# Single Cell RNA-Seq: Advantages and Challenges via Overview and a Real-World Analysis Story

*Oleg Moskvin, May 22, 2017*

# Outline

- Single cell RNA-Seq: technology
- New analytical possibilities
- Challenge: gene dropouts
- Challenge: cell cycle signal
- Challenge: normalization

- **Censored**: a story of single cell analysis of blood progenitor cells

  => 19 out of 41 slides remain

# Single cell RNA-Seq: technology

- Dissociation of cells (if needed)
- Capture of single cells
- RNA extraction
- Reverse transcription
- PCR amplification (or IVT)
- Library construction + sequencing

# New analytical possibilities

- Observe cell differentiation process
- Dissect communities of individually uncultured microbes
- Dissect heterogeneous samples of other nature

# Challenge: gene dropouts

- Starting RNA amount of ~ 10 pg
- Amplification (or IVT) bias
- Zero signal (library-dependent) for actually expressed genes

# Challenge: cell cycle signal

- Cell subset identification is a key application
- The cell clustering is largely influenced by the phase of the cell cycle at which a particular cell was captured
- Methods to compensate for cell cycle signal (scLVM, replaced by ccRemover)

# Challenge: normalization

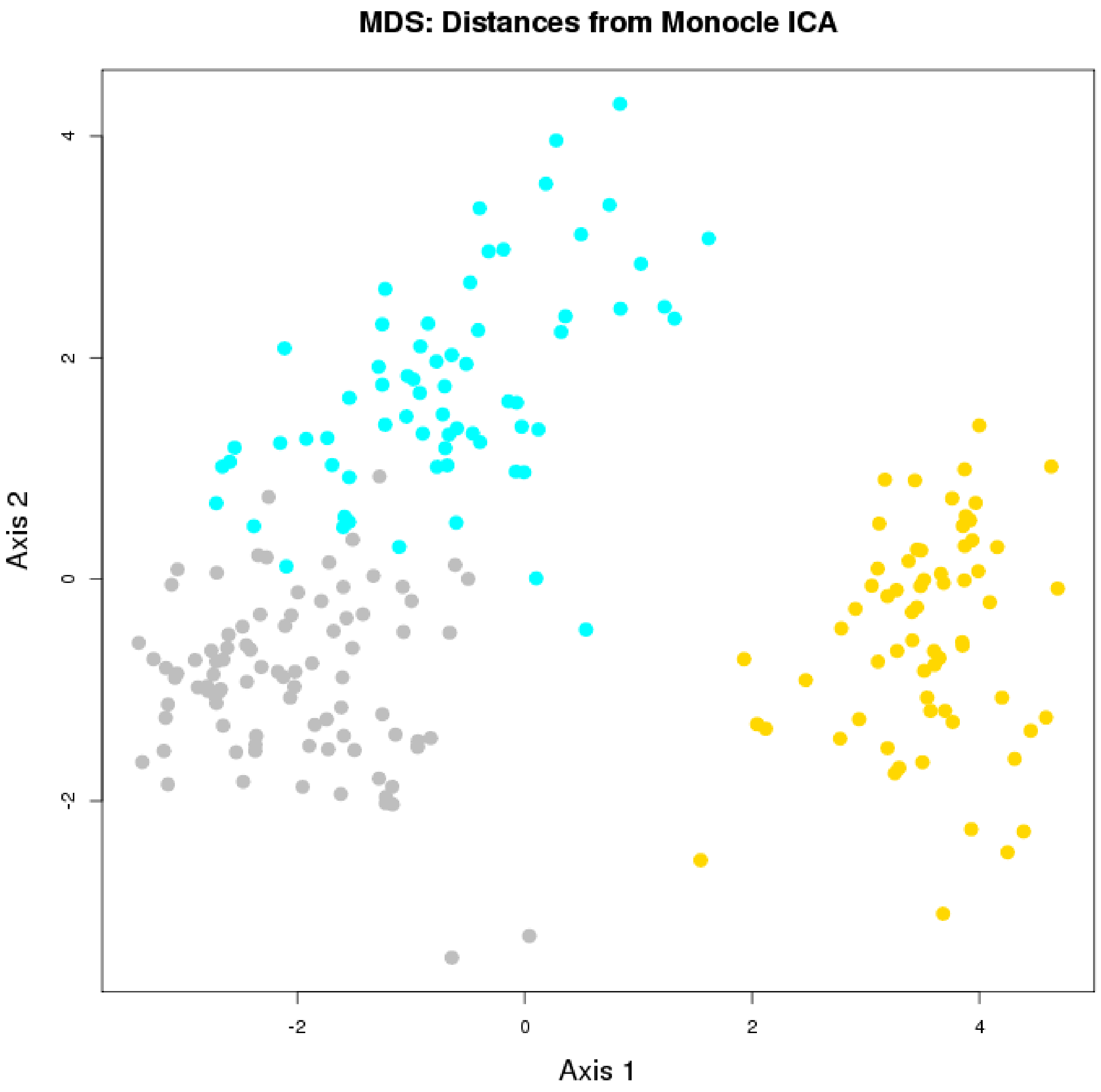
- Median expression in the library is usually zero
- Dependence of the slope of expression vs. sequencing depth relationship on the expression range (low / medium / high)

  *(SCnorm: robust normalization of single-cell RNA-seq data. Nat. Methods, Apr 17, 2017)*

Unpublished data: real-world study

(heavily truncated)

# Population detection: ICA from *monocle*



A

B

C

# SCDE: robust derivative distance measures

Kharchenko et. al. (2014) Bayesian approach to single-cell
differential expression analysis. *Nature Methods* **11**: 740–742

Models expression with a mixture of Negative Binomial and Poisson; cell-specific models

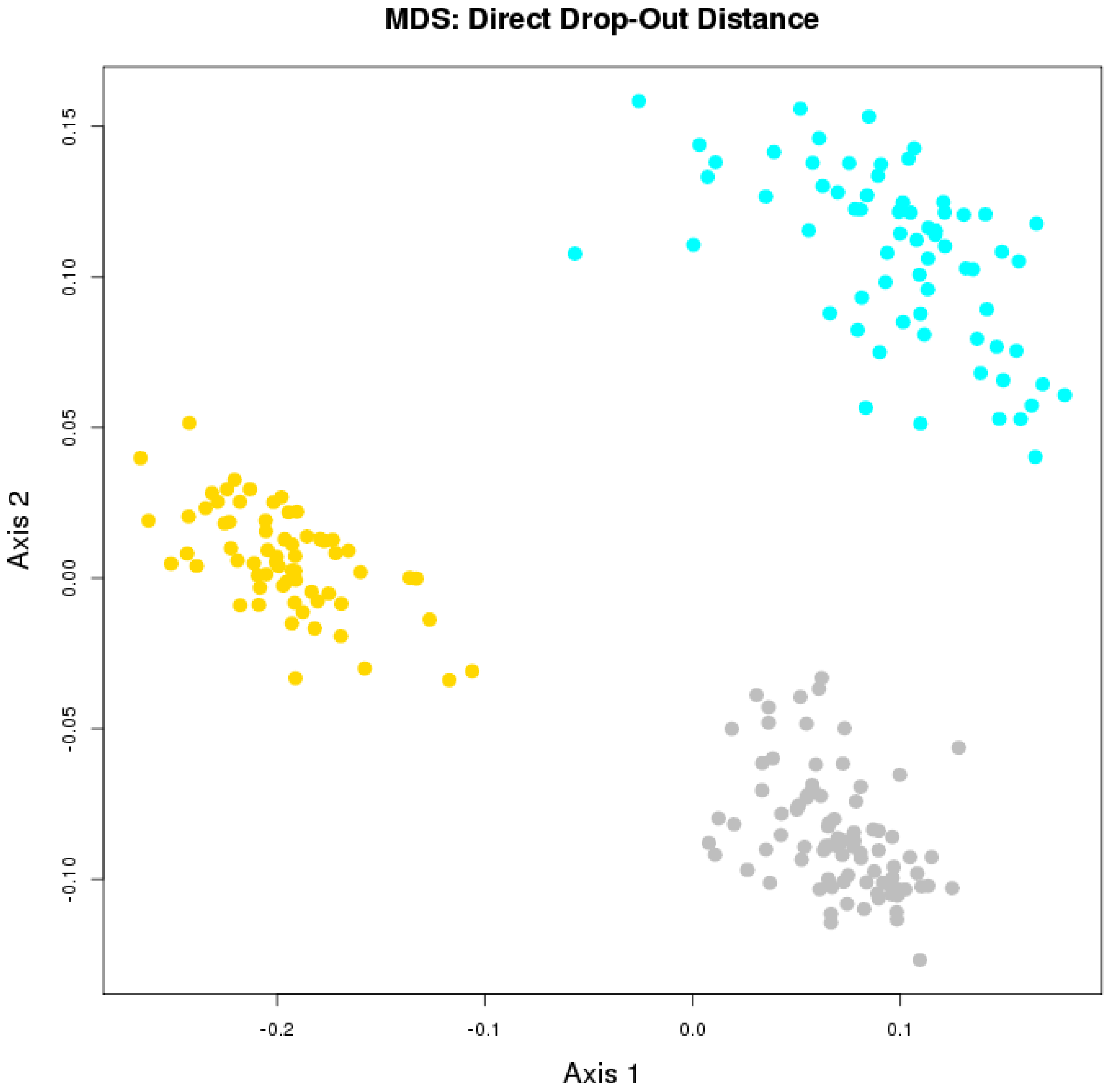
Direct Dropout distance: take dependency of drop-out probability on the average expression value; simulate the drop-out events, replacing them with NA values; calculate correlation using the remaining points

Reciprocal Weighting: give increased weight to pairs of observations where a gene expressed (on average) at a level x1 observed in a cell c1 would not be likely to fail in a cell c2, and vice versa [via “corr” from *boot*]

Mode-Relative Weighting: *combine dropout probabilities computed for individual cells separately and using joint posterior modes for each gene, for correlation weighting*

The definitions of (and the code for) the 3 robust distances
are back online after moving the project to GitHub:
http://hms-dbmi.github.io/scde/diffexp.html

# Population detection:
# Direct Dropout distance from SCDE

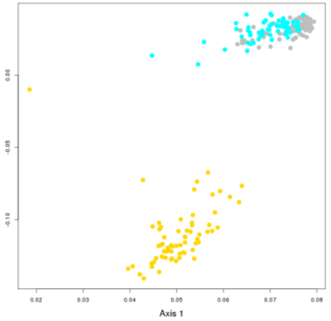


A

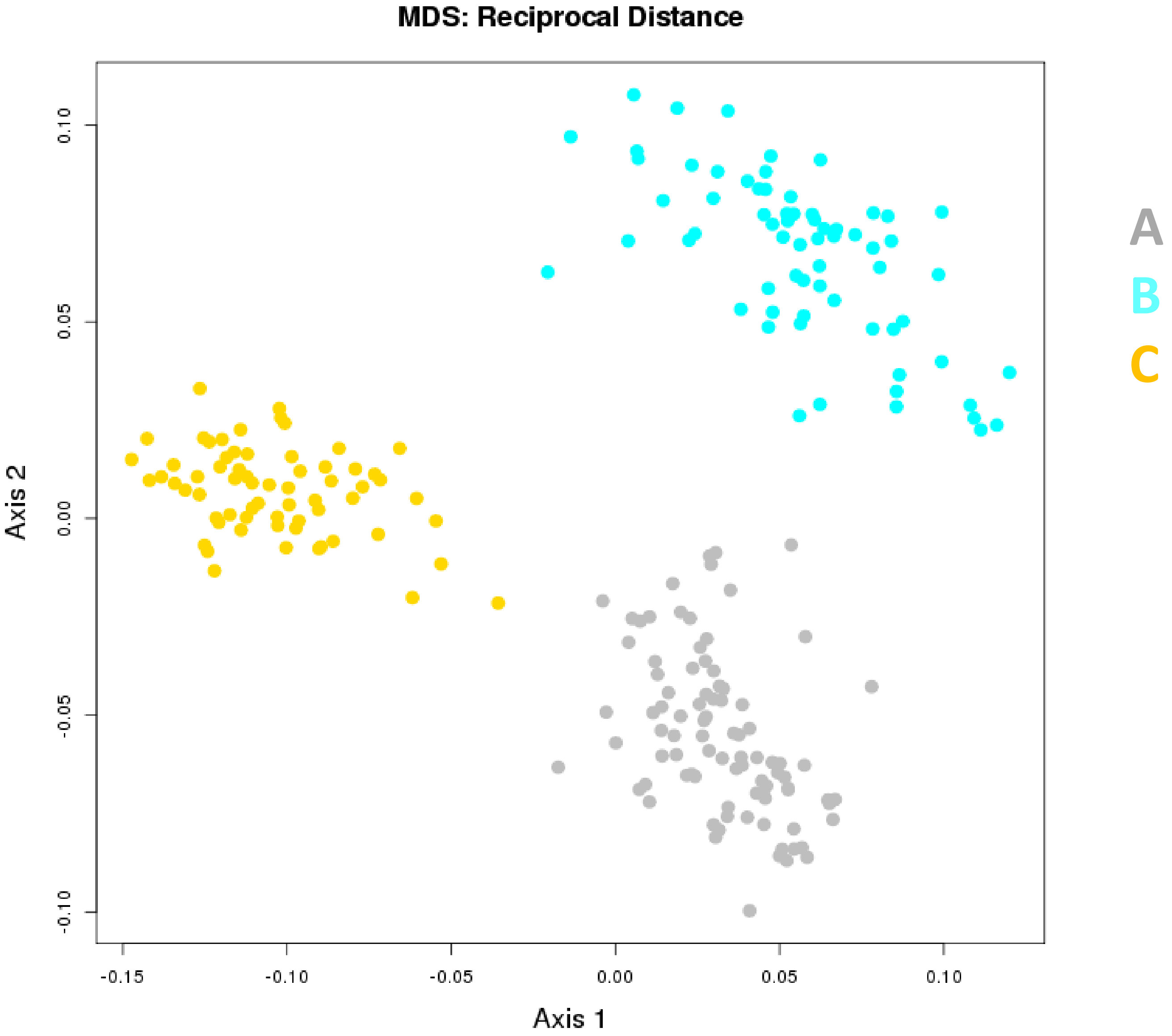
B

C

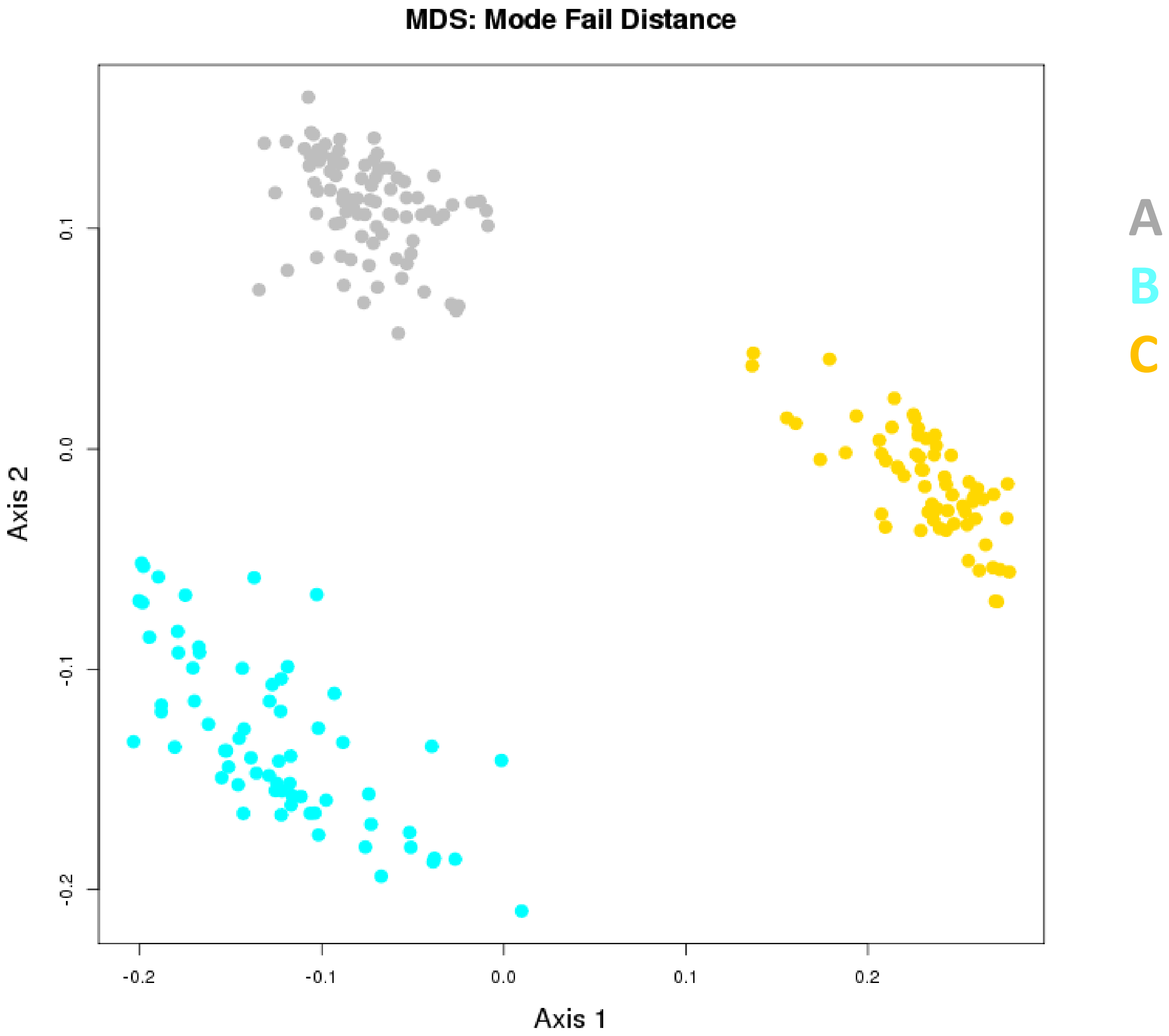
Compare to:
Count-based PCA



# Population detection:
# Reciprocal Weighting distance from SCDE



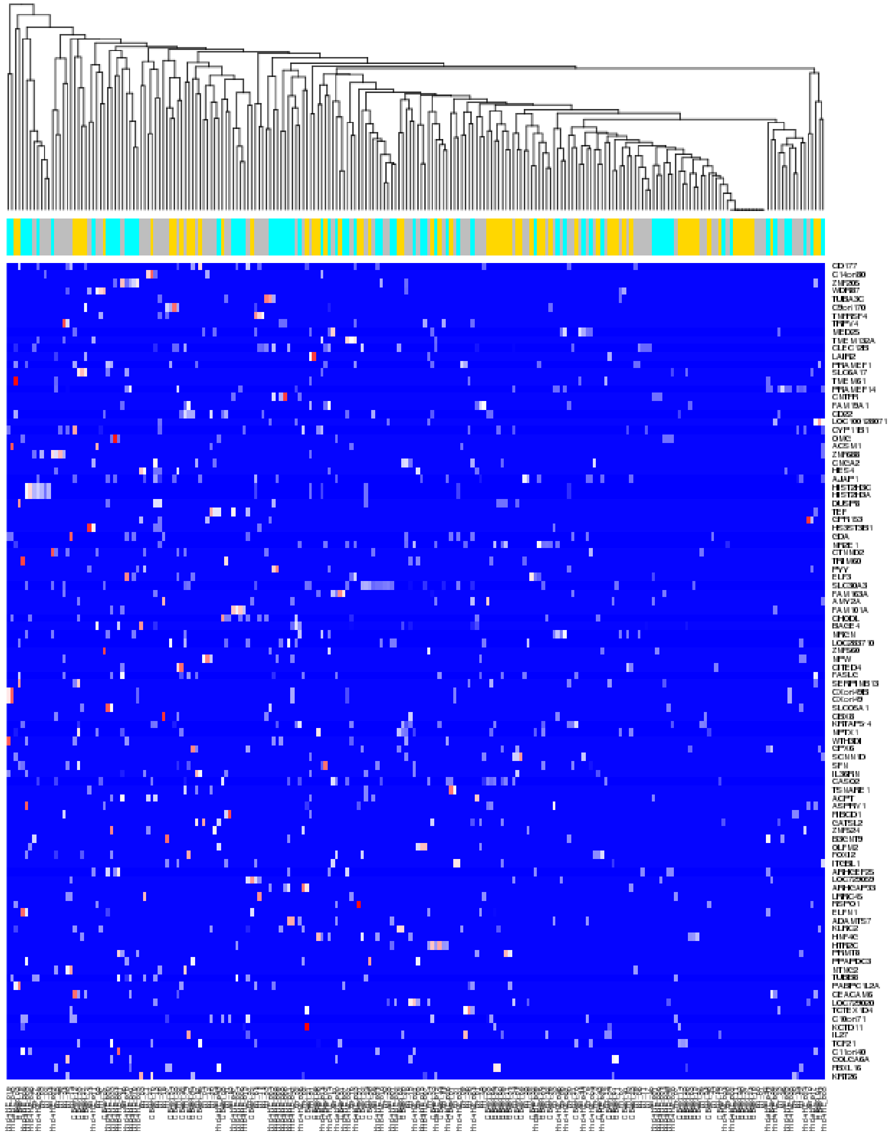
# Population detection:
Mode Relative distance from SCDE



# Another manifestation of apparently general principle of usefulness of low-quality information combined with a probabilistic model?

| Problem | Method | Nature of low-quality information | Advantage visible in |
|---|---|---|---|
| Assigning sequencing reads to transcripts | RSEM | Parts of a sequencing read with lower base calling scores | Better correlation of the resulting vectors of counts between biological replicates |
| Detecting population structure | SCDE | Genes with high dropout rate | Better separation of the phenotypically distinct populations |

# Hierarchical clustering: the popular "top variable genes" approach doesn't work! – example of top 100 genes by variance



A

B

C

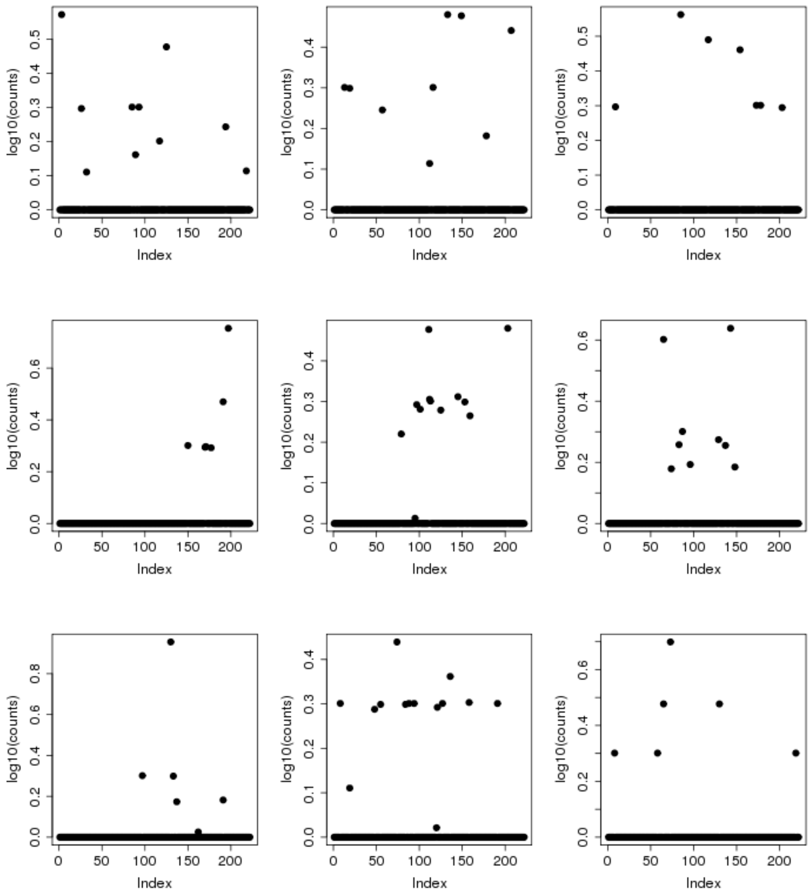
In a recent Kharchenko lab's ssRNA-Seq workshop (Nov. 3, 2015) –

http://hsci.harvard.edu/event/single-cell-genomics-workshops -

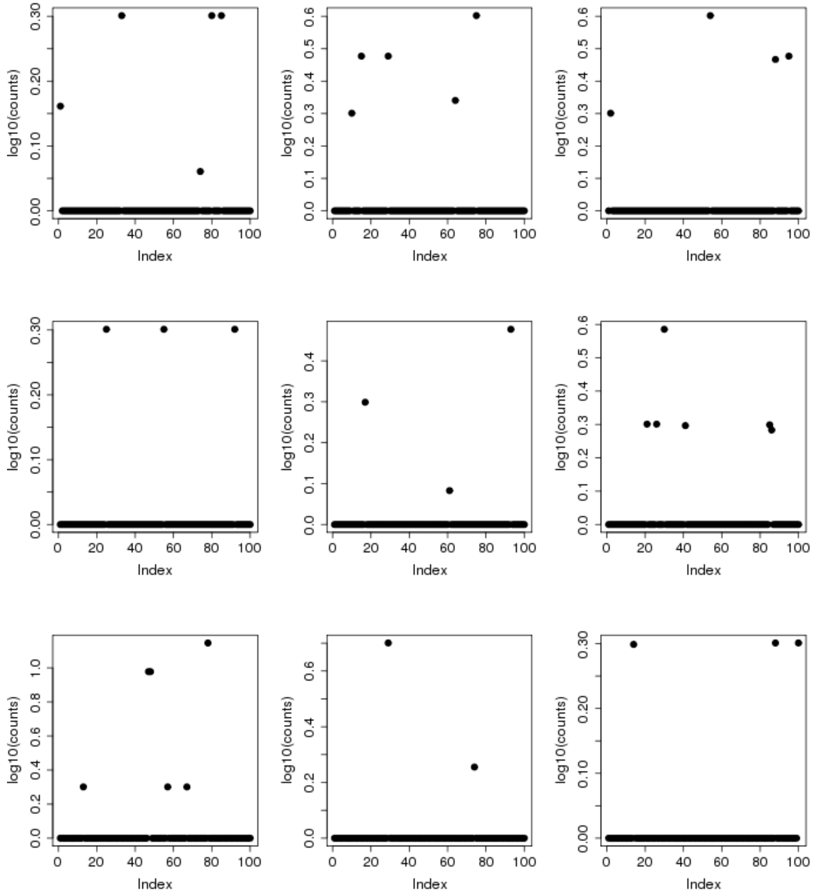
a poor performance of the hierarchical clustering of ssRNA-Seq data based on top variable genes was also pointed out

# Why the "top variable genes" approach doesn't work with ssRNA-Seq?

*Row-wise slice:*
*Expression of top 9 genes by variance across the 221 cells*

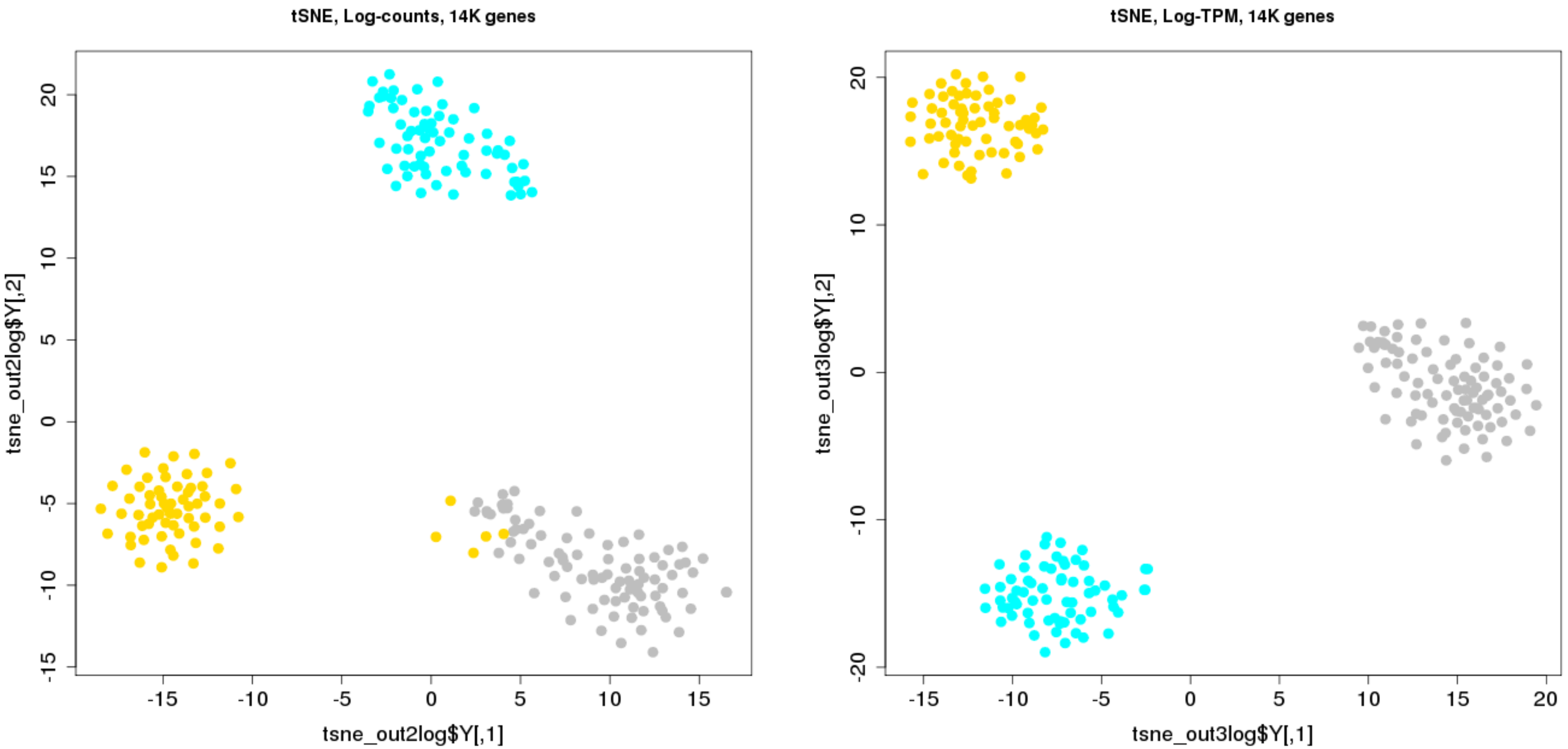


*Column-wise slice:*
*Expression of the top 100 variable genes in 9 randomly selected cells*



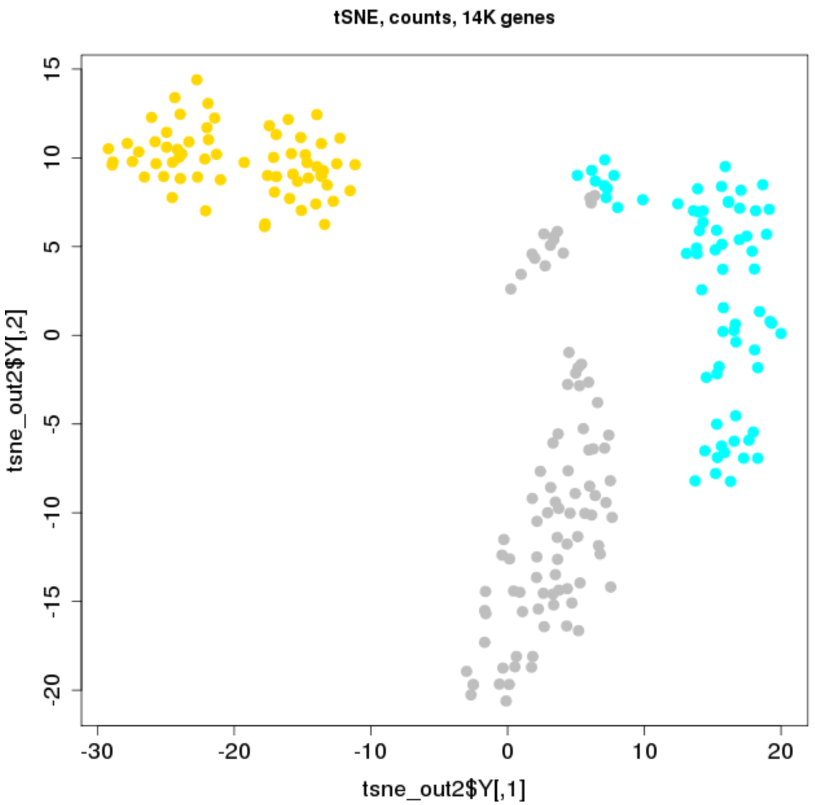
And this is based on the 14k gene dataset (*__after__* the gene pre-filtering)!

# t-SNE, 14K genes, log-transformed data

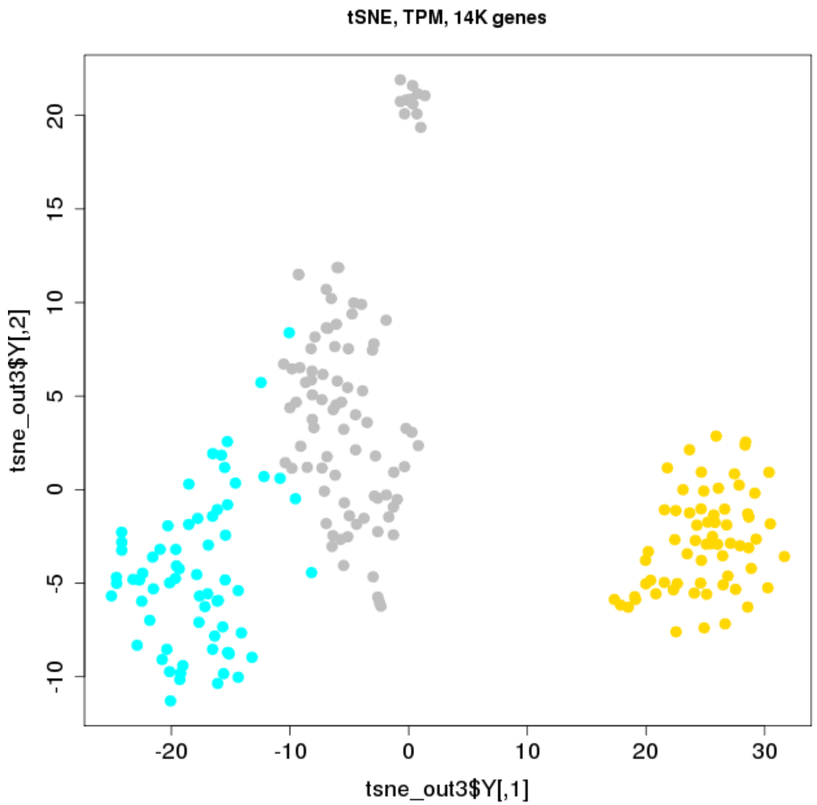


tSNE promise: *“Retaining both the local and the global structure of the data in a single map”*
(van der Maaten, 2008, J. of Machine Learning Research 9: 2579-2605 )

# t-SNE, 14K genes, linear data



tSNE, counts, 14K genes



tSNE, TPM, 14K genes

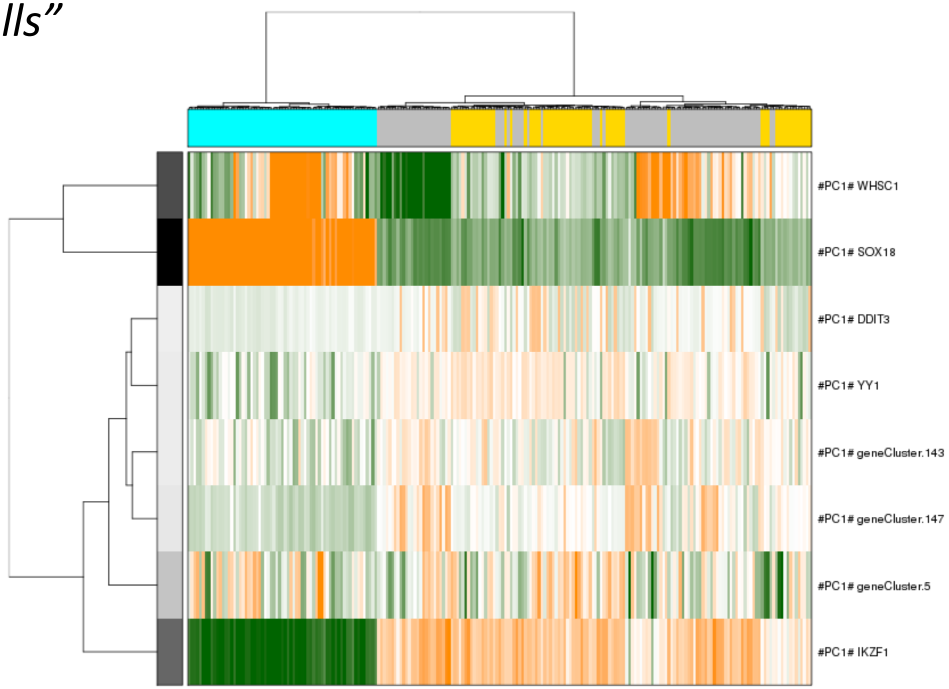
tSNE delivers its promise with NON-log-transformed count data!
(needs high dynamic range to output both global and local structures?)

# PAGODA approach

Latest addition to SCDE package targeted at functional analysis

Fan J et. al. (2016) Characterizing transcriptional heterogeneity
through pathway and gene set overdispersion analysis.
*Nature Methods*, Jan 18.

Address multiple functionality representations with ambition
*“ to resolve multiple, potentially overlapping aspects of transcriptional heterogeneity by testing gene sets for coordinated variability among measured cells”*



*Problem (while testing it in action): Seems to get stuck with PC1*