Submitted to the Annals of Applied Statistics

A COMPOSITIONAL MODEL TO ASSESS EXPRESSION CHANGES FROM SINGLE-CELL RNA-SEQ DATA

By Xiuyu Ma[*], Keegan Korthauer[‡,§], Christina Kendziorski[†], and Michael A. Newton[*,†]

Department of Statistics[*], Department of Biostatistics and Medical Informatics[†], University of Wisconsin - Madison; Department of Data Sciences, Dana-Farber Cancer Institute[‡]; Department of Biostatistics, Harvard T.H. Chan School of Public Health[§]

On the problem of scoring genes for evidence of changes in the distribution of single-cell expression, we introduce an empirical Bayesian mixture approach and evaluate its operating characteristics in a range of numerical experiments. The proposed approach leverages cell-subtype structure revealed in cluster analysis in order to boost gene-level information on expression changes. Cell clustering informs gene-level analysis through a specially-constructed prior distribution over pairs of multinomial probability vectors; this prior meshes with available model-based tools that score patterns of differential expression over multiple subtypes. We derive an explicit formula for the posterior probability that a gene has the same distribution in two cellular conditions, allowing for a gene-specific mixture over subtypes in each condition. Advantage is gained by the compositional structure of the model, in which a host of gene-specific mixture components are allowed, but also in which the mixing proportions are constrained at the whole cell level. This structure leads to a novel form of information sharing through which the cell-clustering results support gene-level scoring of differential distribution. The result, according to our numerical experiments, is improved sensitivity compared to several standard approaches for detecting distributional expression changes.

1. Introduction. The ability to measure genome-wide gene expression at single-cell resolution has accelerated the pace of biological discovery. Overcoming data analysis challenges caused by the scale and unique variation properties of single-cell data will surely fuel further advances in immunology (Papalexi and Satija (2017)), developmental biology (Marioni and Arendt (2017)), cancer (Navin (2015)), and other areas (Nawy (2013)). Computational tools and statistical methodologies created for data of lower-resolution (e.g., bulk RNA-seq) or lower dimension (e.g., flow cytometry) guide our response to the data-science demands of new measurement platforms, but they remain inadequate for efficient knowledge dis-

covery in this rapidly advancing domain (Bacher and Kendziorski (2016)).

An important feature of single-cell studies that could be leveraged better statistically is the fact that cells populate distinct, identifiable subtypes determined by lineage history, epigenetic state, the activity of various transcriptional programs, or other distinguishing factors. Extensive research on clustering cells has produced tools for identifying subtypes, including `SC3` (Kiselev et al. (2017)), `CIDR` (Lin, Troup and Ho (2017)) and `ZIFA` (Pierson and Yau (2015)). We hypothesize that such subtype information may be usefully utilized in other inference procedures in order to improve their operating characteristics.

Assessing the magnitude and statistical significance of changes in gene expression associated with changes in cellular condition has been a central statistical problem in genomics. New tools specific to the single-cell RNA-seq data structure, including `MAST` (Finak et al. (2015)), `scDD` (Korthauer et al. (2016)), and `D3E` (Delmans and Hemberg (2016)), have been deployed to address this problem. These tools respond to scRNA-seq characteristics, such as high prevalence of zero counts and gene-level multimodality, but they do not fully exploit cellular-subtype information and therefore may be underpowered in some settings. The proposed method, which uses negative binomial mixtures to measure changes in a gene's marginal sampling distribution, acquires sensitivity to a variety of distributional effects by how it integrates gene-level data with subtype information. From input data, the associated R package `scDDboost` [1] prioritizes genes with a local false-discovery rate against the null hypothesis of no condition effect on the marginal sampling distribution. The complement of this rate is an empirical Bayesian posterior probability of differential distribution (DD). By incorporating transciptomic information on cell subtypes, scDDboost leverages useful and previously untapped information on each gene's expression sampling distribution.

Through the compositional model underlying `scDDboost`, subtypes inferred by clustering inform the analysis of gene-level expression. The proposed methodology merges two lines of computation after cell clustering: one concerns patterns of differential expression among the cellular subtypes, and here we take advantage of the powerful `EBseq` method for detecting patterns in negative-binomially-distributed expression data (Leng et al. (2013)). The second concerns the counts of cells in various subtypes; for this we propose a Double-Dirichlet-Mixture distribution to model the pair of multinomial probability vectors for subtype counts in two exper-

[1] http://github.com/wiscstatman/scDDboost/

imental conditions. Further elements are developed, on the selection of the number of subtypes and on accounting for uncertainty in the cluster output, in order to provide an end-to-end solution to the differential distribution problem. We note that modularity in the necessary elements provides some methodological advantages. For example, improvements in clustering may be used in place of the default clustering without altering the form of downstream analysis. Also, by avoiding Markov chain Monte Carlo, `scDDboost` computations are relatively inexpensive for a Bayesian procedure.

To set the context by way of example, Figure 1 highlights the ability of `scDDboost` to sense subtype composition changes and thus detect subtle gene expression changes between conditions. The three panels on the left compare expression from 91 human stem cells known to be in the G1 phase of the cell cycle, as well as from 76 such cells known to be in the G2/M phase (Leng et al. (2015)) in three genes (BIRC5, HMMR, and CKAP2), which we happen to know from prior studies have differential activity between G1 and G2/M (Li and Altieri (1999); Sohr and Engeland (2008); Dominguez et al. (2016)). Several standard statistical tools applied to the data behind Figure 1 do not find the observed differences in any of these genes to be statistically significant when controlling the false discovery rate (FDR) at 5%, but `scDDboost` does include these genes on its 5% FDR list. Considering prior studies, these subtle distributional changes are probably not false discoveries. The right panel in Figure 1 shows these three among many other genes also known to be involved in cell-cycle regulation but not identified by standard tools as altered between G1 and G2/M at the 5% FDR level. The color panel provides insight into why `scDDboost` has identified these genes. For this data set, six cellular subtypes were identified in the first step of `scDDboost` (colors red, blue, green, and orange are visible). These subtypes have changed in their proportions between G1 and G2/M; there is a lower proportion of red cells and a greater proportion of orange cells in G2/M, for example. These proportion shifts, which are inferred from genome-wide data, stabilize gene-specific statistics that measure changes between conditions in the mixture distribution of expression, and thereby increase power. We note that `scDDboost` agrees with other statistical tools on very strong differential-distribution signals (not shown), but it has the potential to increase power for subtle signals owing to its unique approach to leveraging cell subtype information.

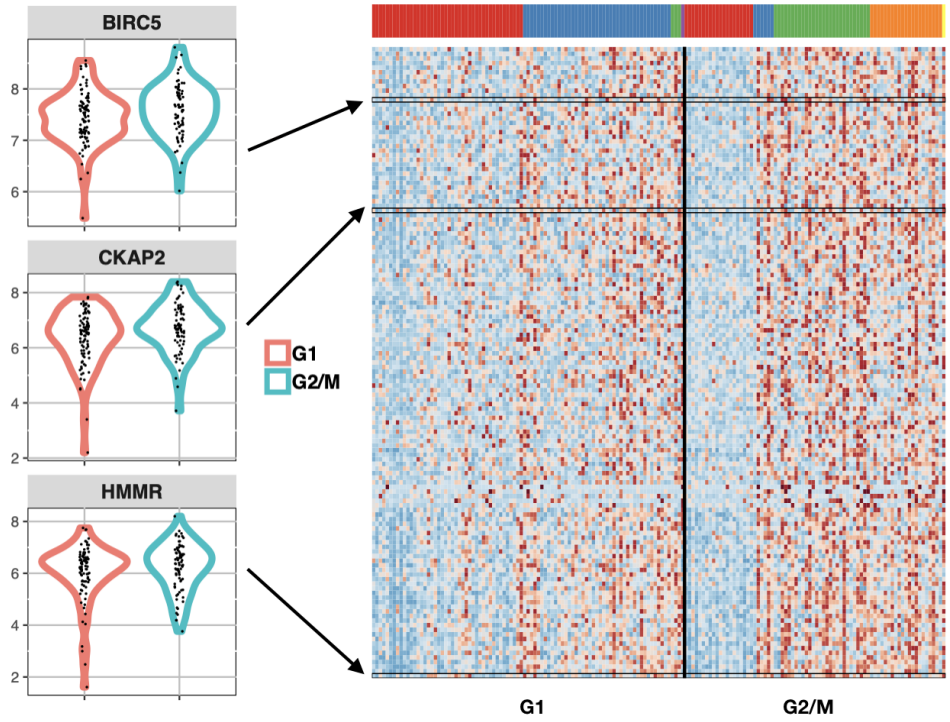

Fig 1: Genes involved in cell-cycle that are identified by scDDboost, but not standard approaches, as differentially distributed between cell-cycle phases G1 and G2/M in human embryonic stem cells (GSE64016). Density estimates on the left show expression data (log2 scale) of three genes identified by scDDboost at 5% FDR, but not similarly identified by MAST, scDD, or DESeq2. Prior studies have shown that the expression of BIRC5, CKAP2, and HMMR is dependent on the phase of cell-cycle, suggesting that these subtle shifts are not false positives. Heatmap (right) shows these three genes among 137 other cell-cycle genes (GO:0007049) identified exclusively by scDDboost, with expression from low (blue) to high (red). Cells (columns) are clustered by their genome-wide expression profiles into distinct cellular subtypes, as indicated by the color panel.

Numerical experiments on both synthetic and published scRNA-seq data bear out the incidental finding in Figure 1, that `scDDboost` has sensitivity for detecting subtle distribution changes. In these experiments we take advantage of `splatter` for generating synthetic data (Zappia, Phipson and Oshlack (2017)) as well as the compendium of scRNA-seq data available through `conquer` (Soneson and Robinson (2017)). Additional numerical experiments show a relationship between `scDDboost` findings and more mechanistic attempts to parameterize transcriptional activation (Delmans and Hemberg (2016)). Finally, we establish first-order asymptotic re-

sults for the methodology.

On manuscript organization, we present the modeling and methodology elements in Section 2, numerical experiments in Section 3, asymptotic analysis in Section 4, and a discussion in Section 5. We relegate some details to an appendix and many others to a Supplementary Material document.

2. Modeling.

2.1. *Data structure, sampling model, and parameters.* In modeling scRNA-seq data, we imagine that each cell c falls into one of $K > 1$ classes, which we think of as subtypes or subpopulations of cells. For notation, $z_c = k$ means that cell c is of subtype k, with the vector $z = (z_c)$ recording the states of all sampled cells. Knowledge of this class structure prior to measurement is not required, as it will be inferred as necessary from available genomic data. We expect that cells arise from multiple experimental conditions, such as by treatment-control status or some other factors measured at the cell level, but we present our development for the special case of two conditions. Notationally, $y = (y_c)$ records the experimental condition, say $y_c = 1$ or $y_c = 2$. Let's say condition j measures $n_j = \sum_c 1[y_c = j]$ cells, and in total we have $n = n_1 + n_2$ cells in the analysis. The examples in Section 3 involve hundreds to thousands of cells. Further let

$$t_k^j = t_k^j(y, z) = \sum_c 1[y_c = j, z_c = k] \tag{1}$$

denote the number of cells of subtype k in condition j and $X_{g,c}$ denote the normalized expression of gene g in cell c. This is one entry in a typically large genes-by-cells data matrix X. Thus, the data structure entails an expression matrix X, a treatment label vector y, and a vector z of latent subtype labels.

We treat subtype counts in the two conditions, $t^1 = (t_1^1, t_2^1, \cdots, t_K^1)$ and $t^2 = (t_1^2, t_2^2, \cdots, t_K^2)$, as independent multinomial vectors, reflecting the experimental design. Explicitly,

$$t^1|y \sim \text{Multinomial}_K(n_1, \phi) \quad \text{and} \quad t^2|y \sim \text{Multinomial}_K(n_2, \psi) \tag{2}$$

for probability vectors $\phi = (\phi_1, \phi_2, \cdots, \phi_K)$ and $\psi = (\psi_1, \psi_2, \cdots, \psi_K)$ that characterize the populations of cells from which the n observed cells are sampled. This follows from the more basic sampling model: $P(z_c = k|y_c = 1) = \phi_k$ and $P(z_c = k|y_c = 2) = \psi_k$.

Our working hypothesis, referred to as the *compositional model*, is that any differences in the distribution of expression $X_{g,c}$ between $y_c = 1$ and $y_c = 2$ (i.e., any condition effects) are attributable to differences between the conditions in the underlying composition of cell types; i.e., owing to $\phi \neq \psi$. We suppose that cells of any given subtype k will present data according to a distribution reflecting technical and biological variation specific to that class of cells, regardless of the condition y_c of the cell. Some care is needed in this, as an overly broad cell subtype (e.g., *epithelial cells*) could have further subtypes that show differential response to some treatment, for example, and so cellular condition (treatment) would then affect the distribution of expression data within the subtype, which is contrary to our working hypothesis. Were that the case, we could have refined the subtype definition to allow a greater number of population classes K in order to mitigate the problem of within-subtype heterogeneity. A risk in this approach is that K could approach n, as if every cell were its own subtype. We find, however, that data sets often encountered do not display this theoretical phenomenon when using a broad class of within-subtype expression distributions. Subtypes are considered such that cellular condition affects their composition but not the sampling distribution of expression within a subtype.

Within the compositional model, let $f_{g,k}$ denote the sampling distribution of expression measurement $X_{g,c}$ assuming that cell c is from subtype k. Then for the two cellular conditions, and at some expression level x, the marginal distributions over subtypes are finite mixtures:

$$f_g^1(x) = \sum_{k=1}^{K} \phi_k f_{g,k}(x) \quad \text{and} \quad f_g^2(x) = \sum_{k=1}^{K} \psi_k f_{g,k}(x).$$

In other words, $X_{g,c}|[y_c = j] \sim f_g^j$ and $X_{g,c}|[z_c = k, y_c = j] \sim f_{g,k}$.

We say that gene g is *differentially distributed*, denoted DD_g and indicated by $f_g^1 \neq f_g^2$, if $f_g^1(x) \neq f_g^2(x)$ for some x, and otherwise it is equivalently distributed (ED_g). Motivated by findings from bulk RNA-seq data analysis, we further set each $f_{g,k}$ to have a a negative-binomial form, with mean $\mu_{g,k}$ and shape parameter σ_g, as in (Leng et al. (2013), Anders and Huber (2010), Love, Huber and Anders (2014) and Chen et al. (2018)). This choice is effective in our numerical experiments though it is not critical to the modeling formulation. The use of mixtures per gene has proven useful in related model-based approaches (e.g., Finak et al. (2015); McDavid et al. (2014); Huang et al. (2018)).

We seek methodology to prioritize genes for evidence of DD_g. Inter-

estingly, even if we have evidence for condition effects on the subtype frequencies, it does not follow that a given gene will have $f_g^1 \neq f_g^2$; that depends on whether or not the subtypes show the right pattern of differential expression at g, to use the standard terminology from bulk RNA-seq. For example, if two subtypes have different frequencies between the two conditions ($\phi_1 \neq \psi_1$ and $\phi_2 \neq \psi_2$) but the same aggregate frequency ($\phi_1 + \phi_2 = \psi_1 + \psi_2$), and also if components are equivalent, $f_{g,1} = f_{g,2}$, then, other things being equal, marginals $f_g^1 = f_g^2$ even though $\phi \neq \psi$. Details confirming such equality are examplified further in Supplementary Material Section 2.1. The fact is so central that we emphasize:

Key issue: A gene that does not distinguish two subtypes will also not distinguish the cellular conditions if those subtypes appear in the same aggregate frequency in the two conditions, regardless of changes in the individual subtype frequencies.

We formalize this issue in order that our methodology has the necessary functionality. To do so, first consider the parameter space $\Theta = \{\theta = (\phi, \psi, \mu, \sigma)\}$, where $\phi = (\phi_1, \phi_2, \cdots, \phi_K)$ and $\psi = (\psi_1, \psi_2, \cdots, \psi_K)$ are as before, where $\mu = \{\mu_{g,k}\}$ holds all the subtype-and-gene-specific expected values, and where $\sigma = \{\sigma_g\}$ holds all the gene-specific negative-binomial shape parameters. Critical to our construction are special subsets of Θ corresponding to partitions of the K cell subtypes. A single partition, π, is a set of mutually exclusive and exhaustive blocks, b, where each block is a subset of $\{1, 2, \cdots, K\}$, and we write $\pi = \{b\}$. Of course, the set Π containing all partitions π of $\{1, 2, \cdots, K\}$ has cardinality that grows rapidly with K. We carry along an example involving $K = 7$ cell types, and one three-block partition taken from the set of 877 possible partitions of $\{1, 2, \cdots, 7\}$ (Figure 2).

For any partition $\pi = \{b\}$, consider aggregate subtype frequencies

$$\Phi_b = \sum_{k \in b} \phi_k \quad \text{and} \quad \Psi_b = \sum_{k \in b} \psi_k,$$

and extend the notation, allowing vectors $\Phi_\pi = \{\Phi_b : b \in \pi\}$ and similarly for Ψ_π. Recall the partial ordering of partitions based on refinement, and note that as long as π is not the most refined partition (every cell type is in its own block), then the mapping from (ϕ, ψ) to (Φ_π, Ψ_π) is many-to-one. Further, define sets

$$A_\pi = \{\theta \in \Theta : \ \Phi_b = \Psi_b \, \forall b \in \pi\}. \tag{3}$$

and

$$M_{g,\pi} = \{\theta \in \Theta : \ \mu_{g,k} = \mu_{g,k'} \iff k, k' \in b, b \in \pi\}. \tag{4}$$

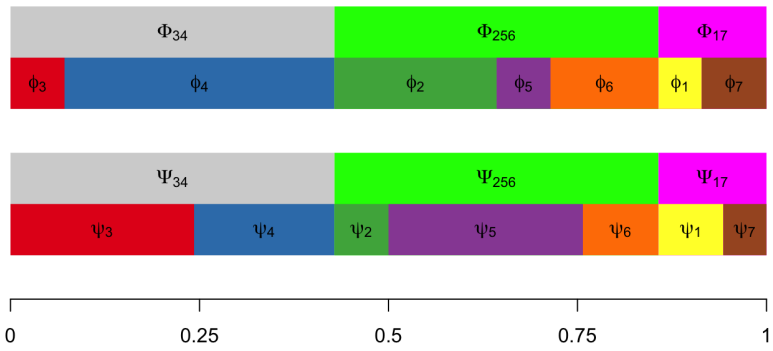

Fig 2: Proportions of $K = 7$ cellular subtypes in two different conditions. Aggregated proportions of subtypes 3 and 4, subtypes 2, 5, and 6, and subtypes 1, and 7 remain same across conditions, while individual subtype frequencies change. Depending on the changes in average expression among subtypes, these frequency changes may or may not induce changes between two conditions in the marginal distribution of some gene's expression.

Under A_π there are constraints on cell subtype frequencies; under $M_{g,\pi}$ there is equivalence in the gene-level distribution of expression between certain subtypes. These sets are precisely the structures needed to address differential distribution DD_g (and it complement, equivalent distribution, ED_g) at a given gene g, since:

THEOREM 1. *Let $C_{g,\pi} = A_\pi \cap M_{g,\pi}$. For partitions $\pi_1 \neq \pi_2$, $C_{g,\pi_1} \cap C_{g,\pi_2} = \emptyset$. Further, at any gene g, equivalent distribution is*

$$\mathrm{ED}_g = \bigcup_{\pi \in \Pi} C_{g,\pi}.$$

With additional probability structure on the parameter space, we immediately obtain from Theorem 1 a formula for local false discovery rates:

$$1 - P(\mathrm{DD}_g | X, y) = P(\mathrm{ED}_g | X, y) = \sum_{\pi \in \Pi} P\left(A_\pi \cap M_{g,\pi} | X, y\right). \tag{5}$$

Local false discovery rates are important empirical Bayesian statistics in large-scale testing (Efron (2007); Muralidharan (2010); Newton et al. (2004)). For example, the conditional false discovery rate of a list of genes is the

arithmetic mean of the associated local false discovery rates. The partition representation guides the construction of a prior distribution (Section 2.3) and a model-based method (Section 2.2) for scoring differential distribution. Setting the stage, Figure 3 shows the dependency structure of the proposed compositional model and the partition-reliant prior specification.

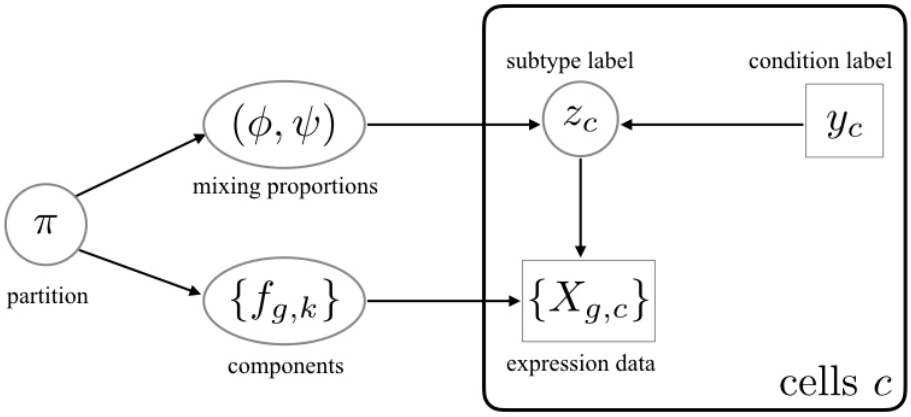

Fig 3: Directed acyclic graph structure of the compositional model and partition-reliant prior. The plate on the right side indicates i.i.d. copies over cells c, conditionally on mixing proportions and mixing components. Observed data are indicated in rectangles/squares, and unobserved variables are in circles/ovals.

Key to computing the gene-specific local false discovery rate $P(\mathrm{ED}_g|X, y)$ is evaluating probabilities $P\left(A_\pi \cap M_{g,\pi}|X, y\right)$. The dependence structure (Figure 3) implies a useful reduction of this quantity, at least conditionally upon subtype labels $z = (z_c)$. For each subtype partition π and gene g,

THEOREM 2. $P\left(A_\pi \cap M_{g,\pi}|X, y, z\right) = P\left(A_\pi|y, z\right) P\left(M_{g,\pi}|X, z\right)$.

In what follows, we develop the modeling and computational elements necessary to efficiently evaluate inference summaries (5) taking advantage of Theorems 1 and 2. Roughly, the methodological idea is that subtype labels z have relatively low uncertainty, and may be estimated from genome-wide clustering of cells in the absence of condition information y (up to an arbitrary label permutation). The modest uncertainty in z we handle through a computationally efficient randomized clustering scheme. Theo-

rem 2 indicates that our computational task then separates into two parts given z. On one hand, cell subtype frequencies combine with condition labels to give $P(A_\pi|y,z)$. Then gene-level data locally drive the posterior probabilities $P\left(M_{g,\pi}|X,z\right)$ that measure differential expression between subtypes. Essentially, the model provides a specific form of information sharing between genes that leverages the compositional structure of single-cell data in order to sharpen our assessments of between-condition expression changes.

2.2. *Method structure and clustering.* To infer subtypes, we leverage the extensive research on how to cluster cells using scRNA-seq data: for example, SC3 (Kiselev et al. (2017)), CIDR (Lin, Troup and Ho (2017)), and ZIFA (Pierson and Yau (2015)). We propose distance-based clustering on the full set of profiles in a way that is blind to the condition label vector y, in order to have as many cells as possible to inform the subtype structure. We investigated several clustering schemes in numerical experiments and allow flexibility in this choice within the scDDboost software. Associating clusters with subtype labels $\hat{z}_c$ estimates the actual subtypes z_c, and prepares us to use Theorems 1 and 2 in order to compute separate posterior probabilities $P(A_\pi|y,\hat{z})$ and $P(M_{g,\pi}|X,\hat{z})$ that are necessary for scoring differential distribution. The first probability concerns patterns of cell counts over subtypes in the two conditions, and has a convenient closed form within the double-Dirichlet model (Section 2.3). The second probability concerns patterns of changes in expected expression levels among subtypes, and this is also conveniently computed for negative-binomial counts using `EBSeq` (Leng et al. (2013)). Algorithm 1 summarizes how these elements combine to get the posterior probability of differential distribution per gene, conditional on an estimate of the subtype labels.

We invoke $K-$medoids (Kaufman and Rousseeuw (1987)) as the default clustering method in `scDDboost`, and customize the cell-cell distance by integrating two measures. The first assembles gene-level information by cluster-based-similarity partitioning (Strehl and Ghosh (2003)). Separately at each gene, modal clustering (Dahl (2009) and Supplementary Material Section 2.2) partitions the cells, and then we define dissimilarity between cells as the Manhattan distance between gene-specific partition labels. A second measure defines dissimilarity by one minus the Pearson correlation between cells, which is computationally inexpensive, less sensitive to outliers than Euclidean distance, and effective at detecting cellular clusters in scRNA-seq (Kim et al. (2018a)). The default clustering in `scDDboost` combines these two measures by weighted average, with $w_C = \frac{\sigma_P}{\sigma_C+\sigma_P}$ and

Algorithm 1 scDDBoost-core

Input:

genes by cells expression data matrix $X = (X_{g,c})$

cell condition labels $y = (y_c)$

cell subtype labels (estimated) $\hat{z} = (\hat{z}_c)$

Output: posterior probabilities of differential distribution from estimated subtypes

procedure scDDBoost-core($X, y, \hat{z}$)
number of cell subtypes $K = \text{length}(\text{unique}(\hat{z}))$
  subtype differential expression: $\forall g, \pi$ compute $P(M_{g,\pi}|X, \hat{z})$ using `EBSeq`
  cell frequency changes: $\forall \pi$ compute $P(A_\pi|y, \hat{z})$ using Double Dirichlet model
  posterior probability: $\forall g$, $P(\text{ED}_g|X, y, \hat{z}) \leftarrow \sum_\pi P(M_{g,\pi}|X, \hat{z})\, P(A_\pi|y, \hat{z})$
  return $\forall g$, $P(\text{DD}_g|X, y, \hat{z}) = 1 - P(\text{ED}_g|X, y, \hat{z})$
end procedure

$w_P = 1 - w_C$, where $w_C, \sigma_C, w_P, \sigma_P$ are the weights and standard deviations of cluster-based distance and Pearson-correlation distance, respectively. The software allows other distances, such as provided by SC3, which we use in some numerical experiments; in any case the final distance matrix is denoted $D = \left(d_{i,j}\right)$.

Any clustering method entails classification errors, and so $\hat{z}_c \neq z_c$ for some cells. To mitigate the effects of this uncertainty, `scDDboost` averages output probabilities from scDDBoost-core over randomized clusterings $\hat{z}^*$. These are not uniformly random, but rather are generated by applying $K-$medoids to a randomized distance matrix $D^* = \left(d_{i,j}/w_{i,j}\right)$, where $w_{i,j}$ are non-negative weights $w_{i,j} = (e_i + e_j)$, and where (e_i) are independent and identically Gamma distributed deviates with shape $\hat{a}/2$ and rate $\hat{a}$, and where $\hat{a}$ is estimated from D. (Thus $w_{i,j}$ is Gamma$(\hat{a}, \hat{a})$ and has unit mean.) The distribution of clusterings induced by this simple computational scheme approximates a Bayesian posterior analysis, as we argue in the Appendix, where we also present pseudo-code for the resulting `scDDboost` Algorithm 2. Averaging over results from randomized clusterings gives additional stability to the posterior probability statistics (Supplementary Figure S10).

Computations become more intensive the larger is the number K of cell subtypes. Version 1.0 of `scDDboost` is restricted to $K \leq 9$; we consider further computational strategies in Section 5. Inferentially, taking K to be too large may inflate the false positive rate (Supplementary Figure S11). The approach taken in `scDDboost` is to set K using the validity score (Ray and Turi (2000)), which measures changes in within-cluster sum of squares

as we increase K. Our implementation, in Supplementary Material Section 2.2, shows good operating characteristics in simulation.

2.3. $P(A_\pi|y,z)$. We introduce the Double Dirichlet Mixture (DDM), which is the partition-reliant prior $p(\phi,\psi)$ indicated in Figure 3, in order to derive an explicit formula for $P(A_\pi|y,z)$. We lose no generality here by defining $A_\pi = \{(\phi,\psi) : \Phi_b = \Psi_b \; \forall b \in \pi\}$, rather than as a subset of the full parameter space as in (3). Each A_π is closed and convex subset of the product space holding all possible pairs of length-K probability vectors.

We propose a spike-slab-style mixture prior with the following form:

$$p(\phi,\psi) = \sum_{\pi\in\Pi} \omega_\pi \, p_\pi(\phi,\psi). \tag{6}$$

Each mixture component $p_\pi(\phi,\psi)$ has support A_π; the mixing proportions ω_π are positive constants summing to one. To specify component p_π, notice that on A_π there is a 1-1 correspondence between pairs (ϕ,ψ) and parameter states:

$$\{(\tilde{\phi}_b, \tilde{\psi}_b, \Phi_b), \; \forall b \in \pi\}, \tag{7}$$

where

$$\tilde{\phi}_b = \frac{\phi_b}{\Phi_b}, \quad \tilde{\psi}_b = \frac{\psi_b}{\Psi_b}, \quad \text{and} \quad \Phi_b = \sum_{k\in b} \phi_k = \sum_{k\in b} \psi_k = \Psi_b.$$

For example, $\tilde{\phi}_b$ is a vector of conditional probabilities for each subtype given that a cell from the first condition is one of the subtypes in b.

We introduce hyperparameters $\alpha_k^1, \alpha_k^2 > 0$ for each subtype k, and set $\beta_b = \sum_{k\in b} \left(\alpha_k^1 + \alpha_k^2\right)$ for any possible block b. Extending notation, let α_b^j be the vector of α_k^j for $k \in b$, β_π be the vector of β_b for $b \in \pi$, ϕ_b and ψ_b be vectors of ϕ_k and ψ_k, respectively, for $k \in b$, and Φ_π and Ψ_π be the vectors of Φ_b and Ψ_b for $b \in \pi$. The proposed double-Dirichlet component p_π is determined in the transformed scale by assuming $\Psi_\pi = \Phi_\pi$ and further:

$$\begin{aligned} \Phi_\pi &\sim \text{Dirichet}_{N(\pi)}[\beta_\pi] \\ \tilde{\phi}_b &\sim \text{Dirichlet}_{N(b)}[\alpha_b^1] \qquad \forall b \in \pi \\ \tilde{\psi}_b &\sim \text{Dirichlet}_{N(b)}[\alpha_b^2] \qquad \forall b \in \pi \end{aligned} \tag{8}$$

where $N(\pi)$ is the number of blocks in π and $N(b)$ is the number of subtypes in b, and where all random vectors in (8) are mutually independent. Mixing over π as in (6), we write $(\phi,\psi) \sim \text{DDM}\left[\omega = (\omega_\pi), \alpha^1 = (\alpha_k^1), \alpha^2 = (\alpha_k^2)\right]$.

We record some properties of the component distributions p_π:

Property 1: In $p_\pi(\phi, \psi)$, ψ and ϕ are dependent, unless π is the null partition in which all subtypes constitute a single block.

Property 2: With $k \in b$, marginal means are:

$$E_\pi(\phi_k) = \frac{\alpha_k^1}{\sum_{k' \in b} \alpha_{k'}^1} \frac{\beta_b}{\sum_{b' \in \pi} \beta_{b'}} \quad \text{and} \quad E_\pi(\psi_k) = \frac{\alpha_k^2}{\sum_{k' \in b} \alpha_{k'}^2} \frac{\beta_b}{\sum_{b' \in \pi} \beta_{b'}}.$$

Recall from (1) the vectors t^1 and t^2 holding counts of cells in each subtype in each condition, computed from y and z. Relative to a block $b \in \pi$, let $t_b^j = \sum_{k \in b} t_k^j$, for cell conditions $j = 1, 2$, and, let t_π^j be the vector of these counts over $b \in \pi$. The following properties refer to marginal distributions in which (ϕ, ψ) have been integrated out of the joint distribution involving (2) and the component p_π.

Property 3: t^1 and t^2 are conditionally independent given y, t_π^1 and t_π^2.

Property 4: For $j = 1, 2$,

$$p_\pi(t^j | t_\pi^j, y) = \prod_{b \in \pi} \left\{ \left[\frac{\Gamma(t_b^j + 1)}{\prod_{k \in b} \Gamma(t_k^j + 1)} \right] \left[\frac{\Gamma(\sum_{k \in b} \alpha_k^j)}{\prod_{k \in b} \Gamma(\alpha_k^j)} \right] \left[\frac{\prod_{k \in b} \Gamma(\alpha_k^j + t_k^j)}{\Gamma(t_b^j + \sum_{k \in b} \alpha_k^j))} \right] \right\}$$

Property 5:

$$p_\pi(t_\pi^1, t_\pi^2 | y) = \left[\frac{\Gamma(n_1 + 1)\Gamma(n_2 + 1)}{\prod_{b \in \pi} \Gamma(t_b^1 + 1)\Gamma(t_b^2 + 1)} \right] \left[\frac{\Gamma(\sum_{b \in \pi} \beta_b)}{\prod_{b \in \pi} \Gamma(\beta_b)} \right] \left[\frac{\prod_{b \in \pi} \Gamma(\beta_b + t_b^1 + t_b^2)}{\Gamma(n_1 + n_2 + \sum_{b \in \pi} \beta_b)} \right].$$

Let's look at some special cases to dissect this result.

Case 1. If π has a single block equal to the entire set of cell types $\{1, 2, \cdots, K\}$, then $t_b^j = n_j$ for both $j = 1, 2$, and Property 5 reduces, correctly, to $p_\pi(t_\pi^1, t_\pi^2 | y) = 1$. Further,

$$p_\pi(t^j | t_\pi^j, y) = \left[\frac{\Gamma(n_j + 1)}{\Gamma(n_j + \sum_{k=1}^K \alpha_k^j)} \right] \left[\frac{\Gamma(\sum_{k=1}^K \alpha_k^j)}{\prod_{k=1}^K \Gamma(\alpha_k^j)} \right] \left[\prod_{k=1}^K \frac{\Gamma(\alpha_k^j + t_k^j)}{\Gamma(t_k^j + 1)} \right]$$

which is the well-known Dirichlet-multinomial predictive distribution for counts t^j (Wagner and Taudes (1986)). E.g, taking $\alpha_k^j = 1$ for all types k we get the uniform distribution

$$p_\pi(t^j | t_\pi^j, y) = \frac{\Gamma(n_j + 1)\Gamma(K)}{\Gamma(n_j + K)}.$$

Case 2. At the opposite extreme, π has one block b for each class k, so $\phi = \psi$. Then $p_\pi(t^j|t^j_\pi, y) = 1$, and further, writing $b = k$,

$$p_\pi(t^1_\pi, t^2_\pi|y) = \left[\frac{\Gamma(n_1+1)\Gamma(n_2+1)}{\prod_{k=1}^K \Gamma(t^1_k+1)\Gamma(t^2_k+1)}\right]\left[\frac{\Gamma(\sum_{k=1}^K \beta_k)}{\prod_{k=1}^K \Gamma(\beta_k)}\right]\left[\frac{\prod_{k=1}^K \Gamma(\beta_k + t^1_k + t^2_k)}{\Gamma(n_1+n_2+\beta_k)}\right].$$

which corresponds to Dirichlet-multinomial predictive distribution for counts $t^1 + t^2$ since t^1 and t^2 are identical distributed given (ϕ, ψ) in this case. These properties are useful in establishing:

THEOREM 3. *DDM is conjugate to multinomial sampling of t^1 and t^2:*

$$(\phi, \psi)|y, z \sim \text{DDM}\left[\omega^{\text{post}} = (\omega^{\text{post}}_\pi), \alpha^1 + t^1, \alpha^2 + t^2\right]$$

where

$$\omega^{\text{post}}_\pi \propto p_\pi(t^1|t^1_\pi, y)\, p_\pi(t^2|t^2_\pi, y)\, p_\pi(t^1_\pi, t^2_\pi|y)\, \omega_\pi. \tag{9}$$

The target probability $P(A_\pi|y, z)$ is an integral of the posterior distribution in Theorem 3. To evaluate it, we need to contend with the fact that sets $\{A_\pi : \pi \in \Pi\}$ are not disjoint. Relevant overlaps have to do with partition refinement. Recall that a partition π^r is a refinement of a partition π^c if for any $b \in \pi^c$ there exists $s \subset \pi^r$ such that $\underset{b' \in s}{\cup} b' = b$. We say π^c coarsens π^r when π^r refines π^c. Any partition both refines and coarsens itself, as a trivial case. Generally, refinements increase the number of blocks. If subtype frequency vectors (ϕ, ψ) satisfy the constraints in A_{π^r} then they also satisfy the constraints of any π^c that coarsens π^r: i.e., $A_{\pi^r} \subset A_{\pi^c}$. Refinements reduce the dimension of allowable parameter states. For the double-Dirichlet component distributions P_π, we find:

Property 6: For two partitions $\tilde{\pi}$ and π, $P_{\tilde{\pi}}(A_\pi|y, z) = 1[\tilde{\pi} \text{ refines } \pi]$.

This supports the main finding of this section:

$$P(A_\pi|y, z) = \sum_{\tilde{\pi} \in \Pi} \omega^{\text{post}}_{\tilde{\pi}}\, 1[\tilde{\pi} \text{ refines } \pi]. \tag{10}$$

2.4. $P(M_{g,\pi}|X, z)$. We leverage well-established modeling techniques for transcript analysis, including (Leng et al. (2013), Kendziorski et al. (2003), and Jensen et al. (2009)), which characterize equivalent or differential expression in terms of shared or independently drawn mean effects. Let $X_{g,b}$ denote the subvector of expression values at gene g over cells c with $z_c = k$ for which subtype k is part of block b of partition π. Conditioning on subtype labels $z = (z_c)$, we assume that under $M_{g,\pi}$:

1. *between blocks:* subvectors $\{X_{g,b} : b \in \pi\}$ are mutually independent,
2. *within blocks:* for cells mapping to block b, observations $X_{g,c}$ are i.i.d.
3. *mean effects:* for each block b, there is a univariate mean, $\mu_{g,b}$, shared by cells mapping to that block. *a priori* these means are i.i.d. between blocks.

These assumptions imply a useful factorization marginally to latent means,

$$P(X_g|M_{g,\pi}, z) = \prod_{b\in\pi} f(X_{g,b}), \tag{11}$$

where f is a customized density kernel. In our case we use `EBseq` from (Leng et al. (2013)): the sampling distribution of $X_{g,c}$ is negative binomial, and f becomes a particular compound multivariate negative binomial formed from integrating uncertainty in the block-specific means (see Supplementary Material Section 2.2). Through its gene-level mixing model, `EBseq` also gives estimates of $\{P(M_{g,\pi}|z)\}$: the proportions of genes governed by any of the different patterns π of equivalent/differential expression among subtypes. With these estimates and (11) we compute by Bayes's rule:

$$P(M_{g,\pi}|X, z) \propto P(M_{g,\pi}|z) \prod_{b\in\pi} f(X_{g,b}).$$

The proportionality is resolved by calculating over all partitions π.

3. Numerical experiments.

3.1. *Synthetic data.* We used `splatter` (v. 1.2.0) to generate synthetic scRNA-seq data for which the DD status of genes is known (Zappia, Phipson and Oshlack (2017)), thereby allowing us to measure operating characteristics of `scDDboost` in a controlled setting. Splatter is a generative system for simulating realistic single-cell RNA-seq data. It accounts for biological and technical sources of variation and is calibrated from a number of published data sets. Our hypothetical two-condition comparison involved $n = 400$ cells and 17421 genes, and mixing over various numbers K of distinct subtypes. To reflect common variation patterns, we adopted default settings of the primary parameters in splatter, and focused our experiments on four settings of splatter's location and scale parameters (θ, γ), which encode distributional shifts between subtypes. We entertained 12 scenarios encoding 4 distributional shift settings for each of 3 different values for the number K of subtypes, with composition parameters ϕ and ψ selected to account for various mixing possibilities. Ten replicate data sets

were simulated on each scenario. These 12 scenarios, encoded by $K/\theta/\gamma$, span states with rather strong signals, like $3/-0.1/1$ to quite weak signals, like $15/0.1/0.4$. Supplementary Figures S6 and S7 provide a view of the global separation between the subtypes and the degree of difficulty of the inference task. We note that the mechanistic sampling model induced by splatter is distinct from the descriptive model underlying scDDboost. We choose it to reflect anticipated technical and biological sources of variation. Further details are in Supplementary Material Section 3.1.

Figures 4 and 5 summarize the true positive rate and false discovery rate of `scDDboost` compared to three other methodologies: `MAST` (v. 1.4.0), `scDD` (v. 1.2.0), and `DESeq2` (v. 1.18.1). `scDDboost` exhibits very good operating characteristics in this study, as it controls the FDR in all cases while also delivering a relatively high rate of true positives in all cases. The beneficial sampling properties are not limited to the 5% FDR threshold, as indicated by receiver operator characteristic (ROC) curves (Supplementary Figure S9).

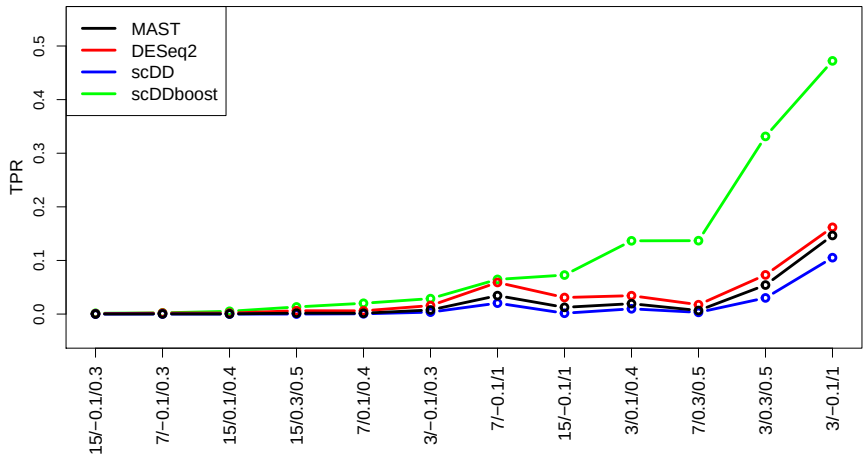

Fig 4: True positive rate (vertical) of four DD detection methods in 12 synthetic-data settings (horizontal). Settings are labeled for $K/\theta/\gamma$ and ranked by scDDboost values. Each method is targeting a 5% false discovery rate (FDR). The plot shows average rates over replicate simulated data in each setting.

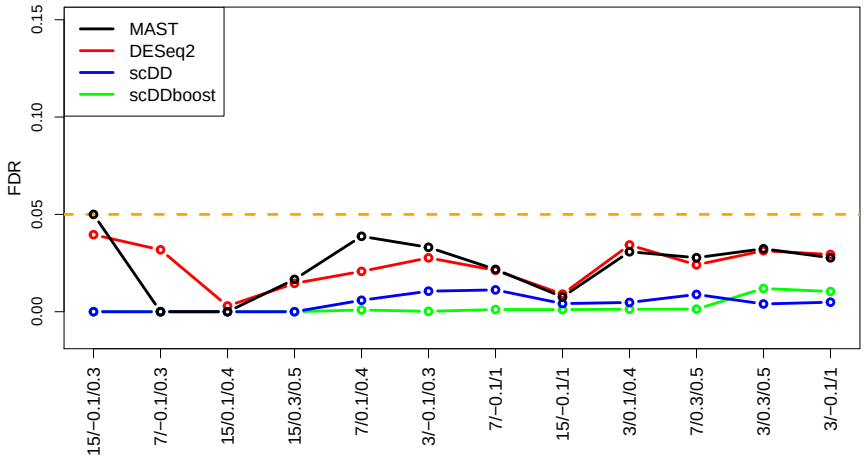

Fig 5: False discovery rate (vertical) of methods in settings (horizontal, same order) from Figure 4

3.2. *Empirical study.* We applied `scDDboost` to a collection of previously published data sets that are recorded at `conquer` (Soneson and Robinson (2017)). Though not knowing the truly DD genes, we can examine how `scDDboost` output compares to output from several standard methods. We selected 12 data sets from `conquer` representing different species and experimental settings and involving hundreds to thousands of cells. Appendix Table A2 provides details. Figure 6 compares methods in terms of the size of the reported list of DD genes at the 5% FDR target level. We see a consistently high yield of `scDDboost` among the evaluated methods. For reference, one of these data sets (GSE64016) happens to be the data behind Figure 1, where we know from other information that some of the uniquely identified genes are likely not to be false positives.

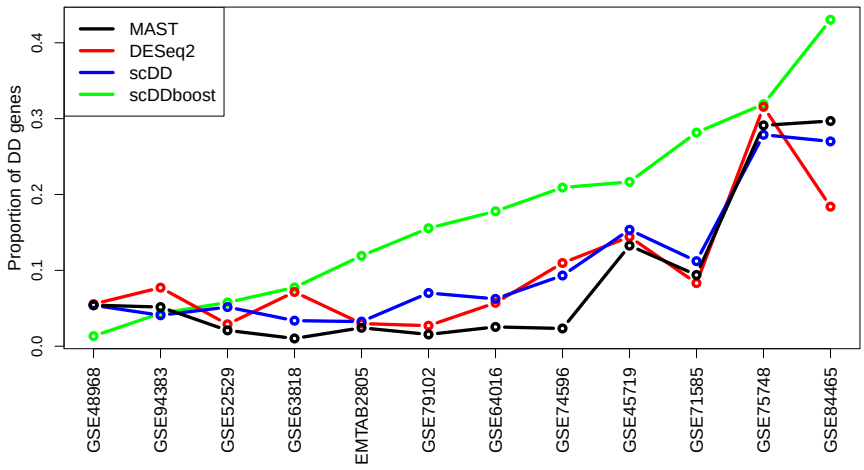

Fig 6: Proportion of DD genes at 5% FDR threshold with respect to total number of genes identified by each method. Ranked by scDDboost list size

To check that the increased discovery rate of `scDDboost` is not associated with an increased rate of false calls, we applied it to a series of random splits of single-condition data sets (Appendix Table A3). Figure 7 confirms a very low call rate in cases where no changes in distribution are expected.

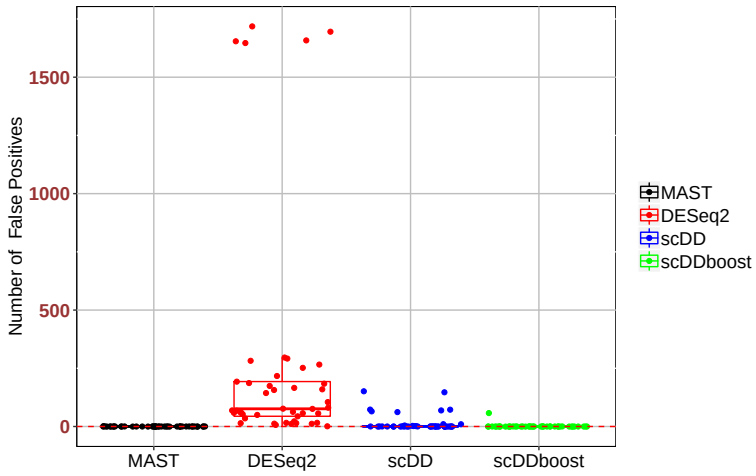

Fig 7: False positive counts at 5% FDR threshold by several methods on 5 random splits of 9 single-condition data sets from Appendix Table A3

We conjecture that `scDDboost` gains power through its novel approach to borrowing strength across genes; i.e., that the genomic data are providing information about cell subtypes and mixing proportions, leaving gene-level data to guide gene-specific mixture components. One way to drill into this idea is to consider how many genes have similar expression characteristics to a given gene. By virtue of the `EBseq` analysis inside `scDDboost`, we may assign each gene to a set of related genes that all have the same highest-probability pattern of equality/inequality of means across the subtypes. Say $\hat{\pi}_g = \text{argmax}_\pi P(M_{g,\pi}|\hat{z}, X)$. In Figure 8, we show that compared to DD genes commonly identified by multiple methods (blue), the set sizes for genes uniquely identified by `scDDboost` (red) tend to be larger. Essentially, the proposed methodology boosts weak DD evidence when a gene's pattern of differential expression among cell subtypes matches a large number of other genes.

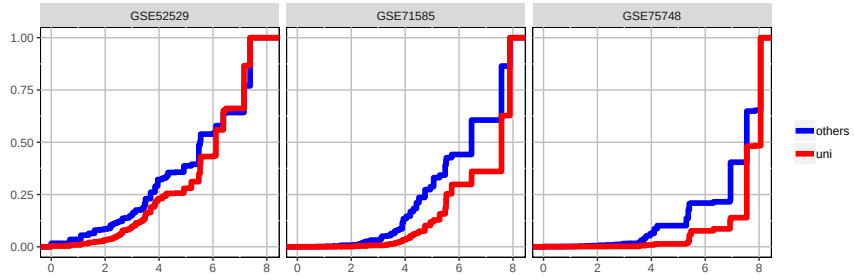

Fig 8: Genes are grouped by their pattern of differential expression across subtypes as inferred by the EBseq computation within scDDboost for three example datasets. Cumulative distribution functions of the log-scale size statistic for all genes identified by scDDboost are plotted; red is the subset uniquely identified by scDDboost; blue are those also identified by the comparison methods (MAST, scDD, or DESeq2). Sets of similarly-patterned genes tend to be larger (horizontal axis, log size) for genes uniquely identified by scDDboost (red) compared to other DD genes (blue), at 5% FDR.

3.3. *Bursting.* Transcriptional bursting is a fundamental property of genes, wherein transcription is either negligible or attains a certain probability of activation (Raj and van Oudenaarden (2008)). D3E (Delmans and Hemberg (2016)) is a computationally intensive method for DE gene analysis rooted in modeling the bursting process. It considers transcripts as in the stationary distribution from an experimentally validated stochastic process of single-cell gene expression (Peccoud and Ycart (1995)). Three mechanistic parameters (rate of promoter activation, rate of promoter inactivation, and the conditional rate of transcription given an active pro-

moter) characterize the model, which allows distributional changes between conditions without changes in mean expression level. For genes identified as DD by `scDDboost` in dataset GSE71585, either uniquely or in common with comparison methods MAST, scDD, and DESeq2, Figure 9 shows changes of these bursting parameters. Interestingly, genes uniquely identified by `scDDboost` are associated with more significant changes between estimated bursting parameters than genes that all methods identify. This finding and similar findings on other data sets (not shown) provide some evidence that `scDDboost` is able to detect biologically meaningful changes in the expression distribution.

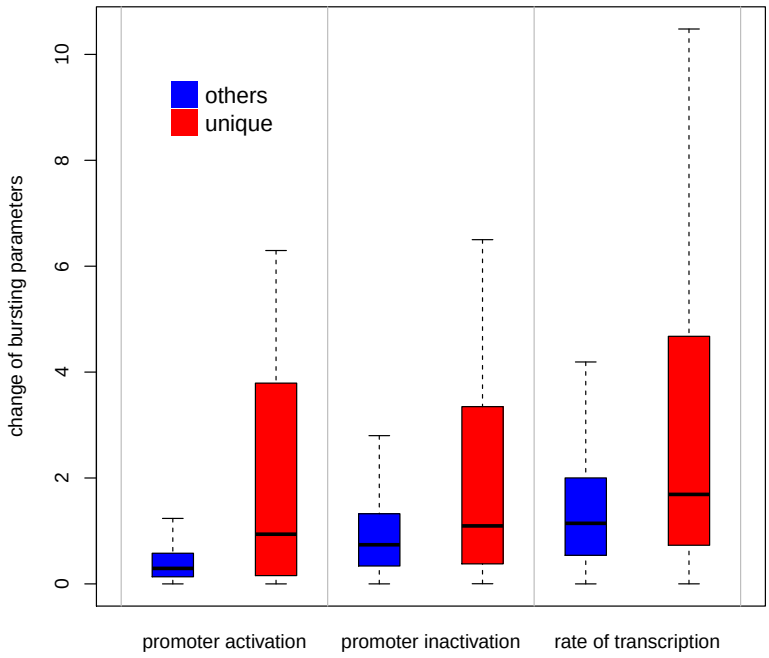

Fig 9: Absolute values of log fold changes of bursting parameters tend to be larger for 1758 genes uniquely identified by scDDboost (red) compared to 2983 genes (blue) that are identified at 5% FDR by scDDboost and other methods: MAST, scDD, and DESeq2.

3.4. *Time Complexity.* Run time complexity of `scDDboost` is dominated by the cost of clustering cells and of running `EBSeq` to measure differ-

ences between subtypes. Recall the notation that n for number of cells, G for number of genes and K for number of subtypes. Our distance-based clustering of n cells measuring G genes requires on the order of $G \times n^2$ operations (see Supplementary Material Section 2.2). Further, `EBSeq` uses summed counts within each subtype for each gene to compute its density kernel, and there are Bell(K) differential patterns to compute, where Bell counts the partitions of K. We impose the computational limit $K \leq 9$ in `scDDboost` (v. 1.0). In a typical case involving 20000 genes and 200 cells, using 50 of randomized distances, `scDDboost` is relatively efficient for $K \leq 6$ requiring less than 15 CPU minutes on, for example, a quad-core 2.2 GHz Intel Core i7 with 16 Gb of RAM. The same data might require 20 to 40 CPU hours when $K = 9$. In Section 5 we mention some opportunities to improve this speed.

3.5. *Diagnostics.* As implemented, scDDboost uses a particular distance matrix to inform subtypes and computes probabilities in a model for which expression is a mixture of constant-shape negative binomials. To check the effect of these assumptions we consider a variety of diagnostic calculations using the data sets presented in Sections 3.1 and 3.2. We first point out that model mis-specification may have a limited impact on Type-I error rates, as evidenced by the permutation study (Figure 7) and also the synthetic-data study (Figure 5), which does not encode the same modeling assumptions as scDDboost.

To check the within-subtype negative binomial (NB) assumption, we deployed a bootstrap goodness-of-fit test in three data sets (Yin and Ma (2013)). Fewer than 1.5% of genes show evidence against a within subtype NB assumption at a 5% FDR. Further, for most of these non-NB genes the inference drawn by `scDDboost` is the same as that drawn by various other methods (Supplementary Material, Table S3), and where there are differences the `scDDboost` call is plausible (Supplementary Material, Figure S12). Among the genes identified by `scDDboost` at 5% FDR in the stem-cell example (recall Figure 1), just six of them fail the NB test, and two of these are uniquely called by `scDDboost`; further, one of the two genes is cell-cycle related.

The constant-shape assumption is less well supported empirically, according to a likelihood-ratio test that we developed (Supplementary Figure S13). More than 15% of genes show evidence against constant shape in the examples considered, though inference on differential distribution is only mildly affected (Supplementary Table S4; Figure S14). In spite of

model mis-specifications, Figure 10 shows for a random set of genes from the stem-cell study that the marginal fit by `scDDboost` is reasonably accurate.

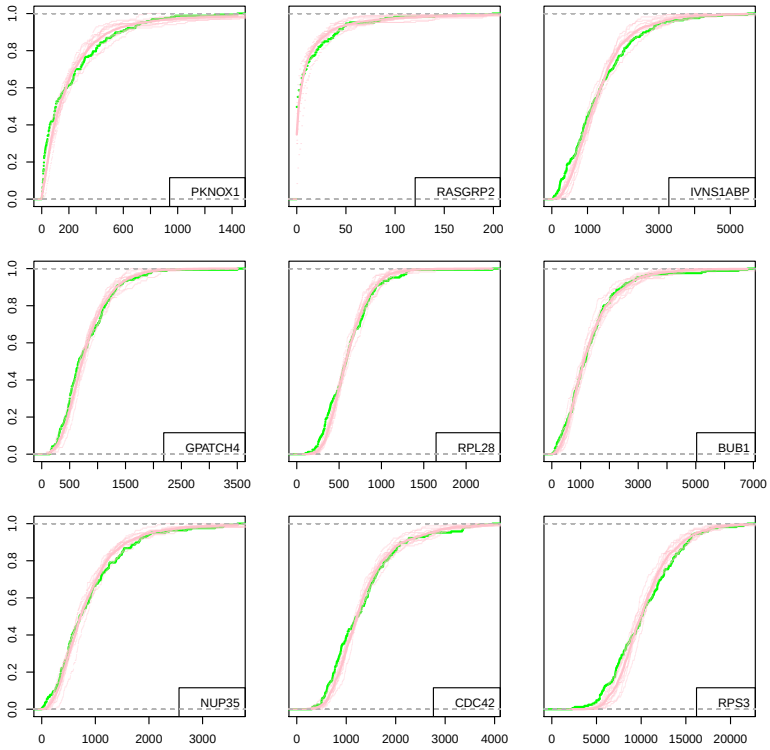

Fig 10: Empirical CDF for observed data (green) vs. empirical CDF for counts simulated from the fitted mixture NB (pink) for data set GSE64016. The top 6 panels are genes randomly selected from those genes being identified as DD by `scDDboost` and not violating constant shape assumption; the last three genes are randomly sampled from genes that fail the test of constant NB shape parameter. Each think pink curve is from one of the randomized distances; the the thicker pink curve represents pointwise averaging over 10 randomized distance matrices. Note horizontal scales differ among the panels. Cells in both conditions are pooled.

on the clustering checks

3.5.1. *clustering method.* We use distance based clustering method, where two distance matrices are combined. One is based on gene level modal clustering, which provides robust estimations when outliers are presented. The other is correlation based, which has been demonstrate to outperform the Euclidean distance(Kim et al. (2018b)). After the distance is obtained,

we perform K-medoids for clustering, where the sizes of clusters are balanced and unlikely to have a cluster with a lot more cells than others.

Our results are averaged over randomly generated clusterings and thus is not sensitive to the choice of the input clustering method. Further, we redeployed `scDDboost` in examples (data sets EBTAB2805, GSE45719) but using clustering method `sc3` in place of the default method. Let p_o be the proportions of commonly identified genes among the top 1000 and n_{DD} be the number of DD genes. Further inference based on default clustering method tends to give conservative result and have a lot in common with `sc3`. We also run `scDDboost` with `sc3` on null case data sets used

Data Set	cor	p_o	n_{DD} default	n_{DD} by sc3	commonly identified
EMTAB2805	0.93	80%	3334	3725	2913
GSE45719	0.94	75%	4805	7591	4851

TABLE 1
similar local FDR under two clustering methods

in Figure 7 and simulation data. We observe very few positives and conclude that our method does not inflate type-I error with different clustering method.(supplementary 3.5.3)

4. Asymptotics of the Double Dirichlet Mixture. Summary statistics $P(A_\pi|y,z)$, from Section 2.3, are amenable to a first-order asymptotic analysis that provides further insight into DDM model behavior. The fact that support sets A_π for component distributions $p_\pi(\phi,\psi)$ are not disjoint becomes an important issue. Consider distinct partitions π_1 and π_2 of subtypes $\{1,2,\cdots,K\}$, and recall that $N(\pi)$ counts the number of blocks in partition π. In case π_2 refines π_1, then $N(\pi_1) < N(\pi_2)$, and we also know that $A_{\pi_2} \subset A_{\pi_1}$, since refinement imposes additional constraints on the pair (ϕ,ψ) of probability vectors. If the data-generating state $(\phi,\psi) \in A_{\pi_2}$, one might ask how posterior probability mass tends to be allocated among the other mixture components whose support sets also contain this state. The question is addressed by the following:

THEOREM 4. *Let π_1 and π_2 denote two partitions for which $N(\pi_1) < N(\pi_2)$ and $A_{\pi_1} \cap A_{\pi_2}$ is non-empty. Let $(\phi,\psi) \in A_{\pi_1} \cap A_{\pi_2}$ denote the data generating state for subtype labels $z_1, z_2, \cdots, z_n$ given i.i.d. Bernoulli condition labels $y_1, y_2, \cdots, y_n$, and recall the posterior mixing proportions ω_π^{post} from equation (9)*

with hyper-parameters $\alpha_i^j \geq 1$ for $i = 1, \cdots, K$, $j = 1, 2$. *Then*

$$\frac{\omega_{\pi_1}^{\text{post}}}{\omega_{\pi_2}^{\text{post}}} \longrightarrow_{a.s.} 0 \qquad \text{as } n \longrightarrow \infty.$$

Essentially, mixing mass is transferred to components associated with the most refined partition consistent with a given parameter state. To be precise, let $H(\phi, \psi) = \{\pi : (\phi, \psi) \in A_\pi\}$ record all the partitions associated with one state. Typically, there is a most refined partition, $\pi^* = \pi^*(\phi, \psi)$, such that

$$A_{\pi^*} = \bigcap_{\pi \in H(\phi,\psi)} A_\pi. \tag{12}$$

This always happens when $K \leq 3$. In Supplementary Material Section 4 we characterize the exceptional set of states where (12) does not hold. Notably, if (12) does hold for state (ϕ, ψ), then for any $\pi \in H(\phi, \psi)$, using Theorem 4 and (10), we have

$$P(A_\pi | y_1, \cdots, y_n; z_1, \cdots z_n) \longrightarrow_{a.s.} 1 \quad \text{as } n \longrightarrow \infty.$$

This provides conditions under which we expect good performance for large numbers of cells.

5. Concluding remarks. We have presented `scDDboost`, a tool for detecting differentially distributed genes from scRNA-seq data, where transcripts are modeled as a mixture of cellular subtypes. The methodology links established model-based techniques with novel empirical Bayesian modeling and computational elements to provide a powerful detection method showing comparatively good operating characteristics in simulation, empirical, and asymptotic studies.

In the software and numerical experiments we made specific choices, such as to use mixtures of negative binomial components per gene, and to use K-medoids clustering on particular cell-cell distances. These choices have evident advantages, but the model structure and theory developed in Section 2 carry through for other cases. Future experiments could study other formulations within the same schema; for example there may be cell-cell distances that better capture the intrinsic dimensionality of expression programs, including, perhaps distances based on diffusions (Haghverdi, Buettner and Theis (2015)) or the longest-leg path distance (Little, Maggioni and Murphy (2017)). Future experiments could also further assess

operating characteristics when the number of cells is very large and the number of reads is relatively small, as may arise with unique molecular identifiers (Chen et al. (2018)). Further, assuming a compositional structure to drive model-based computations may not be restrictive, since it allows great flexibility in the form of each gene/condition-specific expression distribution (as coded, they are finite mixtures of negative binomials).

`EBSeq` currently presents a computational bottleneck for `scDDboost`, since it searches all partitions of K and encodes a hyper-parameter estimation algorithm that scales poorly with K. Several approximations present themselves that may redress the problem, since, in the mixture model context, only patterns π corresponding to relatively probable expression-change patterns over subtypes have a big impact on the final posterior inference. Even resolving this bottleneck there are advantages to having K small compared to n. Numerical experiments show increased false discoveries when K is over-estimated. But accurate estimation with large K would not be expected to provide much improved power, since that depends on accurate estimation of subtypes and their frequencies which relies on K being relatively small compared to n.

Acknowledgement. This research was supported in part by US National Institutes of Health grants P50 DE026787, P30CA14520-45, R01 GM102756, U54AI117924, and US National Science Foundation grant 1740707.

References.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology* **11** R106–R106.

Bacher, R. and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology* **17** 63. .

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C. and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology* **33** 155 EP -.

Chen, W., Li, Y., Easton, J., Finkelstein, D., Wu, G. and Chen, X. (2018). UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biology* **19** 70. .

Chu, L.-F., Leng, N., Zhang, J., Hou, Z., Mamott, D., Vereide, D. T., Choi, J., Kendziorski, C., Stewart, R. and Thomson, J. A. (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biology* **17** 173. .

Dahl, D. B. (2009). Modal clustering in a class of product partition models. *Bayesian Anal.* **4** 243–264.

Darmanis, S., Sloan, S. A., Croote, D., Mignardi, M., Chernikova, S., Samghababi, P., Zhang, Y., Neff, N., Kowarsky, M., Caneda, C., Li, G., Chang, S. D., Connolly, I. D.,

Li, Y., Barres, B. A., Gephart, M. H. and Quake, S. R. (2017). Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. *Cell reports* **21** 1399–1410.

Delmans, M. and Hemberg, M. (2016). Discrete distributional differential expression (D3E) - a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics* **17** 110. .

Deng, Q., Ramsköld, D., Reinius, B. and Sandberg, R. (2014). Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. *Science* **343** 193–196.

Dominguez, D., Tsai, Y.-H., Gomez, N., Jha, D. K., Davis, I. and Wang, Z. (2016). A high-resolution transcriptome map of cell cycle reveals novel connections between periodic genes and cancer. *Cell Research* **26** 946 EP -.

Efron, B. (2007). Size, power and false discovery rates. *Ann. Statist.* **35** 1351–1377.

Engel, I., Seumois, G., Chavez, L., Samaniego-Castruita, D., White, B., Chawla, A., Mock, D., Vijayanand, P. and Kronenberg, M. (2016). Innate-like functions of natural killer T cell subsets result from highly divergent gene programs. *Nature Immunology* **17** 728 EP -.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., Linsley, P. S. and Gottardo, R. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16** 278. .

Guo, F., Yan, L., Guo, H., Li, L., Hu, B., Zhao, Y., Yong, J., Hu, Y., Wang, X., Wei, Y., Wang, W., Li, R., Yan, J., Zhi, X., Zhang, Y., Jin, H., Zhang, W., Hou, Y., Zhu, P., Li, J., Zhang, L., Liu, S., Ren, Y., Zhu, X., Wen, L., Gao, Y. Q., Tang, F. and Qiao, J. (2015). The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells. *Cell* **161** 1437–1452.

Haghverdi, L., Buettner, F. and Theis, F. J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31** 2989–2998.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M. and Zhang, N. R. (2018). SAVER: gene expression recovery for single-cell RNA sequencing. *Nature Methods* **15** 539–542. .

Jensen, S. T., Erkan, I., Arnardottir, E. S. and Small, D. S. (2009). Bayesian testing of many hypotheses x many genes: A study of sleep apnea. *Ann. Appl. Stat.* **3** 1080–1101.

Kaufman, L. and Rousseeuw, P. (1987). *Clustering by means of medoids*. North-Holland.

Kendziorski, C. M., Newton, M. A., Lan, H. and Gould, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22** 3899–3914.

Kim, T., Chen, I. R., Lin, Y., Wang, A. Y.-Y., Yang, J. Y. H. and Yang, P. (2018a). Impact of similarity metrics on single-cell RNA-seq data clustering. *Briefings in Bioinformatics* bby076.

Kim, T., Chen, I. R., Lin, Y., Wang, A. Y.-Y., Yang, J. Y. H. and Yang, P. (2018b). Impact of similarity metrics on single-cell RNA-seq data clustering. *Briefings in Bioinformatics* **20** 2316-2326.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R. and Hemberg, M. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods* **14** 483 EP -.

Korthauer, K. D., Chu, L.-F., Newton, M. A., Li, Y., Thomson, J., Stewart, R. and Kendziorski, C. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology* **17** 222. .

Lane, K., Van Valen, D., DeFelice, M. M., Macklin, D. N., Kudo, T., Jaimovich, A.,

Carr, A., Meyer, T., Pe'er, D., Boutet, S. C. and Covert, M. W. (2017). Measuring Signaling and RNA-Seq in the Same Cell Links Gene Expression to Dynamic Patterns of NF-B Activation. *Cell Systems* **4** 458–469.e5.

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M. G., Haag, J. D., Gould, M. N., Stewart, R. M. and Kendziorski, C. (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29** 1035-1043.

Leng, N., Chu, L.-F., Barry, C., Li, Y., Choi, J., Li, X., Jiang, P., Stewart, R. M., Thomson, J. A. and Kendziorski, C. (2015). Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nature Methods* **12** 947 EP -.

Li, F. and Altieri, D. C. (1999). The Cancer Antiapoptosis Mouse Survivin Gene. *Cancer Research* **59** 3143.

Lin, P., Troup, M. and Ho, J. W. K. (2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biology* **18** 59. .

Little, A. F., Maggioni, M. and Murphy, J. M. (2017). Path-Based Spectral Clustering: Guarantees, Robustness to Outliers, and Fast Algorithms.

Love, M. I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15** 550. .

Marioni, J. C. and Arendt, D. (2017). How Single-Cell Genomics Is Changing Evolutionary and Developmental Biology. *Annual Review of Cell and Developmental Biology* **33** 537-553. PMID: 28813177.

McDavid, A., Dennis, L., Danaher, P., Finak, G., Krouse, M., Wang, A., Webster, P., Beechem, J. and Gottardo, R. (2014). Modeling Bi-modality Improves Characterization of Cell Cycle on Gene Expression in Single Cells. *PLOS Computational Biology* **10** e1003696–.

Muralidharan, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *Ann. Appl. Stat.* **4** 422–438.

Navin, N. E. (2015). The first five years of single-cell cancer genomics and beyond. *Genome Research* **25** 1499–1507.

Nawy, T. (2013). Single-cell sequencing. *Nature Methods* **11** 18 EP -.

Newton, M. A., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–176.

Papalexi, E. and Satija, R. (2017). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology* **18** 35 EP -.

Peccoud, J. and Ycart, B. (1995). Markovian Modeling of Gene-Product Synthesis. *Theoretical Population Biology* **48** 222–234.

Pierson, E. and Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology* **16** 241. .

Raj, A. and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135** 216–226.

Ray, S. and Turi, R. H. (2000). Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation.

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T., Yosef, N., Schwartz, S., Fowler, B., Weaver, S., Wang, J., Wang, X., Ding, R., Raychowdhury, R., Friedman, N., Hacohen, N., Park, H., May, A. P. and Regev, A. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510** 363 EP -.

Sohr, S. and Engeland, K. (2008). RHAMM is differentially expressed in the cell cycle and downregulated by the tumor suppressor p53. *Cell Cycle* **7** 3448–3460.

Soneson, C. and Robinson, M. D. (2017). Bias, Robustness And Scalability In Differential Expression Analysis Of Single-Cell RNA-Seq Data. *bioRxiv*.

Strehl, A. and Ghosh, J. (2003). Cluster Ensembles — a Knowledge Reuse Framework for Combining Multiple Partitions. *J. Mach. Learn. Res.* **3** 583–617.

Tasic, B., Menon, V., Nguyen, T. N., Kim, T. K., Jarsky, T., Yao, Z., Levi, B., Gray, L. T., Sorensen, S. A., Dolbeare, T., Bertagnolli, D., Goldy, J., Shapovalova, N., Parry, S., Lee, C., Smith, K., Bernard, A., Madisen, L., Sunkin, S. M., Hawrylycz, M., Koch, C. and Zeng, H. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience* **19** 335 EP -.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32** 381–386.

Wagner, U. and Taudes, A. (1986). A Multivariate Polya Model of Brand Choice and Purchase Incidence. *Marketing Science* **5** 219–244.

Yakowitz, S. J. and Spragins, J. D. (1968). On the Identifiability of Finite Mixtures. **39** 209–214.

Yin, G. and Ma, Y. (2013). Pearson-type goodness-of-fit test with bootstrap maximum likelihood estimation. *Electronic journal of statistics* **7** 412–427.

Zappia, L., Phipson, B. and Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. *Genome Biology* **18** 174. .

Appendix.

Proof of Theorem 1. If $\theta \in \bigcup_{\pi\in\Pi}\left[A_\pi \cap M_{g,\pi}\right]$, then there exists a partition π for which $\theta \in A_\pi$ and $\theta \in M_{g,\pi}$. By construction

$$f_g^1(x) = \sum_{k=1}^{K} \phi_k f_{g,k}(x) = \sum_{b\in\pi}\sum_{k\in b} \phi_k f_{g,k}(x) = \sum_{b\in\pi} \Phi_b f_{g,k^*(b)}(x),$$

where $k^*(b)$ indexes any component in b, since all components in that block have the same component distribution owing to constraint $M_{g,\pi}$. Continuing, using the constraint $\theta \in A_\pi$,

$$f_g^1(x) = \sum_{b\in\pi} \Psi_b f_{g,k^*(b)}(x) = f_g^2(x) \qquad \forall x.$$

That is, $\theta \in \mathrm{ED}_g$.

If $\theta \in \mathrm{ED}_g$, then $f_g^1(x) = f_g^2(x)$ for all x. Noting that both are mixtures over the same set of components $\{f_{g,k}\}$, let $\{h_{g,l} : l = 1, 2, \cdots, L\}$ be the set of distinct components over this set, and so

$$f_g^1(x) = \sum_{k=1}^{k} \phi_k f_{g,k}(x) = \sum_{l=1}^{L} c_{g,l}(\phi) h_{g,l}(x) = \sum_{l=1}^{L} c_{g,l}(\psi) h_{g,l}(x) = f_g^2(x)$$

where

$$c_{g,l}(\phi) = \sum_{k=1}^{K} \phi_k 1[f_{g,k} = h_{g,l}] \qquad c_{g,l}(\psi) = \sum_{k=1}^{K} \psi_k 1[f_{g,k} = h_{g,l}]. \tag{13}$$

Finite mixtures of distinct negative binomial components are identifiable (Proposition 5 from Yakowitz and Spragins (1968)), and so the equality of f_g^1 and f_g^2 implies $c_{g,l}(\phi) = c_{g,l}(\psi)$ for all $l = 1, 2, \cdots, L$. Identifying the partition blocks $b_l = \{k : f_{g,k} = h_{g,l}\}$, and the partition $\tilde{\pi} = \{b_l\}$, we find $\theta \in A_{\tilde{\pi}} \cap M_{g,\tilde{\pi}}$. The accumulated probabilities in (13) correspond to $\Phi_{\tilde{\pi}}$ and $\Psi_{\tilde{\pi}}$, which are equal on $A_{\tilde{\pi}}$.

Randomizing distances for approximate posterior inference. One way to frame the subtype problem is to suppose that subtype labels $z = (z_i)$ satisfy $z = f(\Delta)$, where $\Delta = \left(\delta_{i,j}\right)$ is a $n \times n$ matrix holding *true*, unobservable distances, such as $\delta_{i,j}$ between cells i and j, and that f is some assignment function, like the one induced by the $K-$medoids algorithm. Then posterior uncertainty in z would follow directly from posterior uncertainty

in Δ. On one hand, we could proceed via formal Bayesian analysis, say under a simple conjugate prior in which $1/\delta_{i,j} \sim \text{Gamma}(a_0, d_0)$, for hyperparameters a_0 and d_0, and in which the observed distance $d_{i,j}|\delta_{i,j} \sim \text{Gamma}(a_1, a_1/\delta_{i,j})$. This would assure that $\delta_{i,j}$ is the expectation of $d_{i,j}$, with shape parameter a_1 affecting variation of measured distances about their expected values. Not accounting for any constraints imposed by both D and Δ being distance matrices, we would have the posterior distribution $1/\delta_{i,j}|D \sim \text{Gamma}(a_0 + a_1, d_0 + a_1 d_{i,j})$. For any threshold $c > 0$, we would find

$$P\left(\delta_{i,j} \leq c|D\right) = P\left(U \geq \frac{d_0 + a_1 d_{i,j}}{c(a_0 + a_1)}\right) \tag{14}$$

where $U \sim \text{Gamma}(a_0 + a_1, a_0 + a_1)$

Alternatively, we could form randomized distances $d^*_{i,j} = d_{i,j}/w_{i,j}$ where $w_{i,j}$ is the analyst-supplied random weight distributed as $\text{Gamma}(\hat{a}, \hat{a})$ as in Section 2.2. Notice that

$$P(d^*_{i,j} \leq c|D) = P(w_{i,j} > d_{i,j}/c|D)$$

which is also an upper tail probability for a unit-mean Gamma deviate with shape and rate equal to $\hat{a}$. Comparing to (14), by setting $\hat{a}$ to equal $a_0 + a_1$, and if a_0 and d_0 are relatively small, we find

$$P(d^*_{i,j} \leq c|D) \approx P(\delta_{i,j} \leq c|D).$$

In other words, the randomized distance procedure is providing approximate posterior draws of the underlying distance matrix. In spite of limitations of this procedure for full Bayesian inference, it provides an elementary scheme to account for uncertainty in subtype allocations. Numerical experiments in Supplementary Material make comparisons to a full, Dirichlet-process-based, posterior analysis.

Algorithm 2 scDDboost

Input:

genes by cells expression data matrix $X = (X_{g,c})$

cell condition labels $y = (y_c)$

number of cell subtypes K

number of randomized clusterings n_r

Output: posterior probabilities of differential distribution

procedure scDDboost(X, y, K, n_r)

distance matrix: $D = \text{dist(X)} \leftarrow$ pairwise distances between cells (columns of X)

hyper-parameters $(a_0, a_1, d_0) \leftarrow \text{hyper}(D)$. Set $\hat{a} = a_0 + a_1$.

repeat

Gamma noise vector: e, with components $\sim \text{Gamma}(\hat{a}/2, \hat{a})$

randomized distance matrix: $D^* \leftarrow D/(e\mathbf{1}^T + \mathbf{1}e^T)$

$\hat{z}^* \leftarrow K-\text{medoids}(D^*)$

$P^* \leftarrow$ scDDboost-core$(X, y, \hat{z}^*)$

until n_r randomized distance matrices

return $\forall$genes g, $P(\text{DD}_g|X, y) = \frac{1}{n_r}\sum_{D^*} P^*_g$

end procedure

Pseudo-code.

Empirical datasets.

Data set	Conditions	# cells	Organism	Ref
GSE94383	0 min unstim vs 75min stim	186,145	human	Lane et al. (2017)
GSE48968-GPL13112	BMDC (2h LPS stimulation) vs 6h LPS	96,96	mouse	Shalek et al. (2014)
GSE52529	T0 vs T72	69,74	human	Trapnell et al. (2014)
GSE74596	NKT1 vs NTK2	46,68	mouse	Engel et al. (2016)
EMTAB2805	G1 vs G2M	96,96	mouse	Buettner et al. (2015)
GSE71585-GPL13112	Gad2tdTpositive vs Cux2tdTnegative	80,140	mouse	Tasic et al. (2016)
GSE64016	G1 vs G2	91,76	human	Leng et al. (2015)
GSE79102	patient1 vs patient2	51, 89	human	Kiselev et al. (2017)
GSE45719	16-cell stage blastomere vs mid blastocyst cell	50, 60	mouse	Deng et al. (2014)
GSE63818	Primordial Germ Cells, develop- mental stage: 7 week gestation vs Somatic Cells, developmental stage: 7 week gestation	40,26	mouse	Guo et al. (2015)
GSE75748	DEC vs EC	64, 64	human	Chu et al. (2016)
GSE84465	neoplastic cells vs non-neoplastic cells	1000, 1000	human	Darmanis et al. (2017)

Appendix Table A2

Data sets used for the empirical study of scDDboost

Data set	Condition	# cells
GSE63818null	7 week gestation	40
GSE75748null	DEC	64
GSE94383null	T0	186
GSE48968-GPL13112null	BMDC (2h LPS stimulation)	96
GSE74596null	NKT1	46
EMTAB2805null	G1	96
GSE71585-GPL13112null	Gad2tdTpositive	80
GSE64016null	G1	91
GSE79102null	patient1	51

APPENDIX TABLE A3

Single-condition data sets used in the random-splitting experiment.

Xiuyu Ma, Michael A. Newton
Department of Statistics
University of Wisconsin Madison
1300 University Ave
Madison, WI 53706
E-mail: ma79@wisc.edu; newton@stat.wisc.edu

Keegan Korthauer
Department of Data Sciences
Dana-Farber Cancer Institute
CLSB 11007 — 450 Brookline Ave
Boston, Massachusetts 02215
E-mail: keegan@jimmy.harvard.edu

Christina Kendziorski
Department of Biostatistics and Medical Informatics
University of Wisconsin Madison
425G Henry Mall
Madison, WI 53706
E-mail: kendzior@biostat.wisc.edu