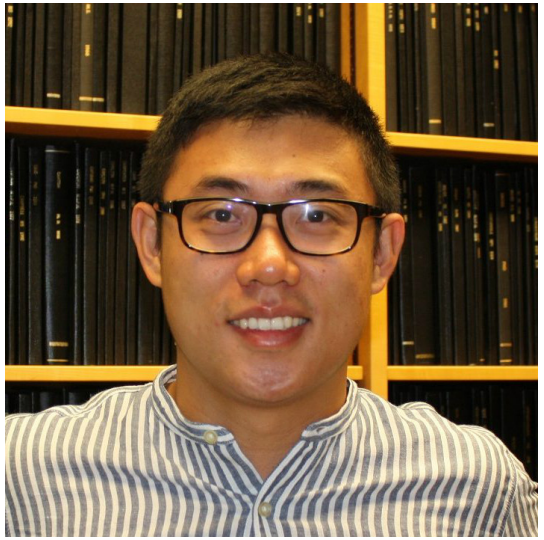
Recent Advances in Statistical Methods and Computational Algorithms for Single-Cell Omics Analysis

ISMB 2019

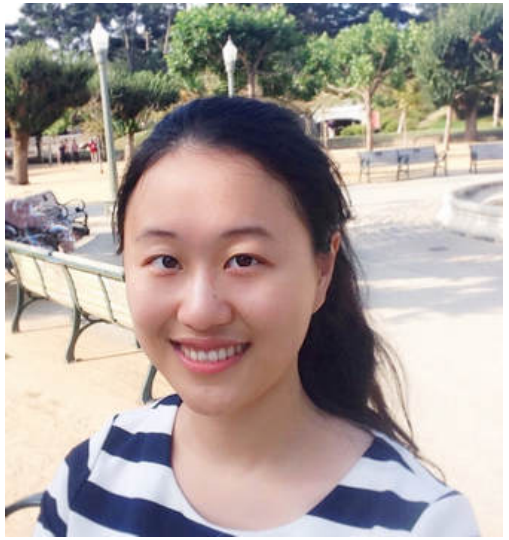
Sunday, July 21, 2019

Introduction

Rhonda Bacher, PhD
Assistant Professor
University of Florida
rbacher@ufl.edu



Yuchao Jiang, PhD
Assistant Professor
University of North Carolina
yuchaoj@email.unc.edu



Jingshu Wang, PhD
Assistant Professor
University of Chicago
jingshuw@upenn.edu

Materials

- All slides presented here are available at Github:

https://github.com/rhondabacher/ISMB2019_SingleCellTutorial

Additional tutorials and links also available.

Course overview

- This tutorial is focused on advanced statistical and computational methods that are recently developed for single-cell omics data.

- This tutorial is intended for an audience with genomics/computational background, who are interested in cutting-edge developments of single-cell research, including both method development and application.

- Advanced tools that are recently developed in the field will be taught from a high-level perspective.

Schedule

9:00 – 9:10 am	Introduction: tutorial infrastructure setup (RB).
9:10 – 9:40 am	Overview of technologies for scRNA-seq data generation; types of analysis that can be carried out; data normalization, spike-ins, and technical artifacts (RB).
9:40 – 10:00 am	Data visualization, including UMAP, t-SNE, etc. (JW).
10:00 – 10:30 am	Denoising, batch correction (JW).
10:30 am – 11:00 am	Autoencoder and transfer learning for scRNA-seq (JW).
11:00 – 11:15 am	**Coffee Break**
11:15 – 11:40 am	Pseudotime reconstruction, cell ordering (RB).
11:40 – 12:00 pm	ScRNA-seq in immunology (VDJ, cell surface protein, RB).
12:00 – 12:30 pm	Methods for scATAC-seq analysis and multimodal alignment of single-cell transcriptomic and epigenomic data (YJ).
12:30 – 1:00 pm	Single-cell omics analysis in cancer, including assessing cancer heterogeneity and inferring tumor phylogeny by scRNA-seq, and profiling copy number changes by scDNA-seq (YJ).

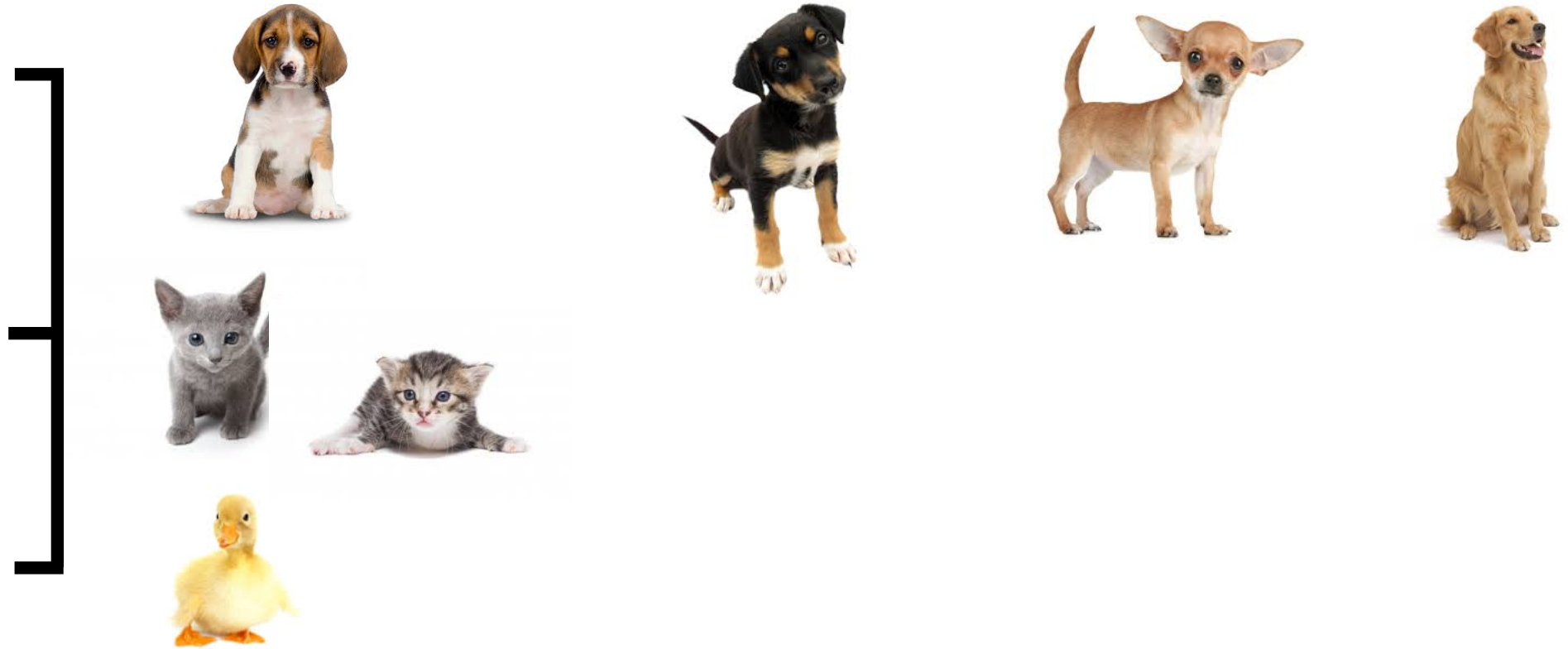
Why single cells?

Population

Bulk RNA-seq sample

Why single cells?

Single-cell RNA-seq



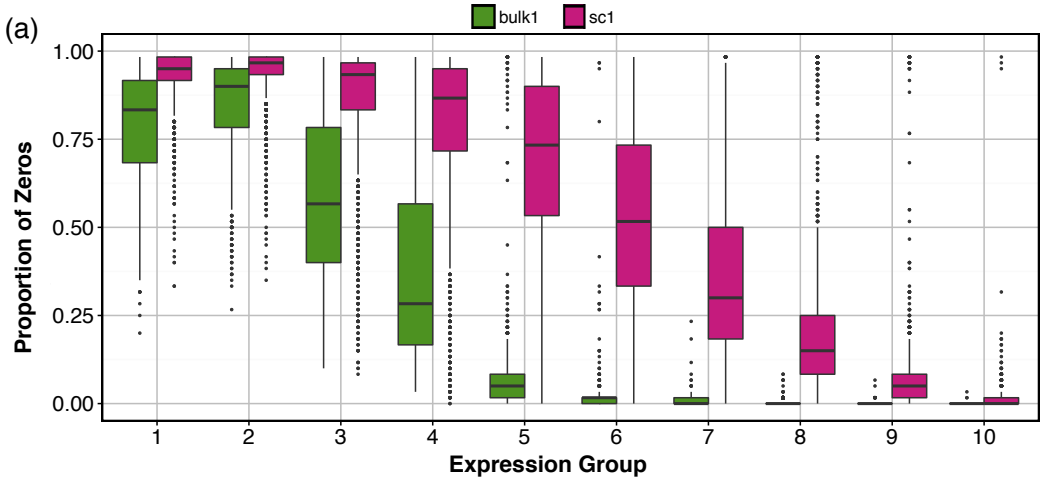
Major applications of single cell RNA-seq

- Cancer
  - Detecting rare tumor cells.
  - Study intratumor heterogeneity (clonal subpopulations).
  - Dynamics of differentiating cancer stem cells.

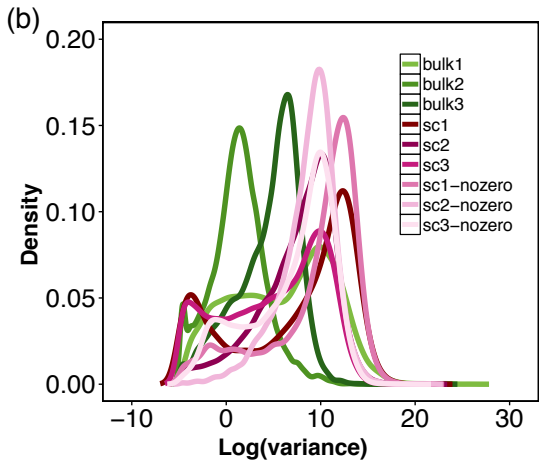
- Immunology
  - Population composition across diseases/states.
  - Study cell transitions and commitments.

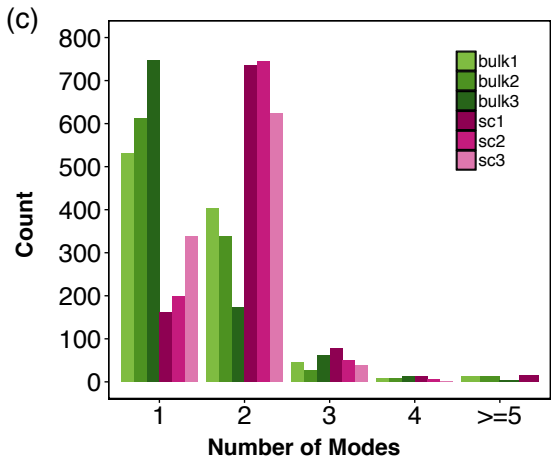
Features of scRNA-seq data

Abundance of zeros



Increased variability



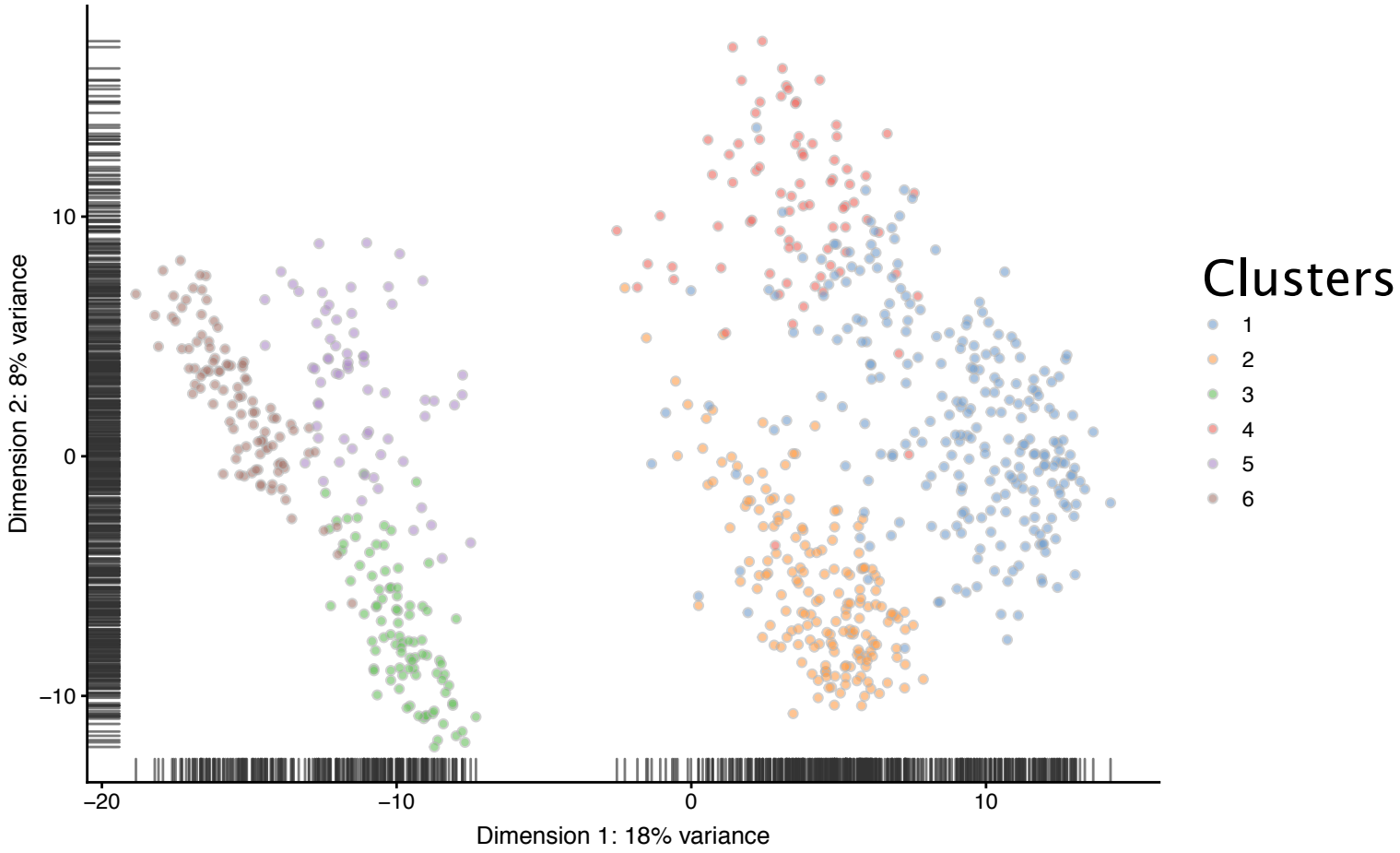


Heterogeneous expression distributions

Bacher, R. & Kendziorski, C., *Genome Biology,* 2016.

Types of analyses for scRNA-seq

Discovery of novel cell subpopulations



Discovery of novel cell subpopulations

- Unsupervised clustering of cells based on their transcriptome profiles.
- Challenges:
  - Large dimensions.
  - Noisy data (technological + biological).
  - Unknown number of true clusters.

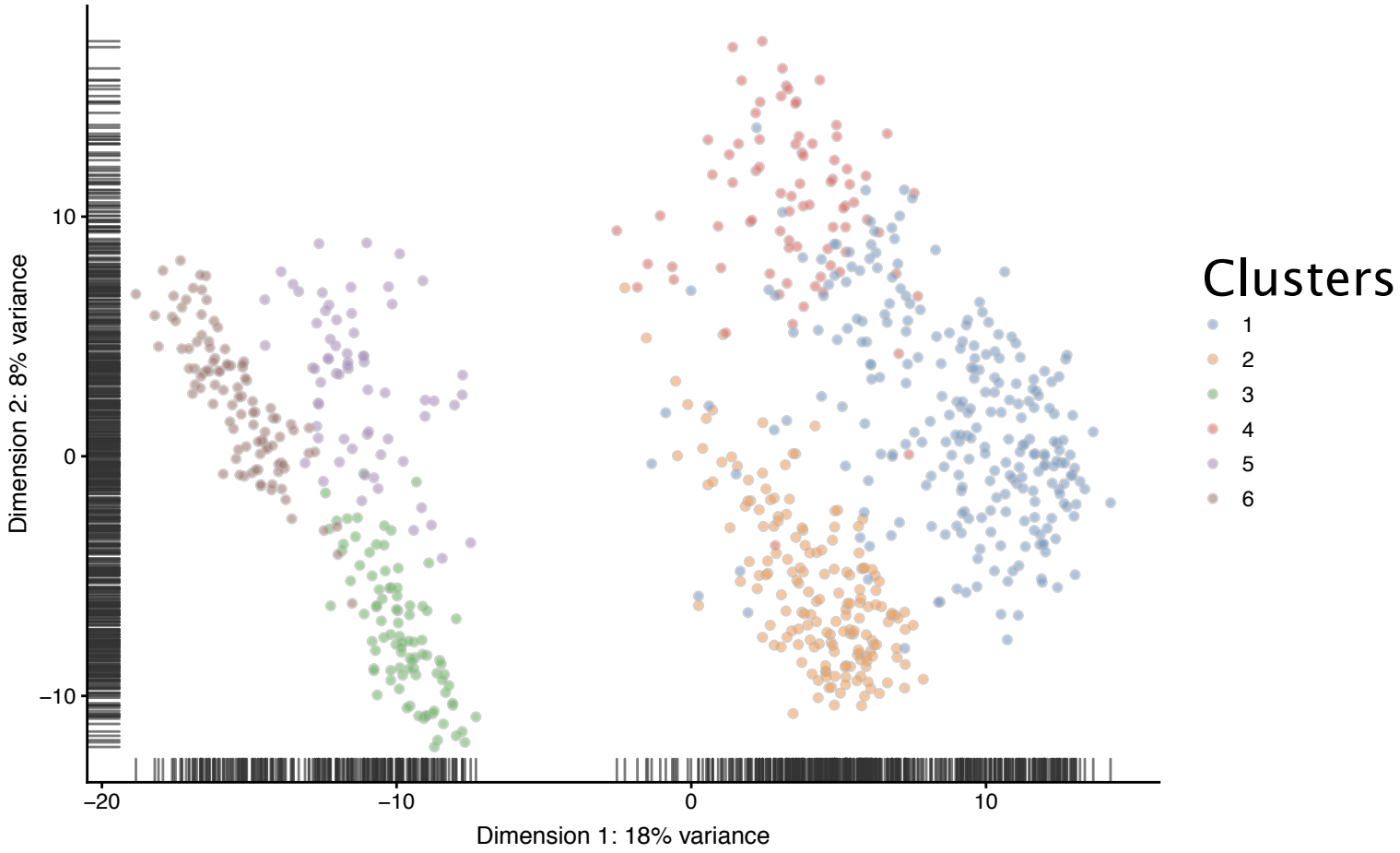
Discovery of novel cell subpopulations

- Denoise the data using statistical methods (**JW**).
- Filter out high noise or low variability genes.
- Reduce the dimensionality using Principle Component Analysis (PCA).
- Cluster cells.
- Visualize the cells and their clusters using t-SNE or UMAP (**JW**).

Clustering methods for scRNA-seq

- Hierarchical clustering and k-means:
  - SINCERA, pcaReduce, CIDR
- Graph-based methods:
  - SNN-Cliq, Seurat
- Semisoft:
  - SOUP
- Consensus Clustering:
  - SC3

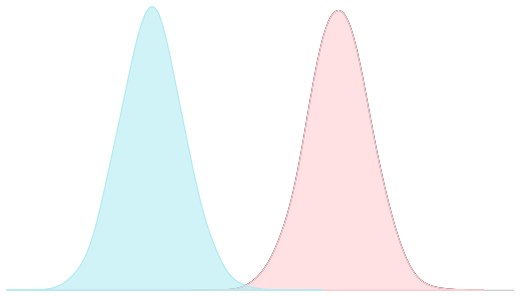
Discovery of novel cell subpopulations



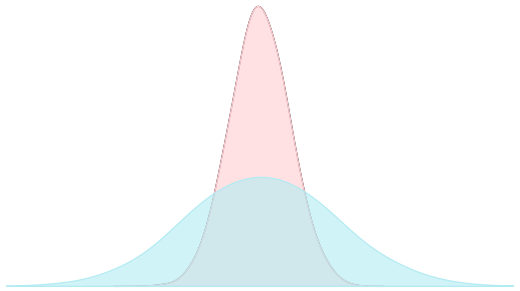
Differential gene expression

- Identify highly or lowly variable genes.
- Identify genes differentially expressed across clusters of cells – beyond means.

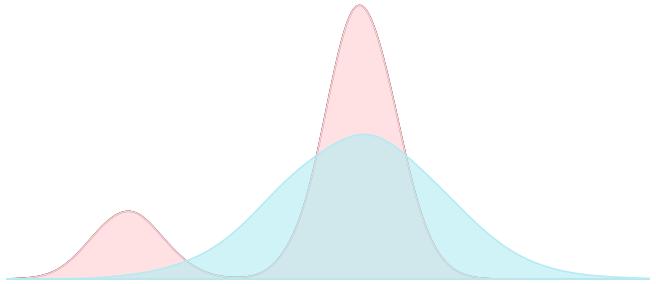
Differential mean



Differential variability



Differential modes



Differential gene expression methods

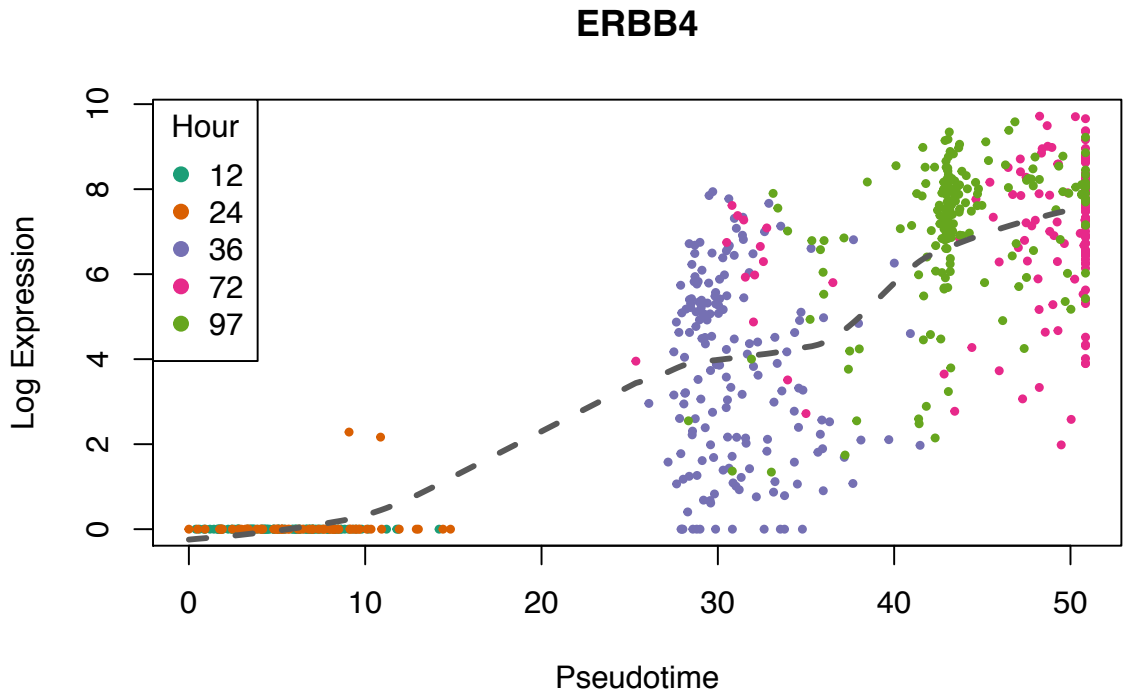
- Means:
  - SCDE, MAST, BASiCS, DECENT
- Variability:
  - BASiCS
- Distribution:
  - scDD, descend

Pseudotime analysis (RB)

- More generally referred to as trajectory inference methods.
- Assuming cells gene expression varies along an underlying dynamic.
  - Time: Respond to stimuli, e.g. drug, growth factors.
  - Space: Embryonic development.

Pseudotime analysis

- Take advantage of the heterogeneity observed to reconstruct a path through the observed cells.
- Identify differentially dynamic genes

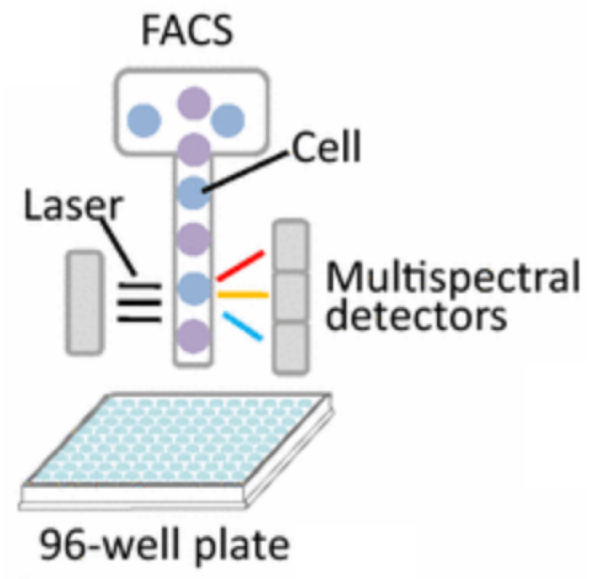


Technologies– Key Differences

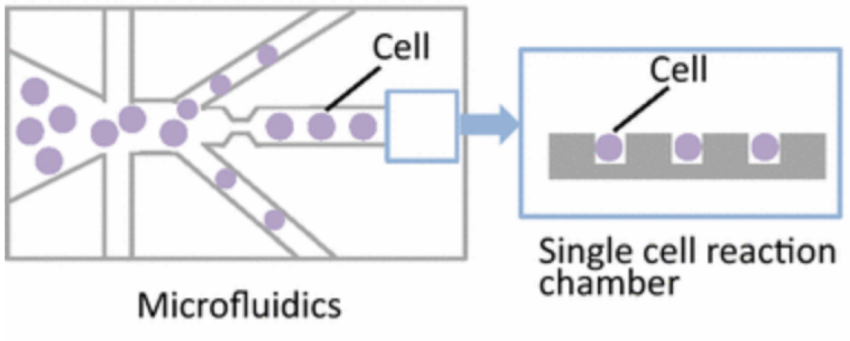
How to capture a single cell

Plate based and Microwells:

- Each cell has a physical location.
- Low throughput but more flexible.



Fluidigm C1

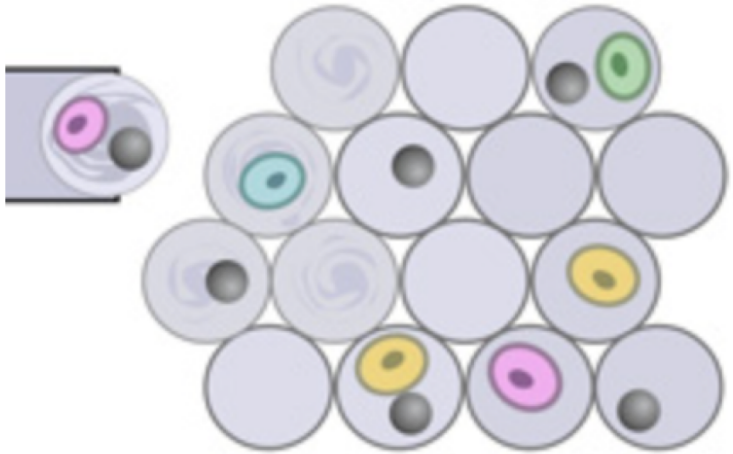


Images from: https://www.mybiosource.com/learn/testing-procedures/single-cell-sequencing/

How to capture a single cell

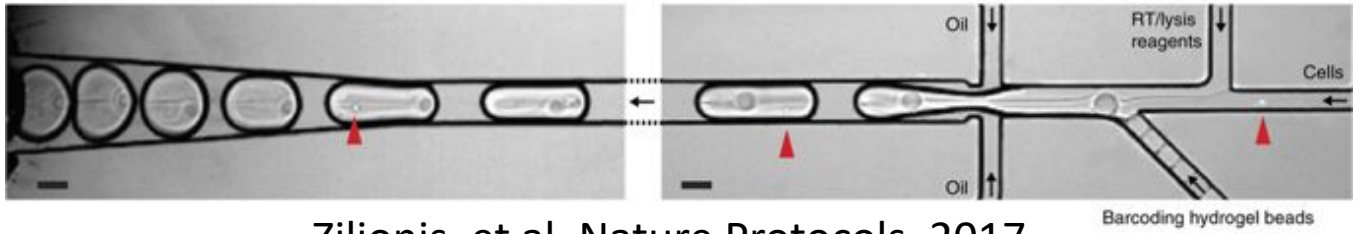
Droplets:

- High throughput and lower cost per cell.
- Capture of cells is about 50% of input.



Droplets

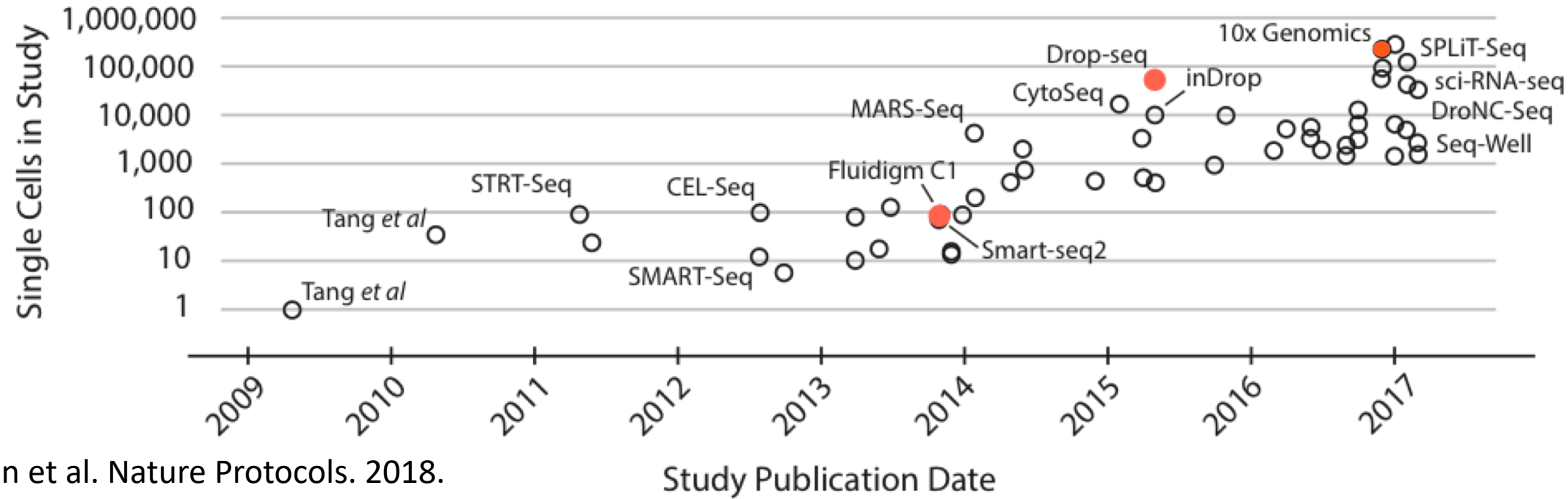
Macosko et al. Cell. 2015.



Zilionis, et al. Nature Protocols. 2017.

Key platform differences

- Throughput: Number of cells prepared in a fixed amount of time.
  - Plate based: 10^2 – 10^3
  - Droplet methods: 10^3 – 10^5



Svensson et al. Nature Protocols. 2018.

Key protocol differences

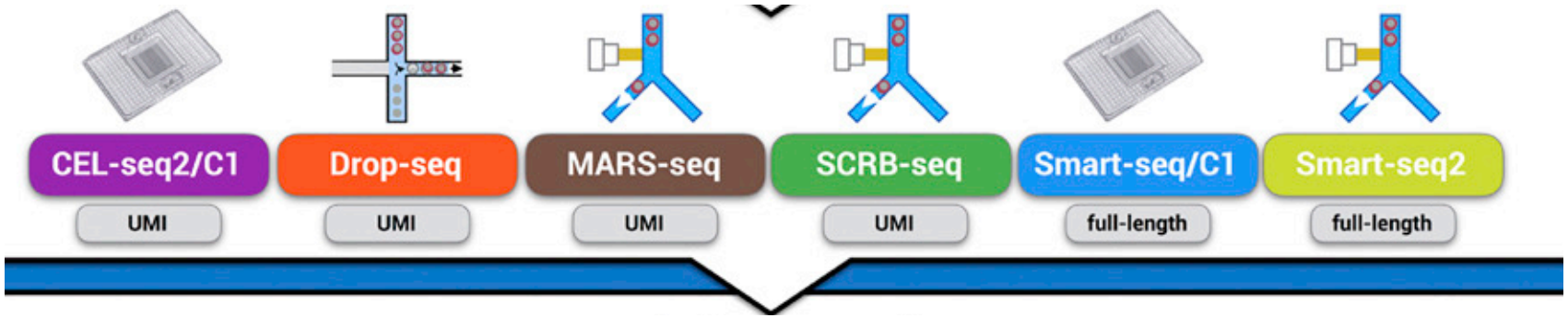
Full-length versus tag-based:

- Full-length captures the entire transcript, and lends itself toward identifying isoforms or allele specific expression.

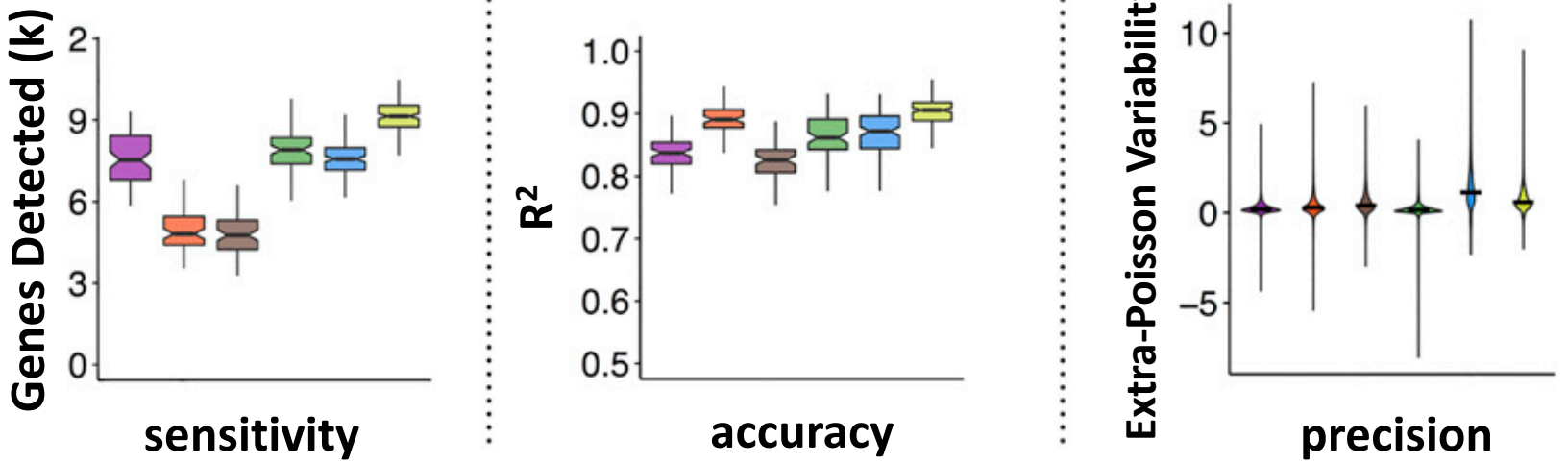
- Tag-based methods allow the use of Unique Molecular Identifiers (UMI) which are a combination of cell- and mRNA- specific barcodes that allow for a unique transcript count and reduce bias.

- Tag-based requires much less sequencing and keeps costs lower.

Reviews/Comparisons of technologies



Descriptive comparison



Ziegenhain et al. Molecular Cell. 2017.

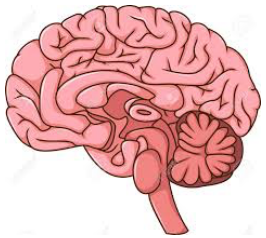
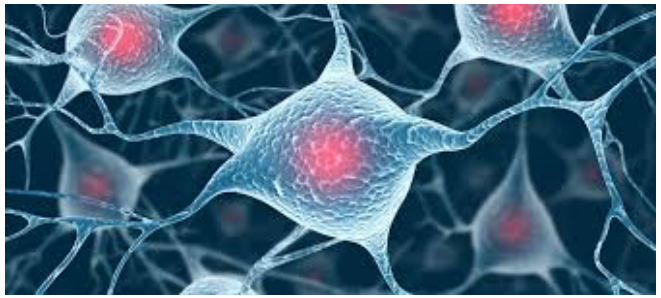
Considerations

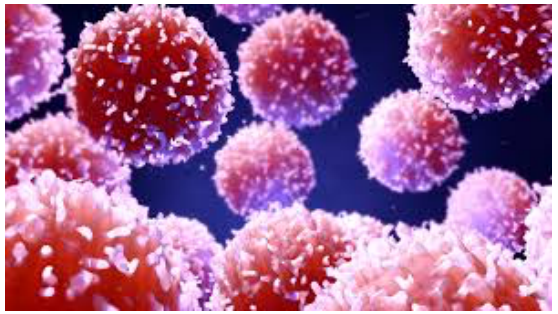
Different biological applications lend themselves to different protocols.

- Isoform analysis and allele specific expression require full length transcripts.
- Droplet based methods are not ideal for precious samples with a small number of cells.
- Discovery of novel cell types requires generating a large number of single-cells.

Considerations

- Decision of protocol/platform will be based on:
  - Type of cells used.
  - Source and availability of cells.



  - Number of cells need.
  - Cost limitations.

Preprocessing

Pre-processing: Quality Control

QC on reads:

- Total reads per sample
- Base quality scores
- GC content
- Adapter sequences
- Over represented sequences

QC on alignment:

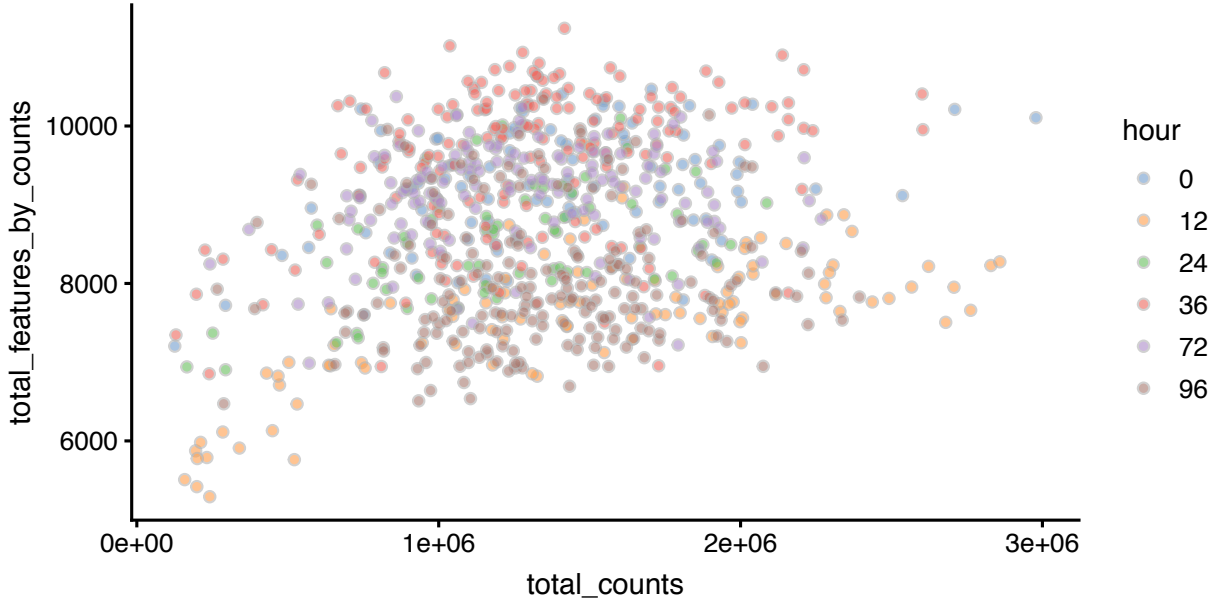
- Total transcripts
- Uniquely mapping reads
- Reads mapping to mitochondria

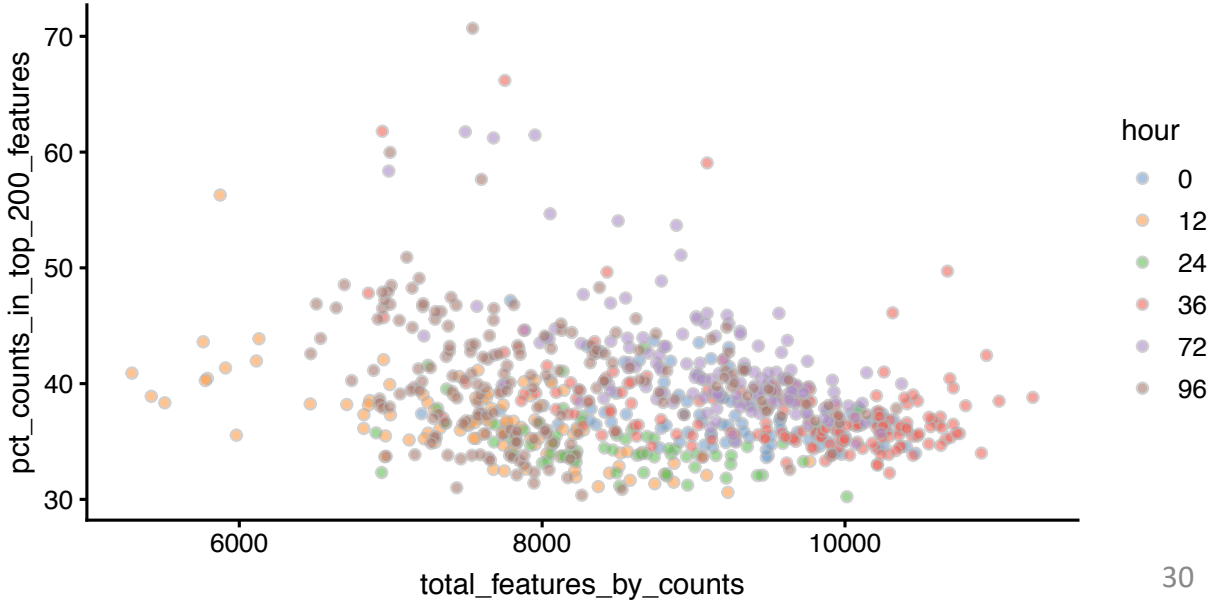
QC across cells:

- Batch effects

Quality Control on aligned reads

- Try to identify technical artifacts.
- Examine total counts, number of detected genes/features, percent of counts represented by top genes.
  - scater





QC across cells

- Will want to check at this stage for any batch effects.
  - E.g. Total number of reads may be very different across batches.

- Batch effects that are not confounded with the biological question can be removed or accounted for.

Pre-processing: Normalization

- Before adjusting for batch effects however, the data must be normalized to adjust for differences in sequencing depth.

- We correct this in order to compare the expression of a gene *across* cells.

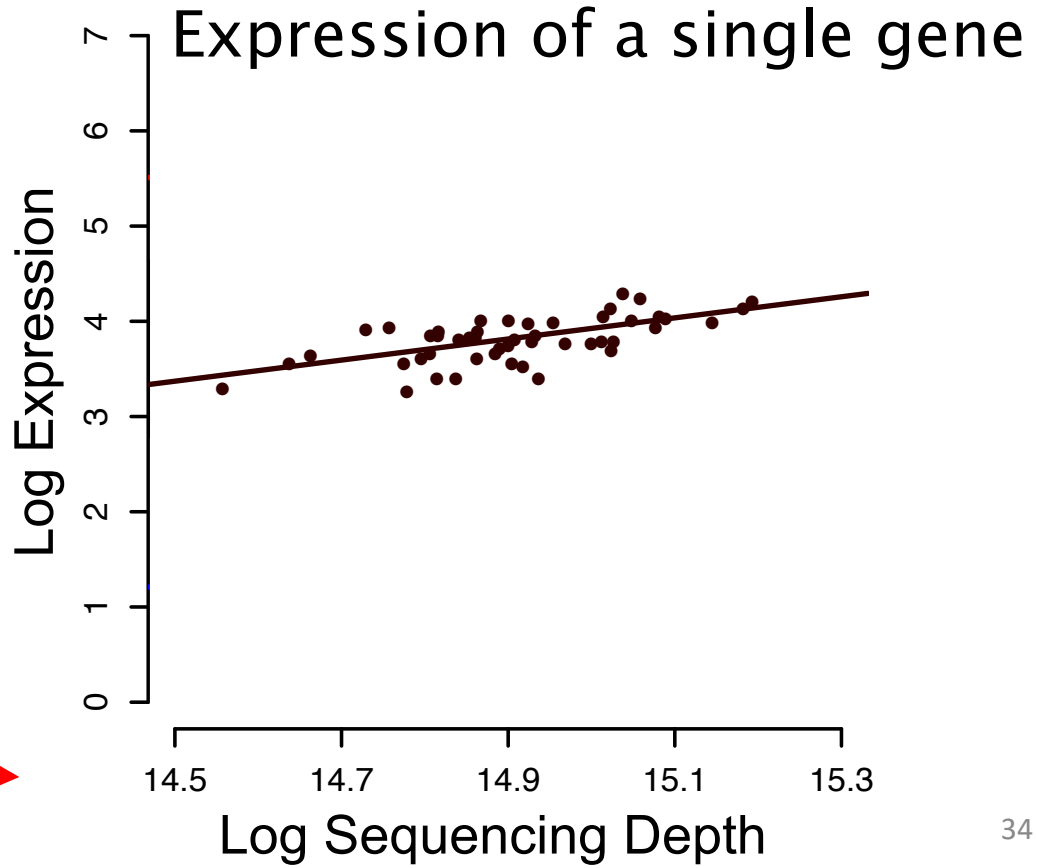
Normalization Assumptions

- What is typically corrected for?
  - Sequencing depth
- Assumption?
  - Cells have different depths due to differences in sampling.
  - If a cell was sequenced twice as much, we would observe twice the expression for every gene (on average).

Count-depth relationship

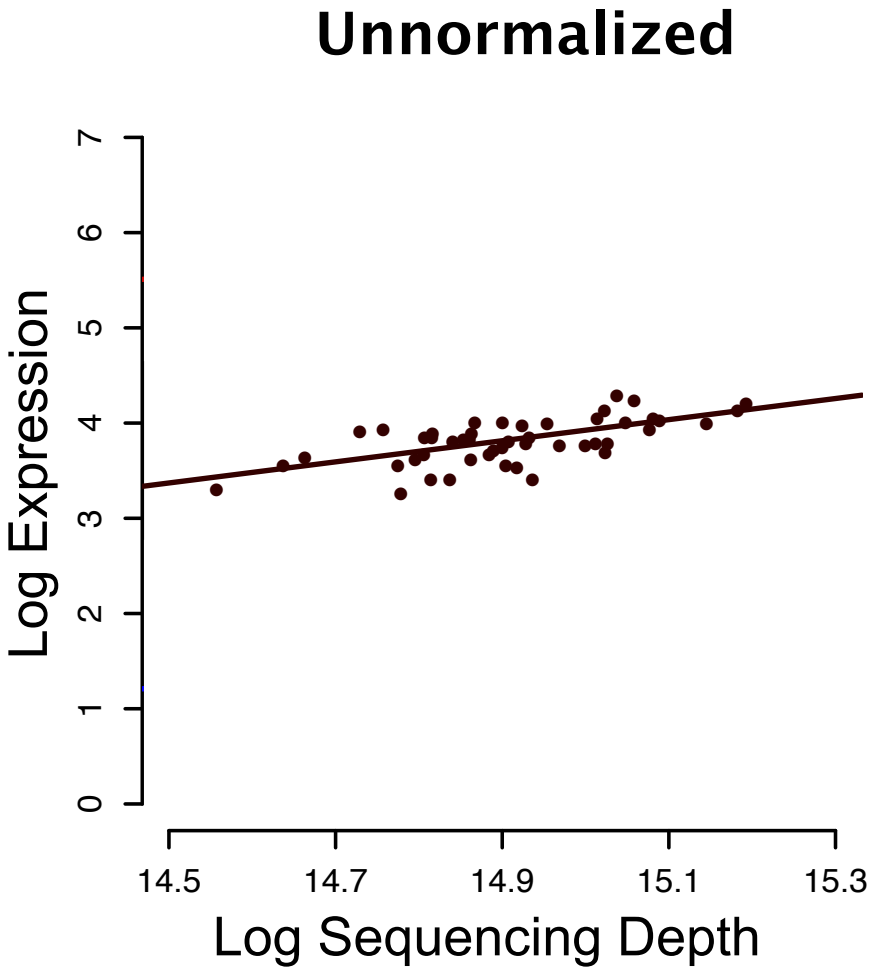
- Relationship between a gene's expression across all cells as a function of each cell's sequencing depth.

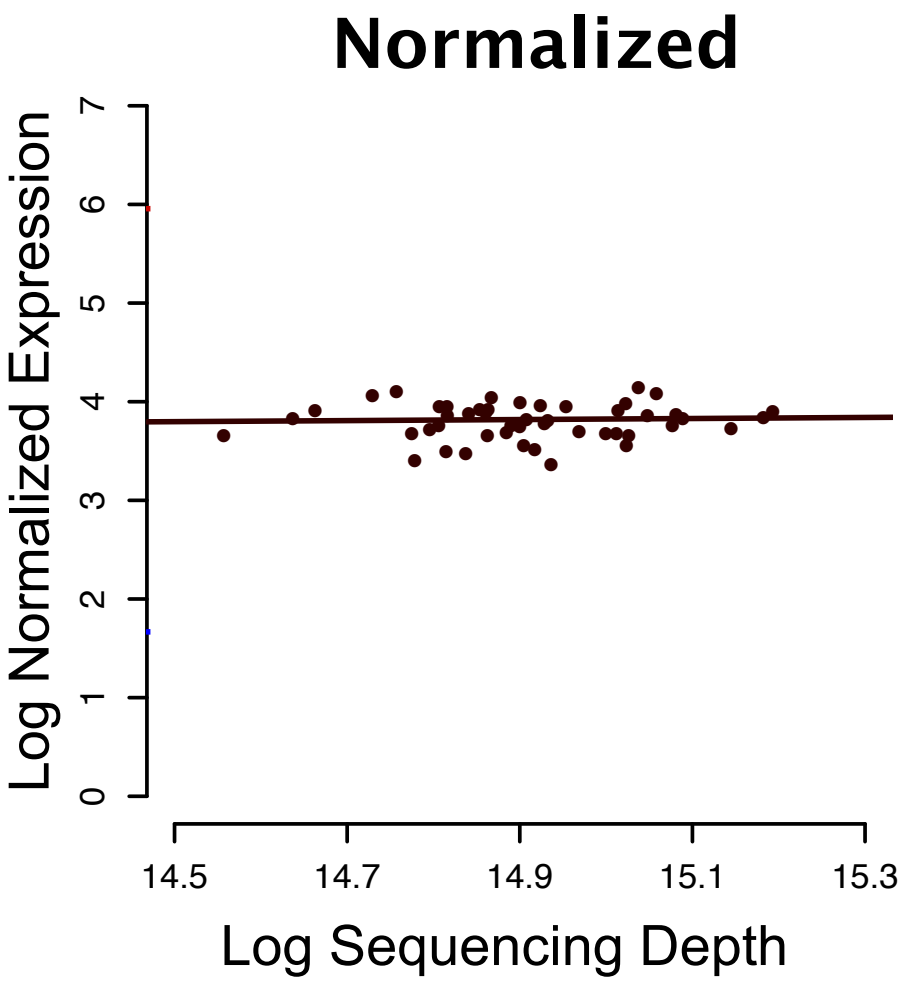
	Sample 1	Sample 2	…	Sample n
Gene 1	62	124	…	42
Gene 2	10	20	…	10
Gene 3	316	632	…	322
…	…	…	$Y_{g,j}$	…
Gene m	85	170	…	73
Sequencing Depth	$\sum_{g=1}^{m} Y_{g,1}$	$\sum_{g=1}^{m} Y_{g,2}$	…	$\sum_{g=1}^{m} Y_{g,n}$



Count-depth relationship - one gene

Bulk data

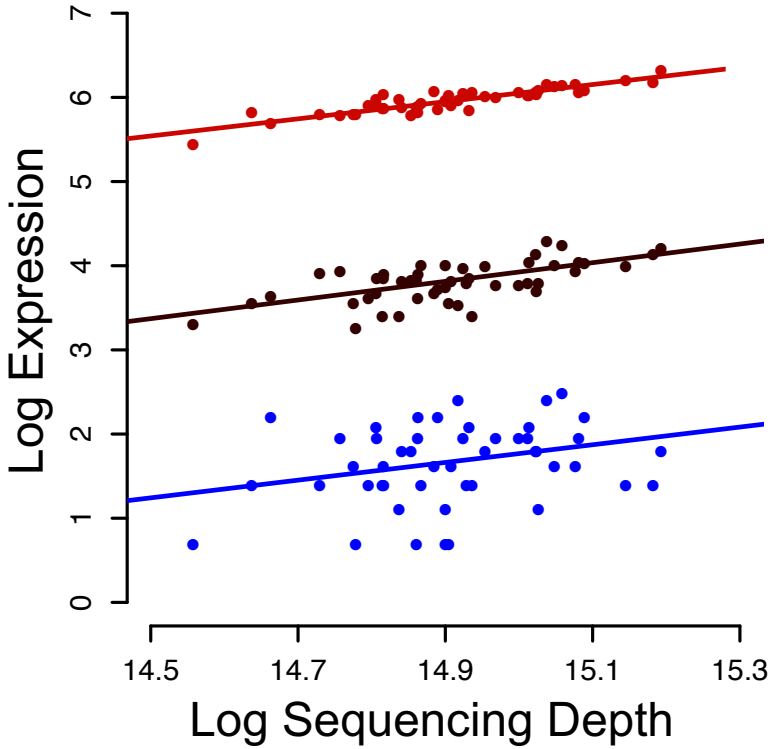




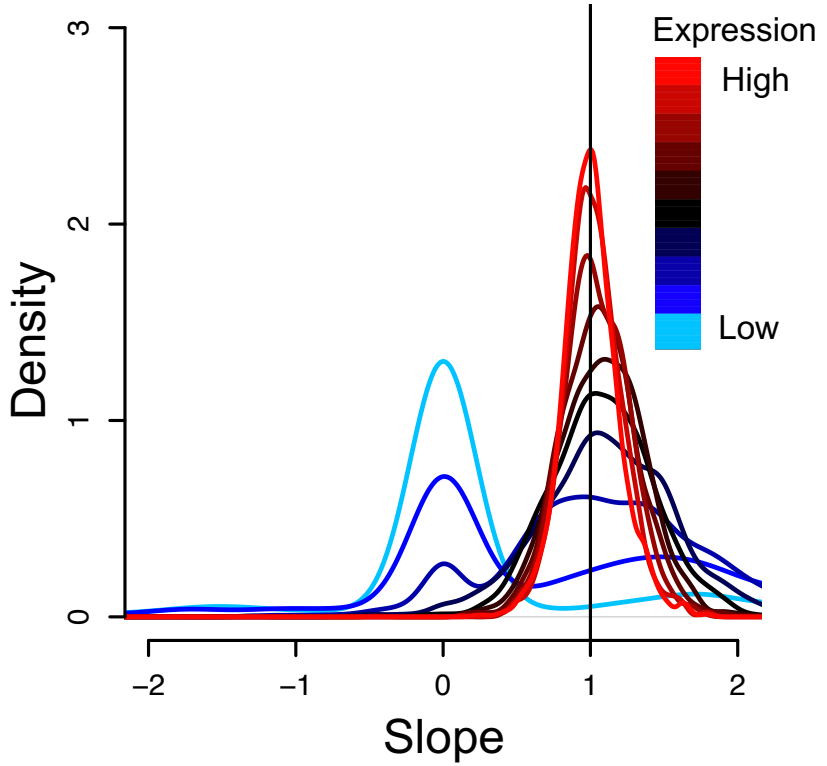
Count-depth relationship in bulk

Unnormalized data:

3 genes having low, moderate, and high expression:

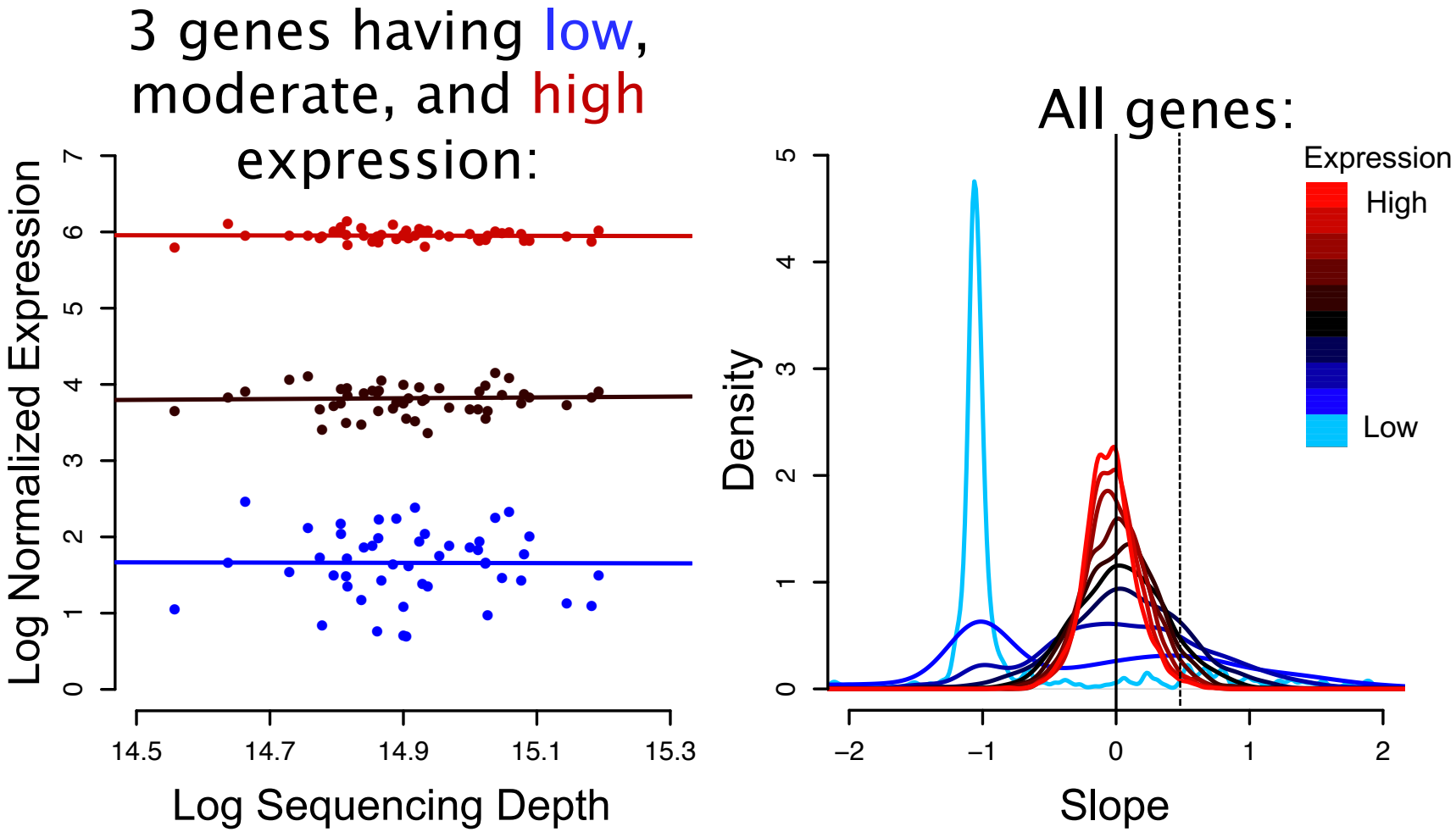


All genes:



Count-depth relationship bulk - post normalization

Normalized data:

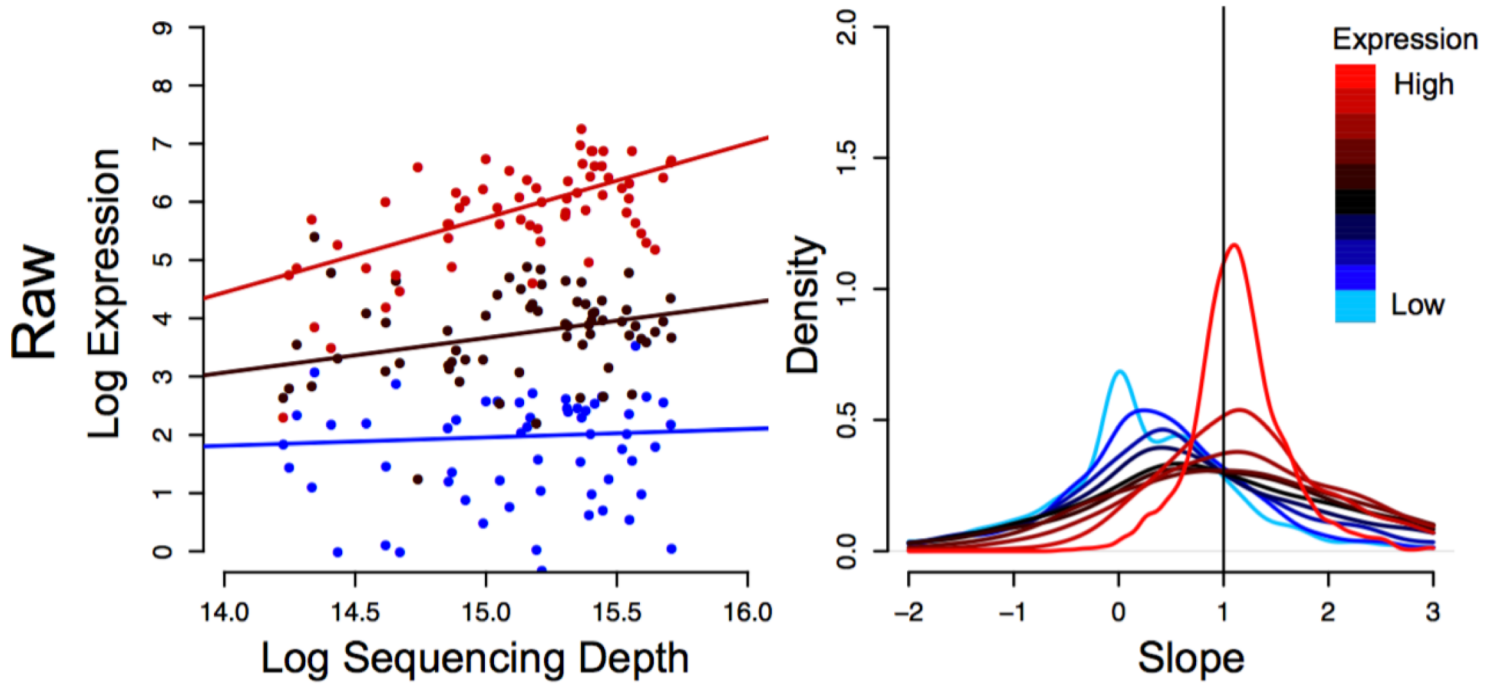


Normalization Procedure

- Majority of normalization methods *are global scale factor* approaches.
- A single scale factor is per sample/cell and applied to all genes.
- This includes: CPM, TMM, Median-Ratio, Upper-Quartile from bulk and scran from single-cell.

Count-depth relationship in scRNA-seq

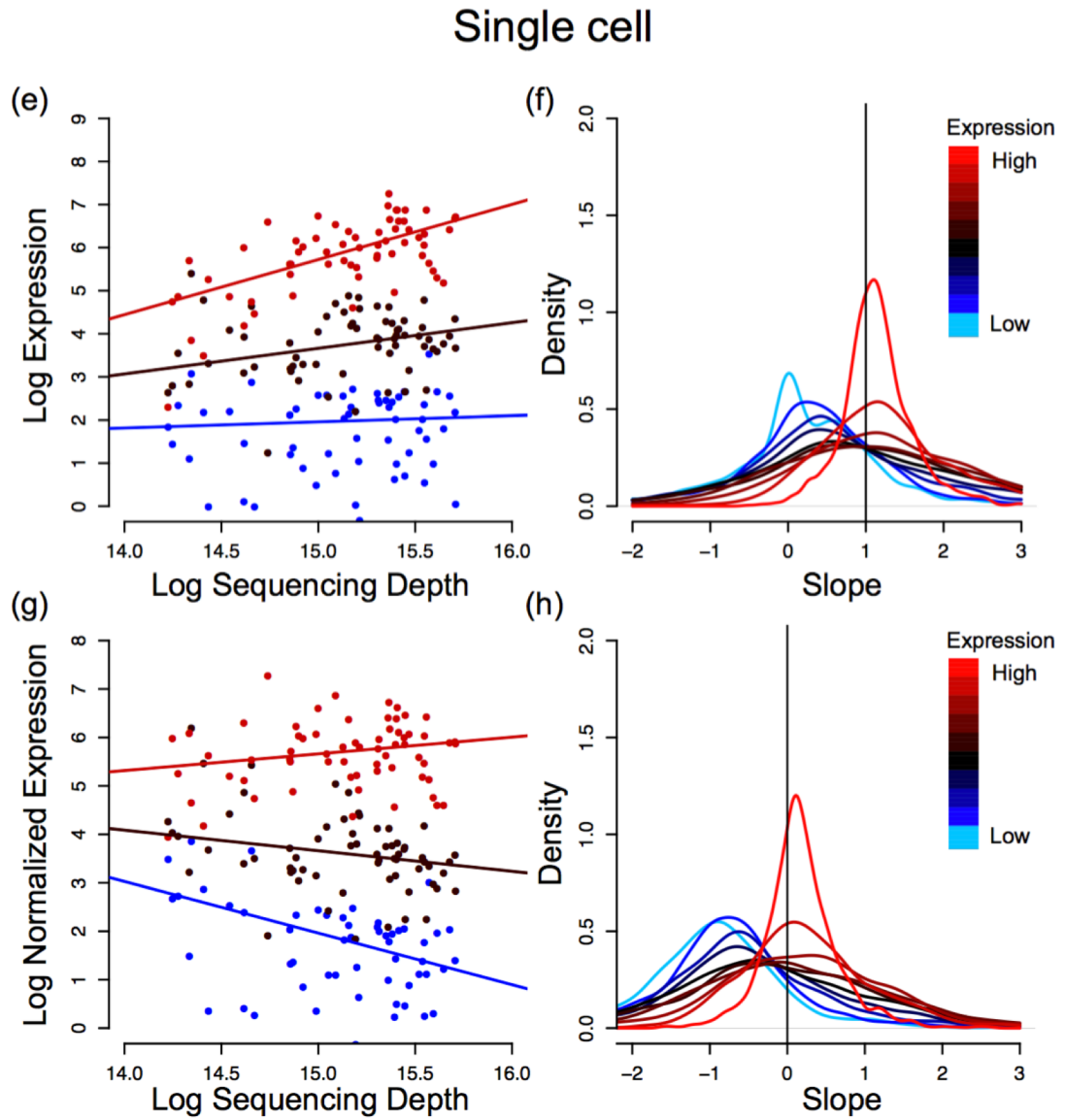
Unnormalized data:



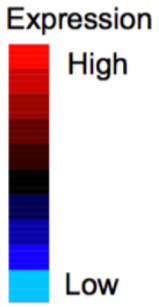
scRNA-seq data using global scale factors

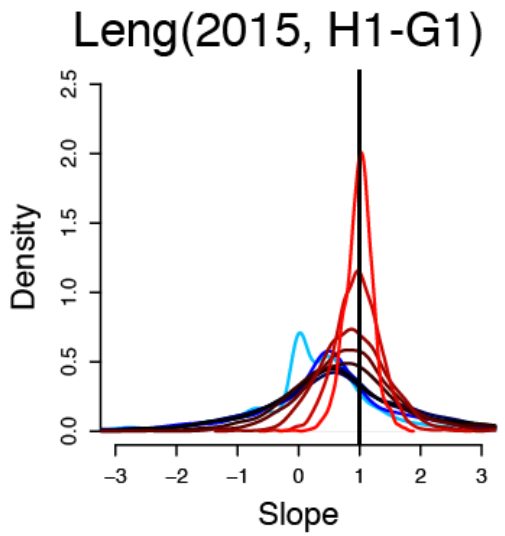
Un-normalized:

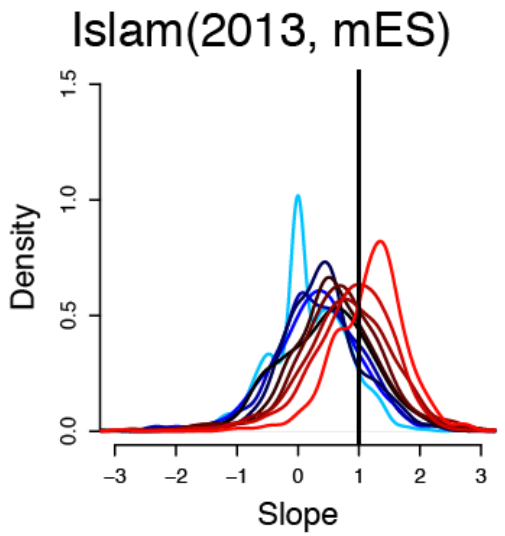
Global scale factor:

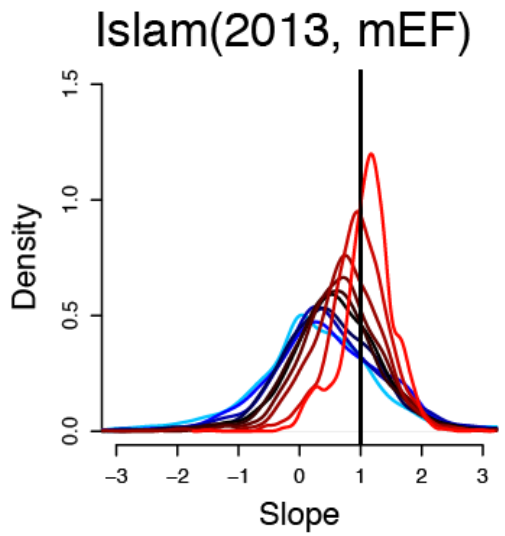


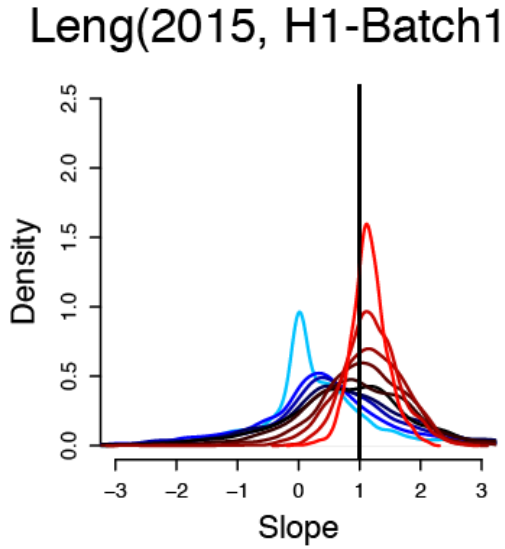
Exists across datasets

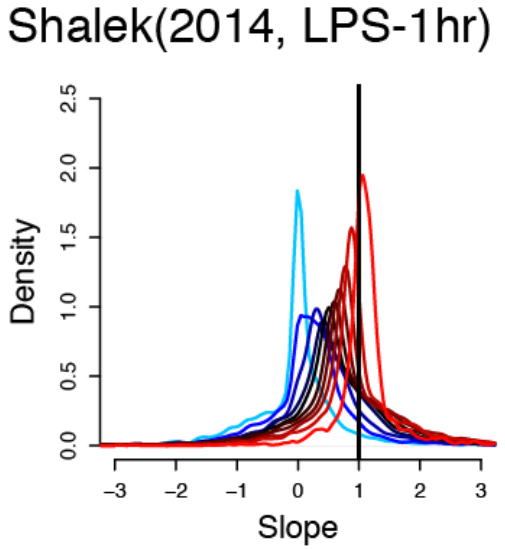


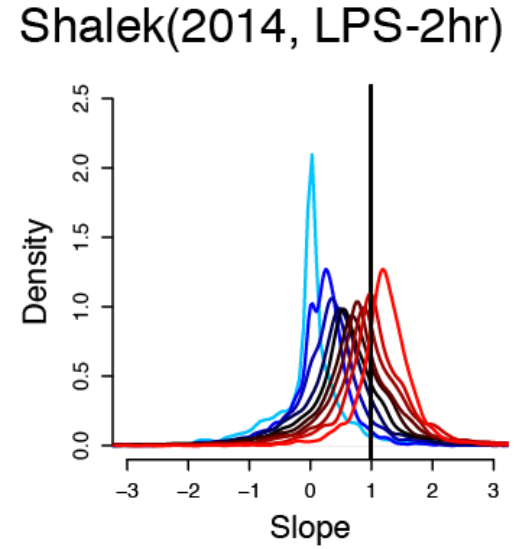












SCnorm: An Overview

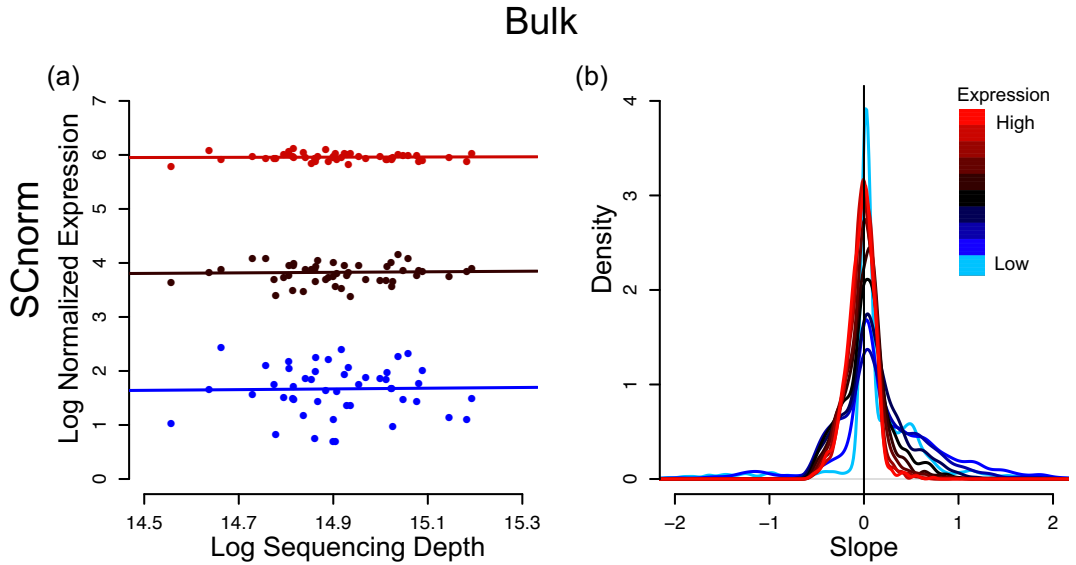
- Step 1: Quantify each gene's relationship with sequencing depth (count-depth relationship) and cluster genes into k groups.

- Step 2: Estimate within group scaling factors and normalize each group separately.

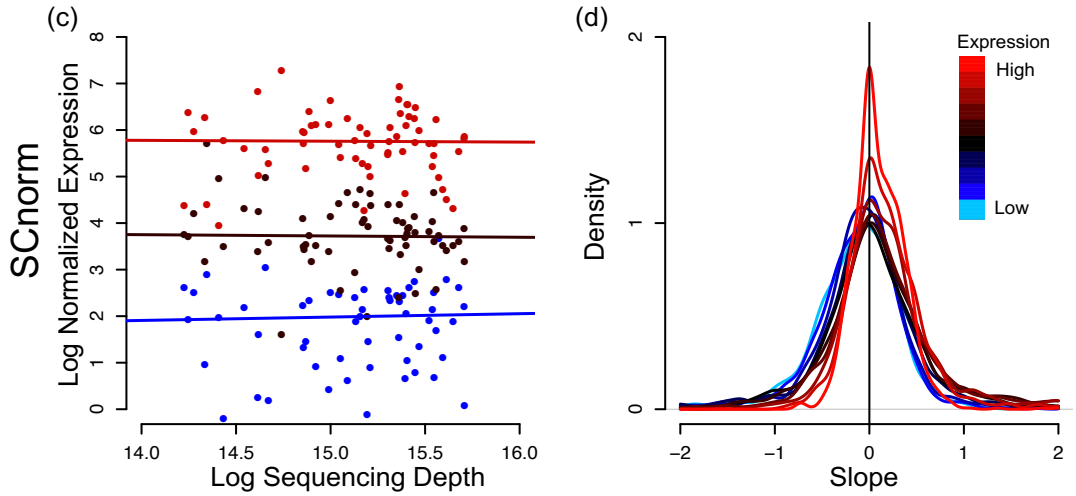
- Step 3 : Evaluate the sufficiency of *k* groups.
  - If the evaluation suggests more groups are needed, step 2 is repeated using $k + 1$ groups until convergence.

SCnorm on bulk and single-cell RNA-seq data

Bulk:



Single-cell:



SCnorm: robust normalization of single-cell RNA-seq data

- Bacher R, Chu LF, Leng N, Gasch AP, Thomson JA, Stewart RM, Newton M, Kendziorski C. “SCnorm: robust normalization of single-cell RNA-seq data.” Nature methods. 2017 Jun;14(6):584.

- https://bioconductor.org/packages/release/bioc/html/SCnorm.html

Pre-processing: Normalization

- Other normalization methods include:
  - scran – global scale factor method, fast.
  - scTransform – similar in principle to SCnorm for UMI data.

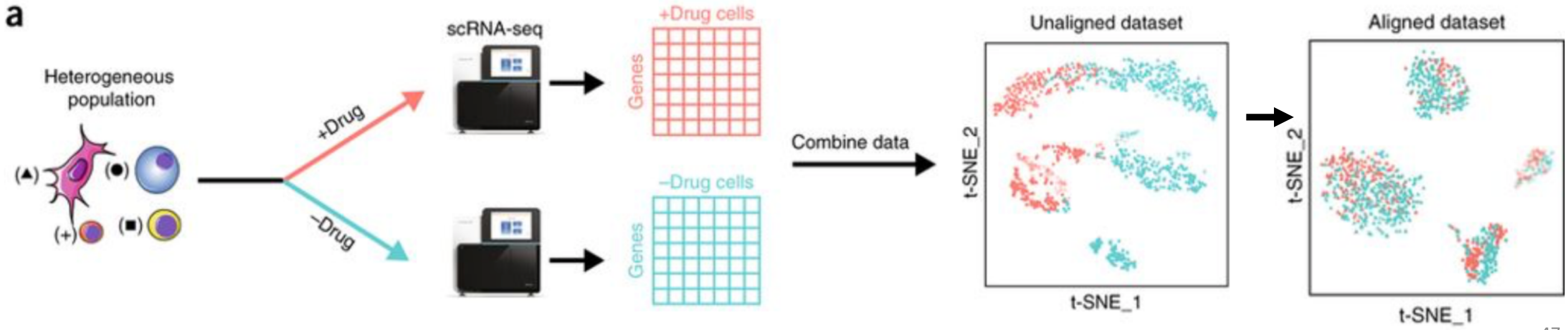
- Normalization can help with batch effects but will not completely remove them.

Pre-processing: Batch effects

- Simple batch correction might be done via:
  - Regression on batch ID.
  - Estimate scale factors using spike-ins (if available)
- Spike-ins are transcripts that are added in equal amounts to each cell.

Pre-processing: Batch effects

- Alignment methods
  - Seurat: Uses canonical correlation analysis
  - mnnCorrect: Uses mutual nearest neighbors



Butler et al. Nature Biotechnology. 2018.