Basic analyses of scRNA-seq data, clustering and inference of cell-types

Acknowledgements:
Timothy Tickle
Brian Haas
Karthik Shekhar
Aviv Regev

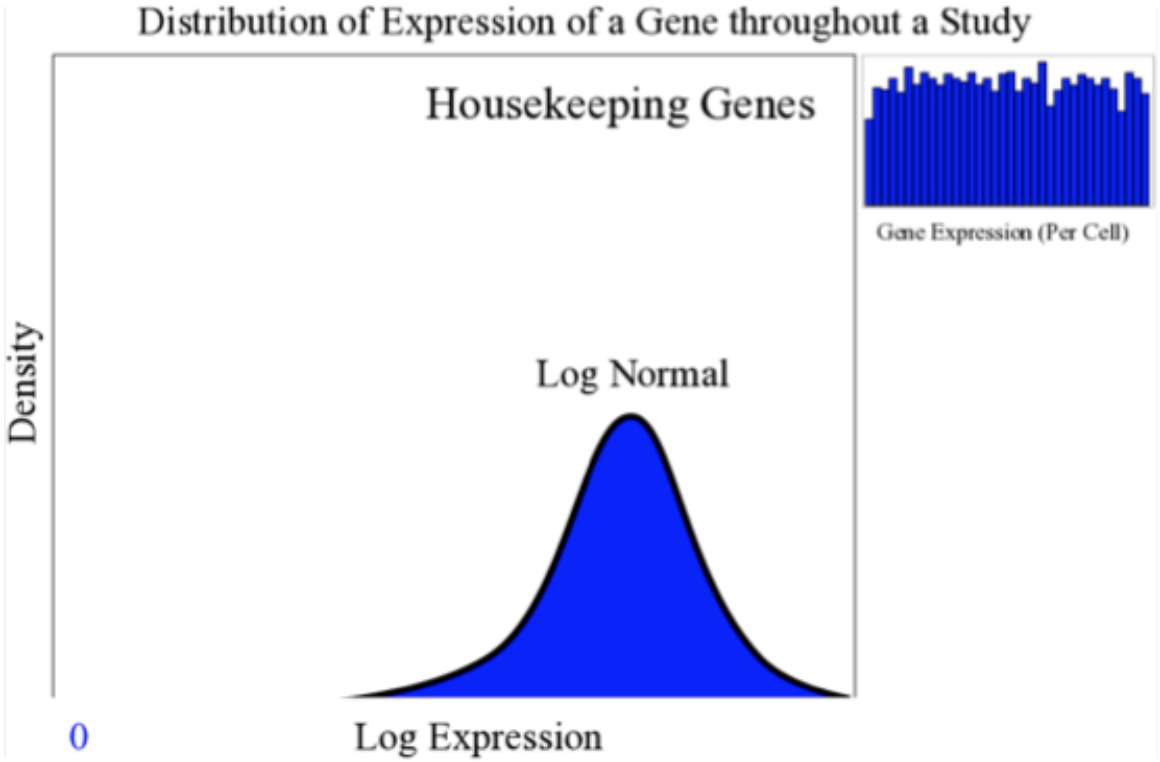
Agenda

- **Characterizing scRNA-Seq (what is in a count?)**
- From Counts to Expression
- Working with Counts
- Metadata and Filtering Cells
- Experimental Design
- Dimensionality Reduction
- Clustering
- Differential Expression

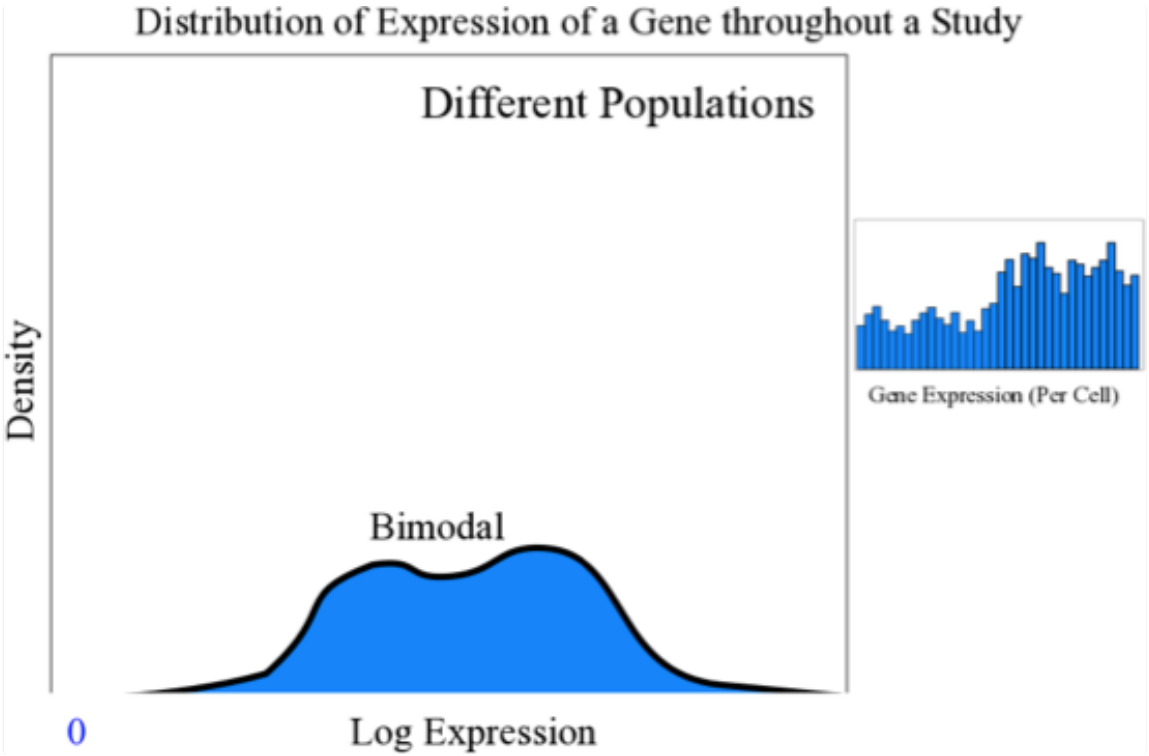
Counts Matrix

	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



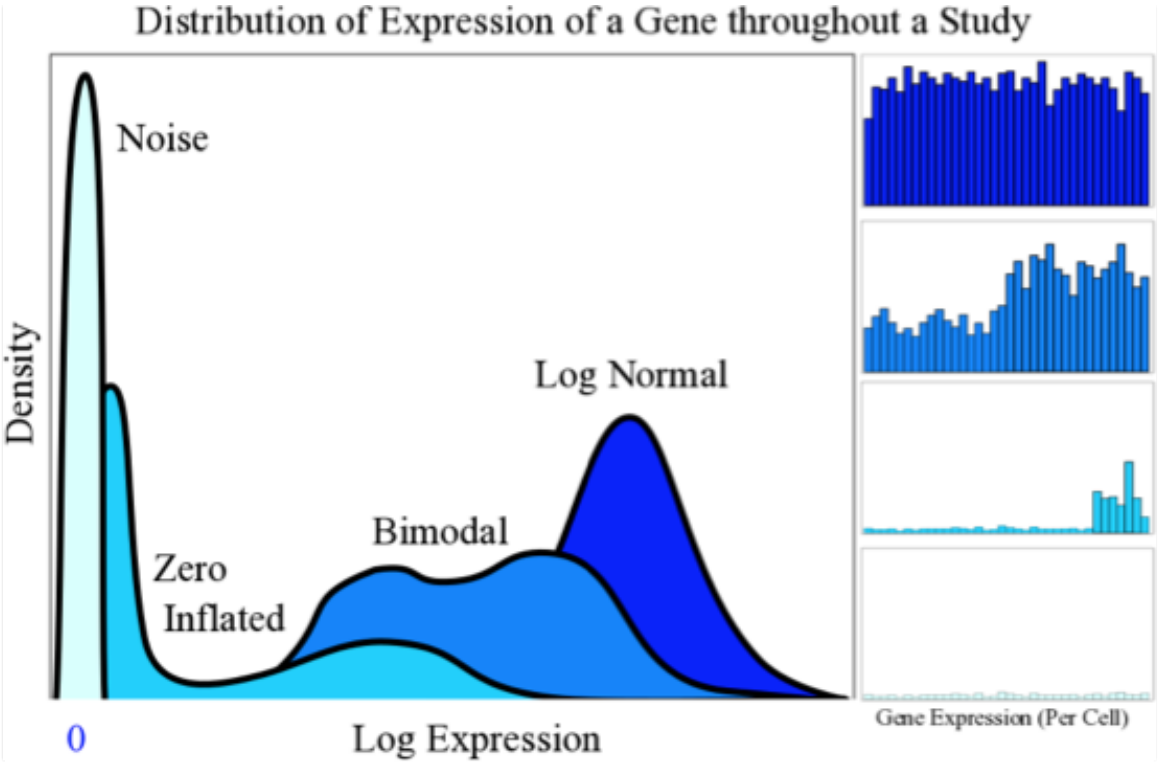
Genes Have Different Distributions



Genes Have Different Distributions

Genes Have Different Distributions

Genes Have Different Distributions



Cell Identity is a Mixture of Multiple Factors

REVIEW

nature biotechnology

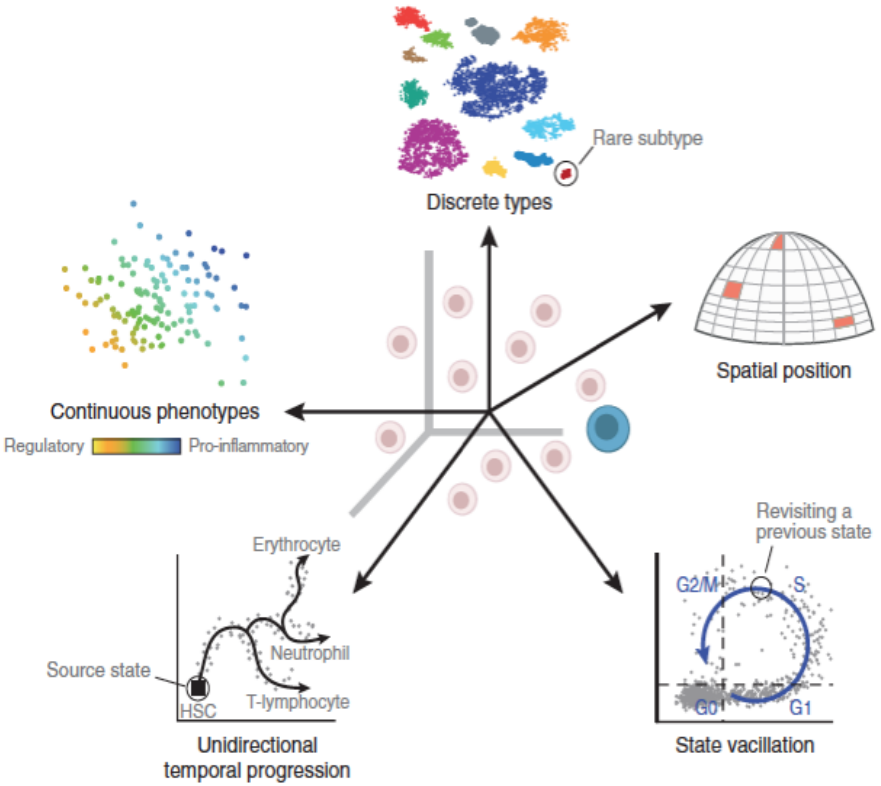
Revealing the vectors of cellular identity with single-cell genomics

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

Multiple factors shape a cell's identity

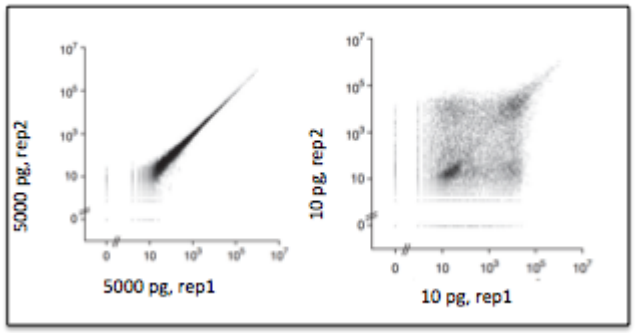
- Membership in a taxonomy of cell types
- Simultaneous time-dependent processes
- Response to the environment
- Spatial positioning

A cell participates in multiple cell contexts.

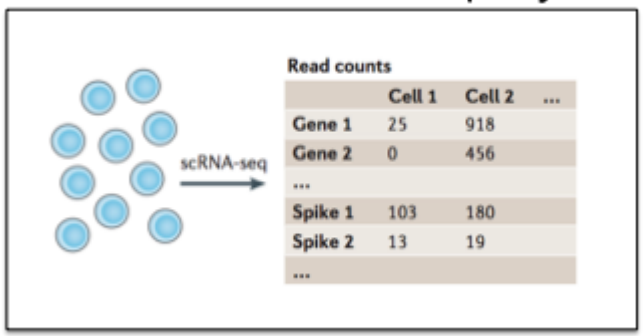


Technical Conceptual Challenges

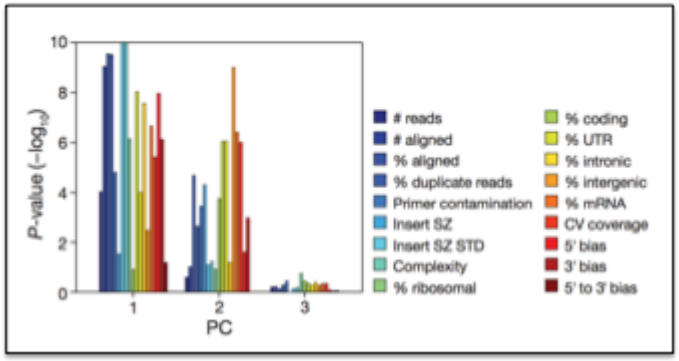
Dropouts



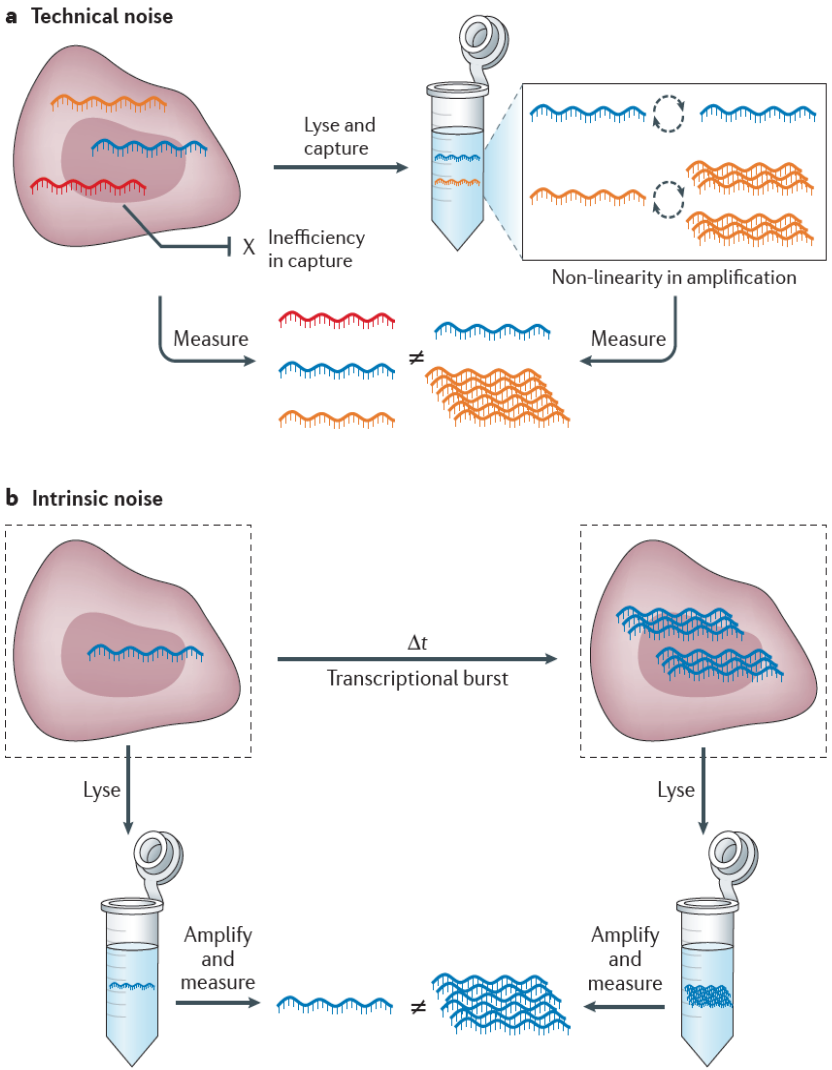
Variation in cell size and quality



Variation dominated by technical factors



Technical vs Intrinsic Noise

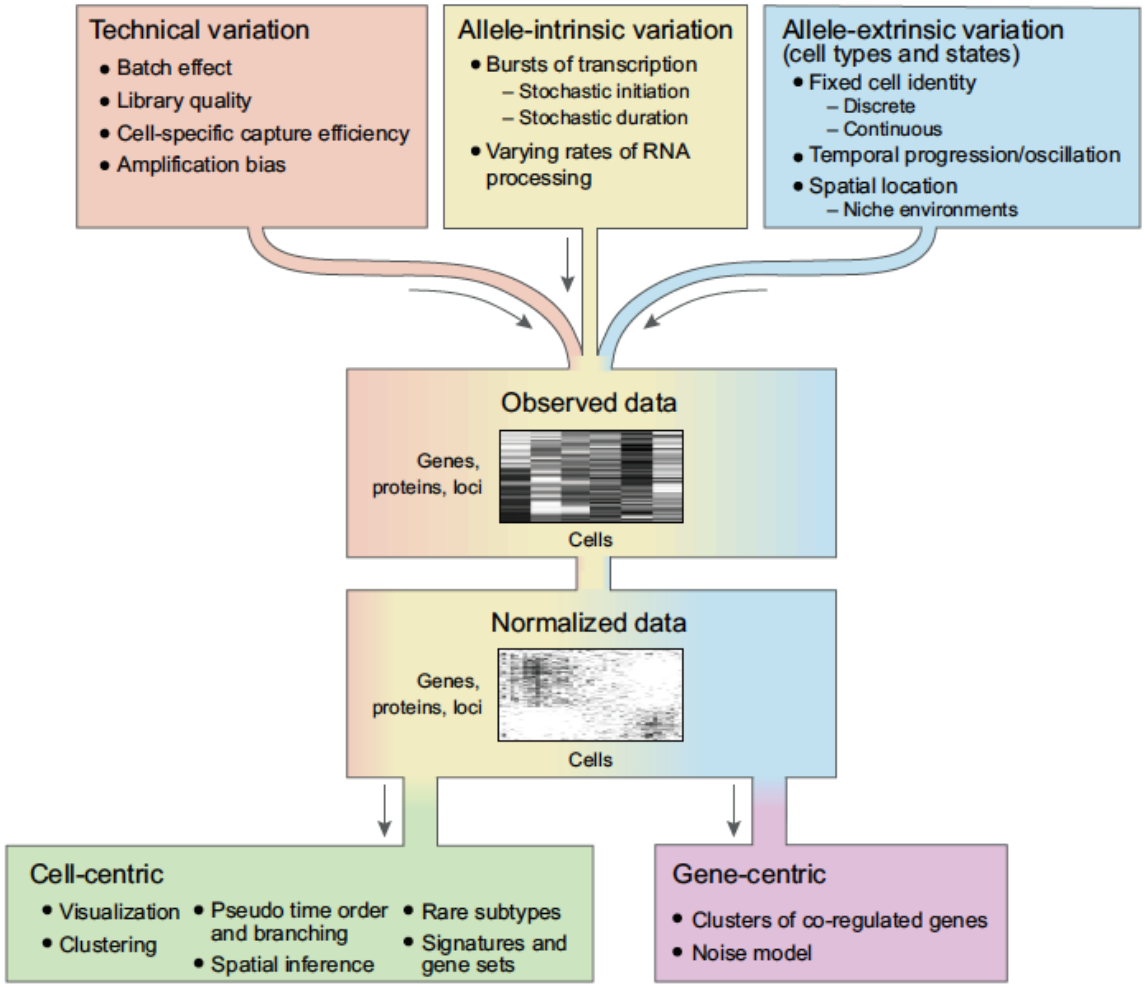


SINGLE-CELL OMICS

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]

Expression has Many Sources



Revealing the vectors of cellular identity with single-cell genomics

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

Agenda

- Characterizing scRNA-Seq (what is in a count?)
- **From Counts to Expression**
- Working with Counts
- Metadata and Filtering Cells
- Experimental Design
- Dimensionality Reduction
- Clustering
- Differential Expression

UMI Collapse

Read Counts

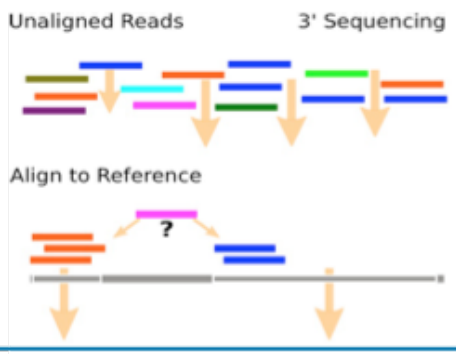
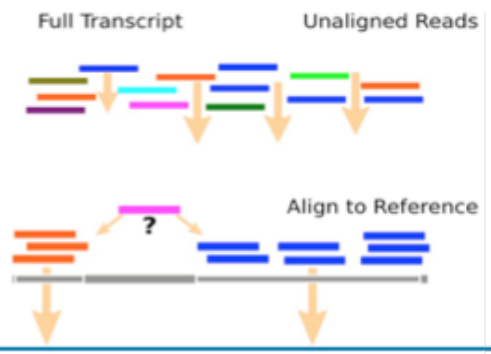
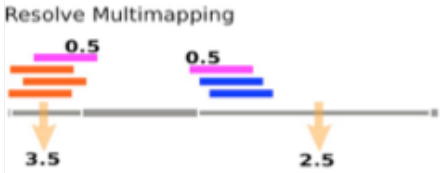
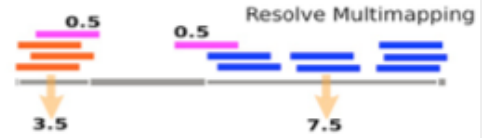
A4GALT	0	0	0	0	0
AAAS	20	22	1	5	9
AACS	15	4	2	3	1
AADAT	14	5	3	5	24
AAED1	33	16	4	46	12
AAGAB	19	19	13	5	0
AAK1	5	5	1	5	0
AAMDC	90	26	10	10	7
AAMP	56	45	28	24	36
AANAT	0	0	0	0	0

Counts by UMI

A4GALT	0	0	0	0	0
AAAS	10	5	1	2	3
AACS	3	2	1	2	1
AADAT	4	2	2	1	8
AAED1	8	7	1	10	4
AAGAB	8	6	3	3	0
AAK1	3	2	1	2	0
AAMDC	27	10	3	4	3
AAMP	21	21	13	11	16
AANAT	0	0	0	0	0

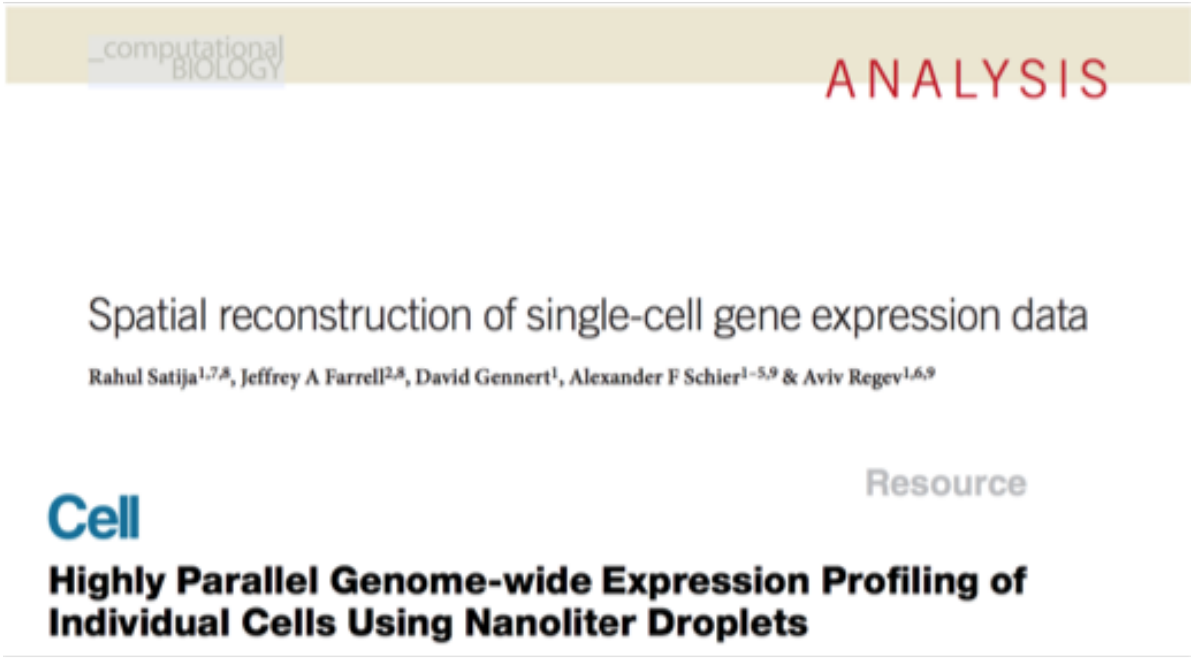
Collapsed but Not Linear

Count Preparation is Different Depending on Assays

	3’ sequencing	Full transcript	
			
RSEM KALLISTO			RSEM KALLISTO
No transcript length correction		3.5 / 1 = 3.5 Correct For Transcript Length 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

Seurat: R scRNA-Seq Analysis Package

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 (using RSEM).
- Log2 transform (in Seurat).
- Sequencing depth is already accounted.

Agenda

- Characterizing scRNA-Seq (what is in a count?)
- From Counts to Expression
- **Working with Counts**
- Metadata and Filtering Cells
- Experimental Design
- Dimensionality Reduction
- Clustering
- Differential Expression

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.

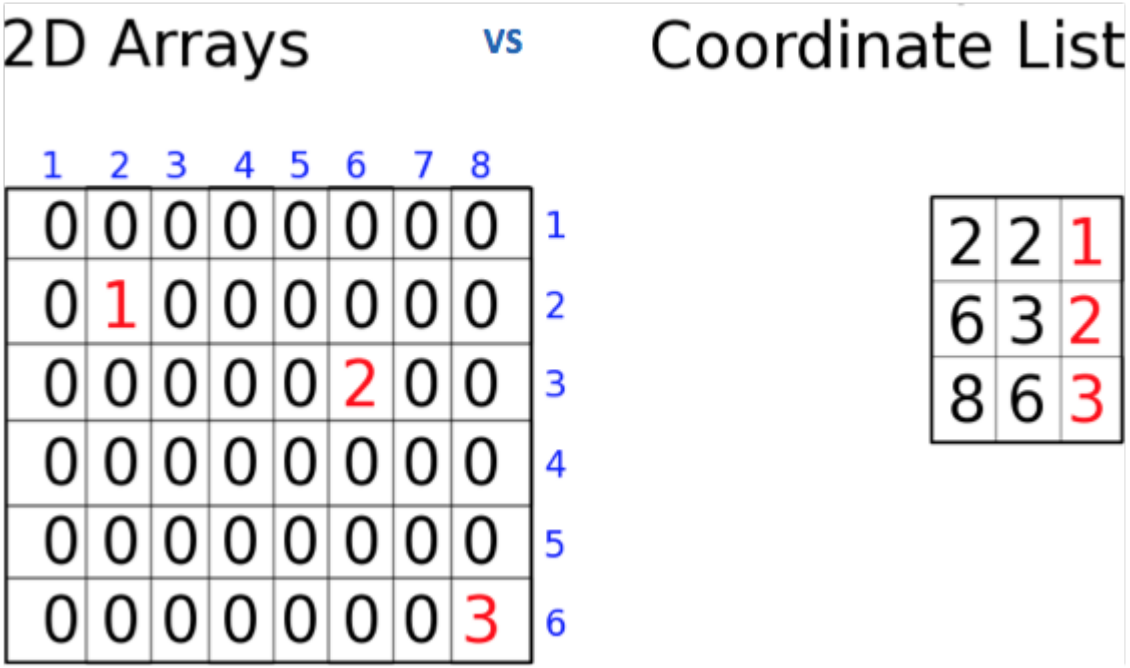
•Many ways to efficiently represent a sparse matrix in memory.

- Here, the underlying data structure is a coordinate list.

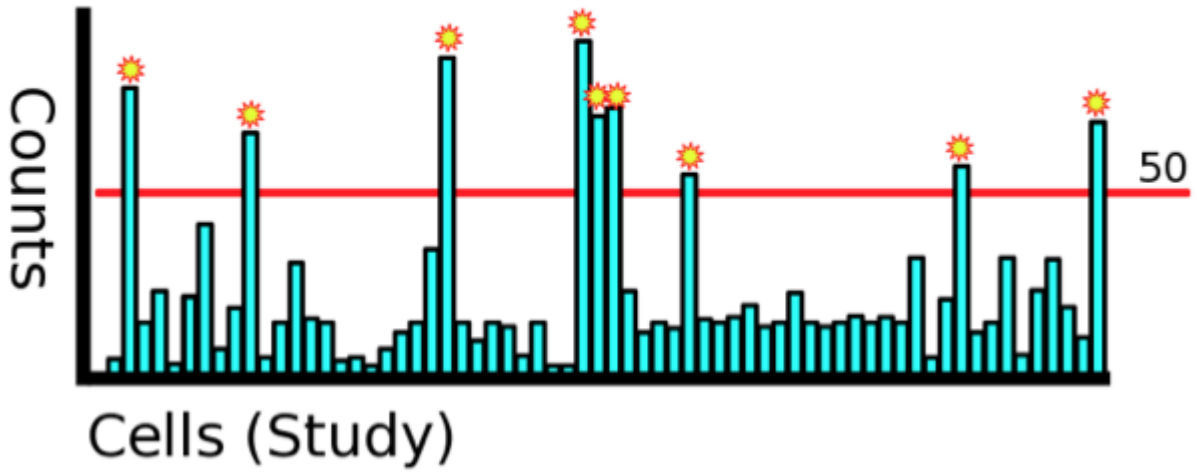
2D Array vs a Coordinate List

Can be optimal for dense matrices

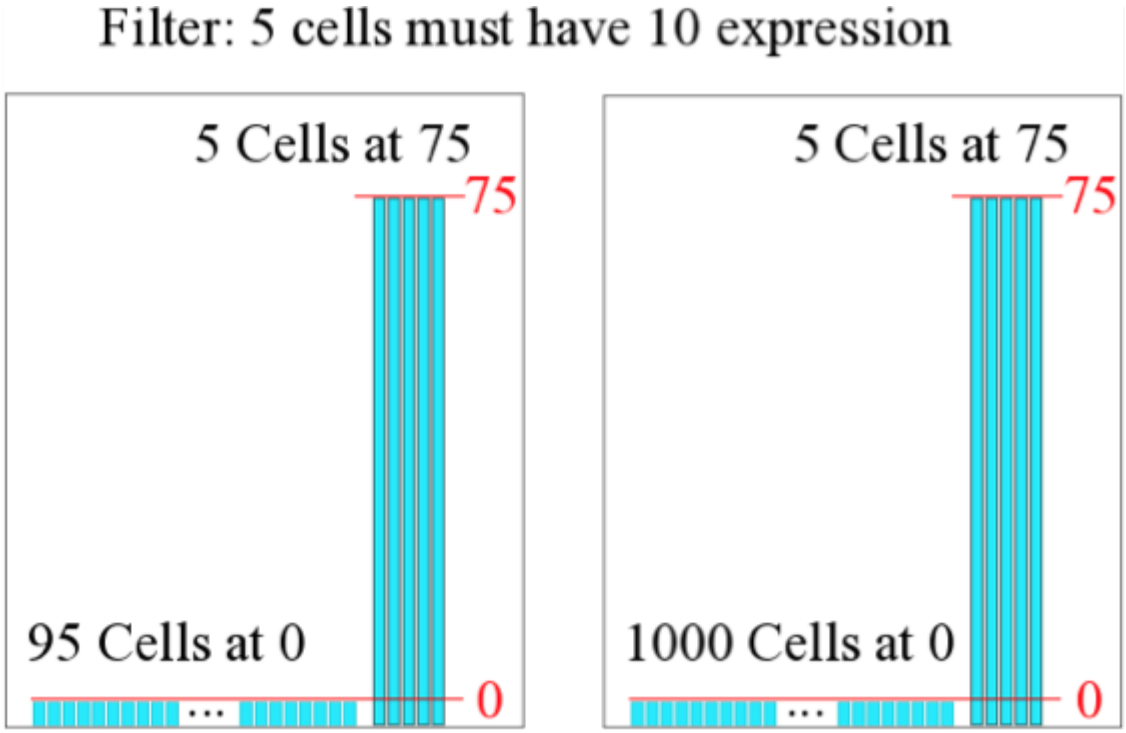
More optimal for sparse matrices



Filtering Genes: Using Prevalence



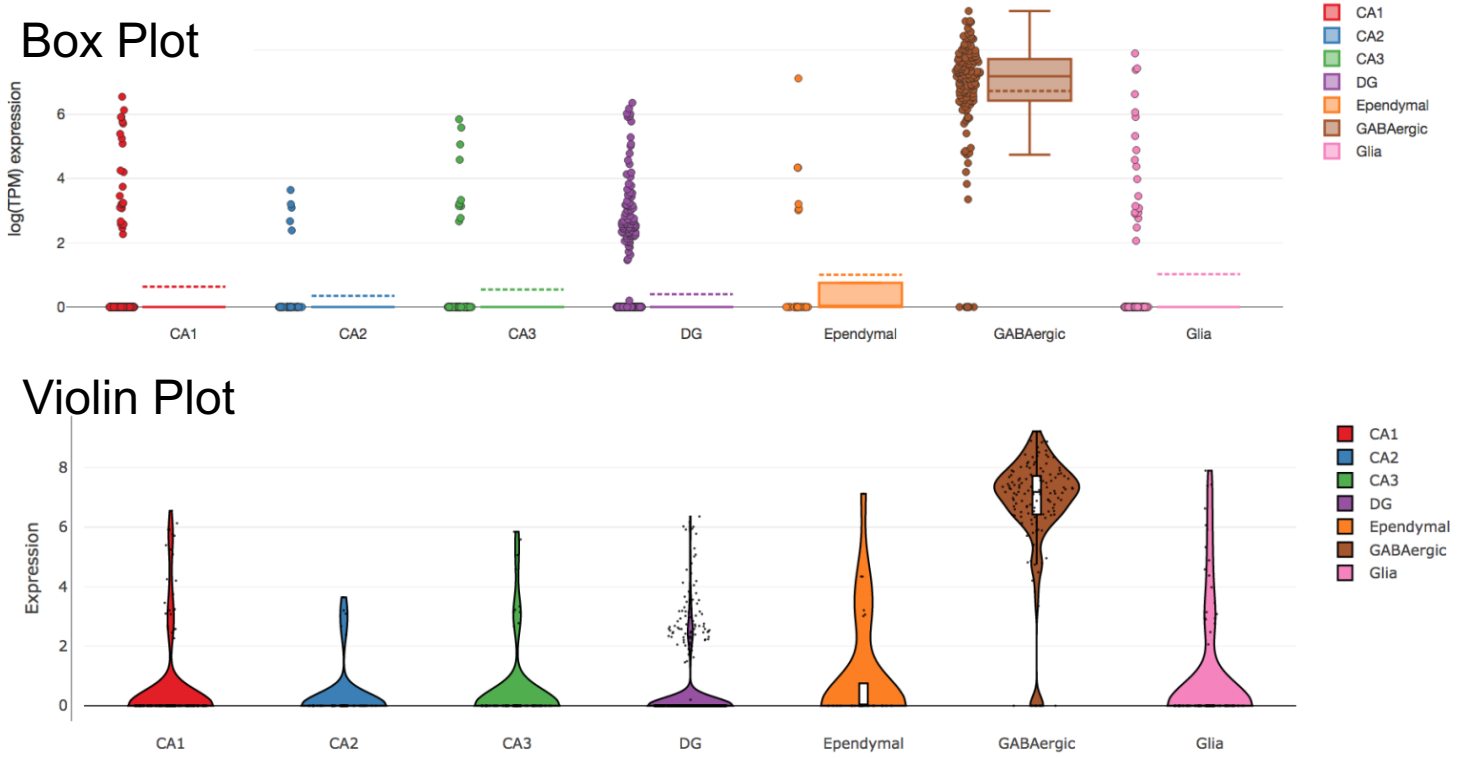
Filtering Genes: Using Prevalence



-Practical approach of removing sparsely expressed genes and focusing analysis on the most informative dimensions.

Representing Genes Throughout Cells

A gene (GAD2) across many groups of cells.



Habib et al. 2016

Agenda

- Characterizing scRNA-Seq (what is in a count?)
- From Counts to Expression
- Working with Counts
- **Metadata and Filtering Cells**
- Experimental Design
- Dimensionality Reduction
- Clustering
- Differential Expression

What is Metadata?

Other information that describes your measurements.

- Patient information—lifestyle (smoking), diet, age, etc.
- Study information—treatment, sequencing date
- Sequence QC on cells, number of expressed genes.

•Useful in filtering and stratifying.

Filtering Cells: Removing Outlier Cells

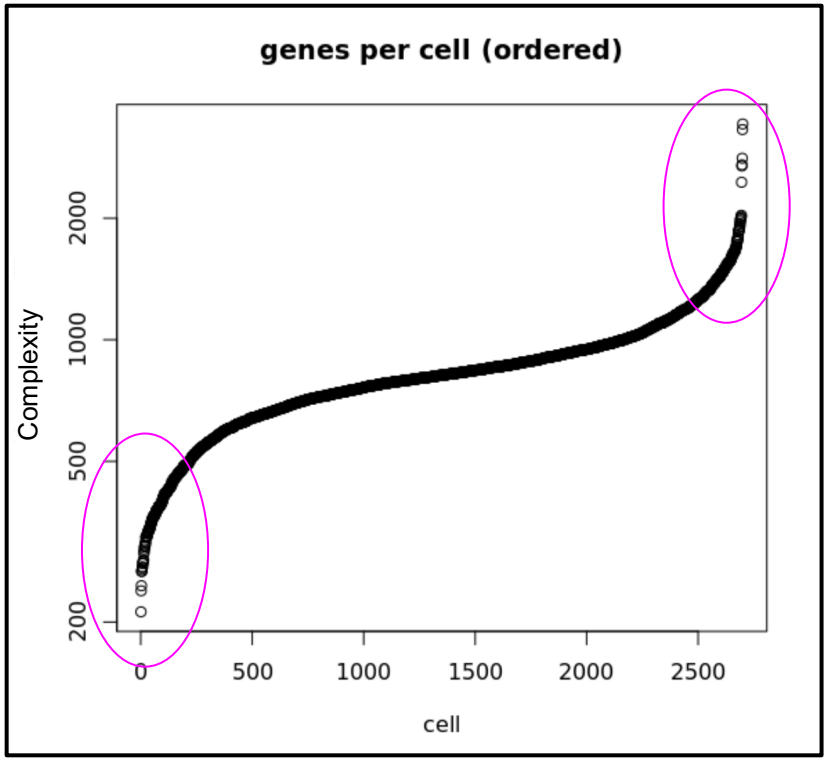
- Bulk RNA-Seq studies often do not remove outliers cells
  - scRNA-Seq often removes “failed libraries”.
- **Percent mitochondrial genes expressed**
- **Number of genes detectably expressed**
- Outlier cells are not just measured by complexity
- Percent reads mapping
- Presence of marker genes
- Intergenic/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

Filtering Cell: Complexity

Complexity:

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

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



Agenda

- Characterizing scRNA-Seq (what is in a count?)
- From Counts to Expression
- Working with Counts
- Metadata and Filtering Cells
- **Experimental Design**
- Dimensionality Reduction
- Clustering

Single Cell RNA-Seq and Batch Affects

New Results

On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data

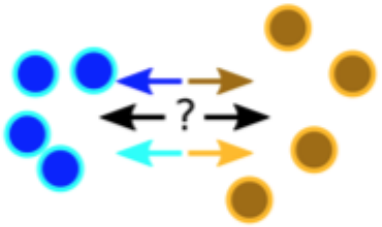
Stephanie C Hicks, Mingxiang Teng, Rafael A Irizarry

doi: http://dx.doi.org/10.1101/025528

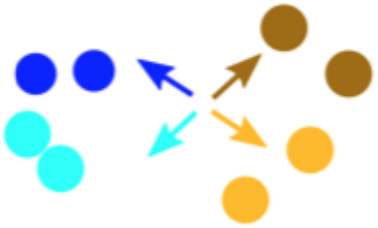
Abstract

What is Study Confounding

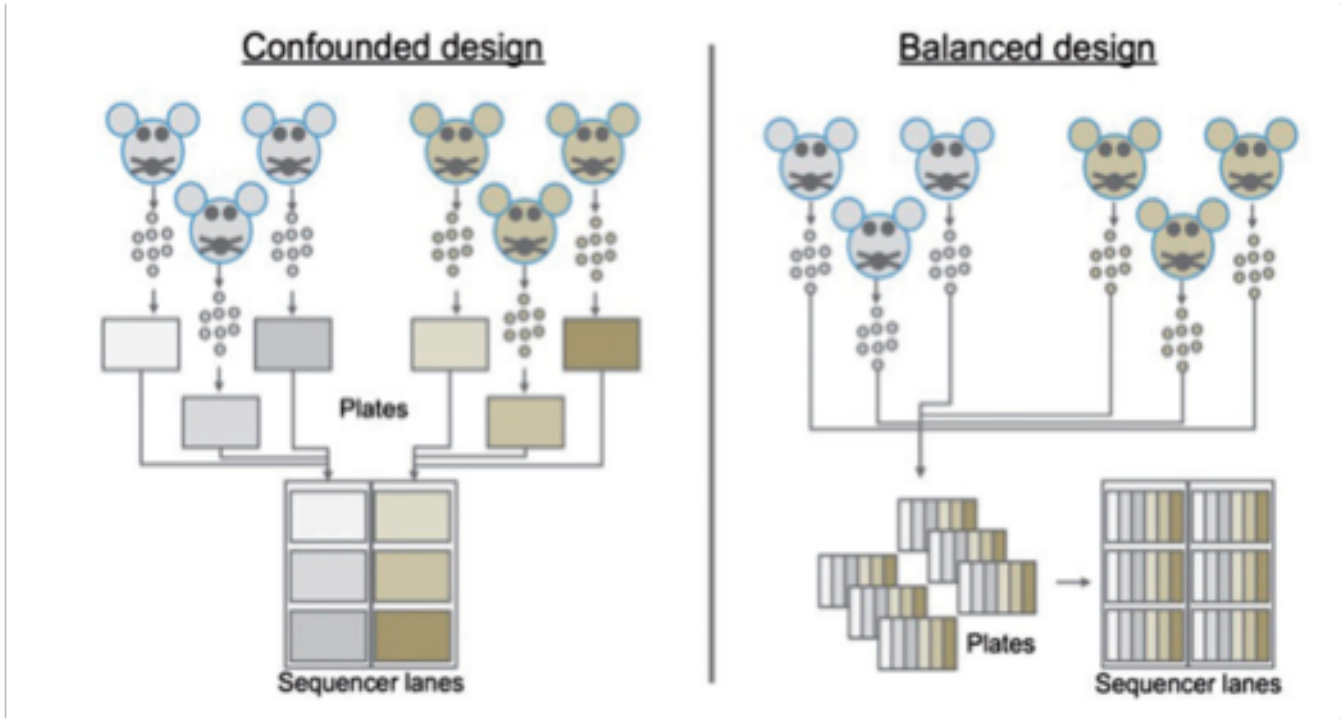
Cell	Site	Treatment
1	Main	A
2	Main	A
3	Main	A
4	Main	A
5	Remote	B
6	Remote	B
7	Remote	B
8	Remote	B



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

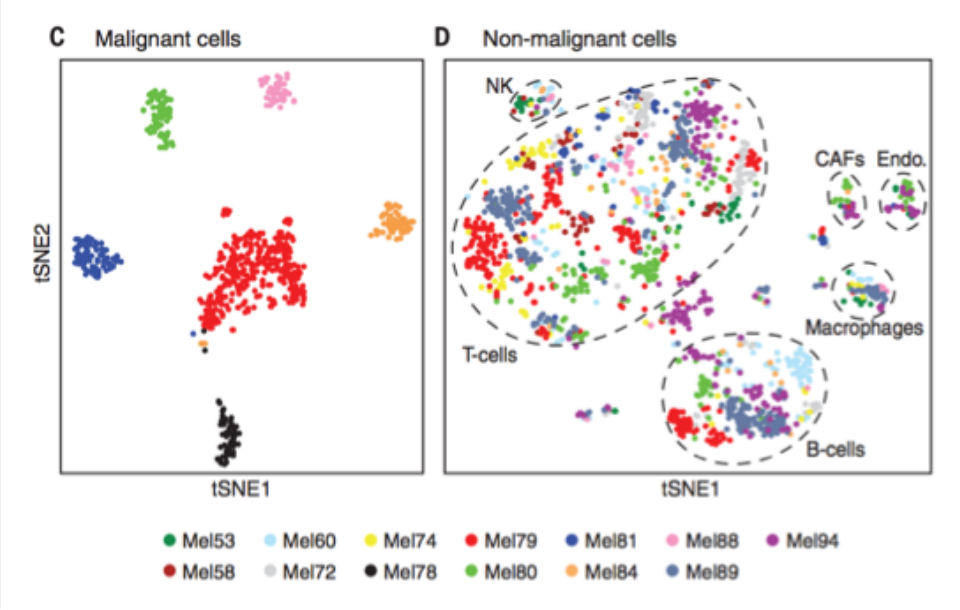


Apply this to Your Studies



- 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

Analysis can provide check and balances



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

Agenda

- Characterizing scRNA-Seq (what is in a count?)
- From Counts to Expression
- Working with Counts
- Metadata and Filtering Cells
- Experimental Design
- **Dimensionality Reduction**
- Clustering
- Differential Expression

Making Sense of Variation

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

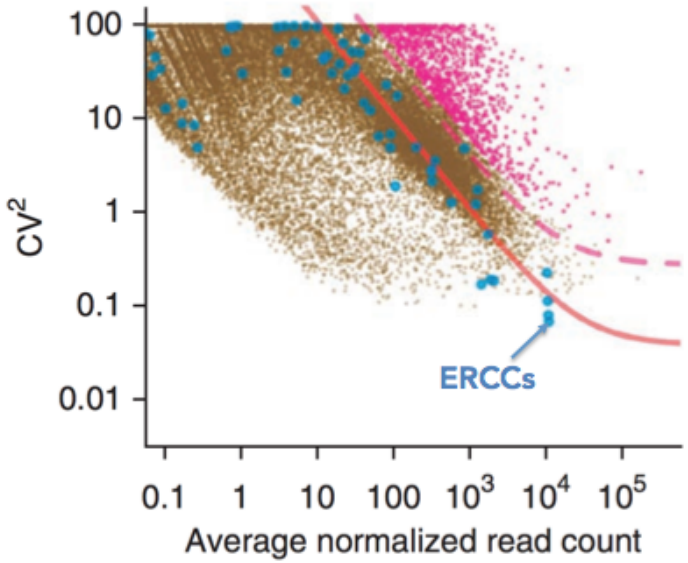
- **Fact 2** : Not everything that exhibits variation in real life, is informative

Identifying Relevant, “Highly Variable” Genes

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

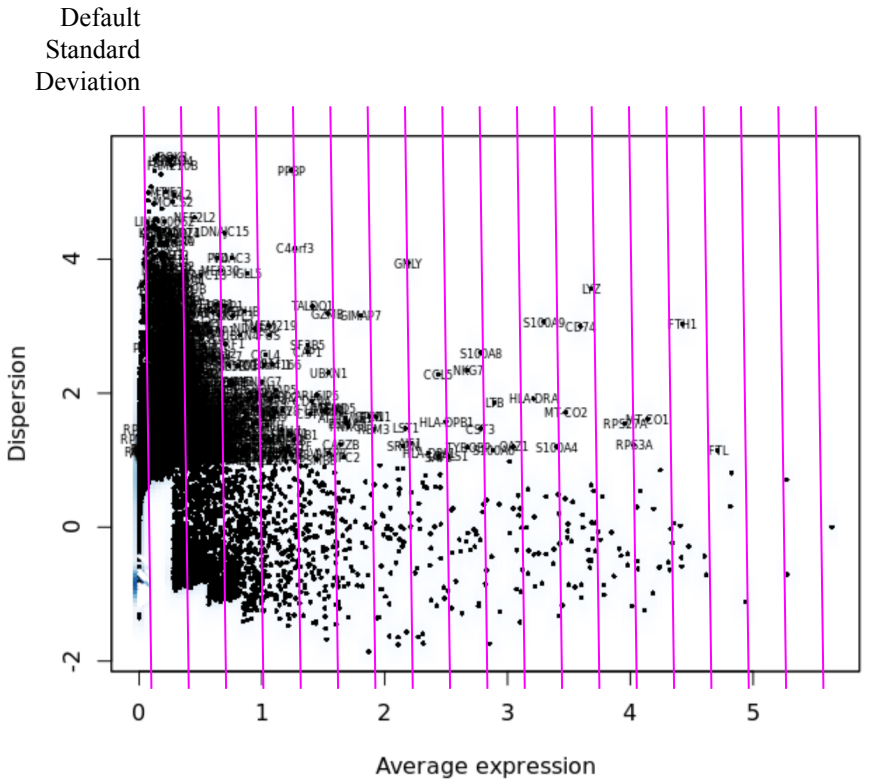
Variable Genes in Seurat

Calculate mean expression.

Calculate dispersion (log VMR in Seurat).

Calculate z-score for dispersions within each mean expression bin.

Stratifies and controls the relationship between variation and mean expression.



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.

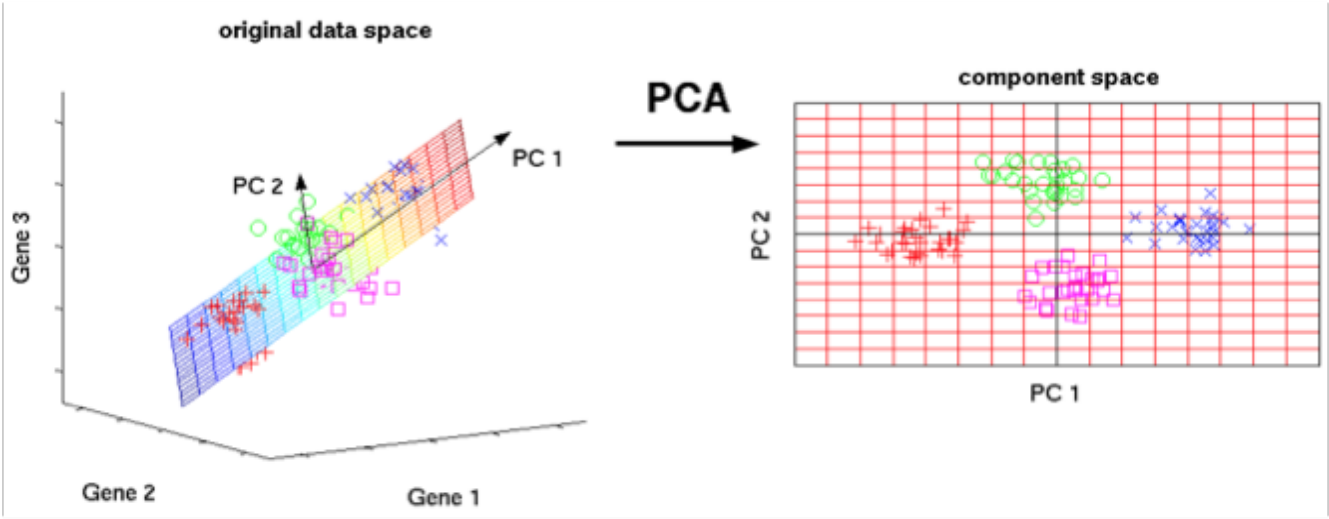
•Another could be to preferentially select some of the current features.

- We have already done this.

•We need this to plot the cells in 2D (or ordinate them)

Dimensionality Reduction—PCA

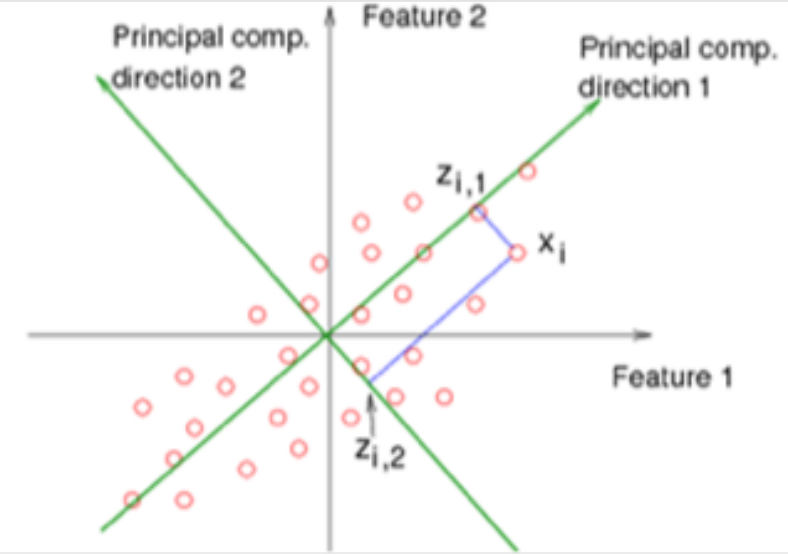
- **Why? :** Genes do not act independently, but as coregulatory "modules". E.g. in a cell type, the activity of a handful of transcription factors might lead to the co-expression of hundreds of genes defining cell-identity
- Cells occupy a low dimensional manifold in gene-expression space defined by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



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

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.



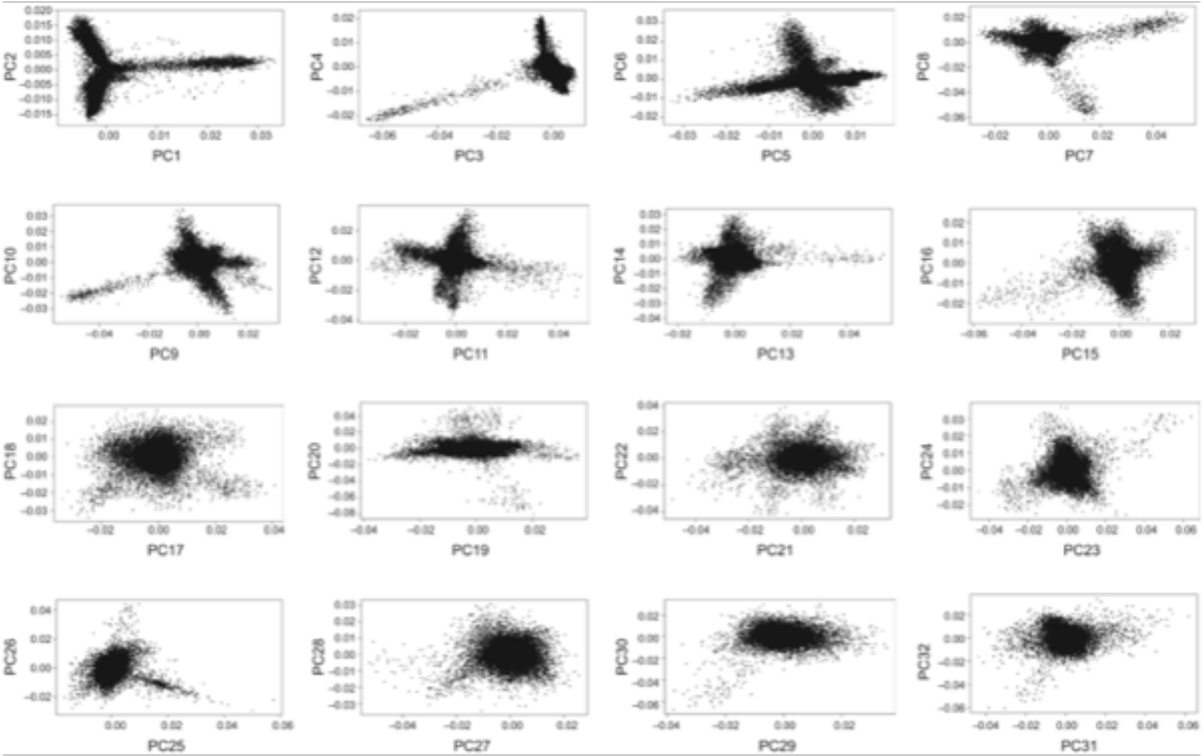
PCA: in Practice

Things to be aware of-

- Genes with different magnitudes will dominate.
  - Data standardized across cells (zero center and divided by SD)
- Can be affected by outliers.
- Data is often first filtered to remove noise.

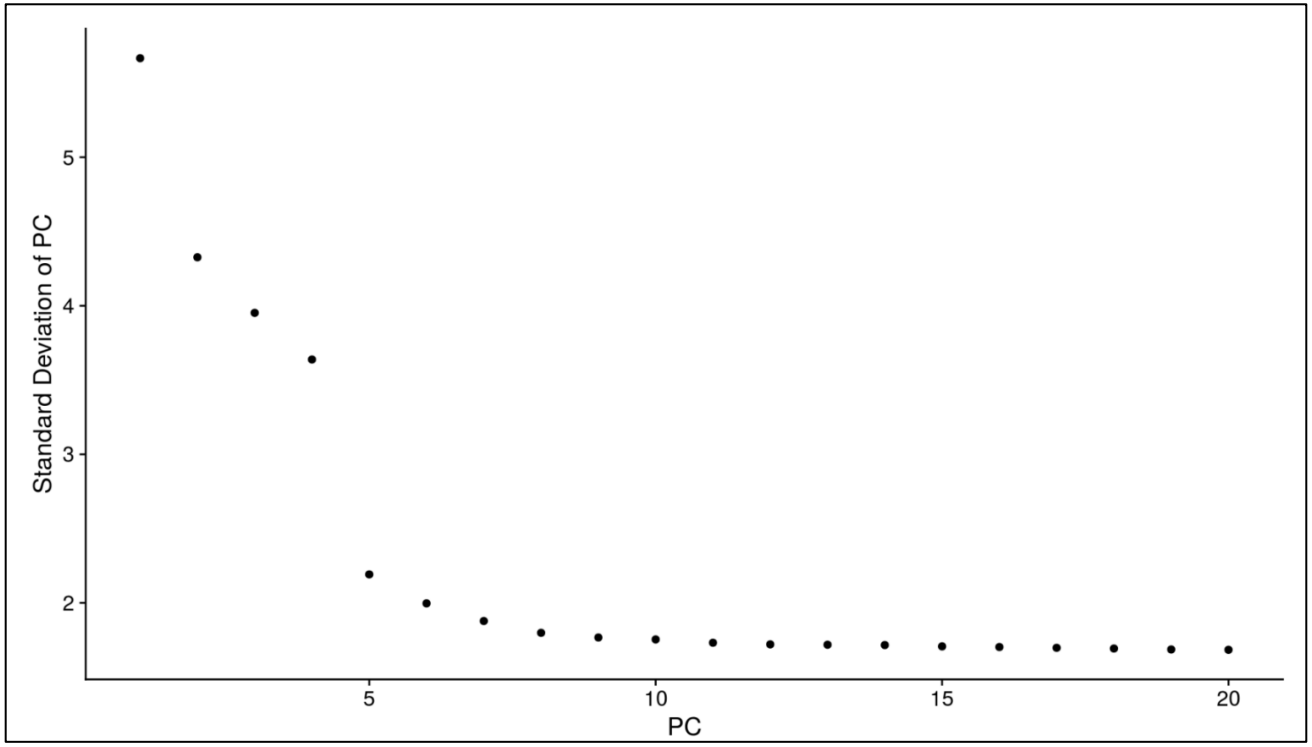
PCs

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

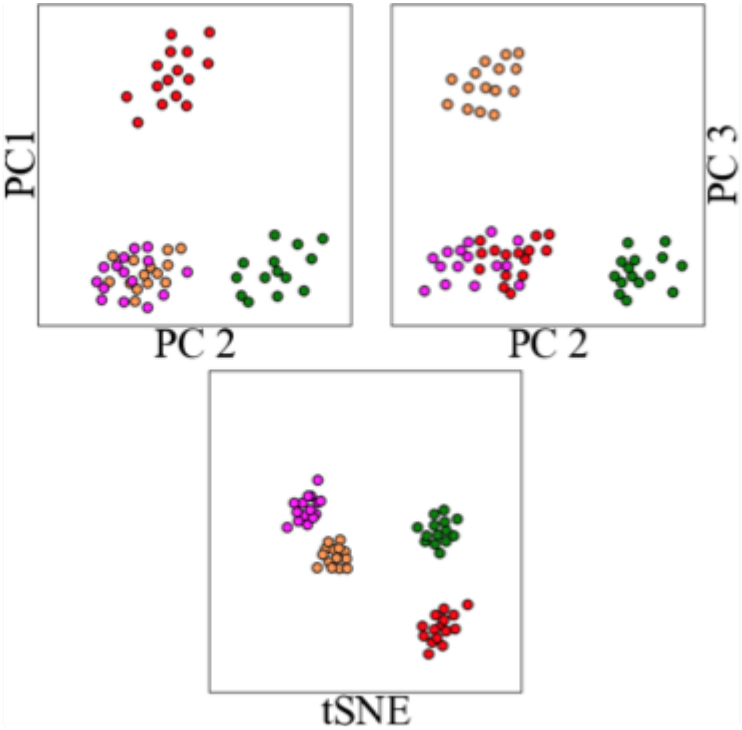


How Many Components Should We Use?

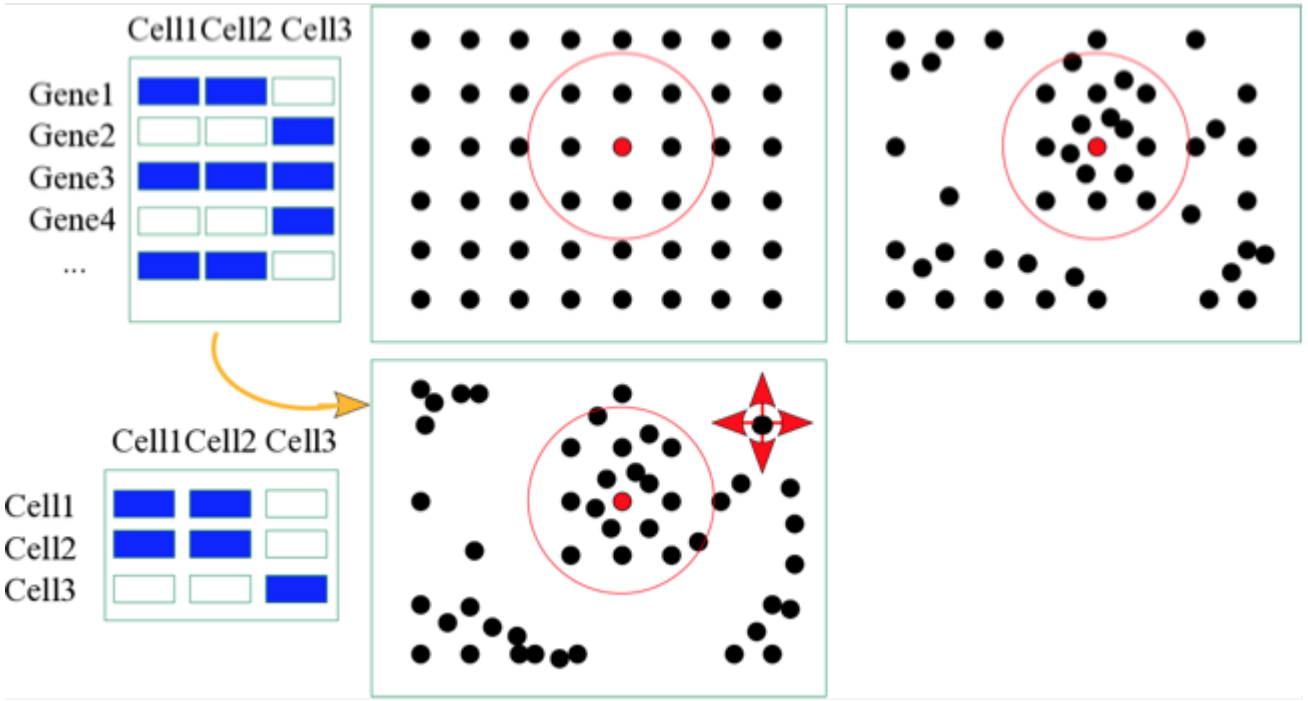
Elbow Plot (Scree Plot)



t-SNE: Collapsing the Visualization to 2D (non-linear)



t-SNE: How it Works



Aims to place cells with similar local neighborhoods in high-dimensional space together in low-dimensional space.

Inherently connected to graph-based structures.

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.

Caution When Interpreting t-SNE

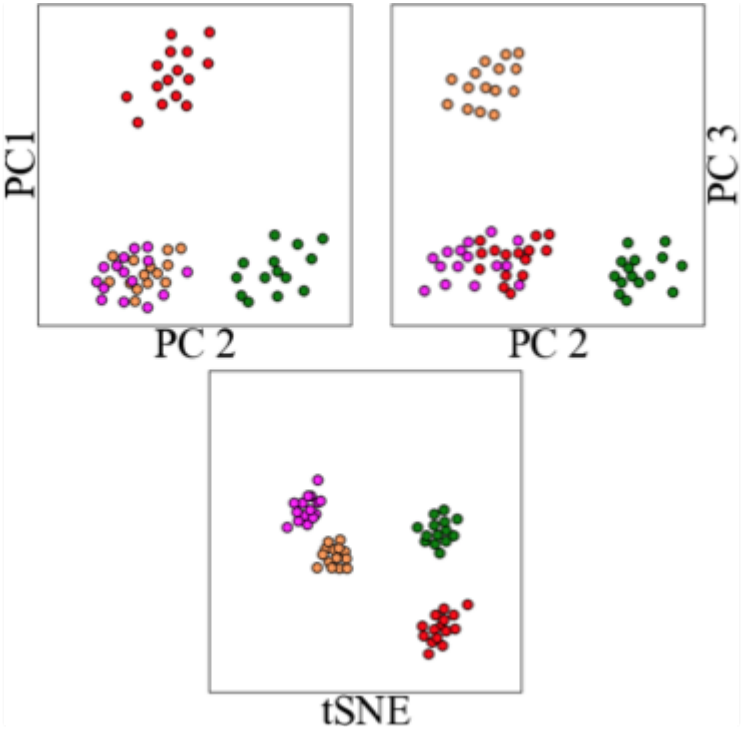
Nonlinear--optimized for local distance

Caveats to be aware of:
Distances between clusters may not mean anything—large distances do not necessarily reflect large dissimilarity

Big clusters can just mean more cells

Perplexity parameter or expected number of neighbors (default 30 in Seurat) can make it hard to find very rare subpopulations (5 cells or less).

Number of iterations run will also affect final visualization



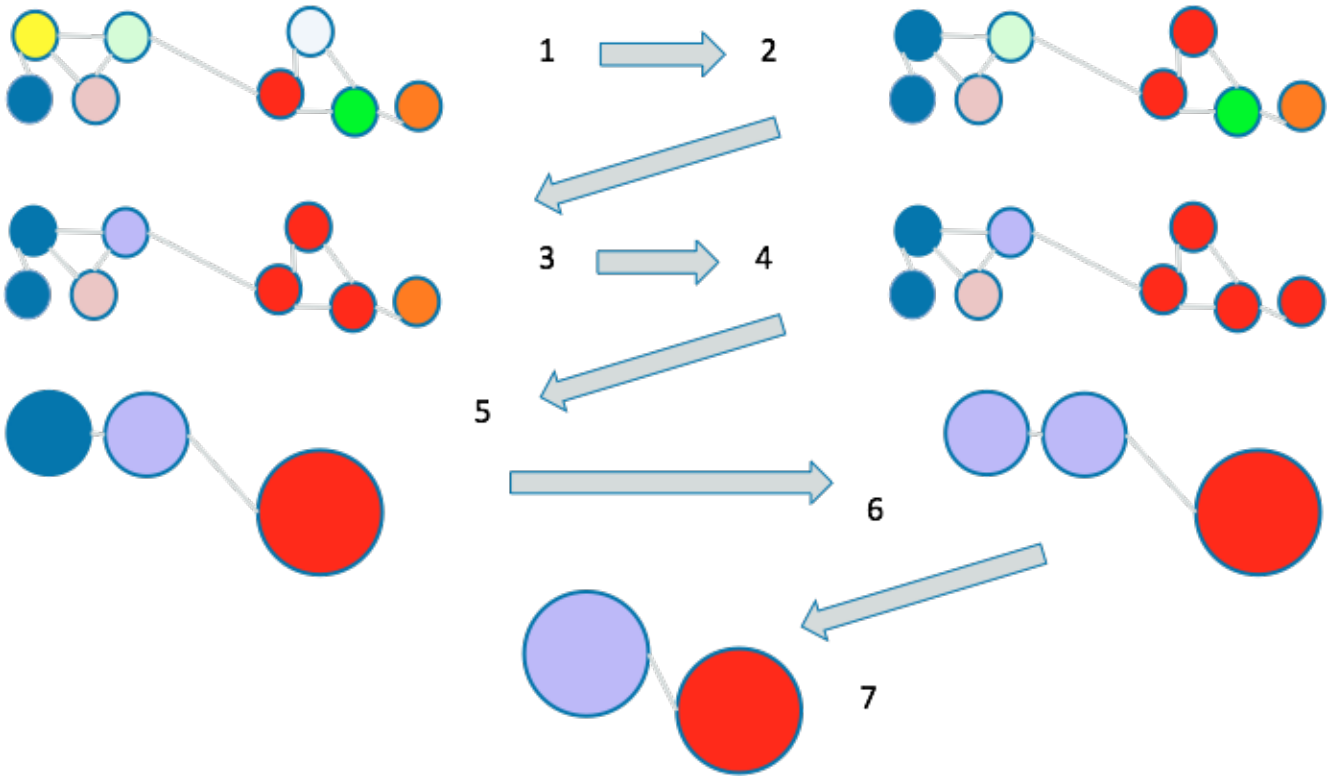
Learn More About t-SNE

- 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/

Agenda

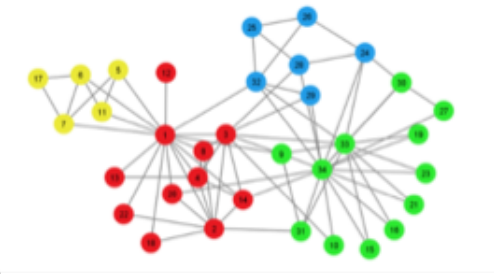
- Characterizing scRNA-Seq (what is in a count?)
- From Counts to Expression
- Working with Counts
- Metadata and Filtering Cells
- Experimental Design
- Dimensionality Reduction
- **Clustering**
- Differential Expression

Defining clusters through graphs—Louvain algorithm



Uses local moving heuristic to increase modularity of network

Smart Local Moving (SLM) algorithm



- 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 level refinement is the idea that one node can switch communities to increase the overall modularity of the network.

SLM iterates over community merging and local level refinement to find a locally optimal solution to both.

Seurat clustering

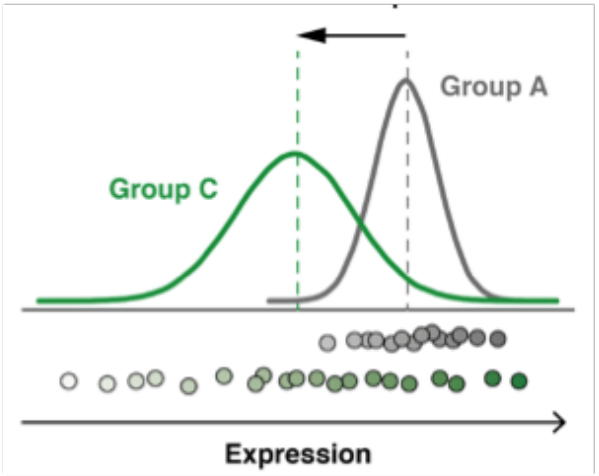
- Inspired by Louvain and SLM algorithms.
- Graph-based clustering approach similar to SNN-Cliq (Xu and Su, 2015) and CyTOF dataPhenoGraph (Levine et al., 2015).
- Partitions graph into highly interconnected communities.
- First builds graph based on euclidean distance between cells in PC space.
- Then refines edge weights based on shared overlap in local neighborhoods (Jaccard distance).
- Use SLM to iteratively group cells together, with the goal of optimizing the standard modularity function.

Agenda

- Characterizing scRNA-Seq (what is in a count?)
- From Counts to Expression
- Working with Counts
- Metadata and Filtering Cells
- Experimental Design
- Dimensionality Reduction
- Clustering
- **Differential Expression**

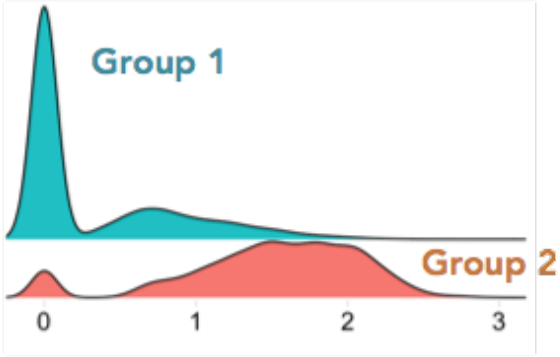
Differential Expression

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

Differential Expression Analysis

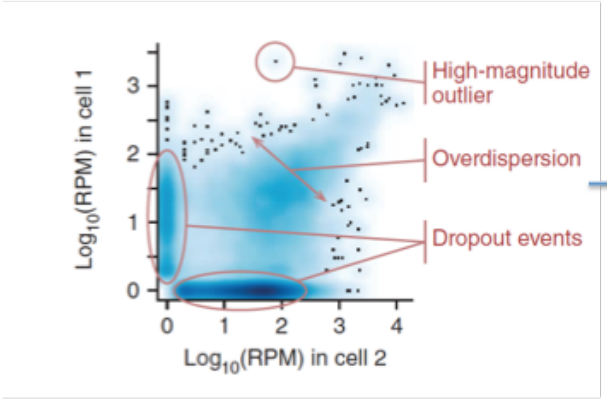
Many of the DE methods developed for bulk RNA-seq (e.g. edgeR, DE-seq) have serious limitations when applied to scRNA-seq data because of dropouts, so apply with caution!

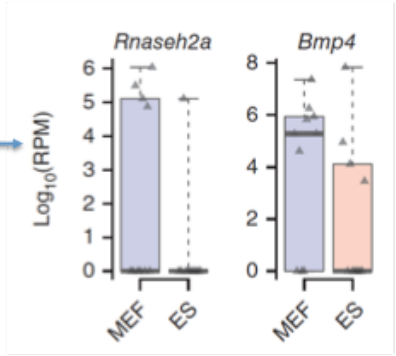
Soneson and Robinson, 2017

	Short name	Method	Software version	Input	Reference
	BPSC	BPSC	BPSC 0.99.0	CPM	[48]
	D3E	D3E	D3E 1.0	raw counts	[49]
	DESeq2	DESeq2	DESeq2 1.14.1	raw counts	[14]
	DESeq2census	DESeq2	DESeq2 1.14.1	census counts	[14]
	DESeq2nofilt	DESeq2 without the built-in independent filtering	DESeq2 1.14.1	raw counts	[14]
	edgeRLRT	edgeR/LRT	edgeR 3.17.5	raw counts	[15, 41, 37]
	edgeRLRTcensus	edgeR/LRT	edgeR 3.17.5	census counts	[15, 41, 37]
	edgeRLRTdeconv	edgeR/LRT with deconvolution normalization	edgeR 3.17.5, scran 1.2.0	raw counts	[15, 37, 42]
	edgeRLRTrobust	edgeR/LRT with robust dispersion estimation	edgeR 3.17.5	raw counts	[15, 41, 37, 40]
	edgeRQLF	edgeR/QLF	edgeR 3.17.5	raw counts	[15, 38, 41]
	limmatrend	limma-trend	limma 3.30.13	raw counts	[57, 16]
	MASTcpm	MAST	MAST 1.0.5	log_2(CPM+1)	[50]
	MASTcpmDetRate	MAST - accounting for detection rate	MAST 1.0.5	log_2(CPM+1)	[50]
	MASTtpm	MAST	MAST 1.0.5	log_2(TPM+1)	[50]
	MASTtpmDetRate	MAST - accounting for detection rate	MAST 1.0.5	log_2(TPM+1)	[50]
	metagenomeSeq	metagenomeSeq	metagenomeSeq 1.16.0	raw counts	[54]
	monocle	monocle	monocle 2.2.0	TPM	[44]
	monoclecensus	monocle	monocle 2.2.0	census counts	[44, 43]
	NODES	NODES	NODES 0.0.0.9010	raw counts	[47]
	ROTScpm	ROTS	ROTS 1.2.0	CPM	[55, 56]
	ROTStpm	ROTS	ROTS 1.2.0	TPM	[55, 56]
	ROTSvoom	ROTS	ROTS 1.2.0	voom-transformed raw counts	[55, 56]
	SAMseq	SAMseq	samr 2.0	raw counts	[45]
	SCDE	SCDE	scde 1.99.4	raw counts	[51]
	SeuratBimod	Seurat (bimod test)	Seurat 1.4.0.7	raw counts	[52, 53]
	SeuratBimodnofilt	Seurat (bimod test) without the internal filtering	Seurat 1.4.0.7	raw counts	[52, 53]
	SeuratBimodIsExpr2	Seurat (bimod test) with internal expression threshold set to 2	Seurat 1.4.0.7	raw counts	[52, 53]
	SeuratTobit	Seurat (tobit test)	Seurat 1.4.0.7	TPM	[52, 44]
	voomlimma	voom-limma	limma 3.30.13	raw counts	[57, 16]
	Wilcoxon	Wilcoxon test	stats (R v 3.3.1)	TMM-normalized TPM	[41, 46]

Single Cell Differential Expression (SCDE)

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





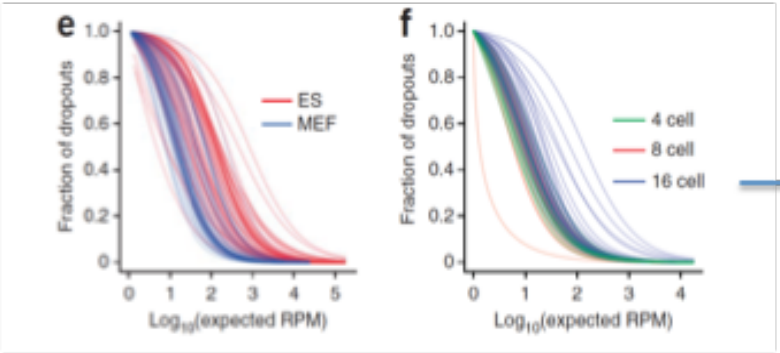
SCDE exchanges information between closely related cells to estimate dropout rates for every cell!

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

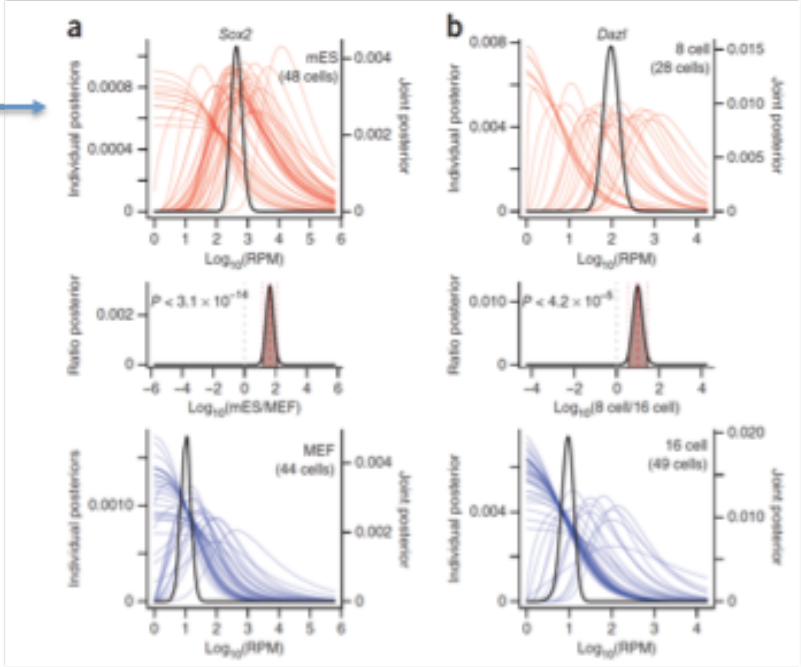
Kharchenko et al., 2014

Single Cell Differential Expression (SCDE)

For every cell, a "dropout curve" is estimated



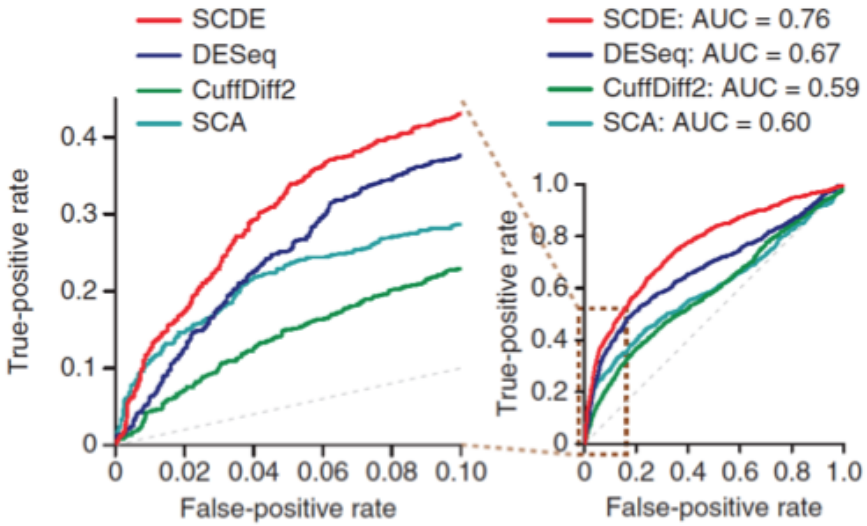
Which is used in a Bayesian framework to estimate posterior distributions for every gene in every cell



Kharchenko et al., 2014

SCDE is Much More Sensitive and Specific

One of the disadvantages of SCDE is its run-time, which does not scale well for large datasets. Newer methods like MAST (Finak et al., 2016) overcome this!



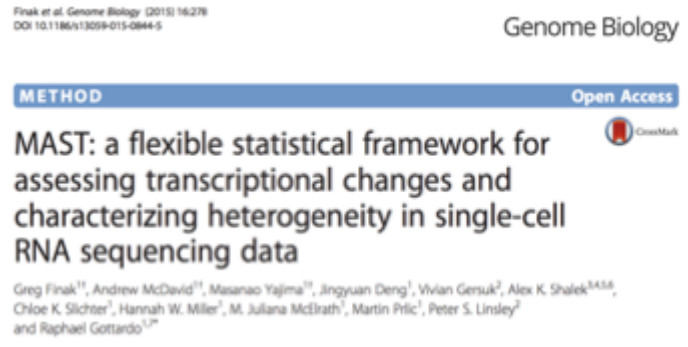
MAST

- Uses linear hurdle model
  - Two part generalized linear model to address both rate of expression (Z) and level of expression (Y).

$$logit\left(P_r\left(Z_{ig}=1\right)\right)=X_i\beta_g^D$$

$$P_r\left(Y_{ig}=y|Z_{ig}=1\right)=N(X_i\beta_g^C,\sigma_g^2)$$

  - GLM means covariates can be used to control for unwanted signal.
- CDR: Cellular detection rate
  - Cellular complexity
  - Values below a threshold are 0

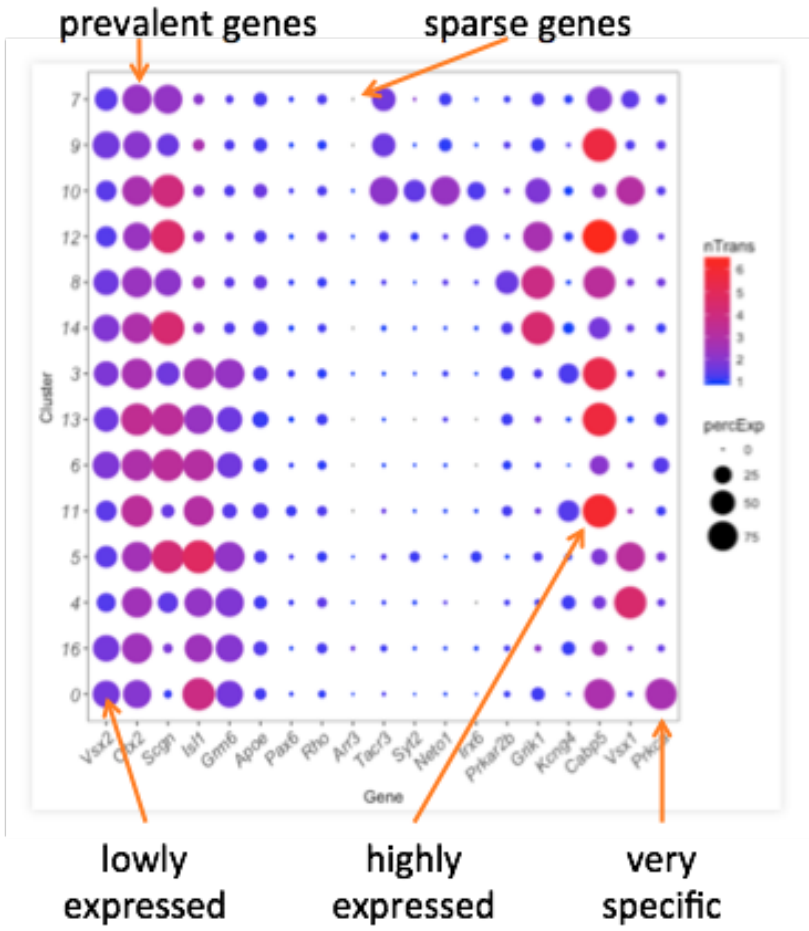


Additionally introduces a GSEA method

https://github.com/RGLab/MAST

Dot Plots

- Size of circle
  - gene prevalence in cluster
- Color of circle
  - gene expression in cluster
  - scales well with any cells.



Seurat: Differential Expression

- Default is one cluster against many tests
- 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

Wilcoxen—default , non-parametric test if two samples came from the same distribution

Bimod—tests differences in mean and proportions.

ROC—uses AUC like definition of separation.

T—Student's T-test (Normal assumption)

Tobit—Tobit regression on a smoothed data.

MAST—hurdle model for zero inflated data

….

Summary

Because of the resolution, scRNA-Seq is very different from population-based RNA-Seq.

scRNA-Seq counts represent biology being study confounded by other biology and technical signal.

Due to the spare (zero-inflated nature) of scRNA-Seq special care is take on handling and visualizing data.

It is important to design experiments so that studies are not confounded with technical batches.

PCA and tSNE for dimensionality reduction

Clustering and differential expression methods for single cell data