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

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

1. Introduction

The ability to measure genome-wide gene expression at single-cell resolution has accelerated the pace of biological discovery[1]. Overcoming data analysis challenges caused by the scale and unique variation properties of single-cell data will surely fuel further advances in immunology[2], developmental biology[3], cancer[4], 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 are not adequate for efficient knowledge discovery in this rapidly advancing domain[5].

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, SC3[6], CIDR[7] and ZIFA[8]. Whether or not a determination of cellular subtypes and their frequencies is a task of interest in a given application, we hypothesize that such subtype information may be injected into other inferences in order to improve their operating characteristics. *cumbersome sentence*

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[9], DESEQ2[10], SCDD[11], etc. These tools respond to scRNAseq characteristics, such as high prevelance of zero counts and gene-level multimodality, but none takes explicit advanatage of cellular subtype information. We present a simple procedure and supporting theoretical analyses for this purpose. A notable technical innovation is a new prior distribution over pairs of multinomial probability vectors that conveys both marginal Dirichlet conjugacy as well as dependence induced through sharp equalities on aggregated subtype probabilities, which turns out to be key in formulating the posterior probability of changes in expression distributions between conditions.

DEPARTMENT OF BIOSTATISTICS AND MEDICAL INFORMATICS, UW MADISON, TECHNICAL REPORT TR***-V1, DECEMBER **, 2018.

Using the proposed compositional model, subtypes inferred from whole genome data improve the analysis of gene-level expression. We utilize the mixture of subtypes to characterize transcripts profile and identify differential distributed genes across conditions in an scRNA-seq experiment. Simulation study suggests that the method provides improved power and precision for identifying differentially distributed genes. Performance on empirical data has been investigated through ten previously published experimental data from conquer[12]. We also obtained asymptotic properties of posterior inference.

The proposed scDDboost methodology extends scDD[11], which similarly treats expression data within a condition as a statistical mixture. The extension provided by the present work is to recognize that this mixture is a mixture over cell subtypes. Thus genome-wide data provide information on mixing proportions, thereby providing useful structural information into each gene-level calculation

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 $y_c = 2$ initially: extensions to multiple conditions are discussed in section 6. 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. Further let $t^j_k = \sum_c 1[y_c = j, z_c = k]$ 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, $t^1 = (t^1_1, t^1_2, \cdots, t^1_K)$ and $t^2 = (t^2_1, t^2_2, \cdots, t^2_K)$, as independent multinomial vectors, reflecting the elementary experimental design. Explicitly,

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

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. 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 discussion section, but for now 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 alternative notation, $X_{g,c}|y_c = j \sim f_g^j$ and also $X_{g,c}|z_c = k \sim f_{g,k}$.

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 ([13];[14];[10]). This choice proves to be effective in our numerical experiments though it is not critical to the modeling formulation. **maybe a sentence citing previous papers that use finite mixtures per gene; our angle is to extend by allowing genome-wide data to inform mixing proportions**

We seek a useful 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 $\phi \neq \psi$. Simply, 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. (A second, pedagogical example is presented in the Supplementary Material file.) We formalize the idea 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).

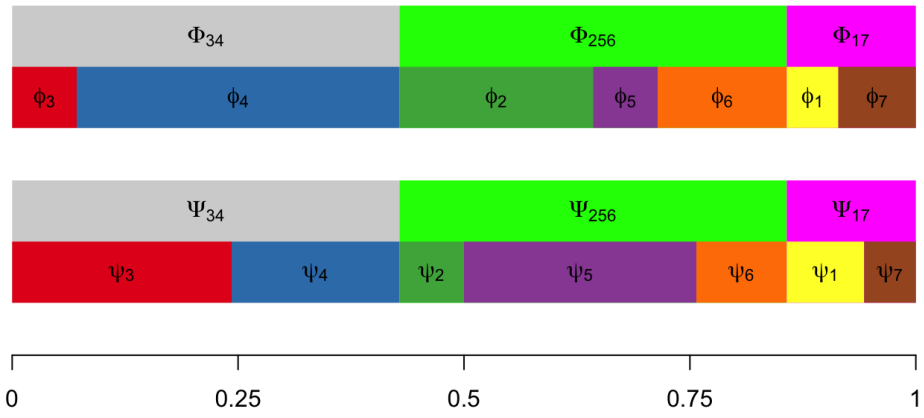

FIGURE 1. Proportions of $K = 7$ cellular subtypes in different conditions. Aggregated proportions of subtypes 3 and 4, subtypes 2, 5, and 6, and subtypes 1, and 7 remain same across conditions, while individual subtype frequencies change.

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 its own block), then the mapping from (ϕ, ψ) to (Φ_π, Ψ_π) is many-to-one. Further, define sets

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

and

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

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:

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

maybe a sentence and citations about lfdr The partition representation guides construction of a prior distribution (Section 3.1) and an empirical Bayesian 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.

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 (1) 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 (Section 5.2). 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 $P\left(A_\pi | y, z\right)$. 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.

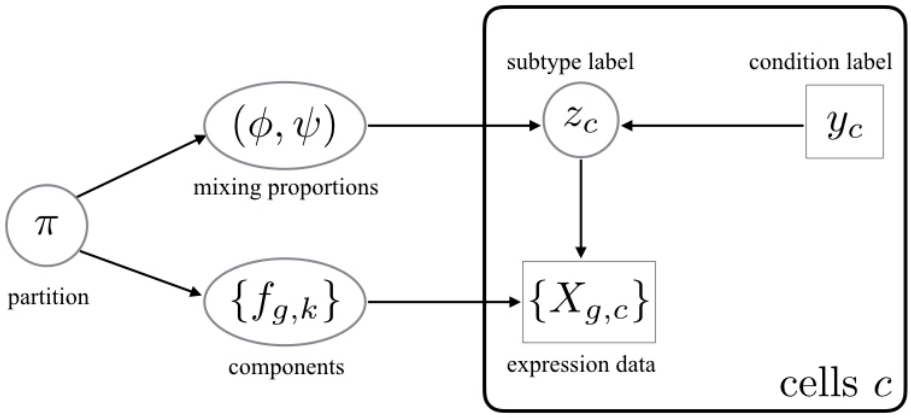

FIGURE 2. Directed acyclic graph structure of compositional model and partition-reliant prior. The plate on the right side indicates i.i.d. copies over cells c, conditionally on mixing proportions and mixing components. Observed data are indicated in rectangles/squares, and unobserved variables are in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

2.2. **Method structure and clustering.** Our approach take transcripts processed by normalizing methods (e.g. SCnorm [15]). The workflow contains two parts, classify cells into subtypes and posterior inference on distributional change. In the first part, recall subtype is a group of cells with distributions of transcripts that are specific to this group, regardless which condition the cells is from. Thus classification process is blind to conditions and can be done by clustering upon similarities between cells(supplementary material).

After identification of subtypes, the second part of our procedure infers patterns of differential expression($M_{g,\pi}$) and aggregated proportions of subtypes(A_π) through empirical Bayes. Specifically, $P(M_{g,\pi}|X, \hat{z})$ is done with EBSeq[13] and we present details of calculating $P(A_\pi|\hat{z})$ in next section. Combining those components, per gene differential distributed probability is obtained.

One advantage of our approach is that the posterior inference can be incorporate with different clustering methods. With the development of technology, clustering methods taking care of newly discovered characteristic of scRNA seq data (e.g. SC3[6], CIDR[7] and ZIFA[8]) could be substituted with our default one. No matter what clustering method is used, we estimate the mixture structure utilizing the whole genome information rather than estimating a gene specific mixture structure solely using information of that gene. Due to this reason, our model is more capable of capturing characteristic of scRNA seq data than scDD and we name our approach SCDDBOOST.

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) $\hat{z}$

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[13]
  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 $\forall g,\ P(\text{DD}_g|X, y, \hat{z}) = 1 - P(\text{ED}_g|X, y, \hat{z})$

We defined distance between cells by weighted average of two metrics. First metrics ensembled gene level information by cluster-based similarity partitioning algorithm([16]) where we applied modal cluster ([17]) for each gene to obtain a partition of cells and defined the dissimilarity between cells as proportion of genes that the two cells are assigned to different clusters. Second metrics defined dissimilarity by 1 - Pearson correlation between cells. There are two reasons we considered these two metrics, first reason is that such cluster based distance is more stable to the outlier than Eculidean distance, faster to compute than many current distance in other clustering method(e.g. sc3 and CIDR). Second reason is that correlation based metrics (e.g Pearson correlation) generally outperformed the distance based metrics (e.g. Euclidean distance)([18]) and considering correlation can add extra information about subtypes between cells. The weights of the two metrics is given by $w_C = \frac{\sigma_C}{\sigma_C + \sigma_P}$ and $w_P = 1 - w_E$. where $w_C, \sigma_C, w_P, \sigma_P$ are the weights and standard deviations of cluster based distance and Pearson correlation distance accordingly. We use K-means to classify cells into subtypes based on the final distance.

Although there are various options for clustering methods, none of them guarantee the accuracy and posterior inference of differential distributed genes can be sensitive to the initial partition. Here we provide a method to make our inferences robust to partitions. Given number of subtypes K, the final posterior probabilities is obtained by averaging results from iterative run of SCDDBOOST-CORE with randomly generated subtype labels z^*. Taking account of information contained in $D = dist(X)$, instead of purely random assigning subtype labels, we generate the distance matrices of cells D^* by dividing weights to the original one and assign labels based on D^*. Specifically, we random sample a noise vector e with length equal to number of cells and components are i.i.d. gamma distributed, then constructing the weighting matrix W by $W_{i,j} = e_i + e_j$ and random distance matrix is obtained by $D^* = D/W$(division is performed componentwisely). The choice of gamma distributed weights and dividing transformation is to matched the probability of two units classified into the same group under random weighting to bayesian framework (supplementary material). After robustification, we select number of clusters K that posterior

probabilities do not vary too much under K and $K+1$. We validate our procedure in simulated and empirical data

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
regularization parameter λ
Output: posterior probabilities of differential distribution
procedure SCDDBOOST(X, y, K, n_r, λ)
distance matrix: $D = \text{dist(X)} \leftarrow$ pairwise distances between cells (columns of X)
repeat
Exponential noise vector: e, with components $\sim \text{Exp}(\lambda)$
randomized distance matrix: $D^* \leftarrow D + e\mathbf{1}^T + \mathbf{1}e^T$
$P^* \leftarrow$ SCDDBOOST-CORE(X, y, D^*, K)
until n_r randomized distance matrices
return $\forall$genes g, $P(\text{DD}_g|X, y) = \frac{1}{n_r}\sum_{D^*} P^*_g$

In general, 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). We view algorithm scDDboost1 as an advanced version of scDDboost0 as it gives improved result depends on finer estimated partition of cells. All the results of scDDboost from simulation and empirical studies are obtained via scDDboost1.

3. Empirical Bayes

3.1. **Double Dirichlet model.** ** ON THE PRIOR INDEPENDENCE** This expresses the idea that subtype proportions (ϕ, ψ) are uninformative about the mean expression levels $\{\mu_{g,i}\}$). Under this assumption: **

Here we describe a prior $p(\phi, \psi)$ that is conjugate to multinomial sampling but that also enables downstream gene-specific inferences about differential distribution when certain cell types do not differ in their expression distributions.

For our purposes, the prior will have a spike-slab structure that mixes over distinct patterns of equality of π-associated accummulated probabilities:

$$p(\phi, \psi) = \sum_{\pi \in \Pi} P(A_\pi)\, p(\phi, \psi | A_\pi)$$

Upon setting up a prior $p(\phi, \psi)$ that can mix over structures A_π, we can obtain posterior inference $P(A_\pi | t^1, t^2)$

Initially, the multitude of $P(A_\pi)$'s will be preset constants.To complete the prior specification $p(\phi, \psi)$, consider further scalers $\alpha_k > 0$ for each class k and $\beta_b > 0$ for each potential block b. (Extending the notational convention, α_b is the vector of α_k for $k \in b$, and β_π is the vector of β_b for $b \in \pi$.) For any block b consider conditional probabilities

$$\tilde{\phi}_b = \frac{\phi_b}{\Phi_b} \qquad \tilde{\psi}_b = \frac{\psi_b}{\Psi_b}$$

which indicate the conditional probability of each class k given that the cell is of one of the types in b. Assume that conditional upon A_π,

$$\Phi_\pi \sim \text{Dirichet}_{N(\pi)}[\beta_\pi]$$

where $N(\pi)$ is the number of blocks b in π, and further that accumulated probabilities are the same between the two source conditions: $\Phi_\pi = \Psi_\pi$. Finally, assume that for each $b \in \pi$,

$$\tilde{\phi}_b, \tilde{\psi}_b \sim_{\text{i.i.d.}} \text{Dirichlet}_{N(b)}[\alpha_b]$$

where $N(b)$ is the number of cell types in block b. In other words, if A_π is the active structure, then accumulated probability vectors Φ_π and Ψ_π are equal between the two source conditions, though the sub-block class-specific rates ϕ_k and ψ_k may differ, as would (re-normalized) independent Dirichlet-distributed vectors. Taken together,

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

with

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

and

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

3.2. **Predictive probabilities:** For notation, we use ϕ_b for the vector of values ϕ_k for $k \in b$, and similarly for ψ_b. Analogously, Φ_π and Ψ_π are vectors of accumulated class probabilities ϕ_b and ψ_b for all $b \in \pi$, respectively.
In order to get the posterior probability $p(A_\pi | t^1, t^2)$, we need to calculate

$$\begin{aligned} p(A_\pi | t^1, t^2) \propto p(A_\pi, t^1, t^2) &= \int_{A_\pi} p(t^1, t^2 | \phi, \psi) p(\phi, \psi) d\phi d\psi \\ &= \sum_{\pi' \in \Pi} \int_{A_\pi} p(t^1, t^2 | \phi, \psi) p(\phi, \psi | A_{\pi'}) p(A_{\pi'}) d\phi d\psi \end{aligned}$$

For simplicity of notation, let $w(\pi_1, \pi_2) = \int_{A_{\pi_1}} p(t^1, t^2|\phi, \psi)p(\phi, \psi|A_{\pi_2})p(A_{\pi_2})d\phi d\psi$, then $p(A_\pi|t^1, t^2) \propto \sum_{\pi' \in \Pi} w(\pi, \pi')$. To calculate component $w(\pi, \pi')$, recall refinement and coarseness relationship between partitions, we say a partition $\tilde{\pi}$ is a refinement of another partition π if $\forall b \in \pi$ there exists $s \subset \tilde{\pi}$ such that $\bigcup_{b' \in s} b' = b$. We say π is a coarseness of $\tilde{\pi}$ when $\tilde{\pi}$ refines π. we have following theorem

Theorem 3. *If π' is a refinement of π then $w(\pi, \pi') = w(\pi', \pi')$ otherwise $w(\pi, \pi') = 0$*

Consequently, let $RF(\pi)$ be the collection of finer partition of π, we have the posterior probability:

$$p(A_\pi|t^1, t^2) \propto \sum_{\pi' \in RF(\pi)} w(\pi', \pi')$$

Using the Dirichlet-Multinomial conjugacy and the collapsing property of these distributions ([19]), we get closed formulas for the predictive probability of cell-type counts t^1 and t^2. Fixing π, let $t^j_b = \sum_{k \in b} t^j_k$, for cell conditions $j = 1, 2$, record the total numbers of cells accumulated over all types in block b. And following our notation convention, t^j_π is the vector of these counts over $b \in \pi$. From the prior and model structure

$$w(\pi, \pi) = p(t^1|t^1_\pi)\, p(t^2|t^2_\pi)\, p(t^1_\pi, t^2_\pi|A_\pi)p(A_\pi).$$

Conditional independence of t^1 and t^2 given the block-level totals t^1_π and t^2_π on A_π reflects the possible differential class proportion structure within blocks but between cell conditions. For either cellular group $j = 1, 2$, we find, after some simplification, the following Dirichlet-Multinomial masses:

$$p(t^j|t^j_\pi) = \prod_{b \in \pi} \left\{ \left[\frac{\Gamma(t^j_b + 1)}{\prod_{k \in b} \Gamma(t^j_k + 1)} \right] \left[\frac{\Gamma(\sum_{k \in b} \alpha_k)}{\prod_{k \in b} \Gamma(\alpha_k)} \right] \left[\frac{\prod_{k \in b} \Gamma(\alpha_k + t^j_k)}{\Gamma(t^j_b + \sum_{k \in b} \alpha_k))} \right] \right\} \tag{2}$$

and

$$p(t^1_\pi, t^2_\pi|A_\pi) = \left[\frac{\Gamma(n_1 + 1)\Gamma(n_2 + 1)}{\prod_{b \in \pi} \Gamma(t^1_b + 1)\Gamma(t^2_b + 1)} \right] \left[\frac{\Gamma(\sum_{b \in \pi} \beta_b)}{\prod_{b \in \pi} \Gamma(\beta_b)} \right] \left[\frac{\prod_{b \in \pi} \Gamma(\beta_b + t^1_b + t^2_b)}{\Gamma(n_1 + n_2 + \sum_{b \in \pi} \beta_b)} \right]. \tag{3}$$

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 the second formula reduces, correctly, to $p(t^1_\pi, t^2_\pi|A_\pi) = 1$. Further,

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

which is the well-known Dirichlet-multinomial predictive distribution for counts t^j [20]. E.g, taking $\alpha_k = 1$ for all types k we get the uniform distribution

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

Case 2. At the opposite extreme, π has one block b for each class k. Then $t^j_b = z^j_k$, and $p(t^j|t^j_\pi) = 1$, and further, assuming $\beta_b = \alpha_k$,

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

which corresponds to Dirichlet-multinomial predictive distribution for counts $t^1 + t^2$ since t^1 and t^2 are identical distributed in this case.

Regardless of the partition, log scale probabilities are readily evaluated given hyper-parameters $\{\alpha_k\}$ and $\{\beta_b\}$ and for cell-type counts t^1 and t^2.
For asymptotic properties of the posterior probabilities, we demonstrated them in section 6.

4. Operating Characteristics

4.1. **Splatter Simulation.** A simulation study was conducted to assess the performance of scDDboost in identifying DD genes. We simulate data by splatter[21] with approximate 200 cells each condition and 7 subtypes with proportions ϕ and ψ from Fig 1 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 subtypes are well separated, which is driven by genes with heterogeneous distribution between subtypes.

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

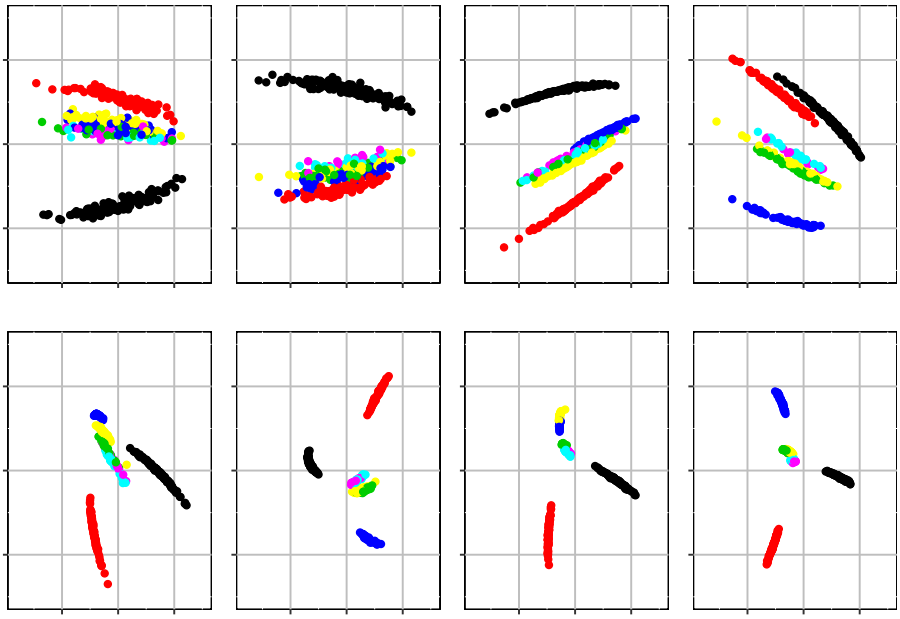

FIGURE 3. first two principal components of transcripts under different parameters for simulated data. Different parameters resulted in different degree of separation of subtypes.

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)

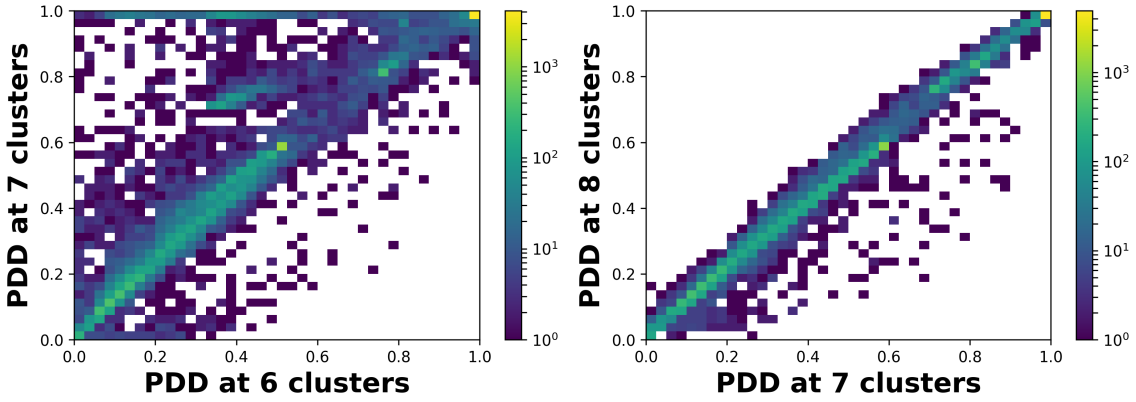

FIGURE 4. comparison of posterior probabilities of being DD among different number of subtypes, when we underestimate the number of subtypes, the difference is huge, see PDD between 6 subtypes and 7 subtypes. There is an approximate horizontal line with massive points at the top of left panel, which indicate that we underestimate lots of DD genes due to underestimate the number of subtypes. While in the case when we overestimate the number of subtypes 7 subtypes vs. 8 subtypes, though inflating PDD but the variation of difference is small, from 6 to 8 subtypes the PDD become more linear related.

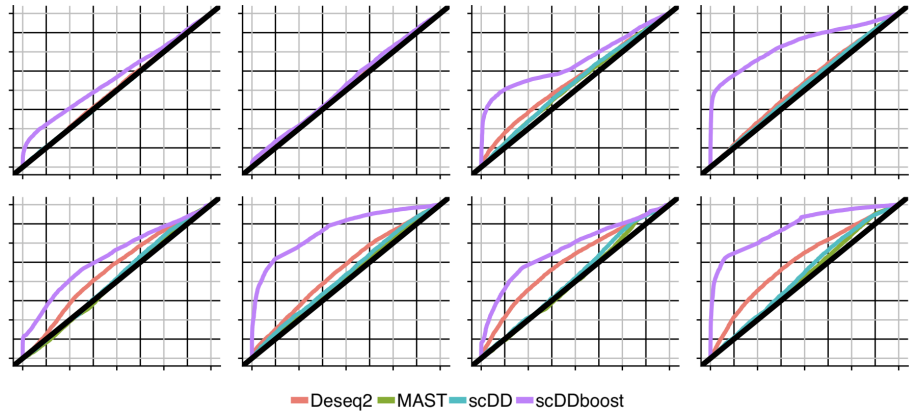

FIGURE 5. Roc curve of scDDboost, scDD, MAST and DESeq2, scDDboost has largest area under the roc curve. Roc curves of other three methods are similar. For those roc curve there is bigger difference at low level of false positive rate, as scDDboost identified twice many true DD genes as other methods.

Since we are modeling gene transcript within each subtype as negative binomial 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.

4.2. **Empirical study.** **Conquer examples...null and alternative**

We use ten datasets from conquer[12] to test performance of our method on empirical data. We compare our results with scDD[11], MAST[9] and DESeq2[10], we have also investigated performance of scDDboost under different clustering method, here sc3[6] is used to couple with scDDboost0, as sc3 did not provide user accessible distance matrix.

We have table of numbers of differentially expressed genes of each dataset by MAST and DESeq2, and numbers of differentially distributed genes of each dataset by scDDboost and scDD.

TEMPORARY..for JSM

Data set	Conditions	Number of cells/condition	Organism
GSE74596	NKT1 vs NKT2	46, 68	mouse
GSE63818-GPL16791	Primordial Germ Cells, developmental stage: 7 week gestation vs Somatic Cells, developmental stage: 7 week gestation	40,26	mouse
GSE48968-GPL13112	BMDC (2h LPS stimulation) vs BMDC(6h LPS stimulation)	96, 96	mouse
GSE45719	16-cell stage blastomere vs Mid blastocyst cell (92-94h post- fertilization)	50, 60	mouse
GSE52529	T0 vs T24	96,96	human
GSE60749	serum+LIF vs 2i+LIF	90,94	mouse
EMTAB2805	G1 vs G2M	96,96	mouse
GSE71585-GPL13112	Gad2tdTpositive vs Cux2tdTnegative	80, 140	mouse
GSE79102	patient1 vs patient2	51, 89	human
GSE57872	patient1 vs patient2	192, 96	human
GSE64016	G1 vs G2	91, 76	human
GSE75748	DEC vs EC	64, 64	human

TABLE 1. single cell transcripts profiles used for differential expression or distribution method evaluation

***

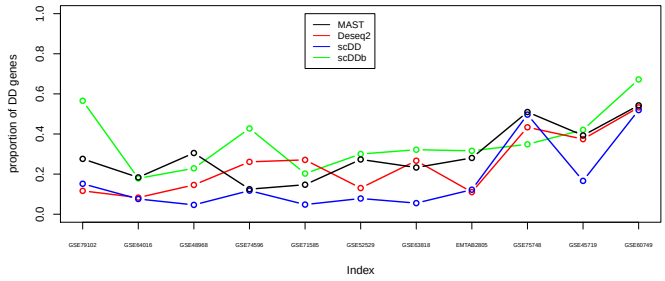

FIGURE 6. proportion of DD genes with respect to total number of genes identified by each method. Ranked by mean list size

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.

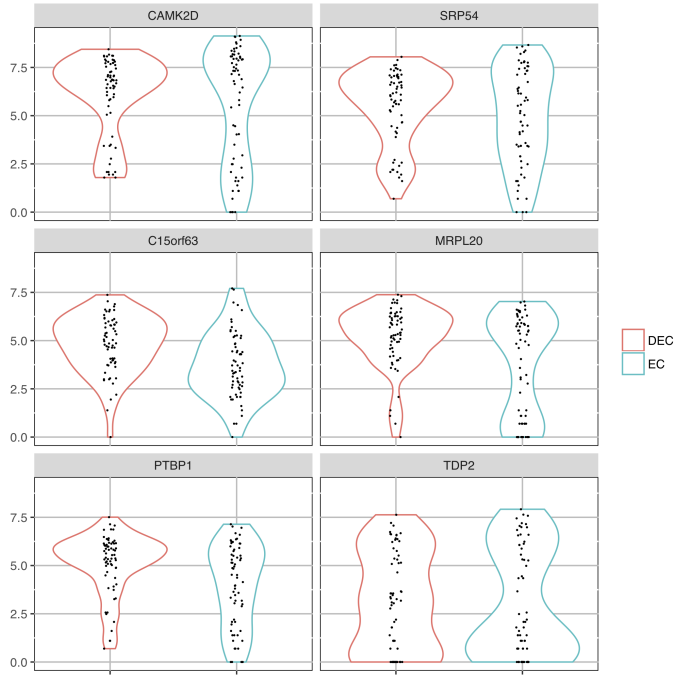

FIGURE 7. Densities of log transformed transcripts, 6 DD genes uniquely identified by scDDboost, for data GSE75748, DEC vs. EC, We observe some of the genes are different distributed across conditions.

4.3. **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	mouse
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 2. 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

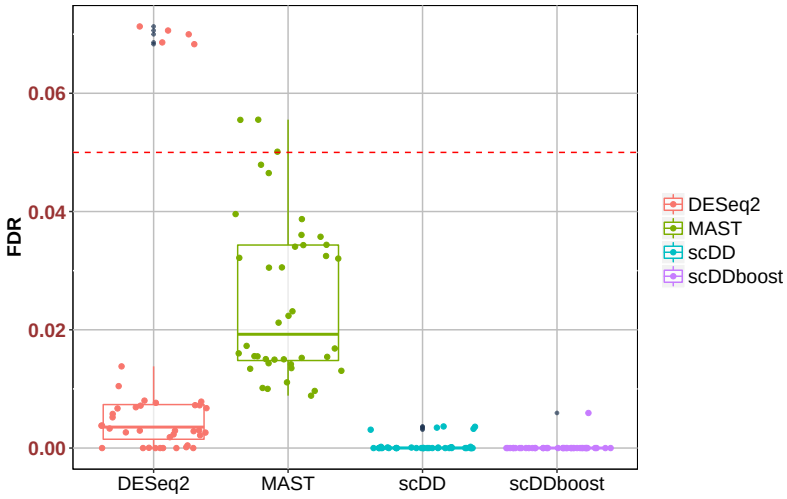

FIGURE 8. FDR of scDDboost, scDD, MAST and DESeq2 on null dataset from table 1, DESeq2 usually identify a lot but may lose the control of type I error. While other single cell methods could control FDR.

4.4. **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.

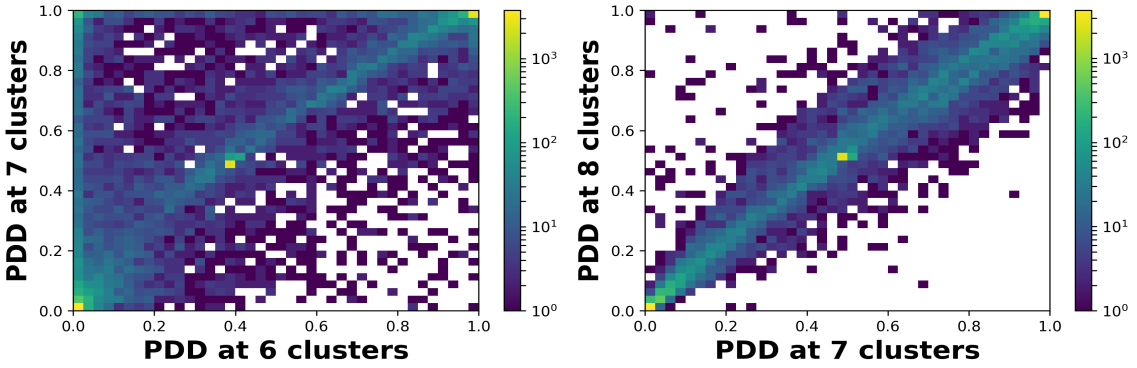

FIGURE 9. selecting number of subtypes for data GSE75748, we observe posterior probabilities become stable at more than 6 subtypes. Since increasing number of subtypes tends to decrease sample size of each subtypes, make complicate constraints for equivalent distribution and inflate estimated PDD. We select number of subtypes to be 7

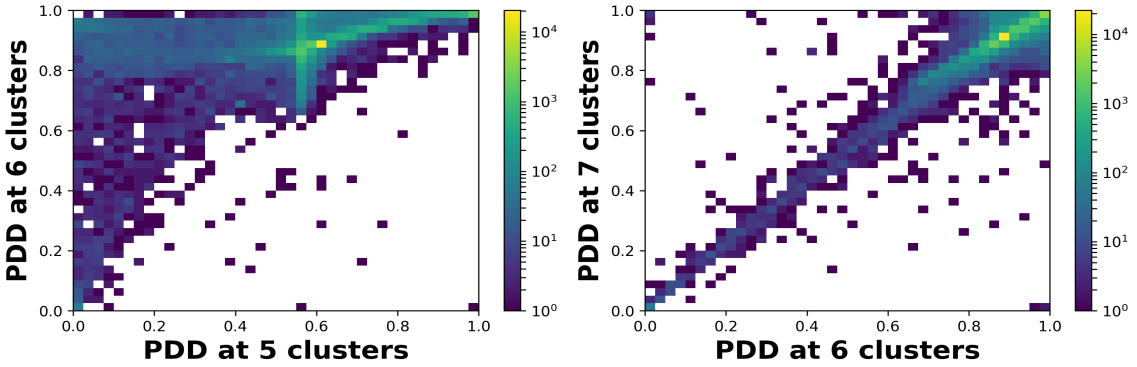

FIGURE 10. selecting number of subtypes for data GSE48968, we observe posterior probabilities become stable at more than 5 subtypes

4.5. **Bursting parameters.** D3E[22] 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 EMTAB2805

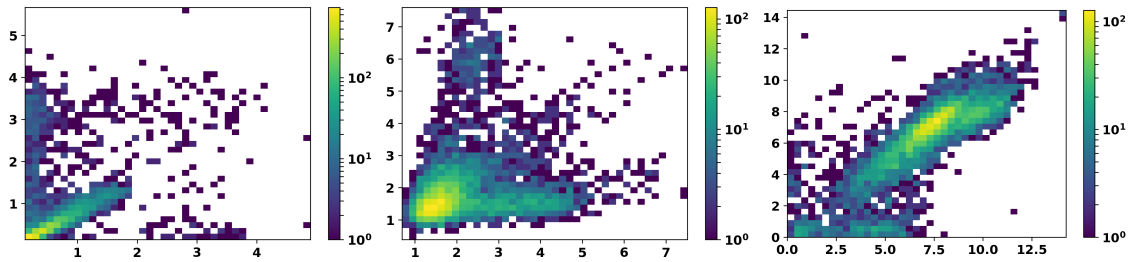

FIGURE 11. 2D histogram for bursting parameters of DD genes identified by scDDboost from dataset EMTAB2805 estimated by D3E. Left panel : comparison of rate of promoter activation between two conditions, similarly, middle panel : rate of promoter inactivation and right panel: rate of transcription when the promoter is in the active state. We observe that difference between transcription rate is smaller compare to difference between the activation and inactivation rate.

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.

5. Theoretical issues

5.1. **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

$$p(A_\pi|t^1,t^2) = c * \sum_{\pi' \text{ refines } \pi} p(t^1|t^1_{\pi'})\, p(t^2|t^2_{\pi'})\, p(t^1_{\pi'}, t^2_{\pi'}|A_{\pi'}) \tag{4}$$

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

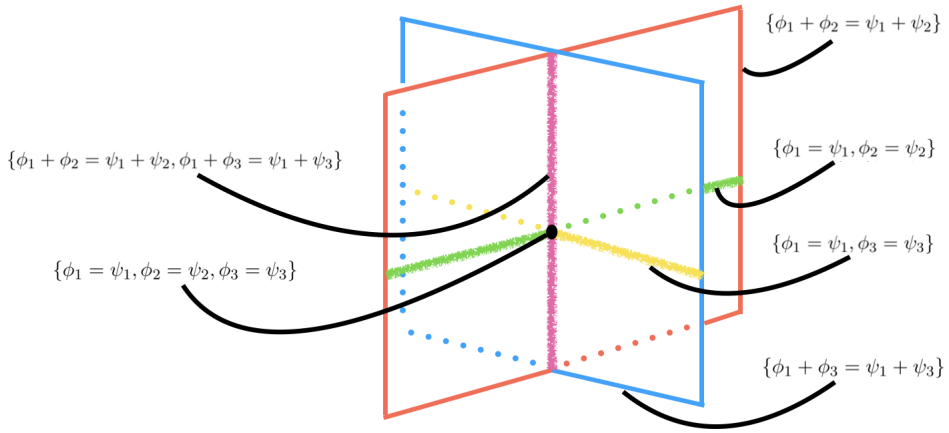

FIGURE 12. Four subtypes of cells, simplexes of (ϕ, ψ) satisfying different constraints.

$A_{\pi_1} = \{(\phi,\psi) : \phi_1+\phi_2 = \psi_1+\psi_2\}$, the rectangle with blue boundary is a simplex $A_{\pi_2} = \{(\phi,\psi) : \phi_1+\phi_3 = \psi_1+\psi_3\}$. 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 lemma 1, $A_{\tilde{\pi}} \cap A_\pi$ 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 $\alpha_i = 1, \forall i$ in (2) and $\beta_b = \Sigma_{i \in b} \alpha_i$ in (3), plug in (4) then we have simplified

$$p(A_\pi|t^1, t^2) = \frac{1}{c'} \sum_{\pi' \in \mathrm{RF}(\pi)} \prod_{b \in \pi'} \frac{\Gamma(\beta_b + t_b^1 + t_b^2)}{\Gamma(\beta_b + t_b^1)\Gamma(\beta_b + t_b^2)} \tag{5}$$

$c' = c/\frac{\Gamma(n+1)\Gamma(n+1)\Gamma(K)}{\Gamma(2n+K)}$ 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, scDDboost 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.

5.2. **Random weighting.**

6. Discussion and future work

We have presented scDDboost, a compositional model for the analysis of scRNA-seq data. scDDboost make whole genome information shared at gene level inference and is flexible of integrating with different cell clustering methods. scDDboost accounts for the over-dispersion and multi-modality of single-cell data by modeling expression data as mixture of subtypes. Information from the change of subtypes' proportions and mean expressions is combined to infer distributional changes of gene expression. We developed a prior distribution that can be used to estimate sharp equalities on aggregated subtype probabilities. Using simulation studies, we have demonstrated that our method outperforms existing approaches when there existing multiple cell types.

6.1. **Multi-conditions.** One direction for future work is to extend the compositional model to multi-conditions case. As scRNA seq data is a time course data, it may be more interested to consider inference of distributional changes across multiple conditions. We outline our procedure in the following.
Assume there are$T$ conditions, K subtypes, Let f_g^c denotes distribution of expression of gene g at condition c. Partition of conditions are $\pi^1, ..., \pi^{n_T}$ with superscripts, and $\pi_1, ..., \pi_{n_K}$ with subscripts refer to partition of subtypes. Then, given a partition of conditions, say π^1, the corresponding distribution change of a gene g is $D_g^{\pi^c} = \{\forall b \in \pi^c, f_g^i = f_g^j, f_g^i \neq f_g^m \forall i, j \in b, \forall m \notin b\}$. Similar in two conditions case, to express distributional change in terms of parameters change, we consider parameter space

$$\Theta = \{\phi^1, ..., \phi^T, \mu^1, ..., \mu^K\}$$

where ϕ^l is proportions of subtypes at condition l and $\mu_k = \{\mu_{g,k}\}$ is gene-specific expected values at subtype k.

Define

$$A_{\pi_k}^{\pi^l} = \{\forall b \in \pi^l, \forall i \in b, \forall m \notin b, \exists b' \in \pi_k, \sum_{s\in b'} \phi_s^i \neq \sum_{s\in b'} \phi_s^m; \forall b' \in \pi_k, \forall i,j \in b. \sum_{s\in b'} \phi_s^i = \sum_{s\in b'} \phi_s^j\}$$

Indeed, $A_{\pi_k}^{\pi^l}$ and M_{g,π_k} are precisely the structures needed to address differential distribution $D_g^{\pi^l}$

Theorem 6. *If $\pi^l \neq \pi^{MF} = \{\{1\}, ..., \{T\}\}$, the most refined partition. Then*

$$D_g^{\pi^l} = \begin{cases} \bigcup\limits_{\pi_k \neq \pi_0} A_{\pi_k}^{\pi^l} \cap M_{g,\pi_k} & \text{if } \pi^l \neq \pi^0 \\ \bigcup\limits_{\pi_k} A_{\pi_k}^{\pi^l} \cap M_{g,\pi_k} & \text{if } \pi^l = \pi^0 \end{cases}$$

where $\pi_0 = \{1, 2, ..., K\}, \pi^0 = \{1, 2, ..., T\}$ are the most coarse partitions.
If $\pi^l = \pi^{MF}$, which corresponds to the case that g in each condition has a distinguished distribution. Then

$$D_g^{\pi^l} = \bigcap_{\pi^j \neq \pi^{MF}} [M_{g,\pi_i}^c \bigcup_{\pi_i} (A_{\pi_i}^{\pi^j})^c]$$

Where A^c refers to the complement set of A

We prioritize $A_{\pi_k}^{\pi^l}$ via following procedure:

$$\phi^b = \{\phi^k, k \in b\} \tag{6}$$

$$P(\phi | A_{\pi_k}^{\pi^l}) = \prod_b P(\phi^b | A_{\pi_k}^{\pi^l}) \tag{7}$$

$$P(\phi^b | A_{\pi_k}^{\pi^l}) = P(\Phi | A_{\pi_k}^{\pi^l}) P(\tilde{\phi} | A_{\pi_k}^{\pi^l}) \tag{8}$$

Where $\Phi = \{\Phi_{b'}, b' \in \pi_k\}, \Phi_{b'} = \sum\limits_{i\in b'} \phi_i^j, j \in b$, $\Phi_{b'} \sim$ dirichlet(β) and $\tilde{\phi} = \{\tilde{\phi}^j, j \in b\}, \tilde{\phi}^j = (\tilde{\phi}_1^j, ..., \tilde{\phi}_K^j), \tilde{\phi}_i^j = \frac{\phi_i^j}{\Phi_b'}$, for $i \in b'$

Since there are a lot of overlapping among $A_{\pi_k}^{\pi^l}$, care must be taken to account for competing issue of prior predictive function. Empirical study may give suggestion of modifying the prior.

6.2. **Randomized distance.** From simulation and empirical studies, we obtained stabilized posterior probabilities by averaging results on random generated distance matrices. It remains unclear how clustering on randomized distance matrices approximate the true partition. We seek a theoretical justification of the benefit from randomized distance.

References

[1] T. Nawy, "Single-cell sequencing," *Nature Methods*, vol. 11, pp. 18 EP –, 12 2013. [Online]. Available: http://dx.doi.org/10.1038/nmeth.2771

[2] E. Papalexi and R. Satija, "Single-cell rna sequencing to explore immune cell heterogeneity," *Nature Reviews Immunology*, vol. 18, pp. 35 EP –, 08 2017. [Online]. Available: http://dx.doi.org/10.1038/nri.2017.76

[3] J. C. Marioni and D. Arendt, "How single-cell genomics is changing evolutionary and developmental biology," *Annual Review of Cell and Developmental Biology*, vol. 33, no. 1, pp. 537–553, Oct 2017, pMID: 28813177. [Online]. Available: https://doi.org/10.1146/annurev-cellbio-100616-060818

[4] N. E. Navin, "The first five years of single-cell cancer genomics and beyond," *Genome Research*, vol. 25, no. 10, pp. 1499–1507, 10 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4579335/

[5] R. Bacher and C. Kendziorski, "Design and computational analysis of single-cell rna-sequencing experiments," *Genome Biology*, vol. 17, no. 1, p. 63, 2016. [Online]. Available: https://doi.org/10.1186/s13059-016-0927-y

[6] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, and M. Hemberg, "Sc3: consensus clustering of single-cell rna-seq data," *Nature Methods*, vol. 14, pp. 483 EP –, 03 2017. [Online]. Available: http://dx.doi.org/10.1038/nmeth.4236

[7] P. Lin, M. Troup, and J. W. K. Ho, "Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data," *Genome Biology*, vol. 18, no. 1, p. 59, 2017. [Online]. Available: https://doi.org/10.1186/s13059-017-1188-0

[8] E. Pierson and C. Yau, "Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis," *Genome Biology*, vol. 16, no. 1, p. 241, 2015. [Online]. Available: https://doi.org/10.1186/s13059-015-0805-z

[9] G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, P. S. Linsley, and R. Gottardo, "Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data," *Genome Biology*, vol. 16, no. 1, p. 278, 2015. [Online]. Available: https://doi.org/10.1186/s13059-015-0844-5

[10] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome Biology*, vol. 15, no. 12, p. 550, 2014. [Online]. Available: https://doi.org/10.1186/s13059-014-0550-8

[11] K. D. Korthauer, L.-F. Chu, M. A. Newton, Y. Li, J. Thomson, R. Stewart, and C. Kendziorski, "A statistical approach for identifying differential distributions in single-cell rna-seq experiments," *Genome Biology*, vol. 17, no. 1, p. 222, 2016. [Online]. Available: https://doi.org/10.1186/s13059-016-1077-y

[12] C. Soneson and M. D. Robinson, "Bias, robustness and scalability in differential expression analysis of single-cell rna-seq data," *bioRxiv*, 2017. [Online]. Available: https://www.biorxiv.org/content/early/2017/05/28/143289

[13] N. Leng, J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. G. Smits, J. D. Haag, M. N. Gould, R. M. Stewart, and C. Kendziorski, "Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments," *Bioinformatics*, vol. 29, no. 8, pp. 1035–1043, 2013. [Online]. Available: + http://dx.doi.org/10.1093/bioinformatics/btt087

[14] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, pp. R106–R106, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3218662/

[15] R. Bacher, L.-F. Chu, N. Leng, A. P. Gasch, J. A. Thomson, R. M. Stewart, M. Newton, and C. Kendziorski, "Scnorm: robust normalization of single-cell rna-seq data," *Nature Methods*, vol. 14, pp. 584 EP –, 04 2017. [Online]. Available: http://dx.doi.org/10.1038/nmeth.4263

[16] A. Strehl and J. Ghosh, "Cluster ensembles — a knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Mar. 2003. [Online]. Available: https://doi.org/10.1162/153244303321897735

[17] D. B. Dahl, "Modal clustering in a class of product partition models," *Bayesian Anal.*, vol. 4, no. 2, pp. 243–264, 06 2009. [Online]. Available: https://doi.org/10.1214/09-BA409

[18] T. Kim, I. R. Chen, Y. Lin, A. Y.-Y. Wang, J. Y. H. Yang, and P. Yang, "Impact of similarity metrics on single-cell rna-seq data clustering," *Briefings in Bioinformatics*, p. bby076, 2018. [Online]. Available: http://dx.doi.org/10.1093/bib/bby076

[19] B. Dickey J., Lientz, "The weighted likelihood ratio, sharp hypotheses, and the order of a markov chain." *Ann. Math. Statist.*, vol. 41, no. 1, p. 214, 1970. [Online]. Available: https://projecteuclid.org/euclid.aoms/1177697203

[20] U. Wagner and A. Taudes, "A multivariate polya model of brand choice and purchase incidence," *Marketing Science*, vol. 5, no. 3, pp. 219–244, Aug. 1986. [Online]. Available: http://dx.doi.org/10.1287/mksc.5.3.219

[21] L. Zappia, B. Phipson, and A. Oshlack, "Splatter: simulation of single-cell rna sequencing data," *Genome Biology*, vol. 18, no. 1, p. 174, 2017. [Online]. Available: https://doi.org/10.1186/s13059-017-1305-0

[22] M. Delmans and M. Hemberg, "Discrete distributional differential expression (d3e) - a tool for gene expression analysis of single-cell rna-seq data," *BMC Bioinformatics*, vol. 17, no. 1, p. 110, 2016. [Online]. Available: https://doi.org/10.1186/s12859-016-0944-6

Appendix A

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$(\phi), t^2 \sim$ 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 $\frac{\partial^2 C}{\partial\Phi_b\partial\Phi_{b'}} = 0$, 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 $\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}\le n!\le en^{n+\frac{1}{2}}e^{-n}$ holds for all nonnegative integers n. Plug in $f(n,b)$, we have:

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

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

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 $\log(1+x)=\log2+\frac{1}{2}(x-1)+O((x-1)^2)$, we have $\log(1+\frac{t_b^2}{\beta_b+t_b^1-1})=\log2+\frac{1}{2}(\frac{t_b^1-t_b^2+1-\beta_b}{\beta_b+t_b^1-1})+O_p((\frac{t_b^1-t_b^2+1-\beta_b}{\beta_b+t_b^1-1})^2)$ 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 $g(n,b)$

$$g(n,b)=\log2*t_b^1+\log2*t_b^2-(\beta_b-\frac{1}{2})\log(\beta_b+t_b^1+t_b^2-1)+O_p(1)$$

and sum up

$$\sum_{b\in\pi}g(n,b)=2n\log2-\sum_{b\in\pi}(\beta_b-\frac{1}{2})\log(\beta_b+t_b^1+t_b^2-1)+O_p(1) \tag{2}$$

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, $\sum_{b' \in \pi_2} g(n, b') - \sum_{b \in \pi_1} g(n, b) = O_p(\log(n))$

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

□

E-mail address: newton@biostat.wisc.edu