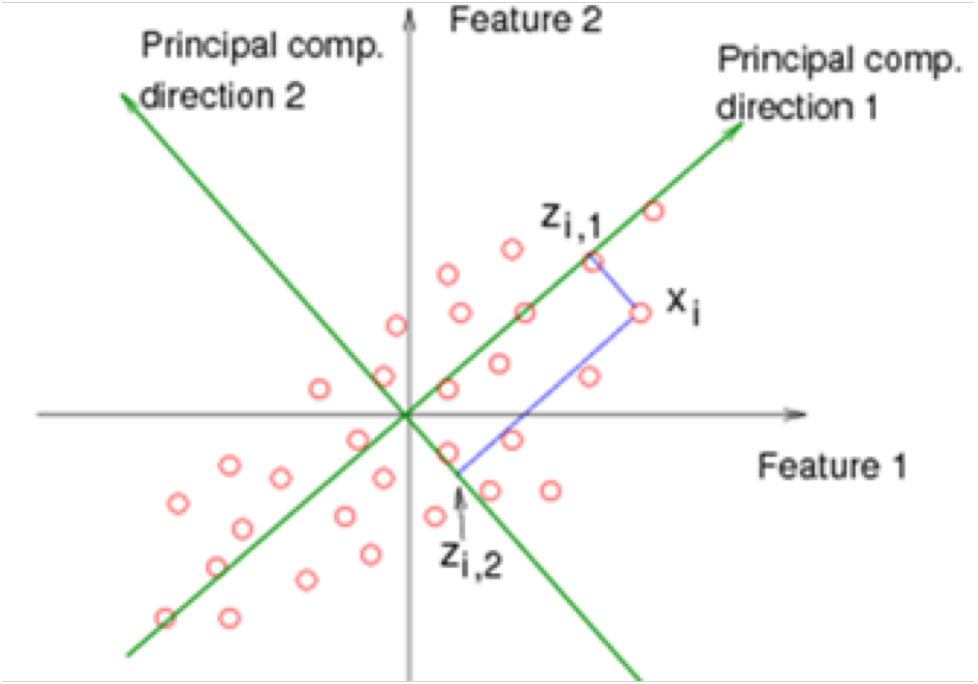
Principal component analysis (PCA) moves us from describing cells with 20,000 gene expression values to 10-100 principal component scores

*** Note that the first principal component often captures technical variability*

Determining cell type, state, and/or function:
2. Dimensionality reduction

- PCA is a dimensionality reduction method that transforms a set of observations into a set of linearly uncorrelated variables called principal components

- The first principal component contains the most variance, and each component after contains as much variance while still being orthogonal to other components

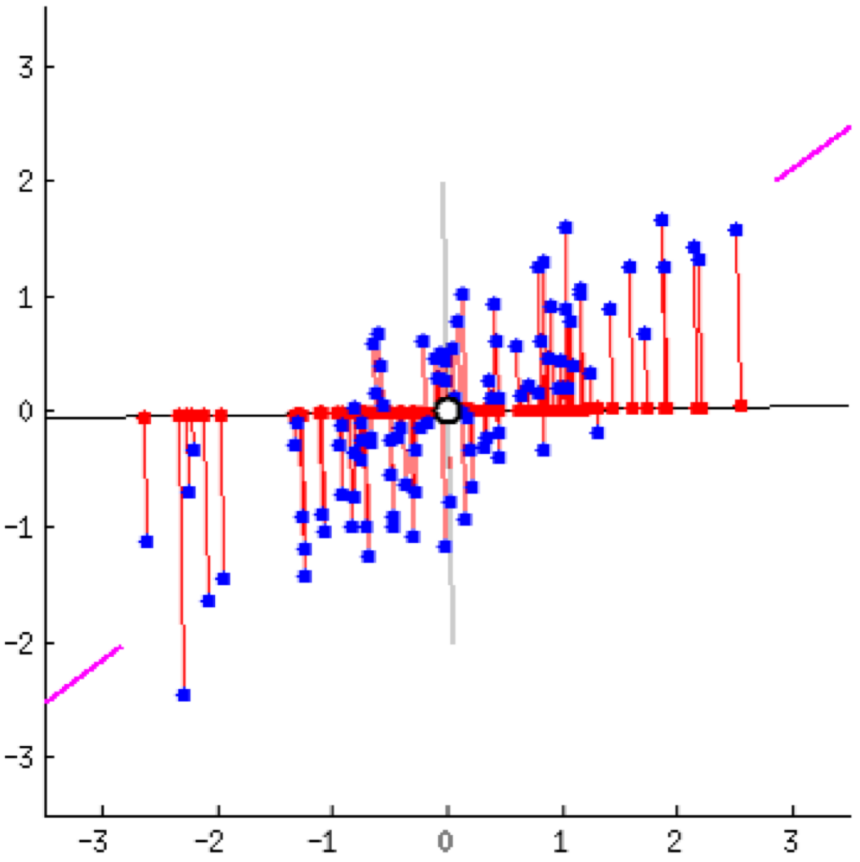


Determining cell type, state, and/or function:
2. Dimensionality reduction

- PCA is a dimensionality reduction method that transforms a set of observations into a set of linearly uncorrelated variables called principal components

- The first principal component contains the most variance, and each component after contains as much variance while still being orthogonal to other components

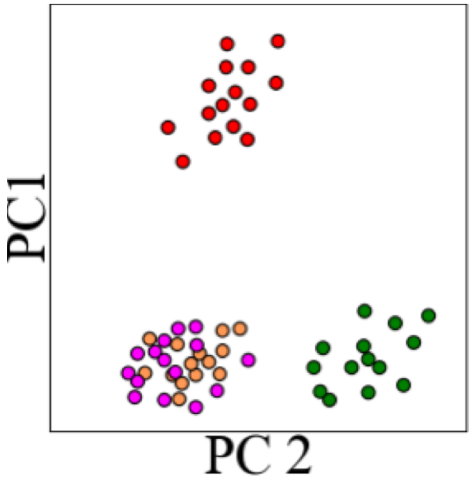
Identifying maximal orthogonal sources of variation



From: https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

Determining cell type, state, and/or function:
2. Dimensionality reduction

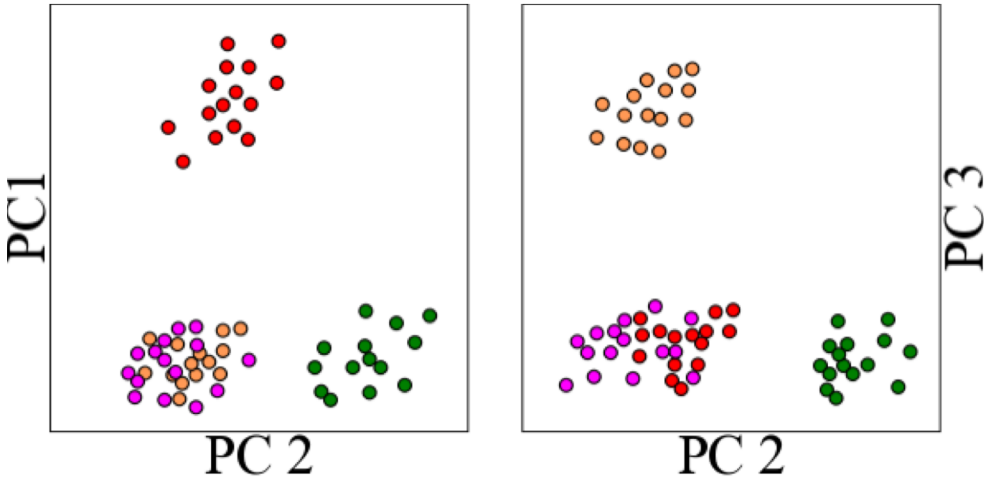
PCA of single cell data



- PC1 separates the red cells from the pink, orange, and green cells
- PC2 separates the green cells from the red, pink, and orange cells

Determining cell type, state, and/or function:
2. Dimensionality reduction

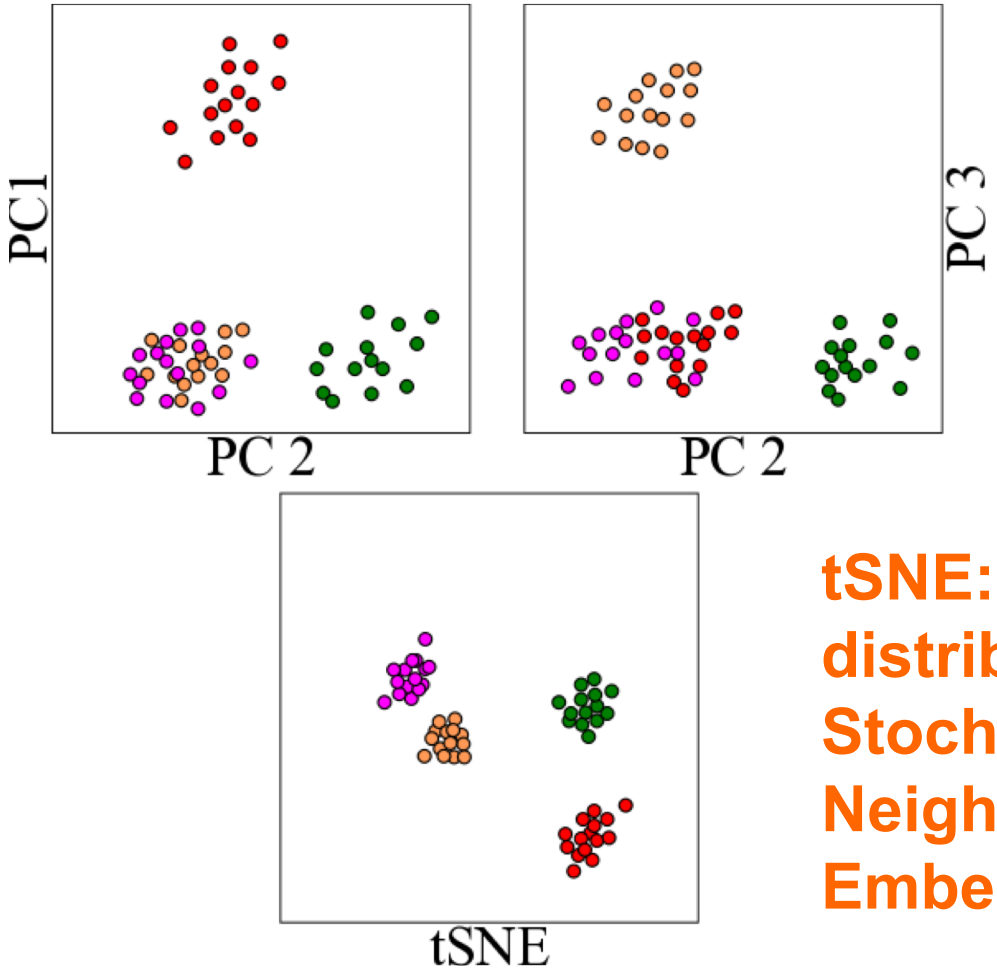
PCA of single cell data



- PC3 further splits off the orange cells

Determining cell type, state, and/or function:
2. Dimensionality reduction

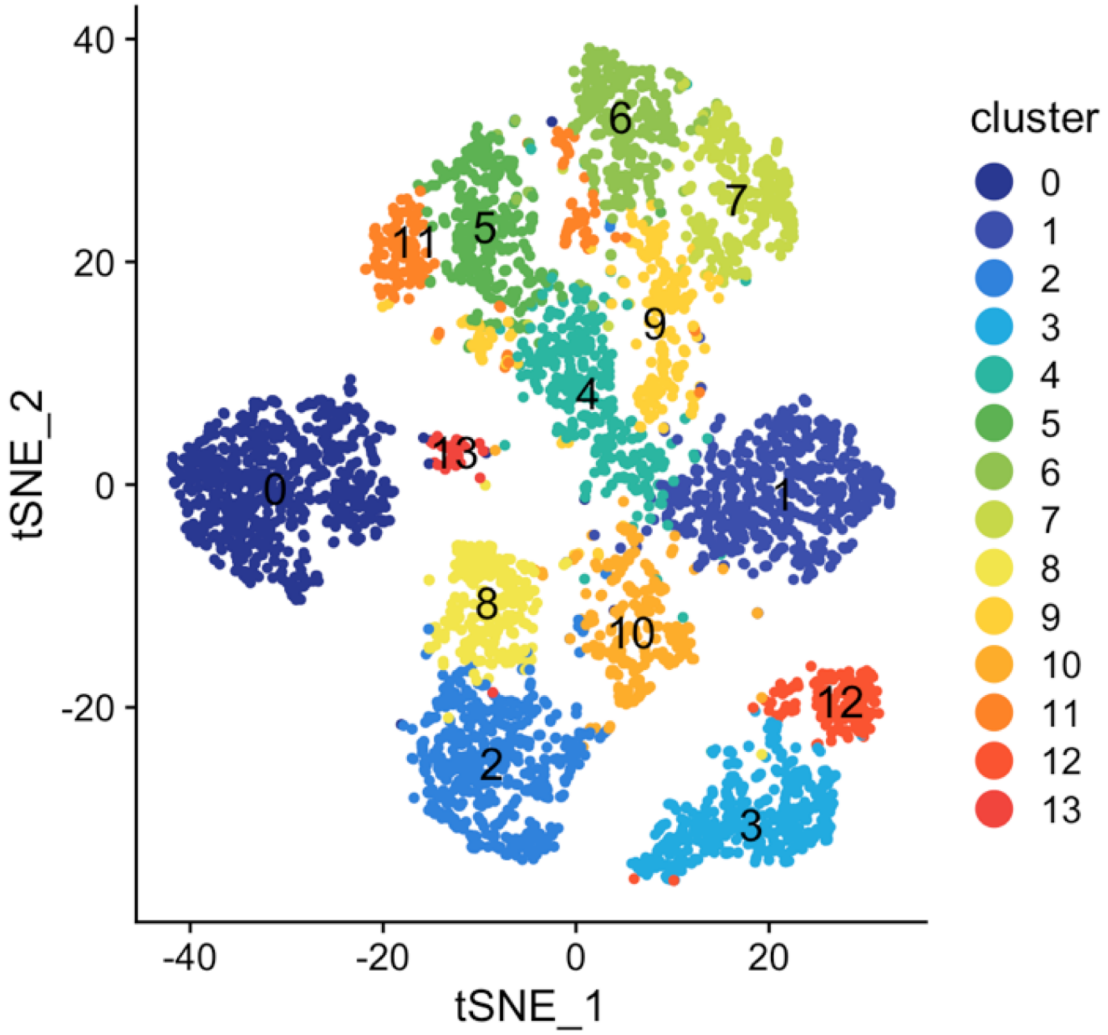
PCA of single cell data



tSNE: t-distributed Stochastic Neighbor Embedding

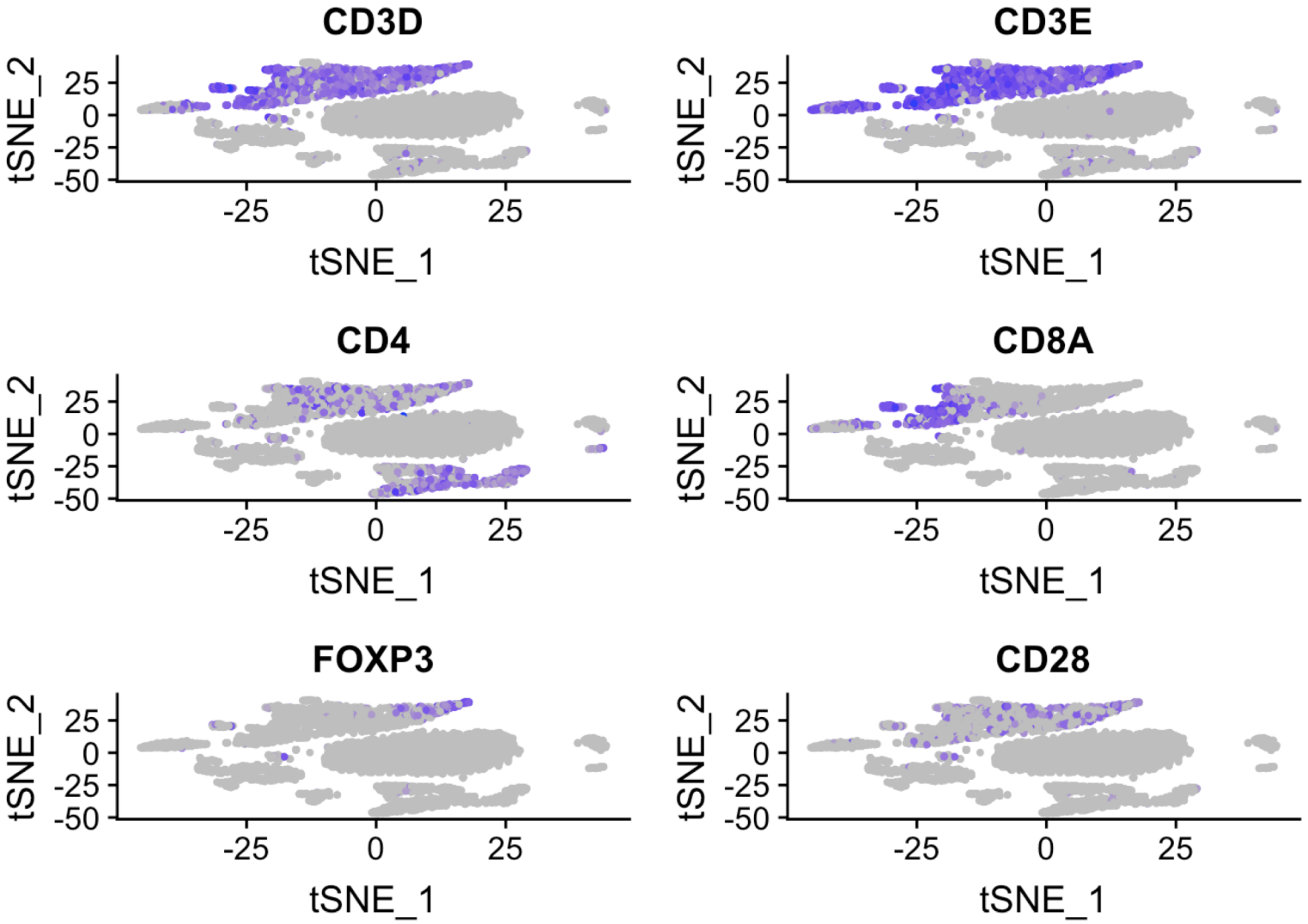
- tSNE is nonlinear dimensionality reduction
- tSNE collapse the visualization to 2D

Determining cell type, state, and/or function:
2. Dimensionality reduction

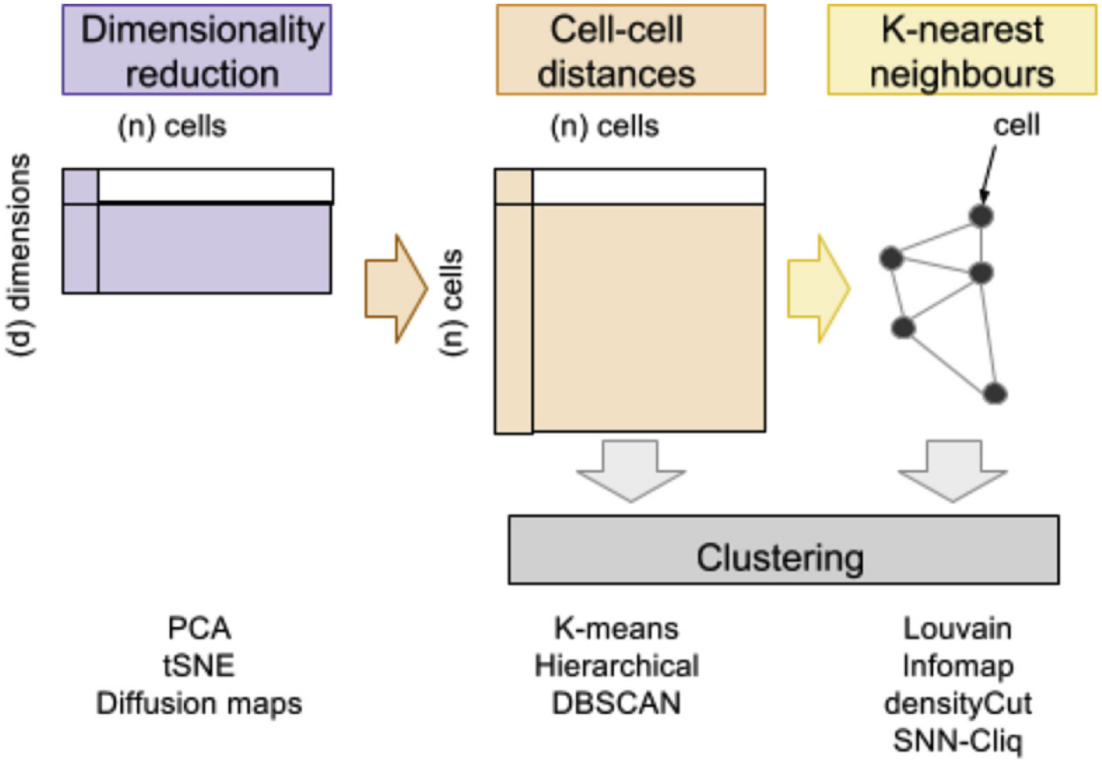


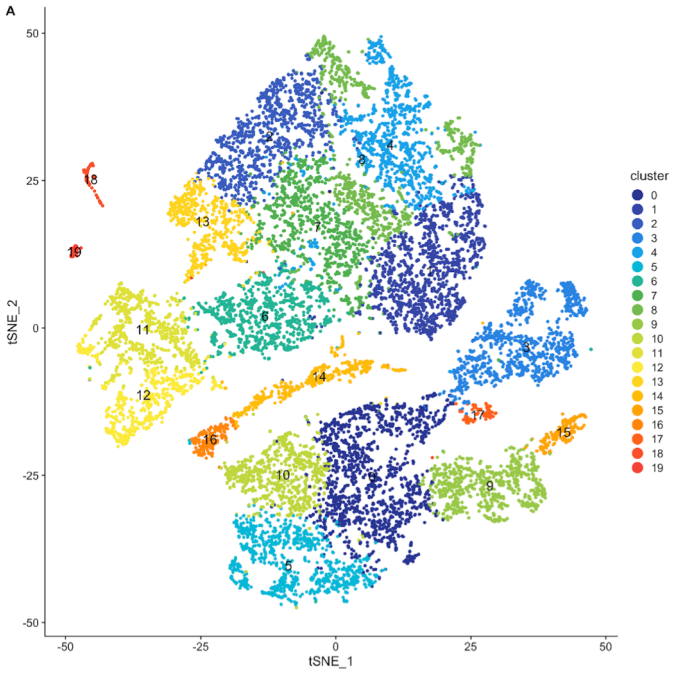
A great tSNE resource! https://distill.pub/2016/misread-tsne/

Determining cell type, state, and/or function:
3. Exploring expression of marker genes



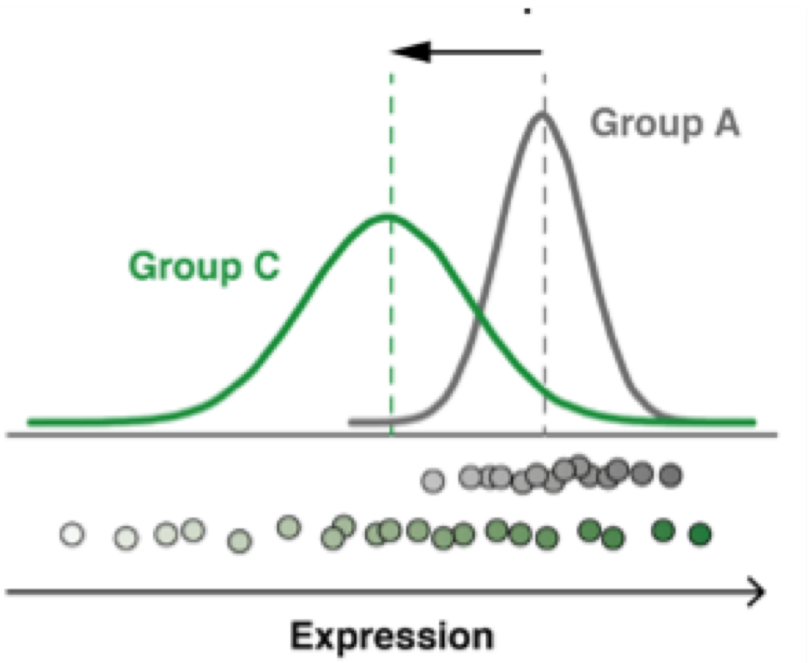
Determining cell type, state, and/or function:
4. Clustering principal components



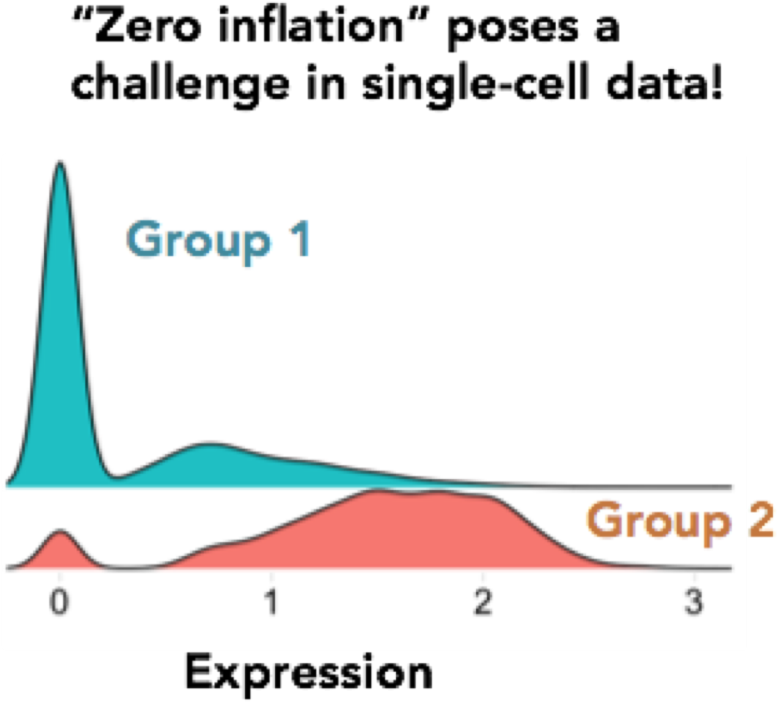


Andrews, Tallulah S. at *al.* (2017) *Molecular aspects of medicine*.

Determining cell type, state, and/or function:
5. Identifying differentially expressed genes

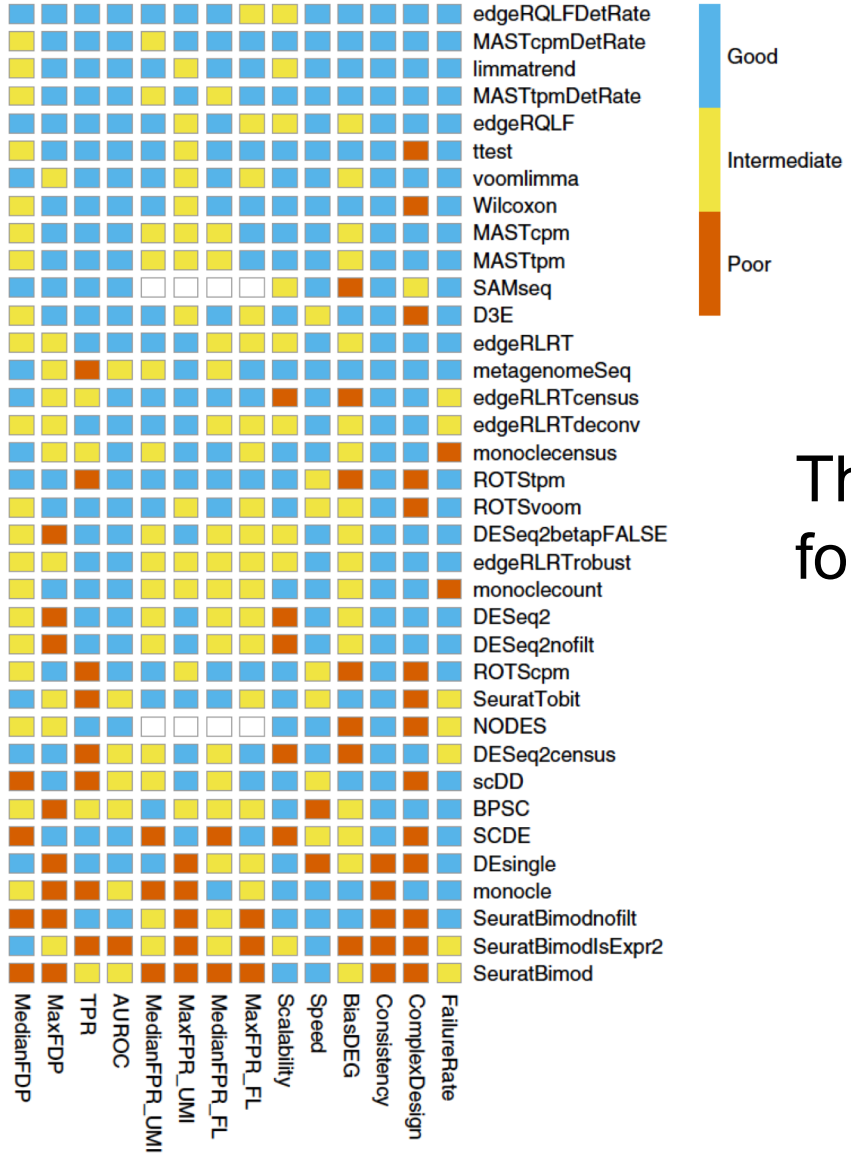


Bulk



Single cell

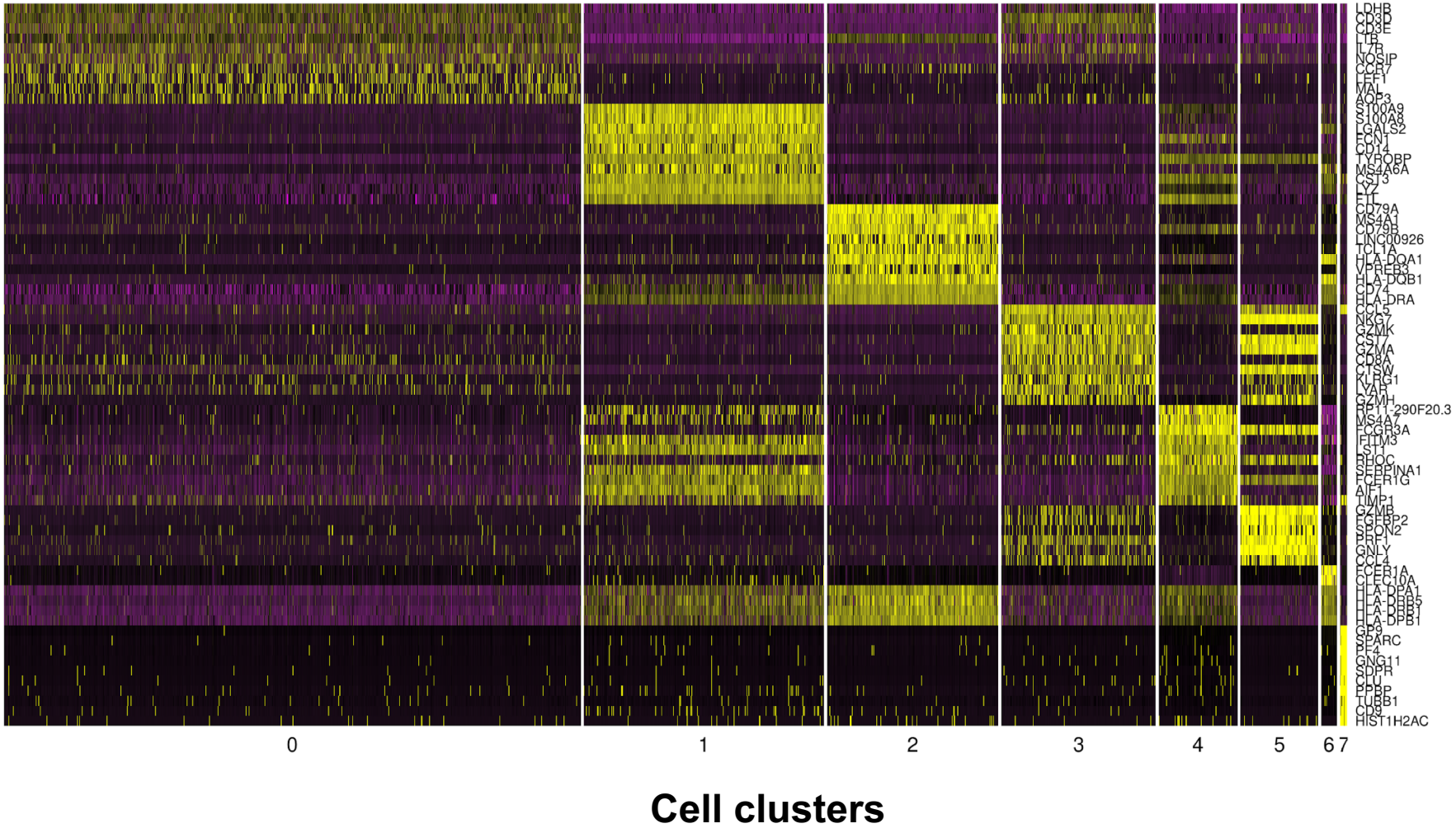
Determining cell type, state, and/or function:
5. Identifying differentially expressed genes



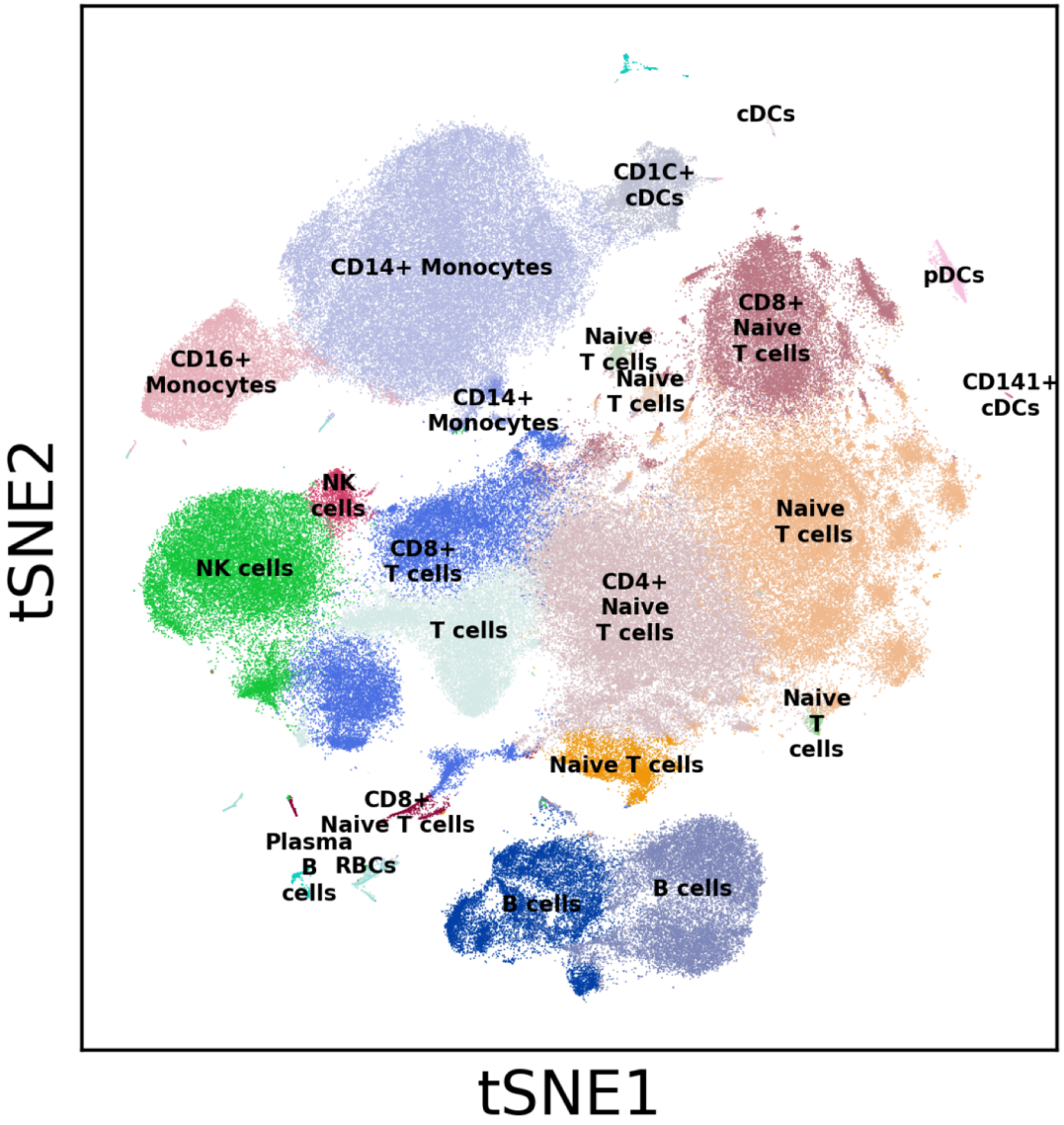
There are many ways to test for differential expression!

Soneson C. at *al.* (2018) *Nature Methods*. 15: 255-261.

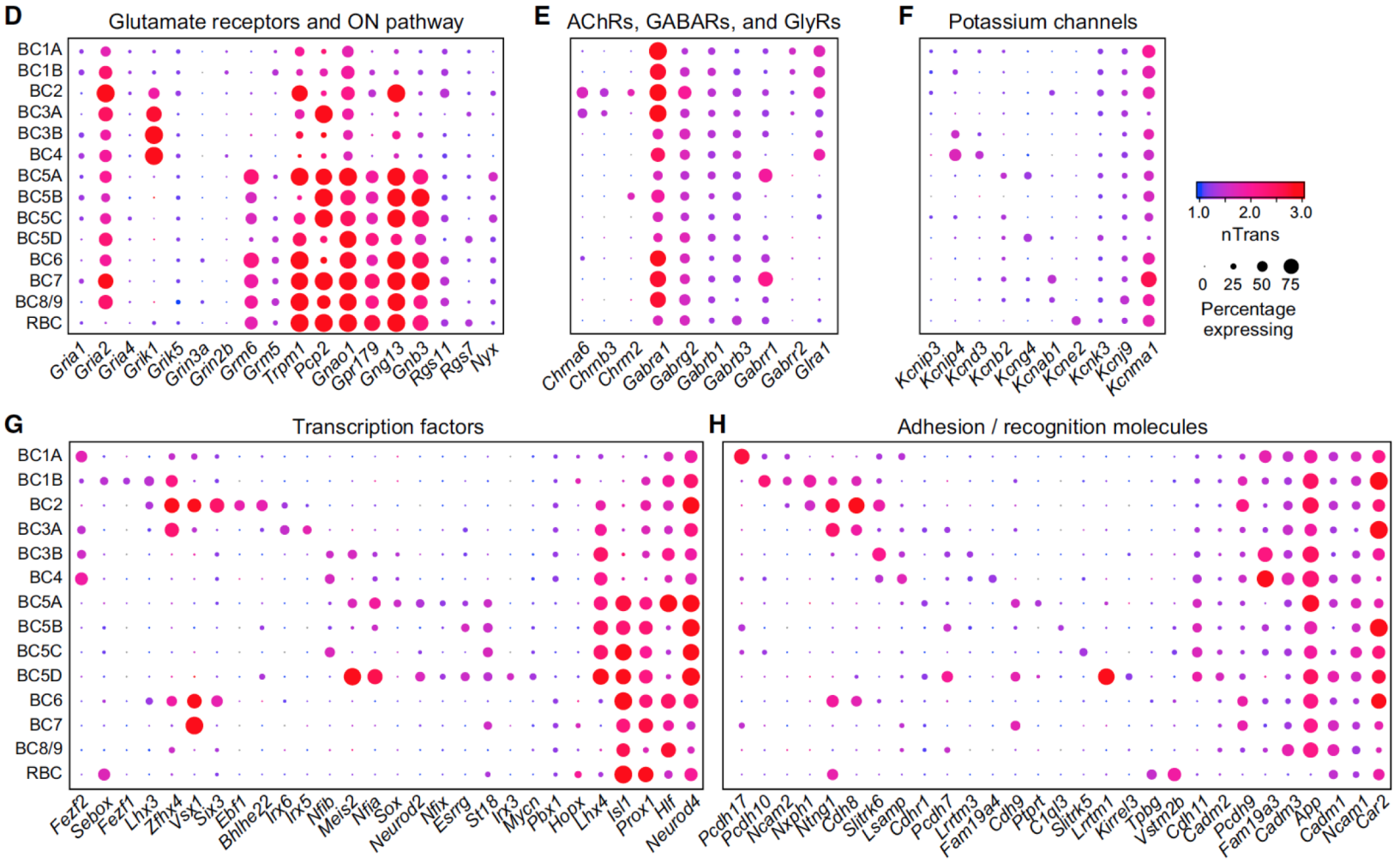
Determining cell type, state, and/or function:
5. Identifying differentially expressed genes



Determining cell type, state, and/or function:
6. Assigning cell type



Determining cell type, state, and/or function:
7. Functional annotation by pathway analysis and gene set enrichment analysis



Shekhar et al. Cell 2016

Visualizing genes of interest

Dot plots, violin plots, feature plots

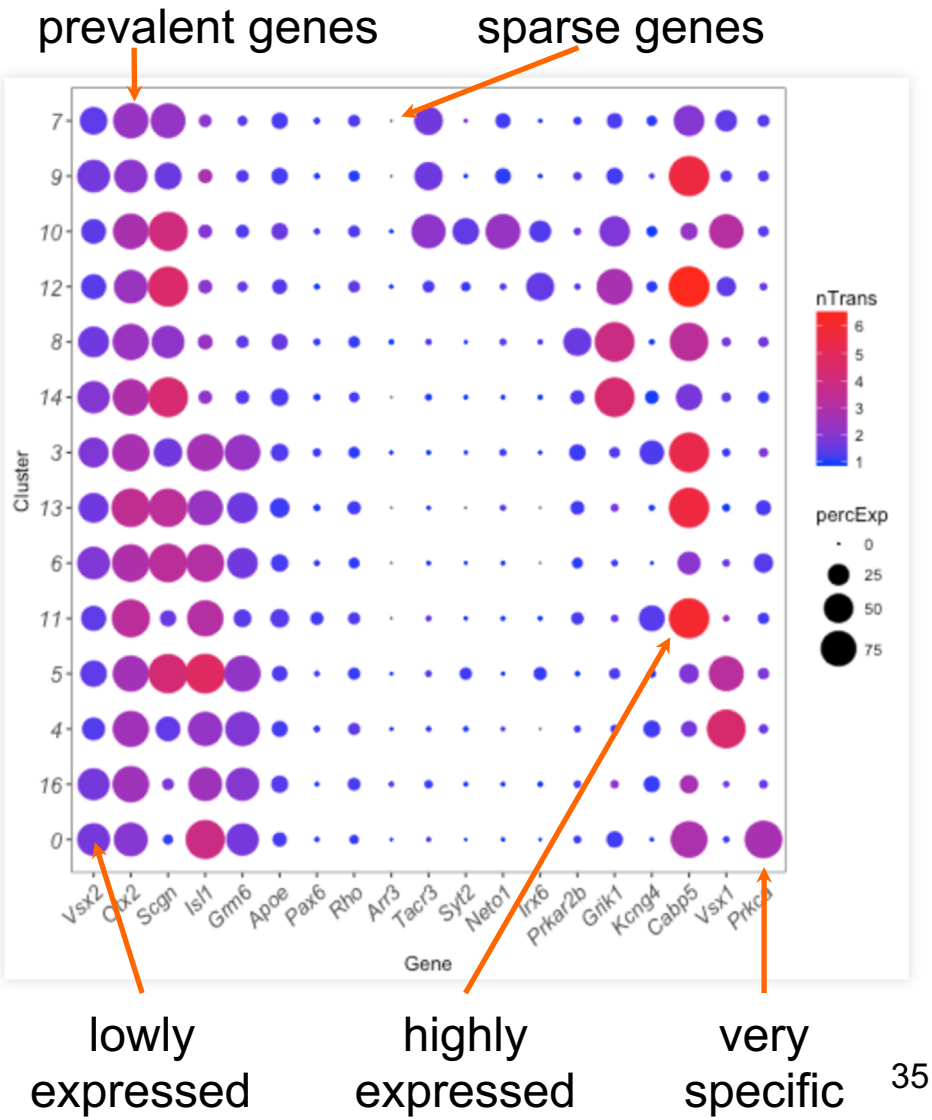
Size of circle

- Gene prevalence in cluster

Color of circle

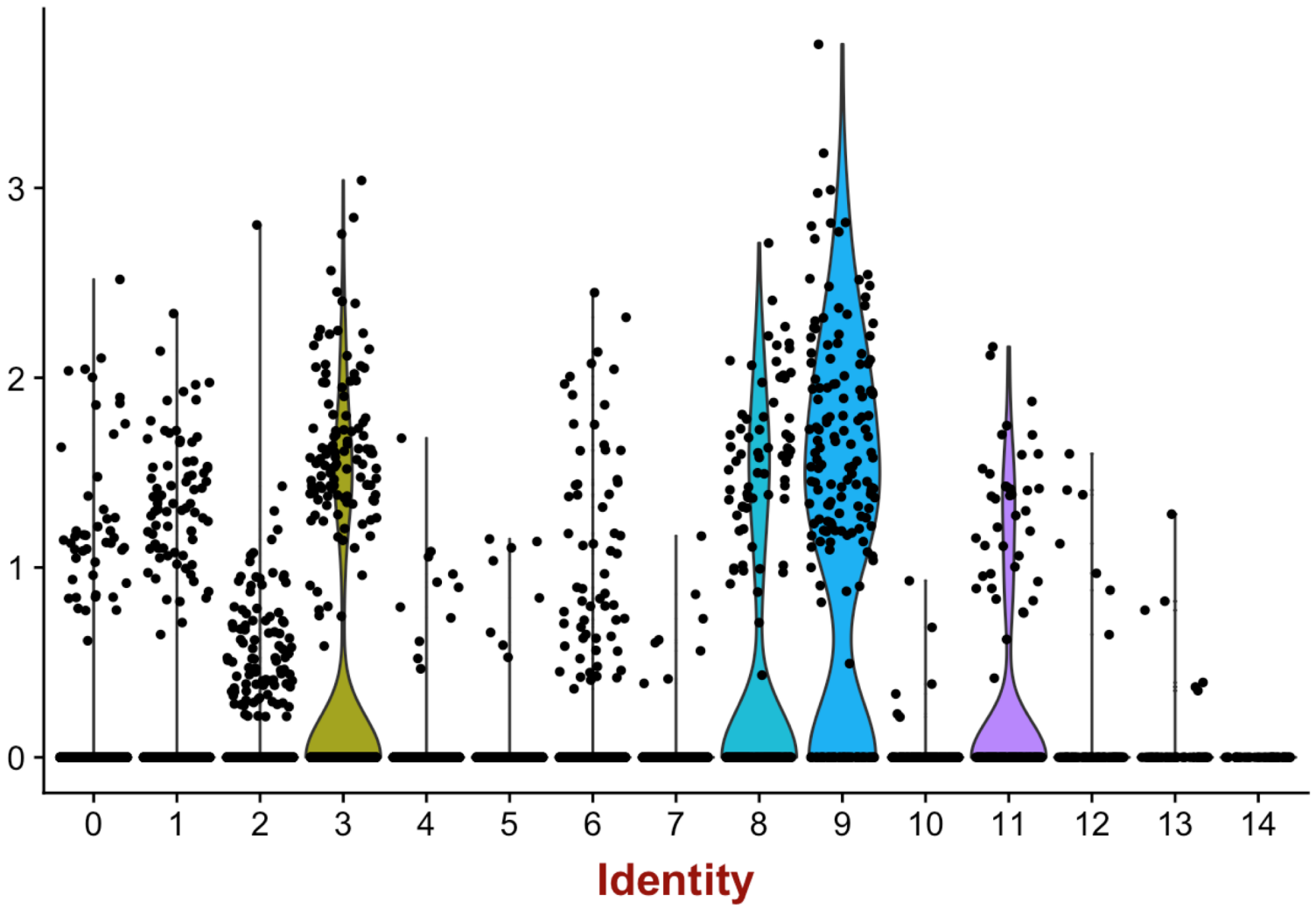
- More red, more expressed in cluster

Scales well with many cells



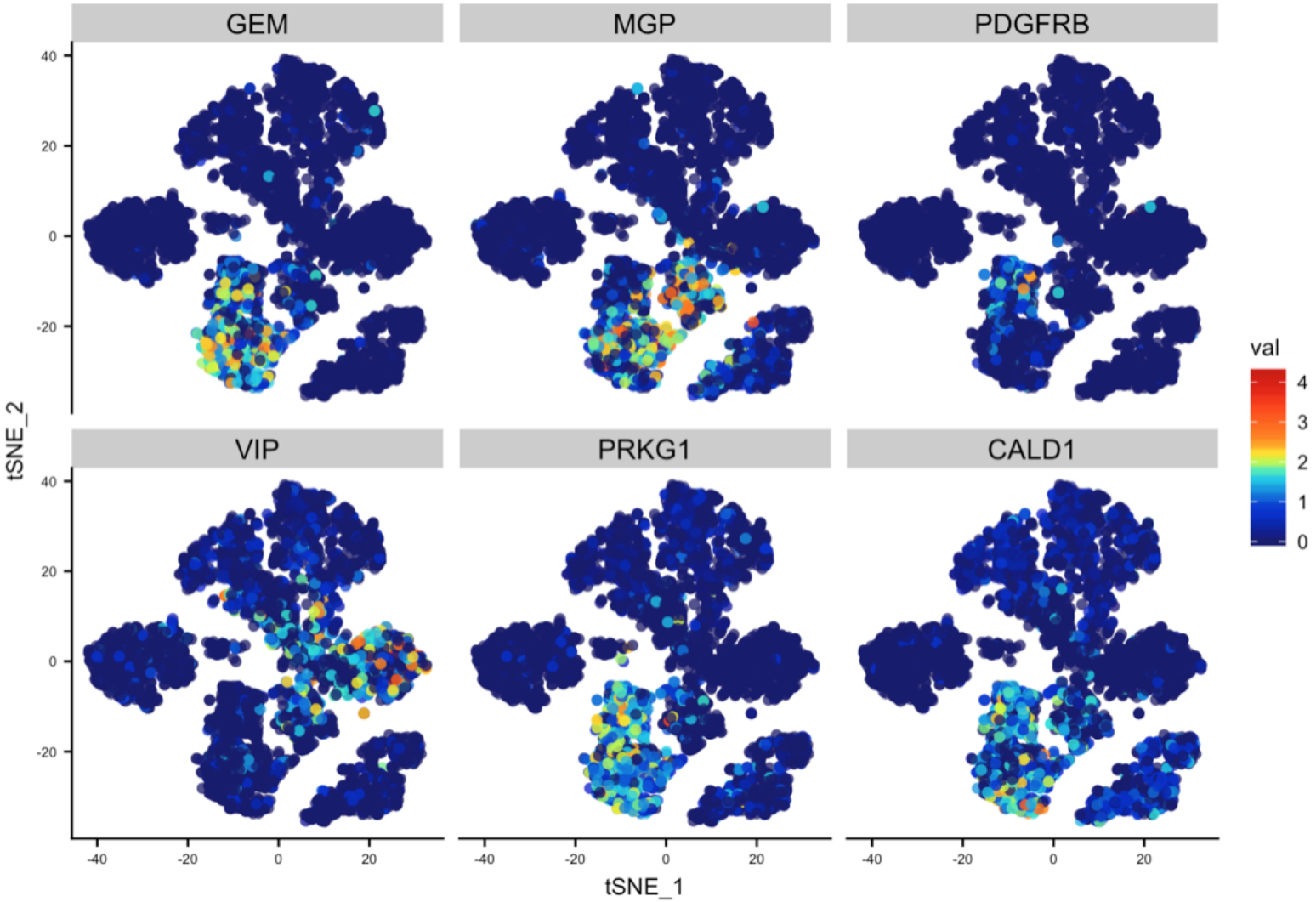
Visualizing genes of interest

Dot plots, violin plots, feature plots



Visualizing genes of interest

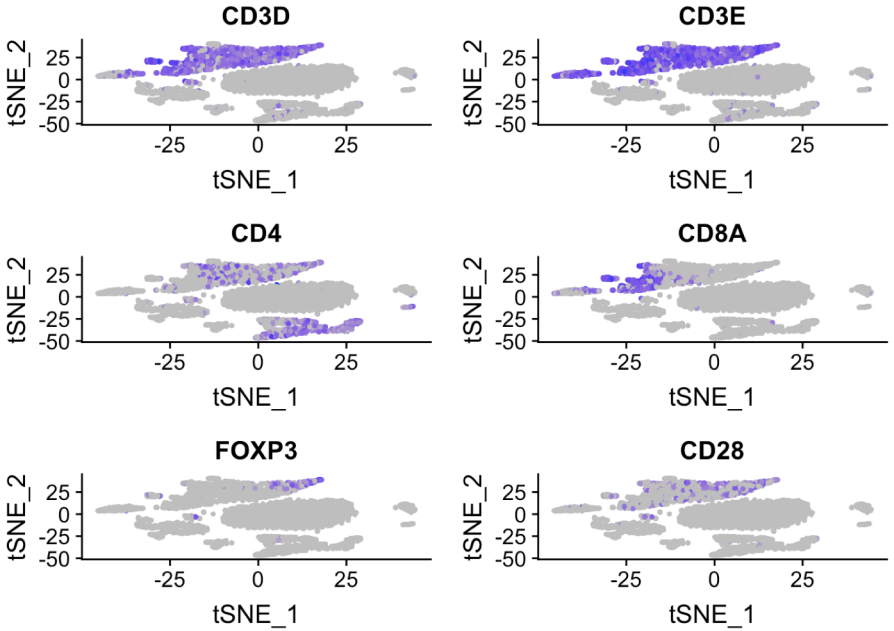
Dot plots, violin plots, feature plots

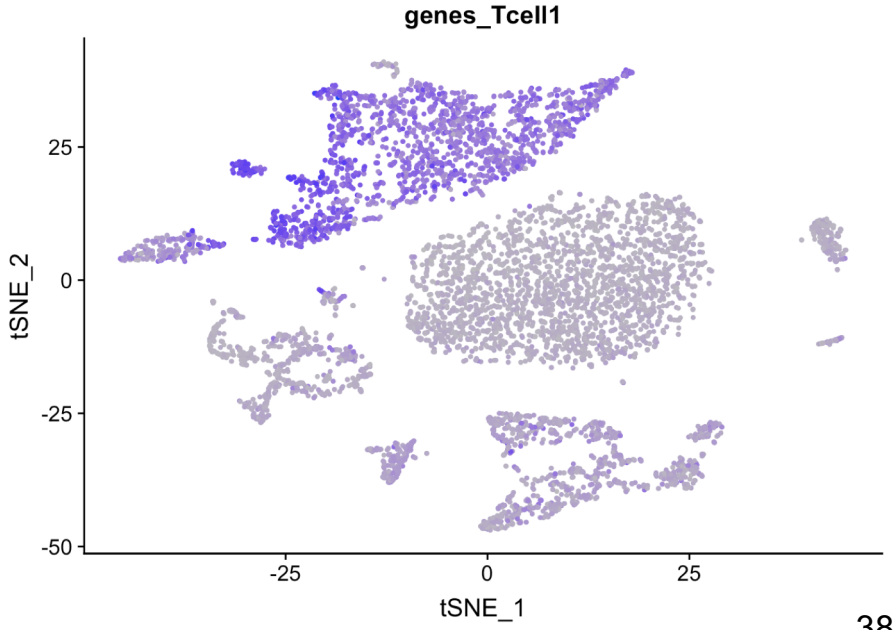


Gene signatures can be used to score each cell based on a set of genes

- Can visualize a score for each cell and look at multiple genes at once
- Done for a gene expression program of interest, e.g, cell-cycle, inflammation, cell type, dissociation
- Reduces the effects of dropouts

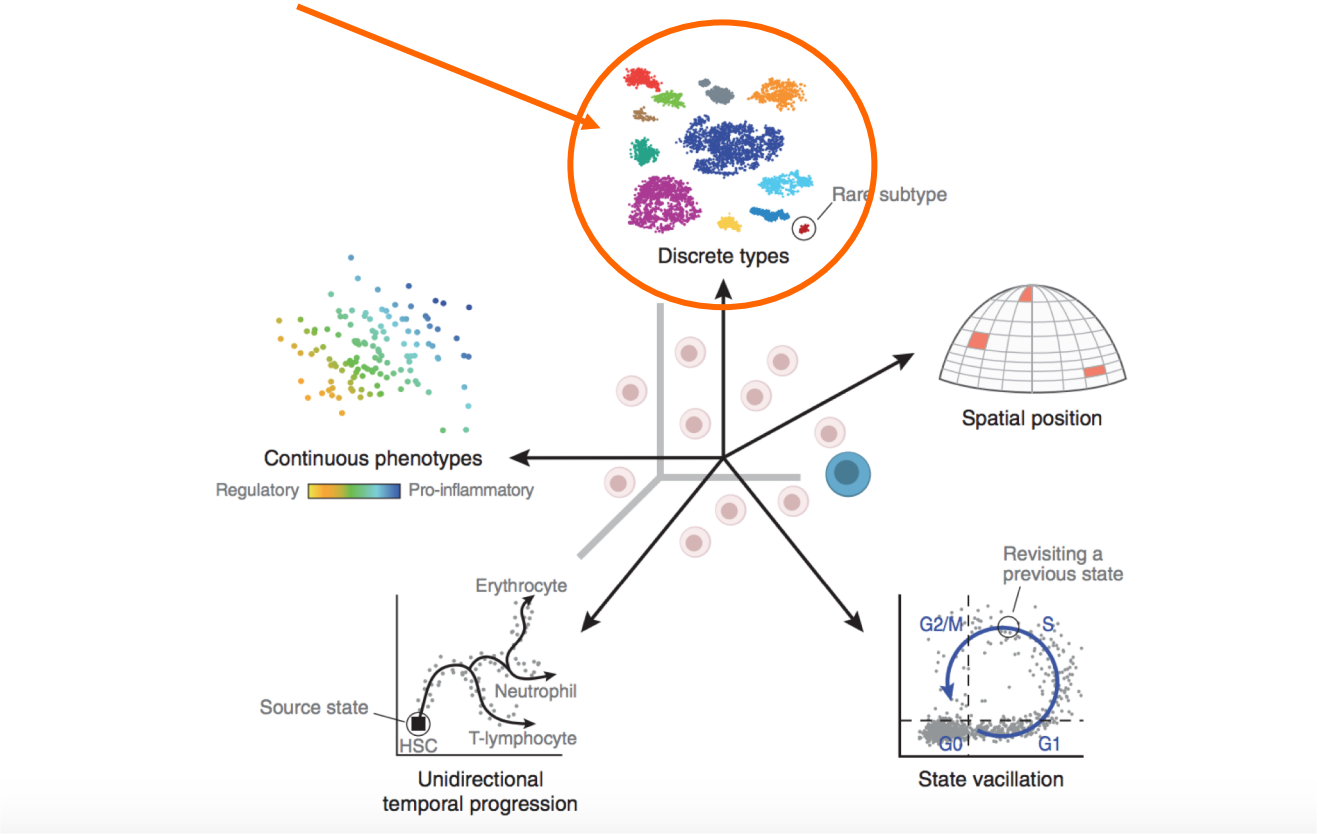
Gene signature for T cells





Recap: what did we just cover?

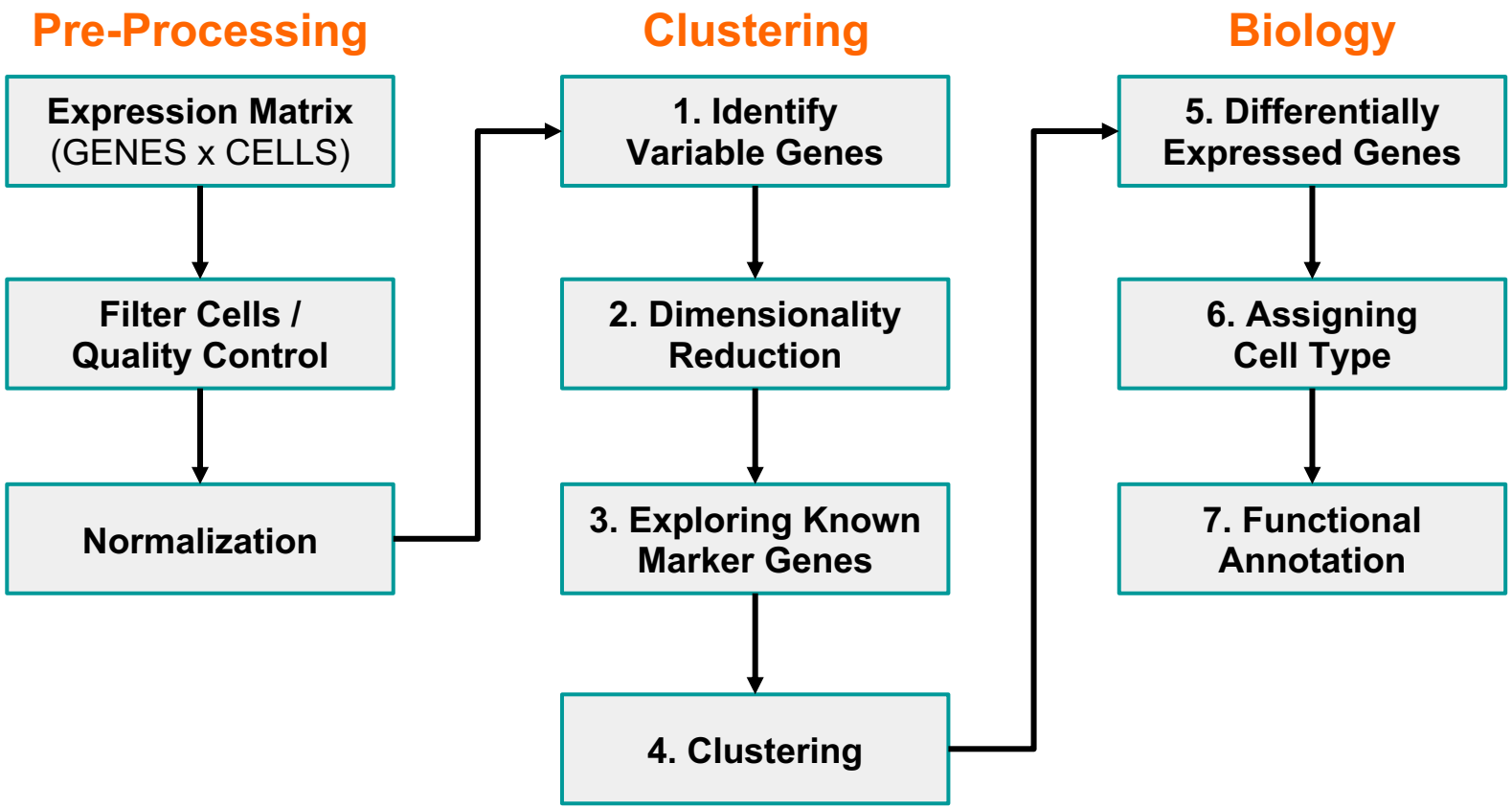
We covered just this.
So much more to learn!



Wagner, Regev and Yosef. Nat Biotech 2016

Recap: what did we just cover?

Time to execute this pipeline in a hands on example!



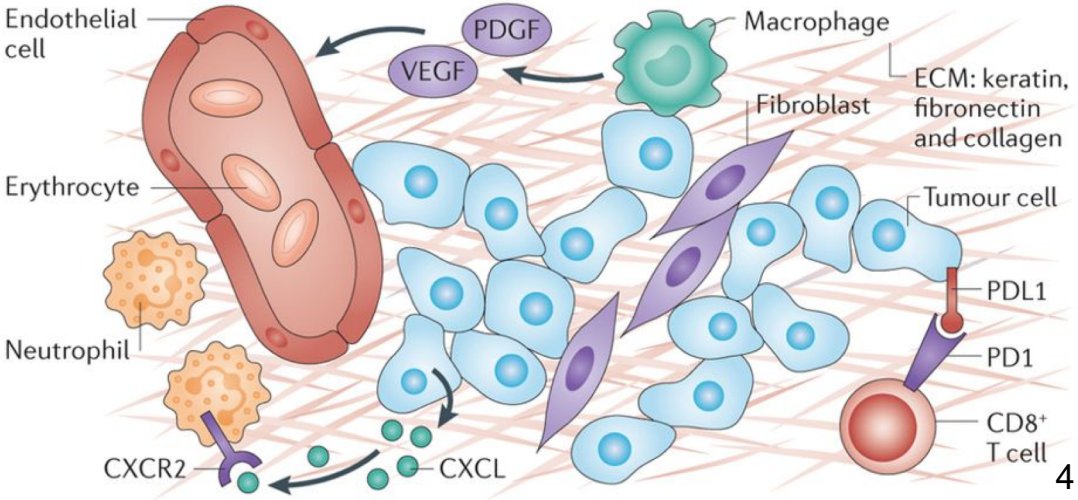
Wagner, Regev and Yosef. Nat Biotech 2016

Today's dataset: Non-Small Cell Lung Cancer

- 5' gene expression data from a fresh surgical resection of a squamous non-small cell lung carcinoma tumor (publically available on 10X website).

- American Cancer Society: Squamous cell (epidermoid) carcinoma: About 25% to 30% of all lung cancers are squamous cell carcinomas. These cancers start in early versions of squamous cells, which are flat cells that line the inside of the airways in the lungs. They are often linked to a history of smoking and tend to be found in the central part of the lungs, near a main airway (bronchus).

- Tumor microenvironment consists of malignant cells, immune cells, stromal cells, vascular networks, and extracellular matrix.



Nature Reviews Cancer volume14, pages 535–546 (2014)

Tutorial is available here:

https://github.com/broadinstitute/CEGS-2018/blob/master/src/cegs_lab.Rmd

BROAD INSTITUTE

Make sure your computer is connected to the “Broad” wireless network.

Open the Chrome Browser.

Access your gcloud Docker instance:

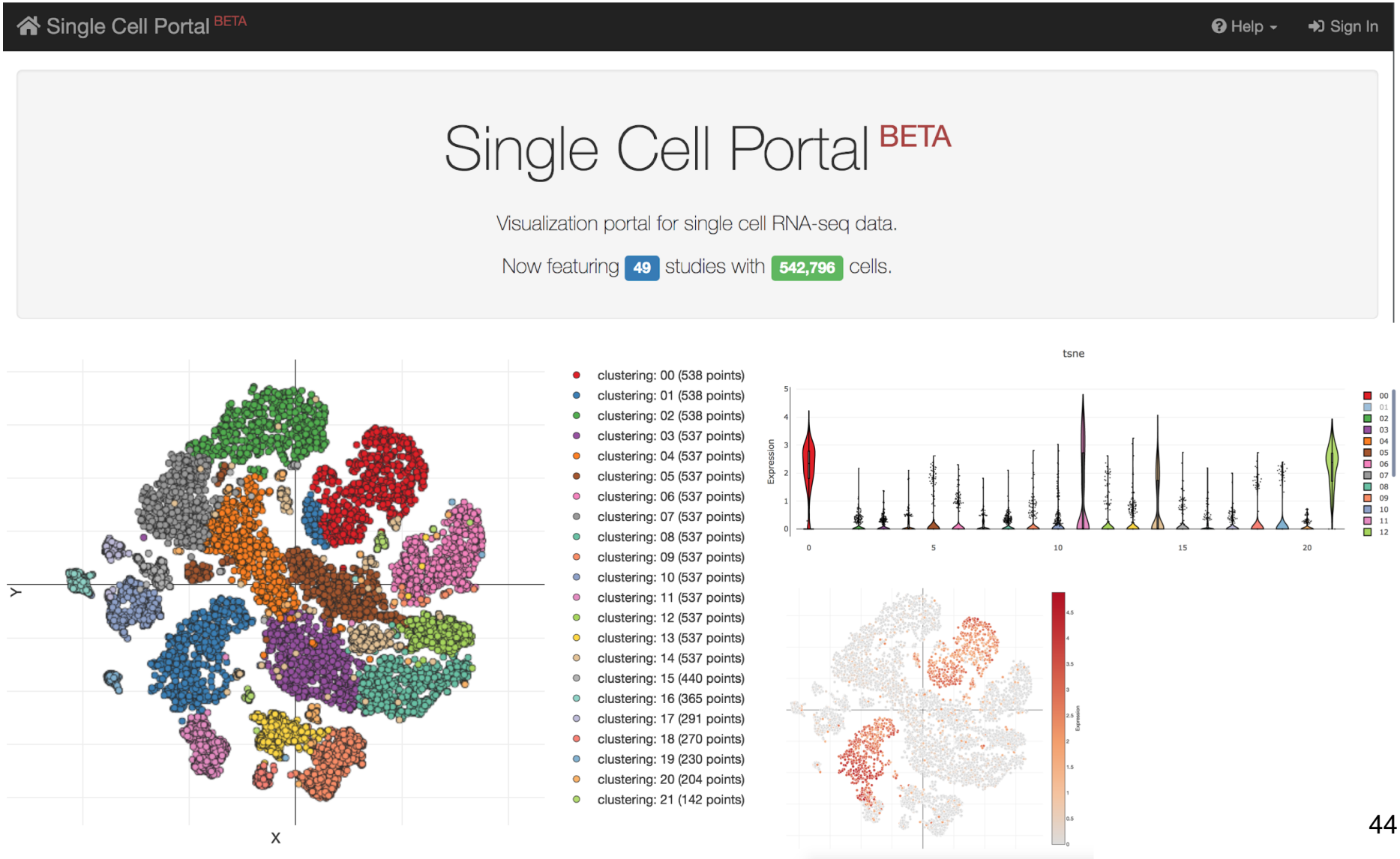
http://35.196.189.242:port

where port = 100xx, where xx is your assigned number

Username and pwd: training

BROAD INSTITUTE

Single-cell portal: facilitates sharing and dissemination of data from single-cell studies



Resources

Learn more about tSNE

- Awesome Blog on t-SNE parameterization: http://distill.pub/2016/misread-tsne
- Publication: https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf
- Nice YouTube Video: https://www.youtube.com/watch?v=RJVL80Gg3lA
- Code: https://lvdmaaten.github.io/tsne/
- Interactive Tensor flow: http://projector.tensorflow.org/

Computational packages for single-cell analysis

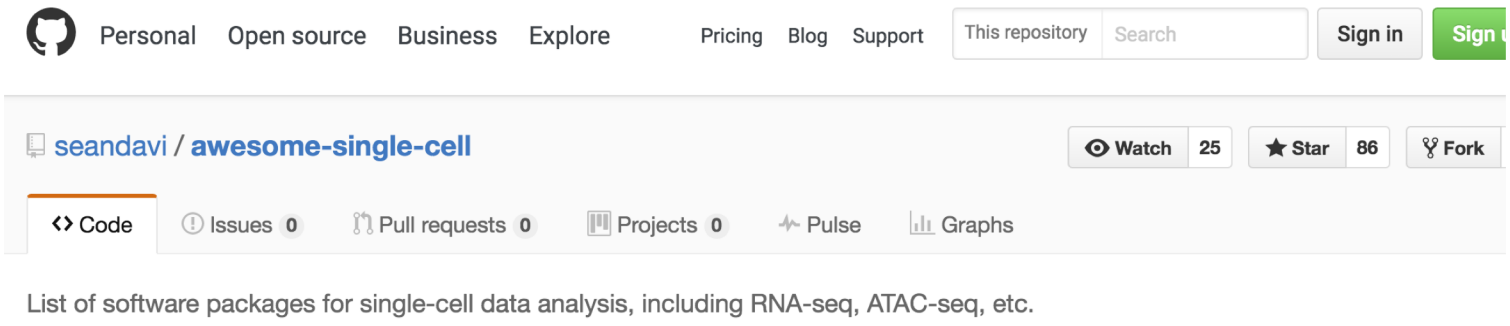
- http://bioconductor.org/packages/devel/workflows/html/simpleSingleCell.html
- https://satijalab.org/seurat/
- https://scanpy.readthedocs.io/

Online courses

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

Resources, cont.

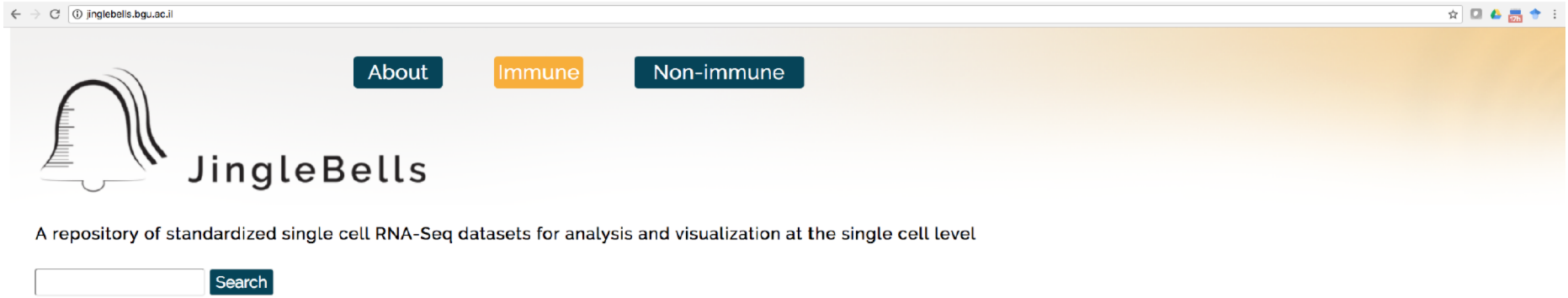
Comprehensive list of single-cell resources:
https://github.com/seandavi/awesome-single-cell



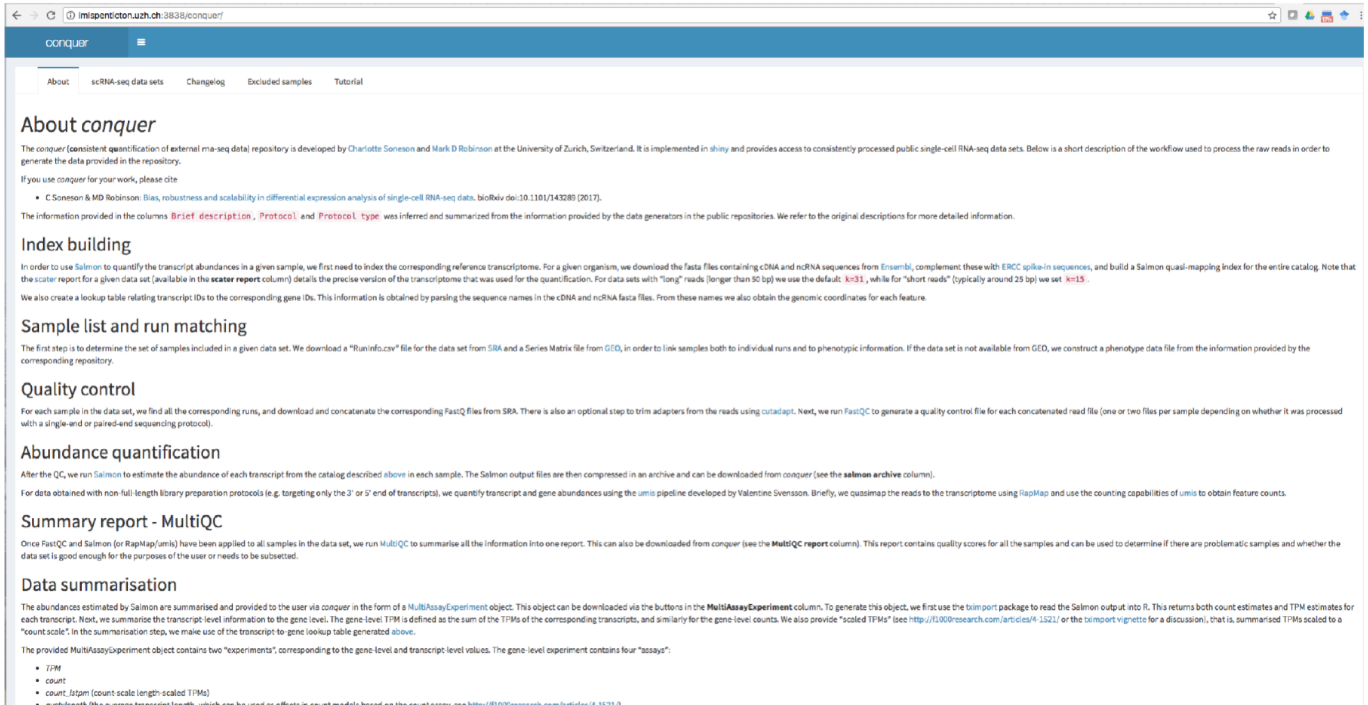
www.singlecellnetwork.org

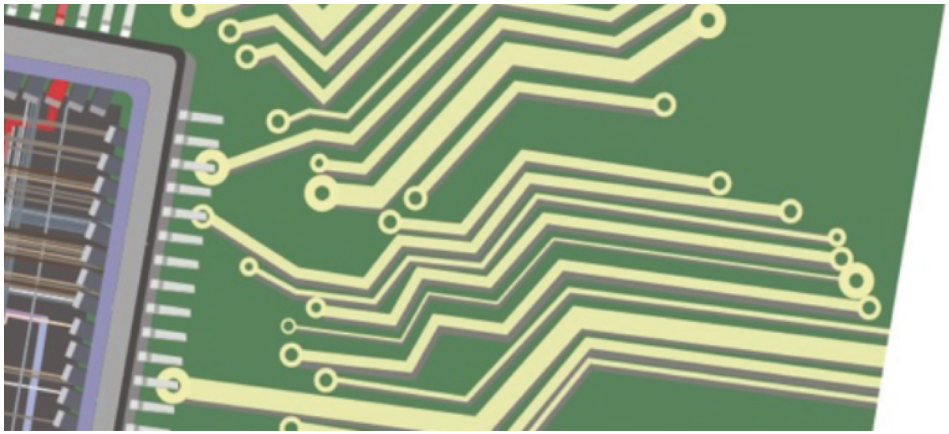
Resources, cont.

Data repositories: JingleBells



Data repositories: Conquer





Computational Genomics Workshop

September 10 & 11, 2018

Acknowledgements

Aviv Regev
Brian Haas
Kristine Schwenck
Elliot Boblitt
Orit Rosen
Vicky Horst
Karthik Shekhar
CEGS Workshop Team

NIH CEGS (Center for Excellence in Genomic Science)

Questions?

BROAD INSTITUTE