Introduction to Single Cell Transcriptomic Analysis

Acknowledgments
Brian Haas
Karthik Shekhar
Timothy Tickle
Caroline Porter

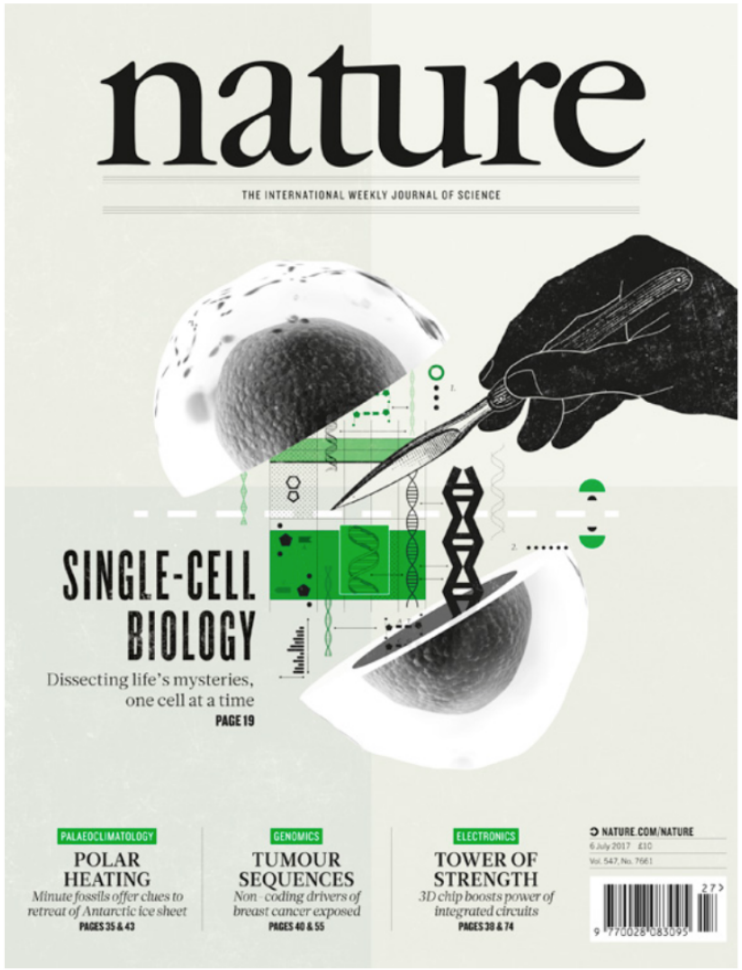
Ayshwarya Subramanian | subraman@broadinstitute.org

BROAD INSTITUTE

Goals for today

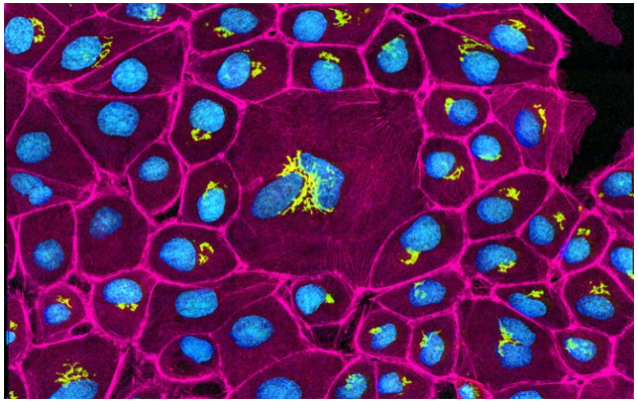
- Overview of single-cell RNA-seq data analysis
- What is in a count matrix?
- Preprocessing data: quality control (QC) and filtering
- Dimensionality reduction & Clustering
- Biology: inferring cell types and further

Why should we study gene expression at the resolution of single cells?

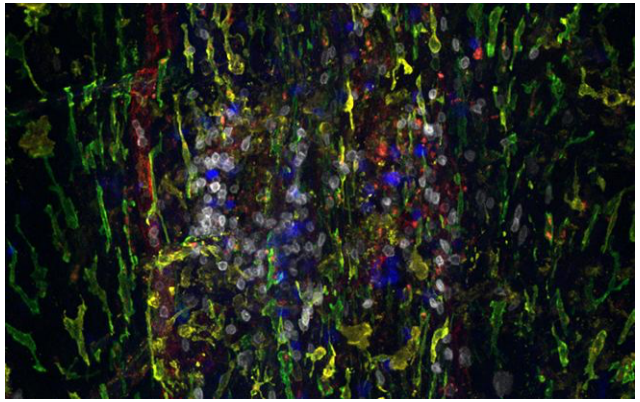


Incredible diversity in cell types, states, & interactions across human tissues

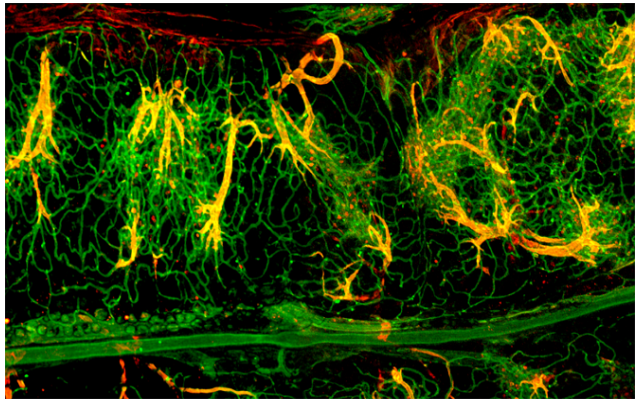
Skin epithelium



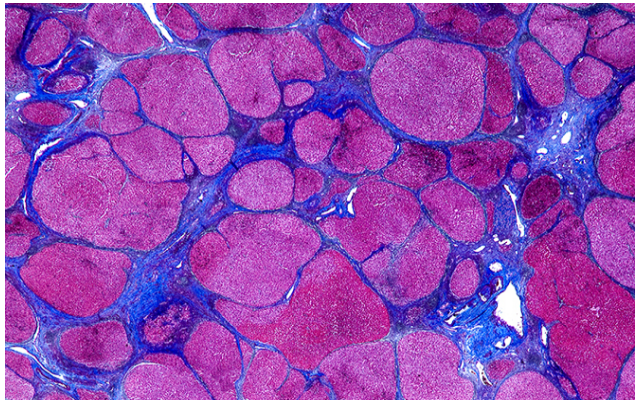
Brain meninges



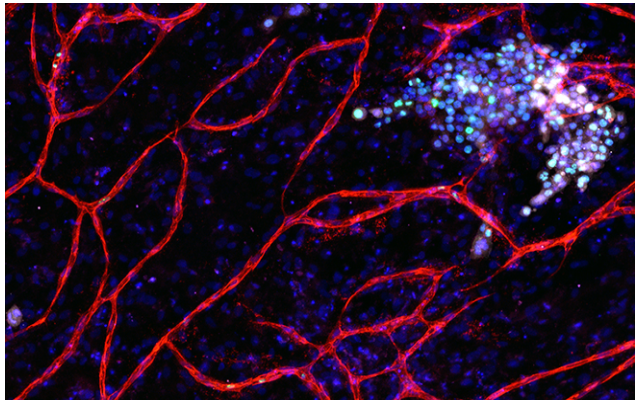
Blood vessels



Small intestine



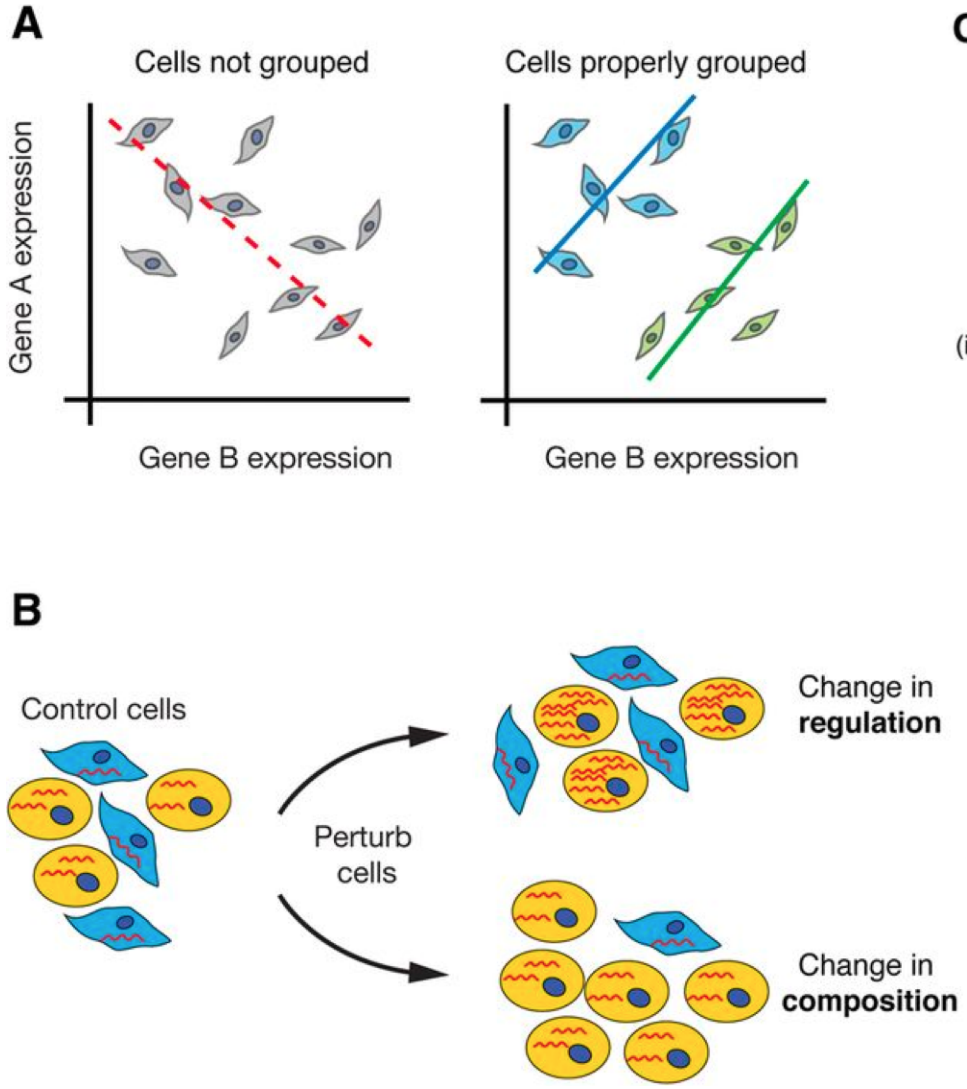
Liver cirrhosis



Breast cancer

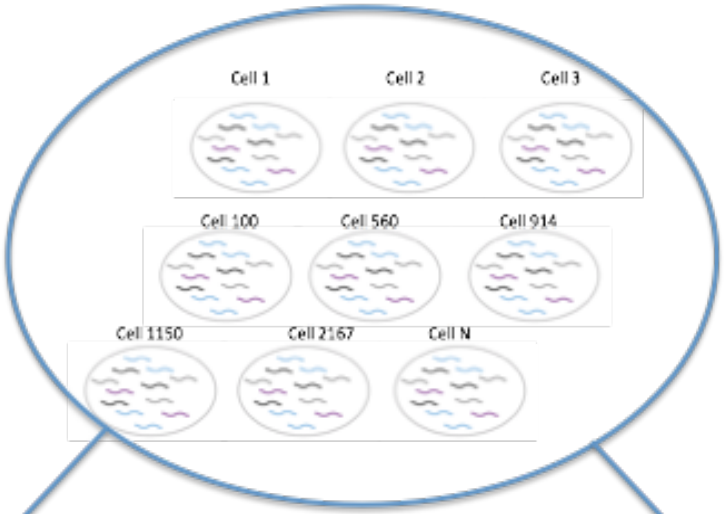
http://www.cell.com/pictureshow/skin | https://library.med.utah.edu/WebPath/webpath.html

Bulk sequencing profiles measure average profiles from tissue samples



Trapnell. Genome Research 2016

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

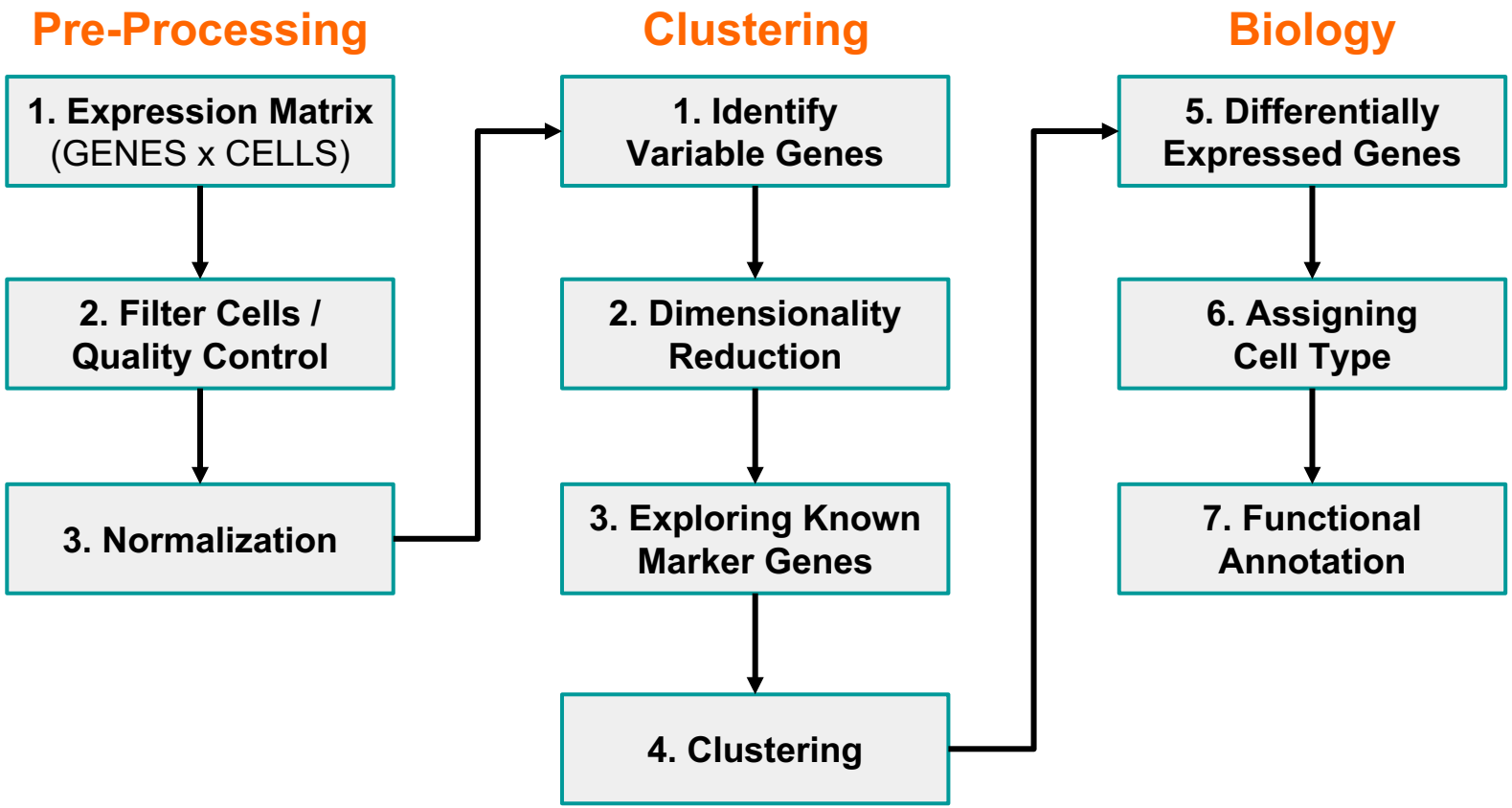
Bulk

Cellular resolution

Single cell

Slide courtesy of Karthik Shekhar

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



Tracking the cell-of-origin of individual transcripts

Slide courtesy of Karthik Shekhar

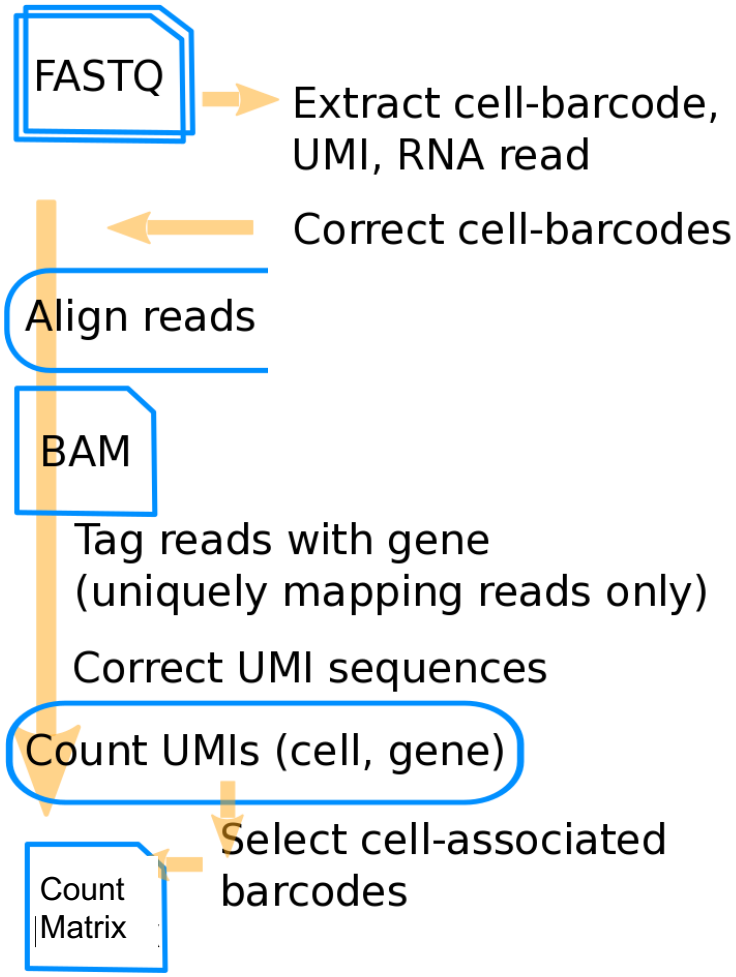
Tracking the cell-of-origin of individual transcripts

Slide courtesy of Karthik Shekhar

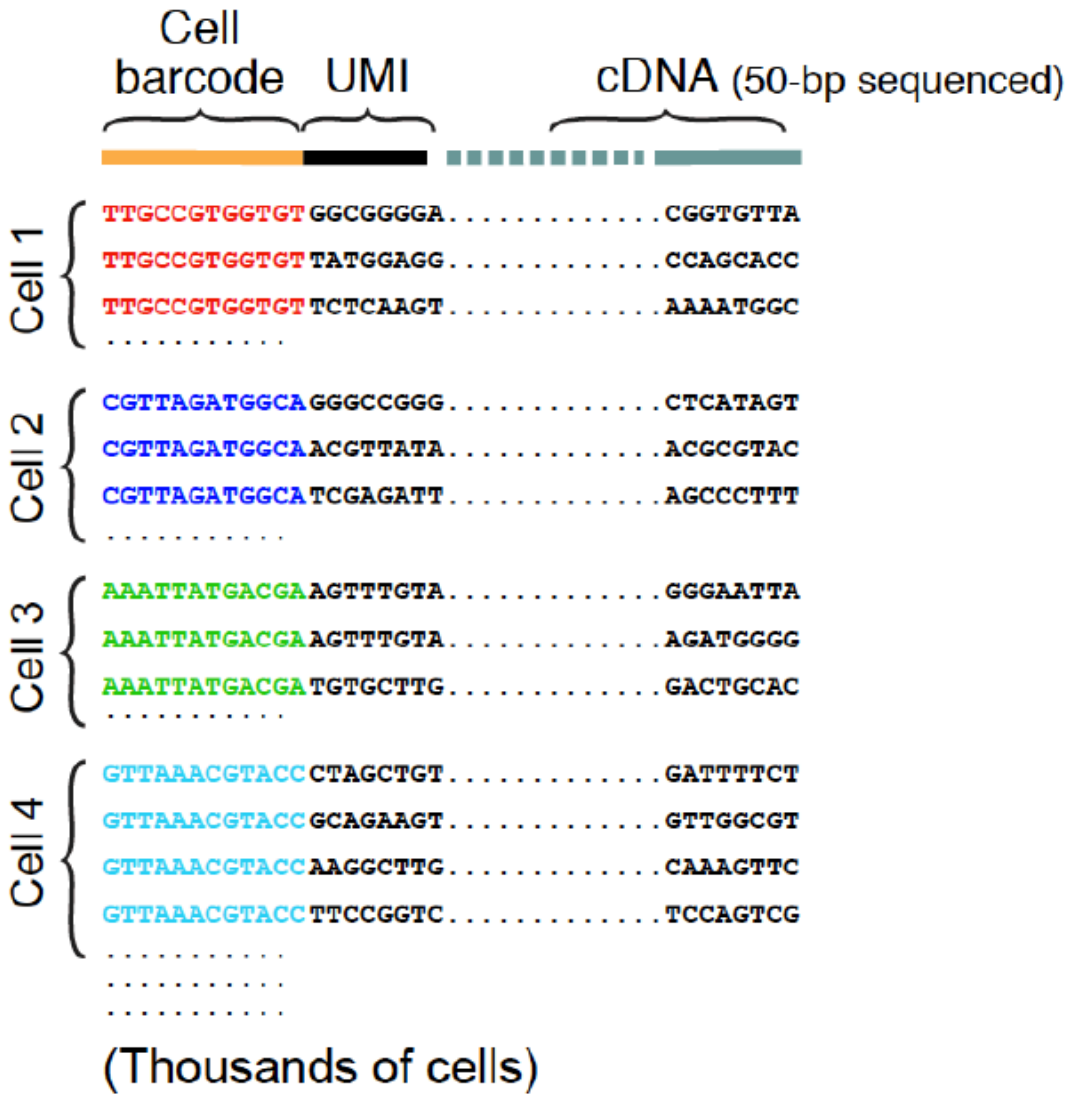
Single-cell RNA-seq pipeline

Slide courtesy of Karthik Shekhar

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



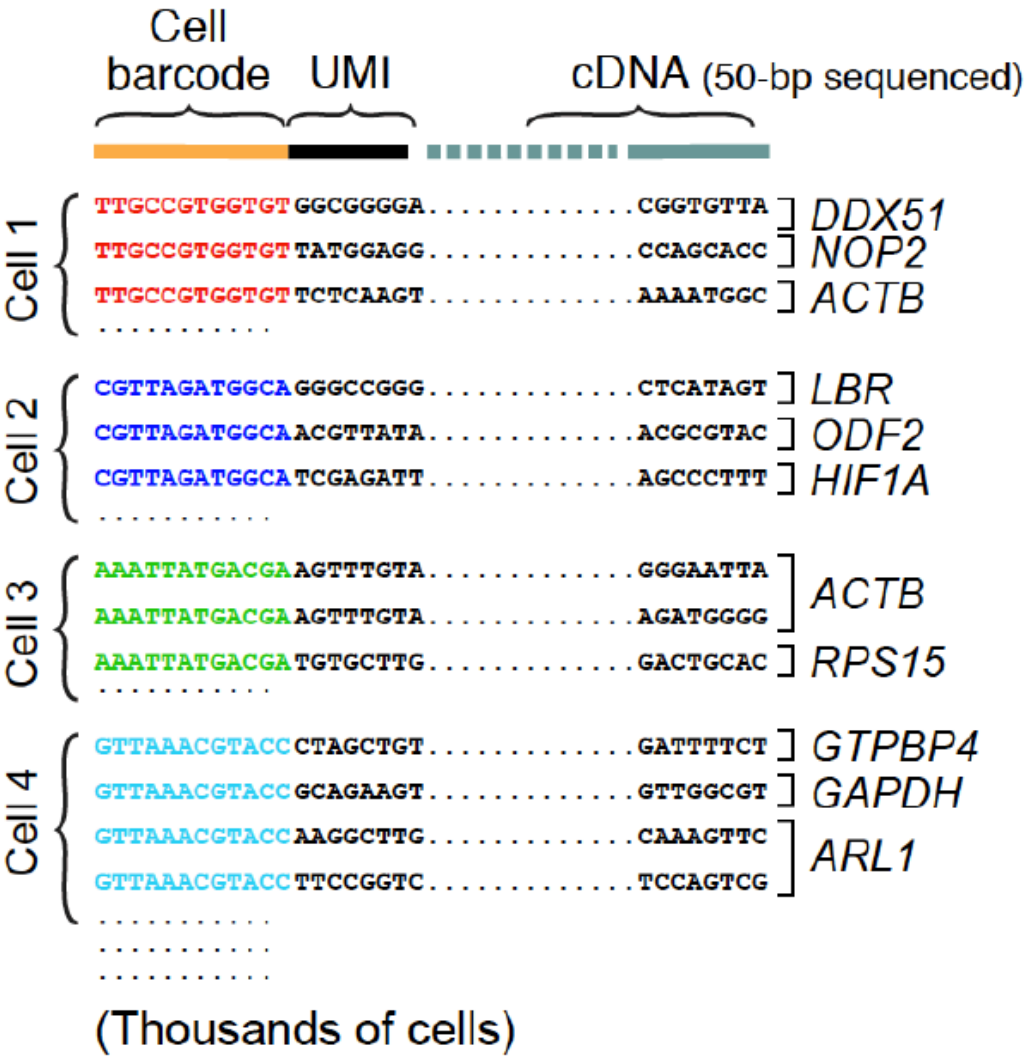
Group reads by cell barcode



Some platforms incorporate "error-correcting" barcodes which makes the pipeline robust to sequencing errors

Slide courtesy of Karthik Shekhar

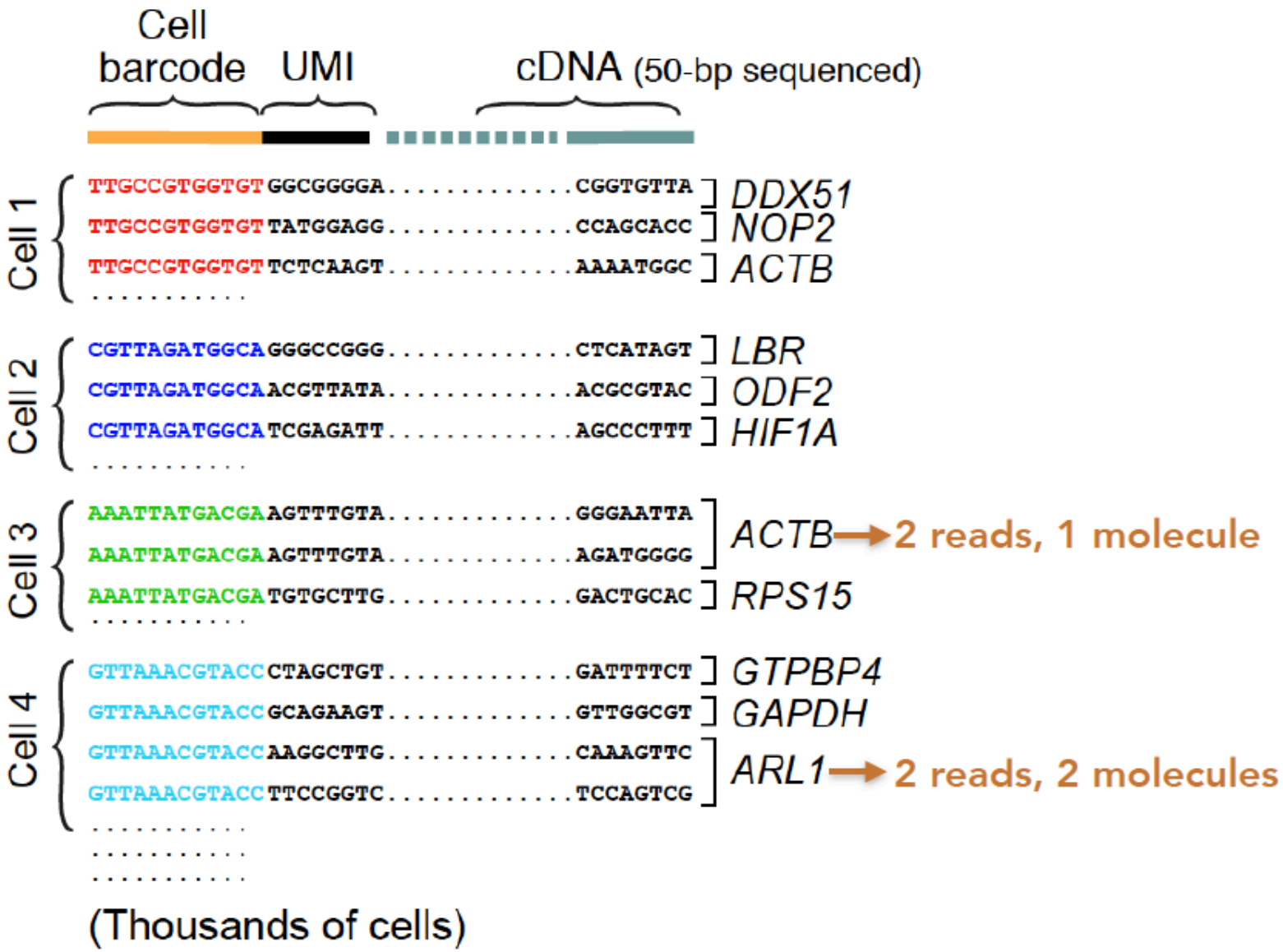
Align reads to the genome using a splice-aware aligner



A lot of pipelines use the STAR aligner, which consumes a lot of memory, but is EXTREMELY fast

Slide courtesy of Karthik Shekhar

Collapse UMIs to count transcripts

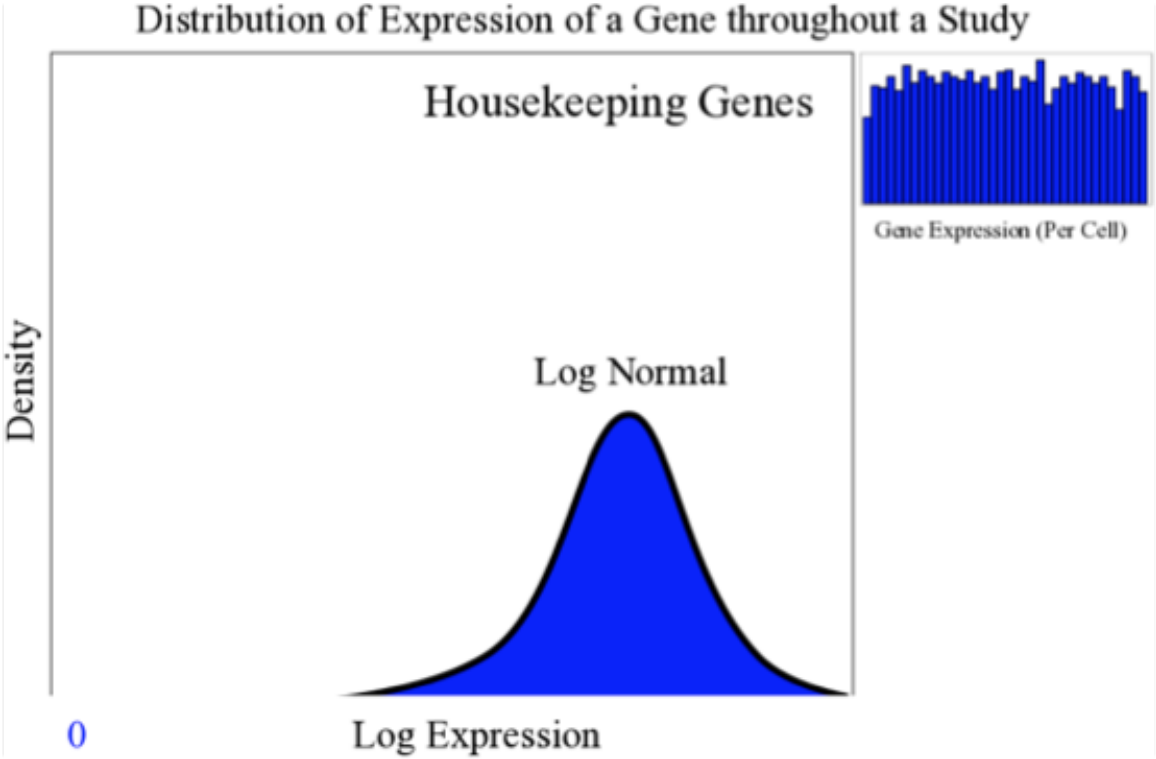


Slide courtesy of Karthik Shekhar

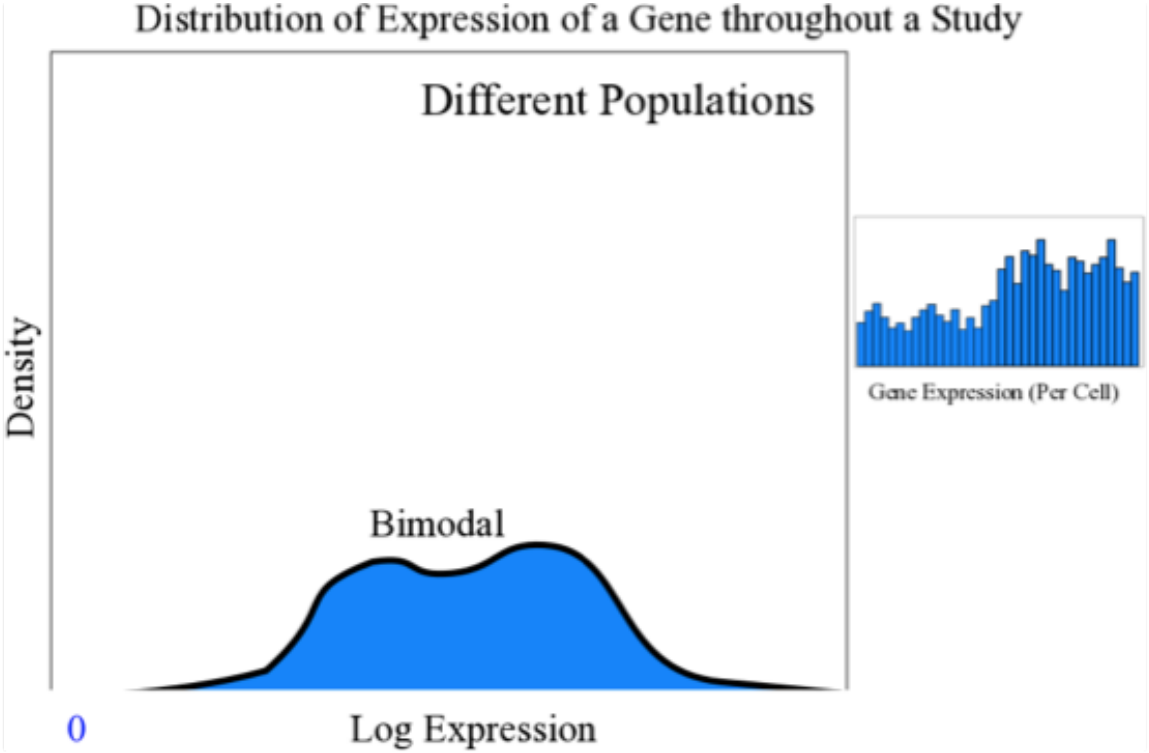
Counts Matrix

	Cell 1	Cell2	Cell3	Cell4	...
Gene 1	0	0	3	10	
Gene 2	24	0	41	12	
Gene 3	175	284	93	162	
Gene 4	0	0	0	0	
Gene 5	36	0	32	21	
...	...	...	...	...	

Genes Have Different Distributions



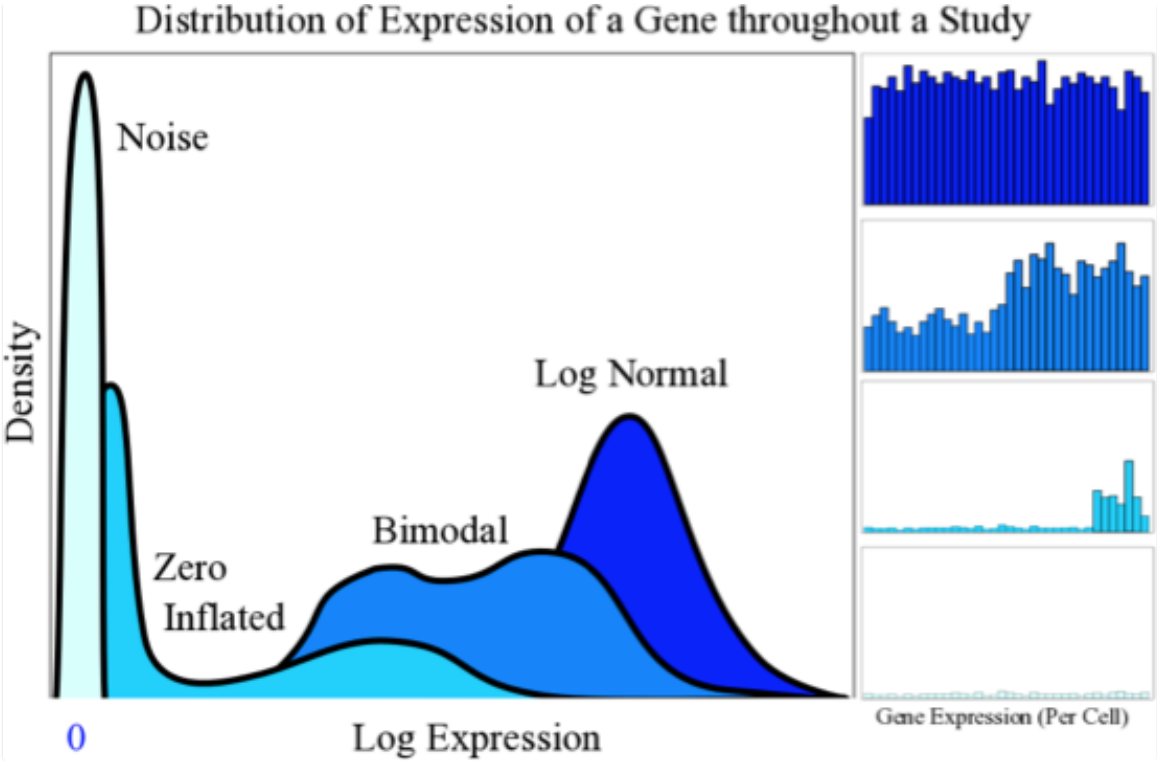
Genes Have Different Distributions



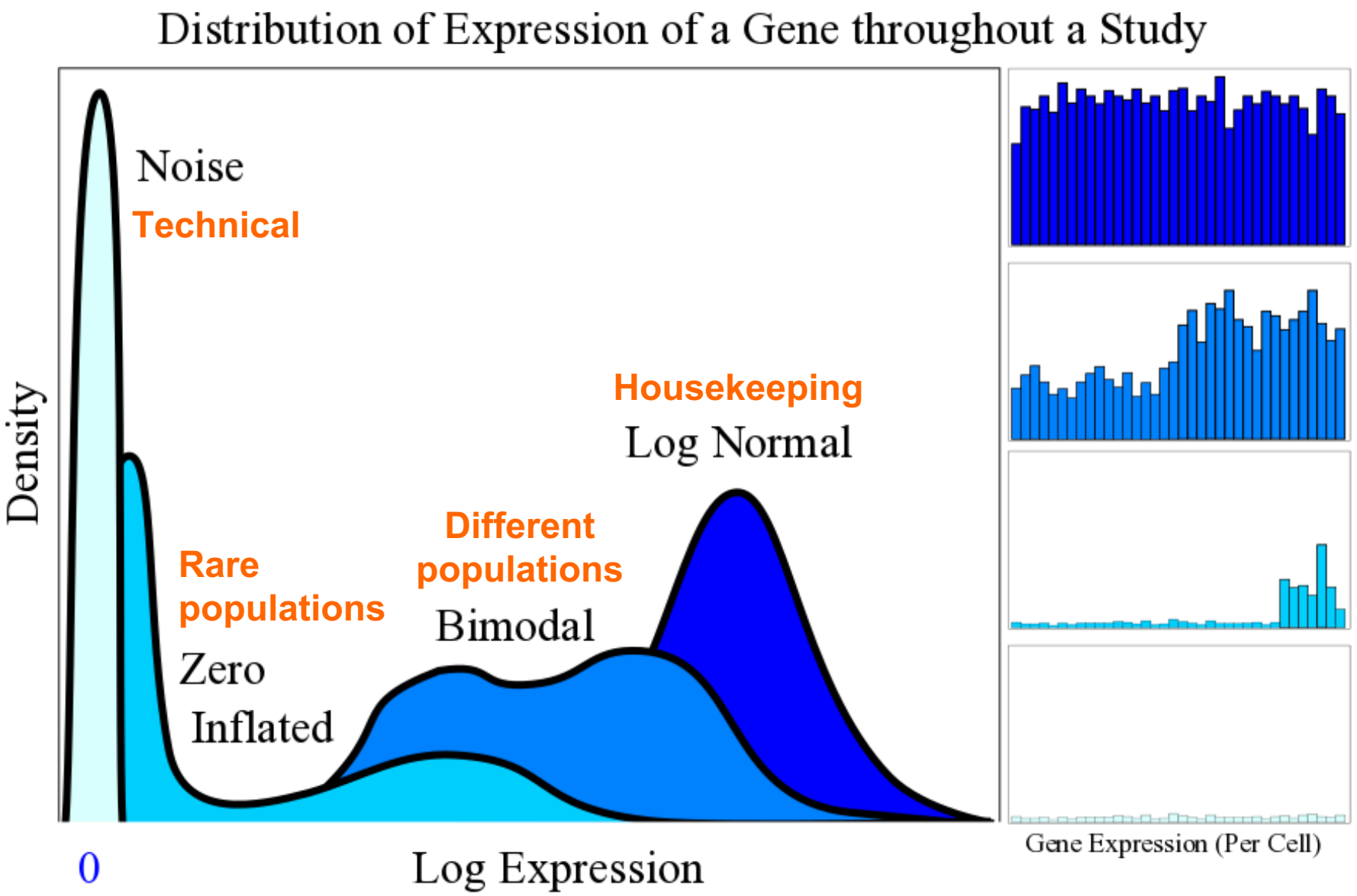
Genes Have Different Distributions

Genes Have Different Distributions

Genes Have Different Distributions



Genes have different distributions



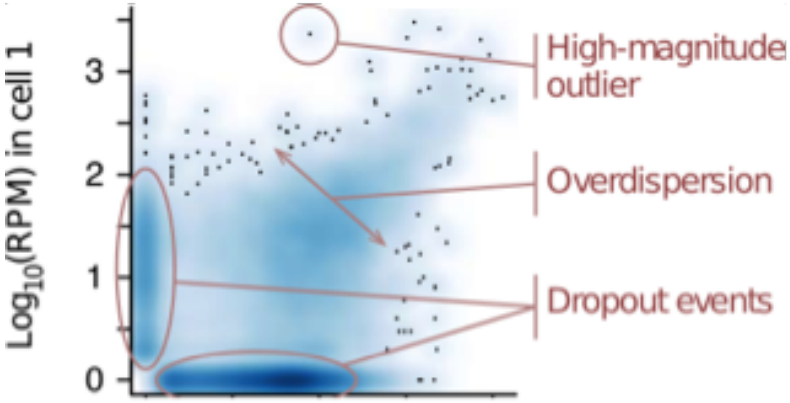
Underlying Biology

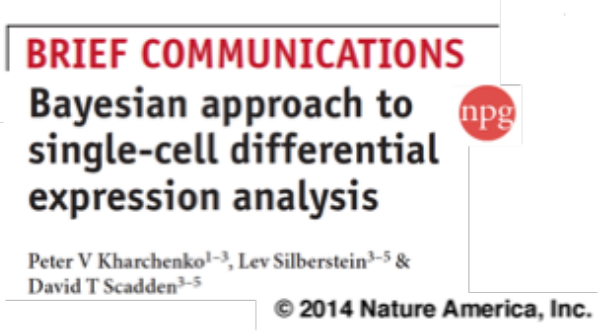
- **Zero inflation.**
  - **Drop-out event during reverse-transcription.**
  - **Genes with more expression have less zeros.**
  - **Complexity varies.**
- **Transcription stochasticity.**
  - **Transcription bursting.**
  - **Coordinated transcription of multigene networks.**

–Over-dispersed counts.

- **Higher Resolution.**

–More sources of signal





Cell Identity is a Mixture of Multiple Factors

REVIEW

nature biotechnology

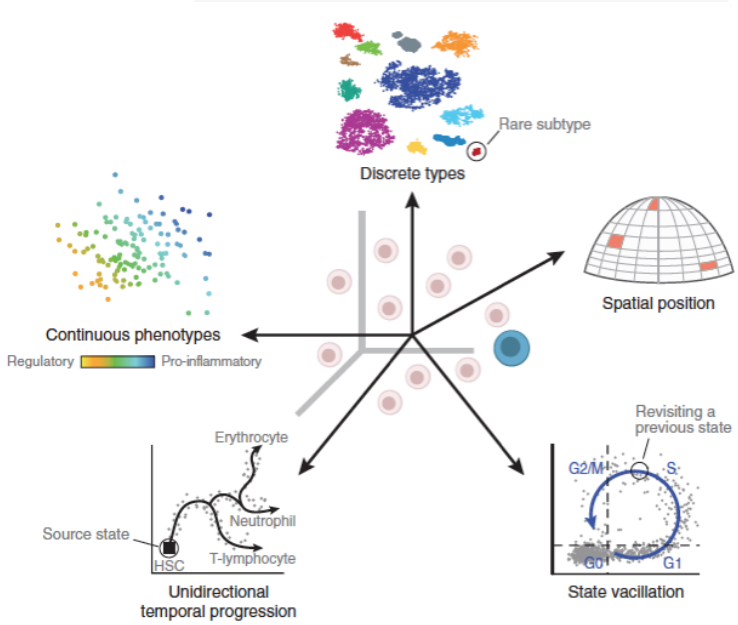
Revealing the vectors of cellular identity with single-cell genomics

Allon Wagner[1], Aviv Regev[2,3,5] & Nir Yosef[1,4,5]

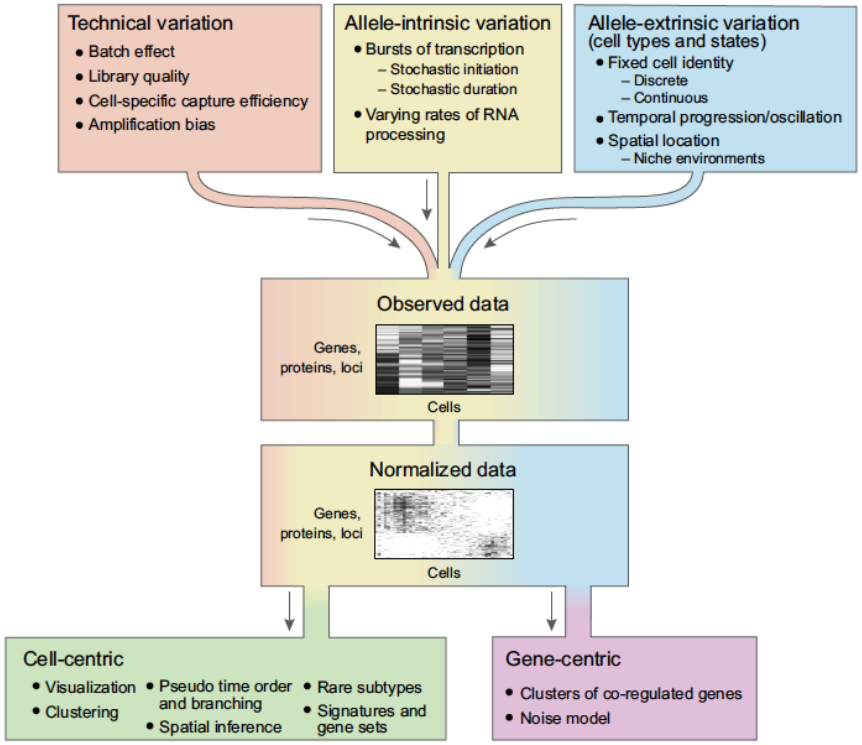
Multiple factors shape a cell's identity

- Membership in a taxonomy of cell types
- Simultaneous time-dependent processes
- Response to the environment
- Spatial positioning

A cell participates in multiple cell contexts.



Expression has Many Sources

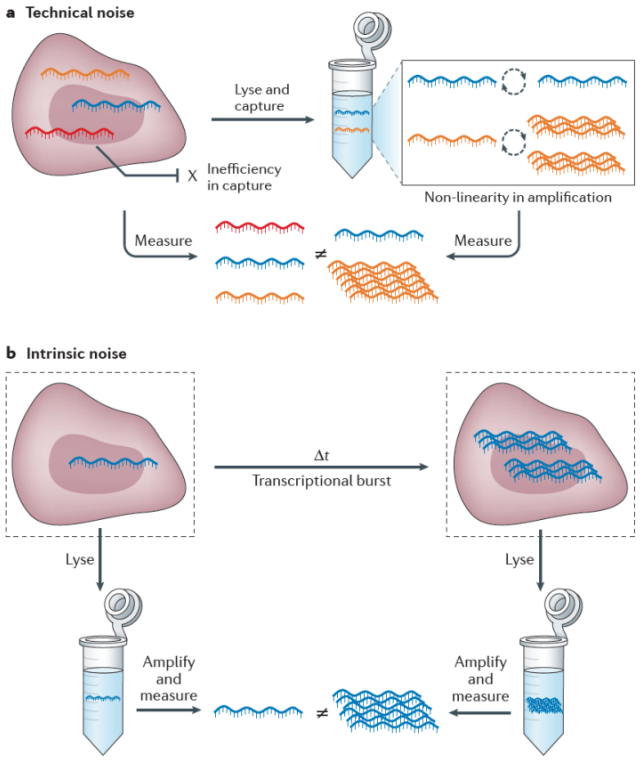


REVIEW
nature biotechnology

Revealing the vectors of cellular identity with single-cell genomics

Allon Wagner[1], Aviv Regev[2,3,5] & Nir Yosef[1,4,5]

Technical vs Intrinsic Noise



SINGLE-CELL OMICS

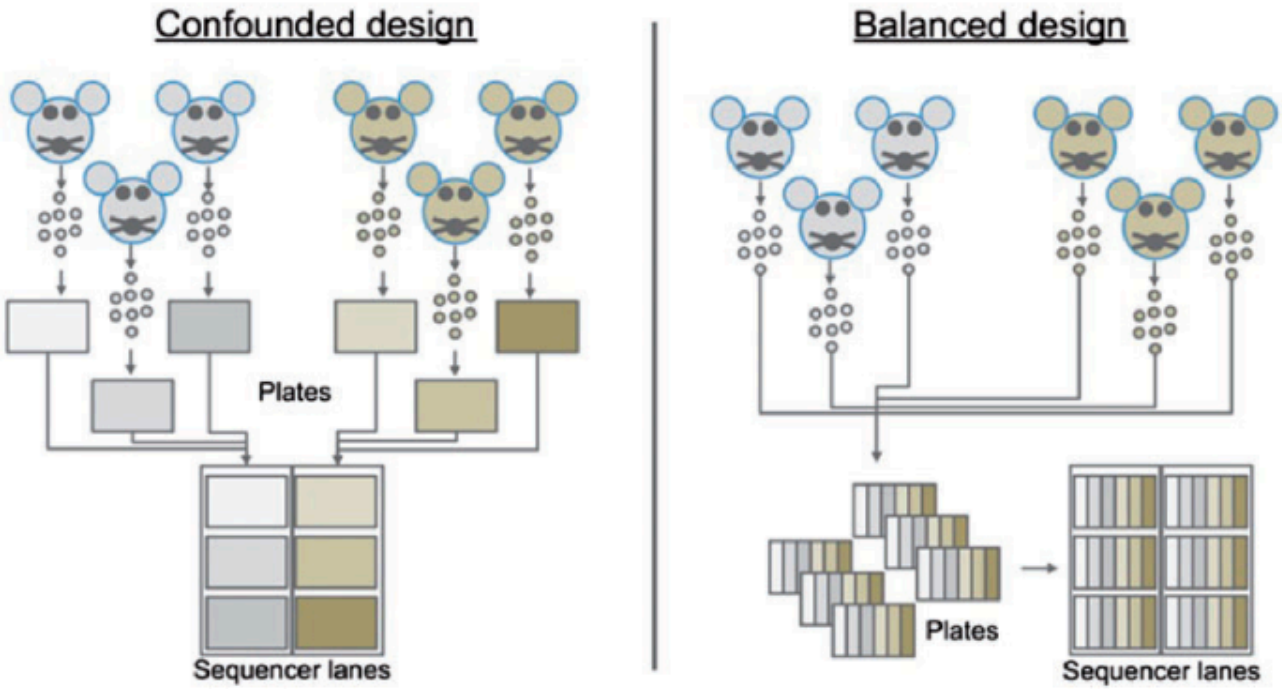
Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices

Sanjay M. Prakadan[1–3], Alex K. Shalek[1–3] and David A. Weitz[4,5]

Experimental Design

Sound experimental design : Replication, Randomization and Blocking

- R. A. Fisher, 1935

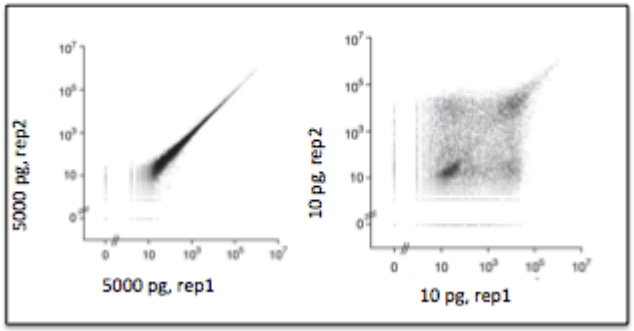


- For example, when analyzing tumor phenotypes in a patient process the tumor sample and a matched control on the same day, using the same reagents!
- Blocking is not always possible because of logistic limitations, in which case ensure that any biological conclusion is supported by multiple, independently collected samples

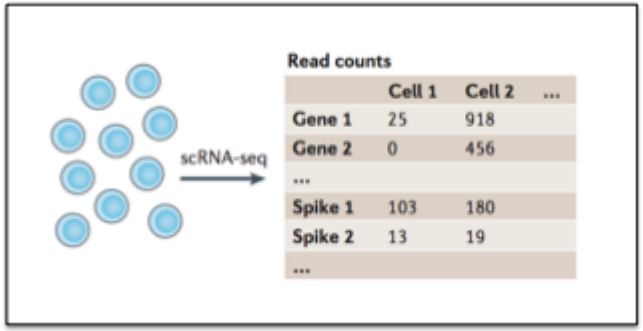
Slide courtesy of Karthik Shekhar

Technical Conceptual Challenges

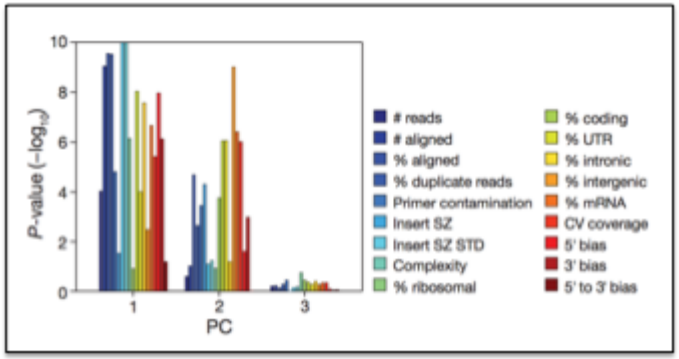
Dropouts



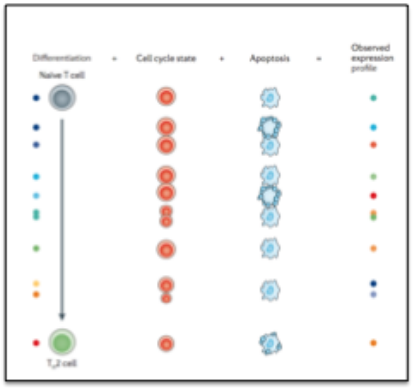
Variation in cell size and quality



Variation dominated by technical factors

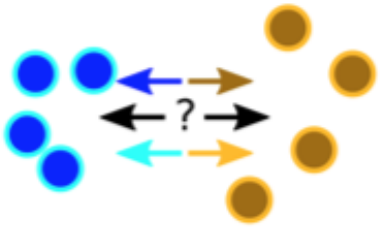


Observed gene expression is a convolution

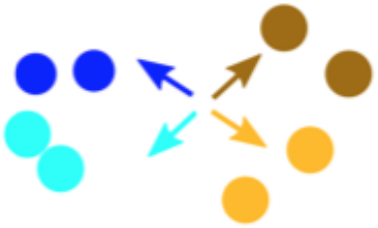


What is Study Confounding

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



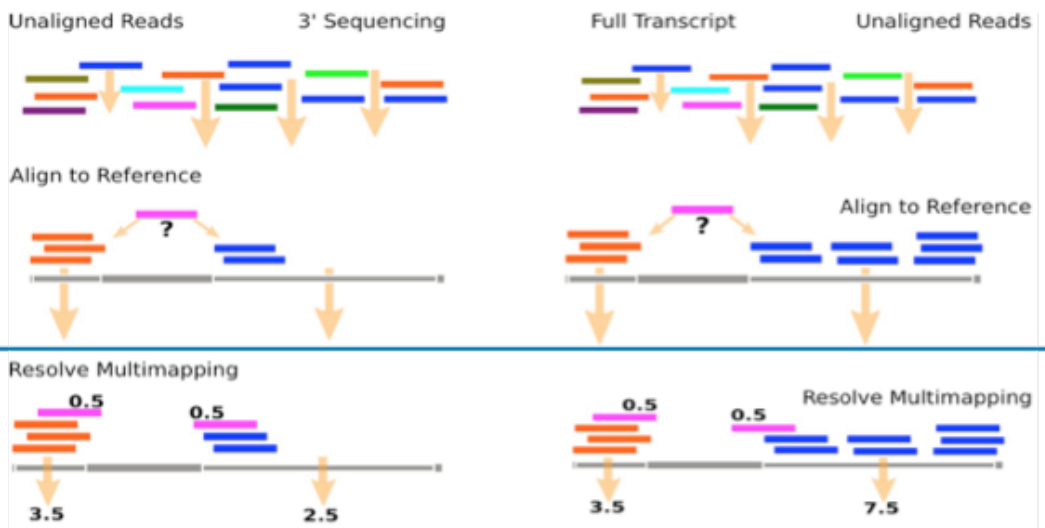
Cell	Site	Treatment
1	Main	A
2	Main	A
3	Main	B
4	Main	B
5	Remote	A
6	Remote	A
7	Remote	B
8	Remote	B



Goals for today

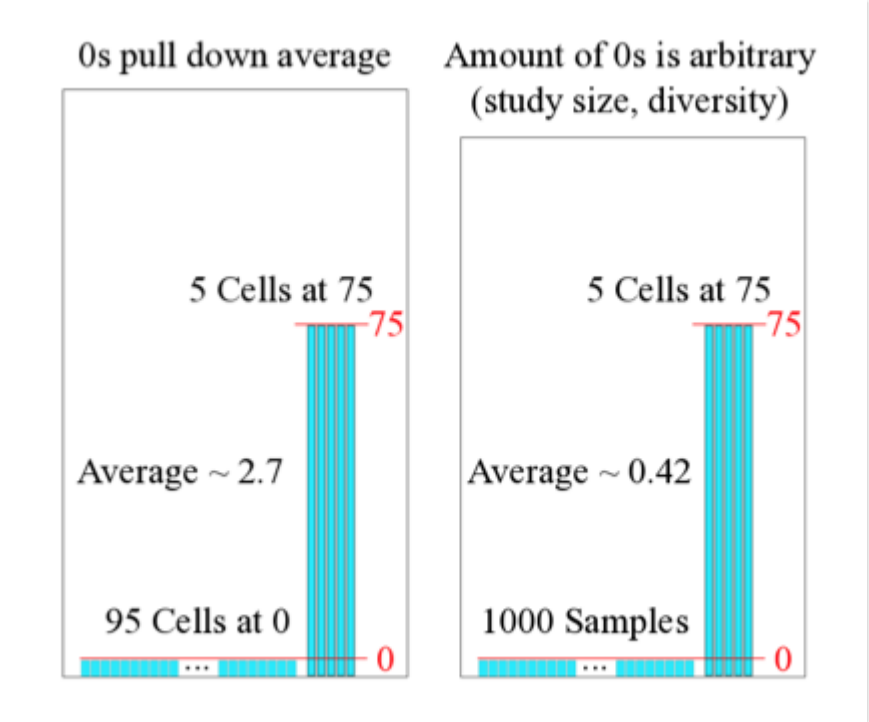
- Overview of single-cell RNA-seq data analysis
- What is in a count matrix?
- Preprocessing data: quality control (QC) and filtering
- Dimensionality reduction & Clustering
- Biology: inferring cell types and further

Count Preparation is Different Depending on Assays

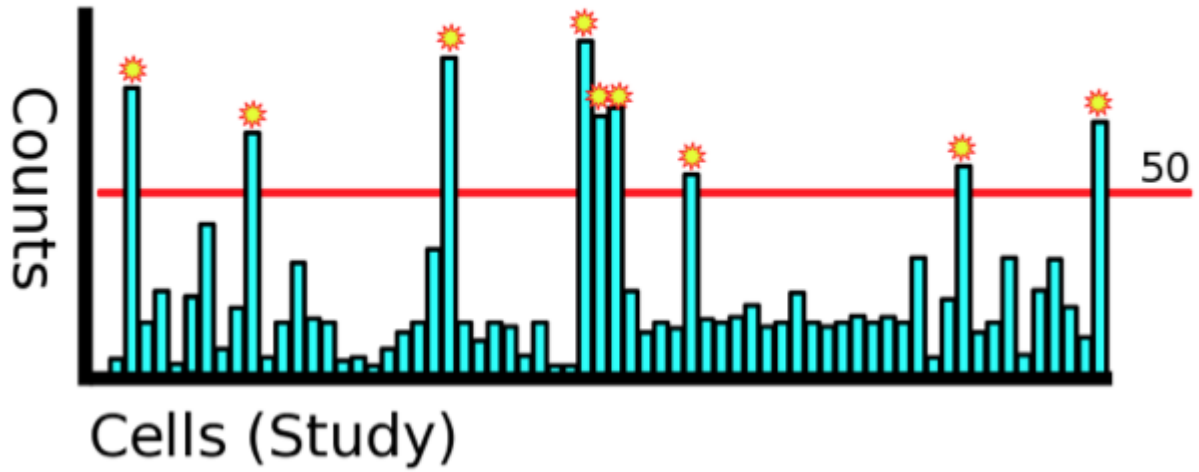


RSEM KALLISTO			RSEM KALLISTO
	No transcript length correction	Correct For Transcript Length 3.5 / 1 = 3.5 7.5 / 3 = 2.5	TPM
Seurat	Correct for Sequencing Depth X / Column Total * 1E5 or 1E6	Correct for Sequencing Depth	TPM
Seurat	Log2() + 1	Log2() + 1	Seurat

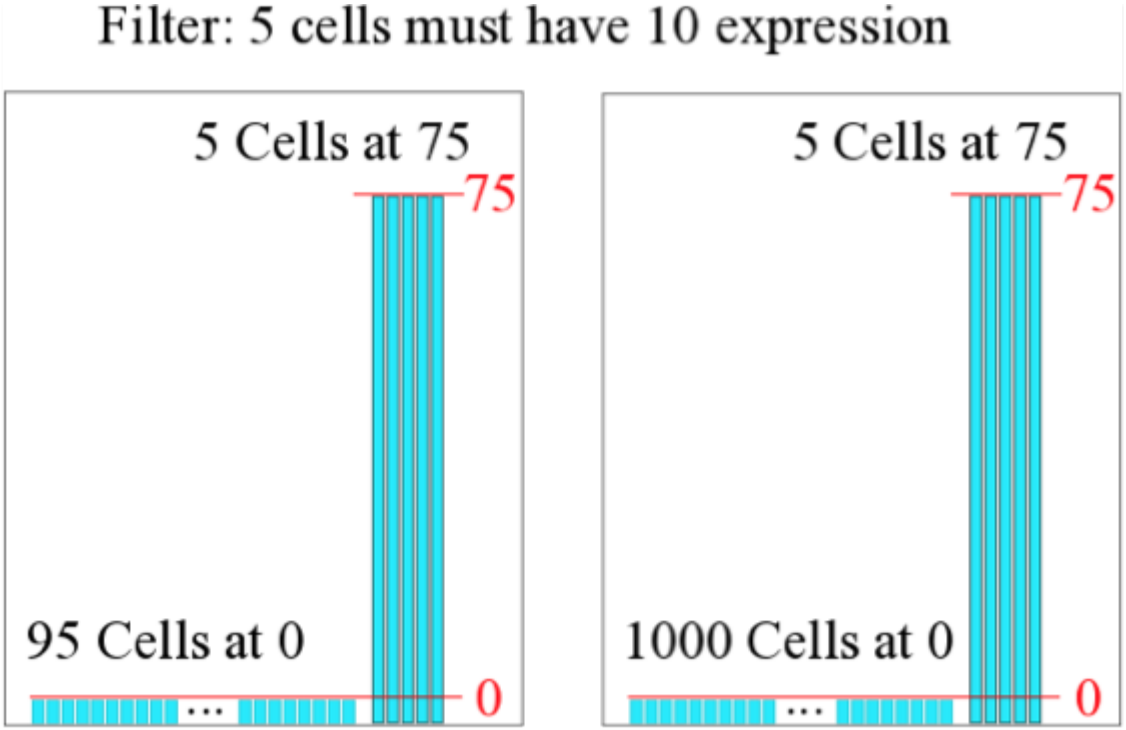
Filtering Genes: Averages are Less Useful



Filtering Genes: Using Prevalence



Filtering Genes: Using Prevalence

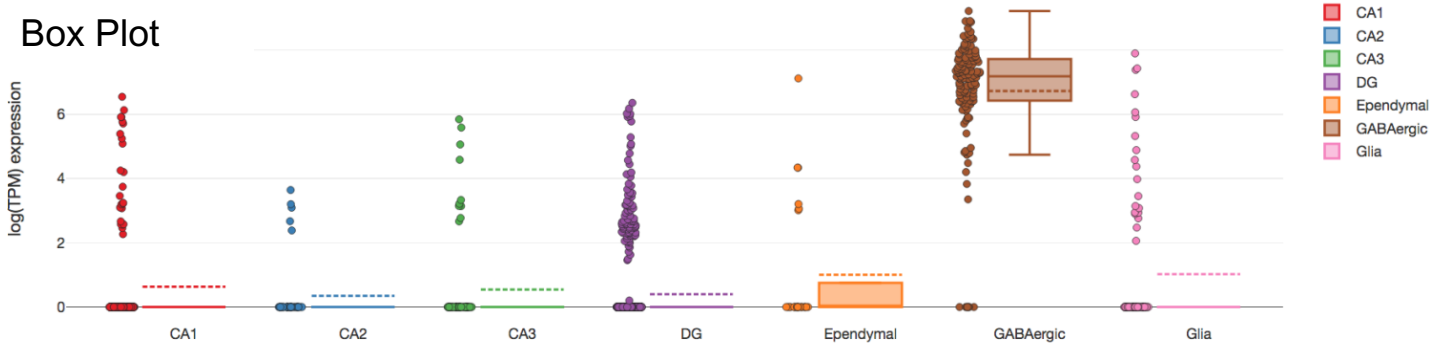


Representing Genes Throughout Cells

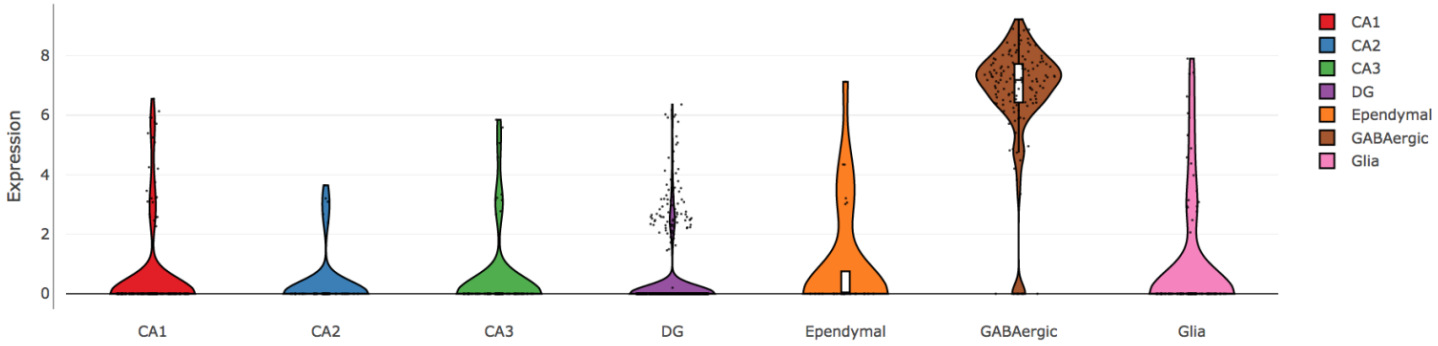
A gene (GAD2) across many groups of cells.

Habib et al. 2016

Box Plot



Violin Plot



What is Metadata?

Other information that describes your measurements.

- **Patient information.**

•Lifestyle (smoking), Patient Biology (age), Comorbidity

- **Study information.**

•Treatment, Cage, Sequencing Site, Sequencing Date

- **Sequence QC on cells.**

•Useful in filtering and stratifying.

Filtering Cells: Removing Outlier Cells

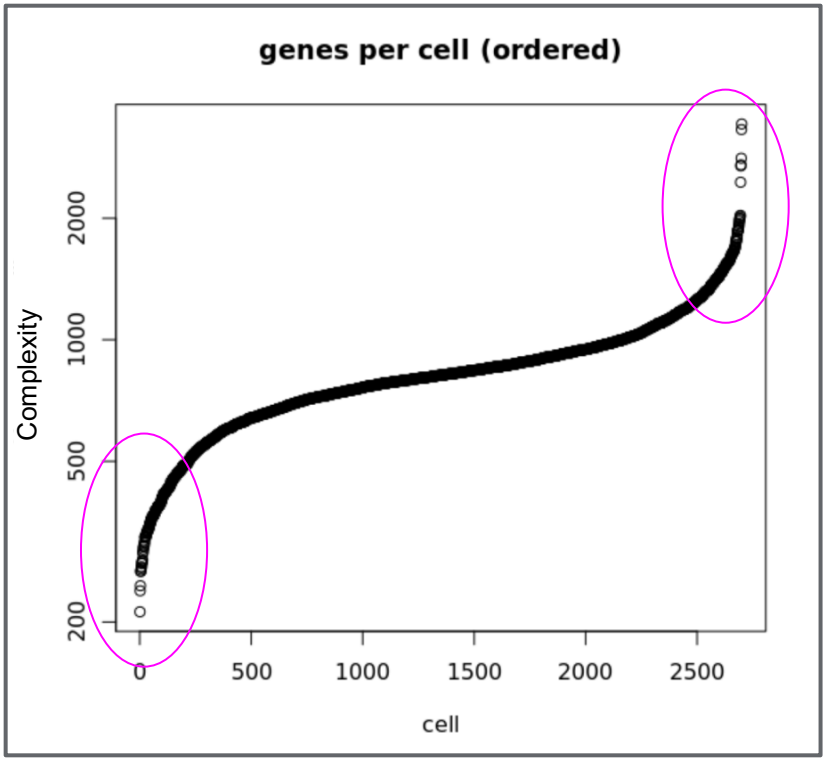
- Bulk RNA-Seq studies often do not remove outliers cells
  - scRNA-Seq often removes “failed libraries”.
- Outlier cells are not just measured by complexity
- Percent Reads Mapping
- Percent Mitochondrial Reads
- Presence of marker genes
- Intergenic/ exonic rate
- 5' or 3' bias
- other metadata …
- Useful Tools
  - Picard Tools and RNASeQC

Filtering Cells: Complexity

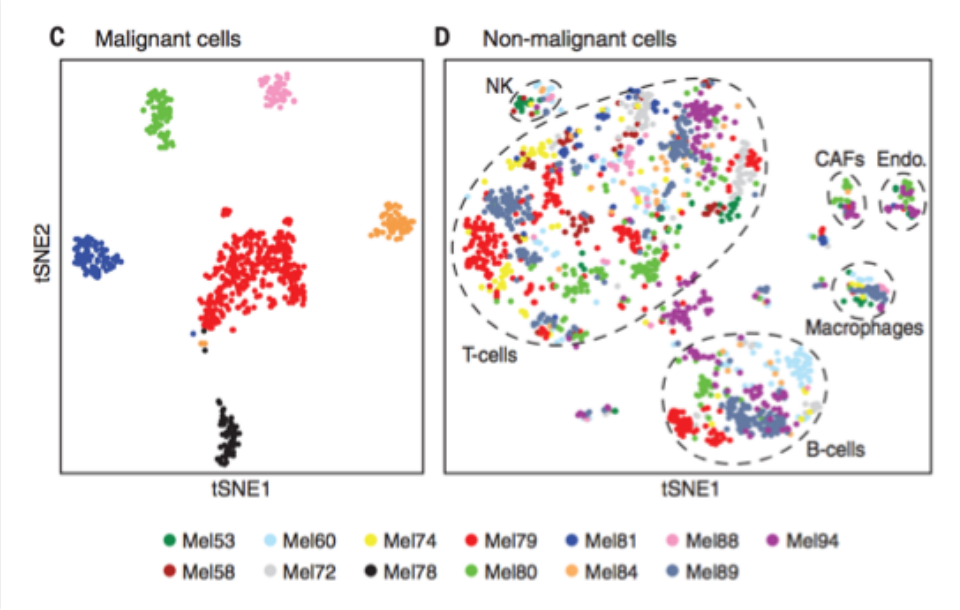
Complexity:

Simplest definition is the number of genes expressing at any amount in a cell.

Filtering both ends.
Lower: Failed libraries?
Higher: Doublets?



Checks and Balances in Analysis



Tumor cells cluster by patient. By itself, this could be simply batch effects!

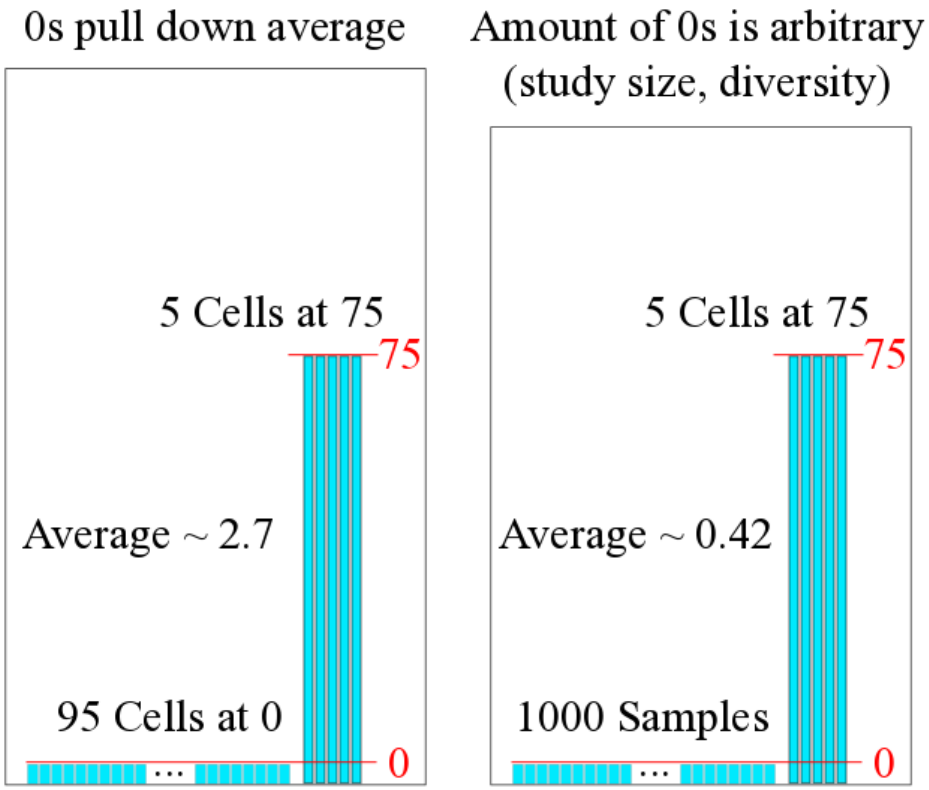
But non-malignant cells cluster by type, rather than patient!

Tirosh et al., Science 2016

Some single-cell RNA-seq data challenges to remember

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

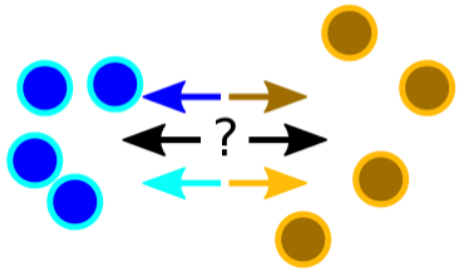
  Zero expression doesn't mean the gene isn't on



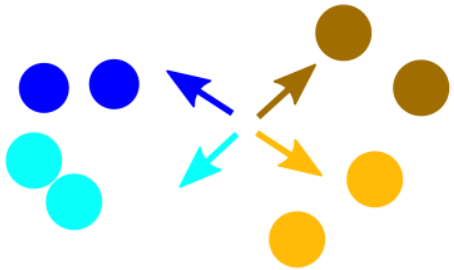
Some single-cell RNA-seq data challenges to remember

- **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



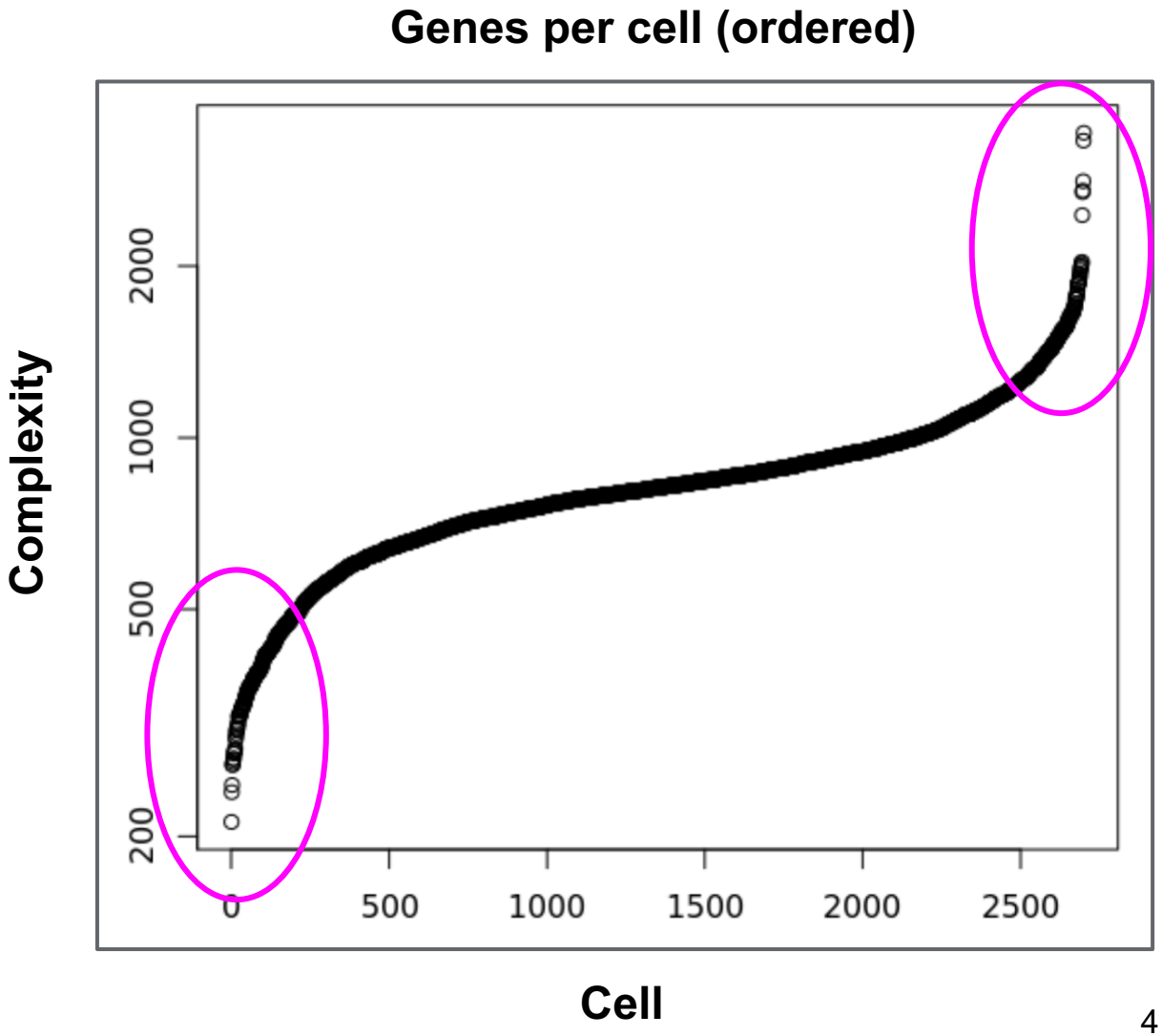
Cell	Site	Treatment
1	Main	A
2	Main	A
3	Main	B
4	Main	B
5	Remote	A
6	Remote	A
7	Remote	B
8	Remote	B



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

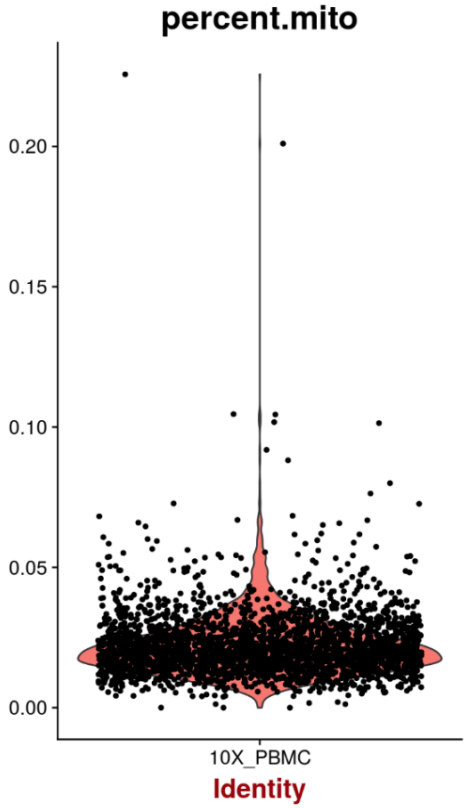
Filtering and quality control:
Number of genes per cell

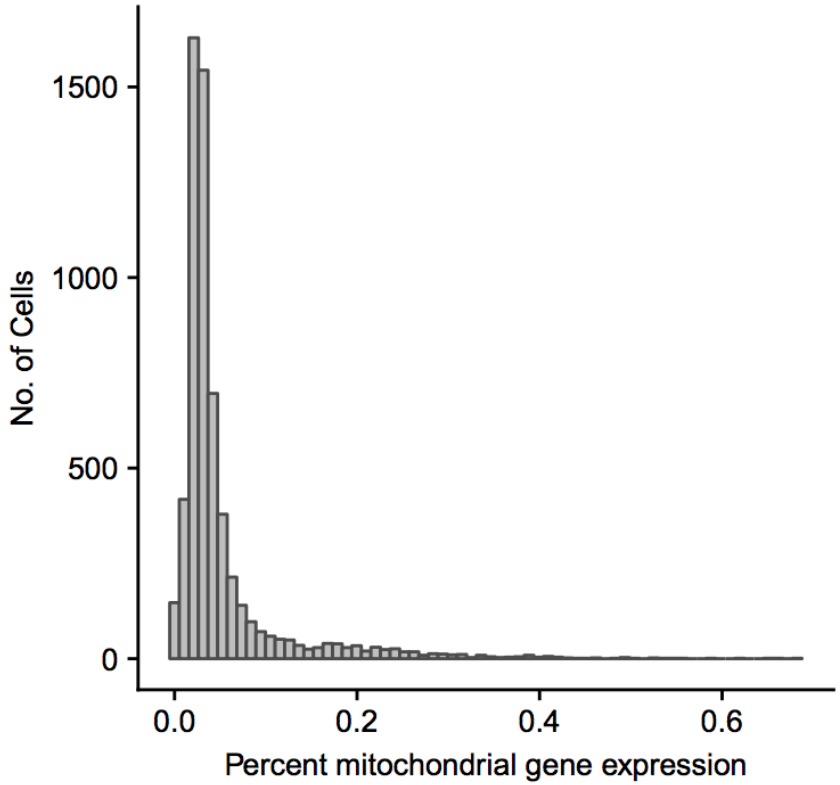
complexity = number of genes detected in a cell



Filtering and quality control:
Mitochondrial gene expression

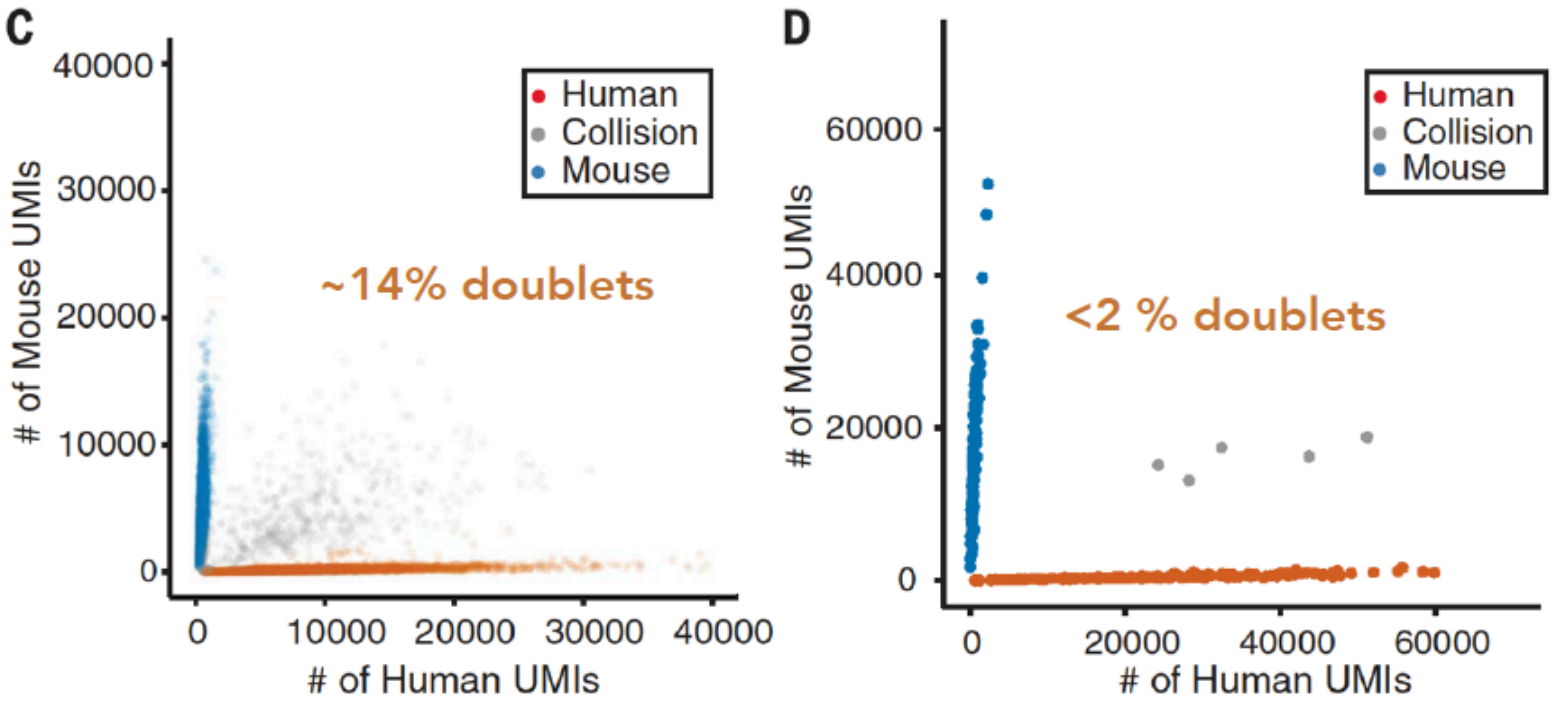
Percent of reads in a cell coming from mitochondrial genes is a good measure of cell quality - high mitochondrial gene expression indicates stressed cells (e..g, from damage during tissue dissociation)





Doublets

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

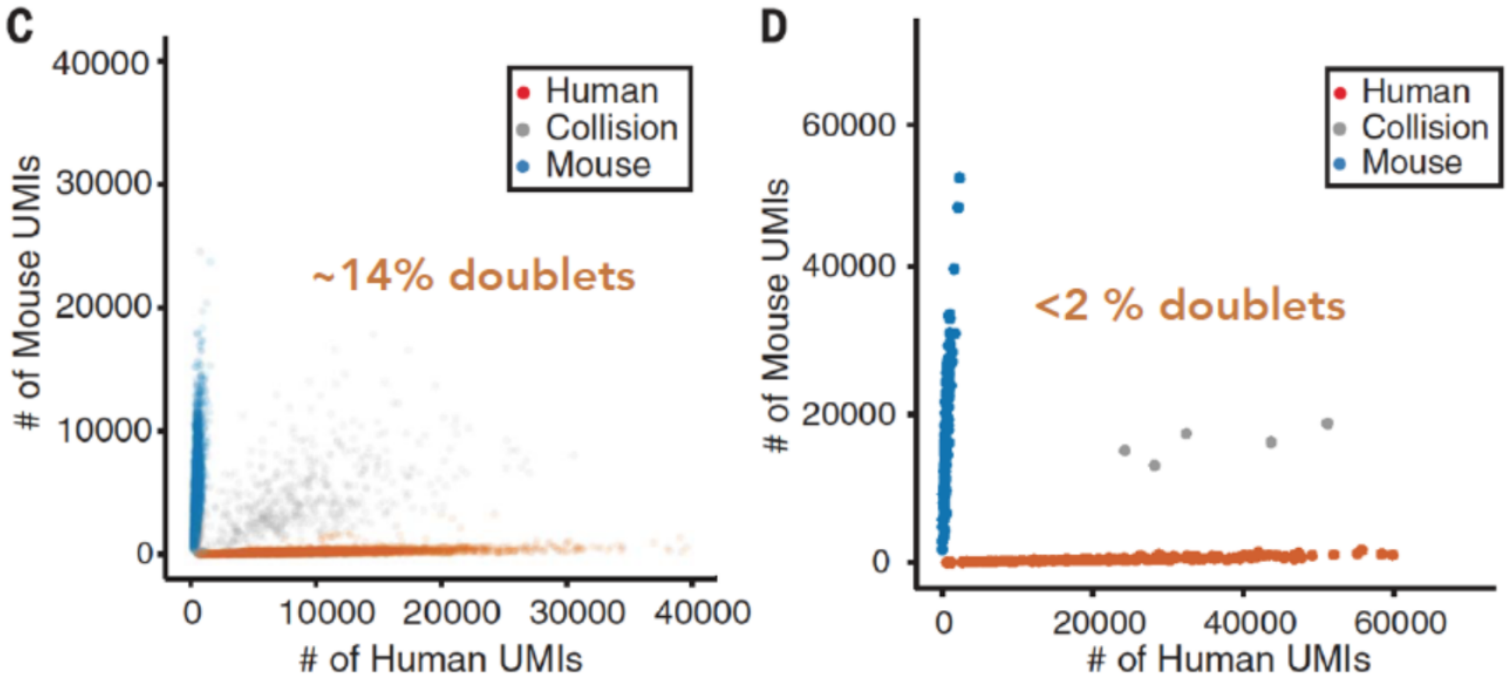


The doublet frequency is +vely correlated with throughput

Slide courtesy of Karthik Shekhar

Filtering and quality control:
Doublets - 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

Filtering and quality control:
Cells vs. reads/cell

- Earlier scRNA-seq studies used to sequence each cell to > 10 million reads / cell - this is now widely accepted as a ridiculous number!

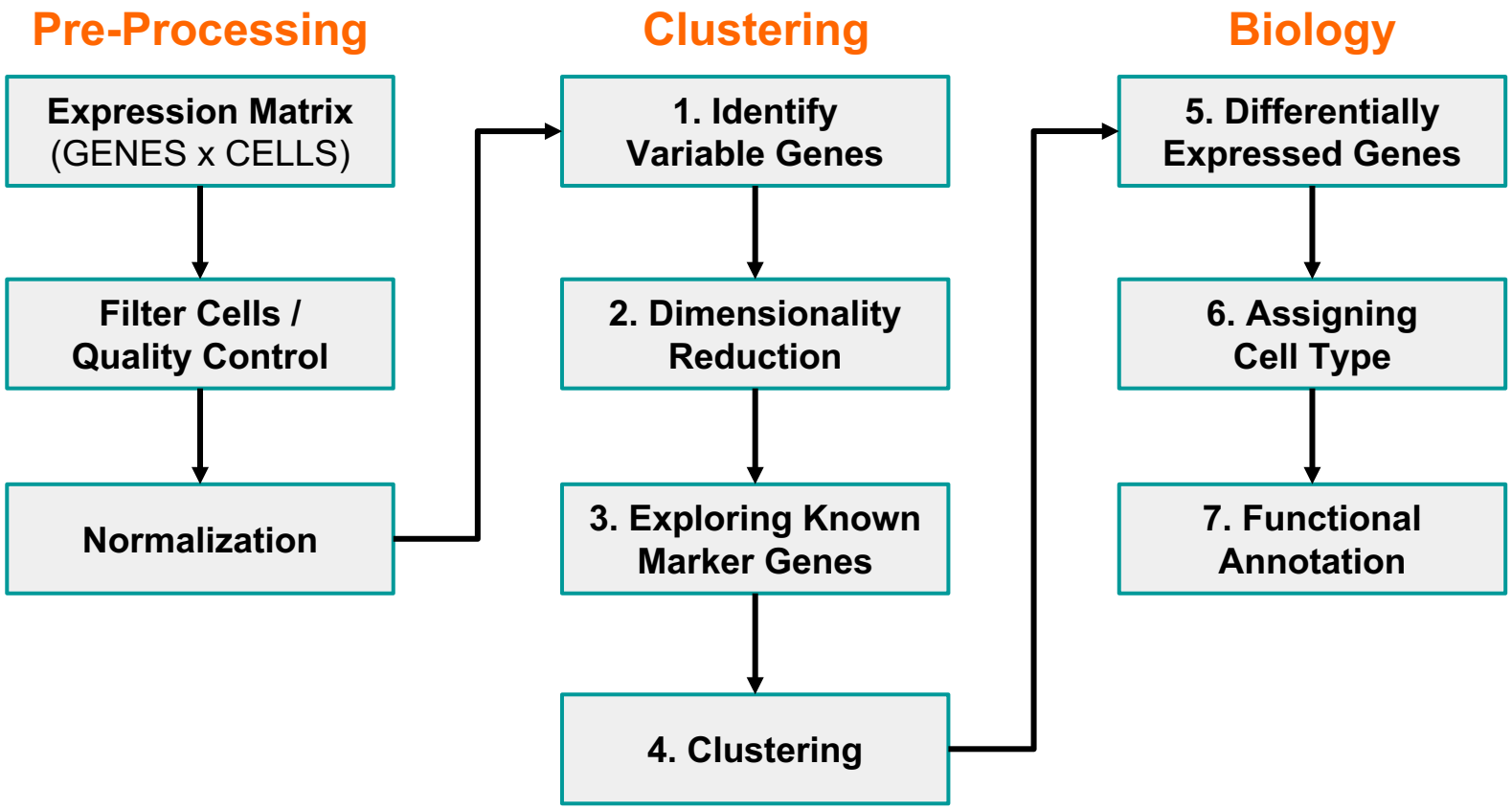
- ~50,000-100,000 reads/cell is now widely regarded as sufficient for most applications. ~1M reads per cell effectively means saturation

- The modular nature of biology "guarantees" that key signals can be recovered at shallow sequencing depth*

Cleary et al., Cell, 2017

Slide courtesy of Karthik Shekhar

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



Data normalization and scaling

Typically, we:

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

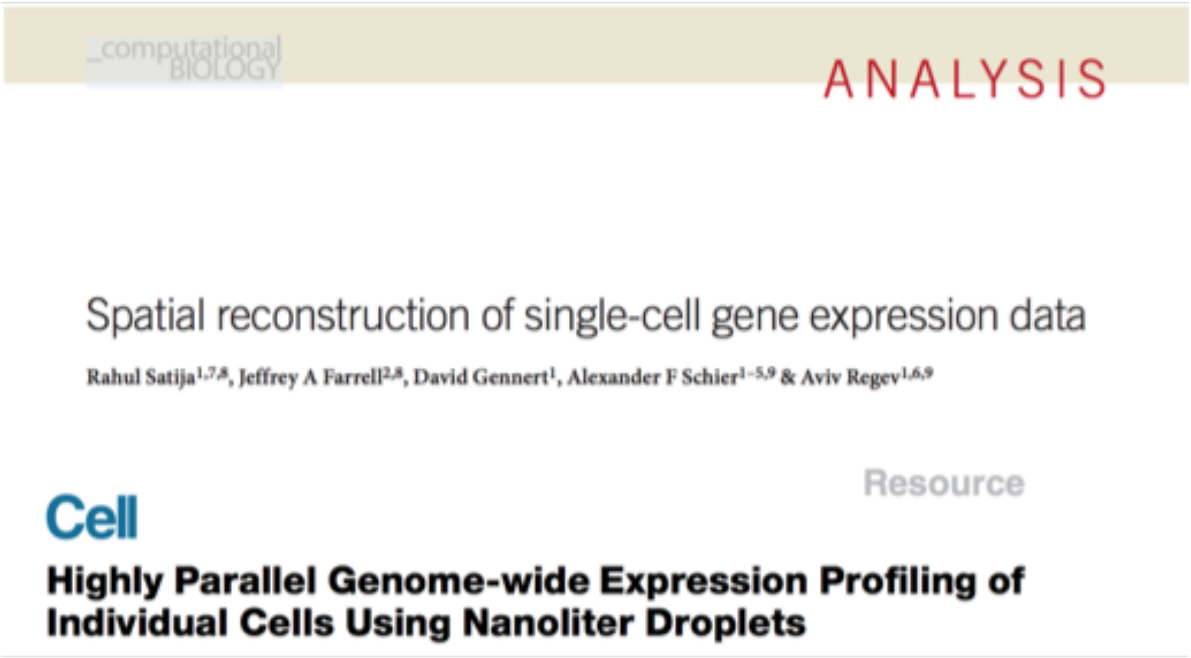
  Objective is to have relative gene expression to eliminate technical factors that impact the variation in the number of molecules per cell

  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

Seurat: R scRNA-Seq Analysis Package



computational BIOLOGY

ANALYSIS

Spatial reconstruction of single-cell gene expression data

Rahul Satija[1,7,8], Jeffrey A Farrell[2,8], David Gennert[1], Alexander F Schier[1–5,9] & Aviv Regev[1,6,9]

Resource

Cell

Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets

https://github.com/satijalab/seurat

Prepping Counts for Seurat

3 prime

- **Expected by Seurat.**
- **Counts collapsed with UMIs.**
- **Log2 transform (in Seurat).**
- **Account for sequencing depth (in Seurat).**

Full Transcript Sequencing

- **Can be used in Seurat.**
- **TPM +1 transformed counts.**
- **Log2 transform (in Seurat).**
- **Sequencing depth is already accounted.**

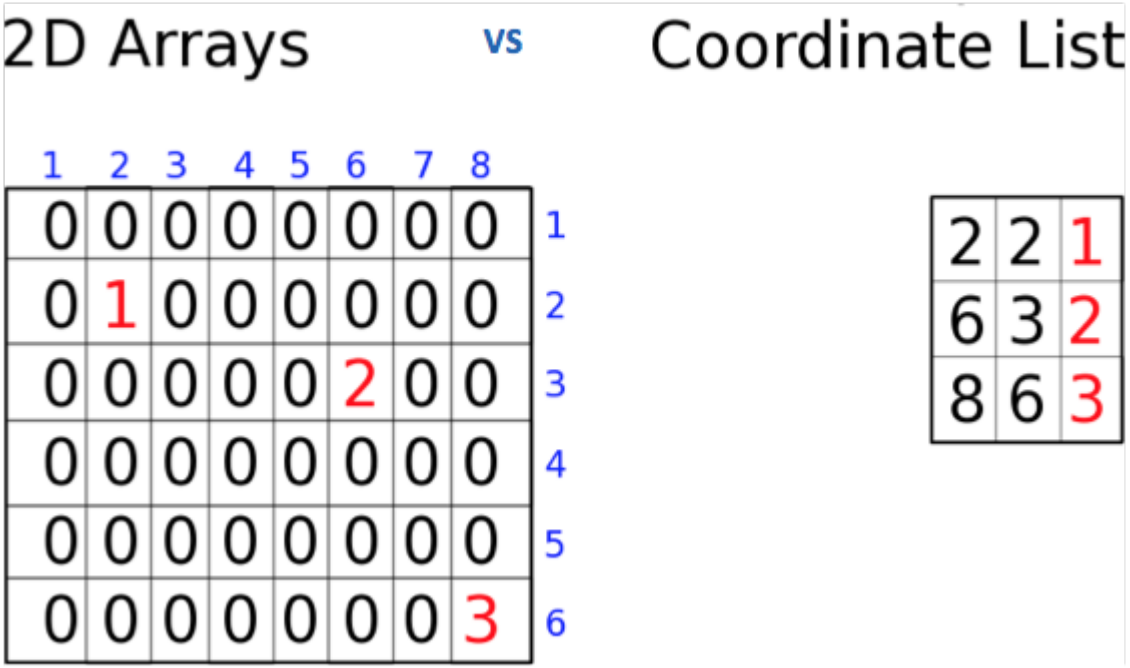
What is a Sparse Matrix?

- **Sparse Matrix**
  - **A matrix where most of the elements are 0.**
- **Dense Matrix**
  - **A matrix where most elements are not 0.**
- **Many ways to efficiently represent a sparse matrix in memory.**
  - **Here, the underlying data structure is a coordinate list.**

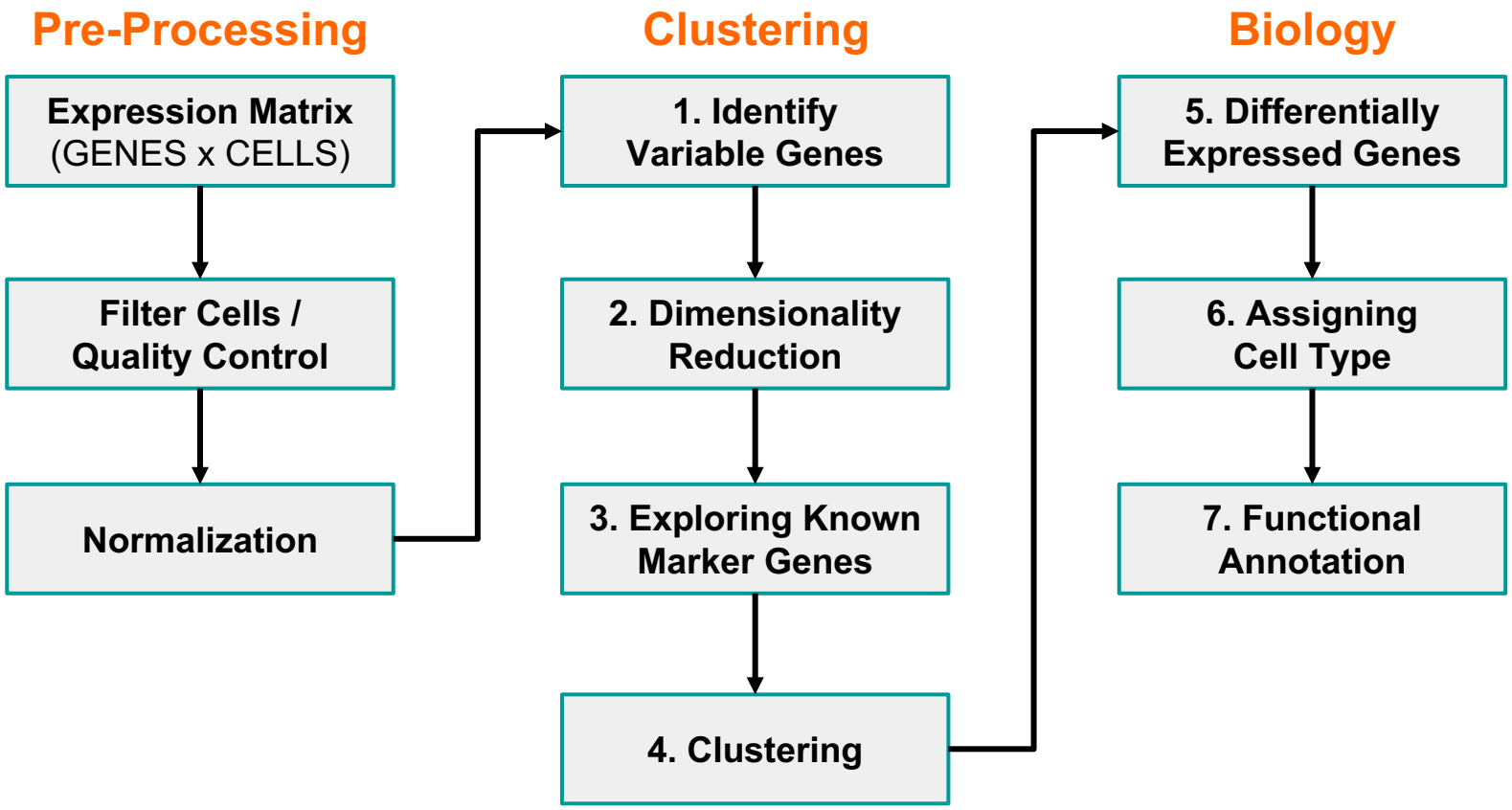
2D Array vs a Coordinate List

Can be optimal for dense matrices

More optimal for sparse matrices



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



1. Making Sense of Variation

- **Fact 1** : For something to be informative, it needs to exhibit variation

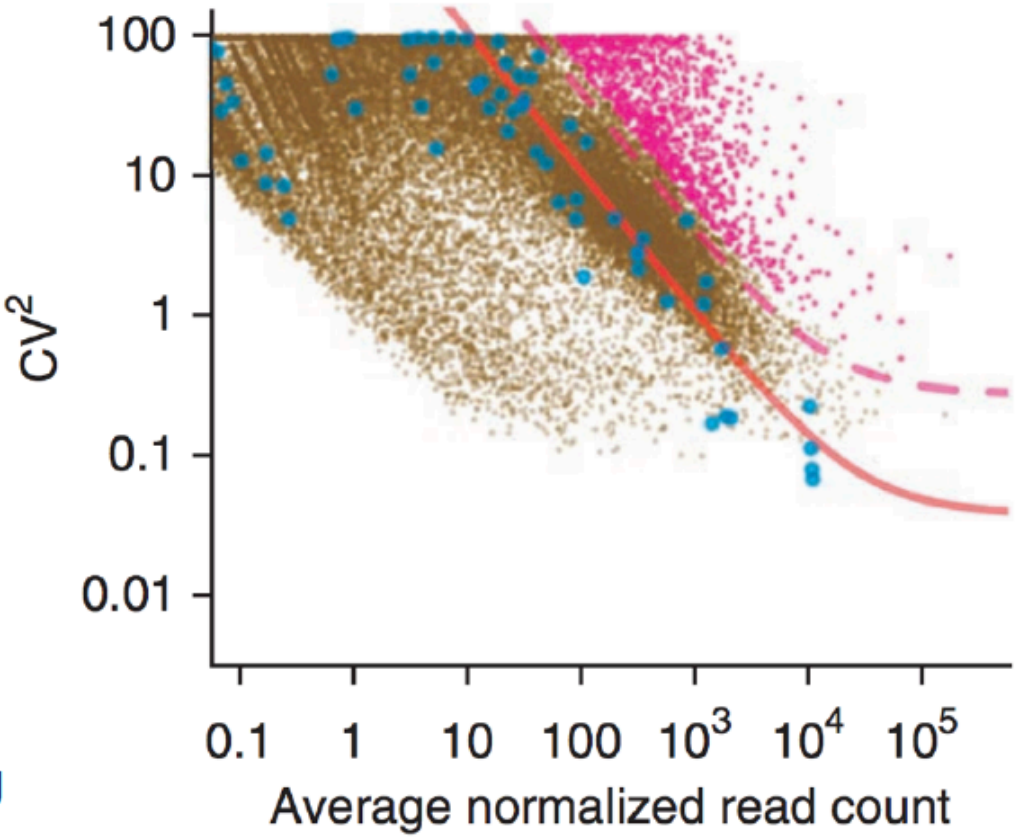
- **Fact 2** : Not everything that exhibits variation in real life, is informative

1. Identifying relevant, "highly variable" features

First filter out,

- Lowly expressed genes
- "Housekeeping" genes

Typical practice to identify "highly" variable genes is to create a null model of statistical variation based on housekeeping or spike-in genes



Brennecke et al., *Nature Methods*, 2013

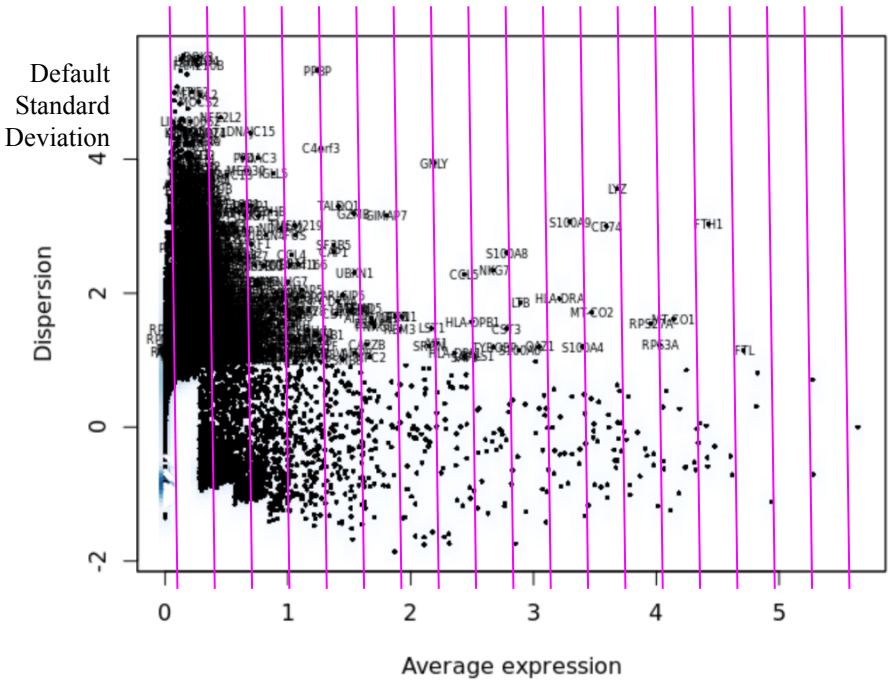
Slide courtesy of Karthik Shekhar

Variable Genes in Seurat

Calculate mean expression.

Calculate disperstion (standard deviation).

Calculate z-score for dispersions within each bin.



Stratifies and controls from the relationship between the variability and mean expression.

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

Cells are in 20,000 dimensional space

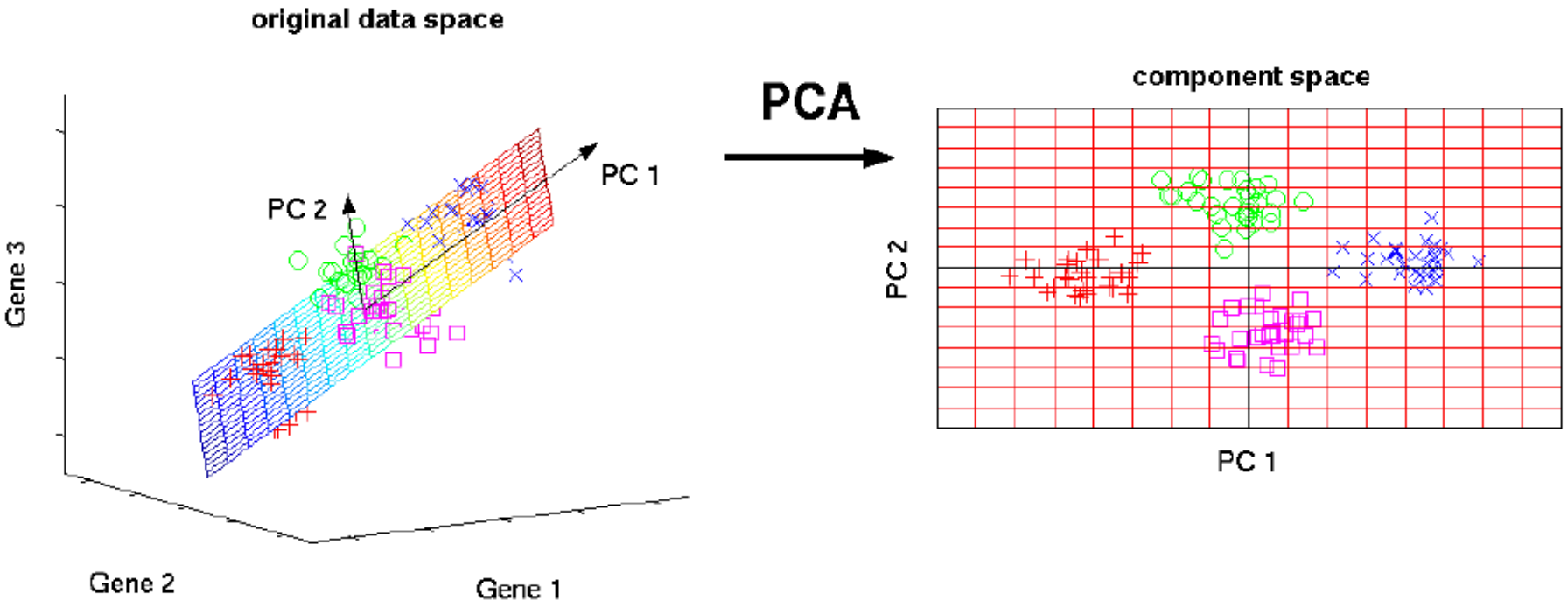
- many genes are lowly detected / noisy measurements
- genes are not independent of one another! rather they operate in coregulatory modules

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*

2. Dimensionality reduction

- **Why? :** Genes do not act independently, but as coregulatory "modules". E.g. in a cell type, the activity of a handful of transcription factors might lead to the co-expression of hundreds of genes defining cell-identity
- Cells occupy a low dimensional manifold in gene-expression space defined by these modules



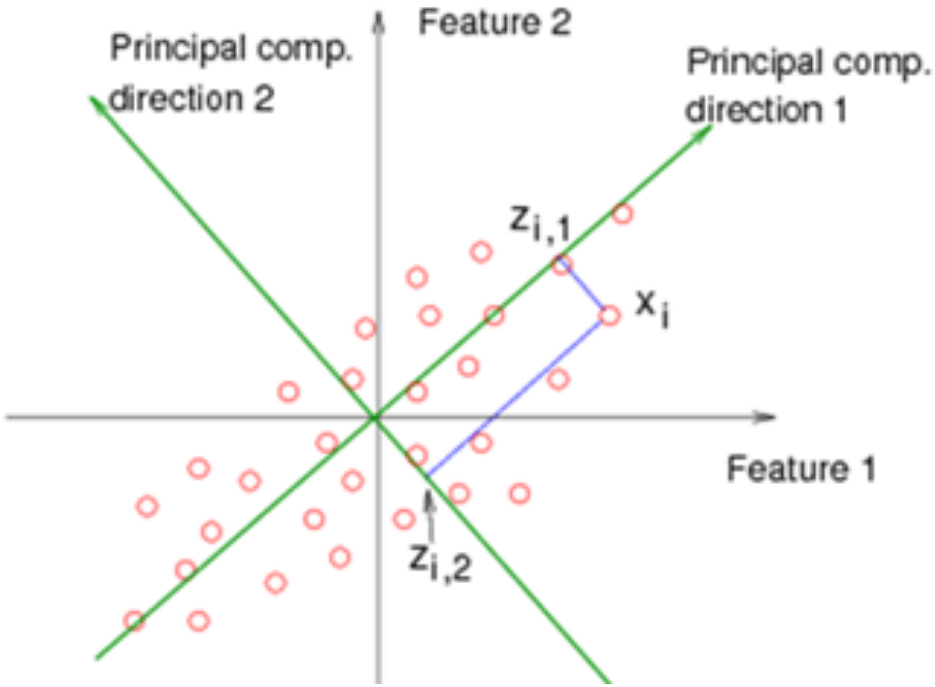
- Principal Component Analysis (PCA) is a **popular linear-method** to identify these modules

Slide courtesy of Karthik Shekhar

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

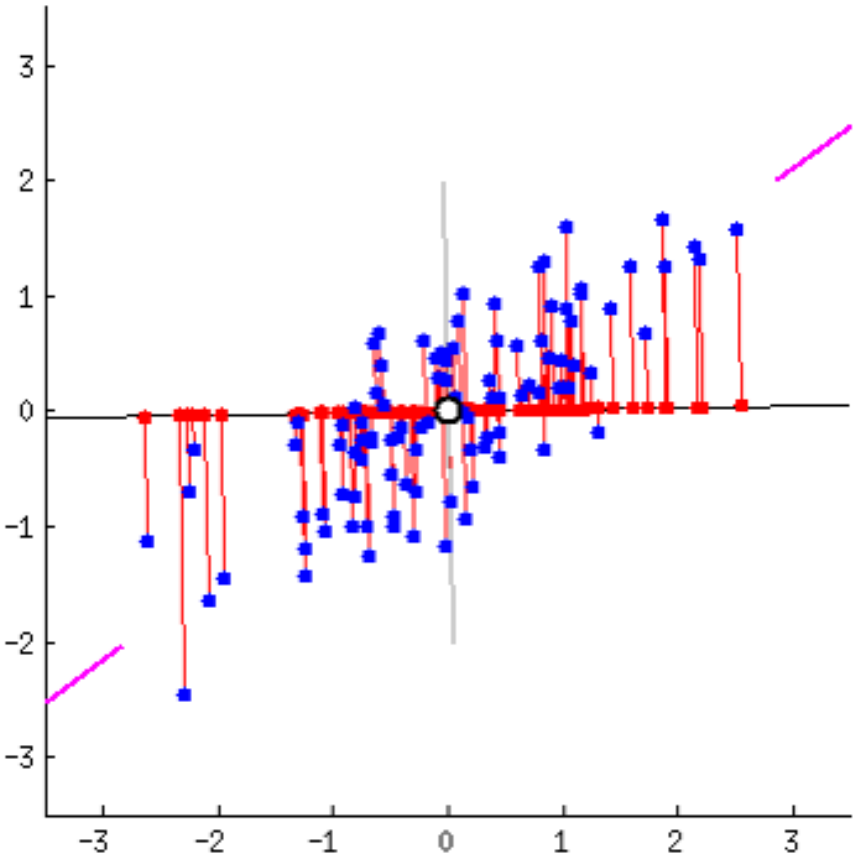


Determining cell type, state, and/or function:
2. Dimensionality reduction

- 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

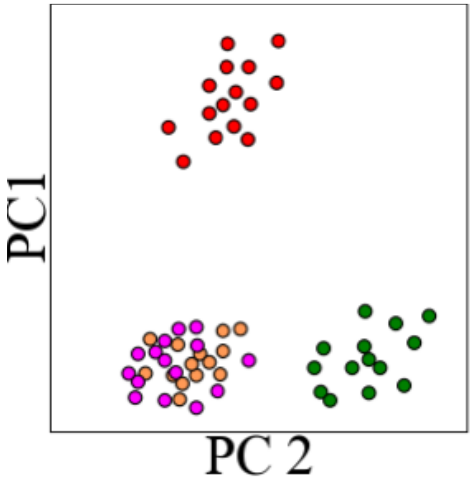
Identifying maximal orthogonal sources of variation



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

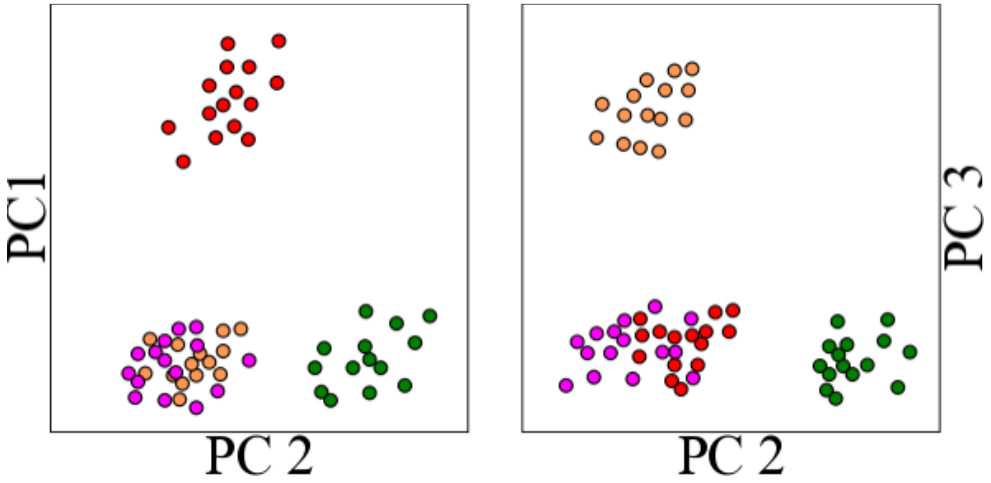
PCA of single cell data



- 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

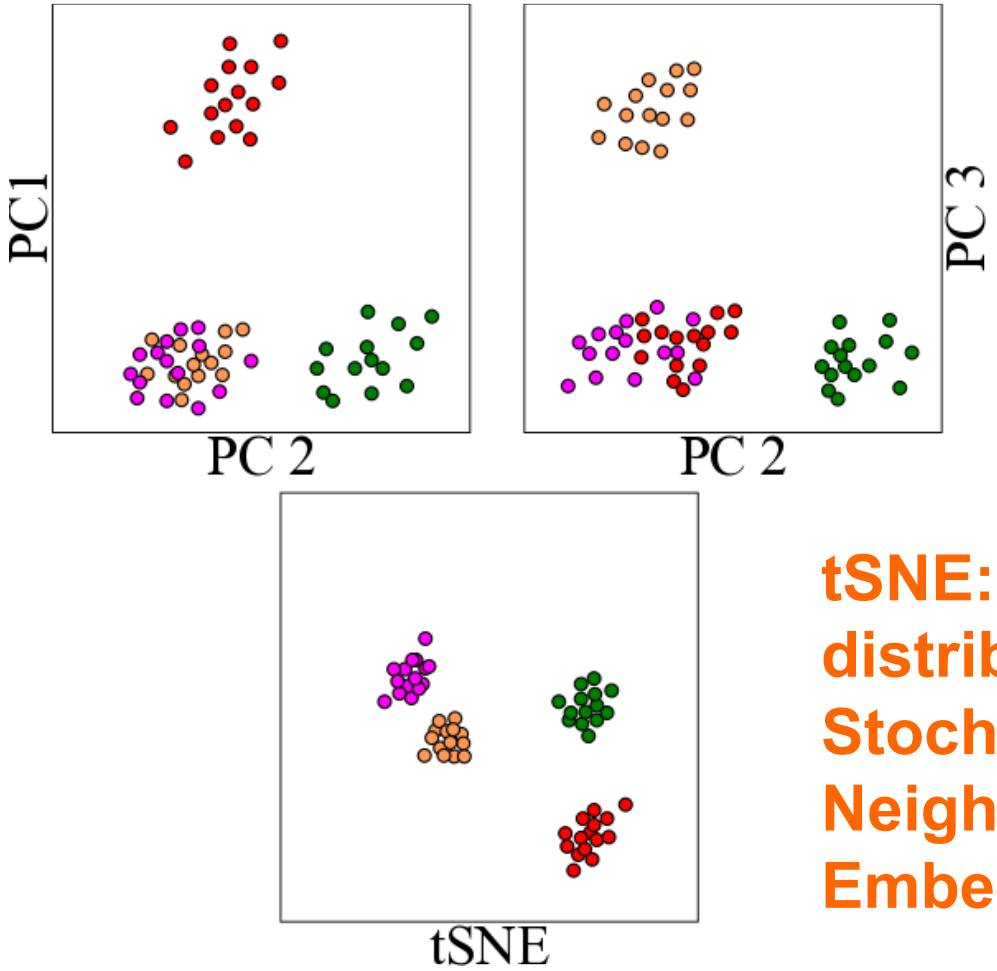
PCA of single cell data



- 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

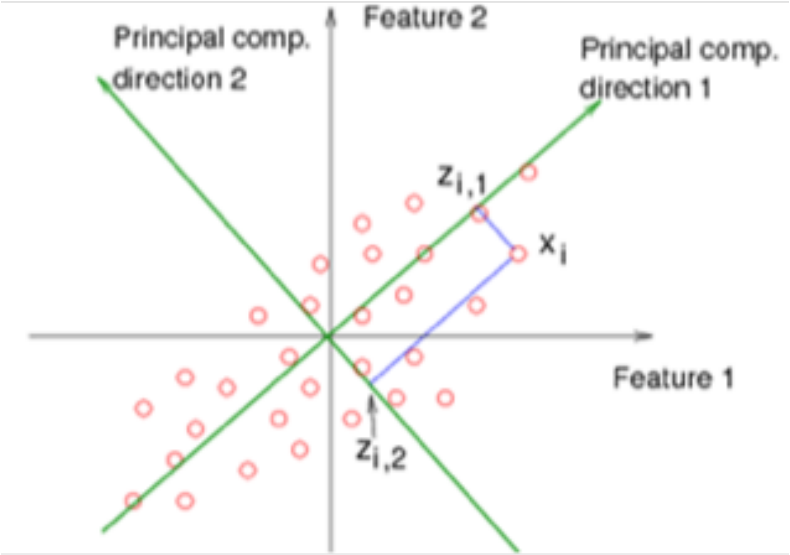
- tSNE is nonlinear dimensionality reduction
- tSNE collapse the visualization to 2D

Dimensionality Reduction

- Start with many measurements (high dimensional).
  - Want to reduce to few features (lower-dimensional space).
- One way is to extract features based on capturing groups of variance.
- Another could be to preferentially select some of the current features.
  - We have already done this.
- We need this to plot the cells in 2D (or ordinate them)
- In scRNA-Seq PC1 may be complexity or technical.

PCA: Overview

- Eigenvectors of covariance matrix.
- Find orthogonal groups of variance.
- Given from most to least variance.
  - Components of variation.
  - Linear combinations explaining the variance.



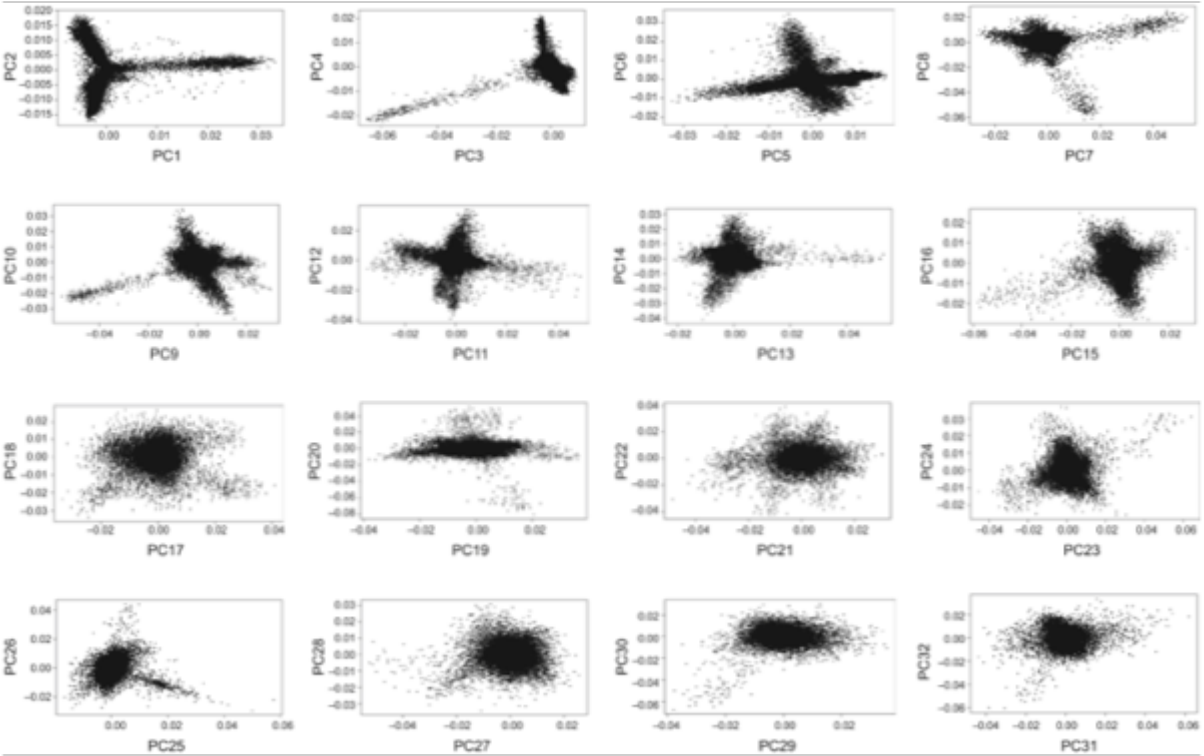
PCA: in Practice

Things to be aware of-

- **Data with different magnitudes will dominate.**
  - **Zero center and divided by SD.**
- **(Standardized).**
- **Can be affected by outliers.**
- **Data is often first filtered to remove noise.**

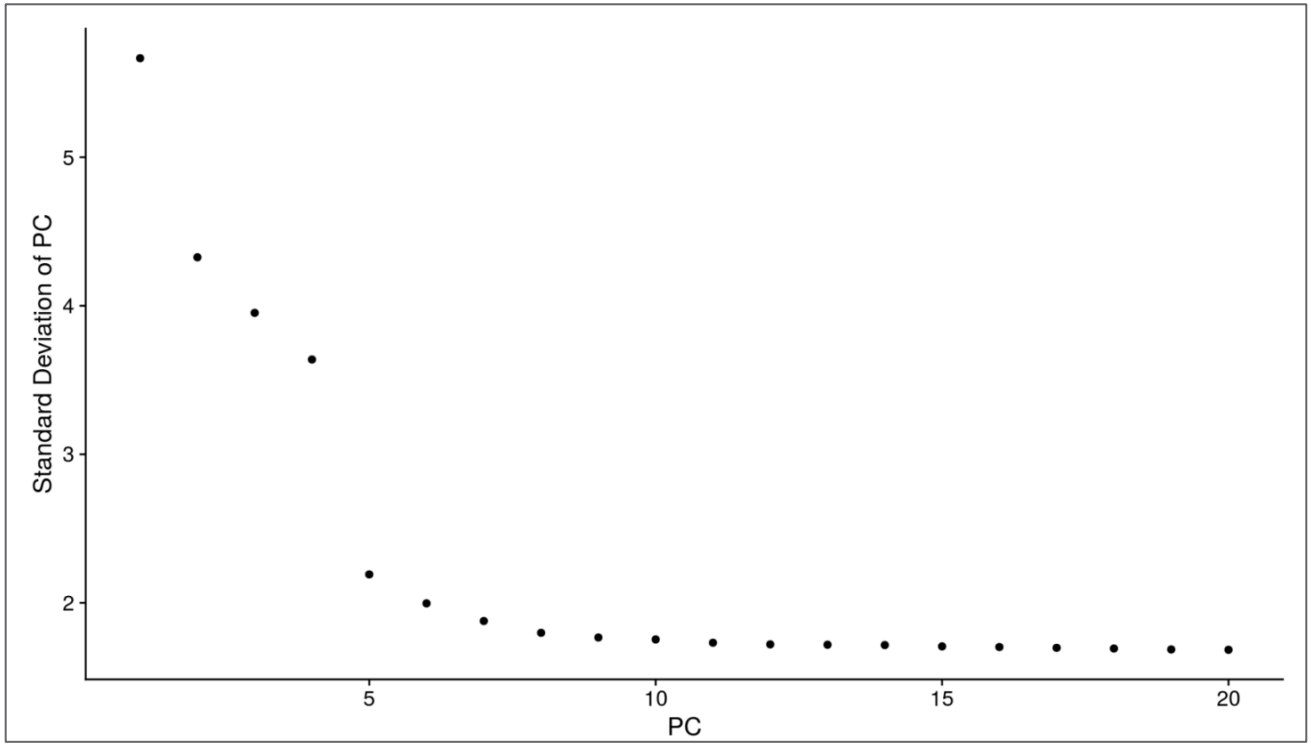
PCs

Notice how lower PCs look more and more “spherical” - this loss of structure indicates that the variation captured by these PCs mostly reflects noise.



How Many Components Should We Use?

Elbow Plot (Scree Plot)



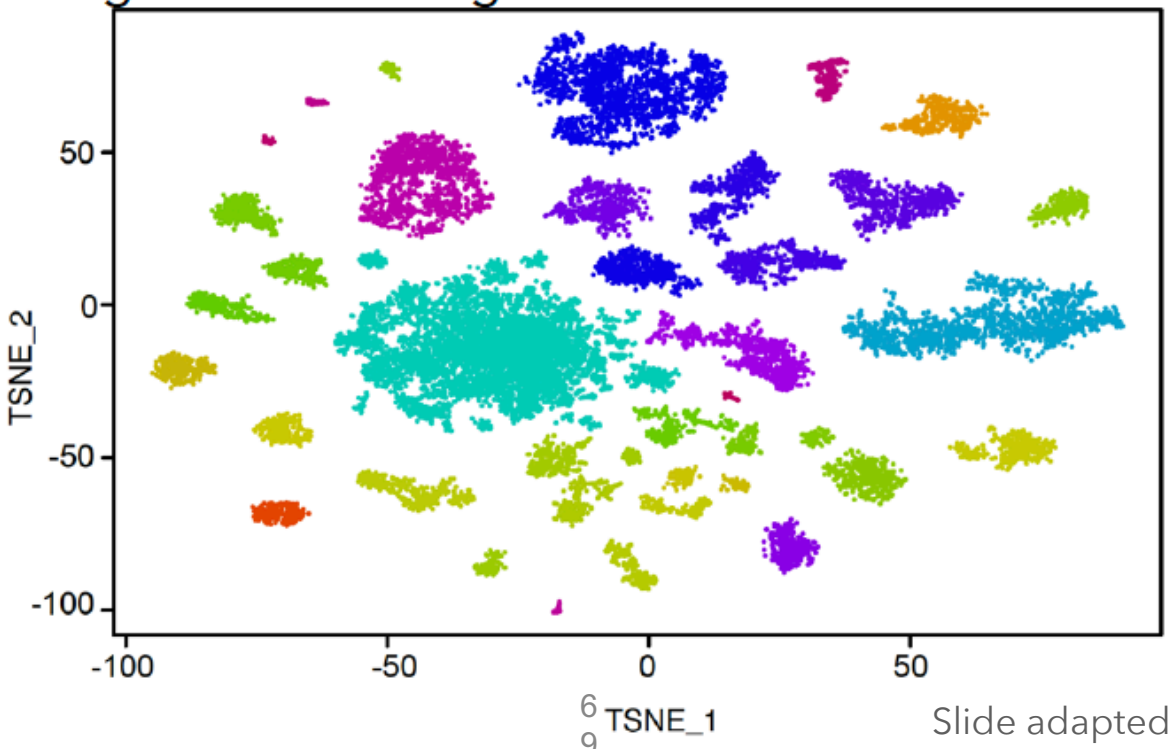
3. Visualization

t-distributed Stochastic Neighbor Embedding

- A non-linear embedding method that preserves local distances between data-points in the low-dimensional space

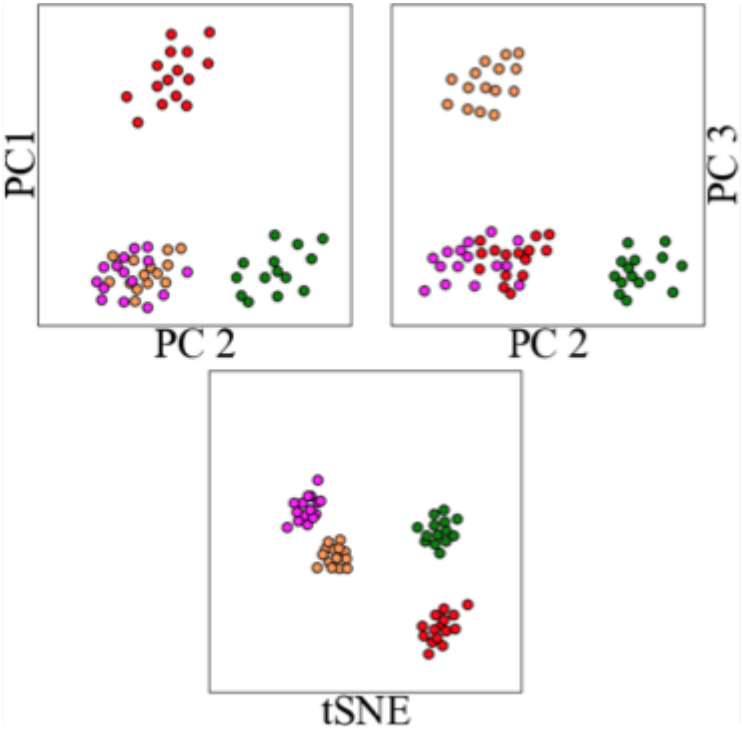
$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}},$$

- t-SNE embedding of 45,000 retinal neurons sequenced using Drop-seq and clustered using the DBScan algorithm



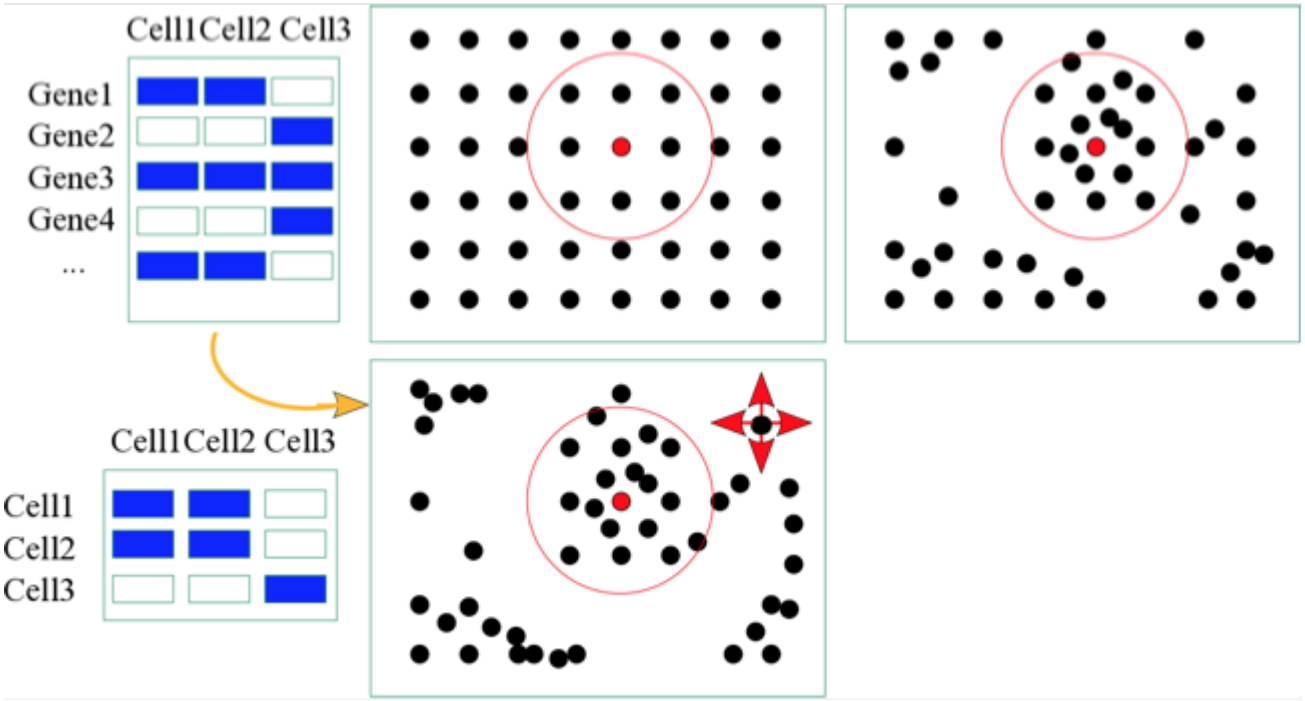
Slide adapted from Karthik Shekhar

t-SNE: Collapsing the Visualization to 2D



t-SNE: Nonlinear Dimensionality Reduction

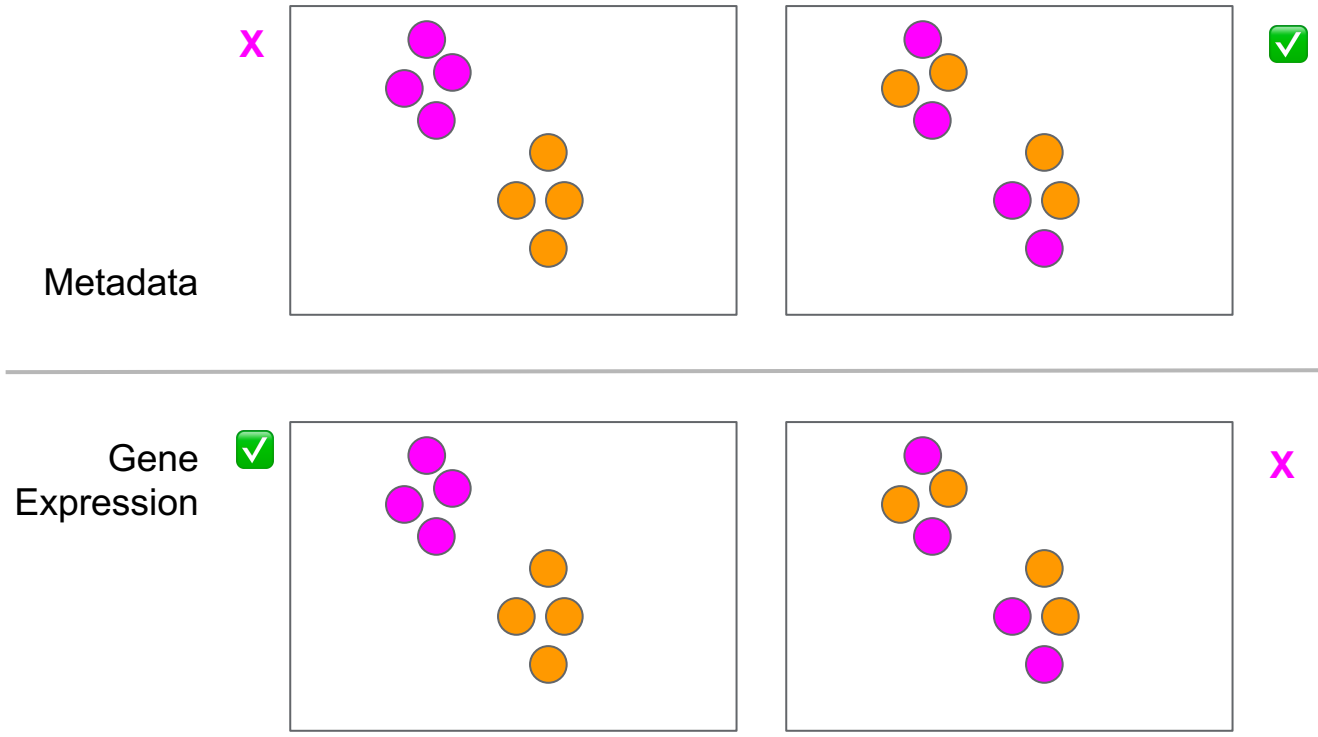
t-SNE: How it Works



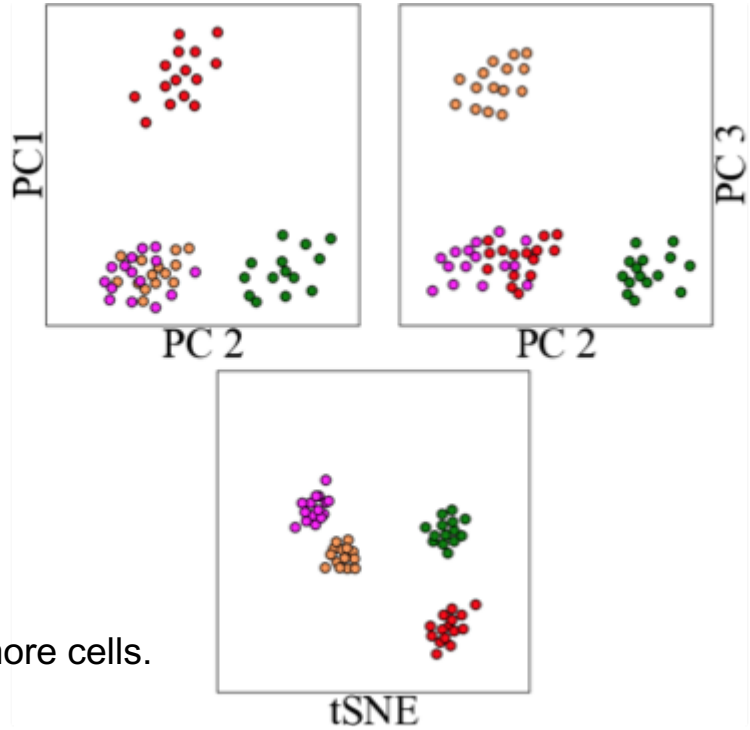
PCA and t-SNE Together

- **Often t-SNE is performed on PCA components**
  - **Liberal number of components.**
  - **Removes mild signal (assumption of noise).**
  - **Faster, on less data but, hopefully the same signal.**

Plotting Metadata on Ordinations



Caution When Interpreting t-SNE

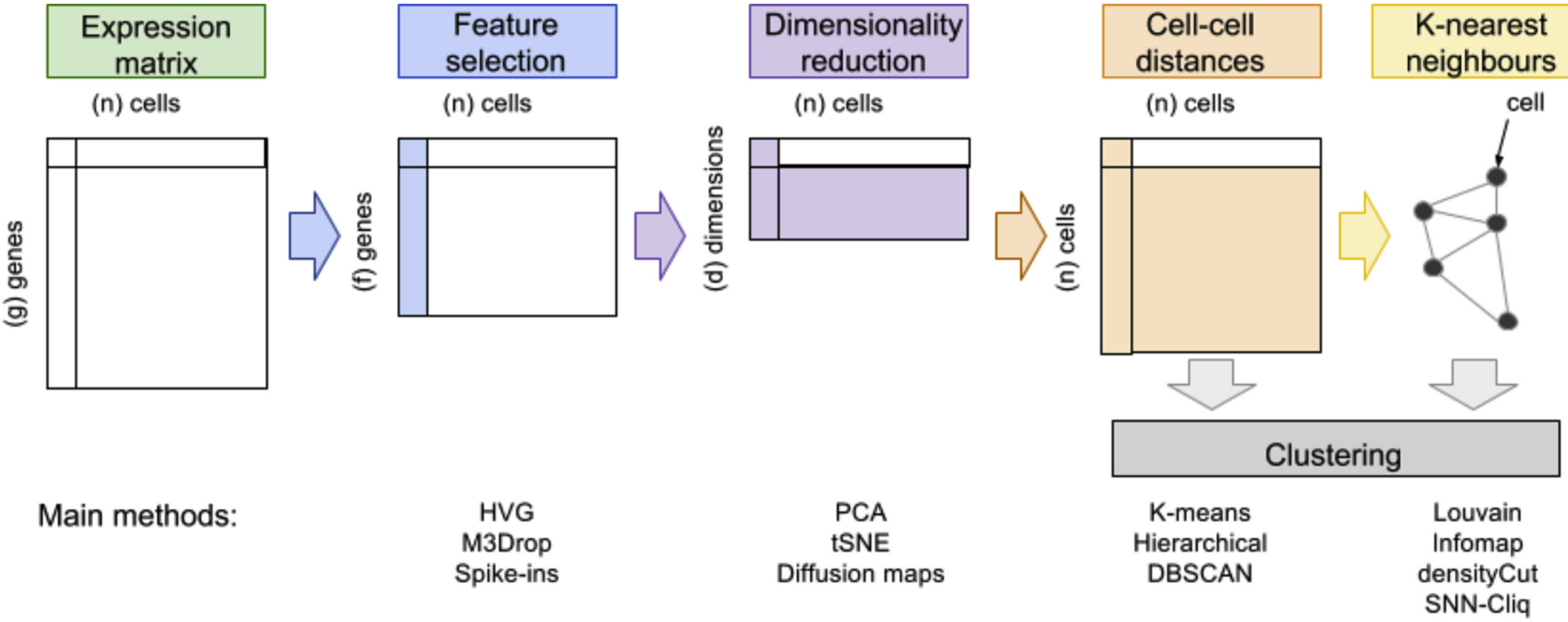


Nonlinear
Optimized for local distanct
Big clusters can just mean more cells.

Learn More About t-SNE

- **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 Tensorflow**
  - **http://projector.tensorflow.org/**

4. Clustering cells to identify cell-types



Andrews TS and Hemberg M. Mol Aspects Med. 2018

Defining Clusters Through Graphs

The European Physical Journal B
November 2013, 86:471

A smart local moving algorithm for large-scale modularity-based community detection

Authors | Authors and affiliations

Ludo Waltman, Nees Jan van Eck

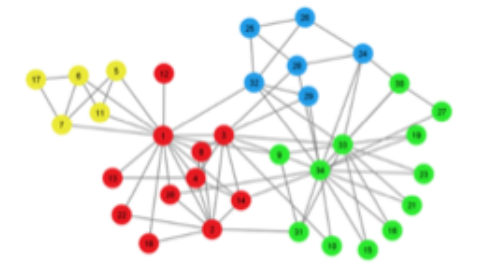
Regular Article

First Online: 13 November 2013

DOI: 10.1140/epjb/e2013-40829-0

Cite this article as:
Waltman, L. & van Eck, N.J. Eur. Phys. J. B (2013) 86: 471. doi:10.1140/epjb/e2013-40829-0

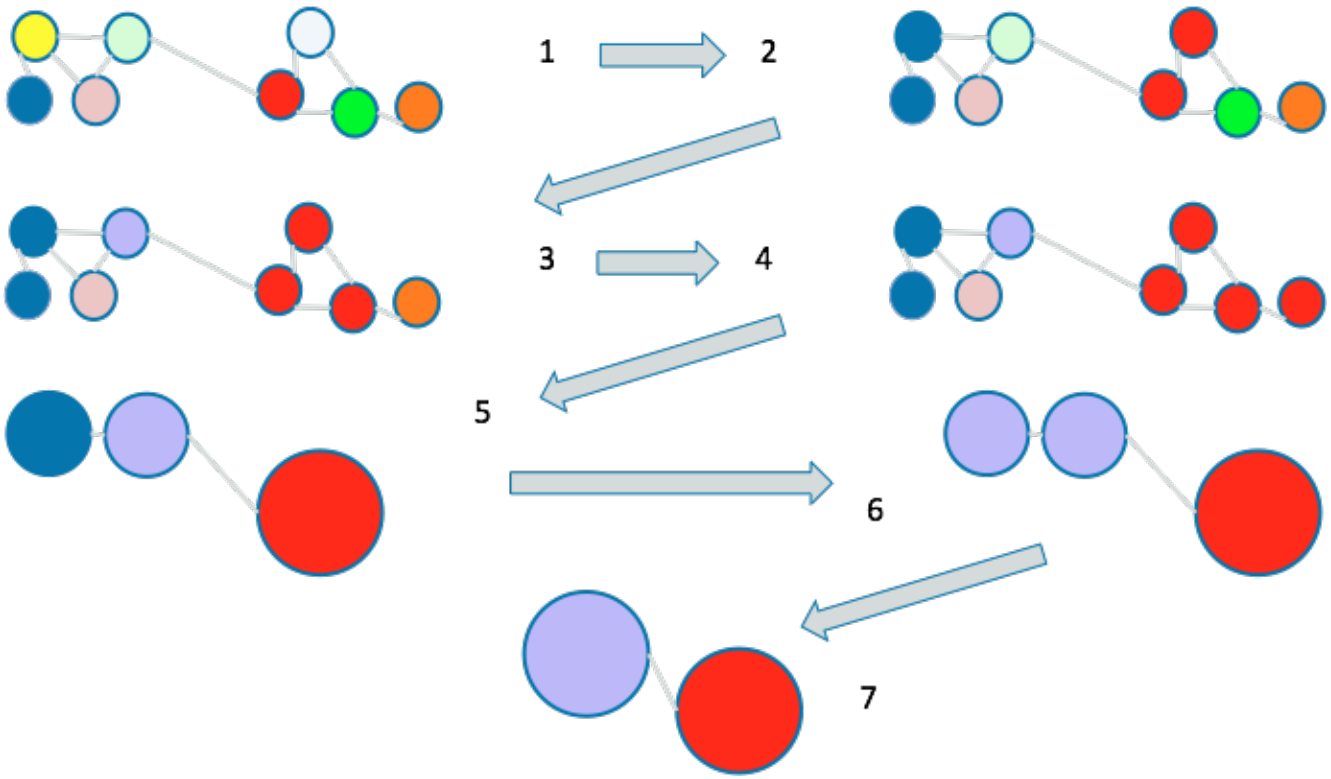
38 Citations | 768 Downloads

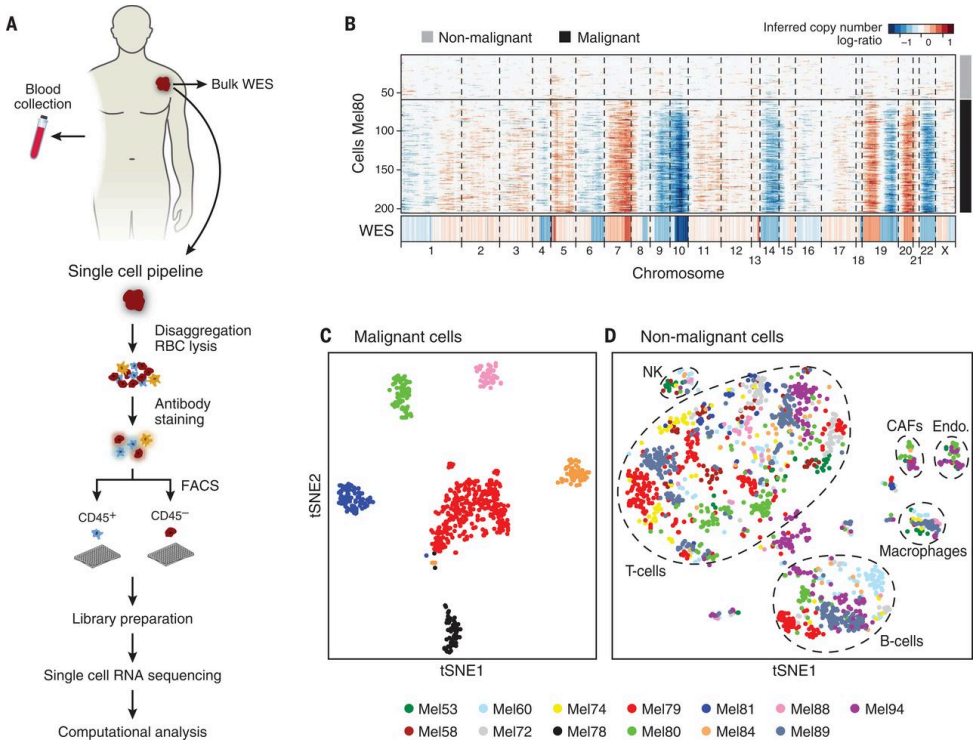


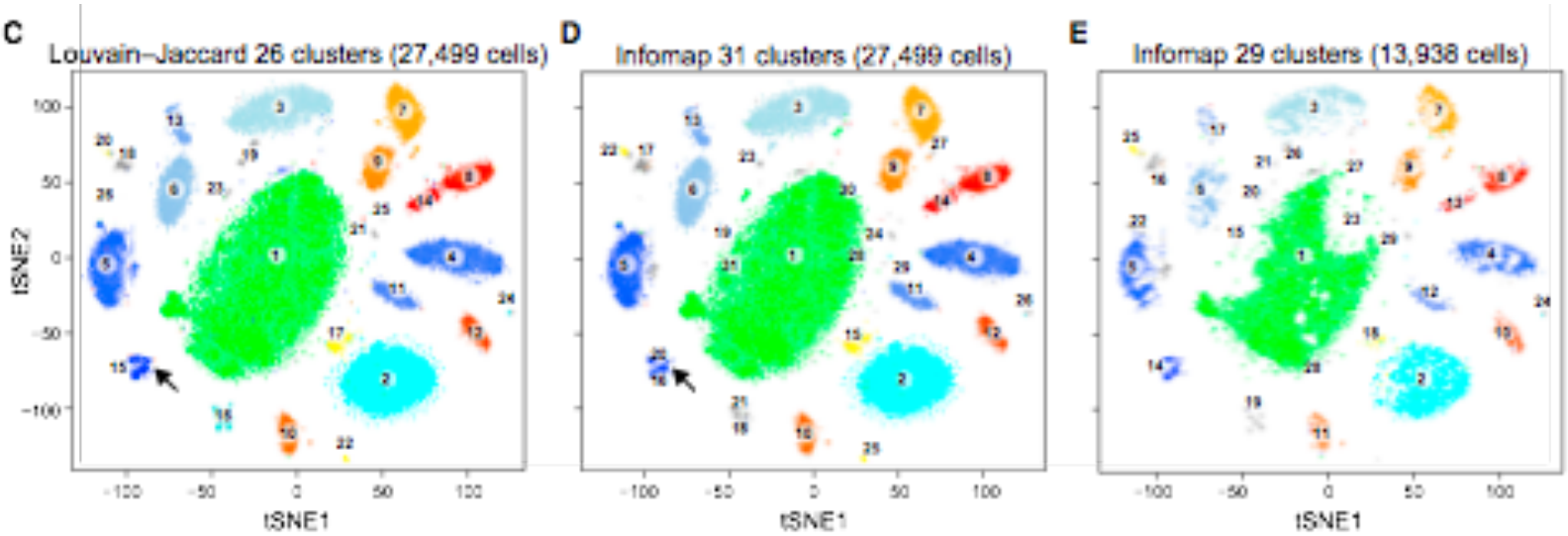
- Smart Local Moving (SLM) algorithm for community (cluster) detection in large networks.
  - Can be applied to 10s of millions cells, 100s of millions of relationships.
  - Evolved from the Louvain algorithm

http://www.ludowaltman.nl/slm/

Local Moving Heuristic

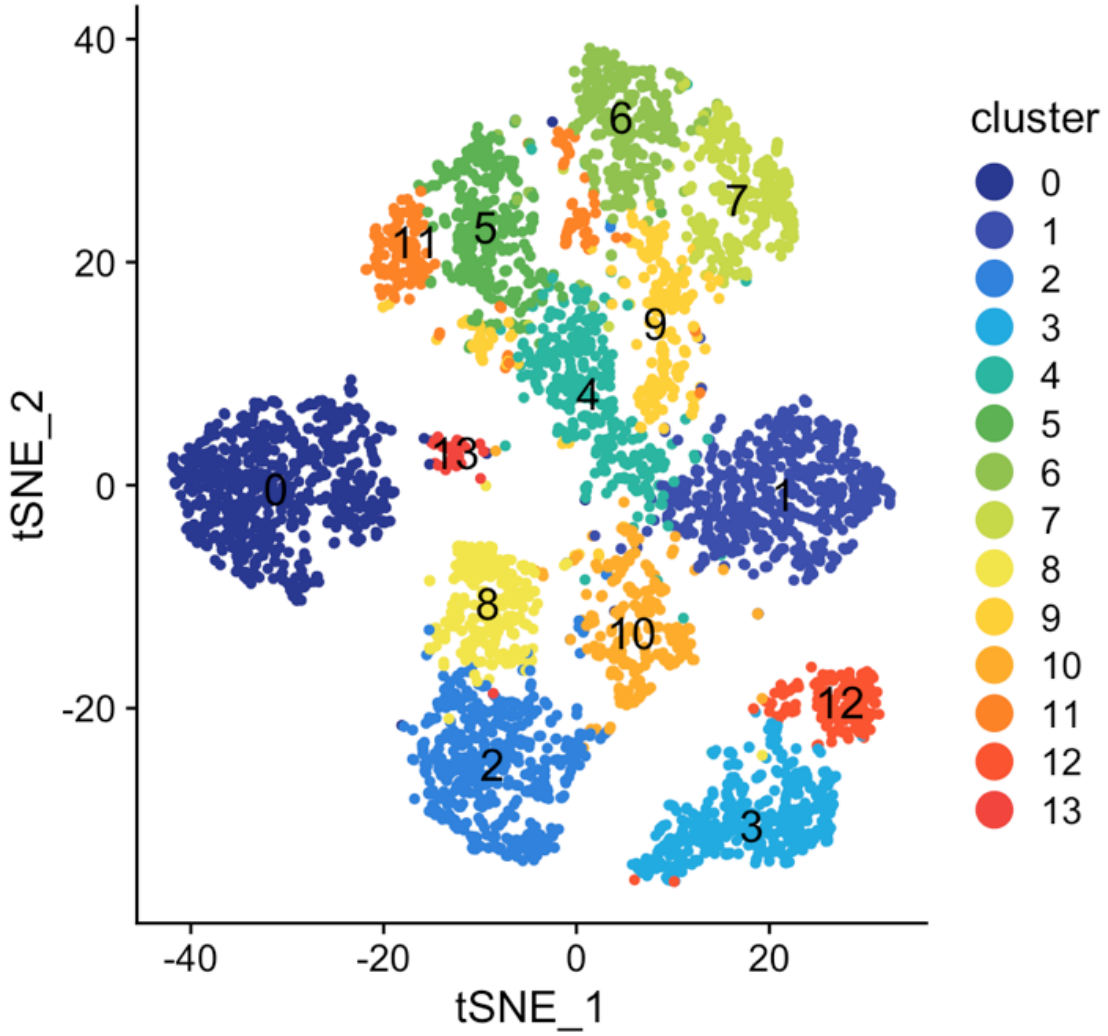






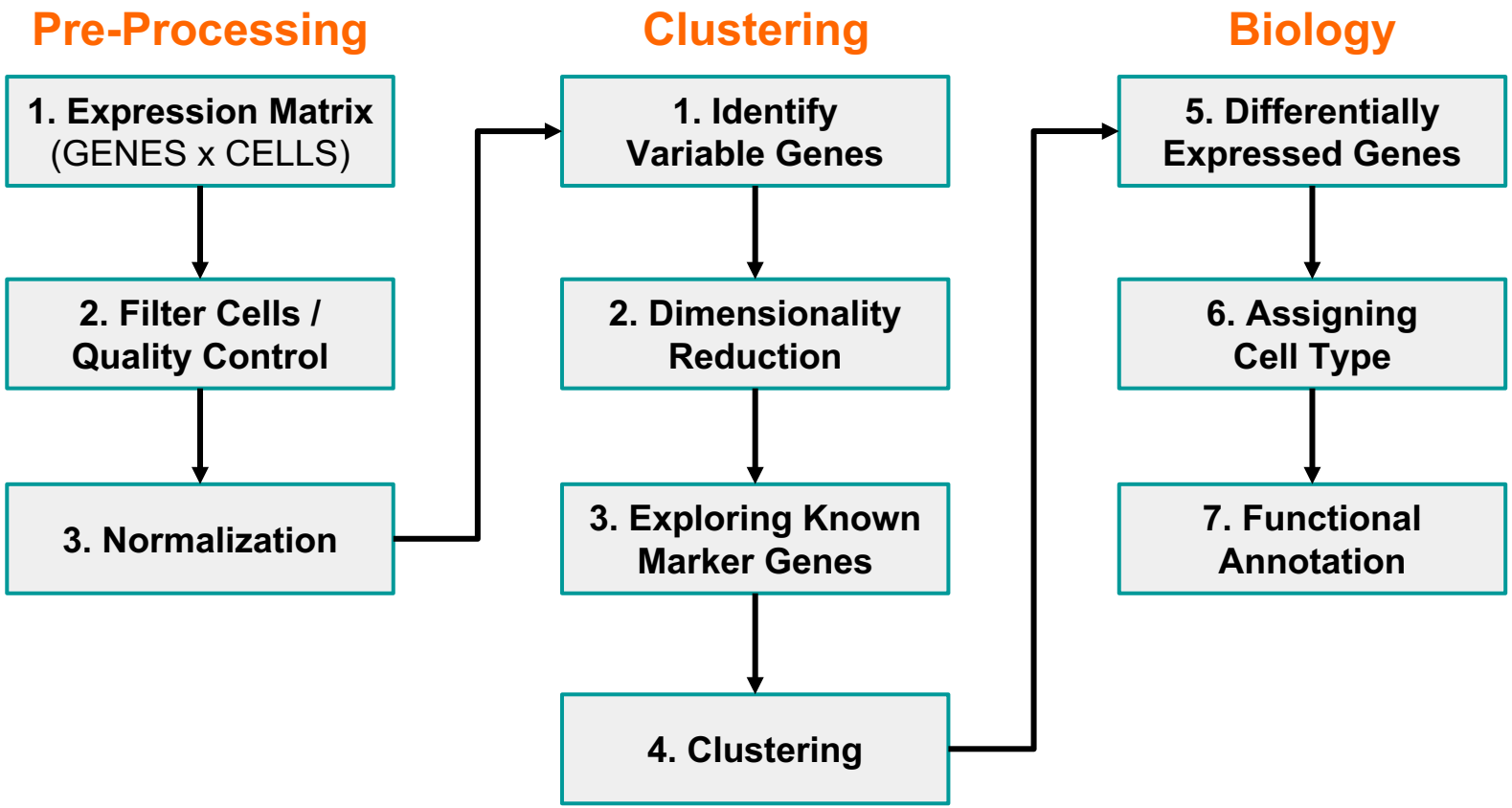
Tirosh and Izar et al. Science 2016
Shekhar et al. Cell 2016

Determining cell type, state, and/or function:
3. Visualization

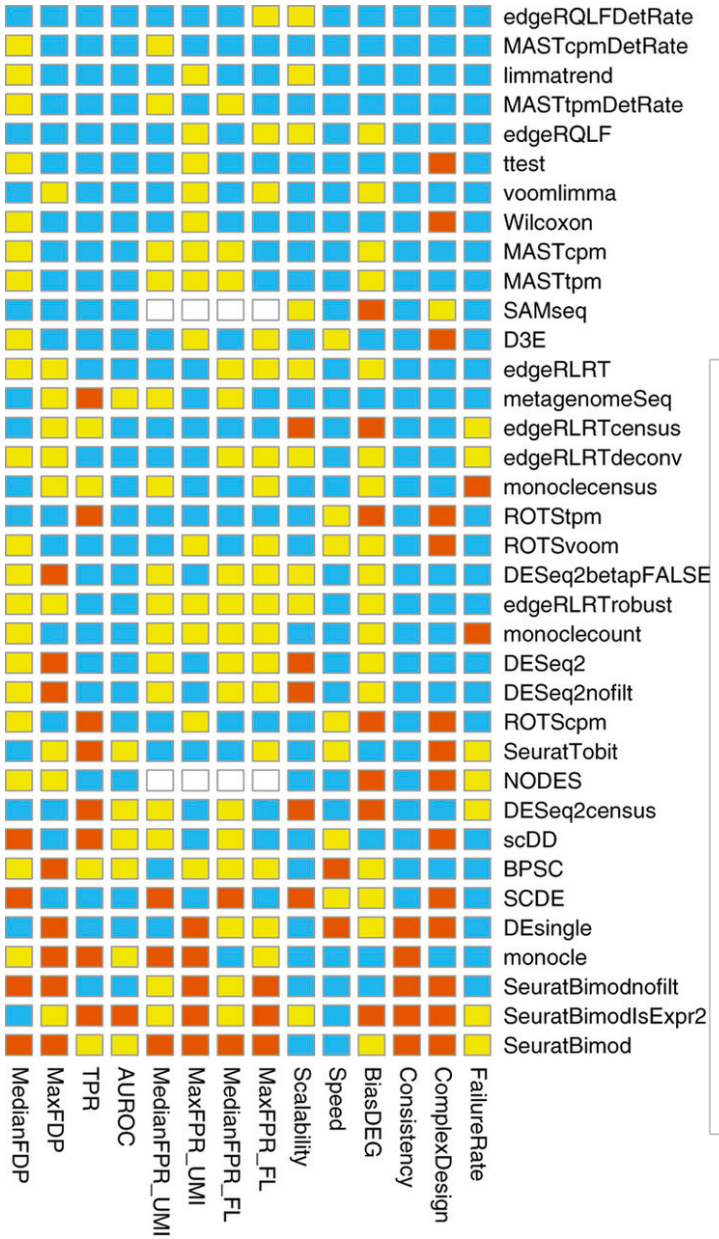


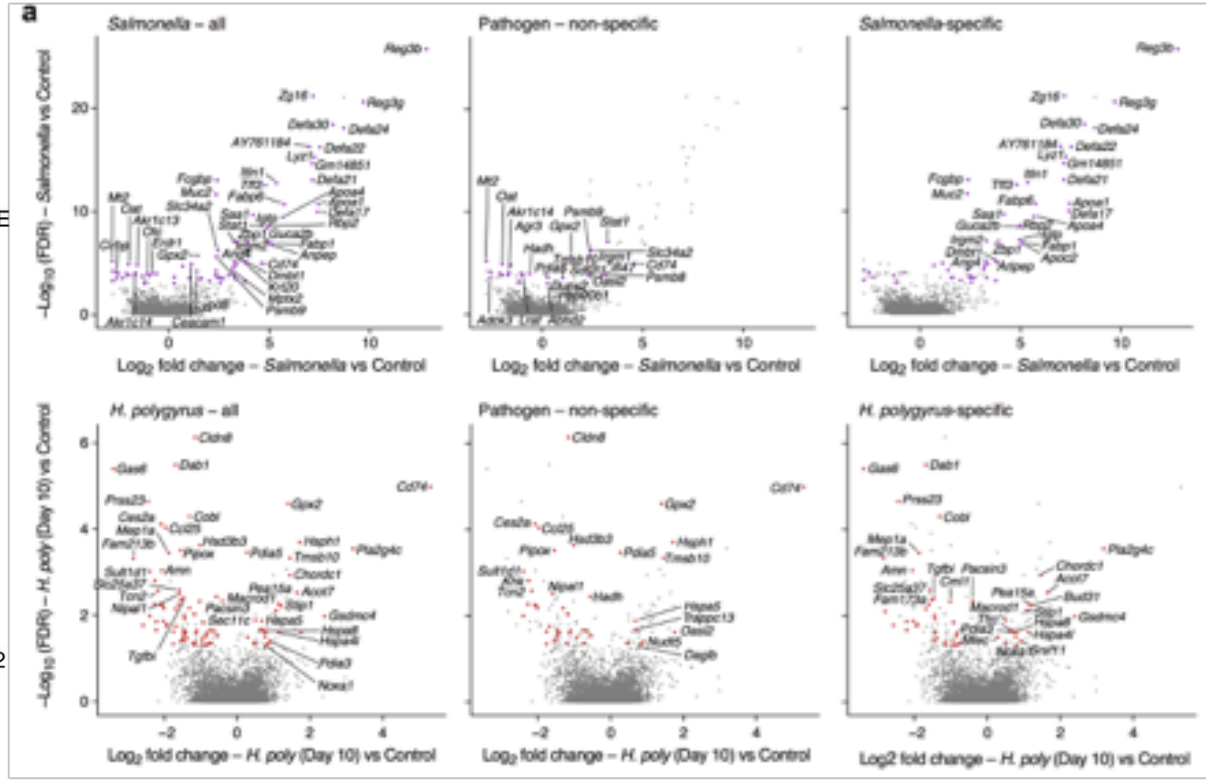
A great tSNE resource! https://distill.pub/2016/misread-tsne/

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



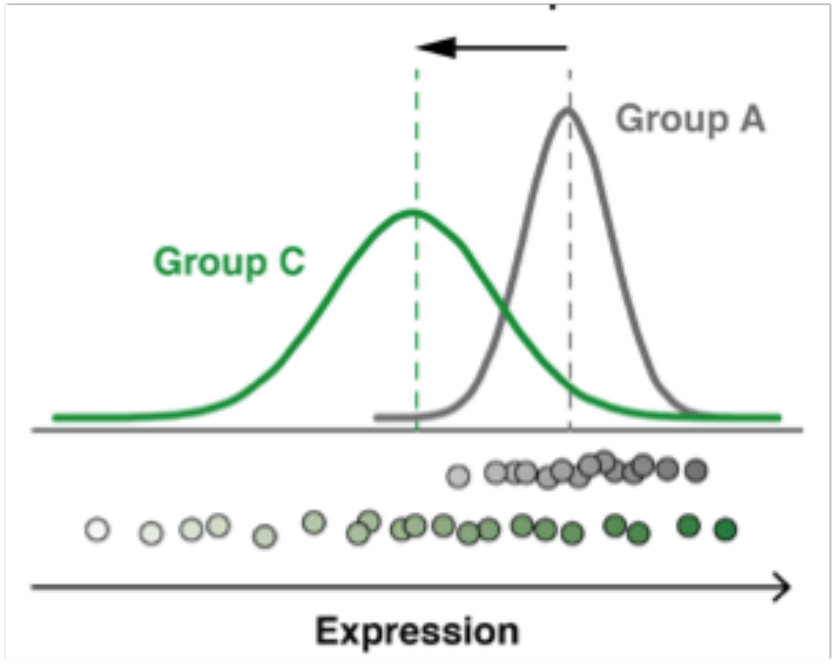
5. Assigning cell identity & comparing across conditions: Differential Expression Analysis



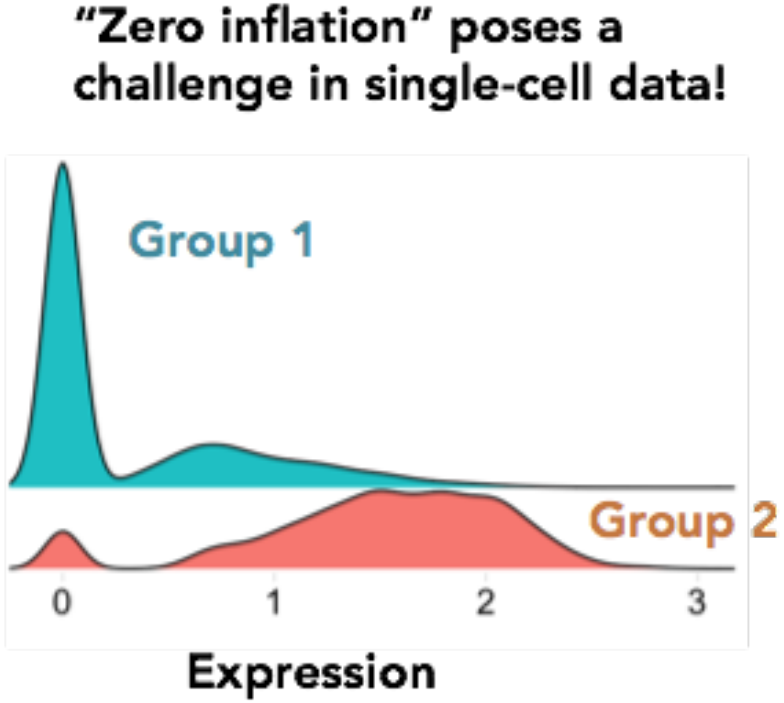


Soneson and Robinson. Nat Methods 2018
Haber, Moshe and Rogel et al. Nature 2017

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



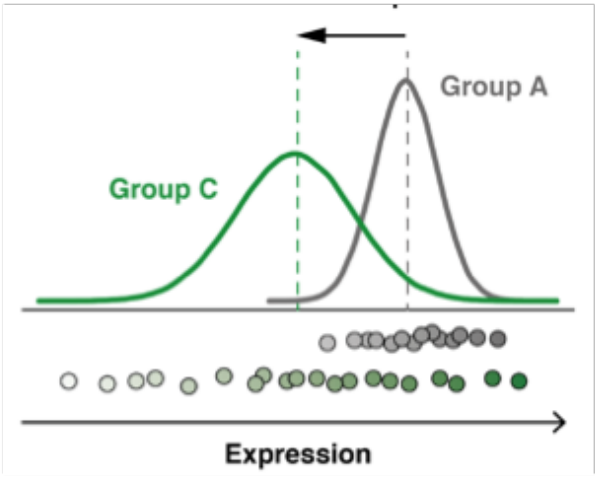
Bulk



Single cell

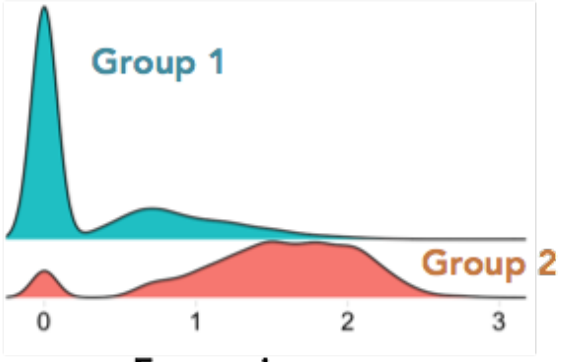
Differential Expression

Group A > Group B (p-value < 0.01)



BUT

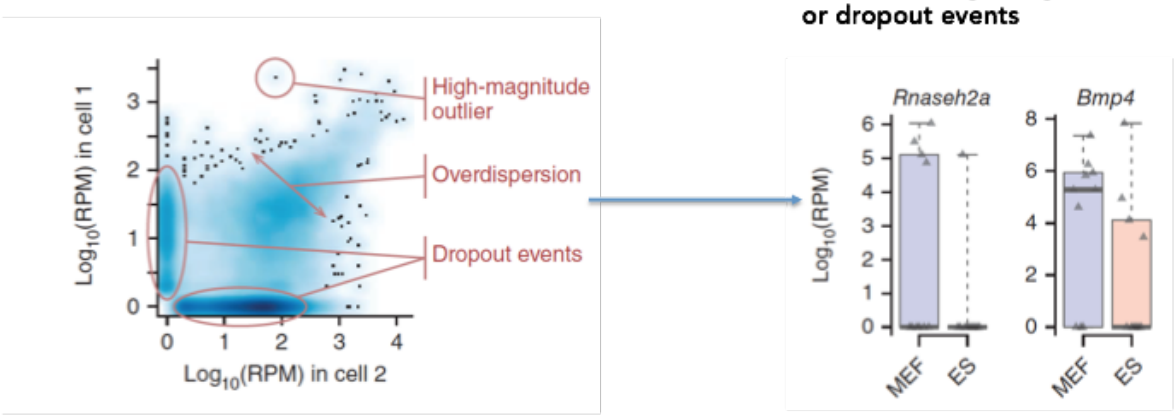
"Zero inflation" poses a challenge in single-cell data!



Conventional statistical tests (e.g. "Student's t"), which assume a unimodal distribution can be underpowered in detecting true genes

Single Cell Differential Expression (SCDE)

DE genes using conventional methods can include high magnitude outliers or dropout events



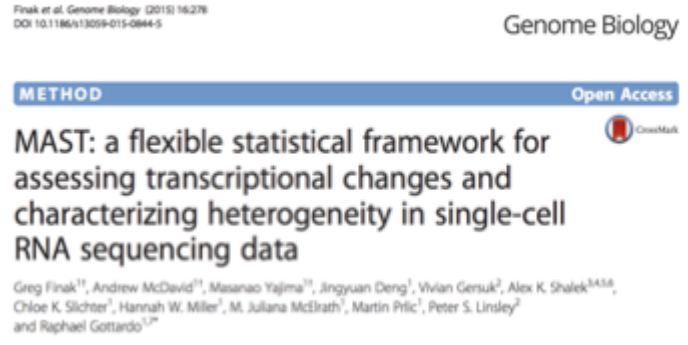
SCDE exchanges information between closely related cells to estimate dropout rates for every cell!

$$\begin{cases} r_1 \approx Poisson(\lambda_0) & \text{Dropout in } c_1 \\ \begin{cases} r_1 \approx NB(r_2) \\ r_2 \approx NB(r_1) \end{cases} & \text{Amplified} \\ r_2 \approx Poisson(\lambda_0) & \text{Dropout in } c_2 \end{cases}$$

Kharchenko et al., 2014

MAST

- **Uses hurdle model**
  - **Two part generalized linear model to address both rate of expression (prevalence) and expression.**
  - **GLM means covariates can be used to control for unwanted signal.**
- **CDR: Cellular detection rate**
  - **Cellular complexity**
  - **Values below a threshold are 0**


Finak et al. Genome Biology (2015) 16:278
DOI 10.1186/s13059-015-0844-5

Genome Biology

METHOD Open Access

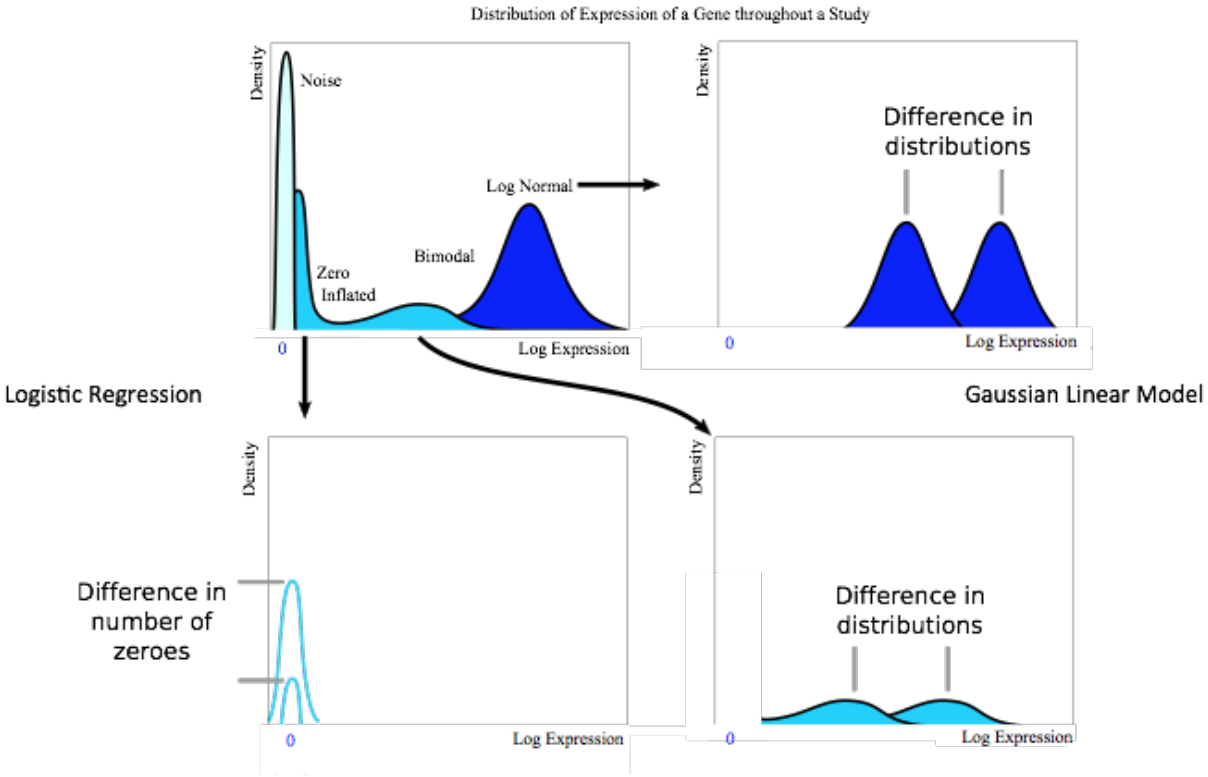
MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data

Greg Finak[1†], Andrew McDavid[1†], Masanao Yajima[1†], Jingyuan Deng[1], Vivian Gersuk[2], Alex K. Shalek[3,4,5,6], Chloe K. Slichter[7], Hannah W. Miller[7], M. Juliana McElrath[7], Martin Prlic[1], Peter S. Linsley[2] and Raphael Gottardo[1,7*]

Additionally introduces a GSEA method

https://github.com/RGLab/MAST

MAST: Hurdle Models



Seurat: Differential Expression

- Default if one cluster again many tests.
  - Can specify an ident.2 test between clusters.
- Adding speed by excluding tests.
  - Min.pct - controls for sparsity
  - Min percentage in a group
  - Thresh.test - must have this difference in averages.

Seurat: Many Choices of DE

Bimod
- Tests differences in mean and proportions.

Roc
- Uses AUC like definition of separation.

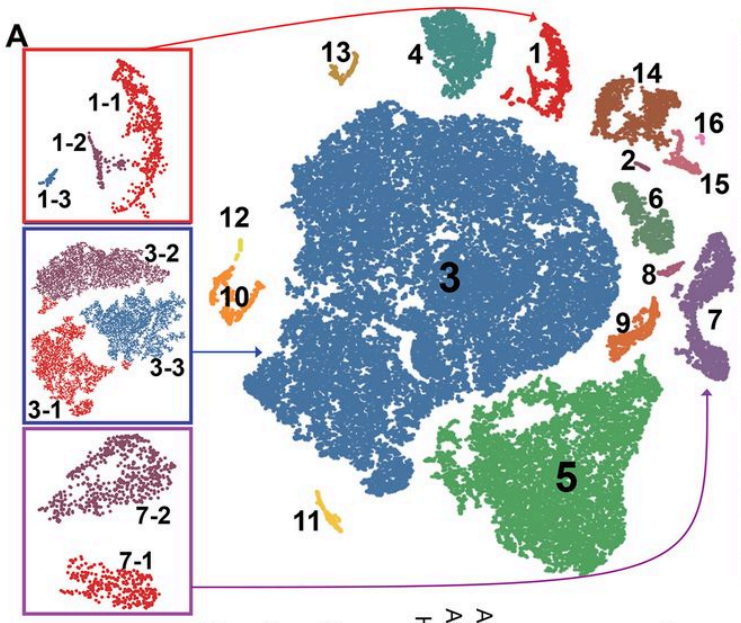
T
- Student's T-test.

Tobit
- Tobit regression on a smoothed data.

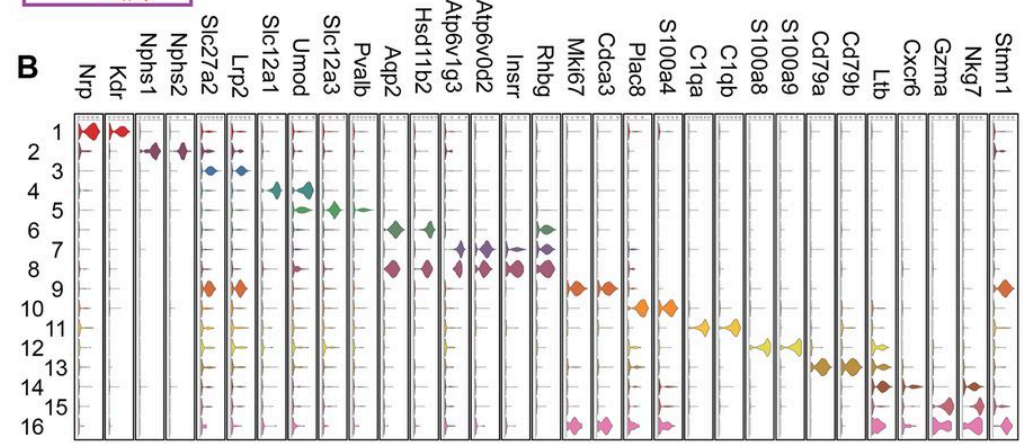
MAST
- Hurdle model for zero inflated data

….

6. Assigning cell identity: Known marker genes

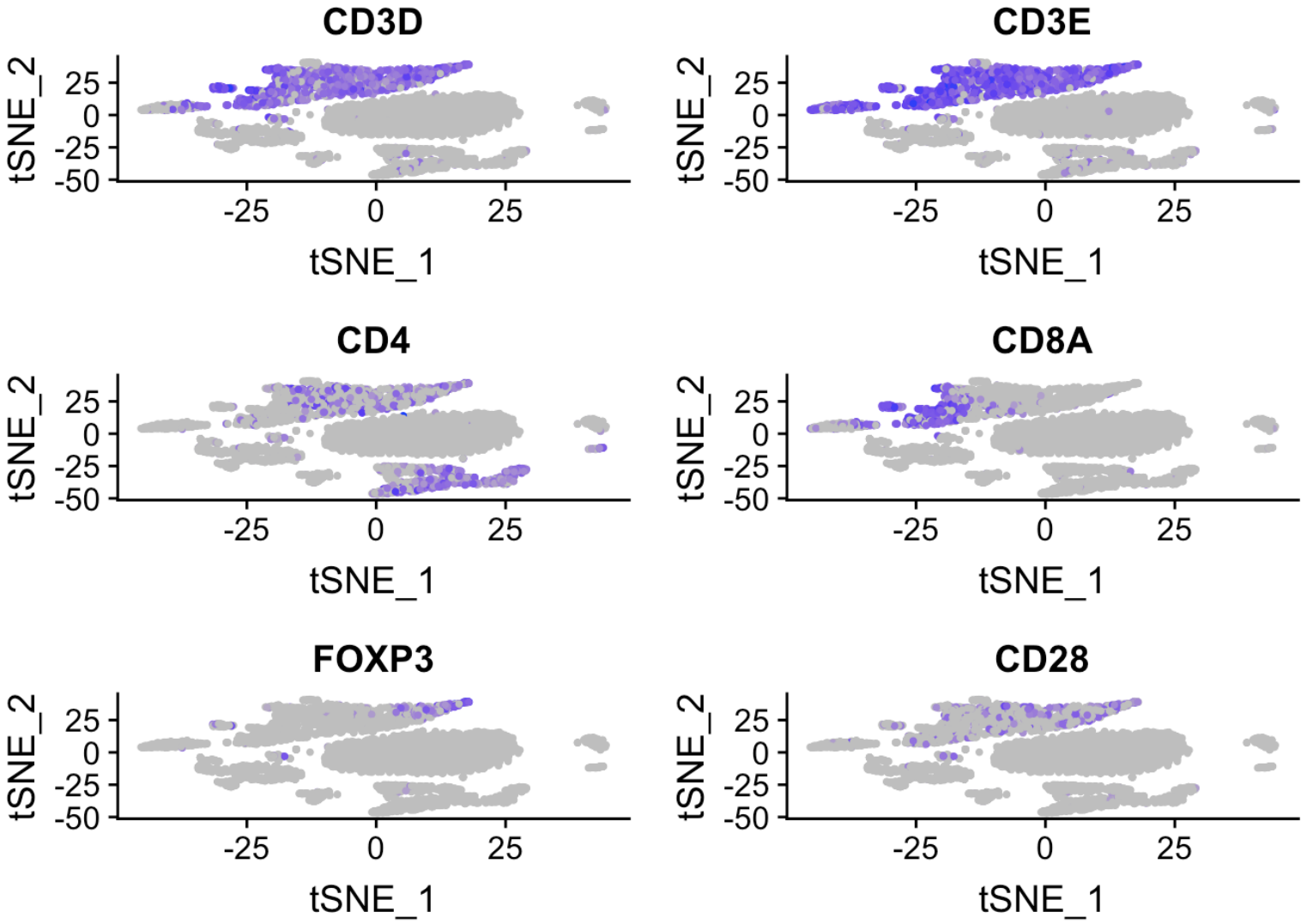


cluster	nCells	%	
1	1,001	2.29	Endo
2	78	0.18	Podo
3	26,482	60.54	PT
4	1,581	3.61	LOH
5	8,544	19.53	DCT
6	870	1.99	CD-PC
7	1,729	3.95	CD-IC
8	110	0.25	CD-Trans
9	601	1.37	Novel1
10	549	1.26	Fib
11	228	0.52	Macro
12	74	0.17	Neutro
13	235	0.54	B lymph
14	1,308	2.99	T lymph
15	313	0.72	NK
16	42	0.1	Novel2

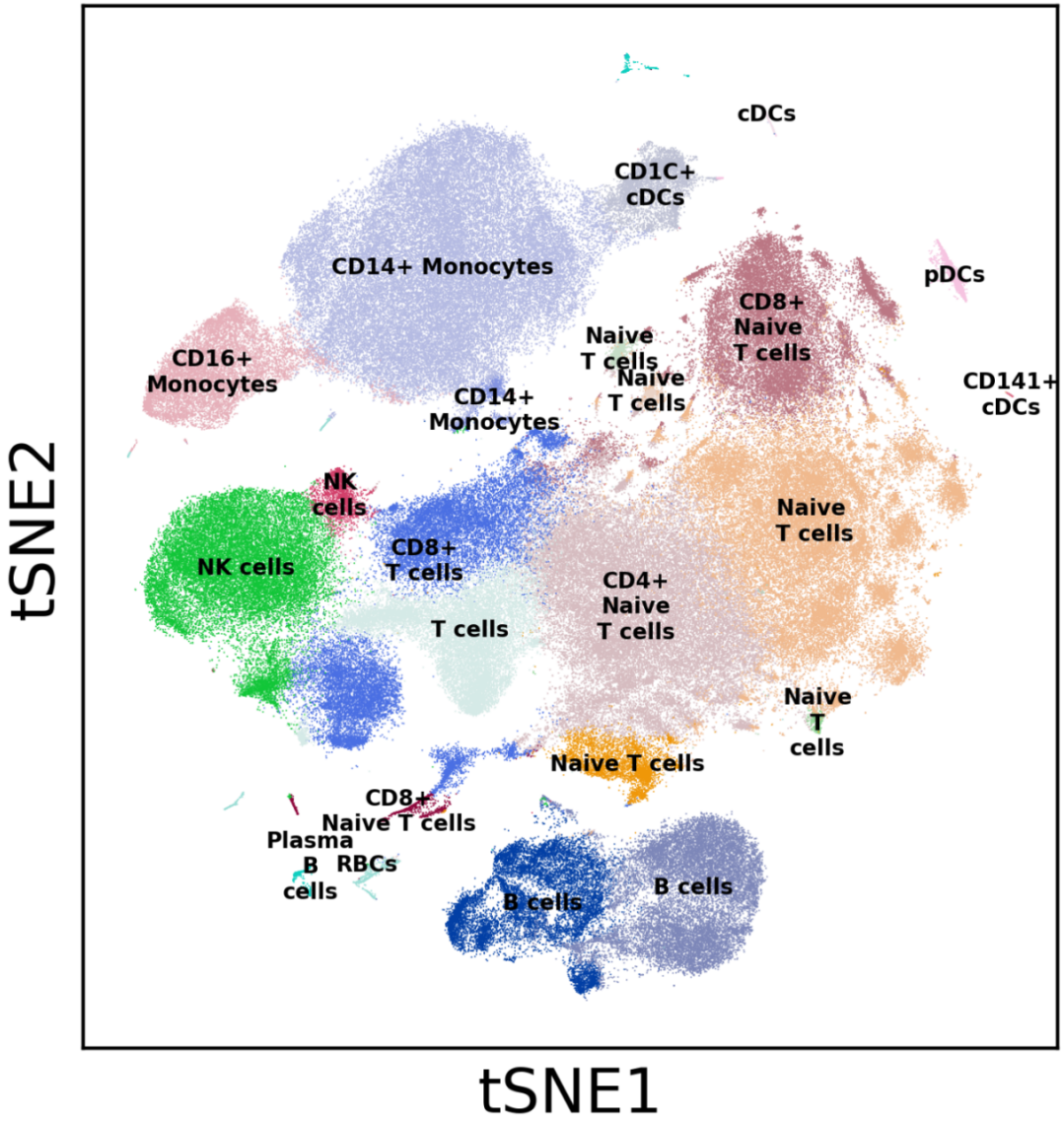


Shekhar et al. Cell 2016
Park and Shreshtha et al. Science 2018

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



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



Visualizing genes of interest

Dot plots, violin plots, feature plots

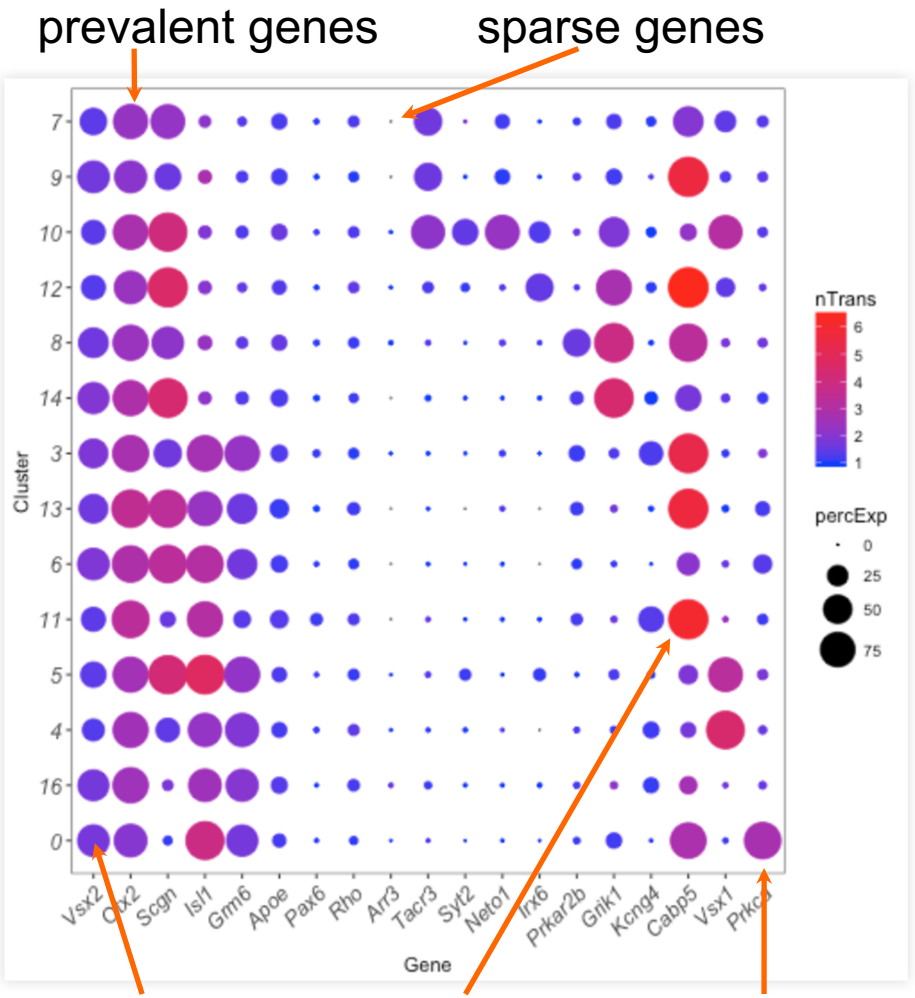
Size of circle

- Gene prevalence in cluster

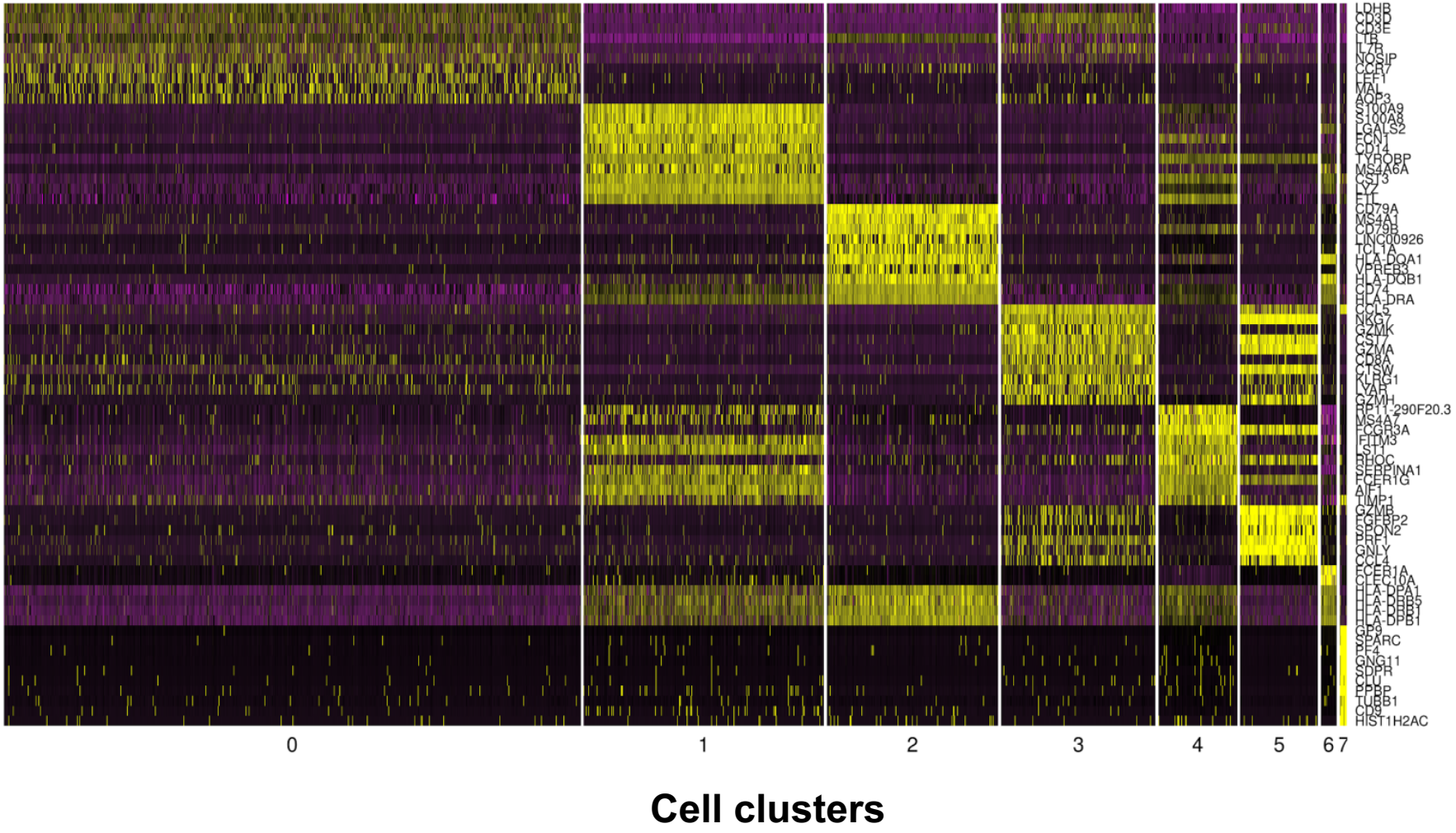
Color of circle

- More red, more expressed in cluster

Scales well with many cells

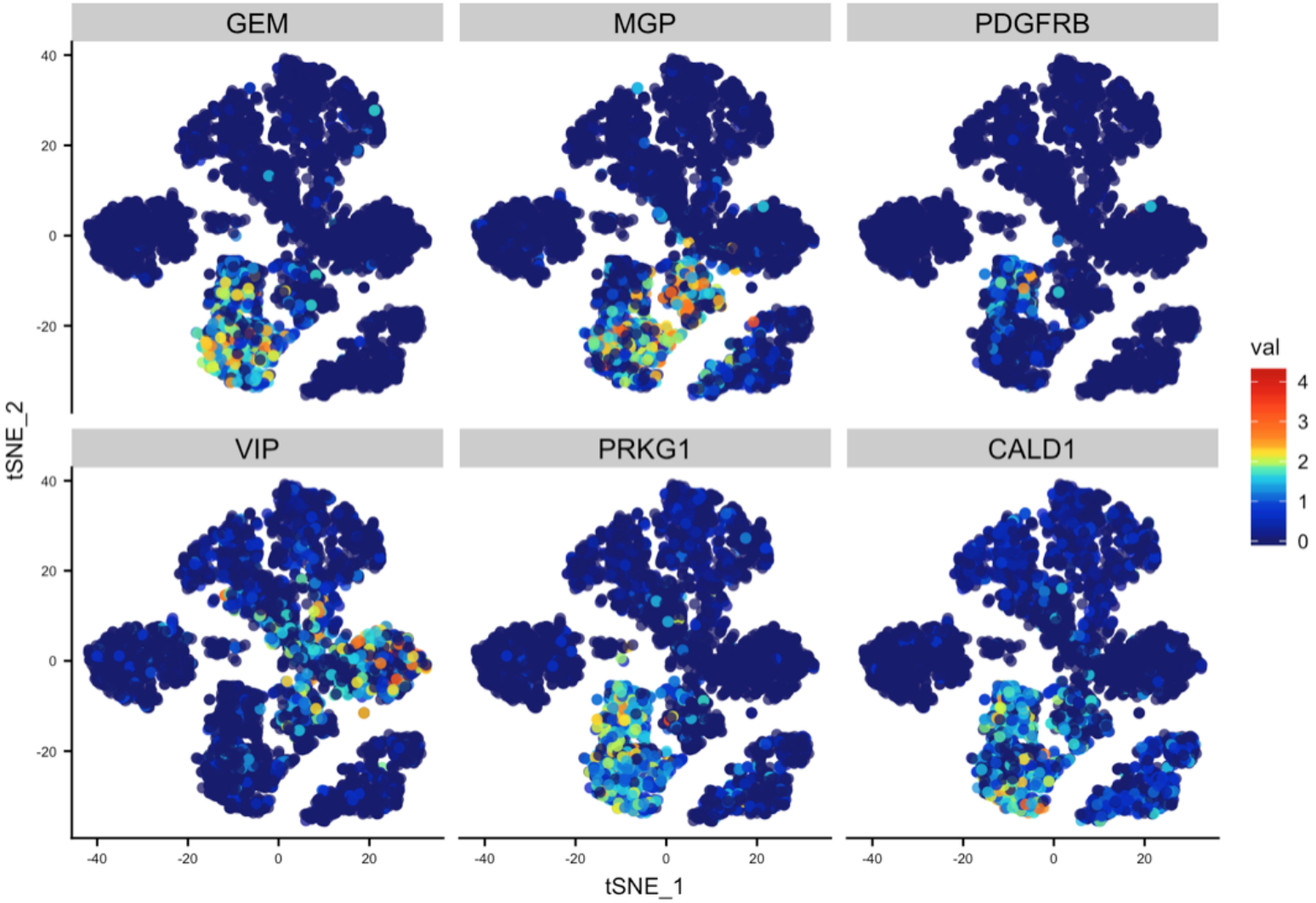


Determining cell type, state, and/or function:
.Identifying differentially expressed genes



Visualizing genes of interest

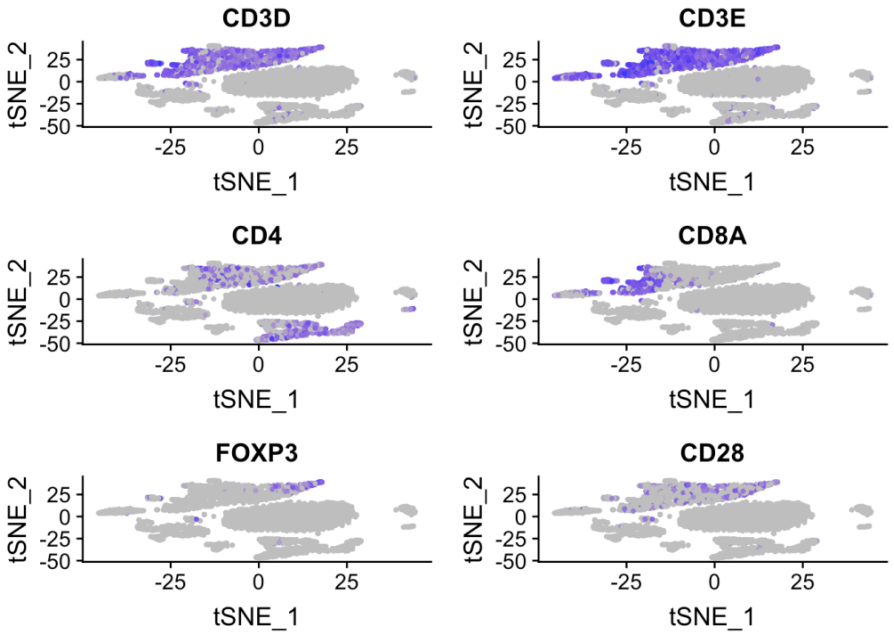
Dot plots, violin plots, feature plots

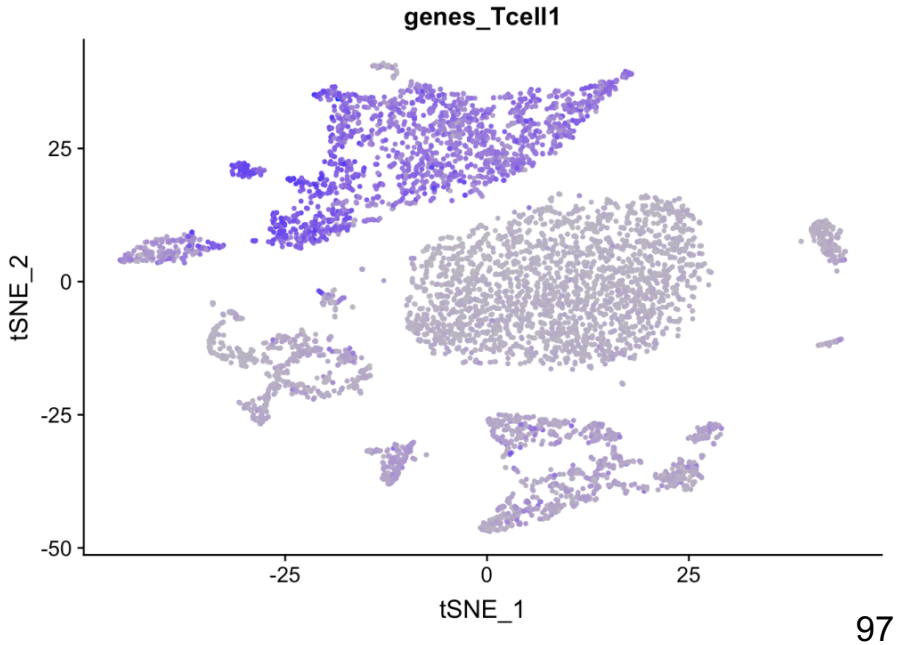


Gene signatures can be used to score each cell based on a set of genes

- 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

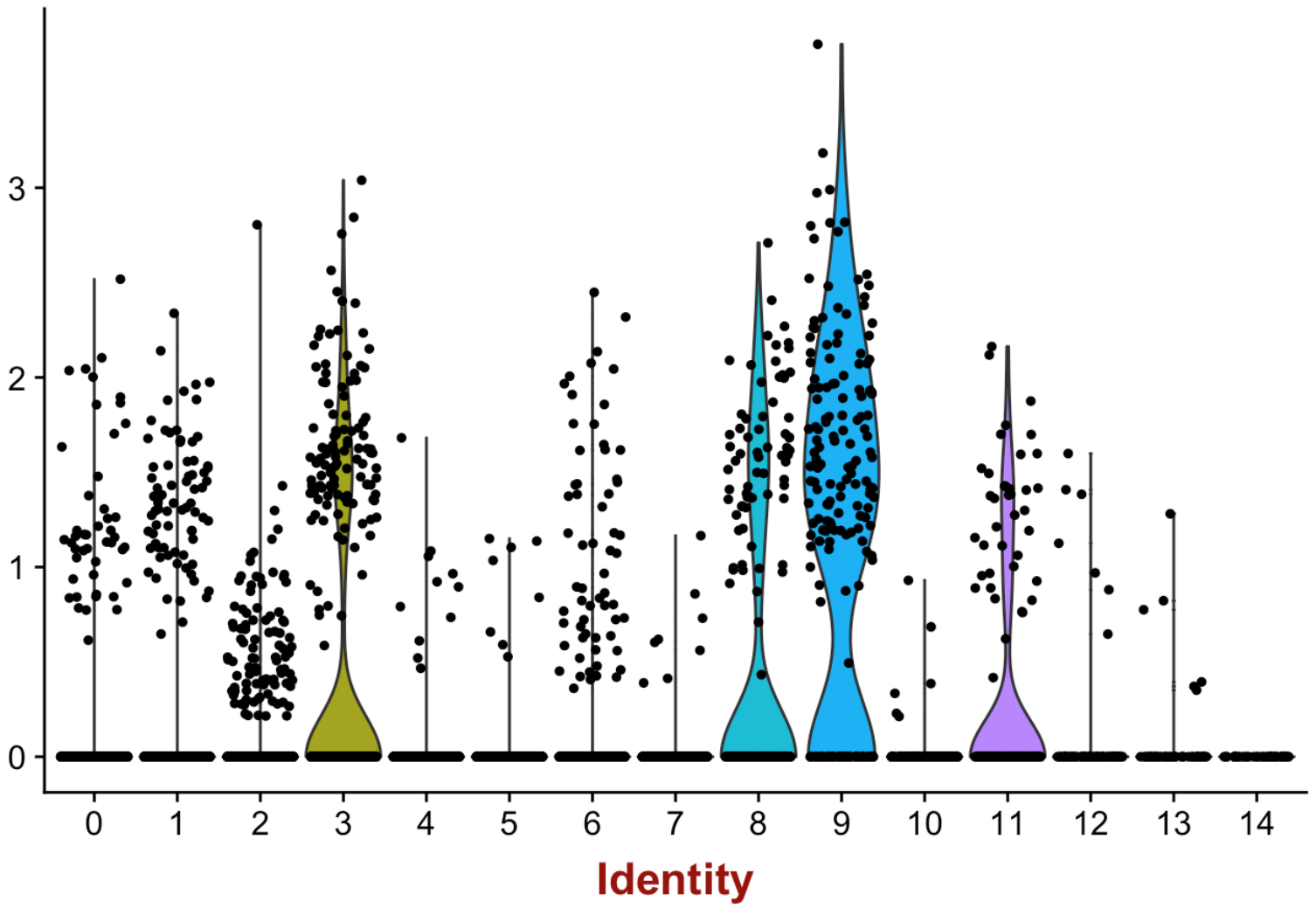
Gene signature for T cells



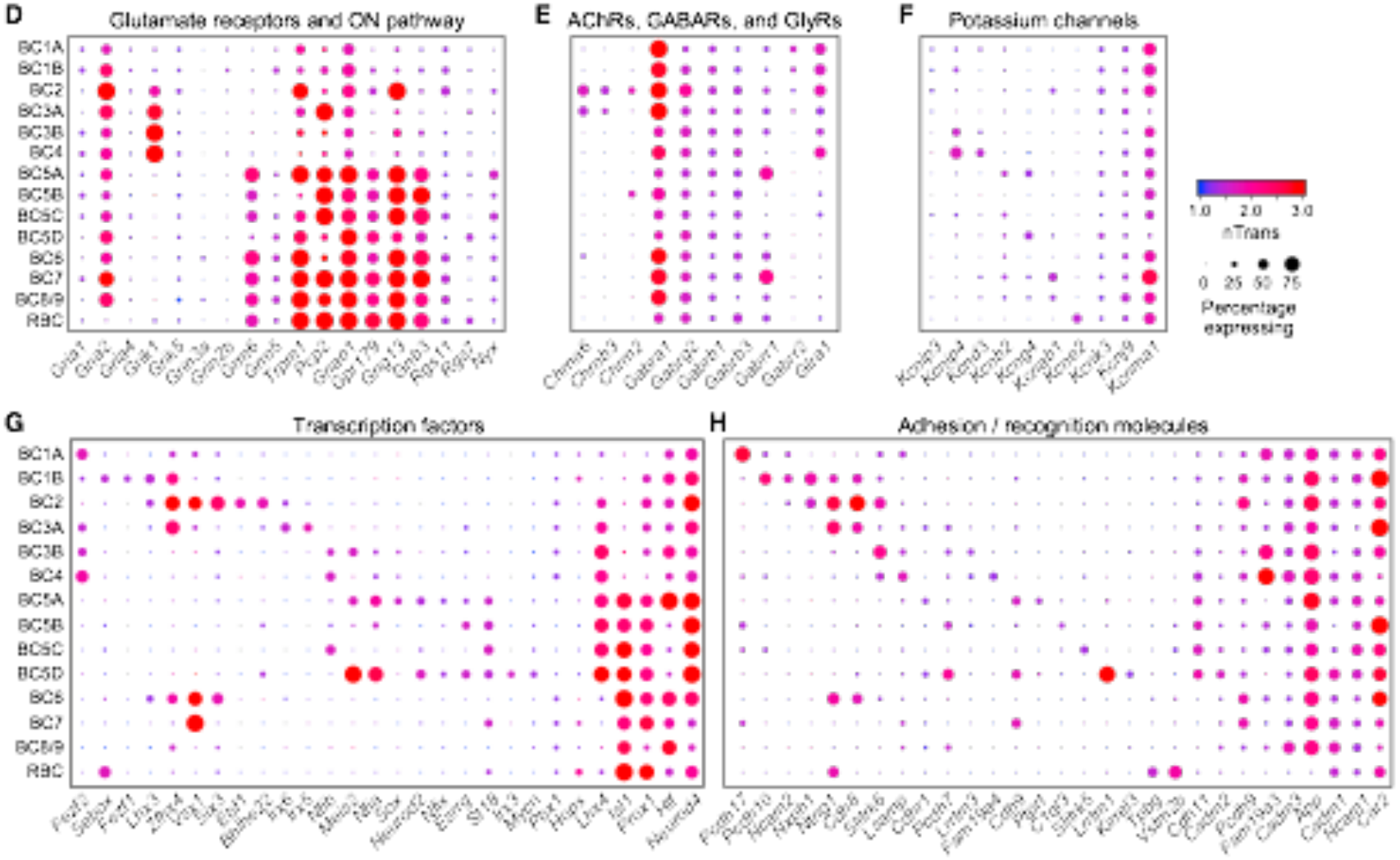


Visualizing genes of interest

Dot plots, violin plots, feature plots

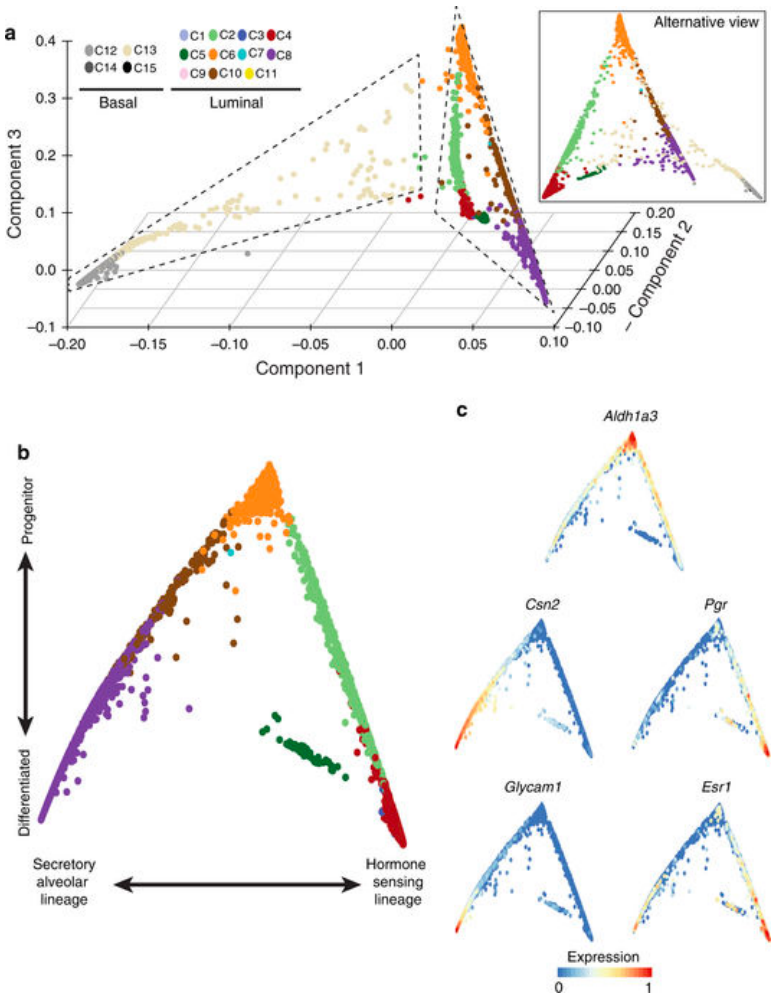


7. Functional annotation by pathway analysis and gene-set enrichment analysis

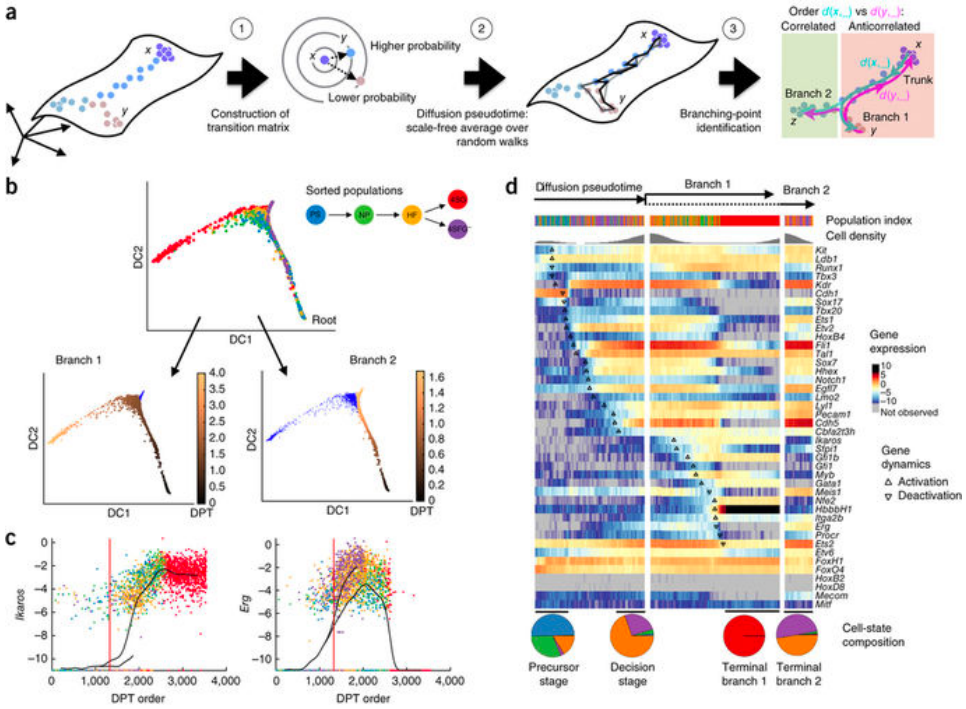


Shekhar et al. Cell 2016

Trajectory inference



Diffusion Maps

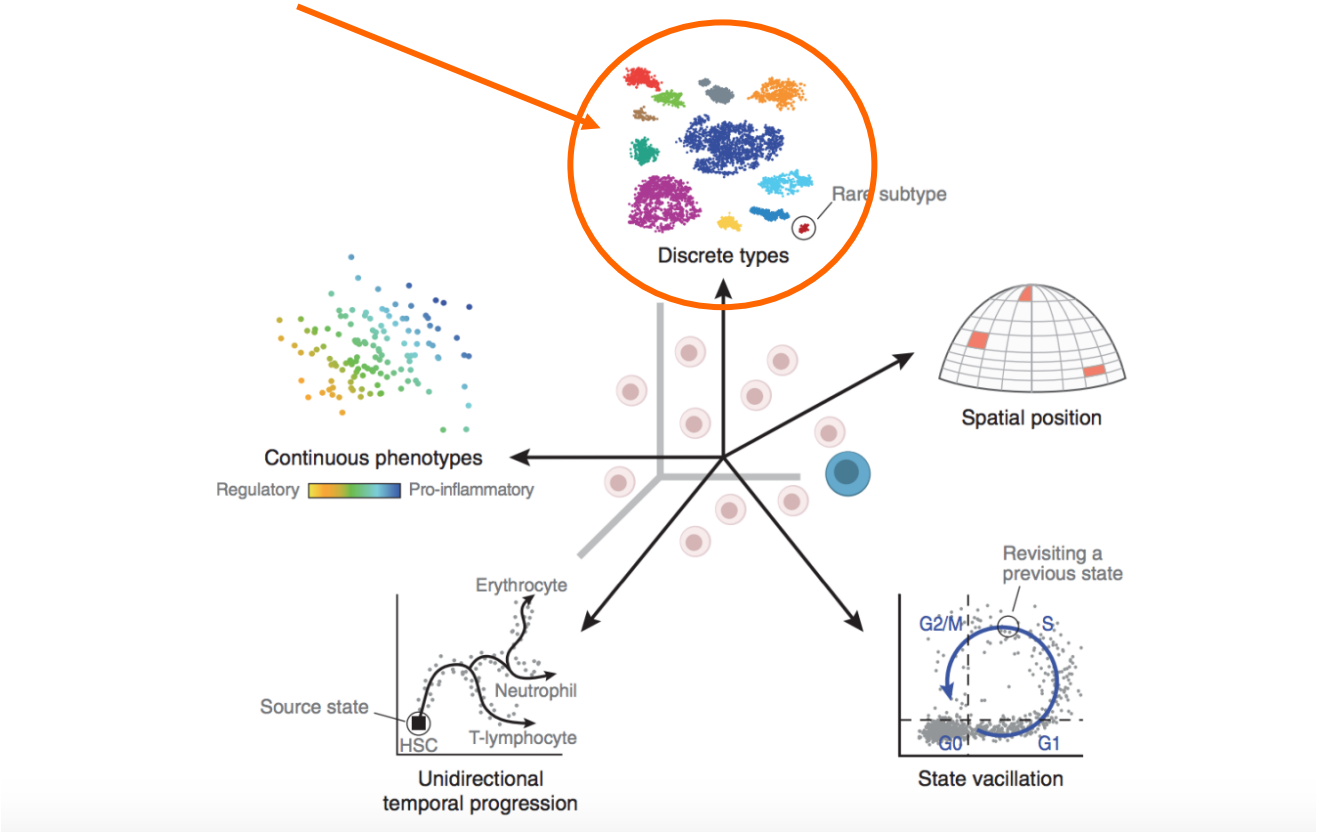


Diffusion pseudotime

Bach et al. Nat Comm 2016
Haghverdi et al. Nat Methods 2016

Recap: what did we just cover?

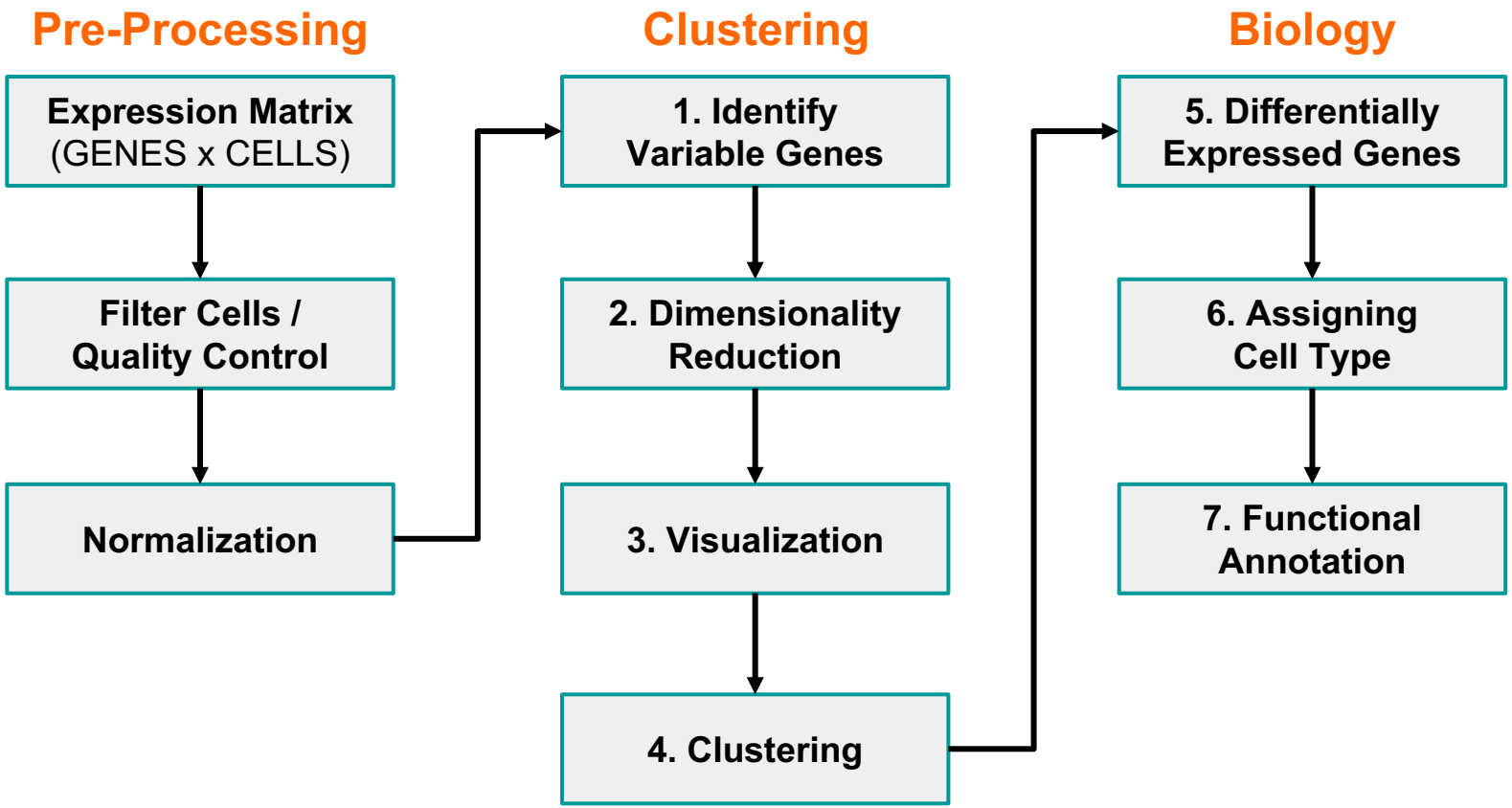
We covered just this.
So much more to learn!



Wagner, Regev and Yosef. Nat Biotech 2016

Recap: what did we just cover?

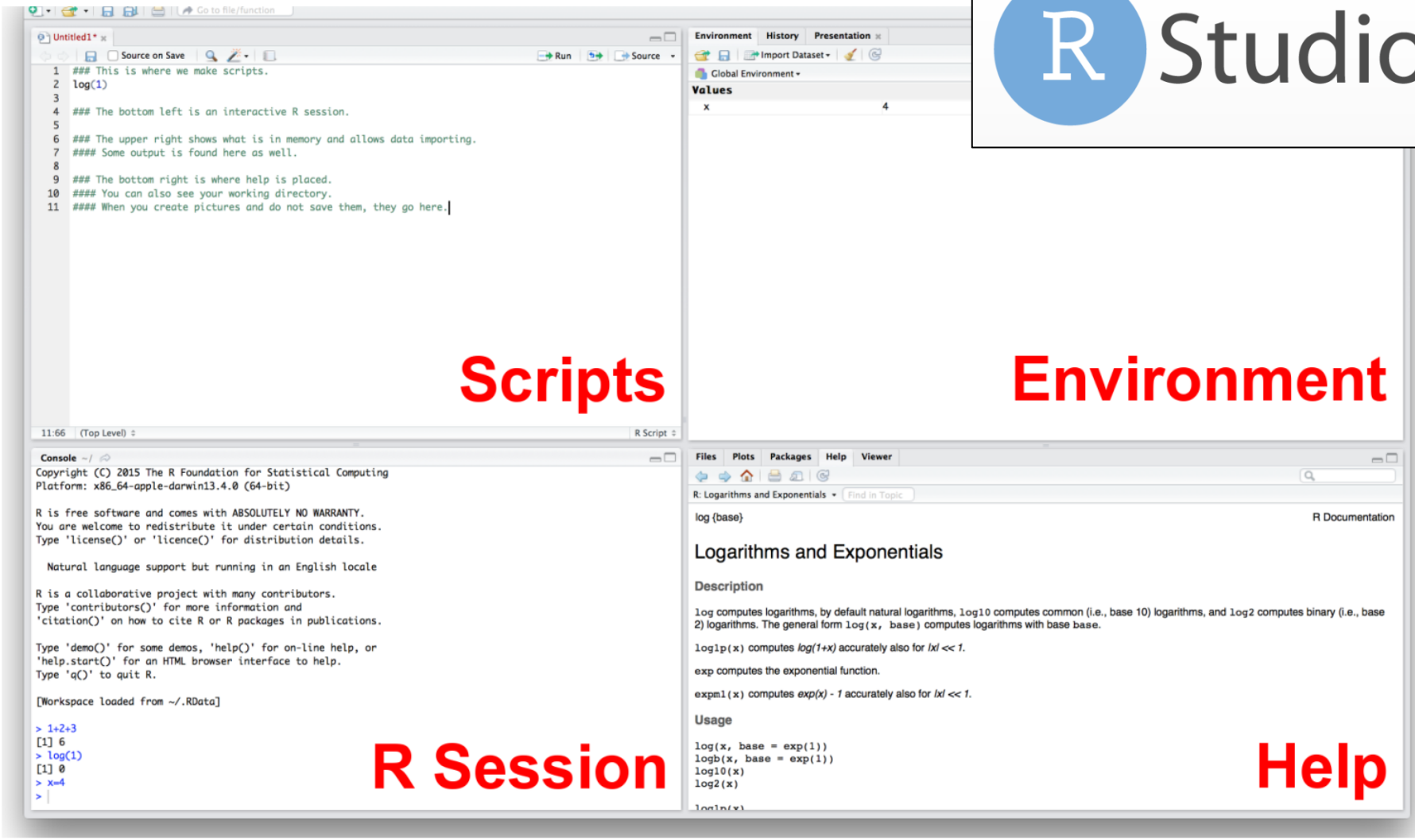
Time to execute this pipeline in a hands on example!



Wagner, Regev and Yosef. Nat Biotech 2016

Tools and resources

RStudio:
integrated development environment for R



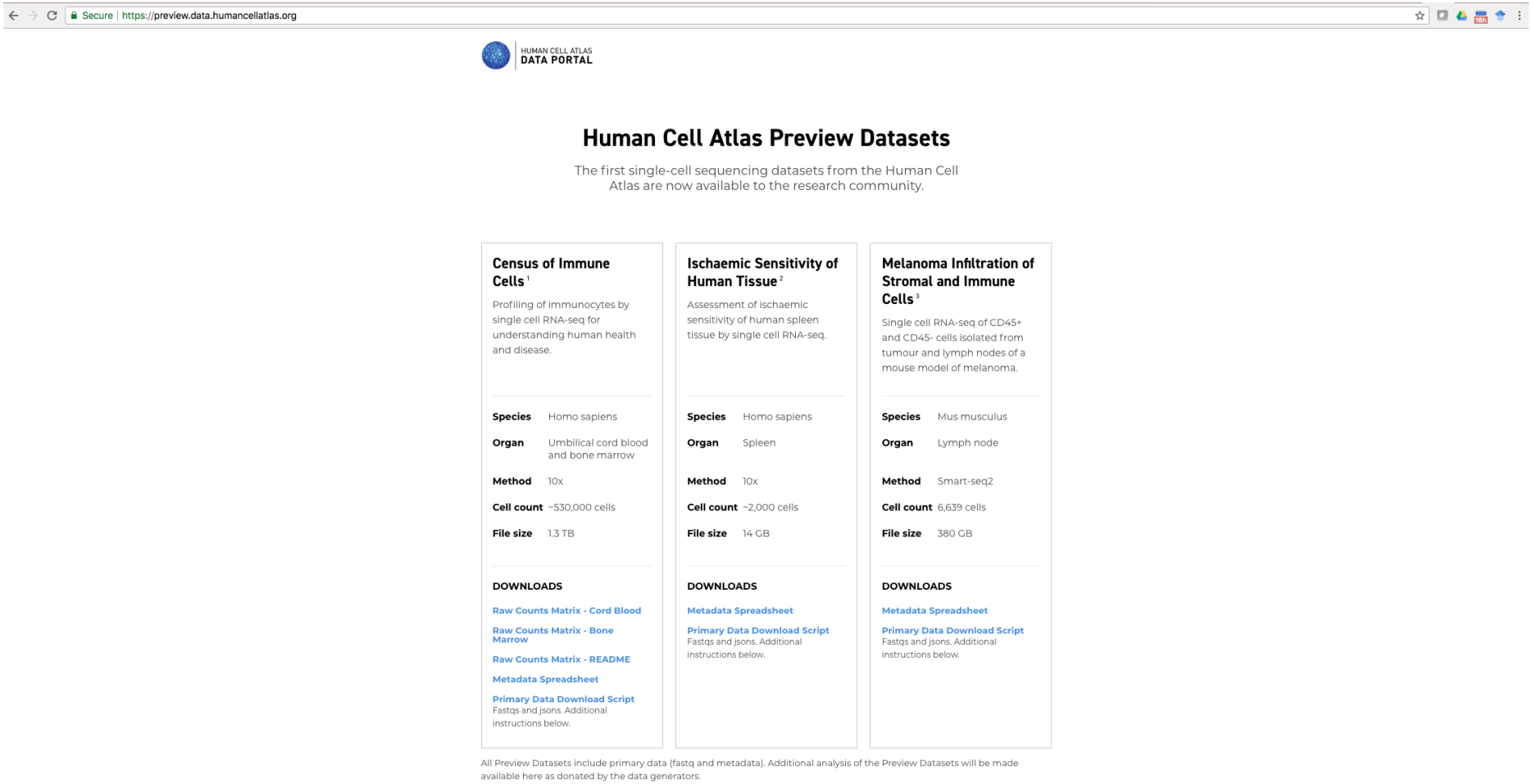
Secure | https://www.humancellatlas.org

HUMAN CELL ATLAS

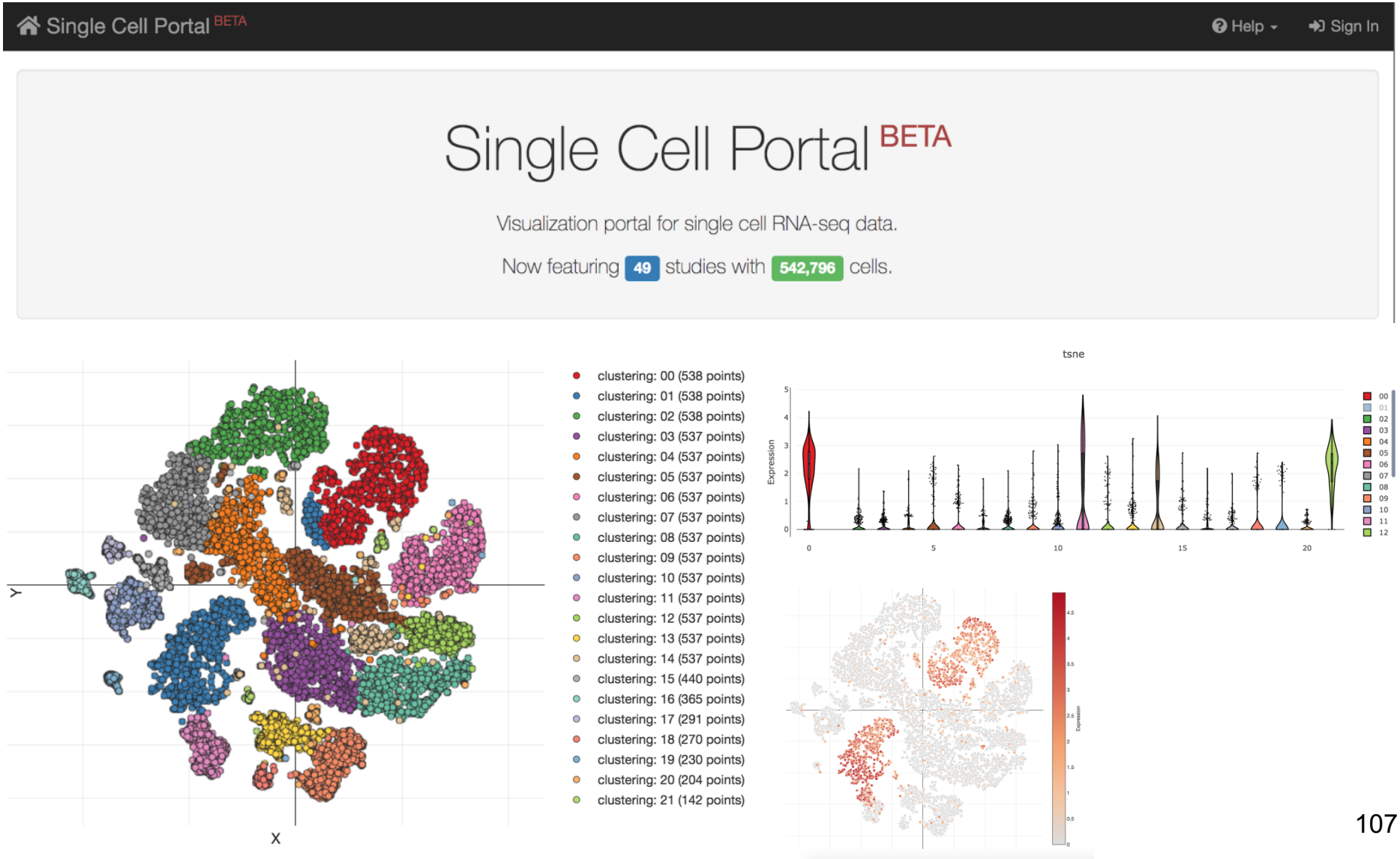
Home HCA Areas of Impact News Publications Data Coordination Join HCA Contact

MISSION

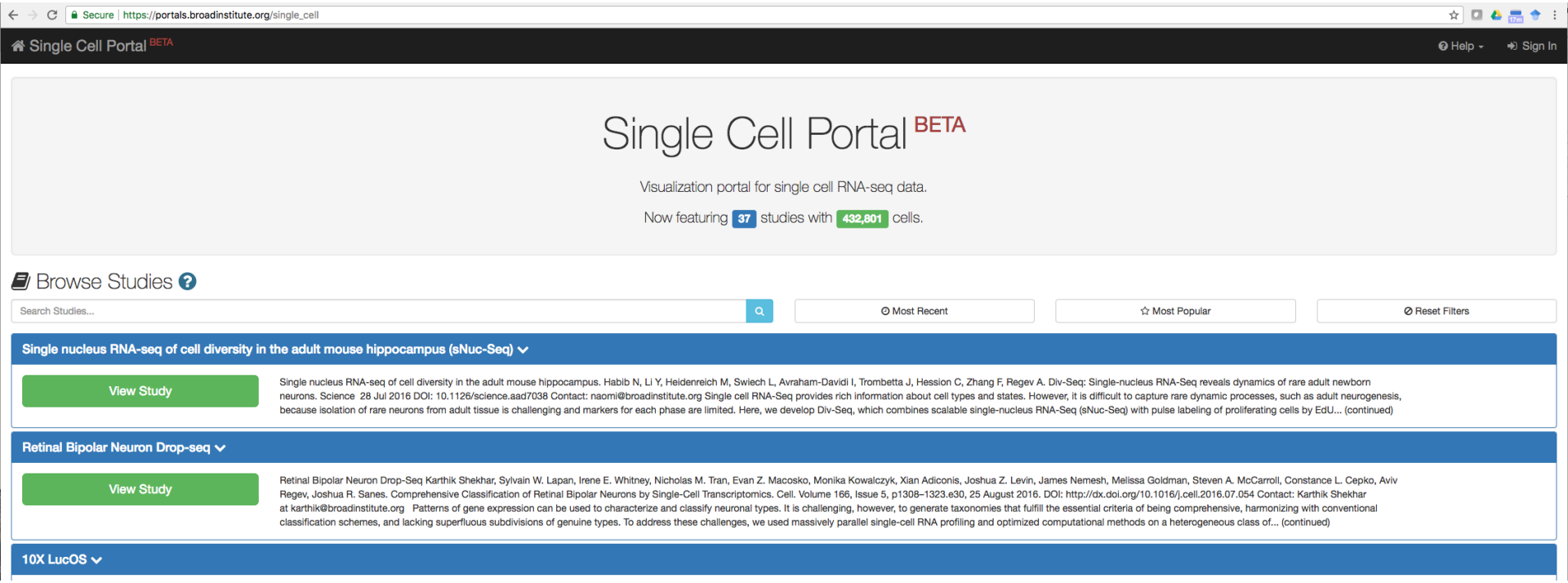
To create comprehensive reference maps of all human cells—the fundamental units of life—as a basis for both understanding human health and diagnosing, monitoring, and treating disease.


Secure | https://preview.data.humancellatlas.org
HUMAN CELL ATLAS
DATA PORTAL
Human Cell Atlas Preview Datasets
The first single-cell sequencing datasets from the Human Cell Atlas are now available to the research community.
Census of Immune Cells [1]
Profiling of immunocytes by single cell RNA-seq for understanding human health and disease.
Species Homo sapiens
Organ Umbilical cord blood and bone marrow
Method 10x
Cell count ~530,000 cells
File size 1.3 TB
DOWNLOADS
Raw Counts Matrix - Cord Blood
Raw Counts Matrix - Bone Marrow
Raw Counts Matrix - README
Metadata Spreadsheet
Primary Data Download Script
Fastqs and jsons. Additional instructions below.
Ischaemic Sensitivity of Human Tissue [2]
Assessment of ischaemic sensitivity of human spleen tissue by single cell RNA-seq.
Species Homo sapiens
Organ Spleen
Method 10x
Cell count ~2,000 cells
File size 14 GB
DOWNLOADS
Metadata Spreadsheet
Primary Data Download Script
Fastqs and jsons. Additional instructions below.
Melanoma Infiltration of Stromal and Immune Cells [3]
Single cell RNA-seq of CD45+ and CD45- cells isolated from tumour and lymph nodes of a mouse model of melanoma.
Species Mus musculus
Organ Lymph node
Method Smart-seq2
Cell count 6,639 cells
File size 380 GB
DOWNLOADS
Metadata Spreadsheet
Primary Data Download Script
Fastqs and jsons. Additional instructions below.
All Preview Datasets include primary data (fastq and metadata). Additional analysis of the Preview Datasets will be made available here as donated by the data generators.

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



Broad Institute Single Cell Portal



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

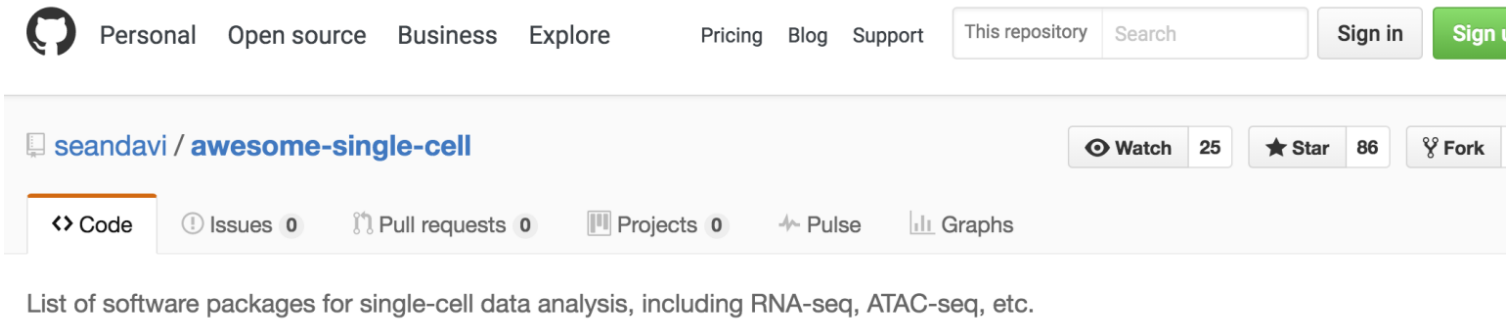
- http://bioconductor.org/packages/devel/workflows/html/simpleSingleCell.html
- https://satijalab.org/seurat/
- https://scanpy.readthedocs.io/

Online courses

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

Resources, cont.

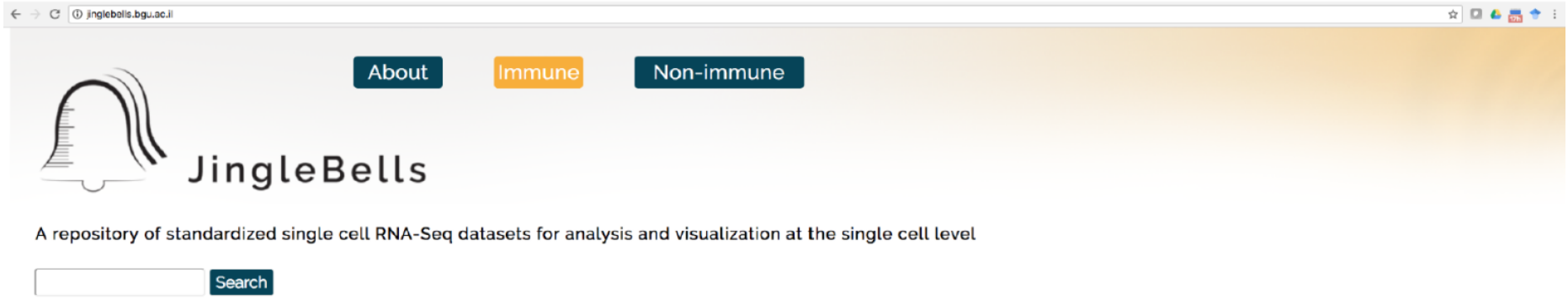
Comprehensive list of single-cell resources:
https://github.com/seandavi/awesome-single-cell



www.singlecellnetwork.org

Resources, cont.

Data repositories: JingleBells



Data repositories: Conquer