Supplementary Material

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

Authors: Ma, Korthauer, Kendziorski, and Newton

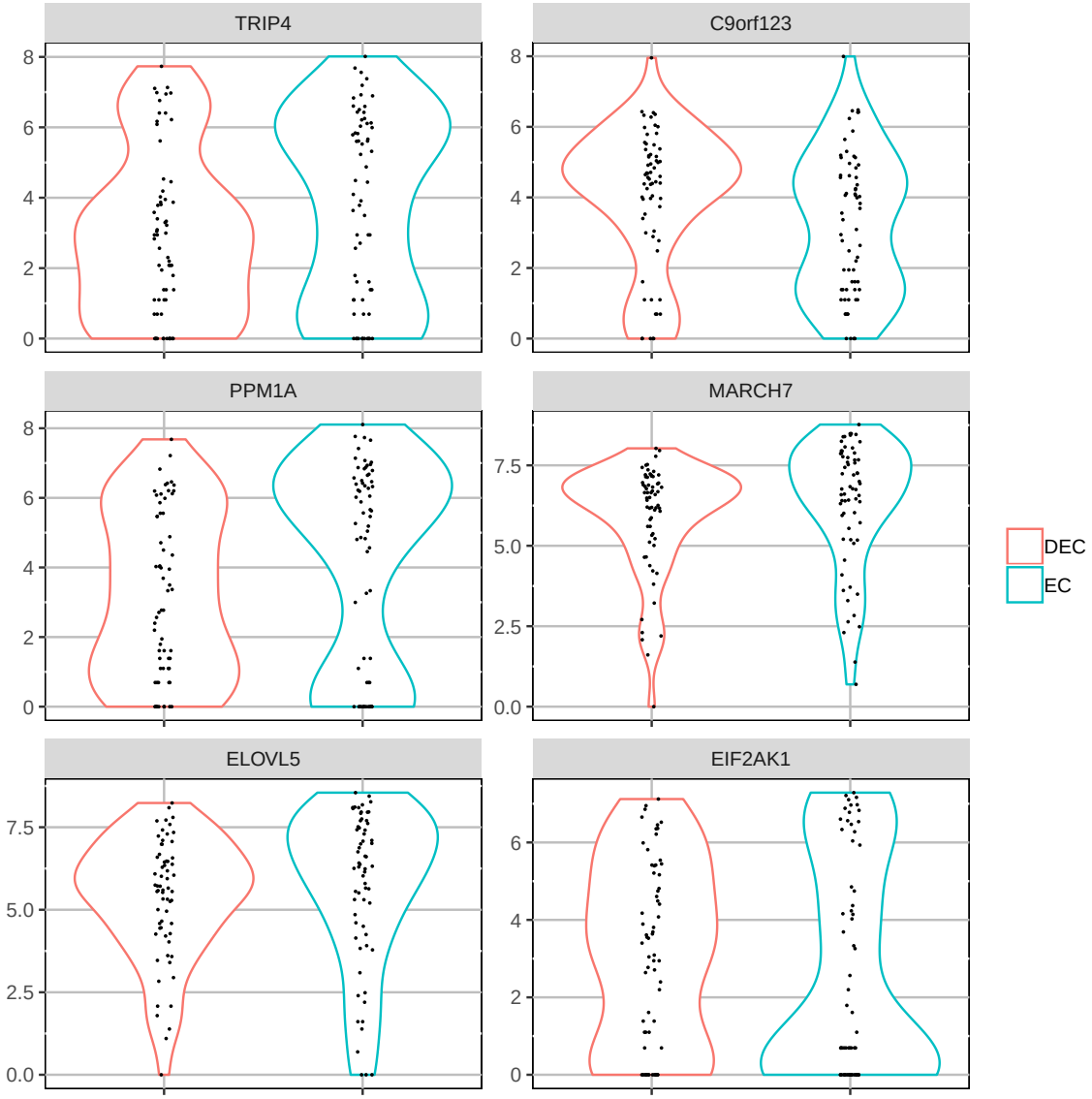
Version: February 20, 2019

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

1. Introduction
   - DEC vs EC data analysis
   - 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 1-8 and Theorem 3
3. Numerical Experiments
   - Synthetic data, `splatter`
   - Empirical study, `conquer`
   - Null cases
4. Other
   - other proofs...

1. Introduction.

1.1. *DEC vs EC.* Using data from GSE75748, comparing conditions DEC vs EC. We have genes potentially to be DD uniquely identified by scDDboost. (supplementary Fig 1)



Supplementary Figure 1: Potentially to be DD genes uniquely identified by us.

1.2. *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. The change of proportion of subtypes A_π is independent with mean expression change between subtypes $M_{g,\pi}$ given expression data X, condition label y and cluster label z. Further, conditioning on condition label y and cluster label z, the change of proportion of subtypes A_π is independent with the expression data X and similarly conditioning on z and X, $M_{g,\pi}$ is independent with y. Thus we have $P\left(A_\pi \cap M_{g,\pi}|X,y,z\right) = P\left(A_\pi|y,z\right)\ P\left(M_{g,\pi}|X,z\right)$. □

2.2. *Method Structure and Clustering.*

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

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

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

The EBSeq model assumed a prior distribution on $q_g^I : q_g^I | \alpha, \beta^{I_g} \sim Beta(\alpha, \beta^{I_g})$. The hyperparameter α is shared by all the isoforms and β^{I_g} is I_g specific. Now we are using EBSeq for expression inference of genes rather than isoforms, we made a modification to the original prior from EBSeq. We still make α to be shared by all the genes and β^g becomes gene specific parameter.

The modeling of $r_{g,s}$ is the same as what EBSeq does. Specifically, we further assume that $r_{g,s} = r_{g,0} * l_s$ where $r_{g,0}$ is an isoform specific parameter common across subtypes and $r_{g,s}$ depends on it through the sample-specific normalization factor l_s

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

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

For any partition π of K elements.

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

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

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

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

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

$$
\begin{aligned}
h_1^g(X_g^{1,2,3}) &= f_0^g(X_g^{1,2,3}) \\
h_2^g(X_g^{1,2,3}) &= f_0^g(X_g^{1,2}) f_0^g(X_g^3) \\
h_3^g(X_g^{1,2,3}) &= f_0^g(X_g^1) f_0^g(X_g^{2,3}) \\
h_4^g(X_g^{1,2,3}) &= f_0^g(X_g^{1,3}) f_0^g(X_g^2) \\
h_5^g(X_g^{1,2,3}) &= f_0^g(X_g^1) f_0^g(X_g^2) f_0^g(X_g^3)
\end{aligned}
$$

Then the marginal distribution of counts $X_g^{1,2,3}$ is $\sum_{k=1}^{5} p_k h_k^g(X_g^{1,2,3})$, where proportion parameters p_k satisfying $\sum_{k=1}^{5} p_k = 1$ and are estimated by EM algorithm. Thus, the posterior probability of an expression pattern k is obtained by:

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

In the optimization steps for determining the hyper parameters α, β^g and proportion of DE patterns p, the computation and memory increase exponentially with the number of subtypes K. We use one-step EM as an approximation for the solution, that is α and β^g are updated through gradient aescent. p can still be updated by the explicit form of the maximizer of the log likelihood.

2.2.2. *modalClust.* **Product Partition Model**

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

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

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

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

Dahl demonstrated by some choice of f and prior of π, we can reduce the time complexity of finding the MAP partition from factorial(n) to $O(n^2)$ (Dahl, 2009), And the crucial condition for f is that if X_{S_1} and X_{S_2} are overlapped in the sense that $\min\{X_{S_2}\} < \max\{X_{S_1}\} < \max\{X_{S_2}\}$ or $\min\{X_{S_1}\} <$

$\max\{X_{S_2}\} < \max\{X_{S_1}\}$, $X_{S_1^*} and X_{S_2^*}$ be the sets of swapping one pair of those overlapped terms and keep the other unchanged. Then $f(X_{S_1})f(X_{S_2}) \leq f(X_{S_1^*})f(X_{S_2^*})$. Under such condition, we know that possible MAP candidates must be those partition that for any two subgroups of data, all the data from subgroup1 has to be either greater or smaller than all the data from subgroup2.

In Poisson-Gamma Model we assuming:

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

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

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

$f(X_S)$ still satisfying the condition mentioned

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

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

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

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

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

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

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

Similarly,

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

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

Similar proof for the case $h_1 > h_2$.

□

2.2.3. *Randomized K−means.*

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

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

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

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

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

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, $P(\phi|\psi)! = P(\phi)$ as there is some subsets b such that $\sum_{i\in b}\phi = \sum_{i\in b}\psi$. So ϕ and ψ are dependent □

Proof of property 2

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

Proof of property 3

PROOF. t^1/t_π^1 is independent with t^2/t_π^2 conditioning on t_π^1 and t_π^2 by the Neutrality property of dirichlet distribution □

Proof of property 4

PROOF. For $j = 1, 2$, let T_b^j be the vector of t_k^j such that $k \in b$. Recall $t_b^j = \sum_{k\in b} t_k^j$. Without loss of generality, we consider the case condition $j = 1$. We can decompose the density to each blocks by the property of multinomial distribution

$$p_\pi(t^1|t_\pi^1, y) = \prod_{b\in\pi}(p(T_b^1|t_b^1, y))$$

and prior predictive function can be obtained via integral out $\tilde{\phi}_b$ given the prior Dirichlet$[\alpha_b^1]$ and $p(T_b^1|\tilde{\phi}_b)$ is multinomial$(\tilde{\phi}_b)$ distributed.

$$\begin{aligned} p(T_b^1|t_b^1, y) &= \int_{\tilde{\phi}_b} p(T_b^1|\tilde{\phi}_b)p(\tilde{\phi}_b)d\tilde{\phi}_b \\ &= \left\{\left[\frac{\Gamma(t_b^j+1)}{\prod_{k\in b}\Gamma(t_k^j+1)}\right]\left[\frac{\Gamma(\sum_{k\in b}\alpha_k^j)}{\prod_{k\in b}\Gamma(\alpha_k^j)}\right]\left[\frac{\prod_{k\in b}\Gamma(\alpha_k^j+t_k^j)}{\Gamma(t_b^j+\sum_{k\in b}\alpha_k^j))}\right]\right\} \end{aligned}$$

□

Proof of property 5

PROOF. t_π^1 and t_π^2 given the condition label y are independent identical distributed. $t_\pi^1|\Phi \sim \text{multinomial}(\Phi)$

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

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

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}*

in order to proof property 6 we gave a lemma of dimensionality of the intersection of two A_πs 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 property 6

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

Proof of theorem 3

PROOF. $p_\pi(t^1|t_\pi^1, y) p_\pi(t^2|t_\pi^2, y) p_\pi(t_\pi^1, t_\pi^2|y)$ The DDM: $p(\phi, \psi) = \sum_{\pi \in \Pi} p_\pi(\phi, \psi)$. 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 $p(y, z|\phi, \psi) p_\pi(\phi, \psi) = p(y, z|\tilde{\phi}, \tilde{\psi}, \Phi_\pi) p(\tilde{\phi}) p(\tilde{\psi}) p(\Phi_\pi)$ consider the support of $p_\pi(\phi, \psi)$.
Right hand side of the above equation is

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

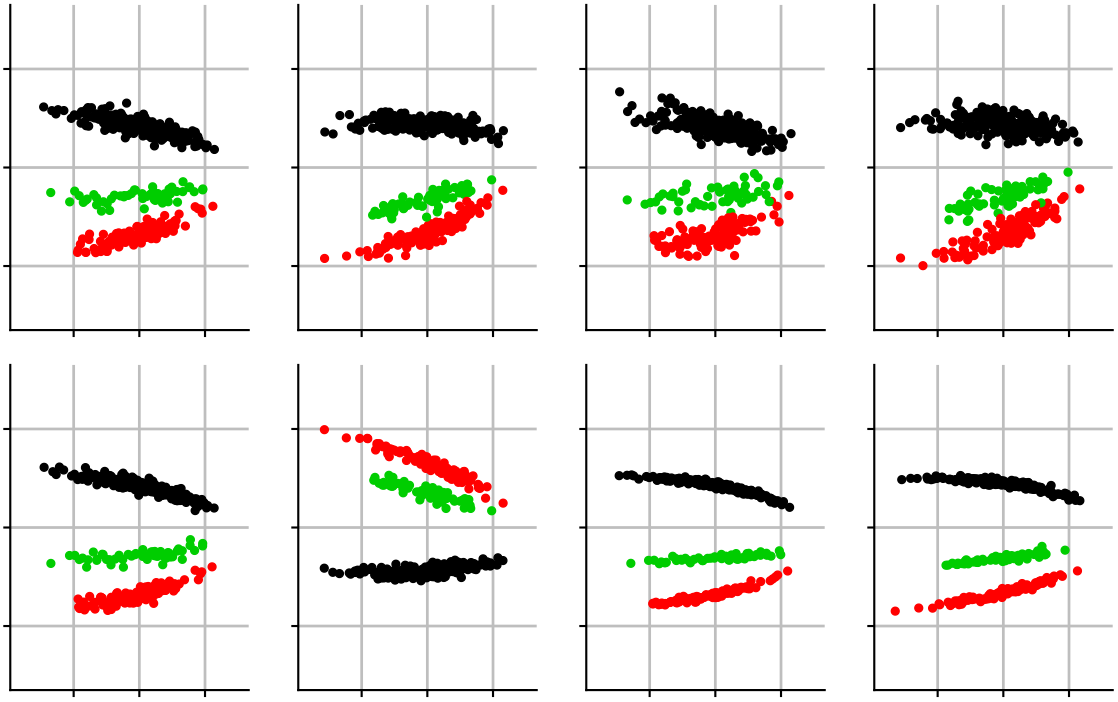
Where A_1 is the product of normalizing terms from multinomial distribution of z^1 and z^2, $A_1 = \frac{\Gamma(n_1+1)\Gamma(n_2+1)}{\prod_{j=1}^{2} \prod_{k=1}^{K} \Gamma(t_k^j + 1)}$

A_2 is the product of normalizing terms from Dirichlet distribution of $\tilde{\phi}$ and $\tilde{\psi}$, $A_2 = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k^1 + 1)\Gamma(\sum_{k=1}^{K} \alpha_k^2 + 1)}{\prod_{j=1}^{2} \prod_{k=1}^{K} \Gamma(\alpha_k^j + 1)}$

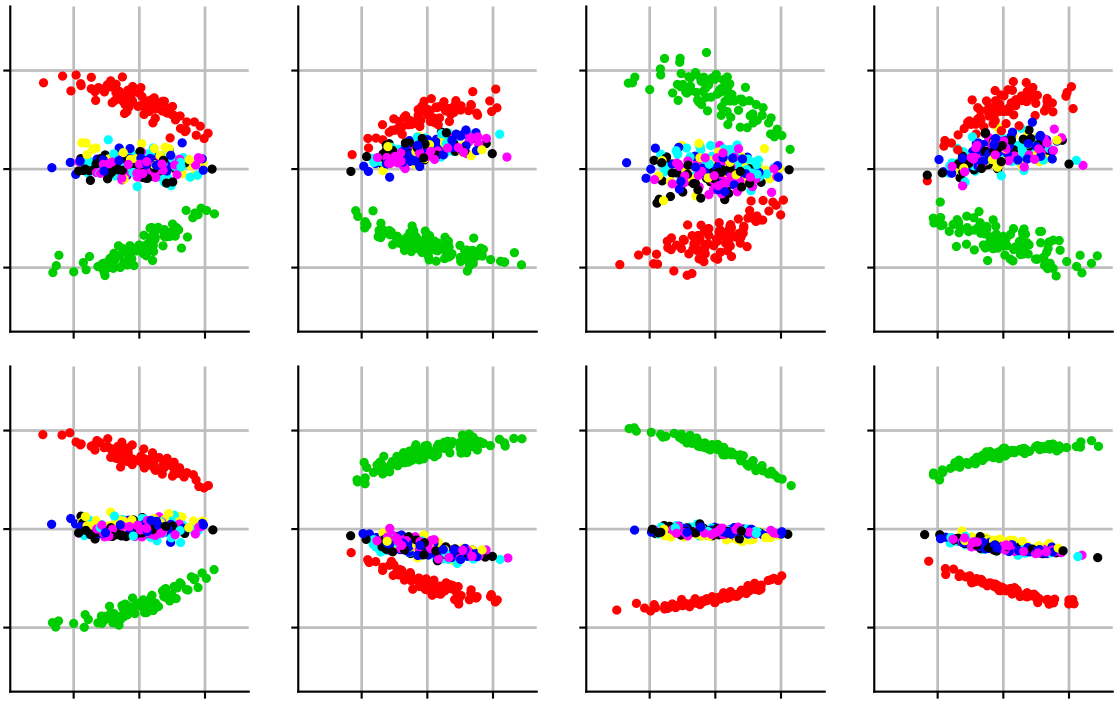
A_3 is the normalizing term from Dirichle distribution of Φ_π, $A_3 = \frac{\Gamma(\sum_{b \in \pi} \beta_b + 1)}{\prod_{b \in \pi} \Gamma(\beta_b + 1)}$
To convert U_π to have form similar of p_π, We need $U_\pi \propto f_1 f_2 f_3$ proportional to product of three Dirichlet densities. We know $f_1 \sim \text{Dirichlet}[\alpha^1 + t^1]$, $f_2 \sim \text{Dirichlet}[\alpha^2 + t^2]$ and $f_3 \sim \text{Dirichlet}[\beta + t^1 + t^2]$. Considering the normalizing factors for f_1, f_2 and f_3, and multiplying them with A_1, A_2 and A_3. We have the $U_\pi = C_\pi * f_1 f_2 f_3$. The final normalizing term is $p_\pi(t^1|t_\pi^1, y)\, p_\pi(t^2|t_\pi^2, y)\, p_\pi(t_\pi^1, t_\pi^2|y)$ Then we have $(\phi, \psi)|y, z \sim \text{DDM}\left[\omega^{\text{post}} = (\omega_\pi^{\text{post}}), \alpha^1 + t^1, \alpha^2 + t^2\right]$ and $\omega_\pi^{\text{post}} \propto p_\pi(t^1|t_\pi^1, y)\, p_\pi(t^2|t_\pi^2, y)\, p_\pi(t_\pi^1, t_\pi^2|y)\, \omega_\pi$.
Notice in DDM, we constrained $\beta = \alpha^1 + \alpha^2$.

□

3. Numerical Experiments.



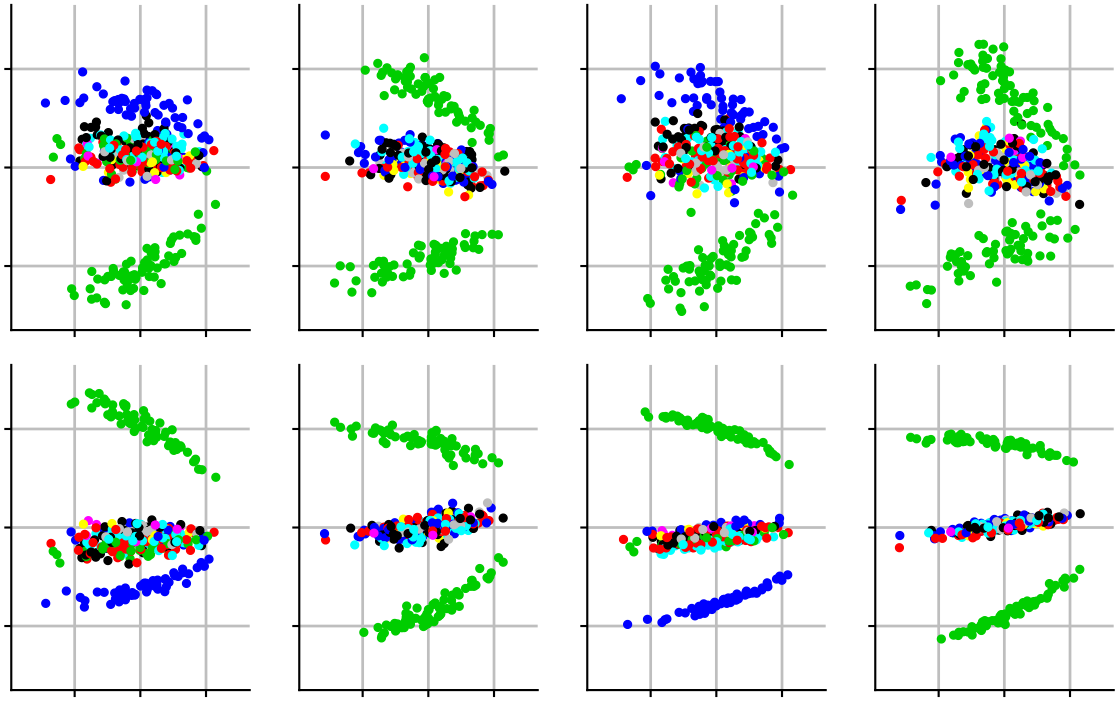
Supplementary Figure 2: first two principal components of transcripts under different parameters for simulated data. Different parameters resulted in different degree of separation of subtypes. We have 4 different settings for hyper-parameters of simulation, each setting has 2 replicates K = 3



Supplementary Figure 3: similar plots K = 7

3.1. *Synthetic Data.* We first look at the pca plots of the simulated data (supplementary figure 2,3,4)

We also have roc curve for the simulated data, each sub-figure is averaged over two replicates under the same parameters setting. scDDboost tends to outperform other methods (supplementary figure 5)



Supplementary Figure 4: similar plots K = 12

3.2. *Empirical Study.* **Data sets** details for the datasets used in the empirical studies of the main paper (supplementary table 1)

3.3. *Null cases.* datasets used for generating the Null cases (supplementary table 2)

scDDboost may lose FDR control if we keep pushing number of subtypes K bigger.

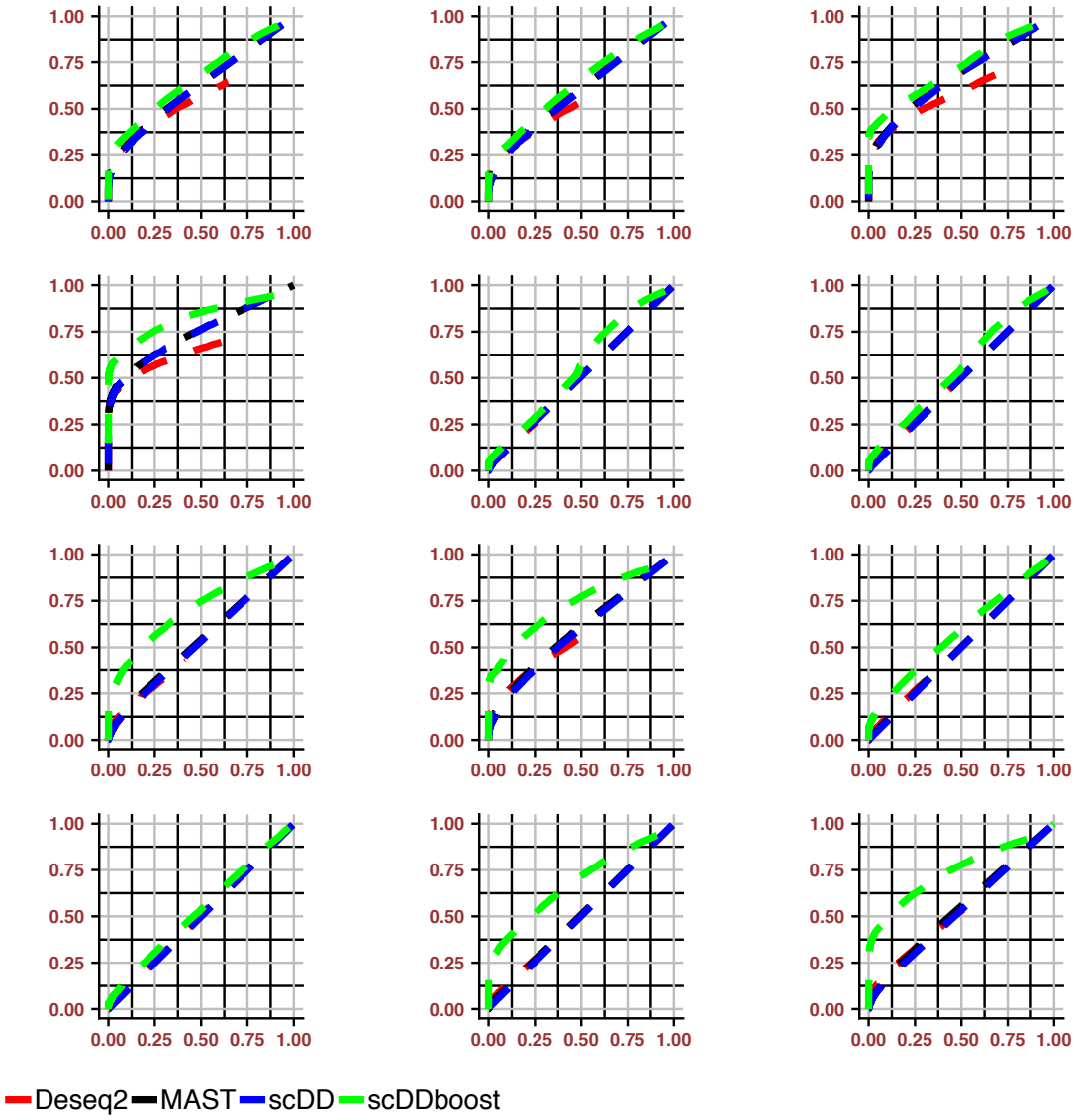
4. Random weighting and proof for consistency theorems.

4.1. *Stability of posterior under random weighting.* Number of subtypes K is a crucial factor controlling the accuracy of our modeling. Too small K may end up in an underfit such that cells within same subtype can still be very different, mean expression change among subtypes is incapable to capture the distribution change for some genes and consequently reducing the power of scDDboost. Too big K may end up in an overfit such that two subtypes can be very similar, given we have fixed number of samples (cells), allowing more clusters will introduce may patterns (both for mean expression change and proportion change) to infer. Also notice the limitation of DDM model (see section 4), overestimating K in scDDboost may losing FDR control (Fig7).

something about change of PDD over K, even though PDD is monotone increasing but it would remain stable in the sense that $\text{PDD}_{K+1} - \text{PDD}_K$ will have small variance over different genes

From our empirical experience, it would be sufficient to capture the heterogeneity underlying cells with number of clusters not greater than 9. And we generally obtain stable in validity score and PDD simultaneously (see supplementary)

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.



Supplementary Figure 5: 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

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

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

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

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	546, 664	human	(Darmanis et al., 2017)	9

SUPPLEMENTARY TABLE 1
datasets used for comparisons of DD analysis under different methods

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

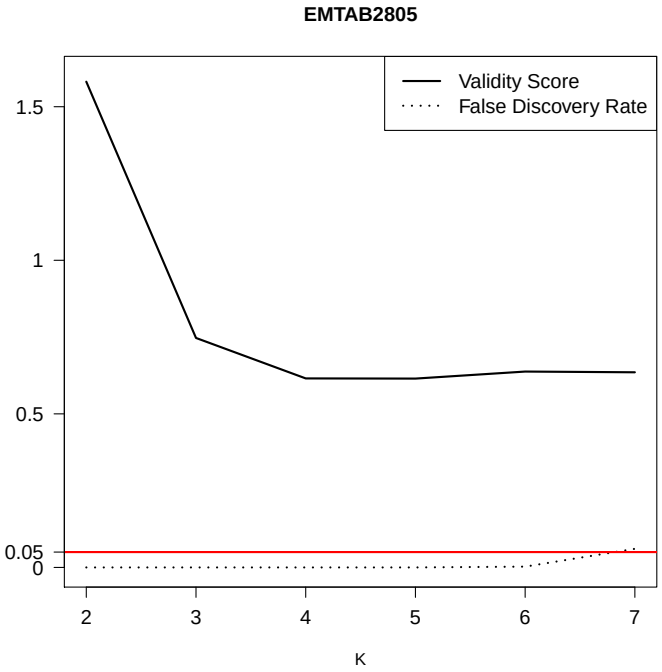
SUPPLEMENTARY TABLE 2
datasets used for null cases, as cells are coming from same biological condition, there should not be any differential distributed genes, any positive call is false positive

is

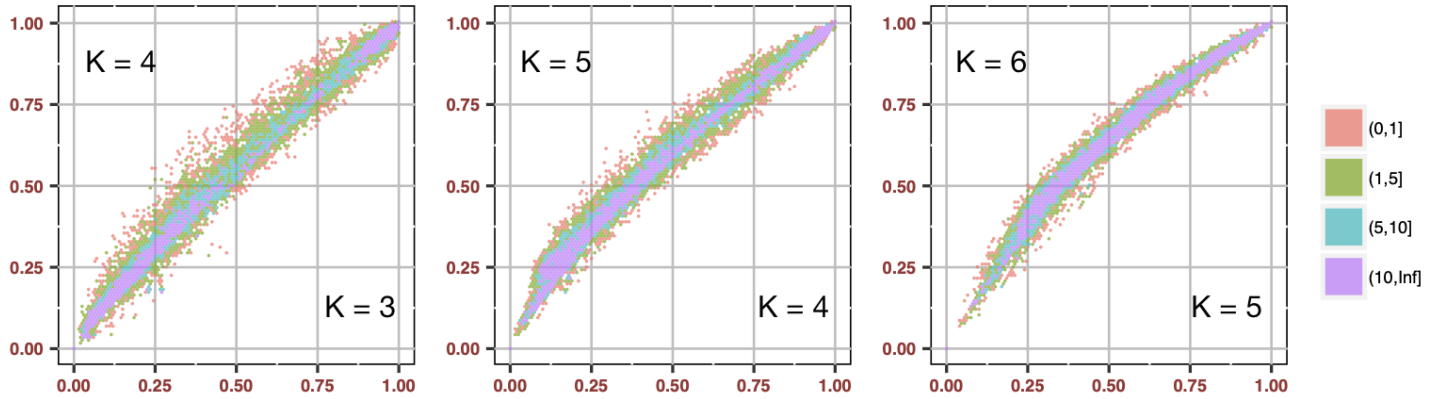
$$p(A_\pi|t^1,t^2) = c * \sum_{\pi' \text{ refines } \pi} p(t^1|t^1_{\pi'})\, p(t^2|t^2_{\pi'})\, p(t^1_{\pi'},t^2_{\pi'}|A_{\pi'}) \tag{3}$$

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:

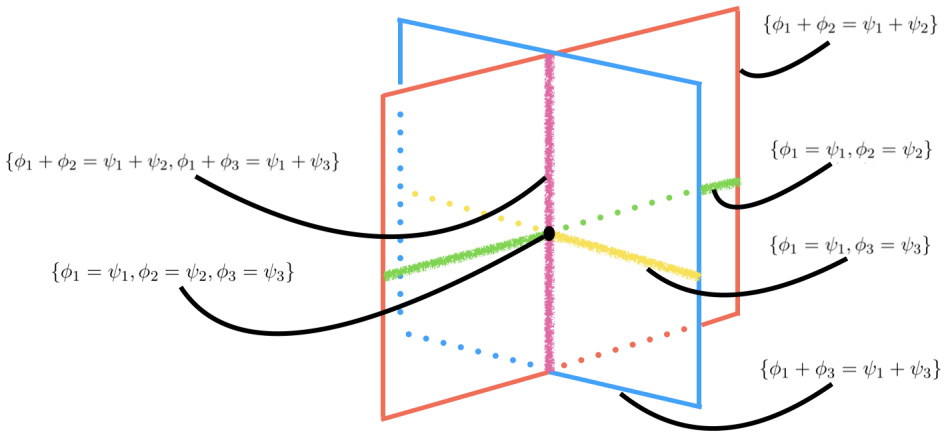


Supplementary Figure 6: under NULL case, using dataset EMTAB2805, when using too big K we may lose FDR control (black dashed line shows proportion of false positive identified by scDDboost under 0.05 threshold, while validity score did not vary too much after K is greater than 2



Supplementary Figure 7: PDD change under different number of subtypes K, dataset used DEC-EC

In Fig 10, there are four subtypes, the rectangle with magenta boundary is a simplex $A_{\pi_1} = \{(\phi, \psi) : \phi_1 + \phi_2 = \psi_1 + \psi_2\}$, the rectangle with blue boundary is a simplex $A_{\pi_2} = \{(\phi, \psi) : \phi_1 + \phi_3 = \psi_1 + \psi_3\}$. 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



Supplementary Figure 8: Four subtypes of cells, simplexes of (ϕ, ψ) satisfying different constraints.

that A^*_π will be disjoint subsets of Ω.

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

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

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

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

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

But for $(\phi, \psi) \in \Omega \setminus Q$, we have consistent posterior inference.

THEOREM 1. *Let* $n = min(n_1, n_2)$ *be the smaller number of cells of two conditions and* $n_1 = O(n_2)$ *namely* $ln(\frac{n_1}{n_2}) = 0$, *and hyper parameters of DDM* α^1, α^2 *be vectors of constants,* $\alpha^j_k \geq 1$, $\forall k, j$ *and* $\beta = \alpha^1 + \alpha^2$. *Then if parameter* $(\phi, \psi) \in \Omega \setminus Q$ *we have*

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

Things become more complicate when (ϕ, ψ) falling into Q, we know $p(Q|y, z)$ vanishes, but $p(A_\pi|y, z)$ 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 $S = \{\pi_1, \pi_2, \pi_3\}$, $N^* = 3$, $S^* = \{\pi_2\}$.

THEOREM 2. *Following the setting in theorem 1, when parameter $(\phi, \psi) \in Q$, and further if $\alpha^j, j = 1, 2$ are vectors of integers, we have*

$$(p(A_\pi|y,z))_{\pi \in S^*} \xrightarrow[n\to\infty]{d} (V_1, ..., V_{N(S^*)})$$

$V_1, ..., V_{N(S^*)}$ are random variables and $V_1 + .. + V_{N(S^*)} = 1$

Still using above example, in limiting case, we have $p(A_{\pi_3}|y,z) = 1$, $p(A_{\pi_2}|y,z) = 1$ and $p(A_{\pi_1}|y,z) = 0$. When the DE pattern is B_{π_1} for some genes and our estimation of $p(A_{\pi_1}|y,z) = 0$, 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.* In this section, we gave an intuitive justification for consistency between bayesian framework clustering analysis and random weighting procedure. A full bayesian analysis for clustering needs to specify the density of data given the partition. Specifically, in single cell analysis we need to know the density of transcripts of genes given the partitions which requires understanding of co-expression and dependence between genes. Instead of trying to untangle the mystery behind the dependence of genes, we consider following approximation

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

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

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

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

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

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

and

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

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

Randomized k-means

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

We are determining the parameter for weights through maximizing the marginal likelihood of our estimated distance $d_{i,j}$. More specifically, we gave a simple bayesian framework for distance in the following way: given the true distance $\Delta_{i,j}$, $d_{i,j}|\Delta_{i,j} \sim \text{Gamma}(a_1, a_1/\Delta_{i,j}))$, so that the sampling mean of $d_{i,j}$ is $\Delta_{i,j}$. We suppose that a-priori $1/\Delta_{i,j} \sim \text{Gamma}(a_0, d_0)$. And our weights $w_{i,j} \sim \text{Gamma}(a, a)$. $a = a_0 + a_1$. Notice that $E(\Delta_{i,j})/Var(\Delta_{i,j}) = d_0$. We approximate d_0 by the ratio of first and second moments of $d_{i,j}$, Once we have the estimated of d_0, we plug it into the marginal likelihood of $d_{i,j}$ and obtain the MLE of a_0 and a_1. An issue is that the frequently used optimizing function in r (optim or nlminb) typically gave extreme large value of a_0 or a_1, which will yield a degenerate weights without randomness. The reason is the stopping criterion as relative tolerance is too small under default setting. We pick a relative big value for the stopping threshold and obtain much smaller a_0 and a_1 while maintain the likelihood close to the optimal value under default threshold.

We also justified our choice of stopping threshold is reasonable. We plot the adjusted random index between the randomly generated clustering to original clustering under the distance without dividing by the random weights across eight datasets. Though the mean varies, the length between the 25% and 75% is wide enough presenting a reasonable variation of our randomly generated clustering.

Proofs:

As the density of DDM is computed by product or ratio over bunches of gamma function and gamma function is not easy to direct work on it and derive limiting theorem. To proof theorem 4 and 5, we need a crucial lemma which gave us an approximation to the gamma function, namely

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

PROOF. By (**?**), we have $\frac{x^{x-c}}{e^{-x}} \leq \Gamma(x) \leq \frac{x^{x-1/2}}{e^{-x}}$ 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,* $\sqrt{2\pi}n^{n+1/2}e^{-n} \leq \Gamma(n+1) \leq en^{n+1/2}e^{-n}$

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

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

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

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 $\frac{\omega_{\pi_1}^{post}}{\omega_{\pi_2}^{post}} = \frac{g(\pi_1,\alpha,\beta,n_1,n_2)}{g(\pi_2,\alpha,\beta,n_1,n_2)} \frac{f(\pi_1,t^1,t^2,\alpha,\beta)}{f(\pi_2,t^1,t^2,\alpha,\beta)}$ 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 $\frac{\omega_{\pi_1}^{post}}{\omega_{\pi_2}^{post}}$, 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, $(x-c)\ln(x) - x \leq h(x) \leq (x-1/2)\ln(x) - x$ we have

$$d(\pi, t^1, t^2, \alpha, \beta) \geq \sum_{b\in\pi} (\beta_b + t_b^1 + t_b^2 - c)\ln(\beta_b + t_b^1 + t_b^2) - \sum_{j=1}^{2}\sum_{b\in\pi} (t_b^j + \gamma_b^j - 1/2)\ln(t_b^j + \gamma_b^j) \tag{4}$$

$$d(\pi, t^1, t^2, \alpha, \beta) \leq \sum_{b\in\pi} (\beta_b + t_b^1 + t_b^2 - 1/2)\ln(\beta_b + t_b^1 + t_b^2) - \sum_{j=1}^{2}\sum_{b\in\pi} (t_b^j + \gamma_b^j - c)\ln(t_b^j + \gamma_b^j) \tag{5}$$

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

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

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

where $X_1 = \frac{(t_b^1 - t_b^2 + \gamma_b^1 - \gamma_b^2)^2}{t_b^1 + \gamma_b^1}$, $X_2 = \frac{(t_b^1 - t_b^2 + \gamma_b^1 - \gamma_b^2)^2}{t_b^2 + \gamma_b^2}$, $Y_1 = \frac{(t_b^1 - t_b^2 + \gamma_b^1 - \gamma_b^2)^3}{(t_b^1 + \gamma_b^1)^2}$, $Y_2 = \frac{(t_b^1 - t_b^2 + \gamma_b^1 - \gamma_b^2)^3}{(t_b^2 + \gamma_b^2)^2}$ 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

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

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

Using above inequalities, we have

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

By CLT we know X_b^1, X_b^2 are asymptotic gamma(χ-square) distributed and Y_b^1, Y_b^2 are $o_p(1)$, $g(\xi_b)$ has bounded variance and $T(\pi_1) - U(\pi_2) = -\ln(n)$ if $N(\pi_2) < N(\pi_2)$ as $\ln(\beta_b + t_b^1 + t_b^2) - \ln(\beta_{b'} + t_{b'}^1 + t_{b'}^2) = \ln(\frac{\beta_b + t_b^1 + t_b^2}{n_1}) - \ln(\frac{\beta_{b'} + t_{b'}^1 + t_{b'}^2}{n_1}) \to 0 \quad a.s.$ 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

(6)

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

(7)

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

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

where $D(\pi) = \Sigma_{b\in\pi}\left[1/2\ln(\beta_b + t_b^1 + t_b^2) - c(\ln(1 + \frac{t_b^2+\gamma_b^2}{t_b^1+\gamma_b^1}) + \ln(1 + \frac{t_b^1+\gamma_b^1}{t_b^2+\gamma_b^2}))\right]$ 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 and theorem 5

Proof. Recall $\Sigma_\pi \omega_\pi^{post} = 1$ and $P(A_\pi|y,z) = \sum_{\tilde{\pi}\in\Pi} \omega_{\tilde{\pi}}^{\text{post}} 1[\tilde{\pi} \text{ refines } \pi]$. 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 1 by lemma 4.

Similarly we use lemma 5 could proof theorem 2.

□

l

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.

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

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

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.

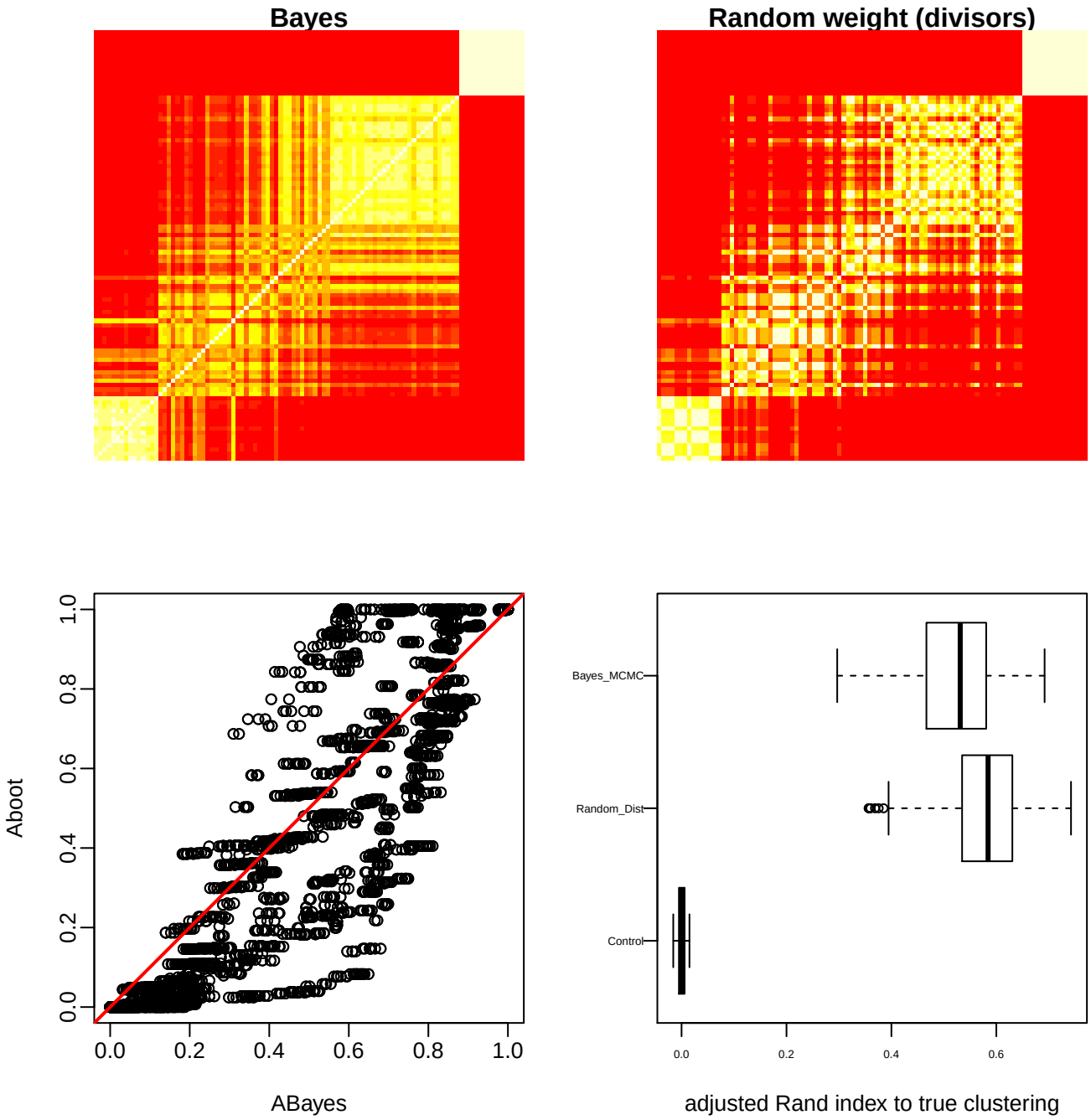
Leng, N., Chu, L.-F., Barry, C., Li, Y., Choi, J., Li, X., Jiang, P., Stewart, R. M., Thomson, J. A. and Kendziorski, C. (2015). Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nature Methods* **12** 947 EP -.

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

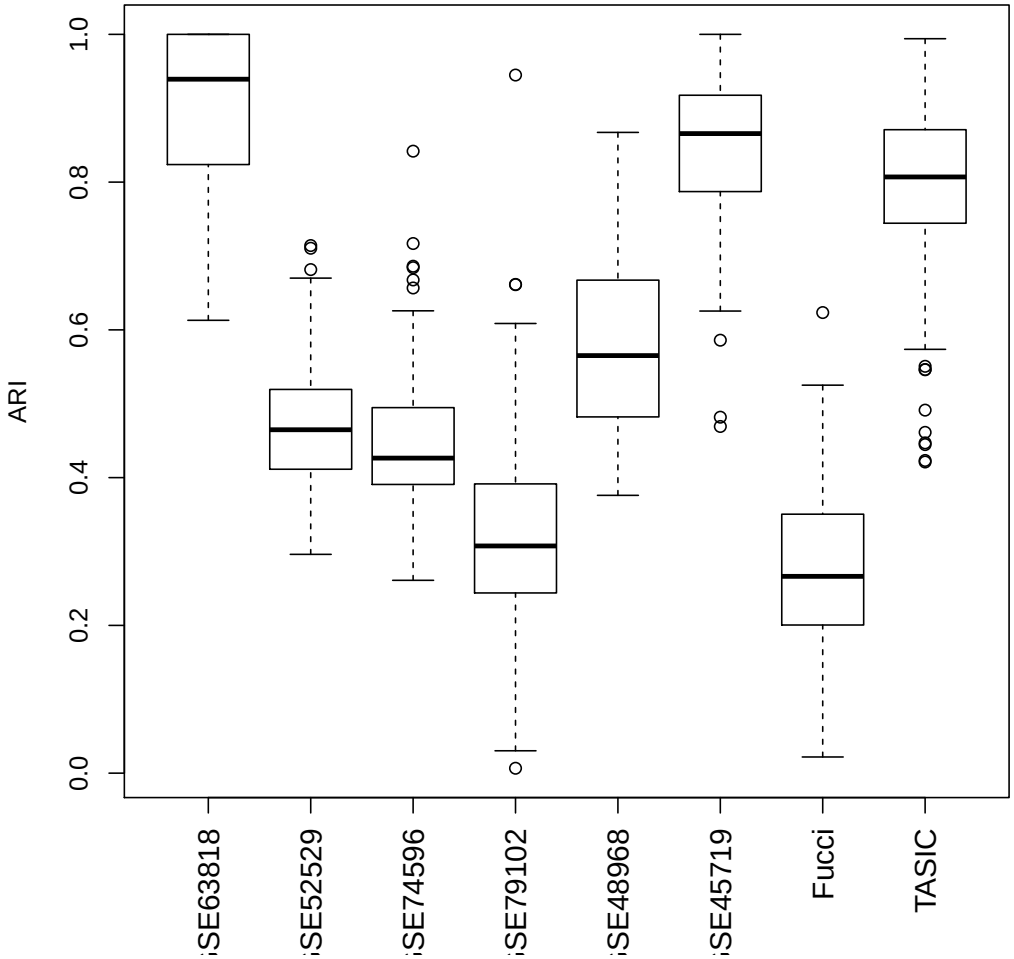
Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T., Yosef, N., Schwartz, S., Fowler, B., Weaver, S., Wang, J., Wang, X., Ding, R., Raychowdhury, R., Friedman, N., Hacohen, N., Park, H., May, A. P. and Regev, A. (2014). Single-cell RNA-seq 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.



Supplementary Figure 9: comparison between random weighting scheme and bayesian clustering procedure in terms of posterior probabilities that two elements belong to the same class given the whole data and adjusted rand index comparing to the underlying true class label



Supplementary Figure 10: Adjusted rand indexes to the clustering based on the original distance matrix without dividing weights. We investigate the randomness of clustering given by our weights through 8 datasets. All have stopping threshold for nlminb optimizing function in r with relative tolerance as 0.001