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

XIUYU MA, AND 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 (*cites*). Overcoming data analysis challenges caused by the scale and unique variation properties of single-cell data will surely fuel further advances in immunology (cite), developmental biology (cite), and cancer (cite). 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 (Bacher and Kendziorski, 2016?; Gottardo?).

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 (e.g. burst states), or other distinguishing factors. [**something about sc methods concerned a lot with clustering cells into different cell subtypes/subpopulations...tsne...]. Whether or not a determination of cellular subtypes and their frequencies is a task of interest in a given application, we hypothesize that such subytpe information may be injected into other inferences 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 ([10]), DESEQ2 ([11]), SCDD ([9]), other. 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. **what else do we do...test on a bunch of data sets... find improved sensitivity sometimes??**

DEPARTMENT OF BIOSTATISTICS AND MEDICAL INFORMATICS, UW MADISON, TECHNICAL REPORT TR***-V1, DECEMBER **, 2017.

2. Modeling

2.1. **Data structure, sampling model, and parameters.** In modeling scRNASeq data, we imagine that each cell falls into one of $K > 1$ classes, which we think of as subtypes or subpopulations of cells. Knowledge of this class structure prior to measurement is not required, as it will be inferred as necessary from available genomic data. We also assume that cells arise from multiple experimental conditions, such as by treatment-control status or some other factor measured at the cell level, and we present our development for the special case of two conditions, noting in Section ** how to proceed more generally. Let's say conditions 1 and 2 contain n_1 and n_2 cells, respectively, and let z_k^j denote the number of cells of subtype k in condition j. Count vectors $z^1 = (z_1^1, z_2^1, \cdots, z_K^1)$ and $z^2 = (z_1^2, z_2^2, \cdots, z_K^2)$ we treat as independent multinomial vectors, reflecting the common, two-condition experimental design. Explicitly,

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

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 $n_1 + n_2$ observed cells are sampled. As for data, the normalized expression of gene g in cell c, say $X_{g,c}$, is one entry in a typically large data matrix; and we record cell condition with the binary label y_c.

Our working hypothesis is that any differences in the distribution of $X_{g,c}$ 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. On the other hand, we could then refine the subtype definition to allow more population classes K in order to mitigate that problem. We revisit the issue in Section **, but for now proceed assuming that cellular condition affects the composition of subtypes but not the distribution of expression within a subtype.

With this working hypothesis, let $f_{g,k}(x)$ denote the sampling distribution of expression measurement $X_{g,c}$ assuming that cell c is from subtype k. Then in the two cellular conditions, the marginal distributions over subtypes are

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

We say that gene g is *differentially distributed*, denote DD_g, 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 [cites, including Leng et al 2013?]. This choice

proves to be effective in our numerical experiments though it is not critical to the modeling formulation.

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. We formalize this idea in order that our methodology has the necessary functionality. First, consider the parameter space

$$\Theta = \{(\phi, \psi, \mu, \sigma)\}$$

where $\phi = (\phi_1, \phi_2, \cdots, \phi_K)$ and $\psi = (\psi_1, \psi_2, \cdots, \psi_K)$, as before, where $\mu = \{\mu_{g,k}\}$, all the subtype-and-gene-specific expected values, and where $\sigma = \{\sigma_g\}$ holds all the gene-specific Negative binomial shape parameters. We define special subsets of Θ using 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 $\pi = \{b\}$. We recall that the set Π containing all partitions π of $\{1, 2, \cdots, K\}$ has cardinality that grows rapidly with K. We'll 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 $\pi = \{b\}$ we have aggregate subtype frequencies

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

We'll also use the notation $\Phi_\pi = \{\Phi_b : b \in \pi\}$ and similarly for Ψ_π. As long as π is not the most refined partition, the mapping from (ϕ, ψ) to (Φ_π, Ψ_π) is many-to-one (Figure 2). Define

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

and

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

Indeed, these are precisely the structures needed to address differential distribution DD_g (and it complement, equivalent distribution, ED_g) at a given gene g:

Theorem 1. *At a given gene, equivalent distribution is*

$$ED_g = \bigcup_{\pi \in \Pi} \left[A_\pi \cap B_\pi \right].$$

**

2.2. **Clustering method.** To identify subtypes of cells, we pool cells from two biological conditions. At each gene level, we do a Poison-Gamma model extended from modal clustering[1]. After gene level clustering, we use the cluster-based similarity partition algorithm (CSPA[2]). For each individual clustering result, a binary similarity matrix is constructed from the corresponding cell labels: if two cells belong to the same cluster, their similarity is 1; otherwise their similarity is 0. We obtain a consensus similarity matrix $M1$ by averaging all similarity matrices of individual clusterings. Another distance matrix $M2$ calculated by Pearson distance between cells. A final similarity matrix is obtained by weighted combining $M1$ and $M2$. Cells are classified into subtypes by K-means clustering based the final similarity matrix.

2.3. **Empirical Bayes prior.** ** Here I 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, and by combining cell-type counts z^1 and z^2 with expression data x_g at a gene, we may compute $P(\mathrm{ED}_g|\mathrm{data})$ at each gene:

$$P(\mathrm{ED}_g|\mathrm{data}) = \sum_{\pi \in \Pi} P(A_\pi | z^1, z^2)\, P(B_\pi | x_g).$$

To discuss the issue of overlapping of parameter space 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 subspace that we do not give inference due to lack of proper prior to make an explicit prior predictive function, we only consider inference on space $\Omega \setminus N$, $N = \{(\phi, \psi), s.t. \exists b_1, b_2 \subset \{1, 2, ..., K\}, b_1 \neq b_2, b_1 \cap b_2 \neq \emptyset \text{ and } \sum_{i \in b_j} \phi_i = \sum_{i \in b_j} \psi_i, j = 1, 2\}$
Next, we define the refinement 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 $\cup_{x \in s} x = b$. Consequently, if $\tilde{\pi}$ refines π we would have $A_{\tilde{\pi}} \subset A_\pi$. Finally, we introduce a thinning constrained parameter space $A_\pi^* = A_\pi \setminus \cup_{\{\tilde{\pi}, \pi \text{ not refines } \tilde{\pi}\}} A_{\tilde{\pi}}$, that is parameter in A_π^* only satisfy constraints given by A_π and does not satisfy any further constraints.
The reason to introduce such notations is to resolve definition issue of $\mathrm{P}(\phi, \psi | A_\pi)$ when (ϕ, ψ) satisfy further constraints, e.g. $(\phi, \psi) \in A_{\tilde{\pi}}$ for a $\tilde{\pi}$ refines π and $A_{\tilde{\pi}}$ is a lower dimensional subset of A_π, naturally, we have $\mathrm{P}((\phi, \psi) \in A_{\tilde{\pi}} | A_\pi) = 0$, then $\mathrm{P}((\phi, \psi) \in A_{\tilde{\pi}}) \propto \mathrm{P}((\phi, \psi) \in A_{\tilde{\pi}} | A_\pi) * \mathrm{P}(A_\pi) = 0$. Instead of defining $\mathrm{P}(\phi, \psi | A_\pi)$, we separate space $\Omega \setminus N$ into disjoint subspace and define $\mathrm{P}(\phi, \psi | A_\pi^*)$ below .

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

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

2.4. **Predictive and posterior 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.

Using the Dirichlet-Multinomial conjugacy and the collapsing property of these distributions ([3]), we get closed formulas for the predictive probability of cell-type counts z^1 and z^2. Fixing π, let $t^j_b = \sum_{k \in b} z^j_k$, 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

$$p(z^1, z^2 | A^*_\pi) = p(z^1 | t^1_\pi)\, p(z^2 | t^2_\pi)\, p(t^1_\pi, t^2_\pi | A^*_\pi).$$

Conditional independence of z^1 and z^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:

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

and

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

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

Check 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,

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

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

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

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

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

[4]

Regardless of the partition, log scale probabilities are readily evaluated given hyper-parameters $\{\alpha_k\}$ and $\{\beta_b\}$ and for cell-type counts z^1 and z^2.

After finishing calculating $p(z^1, z^2|A^*_\pi)$, we obtain posterior $p(A^*_\pi|z^1, z^2)$ by assign weakly informative prior of A^*_π ($p(A^*_\pi)$ is constant not depend on A^*_π) and obtain $p(A_\pi|z^1, z^2)$ by summing over all the posterior units of refinements of π.

2.5. **asymptotic properties.**

Theorem 2. *assuming $\beta_b = \Sigma_{k\in b}\alpha_k$, and $\alpha_k = 1, \forall k$, let $n = min(n_1, n_2)$ be the smaller one of number of cells of two conditions and $\frac{n_1}{n_2} = C(C! = 0)$ when $(\phi, \psi) \in \Omega \setminus N$ we have*

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

Theorem 3. *when* $(\phi, \psi) \in N$, *and* $S = \{\pi, (\phi, \psi) \in A_\pi\}$, *we have*

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

and $\sum_{\pi \in S_\pi} m(\pi) = 1$

proofs are in the appendix.

3. Data analysis workflow

3.1. **simulated data.**

Here's an example using the probabilities ϕ and ψ from Figure 2; We simulate data by

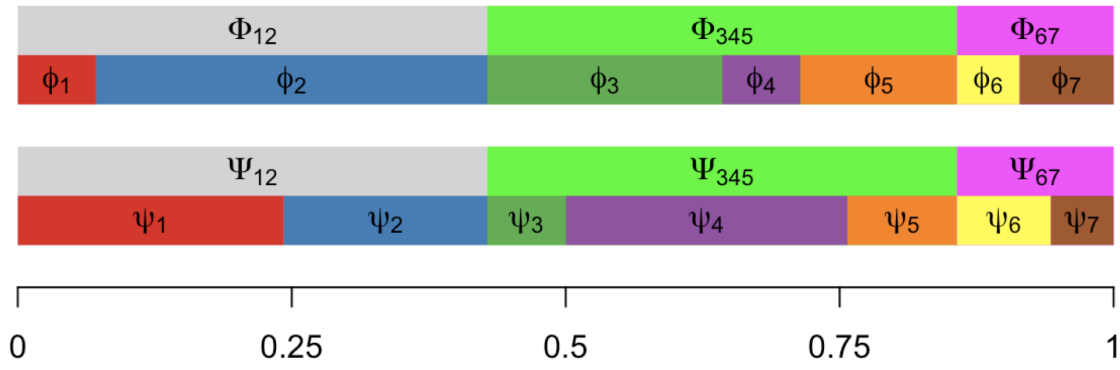

Figure 1. proportion change in different conditions.

splatter[5] with $n_1 = n_2 = 200$, and 7 subtypes among two conditions with proportional constraints: $\phi_1 + \phi_2 = \psi_1 + \psi_2$, $\phi_3 + \phi_4 + \phi_5 = \psi_3 + \psi_4 + \psi_5$ and $\phi_6 + \phi_7 = \psi_6 + \psi_7$

3.2. **The scDDboost modeling framework.**

We normalized raw transcripts count data by SCnorm[6] to adjust for technical sources of variation including amplification bias and sequencing depth. We use the clustering method in section 2 to classify cells into subtypes.

After clustering of cells, we obtain posterior inference on differential expression pattern via EBSeq[7] and posterior inference on proportion change via the method in section 2 by assuming $\alpha_k = 1$ for all k and $\beta_b = \sum_{k \in b} \alpha_k$..

In the simulation data, there are 7 groups and 10% genes each group are DE genes. There are in total 9067 DD genes and 8306 ED genes.

Below are numbers of DD or DE genes identified by four methods with target FDR at 5 %.

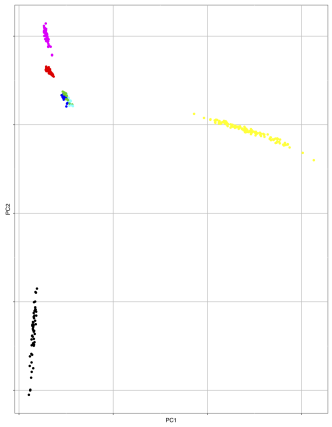

FIGURE 2. pca of cells

	scDDboost	scDD	MAST	DESeq2
DD or DE genes	6073	3038	2468	2076
True positive	4774	2724	2442	2073
false positive	1299	314	26	3

And we compare roc curves of scDDboost, scDD, MAST and DESeq2. (figure 3)

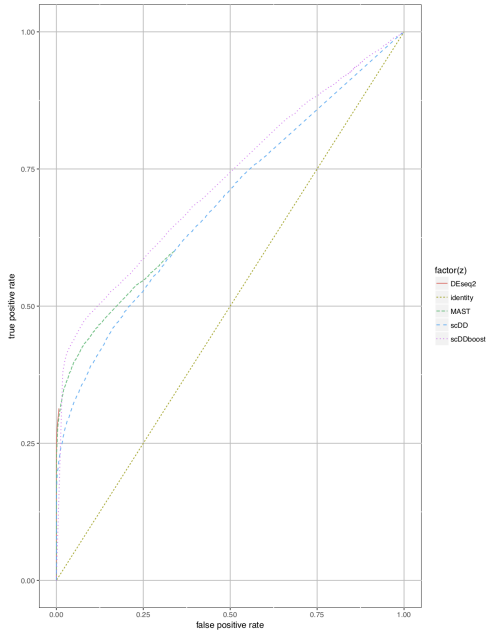

FIGURE 3. roc curve of scDDboost, scDD, MAST and DESeq2

scDDboost may work poorly, when ϕ and ψ satisfy partial overlapping constraints, i.e. there are at least a pair of $b_1, b_2 \in \{1, 2, ..., K\}$ that $b_1 \neq b_2$, $b_1 \cap b_2 \neq \emptyset$ and $\sum_{i \in b_j} \phi_i = \sum_{i \in b_j} \psi_i, j = 1, 2$. For example, parameter space that$\phi_1 + \phi_2 = \psi_1 + \psi_2$ and $\phi_1 + \phi_3 = \psi_1 + \psi_3$. Based on the theorem 3, assume we have sufficiently large n cells. When constraints are partially overlapping, we may underestimate the posterior probability of true proportion pattern, which reduce the posterior probabilities of true negative and enlarge false positive rate.

Another situation that the power of scDDboost could be limited is that even though there is no mean expression change among subtypes but the distribution among subtypes changed, EBSeq would fail to detect the discrepancies between subtypes, thus reduce power of detecting DD genes.

4. Examples

We use ten datasets from conquer[8] to test performance of our method on real data. We compare our results with scDD[9], MAST[10] and DESeq2[11]

Data set	Compared cell subsets	Number of cells per condition	Organism	Ref
GSE45719	16-cell stage blastomere vs Mid blastocyst cell (92-94h post-fertilization)	50, 60	mouse	[12]
GSE45719null	16-cell stage blastomere	50	mouse	[12]
GSE48968-GPL13112	BMDC (1h LPS stimulation) vs BMDC(4h LPS stimulation)	96, 95	mouse	[13]
GSE48968-GPL13112null	BMDC (1h LPS stimulation)	96	mouse	[13]
GSE60749-GPL13112	v6.5 mouse embryonic stem cells, culture conditions: 2i+LIF vs v6.5 mouse embryonic stem cells, culture conditions: serum+LIF	90, 94	mouse	[14]
GSE60749-GPL13112null	v6.5 mouse embryonic stem cells, culture conditions: 2i+LIF	90	mouse	[14]
GSE74596	NKT0 vs NKT17	45,44	mouse	[15]
GSE74596null	NKT0	45	mouse	[15]
EMTAB2805	G1 vs G2m	96,96	mouse	[16]
EMTAB2805null	G1	96	mouse	[16]
GSE63818-GPL16791	Primordial Germ Cells, developmental stage: 7 week gestation vs Somatic Cells, developmental stage: 7 week gestation	39,26	mouse	[17]
GSE71585-GPL13112	Chrna2 tdTpositive vs Cux2 tdTpositive	84, 124	mouse	[18]
GSE71585-GPL13112null	Chrna2 tdTpositive	84	mouse	[18]
GSE75748	NPC vs DEC	64, 87	human	[19]
GSE75748	NPC	64	human	[19]
GSE75748	DEC vs EC	70, 64	human	[19]
GSE75748	DEC	70	human	[19]
GSE64016null	H1 vs H9	64, 87	human	[20]

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.

Data set	scDDboost	scDDboost-sc3	scDD	MAST	DESeq2	total number of genes
GSE45719	4278	4228	6416	5652	11202	45686
GSE48969-GPL13112	11691	9819	2080	3396	9542	45686
GSE60749-GPL13112	19215	19168	18074	13674	23178	45686
GSE74596	1942	1353	1099	540	3796	45686
EMTAB 2805	1194	3748	760	1088	5391	45686
GSE63818-GPL16791	3948	3480	1365	873	8934	45686
GSE71585-GPL13112	2902	1460	1622	2572	7378	24057
NPC-DEC	3237	3211	5982	6666	8439	19037
DEC-EC	3461	3023	3818	5429	8127	19037
H1 exp1-H1 exp2	0	0	1300	2077	2841	16579

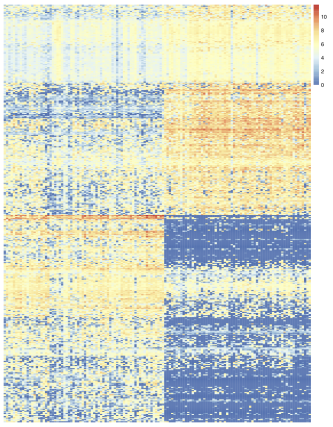

FIGURE 4. heatmap of log transformed transcripts DD genes uniquely identified by scDDboost

We validate false discovery rate on ten null datasets from the table. For each null dataset, we randomly split the cells from one condition into two equal sized subsets and do DE analysis between those subsets for five times . 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 (figure 5).

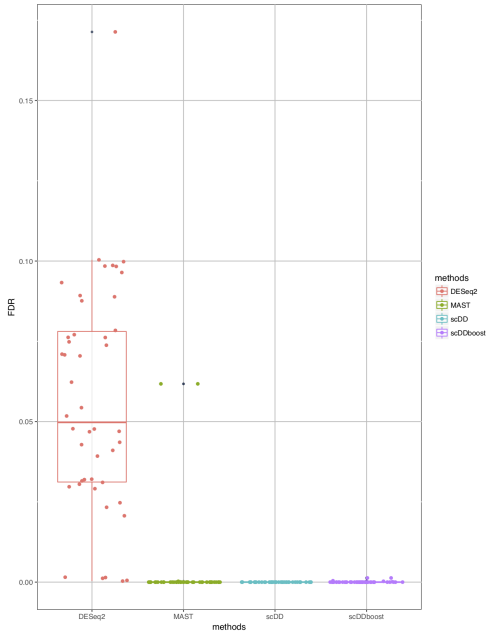

FIGURE 5. FDR of scDDboost, scDD, MAST and DESeq2

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, proportions of subtypes remain unchanged among the two subsets.

D3E[21] 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

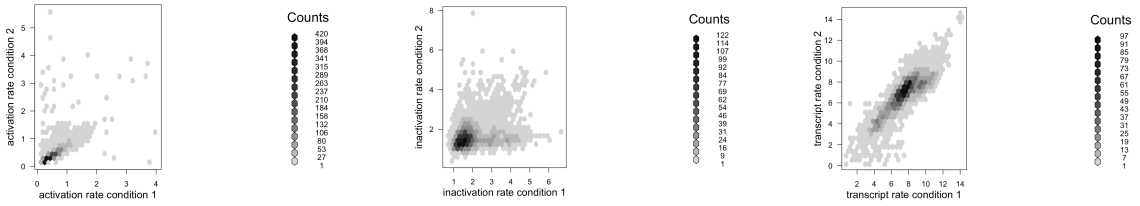

FIGURE 6

We observed that DD genes identified by our methods are driven by the change of activation and inactivation rates.

5. Discussion

Cluster cells is an unsupervised learning, we do not know the true underlying partition and different number of clusters will lead to large differences in posterior probabilities of genes being differentially distributed (figure 7). We propose a modified bootstrap to stabilize our inferences. Instead of resample the cells, we resample the distance matrices of cells by adding noises to original distance matrix. Denote the original distance matrix as $D = (d_{i,j})$, for each time we random sample a vector e with length equal to number of cells and components are i.i.d. exponentially distributed. let w be the standard deviation of $d_{i,j}$. We resample a new $\hat{D}$ by adding nosies: $\hat{d}_{i,j} = d_{i,j} + e_i * w + e_j * w$. For $\hat{D}$ we still have triangle inequality of distance holds as $\hat{d}_{i,j} + \hat{d}_{j,k} \geq \hat{d}_{i,k}$, it is a valid distance matrix. Following our procedure, given one distance matrix of cells we

Appendix A

$\frac{p(A_{\pi_1^*}|z^1,z^2)}{p(A_{\pi_2^*}|z^1,z^2)} \to O(1) \quad a.s.$

proof of theorem 1 and theorem 2: given the condition that $\alpha_k = 1, \forall k$ and $\beta_b = \sum_{k\in b}\alpha_k$, we can simplify $p(z^1, z^2|A_\pi^*)$ and obtain

$$p(z^1, z^2|A_\pi^*) = \prod_{b\in\pi} \frac{\Gamma(\beta_b + t_b^1 + t_b^2)}{\Gamma(\beta_b + t_b^1)\Gamma(\beta_b + t_b^2)} \frac{\Gamma(n+1)\Gamma(n+1)\Gamma(K)}{\Gamma(2n+K)}$$

Assuming there are in total K subgroups, since n_1 and n_2 goes to infinite at same rate, for simplicity we assume $n = \sum_{i=1}^K z_i^1 = \sum_{i=1}^K z_i^2$, $z^1 \sim$ multinomial(ϕ), $z^2 \sim$ 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

$$p(z^1, z^2|A_\pi^*) \propto \prod_{b\in\pi} f(n,b)$$

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

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

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

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

To find the maxima (Φ, Ψ), we fix Ψ and let $C = \frac{\ln(\prod_{b\in\pi} f(n,b))}{n} + \lambda(\sum_b \Phi_b - 1)$, we have $\frac{\partial C}{\partial \Phi_b} = \ln(1 + \frac{\Psi_b}{\Phi_b}) + \lambda$, 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 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 $\Phi = \Psi$, $\frac{\ln(\prod_{b\in\pi} f(n,b))}{n} = 2\ln(2)$ 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 $\lim_{n\to\infty} \frac{\ln(\prod_{b\in\pi_1} f(n,b))}{n} = \lim_{n\to\infty} \frac{\ln(\prod_{b\in\pi_2} f(n,b))}{n}$ and $\lim_{n\to\infty}[\frac{\ln(\prod_{b\in\pi_1} f(n,b))}{n} - \frac{\ln(\prod_{b\in\pi_3} f(n,b))}{n}] = c > 0$, which implies:

$$\frac{p(A_{\pi_3^*}|z^1,z^2)}{p(A_{\pi_1^*}|z^1,z^2)} \to 0 \quad a.s.$$

To investigate the limit of $\frac{p(A_{\pi_1^*}|z^1,z^2)}{p(A_{\pi_2^*}|z^1,z^2)}$, 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:

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

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

By Taylor's expansion $\ln(1+x) = \ln 2 + \frac{1}{2}(x-1) + O((x-1)^2)$, we have $\ln(1+\frac{t_b^2}{\beta_b+t_b^1-1}) = \ln 2 + \frac{1}{2}(\frac{t_b^1-t_b^2+1-\beta_b}{\beta_b+t_b^1-1}) + O_p((\frac{t_b^1-t_b^2+1-\beta_b}{\beta_b+t_b^1-1})^2)$ and under condition $\Phi_b = \Psi_b$, $\frac{t_b^1-t_b^2+1-\beta_b}{\beta_b+t_b^1-1}$ is $O_p(\frac{1}{n})$. Plug in $g(n,b)$

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

and sum up

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

we examine log fold change of mean expression across conditions

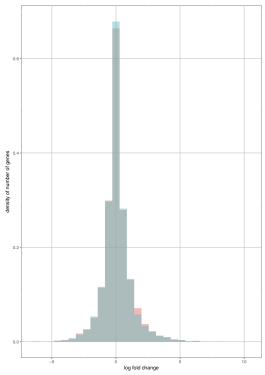

FIGURE 7. MAST

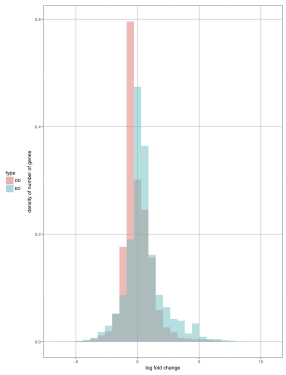

FIGURE 8. scDD

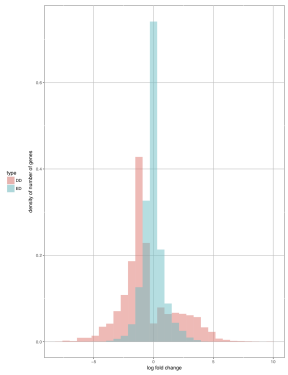

FIGURE 9. scDDboost

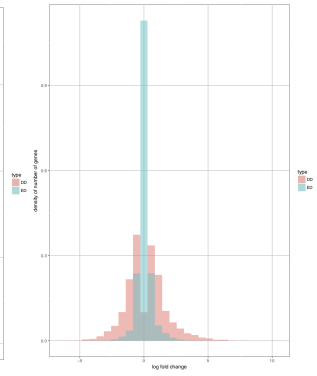

FIGURE 10. DESeq2

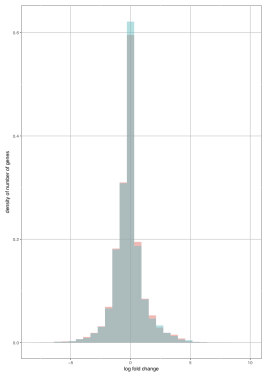

FIGURE 11. MAST

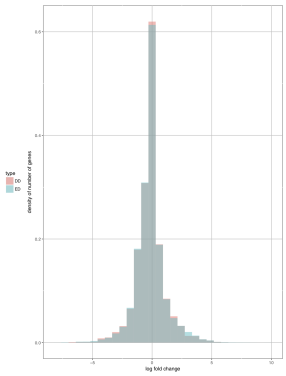

FIGURE 12. scDD

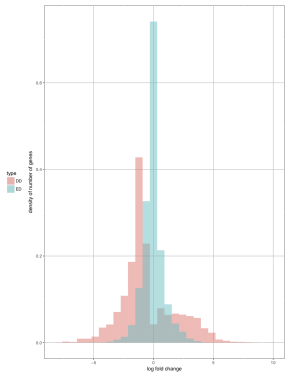

FIGURE 13. scDDboost

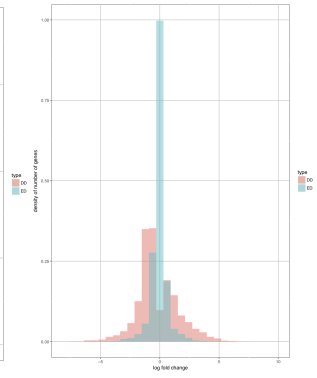

FIGURE 14. DESeq2

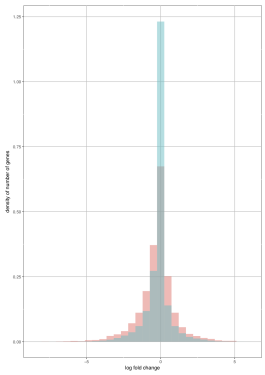

FIGURE 15. MAST

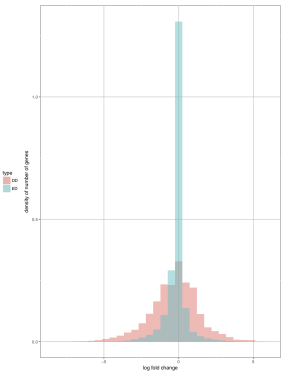

FIGURE 16. scDD

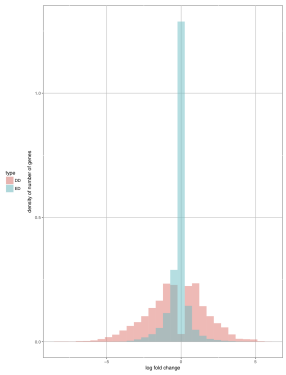

FIGURE 17. scDDboost

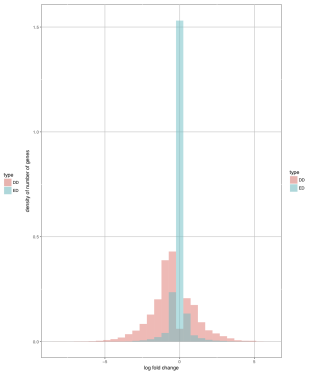

FIGURE 18. DESeq2

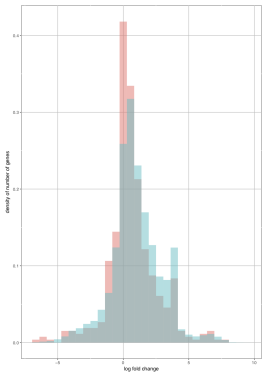

FIGURE 19. MAST

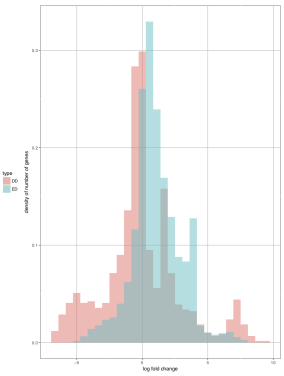

FIGURE 20. scDD

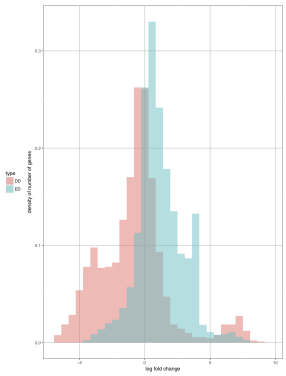

FIGURE 21. scDDboost

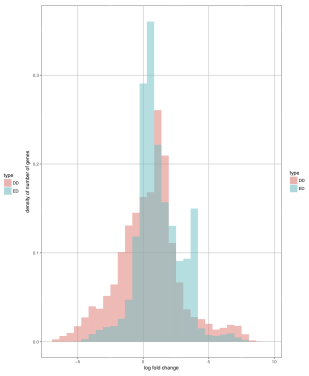

FIGURE 22. DESeq2

References

[1] 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

[2] 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

[3] 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

[4] H. I. Weisberg, "Bayesian comparison of two ordered multinomial populations," *Biometrics*, vol. 28, no. 3, pp. 859–867, 1972. [Online]. Available: http://www.jstor.org/stable/2528768

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] Q. Deng, D. Ramsköld, B. Reinius, and R. Sandberg, "Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian

cells," *Science*, vol. 343, no. 6167, pp. 193–196, 2014. [Online]. Available: http://science.sciencemag.org/content/343/6167/193

[13] A. K. Shalek, R. Satija, J. Shuga, J. J. Trombetta, D. Gennert, D. Lu, P. Chen, R. S. Gertner, J. T. Gaublomme, N. Yosef, S. Schwartz, B. Fowler, S. Weaver, J. Wang, X. Wang, R. Ding, R. Raychowdhury, N. Friedman, N. Hacohen, H. Park, A. P. May, and A. Regev, "Single-cell rna-seq reveals dynamic paracrine control of cellular variation," *Nature*, vol. 510, pp. 363 EP –, 06 2014. [Online]. Available: http://dx.doi.org/10.1038/nature13437

[14] R. M. Kumar, P. Cahan, A. K. Shalek, R. Satija, A. Jay DaleyKeyser, H. Li, J. Zhang, K. Pardee, D. Gennert, J. J. Trombetta, T. C. Ferrante, A. Regev, G. Q. Daley, and J. J. Collins, "Deconstructing transcriptional heterogeneity in pluripotent stem cells," *Nature*, vol. 516, pp. 56 EP –, 12 2014. [Online]. Available: http://dx.doi.org/10.1038/nature13920

[15] I. Engel, G. Seumois, L. Chavez, D. Samaniego-Castruita, B. White, A. Chawla, D. Mock, P. Vijayanand, and M. Kronenberg, "Innate-like functions of natural killer t cell subsets result from highly divergent gene programs," *Nature Immunology*, vol. 17, pp. 728 EP –, 04 2016. [Online]. Available: http://dx.doi.org/10.1038/ni.3437

[16] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, "Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells," *Nature Biotechnology*, vol. 33, pp. 155 EP –, 01 2015. [Online]. Available: http://dx.doi.org/10.1038/nbt.3102

[17] F. Guo, L. Yan, H. Guo, L. Li, B. Hu, Y. Zhao, J. Yong, Y. Hu, X. Wang, Y. Wei, W. Wang, R. Li, J. Yan, X. Zhi, Y. Zhang, H. Jin, W. Zhang, Y. Hou, P. Zhu, J. Li, L. Zhang, S. Liu, Y. Ren, X. Zhu, L. Wen, Y. Q. Gao, F. Tang, and J. Qiao, "The transcriptome and dna methylome landscapes of human primordial germ cells," *Cell*, vol. 161, no. 6, pp. 1437–1452, 2017/12/05. [Online]. Available: http://dx.doi.org/10.1016/j.cell.2015.05.015

[18] B. Tasic, V. Menon, T. N. Nguyen, T. K. Kim, T. Jarsky, Z. Yao, B. Levi, L. T. Gray, S. A. Sorensen, T. Dolbeare, D. Bertagnolli, J. Goldy, N. Shapovalova, S. Parry, C. Lee, K. Smith, A. Bernard, L. Madisen, S. M. Sunkin, M. Hawrylycz, C. Koch, and H. Zeng, "Adult mouse cortical cell taxonomy revealed by single cell transcriptomics," *Nature Neuroscience*, vol. 19, pp. 335 EP –, 01 2016. [Online]. Available: http://dx.doi.org/10.1038/nn.4216

[19] L.-F. Chu, N. Leng, J. Zhang, Z. Hou, D. Mamott, D. T. Vereide, J. Choi, C. Kendziorski, R. Stewart, and J. A. Thomson, "Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm," *Genome Biology*, vol. 17, no. 1, p. 173, 2016. [Online]. Available: https://doi.org/10.1186/s13059-016-1033-x

[20] N. Leng, L.-F. Chu, C. Barry, Y. Li, J. Choi, X. Li, P. Jiang, R. M. Stewart, J. A. Thomson, and C. Kendziorski, "Oscope identifies oscillatory genes in unsynchronized single-cell rna-seq experiments," *Nature Methods*, vol. 12, pp. 947 EP –, 08 2015. [Online]. Available: http://dx.doi.org/10.1038/nmeth.3549

[21] 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

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