scDD: A Statistical Approach for Identifying Differential Distributions in Single-Cell RNA-Seq Experiments

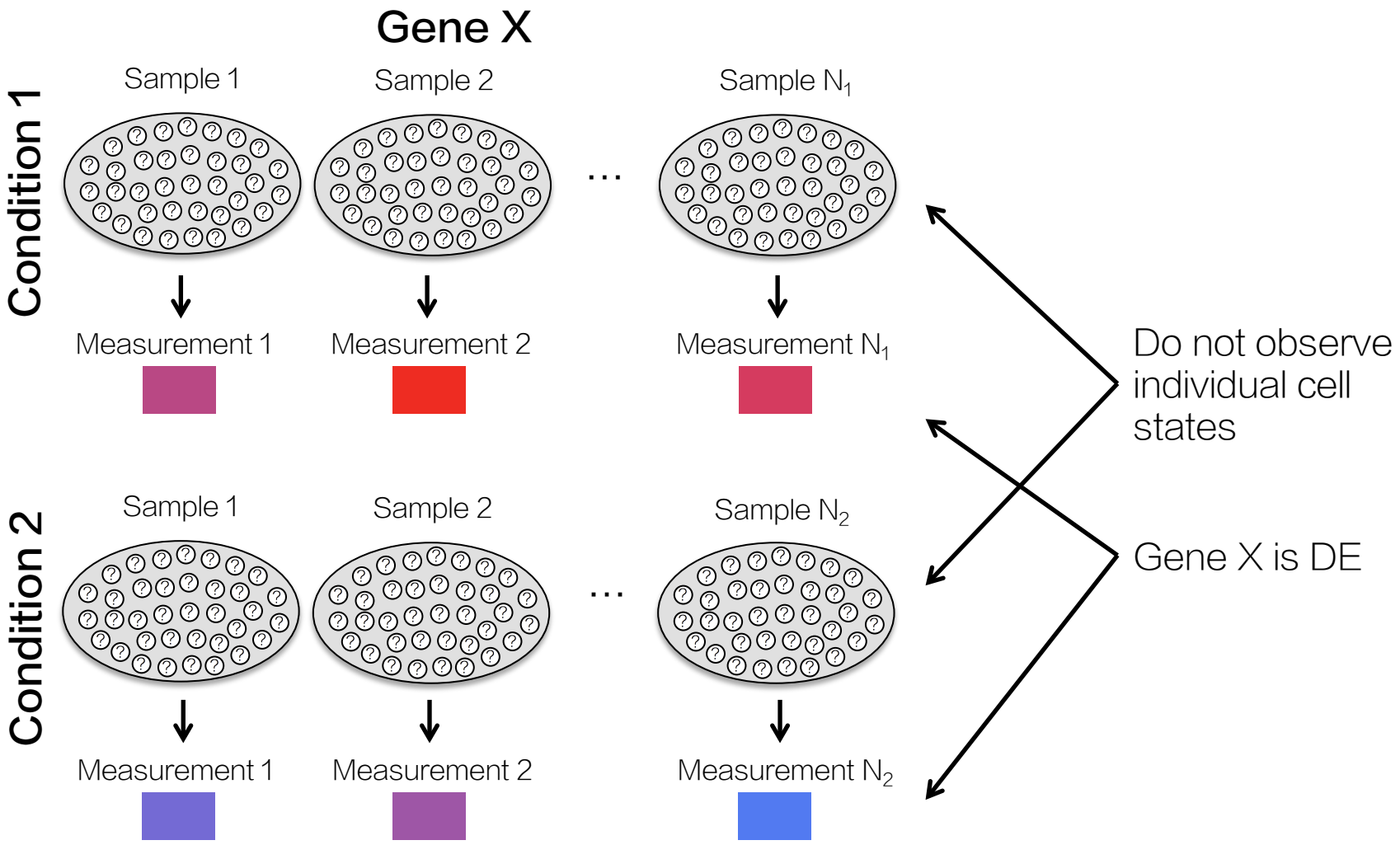
Keegan Korthauer, PhD
Postdoctoral Research Fellow, Irizarry Lab
Dana-Farber Cancer Institute
Harvard T. H. Chan School of Public Health
keegan@jimmy.harvard.edu

Joint Statistical Meetings, 8/2/2016

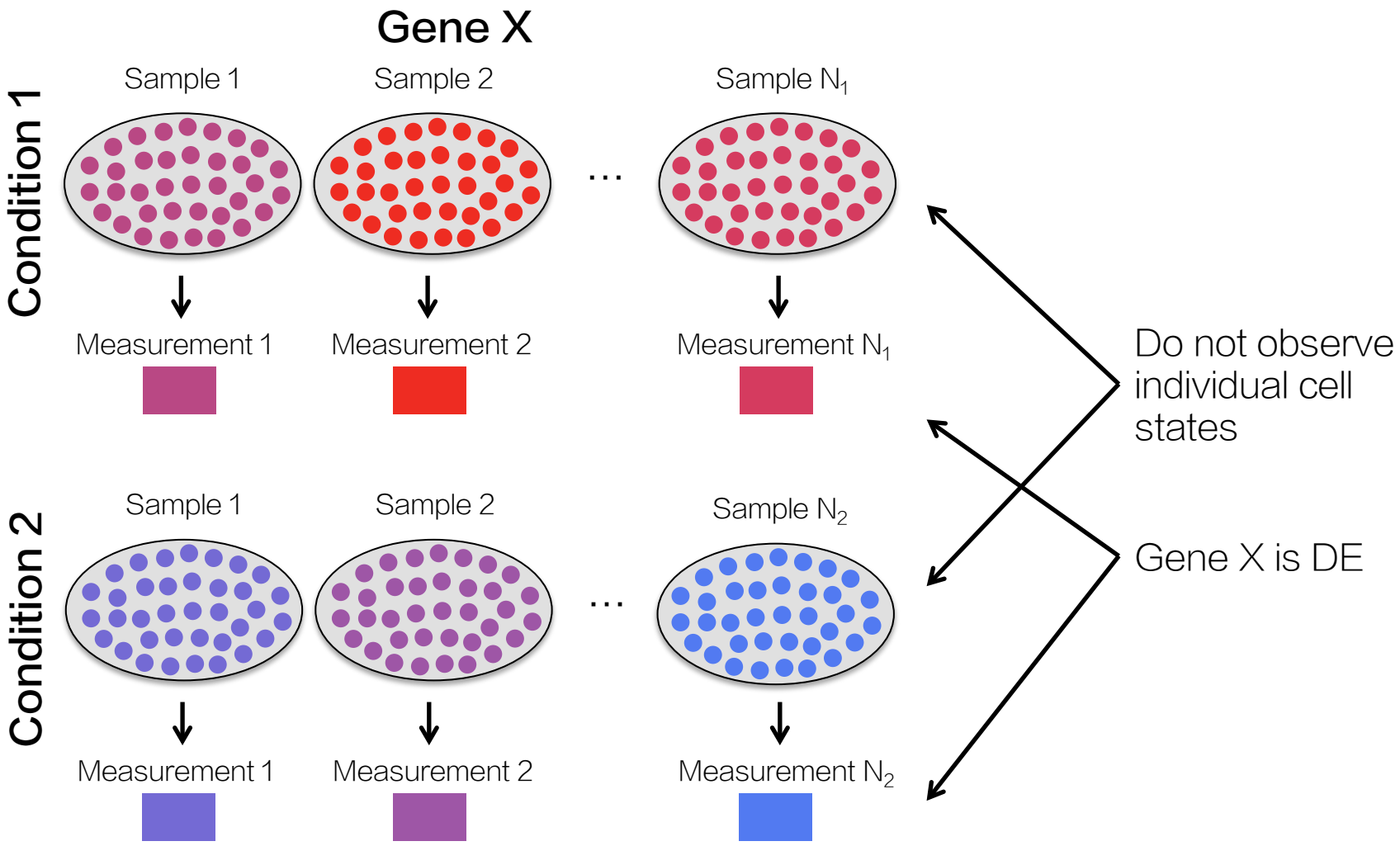
DANA-FARBER CANCER INSTITUTE

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

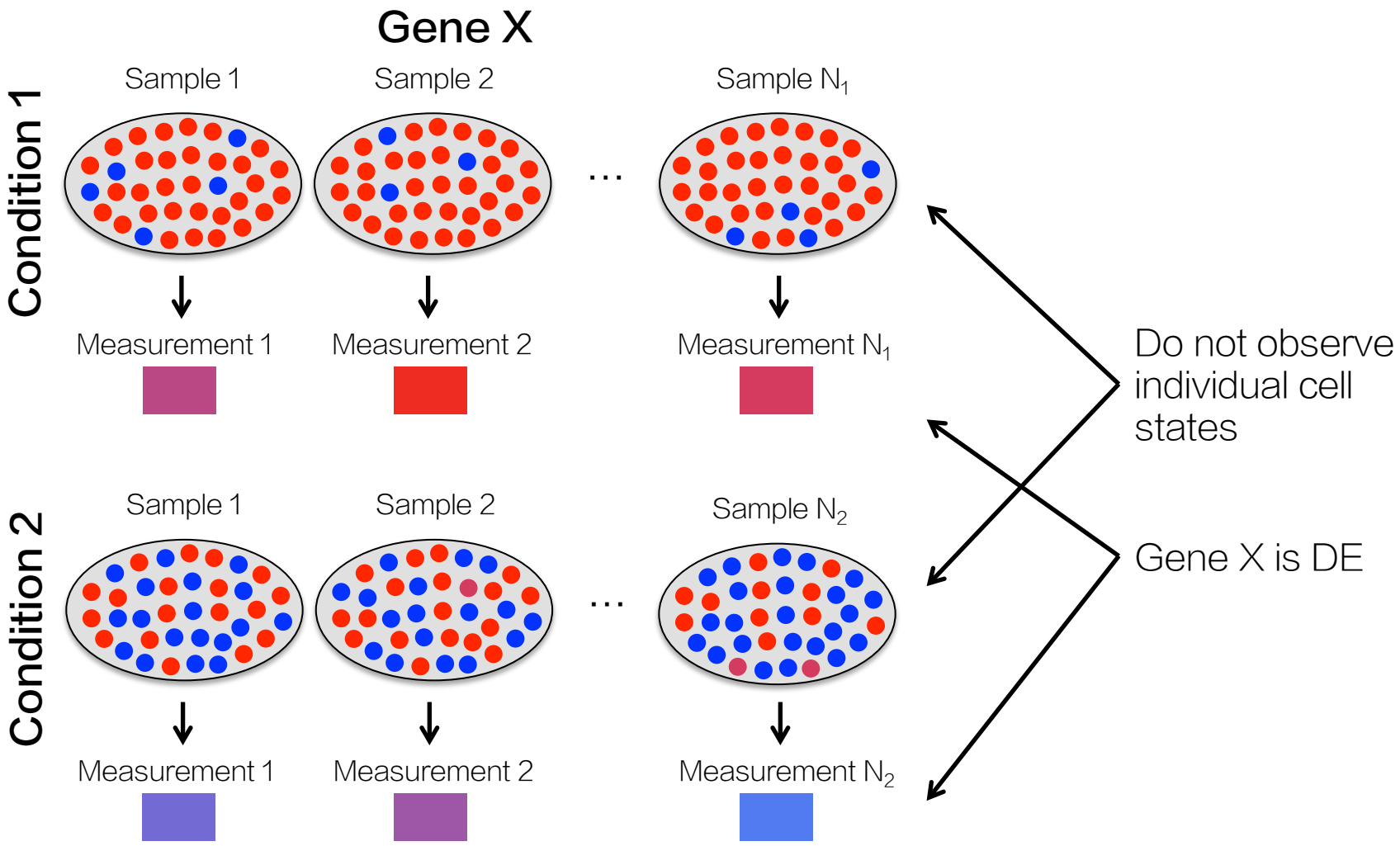
Differential Expression Analysis in bulk is blind to cellular heterogeneity



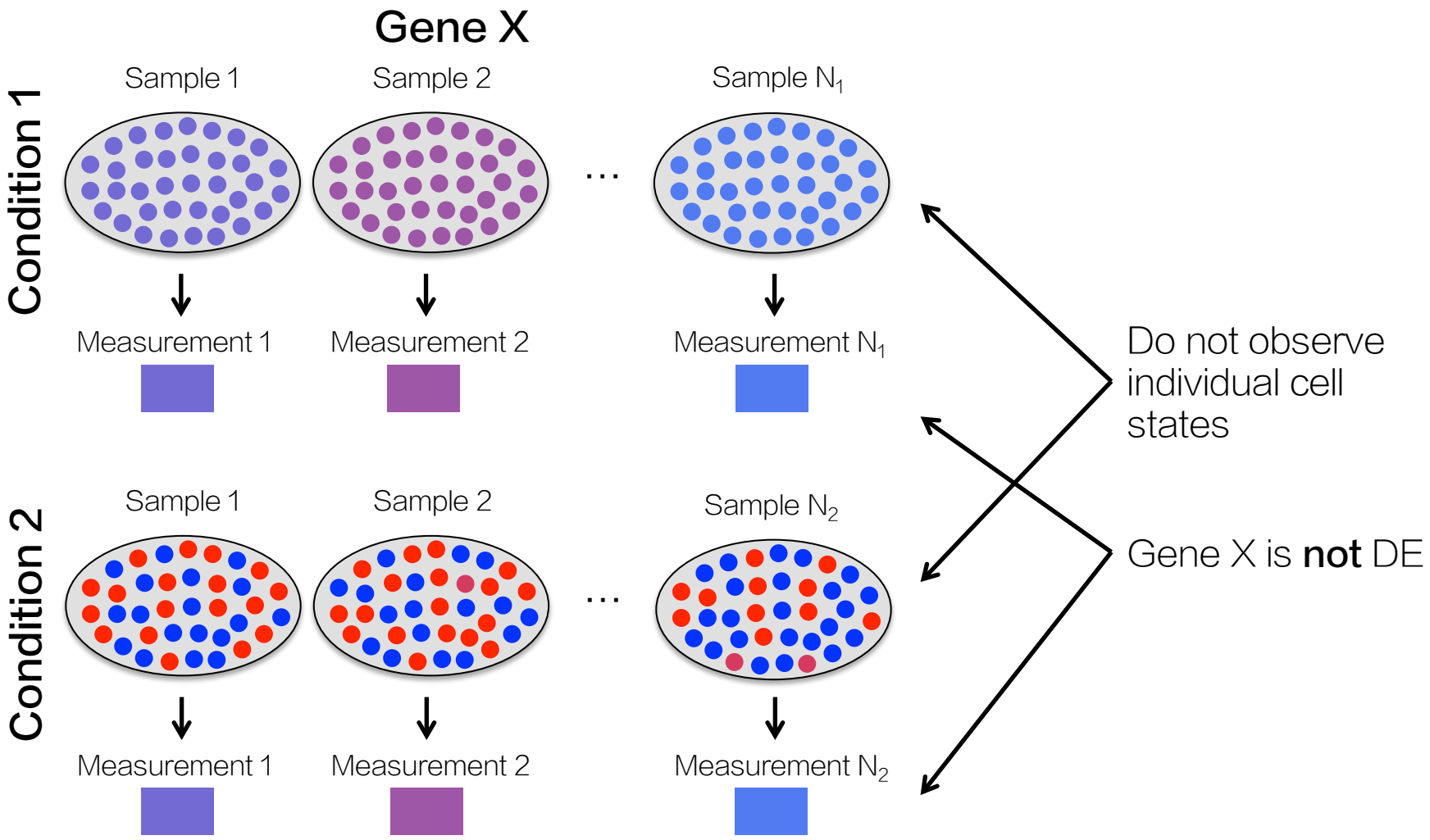
Differential Expression Analysis in bulk is blind to cellular heterogeneity



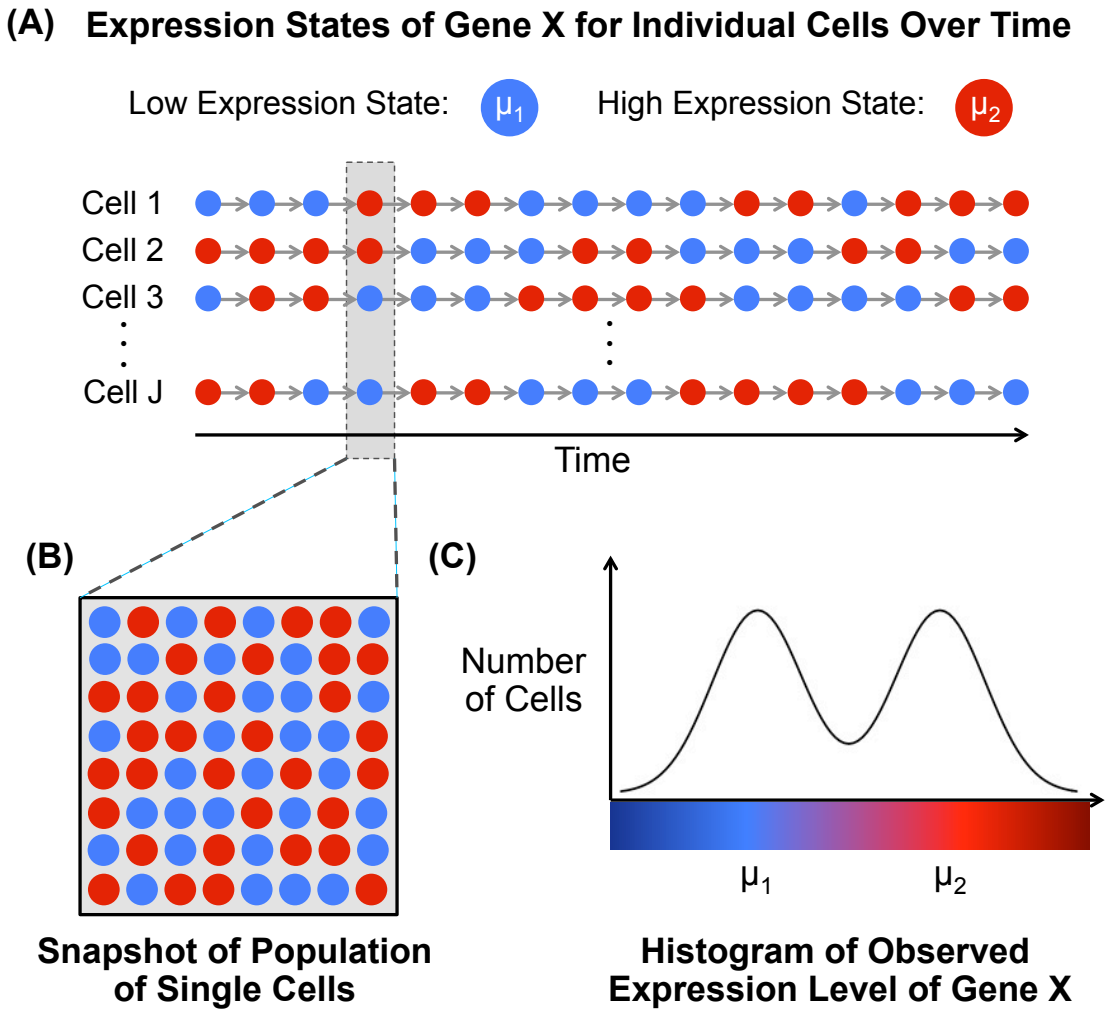
Differential Expression Analysis in bulk is blind to cellular heterogeneity



Differential Expression Analysis in bulk is blind to cellular heterogeneity

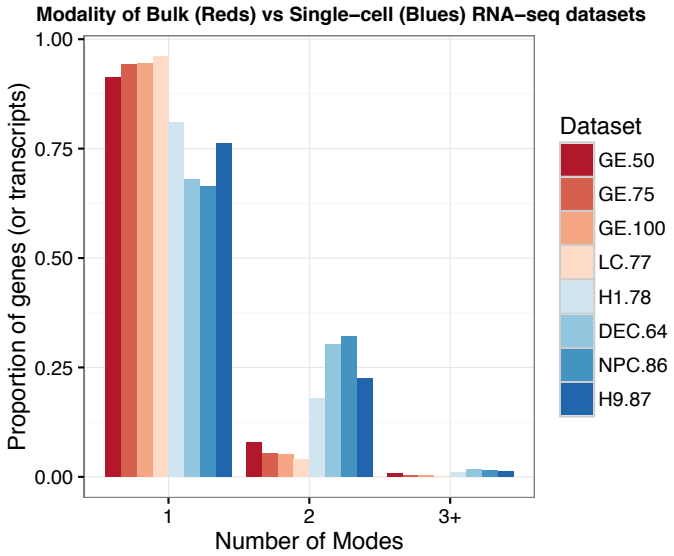


Cellular heterogeneity can lead to multi-modal expression distributions



Possible mechanisms

- Multiple stable underlying cell states
- Stochastic 'burst' fluctuations
- Oscillatory patterns



Need to reassess evaluation of DE methods in single-cell

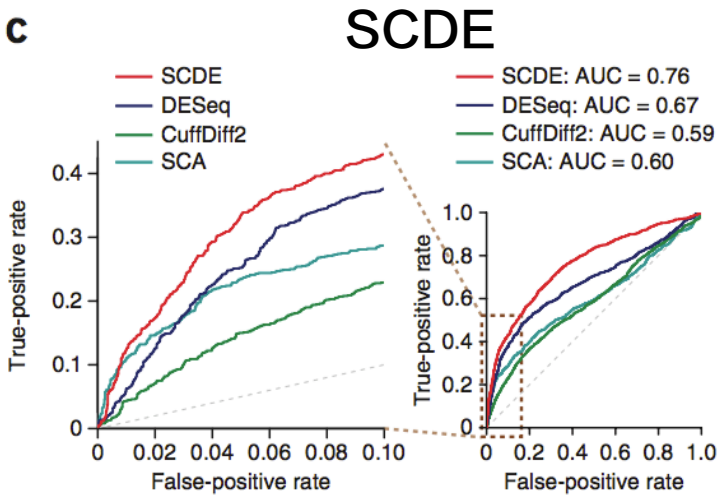


Fig 2C, Kharchenko et al. 2014, Nature Methods

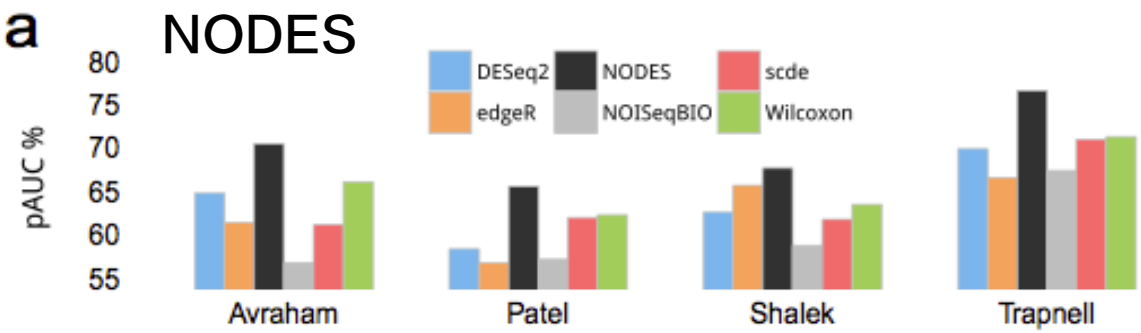


Fig 2A, Sengupta et al. 2016, BioRxiv

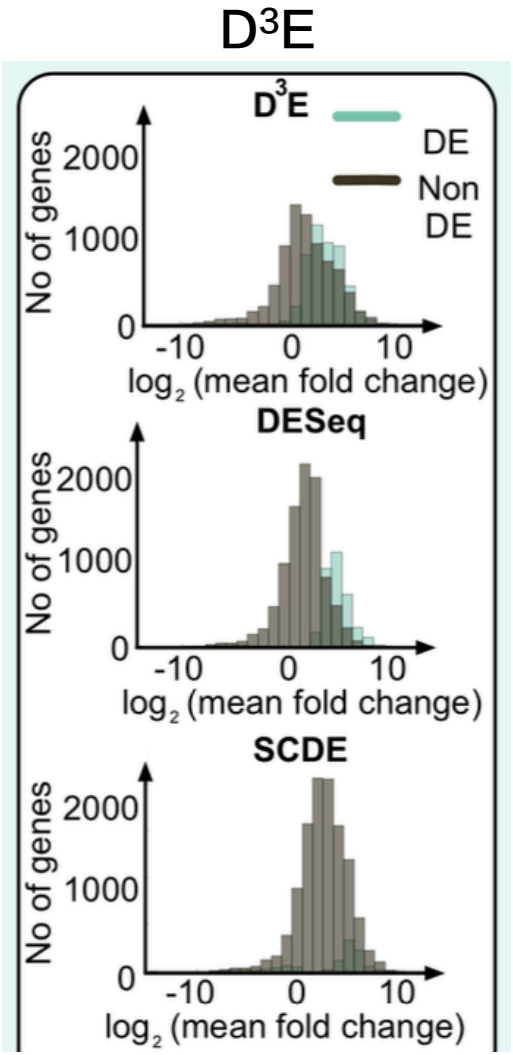


Fig 5A, Delmans et al. 2015, BioRxiv

scDD Framework

Preprocessing

1. Obtain log transformed counts normalized for library size
2. Filter genes that are detected in fewer than 25% of cells

scDD Framework

Preprocessing

1. Obtain log transformed counts normalized for library size
2. Filter genes that are detected in fewer than 25% of cells

Detection

1. Model expressed cells for each gene: DPM of Normals
2. Quantify evidence of Differential Distributions (DD):
   - BF with permutation for expressed component
   - GLM LRT for dropout component

scDD Framework

Preprocessing

1. Obtain log transformed counts normalized for library size
2. Filter genes that are detected in fewer than 25% of cells

Detection

1. Model expressed cells for each gene: DPM of Normals
2. Quantify evidence of Differential Distributions (DD):
   - BF with permutation for expressed component
   - GLM LRT for dropout component

Conditional on partition, likelihood is a product over component-specific distributions:

$$y_j \,|z_j = k, \mu_k, \tau_k \sim N(\mu_k, \tau_k)$$

$$\mu_k, \tau_k \sim NG(m_0, s_0, a_0/2, 2/b_0)$$

$$z \sim \frac{\alpha^K \Gamma(\alpha)}{\Gamma(\alpha + J)} \prod_{k=1}^{K} \Gamma(n^{(k)})$$

Partition estimate by BIC with additional merge/split step based on Bimodal Index:

$$BI = 2 * \sqrt{\frac{n_1 n_2}{(n_1 + n_2)^2}} \left(\frac{|\mu_1 - \mu_2|}{\sigma} \right)$$

scDD Framework

Preprocessing

1. Obtain log transformed counts normalized for library size
2. Filter genes that are detected in fewer than 25% of cells

Detection

1. Model expressed cells for each gene: DPM of Normals
2. Quantify evidence of Differential Distributions (DD):
   - BF with permutation for expressed component
   - GLM LRT for dropout component

Approximate Bayes Factor score for DD of expressed cells between conditions:

$$Score_g = log\left(\frac{f(Y_g, Z_g|\mathcal{M}_{DD})}{f(Y_g, Z_g|\mathcal{M}_{ED})}\right)$$

$$= log\left(\frac{f_{C1}(Y_g^{C1}, Z_g^{C1}) f_{C1}(Y_g^{C2}, Z_g^{C2})}{f_{C1,C2}(Y_g, Z_g)}\right)$$

Assess significance via permutation of samples to conditions to obtain gene-specific empirical p-values

scDD Framework

Preprocessing

1. Obtain log transformed counts normalized for library size
2. Filter genes that are detected in fewer than 25% of cells

Detection

1. Model expressed cells for each gene: DPM of Normals
2. Quantify evidence of Differential Distributions (DD):
   - BF with permutation for expressed component
   - GLM LRT for dropout component

If expressed component does not display significant DD, assess evidence for differential proportion of zeroes (dropout):

Logistic regression adjusted for overall cellular rate of dropout

scDD Framework

Preprocessing

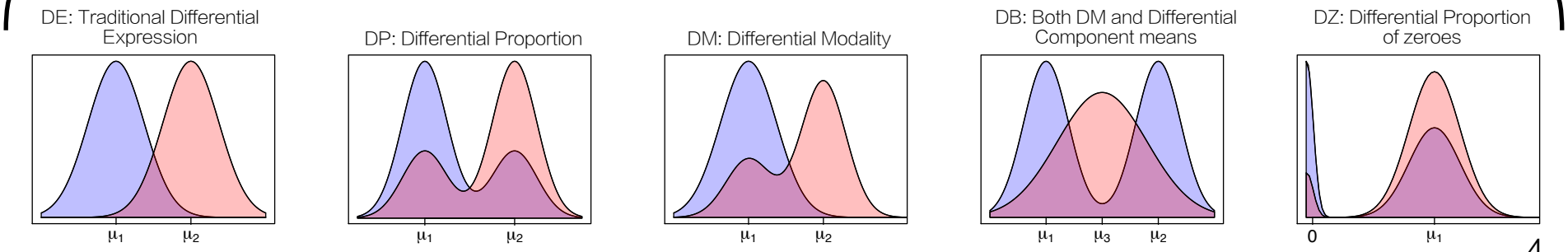
1. Obtain log Expected Counts normalized for library size
2. Filter genes that are detected in fewer than 25% of cells

Detection

1. Model expressed cells for each gene: DPM of Normals
2. Quantify evidence of Differential Distributions (DD):
   - BF with permutation for expressed component
   - GLM LRT for dropout component

Classification

Classify significant DD genes into patterns DE, DP, DM, DB, DZ



scDD Framework

Preprocessing

1. Obtain log Expected Counts normalized for library size
2. Filter genes that are detected in fewer than 25% of cells

Detection

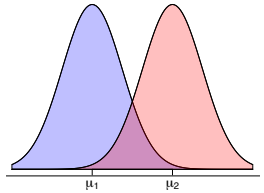
1. Model expressed cells for each gene: DPM of Normals
2. Quantify evidence of Differential Distributions (DD):
   - BF with permutation for expressed component
   - GLM LRT for dropout component

Classification

Classify significant DD genes into patterns DE, DP, DM, DB, DZ

Classification algorithm considers number of components in each condition as well as their overlap

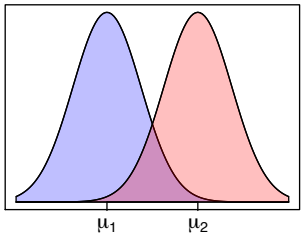
e.g. if there is one component in both conditions, and they do not overlap => DE



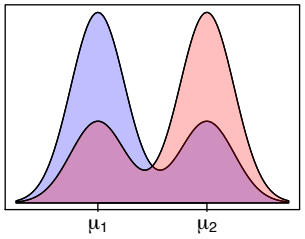
Overlap is assessed via posterior sampling of component-specific parameters:

$$(\mu_k, \tau_k)|Y, Z \sim NG(m_k, s_k, a_k/2, 2/b_k)$$

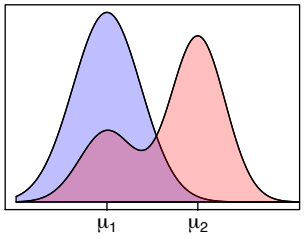
DE: Traditional Differential Expression



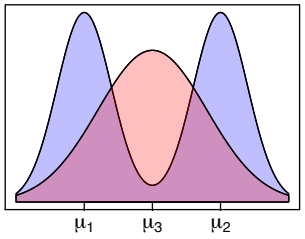
DP: Differential Proportion



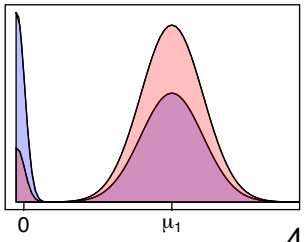
DM: Differential Modality



DB: Both DM and Differential Component means

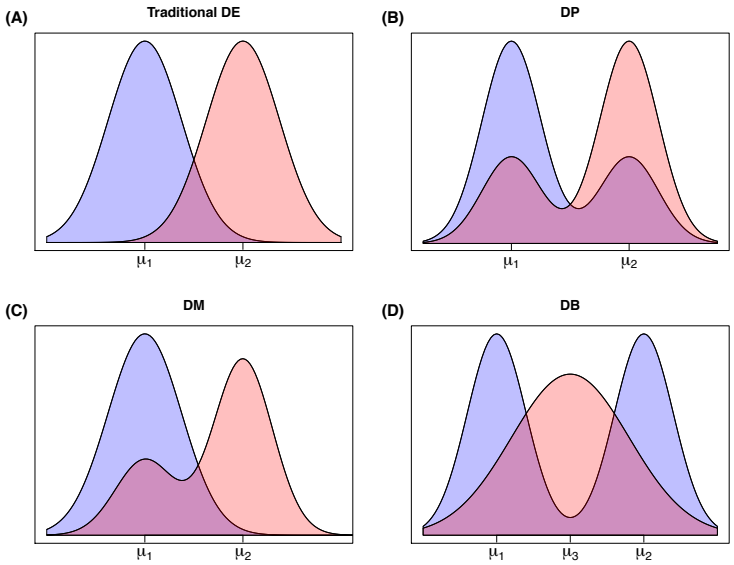


DZ: Differential Proportion of zeroes



Simulation

scDD detects and classifies complex patterns



Sample Size	Method	True Gene Category DE	DP	DM	DB	Overall (FDR)
50	scDD	0.893	**0.418**	**0.898**	**0.572**	**0.695** (0.029)
	SCDE	0.872	0.026	0.817	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**	**0.633**	0.943	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**	0.983	**1.000**	**0.905**	**0.972** (0.035)
	SCDE	**1.000**	0.855	0.998	0.787	0.910 (0.004)
	MAST	**1.000**	**0.993**	**1.000**	0.170	0.791 (0.022)

- 500 DD genes from each category, 8000 null genes
- Observations generated from mixtures of negative binomial distributions

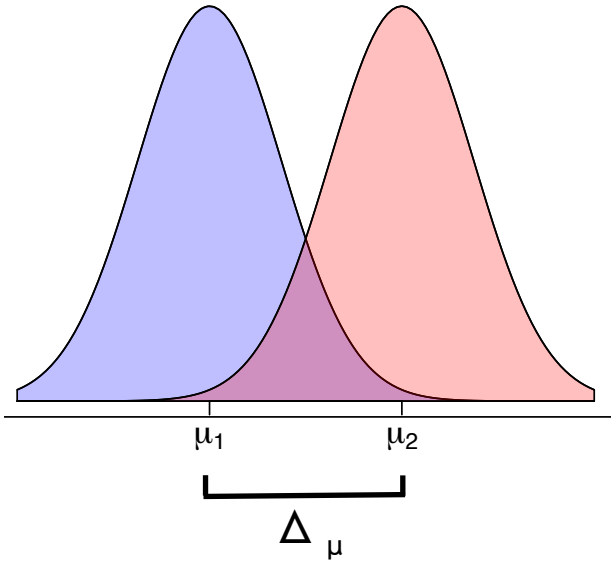
Simulation

scDD detects and classifies complex patterns

Correct Classification Rate

	Gene Category			
Sample Size	DE	DP	DM	DB
50	0.719	0.801	0.557	0.665
75	0.760	0.732	0.576	0.698
100	0.782	0.678	0.599	0.706
500	0.816	0.550	0.583	0.646

Ability to correctly classify DD genes depends on the ability to detect the correct number of components



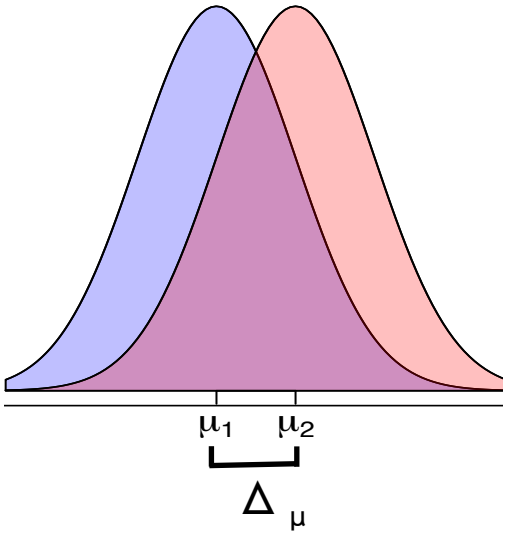
Simulation

scDD detects and classifies complex patterns

Correct Classification Rate

Sample Size	Gene Category DE	DP	DM	DB
50	0.719	0.801	0.557	0.665
75	0.760	0.732	0.576	0.698
100	0.782	0.678	0.599	0.706
500	0.816	0.550	0.583	0.646

Ability to correctly classify DD genes depends on the ability to detect the correct number of components

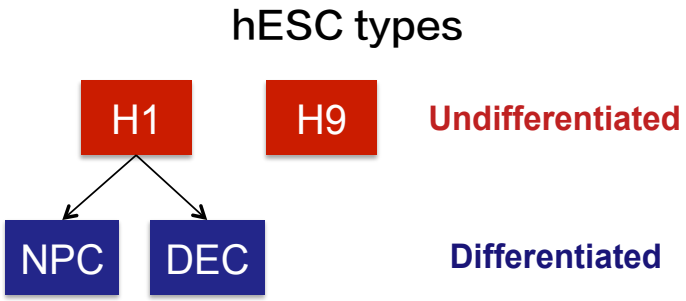


Power to detect correct number of components

Sample Size	Bimodal component mean distance Δ_μ 2	3	4	5	6	Unimodal
50	0.056	0.196	0.579	0.848	0.922	0.907
75	0.052	0.252	0.719	0.917	0.957	0.908
100	0.050	0.302	0.811	0.950	0.974	0.905
500	0.073	0.417	0.959	0.995	0.991	0.884

Case Study

scDD detects and classifies complex patterns

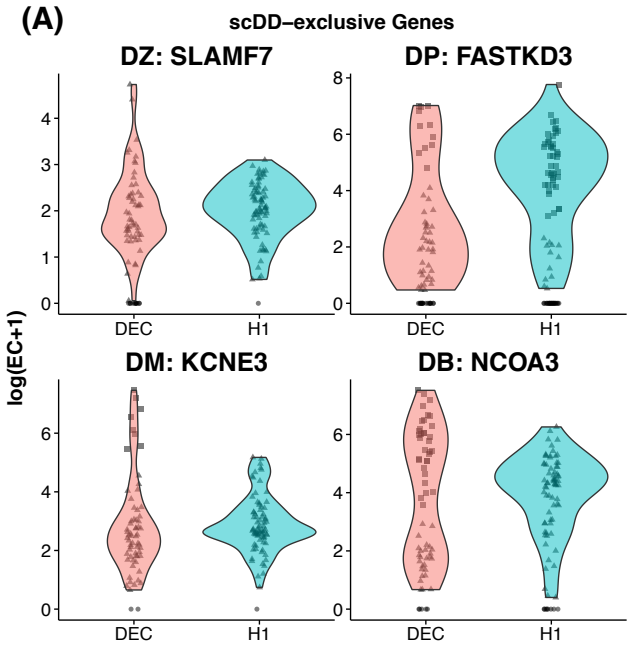


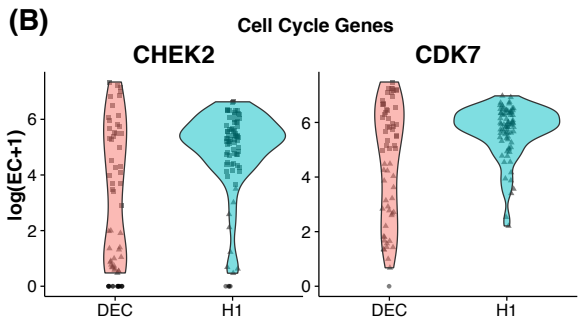
Differentially expressed genes detected by each method

Comparison	DE	DP	DM	DB	DZ	Total	SCDE	MAST
H1 vs NPC	1686	270	902	440	1603	5555	2921	5887
H1 vs DEC	913	254	890	516	911	5295	1616	3724
NPC vs DEC	1242	327	910	389	2021	5982	2147	5624
H1 vs H9	260	55	85	37	145	739	111	1119

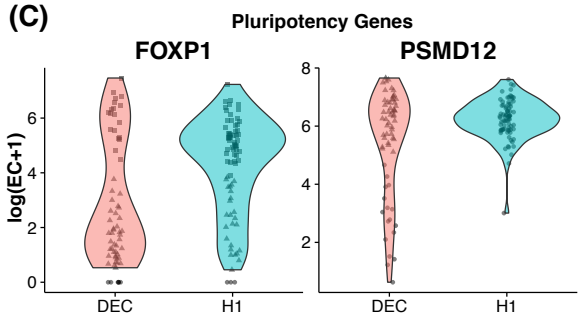
471 DD genes not detected by SCDE or MAST are enriched for complex patterns (1 gene categorized as DE)

H1 vs DEC





Cyclin genes expressed constitutively in hESCs, oscillatory in differentiated cell types



PSMD12 encodes a subunit of the proteasome complex vital to maintenance of pluripotency and has shown decreased expression in differentiating hESCs

Summary: Advantages & Limitations

- scDD is a novel statistical method that detects gene expression differences in scRNA-seq experiments while **explicitly accounting for potential multimodality** among expressed cells

- Comparable performance to existing methods at detecting mean shifts, but able to **detect and characterize more complex differences** that are masked under unimodal assumptions

- Modeling framework does not directly incorporate covariates and is limited to pairwise comparisons of biological conditions

- Genes are evaluated independently; does not aim to cluster cells into subtypes based on **global gene expression changes**

Learn More

Preprint available on BioRxiv

http://biorxiv.org/content/early/2016/05/13/035501

R package scDD available on GitHub

https://github.com/kdkorthauer/scDD

Contact

keegan@jimmy.harvard.edu

@keegsdur

Acknowledgements

UW Madison Biostatistics

Christina Kendziorski

Yuan Li

Rhonda Bacher

Morgridge Institute

Li-Fang Chu

Ron Stewart

James Thomson

UW Madison Statistics

Michael Newton

DFCI/HSPH

Rafael Irizarry Lab