Clustering and Differential Expression

Introduction to Single Cell RNA-Seq (45)

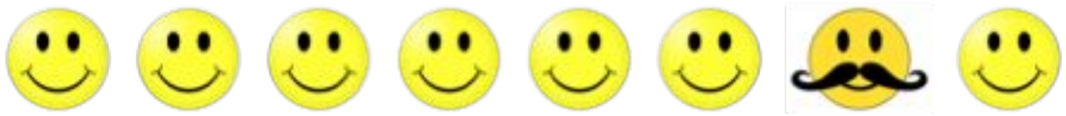
Timothy Tickle
Brian Haas

Agenda (Clustering and Differential Expression)

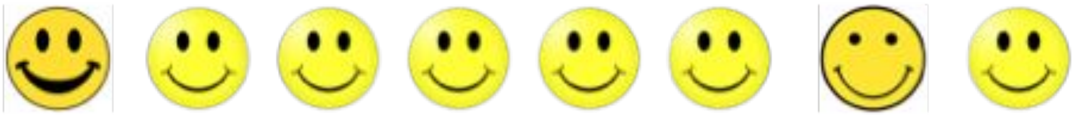
- **Dimensionality Reduction**
  - **PCA**
  - **t-SNE**
- Differential Expression
  - SCDE
  - MAST

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

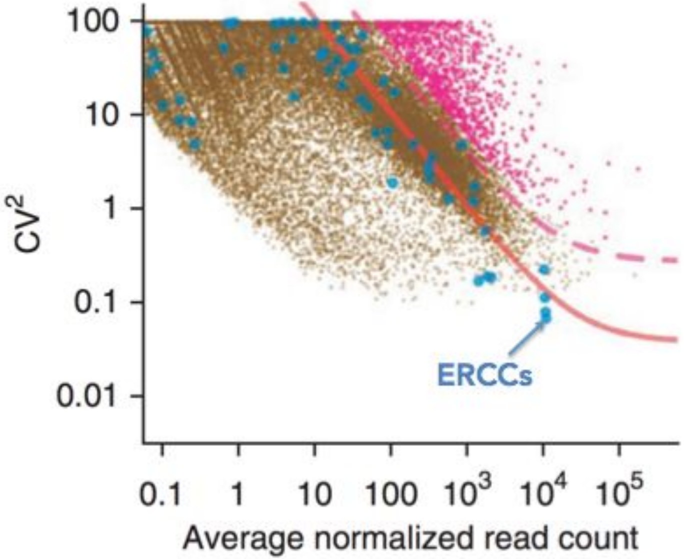


Identifying Relevant, “Highly Variable” Genes

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

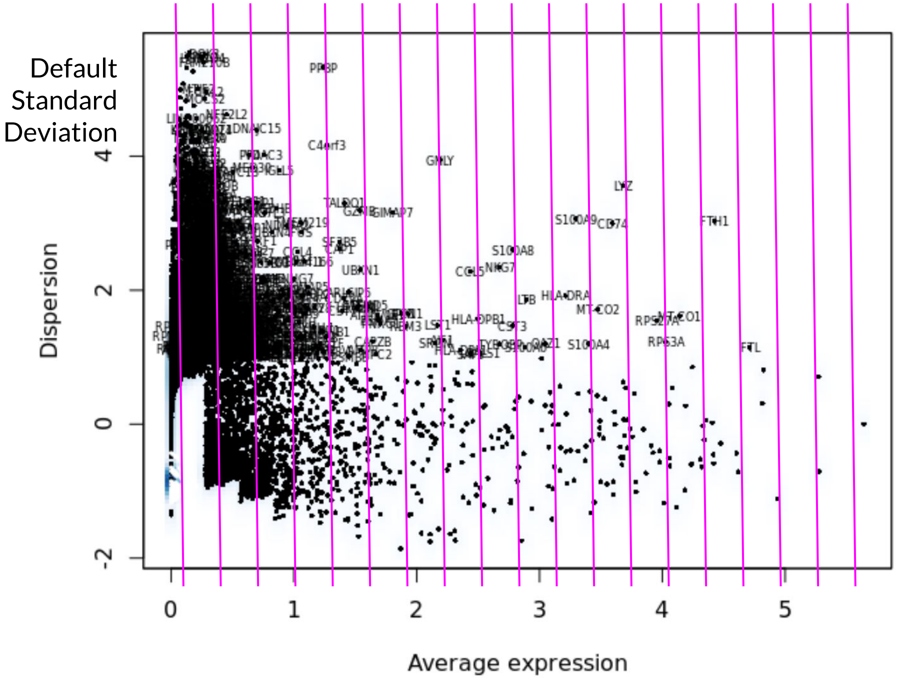
Variable Genes in Seurat

Calculate mean expression.

Calculate disperstion (standard deviation).

Calculate z-score for dispersions within each bin.

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



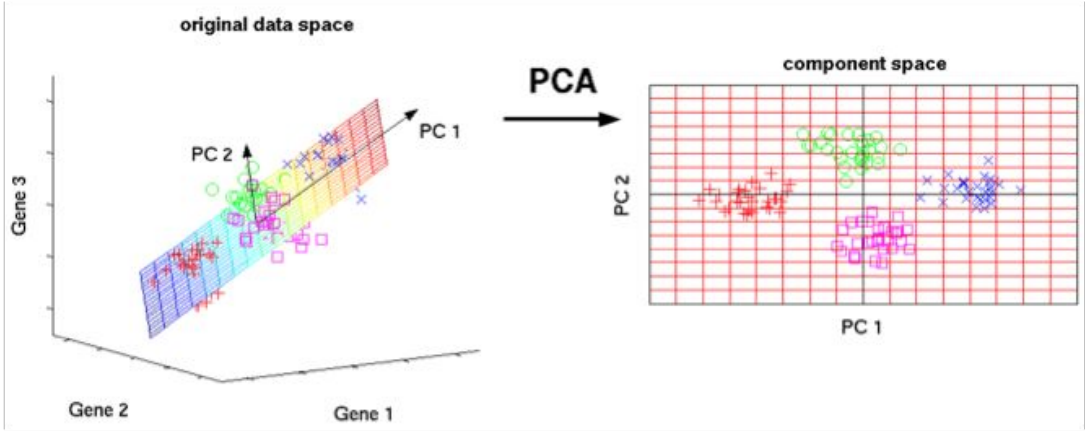
Default Standard Deviation

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.

Dimensionality Reduction

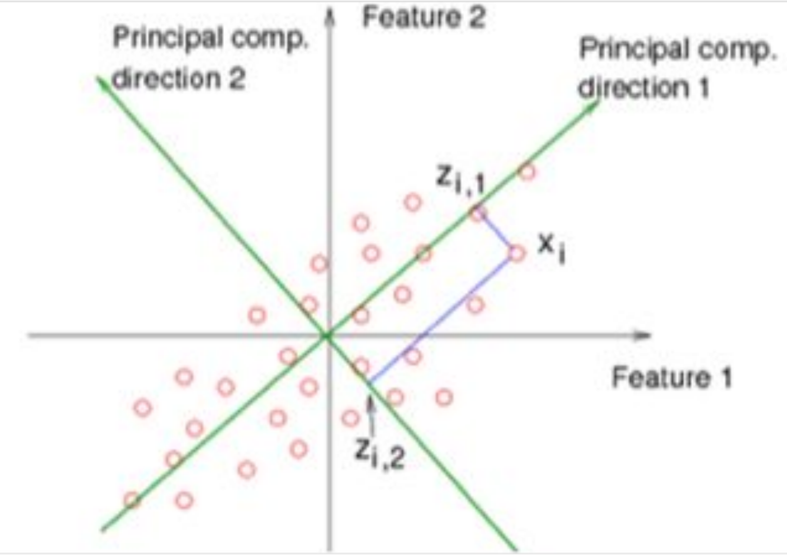
- **Why? :** Genes do not act independently, but as coregulatory "modules". E.g. in a cell type, the activity of a handful of transcription factors might lead to the co-expression of hundreds of genes defining cell-identity
- Cells occupy a low dimensional manifold in gene-expression space defined by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



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

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: an Interactive Example

PCA Explained Visually

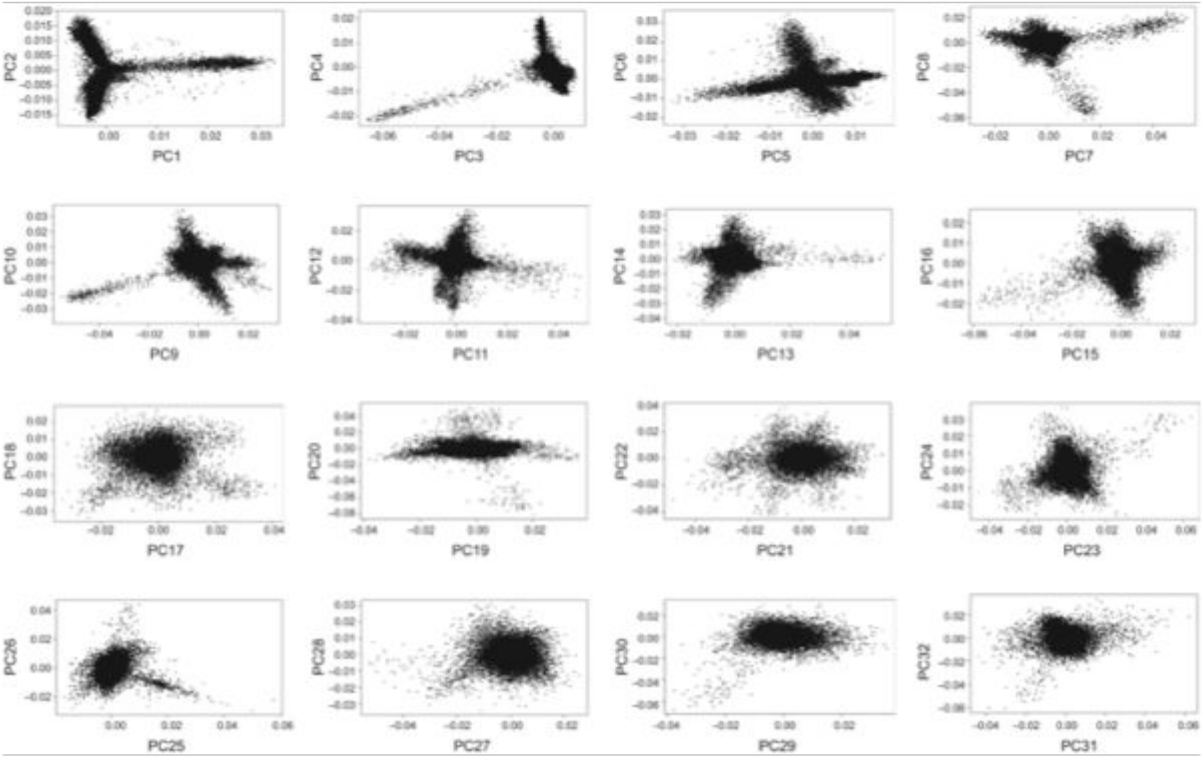
PCA: in Practice

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.

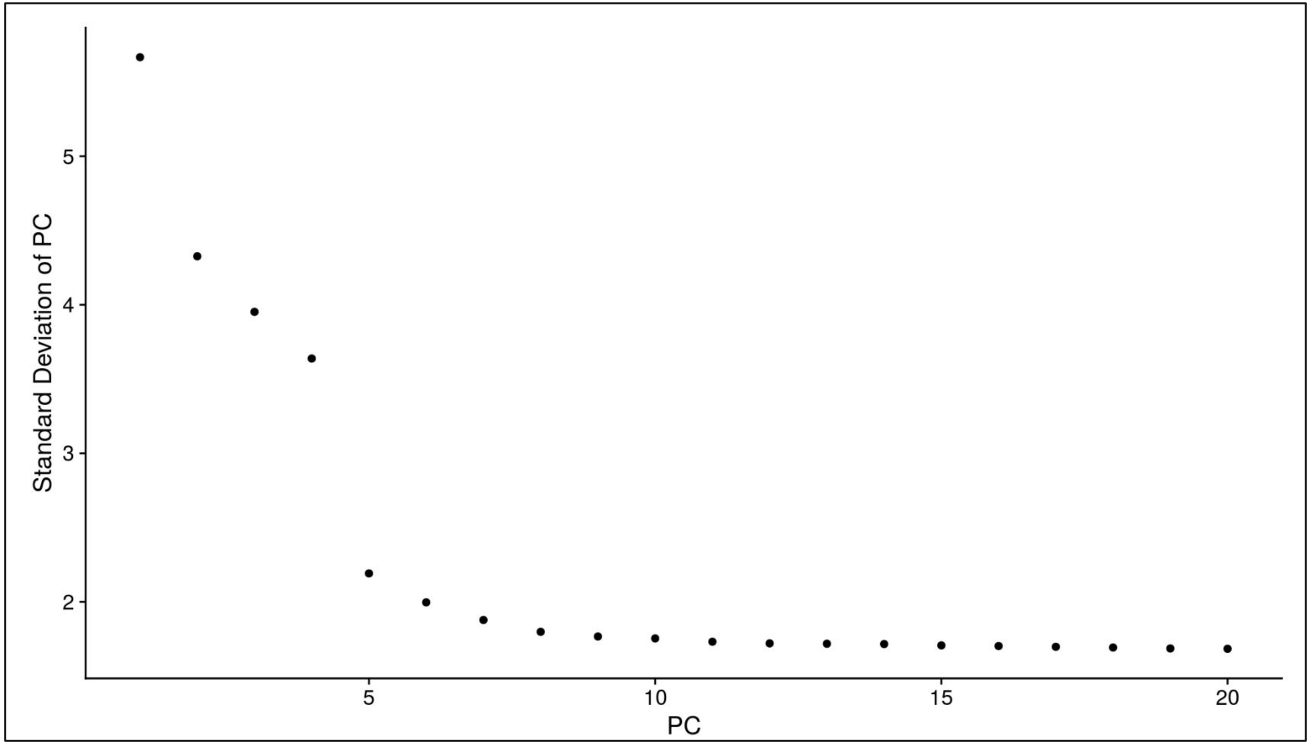
PCs

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

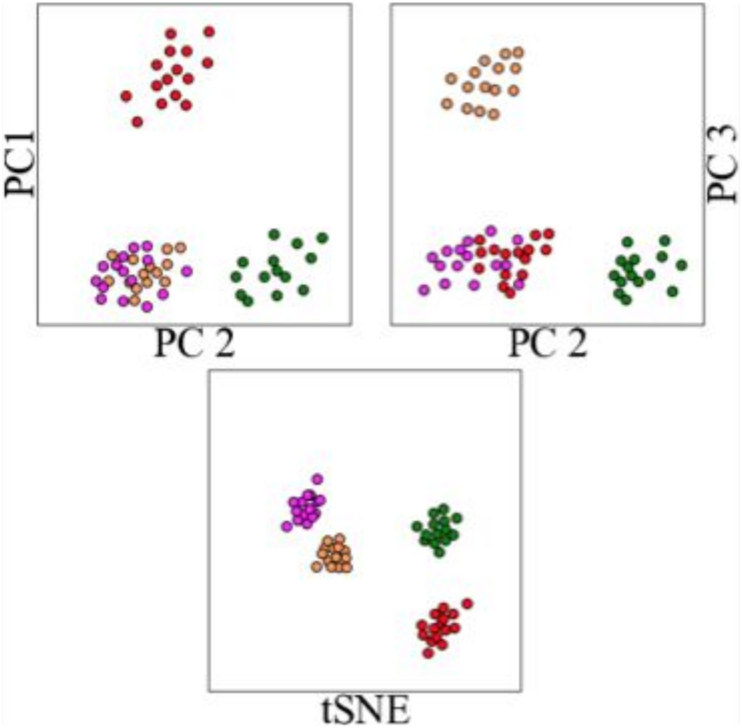


How Many Components Should We Use?

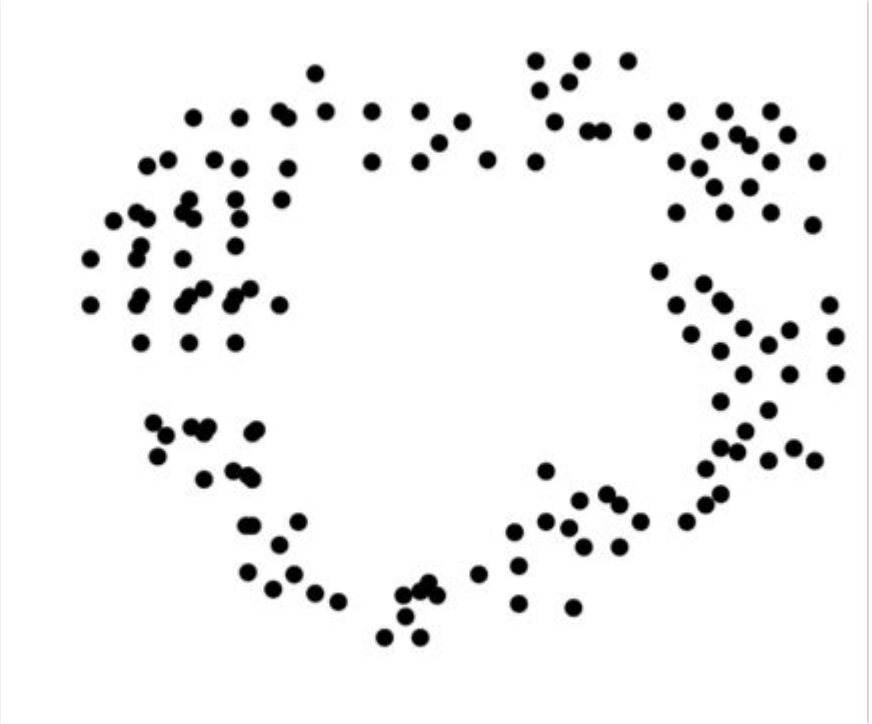
Elbow Plot (Scree Plot)



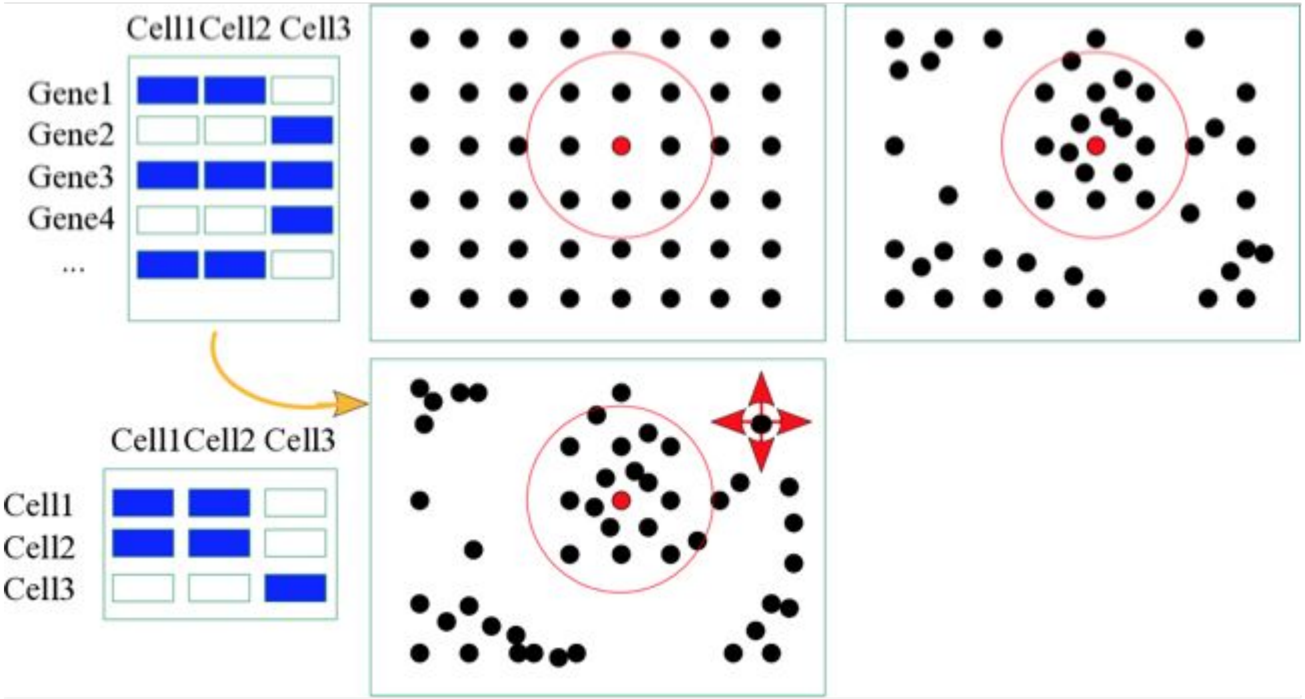
t-SNE: Collapsing the Visualization to 2D



t-SNE: Nonlinear Dimensionality Reduction



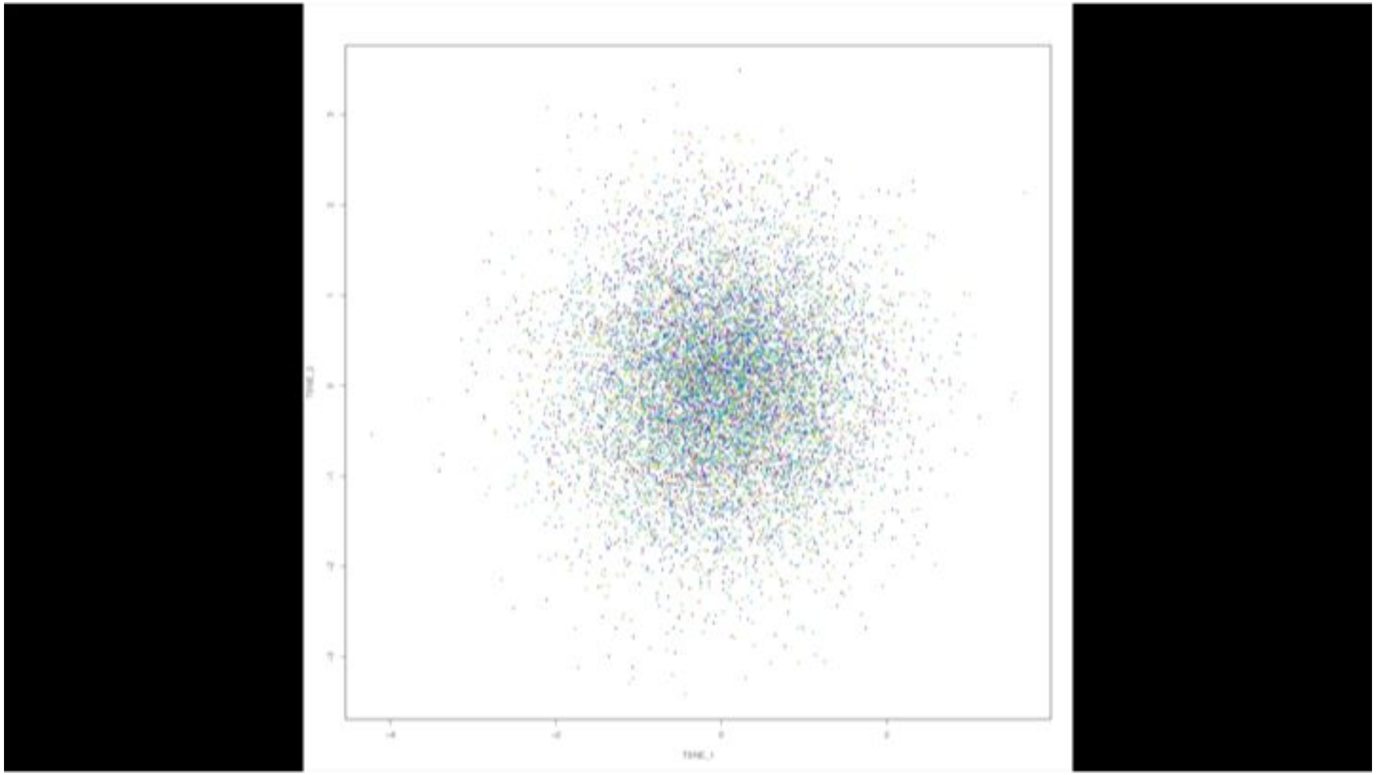
t-SNE: How it Works



	Cell1	Cell2	Cell3
Gene1	■	■	□
Gene2	□	□	■
Gene3	■	■	■
Gene4	□	□	■
...	■	■	□

	Cell1	Cell2	Cell3
Cell1	■	■	□
Cell2	■	■	□
Cell3	□	□	■

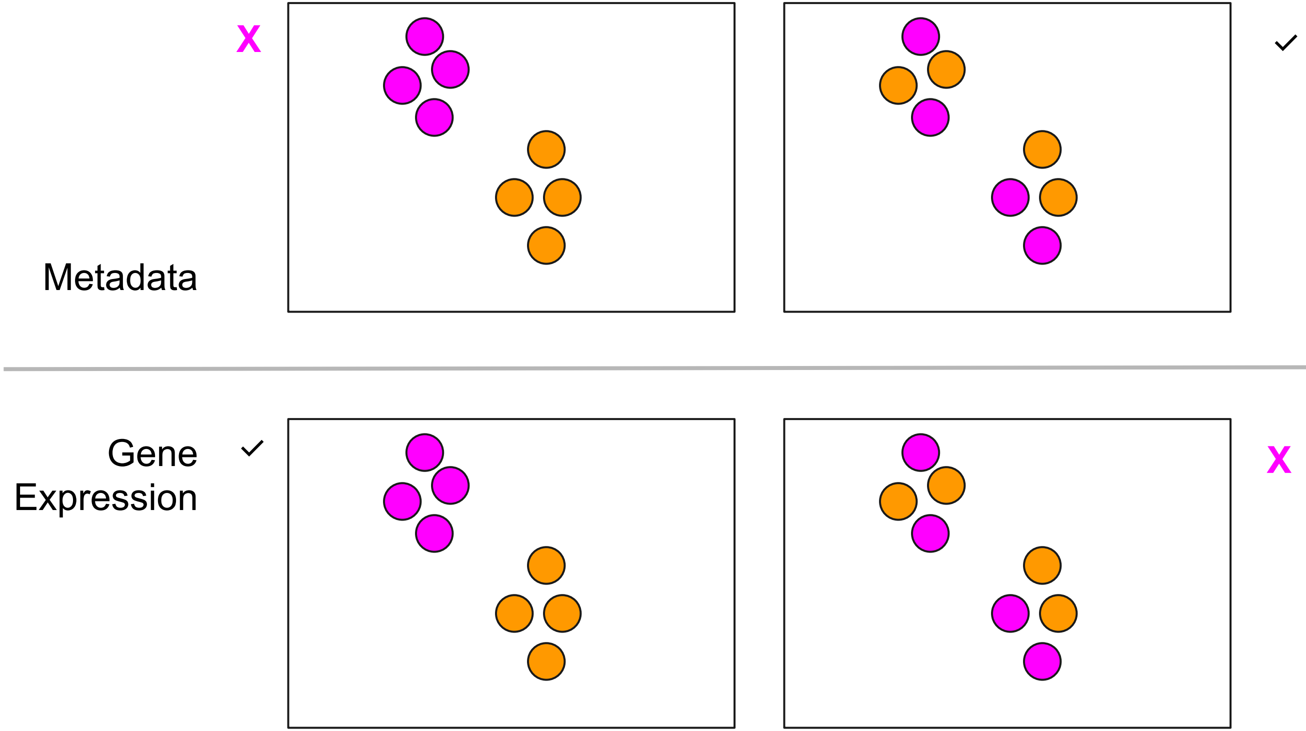
Visualizing t-SNE



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

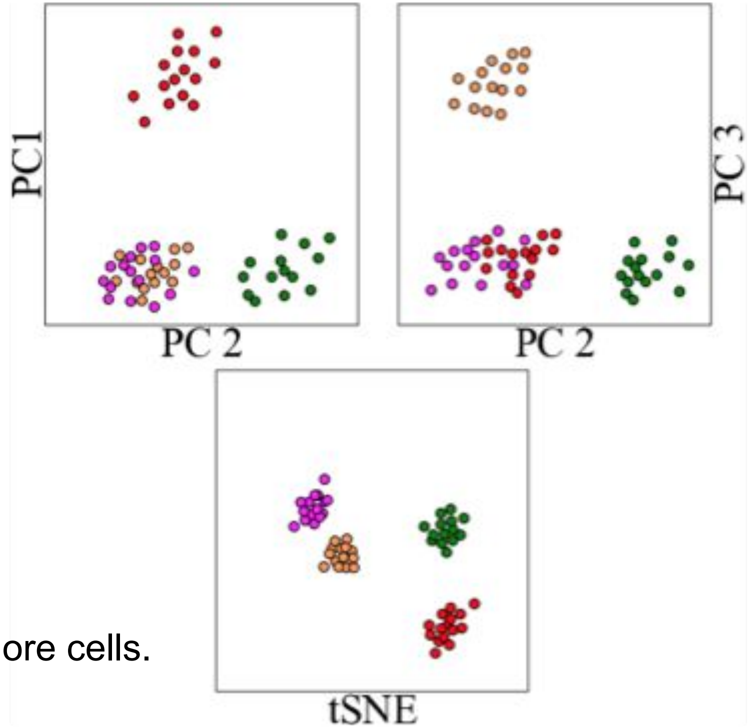
✓

Gene Expression

✓

X

Caution When Interpreting t-SNE

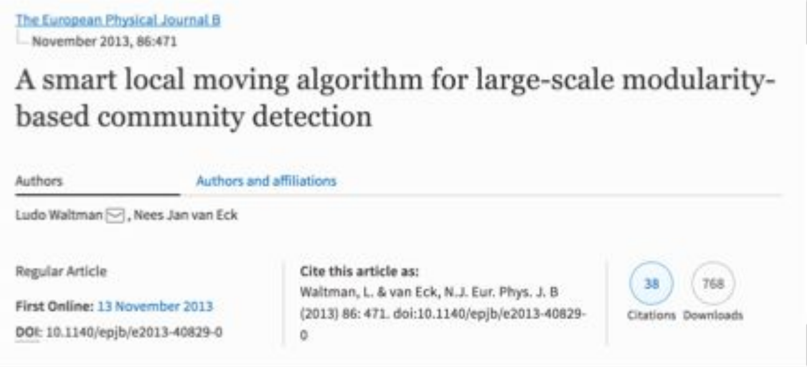


Nonlinear
Optimized for local distanct
Big clusters can just mean more cells.

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/

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

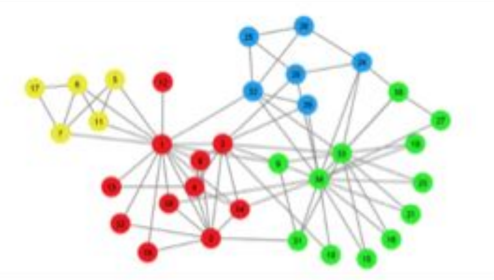
Regular Article

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

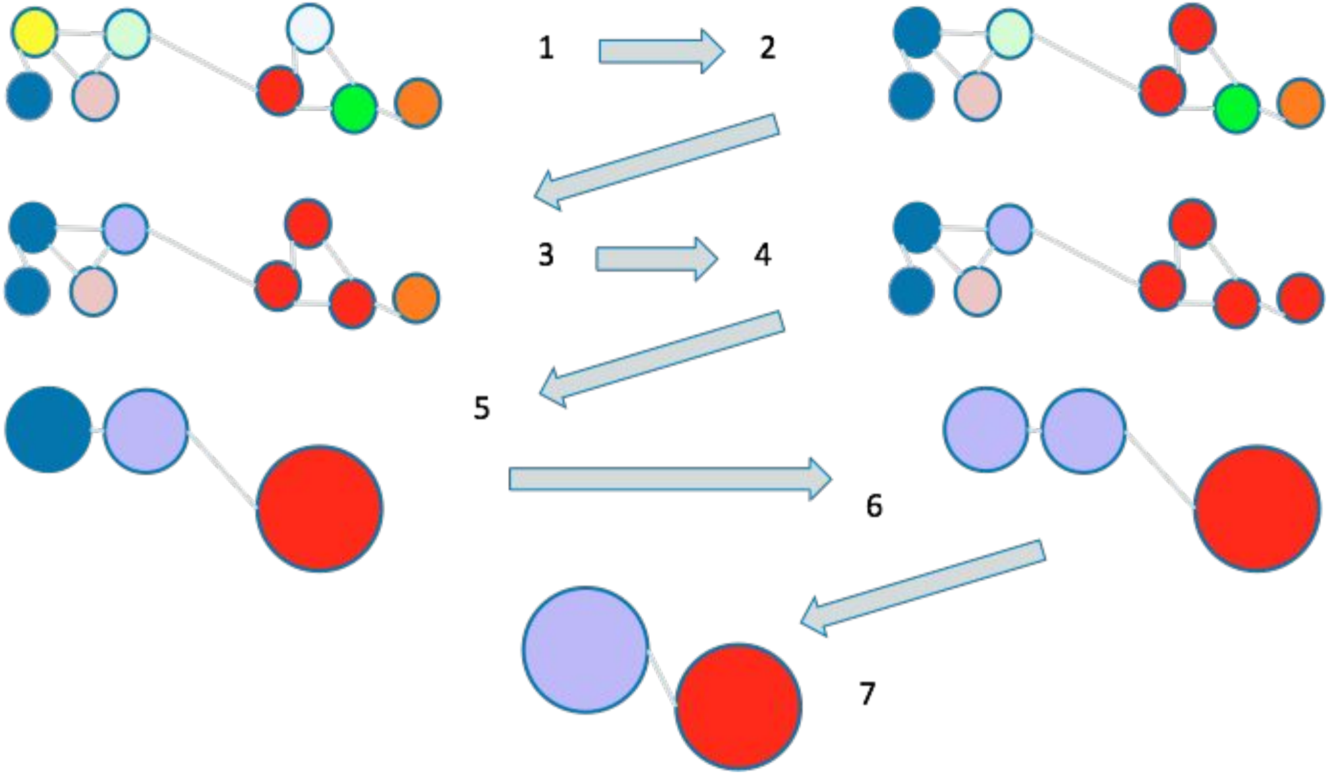
38 Citations 768 Downloads



- 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

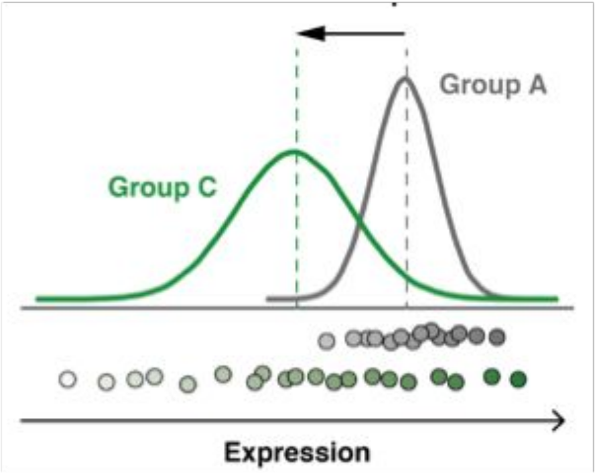


Agenda (Clustering and Differential Expression)

- Dimensionality Reduction
    - PCA
    - t-SNE
- **Differential Expression**
    - **SCDE**
    - **MAST**

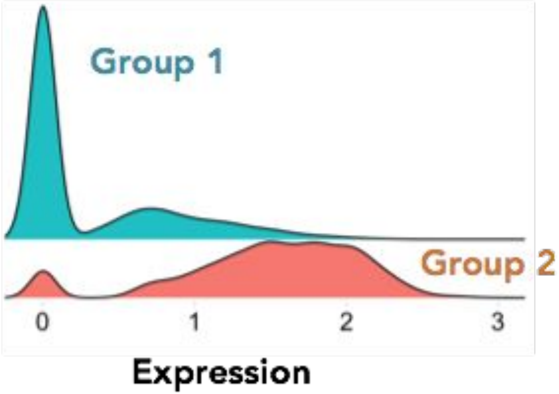
Differential Expression

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



BUT

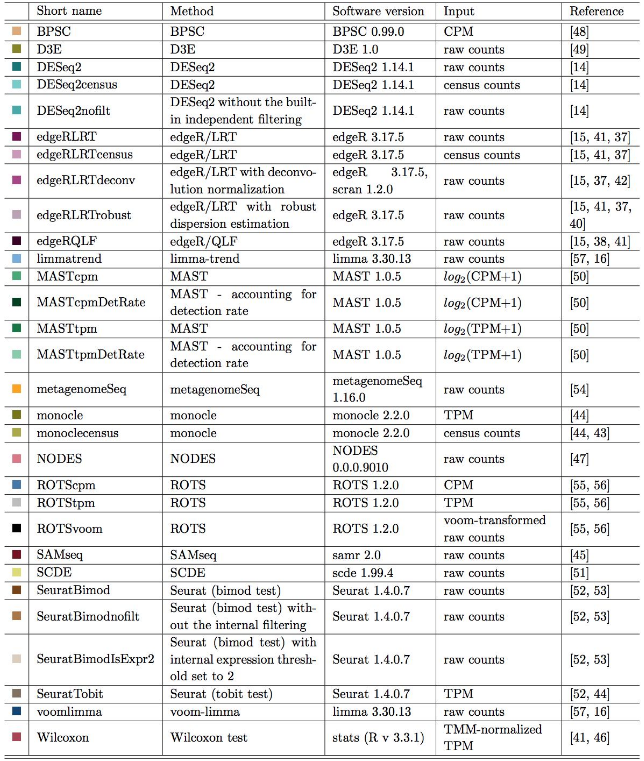
"Zero inflation" poses a challenge in single-cell data!



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

Differential Expression Analysis

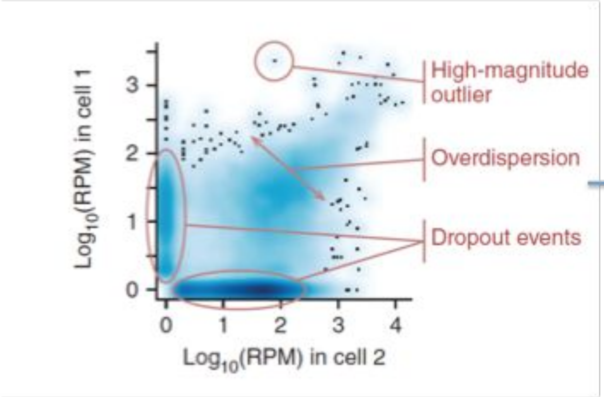
Many of the DE methods developed for bulk RNA-seq (e.g. edgeR, DE-seq) have serious limitations when applied to scRNA-seq data because of dropouts, so apply with caution!



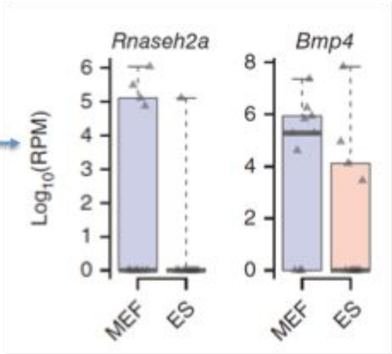
	Short name	Method	Software version	Input	Reference
■	BPSC	BPSC	BPSC 0.99.0	CPM	[48]
■	D3E	D3E	D3E 1.0	raw counts	[49]
■	DESeq2	DESeq2	DESeq2 1.14.1	raw counts	[14]
■	DESeq2census	DESeq2	DESeq2 1.14.1	census counts	[14]
■	DESeq2nofilt	DESeq2 without the built-in independent filtering	DESeq2 1.14.1	raw 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]

Soneson and Robinson, 2017

Single Cell Differential Expression (SCDE)



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



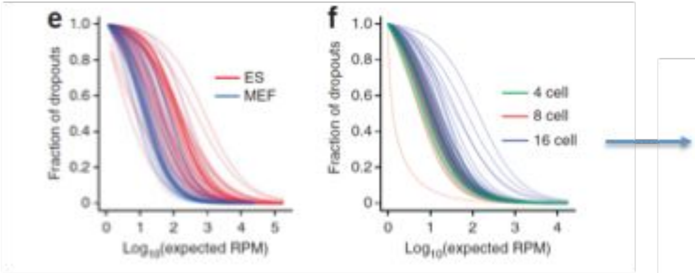
SCDE exchanges information between closely related cells to estimate dropout rates for every cell!

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

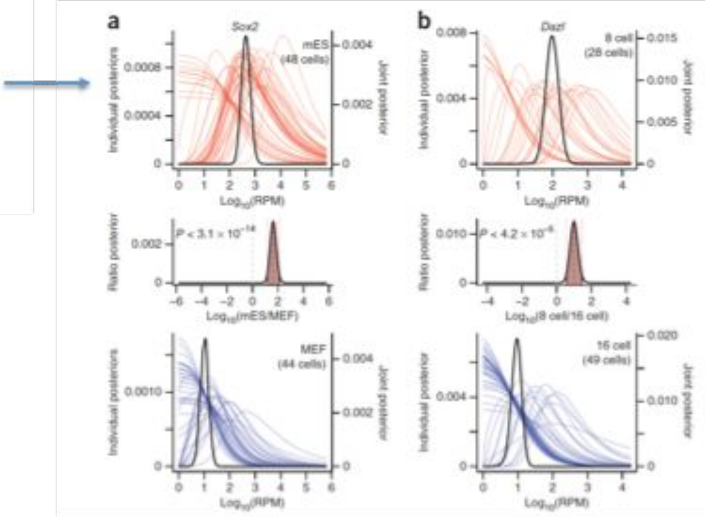
Kharchenko et al., 2014

Singe Cell Differential Expression (SCDE)

For every cell, a "dropout curve" is estimated

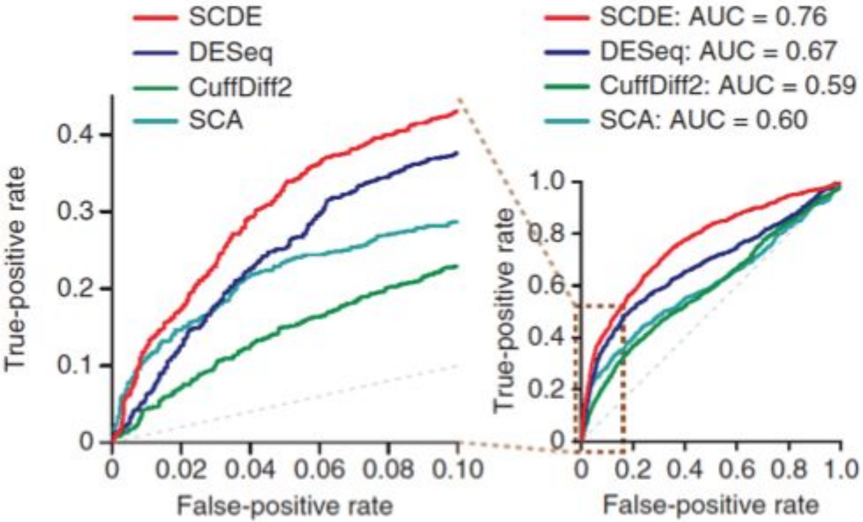


Which is used in a Bayesian framework to estimate posterior distributions for every gene in every cell



Kharchenko et al., 2014

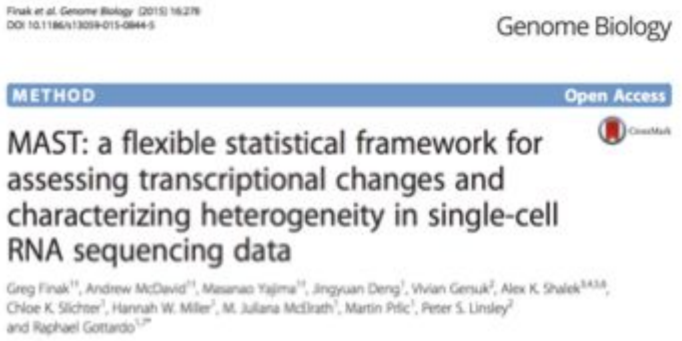
SCDE is Much More Sensitive and Specific



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 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

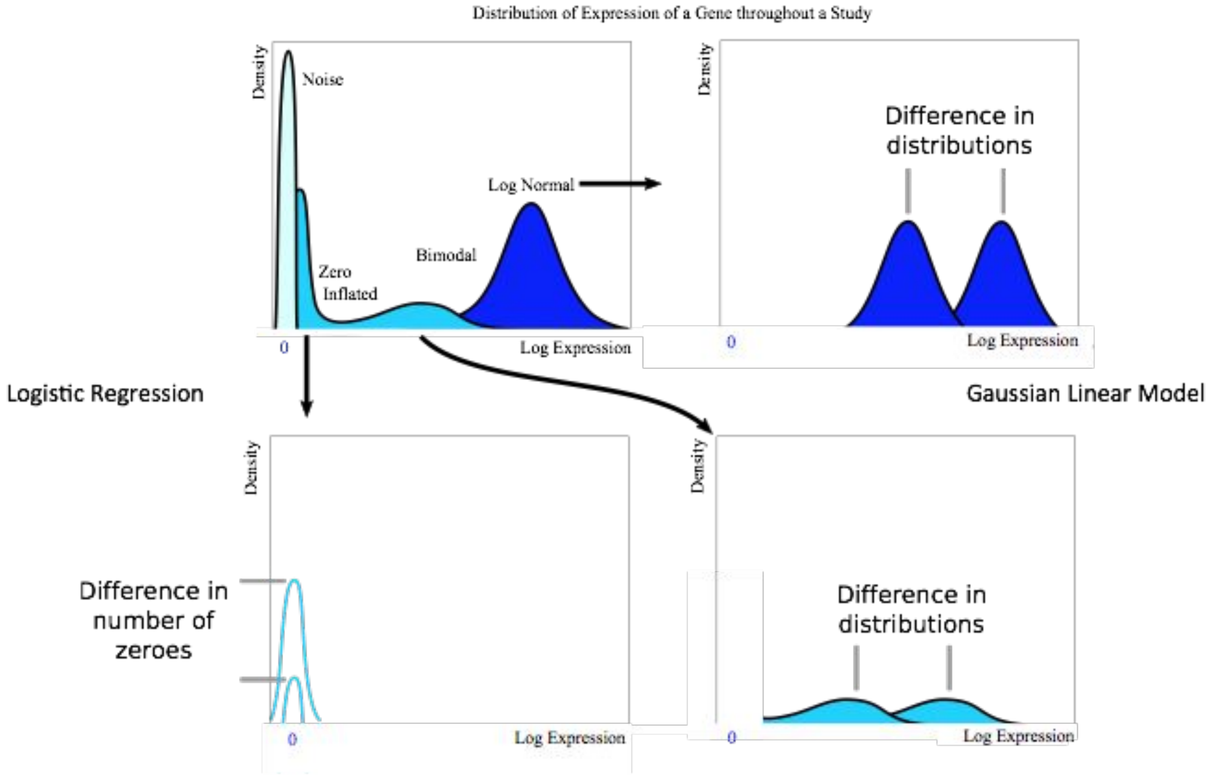
MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data

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[1], Hannah W. Miller[2], M. Juliana McElrath[1], Martin Prlic[1], Peter S. Linsley[2] and Raphael Gottardo[1,7*]

Additionally introduces a GSEA method

https://github.com/RGLab/MAST

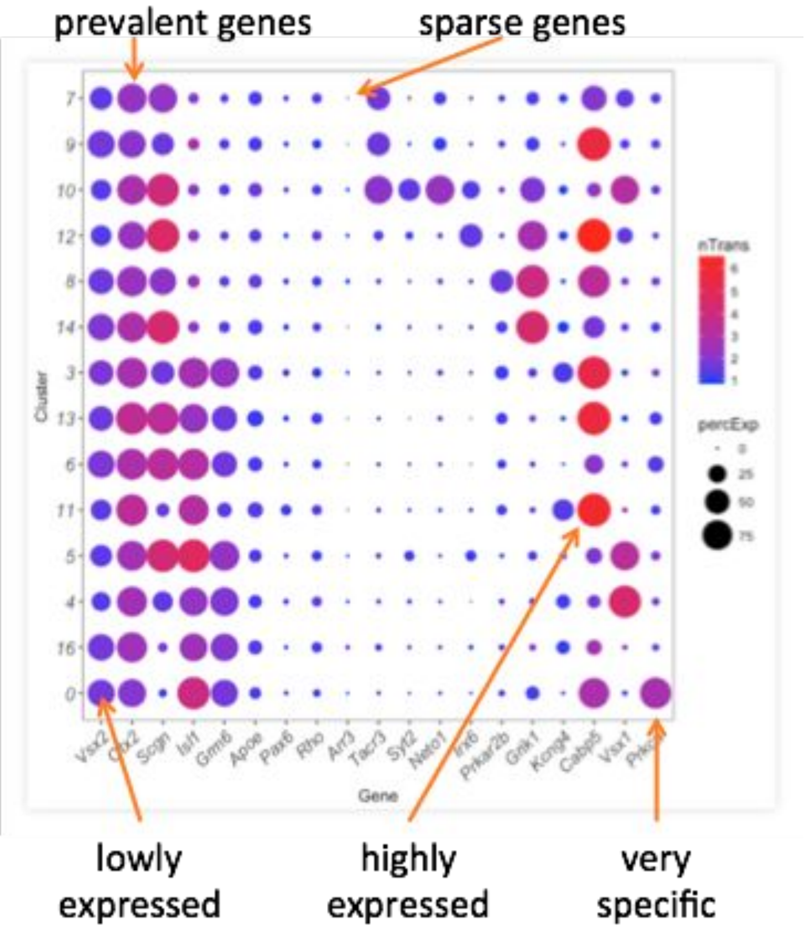
MAST: Hurdle Models



Dot Plots

Size of circle

- Gene prevalence in cluster.
  - Color of circle
- More red, more expressed in cluster.
  - Scales well with many cells.



prevalent genes

sparse genes

lowly expressed

highly expressed

very specific

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

....

Section Summary

We motivated dimensionality reduction with the helpfulness of focusing on higher variability.

We explored several methods for dimensionality reduction.

- Contrasted and showed how to leverage together.

Explored differential expression.