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 (Bacher and Kendziorski, 2016). However, like the other high-throughput technologies (Irizarry et al., 2005; Leek et al., 2010; Taub et al., 2010), scRNA-seq experiments can suffer from severe batch effects (Hicks et al., 2018). 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 (Kharchenko et al., 2014). 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 (Bacher and Kendziorski, 2016). The cell type of each individual cell is unknown and is often the target of inference. Classic batch effects correction methods, such as Combat (Johnson et al., 2007) and SVA (Leek and Storey, 2007; Leek, 2014), are designed for bulk experiments and require knowledge of the subtype information of each sample a prior. For scRNA-seq data, this subtype information corresponds to the cell type of each individual cell. Clearly, these methods are thus infeasible for scRNA-seq data. Alternatively, if one has knowledge of a set of control genes whose expression levels are constant across cell types, then it is possible to apply RUV (Risso et al., 2014; Jacob et al., 2015). However, selecting control genes is often difficult for scRNA-seq experiments.

To jointly cluster samples across batches, Huo et al. (2016) 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) (Haghverdi et al., 2018) 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" (Haghverdi et al., 2018). These are very strong assumptions and cannot be validated at the design stage of the experiments. Scanorama (Hie et al., 2019) generalizes MNN, first reducing dimensions using randomized singular value decomposition (SVD) and then efficiently searching for nearest neighbors across all datasets using locality sensitive hashing. As Scanorama builds upon MNN, it relies on the same strong assumptions as MNN. Seurat adopts canonical correlation analysis (CCA) to identify shared variations across batches and treats them as shared cell types (Butler et al., 2018). The latest version of Seurat, Seurat 3.0, also applies MNN to the low dimensional representation learned by CCA to identify cells that are likely to belong to the same cell types in different batches (Stuart et al., 2019). LIGER (Welch et al., 2019) adopts integrative non-negative matrix factorization (iNMF), as opposed to CCA, to identify data-specific and shared factors. However, if some batches share certain technical noises, for example when each patient is measured by several batches, CCA and iNMF can mistake the technical variability as biological variability of interest. Recently, Luo and Wei (2019) developed BUS, a hierarchical model that is able to simultaneously cluster samples across multiple batches and correct for severe batch effects for microarray data. In addition, Luo and Wei (2019) mathematically showed flexible experimental designs under which batch effects can be corrected when subtype information is unknown. However, as is the case for MNN, Scanorama, Seurat and LIGER, BUS does not consider features unique to scRNA-seq data, such as the count nature of the data, over-dispersion (Vallejos et al., 2015), dropout events (Kharchenko et al., 2014), or cell-specific size factors (Wang et al., 2018).

ZIFA (Pierson and Yau, 2015) and ZINB-WaVE (Risso et al., 2018) are two factor models that account for dropout events. As factor models are only approximations to the true mixture distributions of distinct cell types, they lose statistical efficiency. scVI (Lopez et al., 2018) models the mean expression levels and dropout rates more flexibly via neural networks. However, none of these authors discuss the experimental designs under which their methods are applicable. Nevertheless, 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 (Luo and Wei, 2019) are also valid for scRNA-seq experiments. Specifically, in addition to the commonly advocated completely randomized design (Bacher and Kendziorski, 2016; Baran-Gale et al., 2017; Hicks et al., 2018; Dal and Di, 2018), in which each batch measures all cell types, it is also legitimate to conduct scRNA-seq experiments following the "reference panel" design and the "chain-type" design, which allow some cell types to be missing from some batches. We demonstrate that BUSseq outperforms the existing approaches in both simulation data and 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 cells each with a sample size of n_b, let us denote the 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

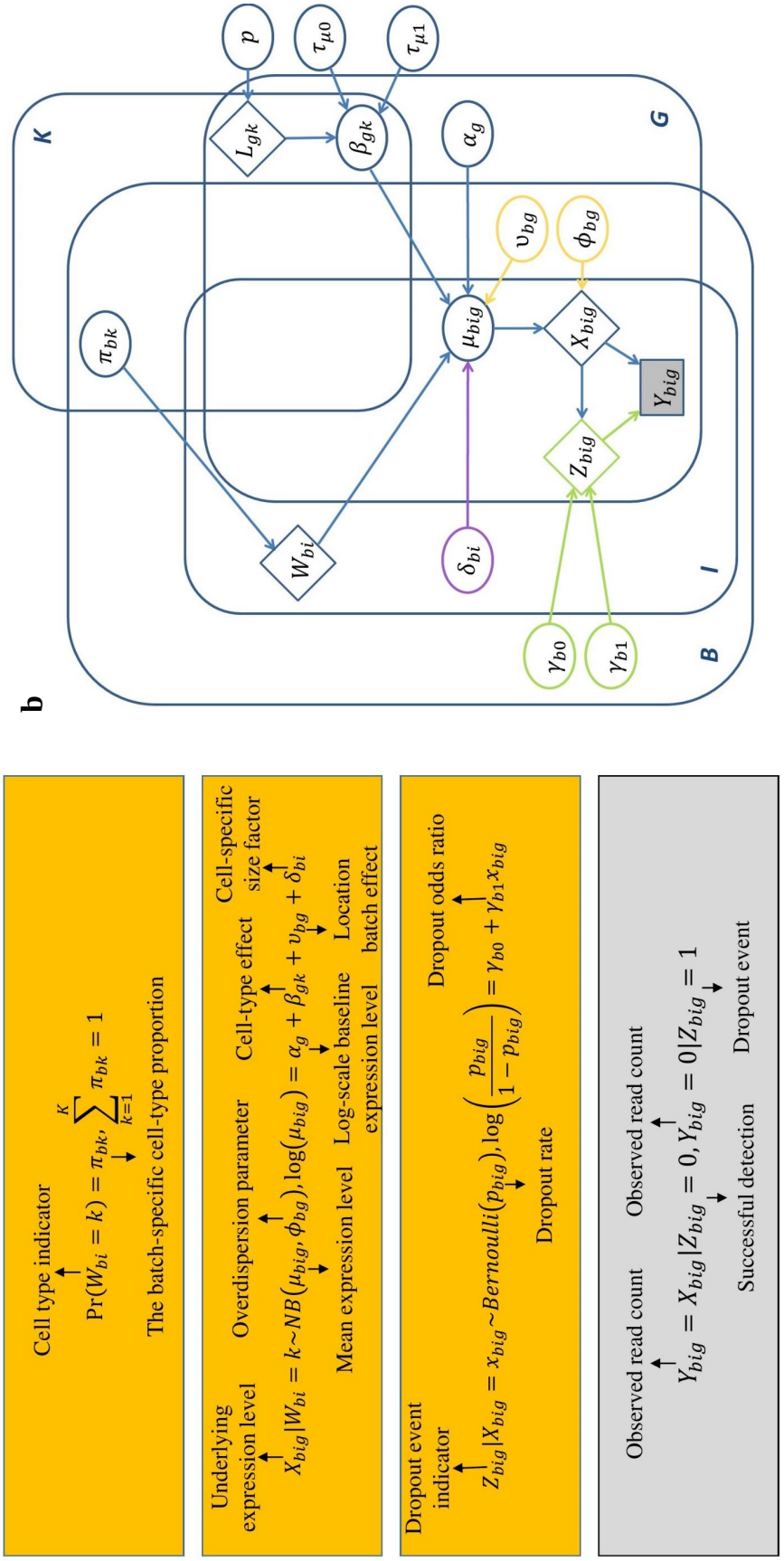

Figure 1: The graphical representation of the BUSseq model. **(a)** The hierarchical structure of BUSseq model. Only Y_{big} in the grey rectangle is observed. **(b)**The yellow color corresponds to batch effects; the green color models the dropout events; the purple color indicates the cell-specific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

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, it is not always possible to observe the expression level X_{big}. Without dropout ($Z_{big} = 0$), we can directly observe $Y_{big} = X_{big}$. However, if a dropout event occurs ($Z_{big} = 1$), then we observe $Y_{big} = 0$ instead of the true level $Y_{big} = X_{big}$. It has been noted that highly expressed genes are less-likely to suffer from dropout events (Kharchenko et al., 2014). We thus model the dependence of the dropout rate $P(Z_{big} = 1|X_{big})$ on the expression level using a logistic regression with batch-specific intercept γ_{b0} and odds ratio γ_{b1}. Noteworthy, BUSseq includes the negative binomial distribution without zero inflation as a special case. When all cells are from a single cell type and the cell-specific size factor δ_{bi} is estimated a priori according to spike-in genes, BUSseq can reduce to a form similar to BASiCS (Vallejos et al., 2015).

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 develop a Markov chain Monte Carlo (MCMC) algorithm (Robert and Casella, 2013). Based on the parameter estimates, we can learn the cell type for each individual cell, impute the missing underlying 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) (Schwarz et al., 1978). 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 can separate biological variability from technical artifacts, because different combinations of batch-effect and cell-type-effect values can lead to the same probabilistic distribution for the observed data, which in statistics is termed a *non-identifiable* model. Formally, a model is said to be *identifiable* if each probability distribution can arise from only one set of parameter values (Casella and Berger, 2002). Statistical inference is impossible for non-identifiable models because two sets of distinct parameter values can give rise to the same probabilistic function. We prove that the BUSseq model is identifiable under conditions that are very

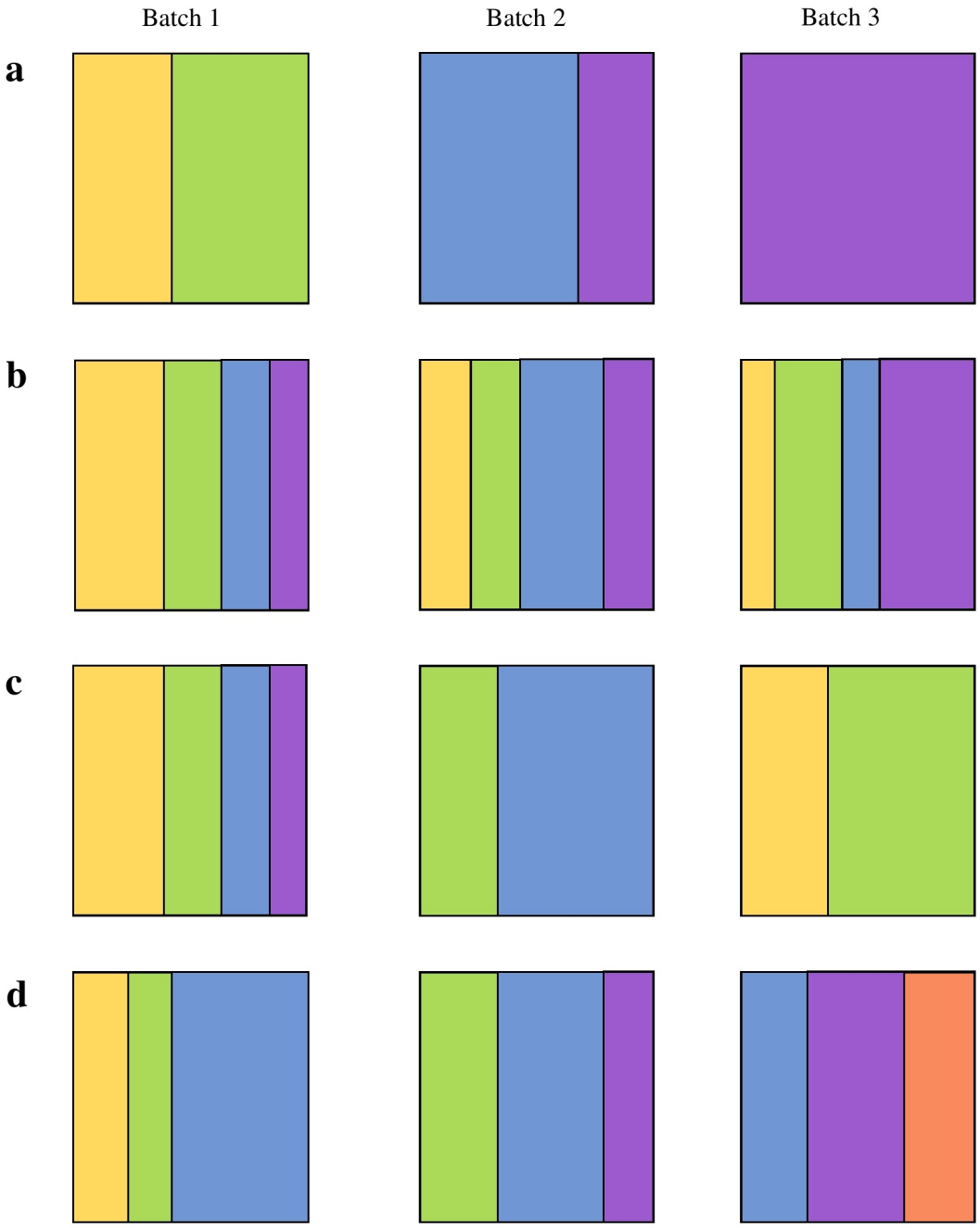

Figure 2: Various types of experimental designs. Each subpanel represents three batches with genes in rows and cells in columns; and each color indicates a cell type. **(a)** A confounded design. Batch 1 assays cells from cell types 1 and 2; batch 2 profiles cells from cell types 3 and 4; and batch 3 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 (see also **Figure S1**).

easily met. It is thus 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)**), 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 Information). Condition (I) requires that the highly expressed genes are less likely to have dropout events, which is routinely observed for scRNA-seq data (Kharchenko et al., 2014). 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 distinct cell-type pairs (k_1, k_2) and (k_3, k_4) their 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 almost 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 (Haghverdi et al., 2018). 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, it is always very challenging 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, 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 Information). 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(a)**). 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 Information).

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 requirement of the chain-type design.

The key insight is that despite batch effects, differences between cell types remain constant across batches. The differences between a pair of cell types allow us to distinguish batch effects from biological variability for those batches that measure both cell types. Once batch effects have been identified, we can conduct joint clustering across batches with batch effects removed. In fact, BUSseq can separate batch effects from cell type effects under more general designs beyond the easily understood and commonly encountered reference panel design and chain-type design. If we regard each batch as a node in a graph and connect two nodes with an edge if the two batches share at least two cell types, then BUSseq is identifiable as long as the resulting graph is connected (see Theorem 4 in Methods and its proof in Supplementary Information).

For scRNA-seq data, dropout rates depend on the 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. Miao et al. (2016) 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 (Luo and Wei, 2019) 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.

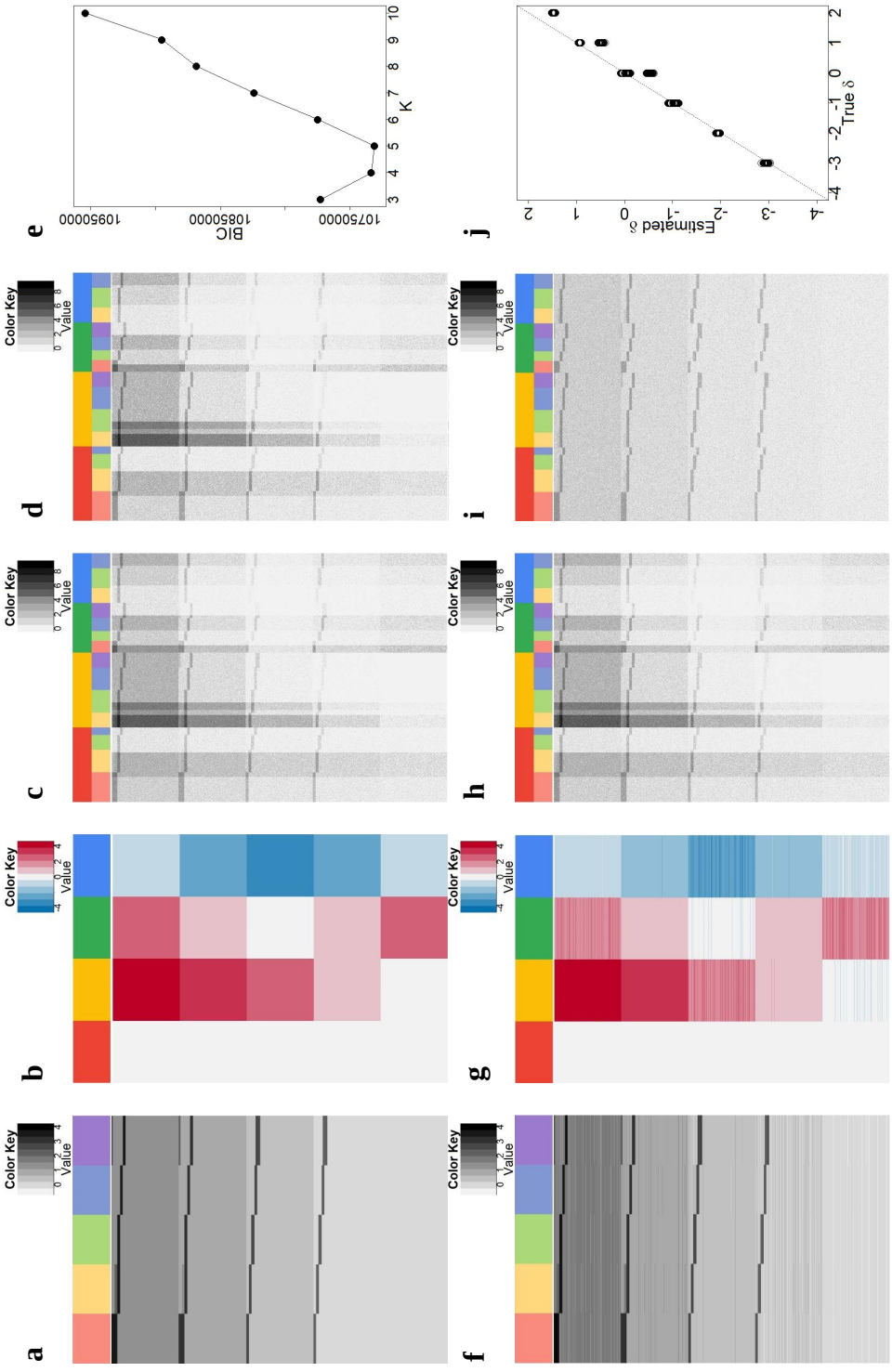

Figure 3: Patterns of the simulation study. **(a)** True log-scale mean expression level for each cell type $\boldsymbol{\alpha}+\boldsymbol{\beta}$. Each row represents a gene, and each column corresponds to a cell type. The intrinsic genes that are differentially expressed between cell types can have high, medium high, median low or low baseline expression levels. **(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 3,000 genes. The sample sizes for each batch are 300, 300, 200 and 200, respectively. **(d)** The simulated observed data $\boldsymbol{Y}$. The overall dropout rate is 27.3%, whereas the overall zero rate is 50.8%. **(e)** The BIC plot. The BIC attains the minimum at $K = 5$, identifying the true cell type number. **(f)** The estimated log-scale mean expression level for each cell type $\hat{\boldsymbol{\alpha}}+\hat{\boldsymbol{\beta}}$. **(g)** Estimated batch effects. **(h)** Imputed expression levels $\widehat{\boldsymbol{X}}$. **(i)** Corrected count data $\widetilde{\boldsymbol{X}}$ grouped by batches. **(j)** Scatter plot of the estimated cell-specific size factor versus the true cell-specific size factor.

BUSseq accurately estimates the parameters and imputes the missing data

We first evaluate the performance of BUSseq via a simulation study. We simulate 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. The sample sizes for each batch are $(n_1, n_2, n_3, n_4) = (300, 300, 200, 200)$, and there are a total of 3,000 genes. The magnitude of the batch effects, cell type effects, the dropout rates, and the cell-specific size factors are chosen to mimic real data scenarios (see **Figure 4(a)** and Data Availability). **Figure 3(d)** shows that the observed data suffer from severe batch effects and dropout events. This is also illustrated by the t-SNE plot (**Figure 4(c)**). The dropout rates for the four batches are $26.79\%, 24.53\%, 28.36\%$ and 31.29%, with the corresponding total zero proportions given by $44.13\%, 48.85\%, 53.07\%$ and 61.38%.

BUSseq correctly identifies the presence of five cell types among the cells (**Figure 3(e)**). Moreover, despite the dropout events, BUSseq accurately estimates the 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(j)**). When controlling the Bayesian False Discovery Rate (FDR) at 0.05 (Newton et al., 2004; Peterson et al., 2015), we identify all intrinsic genes that differentiate cell types with the true FDR being 0.020 (Methods). The total running time of BUSseq for a given cell type number K using 8 cores of two 3.4GHz Intel Gold 6128 processors was 1.01 hours.

In the simulation study, we know the underlying 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 underlying 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 (Johnson et al., 2007). 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)**).

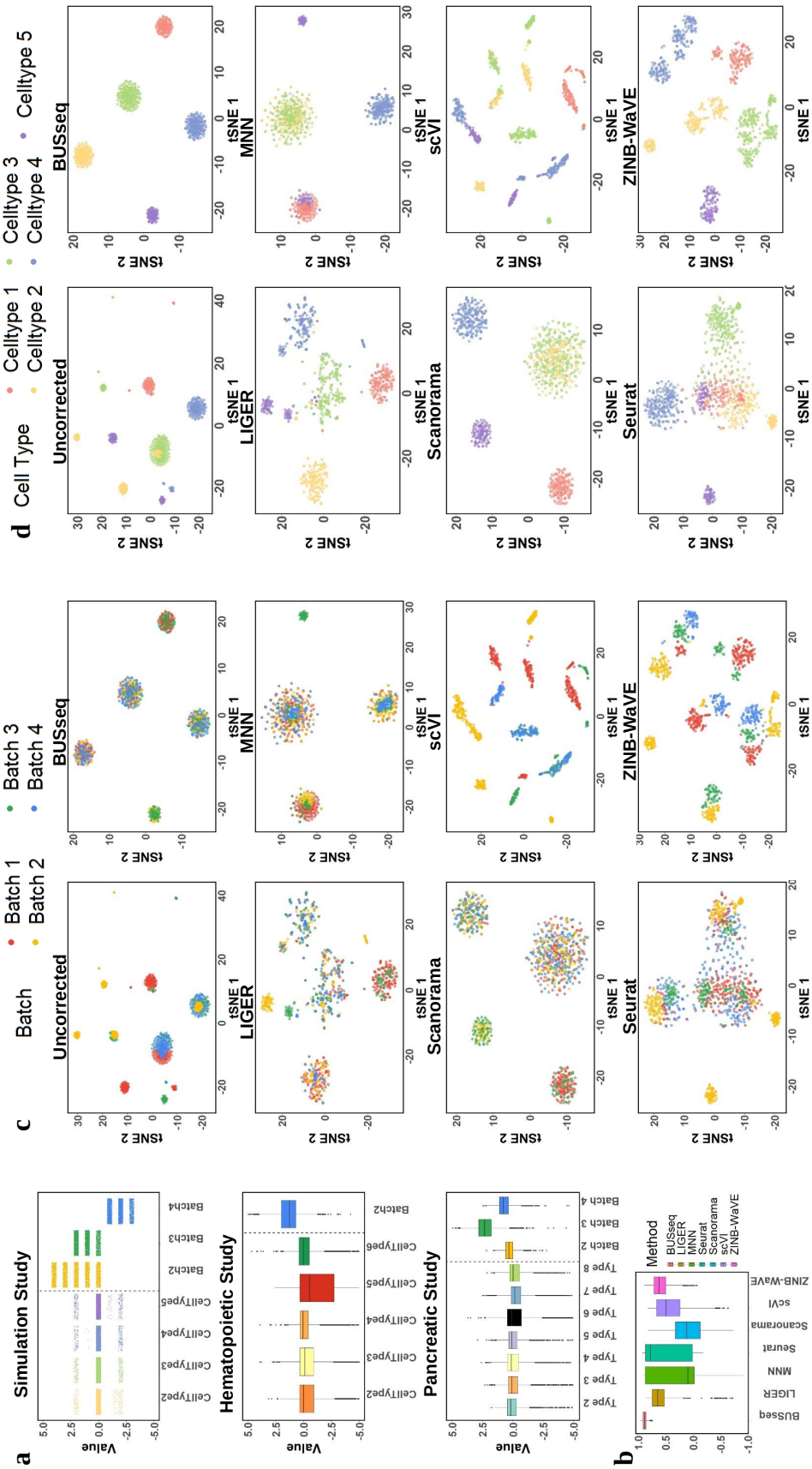

Figure 4: Comparison of batch effects correction methods in the simulation study. **(a)** Comparison of the magnitude of cell type effects and batch effects in the simulation study and the two real applications. The boxplots show the distributions of the estimated cell type effects $\widehat{\boldsymbol{\beta}}$ and batch effects $\widehat{\boldsymbol{\nu}}$ by BUSseq in the two real studies. The magnitude of the batch effects and cell type effects in the simulation study were chosen to mimic the real data scenarios. **(b)** The boxplots of silhouette coefficients for all compared methods. **(c)** T-distributed Stochastic Neighbor Embedding (t-SNE) plots colored by batch for each compared method. **(d)** t-SNE plots colored by true cell type labels for each compared method.

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-art methods for batch effects correction for scRNA-seq data—LIGER (Welch et al., 2019), MNN (Haghverdi et al., 2018), Scanorama (Hie et al., 2019), scVI (Lopez et al., 2018), Seurat (Stuart et al., 2019) and ZINB-WaVE (Risso et al., 2018). 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 0.837 for LIGER, 0.654 for MNN, 0.521 for Scanorama, 0.480 for scVI, 0.632 for Seurat and 0.571 for ZINB-WaVE. Moreover, the t-SNE plots (**Figure 4(c-d)**) show that only BUSseq can perfectly 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(b)** 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. (2018), one profiled by the SMART-seq2 protocol for a population of hematopoietic stem and progenitor cells (HSPC) from 12-week-old female mice (Nestorowa et al., 2016) 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 (Paul et al., 2015). Although the two datasets were generated in two different laboratories (**Figure 5(a)**), both datasets have cell-type label for each cell that is annotated according to the expression levels of marker genes

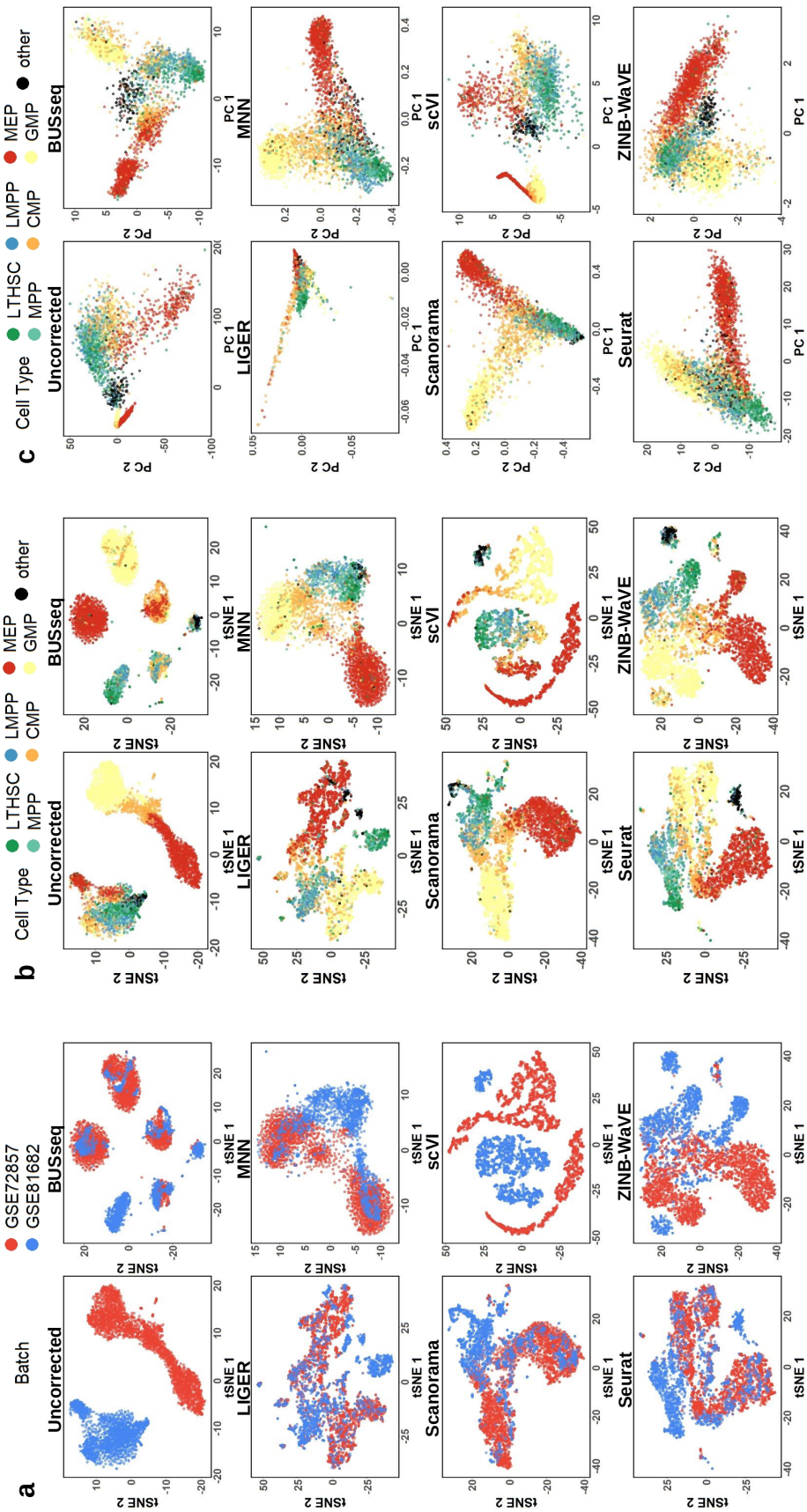

Figure 5: t-SNE and Principal Component Analysis (PCA) plots for the hematopoietic data. **(a)** t-SNE plots colored by batch. **(b)** t-SNE plots colored by FACS cell type labels. (c) PCA plots colored by FACS cell type labels.

(Paul et al., 2015; Haghverdi et al., 2018) from fluorescence-activated cell sorting (FACS) (Methods).

In order to compare BUSseq with existing methods, we compute the ARI between the clustering of each method and the FACS labels. The resulting ARIs are 0.582 for BUSseq, 0.307 for LIGER, 0.575 for MNN, 0.518 for Scanorama, 0.197 for scVI, 0.266 for Seurat and 0.348 for ZINB-WaVE. BUSseq thus outperforms all of the other methods in being consistent with FACS labeling. BUSseq also has Silhouette coefficients that are comparable to those of MNN, which are better than those of all the other methods (**Figure S2**). Furthermore, t-SNE plots confirm that BUSseq performs the best in segregating cells into different cell types (**Figure 5(b)**).

Specifically, BUSseq learns 6 cell types from the dataset. According to the FACS labels (Methods), Cluster 2, Cluster 5, and Cluster 6 correspond to the common myeloid progenitors (CMP), megakaryocyte-erythrocyte progenitors (MEP) and granulocyte-monocyte progenitors (GMP), respectively (**Figure 5(c) and Figure 6(a-c)**). Cluster 1 is composed of long-term hematopoietic stem and progenitor cells (LTHSC) and multi-potent progenitors (MPP). These are cells from the early stage of differentiation. Cluster 4 consists of a mixture of MEP and CMP, while Cluster 3 is dominated by cells labeled as "other". Comparison between the subpanel for BUSseq in **Figure 5(c)** and **Figure 6(b)** indicates that Cluster 4 are cells from an intermediate cell type between CMP and MEP. In particular, according to **Figure 6(e)**, the marker genes *Apoe* and *Gata2* are highly expressed in Cluster 4 but not in CMP (Cluster 2) and MEP (Cluster 6), and the marker gene *Ctse* is expressed in MEP (Cluster 6) but not in Cluster 4 and CMP (Cluster 2). Therefore, cells in Cluster 4 do form a unique group with distinct expression patterns. This intermediate cell stage between CMP and GMP is missed by all of the other methods considered. Moreover, we find that well known B-cell lineage genes (Herman et al., 2018), *Ebf1*, *Vpreb1*, *Vpreb3*, and *Igll1*, are highly expressed in Cluster 3, but not in the other clusters (**Figure 6 (c, e)**). To identify Cluster 3, which is dominated by cells labeled as "other" by Nestorowa et al. (2016), we map the mean expression profile of each cluster learned by BUSseq to the Haemopedia RNA-seq dataset (Choi et al., 2018). It turns out that Cluster 3 aligns well to common lymphoid progenitors (CLP) that give rise to T-lineage cells, B-lineage cells and natural killer cells (**Figure 6(d)**). Therefore, Cluster 3 represents cells that differentiate from lymphoid-primed multipotent progenitors (LMPP)

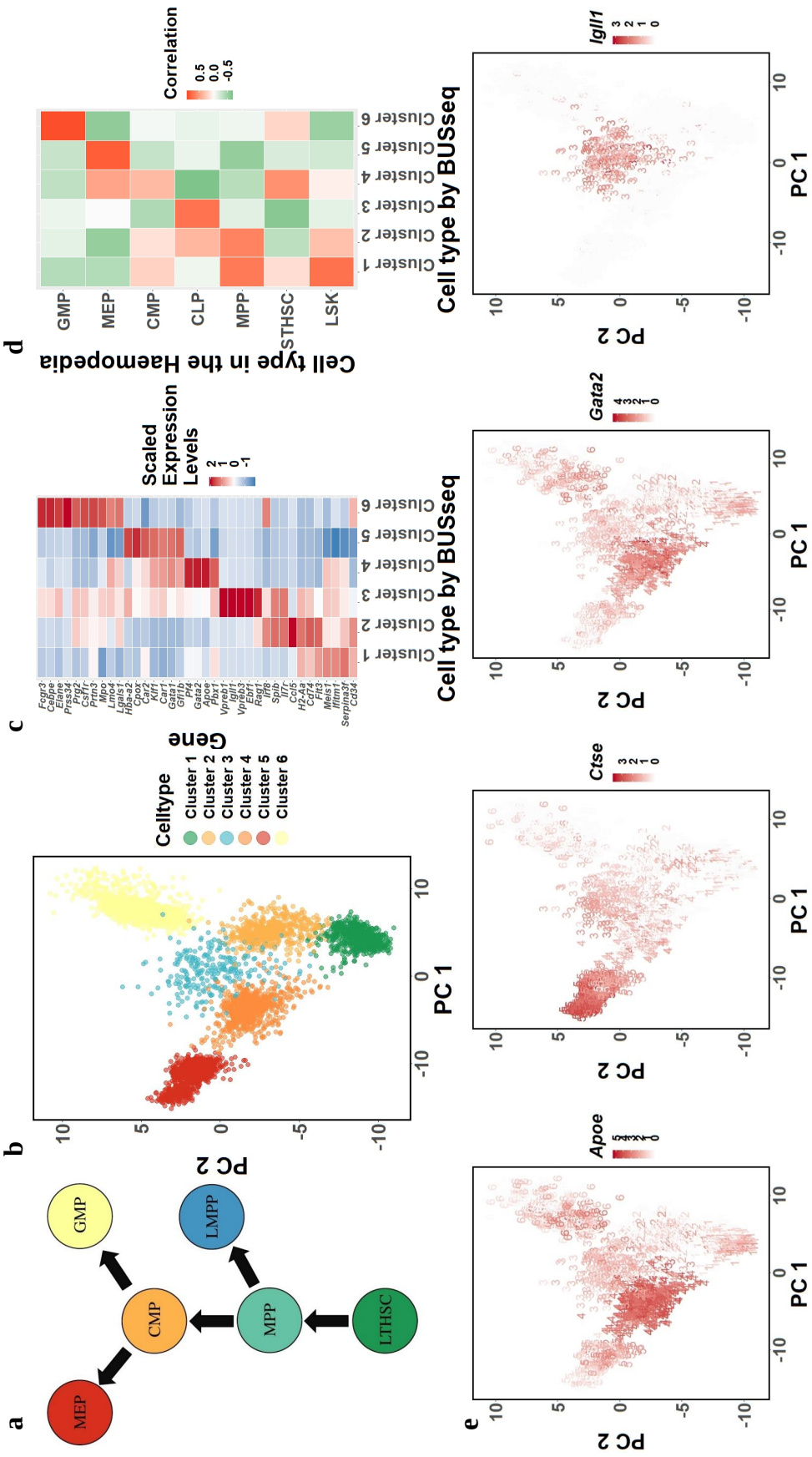

Figure 6: BUSseq preserve the hematopoietic stem and progenitor cells (HSPC) differentiation trajectories. **(a)** The diagram of HSPC differentiation trajectories. **(b)** The PCA plot of the corrected count matrix from BUSseq colored according to the estimated cell types by BUSseq. **(c)** The heatmap of scaled expression levels of key genes for HSPC. **(d)** The heatmap of correlation between gene expression profiles of each cell type inferred by BUSseq and those in the Haemopedia RNA-seq datasets. **(e)** The expression levels of four marker genes, *Apoe*, *Gata2*, *Ctse* and *Igll1*, shown in the PCA plots of corrected ount data by BUSseq, respectively. The digit labels denote the corresponding clusters identified by BUSseq.

(Paul et al., 2015). Once again, all the other methods fail to identify these cells as a separate group. Thus, although BUSseq does not assume any temporal ordering between cell types, it is able to preserve the differentiation trajectories (**Figure 6(a-b)**); although BUSseq assumes each cell belongs to one cell type, it is capable of capturing the subtle changes across cell types and within a cell type due to continuous processes such as development and differentiation.

We further inspect the functions of the intrinsic genes that distinguish different cell types. BUSseq detects 1419 intrinsic genes at the Bayesian FDR cutoff of 0.05 (Methods). The gene set enrichment analysis (Huang et al., 2009) shows that 51 KEGG pathways (Kanehisa and Goto, 2000) are enriched among the intrinsic genes (p-values < 0.05). The highest ranked pathway is the Hematopoietic Cell Lineage Pathway, which corresponds to the exact biological process studied in the two datasets. Among the remaining 50 pathways, thirteen are related to the immune system, and another nine are associated with cell growth and differentiation (Supplementary Table S1). Therefore, the pathway analysis demonstrates that BUSseq is able to capture the underlying true biological variability, even if the batch effects are severe, as shown in **Figure 4(a)** and **Figure 5(a)**.

BUSseq outperforms existing method on pancreas data

We further studied the four scRNA-seq datasets of human pancreas cells analyzed in Haghverdi et al. (2018), two profiled by CEL-seq2 protocol (Grün et al., 2016; Lawlor et al., 2017) and two assayed by SMART-seq2 protocol (Segerstolpe et al., 2016; Lawlor et al., 2017). 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. (2018).

For the two datasets profiled by the SMART-seq2 protocol, Segerstolpe et al. (2016) and Lawlor et al. (2017) provide cell-type labels; for the other two datastes assayed by the CEL-seq2 protocol, Haghverdi et al. (2018) provide the cell-type labels based on the marker genes in the original publications (Grün et al., 2016; Lawlor et al., 2017). We can thus compare the clustering results from each batch effects correction method with the labeled cell types (**Figure 7(a,b)**). 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

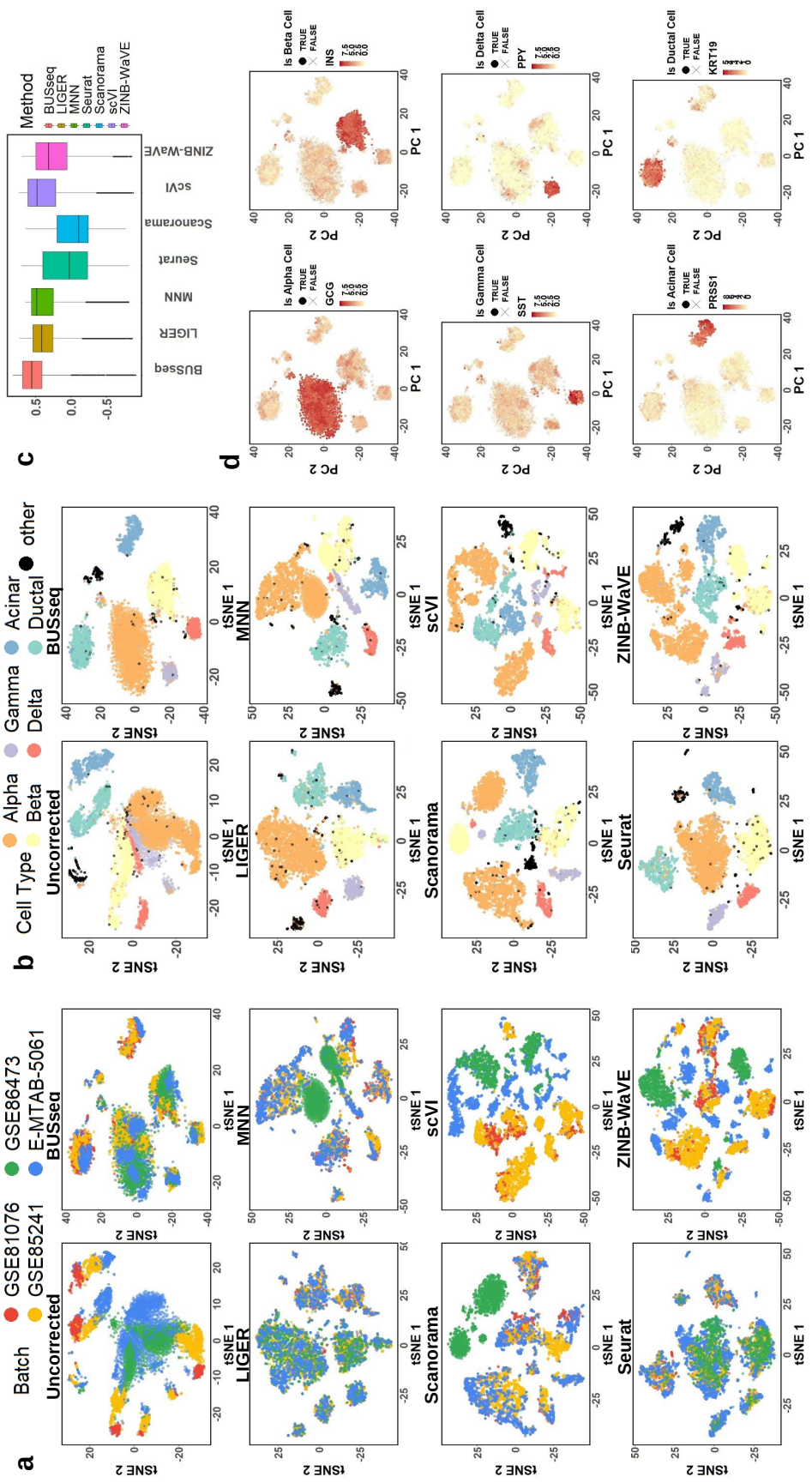

Figure 7: t-SNE plots for the pancreas data. **(a)** t-SNE plots colored by batch. **(b)** t-SNE plot colored by FACS cell type labels. **(c)** The boxplot of silhouette coefficients for all of the compared methods. **(d)** The expression levels of six marker genes, *GCG* for alpha cells, *INS* for beta cells, *SST* for gamma cells, *PPY* for delta cells, *PRSS1* for acinar cells, and *KRT19* for ductal cells, shown in the t-SNE plot of the corrected count data of BUSseq, respectively.

delta cells, while non-islet cells include acinar and ductal cells. BUSseq identifies a total of eight cell types: five for islet cells, two for non-islet cells and one for the labeled "other" cells. Specifically, the five islet cell types identified by BUSseq correspond to three groups of alpha cells, a group of beta cells, and a group of delta and gamma cells. The two non-islet cell types identified by BUSseq correspond exactly to the acinar and ductal cells. Compared to all of the other methods, BUSseq gives the best separation between islet and non-islet cells, as well as the best segregation within islet cells. In particular, the median silhouette coefficient by BUSseq is higher than that of any other method (**Figure 7(c)**).

The ARIs of all methods are 0.608 for BUSseq, 0.542 for LIGER, 0.279 for MNN, 0.527 for Scanorama, 0.282 for scVI, 0.287 for Seurat and 0.380 for ZINB-WaVE. Thus, BUSseq outperforms all of the other methods in being consistent with the cell-type labels according to marker genes. In **Figure 7(d)**, the locally high expression levels of marker genes for each cell type show that BUSseq correctly clusters cells according to their biological cell types.

BUSseq identifies 426 intrinsic genes at the Bayesian FDR cutoff of 0.05 (Methods). We conducted the gene set enrichment analysis (Huang et al., 2009) on the KEGG pathway database (Kanehisa and Goto, 2000). There are 14 enriched pathways (p-values < 0.05). Among them, three pathways are diabetes pathways; two are pancreatic and insulin secretion pathways; and another two pathways are related to metabolism (Supplementary Table S2). 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 have been corrected and the differential expressed genes can only be called after the cells have been clustered. The major issue with multi-stage methods is that uncertainties in the previous stages are often ignored. For instance, when cells have been first clustered into different cell types and then differential gene expression identification is conducted, the clustering results are taken as if they were the underlying truth. As the clustering results may be prone to errors in practice, this 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. BUSseq thus accounts for all uncertainties and fully exploits the information embedded in the data. As a result, BUSseq is able to capture more subtle changes between cell types, such as the cluster corresponding to LMPP lineage that is missed by all the state-of-the-art methods.

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.

Acknowledgment

We acknowledge Dr. Xiangyu Luo for helpful comments on an early version of our paper.

METHODS

BUSseq model

The hierarchical model of BUSseq can be summarized as:

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

Collectively, $\boldsymbol{Y}=\{Y_{big}\}_{b=1,\cdots,B;i=1,\cdots,n_b}^{g=1,\cdots,G}$ are the observed data; the underlying 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:

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

Consequently, the observed data likelihood function becomes

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

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

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 sets of 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 Information).

Theorem 1. *(The Complete Setting)*
If $\pi_{bk}>0$ for every batch b and cell type 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 distinct 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 exists 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. Both the reference panel design and the chain-type design belong to the more general connected design.

Theorem 4. *(The Connected Design)*
We define a batch graph $G = (V, E)$*. Each node* $b \in V$ *represents a batch. There is an edge* $e \in E$ *between two nodes* b_1 *and* b_2 *if and only if batches* b_1 *and* b_2 *share at least two cell types. If the batch graph is connected and conditions (I)-(III) hold, then BUSseq is identifiable (up to label switching).*

Statistical inference

We conduct the statistical inference under the Bayesian framework. We assign independent priors to each component of $\mathbf{\Theta}$ as follows: $\boldsymbol{\pi}_b = (\pi_{b1}, \cdots, \pi_{bK}) \sim Dirichlet(\xi, \cdots, \xi), 1 \leq b \leq B; \gamma_{b0} \sim N(0, \sigma^2_{z0}), 1 \leq b \leq B; -\gamma_{b1} \sim Gamma(a_\gamma, b_\gamma), 1 \leq b \leq B; \alpha_g \sim N(m_a, \sigma^2_a), 1 \leq g \leq G; \nu_{bg} \sim N(m_c, \sigma^2_c), 2 \leq b \leq B, g = 1, \cdots, G; \delta_{bi} \sim N(m_d, \sigma^2_d), 1 \leq b \leq B, 2 \leq i \leq n_b; \phi_{bg} \sim Gamma(\kappa, \tau), 1 \leq b \leq B, 1 \leq g \leq G.$

We are interested in detecting genes that differentiate cell types. Therefore, we impose a spike-and-slab prior (George and McCulloch, 1993) using a normal mixture to the cell-type effect β_{gk}. The spike component concentrates on zero with a small variance $\tau^2_{\beta 0}$, whereas the slab component tends to deviate from zero, thus having a larger variance $\tau^2_{\beta 1}$. 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. (2016), 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 Information). 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) (Newton et al., 2004; Peterson et al., 2015) defined as

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

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 (Peterson et al., 2015) 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.

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

Batch-effects-corrected values

To facilitate further downstream analysis, we also provide a version of count data $\widetilde{\boldsymbol{X}} = \{\widetilde{X}_{big}\}_{b=1,\cdots,B;i=1,\cdots,n_b}^{g=1,\cdots,G}$ 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 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 distri-

bution 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 $\widetilde{x}_{big}$.

The corrected data $\widetilde{\boldsymbol{X}}$ 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 $\widetilde{\boldsymbol{X}}$.

The benchmarked methods

To ensure a fair comparison, we follow the preprocessing steps of each of the methods used for benchmarking according to their original publications. LIGER (Welch et al., 2019) normalizes the raw read count of each cell by the cell's total read counts (see `https://github.com/MacoskoLab/liger`). MNN (Haghverdi et al., 2018) takes the first batch as the reference batch and normalizes the other batches to adjust for difference in sequencing depths (see `https://github.com/MarioniLab/MNN2017`). Scanorama (Hie et al., 2019) conducts L_2-normalization in the preprocessing steps (see `https://github.com/brianhie/scanorama`). Seurat (Stuart et al., 2019) log-transforms and scales the observed read count data (see `https://satijalab.org/seurat/`). scVI (Lopez et al., 2018) and ZINB-WaVE (Risso et al., 2018) directly work on the raw read count data (see `https://github.com/YosefLab/scVI` and `https://github.com/drisso/zinbwave`, respectively). For BUSseq, we run the MCMC algorithm for 4,000, 8,000 and 8,000 iterations for the simulated data, the hematopoietic study and the pancreas study, respectively. In each case, we treat the first half of all the iterations as the burn-in period and use the posterior samples collected from the second half for statistical inference. Please see `https://github.com/songfd2018/BUSseq-1.0_implementation` for the specification of hyperparameters used in this manuscript.

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. (2015) and the 2,729 cells labeled as myeloid progenitor cells by Nestorowa et al. (2016) from the NCBI Gene Expression Omnibus (GEO) with the accession numbers GSE72857 and GSE81682. Following Brennecke et al. (2013), 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 (Grün et al., 2016) and GSE86473 (Lawlor et al., 2017). The two datasets assayed by the SMART-seq2 platform were obtained from GSE85241 (Muraro et al., 2016) and from ArrayExpress accession number E-MATB-5061 (Segerstolpe et al., 2016). Following Haghverdi et al. (2018), 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. (2013) 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 labeled according to Lawlor et al. (2017) and Grün et al. (2016), 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.

Assignment of FACS cell type labels to learned clusters

In the two real data examples, we first identify the cell type of each individual cell according to FACS labeling. Then, for each cluster learned by BUSseq, we calculate the proportion of labeled cell types. If a cell type accounts for more than one-third of the cells in a given cluster, we assign this cell type to the cluster. Although a cluster may be assigned more than one cell type, most identified clusters by BUSseq are dominated by only one cell type.

Mapping clusters to Haemopedia

Haemopedia is a database of gene expression profiles from diverse types of haematopoietic cells (Choi et al., 2018). It collected flow sorted cell populations from healthy mice. To understand Cluster 3 learned by BUSseq for the hematopoietic data, which is dominated by cells classified as "other" according to the FACS labeling, we mapped the cluster means

learned by BUSseq to the Haemopedia RNA-seq dataset.

We first applied TMM normalization (Robinson and Oshlack, 2010) to all the samples in the Haemopedia RNA-seq dataset. Then, we extracted 7 types of hematopoietic stem and progenitor cells from Haemopedia, including $\text{Lin}^{-}\text{Sca-1}^{+}\text{c-Kit}^{+}$ (LSK) cells, short-term hematopoietic stem cells (STHSC), MPP, CLP, CMP, MEP and GMP. Each selected cell type had two RNA-seq samples in Haemopedia, so we averaged over the two replicates for each cell type. Further, we added one to the normalized expression levels as a pseudo read count to handle genes with zero read count and log-transformed the data. Finally, we scaled the data across the 7 cell types for each gene. To be comparable, we transformed the cluster mean learned by BUSseq as $m_{gk} = log(1 + exp(\alpha_g + \beta_{gk}))$ for gene g in the cluster k and scaled m_{gk} across all cell types as well. Finally, we calculated the correlation between the cluster means inferred by BUSseq and the reference expression profiles in Haemopedia for 37 marker genes. The 37 marker genes were retrieved from Paul et al. (2015) (31 maker genes for HSPC) and Herman et al. (2018) (6 maker genes for LMPP).

Silhouette coefficient

To evaluate the separation of different cell types after correction, we calculate the silhouette coefficient of each cell using the R package *cluster* (Kaufman and Rousseeuw, 2009). We regard each cell type, either the truth known in the simulation study or the labeling according to FACS, as a cluster. Let $a(i)$ be the average distance of cell i to all the other cells assigned to the same cluster as cell i, and let $b(i)$ be the average distance of cell i to all cells in the neighboring cluster, i.e., the cluster with the lowest average distance to cell i's cluster. The silhouette coefficient for cell i is defined as:

$$s(i) = \frac{b(i) - a(i)}{min(a(i), b(i))}$$

The silhouette coefficient $s(i)$ ranges from -1 to 1. The larger the values of $s(i)$, the closer cell i is to cells in the same cluster than cells in other clusters. We calculate the silhouette coefficient according to the t-SNE coordinates obtained from the corrected count data matrix (BUSseq and MNN) or from low-dimensional representations (LIGER, Scanorama, scVI, Seurat and ZINBWaVE).

Implementation of pathway analysis

To identify the biological functions of the intrinsic genes, we conduct gene set enrichment analysis for the intrinsic genes on KEGG pathways using DAVID (Huang et al., 2009). We control the Expression Analysis Systematic Explorer Score, a modified version of Fisher exact p-value, at the level of 0.05 to identify enriched pathways.

Software availability

The C++ source code of BUSseq is available on GitHub (`https://github.com/songfd2018/BUSseq-1.0`). All codes for producing results and figures in this manuscript are also available on Github (`https://github.com/songfd2018/BUSseq-1.0_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. (2016); MARS-seq platform hematopoietic data with GEO GSE72857 by Paul et al. (2015); CEL-seq platform pancreas data with GEO GSE81076 by Grün et al. (2016); CEL-seq2 platform pancreas data with GEO GSE85241 by Muraro et al. (2016); SMART-seq2 platform pancreas data with GEO GSE86473 by Lawlor et al. (2017); and SMART-seq2 platform pancreas data with ArrayExpress E-MTAB-5061 by Segerstolpe et al. (2016).

The parameter settings for the simulation study and the simulated data are available on Github (`https://github.com/songfd2018/BUSseq-1.0_implementation`).

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. analyzed the data. YY.W. conceived and supervised the study, developed the method and the proof, and wrote the paper.

Supplementary Information

Proofs for Theorem 1 to 4

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 $f_{k_1}^G = f^G(\boldsymbol{y}|\boldsymbol{\gamma}_{k_1}, \boldsymbol{\phi}_{k_1}, \boldsymbol{\mu}_{k_1}) \in \mathcal{F}^G$, $f_{k_2}^G = f^G(\boldsymbol{y}|\boldsymbol{\gamma}_{k_2}, \boldsymbol{\phi}_{k_2}, \boldsymbol{\mu}_{k_2}) \in \mathcal{F}^G$, there exist at least two dimensions g_1 and g_2 with $\mu_{g_1k_1} \neq \mu_{g_1k_2}$ and $\mu_{g_2k_1} \neq \mu_{g_2k_2}$, 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 $f_1^G, f_2^G \in \mathcal{F}^G$, $f_1^G \succeq f_2^G$ 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 $G_0 \in \{1, 2, \cdots, G\}$,

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

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

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

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

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

More specifically,

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

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

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

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

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

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

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

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 $M_{g+1}(f^{G-g}(\boldsymbol{y}_{-g}|\boldsymbol{\gamma},\boldsymbol{\phi}_{-g},\boldsymbol{p}_{-g}))=\Phi_{g+1}(t_{g+1},\boldsymbol{y}_{-(g+1)})=\sum_{y_{g+1}=1}^{\infty}f^{G-g}(\boldsymbol{y}_{-g}|\boldsymbol{\gamma},\boldsymbol{\phi}_{-g},\boldsymbol{p}_{-g})t_{g+1}^{y_{g+1}}=$

$\sum_{y_{g+1}=1}^{\infty}f^1(y_{g+1}|\gamma,\phi_{g+1},p_{g+1})t_{g+1}^{y_{g+1}}\cdot f^{G-(g+1)}(\boldsymbol{y}_{-(g+1)}|\boldsymbol{\gamma},\boldsymbol{\phi}_{-(g+1)},\boldsymbol{p}_{-(g+1)})$. 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 $r_{g+1,1}=p_{g+1,1},\psi_{g+1}\le\phi_{g+1}$. Similar to Equation (12), we have

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

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

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

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,

$$\sum_{k=1}^{K_{g+1}} \pi_k \frac{\Phi_{g+1,k}(t_{g+1}, \boldsymbol{y}_{-(g+1)})}{\Phi_{g+1,1}(t_{g+1}, \boldsymbol{y}_{-(g+1)})} = \sum_{l=1}^{L_{g+1}} \xi_l \frac{\Psi_{g+1,l}(t_{g+1}, \boldsymbol{y}_{-(g+1)})}{\Phi_{g+1,1}(t_{g+1}, \boldsymbol{y}_{-(g+1)})}$$

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

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 $\boldsymbol{Y} = \{\boldsymbol{Y}_b, 1 \leq b \leq B\}$ and $\boldsymbol{m}_{bik} = exp(\boldsymbol{\alpha} + \boldsymbol{\beta}_k + \boldsymbol{\nu}_b + \delta_{bi}\mathbf{1})$, 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 Teicher (1967), 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 $\pi_{bk} = \pi^*_{b\rho_b(k)}, \boldsymbol{\gamma}_b = \boldsymbol{\gamma}_b^*, \boldsymbol{\alpha} + \boldsymbol{\beta}_k + \boldsymbol{\nu}_b + \delta_{bi}\mathbf{1} = log(\boldsymbol{m}_{bik}) = log(\boldsymbol{m}^*_{bi\rho_b(k)}) = \boldsymbol{\alpha}^* + \boldsymbol{\beta}^*_{\rho_b(k)} + \boldsymbol{\nu}_b^* + \delta_{bi}^*\mathbf{1}$ and $\boldsymbol{\phi}_b = \boldsymbol{\phi}_b^*$ for a permutation ρ_b of $\{1, 2, \cdots, K\}$, where $\mathbf{1}$ denotes a vector of one with length G.

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 $\boldsymbol{\nu}_1 = \boldsymbol{\nu}^*_1 = \mathbf{0}$ 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

$$\boldsymbol{\alpha} + \boldsymbol{\beta}_k + \boldsymbol{\nu}_b = \boldsymbol{\alpha}^* + \boldsymbol{\beta}^*_{\rho(k)} + \boldsymbol{\nu}_b, \boldsymbol{\alpha} + \boldsymbol{\beta}_k + \boldsymbol{\nu}_b + \delta_{bi}\mathbf{1} = \boldsymbol{\alpha}^* + \boldsymbol{\beta}^*_{\rho(k)} + \boldsymbol{\nu}_b + \delta^*_{bi}\mathbf{1}. \tag{24}$$

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

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

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

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

$$\boldsymbol{\alpha}+\boldsymbol{\beta}_k=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_1(k)}$$
$$\boldsymbol{\alpha}+\boldsymbol{\beta}_k+\boldsymbol{\nu}_b=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_b(k)}+\boldsymbol{\nu}^*_b$$
$$\boldsymbol{\alpha}+\boldsymbol{\beta}_k+\boldsymbol{\nu}_b+\delta_{bi}\mathbf{1}=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_b(k)}+\boldsymbol{\nu}^*_b+\delta^*_{bi}\mathbf{1}.$$

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

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

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

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

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

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 $k \in C_b \bigcap C_{\tilde{b}}$:

$$\left.\begin{aligned}\boldsymbol{\alpha}+\boldsymbol{\beta}_k+\boldsymbol{\nu}_b+\delta_{bi}\mathbf{1}&=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_b(k)}+\boldsymbol{\nu}_b+\delta_{bi}\mathbf{1}\\ \boldsymbol{\alpha}+\boldsymbol{\beta}_k+\boldsymbol{\nu}_{\tilde{b}}+\delta_{\tilde{b}i}\mathbf{1}&=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_{\tilde{b}}(k)}+\boldsymbol{\nu}_{\tilde{b}}+\delta_{\tilde{b}i}\mathbf{1}\end{aligned}\right\}\Rightarrow \boldsymbol{\beta}^*_{\rho_b(k)}=\boldsymbol{\beta}^*_{\rho_{\tilde{b}}(k)}$$

Consequently, we have proven Theorem 3.

Theorem 4

Proof. Our object 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. At the same time, $\boldsymbol{\nu}_b=\boldsymbol{\nu}^*_b$ and $\delta_{bi}=\delta^*_{bi}$ for each cell i in batch b.

For any two connected batches (b_1, b_2), we have $|C_{b_1} \cap C_{b_2}| \geq 2$. Thus, for any two shared cell types $k_1, k_2 \in C_{b_1} \cap C_{b_2}$, we have

$$\left.\begin{aligned}\boldsymbol{\alpha}+\boldsymbol{\beta}_{k_1}+\boldsymbol{\nu}_{b_1}+\delta_{b_1,i}\mathbf{1}&=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_{b_1}(k_1)}+\boldsymbol{\nu}^*_{b_1}+\delta^*_{b_1,i}\mathbf{1},\\ \boldsymbol{\alpha}+\boldsymbol{\beta}_{k_2}+\boldsymbol{\nu}_{b_1}+\delta_{b_1,i}\mathbf{1}&=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_{b_1}(k_2)}+\boldsymbol{\nu}^*_{b_1}+\delta^*_{b_1,i}\mathbf{1},\end{aligned}\right\}\Rightarrow\boldsymbol{\beta}_{k_1}-\boldsymbol{\beta}_{k_2}=\boldsymbol{\beta}^*_{\rho_{b_1}(k_1)}-\boldsymbol{\beta}^*_{\rho_{b_1}(k_2)}$$

$$\left.\begin{aligned}\boldsymbol{\alpha}+\boldsymbol{\beta}_{k_1}+\boldsymbol{\nu}_{b_2}+\delta_{b_2,i}\mathbf{1}&=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_{b_2}(k_1)}+\boldsymbol{\nu}^*_{b_2}+\delta^*_{b_2,i}\mathbf{1},\\ \boldsymbol{\alpha}+\boldsymbol{\beta}_{k_2}+\boldsymbol{\nu}_{b_2}+\delta_{b_2,i}\mathbf{1}&=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_{b_2}(k_2)}+\boldsymbol{\nu}^*_{b_2}+\delta^*_{b_2,i}\mathbf{1}.\end{aligned}\right\}\Rightarrow\boldsymbol{\beta}_{k_1}-\boldsymbol{\beta}_{k_2}=\boldsymbol{\beta}^*_{\rho_{b_2}(k_1)}-\boldsymbol{\beta}^*_{\rho_{b_2}(k_2)}$$

Further, according to condition (III), $\boldsymbol{\beta}_{k_1}-\boldsymbol{\beta}_{k_2}=\boldsymbol{\beta}^*_{\rho_{b_1}(k_1)}-\boldsymbol{\beta}^*_{\rho_{b_1}(k_2)}=\boldsymbol{\beta}^*_{\rho_{b_2}(k_1)}-\boldsymbol{\beta}^*_{\rho_{b_2}(k_2)}$ implies that $\rho_{b_1}(k)=\rho_{b_2}(k)$ for $k \in C_{b_1} \bigcap C_{b_2}$.

Consequently, for a cell type $k \in C_{b_1} \bigcap C_{b_2}$ shared by two connected batches b_1 and b_2, similar to Equation (23), we have

$$\left.\begin{aligned}\boldsymbol{\alpha}+\boldsymbol{\beta}_k+\boldsymbol{\nu}_{b_1}&=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_{b_1}(k)}+\boldsymbol{\nu}^*_{b_1}\\ \boldsymbol{\alpha}+\boldsymbol{\beta}_k+\boldsymbol{\nu}_{b_2}&=\boldsymbol{\alpha}^*+\boldsymbol{\beta}^*_{\rho_{b_2}(k)}+\boldsymbol{\nu}^*_{b_2}\end{aligned}\right\}\Rightarrow\boldsymbol{\nu}_{b_1}-\boldsymbol{\nu}^*_{b_1}=\boldsymbol{\nu}_{b_2}-\boldsymbol{\nu}^*_{b_2}.$$

Because of the connectivity of the batch graph G, we can find a path $(1, b_1, b_2, \cdots, b_k, b), k \leq B-2$ between any batch b and the first batch in the batch graph G such that $\boldsymbol{\nu}_b-\boldsymbol{\nu}^*_b=\boldsymbol{\nu}_{b_k}-\boldsymbol{\nu}^*_{b_k}=\cdots=\boldsymbol{\nu}_{b_1}-\boldsymbol{\nu}^*_{b_1}=\boldsymbol{\nu}_1-\boldsymbol{\nu}^*_1$. Notice that $\boldsymbol{\nu}_1=\boldsymbol{\nu}^*_1=0$, so we have $\boldsymbol{\nu}_b=\boldsymbol{\nu}^*_b$. Moreover, similar to Equation (24), we have $\delta_{bi}=\delta^*_{bi}$ for each cell i in the 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 $k \in C_b \bigcap C_{\tilde{b}}$:

$$\left.\begin{aligned} \boldsymbol{\alpha} + \boldsymbol{\beta}_k + \boldsymbol{\nu}_b + \delta_{bi}\mathbf{1} &= \boldsymbol{\alpha}^* + \boldsymbol{\beta}^*_{\rho_b(k)} + \boldsymbol{\nu}_b + \delta_{bi}\mathbf{1} \\ \boldsymbol{\alpha} + \boldsymbol{\beta}_k + \boldsymbol{\nu}_{\tilde{b}} + \delta_{\tilde{b}i}\mathbf{1} &= \boldsymbol{\alpha}^* + \boldsymbol{\beta}^*_{\rho_{\tilde{b}}(k)} + \boldsymbol{\nu}_{\tilde{b}} + \delta_{\tilde{b}i}\mathbf{1} \end{aligned}\right\} \Rightarrow \boldsymbol{\beta}^*_{\rho_b(k)} = \boldsymbol{\beta}^*_{\rho_{\tilde{b}}(k)},$$

which implies that $\rho_b(k) = \rho_{\tilde{b}}(k)$ according to condition (II). Consequently, we have proven Theorem 4.

Diagram of batch graphs

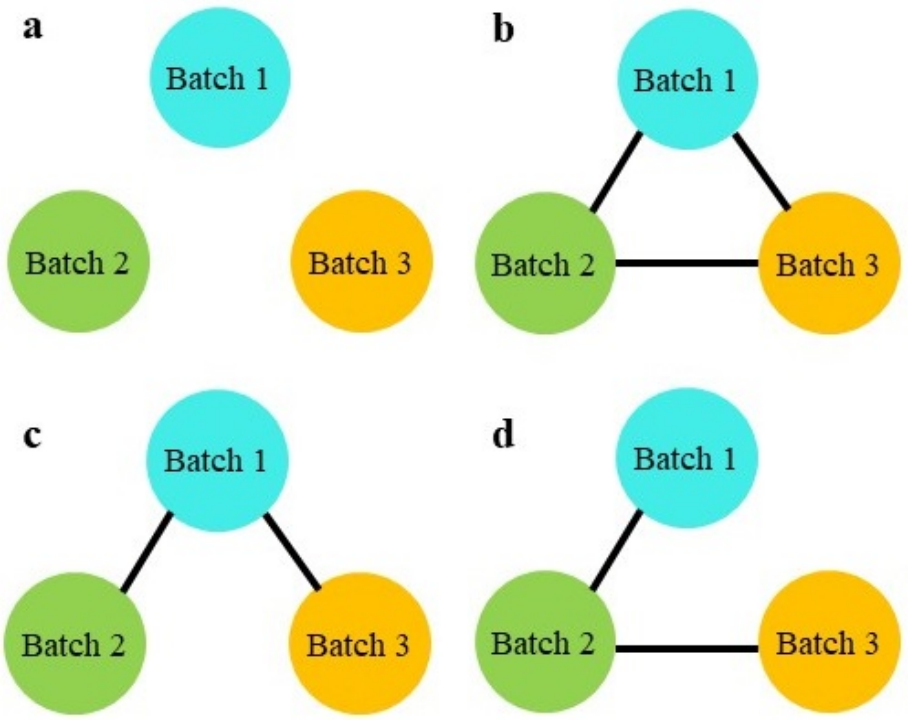

Figure S1: The batch graphs for the experiment designs in **Figure 2**. **(a)** The confounded design. **(b)** The complete setting design. **(c)** The reference panel design. **(d)** The chain-type design.

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 $z_{big}^{[t]}$ and $x_{big}^{[t]}$ sequentially for (b, i, g), if $y_{big} = 0$:

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

$$x_{big}^{[t]} \begin{cases} = 0 & , \text{if } z_{big}^{[t]} = 0; \\ \propto \frac{exp(\gamma_{b0}^{[t-1]}+\gamma_{b1}^{[t-1]}x_{big}^{[t]})}{1+exp(\gamma_{b0}^{[t-1]}+\gamma_{b1}^{[t-1]}x_{big}^{[t]})} \frac{\Gamma(\phi_{bg}^{[t-1]}+x_{big}^{[t]})(\mu_{big}^{[t-1]})^{x_{big}^{[t]}}}{\Gamma(x_{big}^{[t]})(\phi_{bg}^{[t-1]}+\mu_{big}^{[t-1]})^{\phi_{bg}^{[t-1]}+x_{big}^{[t]}}} & , \text{if } z_{big}^{[t]} = 1. \end{cases}$$

where $\mu_{big}^{[t-1]} = exp(\alpha_g^{[t-1]} + \beta_{gw_{bi}^{[t-1]}}^{[t-1]} + \nu_{bg}^{[t-1]} + \delta_{bi}^{[t-1]})$, and $\Gamma(\cdot)$ represents the Gamma function.

When $z_{big}^{[t]} = 1$, we incorporate a Metropolis-Hasting (MH) step (Hastings, 1970). We sample x_{big}^{*} from the proposal distribution $NB(\mu_{big}^{[t-1]}, \phi_{bg}^{[t-1]})$ and accept the proposal with probability

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

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

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

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

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

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

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

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

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:

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

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

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

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

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

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 (Altekar et al., 2004) 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.

Comparison of the silhouette coefficients in the hematopoietic study

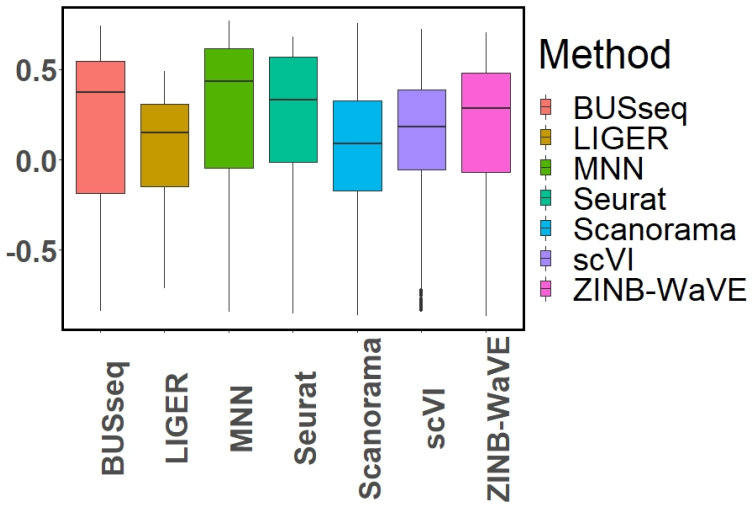

Figure S2: The boxplots of silhouette coefficients for all of the compared methods in the hematopoietic study.

Pathway analysis

Ranking	Pathway	p values	Category
1	Hematopoietic cell lineage	1.73×10^{-14}	
2	Cytokine-cytokine receptor interaction	1.84×10^{-12}	Cell growth and differentiation
3	Cell adhesion molecules (CAMs)	3.29×10^{-9}	Immune system
4	Leukocyte transendothelial migration	1.54×10^{-6}	Immune system
5	Primary immunodeficiency	6.75×10^{-6}	Immune system
6	Rap1 signaling pathway	3.44×10^{-5}	Cell growth and differentiation
7	Transcriptional misregulation in cancer	4.23×10^{-5}	
8	Rheumatoid arthritis	6.59×10^{-5}	
9	Pathways in cancer	1.15×10^{-4}	
10	Tuberculosis	1.40×10^{-4}	
11	Malaria	3.02×10^{-4}	
12	Toll-like receptor signaling pathway	3.60×10^{-4}	
13	Staphylococcus aureus infection	4.53×10^{-4}	
14	PI3K-Akt signaling pathway	5.53×10^{-4}	Cell growth and differentiation
15	Osteoclast differentiation	9.74×10^{-4}	Cell growth and differentiation
16	T cell receptor signaling pathway	9.96×10^{-4}	Immune system
17	Intestinal immune network for IgA production	1.43×10^{-3}	Immune system
18	Leishmaniasis	1.44×10^{-3}	Immune system
19	Platelet activation	1.62×10^{-3}	
20	NF-kappa B signaling pathway	1.65×10^{-3}	Immune system
21	Asthma	2.13×10^{-3}	
22	Jak-STAT signaling pathway	2.61×10^{-3}	Cell growth and differentiation
23	B cell receptor signaling pathway	3.34×10^{-3}	Immune system
24	ECM-receptor interaction	3.99×10^{-3}	Cell growth and differentiation
25	Neuroactive ligand-receptor interaction	4.05×10^{-3}	
26	Ras signaling pathway	4.93×10^{-3}	Cell growth and differentiation
27	Pertussis	5.48×10^{-3}	
28	Inflammatory bowel disease (IBD)	6.51×10^{-3}	Immune system
29	Thyroid hormone signaling pathway	9.06×10^{-3}	
30	Phagosome	9.51×10^{-3}	Immune system
31	Mineral absorption	1.09×10^{-2}	
32	Amoebiasis	1.17×10^{-2}	
33	Focal adhesion	1.43×10^{-2}	Cell growth and differentiation
34	Glycosphingolipid biosynthesis - lacto and neolacto series	1.49×10^{-2}	
35	p53 signaling pathway	1.67×10^{-2}	
36	Calcium signaling pathway	1.70×10^{-2}	
37	Fc epsilon RI signaling pathway	1.86×10^{-2}	
38	Natural killer cell mediated cytotoxicity	1.98×10^{-2}	Immune system
39	Proteoglycans in cancer	2.02×10^{-2}	
40	Chemokine signaling pathway	2.40×10^{-2}	Immune system
41	Gastric acid secretion	2.74×10^{-2}	
42	ABC transporters	2.82×10^{-2}	
43	HIF-1 signaling pathway	3.25×10^{-2}	
44	Chagas disease (American trypanosomiasis)	3.50×10^{-2}	
45	Retrograde endocannabinoid signaling	3.50×10^{-2}	
46	NOD-like receptor signaling pathway	3.63×10^{-2}	Immune system
47	Aldosterone-regulated sodium reabsorption	3.77×10^{-2}	
48	Sphingolipid signaling pathway	3.87×10^{-2}	
49	Progesterone-mediated oocyte maturation	4.41×10^{-2}	
50	MAPK signaling pathway	4.58×10^{-2}	Cell growth and differentiation
51	Carbohydrate digestion and absorption	4.76×10^{-2}	

Supplementary Table S1: The 51 KEGG pathways (p-value < 0.05 (Huang et al., 2009)) significantly enriched among the intrinsic genes identified by BUSseq from the hematopoietic data.

Ranking	Pathway	p values	
1	Maturity onset diabetes of the young	9.09×10^{-9}	Diabetes
2	Pancreatic secretion	6.42×10^{-7}	Protein Secretion
3	Insulin secretion	1.58×10^{-6}	Protein Secretion
4	Protein digestion and absorption	1.89×10^{-3}	Metabolism
5	ECM-receptor interaction	7.08×10^{-3}	
6	Type II diabetes mellitus	7.63×10^{-3}	Diabetes
7	Morphine addiction	8.99×10^{-3}	
8	Proteoglycans in cancer	1.44×10^{-2}	
9	Dopaminergic synapse	1.76×10^{-2}	
10	GABAergic synapse	2.24×10^{-2}	
11	Type I diabetes mellitus	2.26×10^{-2}	Diabetes
12	Tight junction	2.48×10^{-2}	
13	Drug metabolism - cytochrome P450	3.08×10^{-2}	Metabolism
14	Focal adhesion	4.07×10^{-2}	

Supplementary Table S2: The 14 KEGG pathways (p-value$<$ 0.05 (Huang et al., 2009)) significantly enriched among the intrinsic genes identified by BUSseq from the pancreatic data.

References

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

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.

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.

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.

S. C. Hicks, F. W. Townes, M. Teng, and R. A. Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4):562–578, 2018.

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

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.

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

Jeffrey T Leek. Svaseq: removing batch effects and other unwanted noise from sequencing

data. *Nucleic Acids Research*, 42(21):e161–e161, 2014.

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.

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.

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.

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.

Brian Hie, Bryan Bryson, and Bonnie Berger. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology*, 37(6):685, 2019.

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.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 2019.

Joshua D Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z Macosko. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887.e17, 2019.

Xiangyu Luo and Yingying Wei. Batch effects correction with unknown subtypes. *Journal of the American Statistical Association*, 114(526):581–594, 2019.

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

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.

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

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.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053, 2018.

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

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.

Christian Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Science & Business Media, 2013.

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

George Casella and Roger L Berger. *Statistical Inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

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.

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.

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.

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*, 128(8):e20–31, 2016.

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.

Josip S Herman, Dominic Grün, et al. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nature Methods*, 15(5):379, 2018.

Jarny Choi, Tracey M Baldwin, Mae Wong, Jessica E Bolden, Kirsten A Fairfax, Erin C Lucas, Rebecca Cole, Christine Biben, Clare Morgan, Kerry A Ramsay, et al. Haemopedia

RNA-seq: a database of gene expression during haematopoiesis in mice and humans. *Nucleic Acids Research*, 47(D1):D780–D785, 2018.

Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1): 44, 2009.

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

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.

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.

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.

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

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.

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.

Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010.

Leonard Kaufman and Peter J Rousseeuw. *Finding Groups in Data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

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

W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications.

Biometrika, 57(1):97–109, 1970.

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.