Flexible Experimental Designs for Valid Single-cell RNA-sequencing Experiments Allowing Batch Effects Correction

Fangda Song, Ga Ming Chan and Yingying Wei*

Department of Statistics, The Chinese University of Hong Kong

Abstract

Despite their widespread applications, single-cell RNA-sequencing (scRNA-seq) experiments are still plagued by batch effects and dropout events. Although the completely randomized experimental design has frequently been advocated to control for batch effects, it is rarely implemented in real applications due to time and budget constraints. Here, we mathematically prove that under two more flexible and realistic experimental designs—the "reference panel" and the "chain-type" designs—true biological variability can also be separated from batch effects. We develop **B**atch effects correction with **U**nknown **S**ubtypes for scRNA-**seq** data (BUSseq), which is an interpretable Bayesian hierarchical model that closely follows the data-generating mechanism of scRNA-seq experiments. BUSseq can simultaneously correct batch effects, cluster cell types, impute missing data caused by dropout events, and detect differentially expressed genes without requiring a preliminary normalization step. We demonstrate that BUSseq outperforms existing methods with simulated and real data.

Keywords: Batch effects; Experimental design; Single-cell RNA-seq experiments; Model-based clustering; Integrative analysis

*Correspondence should be addressed to Yingying Wei (yweicuhk@gmail.com)

Introduction

Single-cell RNA-sequencing (scRNA-seq) technologies enable the measurement of the transcriptome of individual cells, which provides unprecedented opportunities to discover cell types and understand cellular heterogeneity [1]. However, like the other high-throughput technologies [2–4], scRNA-seq experiments can suffer from severe batch effects [5]. Moreover, compared to bulk RNA-seq data, which measure the average gene expression levels of a cell population, scRNA-seq data can have an excessive number of zeros that result from dropout events—that is, the expressions of some genes are not detected even though they are actually expressed in the cell due to amplification failure prior to sequencing [6]. Consequently, despite the widespread adoption of scRNA-seq experiments, the design of a valid scRNA-seq experiment that allows the batch effects to be removed, the biological cell types to be discovered, and the missing data to be imputed remains an open problem.

One of the major tasks of scRNA-seq experiments is to identify cell types for a population of cells [1]. Therefore, the cell type of each individual cell is always unknown and is the target of inference. However, most existing methods for batch effects correction such as Combat [7] and SVA [8, 9] are designed for bulk experiments and require knowledge of the subtype information, which corresponds to cell type information for scRNA-seq data, of each sample a priori. As a result, they are not feasible for scRNA-seq data, for which clustering is the main interest.

To jointly cluster samples across batches, Huo et al. [10] proposed MetaSparseKmeans. Unfortunately, MetaSparseKmeans requires all subtypes to be present in each batch. Suppose that we conduct scRNA-seq experiments for blood samples from a healthy individual and a leukemia patient, one person per batch. Although we can anticipate that the two batches will share T cells and B cells, we do not expect that the healthy individual will have cancer cells as the leukemia patient. Therefore, MetaSparseKmeans is not applicable to scRNA-seq data.

The mutual nearest neighbors (MNN) [11] approach allows each batch to contain some but not all cell types. However, MNN requires that "the batch effect is almost orthogonal to the biological subspaces" and "the batch-effect variation is much smaller than the biological-effect variation between different cell types." These are very strong assumptions and cannot be validated at the design stage of the experiments. If we know a set of control genes

whose expression levels are supposed to be the same across different subtypes or a group of control samples who come from the same subtypes, then we can also apply RUV [12, 13]. However, the selection of control genes and control samples is not an easy job for scRNA-seq experiments. Seurat adopts Canonical Correlation Analysis (CCA) to identify the shared variations across batches and treats them as the shared cell types [14]. However, if some batches share certain technical noises, for example when each patient is measured by several batches, CCA can also mistake the technical variability as biological variability of interest. Recently, Luo and Wei [15] developed a hierarchal model called BUS to simultaneously cluster samples across multiple batches and correct for batch effects for microarray data. BUS can handle severe batch effects. In addition, Luo and Wei [15] mathematically showed flexible experimental designs under which batch effects can be corrected when subtype information is unknown. Nevertheless, like MetaSparseKmeans, MNN, RUV and Seurat, BUS does not consider features such as the count nature of the data, overdispersion [16], dropout events [6], and the cell-specific size factors [17], which are all unique to scRNA-seq data.

Two recently proposed factor models—ZIFA [18] and ZINB-WaVE [19]— are aware of dropout events and use the first several inferred components for clustering. Factor models are approximations to the true mixture distributions of distinct cell types, and the authors do not discuss the experimental designs under which their methods are applicable. However, it is crucial to understand the conditions under which biological variability can be separated from technical artifacts. Obviously, for completely confounded designs—for example one in which batch 1 measures cell type 1 and 2, whereas batch 2 measures cell type 3 and 4—no method is applicable.

Here, we propose Batch Effects Correction with Unknown Subtypes for scRNA-seq data (BUSseq), an interpretable hierarchical model that simultaneously corrects batch effects, clusters cell types, and takes care of the count data nature, the overdispersion, the dropout events, and the cell-specific size factors of scRNA-seq data. Despite the cell-specific size factors and the dropout rates, we can mathematically prove that the same experimental designs under which batch effects can be corrected when the subtype information is unknown for bulk experiments [15] are also valid for scRNA-seq experiments. Specifically, in addition to the commonly advocated completely randomized design [1, 5, 20, 21], in which each batch measures all cell types, scRNA-seq experiments with the "reference panel" design

and the "chain-type" design, which allow some cell types to be missing from some batches, are also legitimate. We demonstrate that BUSseq outperforms the existing approaches in both simulation data and in real applications. We envision that the proposed experimental designs will be able to guide biomedical researchers and help them to design better scRNA-seq experiments.

Results

BUSseq is an interpretable hierarchical model for scRNA-seq

In this work, we develop a hierarchical model BUSseq that closely mimics the data generating procedure of scRNA-seq experiments (**Figure 1** and Methods). Given that we have measured B batches of data each with a sample size of n_b, let us denote the true underlying gene expression level of gene g in cell i of batch b as X_{big}. X_{big} follows a negative binomial distribution with mean expression level μ_{big} and a gene-specific and batch-specific overdispersion parameter ϕ_{bg}. The mean expression level is determined by the cell type W_{bi} with the cell type effect β_{gk}, the log-scale baseline expression level α_g, the location batch effect ν_{bg}, and the cell-specific size factor δ_{bi}. The cell-specific size factor δ_{bi} characterizes the impact of cell size, library size and sequencing depth. It is of note that the cell type W_{bi} of each individual cell is unknown and is our target of inference. Therefore, we assume that a cell on batch b comes from cell type k with probability $P(W_{bi} = k) = \pi_{bk}$, and the proportions of cell types $(\pi_{b1}, \cdots, \pi_{bK})$ vary among batches.

Unfortunately, we cannot always observe the true expression level X_{big}s. Without dropout, indicated by $Z_{big} = 0$, we can directly see $Y_{big} = X_{big}$. However, if a dropout event occurs—$Z_{big} = 1$, then we observe $Y_{big} = 0$ instead of the true level X_{big}. It has been observed that highly expressed genes are less-likely to suffer from dropout events [6]. Therefore, we model the dependence of the dropout rate $P(Z_{big} = 1|X_{big})$ on the true expression level using a logistic regression with batch-specific intercept γ_{b0} and odds ratio γ_{b1}.

We only observe Y_{big} for all cells in the B batches and the total G genes. We conduct statistical inference under the Bayesian framework, and we develop a Markov chain Monte Carlo algorithm (MCMC) [22] to sample from the posterior distribution. Based on the

parameter estimates, we can learn the cell type for each individual cell, impute the missing underlying true expression levels X_{big} for dropout events, and identify genes that are differentially expressed among cell types. Moreover, our algorithm can automatically detect the total number of cell types K that exists in the dataset according to the Bayesian information criterion (BIC) [23]. BUSseq also provides a batch-effect corrected version of count data, which can be used for downstream analysis as if all of the data were measured in a single batch.

Valid experimental designs for scRNA-seq experiments

If a study design is completely confounded, as shown in (**Figure 2(a)**, then no method is able to separate the biological variability from the technical artifacts, because different combinations of batch-effect and cell-type-effect values can lead to the same probabilistic distributions for the observed data, which in statistics is termed a non-identifiable model. Statistical inference is impossible for non-identifiable models because two sets of distinct parameter values can give rise to the same probabilistic functions. In contrast, we can prove that the very flexible BUSseq model is identifiable under conditions that are very easily met. Therefore, BUSseq is applicable to a wide range of experimental designs.

For the "complete setting," in which each batch measures all of the cell types (**Figure 2(b)**), although the proportions of cell types can still differ from batch to batch, BUSseq is identifiable as long as: (I) the odds ratio γ_{b1}s in the logistic regressions for the dropout rates are negative for all of the batches, (II) every two cell types have more than one differentially expressed gene, and (III) the ratios of mean expression levels between two cell types $(\frac{exp(\beta_{1k})}{exp(\beta_{1\tilde{k}})}, \cdots, \frac{exp(\beta_{Gk})}{exp(\beta_{G\tilde{k}})})$ are different for each cell-type pair $(k, \tilde{k})$ (see Theorem 1 in Methods and its proof in Supplementary Materials). Condition (I) requires that the highly expressed genes are less likely to have dropout events, which is routinely observed for scRNA-seq data [6]. Condition (II) always holds in reality. Because scRNA-seq experiments measure the whole transcriptome of a cell, condition (III) is also always met in real data. For example, if there exists one gene g such that for any two cell-type pairs (k_1, k_2) and (k_3, k_4) its mean expression levels ratios $\frac{exp(\beta_{gk_1})}{exp(\beta_{gk_2})}$ and $\frac{exp(\beta_{gk_3})}{exp(\beta_{gk_4})}$ are not the same, then condition (III) is already satisfied.

The commonly advocated completely randomized experimental design falls into the

"complete setting," whereas the latter further relaxes the assumption implied by the former that the cell-type proportions are about the same for all batches. The identical composition of the cell population within each batch is a crucial requirement for traditional batch effects correction methods developed for bulk experiments such as Combat [11]. In contrast, BUSseq is not limited to this balanced design constraint and is applicable to not only the completely randomized design but also the general complete setting design.

Ideally, we would wish to adopt completely randomized experimental designs. However, in reality, we always fail to implement complete randomization due to time and budget constraints. For example, when we recruit patients sequentially, we often have to conduct scRNA-seq experiments patient-by-patient rather than randomize the cells from all of the patients to each batch, and the patients may not have the same set of cell types. Fortunately, we can prove that BUSseq also applies to two sets of flexible experimental designs, which allow cell types to be measured in only some but not all of the batches.

Assuming that conditions (I)-(III) are satisfied, and if there exists one batch that contains cells from all cell types and the other batches have at least two cell types (**Figure 2(c)**), then BUSseq can tease out the batch effects and identify the true biological variability (see Theorem 2 in Methods and its proof in Supplementary Materials). We call this setting the "reference panel design."

Sometimes, it can still be difficult to obtain a reference batch that collects all cell types. In this case, we can turn to the chain-type design, which requires every two consecutive batches to share two cell types (**Figure 2(d)**). Under the chain-type design, given that conditions (I)-(III) hold, BUSseq is also identifiable and can estimate the parameters well (see Theorem 3 in Methods and its proof in Supplementary Materials).

A special case of the chain-type design is when two common cell types are shared by all of the batches, which is frequently encountered in real applications. For instance, when blood samples are assayed, even if we perform scRNA-seq experiment patient-by-patient with one patient per batch, we know a priori that each batch will contain at least both T cells and B cells, thus satisfying the condition of the chain-type design.

For scRNA-seq data, the dropout rate depends on the true underlying expression levels. Such missing data mechanism is called missing not at random (MNAR) in statistics. It is very challenging to establish identifiability for MNAR [24]. Miao et al. [24] showed that

for many cases even when both the outcome distribution and the missing data mechanism have parametric forms, the model can be nonidentifiable. However, fortunately, despite the dropout events and the cell-specific size factors, by creating a set of functions similar to the probability generating function, we can still arrive at the same experimental designs as those for the bulk experiments [15] under which batch effects can be removed and cell types can be discovered. The reference panel design and the chain-type design liberalize researchers from the ideal but often unrealistic requirement of the completely randomized design.

BUSseq accurately estimates the parameters and imputes the missing data

We first evaluated the performance of BUSseq via a simulation study. We simulated a dataset with four batches and a total of five cell types under the chain-type design (**Figures 3(a-d)**). Every two consecutive batches share at least two cell types, but none of the batches contains all of the cell types. **Figure 3(d)** shows the observed data, which suffer from batch effects and dropout events. The t-SNE plot also confirms the observation (**Figure 4(a)**).

BUSseq correctly identifies the presence of five cell types among the cells (**Figure 3(j)**). Moreover, despite the dropout events, BUSseq accurately estimates the true cell type effects β_{gk}s (**Figures 3(a) and (f)**), the batch effects ν_{bg}s (**Figures 3(b) and (g)**), and the cell-specific size factors δ_{bi}s (**Figure 3(e)**). When controlling the Bayesian false discovery rate (FDR) at 0.05 [25, 26], we identify all intrinsic genes that differentiate cell types with the true FDR being 0.01 (Methods).

In the simulation study, we know the true expression levels X_{big}s. Therefore, we can compare them with our inferred expression levels $\widehat{X}_{big}$s based the observed data Y_{big}s which are subject to dropout events. **Figures 3(h)** demonstrate that BUSseq can learn the true expression levels well. This success arises because BUSseq uses an integrative model to borrow strengths both across genes and across cells from all batches. As a result, BUSseq can achieve accurate estimation and imputation despite the dropout events.

Combat offers a version of data that have been adjusted for batch effects [7]. Here, we also provide batch-effects-corrected count data based on quantile matching (Methods). The adjusted count data no longer suffer from batch effects and dropout events, and they even do

not need further cell-specific normalization (**Figure 3(i)**). Therefore, they can be treated as if measured in a single batch for downstream analysis.

BUSseq outperforms existing methods in batch effects correction and cell type clustering

We benchmarked BUSseq with the state-of-the-arts methods for batch effects correction for scRNA-seq data—MNN, ZINB-WaVE and Seurat. The adjusted Rand index (ARI) measures the consistency between two clustering results and is between zero and one, a higher value indicating better consistency. The ARI between the inferred cell types $\widehat{W}_{bi}$s by BUSseq and the true underlying cell types W_{bi}s is one. Thus, BUSseq can perfectly recover the true cell type of each cell. In comparison, we apply each of the compared methods to the dataset and then perform their own clustering approaches (Methods). The ARI is able to compare the consistency of two clustering results even if the numbers of clusters differ, therefore, we choose the number of cell types by the default approach of each method rather than set it to a common number. The resulting ARIs are 0.896 for MNN, 0.433 for Seurat and 0.961 for ZINB-WaVE. Moreover, the t-SNE plots (**Figure 4**) show that BUSseq and ZINB-WaVE can cluster the cells by cell types rather than batches. We also calculated the Silhouette score for each cell for each compared method. A high Silhouette score indicates that the cell is well matched to its own cluster and separated from neighboring clusters. **Figure 4(k)** shows that BUSseq gives the best segregated clusters.

BUSseq outperforms existing methods on hematopoietic data

We re-analyzed the two hematopoietic datasets previously studied by Haghverdi et al. [11], one profiled by the SMART-seq2 protocol for hematopoietic stem and progenitor cell populations from 12-week-old female mice and another assayed by the massively parallel single-cell RNA-sequencing (MARS-seq) protocol for myeloid progenitors from 6- to 8-week-old female mice. Although the two datasets were generated in two different laboratories, both datasets have cell-type label for each cell that is annotated according to marker expression levels from fluorescence-activated cell sorting (FACS) [11, 27] (Methods).

For each of the compared methods, we calculated the ARI between that method's clustering

and the FACS labels. The resulting ARIs are 0.587 for BUSseq, 0.324 for MNN, 0.478 for Seurat, and 0.470 for ZINB-WaVE. Thus, BUSseq outperforms all of the other methods in being consistent with FACS labeling. Both the t-SNE plots and the Silhouette widths (**Figure 5**) according to the FACS labeling confirm that BUSseq performs the best in segregating cells into different cell types.

BUSseq outperforms existing method on pancreas data

We further studied the four scRNA-seq datasets of human pancreas cells analyzed in Haghverdi et al. [11], two profiled by CEL-seq2 protocol [28, 29] and two assayed by SMART-seq2 protocol [28, 30]. These cells were isolated from deceased organ donors with and without type 2 diabetes. We obtained 7,095 cells after quality control (Methods) and treated each dataset as a batch following Haghverdi et al. [11].

For the two datasets profiled by the SMART-seq2 protocol, Segerstolpe et al. [30] and Lawlor et al. [28] provide cell-type labels; for the other two datastes assayed by the CEL-seq2 protocol, Haghverdi et al. [11] provides the cell-type labels based on the marker genes in the original publications [28, 29]. Therefore, we can compare the clustering results from each batch effects correction method with the labeled cell types. The pancreas is highly heterogeneous and consists of two major categories of cells—islet cells and non-islet cells. Islet cells include alpha, beta, gamma, and delta cells; and non-islet cells include acinar and ductal cells. BUSseq identifies a total of twelve cell types: five for islet cells, two for non-islet cells, two for the labeled "other" cells and three for a small group of outlier cells. Specifically, the five islet cell types identified by BUSseq correspond to two groups of alpha cells, a group of beta cells, a group of delta, and a group of islet cell mixture. The two non-islet cell types identified by BUSseq match exactly with the acinar and ductal cells, respectively. The ARIs with the cell-type labels are 0.381 for BUSseq, 0.370 for MNN, 0.207 for Seurat, and 0.277 for ZINB-WaVE; the means of Silhouette widths are -0.052 for BUSseq, -0.286 for MNN, -0.537 for Seurat and -0.269 for ZINB-WaVE. Once again, BUSseq outperforms all of the other methods in being consistent with the cell-type labels according to marker genes. (**Figure 6**)

BUSseq identifies 935 intrinsic genes at the Bayesian FDR cutoff of 0.05 (Methods). We conducted the gene set enrichment analysis [31] on the KEGG pathway database [32]. There are 30 significant pathways (q-values < 0.025). Among them, three pathways are

diabetes pathways; one is an insulin signaling pathway; and another 12 pathways are related to metabolism (Supplementary Table S1). Recall that the four datasets assayed pancreas cells from type 2 diabetes and healthy individuals, therefore, the pathway analysis once again confirms that BUSseq provides biologically and clinically valid cell typing.

Discussion

For the completely randomized experimental design, it seems that "everyone is talking, but no one is listening." Due to time and budget constraints, it is always difficult to implement a completely randomized design in practice. Consequently, researchers often pretend to be blind to the issue when carrying out their scRNA-seq experiments. In this paper, we mathematically prove and empirically show that under the more realistic reference panel and chain-type designs, batch effects can also be adjusted for scRNA-seq experiments. We hope that our results will alarm researchers of confounded experimental designs and encourage them to implement valid designs for scRNA-seq experiments in real applications.

BUSseq provides one-stop services. In contrast, most existing methods are multi-stage approaches—clustering can only be performed after the batch effects are corrected and the differential expressed genes can only be called after the cells are clustered. The major issue with multi-stage methods is that the uncertainties in the previous stages are often ignored. For instance, when cells are first clustered into different cell types and then differential gene expression identification is conducted for cells clustered into two groups, the clustering results are taken as if they were the underlying truth, even though they may be prone to errors, which can lead to false positives and false negatives. In contrast, BUSseq simultaneously corrects batch effects, clusters cell types, imputes missing data, and identifies intrinsic genes that differentiate cell types. Therefore, BUSseq automatically accounts for all the uncertainties together and fully exploits the information embedded in the data.

BUSseq is computationally efficient. For both our simulated data and real data with thousands of cells, the MCMC algorithm for BUSseq always converges within 5,000 iterations. The computational complexity of BUSseq is $O(\sum_{b=1}^{B} n_b GK)$, which is both linear in the number of batches B and in the number of cell type K. Moreover, most steps of the MCMC algorithm for BUSseq are parallelizable. Therefore, using graphics processing unit (GPU)

computing and cloud computing, we can expect that BUSseq will scale well, even with a larger number of cells.

Practical and valid experimental designs are urgently required for scRNA-seq experiments. We envision that the flexible reference panel and the chain-type designs will be widely adopted in scRNA-seq experiments and BUSseq will greatly facilitate the analysis of scRNA-seq data.

Methods

BUSseq model

The hierarchical model of BUSseq can be summarized as:

$$
\begin{aligned}
& Pr(W_{bi} = k) = \pi_{bk}, \sum_{k=1}^{K} \pi_{bk} = 1; \\
& X_{big} | W_{bi} = k \sim NB(\mu_{big}, \phi_{bg}), \quad log(\mu_{big}) = \alpha_g + \beta_{gk} + \nu_{bg} + \delta_{bi}; \\
& Z_{big} | X_{big} = x_{big} \sim Bernoulli(p_{big}), \quad log(\frac{p_{big}}{1 - p_{big}}) = \gamma_{b0} + \gamma_{b1} x_{big}; \\
& Y_{big} = X_{big} | Z_{big} = 0, \quad Y_{big} = 0 | Z_{big} = 1. \qquad (1)
\end{aligned}
$$

Collectively, $\boldsymbol{Y} = \{Y_{big}\}_{b=1,\cdots,B;i=1,\cdots,n_b}^{g=1,\cdots,G}$ are the observed data; the underlying true expression levels $\boldsymbol{X} = \{X_{big}\}_{b=1,\cdots,B;i=1,\cdots,n_b}^{g=1,\cdots,G}$, the dropout indicators $\boldsymbol{Z} = \{Z_{big}\}_{b=1,\cdots,B;i=1,\cdots,n_b}^{g=1,\cdots,G}$ and the cell type indicators $\boldsymbol{W} = \{W_{bi}\}_{b=1,\cdots,B;i=1,\cdots,n_b}$ are all missing data; the log-scale baseline gene expression levels $\boldsymbol{\alpha} = \{\alpha_g\}_{g=1,\cdots,G}$, the cell type effects $\boldsymbol{\beta} = \{\beta_{gk}\}_{k=2,\cdots,K}^{g=1,\cdots,G}$, the location batch effects $\boldsymbol{\nu} = \{\nu_{bg}\}_{b=2,\cdots,B}^{g=1,\cdots,G}$, the overdispersion parameters $\boldsymbol{\phi} = \{\phi_{bg}\}_{b=1,\cdots,B}^{g=1,\cdots,G}$, the cell-specific size factors $\boldsymbol{\Delta} = \{\delta_{bi}\}_{b=1,\cdots,B}^{i=2,\cdots,n_b}$, the dropout parameters $\boldsymbol{\Gamma} = \{\gamma_{b0}, \gamma_{b1}\}^{b=1,\cdots,B}$ and the cell compositions $\boldsymbol{\pi} = \{\pi_{bk}\}_{b=1,\cdots,B}^{k=1,\cdots,K}$ are the parameters. Without loss of generality, for model identifiability, we assume that the first batch is the reference batch measured without batch effects with $\nu_{1g} = 0$ for every gene and the first cell type is the baseline cell type with $\beta_{g1} = 0$ for every gene. Similarly, we take the cell-specific size factor $\delta_{b1} = 0$ for the first cell of each batch. We gather all the parameters as $\boldsymbol{\Theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\nu}, \boldsymbol{\phi}, \boldsymbol{\Delta}, \boldsymbol{\Gamma}, \boldsymbol{\pi}\}$. Let $f_{NB}(x; \mu, \phi) = C_x^{\phi+x-1} (\frac{\mu}{\mu+\phi})^x (\frac{\phi}{\mu+\phi})^\phi$ denote the probability mass function of the negative binomial distribution $NB(\mu, \phi)$, where C_k^n is the binomial coefficient, then the complete data

likelihood function equals to:

$$L_c(\boldsymbol{\Theta}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{z},\boldsymbol{w}) = \prod_{b=1}^{B}\prod_{i=1}^{n_b}\prod_{k=1}^{K}\{\pi_{bk}\prod_{g=1}^{G}[I(y_{big}=x_{big}(1-z_{big}))\frac{exp[(\gamma_{b0}+\gamma_{b1}x_{big})z_{big}]}{1+exp(\gamma_{b0}+\gamma_{b1}x_{big})}$$
$$\cdot f_{NB}(x_{big};exp(\alpha_g+\beta_{gk}+\nu_{bg}+\delta_{bi}),\phi_{bg})]\}^{I(w_{bi}=k)} \quad (2)$$

Consequently, the observed data likelihood function becomes

$$L_o(\boldsymbol{\Theta}|\boldsymbol{y}) = \prod_{b=1}^{B}\prod_{i=1}^{n_b}[\sum_{k=1}^{K}\pi_{bk}\prod_{g=1}^{G}Pr(Y_{big}=y_{big}|\boldsymbol{\Theta})], \quad (3)$$

$$Pr(Y_{big}=y_{big}|\boldsymbol{\Theta}) = \begin{cases} \sum_{k=1}^{K}\pi_{bk}[\sum_{x=1}^{\infty}\frac{exp(\gamma_{b0}+\gamma_{b1}x)}{1+exp(\gamma_{b0}+\gamma_{b1}x)}f_{NB}(x;exp(\alpha_g+\beta_{gk}+\nu_{bg}+\delta_{bi}),\phi_{bg}) & \\ +f_{NB}(0;exp(\alpha_g+\beta_{gk}+\nu_{bg}+\delta_{bi}),\phi_{bg})] & y_{big}=0, \\ \sum_{k=1}^{K}\pi_{bk}\frac{1}{1+exp(\gamma_{b0}+\gamma_{b1}y_{big})}f_{NB}(y_{big};exp(\alpha_g+\beta_{gk}+\nu_{bg}+\delta_{bi}),\phi_{bg}) & y_{big}>0. \end{cases}$$

Experimental designs

By creating a set of functions similar to the probability generating function, we prove that BUSseq is identifiable, in other words, if two parameters are different, then their probability distribution functions for the observed data are different, for not only the "complete setting" but also the "reference panel" and the "chain-type" designs (see the proofs in the Supplementary Materials).

Theorem 1. *(The Complete Setting)*
If $\pi_{bk}>0$ for every b and k, given that (I) $\gamma_{b1}<0$ for every b, (II) for any two cell types k_1 and k_2, there exist at least two differentially expressed genes g_1 and g_2—$\beta_{g_1k_1}\neq\beta_{g_1k_2}$ and $\beta_{g_2k_1}\neq\beta_{g_2k_2}$, and (III) for any two cell-type pairs $(k_1,k_2)\neq(k_3,k_4)$, their differences in cell-type effects are not the same $\boldsymbol{\beta}_{k_1}-\boldsymbol{\beta}_{k_2}\neq\boldsymbol{\beta}_{k_3}-\boldsymbol{\beta}_{k_4}$, then BUSseq is identifiable (up to label switching) in the sense that $L_o(\boldsymbol{\Theta}|\boldsymbol{y})=L_o(\boldsymbol{\Theta}^*|\boldsymbol{y})$ for any $\boldsymbol{y}$ implies that $\pi_{bk}=\pi^*_{b\rho(k)}, (\gamma_{b0},\gamma_{b1})=(\gamma^*_{b0},\gamma^*_{b1}), \alpha_g+\beta_{gk}=\alpha^*_g+\beta^*_{g\rho(k)}, \nu_{gb}=\nu^*_{gb}, \delta_{bi}=\delta^*_{bi}$ and $\phi_{bg}=\phi^*_{bg}$ for every gene g and batch b, where ρ is a permutation of $\{1,2,\cdots,K\}$.

In the following, we denote the cell types that are present in batch b as C_b and count the number of cell types existing in batch b as $K_b=|C_b|$.

Theorem 2. *(The Reference Panel Design)*
If there are a total of K cell types $\cup_{b=1}^{B} C_b = \{1, 2, \cdots, K\}$, $K_b \geq 2$ for every batch b, and there exist a batch $\tilde{b}$ such that it contains all of the cell types $C_{\tilde{b}} = \{1, 2, \cdots, K\}$, then given that conditions (I)-(III) hold, BUSseq is identifiable (up to label switching).

Theorem 3. *(The Chain-type Design)*
If there are a total of K cell types $\cup_{b=1}^{B} C_b = \{1, 2, \cdots, K\}$ and every two consecutive batches share at least two cell types $|C_b \cap C_{b-1}| \geq 2$ for all $b \geq 2$, then given that conditions (I)-(III) hold, BUSseq is identifiable (up to label switching).

Therefore, even for the "reference panel" and "chain-type" designs that do not assay all cell types in each batch, batch effects can be removed; cell types can be clustered; and missing data due to dropout events can be imputed.

Statistical inference

We conduct the statistical inference under the Bayesian framework. We assign independent priors to each component of $\boldsymbol{\Theta}$ as follows: $\boldsymbol{\pi}_b = (\pi_{b1}, \cdots, \pi_{bK}) \sim Dirichlet(\xi, \cdots, \xi), 1 \leq b \leq B; \gamma_{b0} \sim N(0, \sigma_{z0}^2), 1 \leq b \leq B; \gamma_{b1} \sim N(0, \sigma_{z1}^2), 1 \leq b \leq B; \alpha_g \sim N(m_a, \sigma_a^2), 1 \leq g \leq G; \nu_{bg} \sim N(m_c, \sigma_c^2), 2 \leq b \leq B, g = 1, \cdots, G; \delta_{bi} \sim N(m_d, \sigma_d^2), 1 \leq b \leq B, 2 \leq i \leq n_b; \phi_{bg} \sim Gamma(\kappa, \tau), 1 \leq b \leq B, 1 \leq g \leq G$. We are interested in detecting genes that differentiate cell types. Therefore, we impose a spike-and-slab prior [33] using a normal mixture to the cell-type effect β_{gk}. The spike component concentrates on zero with a small variance $\tau_{\beta 0}^2$, whereas the slab component tends to deviate from zero, thus having a larger variance $\tau_{\beta 1}^2$. We introduce another latent variable L_{gk} to indicate which component β_{gk} comes from. $L_{gk} = 0$ if gene g is not differentially expressed between cell type k and cell type one, and $L_{gk} = 1$, otherwise. We further define $D_g = \sum_{k=2}^{K} L_{gk}$. If $D_g > 0$, then the expression level of gene g does not stay the same across cell types. Following Huo et al. [10], we call such genes intrinsic genes, which are able to differentiate cell types. To control for multiple hypothesis testing, we let $L_{gk} \sim Bernoulli(p)$ and assign a conjugate prior $Beta(a_p, b_p)$ to p. We set $\tau_{\beta 1}$ to a large number and let $\tau_{\beta 0}$ follow an inverse-gamma prior $Inv\Gamma(a_\tau, b_\tau)$ with a small prior mean.

We develop an MCMC algorithm to sample from the posterior distribution (Supplementary Materials). After the burn-in period, we take the mean of the posterior samples to estimate

$\boldsymbol{\gamma}_b, \alpha_g, \beta_{gk}, \nu_{bg}, \delta_{bi}$ and ϕ_{bg} and use the mode of posterior samples of W_{bi} to infer the cell type for each cell.

When inferring the differential expression indicator L_{gk}, we control the Bayesian false discovery rate (FDR) [25, 26] defined as

$$FDR(\kappa) = \frac{\sum_{g=1}^{G}\sum_{k=2}^{K}\xi_{gk}I(\xi_{gk} \leq \kappa)}{\sum_{g=1}^{G}\sum_{k=2}^{K}I(\xi_{gk} \leq \kappa)}, \tag{4}$$

where $\xi_{gk} = Pr(L_{gk} = 0|\boldsymbol{y})$ is the posterior marginal probability that gene g is not differentially expressed between cell type k and cell type one, which can be estimated by the T posterior samples $L_{gk}^{(t)}$s collected after the burn-in period as $\frac{1}{T}\sum_{t=1}^{T}(1 - L_{gk}^{(t)})$. Given a control level α such as 0.1, we search for the largest $\kappa_0 \leq 0.5$ such that the estimated $\widehat{FDR}(\kappa)$ based on $\widehat{\xi}_{gk}$s is smaller than α and declare $\widehat{L}_{gk} = 1$ if $\widehat{\xi}_{gk} \leq \kappa_0$. The upper bound 0.5 for κ_0 [26] prevents us from calling differentially expressed genes with small posterior probability $Pr(L_{gk} = 1|\boldsymbol{y})$. Consequently, we identify the genes with $\widehat{D}_g = \sum_{k=2}^{K}\widehat{L}_{gk} > 0$ as the intrinsic genes.

BUSseq allows the user to input the total number of cell types K according to prior knowledge. When K is unknown, BUSseq selects the number of cell types $\widehat{K}$ such that it achieves the minimum BIC.

$$BIC(K) = -2L_o(\widehat{\boldsymbol{\Theta}}|\boldsymbol{y}) + [K(B+G) + 2B + (2B-1)G + \sum_{b=1}^{B}(n_b - 1)] \cdot log(\sum_{b=1}^{B} n_b G), \tag{5}$$

Batch-effects-corrected values

To facilitate further downstream analysis, we also provide a version of count data $\tilde{\boldsymbol{X}} = \{\tilde{X}_{big}\}_{b=1,\cdots,B;i=1,\cdots,n_b}^{g=1,\cdots,G}$ for which the batch effects are removed and the biological variability is retained. We develop a quantile matching approach based on inverse sampling. Specifically, given the fitted model and the inferred true underlying expression level $\widehat{x}_{big}$, we first sample u_{big} from $Unif[F_{NB}(\widehat{x}_{big}-1; exp(\widehat{\alpha}_g+\widehat{\beta}_{g\widehat{w}_{bi}}+\widehat{\nu}_{bg}+\widehat{\delta}_{bi}), \widehat{\phi}_{bg}), F_{NB}(\widehat{x}_{big}; exp(\widehat{\alpha}_g+\widehat{\beta}_{g\widehat{w}_{bi}}+\widehat{\nu}_{bg}+\widehat{\delta}_{bi}), \widehat{\phi}_{bg})]$ where $F_{NB}(\cdot;\mu,r)$ denotes the cumulative distribution function of a negative binomial distribution with mean μ and overdispersion parameter r. Next, we calculate the u_{big}^{th} quantile of $NB(exp(\widehat{\alpha}_g + \widehat{\beta}_{g\widehat{w}_{bi}}), \widehat{\phi}_{1g})$ as the corrected value $\tilde{x}_{big}$.

The corrected data $\tilde{\boldsymbol{X}}$ are not only protected from batch effects but also impute the

missing data due to dropout events. Moreover, further cell-specific normalization is not needed. Meanwhile, the biological variability is retained thanks to the quantile transformation and sampling step. Therefore, we can directly perform downstream analysis on $\tilde{\boldsymbol{X}}$.

The benchmarked methods

MNN [11] takes the first batch as the reference batch and normalizes the other batches to adjust for difference in sequencing depths. It then log-transforms the normalized data and performs batch effects correction. For clustering cell types, MNN constructs a shared nearest neighbor (SNN) graph and applies the "Walktrap" algorithm to the SNN graph. ZINB-WaVE [19] directly works on the raw read count data and uses a sequential K-means clustering approach implemented in the clusterSingle function in the clusterExperiment package. Seurat [14] first log-transforms and scales the observed read count data. Then, it performs CCA and conducts clustering on the subspace spanned by the first few canonical variables with the FindClusters function in the Seurat package. Following Butler et al. [14], we select the number of canonical variables by looking for a saturation in the relationship between the number of canonical variables and the shared correlation strength. The correlation strength is calculated by the MetageneBicorPlot function in the Seurat package.

Processing of the real datasets

For the two hematopoietic datasets, we downloaded the read count matrix of the 1,920 cells profiled by Paul et al. [27] and the 2,729 cells labeled as myeloid progenitor cells by Nestorowa et al. [34] from the NCBI Gene Expression Omnibus (GEO) with the accession numbers GSE72857 and GSE81682. Following Brennecke et al. [35], we sorted the genes according to their adjusted variance-mean ratio of expression levels in both datasets separately and focused on the 3,470 genes that are highly variable in both datasets.

Two of the pancreas datasets profiled by the CEL-seq2 platform were downloaded from GEO with accession number GSE80176 [29] and GSE86473 [28]. The two datasets assayed by the SMART-seq2 platform were obtained from GSE85241 [36] and from ArrayExpress accession number E-MATB-5061 [30]. Following Haghverdi et al. [11], we excluded cells with low library sizes ($< 100,000$ reads), low numbers of expressed genes ($> 40\%$ total counts

from ribosomal RNA genes), or high ERCC content (> 20% of total counts from spike-in transcripts) resulting in 7,095 cells. We selected the 2,480 highly variable genes shared by the four datasets according to Brennecke et al. [35] by sorting the ratio of variance and mean expression level after adjusting technical noise with the variances of spike-in transcripts. The cell types of the two datasets profiled by the CEL-seq2 platform were labelled according to Lawlor et al. [28] and Grün et al. [29], with the GCG gene marking alpha islets, INS for beta islets, SST for delta islets, PPY for gamma islets, PRSS1 for acinar cells, and KRT19 for ductal cells. The cell types of the other two datasets assayed by the SMART-seq2 platform were provided in their metadata.

Software availability

The R package BUSseq is available on Github (`https://github.com/songfd2018/BUSseq`). All codes for producing results and figures in this manuscript are also available on Github (`https://github.com/songfd2018/BUSseq_implementation`).

Data availability

The published data sets used in this manuscript are available through the following accession numbers: SMART-seq2 platform hematopoietic data with GEO GSE81682 by Nestorowa et al. [34]; MARS-seq platform hematopoietic data with GEO GSE72857 by Paul et al. [27]; CEL-seq platform pancreas data with GEO GSE81076 by Grün et al. [29]; CEL-seq2 platform pancreas data with GEO GSE85241 by Muraro et al. [36]; SMART-seq2 platform pancreas data with GEO GSE86473 by Lawlor et al. [28]; and SMART-seq2 platform pancreas data with ArrayExpress E-MTAB-5061 by Segerstolpe et al. [30].

Acknowledgment

The authors thank Dr. Xiangyu Luo for helpful comments on an early version of our paper.

Author contributions

FD.S developed the method and the proof, implemented the algorithm, prepared the software package, analyzed the data, and wrote the paper. GM.C. conducted data preprocessing. YY.W. conceived and supervised the study, developed the method and the proof, and wrote the paper.

References

[1] Rhonda Bacher and Christina Kendziorski. Design and computational analysis of single-cell rna-sequencing experiments. *Genome Biology*, 17(1):63, 2016.

[2] Rafael A Irizarry, Daniel Warren, Forrest Spencer, Irene F Kim, Shyam Biswal, et al. Multiple-laboratory comparison of microarray platforms. *Nature Methods*, 2(5):345–350, 2005.

[3] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

[4] Margaret A Taub, H Corrada Bravo, and Rafael A Irizarry. Overcoming bias and systematic errors in next generation sequencing data. *Genome Medicine*, 2(12):87, 2010.

[5] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics*, 2017.

[6] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740, 2014.

[7] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[8] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

[9] Jeffrey T Leek. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Research*, page gku864, 2014.

[10] Zhiguang Huo, Ying Ding, Silvia Liu, Steffi Oesterreich, and George Tseng. Meta-analytic framework for sparse k-means to identify disease subtypes in multiple transcriptomic studies. *Journal of the American Statistical Association*, 111(513):27–42, 2016.

[11] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects

in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421, 2018.

[12] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32 (9):896, 2014.

[13] Laurent Jacob, Johann A Gagnon-Bartsch, and Terence P Speed. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics*, 17(1):16–28, 2015.

[14] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411, 2018.

[15] Xiangyu Luo and Yingying Wei. Batch effects correction with unknown subtypes. *Journal of the American Statistical Association*, pages 1–14, 2018.

[16] Catalina A Vallejos, John C Marioni, and Sylvia Richardson. Basics: Bayesian analysis of single-cell sequencing data. *PLoS Computational Biology*, 11(6):e1004333, 2015.

[17] J. Wang, M. Huang, E Torre, H Dueck, S Shaffer, J Murray, A Raj, M. Li, and N. R. Zhang. Gene expression distribution deconvolution in single-cell rna sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 115 (28):E6437–E6446, 2018.

[18] Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16(1):241, 2015.

[19] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature Communications*, 9(1):284, 2018.

[20] J Baran-Gale, T Chandra, and K Kirschner. Experimental design for single-cell rna sequencing. *Briefings in Functional Genomics*, 17(4), 2017.

[21] Molin A Dal and Camillo B Di. How to design a single-cell rna-sequencing experiment: pitfalls, challenges and perspectives. *Briefings in Bioinformatics*, (1), 2018.

[22] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[23] Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6 (2):461–464, 1978.

[24] Wang Miao, Peng Ding, and Zhi Geng. Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, 111(516):1673–1683, 2016.

[25] Michael A Newton, Amine Noueiry, Deepayan Sarkar, and Paul Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176, 2004.

[26] Christine Peterson, Francesco C Stingo, and Marina Vannucci. Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association*, 110 (509):159–174, 2015.

[27] Franziska Paul, Yaara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, David Lara-Astiaso, Meital Gury, Assaf Weiner, et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163(7):1663–1677, 2015.

[28] Nathan Lawlor, Joshy George, Mohan Bolisetty, Romy Kursawe, Lili Sun, V Sivakamasundari, Ina Kycia, Paul Robson, and Michael L Stitzel. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. *Genome Research*, 27(2):208–222, 2017.

[29] Dominic Grün, Mauro J Muraro, Jean-Charles Boisset, Kay Wiebrands, Anna Lyubimova, Gitanjali Dharmadhikari, Maaike van den Born, Johan Van Es, Erik Jansen, Hans Clevers, et al. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell*, 19(2):266–277, 2016.

[30] Åsa Segerstolpe, Athanasia Palasantza, Pernilla Eliasson, Eva-Marie Andersson, Anne-Christine Andréasson, Xiaoyan Sun, Simone Picelli, Alan Sabirsh, Maryam Clausen, Magnus K Bjursell, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metabolism*, 24(4):593–607, 2016.

[31] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102 (43):15545–15550, 2005.

[32] Minoru Kanehisa and Susumu Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[33] Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[34] Sonia Nestorowa, Fiona K Hamey, Blanca Pijuan Sala, Evangelia Diamanti, Mairi Shepherd, Elisa Laurenti, Nicola K Wilson, David G Kent, and Berthold Göttgens. A single cell resolution map of mouse haematopoietic stem and progenitor cell differentiation. *Blood*, pages blood–2016, 2016.

[35] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei

Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, et al. Accounting for technical noise in single-cell rna-seq experiments. *Nature Methods*, 10(11):1093, 2013.

[36] Mauro J Muraro, Gitanjali Dharmadhikari, Dominic Grün, Nathalie Groen, Tim Dielen, Erik Jansen, Leon van Gurp, Marten A Engelse, Francoise Carlotti, Eelco JP de Koning, et al. A single-cell transcriptome atlas of the human pancreas. *Cell Systems*, 3(4): 385–394, 2016.

[37] Henry Teicher. Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics*, 38(4):1300–1302, 1967.

[38] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[39] Gautam Altekar, Sandhya Dwarkadas, John P Huelsenbeck, and Fredrik Ronquist. Parallel metropolis coupled markov chain monte carlo for bayesian phylogenetic inference. *Bioinformatics*, 20(3):407–415, 2004.

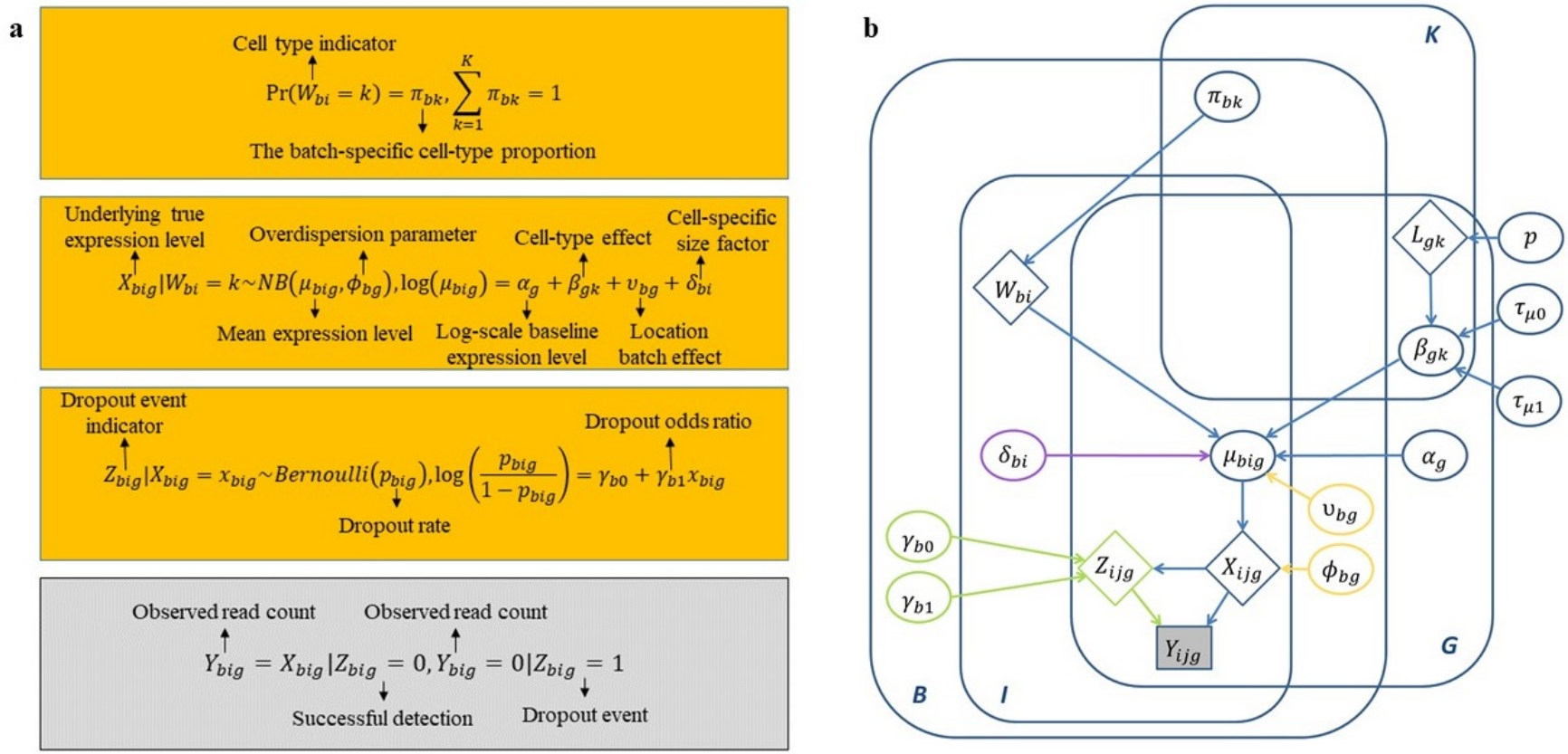

Figure 1: The graphical representation of the BUSseq model. **(a)** The hierarchical structure of BUSseq model. Only Y_{big} in the grey rectangle is observed **(b)**The yellow color corresponds to batch effects; the green color models the dropout events; the purple color indicates the cell-specific size factor. The ellipses are for parameters; the diamonds represent latent variables; and only Y_{big} in the grey rectangle is observed.

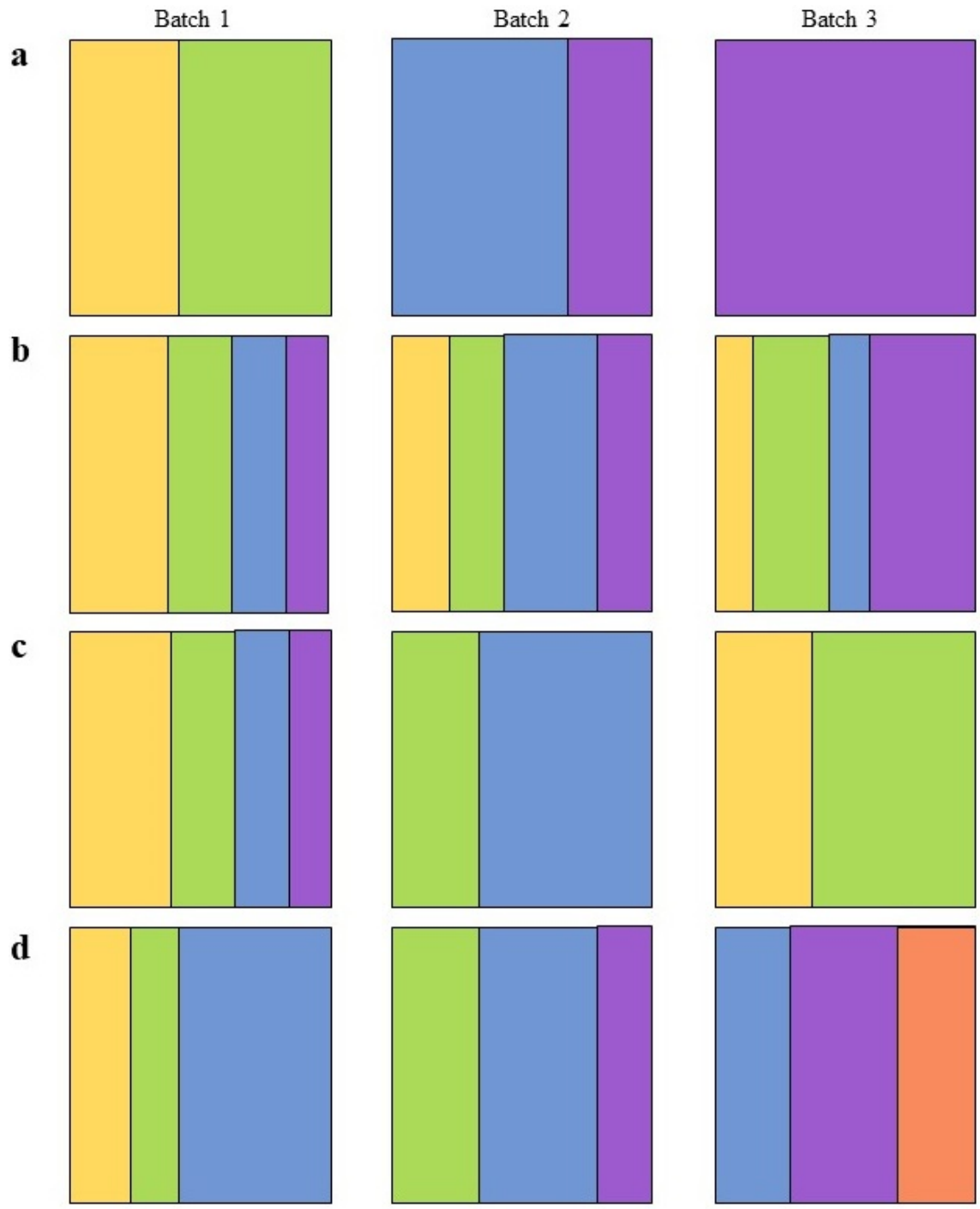

Figure 2: Various types of experimental designs. Each subpanel represents a batch with genes in rows and cells in columns; and each color indicates a cell type. **(a)** A confounded design. Batch 1 assays cells from cell types 1 and 2; batch 2 profiles cells from cell types 3 and 4; and batch 3 only contains cells from cell type 4. **(b)** The complete setting design. Each batch assays cells from all of the four cell types, although the cellular compositions vary across batches. **(c)** The reference panel design. Batch 1 contains cells from all of the cell types, and all the other batches have at least two cell types. **(d)** The chain-type design. Every two consecutive batches share two cell types. Batch 1 and Batch 2 share cell types 2 and 3; Batch 2 and Batch 3 share cell types 3 and 4.

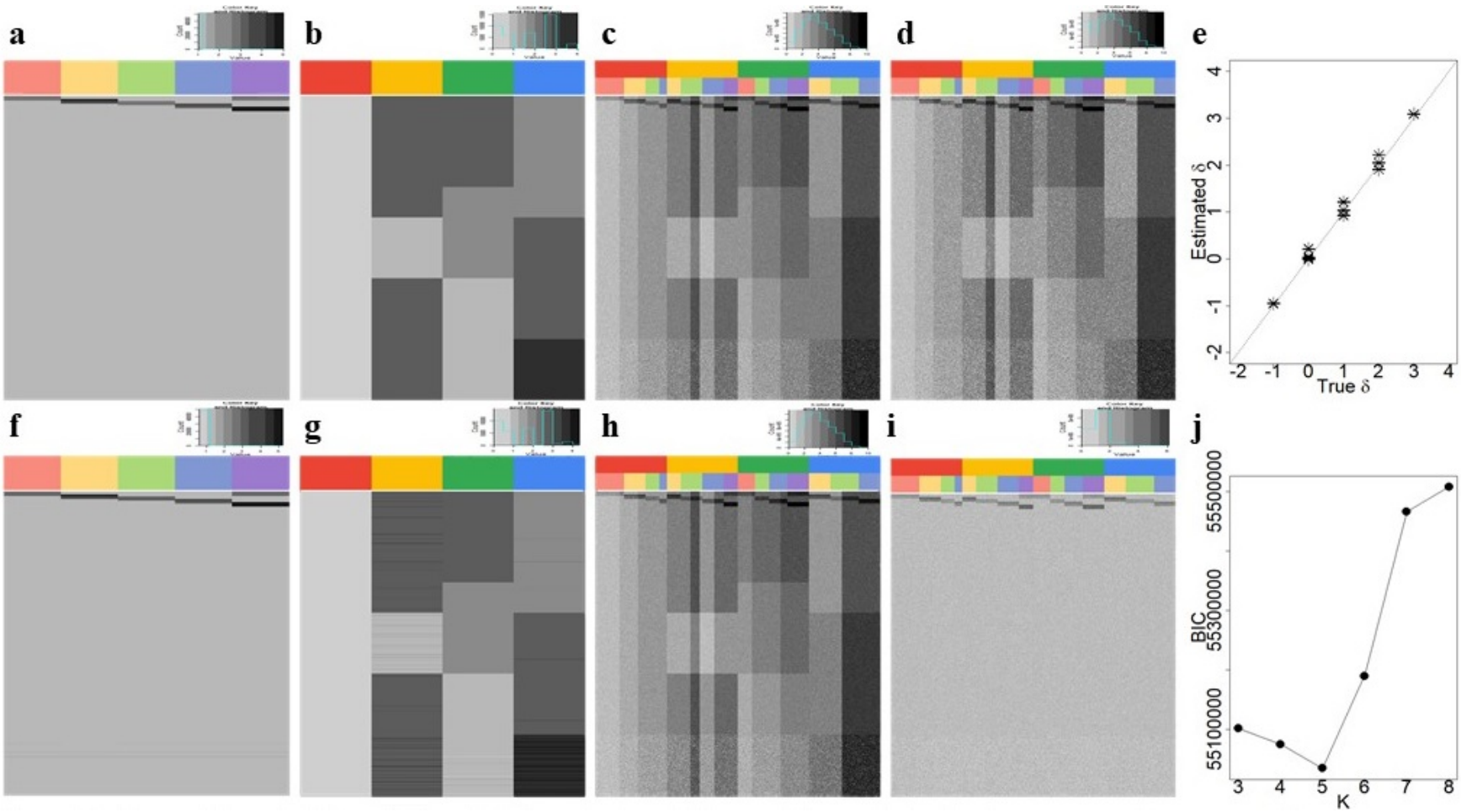

Figure 3: Patterns of the simulation for the chain-type design. **(a)** True cell type effects. Each row represents a gene, and each column corresponds to a cell type. **(b)** True batch effects. Each row represents a gene, and each column corresponds to a batch. **(c)** True underlying expression levels $\boldsymbol{X}$. Each row represents a gene, and each column corresponds to a cell. The upper color bar indicates the batches, and the lower color bar represents the cell types. There are a total of 10,000 genes. The sample size for each batch is 150. **(d)** The simulated observed data $\boldsymbol{Y}$. The dropout rate for each cell varies from 2.7% to 17.8%. **(e)** Scatter plot of the estimated cell-specific size factor versus the true cell-specific size factor. **(f)** Estimated cell type effects. **(g)** Estimated batch effects. **(h)** Imputed expression levels $\widehat{\boldsymbol{X}}$. **(i)** Corrected count data $\tilde{\boldsymbol{X}}$ grouped by batches. **(j)** The BIC plot.

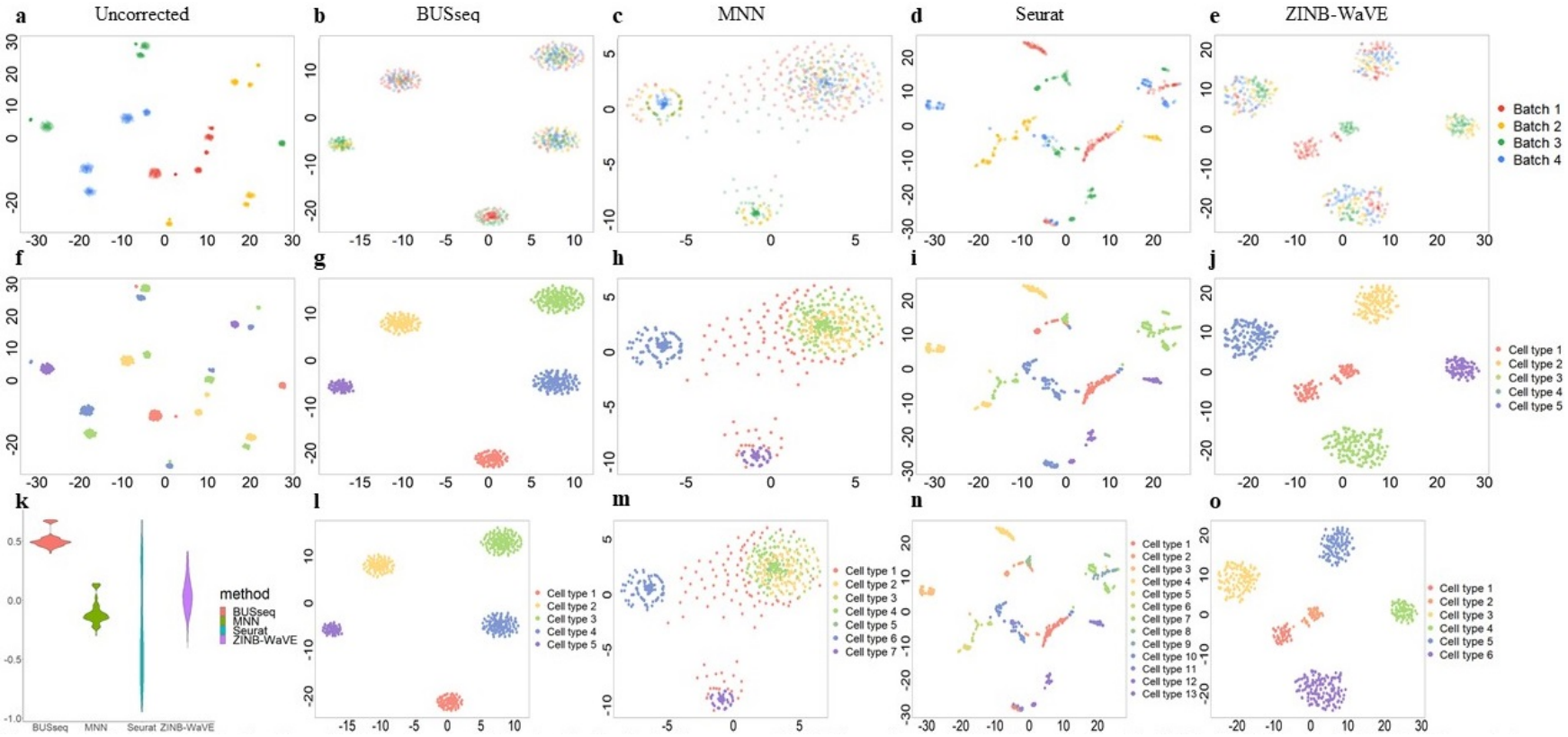

Figure 4: T-distributed Stochastic Neighbor Embedding (t-SNE) plots for the simulated data. **(a)** Data before correction colored by batch. **(b)** Data corrected by BUSseq colored by batch. **(c)** Data corrected by MNN colored by batch. **(d)** The first 15 canonical correlations (CC) from Seurat colored by batch. **(e)** The 10-component W from ZINB-WaVE colored by batch. **(f)** Data before correction colored by true cell type. **(g)** Data corrected by BUSseq colored by true cell type. **(h)** Data corrected by MNN colored by true cell type. **(i)** The first 15 CCs from Seurat colored by true cell type. **(j)** The 10-component W from ZINB-WaVE colored by true cell type. **(k)** Silhouette coefficients for BUSseq (according to the corrected read counts of identified intrinsic genes), MNN (according to the whole corrected read count matrix by MNN), ZINB-WaVE (according to the low-dimensional characteristics inferred by ZINB-WaVE) and Seurat. **(l)** Data corrected by BUSseq colored by the five inferred cell types from BUSseq. **(m)** Data corrected by MNN colored by the 7 inferred cell types from SNN and walktrap algorithm. **(n)** The first 15 CCs from Seurat by the 13 labeled cell types by the FindClusters function in the Seurat. **(o)** The 10-component W from ZINB-WaVE colored by the 6 labeled cell types from k-means clustering in the clusterSingle function in the clusterExperiment R package.

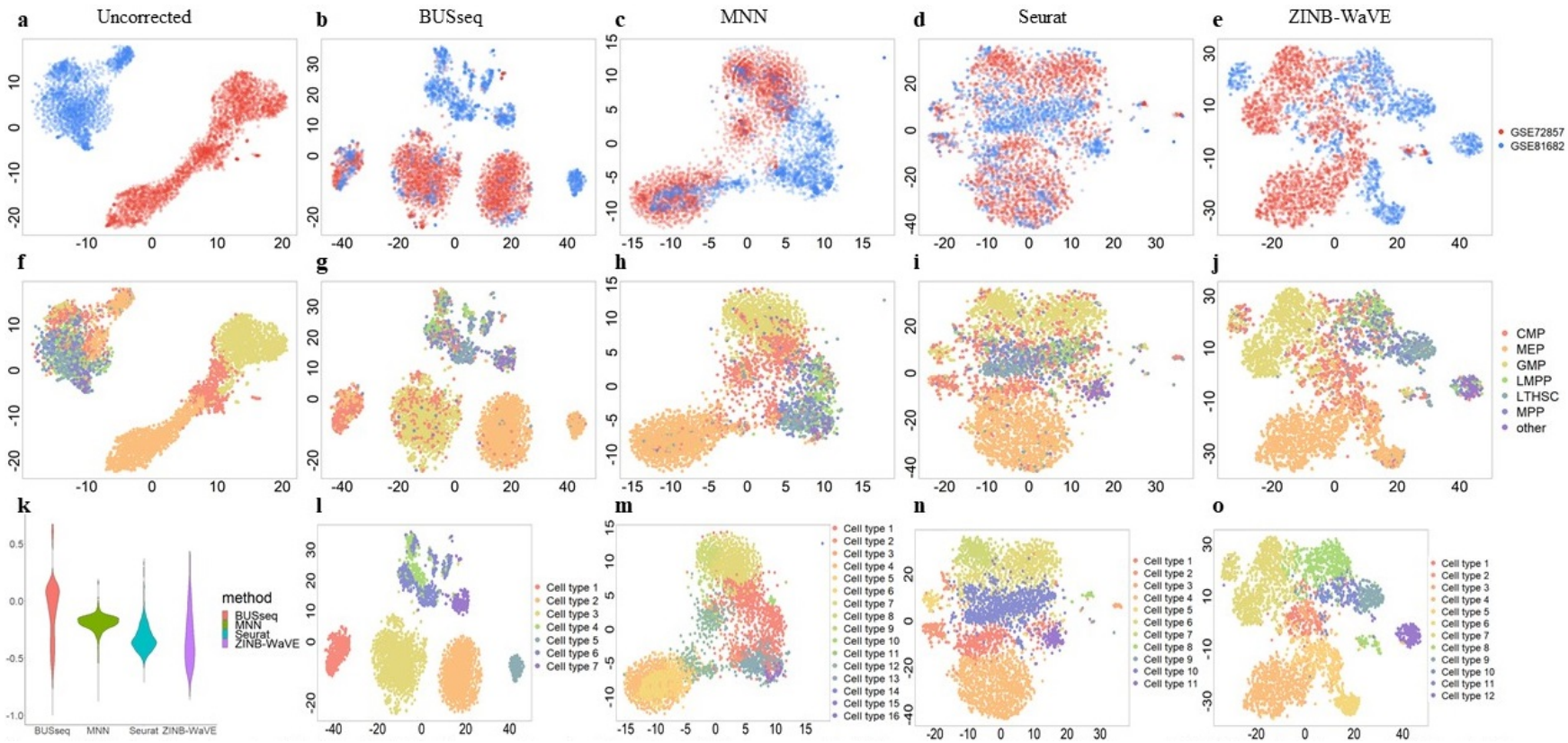

Figure 5: T-distributed Stochastic Neighbor Embedding (t-SNE) plots for the hematopoietic data. **(a)** Data before correction colored by dataset. **(b)** Data corrected by BUSseq colored by dataset. **(c)** Data corrected by MNN colored by dataset. **(d)** The first 25 canonical correlations (CC) from Seurat colored by dataset. **(e)** The 10-component W from ZINB-WaVE colored by dataset. **(f)** Data before correction colored by the 7 labeled cell types from FACS. **(g)** Data corrected by BUSseq colored by the 7 labeled cell types. **(h)** Data corrected by MNN colored by the 7 labeled cell types. **(i)** The first 25 CCs from Seurat colored by the 7 labeled cell types. **(j)** The 10-component W from ZINB-WaVE colored by the 7 labeled cell types. **(k)** Silhouette coefficients for BUSseq (according to the corrected read counts of identified intrinsic genes), MNN (according to the whole corrected read count matrix by MNN), ZINB-WaVE (according to the low-dimensional characteristics inferred by ZINB-WaVE) and Seurat. **(l)** Data corrected by BUSseq colored by the 7 inferred cell types from BUSseq. **(m)** Data corrected by MNN colored by the 16 inferred cell types from walktrap algorithm. **(n)** The first 25 CCs from Seurat by the 11 inferred cell types by the FindClusters function in the Seurat. **(o)** The 10-component W from ZINB-WaVE colored by the 12 inferred cell types from k-means clustering in the clusterSingle function of the clusterExperiment R package.

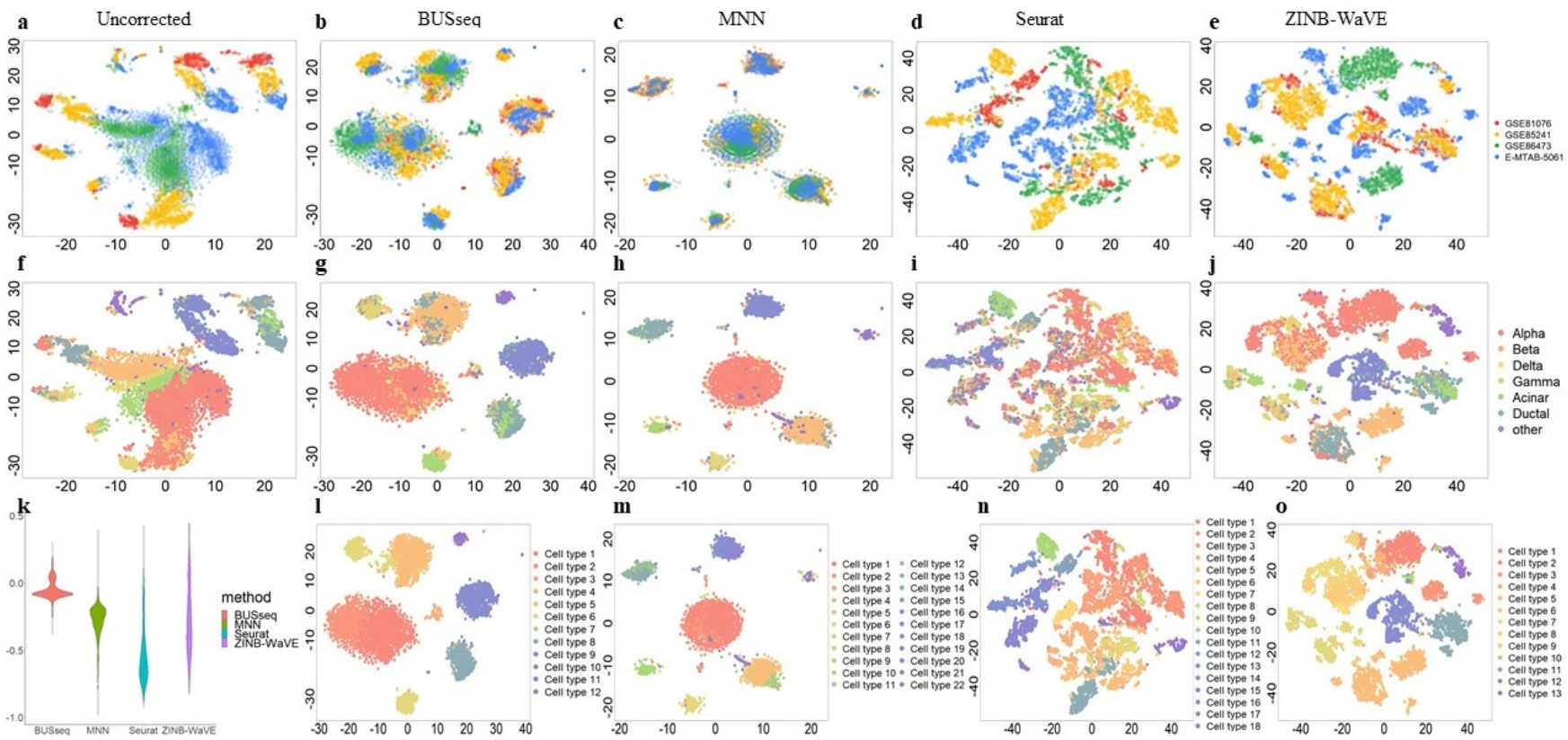

Figure 6: T-distributed Stochastic Neighbor Embedding (t-SNE) plots for the pancreas data. **(a)** Data before correction colored by dataset. **(b)** Data corrected by BUSseq colored by dataset. **(c)** Data corrected by MNN colored by dataset. **(d)** The first 25 canonical correlations (CC) from Seurat colored by dataset. **(e)** The 10-component W from ZINB-WaVE colored by dataset. **(f)** Data before correction colored by the 7 labeled cell types from marker genes. **(g)** Data corrected by BUSseq colored by the 7 labeled cell types. **(h)** Data corrected by MNN colored by the 7 labeled cell types. **(i)** The first 12 CCs from Seurat colored by the 7 labeled cell types. **(j)** The 10-component W from ZINB-WaVE colored by the 7 labeled cell types. **(k)** Silhouette coefficients for BUSseq (according to the corrected read counts of identified intrinsic genes), MNN (according to the whole corrected read count matrix by MNN), ZINB-WaVE (according to the low-dimensional characteristics inferred by ZINB-WaVE) and Seurat. **(l)** Data corrected by BUSseq colored by the 12 inferred cell types from BUSseq. **(m)** Data corrected by MNN colored by the 22 inferred cell types from Walktrap algorithm. **(n)** The first 12 CCs from Seurat by the 18 inferred cell types by the FindClusters function in the Seurat. **(o)** The 10-component W from ZINB-WaVE colored by the 13 inferred cell types from k-means clustering in the clusterSingle function of the clusterExperiment R package.

Supplementary Materials

Proofs for Theorem 1 to 3

Lemma 1. *Let $\mathcal{F}^G$ be the family of $G(\geq 2)$-dimensional multivariate distribution with the probability mass function for $\boldsymbol{y} = (y_1, \cdots, y_G)$ as*

$$f^G(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\mu}) = \prod_{g=1}^{G} \{[\frac{1}{1 + exp(\gamma_0 + \gamma_1 y_g)} f_{NB}(y_g; \mu_g, \phi_g)]^{1(y_g > 0)} \cdot [\sum_{x=1}^{\infty} \frac{exp(\gamma_0 + \gamma_1 x)}{1 + exp(\gamma_0 + \gamma_1 x)} f_{NB}(x; \mu_g, \phi_g) + f_{NB}(0; \mu_g, \phi_g)]^{1(y_g = 0)}\} \tag{6}$$

such that $\gamma_1 < 0$ and for any two distinct elements $f^G_{k_1} = f^G(\boldsymbol{y}|\boldsymbol{\gamma}_{k_1}, \boldsymbol{\phi}_{k_1}, \boldsymbol{\mu}_{k_1}) \in \mathcal{F}^G, f^G_{k_2} = f^G(\boldsymbol{y}|\boldsymbol{\gamma}_{k_2}, \boldsymbol{\phi}_{k_2}, \boldsymbol{\mu}_{k_2}) \in \mathcal{F}^G$, there exist at least two dimensions g_1 and g_2 with $\mu_{g_1 k_1} \neq \mu_{g_1 k_2}$ and $\mu_{g_2 k_1} \neq \mu_{g_2 k_2}$, then the class of all finite mixtures of $\mathcal{F}^G$ is identifiable (up to label switching).

Proof. We reparameterize (μ_g, ϕ_g) as (p_g, ϕ_g) such that $p_g = \frac{\mu_g}{\mu_g + \phi_g}$ for all $g = 1, 2, \cdots, G$. Consequently, the identifiability with respect to $(\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\mu})$ is equivalent to that with respect to $(\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{p})$. With a little bit abuse of notations, we still use $f^G(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{p})$ to indicate the probability mass function of the new parameterization hereafter. Suppose that the finite mixture of $\mathcal{F}^G$ is not identifiable, then we have two different representations of the probability mass function $h(\boldsymbol{y})$ of the same finite mixtures:

$$h(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k f^G(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{p}_k) = \sum_{l=1}^{L} \xi_l f^G(\boldsymbol{y}|\boldsymbol{\delta}, \boldsymbol{\psi}, \boldsymbol{r}_l). \tag{7}$$

where the tuples $(\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{p}_k)$ for $k = 1, 2, \cdots, K$ are mutually distinct, and so are the tuples $(\boldsymbol{\delta}, \boldsymbol{\psi}, \boldsymbol{r}_l)$ for $l = 1, 2, \cdots, L$.

We define a total ordering $(\succeq)$ of $\mathcal{F}^G$. For $f^G_1, f^G_2 \in \mathcal{F}^G$, $f^G_1 \succeq f^G_2$ if:

(i) there exists a $g \geq 1$ such that for all $j < g$, $p_{j1} = p_{j2}$ and $\phi_{j1} = \phi_{j2}$ but $p_{g1} > p_{g2}$;

(ii) or there exists a g such that for all $j < g$, $p_{j1} = p_{j2}$ and $\phi_{j1} = \phi_{j2}$ as well as $p_{g1} = p_{g2}$ but $\phi_{g1} > \phi_{g2}$;

(iii) or $\boldsymbol{p}_1 = \boldsymbol{p}_2$ and $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2$ but $\gamma_{11} < \gamma_{21}$;

(iv) or $\boldsymbol{p}_1 = \boldsymbol{p}_2, \boldsymbol{\phi}_1 = \boldsymbol{\phi}_2$ and $\gamma_{11} = \gamma_{21}$ but $\gamma_{10} \leq \gamma_{20}$.

Without loss of generality, we assume that $f^G(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{p}_1) \succeq f^G(\boldsymbol{y}|\boldsymbol{\delta}, \boldsymbol{\psi}, \boldsymbol{r}_1)$ and the mixture

components on both sides of (7) are ordered:

$$f^G(\boldsymbol{y}|\boldsymbol{\gamma},\boldsymbol{\phi},\boldsymbol{p}_1) \succeq f^G(\boldsymbol{y}|\boldsymbol{\gamma},\boldsymbol{\phi},\boldsymbol{p}_2) \succeq \cdots \succeq f^G(\boldsymbol{y}|\boldsymbol{\gamma},\boldsymbol{\phi},\boldsymbol{p}_K),$$
$$f^G(\boldsymbol{y}|\boldsymbol{\delta},\boldsymbol{\psi},\boldsymbol{r}_1) \succeq f^G(\boldsymbol{y}|\boldsymbol{\delta},\boldsymbol{\psi},\boldsymbol{r}_2) \succeq \cdots \succeq f^G(\boldsymbol{y}|\boldsymbol{\delta},\boldsymbol{\psi},\boldsymbol{r}_L).$$

For $k=1$, we use mathematical induction to prove that for every $G_0 \in \{1,2,\cdots,G\}$,

$$r_{j1} = p_{j1}, \phi_j = \psi_j, \forall j \in \{1,2,\cdots,G_0\}, \tag{*}$$

and there exist a K_{G_0} and an L_{G_0} such that

$$\sum_{k=1}^{K_{G_0}} \pi_k f^{G-G_0}(\boldsymbol{y}_{-G_0}|\boldsymbol{\gamma},\boldsymbol{\phi}_{-G_0},\boldsymbol{p}_{-G_0,k}) = \sum_{l=1}^{L_{G_0}} \xi_l f^{G-G_0}(\boldsymbol{y}_{-G_0}|\boldsymbol{\delta},\boldsymbol{\psi}_{-G_0},\boldsymbol{r}_{-G_0,l}), \tag{**}$$

where the subscript $-G_0$ denotes that the first G_0 entries in the original vectors are excluded. Specifically, $\boldsymbol{y}_{-G_0} = (y_{G_0+1}, y_{G_0+2}, \cdots, y_G)^T$.

We first prove Equations (*) and (**) hold for $G_0 = 1$. We define a linear mapping that maps a probability distribution of $\mathcal{F}^G$ to a function that shares a similar spirit as a probability generating function $M_1 : f^G(\boldsymbol{y}) \in \mathcal{F}^G \to \Phi_1(t_1, \boldsymbol{y}_{-1}) \in \mathcal{G}_1$ such that $M_1(f^G(\boldsymbol{y})) = \Phi_1(t_1,\boldsymbol{y}_{-1}) = \sum_{y_1=1}^{\infty} f^G(\boldsymbol{y}|\boldsymbol{\gamma},\boldsymbol{\phi},\boldsymbol{p})t_1^{y_1} = \sum_{y_1=1}^{\infty} f^1(y_1|\boldsymbol{\gamma},\phi_1,p_1)t_1^{y_1} \cdot f^{G-1}(\boldsymbol{y}_{-1}|\boldsymbol{\gamma},\boldsymbol{\phi}_{-1},\boldsymbol{p}_{-1})$. Notice that $\Phi_1(t_1,\boldsymbol{y}_{-1})$ does not include the term of $y_1 = 0$, that is, $f^1(0|\boldsymbol{\gamma},\phi_1,p_1)t_1^0 \cdot f^{G-1}(\boldsymbol{y}_{-1}|\boldsymbol{\gamma},\boldsymbol{\phi}_{-1},\boldsymbol{p}_{-1})$. Specifically, we denote $\Phi_{1k}(t_1,\boldsymbol{y}_{-1}) = M_1(f^G(\boldsymbol{y}|\boldsymbol{\gamma},\boldsymbol{\phi},\boldsymbol{p}_k))$ and $\Psi_{1l}(t_1,\boldsymbol{y}_{-1}) = M_1(f^G(\boldsymbol{y}|\boldsymbol{\delta},\boldsymbol{\psi},\boldsymbol{r}_l)) \in \mathcal{G}_1$ for $k = 1,2,\cdots,K$ and $l = 1,2,\cdots,L$. It is noteworthy that M_1 is a linear mapping so that if applying M_1 to both sides of Equation (7), then we have

$$\sum_{k=1}^{K} \pi_k \Phi_{1k}(t_1,\boldsymbol{y}_{-1}) = \sum_{l=1}^{L} \xi_l \Psi_{1l}(t_1,\boldsymbol{y}_{-1}). \tag{8}$$

More specifically,

$$\begin{aligned}\Phi_{1k}(t_1,\boldsymbol{y}_{-1}) &= \sum_{y_1=1}^{\infty} \frac{1}{1+exp(\gamma_0+\gamma_1 y_1)} C_{y_1}^{\phi_1+y_1-1}(p_{1k})^{y_1}(1-p_{1k})^{\phi_1}(t_1)^{y_1} \cdot f^{G-1}(\boldsymbol{y}_{-1}|\boldsymbol{\gamma},\boldsymbol{\phi}_{-1},\boldsymbol{p}_{-1,k}) \\ &= [(\frac{1-p_{1k}}{1-p_{1k}t_1})^{\phi_1} - R_{1k}(t_1)] \cdot f^{G-1}(\boldsymbol{y}_{-1}|\boldsymbol{\gamma},\boldsymbol{\phi}_{-1},\boldsymbol{p}_{-1,k}),\end{aligned} \tag{9}$$

where $R_{1k}(t_1) = (1-p_{1k})^{\phi_1} + \sum_{y_1=1}^{\infty} \frac{exp(\gamma_0+\gamma_1 y_1)}{1+exp(\gamma_0+\gamma_1 y_1)} C_{y_1}^{\phi_1+y_1-1}(p_{1k}t_1)^{y_1}(1-p_{1k})^{\phi_1}$ is the residual part. Let $t_1 \to \frac{1}{p_{11}}$, because $\gamma_1 < 0$ so that $\frac{p_{1k}exp(\gamma_1)}{p_{11}} < \frac{p_{1k}}{p_{11}} \leq 1$, we have

$$\begin{aligned}\lim_{t_1 \to \frac{1}{p_{11}}} R_{1k}(t_1) &\leq (1-p_{1k})^{\phi_1} + \lim_{t_1 \to \frac{1}{p_{11}}} \sum_{y_1=1}^{\infty} exp(\gamma_0+\gamma_1 y_1) C_{y_1}^{\phi_1+y_1-1}(p_{1k}t_1)^{y_1}(1-p_{1k})^{\phi_1} \\ &= [1-exp(\gamma_0)](1-p_{1k})^{\phi_1} + exp(\gamma_0)(\frac{1-p_{1k}}{1-p_{1k}exp(\gamma_1)/p_{11}})^{\phi_1} < \infty,\end{aligned} \tag{10}$$

Similarly, $\Psi_{1l}(t_1, \boldsymbol{y}_{-1}) = \sum_{y_1=1}^{\infty} f^G(\boldsymbol{y}|\boldsymbol{\delta}, \boldsymbol{\psi}, \boldsymbol{r}_l) t_1^{y_1} = [(\frac{1-r_{1l}}{1-r_{1l}t_1})^{\psi_1} - S_{1l}(t_1)] \cdot f^{G-1}(\boldsymbol{y}_{-1}|\boldsymbol{\delta}, \boldsymbol{\psi}_{-1}, \boldsymbol{r}_{-1,l})$,

$$\begin{aligned} S_{1l}(t_1) &= (1-r_{1l})^{\psi_1} + \sum_{y_1=1}^{\infty} exp(\delta_0 + \delta_1 y_1) C_{y_1}^{\psi_1+y_1-1} (r_{1l}t_1)^{y_1} (1-r_{1l})^{\psi_1} \\ &\leq [1 - exp(\delta_0)](1-r_{1l})^{\psi_1} + exp(\delta_0)(\frac{1-r_{1l}}{1-r_{1l}exp(\delta_1)t_1})^{\psi_1} \end{aligned} \tag{11}$$

As $t_1 \to \frac{1}{p_{11}}$, because $\frac{r_{1l}exp(\gamma_1)}{p_{11}} < \frac{r_{1l}}{p_{11}} \leq \frac{r_{11}}{p_{11}} \leq 1$, we have $\lim_{t_1 \to \frac{1}{p_{11}}} S_{1l}(t_1) < \infty$.
Notice that $f^G(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{p}_1) \succeq f^G(\boldsymbol{y}|\boldsymbol{\delta}, \boldsymbol{\psi}, \boldsymbol{r}_1)$ implies $r_{11} < p_{11}$ or $r_{11} = p_{11}, \psi_1 \leq \phi_1$. According to (10) and (11), we have

$$\begin{aligned} \lim_{t_1 \to \frac{1}{p_{11}}} \frac{\Psi_{1l}(t_1, \boldsymbol{y}_{-1})}{\Phi_{11}(t_1, \boldsymbol{y}_{-1})} &= \lim_{t_1 \to \frac{1}{p_{11}}} \frac{(\frac{1-r_{1l}}{1-r_{1l}t_1})^{\psi_1} - S_{1l}(t_1)}{(\frac{1-p_{11}}{1-p_{11}t_1})^{\phi_1} - R_{11}(t_1)} \cdot \frac{f^{G-1}(\boldsymbol{y}_{-1}|\boldsymbol{\delta}, \boldsymbol{\psi}_{-1}, \boldsymbol{r}_{-1,l})}{f^{G-1}(\boldsymbol{y}_{-1}|\boldsymbol{\gamma}, \boldsymbol{\phi}_{-1}, \boldsymbol{p}_{-1,1})} \\ &= \frac{f^{G-1}(\boldsymbol{y}_{-1}|\boldsymbol{\delta}, \boldsymbol{\psi}_{-1}, \boldsymbol{r}_{-1,l})}{f^{G-1}(\boldsymbol{y}_{-1}|\boldsymbol{\gamma}, \boldsymbol{\phi}_{-1}, \boldsymbol{p}_{-1,1})} \cdot \lim_{t_1 \to \frac{1}{p_{11}}} \frac{(\frac{1-r_{1l}}{1-r_{1l}t_1})^{\psi_1}(1-p_{11}t_1)^{\phi_1} - S_{1l}(t_1)(1-p_{11}t_1)^{\phi_1}}{(1-p_{11})^{\phi_1} - R_{11}(t_1)(1-p_{11}t_1)^{\phi_1}} \\ &= \begin{cases} \frac{f^{G-1}(\boldsymbol{y}_{-1}|\boldsymbol{\delta}, \boldsymbol{\psi}_{-1}, \boldsymbol{r}_{-1,l})}{f^{G-1}(\boldsymbol{y}_{-1}|\boldsymbol{\gamma}, \boldsymbol{\phi}_{-1}, \boldsymbol{p}_{-1,1})} & , if\ r_{1l} = p_{11}, \psi_1 = \phi_1 \\ 0 & , if\ r_{1l} = p_{11}, \psi_1 < \phi_1 \\ 0 & , if\ r_{1l} < p_{11} \end{cases} \end{aligned} \tag{12}$$

If $r_{11} < p_{11}$ or $r_{11} = p_{11}, \psi_1 < \phi_1$, then dividing $\Phi_{11}(t_1, \boldsymbol{y}_{-1})$ on both sides of Equation (8) and let $t_1 \to \frac{1}{p_{11}}$, we have

$$\lim_{t_1 \to \frac{1}{p_{11}}} \sum_{k=1}^{K} \pi_k \frac{\Phi_{1k}(t_1, \boldsymbol{y}_{-1})}{\Phi_{11}(t_1, \boldsymbol{y}_{-1})} \geq \lim_{t_1 \to \frac{1}{p_{11}}} \pi_1 \frac{\Phi_{11}(t_1, \boldsymbol{y}_{-1})}{\Phi_{11}(t_1, \boldsymbol{y}_{-1})} = \pi_1 > 0 = \lim_{t_1 \to \frac{1}{p_{11}}} \sum_{l=1}^{L} \xi_l \frac{\Psi_{1l}(t_1, \boldsymbol{y}_{-1})}{\Phi_{11}(t_1, \boldsymbol{y}_{-1})},$$

which contradicts with Equation (7). Thus, $r_{11} = p_{11}$ and $\psi_1 = \phi_1$, which means Equation (*) holds. Similar to Equation (12), we have

$$\lim_{t_1 \to \frac{1}{p_{11}}} \frac{\Phi_{1k}(t_1, \boldsymbol{y}_{-1})}{\Phi_{11}(t_1, \boldsymbol{y}_{-1})} = \begin{cases} \frac{f^{G-1}(\boldsymbol{y}_{-1}|\boldsymbol{\gamma}, \boldsymbol{\phi}_{-1}, \boldsymbol{p}_{-1,k})}{f^{G-1}(\boldsymbol{y}_{-1}|\boldsymbol{\gamma}, \boldsymbol{\phi}_{-1}, \boldsymbol{p}_{-1,1})} & , if\ p_{1k} = p_{11} \\ 0 & , if\ p_{1k} < p_{11} \end{cases}$$

Moreover, there exists a $K_1 \leq K$ such that $p_{1k} = p_{11}$ for $k = 1, 2, \cdots, K_1$ but $p_{1k} < p_{11}$ for $k = K_1 + 1, \cdots, K$. There also exists an $L_1 \leq L$ such that $r_{1l} = p_{11}$ for $l = 1, 2, \cdots, L_1$ but $r_{1l} < p_{11}$ for $l = L_1 + 1, \cdots, L$. $p_{11} = p_{11}$ and $r_{11} = p_{11}$, therefore, $K_1 \geq 1$ and $L_1 \geq 1$. Dividing $\Phi_{11}(t_1, \boldsymbol{y}_{-1})$ on both sides of Equation (8) and let $t_1 \to \frac{1}{p_{11}}$, we have

$$\sum_{k=1}^{K_1} \pi_k \frac{\Phi_{1k}(t_1, \boldsymbol{y}_{-1})}{\Phi_{11}(t_1, \boldsymbol{y}_{-1})} = \sum_{l=1}^{L_1} \xi_l \frac{\Psi_{1l}(t_1, \boldsymbol{y}_{-1})}{\Phi_{11}(t_1, \boldsymbol{y}_{-1})}$$

$$\Rightarrow \sum_{k=1}^{K_1} \pi_k f^{G-1}(\boldsymbol{y}_{-1}|\boldsymbol{\gamma}, \boldsymbol{\phi}_{-1}, \boldsymbol{p}_{-1,k}) = \sum_{l=1}^{L_1} \xi_l f^{G-1}(\boldsymbol{y}_{-1}|\boldsymbol{\delta}, \boldsymbol{\psi}_{-1}, \boldsymbol{r}_{-1,l}). \tag{13}$$

Thus, we have proven that Equation (**) holds.

Now let us assume that Equations (*) and (**) hold for for $G_0 = g$. In other words, $p_{j1} = r_{j1}, \phi_j = \psi_j$ for all $j = 1, 2, \cdots, g$ and there are $K_g \geq 1$ and $L_g \geq 1$ such that

$$\sum_{k=1}^{K_g} \pi_k f^{G-g}(\boldsymbol{y}_{-g}|\boldsymbol{\gamma}, \boldsymbol{\phi}_{-g}, \boldsymbol{p}_{-g,k}) = \sum_{l=1}^{L_g} \xi_l f^{G-g}(\boldsymbol{y}_{-g}|\boldsymbol{\delta}, \boldsymbol{\psi}_{-g}, \boldsymbol{r}_{-g,l}). \tag{14}$$

Let $G_0 = g + 1$. Similar to M_1, we define a linear map $M_{g+1} : \mathcal{F}^{G-g} \to \mathcal{G}_{g+1}$ such that $M_{g+1}(f^{G-g}(\boldsymbol{y}_{-g}|\boldsymbol{\gamma}, \boldsymbol{\phi}_{-g}, \boldsymbol{p}_{-g})) = \Phi_{g+1}(t_{g+1}, \boldsymbol{y}_{-g}) = \sum_{y_1=1}^{\infty} f^{G-g}(\boldsymbol{y}_{-g}|\boldsymbol{\gamma}, \boldsymbol{\phi}_{-g}, \boldsymbol{p}_{-g}) t_{g+1}^{y_{g+1}} = \sum_{y_{g+1}=1}^{\infty} f^1(y_{g+1}|\gamma, \phi_{g+1}, p_{g+1}) t_{g+1}^{y_{g+1}} \cdot f^{G-(g+1)}(\boldsymbol{y}_{-(g+1)}|\boldsymbol{\gamma}, \boldsymbol{\phi}_{-(g+1)}, \boldsymbol{p}_{-(g+1)})$. Consequently,

$$M_{g+1}(f^{G-g}(\boldsymbol{y}_{-g}|\boldsymbol{\gamma}, \boldsymbol{\phi}_{-g}, \boldsymbol{p}_{-g,k})) = \Phi_{g+1,k}(t_{g+1}, \boldsymbol{y}_{-(g+1)})$$
$$= [(\frac{1 - p_{g+1,k}}{1 - p_{g+1,k} t_{g+1}})^{\phi_{g+1}} - R_{g+1,k}(t_{g+1})] \cdot f^{G-(g+1)}(\boldsymbol{y}_{-(g+1)}|\boldsymbol{\gamma}, \boldsymbol{\phi}_{-(g+1)}, \boldsymbol{p}_{-(g+1),k}),\ k = 1, \cdots, K_g;$$

$$M_{g+1}(f^{G-g}(\boldsymbol{y}_{-g}|\boldsymbol{\delta}, \boldsymbol{\psi}_{-g}, \boldsymbol{r}_{-g,l})) = \Psi_{g+1,l}(t_{g+1}, \boldsymbol{y}_{-(g+1)})$$
$$= [(\frac{1 - r_{g+1,l}}{1 - p_{g+1,l} t_{g+1}})^{\psi_{g+1}} - S_{g+1,l}(t_{g+1})] \cdot f^{G-(g+1)}(\boldsymbol{y}_{-(g+1)}|\boldsymbol{\delta}, \boldsymbol{\psi}_{-(g+1)}, \boldsymbol{r}_{-(g+1),l}),\ l = 1, \cdots, L_g.$$

If we apply M_{g+1} to both sides of Equation (14), then we have

$$\sum_{k=1}^{K_g} \pi_k \Phi_{g+1,k}(t_{g+1}, \boldsymbol{y}_{-(g+1)}) = \sum_{l=1}^{L_g} \xi_l \Psi_{g+1,l}(t_{g+1}, \boldsymbol{y}_{-(g+1)}). \tag{15}$$

Notice that given $p_{j1} = r_{j1}, \phi_j = \psi_j$ for all $j = 1, 2, \cdots, g$, $f^G(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{p}_1) \succeq f^G(\boldsymbol{y}|\boldsymbol{\delta}, \boldsymbol{\psi}, \boldsymbol{r}_1)$ implies $r_{g+1,1} < p_{g+1,1}$ or $r_{g+1,1} = p_{g+1,1}, \psi_{g+1} \leq \phi_{g+1}$. Similar to Equation (12), we have

$$
\begin{aligned}
&\lim_{t_{g+1}\to\frac{1}{p_{g+1,1}}} \frac{\Psi_{g+1,l}(t_{g+1},\boldsymbol{y}_{-(g+1)})}{\Phi_{g+1,1}(t_{g+1},\boldsymbol{y}_{-(g+1)})} \\
&= \lim_{t_{g+1}\to\frac{1}{p_{g+1,1}}} \frac{(\frac{1-r_{g+1,l}}{1-r_{g+1,l}t_{g+1}})^{\psi_{g+1}} - S_{g+1,l}(t_{g+1})}{(\frac{1-p_{g+1,1}}{1-p_{g+1,1}t_{g+1}})^{\phi_{g+1}} - R_{g+1,1}(t_{g+1})} \cdot \frac{f^{G-(g+1)}(\boldsymbol{y}_{-(g+1)}|\boldsymbol{\delta},\boldsymbol{\psi}_{-(g+1)},\boldsymbol{r}_{-(g+1),l})}{f^{G-(g+1)}(\boldsymbol{y}_{-(g+1)}|\boldsymbol{\gamma},\boldsymbol{\phi}_{-(g+1)},\boldsymbol{p}_{-(g+1),1})} \\
&= \frac{f^{G-(g+1)}(\boldsymbol{y}_{-(g+1)}|\boldsymbol{\delta},\boldsymbol{\psi}_{-(g+1)},\boldsymbol{r}_{-(g+1),l})}{f^{G-(g+1)}(\boldsymbol{y}_{-(g+1)}|\boldsymbol{\gamma},\boldsymbol{\phi}_{-(g+1)},\boldsymbol{p}_{-(g+1),1})} \cdot \lim_{t_{g+1}\to\frac{1}{p_{g+1,1}}} \frac{(\frac{1-r_{g+1,l}}{1-r_{g+1,l}t_{g+1}})^{\psi_{g+1}} - S_{g+1,l}(t_{g+1})}{(\frac{1-p_{g+1,1}}{1-p_{g+1,1}t_{g+1}})^{\phi_{g+1}} - R_{g+1,1}(t_{g+1})} \\
&= \begin{cases} \frac{f^{G-(g+1)}(\boldsymbol{y}_{-(g+1)}|\boldsymbol{\delta},\boldsymbol{\psi}_{-(g+1)},\boldsymbol{r}_{-(g+1),l})}{f^{G-(g+1)}(\boldsymbol{y}_{-(g+1)}|\boldsymbol{\gamma},\boldsymbol{\phi}_{-(g+1)},\boldsymbol{p}_{-(g+1),1})} & , \ if \ r_{g+1,l} = p_{g+1,1}, \psi_{g+1} = \phi_{g+1} \\ 0 & , \ if \ r_{g+1,l} = p_{g+1,1}, \psi_{g+1} < \phi_{g+1} \\ 0 & , \ if \ r_{g+1,l} < p_{g+1,1} \end{cases}
\end{aligned} \tag{16}
$$

If $r_{g+1,l} < p_{g+1,1}$ or $r_{g+1,l} = p_{g+1,1}, \psi_{g+1} < \phi_{g+1}$, then dividing $\Phi_{g+1,1}(t_{g+1},\boldsymbol{y}_{-(g+1)})$ on both sides of Equation (15) and letting $t_{g+1} \to \frac{1}{p_{g+1,1}}$, we have

$$
\begin{aligned}
\lim_{t_{g+1}\to\frac{1}{p_{g+1,1}}} \sum_{k=1}^{K_g} \pi_k \frac{\Phi_{g+1,k}(t_{g+1},\boldsymbol{y}_{-(g+1)})}{\Phi_{g+1,1}(t_{g+1},\boldsymbol{y}_{-(g+1)})} &\geq \lim_{t_{g+1}\to\frac{1}{p_{g+1,1}}} \pi_1 \frac{\Phi_{g+1,1}(t_{g+1},\boldsymbol{y}_{-(g+1)})}{\Phi_{g+1,1}(t_{g+1},\boldsymbol{y}_{-(g+1)})} = \pi_1 \\
&> 0 = \lim_{t_{g+1}\to\frac{1}{p_{g+1,1}}} \sum_{l=1}^{L_g} \xi_l \frac{\Psi_{g+1,l}(t_{g+1},\boldsymbol{y}_{-(g+1)})}{\Phi_{g+1,1}(t_{g+1},\boldsymbol{y}_{-(g+1)})},
\end{aligned} \tag{17}
$$

which contradicts with (14). Thus, $r_{g+1,l} = p_{g+1,1}$ and $\psi_{g+1} = \phi_{g+1}$, which means that Equation $(*)$ holds. Similar to Equation (16), for $k = 1, 2, \cdots, K_g$, we have

$$
\lim_{t_{g+1}\to\frac{1}{p_{g+1,1}}} \frac{\Phi_{g+1,k}(t_{g+1},\boldsymbol{y}_{-(g+1)})}{\Phi_{g+1,1}(t_{g+1},\boldsymbol{y}_{-(g+1)})} = \begin{cases} \frac{f^{G-(g+1)}(\boldsymbol{y}_{-(g+1)}|\boldsymbol{\gamma},\boldsymbol{\phi}_{-(g+1)},\boldsymbol{p}_{-(g+1),k})}{f^{G-(g+1)}(\boldsymbol{y}_{-(g+1)}|\boldsymbol{\gamma},\boldsymbol{\phi}_{-(g+1)},\boldsymbol{p}_{-(g+1),1})} & , \ if \ p_{g+1,k} = p_{g+1,1} \\ 0 & , \ if \ p_{g+1,k} < p_{g+1,1} \end{cases}
$$

Further, there exists a $K_{g+1} \leq K_g$ such that $p_{g+1,k} = p_{g+1,1}$ for $k = 1, 2, \cdots, K_{g+1}$ but $p_{g+1,k} < p_{g+1,1}$ for $k = K_{g+1}+1, K_{g+1}+2, \cdots, K_g$. There also exists an $L_{g+1} \leq L_g$ such that $r_{g+1,l} = p_{g+1,1}$ for $l = 1, 2, \cdots, L_{g+1}$ but $r_{g+1,l} < p_{g+1,1}$ for $l = L_{g+1}+1, L_{g+1}+2, \cdots, L_g$. $p_{g+1,1} = p_{g+1,1}$ and $p_{g+1,1} = r_{g+1,1}$, therefore, $K_{g+1} \geq 1$ and $L_{g+1} \geq 1$. Dividing $\Phi_{g+1,1}(t_{g+1},\boldsymbol{y}_{-(g+1)})$ on both sides of Equation (15) and letting $t_{g+1} \to \frac{1}{p_{g+1,1}}$, we have,

$$
\begin{aligned}
&\sum_{k=1}^{K_{g+1}} \pi_k \frac{\Phi_{g+1,k}(t_{g+1},\boldsymbol{y}_{-(g+1)})}{\Phi_{g+1,1}(t_{g+1},\boldsymbol{y}_{-(g+1)})} = \sum_{l=1}^{L_{g+1}} \xi_l \frac{\Psi_{g+1,l}(t_{g+1},\boldsymbol{y}_{-(g+1)})}{\Phi_{g+1,1}(t_{g+1},\boldsymbol{y}_{-(g+1)})} \\
\Rightarrow &\sum_{k=1}^{K_{g+1}} \pi_k f^{G-(g+1)}(\boldsymbol{y}_{-(g+1)}|\boldsymbol{\gamma},\boldsymbol{\phi}_{-(g+1)},\boldsymbol{p}_{-(g+1),k}) = \sum_{l=1}^{L_{g+1}} \xi_l f^{G-(g+1)}(\boldsymbol{y}_{-(g+1)}|\boldsymbol{\delta},\boldsymbol{\psi}_{-(g+1)},\boldsymbol{r}_{-(g+1),l}),
\end{aligned}
$$

so Equation (**) holds for $G_0 = g+1$.

Consequently, by mathematical induction, we have shown that Equations (*) and (**) hold for any $G_0 \in \{1, \cdots, G\}$, which implies that $\boldsymbol{p}_1 = \boldsymbol{r}_1$ and $\boldsymbol{\phi} = \boldsymbol{\psi}$.

For $G_0 = G$ and $G_0 = G - 1$, Equation (**) gives

$$\sum_{k=1}^{K_G} \pi_k = \sum_{l=1}^{L_G} \xi_l, \tag{18}$$

$$\sum_{k=1}^{K_{G-1}} \pi_k f^1(y_G|\boldsymbol{\gamma}, \phi_G, p_{Gk}) = \sum_{l=1}^{L_{G-1}} \xi_l f^1(y_G|\boldsymbol{\delta}, \phi_G, r_{Gl}). \tag{19}$$

For any two distinct elements $f^G(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{p}_1)$ and $f^G(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{p}_k), k = 2, 3, \cdots, K_{G-2}$, because there exist at least two different dimensions and $p_{g1} = p_{gk}$ with $g = 1, 2, \cdots, G-2$, $p_{G-1,k} \neq p_{G-1,1}$ and $p_{Gk} \neq p_{G1}$. Therefore, $K_G = K_{G-1} = 1$. Similarly, we have $L_G = L_{G-1} = 1$. Thus, Equation (18) and (19) turn to

$$\pi_1 = \xi_1$$
$$f^1(y_G|\boldsymbol{\gamma}, \phi_G, p_{G1}) = f^1(y_G|\boldsymbol{\delta}, \phi_G, r_{G1}), \forall y_G \in \mathbb{N}. \tag{20}$$

Plugging $y_G = 1$ and $y_G = 2$ into Equation (20), we have

$$\frac{1}{1 + exp(\gamma_0 + \gamma_1)} C_1^{\phi_G} p_{G1}(1 - p_{G1})^{\phi_G} = \frac{1}{1 + exp(\delta_0 + \delta_1)} C_1^{\phi_G} p_{G1}(1 - p_{G1})^{\phi_G}$$
$$\frac{1}{1 + exp(\gamma_0 + 2\gamma_1)} C_2^{\phi_G + 1} p_{G1}^2(1 - p_{G1})^{\phi_G} = \frac{1}{1 + exp(\delta_0 + 2\delta_1)} C_2^{\phi_G + 1} p_{G1}^2(1 - p_{G1})^{\phi_G},$$

therefore $\boldsymbol{\gamma} = \boldsymbol{\delta}$.

Plugging $\boldsymbol{\gamma} = \boldsymbol{\delta}, \boldsymbol{\phi} = \boldsymbol{\psi}, \boldsymbol{p}_1 = \boldsymbol{r}_1$ and $\pi_1 = \xi_1$ into Equation (7), we have

$$\sum_{k=2}^{K} \pi_k f^G(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{p}_k) = \sum_{l=2}^{L} \xi_l f^G(\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{r}_l) \tag{21}$$

Similarly, we can apply mathematical induction to prove that $\boldsymbol{p}_k = \boldsymbol{r}_k$ and $\pi_k = \xi_k$ sequentially for $k = 2, 3, \cdots, \min\{K, L\}$. Finally, if $K \neq L$, without loss of generality, let us assume that $K > L$, then $\sum_{k=L+1}^{K} \pi_k = 1 - \sum_{k=1}^{L} \pi_k = 1 - \sum_{l=1}^{L} \xi_l = 0$, which contradicts with $\pi_k > 0$ for all $k = 1, 2, \cdots, K$. Thus, $K = L, \boldsymbol{\gamma} = \boldsymbol{\delta}, \boldsymbol{\phi} = \boldsymbol{\psi}, \boldsymbol{\pi} = \boldsymbol{\xi}$ and $\boldsymbol{p}_k = \boldsymbol{r}_k$ for all $k = 1, 2, \cdots, K$. Therefore, the class of all finite mixtures of $\mathcal{F}^G$ is identifiable.

Theorem 1

Proof. Let $\boldsymbol{Y}_b \in N^{n_b \times G}$ denote the data from batch b and collect $\boldsymbol{Y} = \{\boldsymbol{Y}_b, 1 \leq b \leq B\}$, then the marginal distribution for $f(\boldsymbol{Y}_b|\boldsymbol{\Theta}) = \prod_{i=1}^{n_b} [\sum_{k=1}^{K} \pi_{bk} f^G(\boldsymbol{y}_b|\boldsymbol{\gamma}_b, \boldsymbol{\phi}_b, \boldsymbol{m}_{bik})]$ with

$f^G(\boldsymbol{y}_b|\boldsymbol{\gamma}_b, \boldsymbol{\phi}_b, \boldsymbol{m}_{bik}) \in \mathcal{F}^G$ reduces to a mixture of G-dimensional zero-inflated negative binomial (ZINB) model on batch b. Therefore, we can view the BUSseq model as a combination of B ZINB models with the constraints that $\boldsymbol{\beta}_k^{(1)} = \cdots = \boldsymbol{\beta}_k^{(B)} = \boldsymbol{\beta}_k$ for each k and $\boldsymbol{\alpha}^{(1)} = \cdots = \boldsymbol{\alpha}^{(B)} = \boldsymbol{\alpha}$.

According to conditions (I)-(III), Lemma 1 and [37], the ZINB model for batch b is identifiable up to label switching in the sense that $f(\boldsymbol{Y}_b|\boldsymbol{\Theta}) = f(\boldsymbol{Y}_b|\boldsymbol{\Theta}^*)$ for any $\boldsymbol{Y}_b$ implies that $\pi_{bk} = \pi^*_{b\rho_b(k)}, \boldsymbol{\gamma}_b = \boldsymbol{\gamma}^*_b, \boldsymbol{\alpha} + \boldsymbol{\beta}_k + \boldsymbol{\nu}_b + \delta_{bi} = log(\boldsymbol{m}_{bik}) = log(\boldsymbol{m}^*_{bi\rho_b(k)}) = \boldsymbol{\alpha}^* + \boldsymbol{\beta}^*_{\rho_b(k)} + \boldsymbol{\nu}^*_b + \delta^*_{bi}$ and $\boldsymbol{\phi}_b = \boldsymbol{\phi}^*_b$ for a permutation ρ_b of $\{1, 2, \cdots, K\}$.

We first prove that the permutation ρ_b is the same for all of the batches. Recall that we take the first cell type as the reference cell type with $\boldsymbol{\beta}_1 = 0$. Therefore, the ratio of mean expression levels between cell type k and cell type one is

$$\frac{\boldsymbol{m}_{bik}}{\boldsymbol{m}_{bi1}} = \frac{\boldsymbol{m}^*_{bi\rho_b(k)}}{\boldsymbol{m}^*_{bi\rho_b(1)}} \Rightarrow exp(\boldsymbol{\beta}_k) = exp(\boldsymbol{\beta}^*_{\rho_b(k)} - \boldsymbol{\beta}^*_{\rho_b(1)}) \tag{22}$$

Notice the left hand side of Equation (22) is invariant to the batch indicator b, and therefore $\boldsymbol{\beta}^*_{\rho_b(k)} - \boldsymbol{\beta}^*_{\rho_b(1)} = \boldsymbol{\beta}^*_{\rho_1(k)} - \boldsymbol{\beta}^*_{\rho_1(1)}$ for every k. By condition (III), $\rho_b = \rho_1$ for every b.

Let us then compare $log(\boldsymbol{m}_{b1k})$ with $log(\boldsymbol{m}_{11k})$. Because $\boldsymbol{\nu}_1 = \boldsymbol{\nu}^*_1 = \boldsymbol{0}$ and $\delta_{b1} = \delta^*_{b1} = 0$, we have

$$\boldsymbol{\alpha} + \boldsymbol{\beta}_k = \boldsymbol{\alpha}^* + \boldsymbol{\beta}^*_{\rho(k)}, \boldsymbol{\alpha} + \boldsymbol{\beta}_k + \boldsymbol{\nu}_b = \boldsymbol{\alpha}^* + \boldsymbol{\beta}^*_{\rho(k)} + \boldsymbol{\nu}^*_b. \tag{23}$$

Thus, we have proven $\boldsymbol{\nu}_b = \boldsymbol{\nu}^*_b$.

Next we compare $log(\boldsymbol{m}_{bik})$ with $log(\boldsymbol{m}_{b1k})$ for each batch. Then, we have

$$\boldsymbol{\alpha} + \boldsymbol{\beta}_k + \boldsymbol{\nu}_b = \boldsymbol{\alpha}^* + \boldsymbol{\beta}^*_{\rho(k)} + \boldsymbol{\nu}_b, \boldsymbol{\alpha} + \boldsymbol{\beta}_k + \boldsymbol{\nu}_b + \boldsymbol{\delta}_{bi} = \boldsymbol{\alpha}^* + \boldsymbol{\beta}^*_{\rho(k)} + \boldsymbol{\nu}_b + \delta^*_{bi}. \tag{24}$$

Consequently, $\delta_{bi} = \delta^*_{bi}$ for any cell i in any batch. Therefore, BUSseq is identifiable (up to label switching).

Theorem 2

Proof. In the reference panel design, any batch b shares at least two cell types with the first batch. If we compare the two distinct cell types k_1 and k_2 shared by batch b and batch one

in terms of the log-scale mean expression levels, respectively, then we have

$$\left.\begin{aligned}\boldsymbol{\alpha}+\boldsymbol{\beta}_{k_1}+\boldsymbol{\nu}_b+\delta_{bi}=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_b(k_1)}+\boldsymbol{\nu}^*_b+\delta^*_{bi},\\ \boldsymbol{\alpha}+\boldsymbol{\beta}_{k_2}+\boldsymbol{\nu}_b+\delta_{bi}=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_b(k_2)}+\boldsymbol{\nu}^*_b+\delta^*_{bi},\end{aligned}\right\}\Rightarrow\boldsymbol{\beta}_{k_1}-\boldsymbol{\beta}_{k_2}=\boldsymbol{\beta}^*_{\rho_b(k_1)}-\boldsymbol{\beta}^*_{\rho_b(k_2)},$$

$$\left.\begin{aligned}\boldsymbol{\alpha}+\boldsymbol{\beta}_{k_1}+\delta_{1i}=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_1(k_1)}+\delta^*_{1i},\\ \boldsymbol{\alpha}+\boldsymbol{\beta}_{k_2}+\delta_{1i}=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_1(k_2)}+\delta^*_{1i}.\end{aligned}\right\}\Rightarrow\boldsymbol{\beta}_{k_1}-\boldsymbol{\beta}_{k_2}=\boldsymbol{\beta}^*_{\rho_1(k_1)}-\boldsymbol{\beta}^*_{\rho_1(k_2)}.$$

Further, according to condition (III), $\boldsymbol{\beta}_{k_1}-\boldsymbol{\beta}_{k_2}=\boldsymbol{\beta}^*_{\rho_b(k_1)}-\boldsymbol{\beta}^*_{\rho_b(k_2)}=\boldsymbol{\beta}^*_{\rho_1(k_1)}-\boldsymbol{\beta}^*_{\rho_1(k_2)}$ implies that $\rho_b(k)=\rho_1(k)$ for each cell type $k\in C_b$ $(b\geq 2)$.

Finally, similar to Equations (23) and (24), for a shared cell type k between batch b and batch one, we have

$$\begin{aligned}\boldsymbol{\alpha}+\boldsymbol{\beta}_k&=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_1(k)}\\ \boldsymbol{\alpha}+\boldsymbol{\beta}_k+\boldsymbol{\nu}_b&=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_b(k)}+\boldsymbol{\nu}^*_b\\ \boldsymbol{\alpha}+\boldsymbol{\beta}_k+\boldsymbol{\nu}_b+\delta_{bi}&=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_b(k)}+\boldsymbol{\nu}^*_b+\delta^*_{bi}.\end{aligned}$$

Thus, $\rho_b(k)=\rho_1(k)$ for each $k\in C_b$ $(b\geq 2)$ implies that $\boldsymbol{\nu}_b=\boldsymbol{\nu}^*_b$ and $\delta_{bi}=\delta^*_{bi}$.

Theorem 3

Proof. Our objective is to prove that for any two distinct batches b and $\tilde{b}$, $\rho_b(k)=\rho_{\tilde{b}}(k)$ holds for any cell type $k\in C_b\bigcap C_{\tilde{b}}$ shared by these two batches.

First, we prove that $\rho_b(k)=\rho_{b-1}(k)$ for the shared cell types $k\in C_b\bigcap C_{b-1}, 2\leq b\leq B$ in any two consecutive batches. Notice that $|C_b\cap C_{b-1}|\geq 2$, so for any two shared cell types k_1 and k_2 between batch b and batch $b-1$, we have

$$\left.\begin{aligned}\boldsymbol{\alpha}+\boldsymbol{\beta}_{k_1}+\boldsymbol{\nu}_b+\delta_{bi}=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_b(k_1)}+\boldsymbol{\nu}^*_b+\delta^*_{bi},\\ \boldsymbol{\alpha}+\boldsymbol{\beta}_{k_2}+\boldsymbol{\nu}_b+\delta_{bi}=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_b(k_2)}+\boldsymbol{\nu}^*_b+\delta^*_{bi},\end{aligned}\right\}\Rightarrow\boldsymbol{\beta}_{k_1}-\boldsymbol{\beta}_{k_2}=\boldsymbol{\beta}^*_{\rho_b(k_1)}-\boldsymbol{\beta}^*_{\rho_b(k_2)}$$

$$\left.\begin{aligned}\boldsymbol{\alpha}+\boldsymbol{\beta}_{k_1}+\boldsymbol{\nu}_b+\delta_{b-1,i}=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_{b-1}(k_1)}+\boldsymbol{\nu}^*_{b-1}+\delta^*_{b-1,i},\\ \boldsymbol{\alpha}+\boldsymbol{\beta}_{k_2}+\boldsymbol{\nu}_b+\delta_{b-1,i}=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_{b-1}(k_2)}+\boldsymbol{\nu}^*_{b-1}+\delta^*_{b-1,i}.\end{aligned}\right\}\Rightarrow\boldsymbol{\beta}_{k_1}-\boldsymbol{\beta}_{k_2}=\boldsymbol{\beta}^*_{\rho_{b-1}(k_1)}-\boldsymbol{\beta}^*_{\rho_{b-1}(k_2)}$$

Further, according to condition (III), $\boldsymbol{\beta}_{k_1}-\boldsymbol{\beta}_{k_2}=\boldsymbol{\beta}^*_{\rho_b(k_1)}-\boldsymbol{\beta}^*_{\rho_b(k_2)}=\boldsymbol{\beta}^*_{\rho_{b-1}(k_1)}-\boldsymbol{\beta}^*_{\rho_{b-1}(k_2)}$ implies that $\rho_b(k)=\rho_{b-1}(k)$ for $2\leq b\leq B, k\in C_b\bigcap C_{b-1}$.

Consequently, for a cell type $k\in C_b\bigcap C_{b-1}$ shared by two consecutive batches b and $b-1$,

similar to Equation (23), we have

$$\left.\begin{array}{l}\boldsymbol{\alpha}+\boldsymbol{\beta}_k+\boldsymbol{\nu}_{b-1}=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_{b-1}(k)}+\boldsymbol{\nu}^*_{b-1}\\ \boldsymbol{\alpha}+\boldsymbol{\beta}_k+\boldsymbol{\nu}_b=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_b(k)}+\boldsymbol{\nu}^*_b\end{array}\right\}\Rightarrow \boldsymbol{\nu}_b-\boldsymbol{\nu}_{b-1}=\boldsymbol{\nu}^*_b-\boldsymbol{\nu}^*_{b-1}.$$

Because $\boldsymbol{\nu}_1=\boldsymbol{\nu}^*_1=0$, $\boldsymbol{\nu}_b=\sum_{j=2}^b(\boldsymbol{\nu}_j-\boldsymbol{\nu}_{j-1})=\sum_{j=2}^b(\boldsymbol{\nu}^*_j-\boldsymbol{\nu}^*_{j-1})=\boldsymbol{\nu}^*_b$. Moreover, similar to Equation (24), we have $\delta_{bi}=\delta^*_{bi}$ for each cell i of each batch b.

Now for any two distinct batches b and $\tilde{b}$, we can directly compare the mean expression levels of their shared cell type $k\in C_b\bigcap C_{\tilde{b}}$:

$$\left.\begin{array}{l}\boldsymbol{\alpha}+\boldsymbol{\beta}_k+\boldsymbol{\nu}_b+\delta_{bi}=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_b(k)}+\boldsymbol{\nu}_b+\delta_{bi}\\ \boldsymbol{\alpha}+\boldsymbol{\beta}_k+\boldsymbol{\nu}_{\tilde{b}}+\delta_{\tilde{b}i}=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_{\tilde{b}}(k)}+\boldsymbol{\nu}_{\tilde{b}}+\delta_{\tilde{b}i}\end{array}\right\}\Rightarrow \boldsymbol{\beta}^*_{\rho_b(k)}=\boldsymbol{\beta}^*_{\rho_{\tilde{b}}(k)}$$

Consequently, we have proven Theorem 3.

The Markov chain Monte Carlo (MCMC) Algorithm for BUSseq

To conduct posterior inference, we develop an MCMC algorithm to draw samples from the posterior distribution. At iteration t:

1. Update $z^{[t]}_{big}$ and $x^{[t]}_{big}$ sequentially for (b,i,g), if $y_{big}=0$:

$$z^{[t]}_{big}\begin{cases}=1 & \text{, if } x^{[t-1]}_{big}>0;\\ \sim Bernoulli(\frac{exp(\gamma^{[t-1]}_{b0})}{1+exp(\gamma^{[t-1]}_{b0})}) & \text{, if } x^{[t-1]}_{big}=0.\end{cases}$$

$$x^{[t]}_{big}\begin{cases}=0 & \text{, if } z^{[t]}_{big}=0;\\ \propto \frac{exp(\gamma^{[t-1]}_{b0}+\gamma^{[t-1]}_{b1}x^{[t]}_{big})}{1+exp(\gamma^{[t-1]}_{b0}+\gamma^{[t-1]}_{b1}x^{[t]}_{big})}\frac{\Gamma(\phi^{[t-1]}_{bg}+x^{[t]}_{big})(\mu^{[t-1]}_{big})^{x^{[t]}_{big}}}{\Gamma(x^{[t]}_{big})(\phi^{[t-1]}_{bg}+\mu^{[t-1]}_{big})^{\phi^{[t-1]}_{bg}+x^{[t]}_{big}}} & \text{, if } z^{[t]}_{big}=1.\end{cases}$$

   where $\mu^{[t-1]}_{big}=exp(\alpha^{[t-1]}_g+\beta^{[t-1]}_{gw^{[t-1]}_{bi}}+\nu^{[t-1]}_{bg}+\delta^{[t-1]}_{bi})$, and $\Gamma(\cdot)$ represents the Gamma function.
   When $z^{[t]}_{big}=1$, we incorporate a Metropolis-Hasting (MH) step [38]. We sample x^*_{big} from the proposal distribution $NB(\mu^{[t-1]}_{big},\phi^{[t-1]}_{bg})$ and accept the proposal with probability

$$\rho=\frac{1+exp(-\gamma^{[t-1]}_{b0}-\gamma^{[t-1]}_{b1}x^{[t-1]}_{big})}{1+exp(-\gamma^{[t-1]}_{b0}-\gamma^{[t-1]}_{b1}x^*_{big})}.$$

   On the other hand, if $y_{big}>0$, then $z^{[t]}_{big}=0$ and $x^{[t]}_{big}=y_{big}$.

2. Update $\gamma_{b0}^{[t]}$ and $\gamma_{b1}^{[t]}$ sequentially. Because

$$L(\gamma_b^{[t]}) \propto \prod_{i=1}^{n_b}\prod_{g=1}^{G} \frac{exp[(\gamma_{b1}^{[t]}x_{big}^{[t]} + \gamma_{b0}^{[t]})z_{big}^{[t]}]}{1 + exp(\gamma_{b1}^{[t]}x_{big}^{[t]} + \gamma_{b0}^{[t]})} \cdot exp(-\frac{(\gamma_{b0}^{[t]})^2}{2\sigma_{z0}^2} - \frac{(\gamma_{b1}^{[t]})^2}{2\sigma_{z1}^2}),$$

we update γ_{b0} by an MH step with the symmetric proposal distribution $g(\gamma_{b0}^*|\gamma_{b0}^{[t-1]}) \sim N(\gamma_{b0}^{[t-1]}, \sigma_{MH}^2)$, where σ_{MH}^2 is the pre-defined variance for MH steps to update $\alpha_g^{[t]}$. Consequently, the acceptance rate is

$$\rho = \frac{L(\gamma_{b0}^*|-)}{L(\gamma_{b0}^{[t-1]}|-)} = \prod_{i=1}^{n_b}\prod_{g=1}^{G} \frac{exp(-\gamma_{b0}^{[t-1]}z_{big}^{[t]}) + exp(\gamma_{b1}^{[t-1]}x_{big}^{[t]})}{exp(-\gamma_{b0}^*z_{big}^{[t]}) + exp(\gamma_{b1}^{[t-1]}x_{big}^{[t]})} \cdot exp(-\frac{(\gamma_{b0}^*)^2 - (\gamma_{b0}^{[t-1]})^2}{2\sigma_{z0}^2}).$$

To update γ_{b1}, we incorporate an MH step with the proposal distribution $g(-\gamma_{b1}^*|\gamma_{b1}^{[t-1]}) \sim Gamma(-10\gamma_{b1}^{[t-1]}, 10)$, and the acceptance rate being

$$\rho = \frac{L(\gamma_{b1}^*|-)}{L(\gamma_{b1}^{[t-1]}|-)} = \prod_{i=1}^{n_b}\prod_{g=1}^{G} \frac{exp(-\gamma_{b1}^{[t-1]}x_{big}^{[t]}z_{big}^{[t]}) + exp(\gamma_{b0}^{[t]})}{exp(-\gamma_{b1}^*x_{big}^{[t]}z_{big}^{[t]}) + exp(\gamma_{b0}^{[t]})} \cdot exp(-\frac{(\gamma_{b1}^*)^2 - (\gamma_{b1}^{[t-1]})^2}{2\sigma_{z1}^2})$$
$$\cdot \frac{(-\gamma_{b1}^{[t-1]})^{-\gamma_{b1}^*-1}10^{-10\gamma_{b1}^*}\Gamma(-\gamma_{b1}^{[t-1]})}{(-\gamma_{b1}^*)^{-\gamma_{b1}^{[t-1]}-1}10^{-10\gamma_{b1}^{[t-1]}}\Gamma(-\gamma_{b1}^*)} exp(10(\gamma_{b1}^{[t-1]} - \gamma_{b1}^*)).$$

3. For each gene g, we use an MH step to update α_g. Specifically, we let the proposal distribution be the symmetric $g(\alpha_g^*|\alpha_g^{[t-1]}) \sim N(\alpha_g^{[t-1]}, \sigma_{MH}^2)$ and the acceptance rate be:

$$\rho = \frac{L(\alpha_g^*|-)}{L(\alpha_g^{[t-1]}|-)}$$
$$= \prod_{b=1}^{B}\prod_{i=1}^{n_b} exp((\alpha_g^* - \alpha_g^{[t-1]})x_{big}^{[t]})(\frac{\phi_{bg}^{[t-1]} + exp(\alpha_g^{[t-1]} + \beta_{gw_{bi}^{[t-1]}}^{[t-1]} + \nu_{bg}^{[t-1]} + \delta_{bi}^{[t-1]})}{\phi_{bg}^{[t-1]} + exp(\alpha_g^* + \beta_{gw_{bi}^{[t-1]}}^{[t-1]} + \nu_{bg}^{[t-1]} + \delta_{bi}^{[t-1]})})^{\phi_{bg}^{[t-1]} + x_{big}^{[t]}}$$
$$\cdot exp(-\frac{(\alpha_g^*)^2 - (\alpha_g^{[t-1]})^2}{2\sigma_a^2}).$$

4. For each gene g and for $2 \le k \le K$, we sample the indicator $L_{gk}^{[t]}$ from:

$$L_{gk}^{[t]} \sim Bernoulli(\frac{p^{[t-1]}N(\beta_{gk}^{[t-1]}; 0, (\tau_{\beta 1}^{[t-1]})^2)}{p^{[t-1]}N(\beta_{gk}^{[t-1]}; 0, (\tau_{\beta 1}^{[t-1]})^2) + (1-p)N(\beta_{gk}^{[t-1]}; 0, \tau_{\beta 0}^2)}).$$

5. Update the inclusion probability $p^{[t]}$ for $L_{gk}^{[t]}$s by sampling:

$$p^{[t]} \sim Beta(\sum_{g=1}^{G}\sum_{k=2}^{K} L_{gk}^{[t]} + a_p, G(K-1) - \sum_{g=1}^{G}\sum_{k=2}^{K} L_{gk}^{[t]} + b_p).$$

6. Update the variance of the spike component of the spike-and-slab prior $(\tau_{\beta 0}^{[t]})^2$ by sampling:

$$(\tau_{\beta 0}^{[t]})^2 \sim Inv - Gamma(a_\tau + \frac{1}{2}\#\{(g,k) : L_{gk}^{[t]} = 0, 1 \le g \le G, 2 \le k \le K\},$$
$$b_\tau + \frac{1}{2}\sum_{g=1}^{G}\sum_{k=2}^{K} I(L_{gk}^{[t]} = 0) \cdot (\beta_{gk}^{[t-1]})^2),$$

where $\#\{\cdot\}$ represents the number of elements in the set, and $I(\cdot)$ denotes the indicator function.

7. To update $\beta_{gk}^{[t]}$ for cell type two to K and each gene g, we use an MH step with the symmetric proposal distribution $g(\beta_{gk}^{*}|\beta_{gk}^{[t-1]}) \sim N(\beta_{gk}^{[t-1]}, \sigma_{MH}^2)$ and the acceptance rate

$$\rho = \frac{L(\beta_{gk}^{*}|-)}{L(\beta_{gk}^{[t-1]}|-)}$$
$$= \prod_{(b,i):w_{bi}^{[t-1]}=k} exp((\beta_{gk}^{*} - \beta_{gk}^{[t-1]})x_{big}^{[t]})(\frac{\phi_{bg}^{[t-1]} + exp(\alpha_g^{[t]} + \beta_{gk}^{[t-1]} + \nu_{bg}^{[t-1]} + \delta_{bi}^{[t-1]})}{\phi_{bg}^{[t-1]} + exp(\alpha_g^{[t]} + \beta_{gk}^{*} + \nu_{bg}^{[t-1]} + \delta_{bi}^{[t-1]})})^{\phi_{bg}^{[t-1]} + x_{big}^{[t]}}$$
$$\cdot exp(-\frac{(\beta_{gk}^{*})^2 - (\beta_{gk}^{[t-1]})^2}{2(\tau_{\beta L_{gk}^{[t]}}^{[t]})^2}).$$

8. Update $\nu_{bg}^{[t]}$ by an MH step with the symmetric proposal distribution $g(\nu_{bg}^{*}|\nu_{bg}^{[t-1]}) \sim N(\nu_{bg}^{[t-1]}, \sigma_{MH}^2)$ and the acceptance rate

$$\rho = \frac{L(\nu_{bg}^{*}|-)}{L(\nu_{bg}^{[t-1]}|-)}$$
$$= \prod_{i=1}^{n_b} exp((\nu_{bg}^{*} - \nu_{bg}^{[t-1]})x_{big}^{[t]})(\frac{\phi_{bg}^{[t-1]} + exp(\alpha_g^{[t]} + \beta_{gk}^{[t]} + \nu_{bg}^{[t-1]} + \delta_{bi}^{[t-1]})}{\phi_{bg}^{[t-1]} + exp(\alpha_g^{[t]} + \beta_{gk}^{[t]} + \nu_{bg}^{*} + \delta_{bi}^{[t-1]})})^{\phi_{bg}^{[t-1]} + x_{big}^{[t]}}$$
$$\cdot exp(-\frac{(\nu_{bg}^{*})^2 - (\nu_{bg}^{[t-1]})^2}{2\sigma_c^2}).$$

9. Update $\delta_{bi}^{[t]}$ by an MH step with the symmetric proposal distribution $g(\delta_{bi}^{*}|\delta_{bi}^{[t-1]}) \sim$

$N(\delta_{bi}^{[t-1]}, \sigma_{MH}^2)$ and the acceptance rate

$$\begin{aligned}\rho &= \frac{L(\delta_{bi}^*|-)}{L(\delta_{bi}^{[t-1]}|-)} \\ &= \prod_{g=1}^{G} exp((\delta_{bi}^* - \delta_{bi}^{[t-1]})x_{big}^{[t]})(\frac{\phi_{bg}^{[t-1]} + exp(\alpha_g^{[t]} + \beta_{gk}^{[t]} + \nu_{bg}^{[t]} + \delta_{bi}^{[t-1]})}{\phi_{bg}^{[t-1]} + exp(\alpha_g^{[t]} + \beta_{gk}^{[t]} + \nu_{bg}^{[t]} + \delta_{bi}^*)})^{\phi_{bg}^{[t-1]} + x_{big}^{[t]}} \\ &\cdot exp(-\frac{(\delta_{bi}^*)^2 - (\delta_{bi}^{[t-1]})^2}{2\sigma_d^2}).\end{aligned}$$

10. Update $\phi_{bg}^{[t]}$ by an MH step with the proposal distribution $g(\phi_{bg}^*|\phi_{bg}^{[t-1]}) \sim Gamma(\phi_{bg}^{[t-1]}, 1)$ and the acceptance rate

$$\begin{aligned}\rho &= \frac{L(\phi_{bg}^*)g(\phi_{bg}^{[t-1]}|\phi_{bg}^*)}{L(\phi_{bg}^{[t-1]})g(\phi_{bg}^*|\phi_{bg}^{[t-1]})} \\ &= \prod_{i=1}^{n_b}[\frac{\Gamma(\phi_{bg}^* + x_{big}^{[t]})(\phi_{bg}^*)^{\phi_{bg}^*}}{\Gamma(\phi_{bg}^*)(\phi_{bg}^* + \eta_{big}^{[t]})^{\phi_{bg}^* + x_{big}^{[t]}}} \cdot \frac{\Gamma(\phi_{bg}^{[t-1]})(\phi_{bg}^{[t-1]} + \eta_{big}^{[t]})^{\phi_{bg}^{[t-1]} + x_{big}^{[t]}}}{\Gamma(\phi_{bg}^{[t-1]} + x_{big}^{[t]})(\phi_{bg}^{[t-1]})^{\phi_{bg}^{[t-1]}}}] \\ &\cdot \frac{(\phi_{bg}^*)^{\kappa-1}}{(\phi_{bg}^{[t-1]})^{\kappa-1}} exp(-\tau(\phi_{bg}^* - \phi_{bg}^{[t-1]}))\frac{(\phi_{bg}^{[t-1]})^{\phi_{bg}^*-1}\Gamma(\phi_{bg}^{[t-1]})}{(\phi_{bg}^*)^{\phi_{bg}^{[t-1]}-1}\Gamma(\phi_{bg}^*)} exp(\phi_{bg}^* - \phi_{bg}^{[t-1]}),\end{aligned}$$

where $\eta_{big}^{[t]} = exp(\alpha_g^{[t]} + \beta_{gw_{bi}^{[t-1]}}^{[t]} + \nu_{bg}^{[t]} + \delta_{bi}^{[t]})$ denotes the mean gene expression level for gene g in cell i of batch b.

11. The conditional posterior distribution for the cell type indicator $w_{bi}^{[t]}$ of cell i in batch b is:

$$Pr(w_{bi}^{[t]} = k|-) \propto \pi_{bk}^{[t-1]} \prod_{g=1}^{G} \frac{exp[(\alpha_g^{[t]} + \beta_{gk}^{[t]} + \nu_{bg}^{[t]} + \delta_{bi}^{[t]})x_{big}^{[t]}]}{(exp(\alpha_g^{[t]} + \beta_{gk}^{[t]} + \nu_{bg}^{[t]} + \delta_{bi}^{[t]}) + \phi_{bg}^{[t]})^{x_{big}^{[t]} + \phi_{bg}^{[t]}}}.$$

We implement an MH step with the symmetric proposal distribution $Pr(w_{bi}^{[t]} = k^*|w_{bi}^{[t-1]} = k) \sim Multinomial(1; \frac{1}{K}, \cdots, \frac{1}{K})$ and the acceptance rate

$$\rho = \frac{Pr(w_{bi}^{[t]} = k^*|-)}{Pr(w_{bi}^{[t]} = k|-)} = \frac{\pi_{bk^*}^{[t-1]}}{\pi_{bk}^{[t-1]}} \prod_{g=1}^{G} exp[(\beta_{gk^*}^{[t]} - \beta_{gk}^{[t]})x_{big}^{[t]}](\frac{exp(\alpha_g^{[t]} + \beta_{gk}^{[t]} + \nu_{bg}^{[t]} + \delta_{bi}^{[t]}) + \phi_{bg}^{[t]}}{exp(\alpha_g^{[t]} + \beta_{gk^*}^{[t]} + \nu_{bg}^{[t]} + \delta_{bi}^{[t]}) + \phi_{bg}^{[t]}})^{\phi_{bg}^{[t]}}.$$

12. Update $\pi_b^{[t]}$ by sampling from the Dirichlet distribution

$$Dir(\xi + \sum_{i=1}^{n_b} 1(w_{bi}^{[t]} = 1), \xi + \sum_{i=1}^{n_b} 1(w_{bi}^{[t]} = 2), \cdots, \xi + \sum_{i=1}^{n_b} 1(w_{bi}^{[t]} = K)).$$

The Markov chain of the MCMC algorithm can get stuck in the local modes of the posterior distribution for a long period of time. In principle, we can further incorporate the Metropolis coupled MCMC algorithm [39] to jump out of the local modes more easily. In practice, we recommend running multiple chains with different initial values and then choosing the chain that gives the largest value of the observed data likelihood to conduct the posterior inference. According to our experiences, we can usually achieve good posterior estimations with five Markov chains each with a different initial value by randomly sampling a seed from 1 to 10,000.

Ranking	Gene Set Name	q-value	Category
1	KEGG_MATURITY_ONSET_DIABETES_OF_THE_YOUNG	1.19×10^{-7}	diabetes
2	KEGG_TIGHT_JUNCTION	9.32×10^{-7}	cell adhesion
3	KEGG_TYPE_II_DIABETES_MELLITUS	2.11×10^{-4}	diabetes
4	KEGG_ENDOCYTOSIS	8.11×10^{-4}	
5	KEGG_CHEMOKINE_SIGNALING_PATHWAY	9.86×10^{-4}	
6	KEGG_TGF_BETA_SIGNALING_PATHWAY	1.54×10^{-3}	
7	KEGG_LONG_TERM_DEPRESSION	1.80×10^{-3}	
8	KEGG_PURINE_METABOLISM	1.80×10^{-3}	metabolism
9	KEGG_ARGININE_AND_PROLINE_METABOLISM	1.80×10^{-3}	metabolism
10	KEGG_LYSOSOME	2.60×10^{-3}	
11	KEGG_RNA_DEGRADATION	2.60×10^{-3}	
12	KEGG_FOCAL_ADHESION	2.94×10^{-3}	cell adhesion
13	KEGG_GLYCOLYSIS_GLUCONEOGENESIS	3.02×10^{-3}	metabolism
14	KEGG_AXON_GUIDANCE	3.10×10^{-3}	
15	KEGG_ECM_RECEPTOR_INTERACTION	3.10×10^{-3}	cell adhesion
16	KEGG_GLYCINE_SERINE_AND_THREONINE_METABOLISM	3.72×10^{-3}	metabolism
17	KEGG_INSULIN_SIGNALING_PATHWAY	4.31×10^{-3}	insulin
18	KEGG_BUTANOATE_METABOLISM	5.17×10^{-3}	metabolism
19	KEGG_DRUG_METABOLISM_CYTOCHROME_P450	5.23×10^{-3}	metabolism
20	KEGG_TRYPTOPHAN_METABOLISM	9.98×10^{-3}	metabolism
21	KEGG_NICOTINATE_AND_NICOTINAMIDE_METABOLISM	1.01×10^{-2}	metabolism
22	KEGG_ALDOSTERONE_REGULATED_SODIUM_REABSORPTION	1.09×10^{-2}	
23	KEGG_TYROSINE_METABOLISM	1.09×10^{-2}	metabolism
24	KEGG_ERBB_SIGNALING_PATHWAY	1.19×10^{-2}	
25	KEGG_TYPE_I_DIABETES_MELLITUS	1.19×10^{-2}	diabetes
26	KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION	1.19×10^{-2}	metabolism
27	KEGG_VASCULAR_SMOOTH_MUSCLE_CONTRACTION	1.38×10^{-2}	
28	KEGG_HISTIDINE_METABOLISM	1.57×10^{-2}	metabolism
29	KEGG_O_GLYCAN_BIOSYNTHESIS	1.73×10^{-2}	
30	KEGG_GLUTATHIONE_METABOLISM	1.82×10^{-2}	metabolism

Supplementary Table S1: The 30 significantly enriched KEGG pathways (q-value [31] less than 0.025) according to intrinsic genes identified by BUSseq from the pancreas data.