Bayesian nonparametric mixture modeling of expression dynamics in single-cell RNA-seq experiments

Keegan Korthauer[1]*, Li-Fang Chu[2], Ron M. Stewart[2], and Christina Kendziorski[1,3]

[1]Department of Statistics, UW-Madison, WI; [2]Morgridge Institute for Research, Madison, WI; [3]Department of Biostatistics and Medical Informatics, UW-Madison, WI 53706

ABSTRACT

RNA-seq experiments allow for the quantification of transcript abundance on large collections of cells. Though useful in many settings, traditional RNA-seq (commonly referred to as bulk RNA-seq) quantifies the average signal in the population of cells under study.

In contrast, RNA-seq measurements on the single-cell level will allow us to answer emerging scientific questions such as the quantification of cellular heterogeneity and the study of expression kinetics. Specifically, it is known that transcription is a stochastic process, where genes may switch between distinct cellular states or exhibit bursts of activity separated by periods of dormancy.

When profiling expression levels across individual cells, these dynamic gene expression patterns may exhibit multimodal distributions. Here we develop a Bayesian nonparametric mixture modeling approach to identify genes with these dynamic expression patterns as well as detect differences across biological conditions, which may represent differential regulation.

INTRODUCTION

One of the key features seen in single-cell RNA-seq data is the markedly increased variation from sample to sample compared to bulk RNA-seq, which results from a combination of biological and technical sources. Clearly, measuring a single cell instead of averaging a pool of cells introduces additional biological signal if there is heterogeneity in the amount of transcript in each cell. This heterogeneity is of great interest, for example, in studies of the differentiation potential of individual cells.

Beyond the anticipated heterogeneity among cells from different biological conditions (e.g. differentiation lineages), there is still more variation that can be attributed to the stochasticity of transcription. For example, even in genetically homogeneous cell populations, the total mRNA content within individual cells can vary by a factor of more than five [5].

The focus of this study is to detect expression heterogeneity due to transcriptional stochasticity. Specifically, expression at the single cell level may alternate between distinct states, or may be characterized by bursts punctuated by states of inactivity [7]. As a consequence of these dynamics, expression levels across a population of cells may exhibit multimodal distributions for many genes [2,6], as outlined in Figure 1.

Figure 1

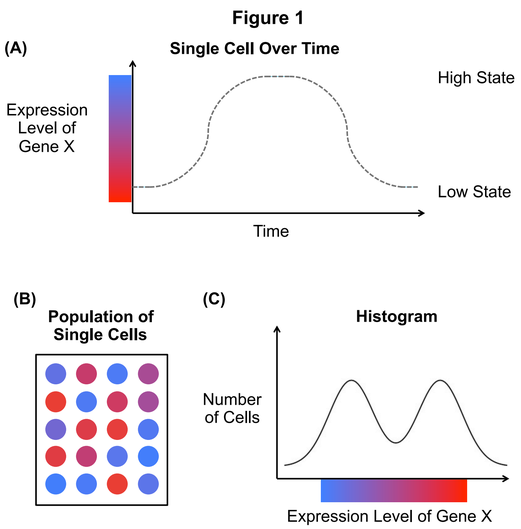

Figure 1 Schematic of single-cell expression dynamics and how they can lead to heterogeneity within cell populations. (A) Time series of the expression of gene X in a single cell, which switches back and forth between a high and low state. (B) Population of individual cells shaded by level of expression of gene X at a snapshot in time. (C) Histogram of the expression of gene X for the cell population in (B).

Our goal is to characterize the expression dynamics in scRNA-seq data, explicitly accounting for multimodal distributions which may represent underlying cellular states. As the regulation of these dynamics is largely uncharacterized for the vast majority of genes, we also aim to detect differences across conditions. We propose a Bayesian nonparametric mixture modeling approach that detects differentially regulated genes exhibiting a variety of patterns (indicated in Figure 2).

Figure 2

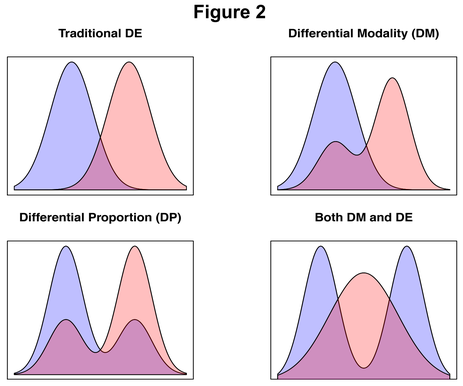

Figure 2 Diagram of plausible differential regulation patterns (histograms), including traditional differential expression (upper left), differential modality (upper right), differential proportion within each mode (lower left), and both differential modality and differential expression (lower right).

METHODS

Modeling Framework

As in [6], we model the nonzero scRNA-seq measurements within a gene as log-normally distributed random variables. Rather than constraining to one component, however, we allow for a mixture of more than one log-normal component to accommodate the patterns in Figure 2. The number of components is determined by first assuming that the log-transformed data arise from a Dirichlet Process Mixture (DPM) model:

Let $Y = (y_1,...,y_J)$ be a vector of log-transformed nonzero expression measurements for a given gene in J cells. Then a conjugate DPM of normals with fixed variance σ^2 is given by

$$y_j|\mu_j \sim N(\mu_j,\sigma^2),\quad \mu_j|G \sim G,\quad G \sim DP(\alpha,G_0),\quad G_0 = N(\mu_0,\sigma_0^2)$$

The posterior distribution of μ_j is intractable even for moderate sample sizes; however, if we let $Z = (z_1,...,z_J)$ be the vector of component memberships for all samples, where the number of unique Z values is r, the likelihood of Y conditional on Z can be viewed as a product over the cluster-specific component likelihoods. This augmented form is known as a Product Partition Model (PPM):

$$f(Y|Z) = \prod_{k=1}^{r} f(Y^{(k)}),\ \text{where } f(Y^{(k)}) \text{ is the component likelihood where the cluster-specific parameter } \mu_k \text{ has been integrated out.}$$

Under this formulation, the joint likelihood of the data Y and partition Z is:

$$f(Y,Z) = f(Y|Z)f(Z) = \frac{\alpha^r\Gamma(\alpha)}{\Gamma(\alpha+J)}\prod_{k=1}^{r}\Gamma(n^{(k)})f(Y^{(k)}),$$

where $n^{(k)}$ is the number of observations in component k

We use *modalclust* [3] to fit the PPM and obtain the optimal partition estimate Z. Since this algorithm relies on component variance estimates, these are obtained using *Mclust* [4]. A sensitivity analysis on the specification of the Dirichlet concentration parameter α is summarized in Figure 4.

Approximate Bayes Factor Score

To assess the evidence for differential regulation (DR) of gene *g* across two conditions, we use an approximate Bayes Factor score that relies on the joint likelihood of the data Y and partition estimate Z:

$$\text{Score}_g = \frac{f(Y_g,\hat{Z}_g|M_{DR})}{f(Y_g,\hat{Z}_g|M_{ER})} = \frac{f_{C1}(Y_g^{C1},\hat{Z}_g^{C1})f_{C2}(Y_g^{C1},\hat{Z}_g^{C1})}{f_{C1,C2}(Y_g,\hat{Z}_g)}$$

The numerator is obtained by fitting the PPM to each condition independently, and the denominator is obtained by fitting one joint PPM to both conditions together.

Permutation and Classification of DR patterns

Significance of the DR pattern is assessed via permutation of cells to condition. Significantly DR genes are classified into the DE, DP, DM, and DB patterns in Figure 2 by examining the number of components in each fit, as summarized below:

- **DE:** 1 mode in each condition, 2 modes overall
- **DP:** modality in each condition the same as overall (>1)
- **DM:** different modality of conditions; equivalent means*
- **DB:** different modality of conditions; different means*

*assessed via pairwise t-tests of observations in each component

SIMULATION RESULTS

Simulation Study I: The sensitivity of Dirichlet concentration parameter α and samples size on ability of the PPM to detect the correct number of components was evaluated under three simulation scenarios: one component, two components with low separation, and two components with high separation. Based on these results, α was set to 0.10.

Figure 3

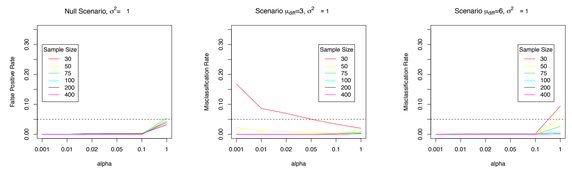

Figure 3 Summary of sensitivity analysis of α and samples size for one component (left), two components with low separation (center), and two components with high separation (right).

Simulation Study II: 2000 genes were simulated via normal mixtures for 100 cells in each of two conditions, with 1200 ER (equivalently regulated) and 200 in each of the four DR patterns. For each DR pattern, the distance between the component means was varied between 3, 4, 5, and 6 (with σ^2=1). P-values were computed for 5000 permutations. Overall power to detect DR was 0.95 (FDR=0.02); classification rates for DR categories are shown in Table 1.

	Component Distance			
DR Pattern	3	4	5	6
DE	0.96	0.98	1.00	1.00
DP	0.79	0.98	0.92	0.96
DM	0.33	0.96	1.00	0.98
DB	0.66	1.00	0.96	1.00

Table 1 Correct classification rates for simulation study (proportion of true positive DR genes assigned to the correct DR category)

CASE STUDY RESULTS

Figure 4

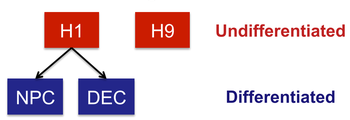

Figure 4 Relationship of cell types used in case study. H1 and H9 are undifferentiated lines; NPC and DEC are differentiated from H1.

Four cell types are compared: two human embryonic stem cell lines (H1 and H9), and two differentiated types (NPC and DEC). Up to 96 cells in each condition passed quality control procedures and expected counts were obtained from RSEM. Counts were normalized using DESeq [1] size factors and genes with more than 75% dropout (zero counts) were not considered. The number of genes found significant by 10,000 permutations in each DR category is shown in Table 2, and example histograms are displayed in Figure 5.

	Cell Line Comparison			
DR Pattern	H1 v H9	H1 v NPC	H1 v DEC	NPC v DEC
DE	821	1049	1548	1066
DP	22	348	645	431
DM	165	1251	1489	808
DB	545	1819	2047	1101

Table 2 Number of genes with significant permutation p-value.

Figure 5

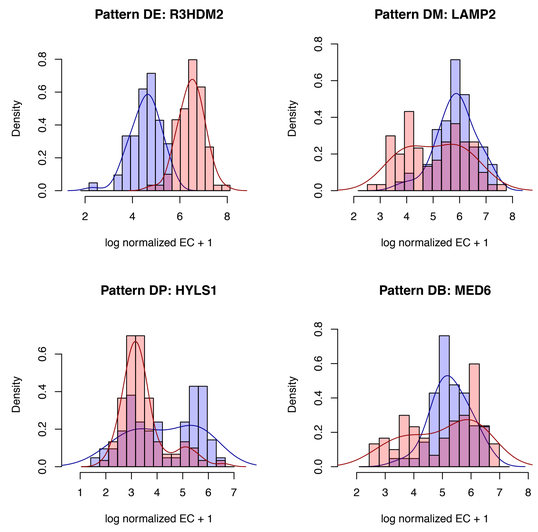

Figure 5 Histograms of genes differentially regulated between H1 (red) and NPC (blue) from each of the four DR categories depicted in Figure 2. Overlap depicted with purple shading, and smoothed kernel density estimate lines are plotted. Top Left: differentially expressed (DE), Top Right: differential modality (DM), Bottom Left: differential proportion (DP), and Bottom Right: both differential modality and differential expression (DB).

SUMMARY & FUTURE WORK

Flexible nonparametric Bayesian mixture models can account for within-condition cellular heterogeneity in scRNA-seq experiments that may arise from the stochasticity of gene expression.

Simulation studies show that the framework has favorable operating characteristics when component means are well-separated. Application to cell lines show that the model identifies significantly differentially regulated genes that exhibit striking patterns that differ between conditions.

Future work will account for differential dropout rates across conditions as a source of technical error by performing a restricted permutation test that utilizes the relationship between dropout rate and average expression level.

REFERENCES

[1] Anders, S and Huber, W. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[2] Birtwistle, MR, et al. Emergence of bimodal cell population responses from the interplay between analog single-cell signaling and protein expression noise. *BMC Systems Biology*, 6(1): 109, 2012.

[3] Dahl, DB. Modal clustering in a class of product partition models. *Bayesian Analysis*, 4(2): 243–264, 2009.

[4] Fraley, C, et al. Mclust version 4 for r: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical report, Technical Report no. 597, Department of Statistics, University of Washington, 2012.

[5] Marinov, GK, et al. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Research*, 24(3):496–510, 2014.

[6] Shalek, AK, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 2013

[7] Singer, ZS, et al. Dynamic heterogeneity and DNA methylation in embryonic stem cells. *Molecular Cell*, 55(2):319–331, 2014.

ACKNOWLEDGEMENTS

James Thomson Lab at the Morgridge Institute for Research

CONTACT

Keegan Korthauer:
kdkorthauer@wisc.edu

Christina Kendziorski:
kendzior@biostat.wisc.edu