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

XIUYU MA, KEEGAN KORTHAUER, CHRISTINA KENDZIORSKI, AND MICHAEL A. NEWTON

Abstract. On the problem of scoring genes for evidence of changes in the distribution of single-cell expression, we introduce a 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 [Nawy, 2013]. 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. 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 discovery 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. Lots of efforts have been made to clustering cells into different cell subtypes,

DEPARTMENT OF BIOSTATISTICS AND MEDICAL INFORMATICS, UW MADISON, TECHNICAL REPORT TR***-V1, FEBRUARY **, 2019.

SC3[Kiselev et al., 2017], CIDR[Lin et al., 2017] and ZIFA[Pierson and Yau, 2015]. We hypothesize that such subtype information may be usefully injected into various inference procedures in order to improve their operating characteristics.

Assessing the magnitude and statistical significance of changes in gene expression associated with different cellular conditions has been a central statistical problem in genomics for which new tools specific to the single-cell RNAseq data structure have been deployed: MAST[Finak et al., 2015], DESEQ2[Love et al., 2014], SCDD[Korthauer et al., 2016], etc. These tools respond to scRNAseq characteristics, such as high prevelance of zero counts and gene-level multimodality, but none takes explicit advantage of cellular subtype information. We present a methodology and supporting theoretical analyses for this purpose. Using a compositional model, subtypes inferred by clustering whole genome data 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 experimental 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. A software implementation of the proposed methodology is available in the R package scDDboost, at `http://github.com/wiscstatman/scDDboost/`. Modularity in the necessary elements provides some methodological advantages. For example, improvements in clustering, such as *sc3?*, may be used in place of the default clustering, without altering the form of downstream analysis. **it's also relatively speedy computationally, not requiring any MCMC**

Numerical experiments on both synthetic and published scRNA-seq data indicate that `scDDboost` has high sensitivity for detecting subtle distribution changes. In these experiments we take advantage of `splatter` for generating synthetic data [Zappia et al., 2017] as well as the compendium of scRNA-seq data available through `conquer` [Soneson and Robinson, 2017]. **something on bursting?** We also establish first-order asymptotic results for the methodology. On organiziation, we present the modeling and methodology elements in Section 2, numerical experiments in Section 3, and a discussion in Section 4. For presentation we move some details to an appendix and many others to a Supplementary Material document, which for ease of reference in the present version is contained as the final part of a single *pdf* file holding the entire manuscript.

2. Modeling

2.1. **Data structure, sampling model, and parameters.** In modeling scRNASeq 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 happens to be 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; we infer something about these counts using genome-wide data. As for molecular data, the normalized expression of gene g in cell c, say $X_{g,c}$, 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 reckon 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 the cell finds itself in. 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 considering a broad class of within-subtype expression distributions. We revisit the issue in Section 4, but for now we proceed assuming that cellular condition affects the composition of subtypes but not the 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*, denote DD_g and indicated $f_g^1 \neq f_g^2$, 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 RNAseq data analysis, we further set each $f_{g,k}$ to have a a negative-binomial form, say with mean $\mu_{g,k}$ and shape parameter α_g ([Leng et al., 2013];[Anders and Huber, 2010];[Love et al., 2014]). This choice proves to be 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.* 2016; Huang *et al.* 2018). Our perspective is that genome-wide data may usefully inform the mixing proportions.

We seek methodology to prioritize genes for evidence of DD_g. Interestingly, 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 RNAseq. 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 $\mu_{g,1} = \mu_{g,2}$ then, other things being equal, $f_g^1 = f_g^2$ even though $\phi \neq \psi$. 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, say π, is a set of mutually exclusive and exhaustive blocks, b, say, each 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 1).

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

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

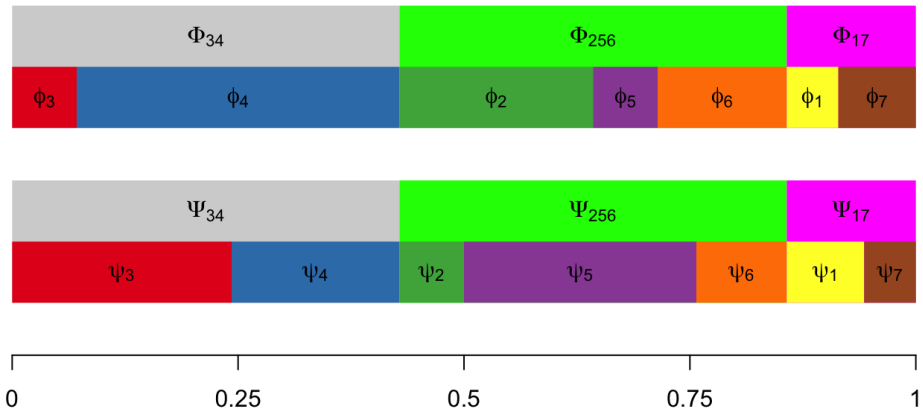

FIGURE 1. Proportions of $K = 7$ cellular subtypes in different conditions. Aggregated proportions of subtypes 3 and 4, subtypes 2, 5, and 6, and subtypes 1, and 7 remain same across conditions, while individual subtype frequencies change.

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 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

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

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 distinct partitions π_1, π_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:

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

Such local false discovery rates are important empirical Bayesian statistics in large-scale testing (e.g., Efron, 2007; Muralidharan, 2010; Newton *et al.* 2004). The partition representation guides construction of a prior distribution (Section 3.1) and a model-based method (Section 2.3) for scoring differential distribution. Setting the stage, Figure 2 shows the dependency structure of the proposed compositional model and the partition-reliant prior specification.

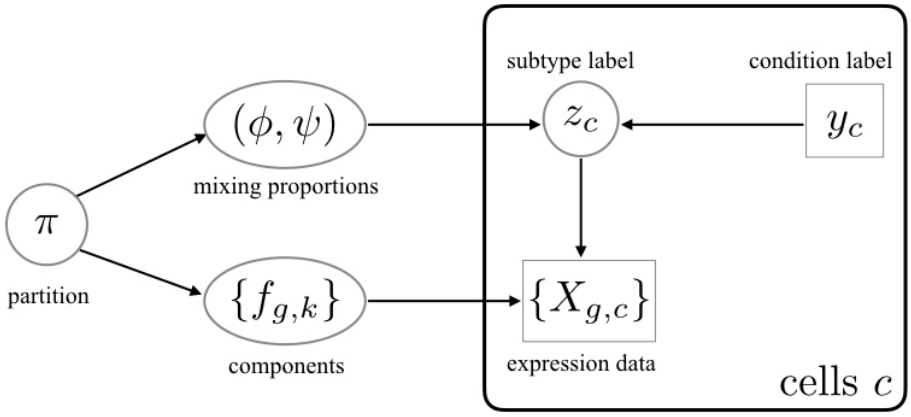

FIGURE 2. Directed acyclic graph structure of compositional model and partition-reliant prior. The plate on the right side indicates i.i.d. copies over cells c, conditionally on mixing proportions and mixing components. Observed data are indicated in rectangles/squares, and unobserved variables are in circles/ovals.

Key to computing the gene-specific local false discovery rate $P(\mathrm{ED}_g|X, y)$ is evaluating probabilities $P\left(A_\pi \cap M_{g,\pi}|X, y\right)$ for any subtype partition π and gene g. The dependence structure (Figure 2) implies a useful reduction of this quantity, at least conditionally upon subtype labels $z = (z_c)$.

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. The modest bit of uncertainty in z we handle through a computationally efficient randomized clustering scheme. Theorem 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 $P(M_{g,\pi}|X,z)$ 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.** We leverage the extensive research on how to cluster cells into subtypes using scRNA-seq data (e.g., SC3[Kiselev et al., 2017], CIDR[Lin et al., 2017], and ZIFA[Pierson and Yau, 2015]). We propose 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]. For completeness we review in Appendix A the empirical Bayes model underlying EBSeq. 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.

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

1: **procedure** scDDBoost-core($X, y, \hat{z}$)
2: number of cell subtypes $K = \text{length}(\text{unique}(\hat{z}))$
3: subtype differential expression: $\forall g, \pi$ compute $P(M_{g,\pi}|X,\hat{z})$ using EBSeqLeng et al. [2013]
4: cell frequency changes: $\forall \pi$ compute $P(A_\pi|y,\hat{z})$ using Double Dirichlet model
5: posterior probability: $\forall g$, $P(\text{ED}_g|X,y,\hat{z}) \leftarrow \sum_\pi P(M_{g,\pi}|X,\hat{z})\, P(A_\pi|y,\hat{z})$
6: **return** $\forall g$, $P(\text{DD}_g|X,y,\hat{z}) = 1 - P(\text{ED}_g|X,y,\hat{z})$

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 Appendix B) partitions the cells, and then we define dissimilarity between cells as the proportion of genes at which the cells are assigned to different clusters. 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., 2018]). We combine the two measures by a weighted average, with $w_C = \frac{\sigma_C}{\sigma_C+\sigma_P}$ and $w_P = 1 - w_C$. where $w_C, \sigma_C, w_P, \sigma_P$ are the weights and standard deviations of cluster based distance and Pearson correlation distance accordingly. The resulting distance matrix is $D = (d_{i,j})$.

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 $D^* = (d_{i,j} \times w_{i,j})$, where $w_{i,j}$ are unit mean, non-negative weights $w_{i,j} = 1/(e_i + e_j)$, and where (e_i) are independent and identically Gamma$(\hat{a}, \hat{b})$ distributed deviates for hyper-parameters $(\hat{a}, \hat{b})$ derived from D. We argue (Appendix C) that the distribution of clusterings induced by this simple scheme approximates a Bayesian posterior analysis. Pseudo-code for the resulting scDDboost is in Algorithm 2.

In order to determine the number of clusters, we consider the change of $validity = \frac{\textbf{intra}}{\textbf{inter}}$ defined in Ray and Turi [2000], where $\textbf{intra} = \frac{1}{N}\sum_{i=1}^{K}\sum_{x\in C_i}||x - z_i||^2$, $\textbf{inter} = mean(||z_i - z_j||^2), i = 1, 2, ..., K-1. j = i+1, ..., K$ and z_i is the center(medoids) of cluster i. **intra** is the average of distance of a point to its corresponding cluster center, which measures the compactness of clusters. We made a small change here, in original paper **inter** was defined as minimum distance between medoids, we use average instead for the purpose of getting a smoother quantity. **inter** is the average distance of two cluster centers, which measures the separation between clusters. We want to have a small intra-cluster distance and a big inter-cluster distance, consequently we wnat to minimize the *validity*. From empirical study, we constantly observe a monotone decreasing relation between number of clusters and *validity*. However this quantity stabilize when K is sufficiently large. The stopping rule for searching K is when $|validity_K - min(validity_K)| < \epsilon$ is satisfied.

In simulations, we observed that averaged adjusted Rand index as well as Rand index of mode of partitions based on randomized distance matrices is higher (better estimation) than that of partition based on original distance matrix (Supplementary Material, **?**). *MAN to XM: please clarify*

2.3. **Double Dirichlet Mixture (DDM).** Here we describe the partition-reliant prior $p(\phi, \psi)$ indicated in Figure 2 and 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 any non-negative 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:

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

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:

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

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^1_π and t^2_π.

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

$$p_\pi(t^j | t^j_\pi, y) = \prod_{b \in \pi} \left\{ \left[\frac{\Gamma(t^j_b + 1)}{\prod_{k \in b} \Gamma(t^j_k + 1)} \right] \left[\frac{\Gamma(\sum_{k \in b} \alpha^j_k)}{\prod_{k \in b} \Gamma(\alpha^j_k)} \right] \left[\frac{\prod_{k \in b} \Gamma(\alpha^j_k + t^j_k)}{\Gamma(t^j_b + \sum_{k \in b} \alpha^j_k))} \right] \right\}$$

Property 5:

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

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 $t^j_b = n_j$ for both $j = 1, 2$, and Property 6 reduces, correctly, to $p_\pi(t^1_\pi, t^2_\pi | y) = 1$. Further,

$$p_\pi(t^j | t^j_\pi, y) = \left[\frac{\Gamma(n_j + 1)}{\Gamma(n_1 + \sum_{k=1}^K \alpha^j_k)} \right] \left[\frac{\Gamma(\sum_{k=1}^K \alpha^j_k)}{\prod_{k=1}^K \Gamma(\alpha^j_k)} \right] \left[\prod_{k=1}^K \frac{\Gamma(\alpha^j_k + t^j_k)}{\Gamma(t^j_k + 1)} \right]$$

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

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

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.

The properties above are useful in establishing:

Theorem 3. *The DDM model 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

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

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 $\forall b \in \pi^c$ there exists $s \subset \pi^r$ such that $\bigcup_{b' \in s} b' = b$. 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 8: For two partitions $\tilde{\pi}$ and π,

$$P_{\tilde{\pi}}\left(A_\pi | y, z\right) = \begin{cases} 1 & \text{if } \tilde{\pi} \text{ refines } \pi \\ 0 & \text{otherwise} \end{cases}$$

This supports the main finding of this section:

$$P(A_\pi | y, z) = \sum_{\tilde{\pi} \in \Pi} \omega_{\tilde{\pi}}^{\text{post}} \, 1[\tilde{\pi} \text{ refines } \pi]. \tag{9}$$

3. Numerical experiments

3.1. **Synthetic data.** A simulation study was conducted to assess the performance of scDDboost in identifying DD genes. We simulate data by splatter[Zappia et al., 2017] with approximate 200 cells each condition and 7 subtypes with proportions ϕ and ψ satisfying constraints: $\phi_1 + \phi_2 = \psi_1 + \psi_2$, $\phi_3 + \phi_4 + \phi_5 = \psi_3 + \psi_4 + \psi_5$ and $\phi_6 + \phi_7 = \psi_6 + \psi_7$. Each subtype has 10% genes to be differential expressed. We view the differences among subtypes by projecting transcripts profiles of cells into its first two principal components(Fig 2). We observed some subtypes are well separated, some subtypes are nested, which would make those simulation setting to be nontrivial for consideration.

See Supplementary Material Figure S1...

We determine the number of subtypes by searching a range of candidates(from 1 to 9 based on our empirical experience). Given number of subtypes, we obtain a subtype structure of cells, which will further be fed into computing the posterior probabilities. We visualize the change between posterior probabilities under number of clusters i and $i+1$(i from 1 to 8). It typically remains stable when number of cluster is above a number that is smaller than 9 (Fig 3) In the simulated data, the posterior probability become stable when we overestimate the number of subtypes. We found the true number of subtypes is 7 and correctly identify the subtypes of cells.

scDDboost identified most true DD genes, the reason is that mean expression shifts between conditions is not as significant as mean expression shifts between subtypes, which limits the power of MAST and DESeq2. Our approach and scDD considered mixture structure underlying the transcrips but scDD did not use the whole genome information to infer mixture components, which leads to inaccurate clustering at gene level and reduce the power. Under randomized distance, scDDboost gave an accurate estimation of subtypes and thus are more sensitive to the mean expression change among subtypes. We also compare roc curves of scDDboost, scDD, MAST and DESeq2. (Fig 4)

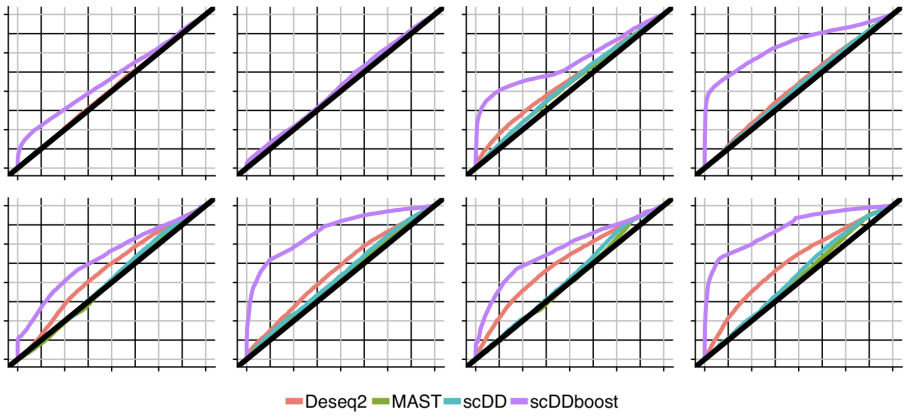

FIGURE 3. Roc curve of scDDboost, scDD, MAST and DESeq2, scDDboost has largest area under the roc curve. Roc curves of other three methods are similar. For those roc curve there is bigger difference at low level of false positive rate, as scDDboost identified twice many true DD genes as other methods. Roc curves are obtained under different settings of hyper parameters of simulations.

Since we are modeling gene transcript within each subtype as negative-binomially distributed and we only test one parameter(mean) change among subtypes. In some scenario, it could be insufficient to model the variability within subtype. Even though there is no mean expression change among subtypes but more subtle distributional change occurred among subtypes changed, EBSeq would fail to detect the discrepancies between subtypes, thus limit power of scDDboost.

3.2. **Empirical study.** We use 13 datasets from conquer[Soneson and Robinson, 2017] to test performance of our method on empirical data. We compare our results with scDD[Korthauer et al., 2016], MAST[Finak et al., 2015] and DESeq2[Love et al., 2014], we have also investigated performance of scDDboost under different clustering method, (sc3[Kiselev et al., 2017], supplimentary) and obtain similar
see Supplementary Material Table S1

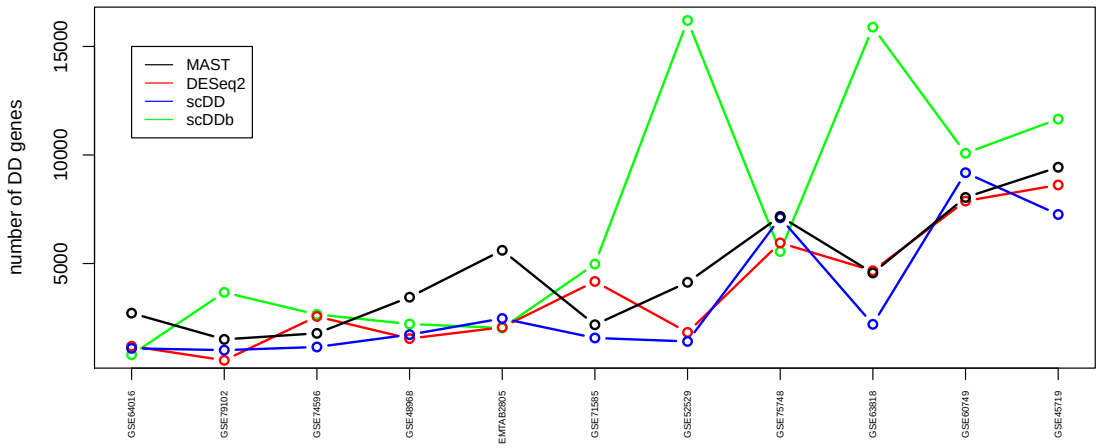

FIGURE 4. number of DD genes with respect to total number of genes identified by each method. Ranked by mean list size

We found that bulk method DESeq2 tends to have the most number of DE genes. But among single cell methods, scDDboost usually identified the most DD genes. Further we observed quite a few genes uniquely identified by scDDboost are likely to have different distribution across conditions. For example, Fig 5, we use violin plot to demonstrate the log expression profiles among DEC and EC.

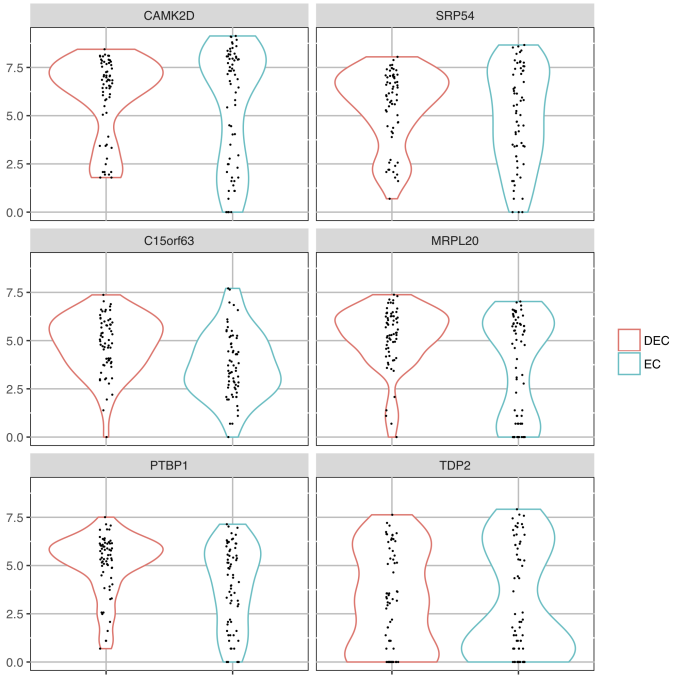

FIGURE 5. Densities of log transformed transcripts, 6 DD genes uniquely identified by scDDboost, for data GSE75748, DEC vs. EC, We observe some of the genes are different distributed across conditions.

3.3. **Empirical study: null cases.** Although bulk methods seems to be the most powerful one, we found it also has a higher false discovery rate comparing to single cell methods. We validate false discovery rate on ten null datasets from table 1. For each null dataset, we randomly split the cells from one condition into two subsets and test difference of gene expression between those subsets. Since the two subsets of cells actually came from same condition, there should not be any differential distributed genes, any positive call would be a false positive. We repeat the random split and testing for five times on each null data set. We evaluate the type I error control for the methods returning nominal p-values, by recording the fraction of genes(with a valid p-value) that are assigned a nomial p-value below 0.05 (Fig 6).
scDDboost could control FDR since we assume cells are sampled from population composed of different subtypes. Cells from one subtype are equal likely to be assigned to either one of the two subsets. Consequently, it is very likely that proportions of subtypes remain unchanged among the two subsets.

Data set	Conditions	Number of cells/condition	Organism
GSE57872null	patient1	96,96	human
GSE52529null	T0	48, 48	human
GSE48968-GPL13112null	BMDC (2h LPS stimulation)	48,48	mouse
GSE60749-GPL13112null	v6.5 mouse embryonic stem cells, culture conditions: 2i+LIF	45,45	mouse
GSE74596null	NKT1	23,23	mouse
EMTAB2805null	G1	48,48	mouse
GSE71585-GPL13112null	Gad2tdTpositive	40,40	mouse
GSE64016null	G1	46,45	human
GSE79102null	patient1	26, 25	human

TABLE 1. datasets used for null cases, as cells are coming from same biological condition, there should not be any differential distributed genes, any positive call is false positive

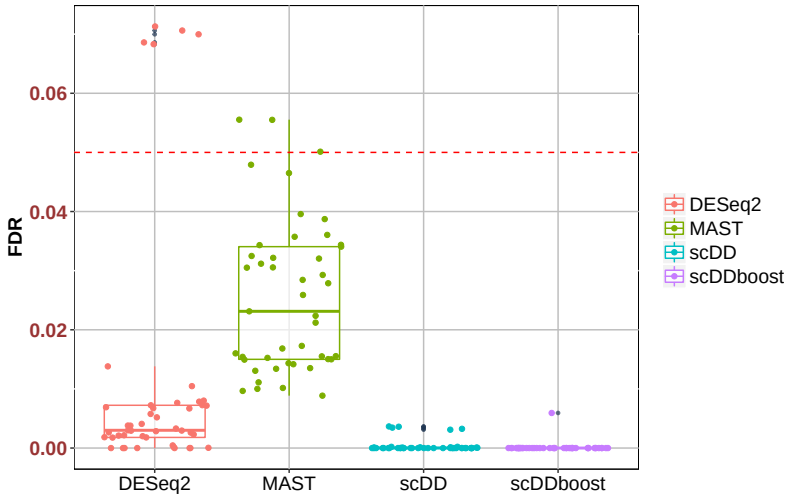

FIGURE 6. FDR of scDDboost, scDD, MAST and DESeq2 on null dataset from table 1, DESeq2 usually identify a lot but may lose the control of type I error. While other single cell methods could control FDR.

3.4. **Number of subtypes.** From our empirical experience, it is typical K will not be larger than 8. We demonstrate the change of posterior probabilities of differential distribution given different number of subtypes at data GSE75748 and GSE48968. In both cases, if allowing one more subtype would result in a lot increases in posterior probabilities, which

suggests that the number of subtypes is underestimated since we found more distribution differences between conditions given one more mixture component. If posterior inference is stable after increasing the number of subtypes, then we consider previous number of subtypes to be optimal.

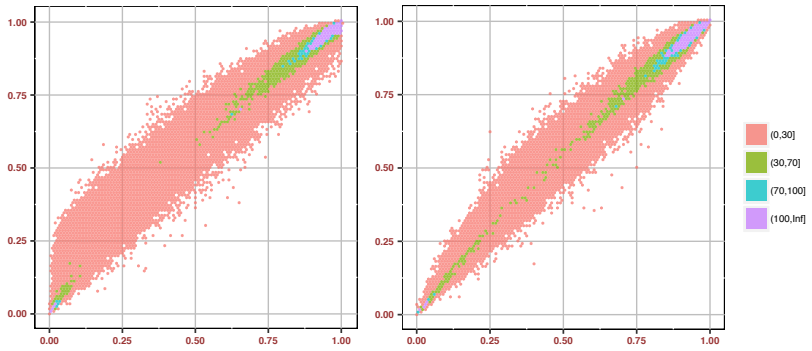

FIGURE 7. selecting number of subtypes for data GSE57872, we observe posterior probabilities become stable at more than 6 subtypes. Since increasing number of subtypes tends to decrease sample size of each subtypes, make complicate constraints for equivalent distribution and inflate estimated PDD. We select number of subtypes to be 7

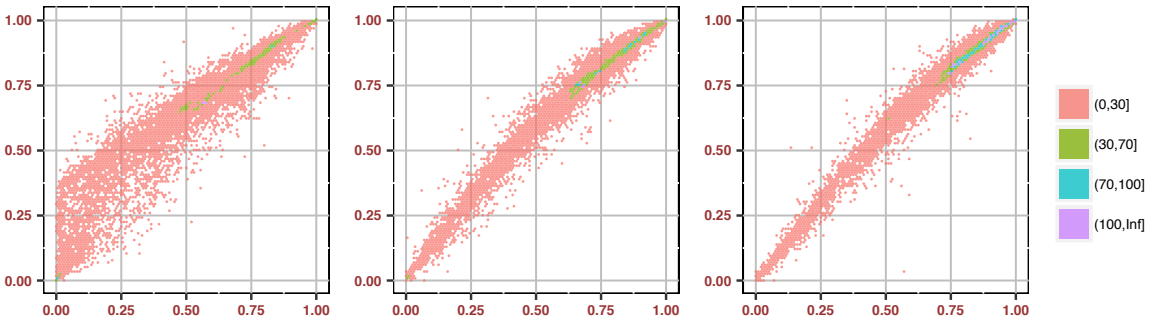

FIGURE 8. selecting number of subtypes for data GSE48968, we observe posterior probabilities become stable at more than 5 subtypes

3.5. **Bursting parameters.** ***on the method estimated p-value, update later***

D3E[Delmans and Hemberg, 2016] is a distributional method that can identify bursting parameters of transcripts. Rate of promoter activation, rate of promoter inactivation and the rate of transcription when the promoter is in the active state are estimated by D3E. We investigate DD genes identified by scDDboost and their change of those three parameters on dataset GSE71585

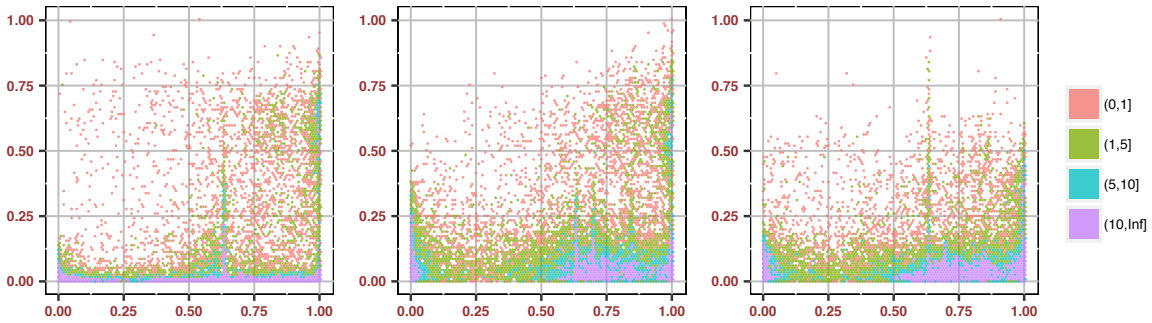

FIGURE 9. D3E method will estimate 3 bursting parameters probability of a gene being on (**a**) and off (**b**) and the expression rate when the gene expression is on (**c**), we plot the hexbin plot of probability of a gene being DD under out method v.s. the absolute value of log fold change of **a** , **b** and **c** across the two conditions accordingly. The log fold change is scaled by dividing the largest log fold change so that ends up in a value between 0 and 1 Here we use the GSE71585 data

We observed that DD genes identified by scDDboost tends to have similar transcription rate when the promoter is active across condition, while there are lots of variabilities in the action and inactivation rate. Estimations from D3E reveals that the major factor to drive DD genes are activation and inactivation rate (proportions of different subtyps), it make sense to consider mixture model like scDDboost.

4. Theoretical issues

4.1. **Posterior consistency.** Under some parameters settings, the double dirichlet prior will have limited resolution and lead to inconsistency of posterior probabilities, which we investigate with the following asymptotic analysis.

We first give the expression of posterior probability. Since there is no information favorable of any particular A_π, we select discrete uniform distribution as the prior for it, then the posterior probability is

$$p(A_\pi|t^1,t^2) = c * \sum_{\pi' \text{ refines } \pi} p(t^1|t^1_{\pi'})\, p(t^2|t^2_{\pi'})\, p(t^1_{\pi'}, t^2_{\pi'}|A_{\pi'}) \tag{10}$$

for a normalizing constant $\frac{1}{c} = \sum_{\pi' \in \Pi} p(t^1|t^1_{\pi'})\, p(t^2|t^2_{\pi'})\, p(t^1_{\pi'}, t^2_{\pi'}|A_{\pi'})$.

Let $\Omega = \{(\phi,\psi) : \sum_{i=1}^{K}\phi_i = \sum_{i=1}^{K}\psi_i = 1, \phi_i \geq 0, \psi_i \geq 0, i = 1,...,K\}$ be the whole space. There is a subset of Ω we lack posterior inference. Let us first see an example:
In Fig 10, there are four subtypes, the rectangle with magenta boundary is a simplex $A_{\pi_1} = \{(\phi,\psi) : \phi_1 + \phi_2 = \psi_1 + \psi_2\}$, the rectangle with blue boundary is a simplex $A_{\pi_2} =$

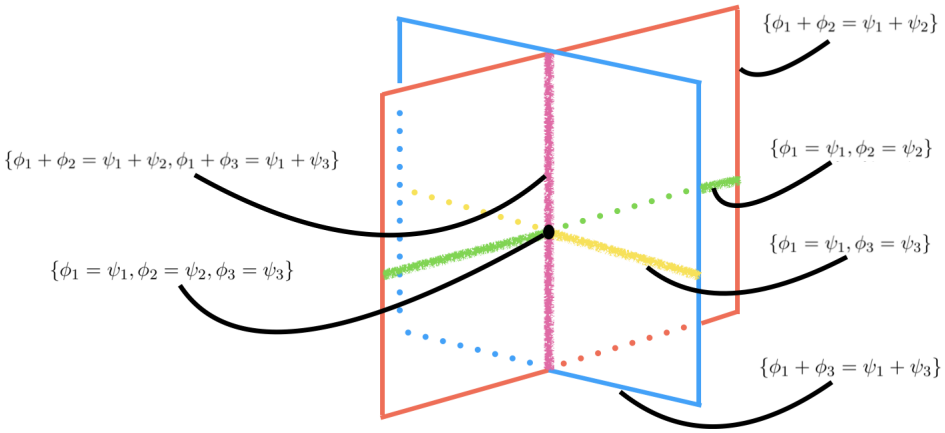

FIGURE 10. Four subtypes of cells, simplexes of (ϕ, ψ) satisfying different constraints.

$\{(\phi,\psi) : \phi_1+\phi_3=\psi_1+\psi_3\}$. The green line refers to $A_{\pi_3}=\{(\phi,\psi) : \phi_1=\psi_1, \phi_2=\psi_2\}$, the yellow line refers to $A_{\pi_4}=\{(\phi,\psi) : \phi_1=\psi_1, \phi_3=\psi_3\}$, the purple line refers to $A_{\pi_5}=\{(\phi,\psi) : \phi_1+\phi_2=\psi_1+\psi_2, \phi_1+\phi_3=\psi_1+\psi_3\}$, which is the intersection of A_{π_1} and A_{π_2}, and finally the black dot which is the intersection of those three lines refers to the simplex with finest partitions, $\phi_i=\psi_i, \forall i=1,..,4$. We lack posterior inference for (ϕ,ψ) along the purple line except the black dot. While on the green line, yellow line and black dot, we have consistent posterior inference(theorem 2). To explain why some space lacking posterior inference and define such space, we define a special subset A^*_π of simplex A_π. $A^*_\pi = A_\pi \setminus \bigcup_{\tilde{\pi} \text{ is not coarser than } \pi} A_{\tilde{\pi}}$, A^*_π is obtained by removing all intersection with other $A_{\tilde{\pi}}$(excluding those $A_{\tilde{\pi}}$ that is superset of A_π) from A_π. Since we removed those intersection parts. It is intuitive that A^*_π will be disjoint subsets of Ω.

Proposition 1. *if* $\pi_1 \neq \pi_2$, *then* $A^*_{\pi_1} \cap A^*_{\pi_2} = \emptyset$

Let $Q = \Omega \setminus \bigcup_{\pi\in\Pi} A^*_\pi$, and we have following proposition of the existence of Q.

Proposition 2. *Let* K *be number of subtypes. When* $K > 3, Q \neq \emptyset$, *when* $K \leq 3, Q = \emptyset$

When the number of subtypes is bigger than three, we lack posterior inference on Q. To

see that we can rewrite A_π^* as $A_\pi^* = A_\pi \setminus \bigcup_{\tilde{\pi} \text{ is not coarser than } \pi} (A_{\tilde{\pi}} \cap A_\pi)$, $\tilde{\pi}$ is not coarser than π, which is equivalently to say π is not refinement of $\tilde{\pi}$. By property 8 in section 2, $A_{\tilde{\pi}} \cap A_\pi$ is a lower dimensional subset of A_π. So $A_\pi \setminus A_\pi^*$ is a lower dimensional subset of A_π. For posterior on Q, it degenerates to integral on a lower dimensional subset of the simplex associating with densities, which will vanish

Proposition 3. *When $K > 3$, $p(Q|z^1, z^2) = 0$*

But for $(\phi, \psi) \in \Omega \setminus Q$, we have consistent posterior inference. Assuming $\alpha_i^j = 1, \forall i$ in (2), $j = 1, 2$ and $\beta_b = \Sigma_{i \in b}(\alpha_i^1 + \alpha_i^2)$ in (3), plug in (4) then we have simplified

$$p(A_\pi|t^1, t^2) = \frac{1}{c'} \sum_{\pi' \in \mathrm{RF}(\pi)} \prod_{b \in \pi'} \frac{\Gamma(\beta_b + t_b^1 + t_b^2)}{\Gamma(\beta_b + t_b^1)\Gamma(\beta_b + t_b^2)} \tag{11}$$

$c' = c/\frac{\Gamma(n+1)\Gamma(n+1)\Gamma(K)}{\Gamma(2n+K)}$ And we have theorem 3.

Theorem 4. *Let $n = min(n_1, n_2)$ be the smaller number of cells of two conditions and $n_1 = O(n_2)$, when parameter $(\phi, \psi) \in \Omega \setminus Q$ we have*

$$p(A_\pi|t^1, t^2) \xrightarrow[n \to \infty]{a.s.} \begin{cases} 1 & \text{if } (\phi, \psi) \in A_\pi \\ 0 & \text{otherwise} \end{cases}$$

Things become more complicate when (ϕ, ψ) falling into Q, we know $p(Q|t^1, t^2)$ vanishes, but $p(A_\pi|t^1, t^2)$ may not.

Recall $N(\pi)$ represents number of blocks b in π. Let $S = \{\pi, (\phi, \psi) \in A_\pi\}$, which is the collection of partitions whose associated simplexes covering (ϕ, ψ). Let $N^* = \max_{\pi \in S} N(\pi)$, which is the max number of blocks of partitions from S. Let $S^* = \{\pi, (\phi, \psi) \in A_\pi \text{ and } N(\pi) = N^*\}$, which is the collection of partitions that covering (ϕ, ψ) with number of blocks equal to the max number N^*.

For example, when $K = 7$, For a $(\phi, \psi) \in A_{\pi_1} \cap A_{\pi_2} \cap A_{\pi_3}$, $\pi_1 = \{\{1, 2, 3\}, \{4, 5, 6, 7\}\}, \pi_2 = \{\{1, 6, 7\}, \{2, 4\}, \{3, 5\}\}, \pi_3 = \{\{1, 2, 3, 4, 5, 6\}\}$, and also (ϕ, ψ) does not belong to any other simplex A_π. Then $S = \{\pi_1, \pi_2, \pi_3\}$, $N^* = 3$, $S^* = \{\pi_2\}$.

Denote components from right hand side of (5): $\frac{1}{c'} \prod_{b \in \pi} \frac{\Gamma(\beta_b + t_b^1 + t_b^2)}{\Gamma(\beta_b + t_b^1)\Gamma(\beta_b + t_b^2)} = J(t^1, t^2, \pi)$. We have theorem 4.

Theorem 5. *Following the setting in theorem 2, when parameter $(\phi, \psi) \in Q$, and we have*

$$J(t^1, t^2, \pi) \xrightarrow[n \to \infty]{a.s.} \begin{cases} m(\pi) & \pi \in S^* \\ 0 & \text{otherwise} \end{cases}$$

$$and \sum_{\pi \in S^*} m(\pi) = 1, m(\pi) > 0$$

proofs are in the appendix.
Still using above example, in limiting case, we have $p(A_{\pi_3}|t^1, t^2) = 1$, $p(A_{\pi_2}|t^1, t^2) = 1$ and $p(A_{\pi_1}|t^1, t^2) = 0$. When the DE pattern is B_{π_1} for some genes. Since our underestimation of $p(A_{\pi_1}|z^1, z^2) = 0$, we will falsely classify those genes as differential distributed.

The asymptotic properties help us gain insight of the performance of our approach, scDD-boost may work poorly, when $(\phi, \psi) \in Q$, we may underestimate the posterior probability of true proportion change pattern, which reduce the posterior probabilities of true negative and enlarge false positive rate.

4.2. **Random weighting.** In this section, we gave an intuitive justification for consistency between bayesian framework clustering analysis and random weighting procedure. A full bayesian analysis for clustering needs to specify the density of data given the partition. Specifically, in single cell analysis we need to know the density of transcripts of genes given the partitions which requires understanding of co-expression and dependence between genes. Instead of trying to untangle the mystery behind the dependence of genes, we consider following approximation

$$P(\text{Partition}|X) \leftarrow P(\text{Partition}|D) \leftarrow P(\Delta|D) \leftarrow D/W$$

where D is the estimated distance matrix of X, Δ is the true distance of X and W is randomly distributed matrix of weights. We conjecture that the probability of partitions given data can be approximated by switching conditioning on data to conditioning on the estimated distance of data. As distance matrix typically gave the geometrical structure between elements which can be used to infer how likely a partition is. In addition, partition can be obtained by distance based clustering algorithm (K-medoids) on true distance matrix Δ. To approximate distribution $(\Delta|D)$, we use our random weighting procedure, namely sampling a weighting matrix W first and then do the component-wisely dividing of original distance matrix D by W.

We gave a brief justification for this approximation, suppose units i and j are merged into a common cluster if (and only if) $d_{i,j} < c$. Then $P(d^*_{i,j} < c) = P(w_{i,j} > c/d_{i,j})$, $w_{i,j} \sim \text{Gamma}(a, b)$. From Bayesian perspective, given the true distance $\Delta_{i,j}$, $d_{i,j}|\Delta_{i,j} \sim \text{Gamma}(a_1, a_1/\Delta_{i,j})$, so that the sampling mean of $d_{i,j}$ is $\Delta_{i,j}$. Further, for simplicity we ignore any issues about the d's or Δ's being true distances. The condition for qualifiable distance matrix is the triangle inequality among the pairwise distances, such condition would not affect our clustering results too much. But, a simple analysis might suppose that a-priori $1/\Delta_{i,j} \sim \text{Gamma}(a_0, d_0)$. The scaling is such that $\text{E}(1/\Delta_{i,j}) = a_0\ d_0$. The posterior, by conjugacy, has $1/\Delta_{i,j}|d_{i,j} \sim \text{Gamma}(a_0 + a_1, d_0 + a_1 d_{i,j})$. Then the posterior

probability that i and j should be clustered is the posterior probability that $\Delta_{i,j} < c$, which is $P(\text{Gamma}((a_0 + a_1), (a_0 + a_1)) > (d_0 + a_1 * d_{i,j})/(a_0 + a_1) * 1/c)$, parameters (a_0, d_0, a_1) are estimated from maximizing the marginal likelihood of $d_{i,j}$.

In order to match the posterior probability that elements i and j belongs to the same cluster through the simple bayesian analysis to random weighting, which is equivalently to match

$$P(\Delta_{i,j} < c|d_{i,j}) = P(1/\Delta_{i,j} > 1/c|d_{i,j})$$

and

$$P(d_{i,j}/w_{i,j} < c|d_{i,j}) = P(w_{i,j}/d_{i,j} > 1/c|d_{i,j})$$

yielding $a = a_0 + a_1$ and $b = a_1$. Therefore, we gave a way of modeling the distribution of weights such that partition based on random generated distance D/W would approximate the partition given data based on a full bayesian framework.

5. Discussion

We have presented scDDboost, a compositional model for detecting differential distributed genes from scRNA-seq data. To account for the over-dispersion and multi-modality of single-cell data, scDDboost modeled transcripts as mixture distributed. Unlike previous invented methods (e.g. Deseq2, MAST and scDD), which conducts genewise DD test in an isolated manner. scDDboost make whole genome information shared at gene level by further assuming the mixture distribution of transcripts is a mixture over the subtypes of cells. Another advantage of scDDboost is its' flexibility to allow user specified clustering methods of cells, with more and more studies of the scRNA-seq data, there will be more accurate distance matrix between cells, which will yield better estimation of subtypes and inference of DD genes. We combine estimations of changes of subtypes' proportions across conditions and changes of mean expressions across subtypes to infer distributional changes of transcripts. To estimate changes of subtypes' proportions across conditions, we use empirical Bayes and developed a double Dirichlet prior distribution. We invented a random weighting scheme that stabilize our DD inference as well as approximating the results as if we have done a fully bayesian clustering analysis based on Dirichlet prior. We demonstrated that scDDboost outperforms existing approaches in simulation and tends to be more powerful than existing methods on a wide range of public available empirical datasets.

One limitation of scDDboost is that current EBseq inference of the DE patterns is computationally not feasible for big number of subtypes. Given the noise level among the single cell data and especially if we want to identify DD genes among conditions containing thousands of cells, allowing a big number of subtypes would make cells under same subtype more homogeneous and result in a more accurate estimations for the distribution of

transcripts. Further research is needed for acceleration of EBseq, one direction is to reduce the calculation on those patterns that would have small posterior probabilities.

This methodology extends `scDD`[Korthauer et al., 2016], which similarly treats expression data within a condition as a statistical mixture, but which does not share information among genes on the mixing proportions.

References

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106–R106, 2010. doi: 10.1186/gb-2010-11-10-r106. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3218662/`.

Rhonda Bacher and Christina Kendziorski. Design and computational analysis of single-cell rna-sequencing experiments. *Genome Biology*, 17(1):63, 2016. doi: 10.1186/s13059-016-0927-y. URL `https://doi.org/10.1186/s13059-016-0927-y`.

Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33:155 EP –, 01 2015. URL `http://dx.doi.org/10.1038/nbt.3102`.

Li-Fang Chu, Ning Leng, Jue Zhang, Zhonggang Hou, Daniel Mamott, David T. Vereide, Jeea Choi, Christina Kendziorski, Ron Stewart, and James A. Thomson. Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biology*, 17(1):173, 2016. doi: 10.1186/s13059-016-1033-x. URL `https://doi.org/10.1186/s13059-016-1033-x`.

David B. Dahl. Modal clustering in a class of product partition models. *Bayesian Anal.*, 4(2):243–264, 06 2009. doi: 10.1214/09-BA409. URL `https://doi.org/10.1214/09-BA409`.

Spyros Darmanis, Steven A Sloan, Derek Croote, Marco Mignardi, Sophia Chernikova, Peyman Samghababi, Ye Zhang, Norma Neff, Mark Kowarsky, Christine Caneda, Gordon Li, Steven D Chang, Ian David Connolly, Yingmei Li, Ben A Barres, Melanie Hayden Gephart, and Stephen R Quake. Single-cell rna-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell reports*, 21(5):1399–1410, 10 2017. doi: 10.1016/j.celrep.2017.10.030. URL `https://www.ncbi.nlm.nih.gov/pubmed/29091775`.

Mihails Delmans and Martin Hemberg. Discrete distributional differential expression (d3e) - a tool for gene expression analysis of single-cell rna-seq data. *BMC Bioinformatics*, 17(1):110, 2016. doi: 10.1186/s12859-016-0944-6. URL `https://doi.org/10.1186/s12859-016-0944-6`.

Qiaolin Deng, Daniel Ramsköld, Björn Reinius, and Rickard Sandberg. Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196, 2014. ISSN 0036-8075. doi: 10.1126/science.1245316. URL `http://science.sciencemag.org/content/343/6167/193`.

Isaac Engel, Grégory Seumois, Lukas Chavez, Daniela Samaniego-Castruita, Brandie White, Ashu Chawla, Dennis Mock, Pandurangan Vijayanand, and Mitchell Kronenberg. Innate-like functions of natural killer t cell subsets result from highly divergent gene programs. *Nature Immunology*, 17:728 EP –, 04 2016. URL http://dx.doi.org/10.1038/ni.3437.

Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, Hannah W. Miller, M. Juliana McElrath, Martin Prlic, Peter S. Linsley, and Raphael Gottardo. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome Biology*, 16(1):278, 2015. doi: 10.1186/s13059-015-0844-5. URL https://doi.org/10.1186/s13059-015-0844-5.

Fan Guo, Liying Yan, Hongshan Guo, Lin Li, Boqiang Hu, Yangyu Zhao, Jun Yong, Yuqiong Hu, Xiaoye Wang, Yuan Wei, Wei Wang, Rong Li, Jie Yan, Xu Zhi, Yan Zhang, Hongyan Jin, Wenxin Zhang, Yu Hou, Ping Zhu, Jingyun Li, Ling Zhang, Sirui Liu, Yixin Ren, Xiaohui Zhu, Lu Wen, Yi Qin Gao, Fuchou Tang, and Jie Qiao. The transcriptome and dna methylome landscapes of human primordial germ cells. *Cell*, 161(6):1437–1452, 2017/12/05 2015. doi: 10.1016/j.cell.2015.05.015. URL http://dx.doi.org/10.1016/j.cell.2015.05.015.

Leonard Kaufman and Peter Rousseeuw. *Clustering by means of medoids*. North-Holland, 1987.

Taiyun Kim, Irene Rui Chen, Yingxin Lin, Andy Yi-Yang Wang, Jean Yee Hwa Yang, and Pengyi Yang. Impact of similarity metrics on single-cell rna-seq data clustering. *Briefings in Bioinformatics*, page bby076, 2018. doi: 10.1093/bib/bby076. URL http://dx.doi.org/10.1093/bib/bby076.

Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, and Martin Hemberg. Sc3: consensus clustering of single-cell rna-seq data. *Nature Methods*, 14:483 EP –, 03 2017. URL http://dx.doi.org/10.1038/nmeth.4236.

Keegan D. Korthauer, Li-Fang Chu, Michael A. Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendziorski. A statistical approach for identifying differential distributions in single-cell rna-seq experiments. *Genome Biology*, 17(1):222, 2016. doi: 10.1186/s13059-016-1077-y. URL https://doi.org/10.1186/s13059-016-1077-y.

Roshan M. Kumar, Patrick Cahan, Alex K. Shalek, Rahul Satija, A. Jay DaleyKeyser, Hu Li, Jin Zhang, Keith Pardee, David Gennert, John J. Trombetta, Thomas C. Ferrante, Aviv Regev, George Q. Daley, and James J. Collins. Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature*, 516:56 EP –, 12 2014. URL http://dx.doi.org/10.1038/nature13920.

Ning Leng, John A. Dawson, James A. Thomson, Victor Ruotti, Anna I. Rissman, Bart M. G. Smits, Jill D. Haag, Michael N. Gould, Ron M. Stewart, and Christina Kendziorski. Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, 29(8):1035–1043, 2013. doi: 10.1093/bioinformatics/btt087. URL + http://dx.doi.org/10.1093/bioinformatics/btt087.

Ning Leng, Li-Fang Chu, Chris Barry, Yuan Li, Jeea Choi, Xiaomao Li, Peng Jiang, Ron M Stewart, James A Thomson, and Christina Kendziorski. Oscope identifies oscillatory genes in unsynchronized single-cell rna-seq experiments. *Nature Methods*, 12:947 EP –, 08 2015. URL http://dx.doi.org/10.1038/nmeth.3549.

Peijie Lin, Michael Troup, and Joshua W. K. Ho. Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. *Genome Biology*, 18(1):59, 2017. doi: 10.1186/s13059-017-1188-0. URL https://doi.org/10.1186/s13059-017-1188-0.

Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550, 2014. doi: 10.1186/s13059-014-0550-8. URL https://doi.org/10.1186/s13059-014-0550-8.

John C. Marioni and Detlev Arendt. How single-cell genomics is changing evolutionary and developmental biology. *Annual Review of Cell and Developmental Biology*, 33(1): 537–553, Oct 2017. ISSN 1530-8995. doi: 10.1146/annurev-cellbio-100616-060818. URL https://doi.org/10.1146/annurev-cellbio-100616-060818. PMID: 28813177.

Nicholas E Navin. The first five years of single-cell cancer genomics and beyond. *Genome Research*, 25(10):1499–1507, 10 2015. doi: 10.1101/gr.191098.115. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4579335/.

Tal Nawy. Single-cell sequencing. *Nature Methods*, 11:18 EP –, 12 2013. URL http://dx.doi.org/10.1038/nmeth.2771.

Efthymia Papalexi and Rahul Satija. Single-cell rna sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, 18:35 EP –, 08 2017. URL http://dx.doi.org/10.1038/nri.2017.76.

Anoop P. Patel, Itay Tirosh, John J. Trombetta, Alex K. Shalek, Shawn M. Gillespie, Hiroaki Wakimoto, Daniel P. Cahill, Brian V. Nahed, William T. Curry, Robert L. Martuza, David N. Louis, Orit Rozenblatt-Rosen, Mario L. Suvà, Aviv Regev, and Bradley E. Bernstein. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014. ISSN 0036-8075. doi: 10.1126/science.1254257. URL http://science.sciencemag.org/content/344/6190/1396.

Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1):241, 2015. doi: 10.1186/s13059-015-0805-z. URL https://doi.org/10.1186/s13059-015-0805-z.

Siddheswar Ray and Rose H. Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. 2000.

Alex K. Shalek, Rahul Satija, Joe Shuga, John J. Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S. Gertner, Jellert T. Gaublomme, Nir Yosef, Schraga Schwartz, Brian Fowler, Suzanne Weaver, Jing Wang, Xiaohui Wang, Ruihua Ding, Raktima Raychowdhury, Nir Friedman, Nir Hacohen, Hongkun Park, Andrew P. May, and Aviv Regev. Single-cell rna-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510: 363 EP –, 06 2014. URL http://dx.doi.org/10.1038/nature13437.

Charlotte Soneson and Mark D. Robinson. Bias, robustness and scalability in differential expression analysis of single-cell rna-seq data. *bioRxiv*, 2017. doi: 10.1101/143289. URL https://www.biorxiv.org/content/early/2017/05/28/143289.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617, March 2003. ISSN 1532-4435. doi: 10.1162/153244303321897735. URL https://doi.org/10.1162/153244303321897735.

Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T Gray, Staci A Sorensen, Tim Dolbeare, Darren Bertagnolli, Jeff Goldy, Nadiya Shapovalova, Sheana Parry, Changkyu Lee, Kimberly Smith, Amy Bernard, Linda Madisen, Susan M Sunkin, Michael Hawrylycz, Christof Koch, and Hongkui Zeng. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience*, 19:335 EP –, 01 2016. URL http://dx.doi.org/10.1038/nn.4216.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386, 04 2014. doi: 10.1038/nbt.2859. URL https://www.ncbi.nlm.nih.gov/pubmed/24658644.

Udo Wagner and Alfred Taudes. A multivariate polya model of brand choice and purchase incidence. *Marketing Science*, 5(3):219–244, August 1986. ISSN 1526-548X. doi: 10.1287/mksc.5.3.219. URL http://dx.doi.org/10.1287/mksc.5.3.219.

Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome Biology*, 18(1):174, 2017. doi: 10.1186/s13059-017-1305-0. URL https://doi.org/10.1186/s13059-017-1305-0.

Appendix

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 $(\hat{a}, \hat{b}) \leftarrow \text{hyper}(D)$
repeat
Gamma noise vector: e, with components $\sim \text{Gamma}(\hat{a}, \hat{b})$
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 $\forall \text{genes}\, g,\ P(\text{DD}_g|X, y) = \frac{1}{n_r} \sum_{D^*} P^*_g$

Supplementary Material

Contents

Synthetic Data

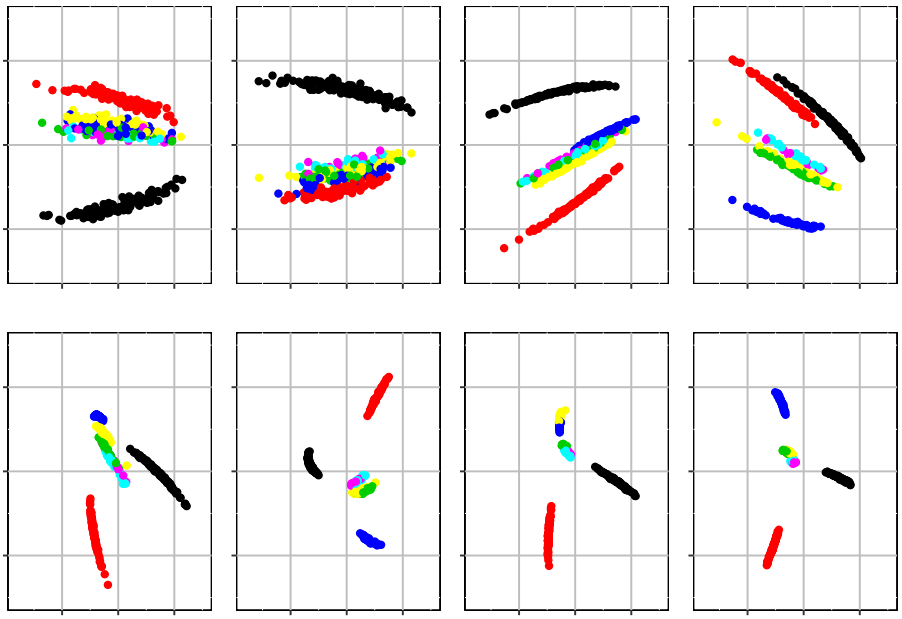

Figure 11. first two principal components of transcripts under different parameters for simulated data. Different parameters resulted in different degree of separation of subtypes. We have 4 different settings for hyper-parameters of simulation, each setting has 2 replicates

Data sets

EBSeq

Data set	Conditions	Number of cells/condition	Organism	Ref	K
GSE52529	T0 vs T24	96,96	human	[Trapnell et al., 2014]	6
GSE57872	patient1 vs patient2	192,96	human	[Patel et al., 2014]	7
GSE48968-GPL13112	BMDC (2h LPS stimulation) vs 6h LPS	96,96	mouse	[Shalek et al., 2014]	8
GSE60749-GPL13112	serum + LIF vs 2i + LIF	90,94	mouse	[Kumar et al., 2014]	3
GSE74596	NKT1 vs NTK2	46,68	mouse	[Engel et al., 2016]	5
EMTAB2805	G1 vs G2M	95,96	mouse	[Buettner et al., 2015]	7
GSE71585-GPL13112	Gad2tdTpositive vs Cux2tdTnegative	80,140	mouse	[Tasic et al., 2016]	7
GSE64016	G1 vs G2	91,76	human	[Leng et al., 2015]	8
GSE79102	patient1 vs patient2	51, 89	human	Kiselev et al. [2017]	4
GSE45719	16-cell stage blastomere vs mid blastocyst cell	50, 60	mouse	[Deng et al., 2014]	5
GSE63818	Primordial Germ Cells, developmental stage: 7 week gestation vs Somatic Cells, developmental stage: 7 week gestation	40,26	mouse	[Guo et al., 2015]	6
GSE75748	DEC vs EC	64, 64	human	[Chu et al., 2016]	9
GSE84465	neoplastic cells vs non-neoplastic cells	546, 664	human	[Darmanis et al., 2017]	9

TABLE 2. datasets used for comparisons of DD analysis under different methods

Suppose we have K subtypes, let $X_g^I = X_{g,1}^I, ..., X_{g,S_1}^I$ denote transcripts at gene g from subtype $I, I = 1, ...K$. In the EBSeq model it assumed that counts within subtype I are distributed as Negative Binomial: $X_{g,s}^I | r_{g,s}, q_g^I \sim NB(r_{g,s}, q_g^I)$ Where

$$P(X_{g,s}^I | r_{g,s}, q_g^I) = \binom{X_{g,s} + r_{g,s} - 1}{X_{g,s}} (1 - q_g^I)^{X_{g,s}^I} (q_g^I)^{r_{g,s}}$$

and $\mu_{g,s}^I = r_{g,s}(1 - q_g^I)/q_g^I; \sigma_{g,s}^I = r_{g,s}(1 - q_g^I)/(q_g^I)^2$.

The EBSeq model assumed a prior distribution on q_g^I : $q_g^I | \alpha, \beta^{I_g} \sim Beta(\alpha, \beta^{I_g})$. The hyperparameter α is shared by all the isoforms and β^{I_g} is I_g specific. We further assume

that $r_{g,s} = r_{g,0} * l_s$ where $r_{g,0}$ is an isoform specific parameter common across subtypes and $r_{g,s}$ depends on it through the sample-specific normalization factor l_s

What we are interested at those K groups comparison is the expression pattern, through EBSeq modeling we are able to obtain posterior probabilities over

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

For any partition π of K elements.

For example $K = 3$, there are 5 expression pattern, $P_1, P_2, ..., P_5$

$$\begin{aligned}
&P1 : q_g^1 = q_g^2 = q_g^3 \\
&P2 : q_g^1 = q_g^2 \neq q_g^3 \\
&P3 : q_g^1 \neq q_g^2 = q_g^3 \\
&P4 : q_g^1 = q_g^3 \neq q_g^2 \\
&P5 : q_g^1 \neq q_g^2 \neq q_g^3 \text{ and } q_g^1 \neq q_g^3
\end{aligned}$$

Under the assumption that two groups I and J share the same q_g we can pool the counts from the two groups by viewing them come from same distribution i.e. $X_g^{I,J}|r_{g,s}, q_g \sim NB(r_{g,s}, q_g)$, $q_g|\alpha, \beta^{I_g} \sim Beta(\alpha, \beta^{I_g})$ and obtained the prior predictive function $f_0^{I_g}(X_g^{I,J}) = \int_0^1 P(X_g^{I,J}|r_{g,s}, q_g) * P(q_g|\alpha, \beta^{I_g}) dq_g = \Big[\prod_{s=1}^{S} \binom{X_{g,s}+r_{g,s}-1}{X_{g,s}}\Big] \frac{Beta(\alpha+\Sigma_{s=1}^S r_{g,s}, \beta^{I_g}+\Sigma_{s=1}^S X_{g,s})}{Beta(\alpha,\beta^{I_g})}$. Consequently, we have prior predictive function for $P1, ..., P5$ as

$$\begin{aligned}
g_1^{I_g}(X_g^{1,2,3}) &= f_0^{I_g}(X_g^{1,2,3}) \\
g_2^{I_g}(X_g^{1,2,3}) &= f_0^{I_g}(X_g^{1,2}) f_0^{I_g}(X_g^3) \\
g_3^{I_g}(X_g^{1,2,3}) &= f_0^{I_g}(X_g^1) f_0^{I_g}(X_g^{2,3}) \\
g_4^{I_g}(X_g^{1,2,3}) &= f_0^{I_g}(X_g^{1,3}) f_0^{I_g}(X_g^2) \\
g_5^{I_g}(X_g^{1,2,3}) &= f_0^{I_g}(X_g^1) f_0^{I_g}(X_g^2) f_0^{I_g}(X_g^3)
\end{aligned}$$

Then the marginal distribution of counts $X_g^{1,2,3}$ is $\sum_{k=1}^{5} p_k g_k^{I_g}(X_g^{1,2,3})$, where proportion parameters p_k satisfying $\sum_{k=1}^{5} p_k = 1$ and are estimated by EM algorithm. Thus, the posterior

probability of an expression pattern k is obtained by:

$$\frac{p_k g_k(X_g^{1,2,3})}{\sum_{k=1}^{5} p_k g_k^{I_g}(X_g^{1,2,3})}$$

modalclust

Product Partition Model Let $X = (X_1, X_2, ..., X_n)$ be n one dimension observed data, given a partition for the data $\pi = \{S_1, ..., S_q\}$, where S_i are disjoint subsets of $\{1, 2, ..., n\}$ and $\bigcup_{i=1}^{q} S_i = \{1, 2, ..., n\}$. The likelihood for X satisfying such partition is

$$p(X|\pi) = \prod_{i=1}^{q} f(X_{S_i})$$

where X_{S_i} is the vector of observations corresponding to the items of component S_i, The component likelihood $f(X_S)$ is defined for any non-empty component S and can take any form. The partition π is the only parameter we are interested at. Any other parameters that may have been involved in the model have been integrated over their prior.

The prior distribution for a partition π is also taken as a product form. We use the partition that maximize the posterior $p(\pi|X) \propto p(X|\pi)p(\pi)$ as the estimated clustering of X.

Dahl demonstrated by some choice of f and prior of π, we can reduce the time complexity of finding the MAP partition from factorial(n) to $O(n^2)$ Dahl [2009], And the crucial condition for f is that if X_{S_1} and X_{S_2} are overlapped in the sense that $\min\{X_{S_2}\} < \max\{X_{S_1}\} < \max\{X_{S_2}\}$ or $\min\{X_{S_1}\} < \max\{X_{S_2}\} < \max\{X_{S_1}\}$, $X_{S_1^*} and X_{S_2^*}$ be the sets of swapping one pair of those overlapped terms and keep the other unchanged. Then $f(X_{S_1})f(X_{S_2}) \leq f(X_{S_1^*})f(X_{S_2^*})$. Under such condition, we know that possible MAP candidates must be those partition that for any two subgroups of data, all the data from subgroup1 has to be either greater or smaller than all the data from subgroup2.

In Poisson-Gamma Model we assuming:

$$\begin{aligned} X_i|\pi, \lambda &\sim Poisson(X_i|\lambda_1 \mathbf{I}\{i \in S_1\} + ... + \lambda_q \mathbf{I}\{i \in S_q\}) \\ \pi &\sim p(\pi) \\ \lambda_j &\sim Gamma(\alpha_0, \beta_0) \end{aligned}$$

where $p(\pi) \propto \prod_{i=1}^{q} \eta_0 \Gamma(|S_i|)$. Integrate out λ, $f(X_S)$ is obtained as:

$$f(X_S) = \frac{\beta^\alpha}{(|S| + \beta)^{\sum_{i \in S} X_i + \alpha}} \frac{\Gamma(\sum_{i \in S} X_i + \alpha)}{\Gamma(\alpha)} \frac{1}{\prod_{i \in S} X_i}$$

$f(X_S)$ still satisfying the condition mentioned

Proof. if X_{S_1} and X_{S_2} are overlapped, without loss of generality, we assume $\min\{X_{S_2}\} < \max\{X_{S_1}\} < \max\{X_{S_2}\}$, and we swap $\max\{X_{S_1}\}$ with $\min\{X_{S_2}\}$ and keep the rest unchanged or we could also swap $\max\{X_{S_1}\}$ with $\max\{X_{S_2}\}$. We denote the new set forming by swap of $\max\{X_{S_1}\}$ with $\min\{X_{S_2}\}$ as S_1^* and S_2^* and swap of $\max\{X_{S_1}\}$ with $\max\{X_{S_2}\}$ as S_1^{**}, S_2^{**} accordingly.

Then we need to show at least one of the following happens

$$f(X_{S_1^*})f(X_{S_2^*}) \geq f(X_{S_1})f(X_{S_2}) \tag{12}$$

$$f(X_{S_1^{**}})f(X_{S_2^{**}}) \geq f(X_{S_1})f(X_{S_2}) \tag{13}$$

Let $a = \max\{X_{S_1}\}$, $b = \min\{X_{S_2}\}$ and $c = \max\{X_{S_2}\}$. $h_1 = \sum_{i \in S_1} X_i - a$ and $h_2 = \sum_{i \in S_2} X_i - b$, n_1 and n_2 are the number of elements in S_1 and S_2. Then

$$\begin{gathered} f(X_{S_1^*})f(X_{S_2^*}) \geq f(X_{S_1})f(X_{S_2}) \\ \Longleftrightarrow \\ \frac{\Gamma(h_1+a+\alpha)}{(n_1+\beta)^{h_1+a+\alpha}} \frac{\Gamma(h_2+b+\alpha)}{(n_2+\beta)^{h_2+b+\alpha}} \leq \frac{\Gamma(h_2+a+\alpha)}{(n_2+\beta)^{h_2+a+\alpha}} \frac{\Gamma(h_1+b+\alpha)}{(n_2+\beta)^{h_1+b+\alpha}} \\ \Longleftrightarrow \\ \frac{\Gamma(h_1+a+\alpha)}{\Gamma(h_1+b+\alpha)} \frac{\Gamma(h_2+b+\alpha)}{\Gamma(h_2+a+\alpha)} \leq (\frac{n_1+\beta}{n_2+\beta})^{a-b} \end{gathered}$$

Left hand side of above formula is $\text{LHS}_1 = \frac{(h_1+b+\alpha)\ldots(h_1+a-1+\alpha)}{(h_2+b+\alpha)\ldots(h_2+a-1+\alpha)}$ by the property of Gamma function and X_i are integer.

Similarly,

$$\begin{gathered} f(X_{S_1^{**}})f(X_{S_2^{**}}) \geq f(X_{S_1})f(X_{S_2}) \\ \Longleftrightarrow \\ \frac{\Gamma(h_2+c+\alpha)}{\Gamma(h_2+a+\alpha)} \frac{\Gamma(h_1+a+\alpha)}{\Gamma(h_1+c+\alpha)} \leq (\frac{n_2+\beta}{n_1+\beta})^{c-a} \end{gathered}$$

Left hand side of above formula is $\text{LHS}_2 = \frac{(h_2+a+\alpha)\ldots(h_2+c-1+\alpha)}{(h_1+a+\alpha)\ldots(h_1+c-1+\alpha)}$

If $h_1 \leq h_2$, then $\text{LHS}_1 \leq (\frac{h_1+a-1+\alpha}{h_2+a-1+\alpha})^{a-b}$ and $\text{LHS}_2 \leq (\frac{h_2+c-1+\alpha}{h_1+c-1+\alpha})^{a-b}$

So if $\frac{h_1+a-1+\alpha}{h_2+a-1+\alpha} \leq \frac{n_1+\beta}{n_2+\beta}$ then (12) holds, if $\frac{h_2+c-1+\alpha}{h_1+c-1+\alpha} \leq \frac{n_1+\beta}{n_2+\beta}$ then (13) holds

We multiply those two inequalities, we found that $\frac{h_1+a-1+\alpha}{h_2+a-1+\alpha} * \frac{h_2+c-1+\alpha}{h_1+c-1+\alpha} = \frac{h_1+a-1+\alpha}{h_1+c-1+\alpha} * \frac{h_2+c-1+\alpha}{h_2+a-1+\alpha} \leq 1$ as $c > a$ and $h_1 \leq h_2$ But $\frac{n_1+\beta}{n_2+\beta} * \frac{n_1+\beta}{n_2+\beta} = 1$. At least one equality holds, consequently at least one of (12) and (13) holds.

Similar proof for the case $h_1 > h_2$.

□

Randomized k-means

5.1. **simulation.** We random generate one-dimensional data X from a mixture of 5 normal distributions with different means and same variance. We compare clustering results between random weighting and bayesian clustering with Dirichlet process as prior in terms of posterior probabilities that two elements belong to the same class given the whole data and adjusted rand index comparing to the underlying true class label (Fig 12).

5.2. **empirical study. Double Dirichlet Model:**

On the double Dirichlet masses, using notation as in Section 2.3 we have density functions:

$$p_\pi(\phi, \psi) = q_\pi(\Phi_\pi, \Psi_\pi) \prod_{b \in \pi} \left[p(\tilde{\phi}_b) p(\tilde{\psi}_b) \right]$$

with

$$q_\pi(\Phi_\pi, \Psi_\pi) = \frac{\Gamma(\sum_{b \in \pi} \beta_b)}{\prod_{b \in \pi} \Gamma(\beta_b)} \left[\prod_{b \in \pi} \Phi_b^{\beta_b - 1} \right] 1 \left[\Phi_\pi = \Psi_\pi \right]$$

and

$$p(\tilde{\phi}_b) = \frac{\Gamma(\sum_{k \in b} \alpha_k)}{\prod_{k \in b} \Gamma(\alpha_k)} \prod_{k \in b} \tilde{\phi}_k^{\alpha_k - 1}, \qquad p(\tilde{\psi}_b) = \frac{\Gamma(\sum_{k \in b} \alpha_k)}{\prod_{k \in b} \Gamma(\alpha_k)} \prod_{k \in b} \tilde{\psi}_k^{\alpha_k - 1}.$$

Proofs:

Lemma 1. *If π_2 is not refinement of π_1 then $A_{\pi_1} \cap A_{\pi_2}$ is a lower dimensional subset of A_{π_2}*

Proof of theorem 2

Proof. by lemma 1, it is easy to verify. □

where $p(t^1, t^2|\phi, \psi) = p(t^1|\phi)p(t^1|\psi)$, $t^1|\phi \sim \text{multinomial}(n_1, \phi), t^2|\psi \sim \text{multinomial}(n_2, \psi)$. Recall the definition of $A_\pi = \{(\phi, \psi) : \Phi_b = \Psi_b\}$ and A_π is a simplex. Denote the finest partition as $\pi_F = \{\{1\}, \{2\}, ..., \{K\}\}$, associated simplex $A_{\pi_F} = \{(\phi, \psi) : \phi_i = \psi_i, i = 1, ..., K\}$ for any two partition π_1 and π_2, intersection of their associated simplex must not be empty

since $A_{\pi_F} \subset A_{\pi_1} \cap A_{\pi_2} \neq \emptyset$. To discuss the issue of overlapping of simplex A_π, we first introduce some notations. The whole space $\Omega = \{(\phi, \psi), \phi_i, \psi_i > 0 \text{ and } \sum_{i=1}^K \phi_i = \sum_{i=1}^K \psi_i = 1\}$ and we define the refinement and coarseness relationship between partitions, we say a partition $\tilde{\pi}$ refines another partition π if $\forall b \in \pi$ there exists $s \subset \tilde{\pi}$ such that $\cup_{b' \in s} b' = b$. When $\tilde{\pi}$ refines π, we say $\tilde{\pi}$ is a refinement of(finer than) π or π is a coarseness of(coarser than) $\tilde{\pi}$. Observe that if π' refines π, then $A_\pi \cap A_{\pi'} = A_{\pi'}$, $\int_{A_\pi \cap A_{\pi'}} p(z^1, z^2|\phi, \psi) p(\phi, \psi|A_{\pi'}) d\phi d\psi = \int_{A_{\pi'}} p(t^1, t^2|\phi, \psi) p(\phi, \psi|A_{\pi'}) d\phi d\psi$. When π' is not refinement of π, we need to know the dimension of $A_\pi \cap A_{\pi'}$. Consider a map $f : b \to v$, which maps the block b to a vector $v \in \{0,1\}^K$, the ith component of v is $1_{\{i \in b\}}$. And denote $\dim(S)$ be the dimension of space S. A_π can be equivalently defined as $A_\pi = \{(\phi, \psi) : M_\pi * (\phi - \psi) = 0\}$, M_π is a matrix with rows be $v_b = f(b), \forall b \in \pi$, that is to say (ϕ, ψ) are in the null space of linear transformation M_π. We have following lemma
Proof of lemma 1

Proof. Let V denote the orthogonal space of $\phi - \psi$, when $(\phi, \psi) \in A_{\pi_1} \cap A_{\pi_2}$, and $\dim(A_{\pi_1} \cap A_{\pi_2}) = \dim(\phi - \psi) + \dim(\psi) = 2K - \dim(V) - 1$. Also let $\pi_1 = \{b_1^1, ..., b_s^1\}, \pi_2 = \{b_1^2, ..., b_t^2\}$. The corresponding vectors are $v_1^1, ..., v_s^1$ and $v_1^2, ..., v_t^2$. We claim there must be a $b_i^1 \in \pi$ whose corresponding v_i^1 is linear independent with $v_1^2, ..., v_t^2$. If not, for every v_i^1 there exists $\alpha_1^i, ..., \alpha_t^i$ such that

$$v_i^1 = \sum_{j=1}^t \alpha_j^i v_j^2 \qquad (*)$$

If $b_j^2 \cap b_i^1 \neq \emptyset$, then multiply v_j^2 on both sides of (*), we obtain $v_i^1 * v_j^2 = \alpha_j^i (v_j^2)^2$, as v_j^2 are orthogonal vectors, and $v_i^1 * v_j^2 > 0$ implies $\alpha_j^i > 0$. Consider $x = f(b_j^2 \setminus b_i^1)$, we have $x * v_i^1 = 0$ and we multiply x on both sides of (*) to obtain $\alpha_j^i v_j^2 * x = 0$, thus x must be zero vector and $b_j^2 \setminus b_i^1 = \emptyset$, which implies $b_j^2 \subset b_i^1$. That is to say when $b_j^2 \cap b_i^1 \neq \emptyset$, b_j^2 must be subset of b_i^1. So b_i^1 is union of some blocks in π_2. Which implies π_2 is refinement of π_1, contradiction.
Consequently there exists $b \in \pi_1$ with $v(b)$ linear independent with $v(b'), b' \in \pi_2$. $\dim(V)$ is at least $N(\pi_2) + 1, \dim(A_{\pi_1} \cap A_{\pi_2}) < \dim(A_{\pi_2})$ □

Proof of theorem 3 and theorem 4

Proof. Given the condition that $\alpha_k = 1, \forall k$ and $\beta_b = \sum_{k \in b} \alpha_k$, recall $p(A_\pi|t^1, t^2) = \sum_{\pi' \in \mathrm{RF}(\pi)} J(t^1, t^2, \pi')$ and $J(t^1, t^2, \pi) = \frac{1}{c'} \prod_{b \in \pi} \frac{\Gamma(\beta_b + t_b^1 + t_b^2)}{\Gamma(\beta_b + t_b^1)\Gamma(\beta_b + t_b^2)}$
Assuming there are K subgroups, since n_1 and n_2 goes to infinite at same rate, for simplicity we assume $n = \sum_{i=1}^K t_i^1 = \sum_{i=1}^K t_i^2$, $t^1 \sim$ multinomial(ϕ), $t^2 \sim$ multinomial(ψ) and $t_b^1 = \sum_{i \in b} z_i^1$ and $t_b^2 = \sum_{i \in b} z_i^2$, so $t_b^1 \sim$ binomial (n, Φ_b) and $t_b^2 \sim$ binomial (n, Ψ_b), where

$\Phi_b = \sum_{i\in b}\phi_i$ and $\Psi_b = \sum_{i\in b}\psi_i$. Let $f(n,b) = \frac{\Gamma(\beta_b+t_b^1+t_b^2)}{\Gamma(\beta_b+t_b^1)\Gamma(\beta_b+t_b^2)}$, then

$$J(z^1, z^2, \pi) \propto \prod_{b\in\pi} f(n,b)$$

$\log f(n,b)$ =log($\Gamma(\beta_b+t_b^1+t_b^2)$) - log($\Gamma(\beta_b+t_b^1)$) - log($\Gamma(\beta_b+t_b^2)$), notice that t_b^1, t_b^2 and β_b are integers, and when x is integer, $\Gamma(x)$ is the factorial of $(x-1)$. We have $\log f(n,b)$ =log($(\beta_b+t_b^1+t_b^2-1)!$) − log($(\beta_b+t_b^1-1)!$) − log($(\beta_b+t_b^2-1)!$) and when n is large we could use Stirling's approximation, i.e. $\log(n!) = n\log(n) - n + O(\log(n))$, we have $\log((\beta_b+t_b^1+t_b^2-1)!) - \log((\beta_b+t_b^1-1)!) - \log((\beta_b+t_b^2-1)!) \approx (\beta_b+t_b^1+t_b^2-1)\log(\beta_b+t_b^1+t_b^2-1) - (\beta_b+t_b^1-1)\log(\beta_b+t_b^1-1) - (\beta_b+t_b^2-1)\log(\beta_b+t_b^2-1) + O(\log(n))$.
Plug into $f(n,b)$ we have:

$$\log f(n,b) \approx (\beta_b+t_b^1-1)\log(1+\frac{t_b^2}{\beta_b+t_b^1-1}) + (\beta_b+t_b^2-1)\log(1+\frac{t_b^1}{\beta_b+t_b^2-1}) + O(\log(n))$$

as $\beta_b\log(\beta_b+t_b^1+t_b^2-1) \sim O(\log(n))$ and by law of large number and slutsky's theorem, $\log(1+\frac{t_b^2}{\beta_b+t_b^1-1}) \to \log(1+\frac{\Psi_b}{\Phi_b})$, $\log(1+\frac{t_b^1}{\beta_b+t_b^2-1}) \to \log(1+\frac{\Phi_b}{\Psi_b})$ *a.s.* and $\frac{\log f(n,b)}{n} \to \Phi_b\log(1+\frac{\Psi_b}{\Phi_b}) + \Psi_b\log(1+\frac{\Phi_b}{\Psi_b})$ a.s. We have:

$$\frac{\log(\prod_{b\in\pi} f(n,b))}{n} \to \sum_b [\Phi_b\log(1+\frac{\Psi_b}{\Phi_b}) + \Psi_b\log(1+\frac{\Phi_b}{\Psi_b})] \quad a.s.$$

To find the maxima (Φ,Ψ), we fix Ψ and let $C = \frac{\log(\prod_{b\in\pi} f(n,b))}{n} + \lambda(\sum_{b\in\pi}\Phi_b - 1)$, we have $\frac{\partial C}{\partial\Phi_b} = \log(1+\frac{\Psi_b}{\Phi_b}) + \lambda$, stationary point is $\Phi_b = \Psi_b, \forall b$. and for the hessian matrix $\frac{\partial^2 C}{\partial\Phi_b^2} = -\frac{\Psi_b}{\Phi_b^2+\Phi_b\Psi_b} < 0$ and $\frac{\partial^2 C}{\partial\Phi_b\partial\Phi_{b'}} = 0, \text{if } b \neq b'$, that is to say the hessian matrix is a diagonal matrix with every diagonal elements to be negative, so it is negative definite, and our objective function is concave. The maxima is the stationary point $\Phi = \Psi$. And when $\Phi = \Psi$, $\frac{\log(\prod_{b\in\pi} f(n,b))}{n} = 2\ln(2)$ a constant not dependent on partition π and Φ. That is to say if $(\phi,\psi) \in A_{\pi_1} \cap A_{\pi_2}$ and $(\phi,\psi) \notin A_{\pi_3}$. Then we would have $\lim_{n\to\infty} \frac{\log(\prod_{b\in\pi_1} f(n,b))}{n} = \lim_{n\to\infty} \frac{\log(\prod_{b\in\pi_2} f(n,b))}{n}$ and $\lim_{n\to\infty}[\frac{\ln(\prod_{b\in\pi_1} f(n,b))}{n} - \frac{\log(\prod_{b\in\pi_3} f(n,b))}{n}] = c > 0$, which implies:

$$\frac{J(t^1,t^2,\pi_3)}{J(t^1,t^2,\pi_1)} \to 0 \quad a.s. \tag{A}$$

To investigate the limit of $\frac{J(t^1,t^2,\pi_1)}{J(t^1,t^2,\pi_2)}$, We use inequalities that $\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n} \leq n! \leq en^{n+\frac{1}{2}}e^{-n}$ holds for all nonnegative integers n. Plug in $f(n,b)$, we have:

$$\beta_b + \log\sqrt{2\pi} - 3 + g(n,b) \leq f(n,b) \leq \beta_b - 2\log\sqrt{2\pi} + g(n,b) \tag{1}$$

$$g(n,b) = (\beta_b+t_b^1-\frac{1}{2})\log(1+\frac{t_b^2}{\beta_b+t_b^1-1})+(\beta_b+t_b^2-\frac{1}{2})\log(1+\frac{t_b^1}{\beta_b+t_b^2-1})-(\beta_b-\frac{1}{2})\log(\beta_b+t_b^1+t_b^2-1)$$

Based on inequalities (1), $\sum_{b\in\pi} f(n,b)$ only differ with $\sum_{b\in\pi} g(n,b)$ by a constant. By Taylor's expansion $\log(1+x) = \log 2 + \frac{1}{2}(x-1) + O((x-1)^2)$, we have $\log(1+\frac{t_b^2}{\beta_b+t_b^1-1}) = \log 2 + \frac{1}{2}(\frac{t_b^1-t_b^2+1-\beta_b}{\beta_b+t_b^1-1}) + O_p((\frac{t_b^1-t_b^2+1-\beta_b}{\beta_b+t_b^1-1})^2)$ and under condition $\Phi_b = \Psi_b$, $\frac{(t_b^1-t_b^2+1-\beta_b)^2}{\beta_b+t_b^1-1}$ is $O_p(1)$. Plug in $g(n,b)$

$$g(n,b) = \log 2 * t_b^1 + \log 2 * t_b^2 - (\beta_b - \frac{1}{2})\log(\beta_b + t_b^1 + t_b^2 - 1) + O_p(1)$$

and sum up

$$\sum_{b\in\pi} g(n,b) = 2n\log 2 - \sum_{b\in\pi}(\beta_b - \frac{1}{2})\log(\beta_b + t_b^1 + t_b^2 - 1) + O_p(1) \tag{2}$$

Notice that when two partition π_1, π_2 have same number of blocks b and $\Phi_b = \Psi_b$, $\forall b \in \pi_1 \cup \pi_2$,

$$\begin{aligned}
\sum_{b\in\pi_1} g(n,b) - \sum_{b'\in\pi_2} g(n,b') &= \sum_{b'\in\pi_2}(\beta'_b - \frac{1}{2})\log(\beta'_b + t_{b'}^1 + t_{b'}^2 - 1) - \sum_{b\in\pi_1}(\beta_b - \frac{1}{2})\log(\beta_b + t_b^1 + t_b^2 - 1) + O_p(1) \\
&= \sum_{b'\in\pi_2}(\beta_{b'} - \frac{1}{2})\log(\frac{\beta'_b + t_{b'}^1 + t_{b'}^2 - 1}{n}) - \sum_{b\in\pi_1}(\beta_b - \frac{1}{2})\log(\frac{\beta_b + t_b^1 + t_b^2 - 1}{n}) \\
&+ \sum_{b'\in\pi_2-\frac{1}{2}}(\beta_{b'} - \frac{1}{2})\log(n) - \sum_{b\in\pi_1-\frac{1}{2}}(\beta_b - \frac{1}{2})\log(n) + O_p(1) \\
&= O_p(1) + \sum_{b\in\pi_1}\frac{1}{2}\log(n) - \sum_{b'\in\pi_2}\frac{1}{2}\log(n) \\
&= O_p(1)
\end{aligned}$$

When π_1 and π_2 have same number of blocks,

$$\frac{J(t^1,t^2,\pi_1)}{J(t^1,t^2,\pi_2)} \to O_p(1) \quad a.s. \tag{B}$$

When π_1 have less blocks than π_2, $\sum_{b'\in\pi_2} g(n,b') - \sum_{b\in\pi_1} g(n,b) = O_p(\log(n))$

$$\frac{J(t^1,t^2,\pi_1)}{J(t^1,t^2,\pi_2)} \to 0 \quad a.s. \tag{C}$$

□

Email address: newton@biostat.wisc.edu

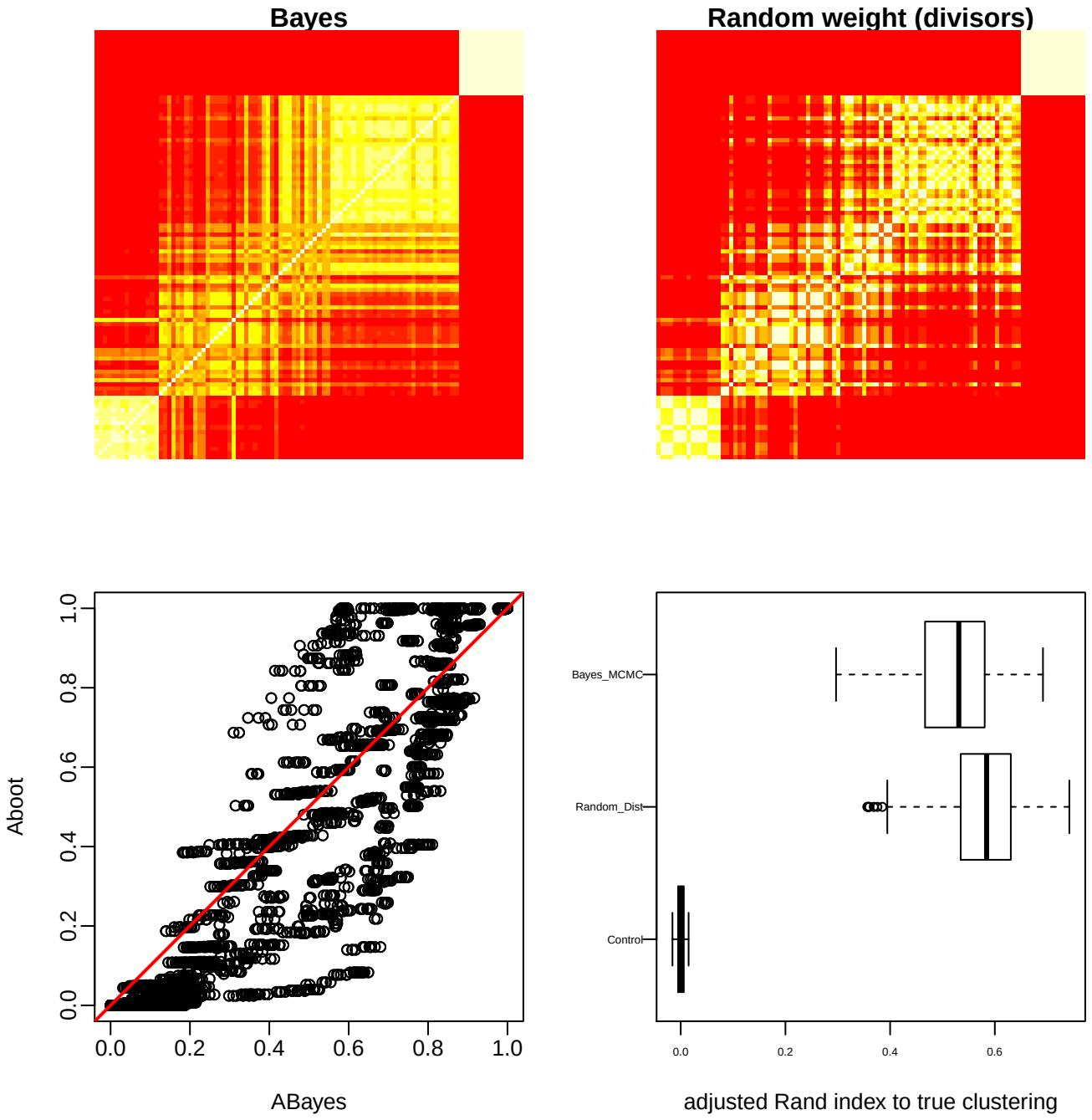

FIGURE 12. comparison between random weighting scheme and bayesian clustering procedure in terms of posterior probabilities that two elements belong to the same class given the whole data and adjusted rand index comparing to the underlying true class label