

Introduction to Single Cell Transcriptomic Analysis

Acknowledgments

Brian Haas

Karthik Shekhar

Timothy Tickle

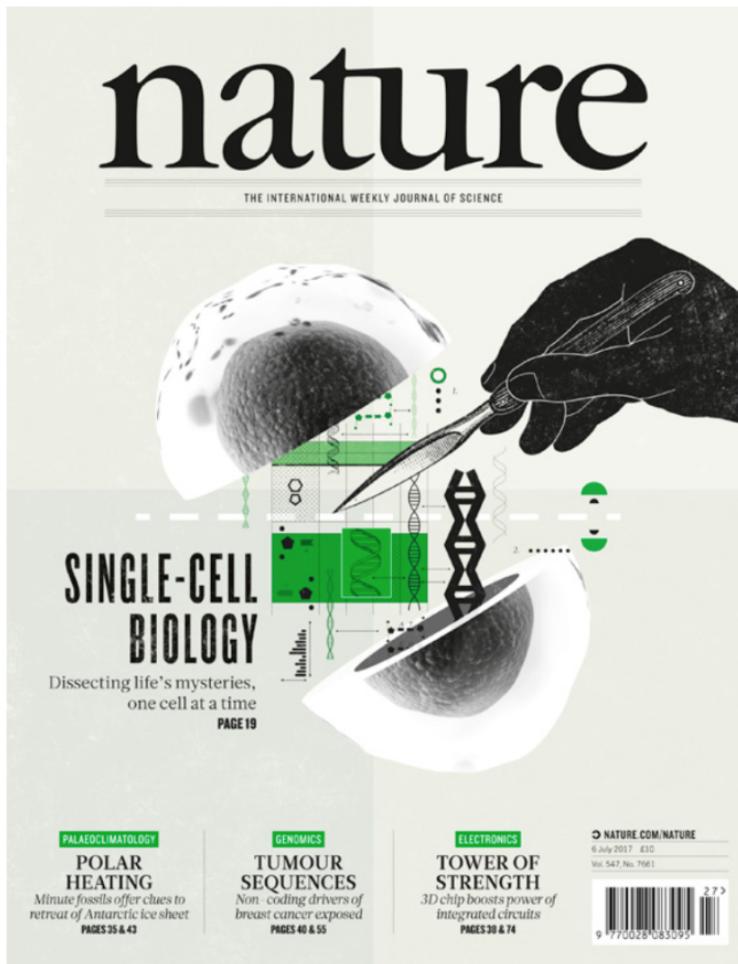
Caroline Porter

Ayshwarya Subramanian | subraman@broadinstitute.org

Goals for today

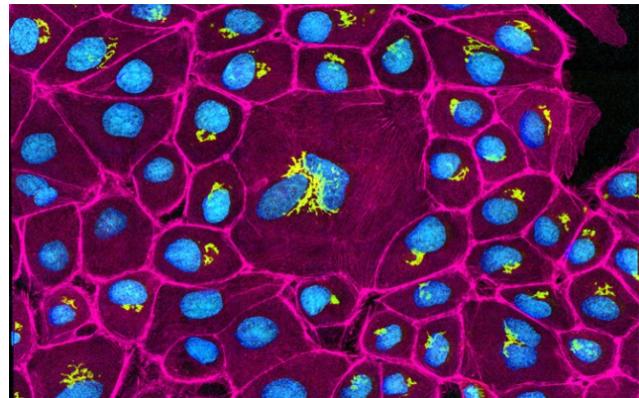
- Overview of single-cell RNA-seq data analysis
- What is in a count matrix?
- Preprocessing data: quality control (QC) and filtering
- Dimensionality reduction & Clustering
- Biology: inferring cell types and further

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

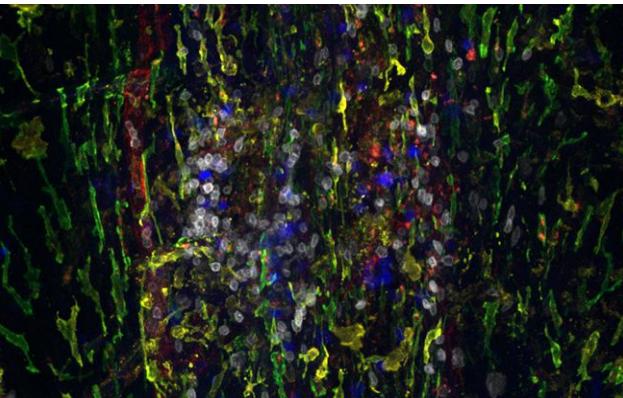


Incredible diversity in cell types, states, & interactions across human tissues

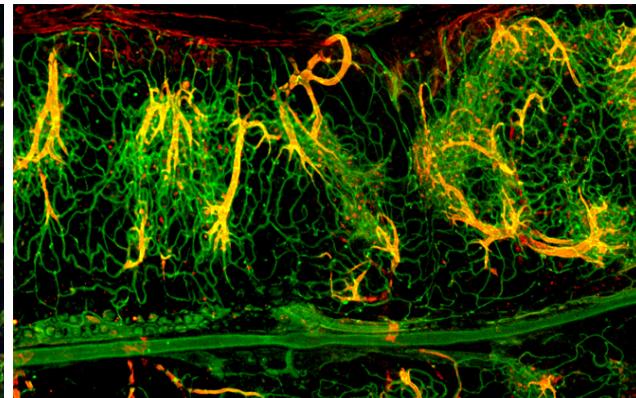
Skin epithelium



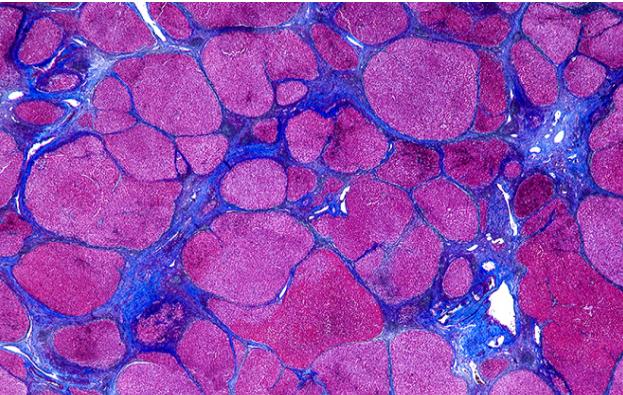
Brain meninges



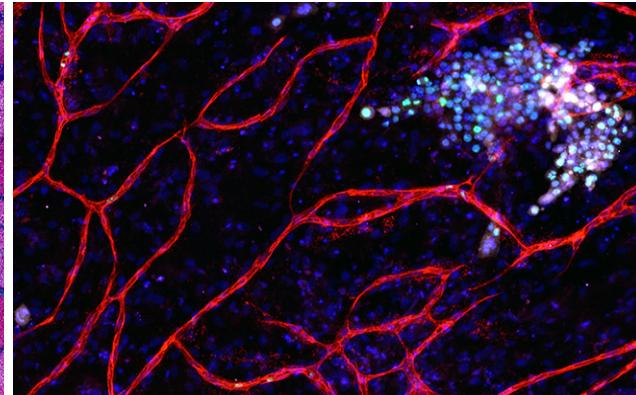
Blood vessels



Small intestine



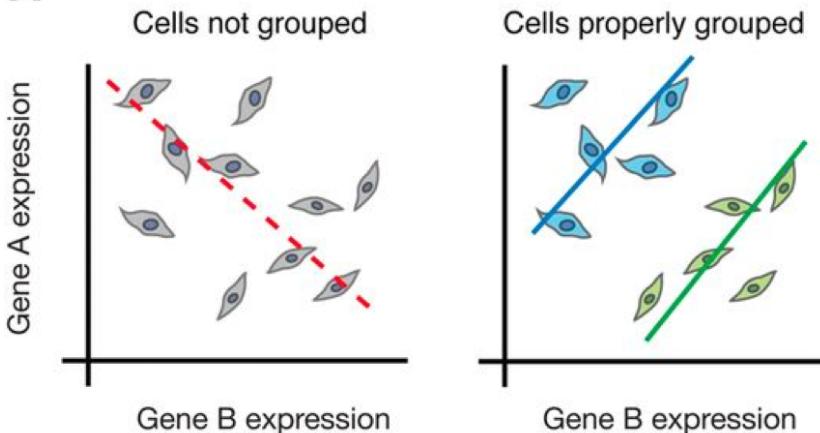
Liver cirrhosis



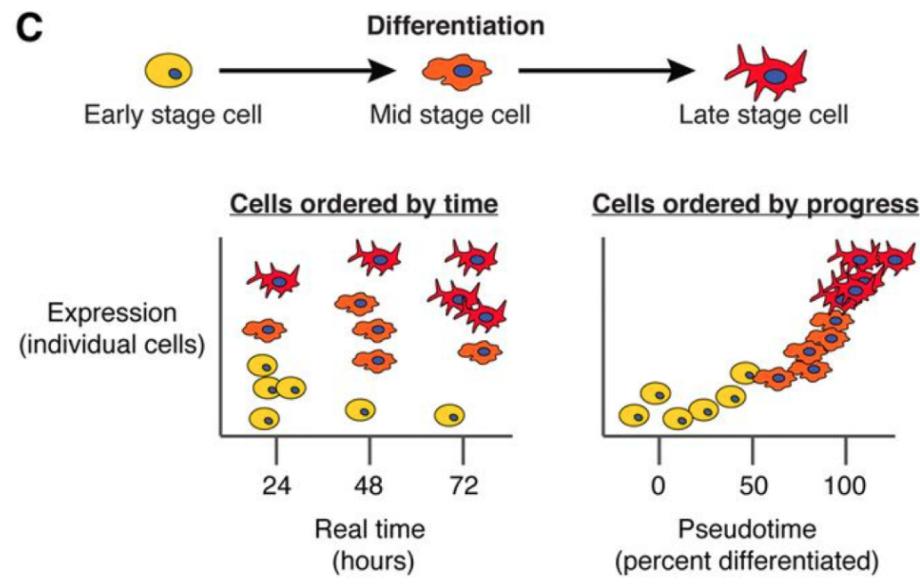
Breast cancer

Bulk sequencing profiles measure average profiles from tissue samples

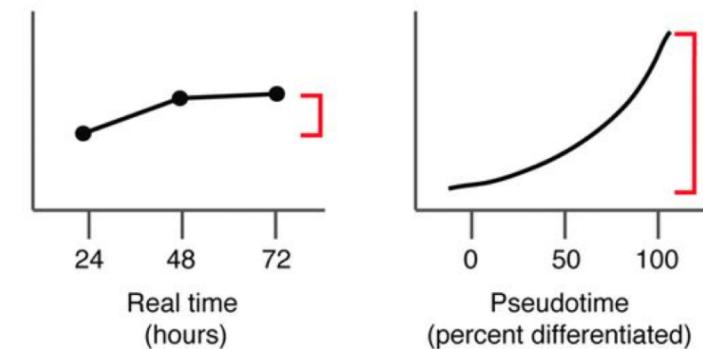
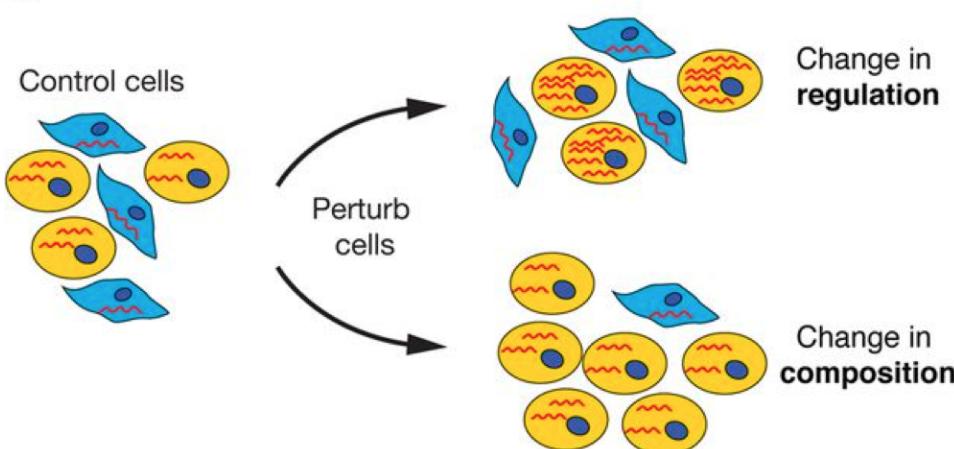
A



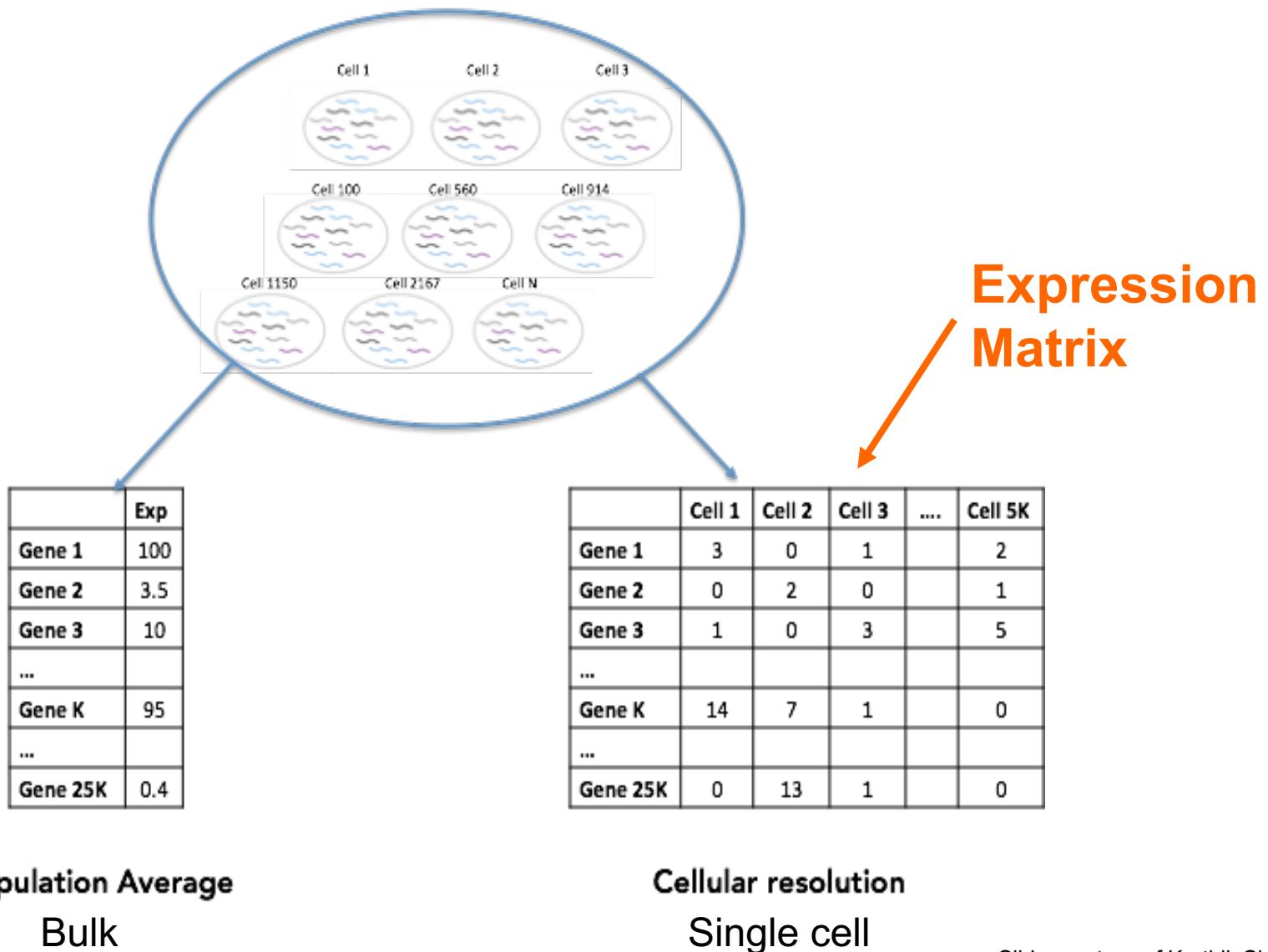
C



B



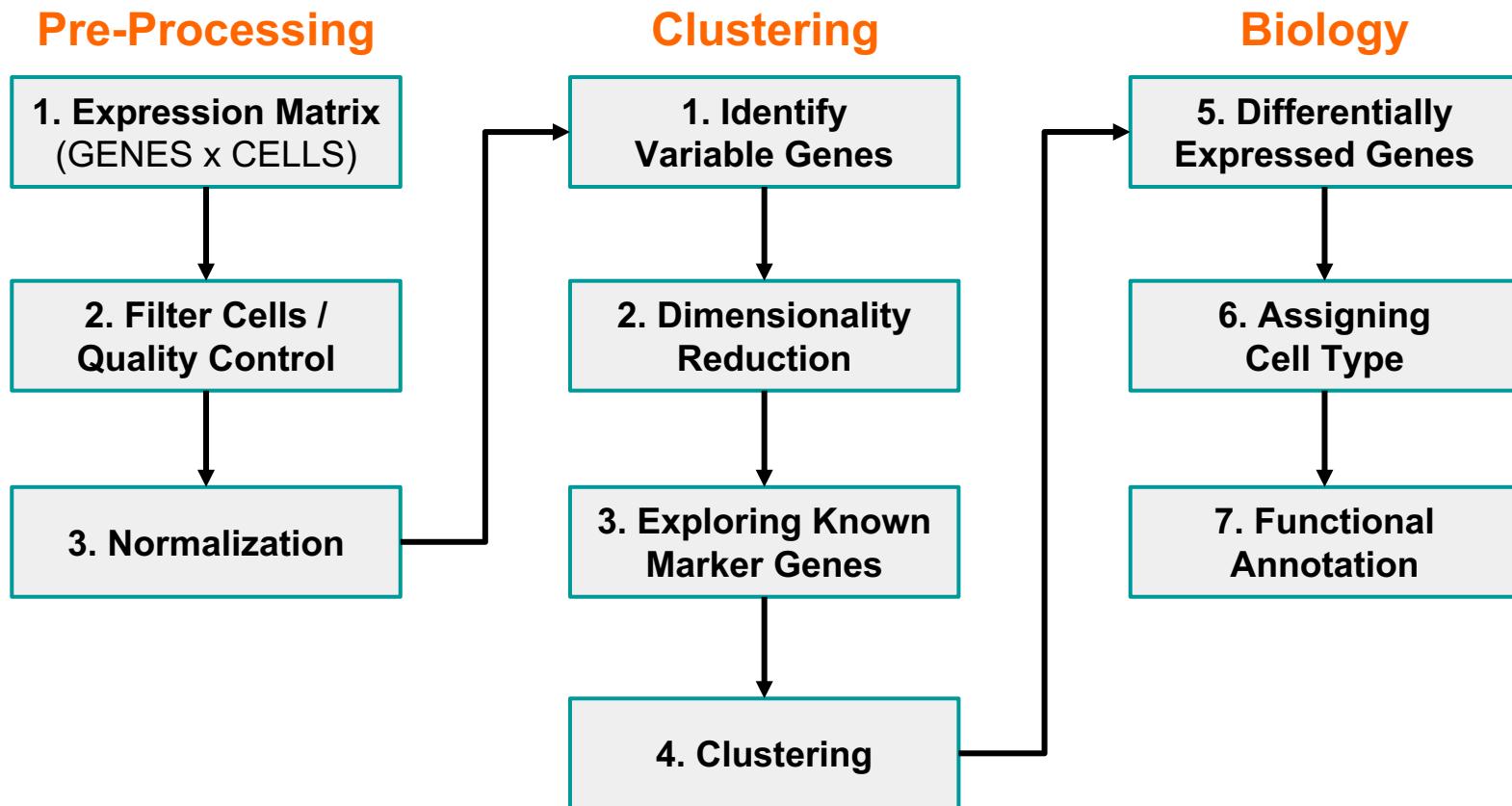
Single-cell and bulk gene expression distributions are very different



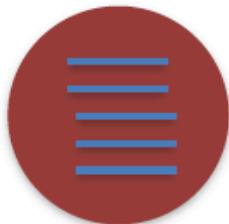
Population Average
Bulk

Cellular resolution
Single cell

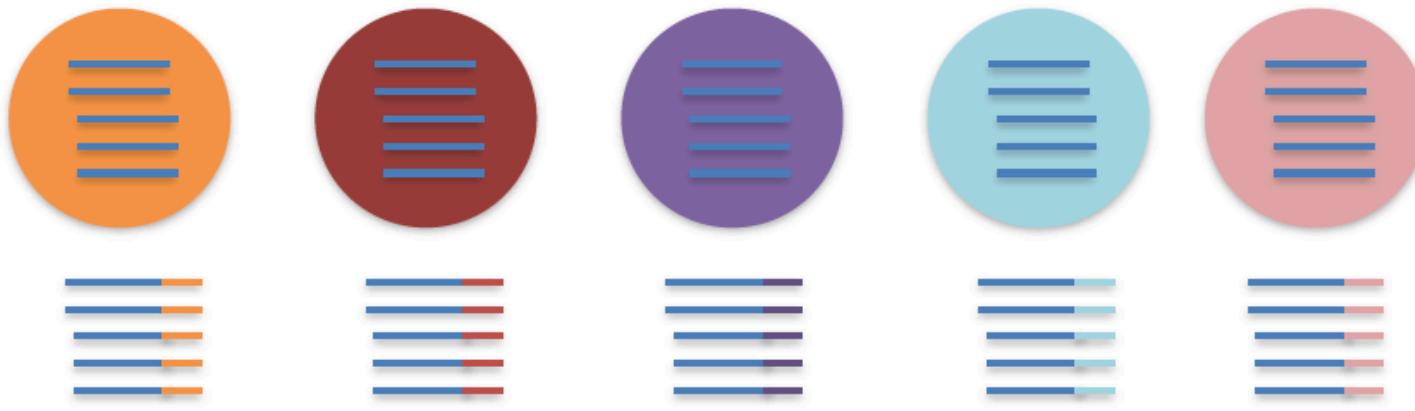
Single-cell RNA-seq analysis pipeline: Analyzing the expression data



Tracking the cell-of-origin of individual transcripts



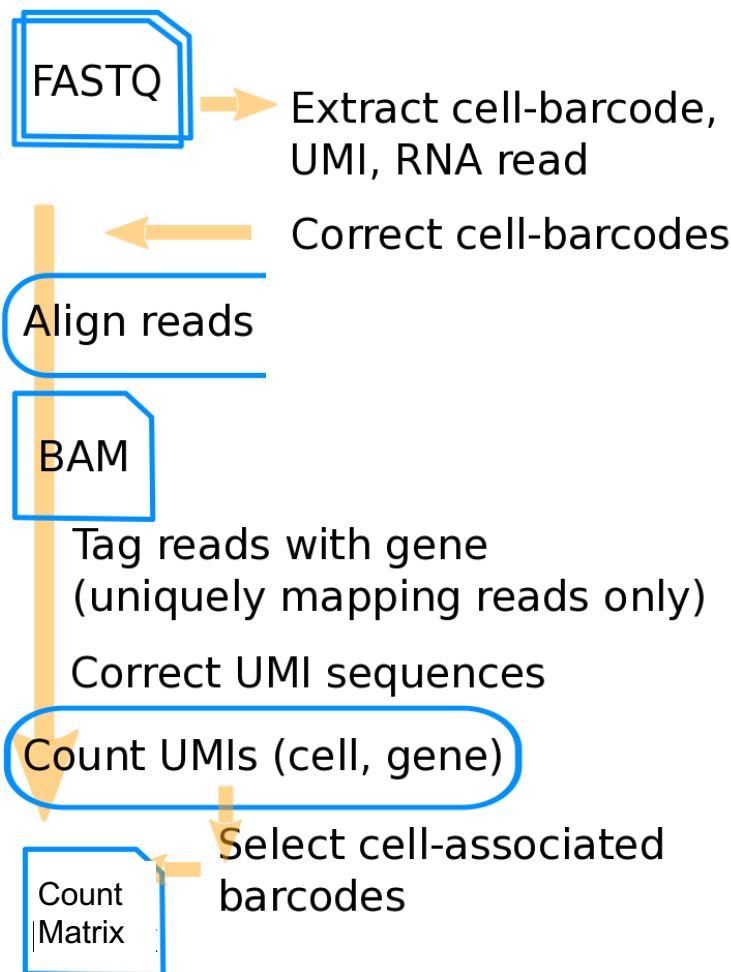
Tracking the cell-of-origin of individual transcripts



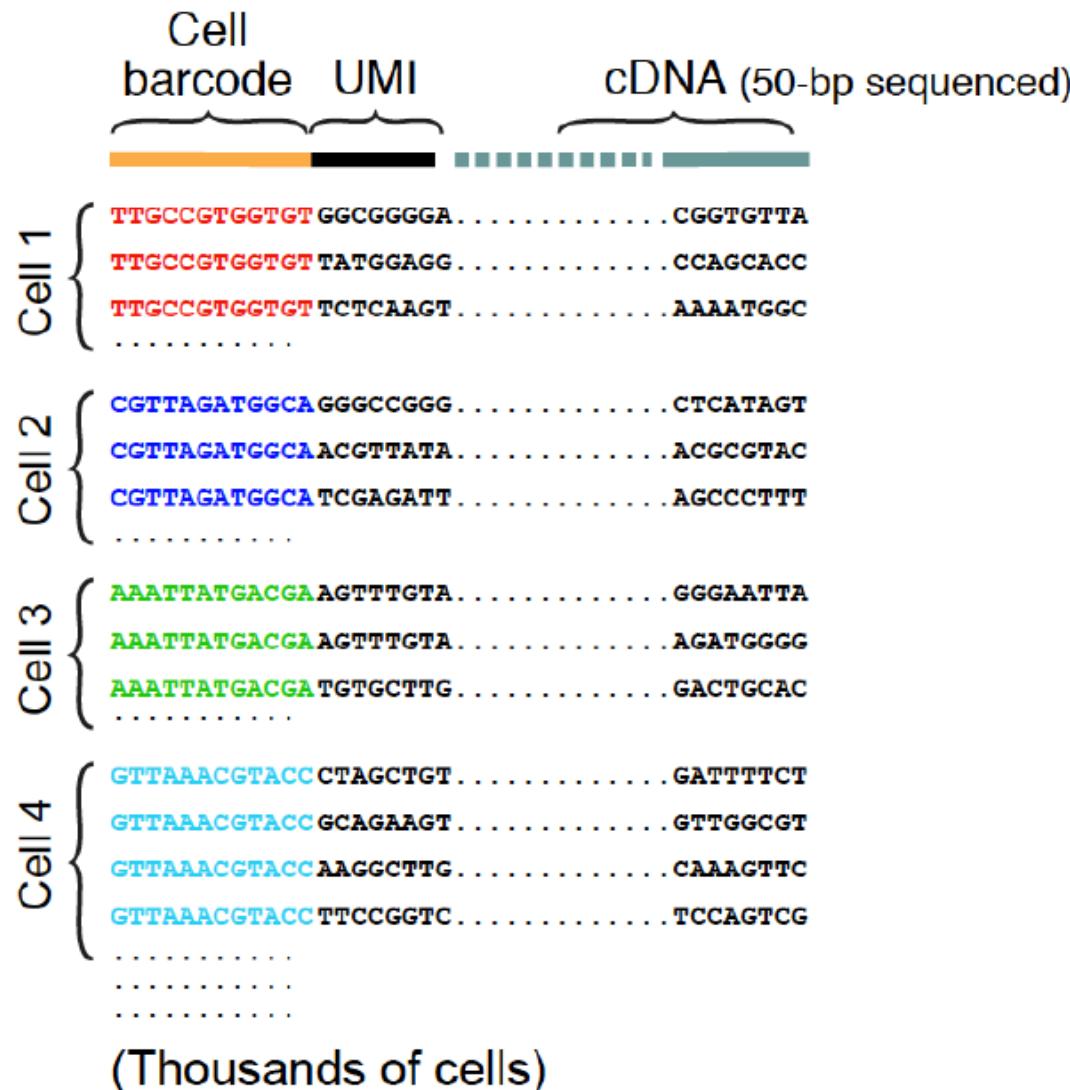
Single-cell RNA-seq pipeline



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

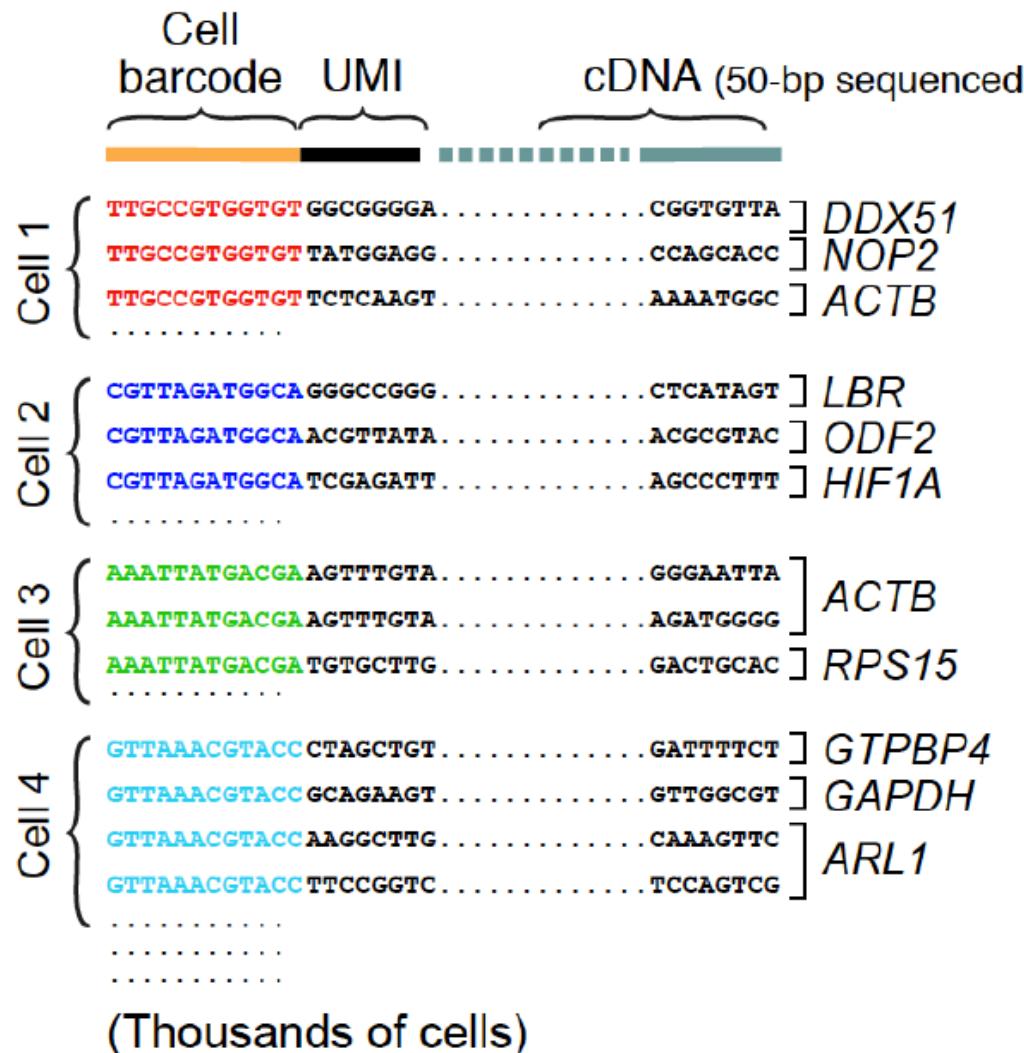


Group reads by cell barcode



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

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

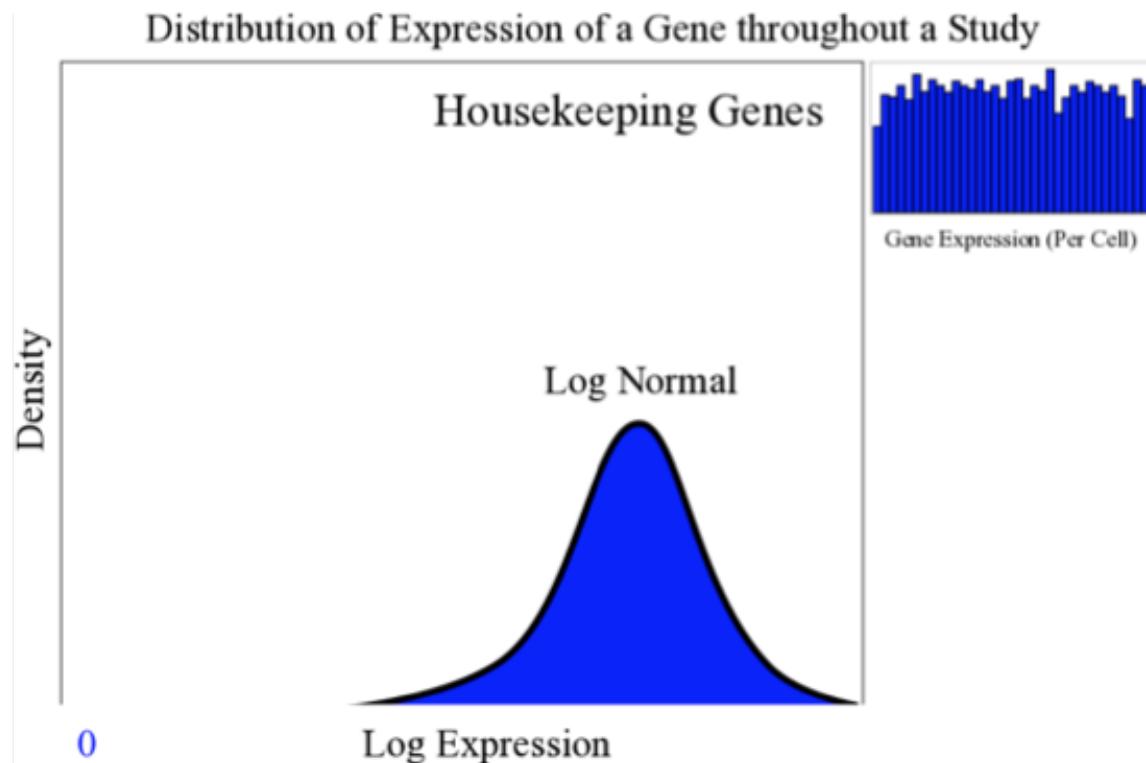
Collapse UMIs to count transcripts

	Cell barcode	UMI	cDNA (50-bp sequenced)	
Cell 1	{	TTGCCGTGGTGT	GGCGGGGA.....CGGTGTTA] <i>DDX51</i>
		TATGGAGG.....	CCAGCACC] <i>NOP2</i>
		TCTCAAGT.....	AAAATGGC] <i>ACTB</i>
Cell 2	{	CGTTAGATGGCA	GGGCCGGG.....CTCATAGT] <i>LBR</i>
		ACGTTATA.....	ACCGGTAC] <i>ODF2</i>
		TCGAGATT.....	AGCCCTTT] <i>HIF1A</i>
Cell 3	{	AAATTATGACGA	AGTTTGTA.....GGGAATTAA] <i>ACTB</i> → 2 reads, 1 molecule
		AGTTTGTA.....	AGATGGGG	
		TGTGCTTG.....	GACTGCAC] <i>RPS15</i>
Cell 4	{	GTTAACGTACC	CTAGCTGT.....GATTTTCT] <i>GTPBP4</i>
		GCAGAAAGT.....	GTTGGCGT] <i>GAPDH</i>
		AAGGCTTG.....	CAAAGTTC] <i>ARL1</i> → 2 reads, 2 molecules
		TTCCGGTC.....	TCCAGTCG	
		
		(Thousands of cells)		

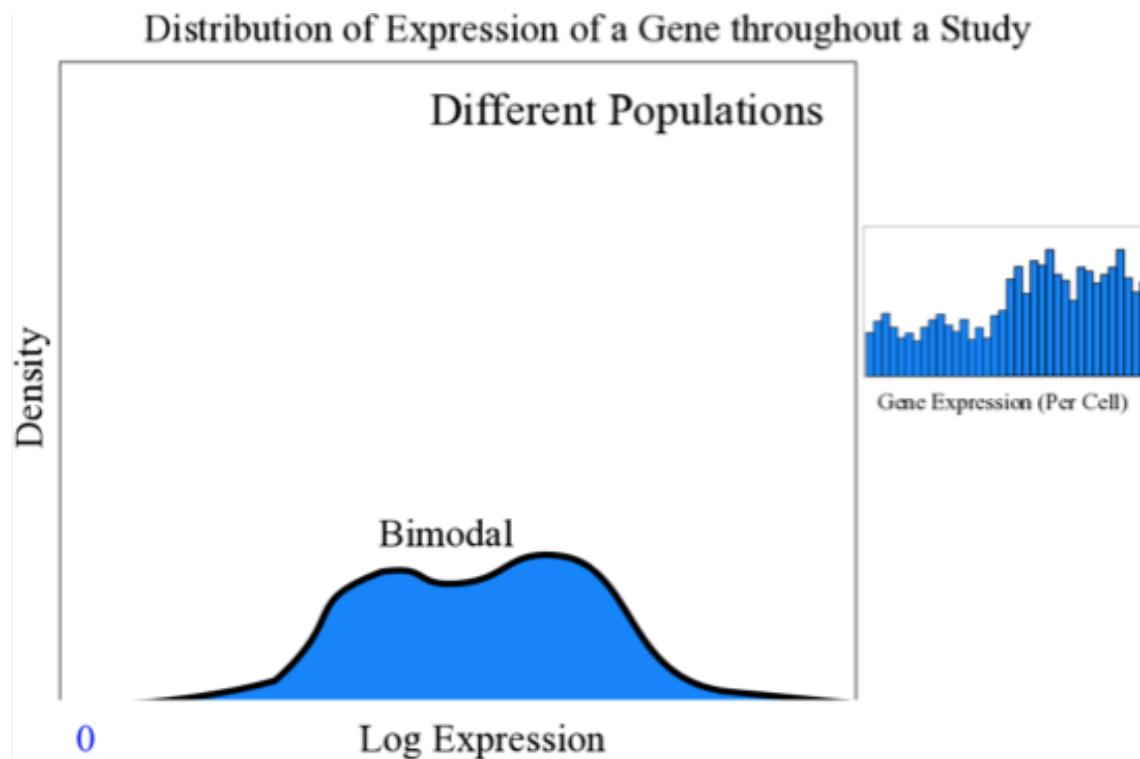
Counts Matrix

	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

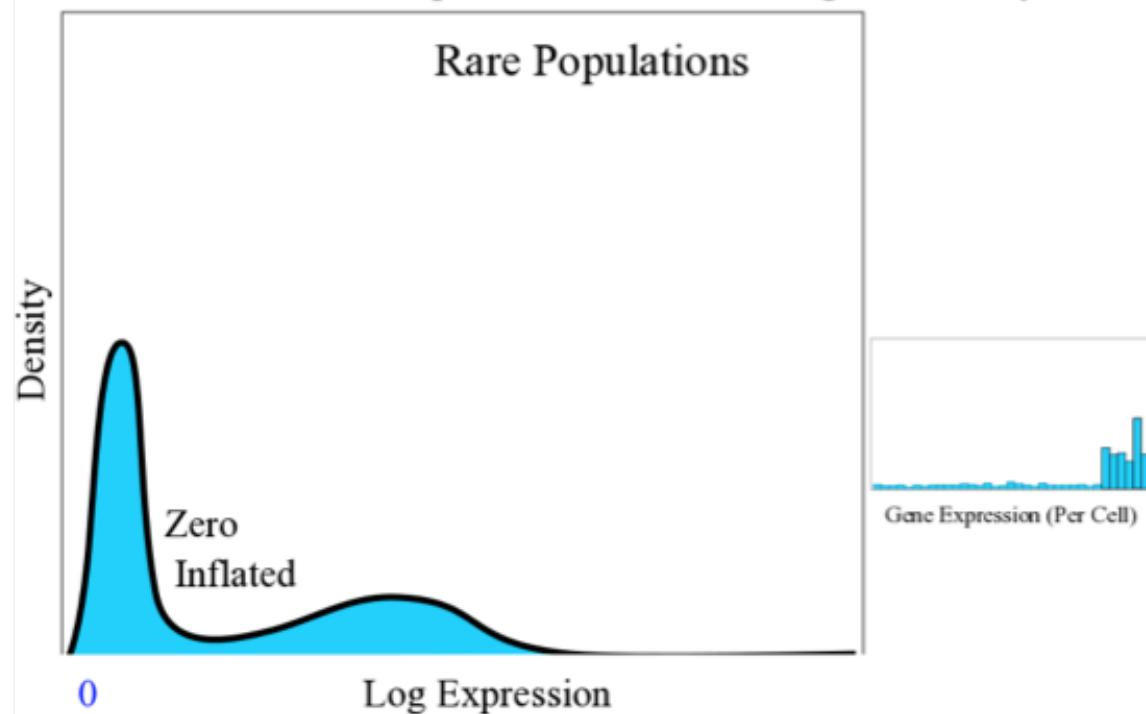


Genes Have Different Distributions



Genes Have Different Distributions

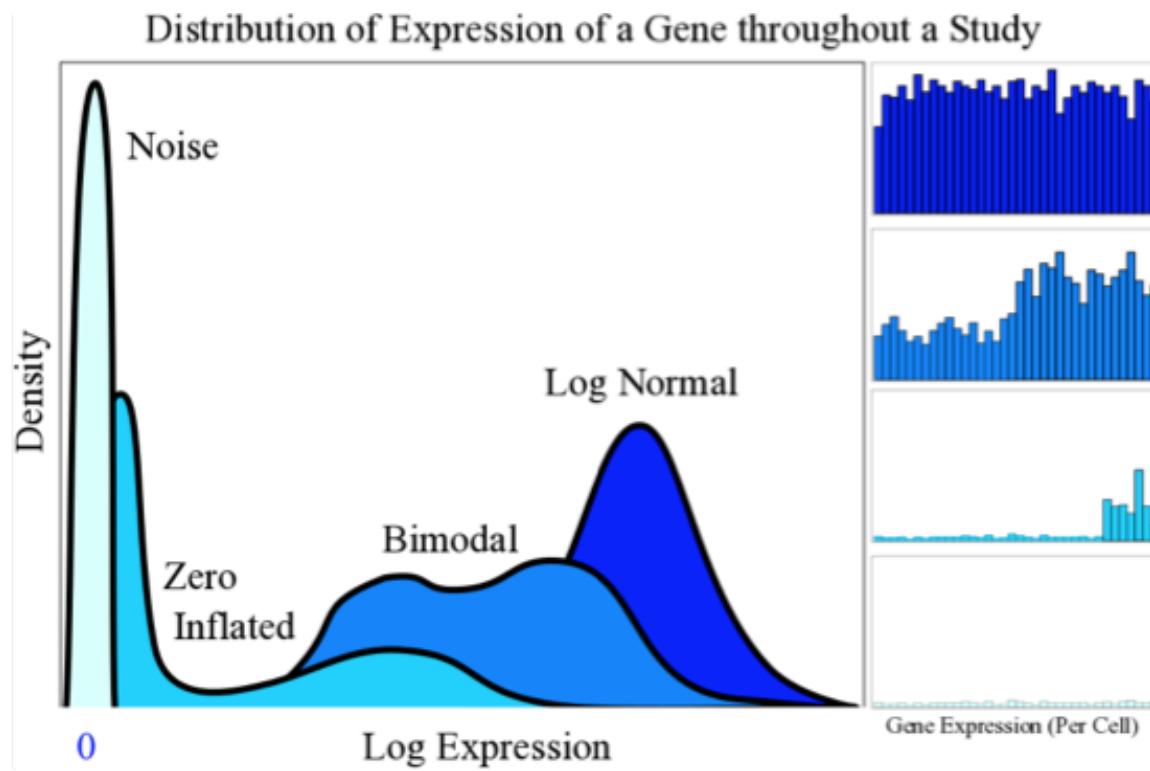
Distribution of Expression of a Gene throughout a Study



Genes Have Different Distributions

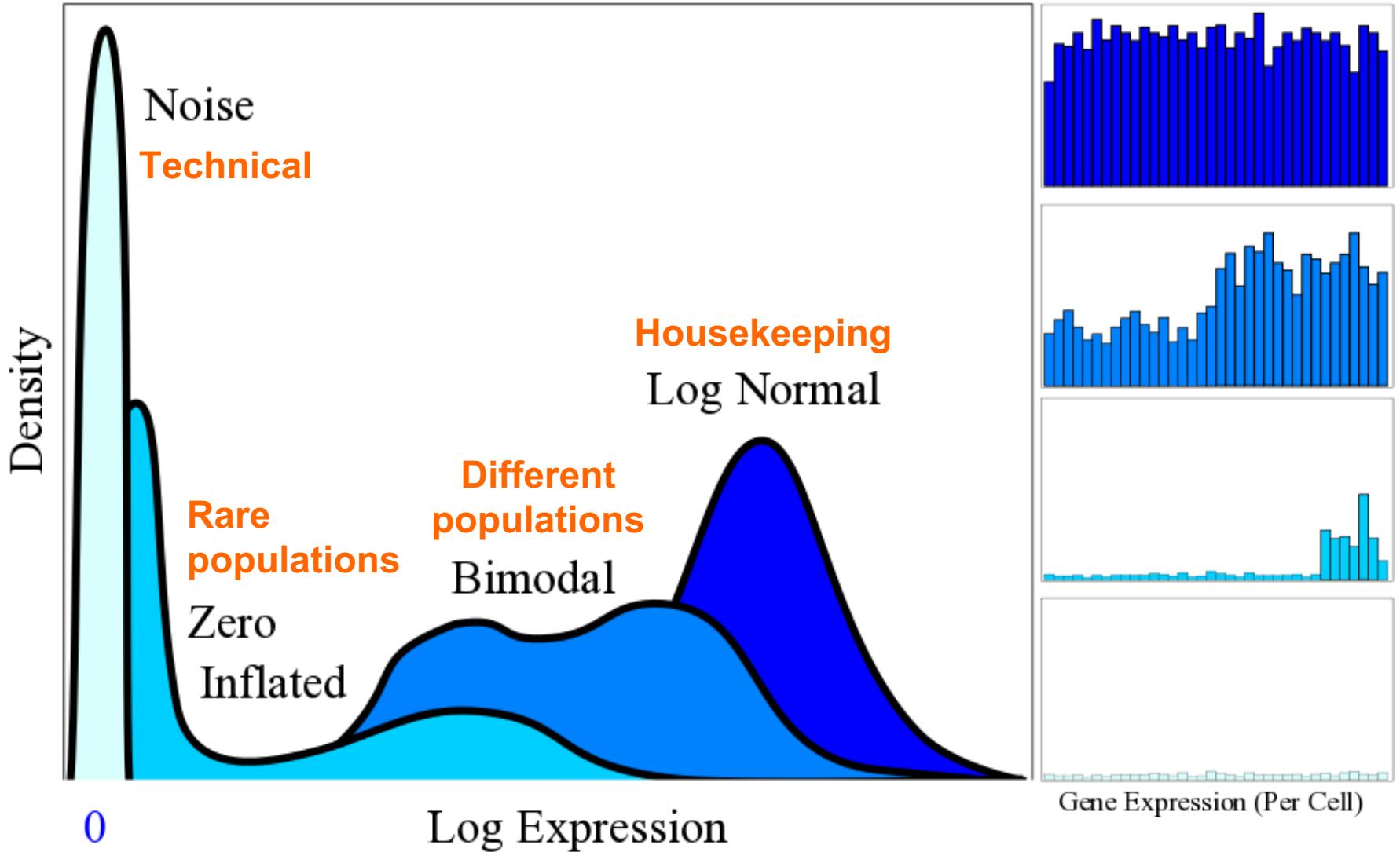


Genes Have Different Distributions



Genes have different distributions

Distribution of Expression of a Gene throughout a Study



Underlying Biology

- **Zero inflation.**

- Drop-out event during reverse-transcription.
- Genes with more expression have less zeros.
- Complexity varies.

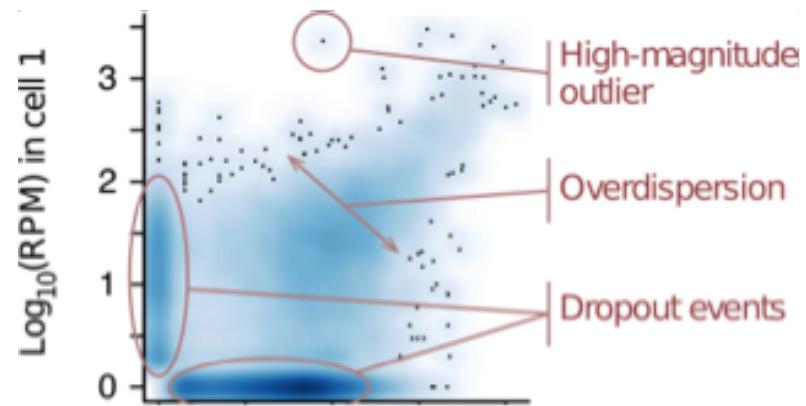
- **Transcription stochasticity.**

- Transcription bursting.
- Coordinated transcription of multigene networks.

- Over-dispersed counts.

- Higher Resolution.

- More sources of signal



BRIEF COMMUNICATIONS

Bayesian approach to
single-cell differential
expression analysis



Peter V Kharchenko¹⁻³, Lev Silberstein³⁻⁵ &
David T Scadden³⁻⁵

© 2014 Nature America, Inc.

Cell Identity is a Mixture of Multiple Factors

nature
biotechnology

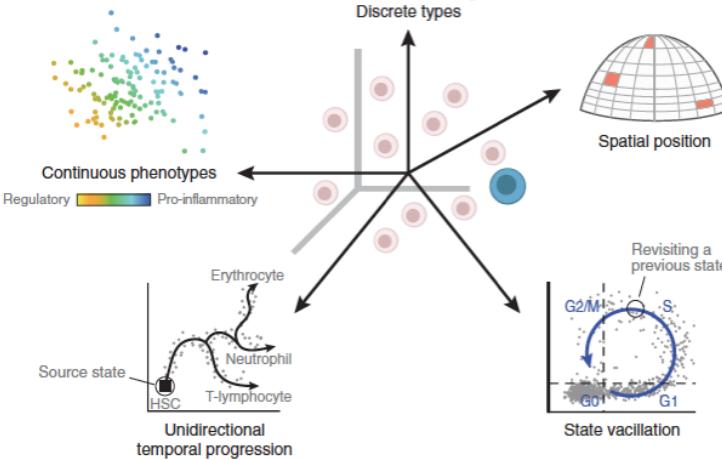
REVIEW

Revealing the vectors of cellular identity with single-cell genomics

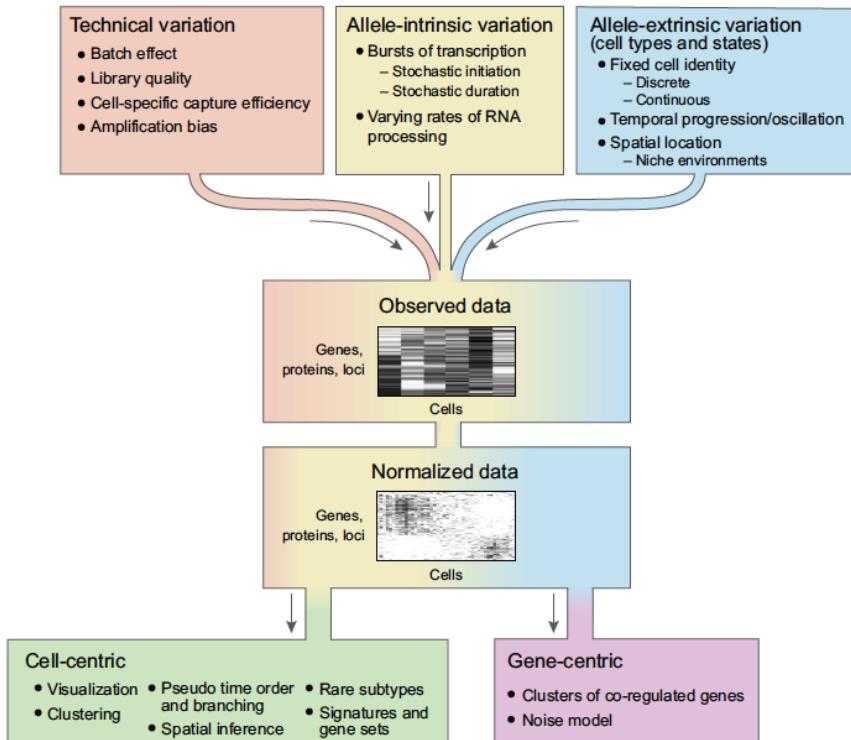
Allon Wagner¹, Aviv Regev^{2,3,5} & Nir Yosef^{1,4,5}

Multiple factors shape a cell's identity

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



Expression has Many Sources



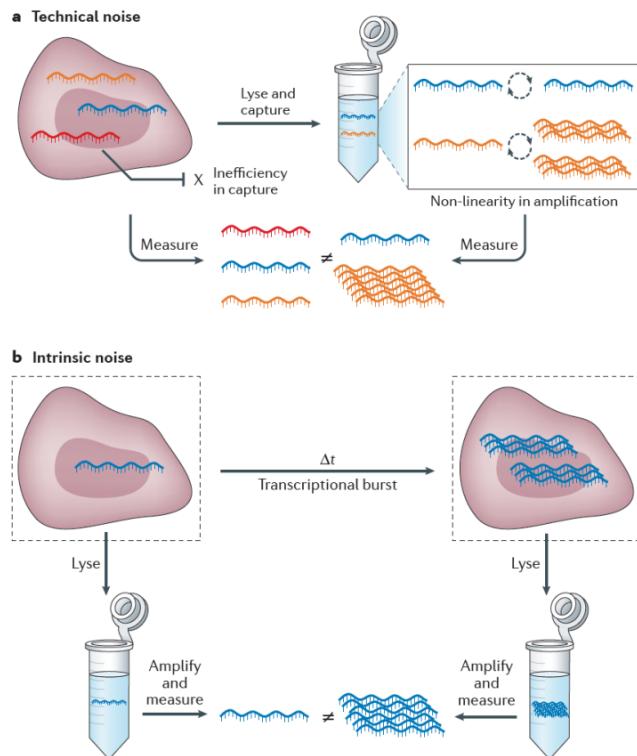
REVIEW

nature
biotechnology

Revealing the vectors of cellular identity with single-cell genomics

Allon Wagner¹, Aviv Regev^{2,3,5} & Nir Yosef^{1,4,5}

Technical vs Intrinsic Noise



SINGLE-CELL OMICS

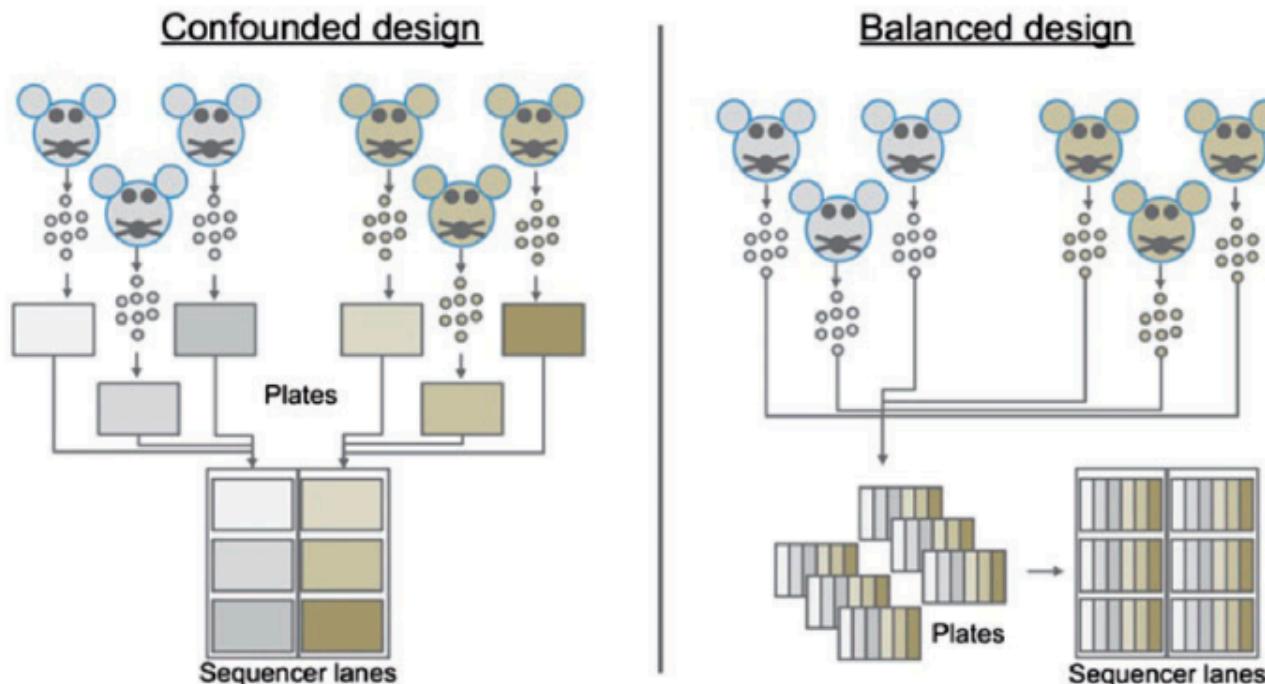
Scaling by shrinking: empowering single-cell ‘omics’ with microfluidic devices

Sanjay M. Prakadan^{1–3}, Alex K. Shalek^{1–3} and David A. Weitz^{4,5}

Experimental Design

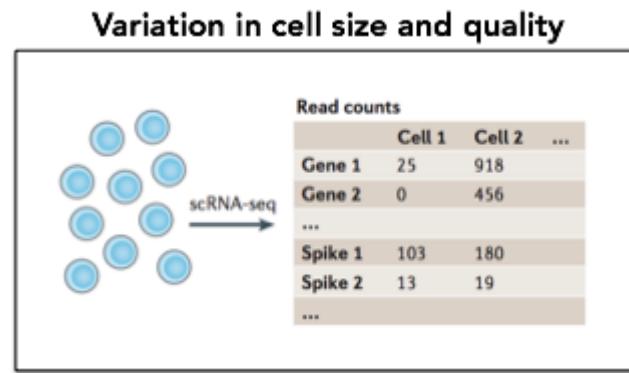
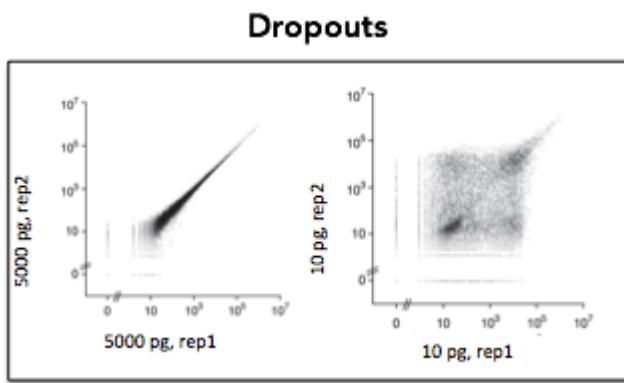
Sound experimental design : Replication, Randomization and Blocking

- R. A. Fisher, 1935

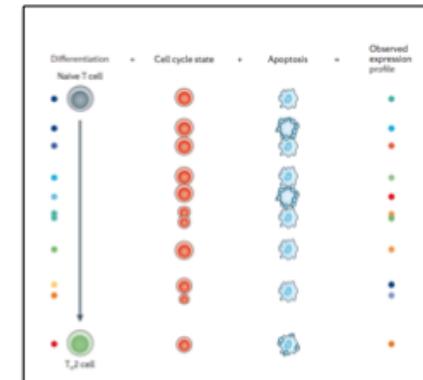
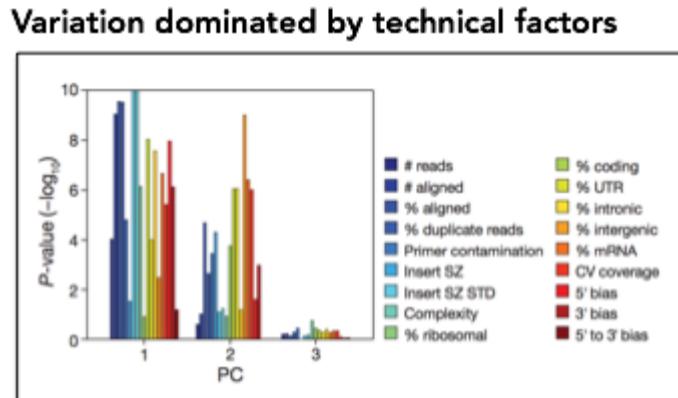


- For example, when analyzing tumor phenotypes in a patient process the tumor sample and a matched control on the same day, using the same reagents!
- Blocking is not always possible because of logistic limitations, in which case ensure that any biological conclusion is supported by multiple, independently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

Technical Conceptual Challenges



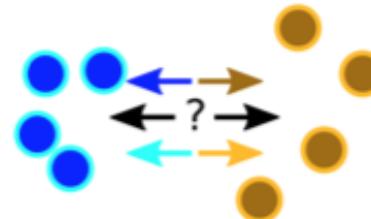
Observed gene expression is a convolution



What is Study Confounding

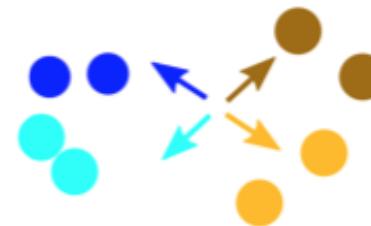
Cell | Site | Treatment

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

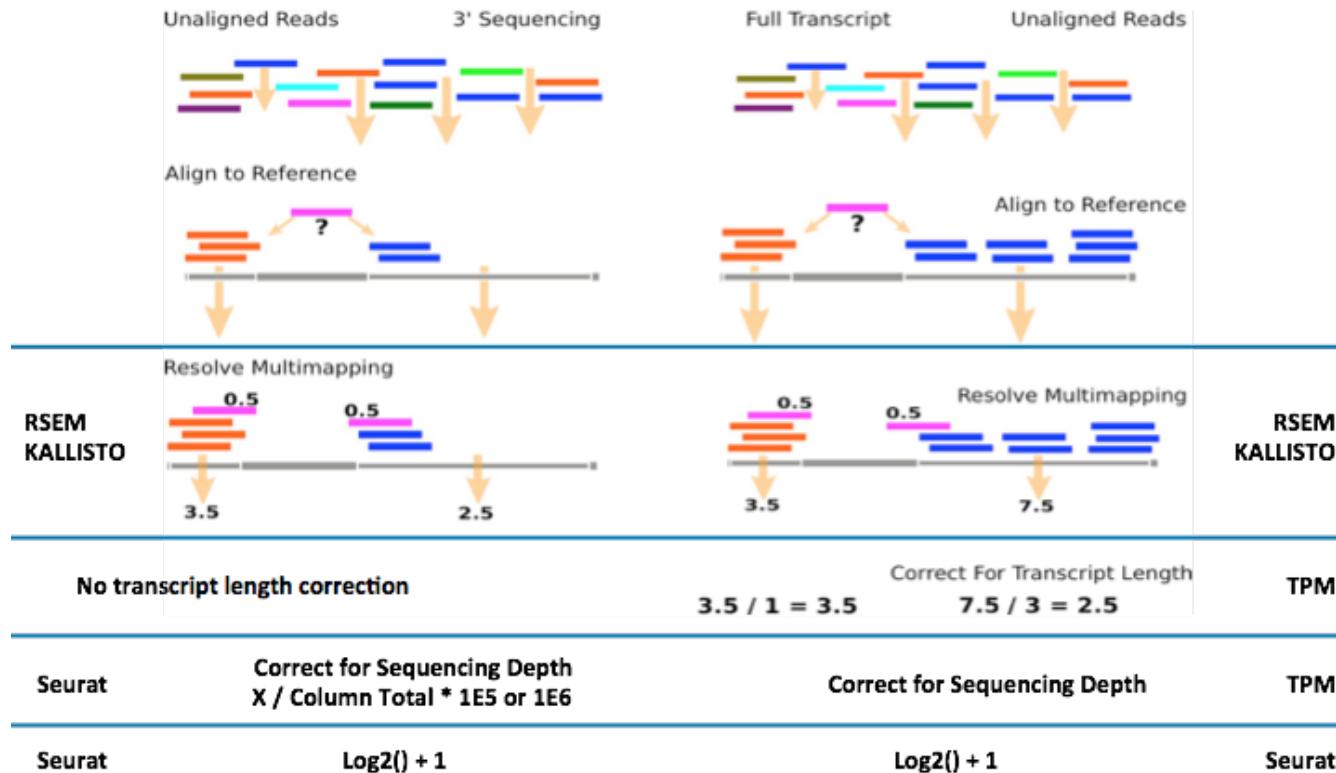
1	Main	A
2	Main	A
3	Main	B
4	Main	B
5	Remote	A
6	Remote	A
7	Remote	B
8	Remote	B



Goals for today

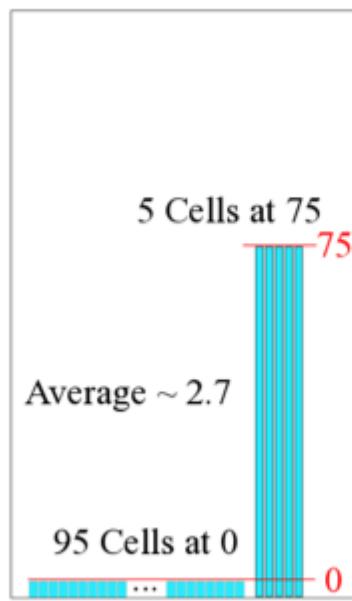
- Overview of single-cell RNA-seq data analysis
- What is in a count matrix?
- Preprocessing data: quality control (QC) and filtering
- Dimensionality reduction & Clustering
- Biology: inferring cell types and further

Count Preparation is Different Depending on Assays

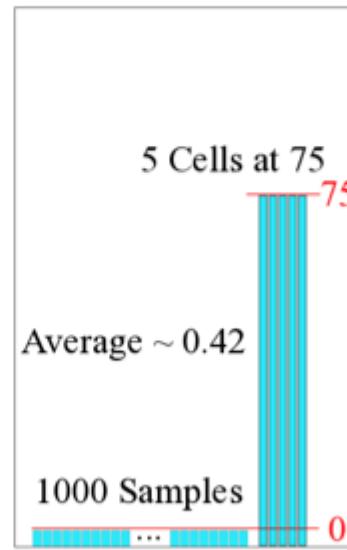


Filtering Genes: Averages are Less Useful

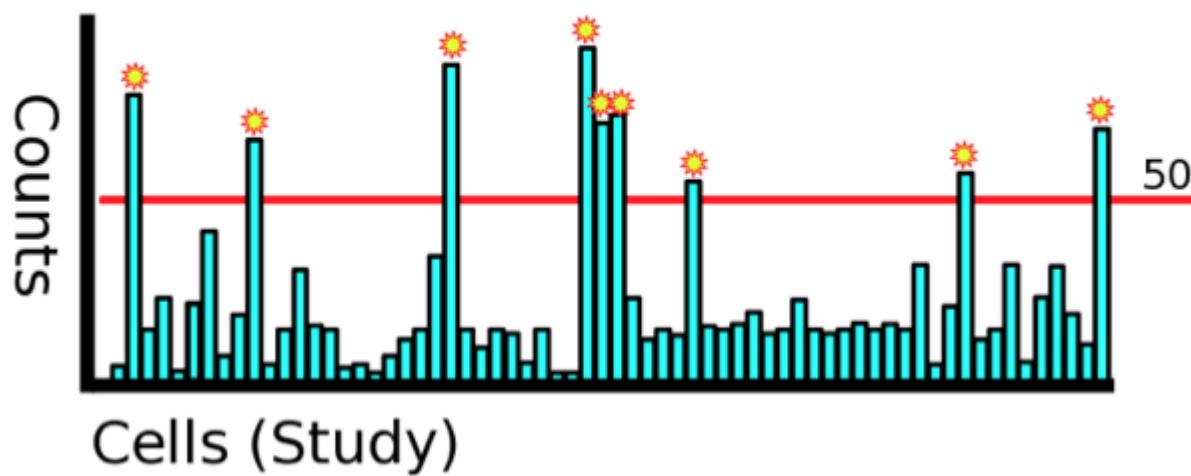
0s pull down average



Amount of 0s is arbitrary
(study size, diversity)

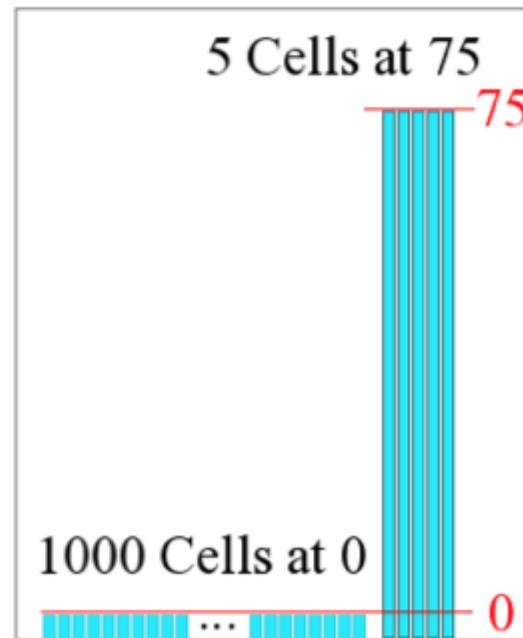
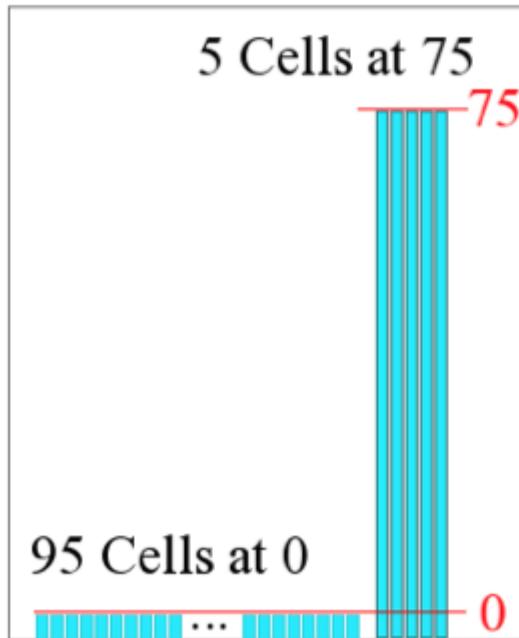


Filtering Genes: Using Prevalence



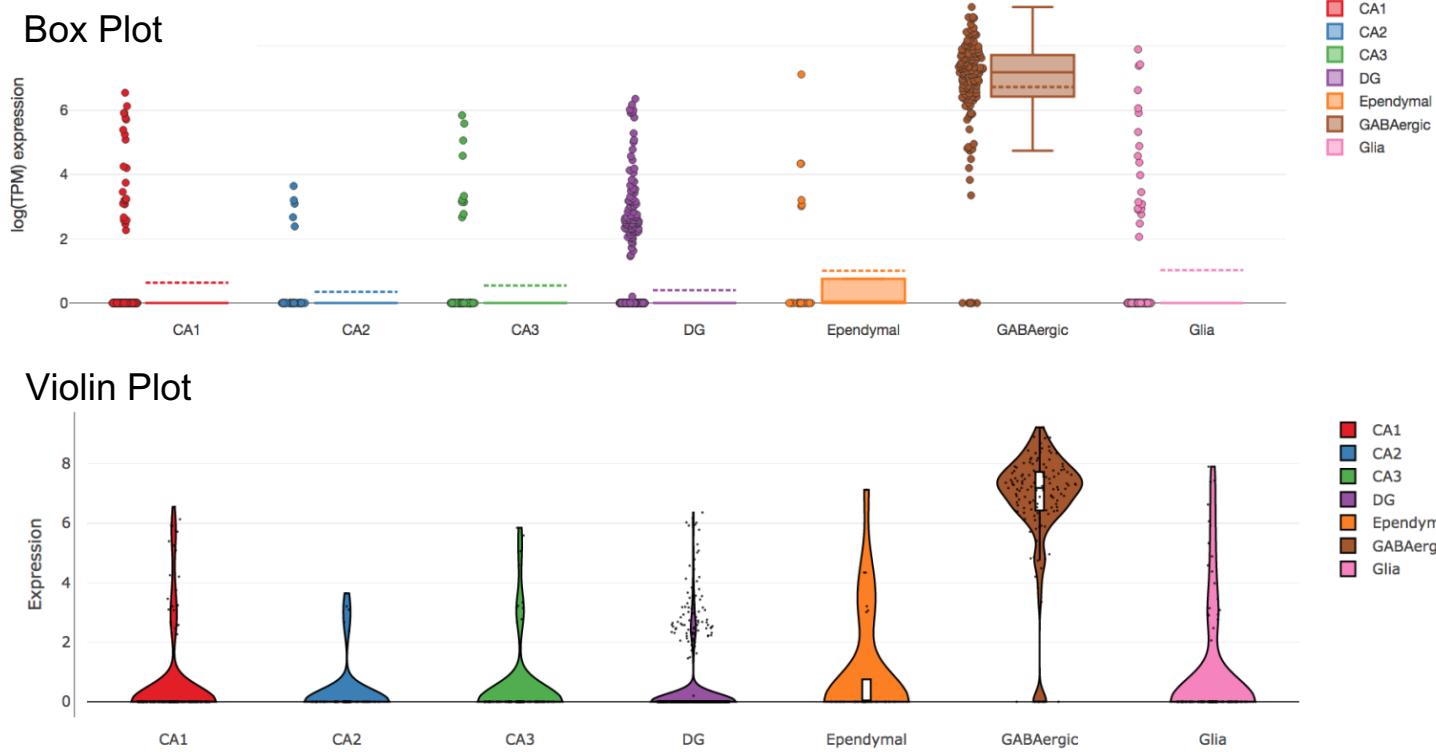
Filtering Genes: Using Prevalence

Filter: 5 cells must have 10 expression



Representing Genes Throughout Cells

A gene (GAD2)
across many
groups of cells.



Habib et al. 2016

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

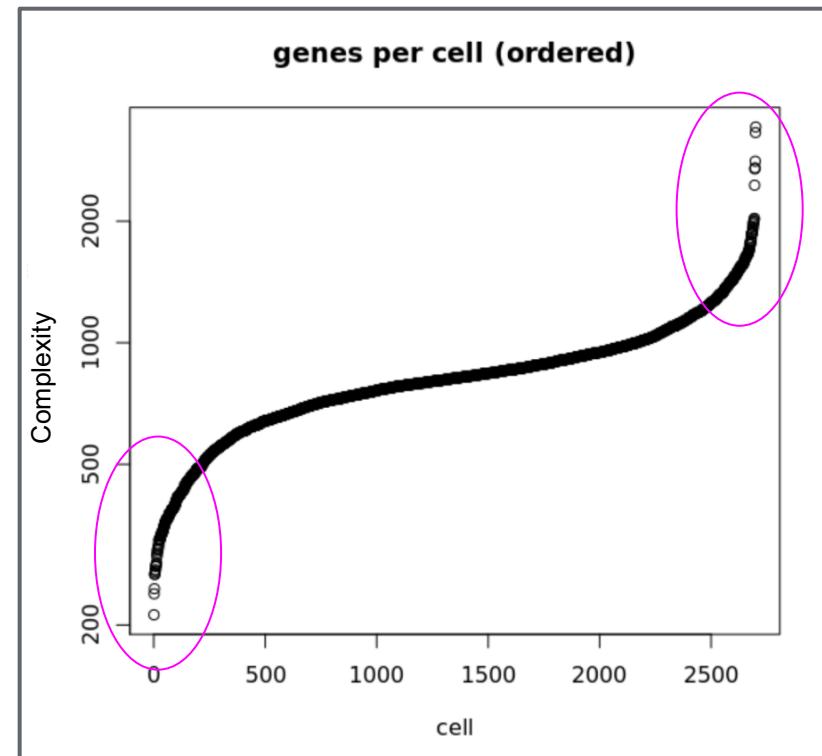
Complexity:

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

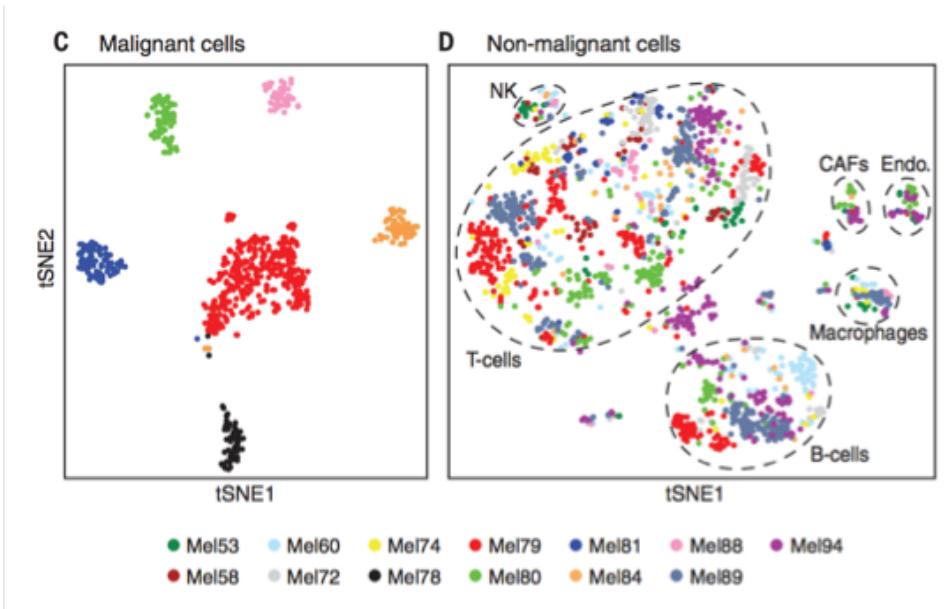
Filtering both
ends.

Lower: Failed
libraries?

Higher:
Duplicates?



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!*

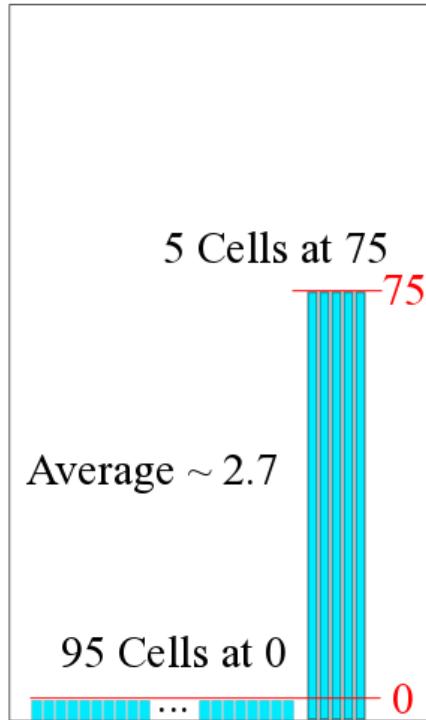
Tirosh et al., Science 2016

Some single-cell RNA-seq data challenges to remember

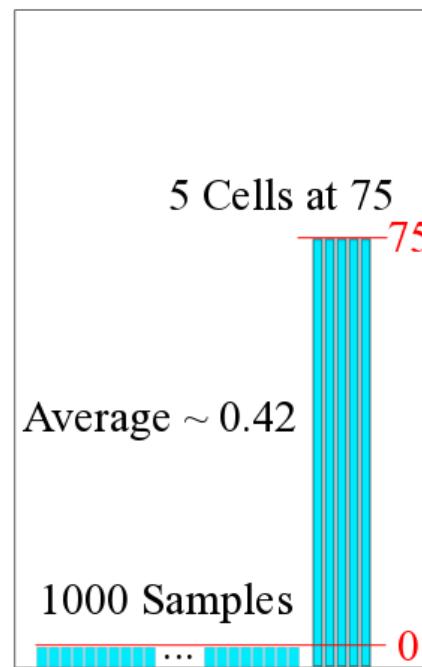
- **Drop out:** data has an excessive amount of zeros due to limiting mRNA

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

0s pull down average



Amount of 0s is arbitrary
(study size, diversity)

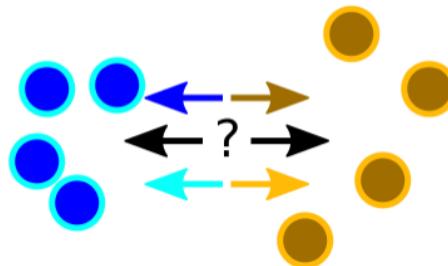


Some single-cell RNA-seq data challenges to remember

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

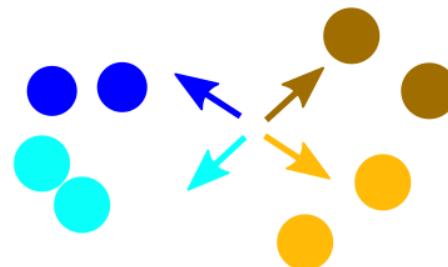
Cell | Site | Treatment

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

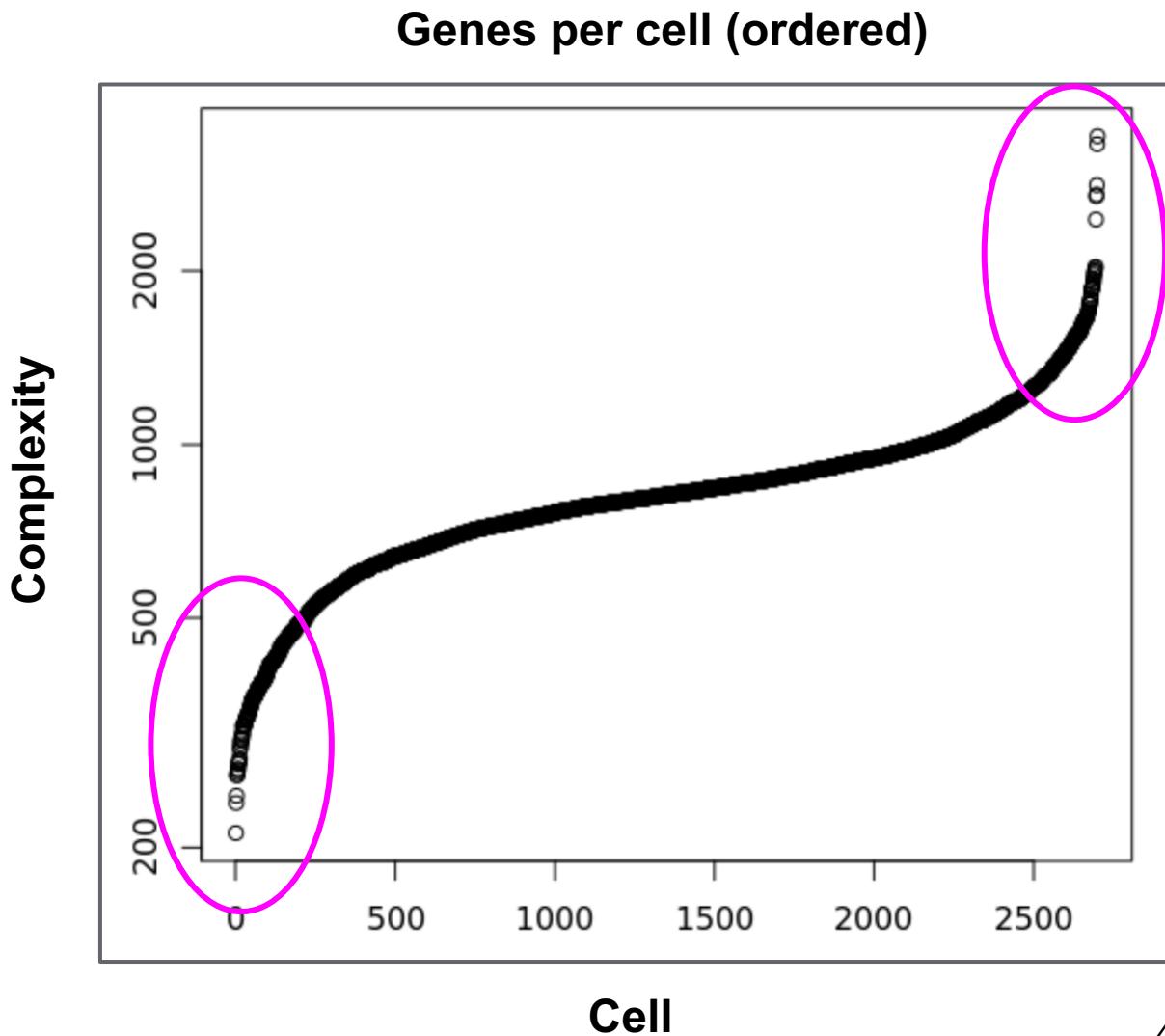
1	Main	A
2	Main	A
3	Main	B
4	Main	B
5	Remote	A
6	Remote	A
7	Remote	B
8	Remote	B



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

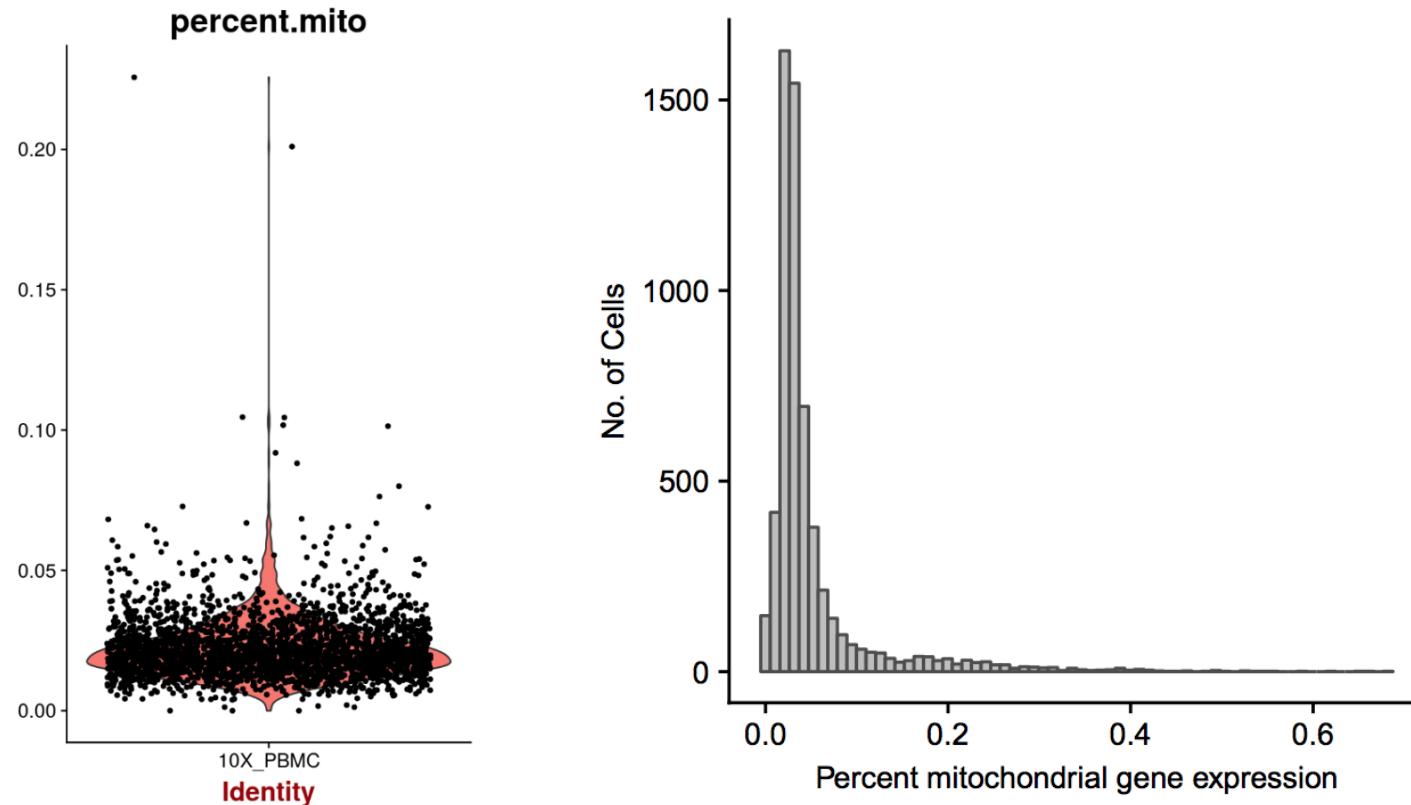
Filtering and quality control: Number of genes per cell

complexity =
number of genes
detected in a cell



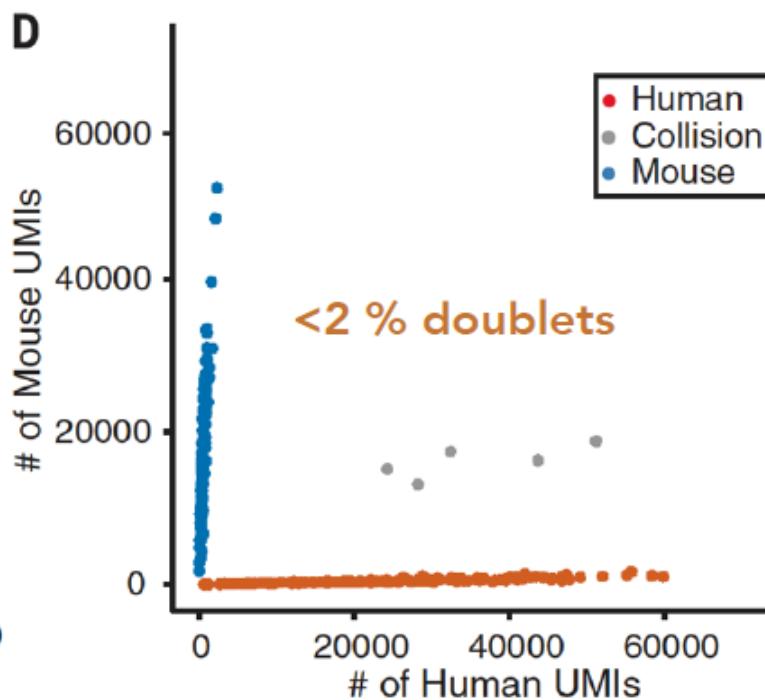
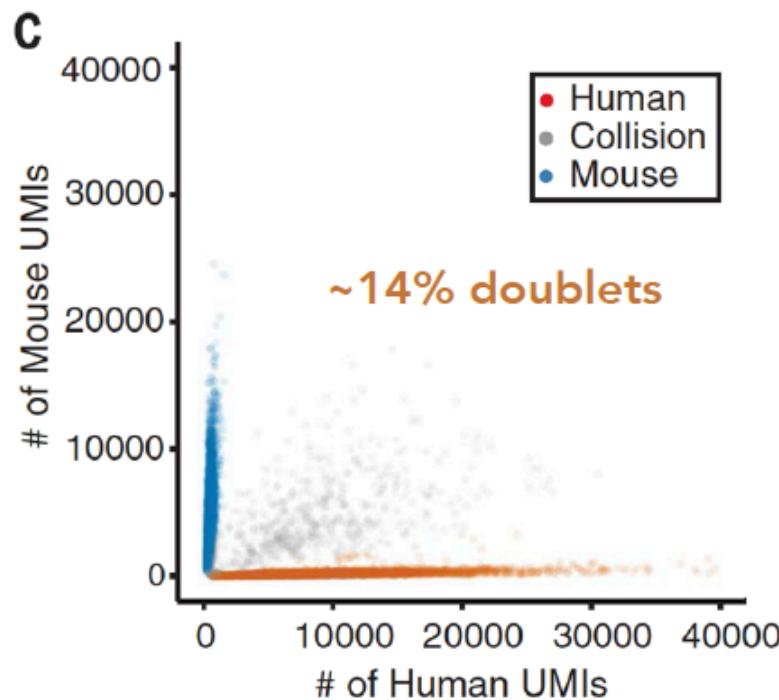
Filtering and quality control: Mitochondrial gene expression

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

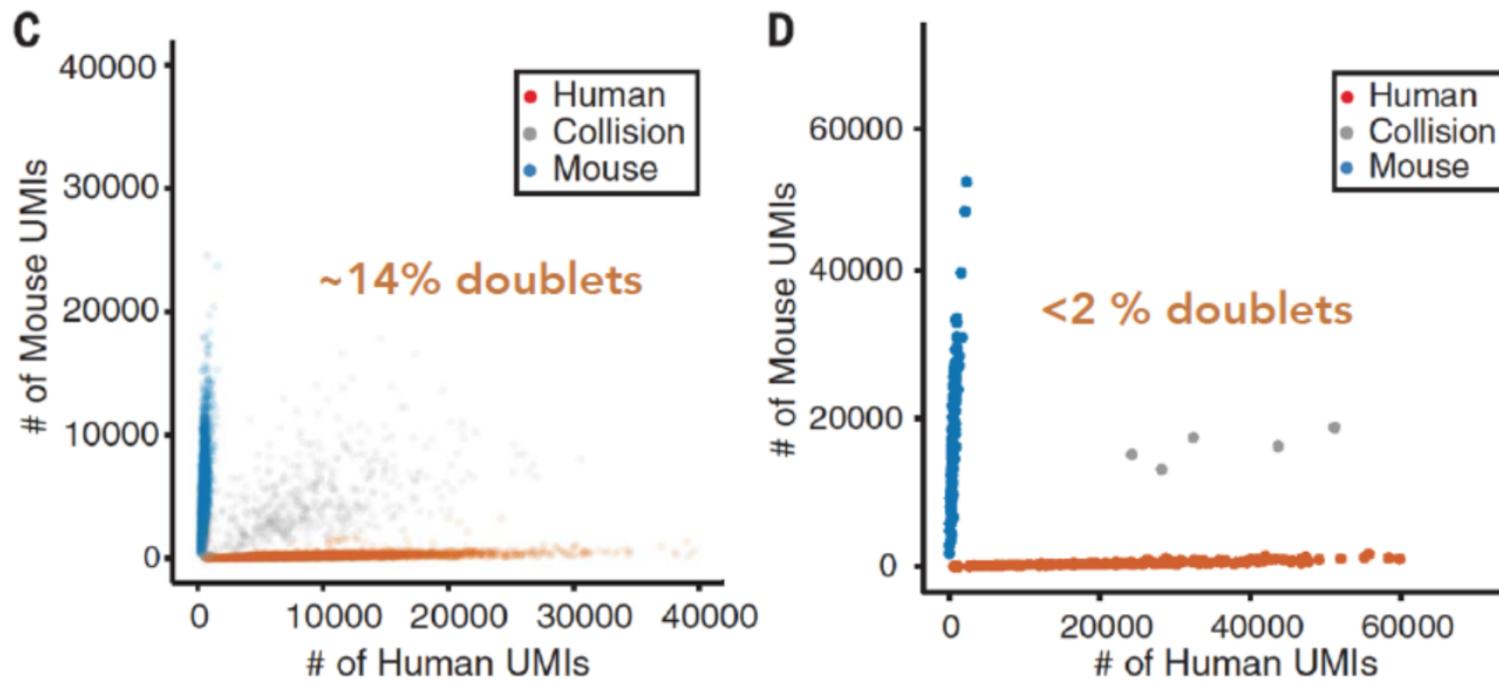
Because of the setup, it is possible that two or more cells can enter the same droplet. Studies estimate doublet frequency through a “mixed-species” experiment



The doublet frequency is +vely correlated with throughput

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

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

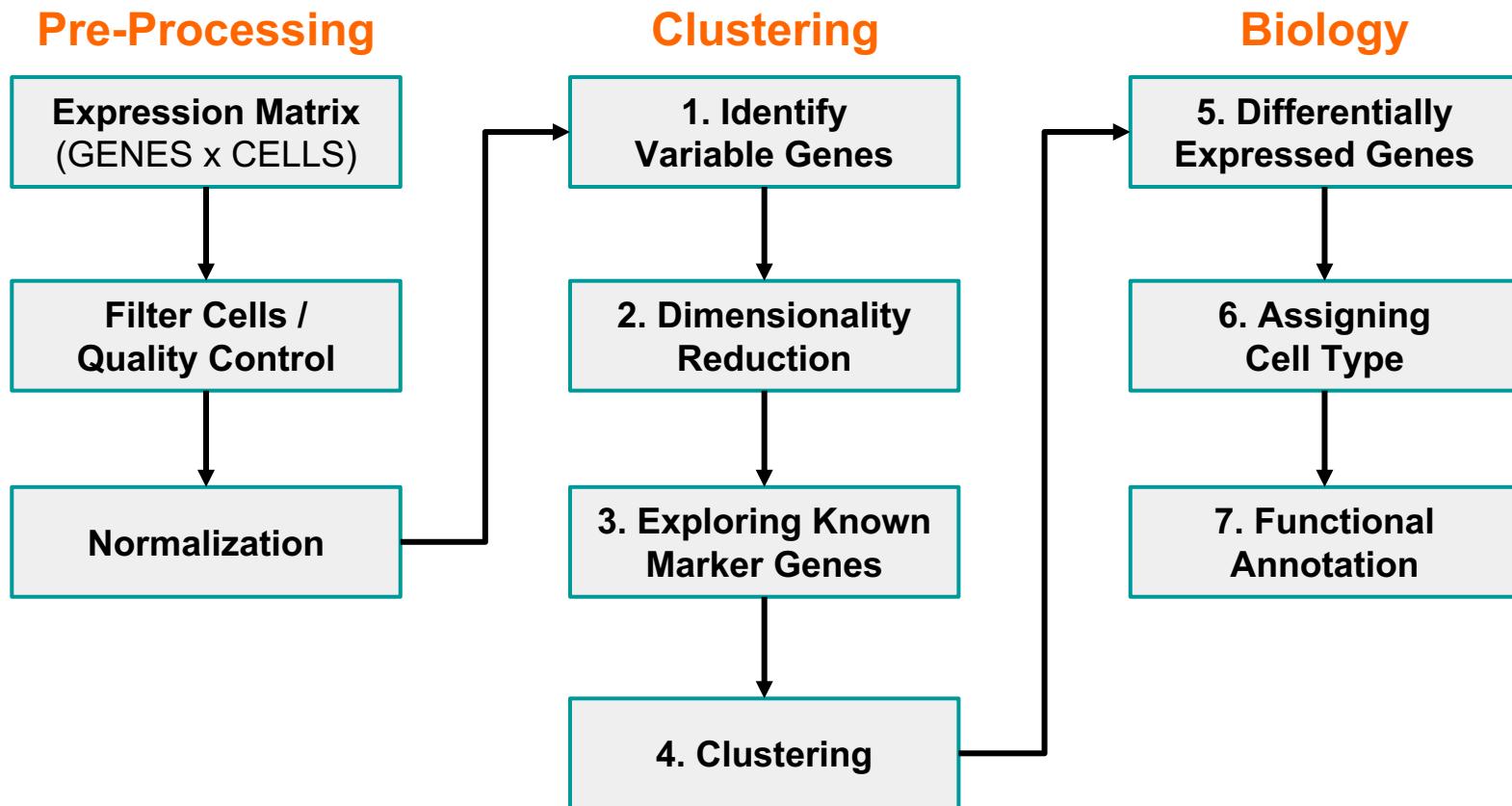
 Christopher S McGinnis,  Lyndsay M Murrow,  Zev J Gartner

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*

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



Spatial reconstruction of single-cell gene expression data

Rahul Satija^{1,7,8}, Jeffrey A Farrell^{2,8}, David Gennert¹, Alexander F Schier^{1-5,9} & Aviv Regev^{1,6,9}

Cell

Resource

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

2D Arrays

vs

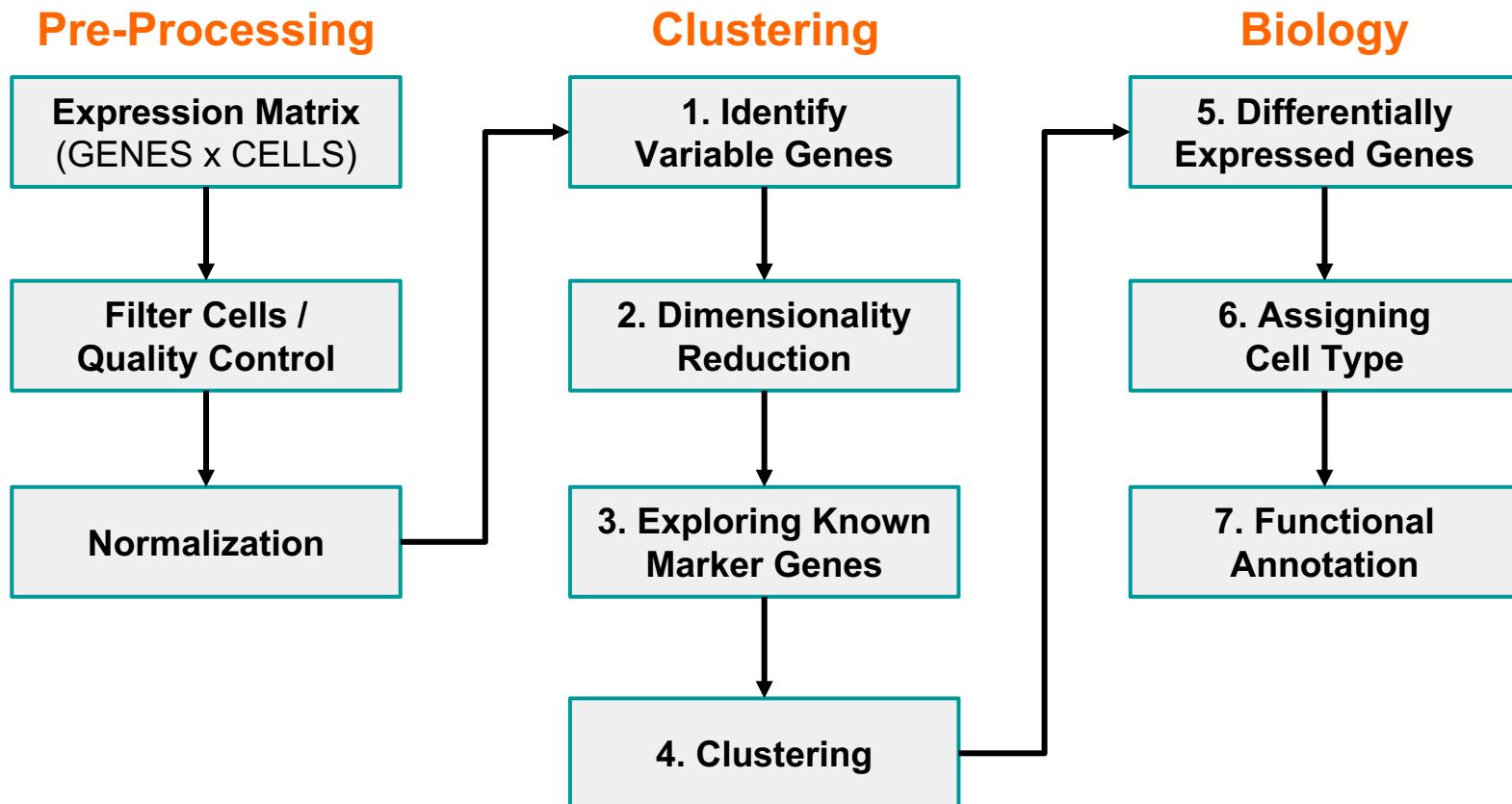
Coordinate List

1	2	3	4	5	6	7	8
0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	0	0	0	0	2	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	3

1
2
3
4
5
6

2	2	1
6	3	2
8	6	3

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

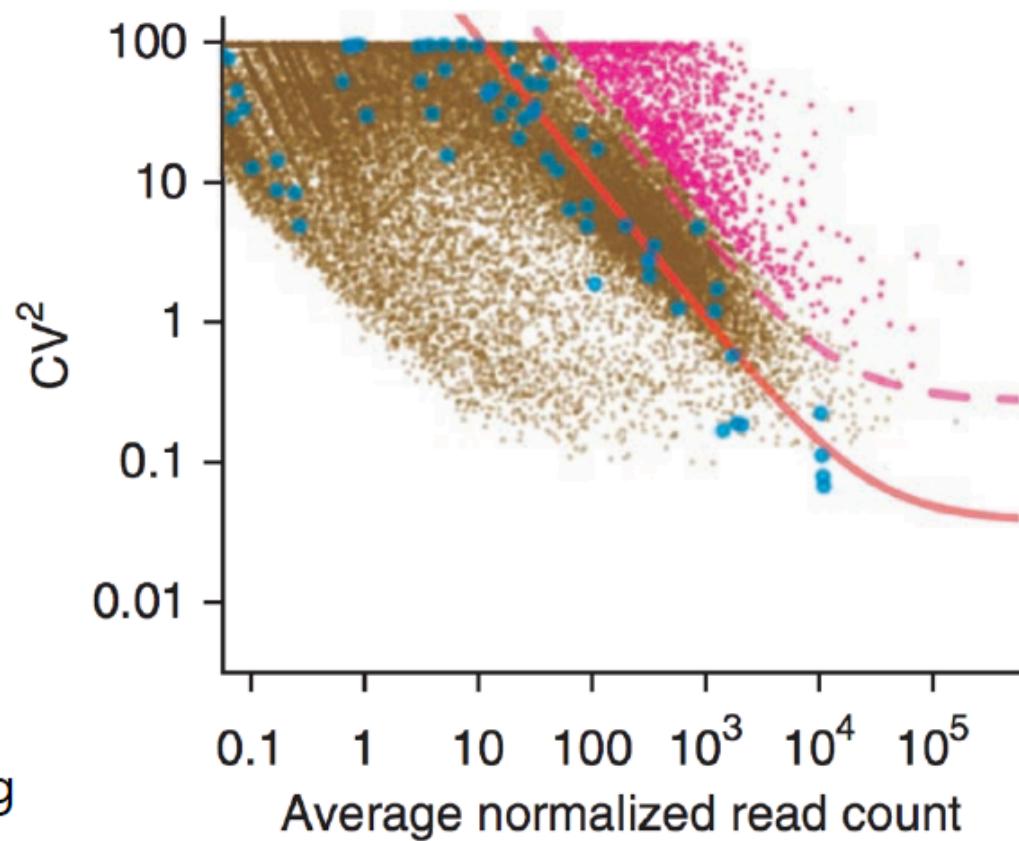


1. Identifying relevant, "highly variable" features

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

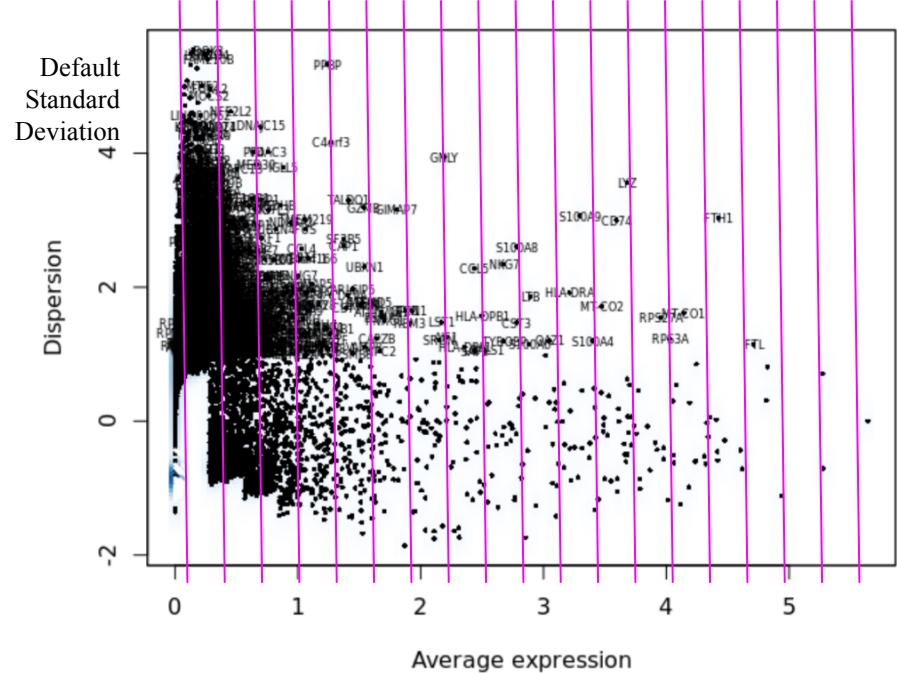
Variable Genes in Seurat

Calculate mean expression.

Calculate dispersion (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

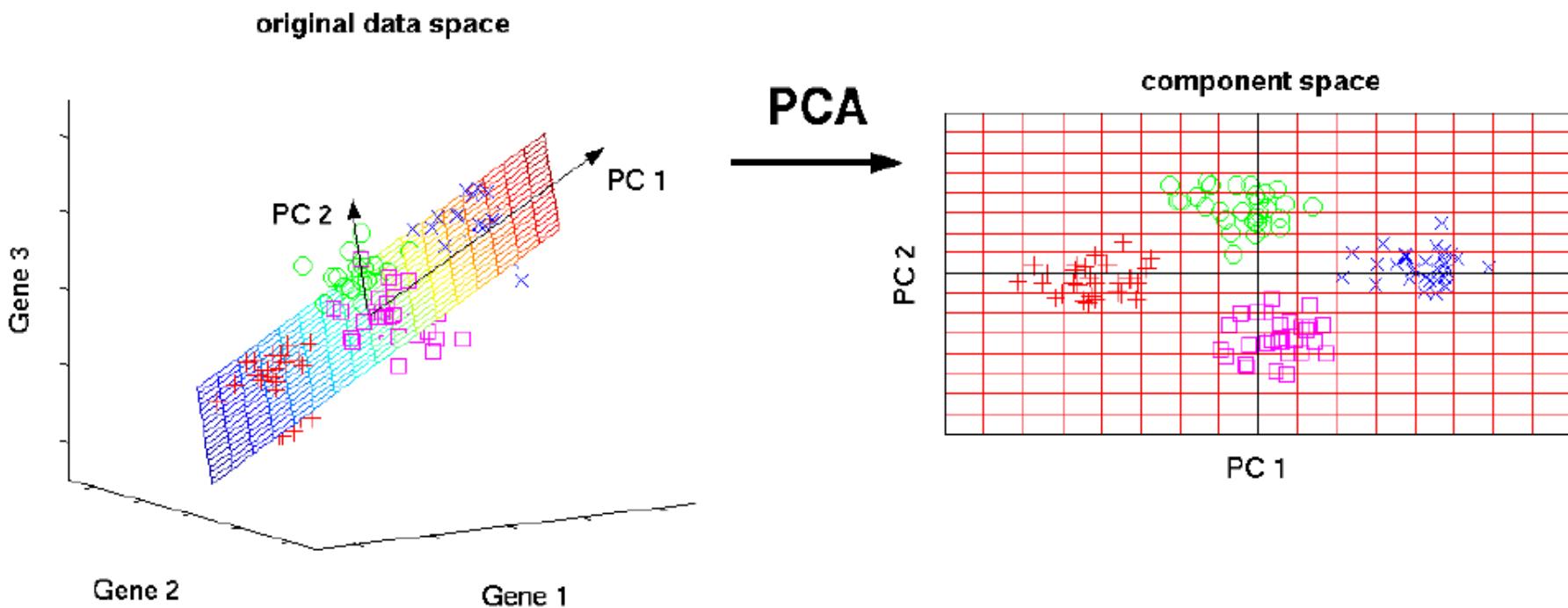
- many genes are lowly detected / noisy measurements
- genes are not independent of one another! rather they operate in coregulatory modules

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

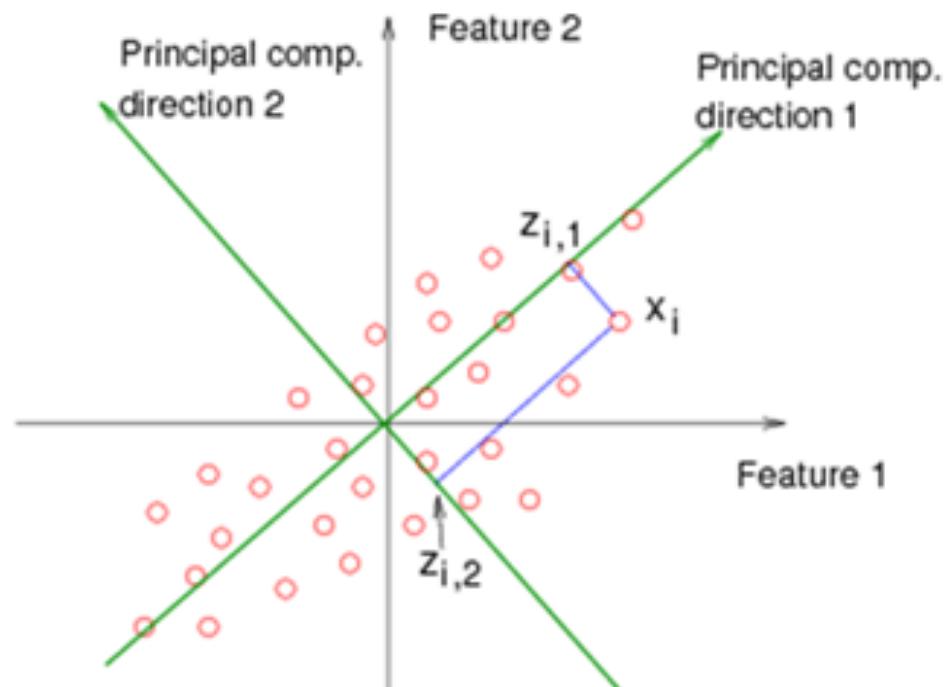


- Principal Component Analysis (PCA) is a **popular linear-method** to identify these modules

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

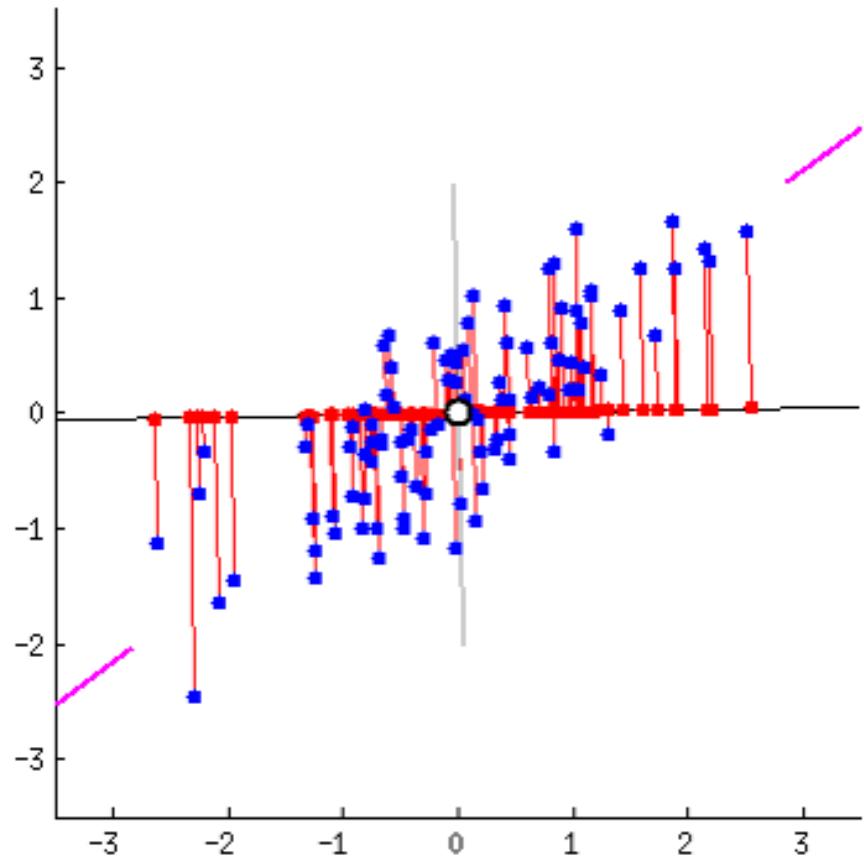


Determining cell type, state, and/or function:

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

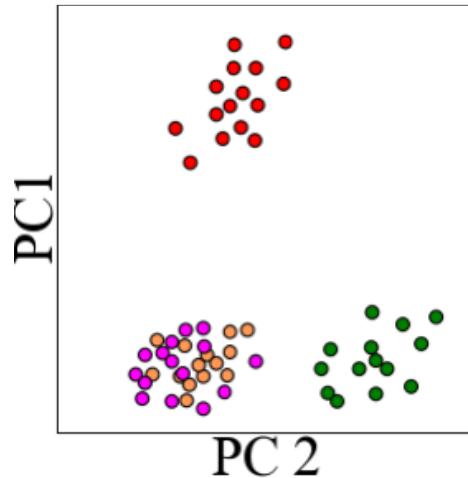
Identifying maximal orthogonal sources of variation



Determining cell type, state, and/or function:

2. Dimensionality reduction

PCA of
single cell
data

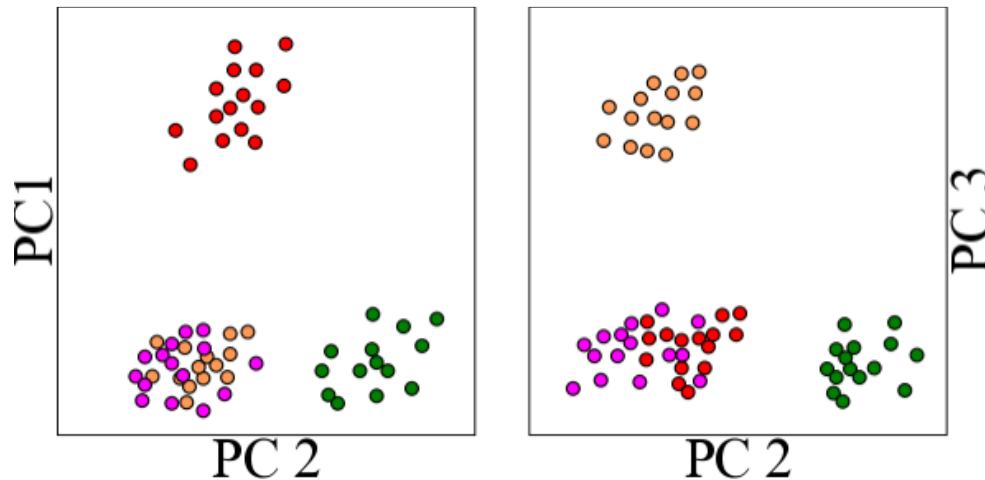


- PC1 separates the red cells from the pink, orange, and green cells
- PC2 separates the green cells from the red, pink, and orange cells

Determining cell type, state, and/or function:

2. Dimensionality reduction

PCA of
single cell
data

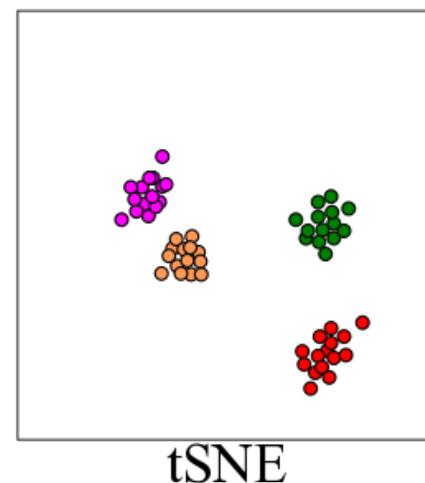
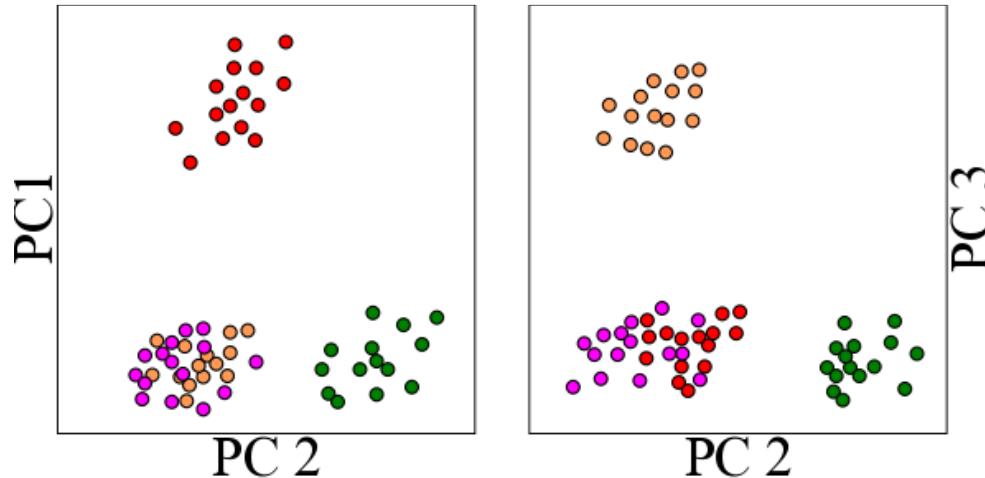


- PC3 further splits off the orange cells

Determining cell type, state, and/or function:

2. Dimensionality reduction

PCA of
single cell
data



tSNE: t-
distributed
Stochastic
Neighbor
Embedding

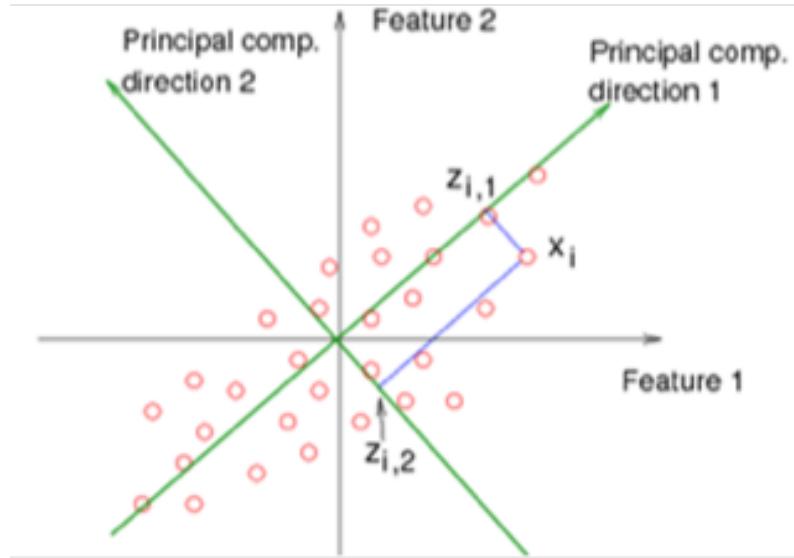
- tSNE is nonlinear dimensionality reduction
- tSNE collapse the visualization to 2D

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.



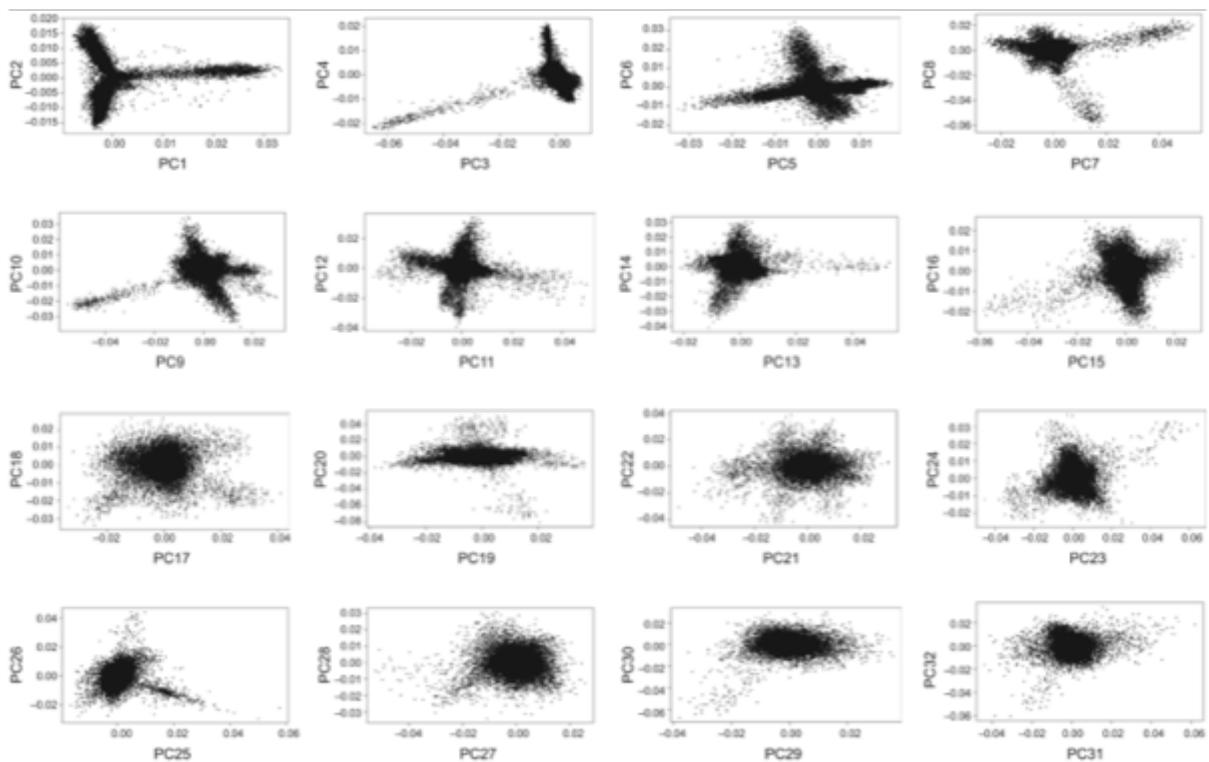
PCA: in Practice

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.**

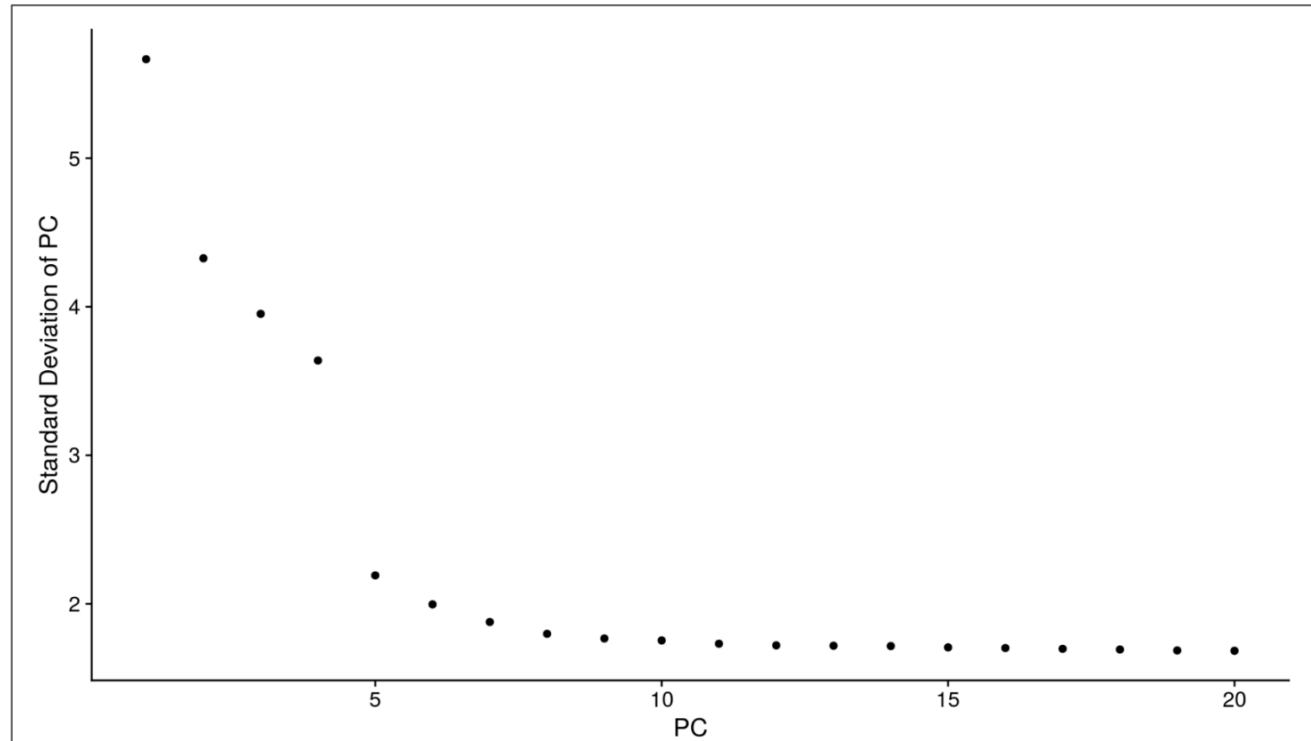
PCs

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)**



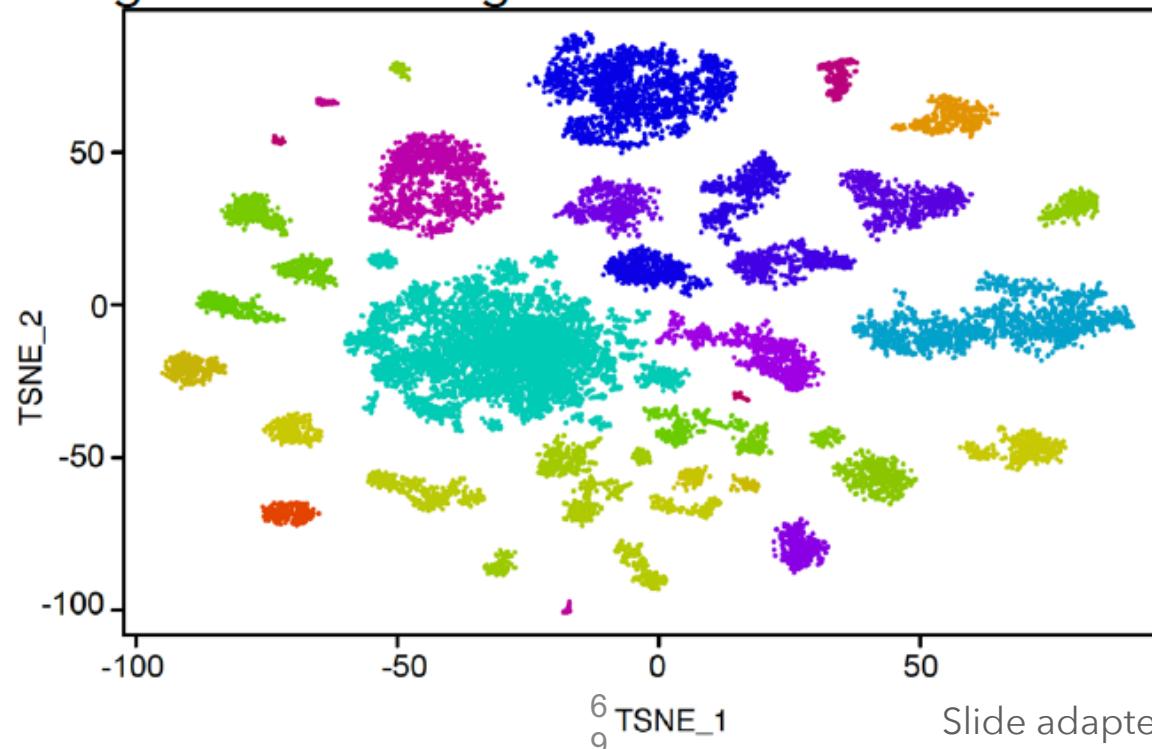
3. Visualization

t-distributed Stochastic Neighbor Embedding

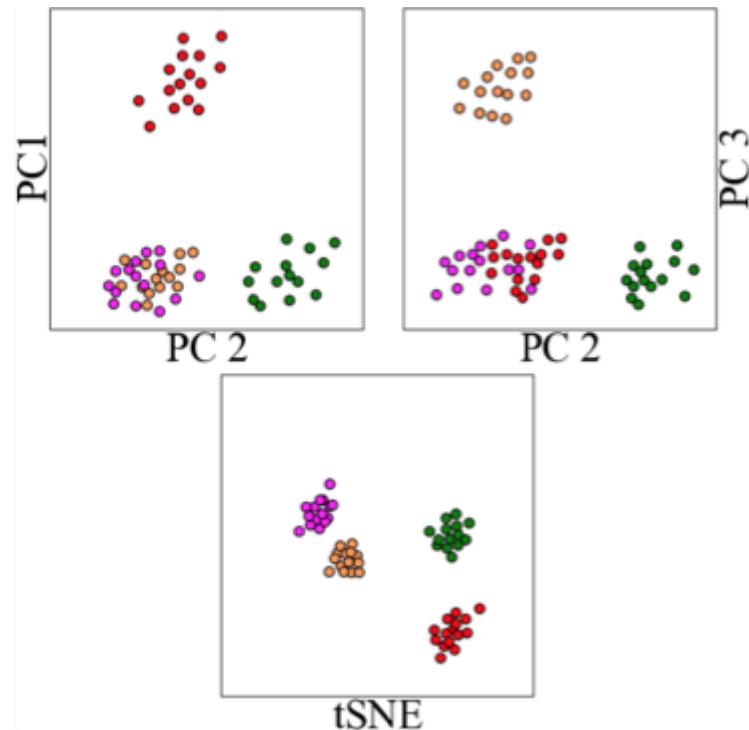
- A non-linear embedding method that preserves local distances between data-points in the low-dimensional space

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

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



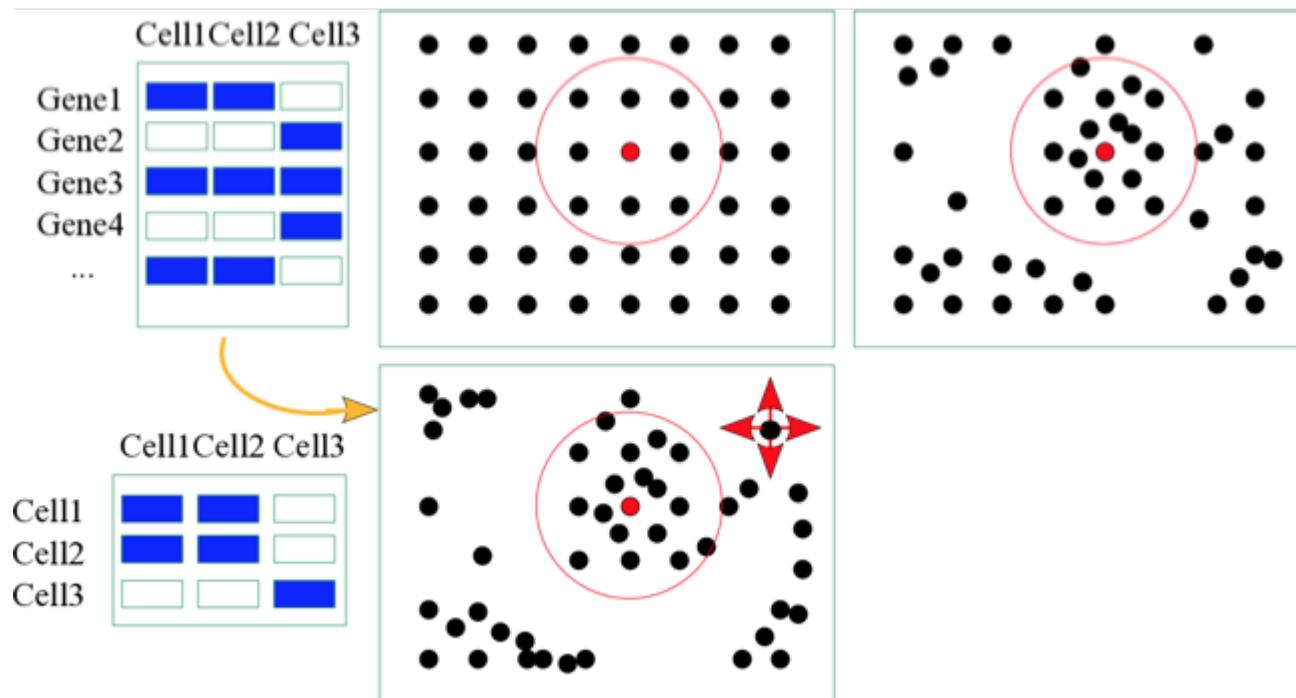
t-SNE: Collapsing the Visualization to 2D



t-SNE: Nonlinear Dimensionality Reduction



t-SNE: How it Works



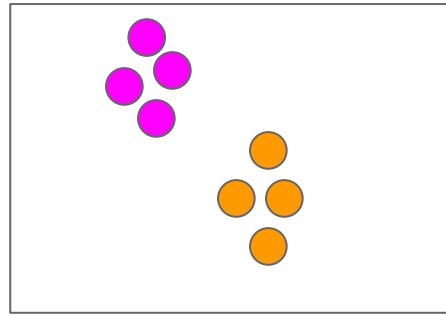
PCA and t-SNE Together

- 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.

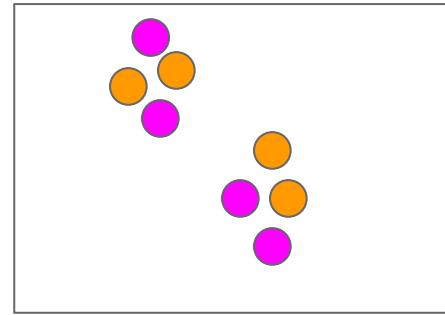
Plotting Metadata on Ordinations

Metadata

X

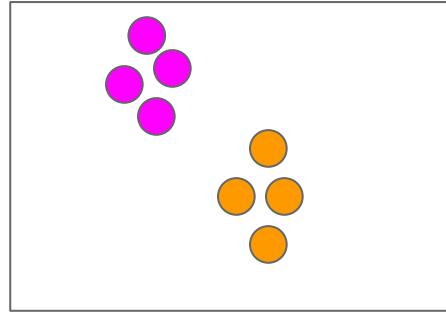


✓

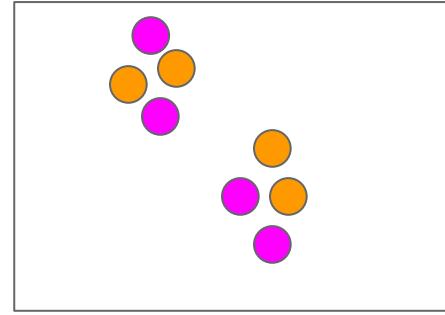


Gene Expression

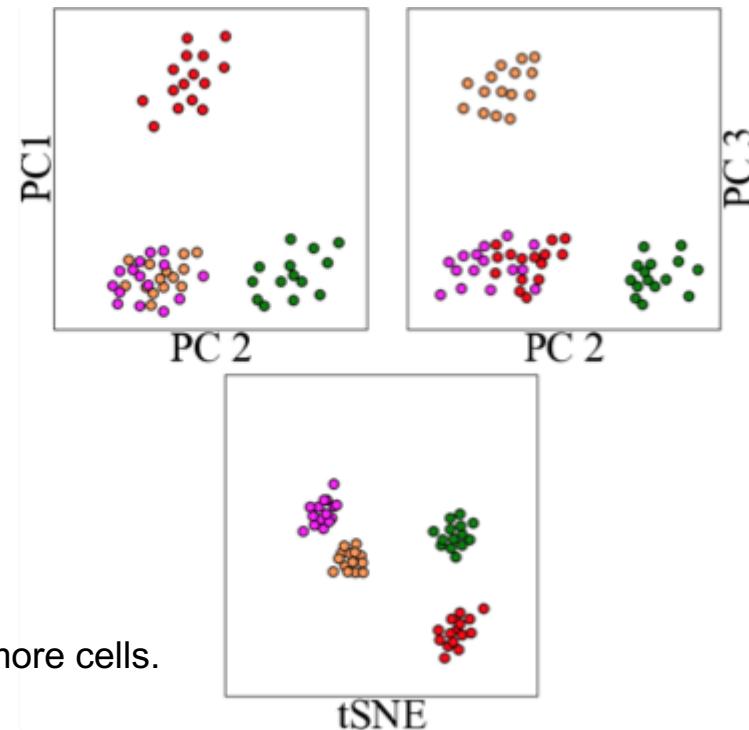
✓



X



Caution When Interpreting t-SNE



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=RJVL80Gg3IA>

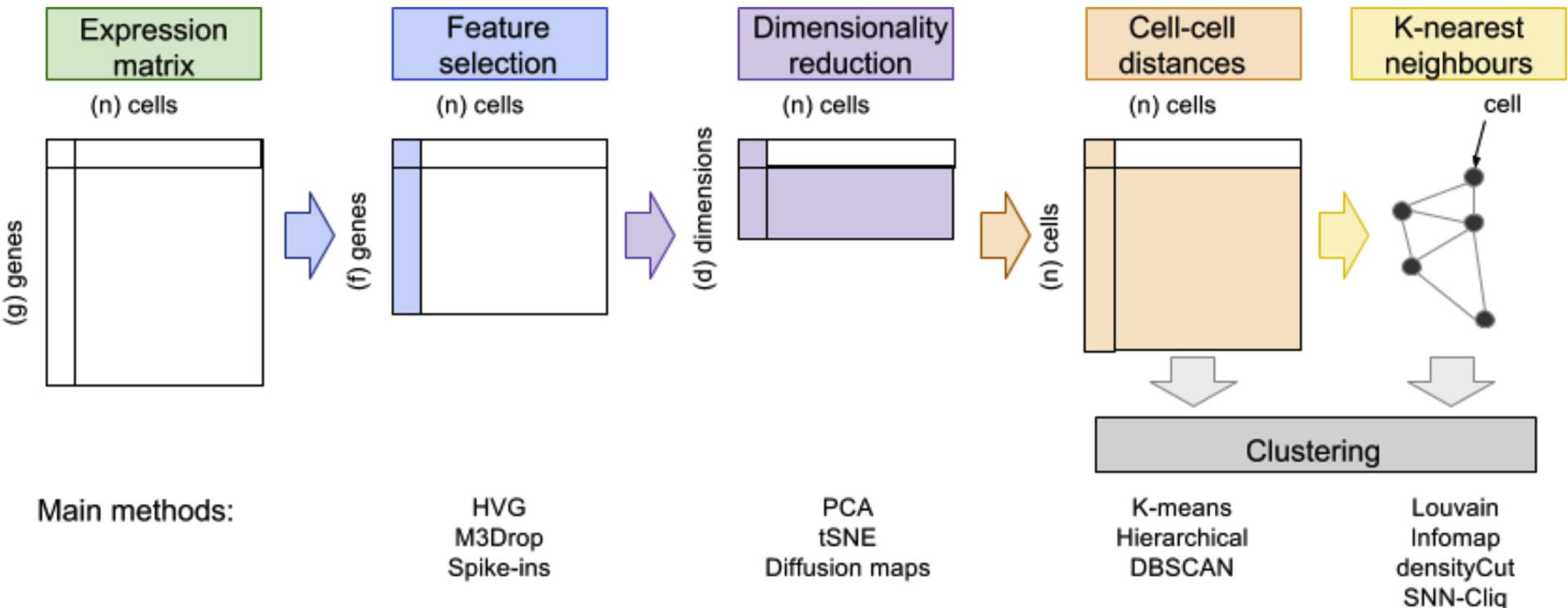
- Code

- <https://lvdmaaten.github.io/tsne/>

- Interactive Tensorflow

- <http://projector.tensorflow.org/>

4. Clustering cells to identify cell-types



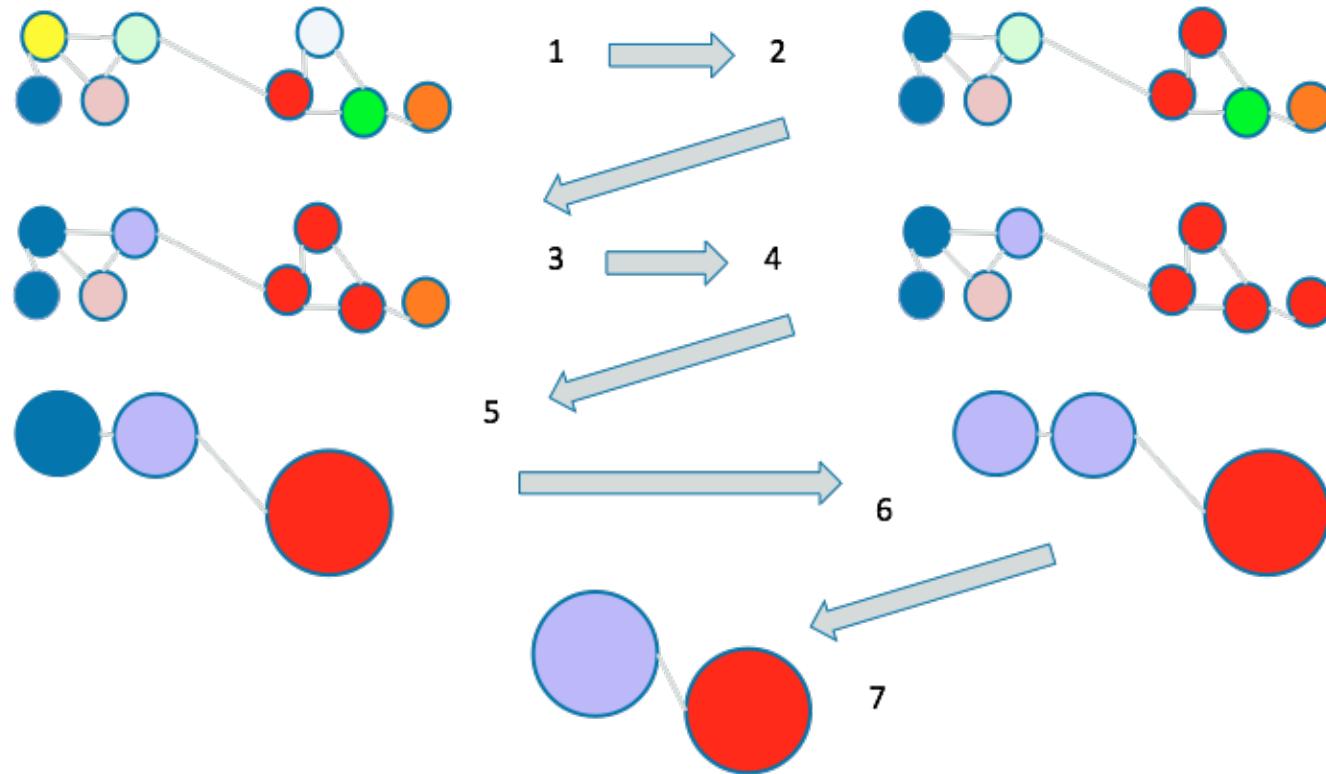
Defining Clusters Through Graphs

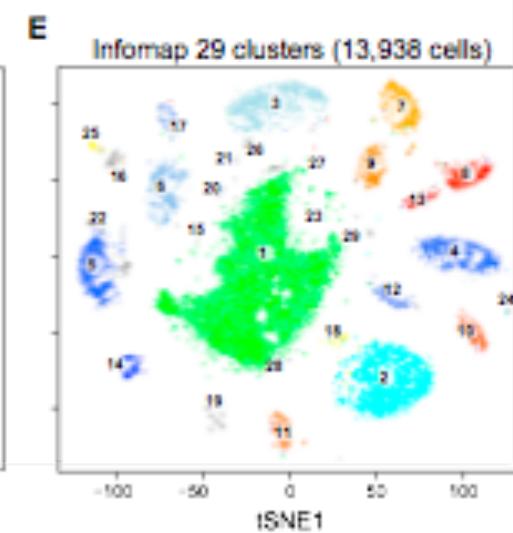
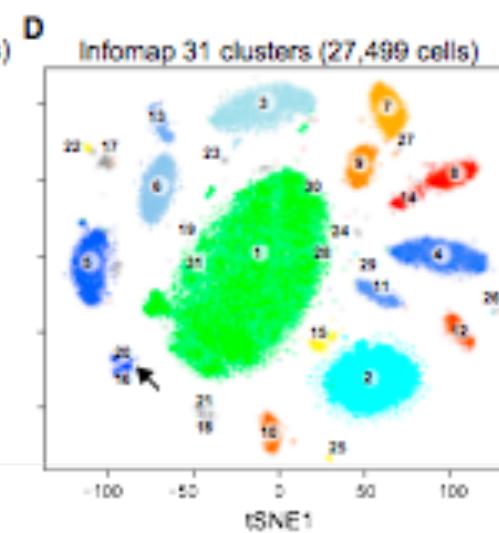
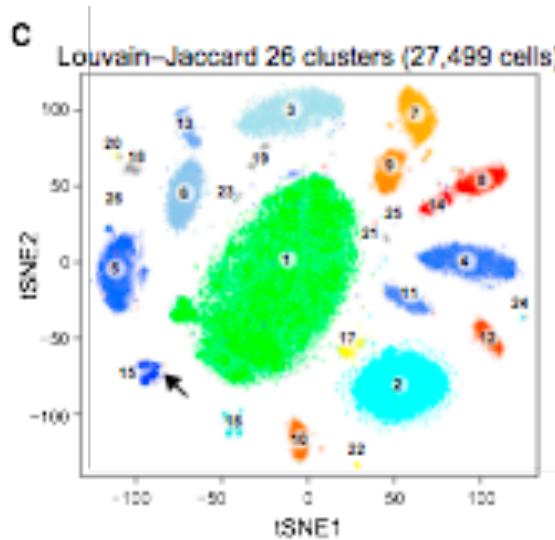
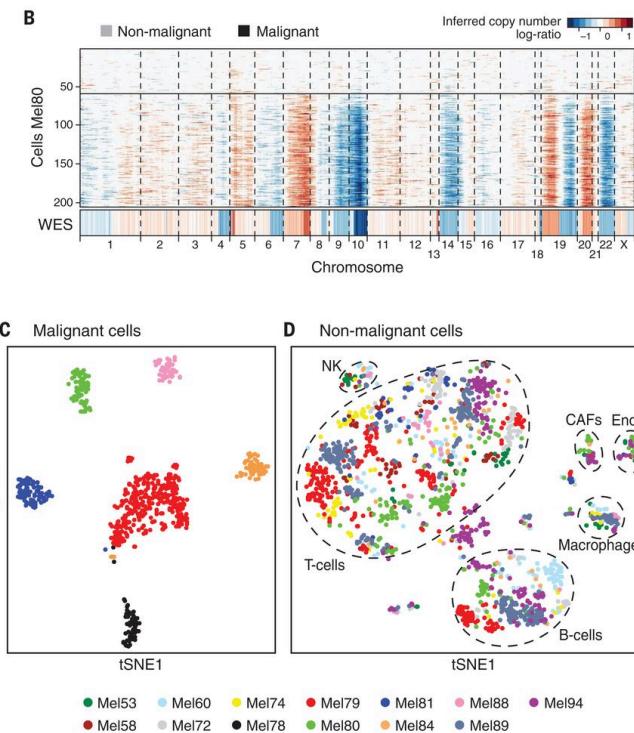
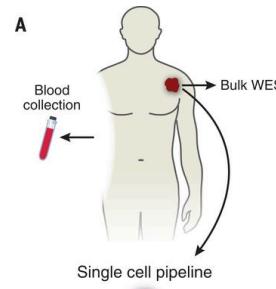


- Smart Local Moving (SLM) algorithm for community (cluster) detection in large networks.
 - Can be applied to 10s of millions cells, 100s of millions of relationships.
 - Evolved from the Louvain algorithm

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

Local Moving Heuristic

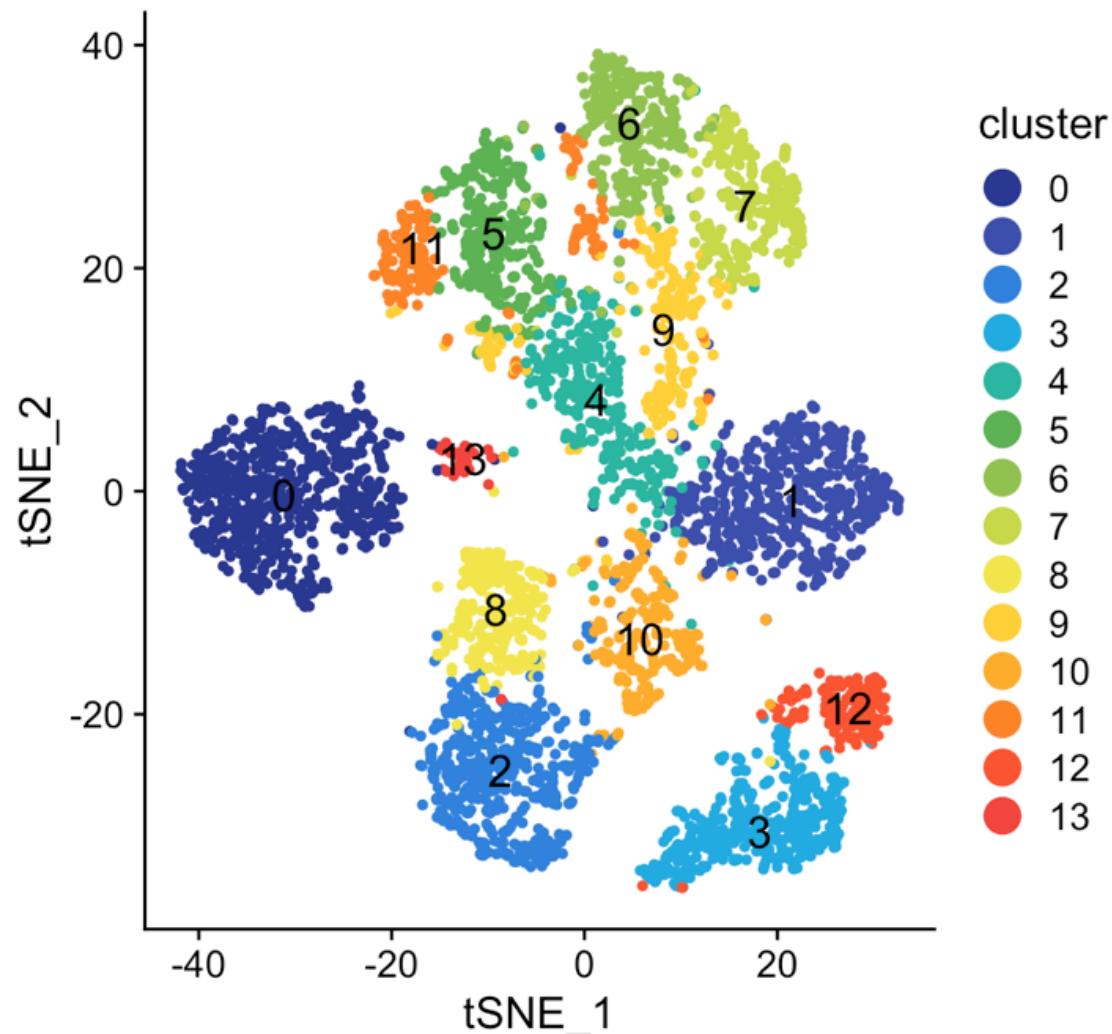




Tirosh and Izar et al. Science 2016
Shekhar et al. Cell 2016

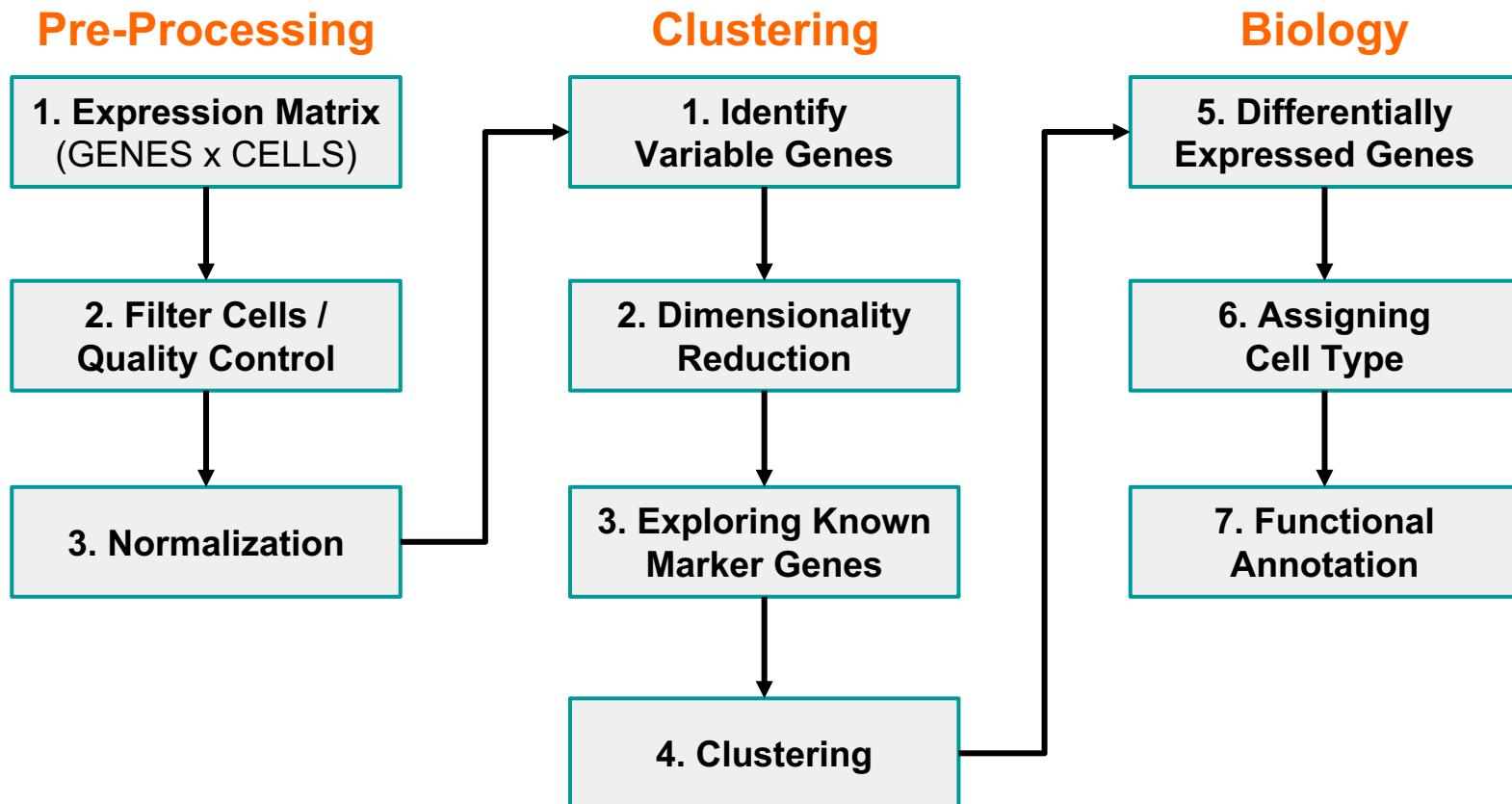
Determining cell type, state, and/or function:

3. Visualization

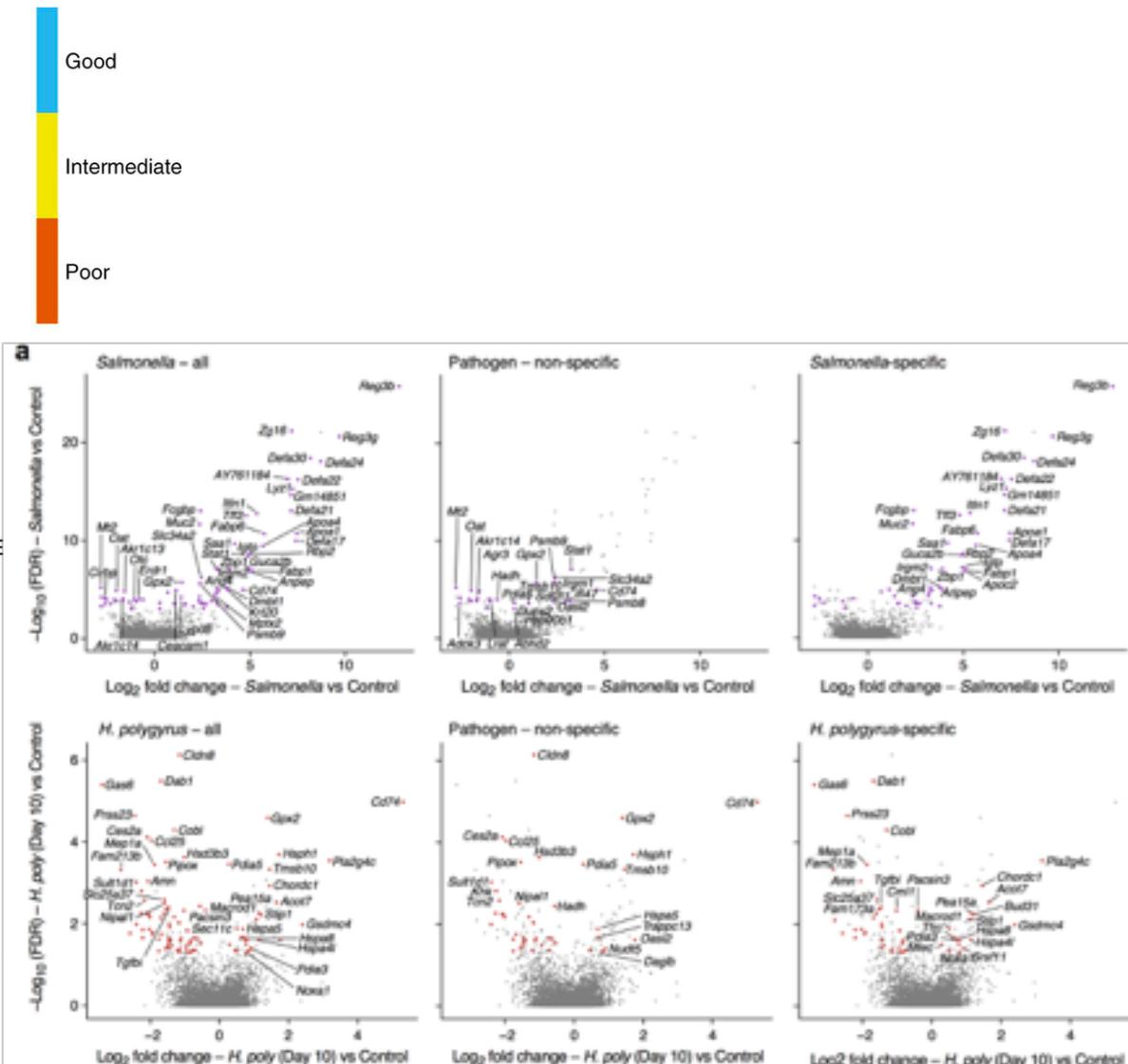
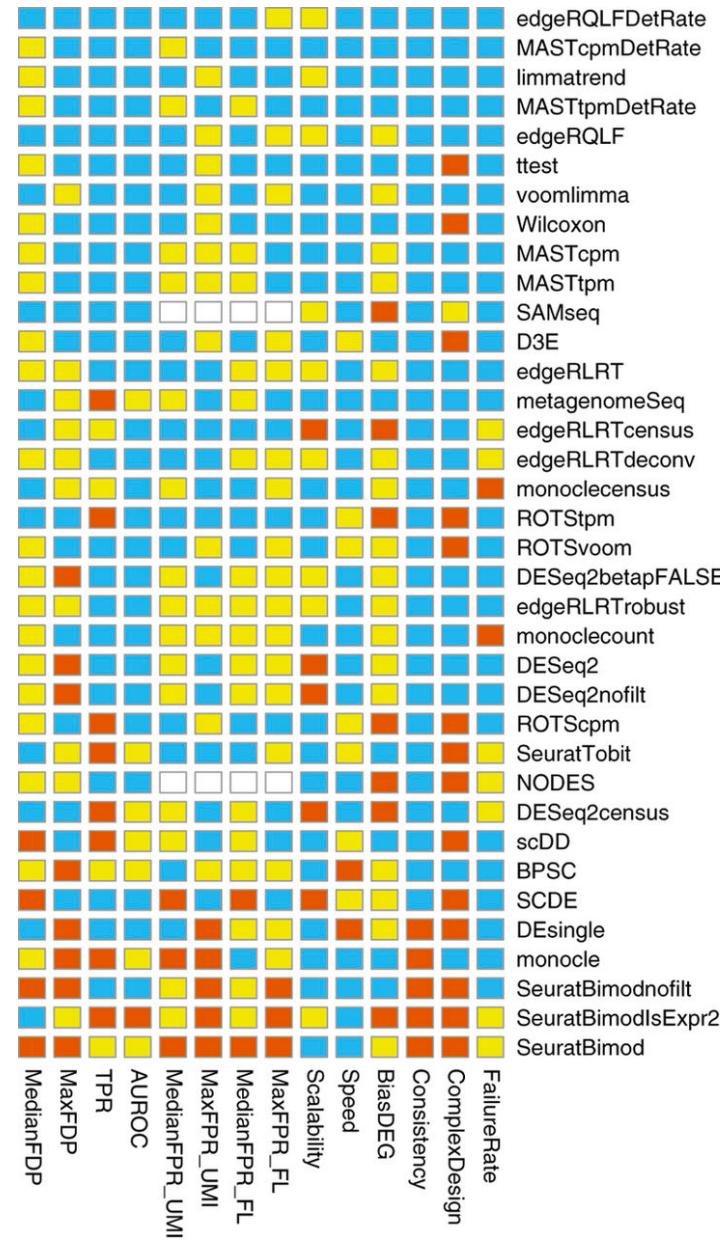


A great tSNE resource! <https://distill.pub/2016/misread-tsne/>

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

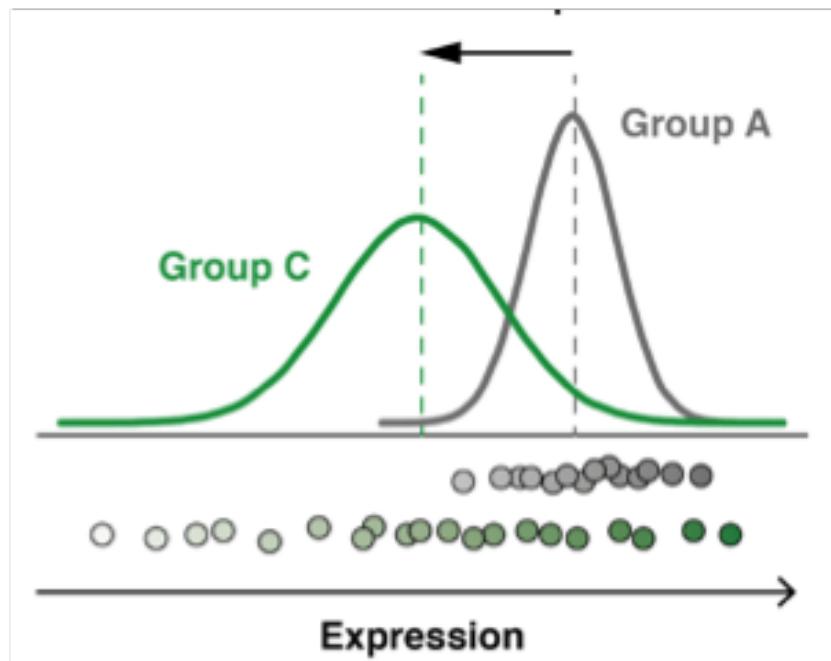


5. Assigning cell identity & comparing across conditions: Differential Expression Analysis



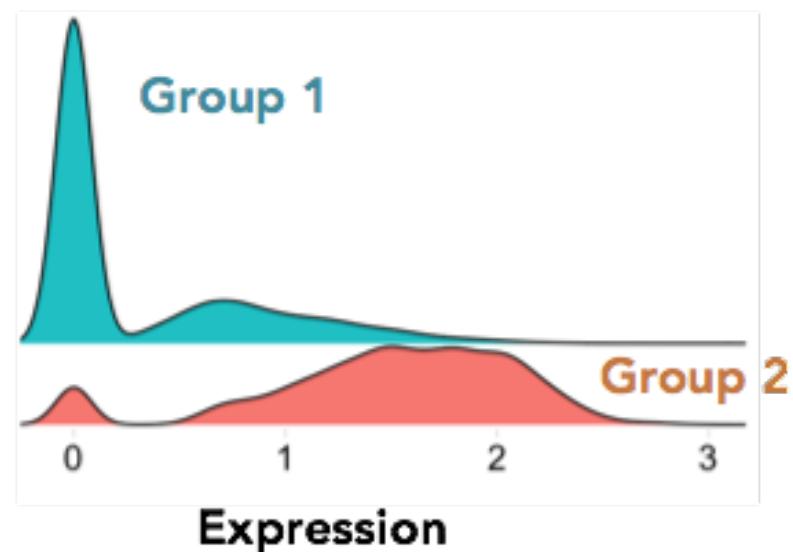
Determining cell type, state, and/or function:

5. Identifying differentially expressed genes



Bulk

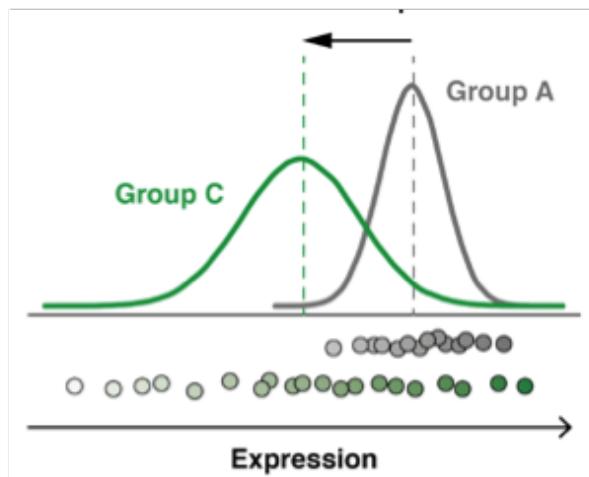
"Zero inflation" poses a challenge in single-cell data!



Single cell

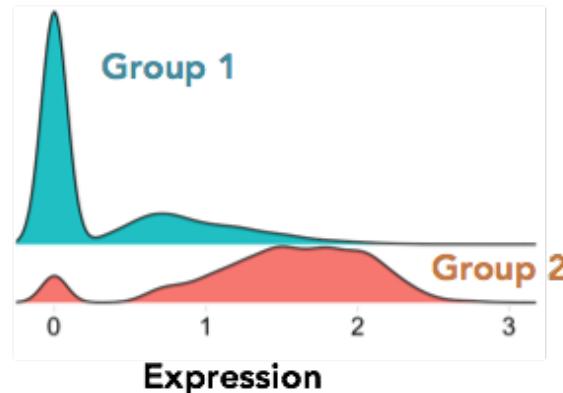
Differential Expression

Group A > Group B ($p\text{-value} < 0.01$)



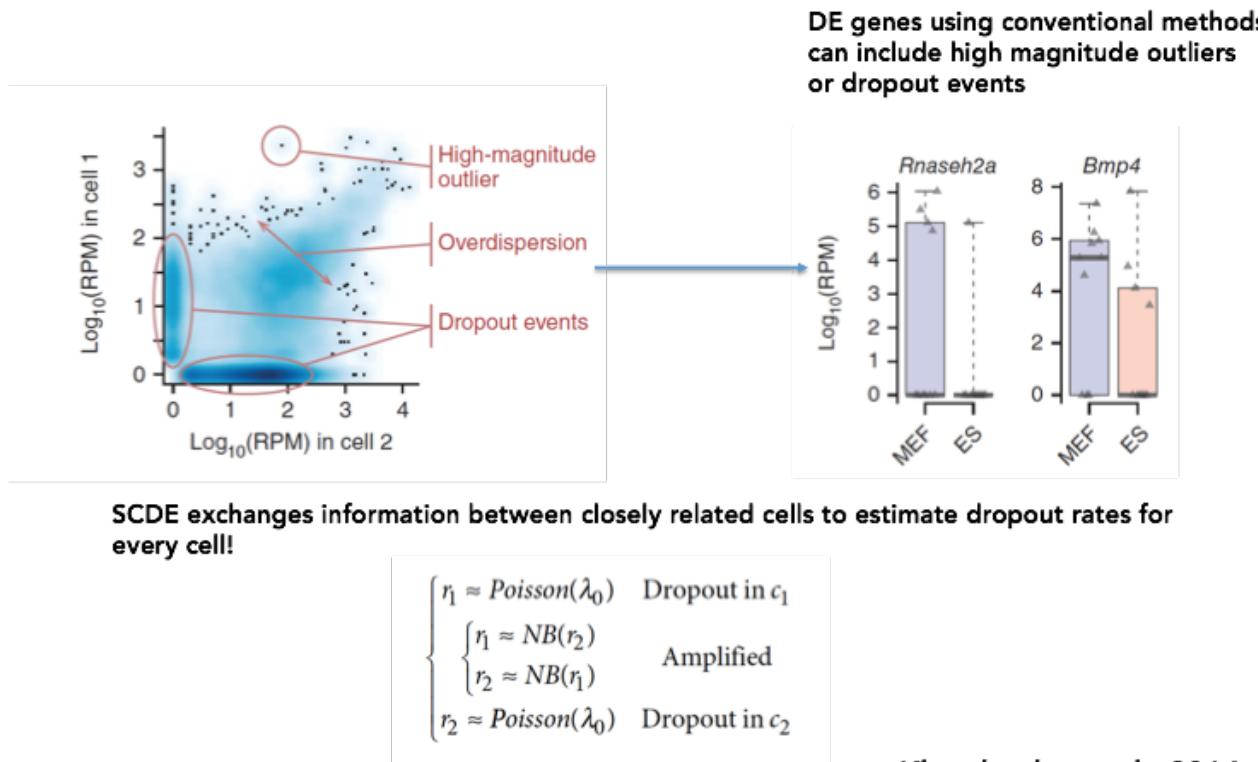
BUT

"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)



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) 16:276
DOI 10.1186/s13059-015-0844-5

Genome Biology

METHOD

Open Access



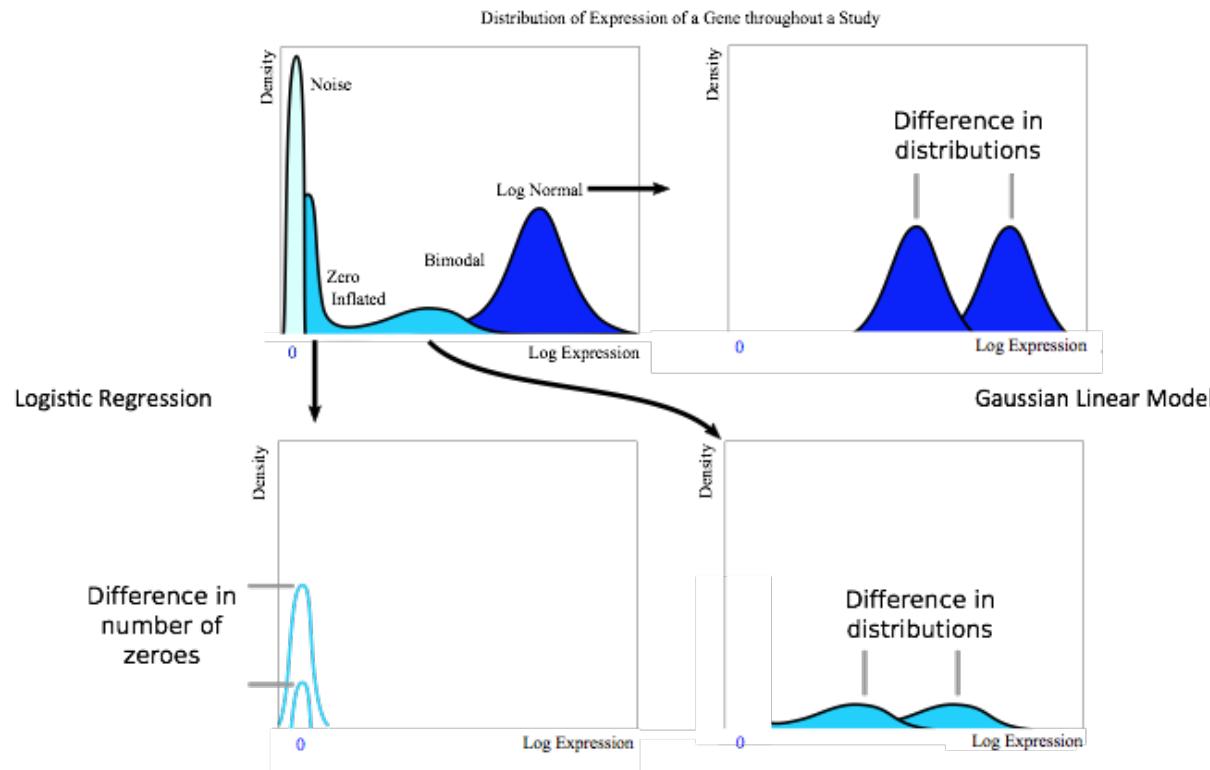
MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data

Greg Finak^{1†}, Andrew McDavid^{1†}, Masanao Yajima^{1†}, Jingyuan Deng¹, Vivian Gensuk², Alex K. Shalek^{3,4,5,6}, Chloe K. Slichter³, Hannah W. Miller³, M. Julianne McElrath⁷, Martin Prlic¹, Peter S. Linsley² 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

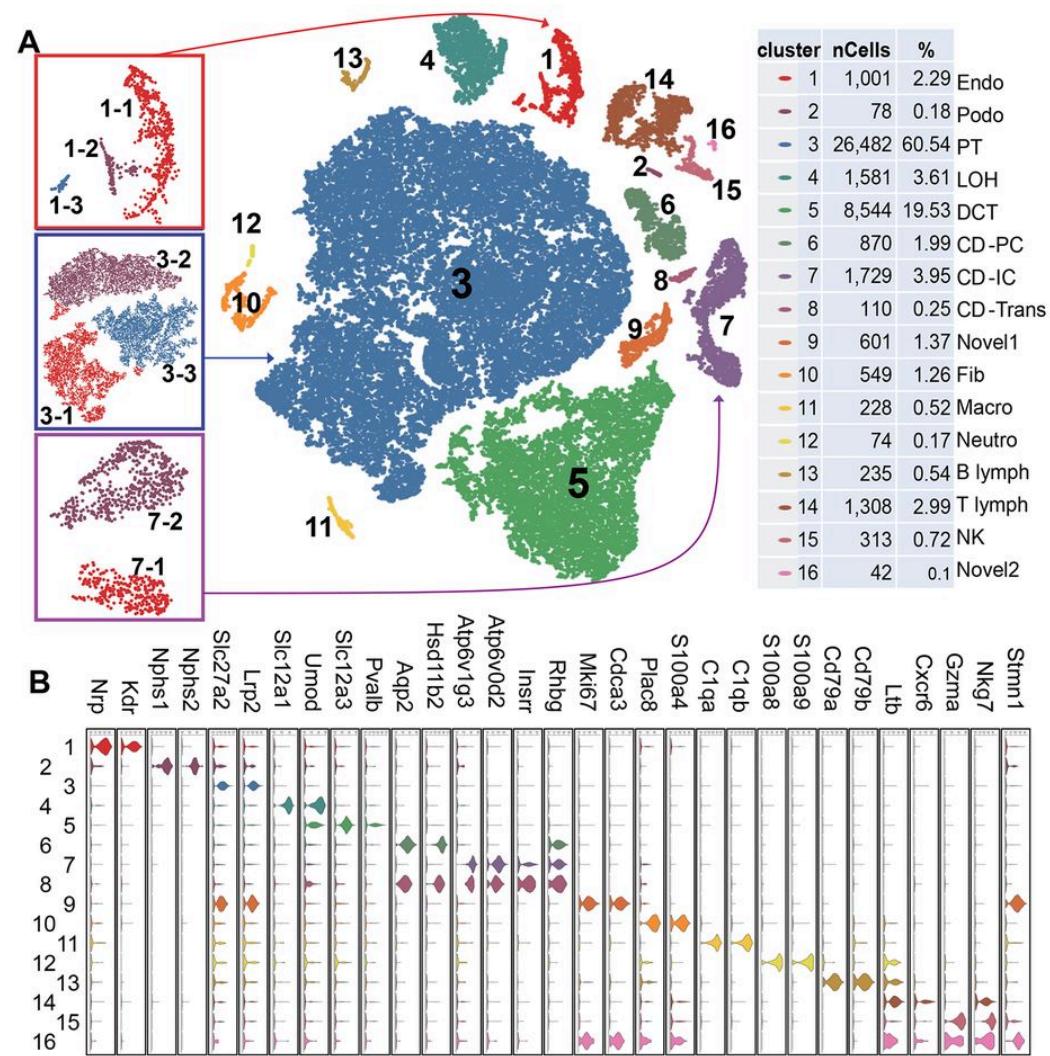
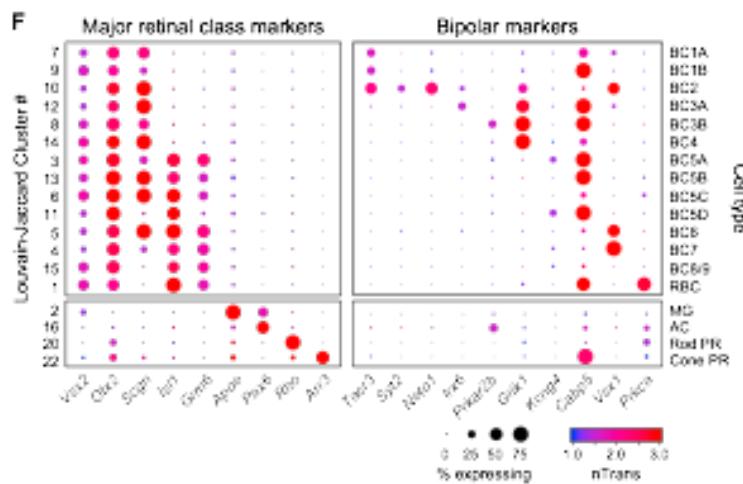
- Tobit regression on a smoothed data.

MAST

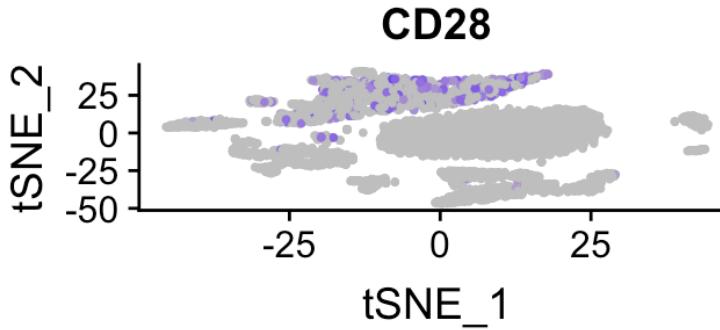
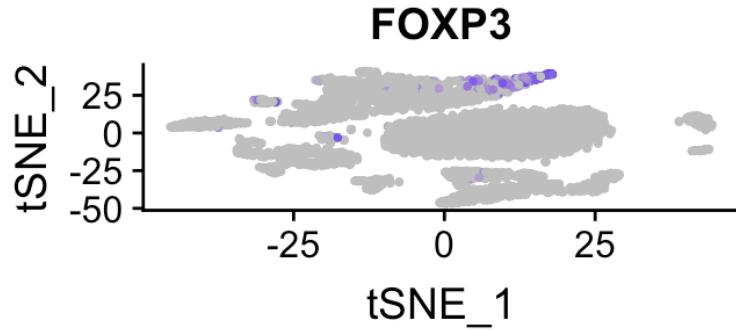
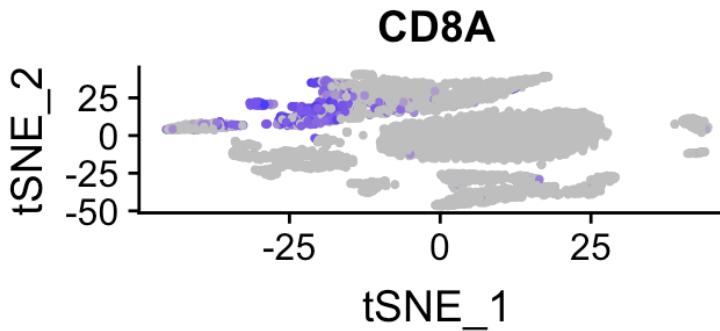
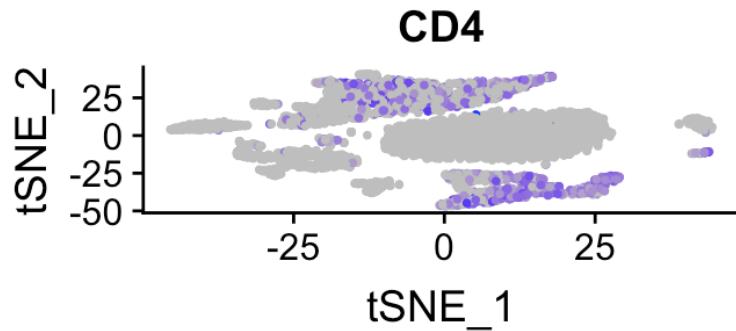
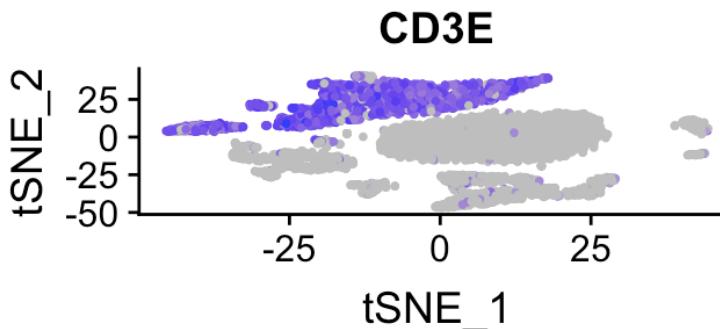
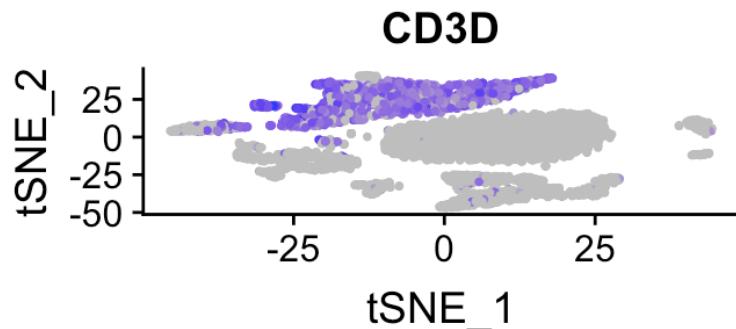
- Hurdle model for zero inflated data

....

6. Assigning cell identity: Known marker genes

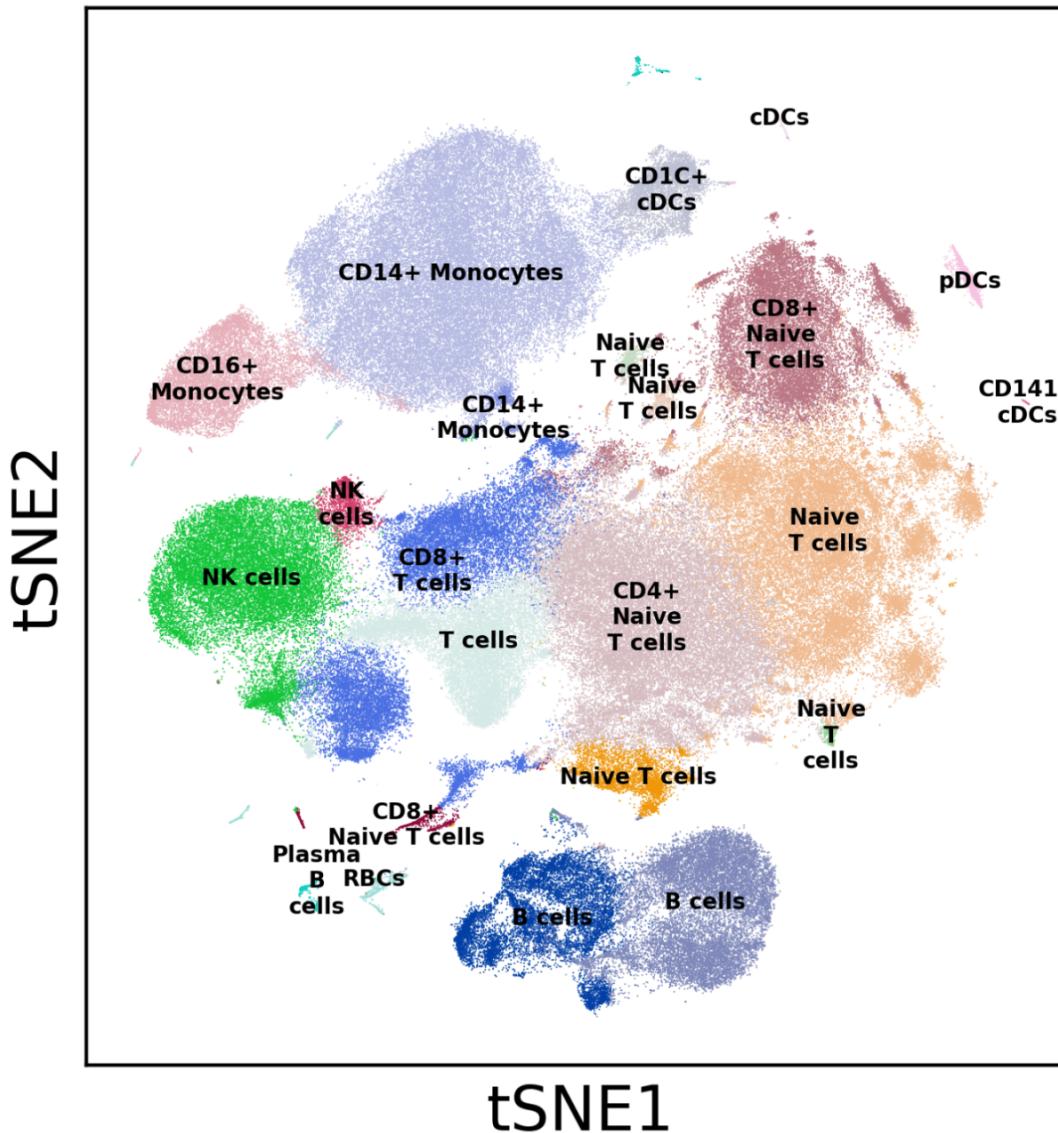


Determining cell type, state, and/or function: Exploring expression of marker genes



Determining cell type, state, and/or function:

6. Assigning cell type



Visualizing genes of interest

Dot plots, violin plots, feature plots

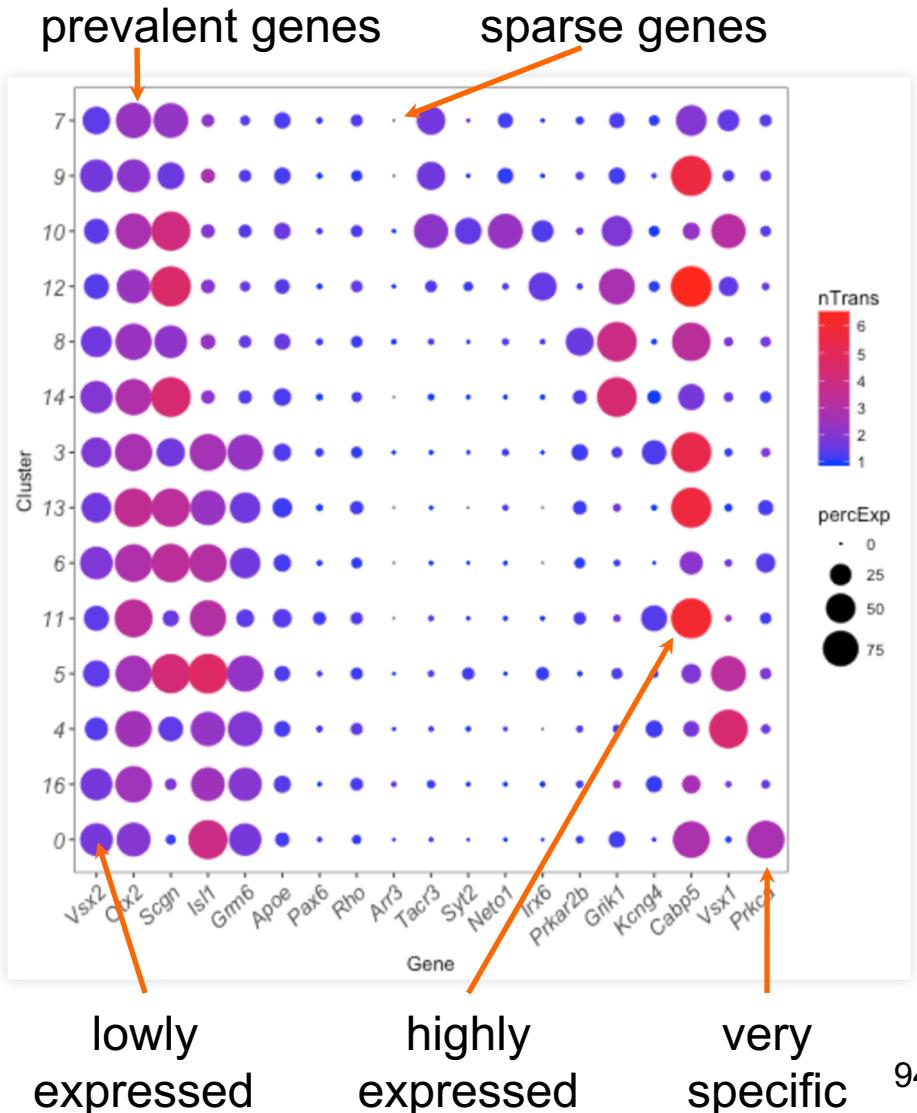
Size of circle

- Gene prevalence in cluster

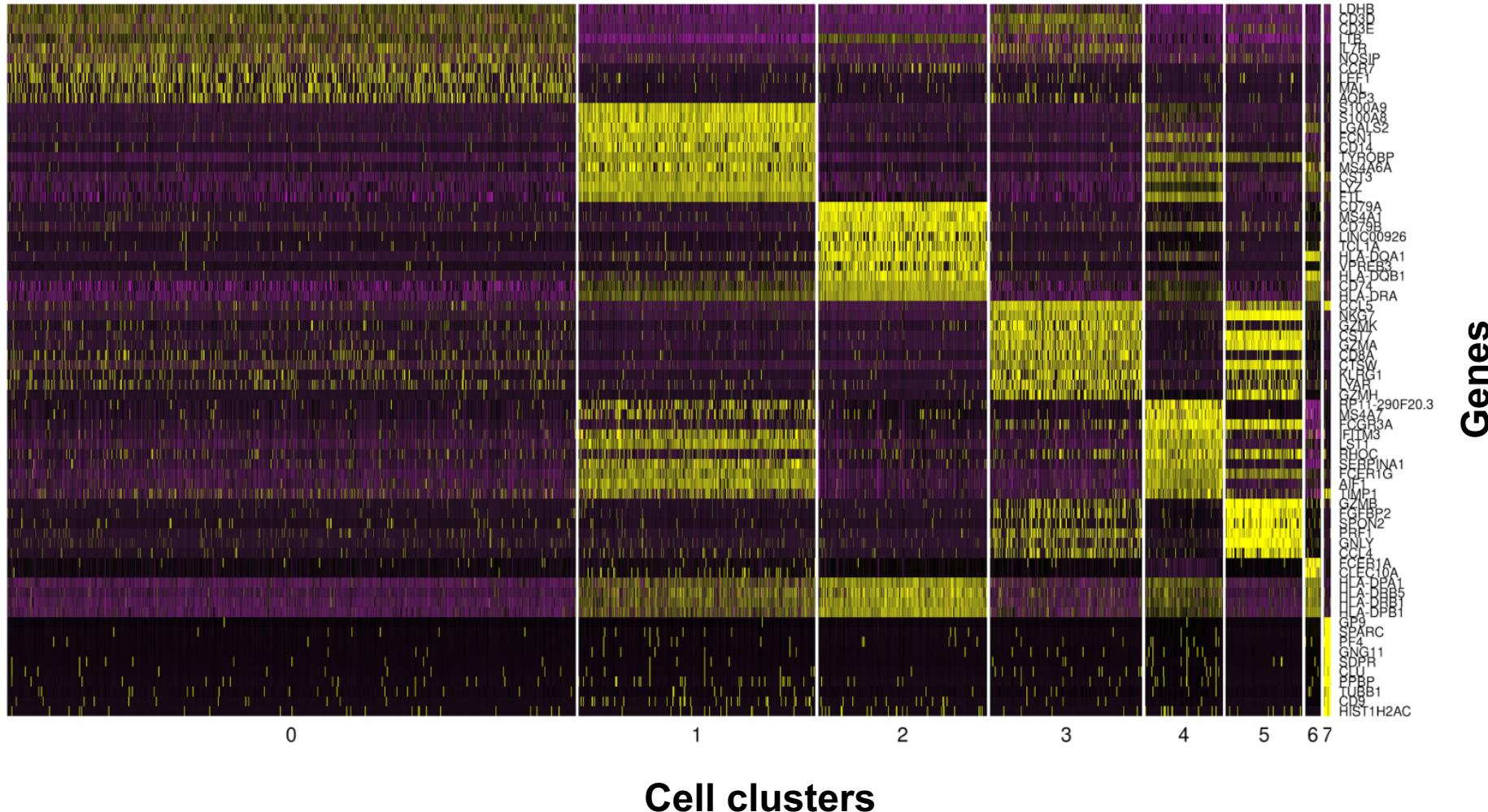
Color of circle

- More red, more expressed in cluster

Scales well with many cells

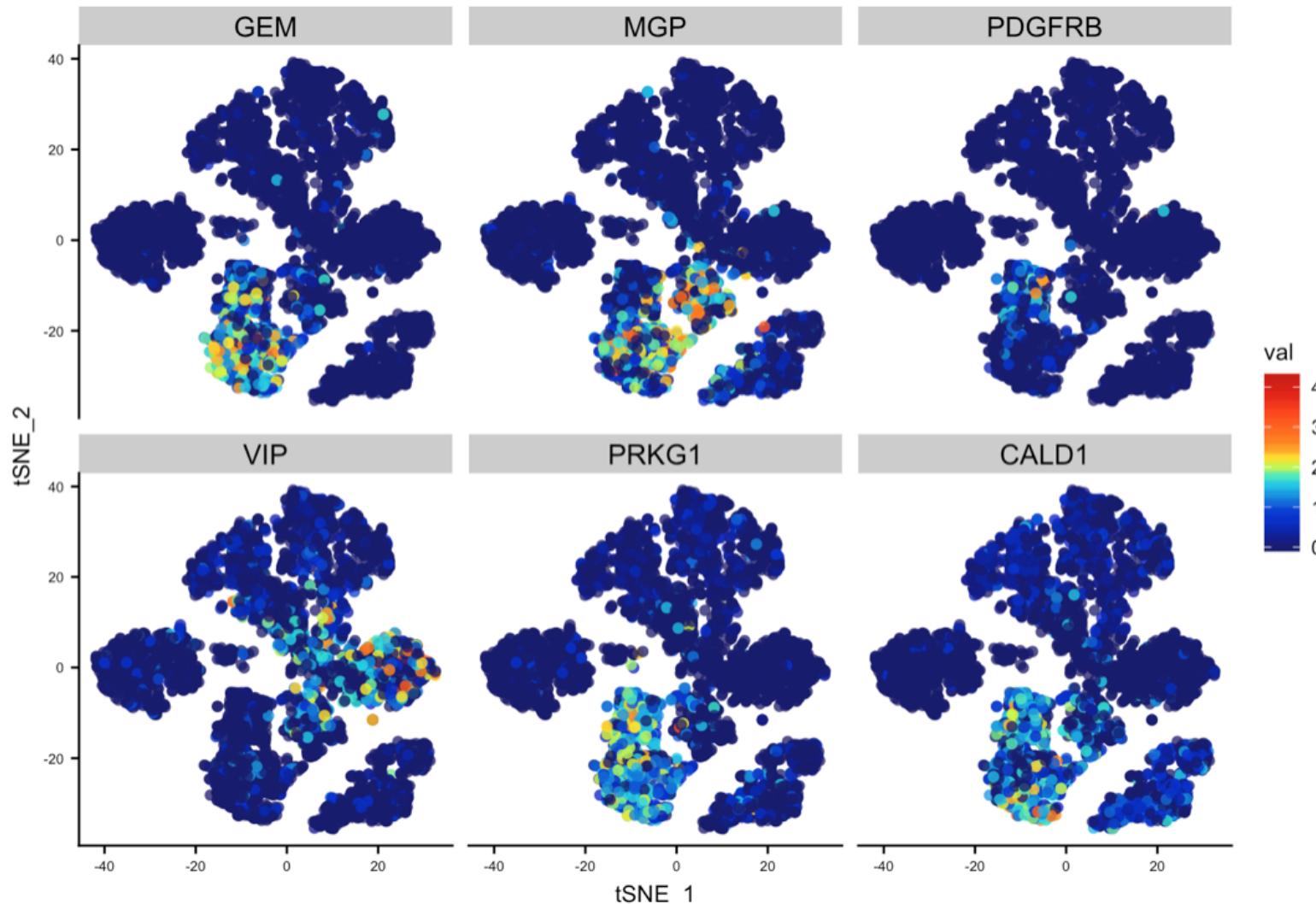


Determining cell type, state, and/or function: .Identifying differentially expressed genes



Visualizing genes of interest

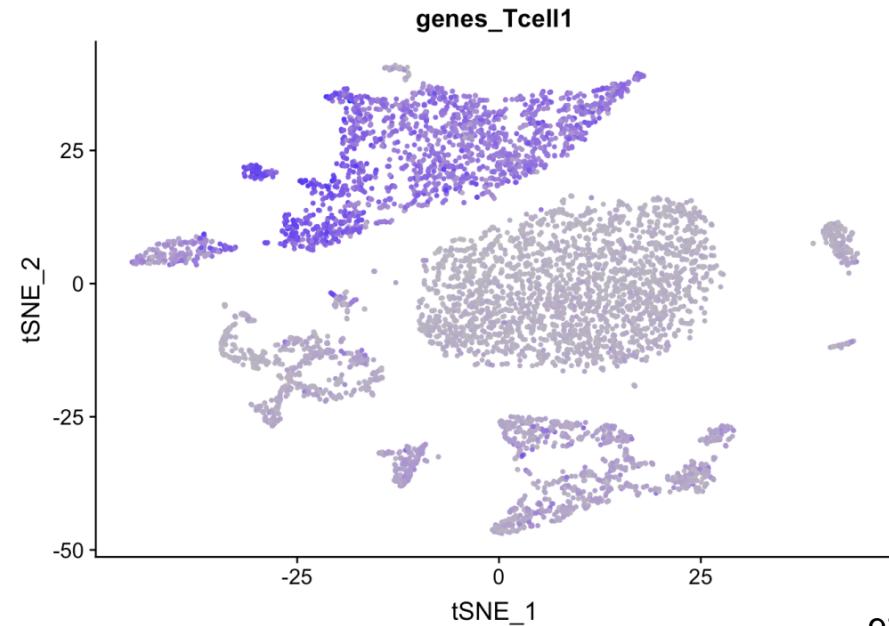
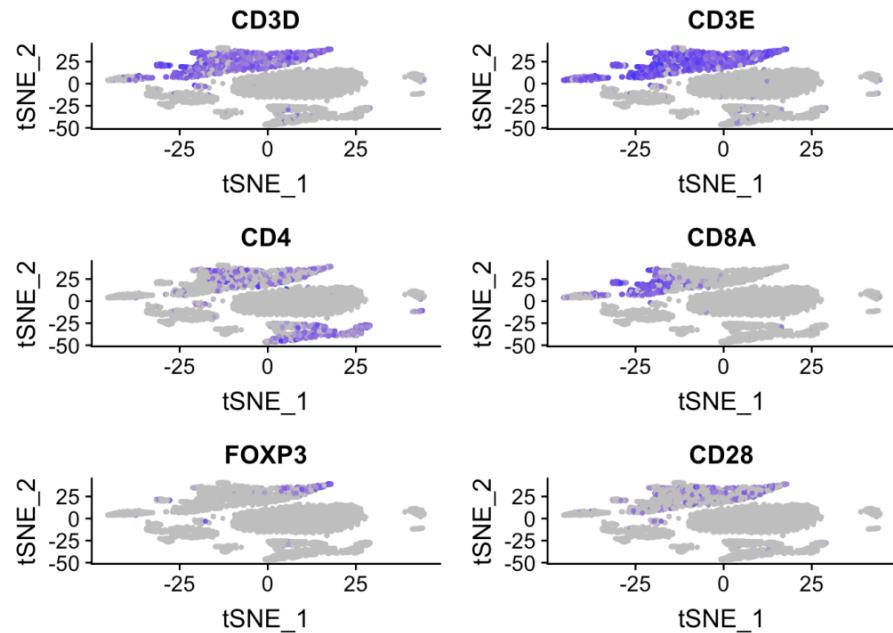
Dot plots, violin plots, feature plots



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

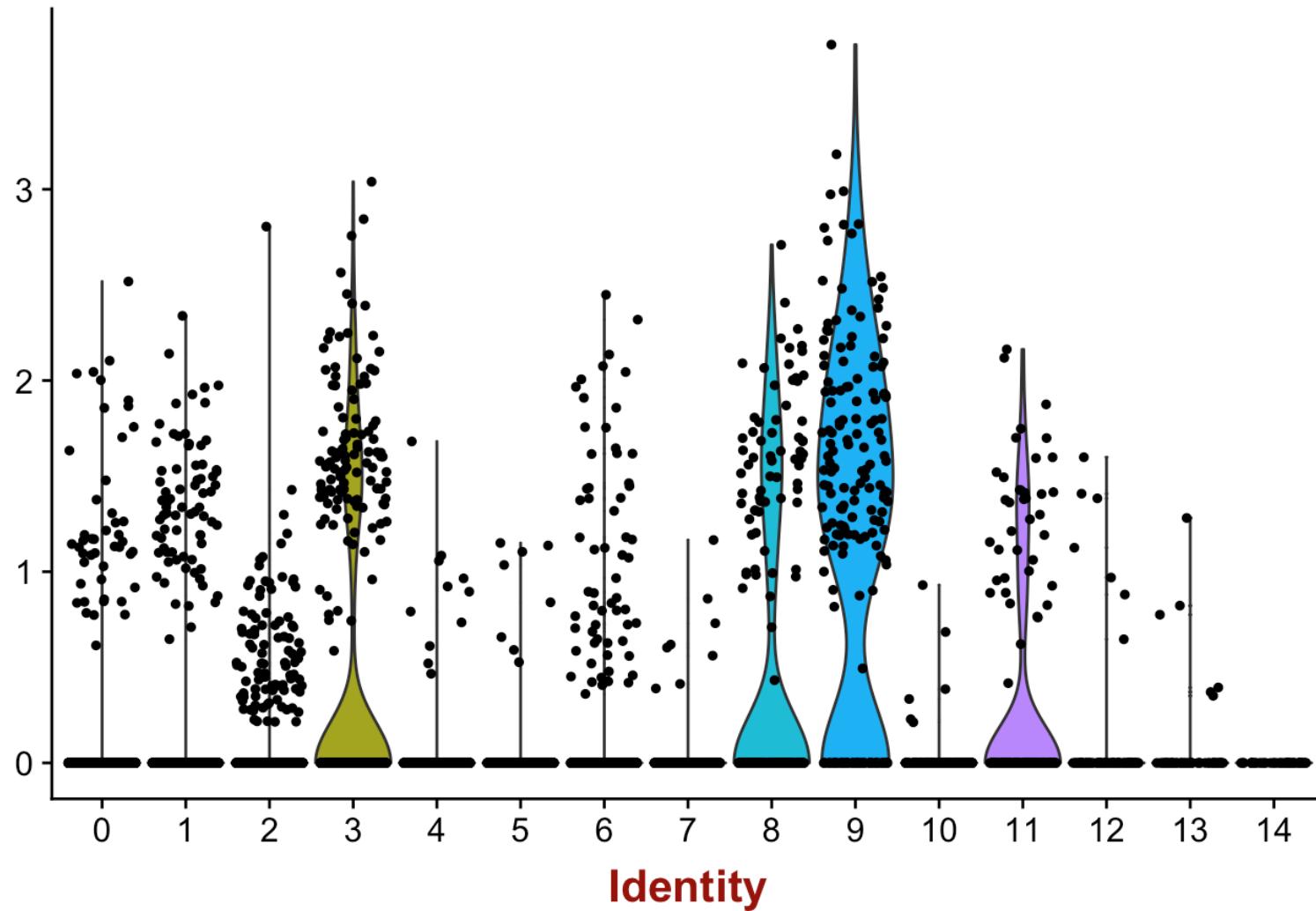
- Can visualize a score for each cell and look at multiple genes at once
- Done for a gene expression program of interest, e.g, cell-cycle, inflammation, cell type, dissociation
- Reduces the effects of dropouts

Gene signature for T cells

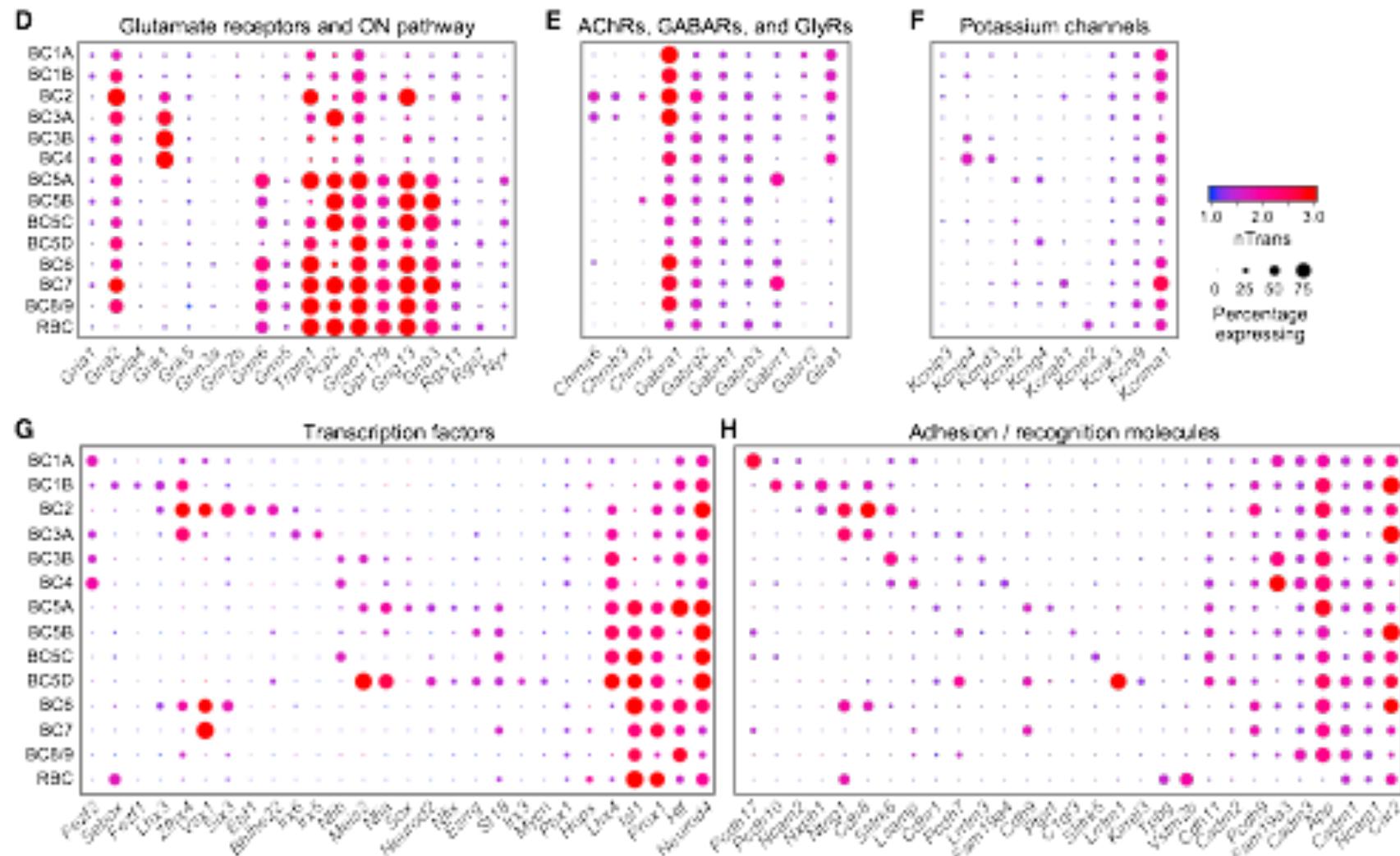


Visualizing genes of interest

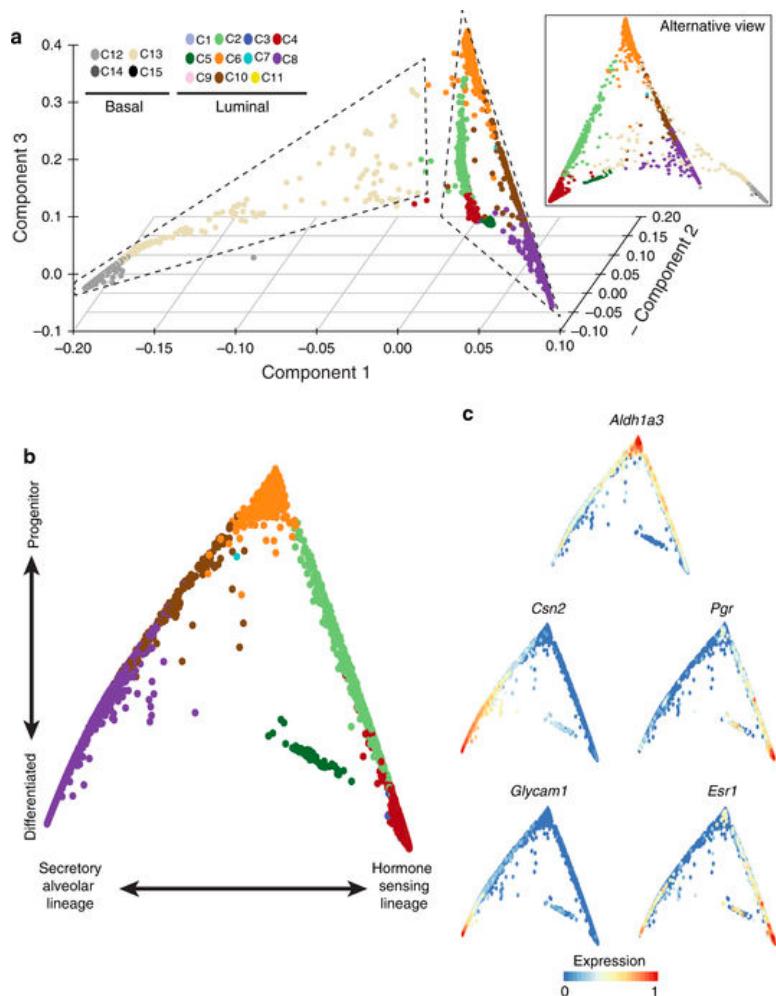
Dot plots, **violin plots**, feature plots



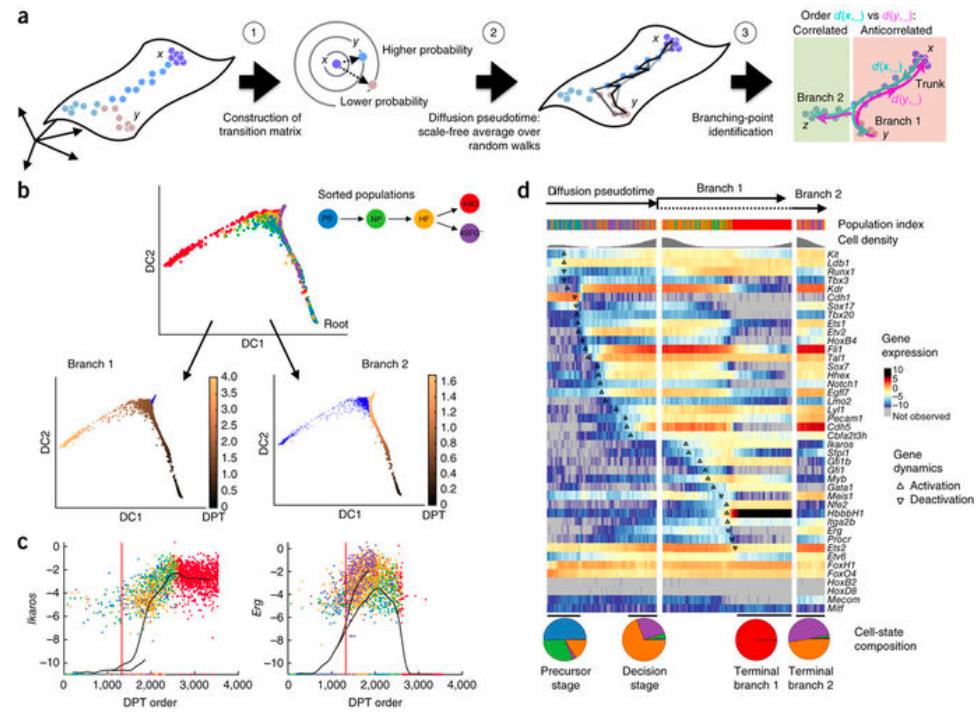
7. Functional annotation by pathway analysis and gene-set enrichment analysis



Trajectory inference



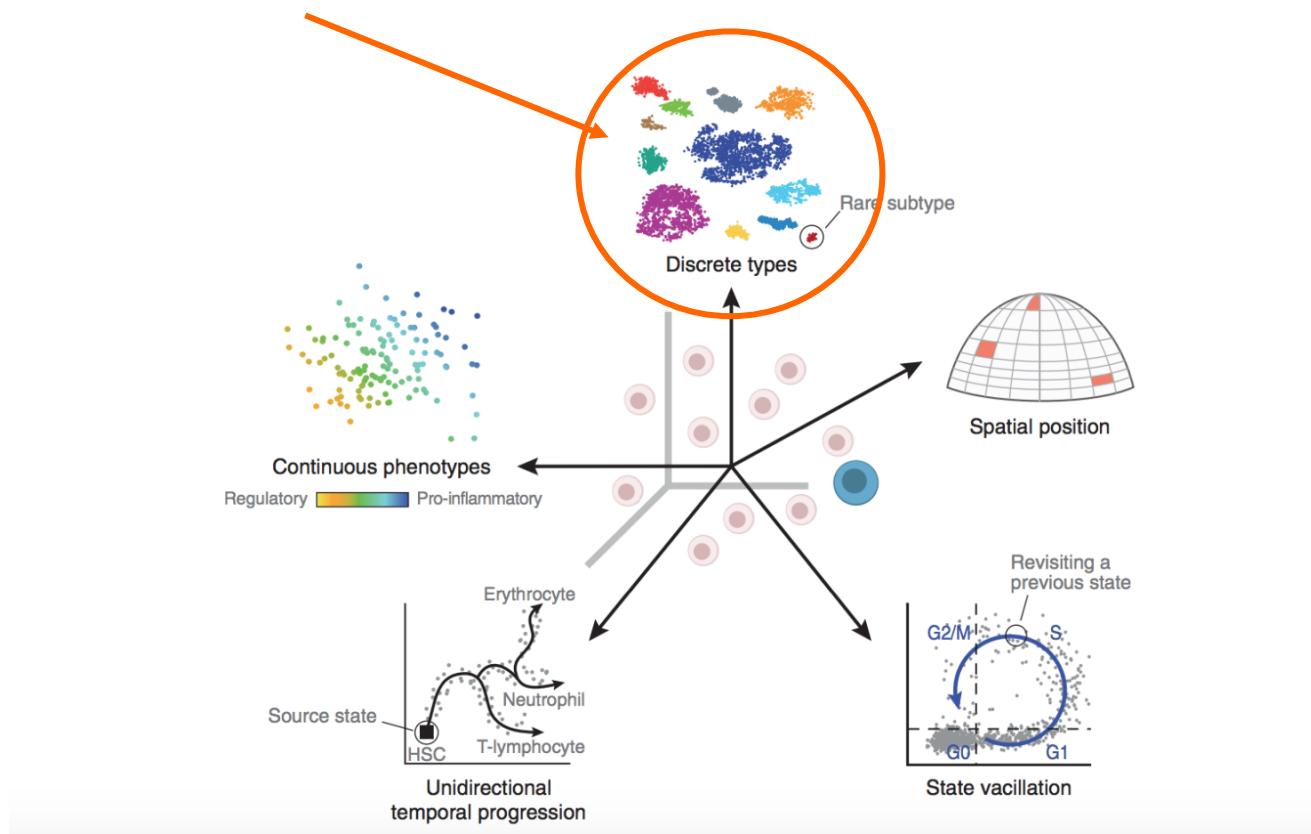
Diffusion Maps



Diffusion pseudotime

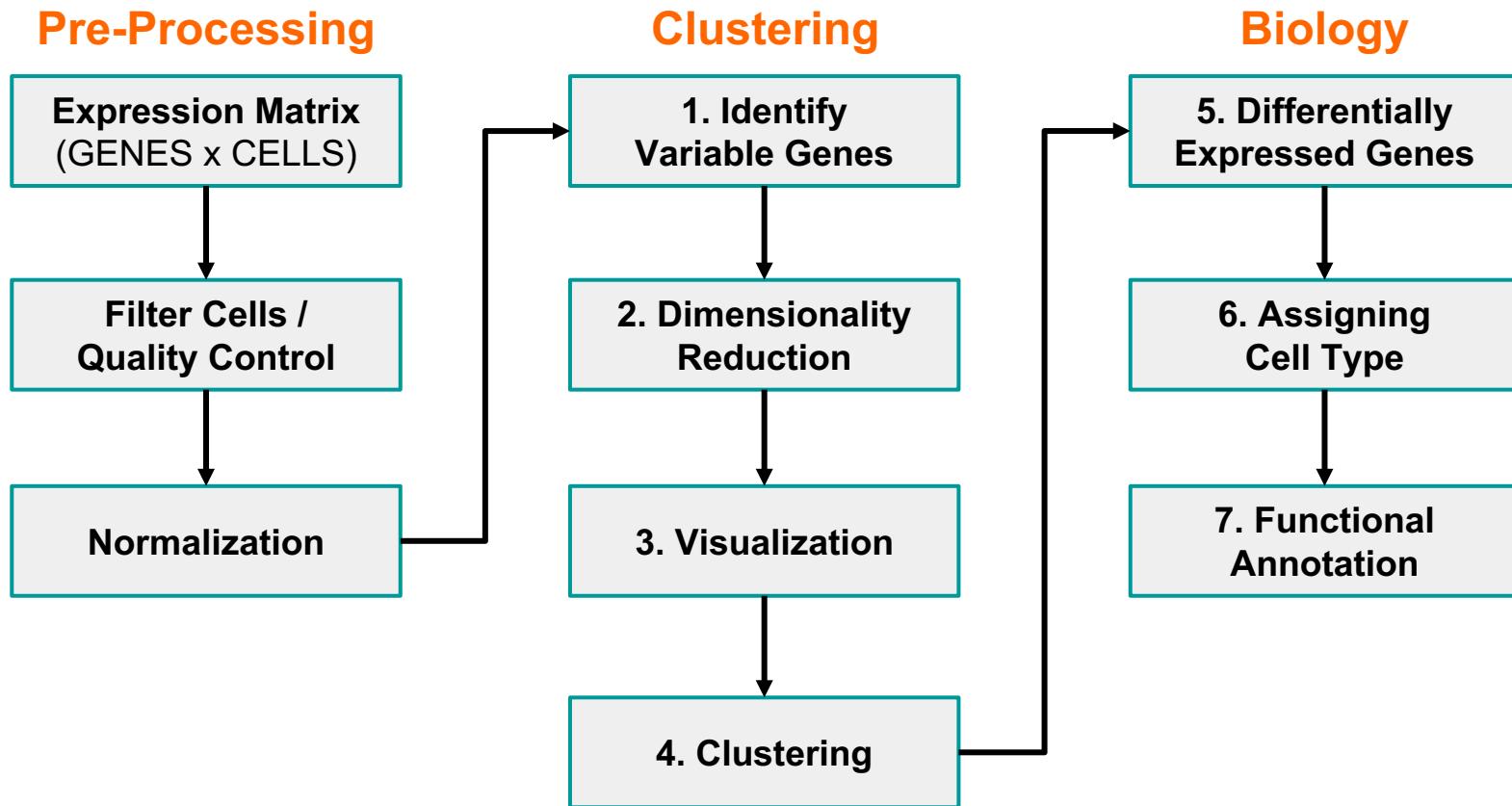
Recap: what did we just cover?

We covered just this.
So much more to learn!



Recap: what did we just cover?

Time to execute this pipeline in a hands on example!



Tools and resources

RStudio: integrated development environment for R

The image shows a screenshot of the RStudio integrated development environment (IDE). The interface is divided into several panes:

- Scripts:** A large pane on the left containing an R script named "Untitled1". The code in the script is as follows:

```
1 ## This is where we make scripts.  
2 log(1)  
3  
4 ### The bottom left is an interactive R session.  
5  
6 ### The upper right shows what is in memory and allows data importing.  
7 ##### Some output is found here as well.  
8  
9 ### The bottom right is where help is placed.  
10 ##### You can also see your working directory.  
11 ##### When you create pictures and do not save them, they go here.|
```

- Environment:** A pane in the top right showing the global environment. It contains a single entry "x" with the value "4".
- R Session:** A large pane at the bottom left showing the R console output. It includes the R startup message, information about the license, natural language support, and a workspace summary.

```
11:56 (Top Level) :  
Copyright (C) 2015 The R Foundation for Statistical Computing  
Platform: x86_64-apple-darwin13.4.0 (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
Natural language support but running in an English locale  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
[Workspace loaded from ~/RData]
```

```
> 1+2+3  
[1] 6  
> log(1)  
[1] 0  
> x=4  
> |
```

- Help:** A large pane on the right showing the documentation for the "log" function. It includes the function's name, description, usage, and examples.

```
R: Logarithms and Exponentials < Find in Topic  
  
log (base)  
  
Logarithms and Exponentials  
  
Description  
  
log computes logarithms, by default natural logarithms, log10 computes common (i.e., base 10) logarithms, and log2 computes binary (i.e., base 2) logarithms. The general form log(x, base) computes logarithms with base base.  
  
log1p(x) computes log(1+x) accurately also for |x| << 1.  
  
exp computes the exponential function.  
  
expm1(x) computes exp(x) - 1 accurately also for |x| << 1.  
  
Usage  
  
log(x, base = exp(1))  
logb(x, base = exp(1))  
log10(x)  
log2(x)  
  
log1p(x)
```



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.





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 Cells¹

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](#)
Fastq and jsons. Additional instructions below.

Ischaemic Sensitivity of Human Tissue²

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

DOWNLOADS

- [Metadata Spreadsheet](#)
- [Primary Data Download Script](#)
Fastq and jsons. Additional instructions below.

Melanoma Infiltration of Stromal and Immune Cells³

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

DOWNLOADS

- [Metadata Spreadsheet](#)
- [Primary Data Download Script](#)
Fastq 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

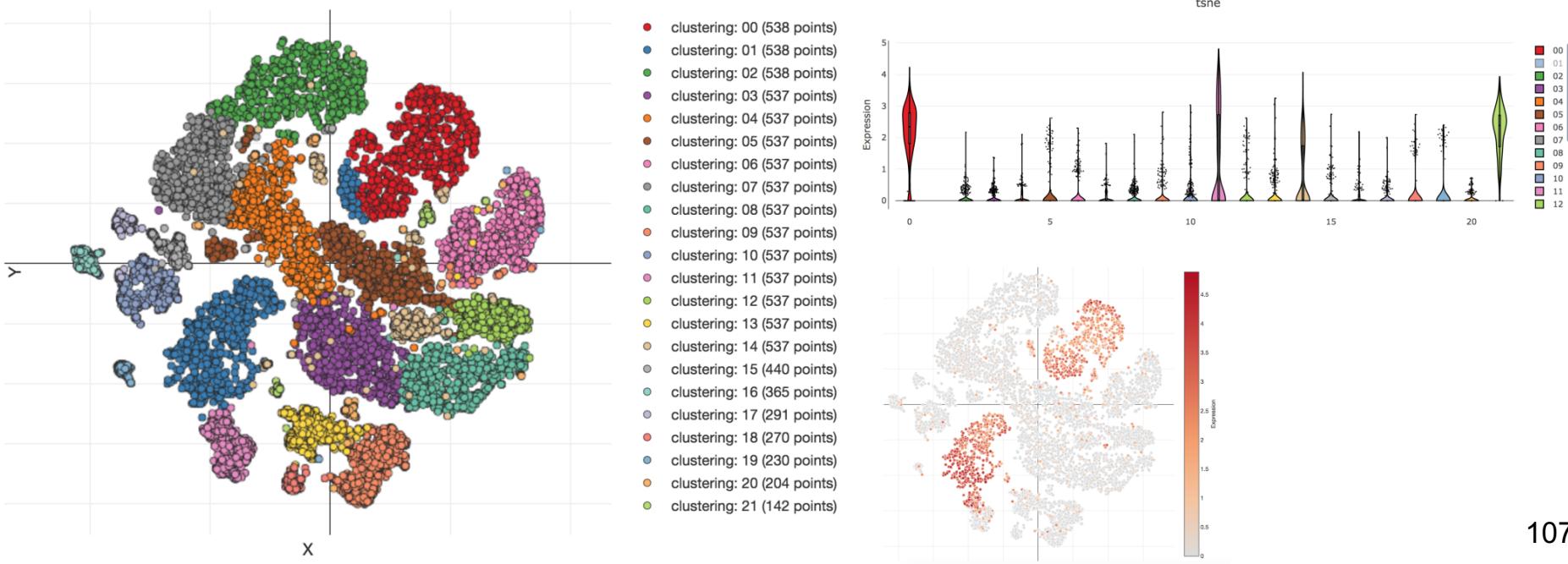
Single Cell Portal BETA

? Help ▾ Sign In

Single Cell Portal BETA

Visualization portal for single cell RNA-seq data.

Now featuring **49** studies with **542,796** cells.



Broad Institute Single Cell Portal

Secure | https://portals.broadinstitute.org/single_cell

Single Cell Portal **BETA** Help Sign In

Single Cell Portal **BETA**

Visualization portal for single cell RNA-seq data.

Now featuring **37** studies with **432,801** cells.

Browse Studies ?

Search Studies...

Most Recent Most Popular Reset Filters

Single nucleus RNA-seq of cell diversity in the adult mouse hippocampus (sNuc-Seq) ▾

View Study

Single nucleus RNA-seq of cell diversity in the adult mouse hippocampus. Habib N, Li Y, Heidenreich M, Swiech L, Avraham-David I, Trombetta J, Hession C, Zhang F, Regev A. Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science*. 28 Jul 2016 DOI: 10.1126/science.aad7038 Contact: naomi@broadinstitute.org Single cell RNA-Seq provides rich information about cell types and states. However, it is difficult to capture rare dynamic processes, such as adult neurogenesis, because isolation of rare neurons from adult tissue is challenging and markers for each phase are limited. Here, we develop Div-Seq, which combines scalable single-nucleus RNA-Seq (sNuc-Seq) with pulse labeling of proliferating cells by EdU... (continued)

Retinal Bipolar Neuron Drop-seq ▾

View Study

Retinal Bipolar Neuron Drop-Seq Karthik Shekhar, Sylvain W. Lapan, Irene E. Whitney, Nicholas M. Tran, Evan Z. Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z. Levin, James Nemesh, Melissa Goldman, Steven A. McCarroll, Constance L. Cepko, Aviv Regev, Joshua R. Sanes. Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell*. Volume 166, Issue 5, p1308–1323.e30, 25 August 2016. DOI: http://dx.doi.org/10.1016/j.cell.2016.07.054 Contact: Karthik Shekhar at karthik@broadinstitute.org Patterns of gene expression can be used to characterize and classify neuronal types. It is challenging, however, to generate taxonomies that fulfill the essential criteria of being comprehensive, harmonizing with conventional classification schemes, and lacking superfluous subdivisions of genuine types. To address these challenges, we used massively parallel single-cell RNA profiling and optimized computational methods on a heterogeneous class of... (continued)

10X LucOS ▾

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=RJVL80Gg3IA>
- 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**](https://github.com/seandavi/awesome-single-cell)

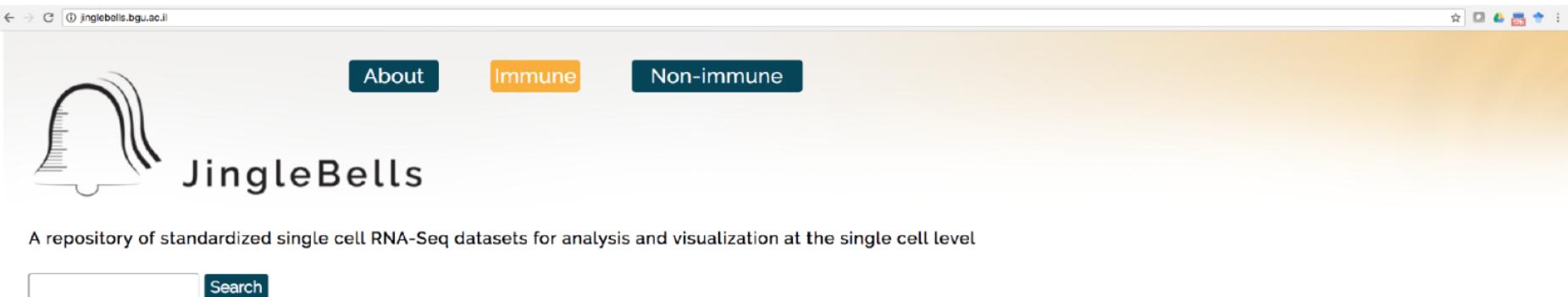
The screenshot shows the GitHub repository page for 'seandavi / awesome-single-cell'. At the top, there are navigation links for Personal, Open source, Business, Explore, Pricing, Blog, and Support. On the right, there are buttons for 'Sign in' and 'Sign up'. Below the header, the repository name 'seandavi / awesome-single-cell' is displayed, along with a 'Watch' button (25), a 'Star' button (86), and a 'Fork' button. A navigation bar below the repository name includes 'Code' (selected), 'Issues 0', 'Pull requests 0', 'Projects 0', 'Pulse', and 'Graphs'. The main content area contains the text: 'List of software packages for single-cell data analysis, including RNA-seq, ATAC-seq, etc.'

[**www.singlecellnetwork.org**](http://www.singlecellnetwork.org)

The screenshot shows the homepage of the Single Cell Network. It features a logo with the letters 'SCN' inside a blue circle. The text 'Single Cell Network' is prominently displayed, followed by the tagline 'Connecting people. Advancing science.' To the right, there is a 'Welcome, Guest' link, a 'Join' button, and a 'Log In' button. A search bar is located at the bottom right.

Resources, cont.

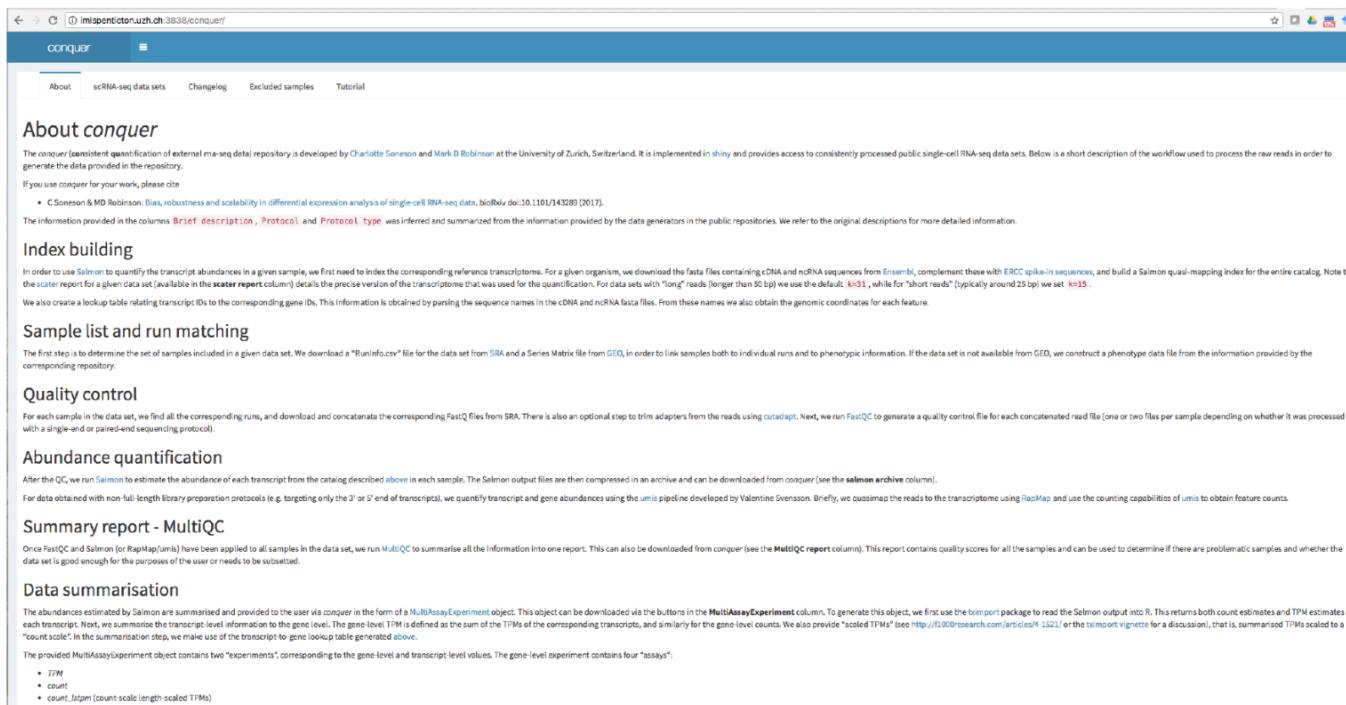
Data repositories: JingleBells



A repository of standardized single cell RNA-Seq datasets for analysis and visualization at the single cell level

Search

Data repositories: Conquer



The *conquer* (consistent quantification of external m-a-seq data) repository is developed by Charlotte Soneson and Mark D Robinson at the University of Zurich, Switzerland. It is implemented in *shiny* and provides access to consistently processed public single-cell RNA-seq data sets. Below is a short description of the workflow used to process the raw reads in order to generate the data provided in the repository.

If you use *conquer* for your work, please cite

- C Soneson & MD Robinson: Bias, robustness and scalability in differential expression analysis of single-cell RNA-seq data. *BioRxiv* doi:10.1101/143289 (2017).

The information provided in the columns *Brief description*, *Protocol* and *Protocol type* was inferred and summarized from the information provided by the data generators in the public repositories. We refer to the original descriptions for more detailed information.

Index building

In order to use *Salmon* to quantify the transcript abundances in a given sample, we first need to index the corresponding reference transcriptome. For a given organism, we download the fasta files containing cDNA and ncRNA sequences from *Ensembl*, complement these with ERCC spike-in sequences, and build a *Salmon* quasi-mapping index for the entire catalog. Note that the *salmon* report for a given data set (available in the *Master report* column) details the precise version of the transcriptome that was used for the quantification. For data sets with "long" reads (longer than 50 bp) we use the default *kr31*, while for "short reads" (typically around 25 bp) we set *kr35*.

We also create a lookup table relating transcript IDs to the corresponding gene IDs. This information is obtained by parsing the sequence names in the cDNA and ncRNA fasta files. From these names we also obtain the genomic coordinates for each feature.

Sample list and run matching

The first step is to determine the set of samples included in a given data set. We download a "RunInfo.csv" file for the data set from *SRA* and a Series Matrix file from *GEO*, in order to link samples both to individual runs and to phenotypic information. If the data set is not available from *GEO*, we construct a phenotype data file from the information provided by the corresponding repository.

Quality control

For each sample in the data set, we find all the corresponding runs, and download and concatenate the corresponding FastQ files from *SRA*. There is also an optional step to trim adapters from the reads using *cutadapt*. Next, we run *FastQC* to generate a quality control file for each concatenated read file (one or two files per sample depending on whether it was processed with a single-end or paired-end sequencing protocol).

Abundance quantification

After the QC, we run *Salmon* to estimate the abundance of each transcript from the catalog described above in each sample. The *Salmon* output files are then compressed in an archive and can be downloaded from *conquer* (see the *salmon archive* column).

For data obtained with non full-length library preparation protocols (e.g. targeting only the 3' or 5' end of transcripts), we quantify transcript and gene abundances using the *umis* pipeline developed by Valentine Svensson. Briefly, we *quasimap* the reads to the transcriptome using *RapMap* and use the counting capabilities of *umis* to obtain feature counts.

Summary report - MultiQC

Once *FastQC* and *Salmon* (or *RapMap/umis*) have been applied to all samples in the data set, we run *MultiQC* to summarise all the information into one report. This can also be downloaded from *conquer* (see the *MultiQC report* column). This report contains quality scores for all the samples and can be used to determine if there are problematic samples and whether the data set is good enough for the purposes of the user or needs to be subsetted.

Data summarisation

The provided *MultiAssayExperiment* object contains two "experiments", corresponding to the gene-level and transcript-level values. The gene-level experiment contains four "assays":

- *TPM*
- *count*
- *count scaled TPM length scaled TPM*, which can be used to reflect the transcript length, which is often used in gene models based on the count entry, see https://bioRxiv.org/articles/14_1421/
- *rank* (the *rank* function in *bioassay* package)