Supplementary Material

Main document: *A compositional model to assess expression changes from single-cell RNA-seq data*

Authors: Ma, Korthauer, Kendziorski, and Newton

Version: May 5, 2019

This supplement is organized to match the sectioning of the main document. In summary,

1. Introduction
   - R package
2. Modeling
   - Data Structure, Sampling Model, and Parameters
     - Proof of Theorem 2
   - Method Structure and Clustering
     - `EBSeq`
     - `modalClust`
     - Randomized $K-$means
     - Selecting K
   - Double Dirichlet Mixture
     - Proof of Properties 1-6 and Theorem 3
3. Numerical Experiments
   - Synthetic data, `splatter`
   - Empirical study, `conquer` and Null case
   - Robustness
4. Posterior consistency
   - proof of theorem 4

1. Introduction.

1.1. *R package.* Reference can be found at github site ...

on scDDboost, web page, etc

2. Modeling.

2.1. *Data Structure, Sampling Model, and Parameters.* Proof of Theorem 2.

PROOF. Recall $\theta = (\phi, \psi, \mu, \sigma)$. Through the sampling procedure of our model (Figure 3), assuming a known number of cells within each condition (n_1, n_2). We have z^1, z^2 are multinomial draw given ϕ and ψ, thus the generation of y, z only depends on (ϕ, ψ). Also given z, $X_{g,c}$ is sampled through $\mathrm{NB}(\mu_{g,z_c}, \sigma_g)$, and only depends on (μ, σ). Thus $P(X, y, z|\theta) = P(y, z|\phi, \psi)P(X|z, \mu, \sigma)$, and we independently give priors for (μ, σ) and (ϕ, ψ). By the Baye's rule,

$$\begin{aligned}
P(\theta|X, y, z) &\propto P(X, y, z|\theta)P(\theta) \\
P(X, y, z|\theta)P(\theta) &= P(y, z|\phi, \psi)P(X|z, \mu, \sigma)P(\mu, \sigma|z)P(\phi, \psi) \\
P(\phi, \psi|y, z) &\propto P(y, z|\phi, \psi)P(\phi, \psi) \\
P(\mu, \sigma|X, z) &\propto P(X|z, \mu, \sigma)P(\mu, \sigma|z) \\
\text{Thus } P(\theta|X, y, z) &\propto P(\phi, \psi|y, z)P(\mu, \sigma|X, z)
\end{aligned}$$

From this we know
1. Given X, y and z, $(\phi, \psi) \perp (\mu, \sigma)$
2. Given condition and subtypes label y, z, $(\phi, \psi) \perp$X
3.*GivenXandz*,$(\mu, \sigma) \perp$y
Thuswehave $\mathrm{P}\left(A_\pi \cap M_{g,\pi}|\mathrm{X}, y, z\right) = P\left(A_\pi|y, z\right)\, P\left(M_{g,\pi}|\mathrm{X}, z\right).$ □

2.2. *Method Structure and Clustering.*

2.2.1. *EBSeq.* In this subsection, we go through how we implemented and modified `EBSeq` to get $P(M_{g,\pi}|X, z)$.

Suppose we have K subtypes, let $X_g^I = X_{g,1}^I, ..., X_{g,S_1}^I$ denote transcripts at gene g from subtype $I, I = 1, ...K$. The `EBSeq` model assumes that counts within subtype I are distributed as Negative Binomial: $X_{g,s}^I|r_{g,s}, q_g^I \sim NB(r_{g,s}, q_g^I)$. Due to sample-specific size factor in the raw counts, r is made sample-specific. However, since we are dealing with normalized counts rather than raw counts as in `EBSeq`, we instead make r shared at gene level across all samples, i.e. $X_{g,s}^I|\sigma_g, q_g^I \sim NB(\sigma_g, q_g^I)$

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

and $\mu_{g,s}^I = \sigma_g(1 - q_g^I)/q_g^I$; For ease in later deriving the density kernel f, we use q rather than μ to parameterize the NB.

Following EBSeq, we assume a prior distribution on q_g^I : $q_g^I|\alpha, \beta^g \sim Beta(\alpha, \beta^g)$. The hyperparameter α is shared by the whole genome and β^g is gene-specific.

We force the size factor to be 1 for all cells and use the same procedure as EBSeq to estimate the failure parameter σ_g. Namely, we have

1. gene-level sample mean $m_g = \frac{1}{n}\sum_{s=1}^{n} X_{g,s}$, where $n = n_1 + n_2$ is the total number of cells
2. average of sample variances over subtypes $v_g = \frac{1}{K}\sum_{I=1}^{K} v_g^I$.
3. v_g^I is the unadjusted sample variance for subtype I, i.e. $v_g^I = \frac{1}{n^I}\sum_{s,z_s=I}(X_{g,s} - m_g^I)^2$ where m_g^I is the sample mean within subtype I and n^I is the number of cells within subtype I.

We estimate the pooled over-dispersion rate by $o_g = \frac{v_g}{m_g}$ and obtain $\sigma_g = m_g \frac{o_g}{1-o_g}$ from the first moment of NB.
Our target is to quantify the expression pattern between K groups,

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

For example, if $K = 3$, there are 5 expression patterns: $P_1, P_2, ..., P_5$. Comparison between μ is equivalent to comparison between q.

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

Under the assumption that two groups I and J share the same q_g we can pool the counts from the two groups by viewing them as sampled from same distribution, i.e. $X_g^{I,J}|\sigma_g, q_g \sim NB(\sigma_g, q_g)$, $q_g|\alpha, \beta^g \sim \text{Beta}(\alpha, \beta^g)$. Then we obtain the prior predictive function $f(X_g^{I,J}) = \int_0^1 P(X_g^{I,J}|r_g, q_g) * P(q_g|\alpha, \beta^g) dq_g = \left[\prod_{s=1}^{S} \binom{X_{g,s}+\sigma_g-1}{X_{g,s}}\right] \frac{Beta(\alpha+\Sigma_{s=1}^{S}\sigma_g, \beta^g+\Sigma_{s=1}^{S}X_{g,s})}{Beta(\alpha,\beta^g)}$.
Consequently, the prior predictive function for $P1, ..., P5$ is

$$
\begin{aligned}
h_1^g(X_g) &= f(X_g^{1,2,3}) \\
h_2^g(X_g) &= f(X_g^{1,2})f(X_g^3) \\
h_3^g(X_g) &= f(X_g^1)f(X_g^{2,3}) \\
h_4^g(X_g) &= f(X_g^{1,3})f(X_g^2) \\
h_5^g(X_g) &= f(X_g^1)f(X_g^2)f(X_g^3)
\end{aligned}
$$

where $h_i^g(X_g) = P(X_g|M_{g,\pi_i}, z)$ for associated π_i. Then the marginal distribution of counts X_g is $\sum_{k=1}^{5} p_k h_k^g(X_g)$, where the marginal $p_k = P(M_{g,\pi}|z)$(shared by all genome). Thus, the posterior probability of an expression pattern k is obtained by:

$$\frac{p_k h_k(X_g)}{\sum_{k=1}^{5} p_k h_k^g(X_g)}$$

In the optimization steps for determining the hyperparameters (α, β^g, p), the computation and memory increase exponentially with the number of subtypes K. We use one-step EM as an approximation for the solution, α and β^g are updated through gradient ascent while p is updated by the explicit form of the maximizer of the log likelihood.

2.2.2. *modalClust.* In this section, we review the procedure of modal-clustering and proof the compatibility of density kernel from Poisson-Gamma model.

Product Partition Model

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

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

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

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

Dahl (2009) demonstrated that by some choice of f and prior of π, we can reduce the time complexity of finding the MAP partition from factorial(n) to $O(n^2)$. The crucial condition for f is the **non-overlapping** condition: if X_{S_1} and X_{S_2} are overlapped in the sense that $\min\{X_{S_2}\} < \max\{X_{S_1}\} < \max\{X_{S_2}\}$ or $\min\{X_{S_1}\} < \max\{X_{S_2}\} < \max\{X_{S_1}\}$, let $X_{S_1^*}$ and $X_{S_2^*}$ be the sets of swapping one pair of those overlapped terms and keep the other unchanged. Then $f(X_{S_1})f(X_{S_2}) \leq f(X_{S_1^*})f(X_{S_2^*})$.

Under the **non-overlapping** condition of density kernel f, we know that in order to be MAP, a partition π must satisfy that for any two blocks $b_1, b_2 \in \pi$, either $\max\limits_{i\in b_1}(X_i) \leq \min\limits_{j\in b_2}(X_j)$ or $\min\limits_{i\in b_1}(X_i) \geq \max\limits_{j\in b_2}(X_j)$. Thus we reduce the solution space and reduce the time complexity.
In the Poisson-Gamma Model we assume:

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

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

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

To apply modal-clustering on Poisson-Gamma model, we need to show the kernel $f(X_S)$ satisfies the **non-overlapping** condition.

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

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

$$f(X_{S_1^*})f(X_{S_2^*}) \geq f(X_{S_1})f(X_{S_2}) \tag{1}$$

$$f(X_{S_1^{**}})f(X_{S_2^{**}}) \geq f(X_{S_1})f(X_{S_2}) \tag{2}$$

Let $a = \max\{X_{S_1}\}$, $b = \min\{X_{S_2}\}$ and $c = \max\{X_{S_2}\}$. $h_1 = \sum\limits_{i\in S_1} X_i - a$ and $h_2 =$

$\sum\limits_{i\in S_2} X_i - b$, n_1 and n_2 are the number of elements in S_1 and S_2. Then

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

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

Similarly,

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

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

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

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

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

The proof for case $h_1 > h_2$ follows similarly.

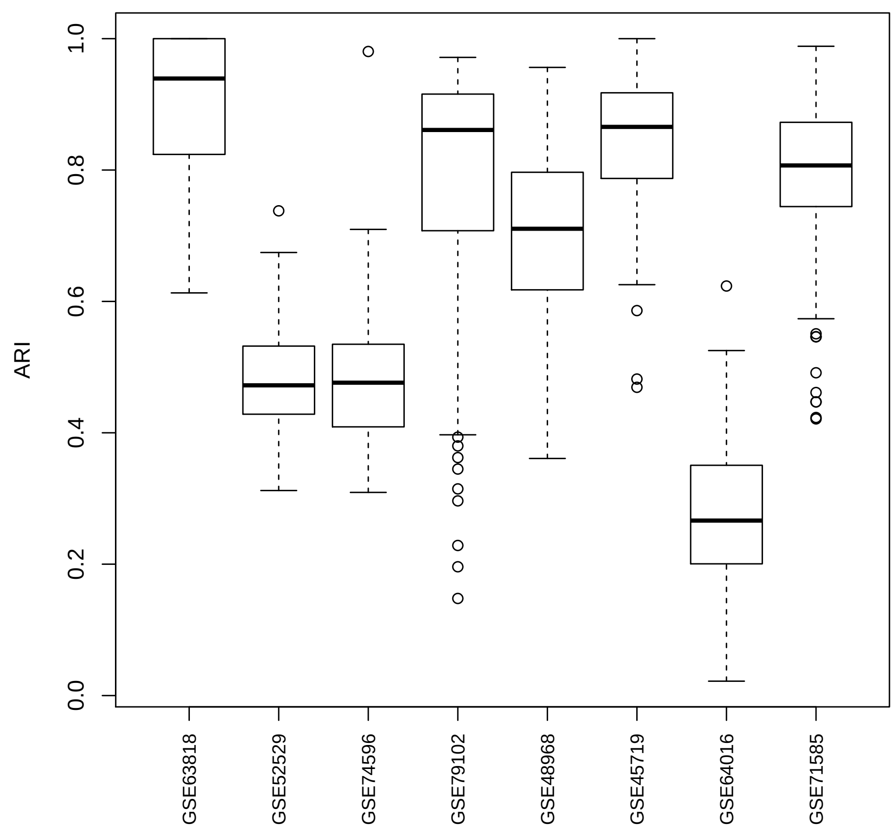
□

2.2.3. *Randomized $K-$means.* In this section, we illustrate how we estimate parameters for the distribution of weights and demonstrate variability of generated clustering on empirical data. We also investigate the accuracy of random weighting and how it approaches the fully Bayesian scheme in simulation.

To find the value of a_0, a_1 and d_0, we have the marginal likelihood of $d_{i,j}$

$$P(d_{i,j}|a_0,a_1,d_0) = \frac{\Gamma(a_0+a_1)}{\Gamma(a_0)\Gamma(a_1)}\frac{d_0^{a_0}d_{i,j}^{a_1-1}a_1^{a_1}}{(d_0+a_1*d_{i,j})^{a_0+a_1}}$$

We estimate d_0 by treating $d_{i,j} \approx \Delta_{i,j}$ and based on the mean-variance ratio ($\frac{E(1/\Delta_{i,j})}{Var(1/\Delta_{i,j})} = d_0$), d_0 can be approximately estimated by moments of $1/d_{i,j}$. Then we obtain a_0, a_1 from MLE of marginal density of $d_{i,j}$. The MLE estimators are obtained through `nlminb` function in `R`. One issue that arises is that the default value for tolerance rate of stopping is 1e-10, which yields large value of $a_1 + a_0$ and results in non-randomness of our weighting matrix. To avoid this issue, we set tolerance rate as 1e-3 to obtain moderate deviation from D (Supplementary Figure S1).

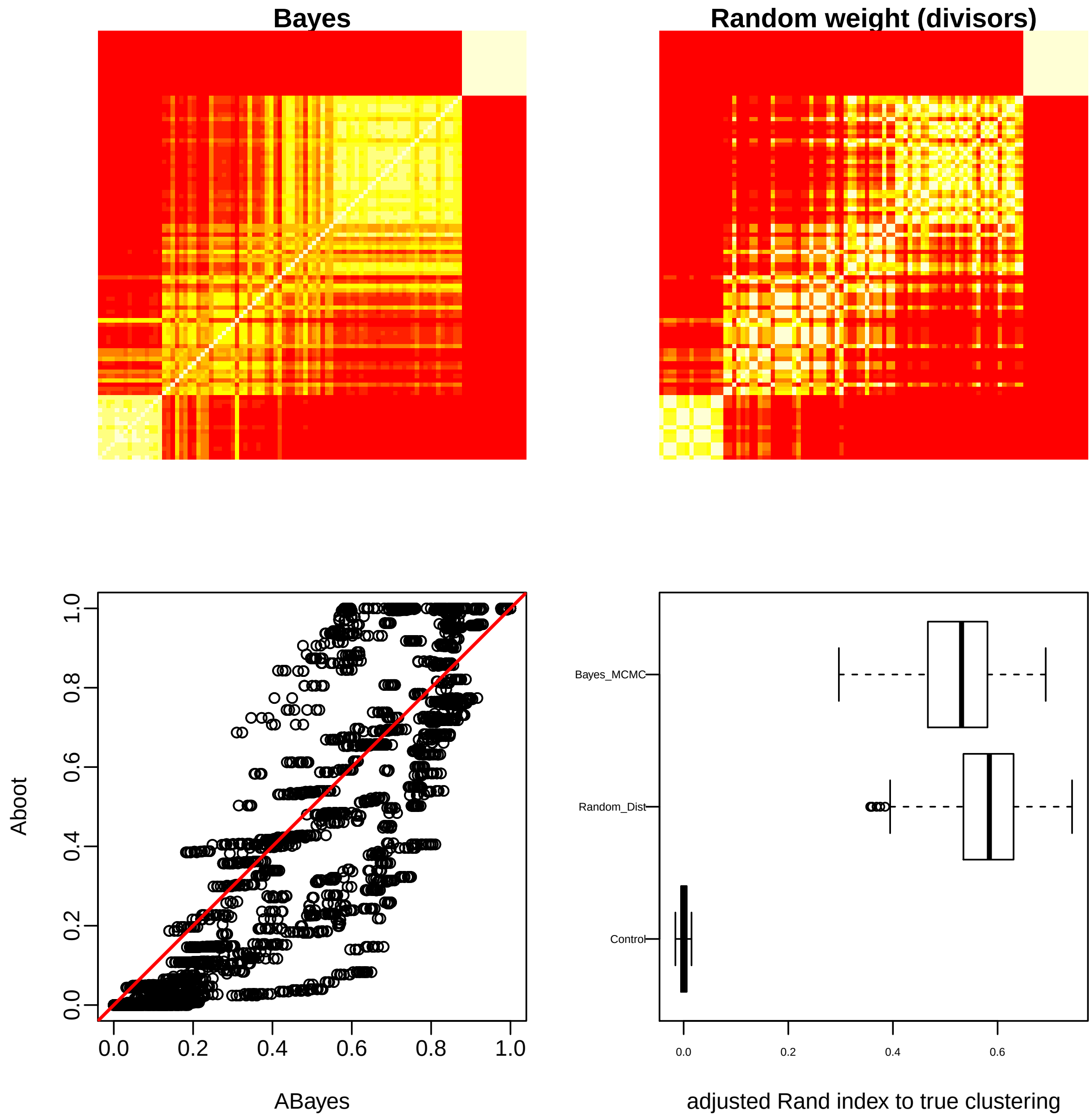


Supplementary Figure S1: Adjusted rand indexes of randomly generated clustering to the one generated by the original distance matrix. We investigate the variation of clustering given by random weighting through 8 datasets and each dataset we are using 100 random distances.

We plot the ARI (adjusted random index) between the randomly generated clustering to clustering under the original distance across eight datasets. Though the mean varies, the interquartile range is wide enough to present a reasonable variation of our random weighting scheme.

We also check validity of random weighting on simulated dataset. We simulate one-dimensional data X from a mixture of 5 normal distributions with different means and same variance ($\mu = (-3, -1, 0, 1, 5), \sigma = 1$). We compare clustering results between random weighting and Bayesian clustering using the Dirichlet process as a prior in terms of posterior probabilities that two elements belong to the same class given the whole data. We also compare accuracy of the two procedures by looking at the ARI comparing to true class label (Supplementary Figure S2). We found that random weighting scheme tends to give better results than classical Bayesian clustering.

Fig S2: missing heatmap legend; labels on scatter-plot are vague. Added more detail to legend - please double check it's accurate.



Supplementary Figure S2: Comparison between random weighting scheme and bayesian clustering procedure. Top: heatmap of posterior probabilities that two elements belong to the same class given the whole data. Bottom: scatterplot of these posterior probabilities (left), and adjusted rand index comparing to the underlying true class label (right).

2.2.4. *Selecting K.* In this section, we give the criterion to select K.

To determine the number of clusters, we implement a procedure inspired by `validity`, as defined in Ray and Turi (2000). We consider a modified `validity`$= \frac{\textbf{intra}}{\textbf{inter}}$, where **intra** $=$

$\frac{1}{N}\sum_{i=1}^{K}\sum_{x\in C_i}||x-z_i||^2$, **inter** $= mean(||z_i - z_j||^2), i, j = 1, 2, ...K$, and z_i is the center (medoid) of cluster i. **intra** is the average of distance of a point to its corresponding cluster center, which measures the compactness of clusters. **inter** is the average distance of two cluster centers, which measures the separation between clusters. In the original paper **inter** was defined as minimum distance between medoids (Ray and Turi, 2000). Here, we instead use the average for the purpose of getting a smoother quantity. We want to have a small intra-cluster distance and a big inter-cluster distance. Consequently we want to minimize the `validity`.From empirical study, we typically observe a monotonically decreasing relationship between the number of clusters and `validity`.However, this trend stabilizes when K is sufficiently large. The stopping rule for searching K is when $\texttt{validity}_K < \epsilon$ is satisfied. We set the default value of ϵ to be 1, as we found this yields DD analysis results most consistent with other scRNAseq methods.

2.3. *Double Dirichlet Mixture.* In this section, we give proofs for the properties and theorem for DDM in section 2.3 of main paper.

On the double Dirichlet masses, using notations in the main paper we have density functions:

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

with

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

and

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

Those computing units will serve as key components for proving property $1 \sim 6$ in section 2.3

Proof of property 1

Proof. When ϕ and ψ only satisfy the coarsest constrants: $\sum_{i=1}^K \phi_i = \sum_{i=1}^K \psi_i = 1$. ϕ and ψ are independently Dirichlet distributed. When ϕ and ψ satisfy finer constraints, $P(\phi|\psi) \neq P(\phi)$ as there is some subsets $b \neq \pi$ such that $\sum_{i\in b}\phi = \sum_{i\in b}\psi$. So ϕ and ψ are dependent. □

Proof of property 2

Proof. $E_\pi(\phi_k) = E_\pi(\phi_k|\Phi_b)E_\Phi(\Phi_b) = E_{\tilde{\phi}_b}(\tilde{\phi}_k)E_\Phi(\Phi_b)$ where b is the block containing subtype index k. As $\tilde{\phi}_b \sim \text{Dirichlet}_{N(b)}[\alpha_b^1]$ and $\Phi_\pi \sim \text{Dirichlet}_{N(\pi)}[\beta_\pi]$ We have

$E_{\tilde{\phi}_b}(\tilde{\phi}_k) = \frac{\alpha^1_k}{\sum_{k' \in b} \alpha^1_{k'}}$ and $E_{\Phi}(\Phi_b) = \frac{\beta_b}{\sum_{b' \in \pi} \beta_{b'}}$. Similarly we could prove the case for $E_\pi(\psi_k)$. □

Proof of property 3

Proof. t^1/t^1_π is independent of t^2/t^2_π conditioning on t^1_π and t^2_π by the Neutrality property of Dirichlet distribution □

Proof of property 4

Proof. For $j = 1, 2$, let T^j_b be the vector of t^j_k such that $k \in b$. Recall $t^j_b = \sum_{k \in b} t^j_k$. Without loss of generality, we consider the case condition $j = 1$.
At the support of p_π, for different blocks, $T^1_b|\tilde{\phi}_b$ are mutually independent. Then we have factorization:

$$p_\pi(t^1|t^1_\pi, y) = \prod_{b \in \pi}(p(T^1_b|t^1_b, y))$$

and right hand side prior predictive function can be obtained by integrating out $\tilde{\phi}_b$ given the prior Dirichlet$[\alpha^1_b]$, and $p(T^1_b|\tilde{\phi}_b)$ is multinomial$(\tilde{\phi}_b)$ distributed.

$$\begin{aligned} p(T^1_b|t^1_b, y) &= \int_{\tilde{\phi}_b} p(T^1_b|\tilde{\phi}_b)p(\tilde{\phi}_b)d\tilde{\phi}_b \\ &= \left\{ \left[\frac{\Gamma(t^j_b + 1)}{\prod_{k \in b} \Gamma(t^j_k + 1)} \right] \left[\frac{\Gamma(\sum_{k \in b} \alpha^j_k)}{\prod_{k \in b} \Gamma(\alpha^j_k)} \right] \left[\frac{\prod_{k \in b} \Gamma(\alpha^j_k + t^j_k)}{\Gamma(t^j_b + \sum_{k \in b} \alpha^j_k))} \right] \right\} \end{aligned}$$

□

Proof of property 5

Proof. t^1_π and t^2_π, given the condition label y, are independent and identically distributed. $t^1_\pi|\Phi \sim$ multinomial(Φ)

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

As prior of Φ is Dirchlet$[\beta]$ and $n_j = \sum_{b \in \pi} t^j_b$ for $j = 1, 2$. □

To prove property 6, we need a lemma of dimensionality of the intersection of two A_πs.

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 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 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. This implies π_2 is refinement of π_1: a 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 property 6

Proof. For a π, $P(A_\pi, |y, z) = \sum_{\tilde{\pi} \in \Pi} \int_{A_\pi} \omega_{\tilde{\pi}}^{\text{post}} d\phi d\psi$, notice the support of $\omega_{\tilde{\pi}}^{\text{post}}$ is $A_{\tilde{\pi}}$. By lemma 1, we know if $\tilde{\pi}$ does not refine π, then $\int_{A_\pi} \omega_{\tilde{\pi}}^{\text{post}} d\phi d\psi$ is an integral on lower dimension set and vanishes. if $\tilde{\pi}$ refines π, then $\int_{A_\pi} \omega_{\tilde{\pi}}^{\text{post}} d\phi d\psi = \int_{A_{\tilde{\pi}}} \omega_{\tilde{\pi}}^{\text{post}} d\phi d\psi = \omega_{\tilde{\pi}}^{\text{post}}$. We have $P(A_\pi, |y, z) = \sum_{\tilde{\pi} \in \Pi} \omega_{\tilde{\pi}}^{\text{post}} 1[\tilde{\pi} \text{ refines } \pi]$. □

Proof of theorem 3

Proof. Recall the DDM prior: $p(\phi, \psi) = \sum_{\pi \in \Pi} p_\pi(\phi, \psi)$. By bayes' rule we know $p(\phi, \psi | y, z) \propto p(\phi, \psi, y, z) = \sum_{\pi \in \Pi} p(y, z | \phi, \psi) p_\pi(\phi, \psi) \omega_\pi$ and we use the 1-1 map from (ϕ, ψ) to $(\tilde{\phi}, \tilde{\psi}, \Phi)$ to get

$$p(y, z | \phi, \psi) p_\pi(\phi, \psi) = p(y, z | \tilde{\phi}, \tilde{\psi}, \Phi_\pi) p(\tilde{\phi}) p(\tilde{\psi}) p(\Phi_\pi)$$

when $(\phi, \psi) \in A_\pi$. Let us denote right hand side of the above equation as U_π, then

$$U_\pi = \omega_\pi A_1 A_2 A_3 \prod_{k=1}^{K} (\tilde{\phi}_k)^{t_k^1 + \alpha_k^1} (\tilde{\psi}_k)^{t_k^2 + \alpha_k^2} \prod_{b \in \pi} (\Phi_b)^{t_b^1 + t_b^2 + \beta_b}$$

Where A_1 is the product of normalizing terms from multinomial distribution of z^1 and z^2, $A_1 = \frac{\Gamma(n_1+1)\Gamma(n_2+1)}{\prod_{j=1}^2 \prod_{k=1}^K \Gamma(t_k^j + 1)}$.

A_2 is the product of normalizing terms from Dirichlet distribution of $\tilde{\phi}$ and $\tilde{\psi}$, $A_2 = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k^1+1)\Gamma(\sum_{k=1}^{K}\alpha_k^2+1)}{\prod_{j=1}^{2}\prod_{k=1}^{2}\Gamma(\alpha_k^j+1)}$.
A_3 is the normalizing term from Dirichle distribution of Φ_π, $A_3 = \frac{\Gamma(\sum_{b\in\pi}\beta_b+1)}{\prod_{b\in\pi}\Gamma(\beta_b+1)}$.
Looking at the indexes of $\tilde{\phi}, \tilde{\psi}$ and Φ, we can decompose U_π as a product of three Dirichlet densities with a normalizing term. Namely $U_\pi = C_\pi * f_1 f_2 f_3$, where $f_1 \sim \text{Dirichlet}[\alpha^1 + t^1]$, $f_2 \sim \text{Dirichlet}[\alpha^2 + t^2]$ and $f_3 \sim \text{Dirichlet}[\beta + t^1 + t^2]$. Considering the normalizing factors for densities f_1, f_2 and f_3, and multiplying them with A_1, A_2 and A_3, we have $C_\pi = p_\pi(t^1|t^1_\pi, y)\, p_\pi(t^2|t^2_\pi, y)\, p_\pi(t^1_\pi, t^2_\pi|y)\omega_\pi$.
Consequently, we have

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

Notice in DDM, we restricted $\beta = \alpha^1 + \alpha^2$.

□

3. Numerical Experiments.

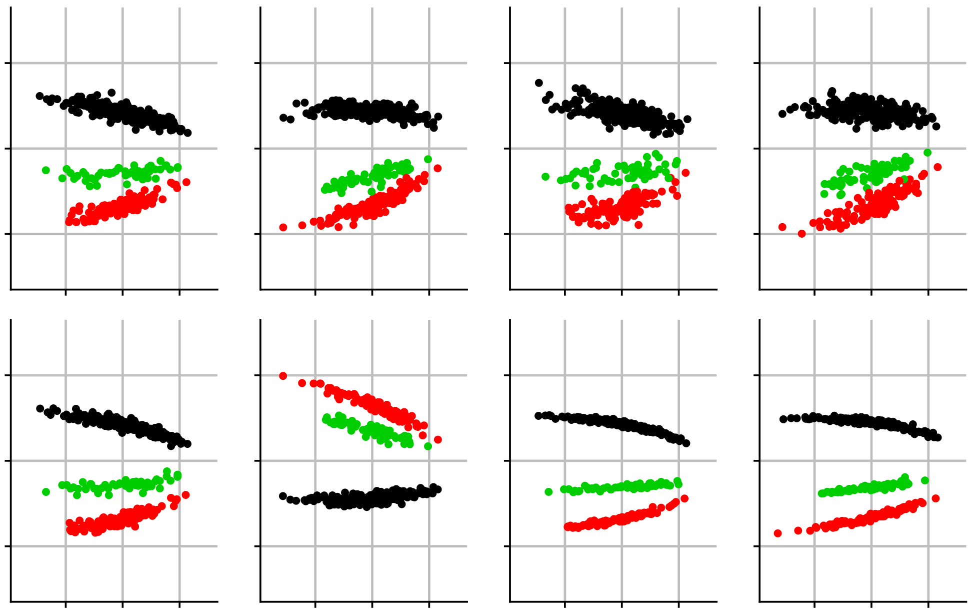
3.1. *Synthetic Data.* In this section, we use PCA plots to show the subtle changes underlying each subtype of simulated data and we demonstrate consistency of estimated distributional changes based on `scDDboost` and Wasserstein distance. Finally, ROC curves illustrate that scDDboost has favorable operating characteristics.

We first look at the PCA plots of the simulated data (Supplementary Figure S3, S4, S5). For $K = 7$ and 12, in each scenario there were some subtypes nested in the 2d PCA projection. Here, the distributional change of transcripts becomes difficult to detect. `scDDboost` benefits from the compositional structure and is more sensitive to those subtle changes.

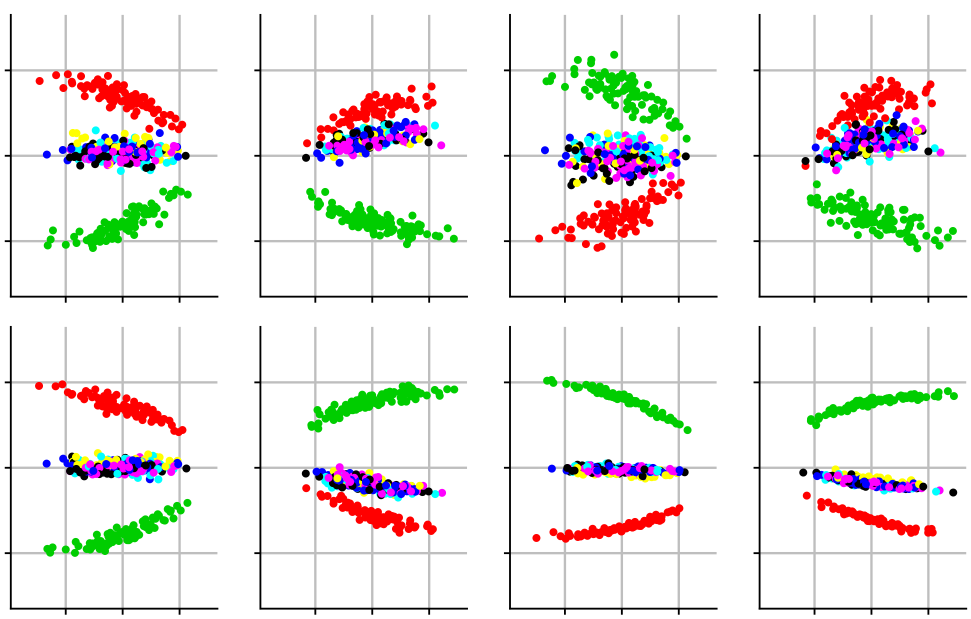
Figs S3-S5 missing axis labels

Do they separate in higher order PCs?

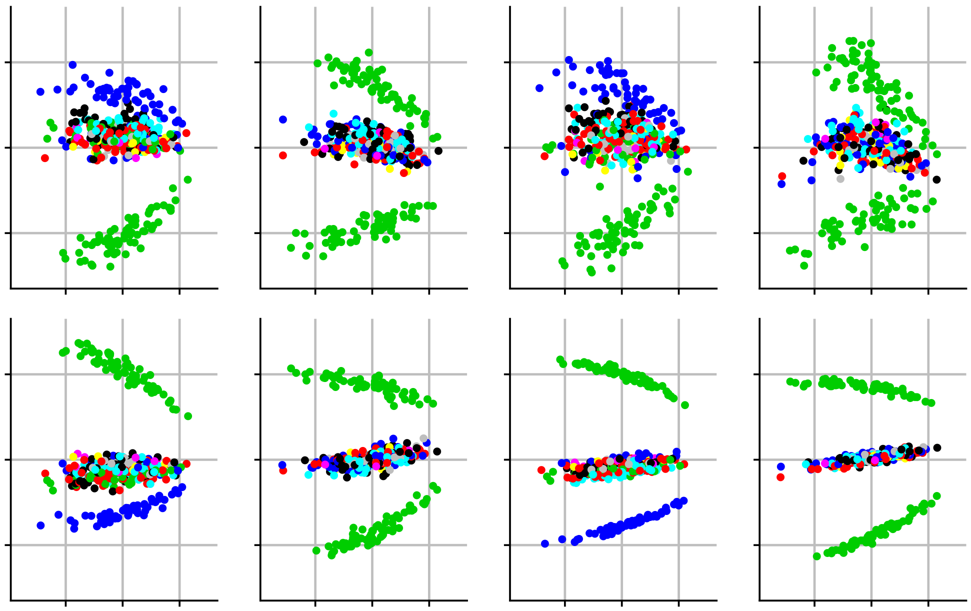
We observed consistent measurements of distributional change based on `scDDboost` and Wasserstein distance between the empirical distribution of transcripts (Supplementary Figure S6). Lower probabilities of equivalent distributed are associated with bigger distances.



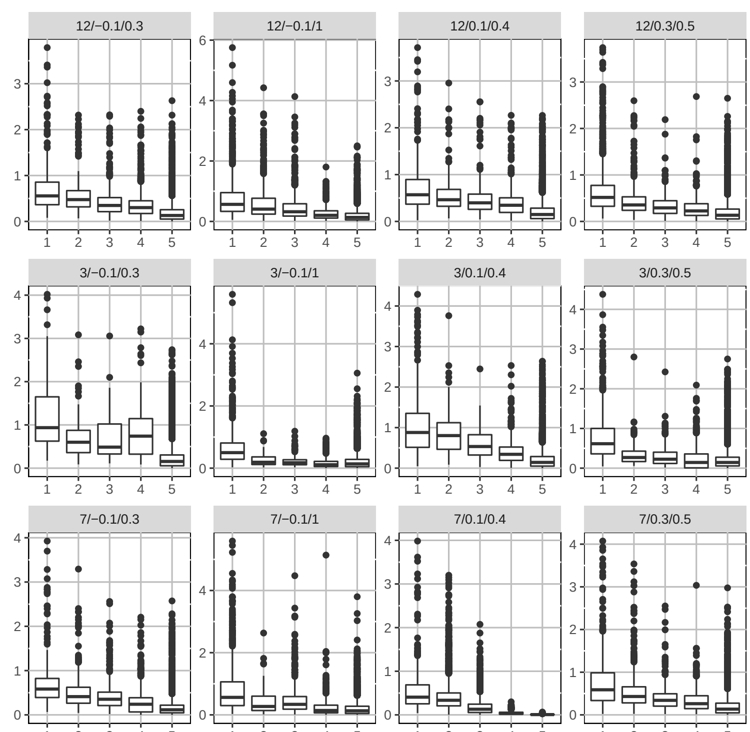
Supplementary Figure S3: first two principal components of transcripts under different parameters for simulated data. Different parameters resulted in different degree of separation of subtypes. We have 4 different settings for hyper-parameters of simulation, each setting has 2 replicates and K = 3



Supplementary Figure S4: Same as Supplementary Figure S3, but for K = 7

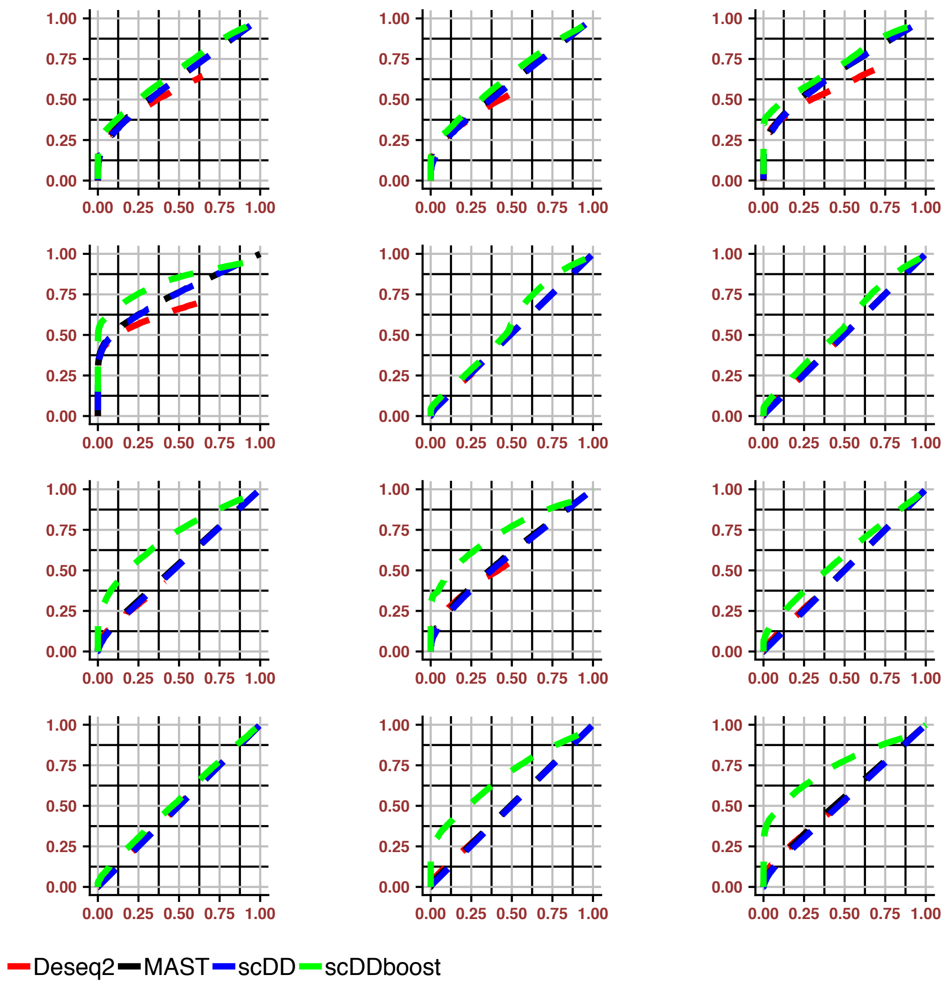


Supplementary Figure S5: Same as Sumpplementary Figure S3, but for K = 12



Supplementary Figure S6: $P(ED_g|X, y)$ given by scDDboost versus empirical Wasserstein distance. Genes associated with boxes from left to right having $P(ED_g|X, y)$ range from 0 - 0.2, 0.2 - 0.4, 0.4 - 0.6, 0.6 - 0.8, 0.8 - 1. For the simulation cases

We also show ROC curves for the simulated data in Supplementary Figure S7. here each sub-figure is averaged over two replicates under the same parameters setting. scDDboost tends to outperform other methods .



Supplementary Figure S7: Roc curve of the 12 simulation settings, under each setting, TPR and FPR are averaged over two replicates, generally we found scDDboost perform better than other methods

3.2. *Empirical Study.* In this section, we provide details of the empirical datasets and also demonstrate consistency to Wasserstein distance on one dataset (FUCCI). citation?

Data sets Details for the datasets used in the empirical studies of the main paper and the estimated number of subtypes K are shown in Supplementary Table S1.

Data set	Conditions	Number of cells/condition	Organism	Ref	K
GSE94383	0 min unstim vs 75min stim	186,145	human	(Lane et al., 2017)	9
GSE48968-GPL13112	BMDC (2h LPS stimulation) vs 6h LPS	96,96	mouse	(Shalek et al., 2014)	4
GSE52529	T0 vs T72	69,74	human	(Trapnell et al., 2014)	7
GSE74596	NKT1 vs NTK2	46,68	mouse	(Engel et al., 2016)	7
EMTAB2805	G1 vs G2M	95,96	mouse	(Buettner et al., 2015)	6
GSE71585-GPL13112	Gad2tdTpositive vs Cux2tdTnegative	80,140	mouse	(Tasic et al., 2016)	4
GSE64016	G1 vs G2	91,76	human	(Leng et al., 2015)	6
GSE79102	patient1 vs patient2	51, 89	human	Kiselev et al. (2017)	4
GSE45719	16-cell stage blastomere vs mid blastocyst cell	50, 60	mouse	(Deng et al., 2014)	4
GSE63818	Primordial Germ Cells, develop- mental stage: 7 week gestation vs Somatic Cells, developmental stage: 7 week gestation	40,26	mouse	(Guo et al., 2015)	6
GSE75748	DEC vs EC	64, 64	human	(Chu et al., 2016)	5
GSE84465	neoplastic cells vs non-neoplastic cells	1000, 1000	human	(Darmanis et al., 2017)	9

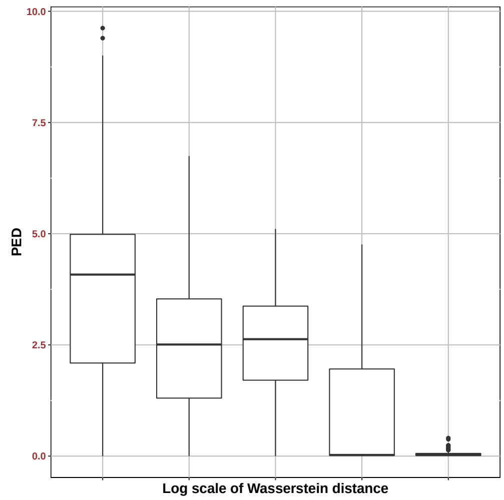
Supplementary Table S1
datasets used for comparisons of DD analysis under different methods

The largest dataset we explored is GSE84465, which contains 3500 cells in total. For this dataset, we randomly sampled 1000 cells from each condition, because it takes more time and memory to compute using all the samples. For other datasets we use all the cells within that condition under same batch. 1000 cells each condition seems to be enough to represent the heterogeneity. We found `DESeq` found significantly smaller numbers of discoveries than others. It is intuitive that we are more likely to encounter subtle changes when we have large samples, and only considering mean shifts would have limited power.

How different are results on a different random subset of 1000?

We also observed consistent distributional change measurements by `scDDboost` and Wasserstein distance between the empirical distribution of transcripts (Supplementary Figure S8).

Datasets used for generating the Null cases are shown in Supplementary Table S2.



Supplementary Figure S8: $P(ED_g|X,y)$ given by scDDboost versus empirical Wasserstein distance. Genes associated with boxes from left to right having $P(ED_g|X,y)$ range from 0 - 0.2, 0.2 - 0.4, 0.4 - 0.6, 0.6 - 0.8, 0.8 - 1, data used: FUCCI

.

Data set	Conditions	Number of cells/condition	Organism
GSE63818null	7 week gestation	20,20	mouse
GSE75748null	DEC	32, 32	human
GSE94383null	T0	93, 93	human
GSE48968-GPL13112null	BMDC (2h LPS stimulation)	48,48	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

Supplementary Table S2

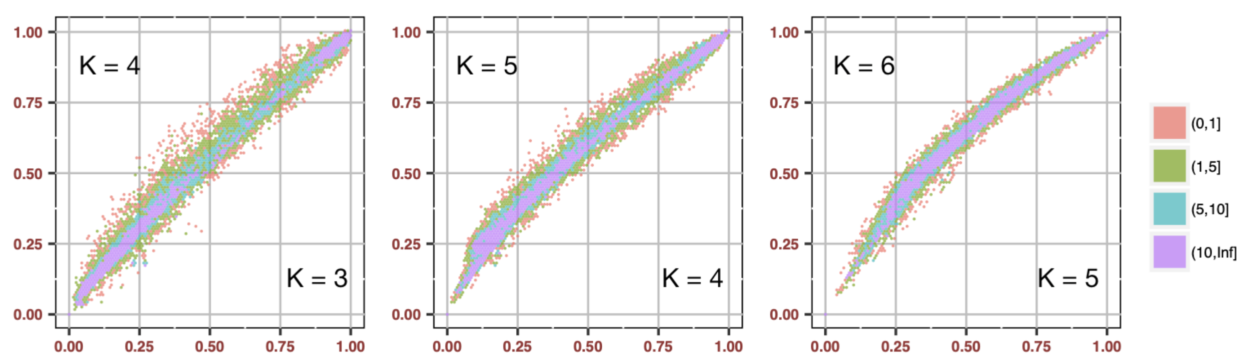
datasets used for null cases, as cells are coming from same biological condition, there should not be any differential distributed genes, any positive call is false positive

3.3. *Robustness.* In this section, we demonstrate change of PDD under different K and the robustness we gain through random weighting. We also give a warning that using arbitrarily large K will inflate FDR.

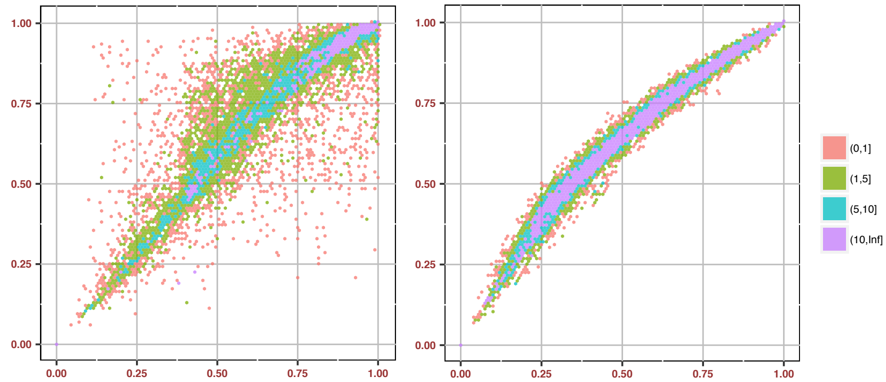
the number of subtypes K is a crucial parameter controlling the accuracy of our modeling. Too small K may end up underfitting such that cells within same subtype can still be very different. If the mean expression change among subtypes is incapable to capture the distribution change for some genes, the power of scDDboost will be reduced. Too big K may end up overfitting such that two subtypes can be very similar. Given that we have a fixed number of samples (cells), allowing more clusters will introduce many patterns (both for mean expression change and proportion change) to infer. Also note the limitation of DDM model (see Section 4). Overestimating K in scDDboost may losing FDR control (Supplementary Figure S11).

what limitation is this referring to?

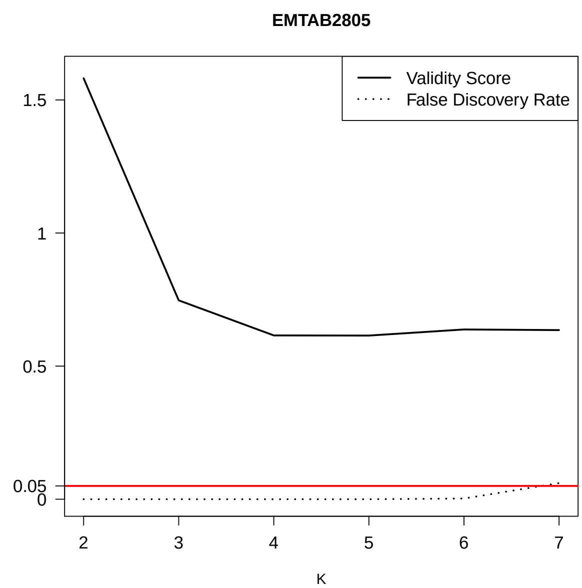
From our empirical experience, it is sufficient to capture the heterogeneity underlying cells with number of clusters less than 10 (Supplementary TableS1). To demonstrate the change of PDD given different K, we present an example using dataset GSE75748. When we increase K, the variance of the differential term $\text{PDD}_{K+1} - \text{PDD}_K$ keeps decreasing and PDD keeps increasing. Our selection criterion (K = 5) happens to choose K such thatchange between PDD_{K+1} and PDD_K is small while not inflating PDD. We generally obtain stable validity score and PDD simultaneously (Supplementary Figure S9). In addition, the random weighting scheme help us obtain robust PDD (Supplementary Figure S10). Finally, we note that scDDboost may lose FDR control if we keep increasing K (Supplementary Figure).



Supplementary Figure S9: PDD change under different number of subtypes K for dataset DEC-EC (GSE75748). Our rule for selecting K tends also to make PDD stabilize.4



Supplementary Figure S10: PDD under K = 5 vs. K = 6 for dataset DEC-EC (GSE75748). Left panel is without the randomized distance and right panel is with randomized distance. We increase robustness of our methods through random weighting.



Supplementary Figure S11: under NULL case, using dataset EMTAB2805, when using too big K we may lose FDR control (black dashed line shows proportion of false positive identified by scDDboost under 0.05 threshold, while validity score did not vary too much after K is greater than 2.

4. Posterior consistency. In this section, we prove theorem 4.

The density of DDM is computed by product or ratio over several gamma functions and the gamma function is not easy to directly derive limiting theorem. To prove theorem 4, we need a crucial lemma which gives us an approximation to the gamma function, namely

LEMMA 2. *For $x \geq 1$, $\frac{x^{x-c}}{e^{x-1}} \leq \Gamma(x) \leq \frac{x^{x-1/2}}{e^{x-1}}$, where $c = 0.577215...$ is the Euler-Mascheroni constant.*

PROOF. By (Li and ping Chen, 2007), we have $\frac{x^{x-c}}{e^{x-1}} \leq \Gamma(x) \leq \frac{x^{x-1/2}}{e^{x-1}}$ for $x > 1$ and now we added the case when $x = 1, \Gamma(x) = 1$ so that both sides will include the equality case. □

LEMMA 3. *For positive integer n, $\sqrt{2\pi} n^{n+1/2} e^{-n} \leq \Gamma(n+1) \leq e n^{n+1/2} e^{-n}$*

We have another two lemmas and theorem 1 and 2 are just proposition of the lemma

LEMMA 4. *If $(\phi, \psi) \in A_{\pi_1} \cap A_{\pi_2}$, follow the conditions in theorem 1 then*

$$\frac{\omega_{\pi_1}^{post}}{\omega_{\pi_2}^{post}} \xrightarrow[n\to\infty]{a.s.} 0 \quad \textit{if } N(\pi_1) < N(\pi_2)$$

PROOF. Recall $\omega_\pi^{\text{post}} \propto p_\pi(t^1|t_\pi^1, y)\, p_\pi(t^2|t_\pi^2, y)\, p_\pi(t_\pi^1, t_\pi^2|y)\, \omega_\pi$ and RHS $= g(\pi, \alpha, \beta, n_1, n_2) f(\pi, t^1, t^2, \alpha, \beta)$ and $\frac{\omega_{\pi_1}^{\text{post}}}{\omega_{\pi_2}^{\text{post}}} = \frac{g(\pi_1,\alpha,\beta,n_1,n_2)}{g(\pi_2,\alpha,\beta,n_1,n_2)} \frac{f(\pi_1,t^1,t^2,\alpha,\beta)}{f(\pi_2,t^1,t^2,\alpha,\beta)}$ where

$$g(\pi, t^1, t^2, \alpha, \beta) = [\prod_{j=1}^{2} \prod_{b\in\pi} \frac{\Gamma(\Sigma_{k\in b}\alpha_k^j)}{\prod_{k\in b}\Gamma(\alpha_k^j)}] \frac{\Gamma(n_1+1)\Gamma(n_2+1)}{\prod_{b\in\pi}\Gamma(\beta_b)} \frac{\Gamma(\Sigma_{b\in\pi}\beta_b)}{\Gamma(n_1+n_2+\Sigma_{b\in\pi}\beta_b)}$$

$$f(\pi, t^1, t^2, \alpha, \beta) = [\prod_{j=1}^{2} \prod_{b\in\pi} \frac{1}{\prod_{k\in b}\Gamma(t_k^j+1)} \frac{\prod_{k\in b}\Gamma(\alpha_k^j + t_k^j)}{\Gamma(t_b^j + \Sigma_{k\in b}\alpha_k^j)}] \prod_{b\in\pi}\Gamma(\beta_b + t_b^1 + t_b^2)$$

For notation simplicity, we use the abbreviation $g(\pi), f(\pi)$ to substitute $g(\pi, \alpha, \beta, n_1, n_2), f(\pi, t^1, t^2, \alpha, \beta)$. We take log on $\frac{\omega_{\pi_1}^{\text{post}}}{\omega_{\pi_2}^{\text{post}}}$, denote it as LR. $LR = \ln g(\pi_1) - \ln g(\pi_2) + \ln f(\pi_1) - \ln f(\pi_2)$. Denote $C(\pi_1, \pi_2, \alpha, \beta) = \ln g(\pi_1) - \ln g(\pi_2)$, $C(\pi_1, \pi_2, \alpha, \beta)$ does not change with sample size n_1, n_2 and is a constant determined by partition π_1, π_2 and hyper parameters α, β. For further convenience of notation let $h(x) = \ln\Gamma(x)$ and $\gamma_b^j = \Sigma_{k\in b}\alpha_k^j$. Denote $R(\pi_1, \pi_2, t^1, t^2, \alpha, \beta) = \ln f(\pi_1) - \ln f(\pi_2)$. And removing the common part of $f(\pi_1)$ and $f(\pi_2)$, we have

$$R(\pi_1, \pi_2, t^1, t^2, \alpha, \beta) = d(\pi_1, t^1, t^2, \alpha, \beta) - d(\pi_2, t^1, t^2, \alpha, \beta)$$

where

$$d(\pi, t^1, t^2, \alpha, \beta) = \sum_{b\in\pi} h(\beta_b + t_b^1 + t_b^2) - \sum_{j=1}^{2}\sum_{b\in\pi} h(t_b + \gamma_b^j)$$

Recall $\beta_b = \gamma_b^1 + \gamma_b^2$ and from lemma 2, $(x-c)\ln(x) - x + 1 \le h(x) \le (x-1/2)\ln(x) - x + 1$ we have

(3)

$$d(\pi, t^1, t^2, \alpha, \beta) \ge \sum_{b\in\pi} (\beta_b + t_b^1 + t_b^2 - c)\ln(\beta_b + t_b^1 + t_b^2) - \sum_{j=1}^{2}\sum_{b\in\pi} (t_b^j + \gamma_b^j - 1/2)\ln(t_b^j + \gamma_b^j) + N(\pi)$$

(4)

$$d(\pi, t^1, t^2, \alpha, \beta) \le \sum_{b\in\pi} (\beta_b + t_b^1 + t_b^2 - 1/2)\ln(\beta_b + t_b^1 + t_b^2) - \sum_{j=1}^{2}\sum_{b\in\pi} (t_b^j + \gamma_b^j - c)\ln(t_b^j + \gamma_b^j) + N(\pi)$$

$$\begin{aligned}\text{RHS of (4)} = \Sigma_b\big[&(t_b^1 + \gamma_b^1)\ln(1 + \frac{t_b^2 + \gamma_b^2}{t_b^1 + \gamma_b^1}) + (t_b^2 + \gamma_b^2)\ln(1 + \frac{t_b^1 + \gamma_b^1}{t_b^2 + \gamma_b^2}) \\ &+(1-c)\ln(\beta_b + t_b^1 + t_b^2) - 1/2(\ln(1 + \frac{t_b^2 + \gamma_b^2}{t_b^1 + \gamma_b^1}) + \ln(1 + \frac{t_b^1 + \gamma_b^1}{t_b^2 + \gamma_b^2}))\big] + N(\pi)\end{aligned}$$

By Taylor expansion at $x = 1$, $\ln(x+1) = \ln 2 + 1/2(x-1) - 1/8(x-1)^2 + g(\xi)(x-1)^3$, where $g(\xi)$ is the reminder term of form $\frac{1}{3(1+\xi)^3}$ for $0 < \xi < x$ For a fixed n_1, n_2, we have

$$\begin{aligned}\text{RHS of (4)} = (n_1 + n_2)\ln 2 - \Sigma_{b\in\pi}(1/8(X_b^1 + X_b^2) \\ +g(\xi_b)(Y_b^1 + Y_b^2)) + T(\pi) + N(\pi)\end{aligned}$$

where $X_b^1 = \frac{(t_b^1 - t_b^2 + \gamma_b^1 - \gamma_b^2)^2}{t_b^1 + \gamma_b^1}$, $X_b^2 = \frac{(t_b^1 - t_b^2 + \gamma_b^1 - \gamma_b^2)^2}{t_b^2 + \gamma_b^2}$, $Y_b^1 = \frac{(t_b^1 - t_b^2 + \gamma_b^1 - \gamma_b^2)^3}{(t_b^1 + \gamma_b^1)^2}$, $Y_b^2 = \frac{(t_b^1 - t_b^2 + \gamma_b^1 - \gamma_b^2)^3}{(t_b^2 + \gamma_b^2)^2}$ and $T(\pi) = \Sigma_{b\in\pi}\big[(1-c)\ln(\beta_b + t_b^1 + t_b^2) - 1/2(\ln(1 + \frac{t_b^2 + \gamma_b^2}{t_b^1 + \gamma_b^1}) + \ln(1 + \frac{t_b^1 + \gamma_b^1}{t_b^2 + \gamma_b^2}))\big]$

Similarly

$$\begin{aligned}\text{RHS of (5)} = (n_1 + n_2)\ln 2 - \Sigma_{b\in\pi}(1/8(X_b^1 + X_b^2) \\ +g(\xi_b)(Y_b^1 + Y_b^2)) + U(\pi) + N(\pi)\end{aligned}$$

$$U(\pi) = \Sigma_{b\in\pi}\big[(2c - 1/2)\ln(\beta_b + t_b^1 + t_b^2) - c(\ln(1 + \frac{t_b^2 + \gamma_b^2}{t_b^1 + \gamma_b^1}) + \ln(1 + \frac{t_b^1 + \gamma_b^1}{t_b^2 + \gamma_b^2}))\big]$$

Using above inequalities, we have

$$\begin{aligned}R(\pi_1, \pi_2, t^1, t^2, \alpha, \beta) \le U(\pi_1) - T(\pi_2) - 1/8(\Sigma_{b\in\pi_1}(X_b^1 + X_b^2) - \Sigma_{b\in\pi_2}(X_b^1 + X_b^2)) \\ +\Sigma_{b\in\pi_1} g(\xi_b)(Y_b^1 + Y_b^2) - \Sigma_{b\in\pi_2} g(\xi_b)(Y_b^1 + Y_b^2)\end{aligned}$$

$Y_b^j = \frac{((t_b^1-t_b^2+\gamma_b^1-\gamma_b^2)/\sqrt{n}))^3/\sqrt{n}}{((t_b^j+\gamma_b^j)/n)^2}$, by LLN the denominator goes to a constant and by CLT in the numerator $(t_b^1 - t_b^2 + \gamma_b^1 - \gamma_b^2)/\sqrt{n} \to (t_b^1 - t_b^2)/\sqrt{n} \to \sqrt{n}[(t_b^1/n - \Phi_b) - (t_b^2/n - \Psi_b)]$, which goes to a normal distributed random variable when $\Phi_b = \Psi_b$. So Y_b^j is $o_p(1)$. Similarly, $X_b^j = \frac{((t_b^1-t_b^2+\gamma_b^1-\gamma_b^2)/\sqrt{n})^2}{t_b^j+\gamma_b^j/n}$ is asymptotic gamma(χ-square) distributed. $g(\xi_b)$ has bounded variance, $U(\pi_1) - T(\pi_2) = -\ln(n)$ if $N(\pi_2) < N(\pi_1)$ as $\ln(\beta_b + t_b^1 + t_b^2) - \ln(\beta_{b'} + t_{b'}^1 + t_{b'}^2) = \ln(\frac{\beta_b+t_b^1+t_b^2}{n}) - \ln(\frac{\beta_{b'}+t_{b'}^1+t_{b'}^2}{n}) \to O(1) \quad a.s.$ so we complete the proof.

□

Lemma 5. *If $(\phi, \psi) \in A_{\pi_1} \cap A_{\pi_2}$, follow the conditions in theorem 1 and further we have $\alpha^j, j = 1, 2$ be vectors of integers then*

$$\frac{\omega_{\pi_1}^{post}}{\omega_{\pi_2}^{post}} \xrightarrow[n\to\infty]{d} v \quad \text{if } N(\pi_1) = N(\pi_2)$$

v is a random variable

Proof. follow almost same procedure in lemma 4, but instead of using inequalities in lemma 2, we use lemma 3. And we still have

$$d(\pi, t^1, t^2, \alpha, \beta) = \sum_{b\in\pi} h(\beta_b + t_b^1 + t_b^2) - \sum_{j=1}^{2}\sum_{b\in\pi} h(t_b + \gamma_b^j)$$

and by lemma 3

$$\begin{aligned} d(\pi, t^1, t^2, \alpha, \beta) \quad &\geq \sum_{b\in\pi}(\beta_b + t_b^1 + t_b^2 - 1/2)\ln(\beta_b + t_b^1 + t_b^2) - \\ &\sum_{j=1}^{2}\sum_{b\in\pi}(t_b^j + \gamma_b^j - 1/2)\ln(t_b^j + \gamma_b^j) + \ln(\sqrt{2\pi}) - 1 \\ d(\pi, t^1, t^2, \alpha, \beta) \quad &\leq \sum_{b\in\pi}(\beta_b + t_b^1 + t_b^2 - 1/2)\ln(\beta_b + t_b^1 + t_b^2) - \\ &\sum_{j=1}^{2}\sum_{b\in\pi}(t_b^j + \gamma_b^j - 1/2)\ln(t_b^j + \gamma_b^j) + 1 - \ln(\sqrt{2\pi}) \end{aligned}$$

$$\begin{aligned} R(\pi_1, \pi_2, t^1, t^2, \alpha, \beta) \approx D(\pi_1) - D(\pi_2) - 1/8(\Sigma_{b\in\pi_1}(X_b^1 + X_b^2) - \Sigma_{b\in\pi_2}(X_b^1 + X_b^2)) \\ -\Sigma_{b\in\pi_1} g(\xi_b)(Y_b^1 + Y_b^2) - \Sigma_{b\in\pi_2} g(\xi_b)(Y_b^1 + Y_b^2) \end{aligned}$$

where $D(\pi) = \Sigma_{b\in\pi}\big[1/2\ln(\beta_b + t_b^1 + t_b^2) - c(\ln(1 + \frac{t_b^2+\gamma_b^2}{t_b^1+\gamma_b^1}) + \ln(1 + \frac{t_b^1+\gamma_b^1}{t_b^2+\gamma_b^2}))\big]$ And $D(\pi_1) - D(\pi_2)$ is O(1) if $N(\pi_1) = N(\pi_2)$ as $\ln(\beta_b + t_b^1 + t_b^2) - \ln(\beta_{b'} + t_{b'}^1 + t_{b'}^2) = \ln(\frac{\beta_b+t_b^1+t_b^2}{n_1}) - \ln(\frac{\beta_{b'}+t_{b'}^1+t_{b'}^2}{n_1}) \to 0 \quad a.s.$

□

Proof of theorem 4

Proof. Recall $\sum_{\pi\in\Pi}\omega_{\pi}^{\text{post}} = 1$ and $P(A_{\pi}|y,z) = \sum_{\tilde{\pi}\in\Pi}\omega_{\tilde{\pi}}^{\text{post}} 1[\tilde{\pi} \text{ refines } \pi]$. If $(\phi,\psi) \notin$ Q, for all the A_{π} covers (ϕ,ψ) there is one finest π^* with the largest $N(\pi^*)$ and every other π that $(\phi,\psi) \in A_{\pi}$ is coarser than π^*. We get the results of theorem 4 by lemma 4. $\square$

References.

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

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

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

Darmanis, S., Sloan, S. A., Croote, D., Mignardi, M., Chernikova, S., Samghababi, P., Zhang, Y., Neff, N., Kowarsky, M., Caneda, C., Li, G., Chang, S. D., Connolly, I. D., Li, Y., Barres, B. A., Gephart, M. H. and Quake, S. R. (2017). Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. *Cell reports* **21** 1399–1410.

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

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

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

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

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

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

Li, X. and ping Chen, C. (2007). Inequalities for the gamma function. In *2007), Art. 28. [ONLINE: http://jipam.vu.edu.au/article.php?sid=842*.

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

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

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

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