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

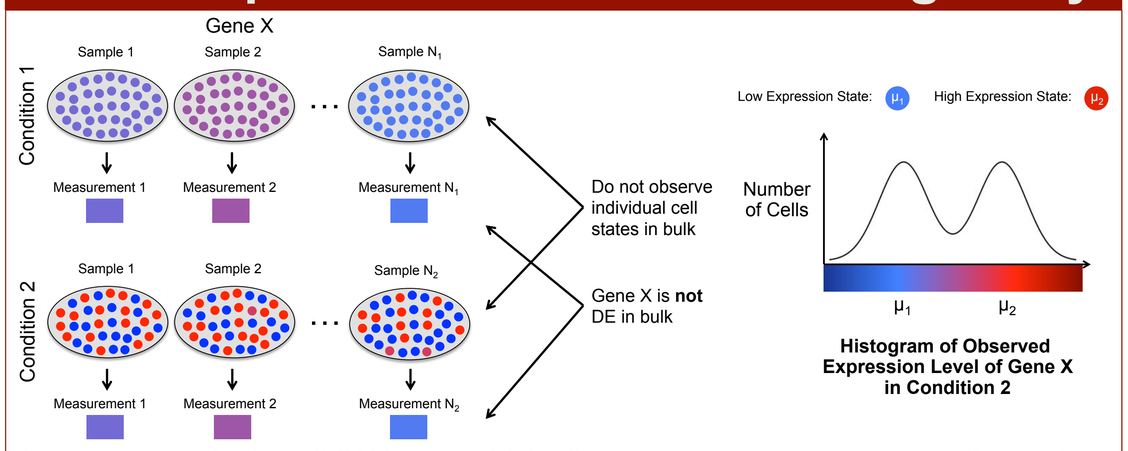
Keegan Korthauer[1,2*], Li-Fang Chu[4], Michael A. Newton[3], Yuan Li[3], James Thomson[4], Ron M. Stewart[4], and Christina Kendziorski[3]

[1]Harvard T.H. Chan School of Public Health, [2]Dana-Farber Cancer Institute, [3]University of Wisconsin-Madison, [4]Morgridge Institute for Research

Abstract

The ability to quantify cellular heterogeneity is a major advantage of single-cell technologies. It is now possible to elucidate gene expression dynamics that were invisible using bulk RNA-seq, such as the presence of distinct expression states. However, statistical methods often treat cellular heterogeneity as a nuisance. We have developed a novel method to characterize differences in expression in the presence of distinct expression states within and among biological conditions. This framework can detect differential expression patterns under a wide range of settings. Compared to alternative approaches, this method has higher power to detect subtle differences in gene expression distributions that are more complex than a mean shift, and can characterize those differences. The R package scDD implements the approach, and is available on Bioconductor [2].

Differential Expression Analysis in Bulk RNA-seq is blind to cellular heterogeneity



In contrast to single-cell RNA-seq, which allows us to get a measurement for each cell, differential expression (DE) analysis in traditional (or bulk) RNA-seq is blind to any cellular heterogeneity. The illustration above shows an example where the bulk RNA-seq experiment would not detect differential expression, but there is clearly a different pattern of expression between the two populations. This type of pattern may be of great biological significance, so it is important that DE methods for scRNA-seq account for it. However, doing so is complicated by the fact that these types of patterns result in multi-modal expression distributions, which are generally not accommodated for in existing approaches.

Biological mechanisms leading to multi-modality

Stochastic burst fluctuations

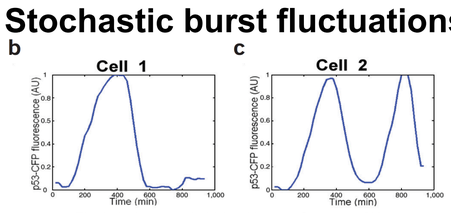


Fig 2, Lahav et al. 2004, Nature Genetics [3]

Unsynchronized Oscillations

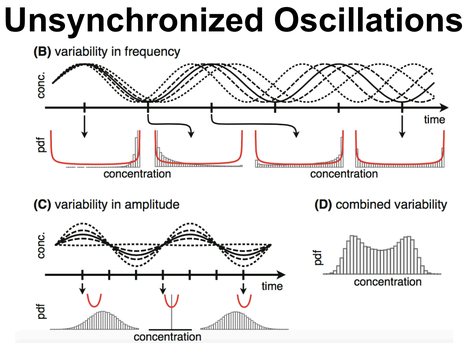


Fig 2, Dobrzynski et al. 2012, CSMB [4]

Bistable Feedback loops

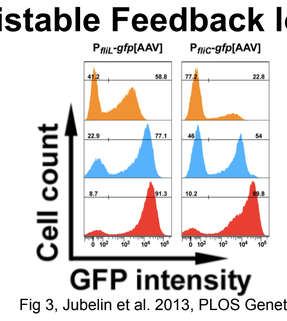


Fig 3, Jubelin et al. 2013, PLOS Genetics [5]

Modality in scRNA-seq

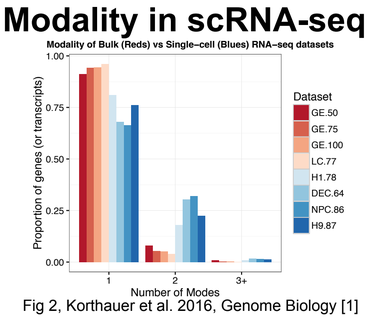


Fig 2, Korthauer et al. 2016, Genome Biology [1]

Biological mechanisms such as stochastic burst-like fluctuations [3], unsynchronized oscillations [4], and bistable feedback loops [5] (illustrated above) can give rise to a mixed population of cells at multiple different expression states, which manifests as multi-modal distributions. This multimodality complicates DE analysis methods for single-cell, since most assume a parametric distribution with one mode representing the expressed cells (such as SCDE [6] and MAST [7]).

scDD Algorithm

Evaluate evidence of DD of expressed cells using an approximate Bayes Factor score comparing:

- Global model for all cells
- Independent models for each biological condition

Assess significance via permutation test or (for large datasets) the Kolmogorov-Smirnov test.

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

Preprocessing

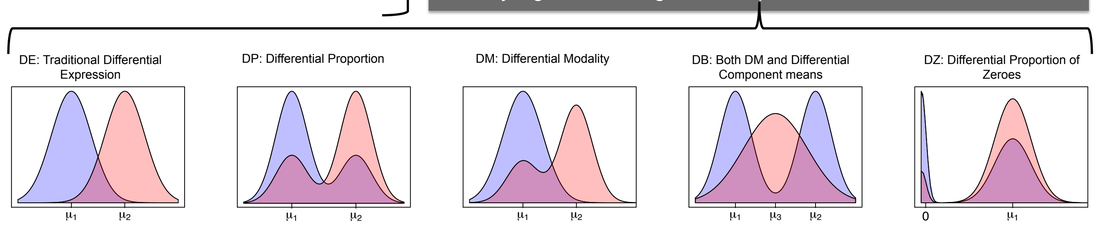
1. Obtain log Expected Counts normalized for library size
2. Filter genes that are detected in fewer than 25% of cells

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



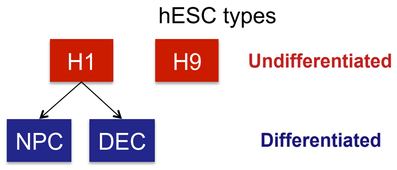
The scDD [1] algorithm (summarized above) tests whether the distribution (possibly multi-modal) of expression is different between biological conditions and classifies genes into categories that summarize the salient characteristics of the differences.

scDD detects and classifies complex patterns

When simulating gene expression from mixtures of negative binomial distributions that represent the patterns depicted above, scDD is comparable or slightly better at detecting the DD genes that have an overall mean shift (as shown in the table to the right). As expected, however, it is superior at detecting the DB category, which has no overall mean shift.

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)

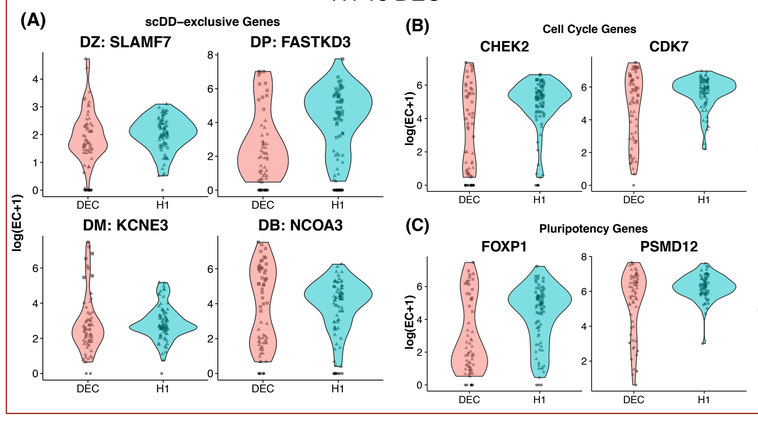
In an analysis of human embryonic stem cell (hESC) types (detailed below), we evaluated pairwise comparisons of four cell lines. Identifying which genes are expressed differently between these conditions can give insight into the differentiation process. scDD generally detects more differential genes than other methods, but the additional are enriched for complex patterns. As expected, cell cycle and pluripotency genes are among those detected only by scDD.



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

H1 vs DEC



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

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

scDD is a novel statistical framework and R package that detects gene expression differences in scRNA-seq experiments while **explicitly accounting for potential multimodality** among expressed cells. It has comparable performance to alternative methods at detecting mean shifts, but is able to **detect and characterize more complex differences** that are masked under unimodal assumptions.

Contact

Keegan Korthauer
Harvard T.H. Chan School of Public Health
Dana-Farber Cancer Institute
keegan@jimmy.harvard.edu
@keegankorthauer

Website:

References

[1] Korthauer, K. D., Chu, L. F., Newton, M. A., Li, Y., Thomson, J., Stewart, R., & Kendziorski, C. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biology, 17(1), 222.
[2] Korthauer, K.D. (2017) scDD: Mixture modeling of single-cell RNA-seq data to identify genes with differential distributions. Bioconductor R Package version 1.0.0 (BioC Release 3.5), https://bioconductor.org/Packages/scDD.
[3] Lahav, G., et al. (2004). Dynamics of the p53-Mdm2 feedback loop in individual cells. Nature genetics, 36(2), 147-150.
[4] Dobrzyński, M., et al. Nonlinear signalling networks and cell-to-cell variability transform external signals into broadly distributed or bimodal responses. Journal of The Royal Society Interface 11.98 (2014): 20140383.
[5] Jubelin, G., et al. FliZ is a global regulatory protein affecting the expression of flagellar and virulence genes in individual Xenorhabdus nematophila bacterial cells. PloS Genet 9.10 (2013): e1003915.
[6] Kharchenko, P. V., et al. Bayesian approach to single-cell differential expression analysis. Nature methods 11.7 (2014): 740-742.
[7] Finak, Greg, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome biology 16.1 (2015): 278.