normjam session 2: The SCnorm approach

Rhonda Bacher, PhD
Assistant Professor
Department of Biostatistics

UF | UNIVERSITY of FLORIDA

What is normalization?

What is normalization?

- Adjustment of the expression counts to remove systematic variations and render counts comparable across genes and/or samples (via direct count adjustment or scale factors).

- Types of systematic variations:
  - Gene length
  - GC content
  - Sequencing depth

Between-sample normalization

- Correct sample-specific features, e.g. sequencing depth.
- Compare the same gene across samples.

	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	$\sum_{g=1}^{m} Y_{g,1}$	$\sum_{g=1}^{m} Y_{g,2}$	…	$\sum_{g=1}^{m} Y_{g,n}$

Global scale factor normalization methods

Global scale factor approaches:

- Counts Per Million (CPM): $SF_j = \frac{\sum_{g=1}^{m} Y_{g,j}}{10^6}$
- Anders and Huber, 2010: $SF_j = \underset{g}{\text{median}} \frac{Y_{g,j}}{(\prod_{j=1}^{n} Y_{g,j})^{1/n}}$

- Normalized expression is given as:

$$Y'_{g,j} = Y_{g,j} / SF_j$$

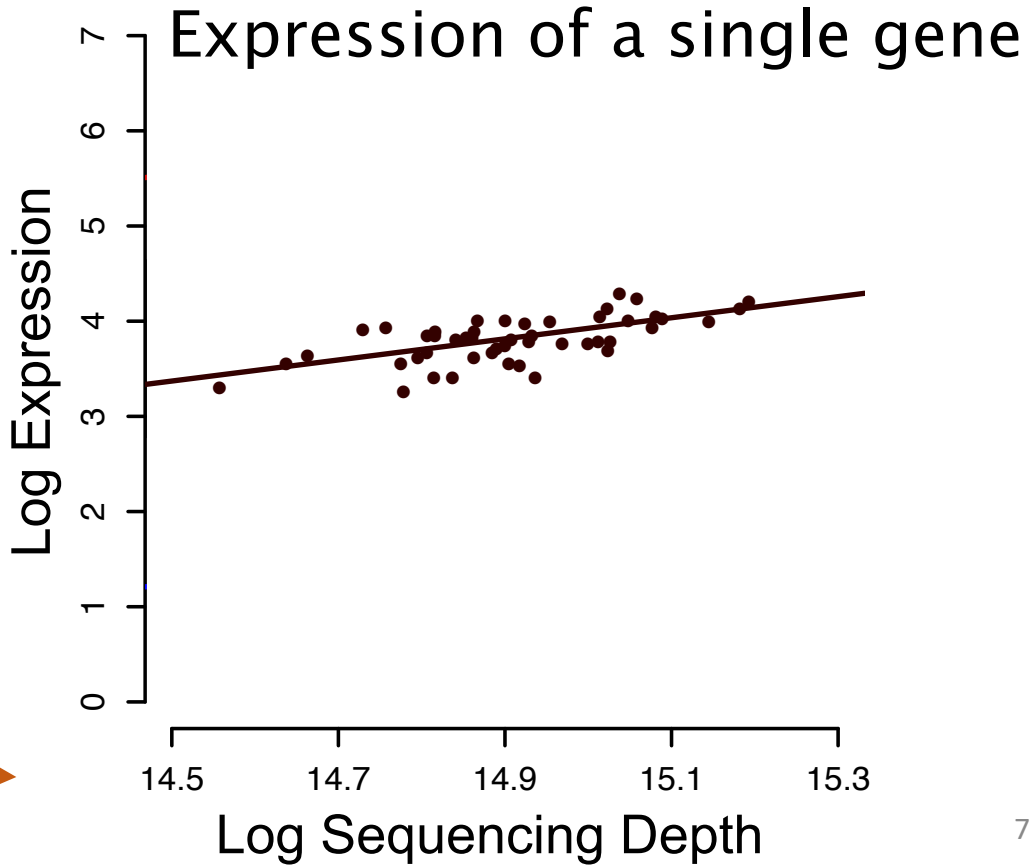
Assumptions of global SF normalizations

- Sequencing depth is a technical artifact:
  - due to differences in sampling during sequencing.
  - affects expression similarly for all genes.
- If a sample was sequenced twice as much, we would observe twice the expression for every gene (on average).

Count-depth relationship

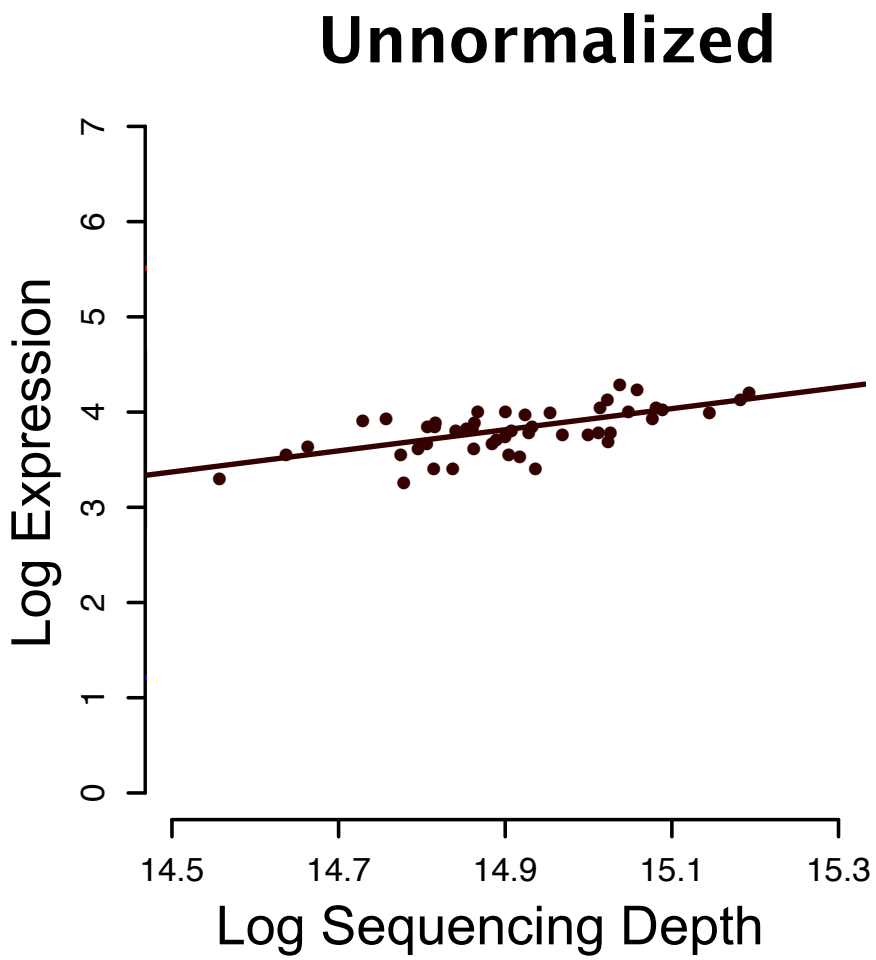
- Relationship between a gene's expression across all cells as a function of each sample's sequencing depth.

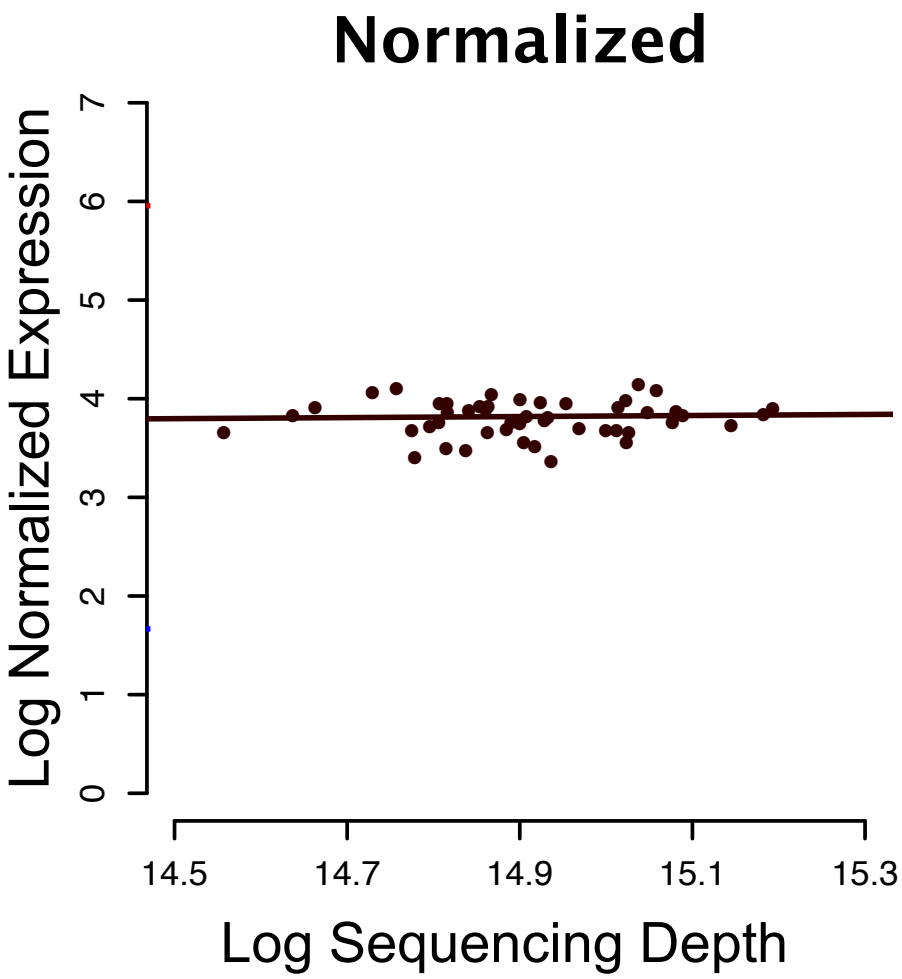
	Sample 1	Sample 2	…	Sample n
Gene 1	62	124	…	42
Gene 2	10	20	…	10
Gene 3	316	632	…	322
…	…	…	$Y_{g,j}$	…
Gene m	85	170	…	73
Sequencing Depth	$\sum_{g=1}^{m} Y_{g,1}$	$\sum_{g=1}^{m} Y_{g,2}$	…	$\sum_{g=1}^{m} Y_{g,n}$



Count-depth relationship - one gene

Bulk data

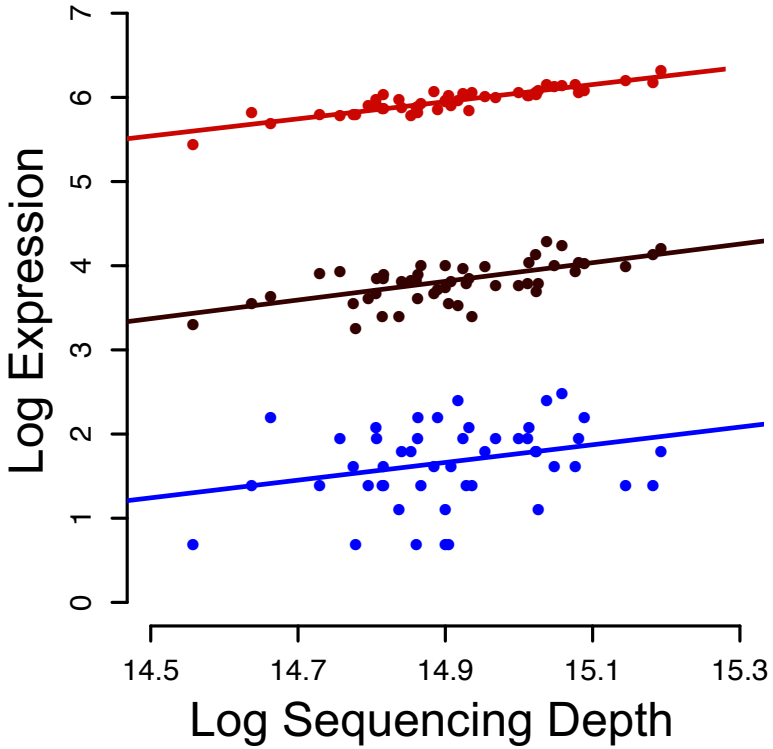




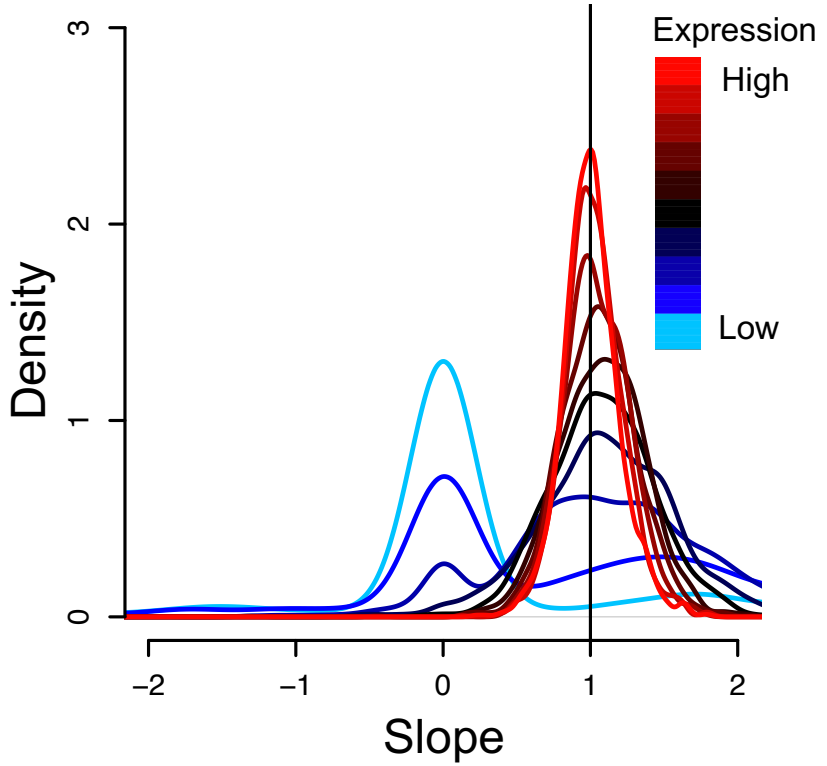
Count-depth relationship in bulk

Unnormalized data:

3 genes having low, moderate, and high expression:

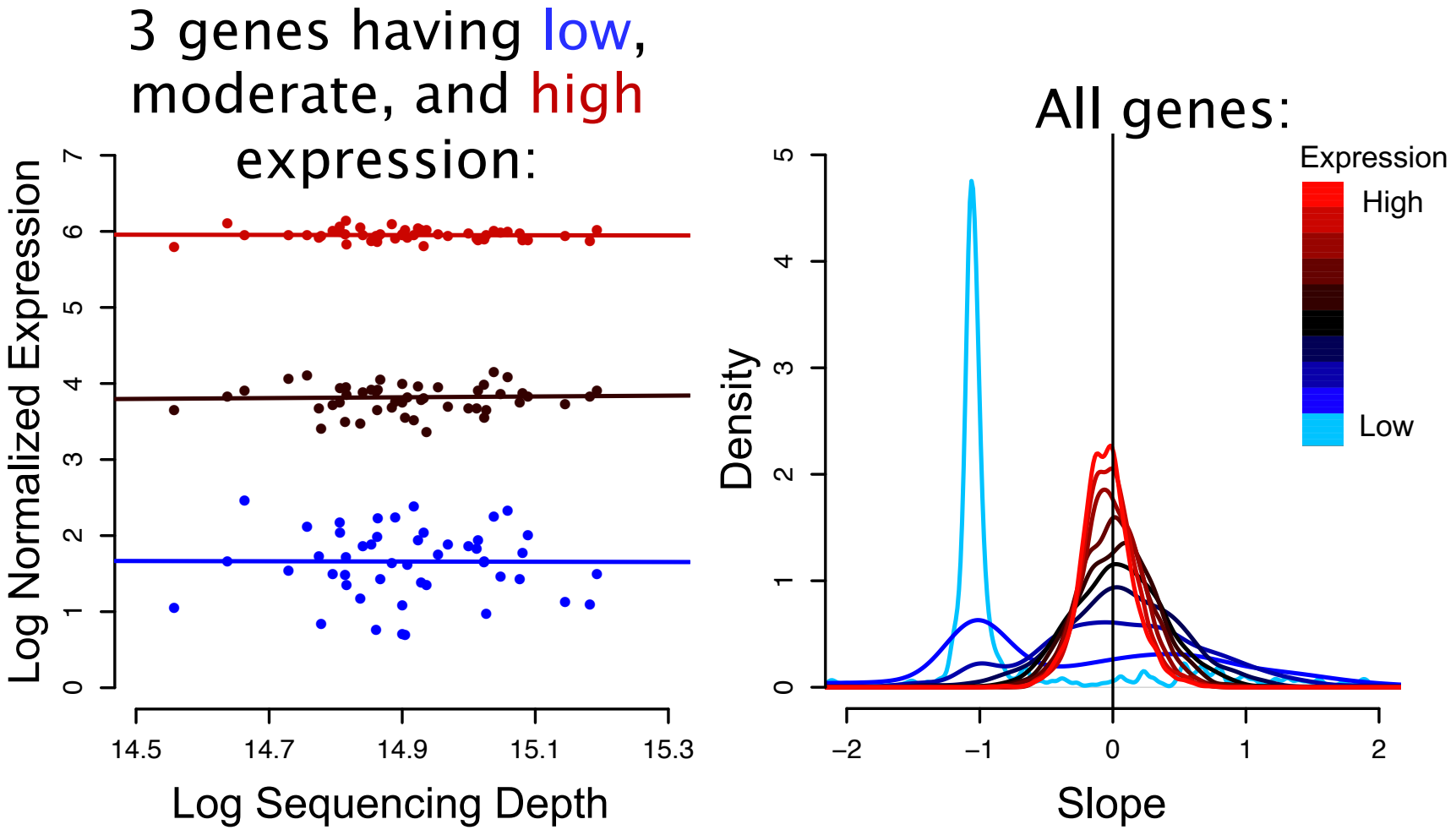


All genes:



Count-depth relationship bulk - post normalization

Normalized data:



How do you normalize?

The SCnorm approach

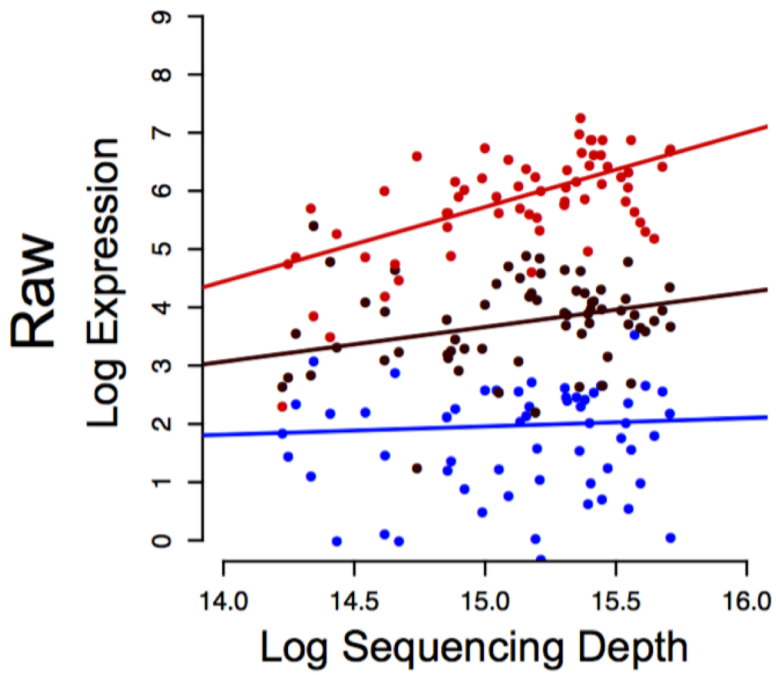
- Bacher R, Chu LF, Leng N, Gasch AP, Thomson JA, Stewart RM, Newton M, Kendziorski C. “SCnorm: robust normalization of single-cell RNA-seq data.” Nature methods. 2017 Jun;14(6):584.

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

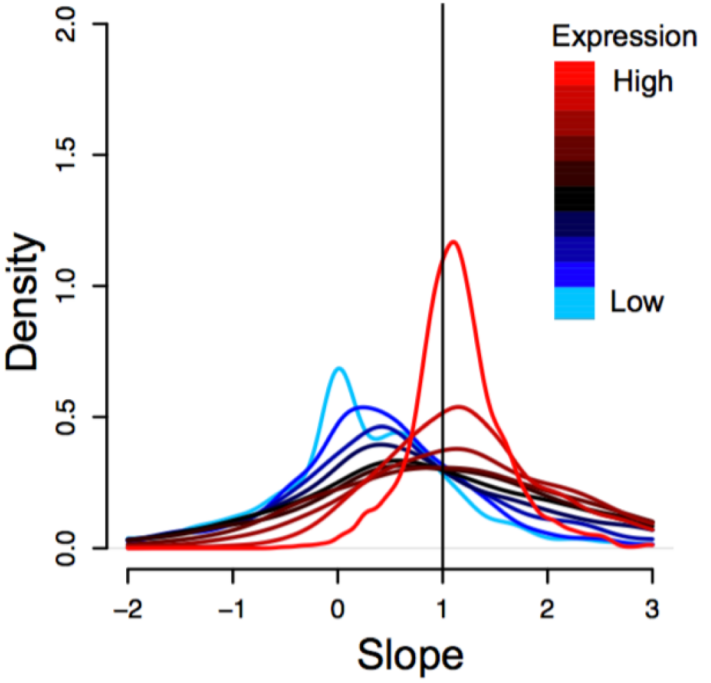
Count-depth relationship in scRNA-seq

Unnormalized data:

3 genes having low, moderate, and high expression:



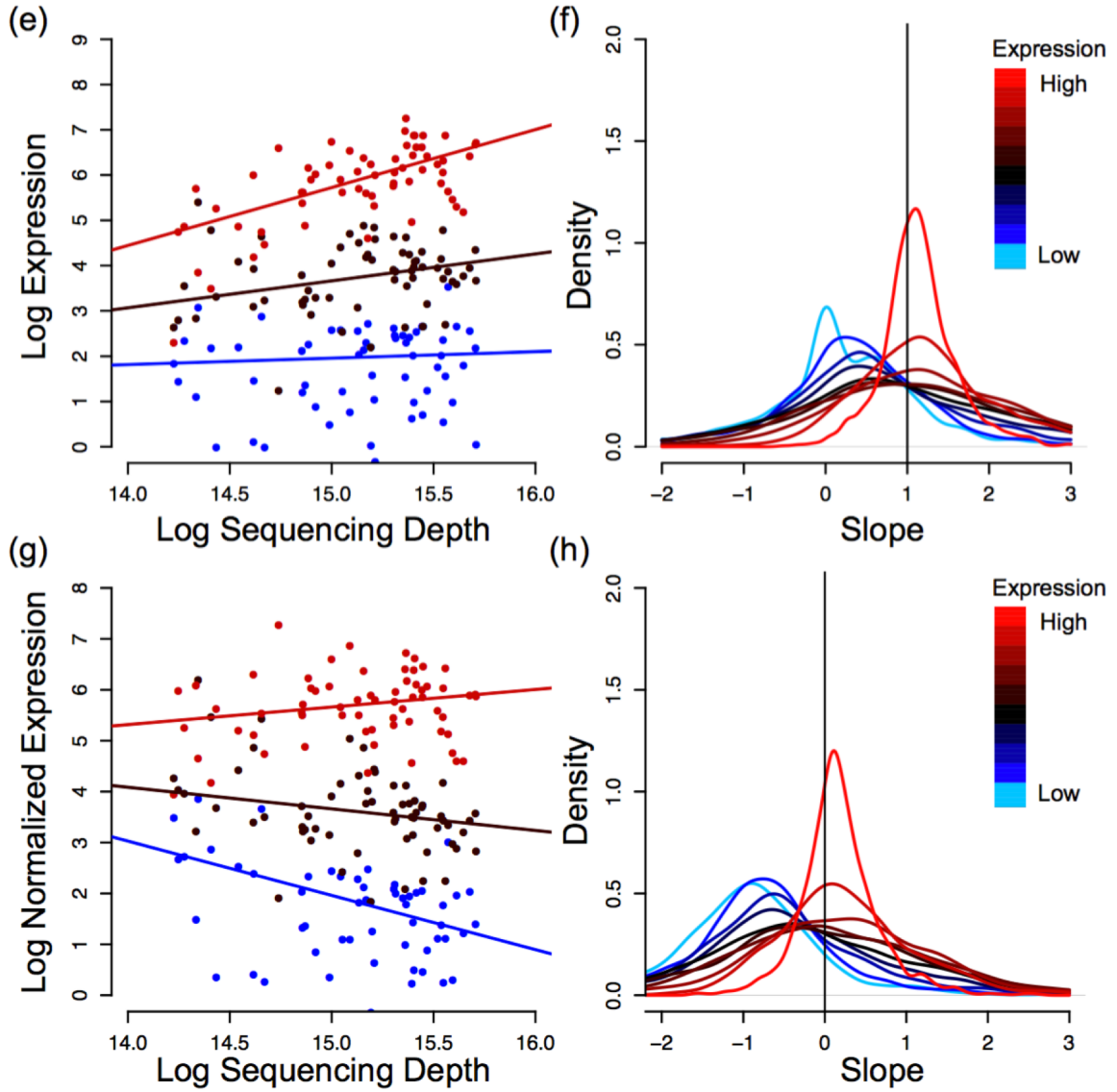
All genes:



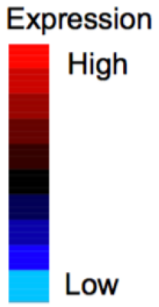
scRNA-seq data using global scale factors

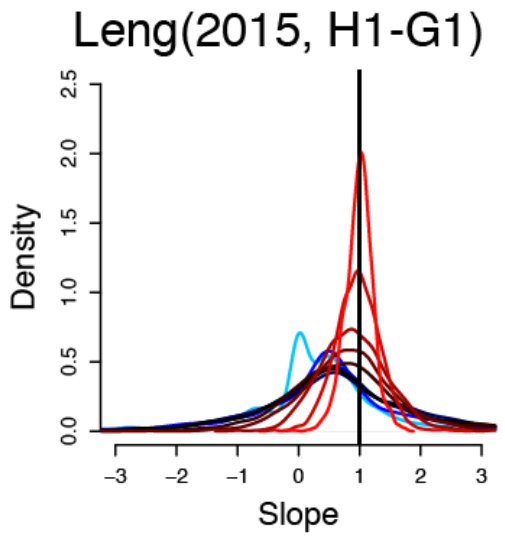
Un-normalized:

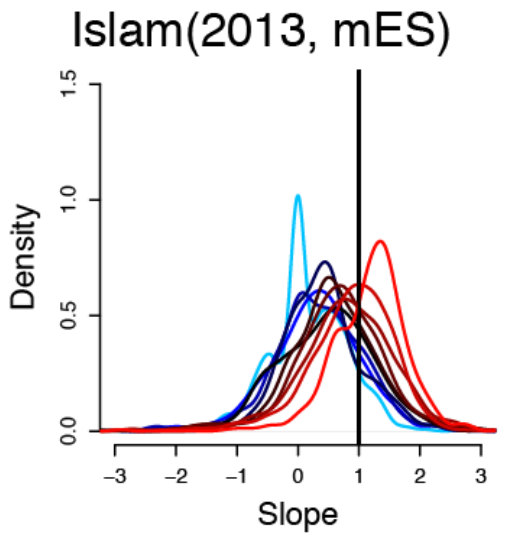
Global scale factor:

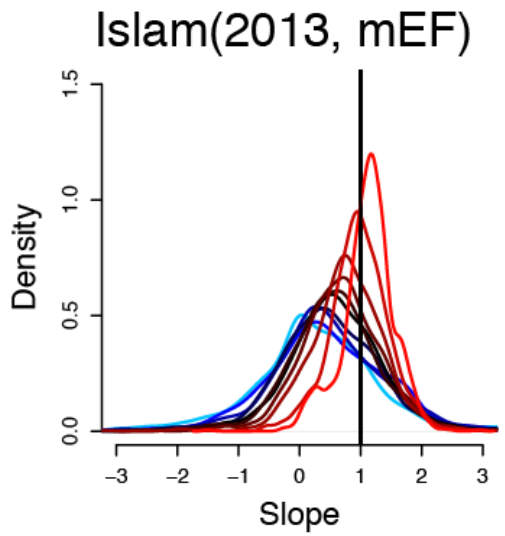


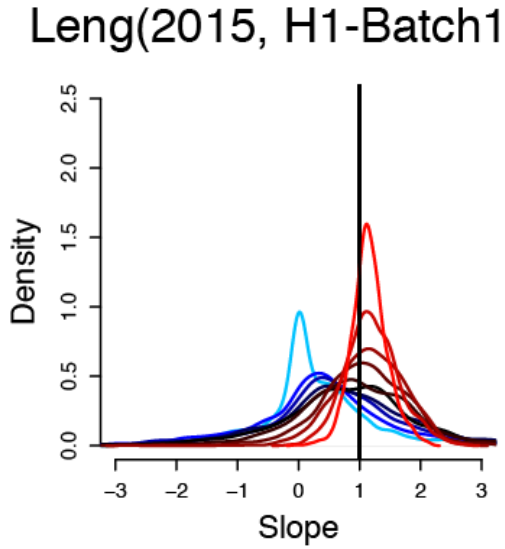
Exists across datasets

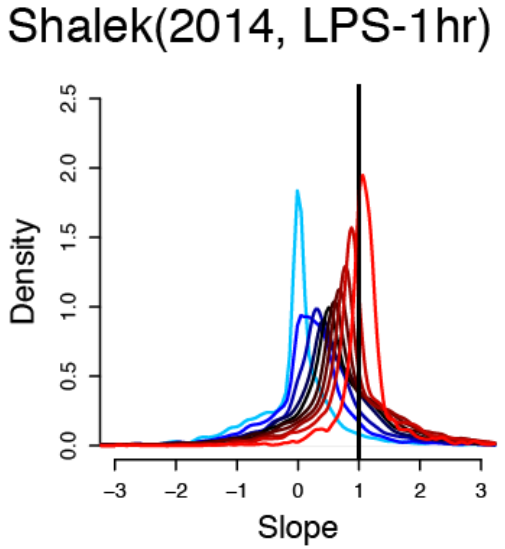


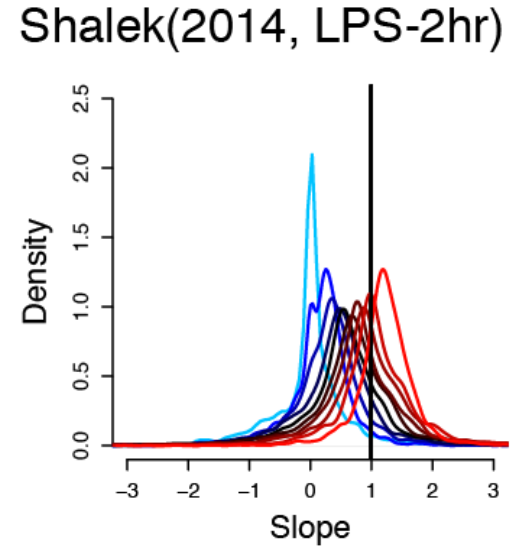












SCnorm: An Overview

- 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.

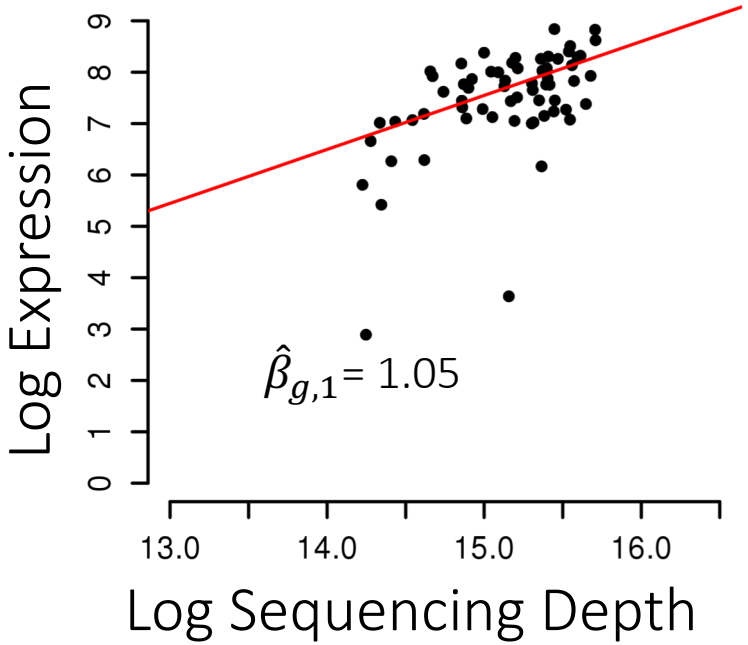
- Step 3 : Evaluate the sufficiency of *k* groups.
  - If the evaluation suggests more groups are needed, step 2 is repeated using $k + 1$ groups until convergence.

SCnorm: Step 1

- $Y_{g,j}$ denote log expression estimates for gene g in cell j.
- X_j denote the log sequencing depth for cell j.
- Estimate count-depth relationship, $\hat{\beta}_{g,1}$, using median quantile regression:

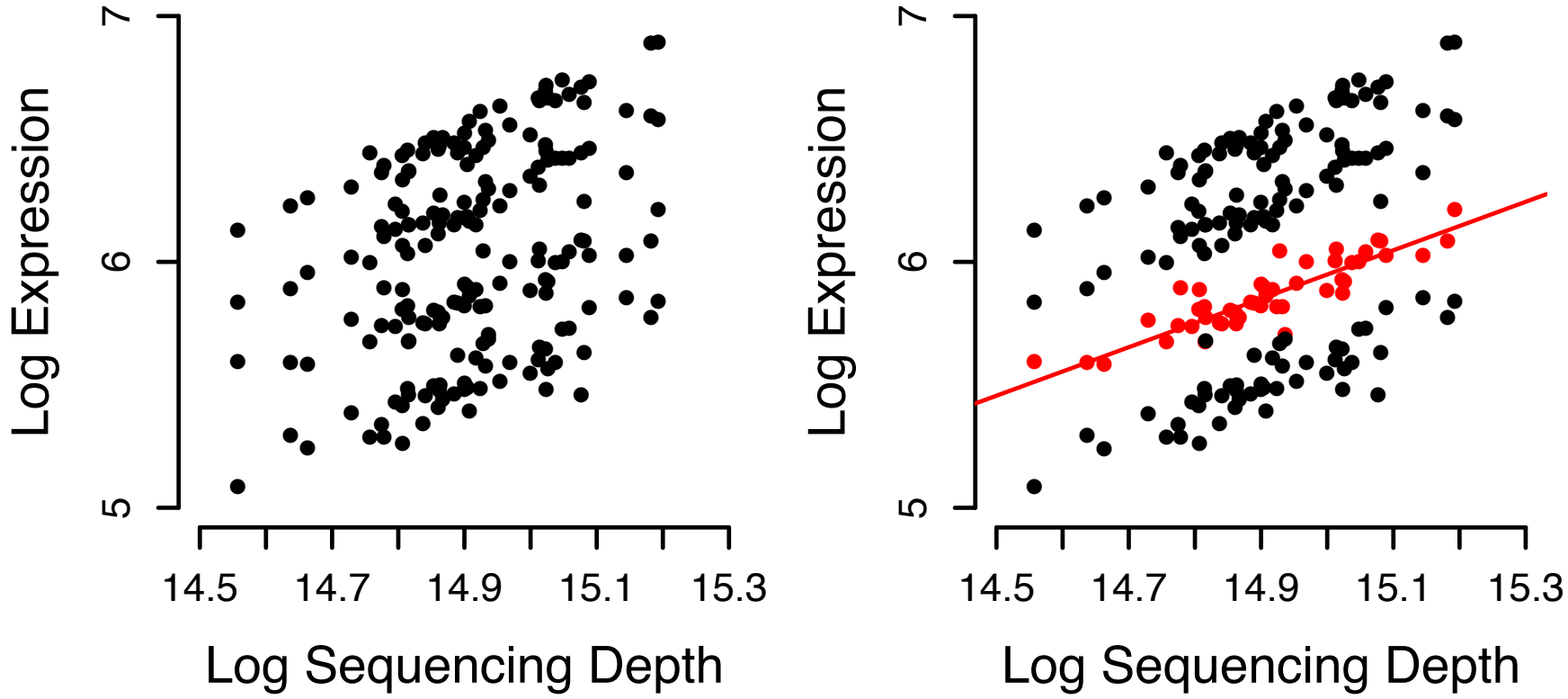
$$Q^{.5}(Y_{g,j}|X_j) = \beta_{g,0} + \beta_{g,1}X_j$$

- Group genes using the K-medoids algorithm.



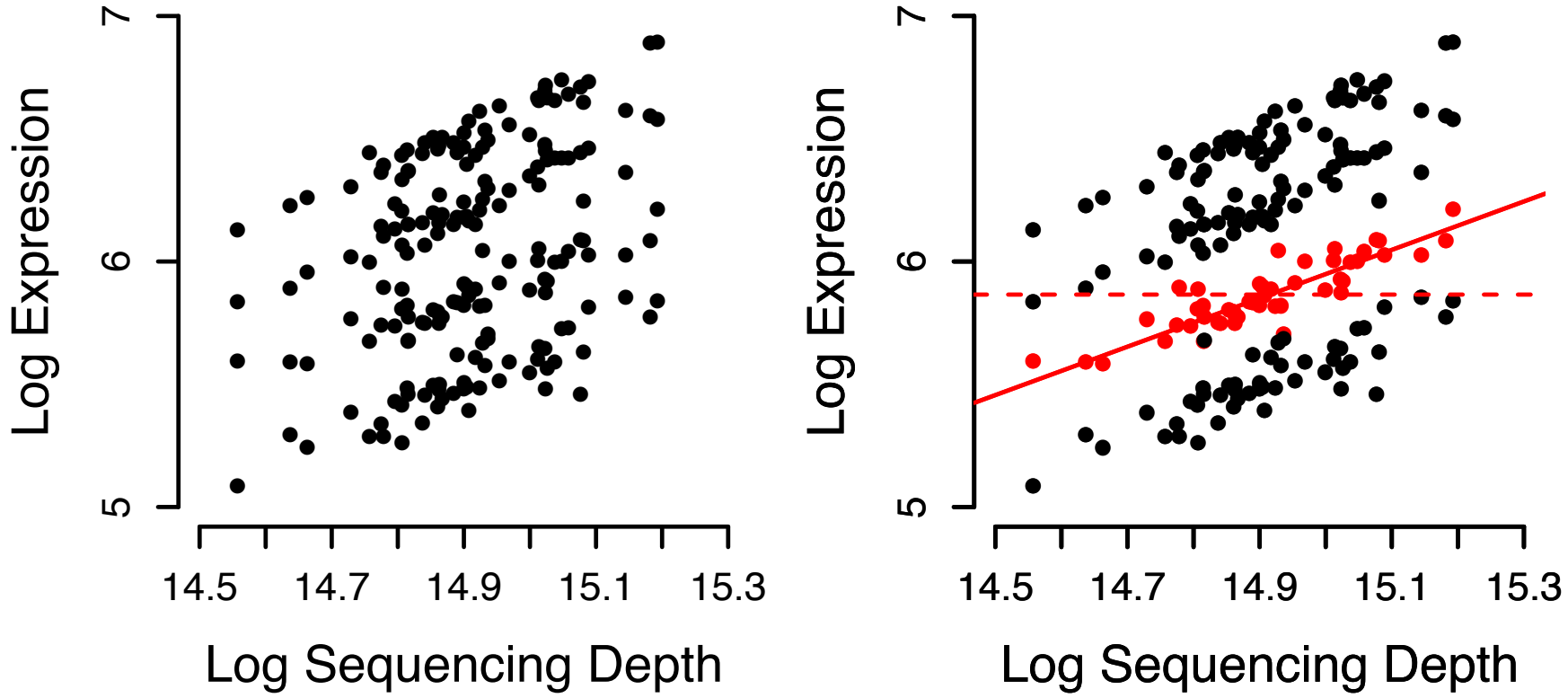
SCnorm: Step 2

- Within each group (cartoon):



SCnorm: Step 2

- Within each group (cartoon):



SCnorm: Step 2

- Within each group :
  - A τ^{th} quantile polynomial regression of degree d is fit over all genes in the group:

$$Q^{\tau_k,d_k}\left(Y_{g_k,j}\middle|X_j\right) = \beta_0^{\tau_k} + \beta_1^{\tau_k}X_j + \cdots + \beta_d^{\tau_k}X_j^{d_k} \qquad (1)$$

  - The predicted values from this regression, $\hat{Y}_j^{\tau_k,d_k}$, can be viewed as values from a stable gene from which we estimate scale factors.

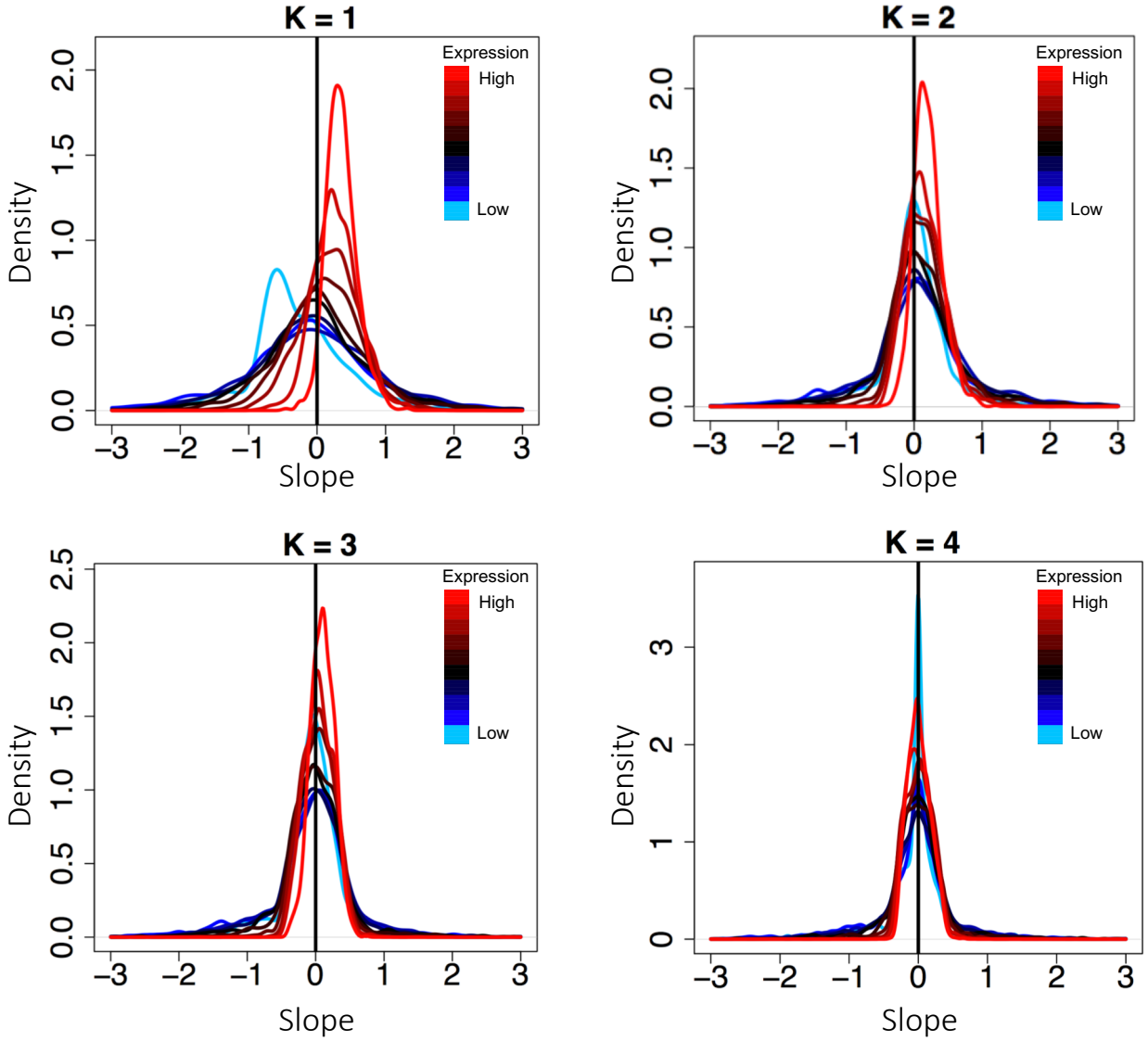
SCnorm: Step 2 (estimation)

- Scale factors for each cell are estimated as:

$$SF_j = e^{\hat{Y}_j^{\tau_k^*, d_k^*}} / e^{Y^{\tau_k^*}}$$

where $Y^{\tau_k^*}$ is the τ^{th} quantile of expression counts in the kth group.

SCnorm: Step 3 (in pictures)

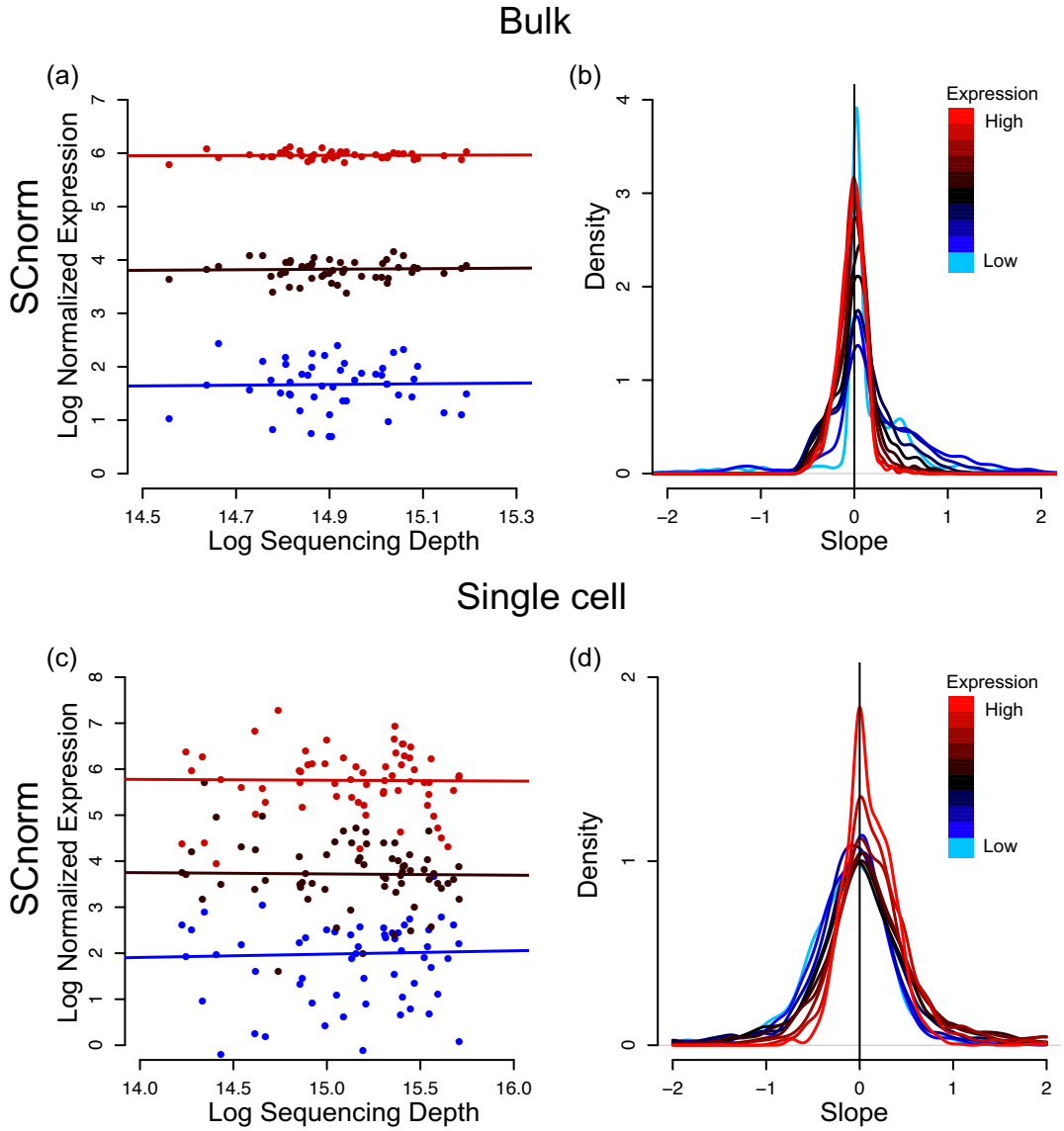


Defining successful normalization

Normalized count-depth plots

Bulk:

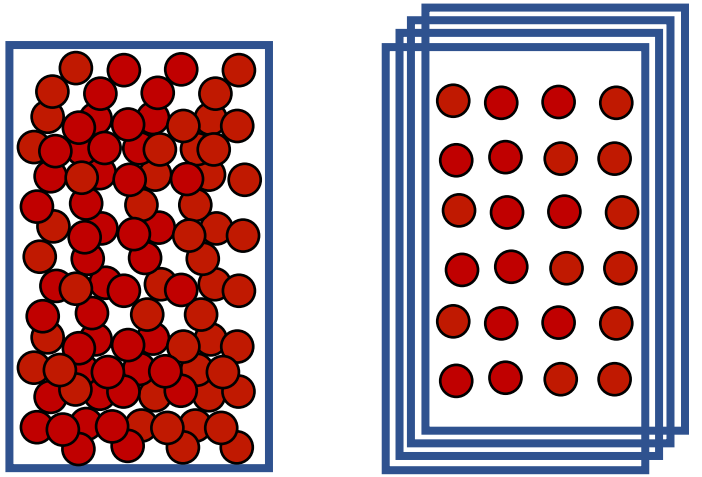
Single-cell:



Case study dataset

- Experiment on H1 embryonic stem cells.
  - Prior to sequencing, the fragmented and indexed cDNA for each cell was split into two groups.
    - Group H1-1M: 96 cells per lane.
    - Group H1-4M: 24 cells per lane.

Reads per lanes is fixed: ~ 100 million

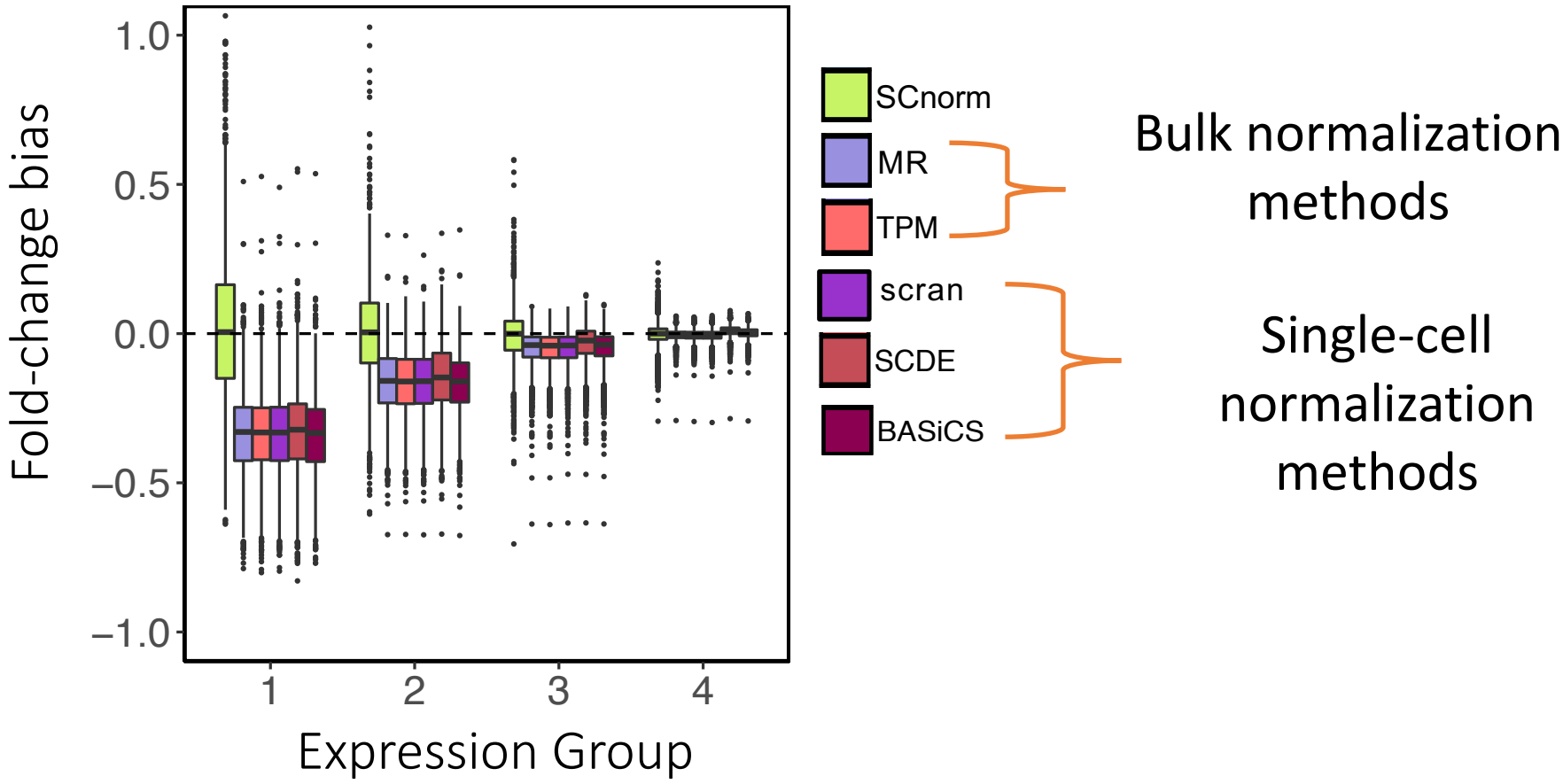


Case study dataset

- Prior to normalization, gene expression in H1-4M will appear on average four times higher than expression in H1-1M.

- Fold-change bias $= \frac{\text{mean}(\text{H1}-\text{4M})}{\text{mean}(\text{H1}-\text{1M})} - 1$

Case study dataset results



Acknowledgements

Kendziorski Lab

Christina Kendziorski, PhD

Ning Leng, PhD

Keegan Korthauer, PhD

Ziyue Wang

Michael Newton, PhD

Audrey Gasch, PhD

Thomson Lab

Li-Fang Chu, PhD

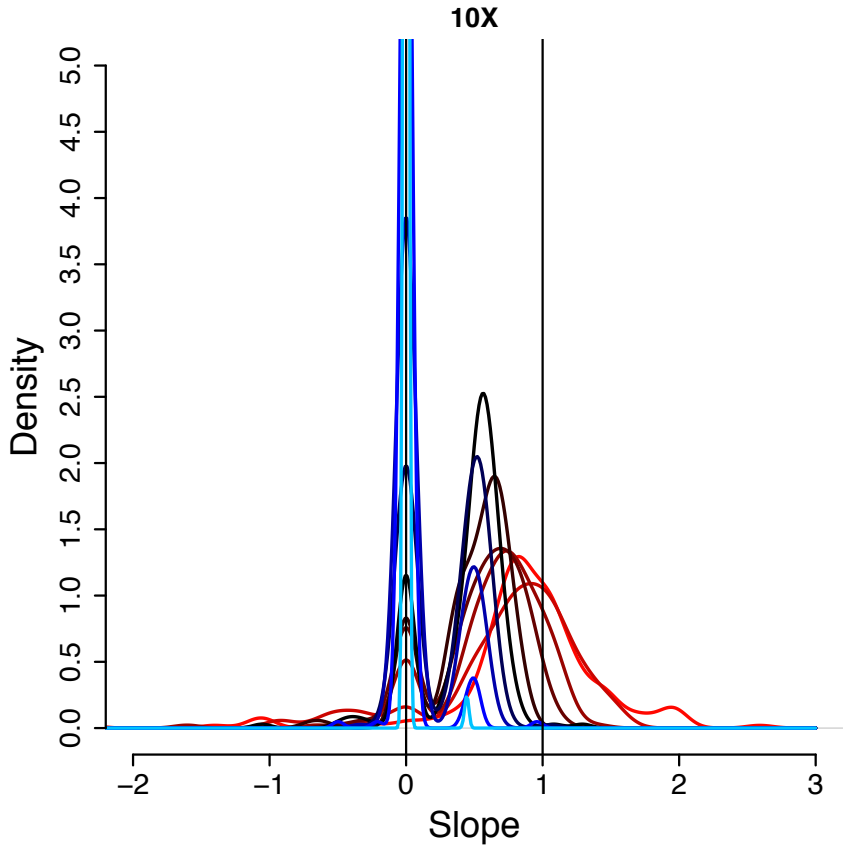
Ron Stewart, PhD

James Thomson, PhD

Limitations and challenges

Current limitations

- Note scalable to tens of thousands of cells.
- SCnorm ignores zeros.



Ideas for day 2

- Benchmarks of a good normalization – metrics.
- Existing datasets for benchmarking?
- Are methods scalable?