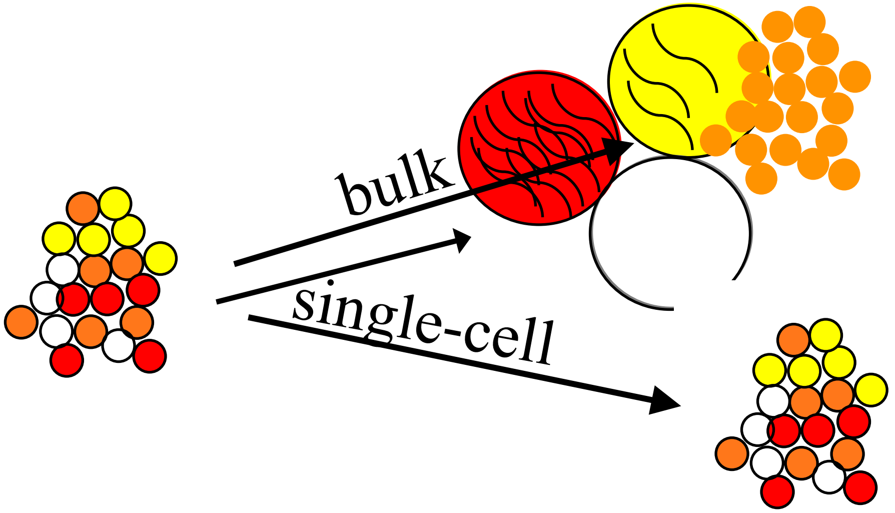
Statistical methods for single-cell RNA sequencing data

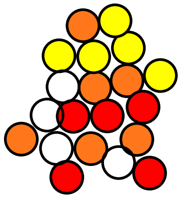
Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison

http://www.biostat.wisc.edu/~kendzior/

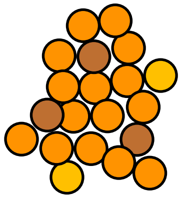
Single-cell vs. bulk RNA-seq



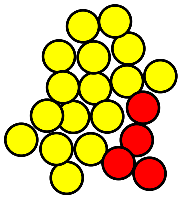
Heterogeneous



Homogeneous

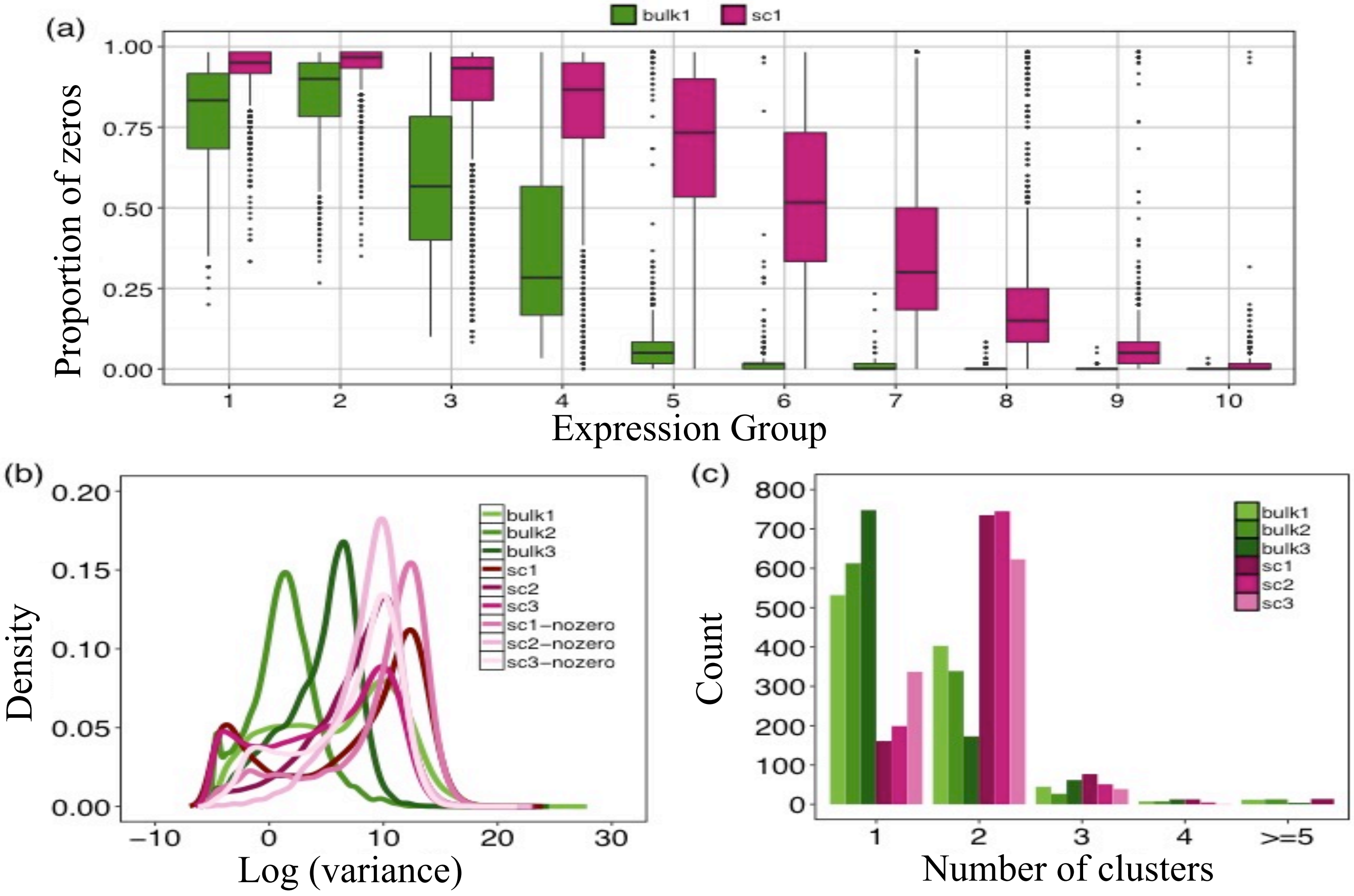


Sub-population



Features of single-cell RNA-seq data

- Abundance of zeros, increased variability, complex distributions



Bacher and Kendziorski, *Genome Biology*, 2016.

Challenges in scRNA-seq

- Normalization
- Technical vs. biological zeros
- Clustering; Identifying sub-populations
- De-noising
    - Adjusting for technical variability
    - Adjusting for biological variability (oscillatory genes)
- Identifying and characterizing differences in gene-specific expression distributions (aka. identifying differential distributions)
- Pseudotime reordering
- Network reconstruction

Challenges in scRNA-seq

- Normalization
- Technical vs. biological zeros
- Clustering; Identifying sub-populations
- De-noising
    - Adjusting for technical variability
    - Adjusting for biological variability (oscillatory genes)
- Identifying and characterizing differences in gene-specific expression distributions (aka. identifying differential distributions)
- Pseudotime reordering
- Network reconstruction

Variability induced by oscillatory genes is substantial in single-cell RNA-seq and can mask effects of interest

We developed an approach called Oscope to identify and characterize oscillations in single-cell RNA-seq experiments

Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments

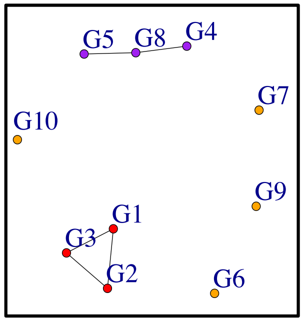
Leng *et al*., *Nature Methods,* 2015

Oscope: Identify and characterize oscillatory genes in an scRNA-seq experiment

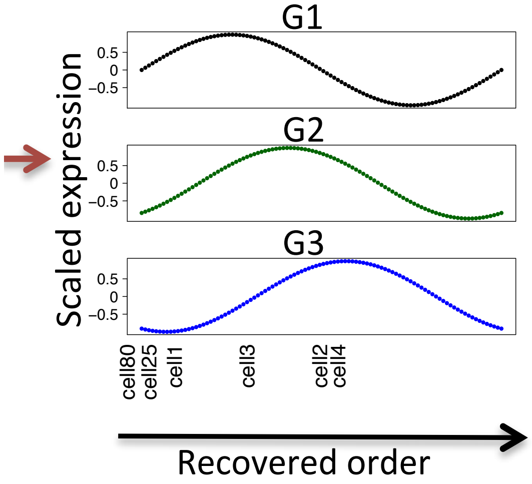
Step 1: Paired-sine model identifies candidate oscillators

Top genes pairs
G1-G2
G3-G2
G1-G3

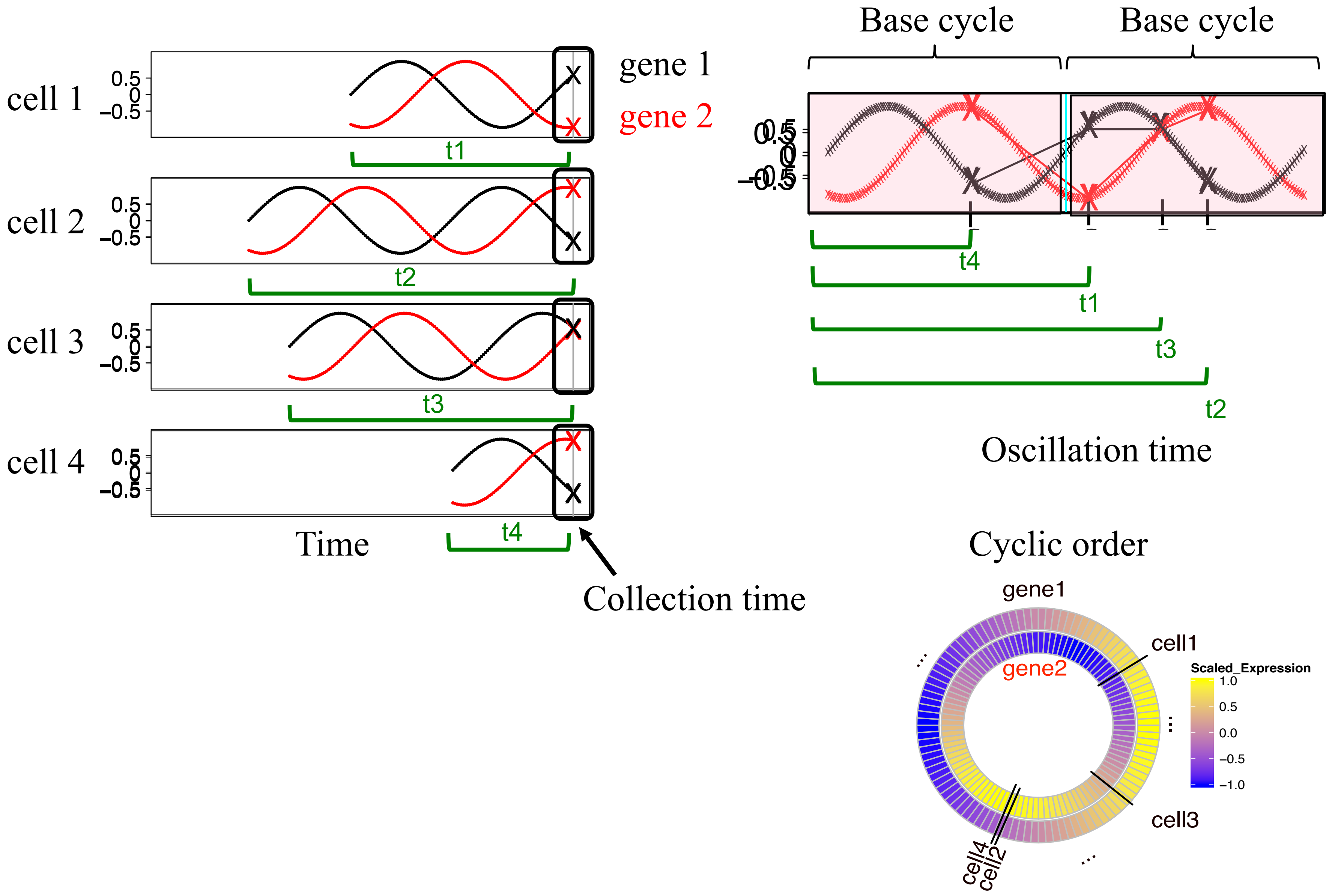
Step 2: Residuals from paired-sine model are used in clustering to group genes with similar frequency, varying phase



Step 3: Extension of the nearest insertion algorithm reconstructs one base cycle for every group



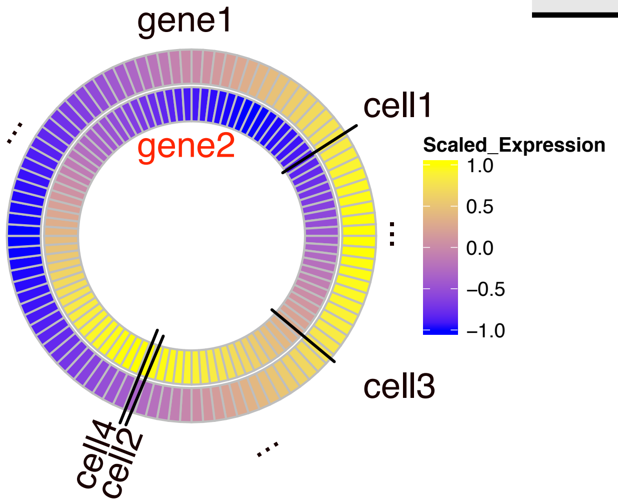
Oscope: Identifying oscillatory genes using scRNA-seq



Oscope: Connection to TSP

- Given a list of cities, only know distances between each pair of cities
- Goal is to find an optimal route that visits each city exactly once and returns to the origin city
- The optimal route will minimize overall distance travelled

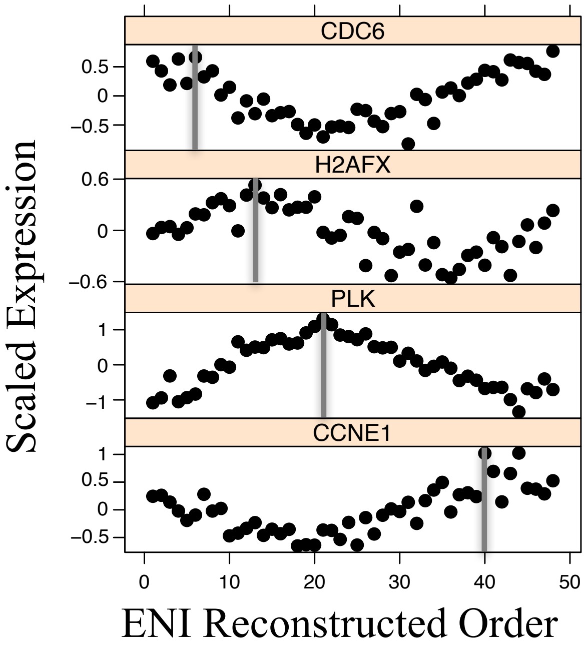
	Distance (mile)
Madison-Chicago	148
Madison-Iowa City	175
Madison-Atlanta	847
Madison-Minneapolis	273
Chicago-Iowa City	223
Chicago-Atlanta	716
…	…

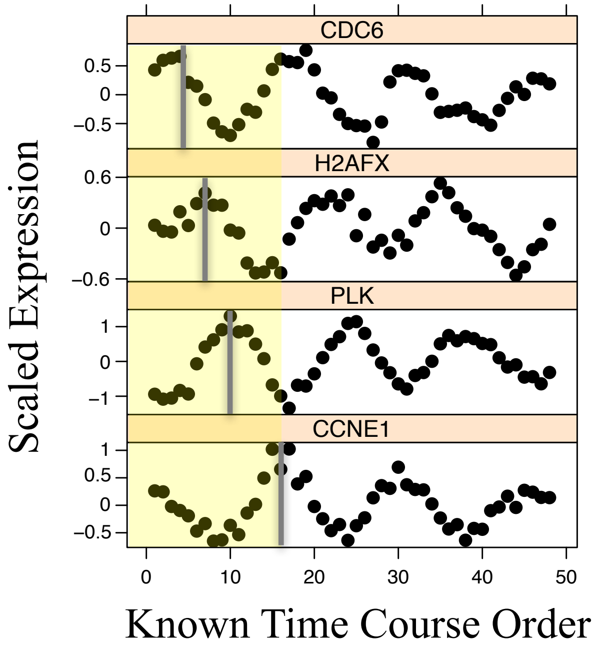


- In our case, for each gene cluster, we want to find an optimal "route" through all cells and return to the first cell
- The optimal route will minimize expression differences between observed and baseline oscillation

Oscope: Results from Whitfield data

- Whitfield data: microarray time course of HeLa cells synchronized for cell cycle. 48 samples; one every hour (~3 cell cycles).
- Applied Oscope on Whitfield data with permuted sample order
  - Top cluster has 69 genes (65 of 69 validated as oscillating in Whitfield).

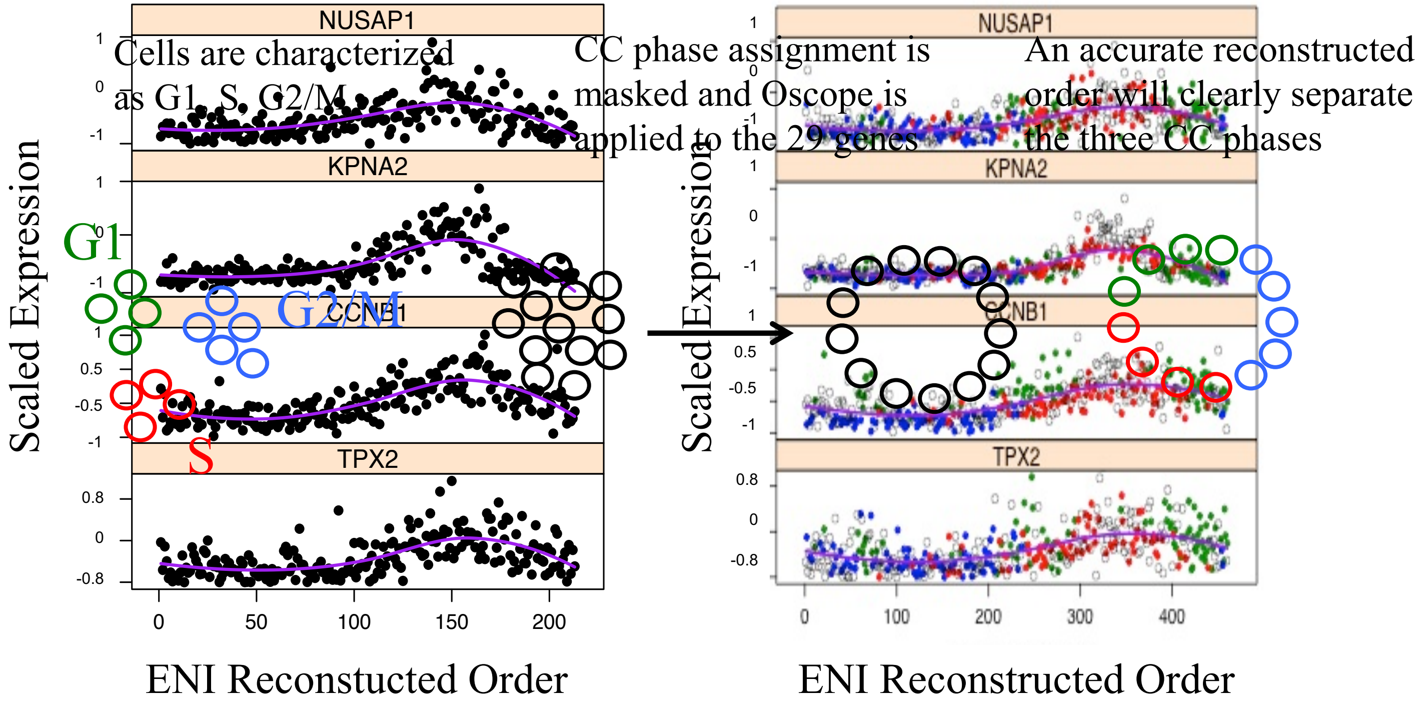




Oscope: Results from H1 hESCs (with Thomson lab)

- Oscope applied to 213 H1 hESCs identified a 29 gene group
  - 21 of 29 genes annotated as cell-cycle by GO.

- To investigate this group, Oscope was reapplied to 460 H1 hESCs
  - 213 unlabeled and 247 FUCCI labeled (cell cycle phase is known).

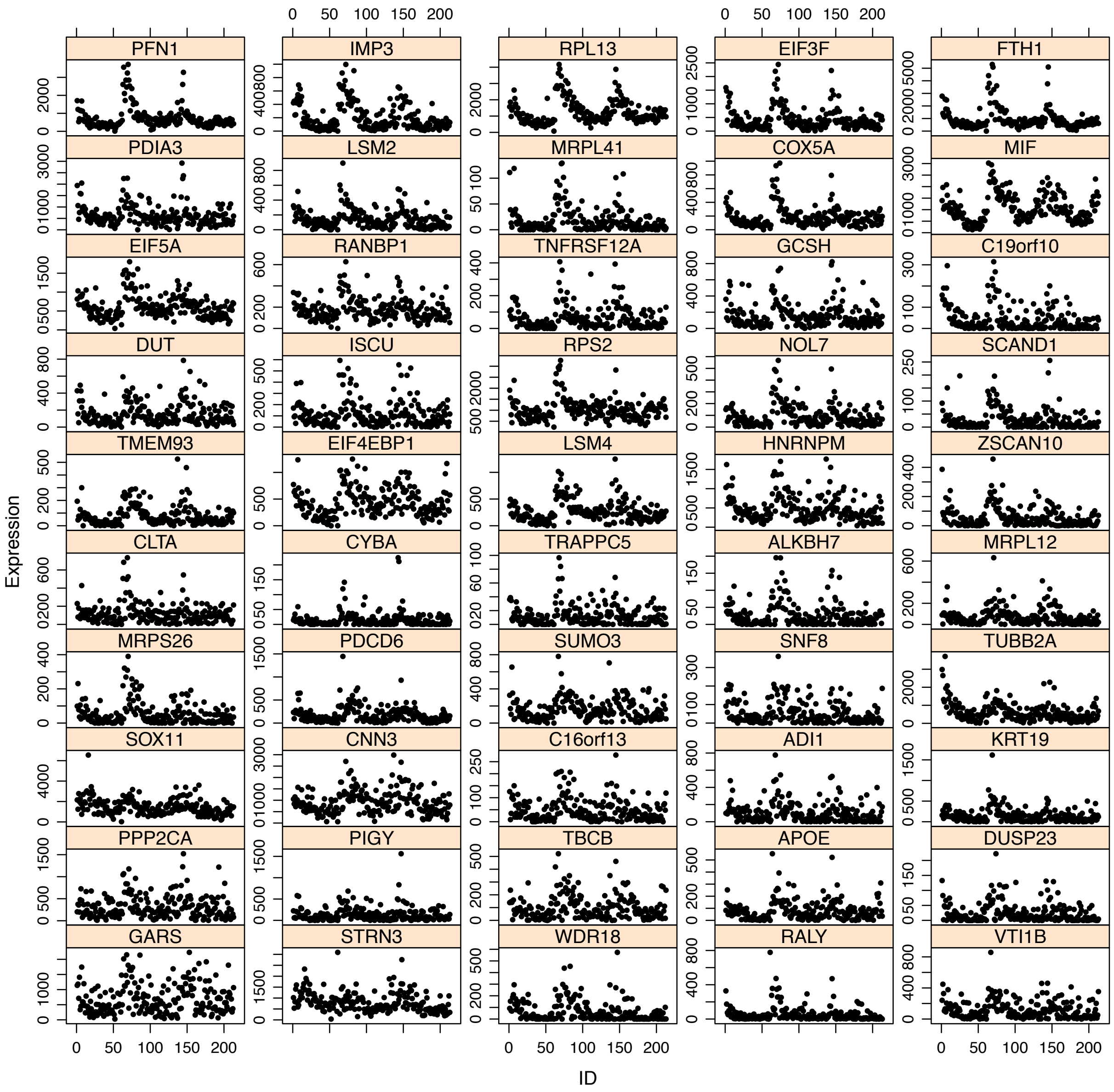


Cells are characterized as G1, S, G2/M

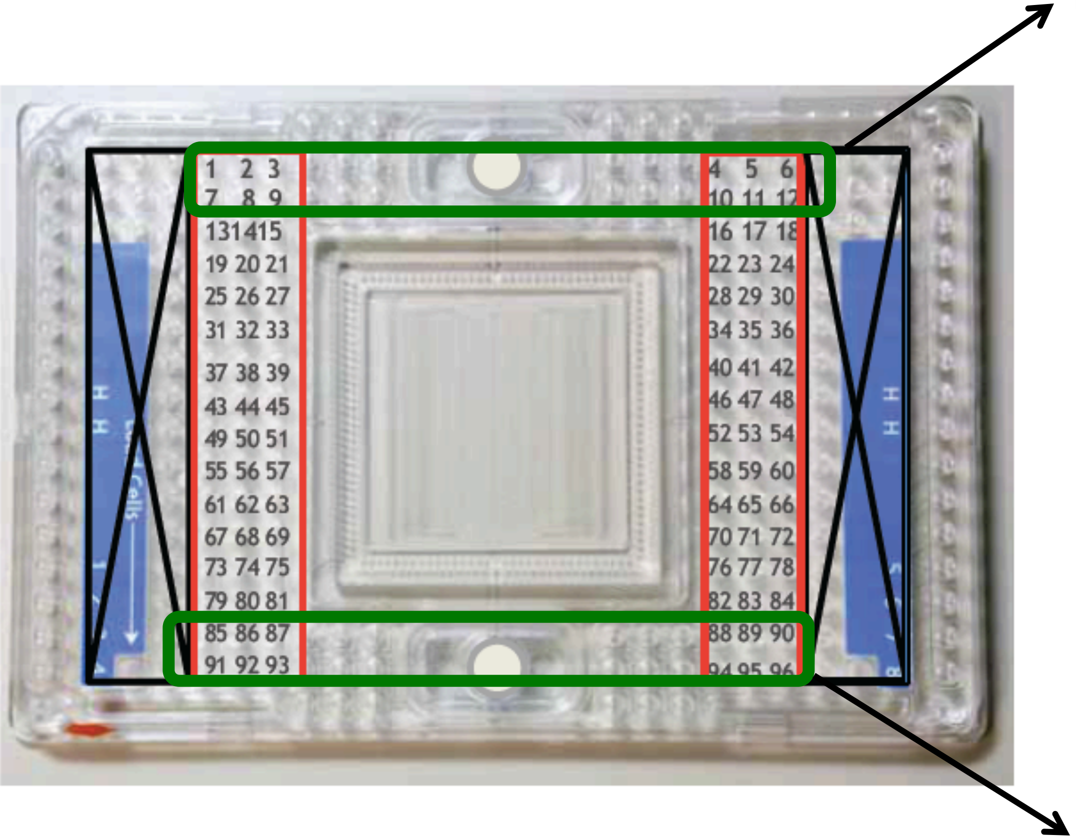
CC phase assignment is masked and Oscope is applied to the 29 genes

An accurate reconstructed order will clearly separate the three CC phases

Oscope identifies potential artifact in Fluidigm C1 platform

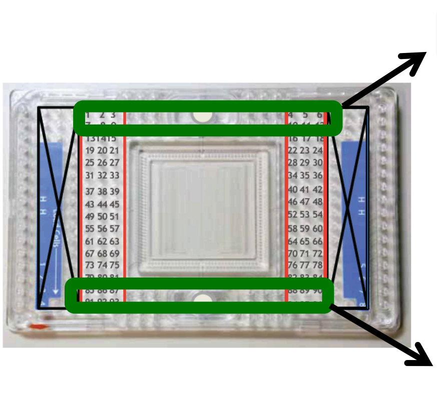


Schematic of Fluidigm’s C1 platform

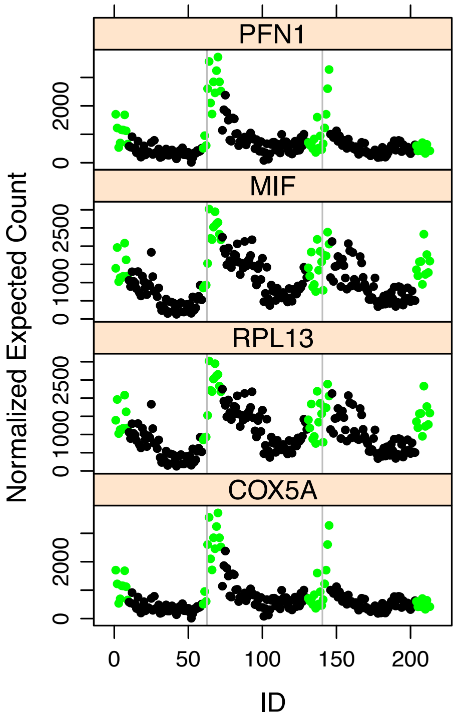


C01 (P03)	C49 (P04)
C02 (P02)	C50 (P05)
C03 (P01)	C51 (P06)
C04 (P09)	C52 (P10)
C05 (P08)	C53 (P11)
C06 (P07)	C54 (P12)
C07 (P15)	C55 (P16)
C08 (P14)	C56 (P17)
C09 (P13)	C57 (P18)
C10 (P21)	C58 (P22)
C11 (P20)	C59 (P23)
C12 (P19)	C60 (P24)
C13 (P27)	C61 (P28)
C14 (P26)	C62 (P29)
C15 (P25)	C63 (P30)
C16 (P33)	C64 (P34)
C17 (P32)	C65 (P35)
C18 (P31)	C66 (P36)
C19 (P39)	C67 (P40)
C20 (P38)	C68 (P41)
C21 (P37)	C69 (P42)
C22 (P45)	C70 (P46)
C23 (P44)	C71 (P47)
C24 (P43)	C72 (P48)
C25 (P49)	C73 (P54)
C26 (P50)	C74 (P53)
C27 (P51)	C75 (P52)
C28 (P55)	C76 (P60)
C29 (P56)	C77 (P59)
C30 (P57)	C78 (P58)
C31 (P61)	C79 (P66)
C32 (P62)	C80 (P65)
C33 (P63)	C81 (P64)
C34 (P67)	C82 (P72)
C35 (P68)	C83 (P71)
C36 (P69)	C84 (P70)
C37 (P73)	C85 (P78)
C38 (P74)	C86 (P77)
C39 (P75)	C87 (P76)
C40 (P79)	C88 (P84)
C41 (P80)	C89 (P83)
C42 (P81)	C90 (P82)
C43 (P85)	C91 (P90)
C44 (P86)	C92 (P89)
C45 (P87)	C93 (P88)
C46 (P91)	C94 (P96)
C47 (P92)	C95 (P95)
C48 (P93)	C96 (P94)

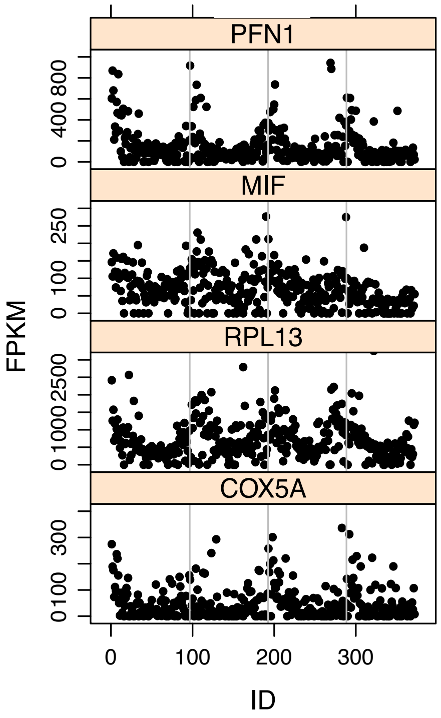
Increased expression related to capture site ID



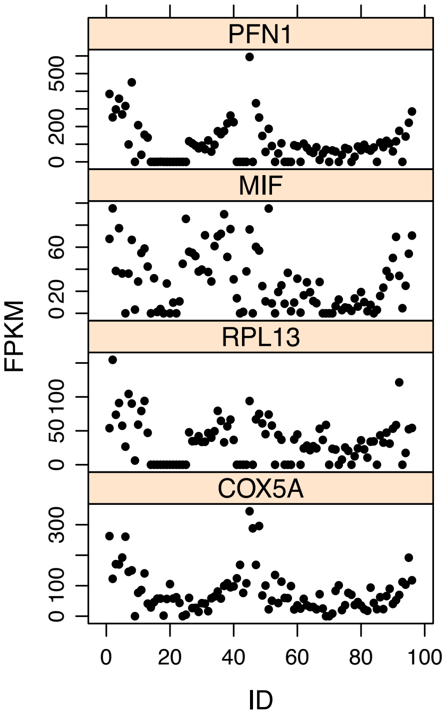
C01 (P03)	C49 (P04)
C02 (P02)	C50 (P05)
C03 (P01)	C51 (P06)
C04 (P09)	C52 (P10)
C05 (P08)	C53 (P11)
C06 (P07)	C54 (P12)
C07 (P15)	C55 (P16)
C08 (P14)	C56 (P17)
C09 (P13)	C57 (P18)
C10 (P21)	C58 (P22)
C11 (P20)	C59 (P23)
C12 (P19)	C60 (P24)
C13 (P27)	C61 (P28)
C14 (P26)	C62 (P29)
C15 (P25)	C63 (P30)
C16 (P33)	C64 (P34)
C17 (P32)	C65 (P35)
C18 (P31)	C66 (P36)
C19 (P39)	C67 (P40)
C20 (P38)	C68 (P41)
C21 (P37)	C69 (P42)
C22 (P45)	C70 (P46)
C23 (P44)	C71 (P47)
C24 (P43)	C72 (P48)
C25 (P49)	C73 (P54)
C26 (P50)	C74 (P53)
C27 (P51)	C75 (P52)
C28 (P55)	C76 (P60)
C29 (P56)	C77 (P59)
C30 (P57)	C78 (P58)
C31 (P61)	C79 (P66)
C32 (P62)	C80 (P65)
C33 (P63)	C81 (P64)
C34 (P67)	C82 (P72)
C35 (P68)	C83 (P71)
C36 (P69)	C84 (P70)
C37 (P73)	C85 (P78)
C38 (P74)	C86 (P77)
C39 (P75)	C87 (P76)
C40 (P79)	C88 (P84)
C41 (P80)	C89 (P83)
C42 (P81)	C90 (P82)
C43 (P85)	C91 (P90)
C44 (P86)	C92 (P89)
C45 (P87)	C93 (P88)
C46 (P91)	C94 (P96)
C47 (P92)	C95 (P95)
C48 (P93)	C96 (P94)



Trapnell *et al.*, 2014



Wu *et al.*, 2014



Challenges in scRNA-seq

- Normalization
- Technical vs. biological zeros
- De-noising
- Clustering; Identifying sub-populations
- Identifying oscillatory genes
- Identifying and characterizing differences in gene-specific expression distributions (aka. identifying differential distributions)
- Pseudotime reordering
- Network reconstruction

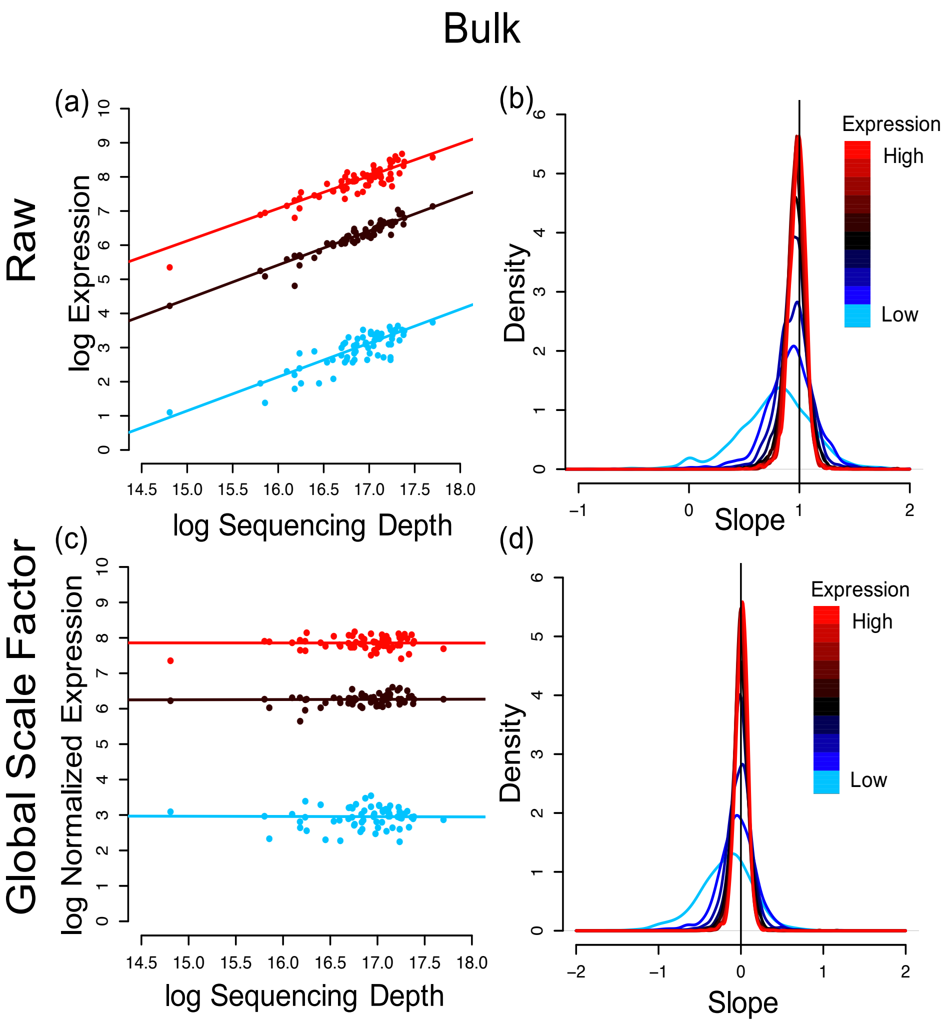
SCnorm: A quantile-regression based approach for robust normalization of single-cell RNA-seq data

Bacher, Chu *et al*., *Nature Methods*, 2017

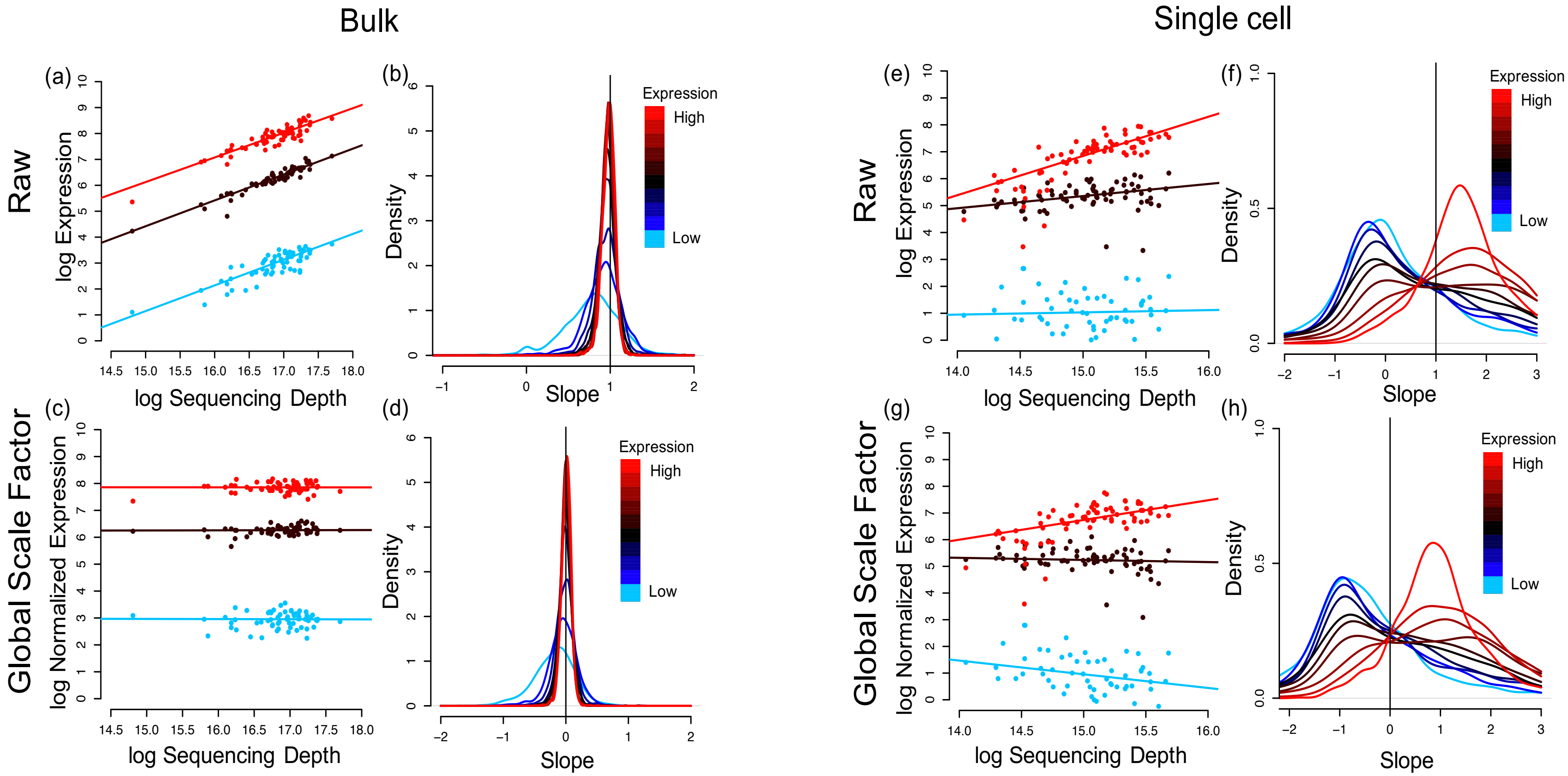
Background

- Goal: correct for technical artifacts and/or gene-specific features
  - Sequencing depth
  - Length, GC content
  - Amplification and other technical biases
- Without UMIs/spike-ins, most single-cell methods calculate global scale factors as in bulk RNA-seq
  - One scale factor is calculated per sample and applied to all genes in that sample.

Bulk: Global scale-factor normalization for sequencing depth



Expression vs. depth varies with expression in scRNA-seq



We see the count-depth relationship varying with expression in many datasets

Overview of SCnorm

- Identify gene groups based on the count-depth relationship.

Within each group,

- Quantile polynomial regression is used to quantify the group-specific relationship between expression and sequencing depth. The quantile is chosen iteratively.
- Predicted values are used to calculate group-specific scale factors for each cell.

SCnorm

- Filter: genes having greater than 10% expression values nonzero and median nonzero expression greater than 2.
- Let $Y_g = (y_{g1},\ldots,y_{gJ})$ denote log non-zero expression for gene g in cell j ; X_j denote log sequencing depth.
- The gene-specific count-depth relationship is estimated by:

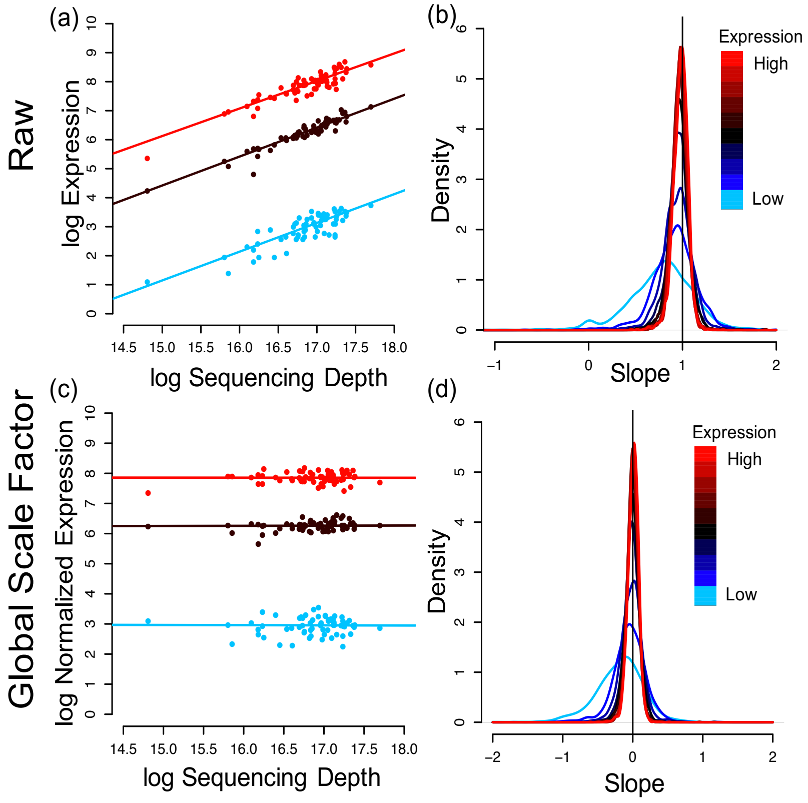
$$Q^{0.5}\left(Y_{g,j}|X_j\right) = \beta_{g,0} + \beta_{g,1}X_j$$

- Genes are split into K groups. The group specific count-depth relationship is estimated by:

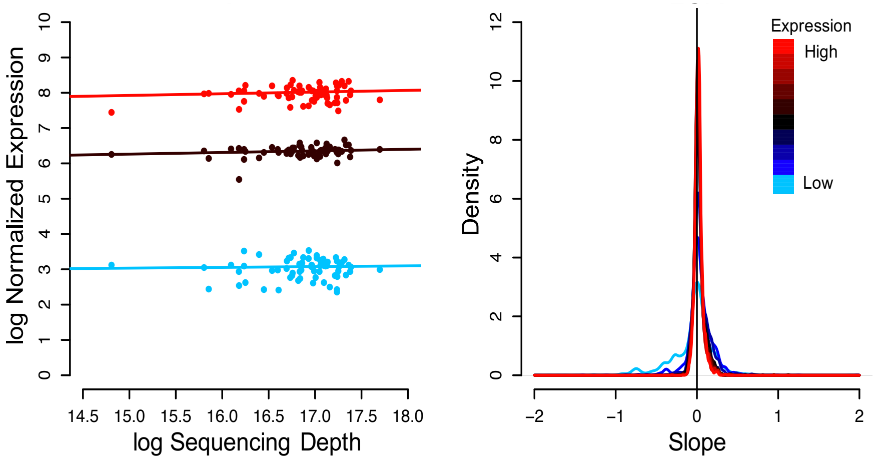
$$Q^{\tau_k,d_k}\left(Y_j|X_j\right) = \beta_0^{\tau_k} + \beta_1^{\tau_k}X_j + \cdots + \beta_d^{\tau_k}X_j^{d_k}$$

- Estimates of τ_k and d_k minimize $|\hat{\eta}_1^{\tau_k} - {}^{mode}_{g}\hat{\beta}_{g,1}|$; where $\hat{\eta}_1^{\tau_k}$ represents the count-depth relationship among predicted values.
- K is chosen so that the absolute value of the maximum normalized slope mode is < 0.1 within each of ten groups.

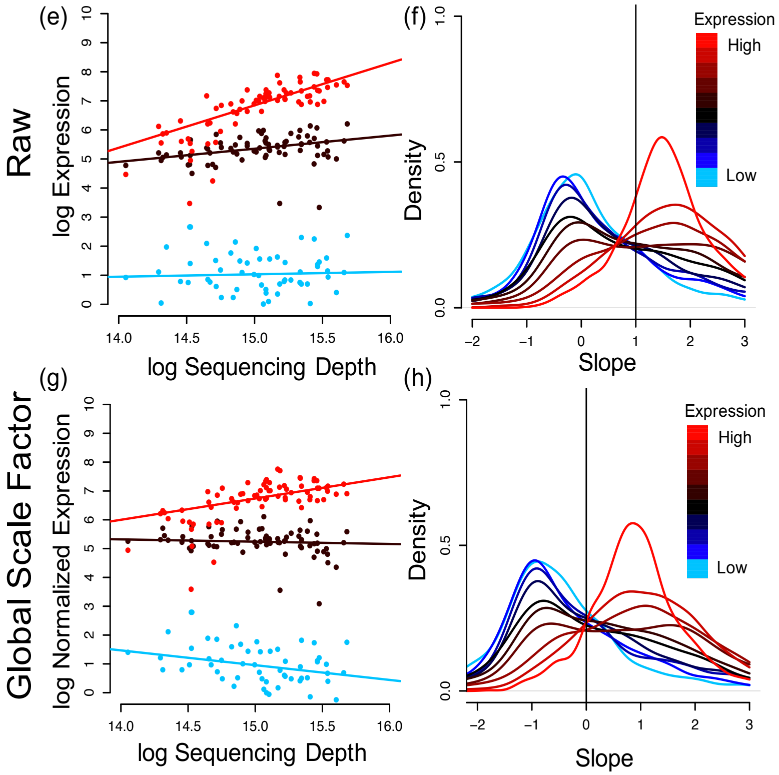
Bulk RNA-seq



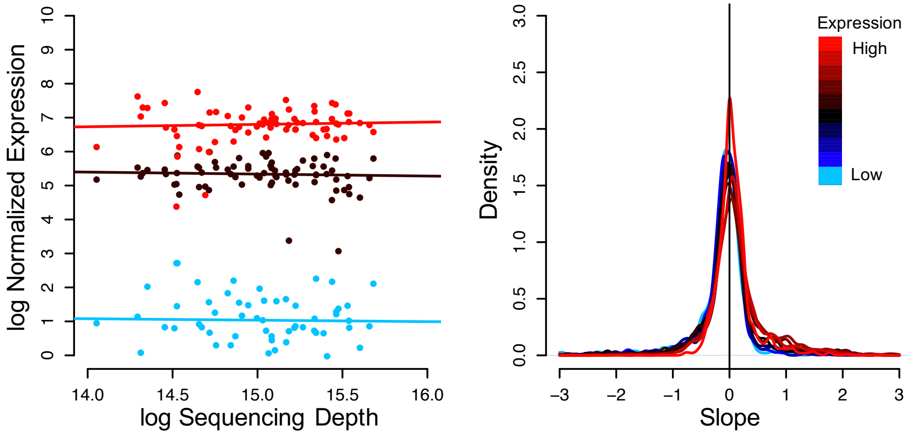
SCnorm



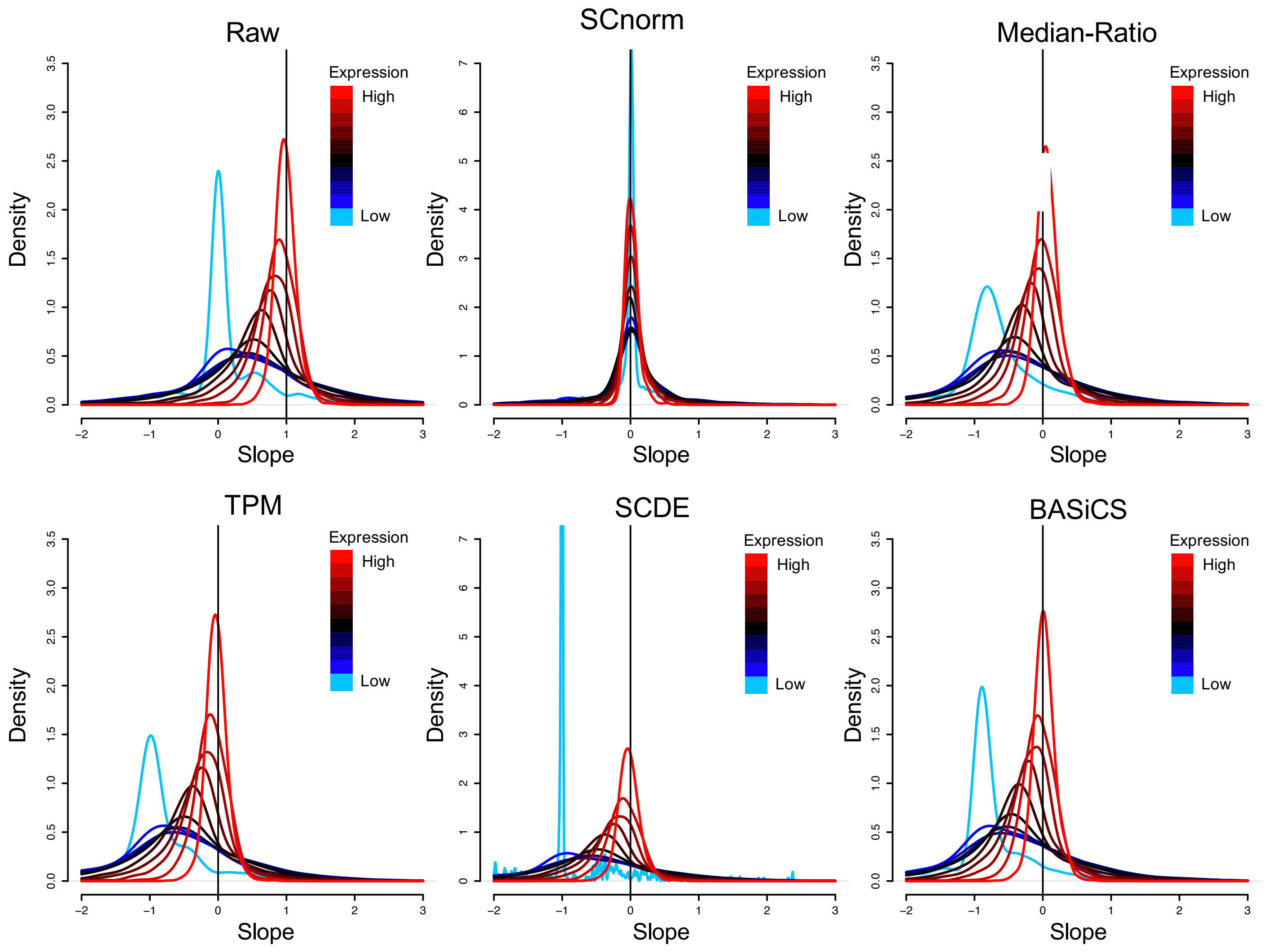
Single-cell RNA-seq



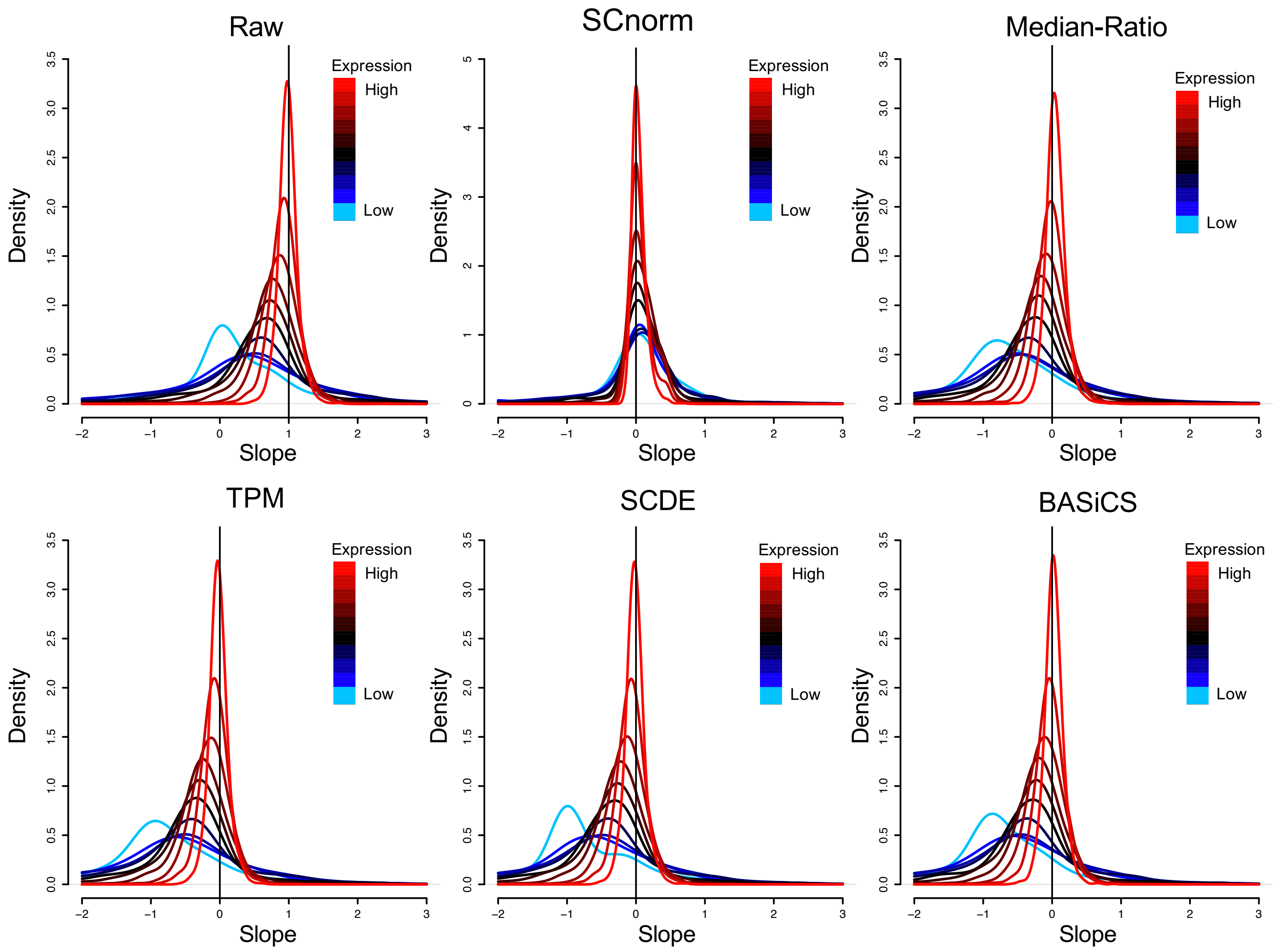
SCnorm



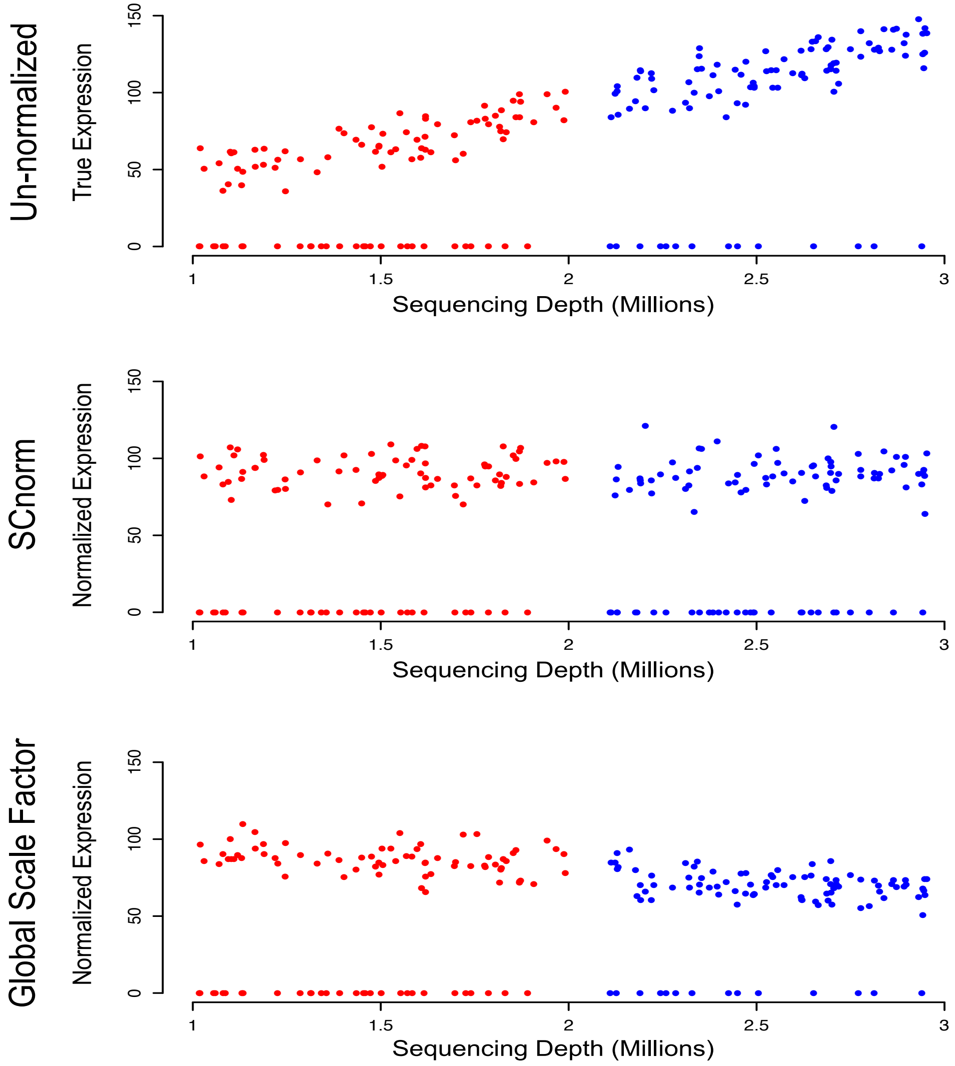
H1 - 1 (~ 1 million reads per cell)



H1 - 4 (~4 million reads per cell)

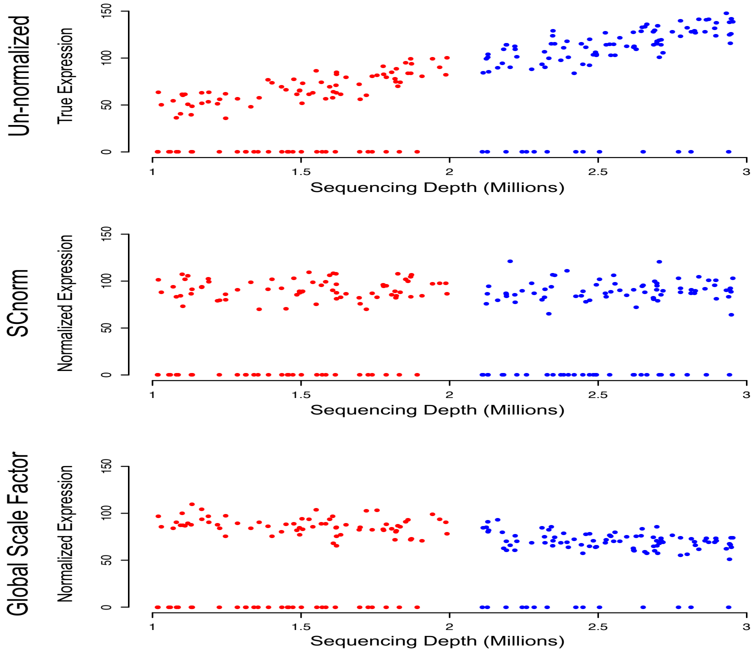


Implications for DE analysis



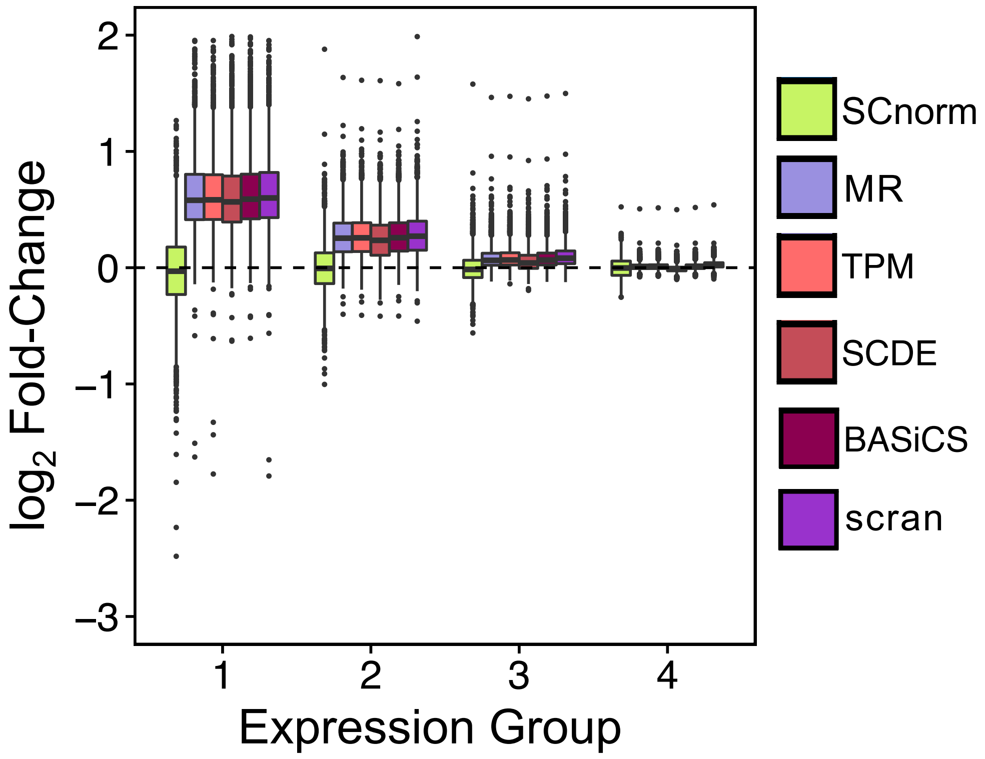
FC= H1-1/H1-4

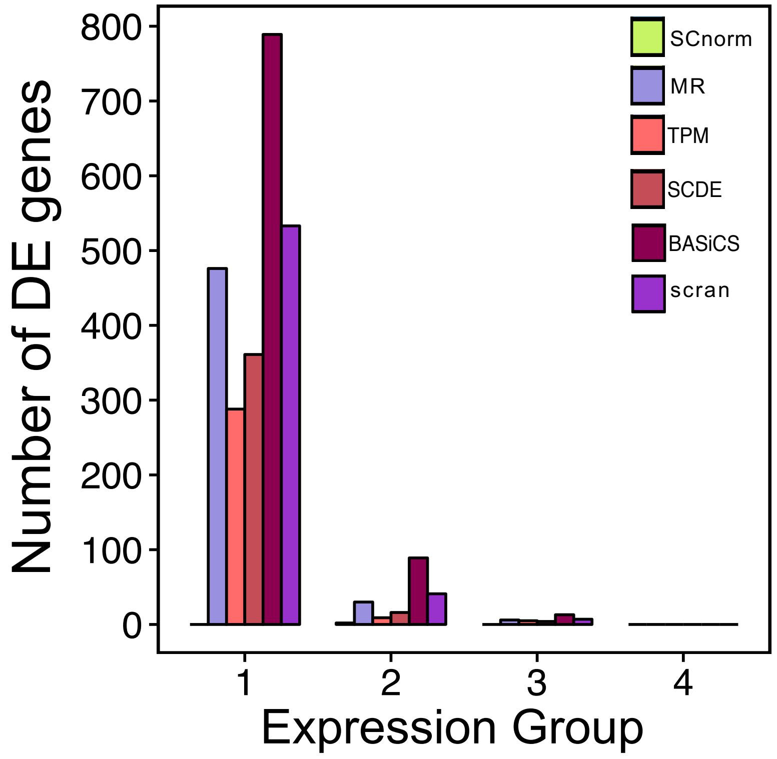
- H1-1: ~100 H1 cells profiles at ~1 million reads per cell
- H1-4: Same H1 cells profiled at ~4 million reads per cell
- Prior to normalization, H1-1/H1-4 should be about ¼
- Post normalization, H1-1/H1-4 should be about 1
- If over-normalization is going on, H1-1/H1-4 will be greater than 1.



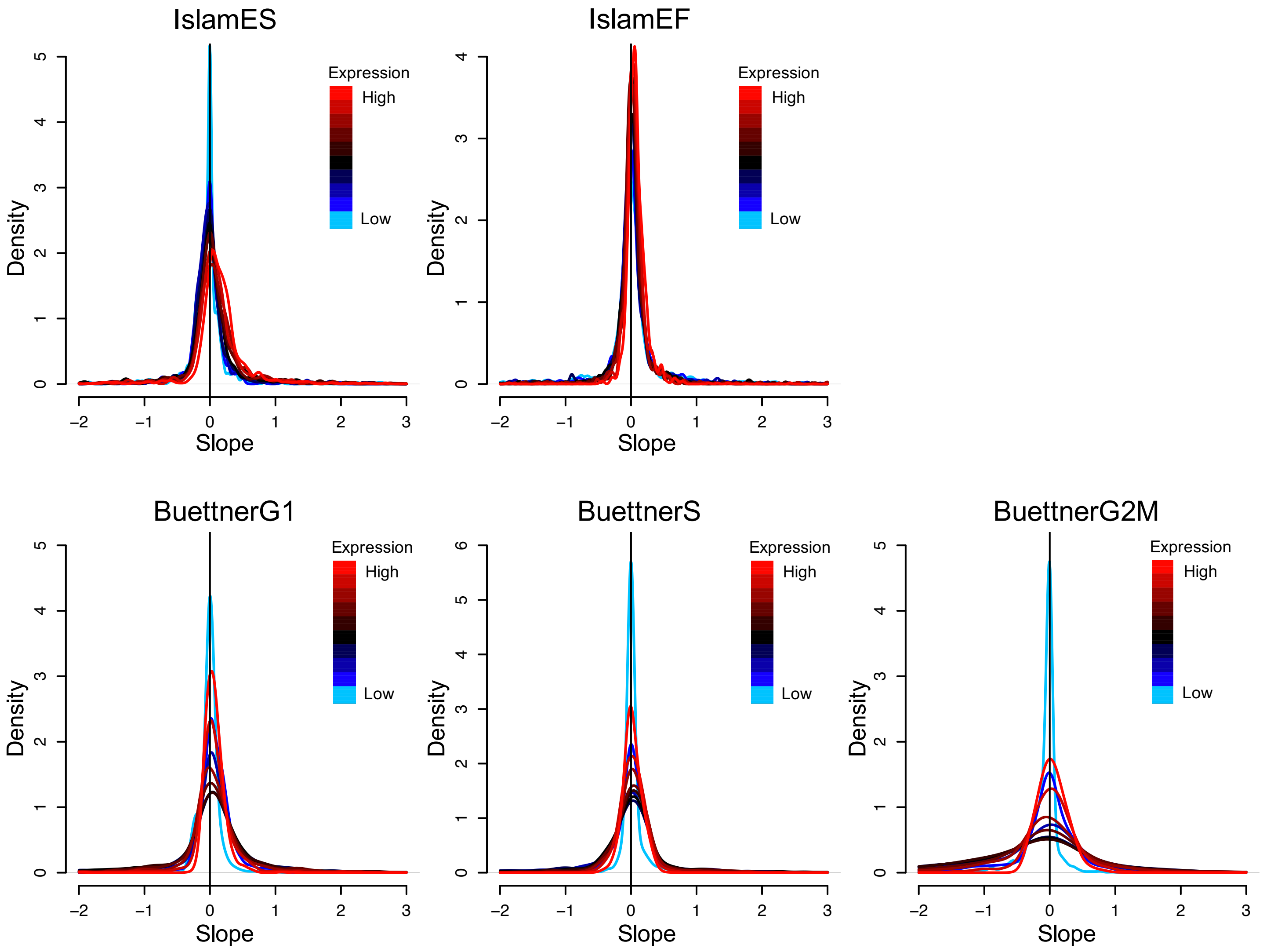
FC= H1-1/H1-4

- H1-1: ~100 H1 cells profiles at ~1 million reads per cell
- H1-4: Same H1 cells profiled at ~4 million reads per cell





Normalization via SCnorm



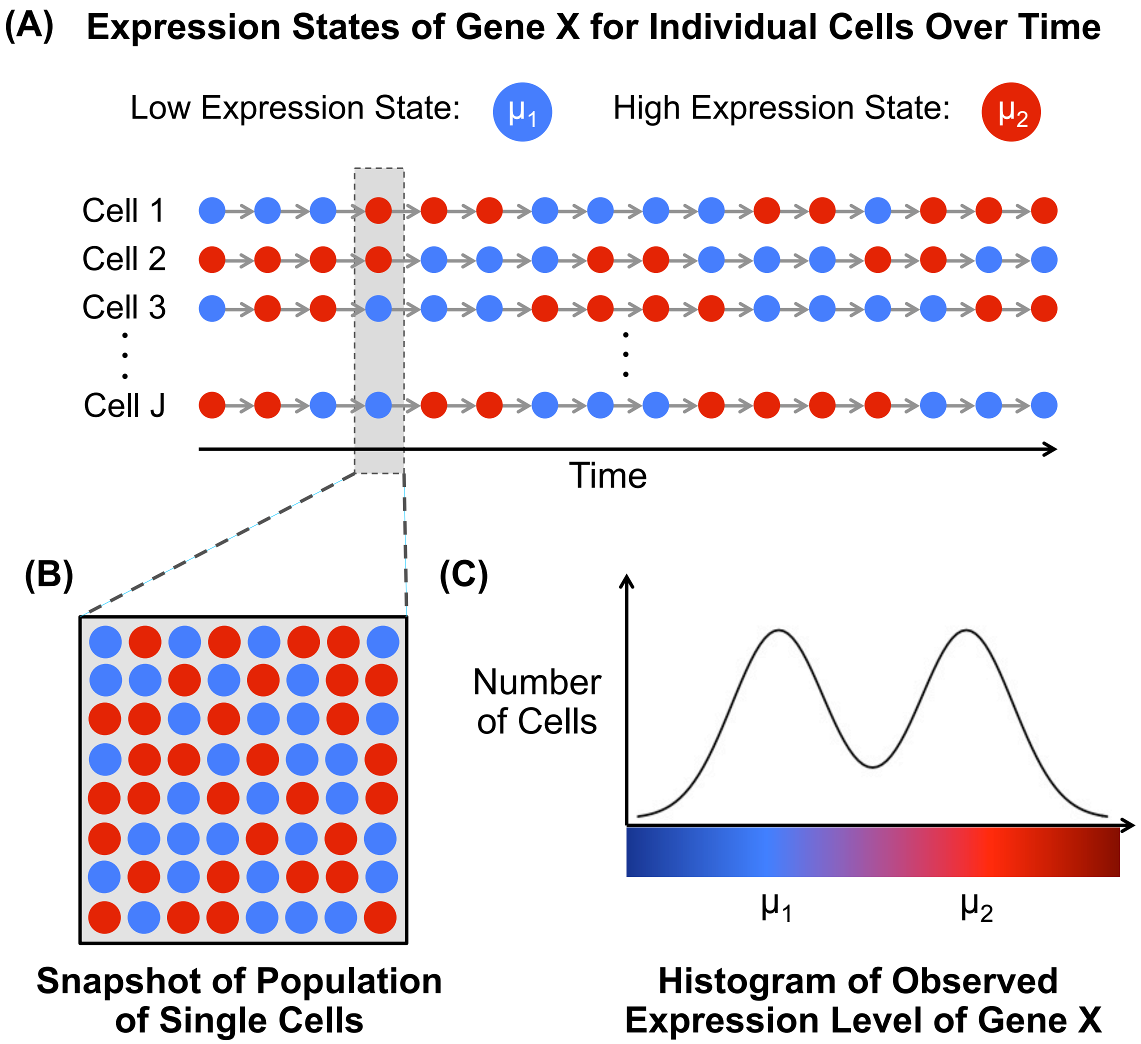
Challenges in scRNA-seq

- Normalization
- Technical vs. biological zeros
- De-noising
- Clustering; Identifying sub-populations
- Identifying oscillatory genes
- Identifying and characterizing differences in gene-specific expression distributions (aka. identifying differential distributions)
- Pseudotime reordering
- Network reconstruction

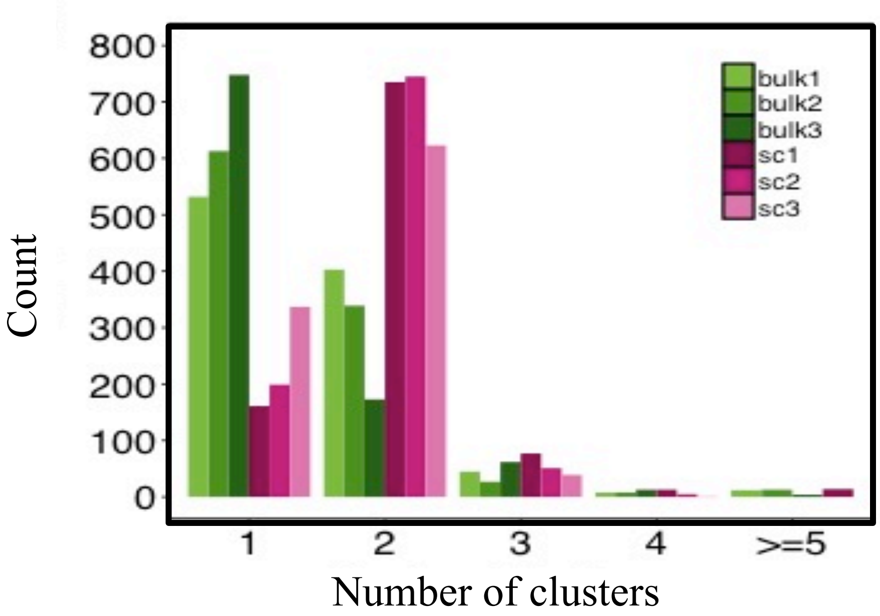
scDD: A Dirichlet mixture model based approach for identifying differential distributions in scRNA-seq experiments

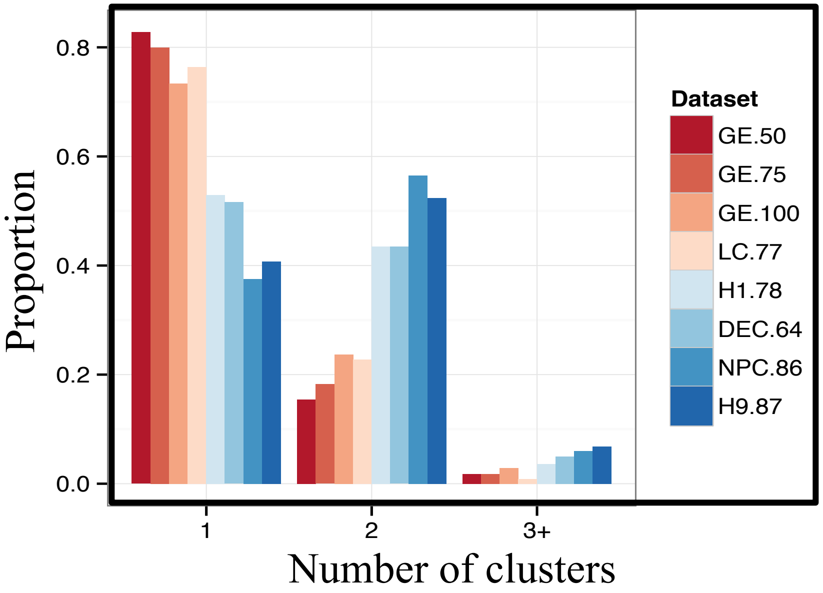
Korthauer *et al*., *Genome Biology,* to appear, 2016

Gene-specific multi-modality

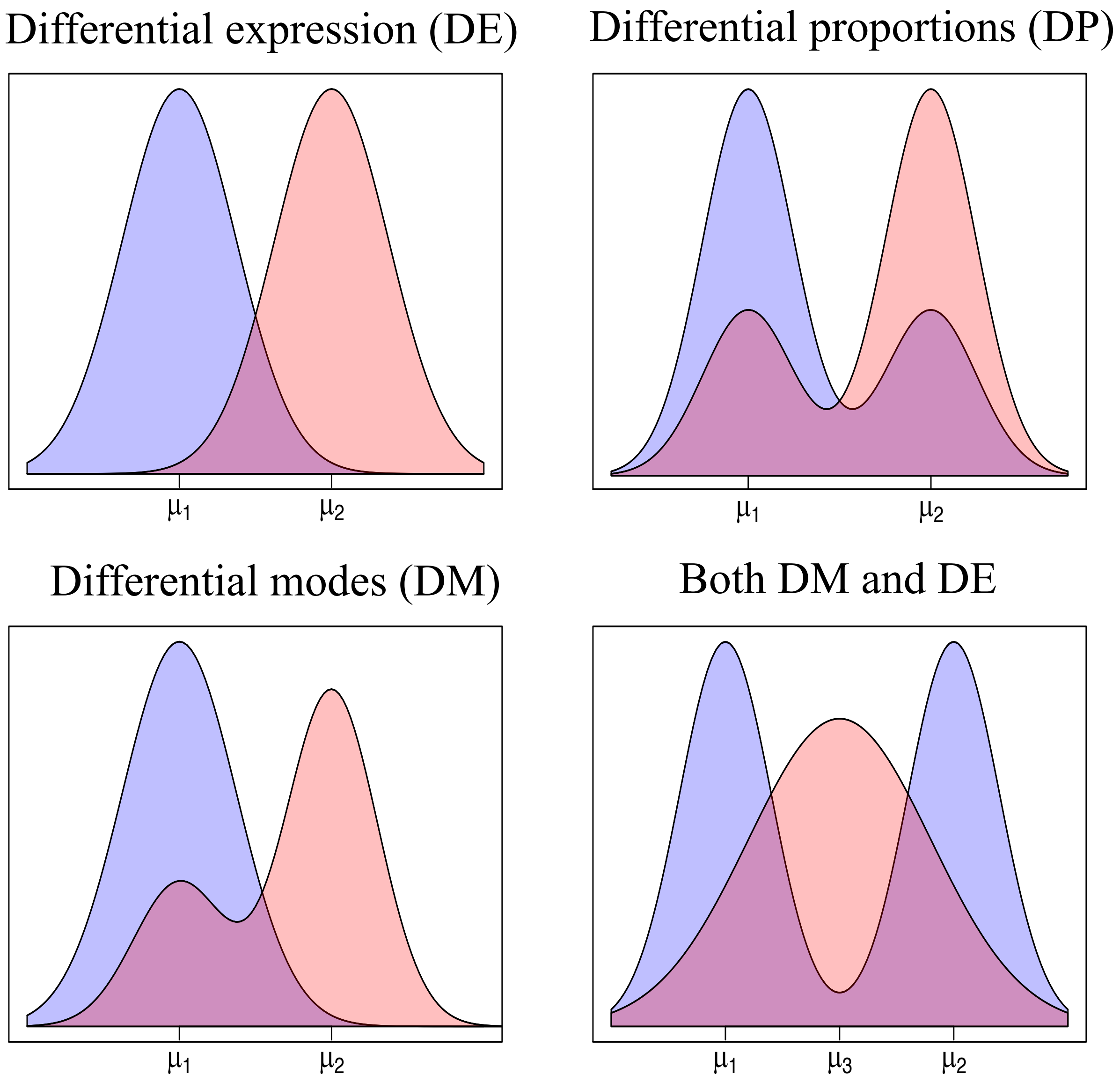

(A) Expression States of Gene X for Individual Cells Over Time
Low Expression State: μ_1
High Expression State: μ_2
Cell 1
Cell 2
Cell 3
Cell J
Time
(B)
Snapshot of Population of Single Cells
(C)
Number of Cells
μ_1
μ_2
Histogram of Observed Expression Level of Gene X

Many genes show multi-modal expression distributions





Opportunity to identify differences beyond traditional DE



Differential expression (DE)

Differential proportions (DP)

Differential modes (DM)

Both DM and DE

scRNA-seq DE Analysis

- Recent methods use mixture modeling to account for ‘on’ and ‘off’ components
  - Shalek et al. (2014)
  - SCDE (Kharchenko *et al.*, 2014)
  - MAST (Finak *et al.*, 2015)

- When detected, each gene has a latent level of expression within a biological condition, and measurements fluctuate around that level due to biological and technical sources of variability

scDD: Goal

- Model expression profiles while accommodating the often multimodal distributions in the detected cells

- Find genes with Differential Distributions (DD) of expression across two conditions:
  - differential means
  - differential proportion within modes
  - differential modality (number of modes)
  - combination thereof
  - differential zeroes (detection rate)

scDD: Overview

- Log non-zero normalized, de-noised, expression arises out of a fixed variance Dirichlet Process Mixture of normals model.

- For each gene, obtain maximum a posteriori (MAP) partition of the samples to components using the *modalclust* algorithm (Dahl 2009).
  - fast and deterministic
  - requires point estimate of cluster variance (obtain via *mclust*).

- To evaluate evidence of DD, fit under two different hypotheses:
  - ignoring condition ($\mathcal{M}_{ED}$: equivalent regulation)
  - separately for each condition ($\mathcal{M}_{DD}$: differential regulation)

scDD: Overview (continued)

- Assume that log non-zero normalized, de-noised, expression measurements $Y_g = (y_{g1},\ldots,y_{gJ})$ for gene g in J cells arise from a conjugate Dirichlet Process Mixture (DPM) of normals model:

$$\begin{aligned} y_j &\sim N(\mu_j, \tau_j) \\ \mu_j, \tau_j &\sim G \\ G &\sim DP(\alpha, G_0) \\ G_0 &= NG(m_0, s_0, a_0/2, 2/b_0) \end{aligned}$$

- Let K denote the number of components (unique values in $\{\mu_{j,}\ \tau_j\}$, j=1,.., J). Of primary interest is the posterior of (μ,τ), which is intractable for moderate sample sizes.

- Let $Z = (z_1, \ldots, z_J)$ denote component memberships. Then $f(Y \mid Z)$ is a PPM.

$$\begin{aligned} f(Y|Z) &= \prod_{k=1}^{K} f(y^{(k)}) \\ &\propto \prod_{k=1}^{K} \frac{\Gamma(a_k/2)}{(b_k/2)^{a_k/2}} s_k^{-1/2} \end{aligned}$$

scDD: Overview (continued)

- To quantify the evidence of DD for gene *g*, obtain MAP partition estimate, $\hat{Z}$, and evaluate $f\left(Y,\hat{Z}\right)$ under competing hypotheses:
  - ignoring condition ($\mathcal{M}_{ED}$: equivalent distributions)
  - separately within condition ($\mathcal{M}_{DD}$: differential distributions)

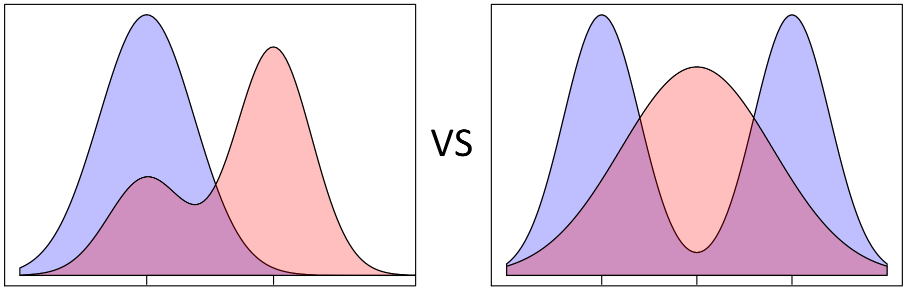
- Evaluate $\mathcal{M}_{DD}$ using a pseudo-Bayes Factor score:

$$Score_g = \log\left(\frac{f\left(Y_g,\hat{Z}_g|M_{DD}\right)}{f\left(Y_g,\hat{Z}_g|M_{ED}\right)}\right)$$

- Assess significance via permutation.

scDD: Classification of DD genes

- Classify DD genes into categories based on
  - number of components detected in each condition
  - whether clusters overlap



- Overlap is determined by sampling from the marginal posterior distribution of cluster means

$$\mu_k | Y, Z \sim t_{a_k}\left(m_k, \frac{b_k}{a_k s_k}\right)$$

scDD: Evaluation via simulation studies

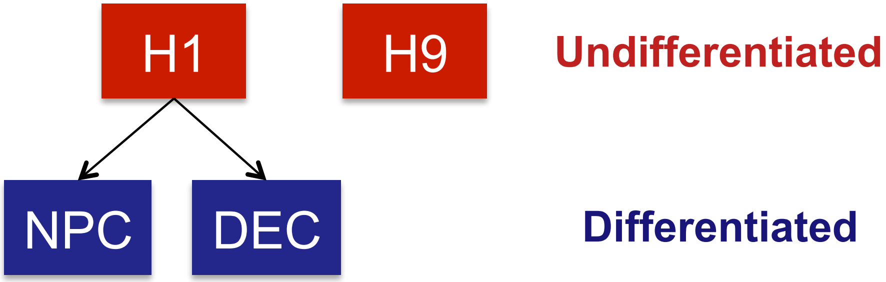
- 8000 ED genes:
  - 4000 from single Negative Binomial component
  - 4000 from two component mixture of Negative Binomial
- 2000 DD genes:
  - 500 DE genes
  - 500 DP genes (0.33/0.66 proportion difference)
  - 500 DM genes (0.50 belong to second mode)
  - 500 DB genes (mean in second condition is average of means in the first)
- Sample sizes varied $\in \{50, 75, 100\}$
- Component distances Δ_μ for multimodal conditions varied $\in \{2, 3, 4, 5, 6\}$ SDs
- Means, variances, and detection rates sampled empirically

Evaluate: Power to identify DD genes
Rate at which DD genes are correctly classified
Rate at which correct # components are identified

scDD: Power to detect DD genes within each category

		True Gene Category				
Sample Size	Method	DE	DP	DM	DB	Overall (FDR)
50	scDD	0.893	**0.418**	**0.898**	**0.572**	**0.695** (0.030)
	SCDE	0.872	0.026	0.816	0.260	0.494 (0.004)
	MAST	**0.908**	0.400	0.871	0.019	0.550 (0.026)
75	scDD	0.951	0.590	**0.960**	**0.668**	**0.792** (0.031)
	SCDE	0.948	0.070	0.903	0.387	0.577 (0.003)
	MAST	**0.956**	**0.632**	0.942	0.036	0.642 (0.022)
100	scDD	0.972	0.717	**0.982**	**0.727**	**0.850** (0.033)
	SCDE	0.975	0.125	0.946	0.478	0.631 (0.003)
	MAST	**0.977**	**0.752**	0.970	0.045	0.686 (0.022)
500	scDD	**1.000**	0.985	**1.00**	**0.903**	**0.972** (0.034)
	SCDE	**1.000**	0.858	0.998	0.785	0.910 (0.004)
	MAST	**1.000**	**0.992**	**1.00**	0.174	0.792 (0.021)

Comparison of hESCs

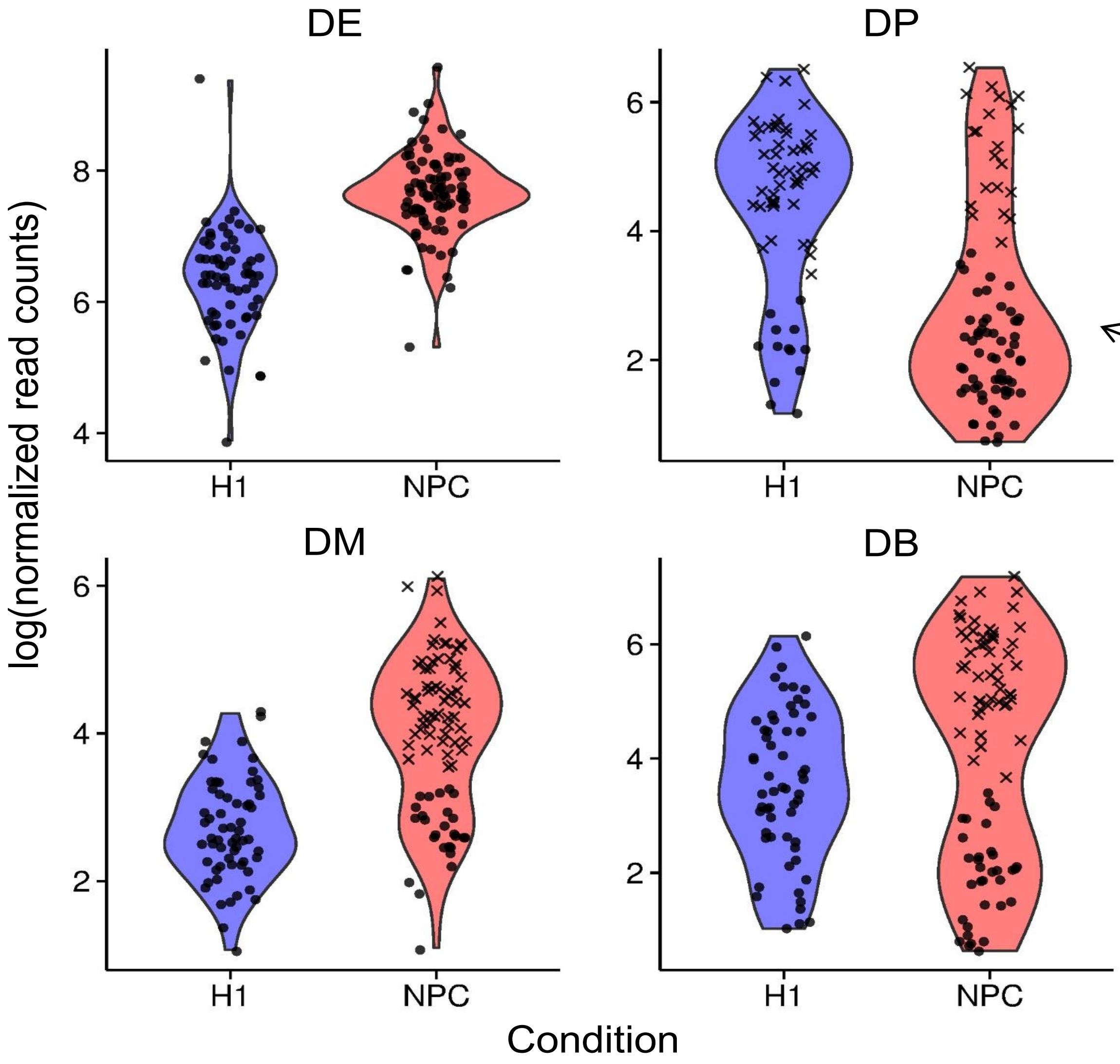


Number of DD genes identified in each cell type comparison

	scDD							
Comparison	DE	DP	DM	DB	DZ	Total	SCDE	MAST
H1 vs NPC	1342	429	739	406	1590	4506	2938	5729
H1 vs DEC	1408	404	939	345	880	3976	1581	3523
NPC vs DEC	1245	449	700	298	2052	4744	1881	5383
H1 vs H9	194	84	55	32	145	510	102	1091

scDD only: 2% 21% 38% 24% 15%

Genes identified in H1 vs. NPC comparison

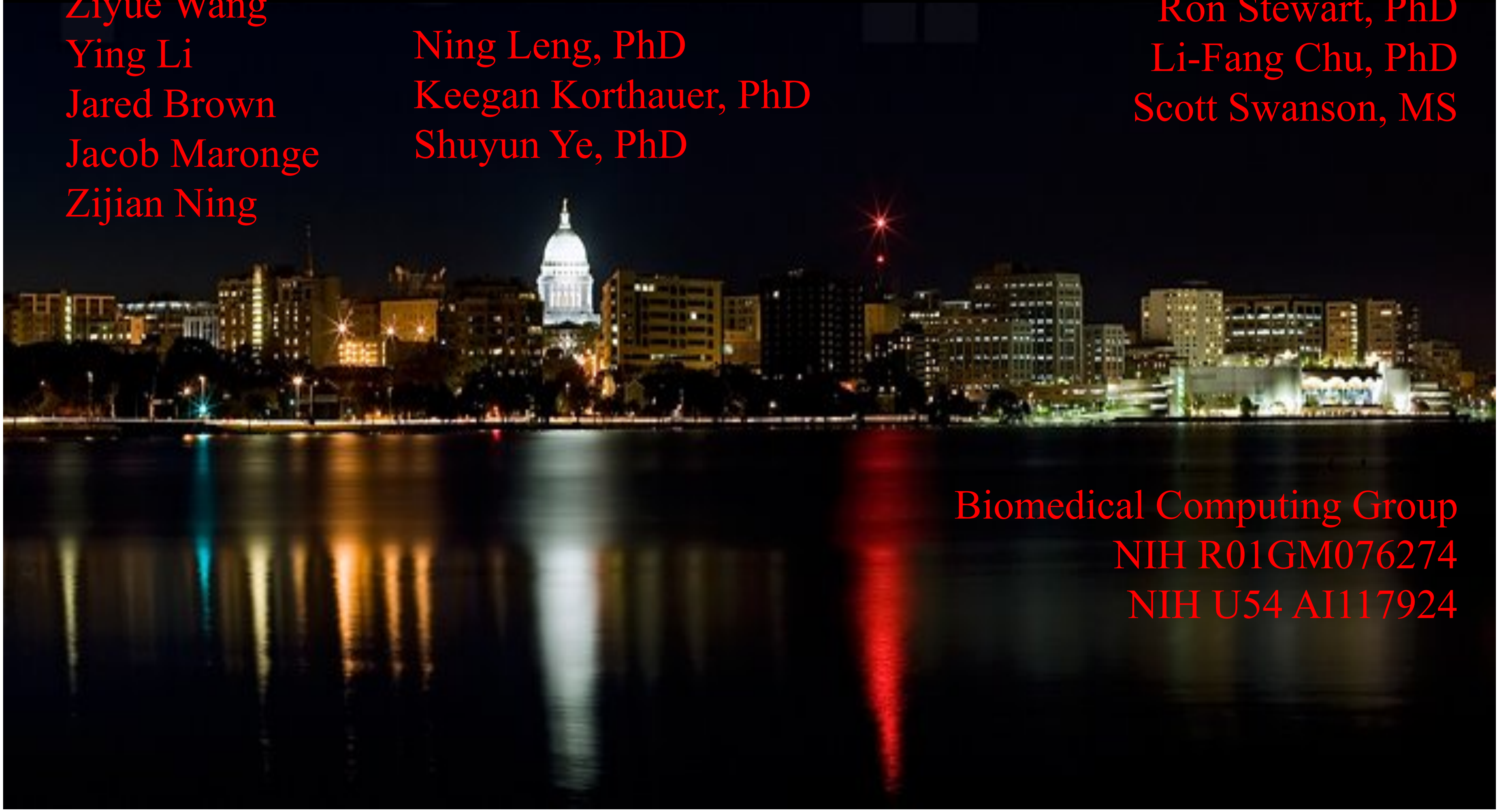


Not DE by SCDE or MAST

Summary

- Challenges due to zeros, increased variability, gene-specific distributions
  - Oscope: for identifying and characterizing oscillations in scRNA-seq experiments. Leng *et al*., *Nature Methods*, 2015.
  - OEFinder: for identifying genes with ordering effects due to position on IFC. Leng et al., *Bioinformatics*, 2016.
  - SCDC: for reducing the variation imposed by identified oscillators.
  - SCnorm: for scRNA-seq normalization. Bacher, Chu *et al*., *Nature Methods*, 2017.
- Opportunities
  - scDD for identifying differential distributions in scRNA-seq data. Korthauer *et al*., *Genome Biology*, 2016.
  - Wavecrest for identifying cell lineage. Chu *et al*., *Genome Biology*, 2016.

Acknowledgements



Rhonda Bacher, PhD
Ziyue Wang
Ying Li
Jared Brown
Jacob Maronge
Zijian Ning

Ning Leng, PhD
Keegan Korthauer, PhD
Shuyun Ye, PhD

Jamie Thomson, VMD, PhD
Ron Stewart, PhD
Li-Fang Chu, PhD
Scott Swanson, MS

Biomedical Computing Group
NIH R01GM076274
NIH U54 AI117924