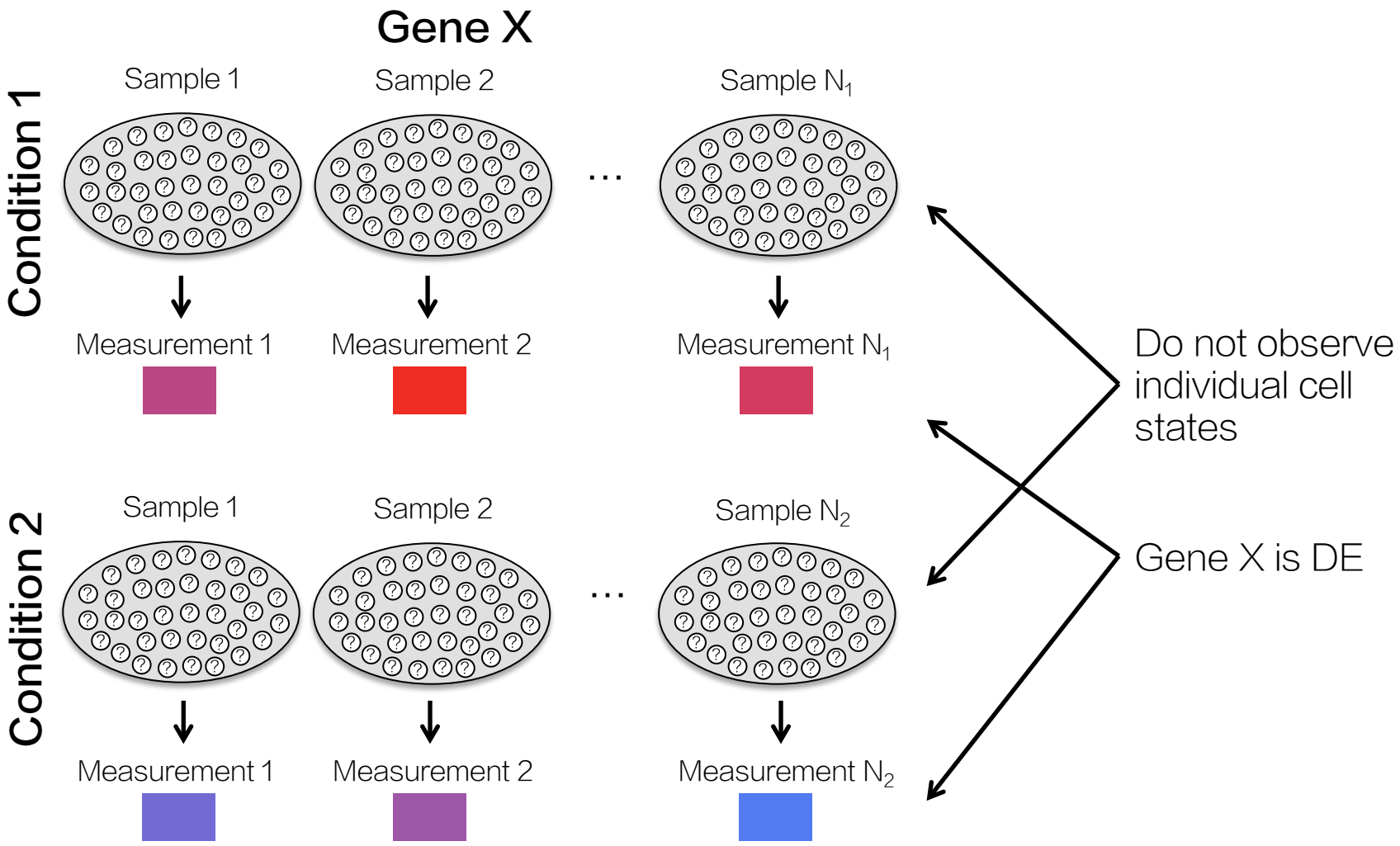
Exploiting heterogeneity in single-cell transcriptomic analyses: how to move beyond comparisons of averages

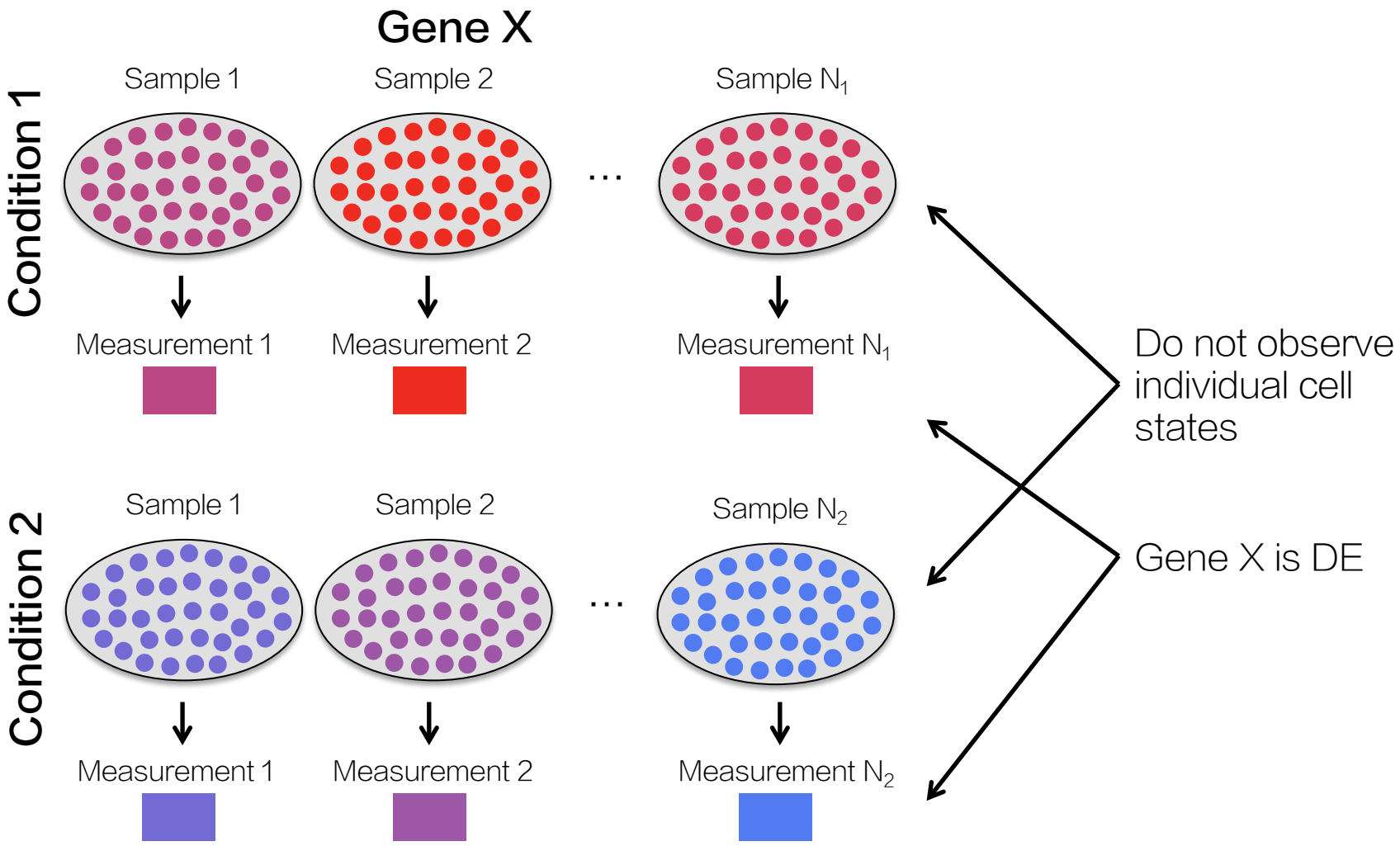
Keegan D. Korthauer, PhD
Postdoctoral Research Fellow
Dana-Farber Cancer Institute
Harvard T. H. Chan School of Public Health
@keegsdur

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

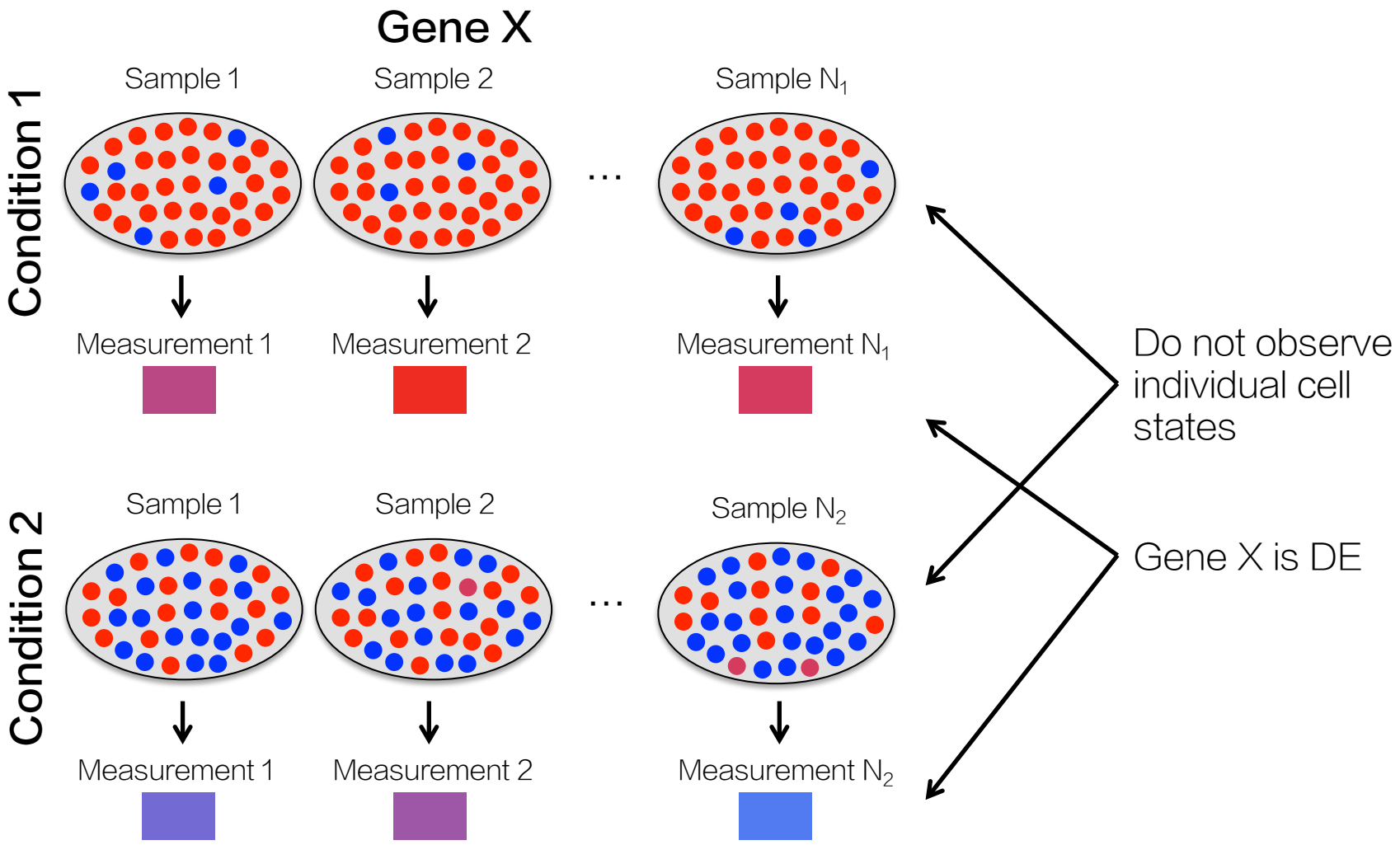
Differential Expression Analysis in bulk is blind to cellular heterogeneity



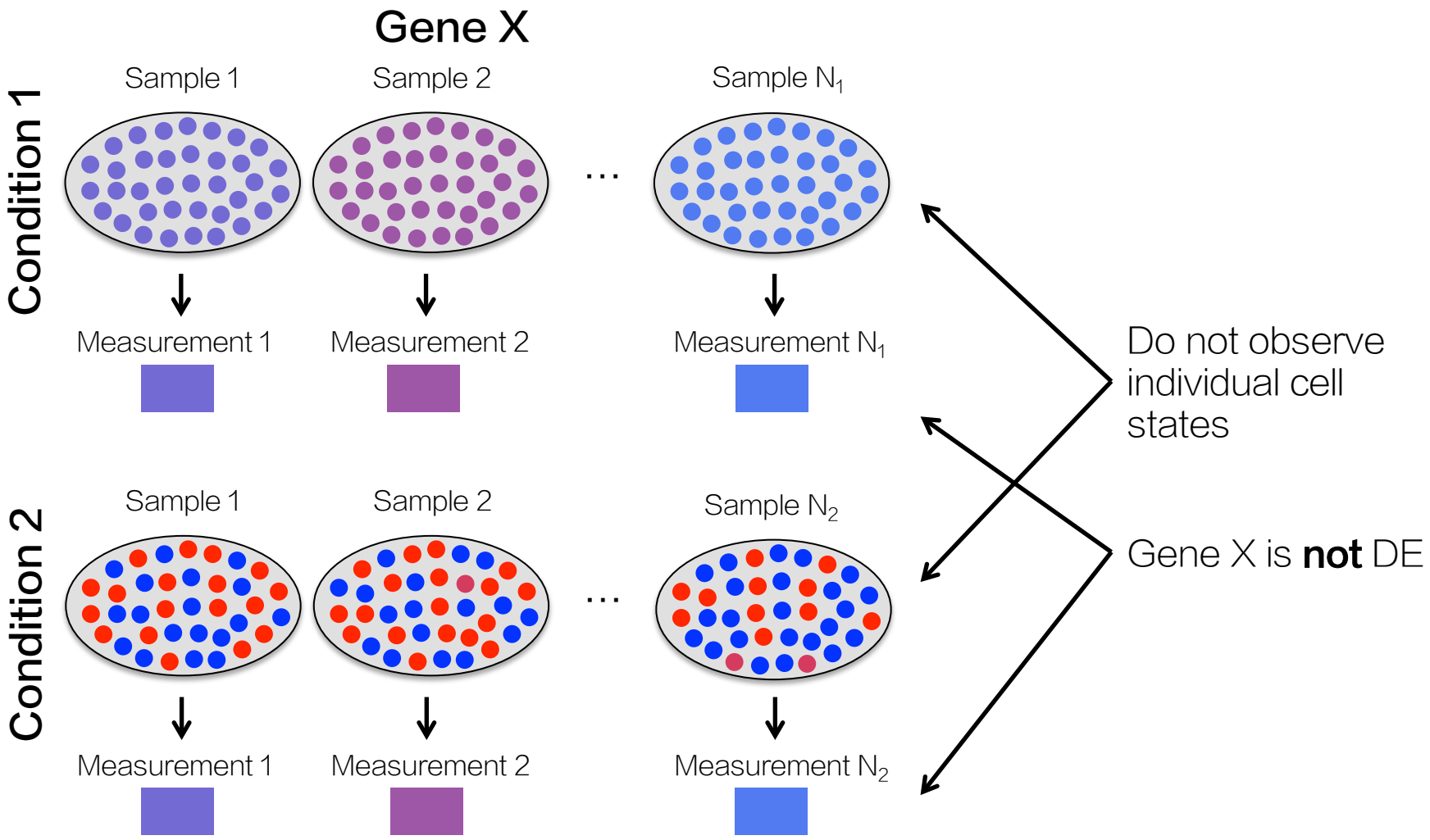
Differential Expression Analysis in bulk is blind to cellular heterogeneity



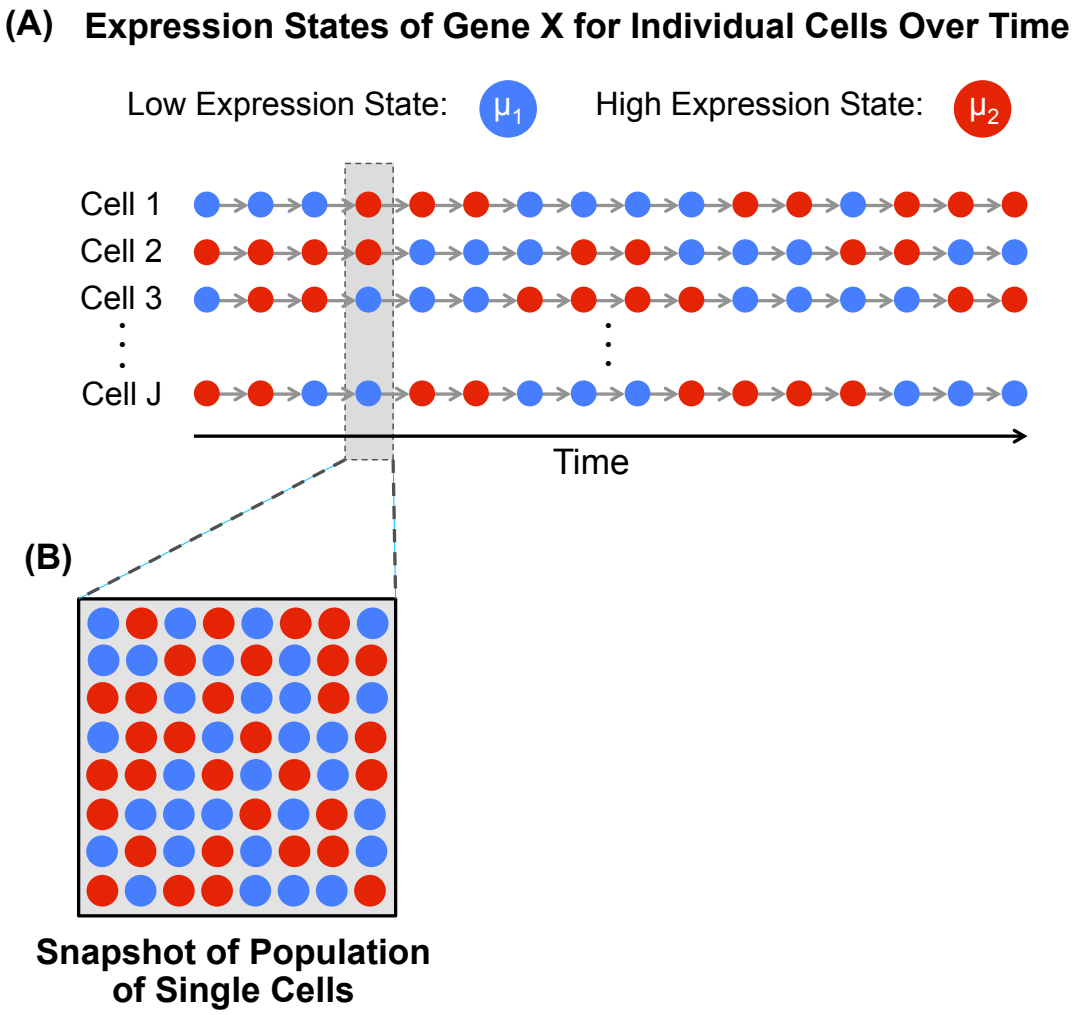
Differential Expression Analysis in bulk is blind to cellular heterogeneity



Differential Expression Analysis in bulk is blind to cellular heterogeneity



Cellular heterogeneity can lead to multi-modal expression distributions



Mechanisms leading to multi-modality

Stochastic burst fluctuations

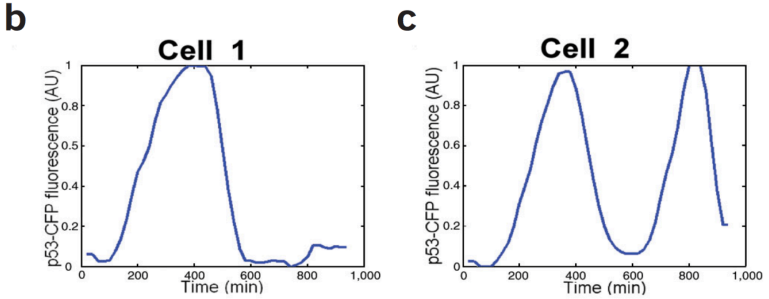


Fig 2, Lahav et al. 2004, Nature Genetics

Unsynchronized Oscillations

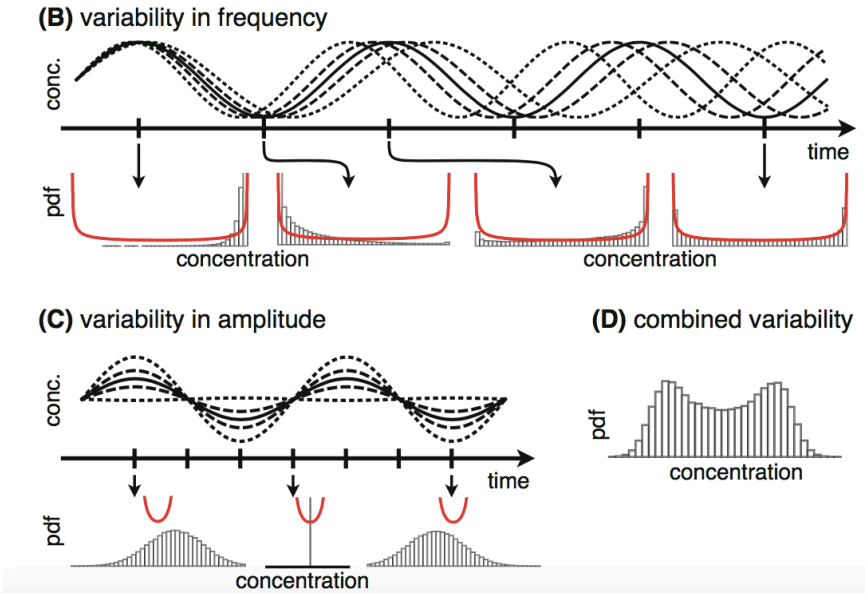


Fig 2, Dobrzynski et al. 2012, CSMB

Bistable Feedback loops

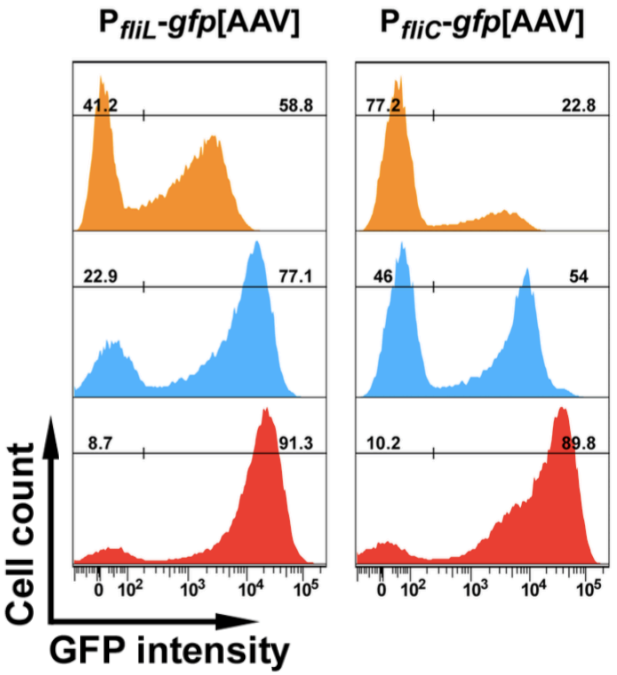
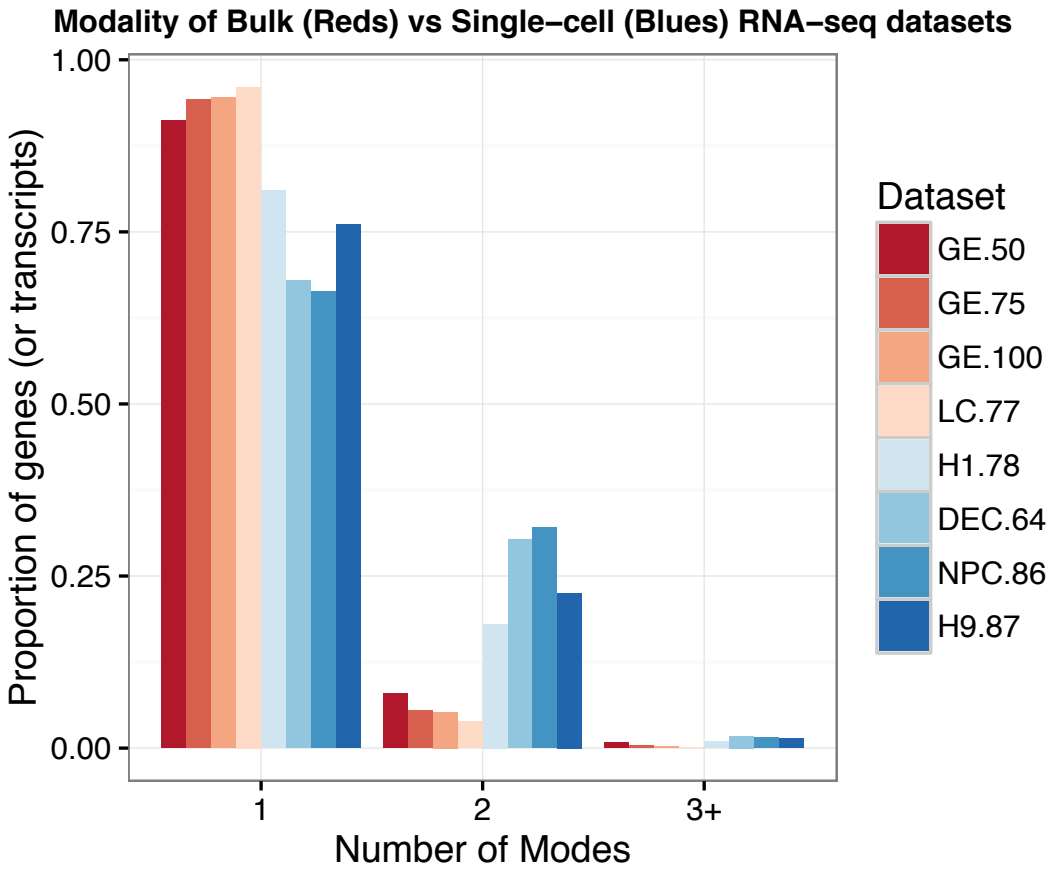


Fig 3, Jubelin et al. 2013, PLOS Genetics

scRNA−seq exhibits substantial multi−modality



Need to reassess the aim of single-cell DE analysis

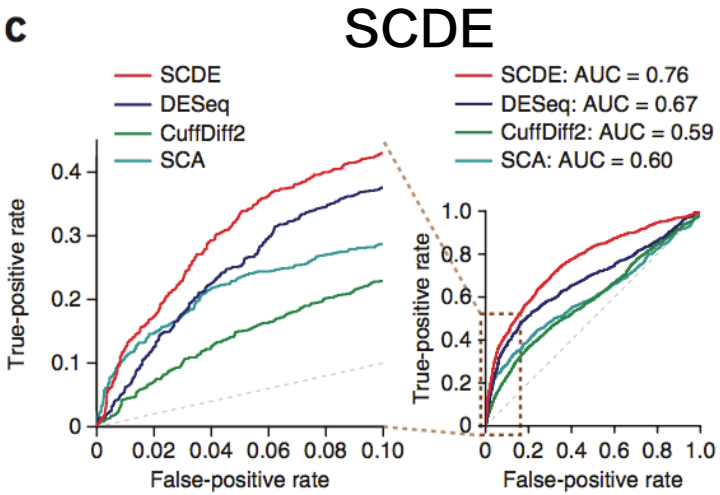


Fig 2C, Kharchenko et al. 2014, Nature Methods

Want to move beyond recapitulating what we can find in a bulk experiment

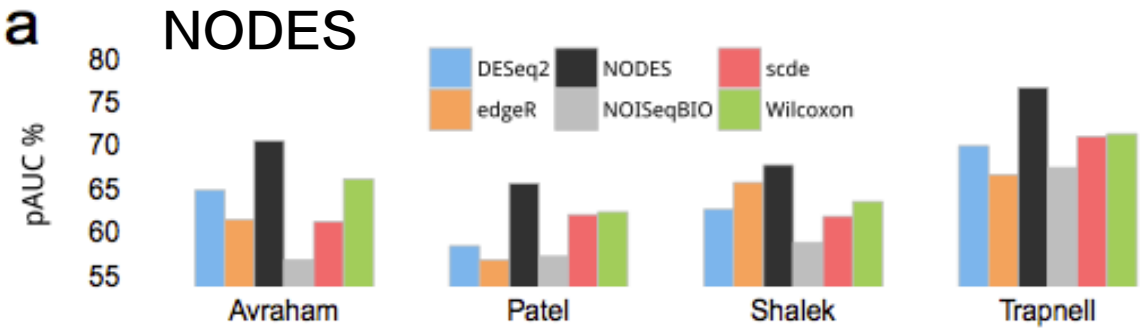


Fig 2A, Sengupta et al. 2016, BioRxiv

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: Normal DPM
2. Quantify evidence of Differential Distributions (DD)

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: Normal DPM
2. Quantify evidence of Differential Distributions (DD)

Dirichlet process mixture of normal distributions

- **Flexible** to account for multiple modes
- Incorporates **uncertainty** over the number of modes
- Number of modes **inferred** from the data

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: Normal DPM
2. Quantify evidence of Differential Distributions (DD)

Compare two competing models:

1. Global model for all cells in both populations
2. Independent models for each biological condition

scDD Framework

Preprocessing

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

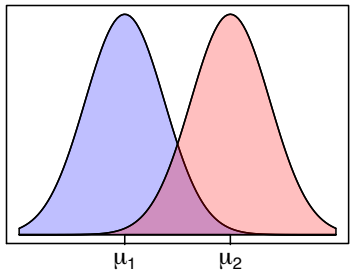
Detection

1. Model expressed cells for each gene: Normal DPM
2. Quantify evidence of Differential Distributions (DD)

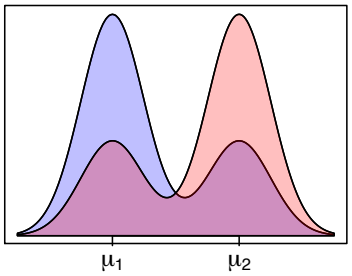
Classification

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

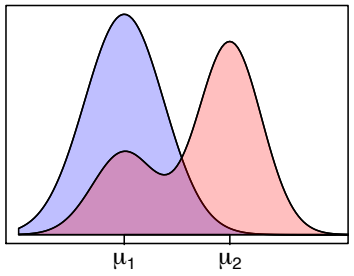
DE: Traditional Differential Expression



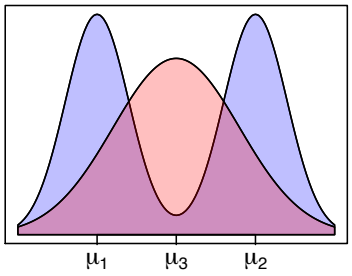
DP: Differential Proportion



DM: Differential Modality

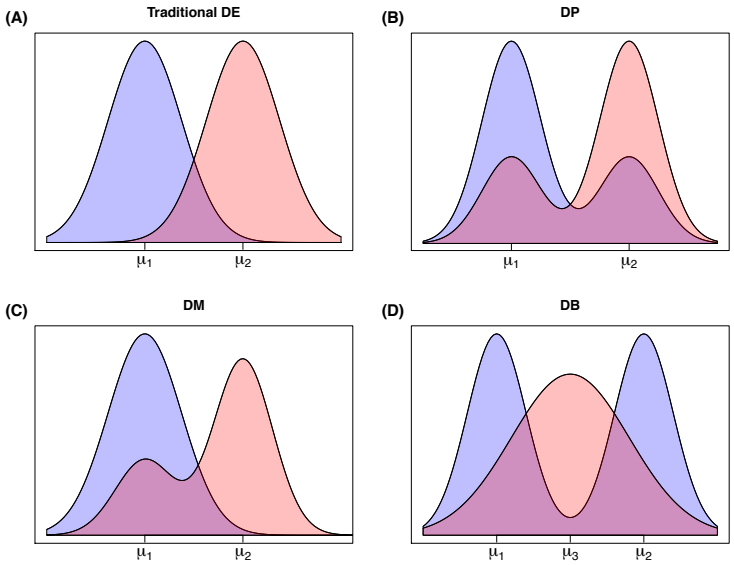


DB: Both DM and Differential Component means



Simulation

scDD detects and classifies complex patterns

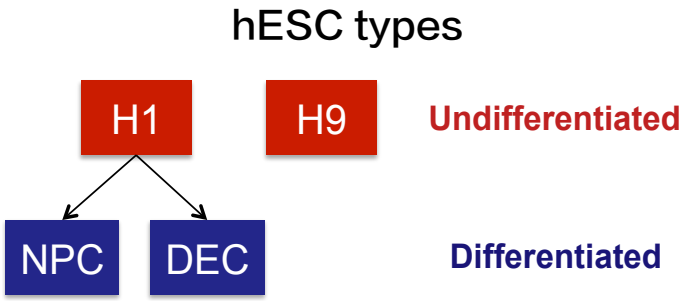


- 500 DD genes from each category, 8000 null genes
- Observations generated from mixtures of negative binomial distributions

		True Gene Category				
Sample Size	Method	DE	DP	DM	DB	Overall (FDR)
50	scDD	0.893	**0.418**	**0.898**	**0.572**	**0.695** (0.029)
	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)

Case Study

scDD detects and classifies complex patterns

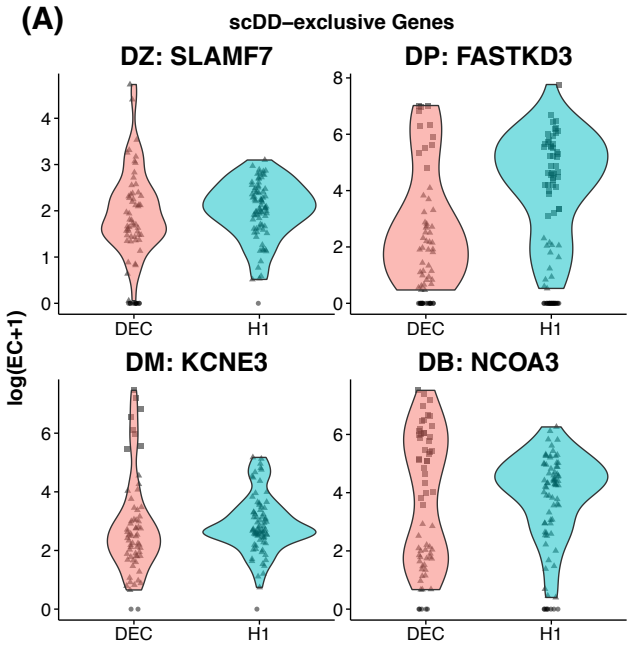


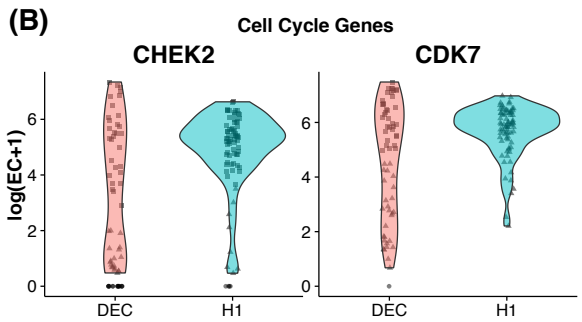
Differentially expressed genes detected by each method

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

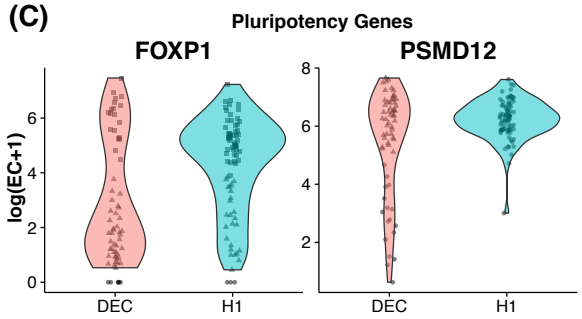
471 DD genes not detected by SCDE or MAST are enriched for complex patterns (1 gene categorized as DE)

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

Take-aways

- Bulk RNA-seq is **blind to cellular heterogeneity**, so differential expression analysis is only aimed at detecting changes in **average** expression level

- Single-cell data exhibits substantial multimodality; possible mechanisms include **stochasticity, bistability, and oscillations**

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

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

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

Morgridge Institute

Li-Fang Chu

Ron Stewart

James Thomson

UW Madison Statistics

Michael Newton

DFCI/HSPH

Rafael Irizarry Lab