OXFORD

Briefings in Bioinformatics, 2022, **23(2)**, 1–17

https://doi.org/10.1093/bib/bbab565

Review

Comprehensive evaluation of noise reduction methods for single-cell RNA sequencing data

Shih-Kai Chu, Shilin Zhao, Yu Shyr and Qi Liu

Corresponding authors: Qi Liu, 2220 Pierce Avenue, 497A PRB, Nashville, TN 37232, USA. Tel.: +1-615-322-6618; Fax: +1-615-936-2602. E-mail: qi.liu@vumc.org; Yu Shyr, 2525 West End Avenue, Suite 1100, Rm 11132, Nashville, TN 37203, USA. Tel.: +1-615-936-6760; Fax: +1-615-936-2602. E-mail: yu.shyr@vumc.org

Abstract

Normalization and batch correction are critical steps in processing single-cell RNA sequencing (scRNA-seq) data, which remove technical effects and systematic biases to unmask biological signals of interest. Although a number of computational methods have been developed, there is no guidance for choosing appropriate procedures in different scenarios. In this study, we assessed the performance of 28 scRNA-seq noise reduction procedures in 55 scenarios using simulated and real datasets. The scenarios accounted for multiple biological and technical factors that greatly affect the denoising performance, including relative magnitude of batch effects, the extent of cell population imbalance, the complexity of cell group structures, the proportion and the similarity of nonoverlapping cell populations, dropout rates and variable library sizes. We used multiple quantitative metrics and visualization of low-dimensional cell embeddings to evaluate the performance on batch mixing while preserving the original cell group and gene structures. Based on our results, we specified technical or biological factors affecting the performance of each method and recommended proper methods in different scenarios. In addition, we highlighted one challenging scenario where most methods failed and resulted in overcorrection. Our studies not only provided a comprehensive guideline for selecting suitable noise reduction procedures but also pointed out unsolved issues in the field, especially the urgent need of developing metrics for assessing batch correction on imperceptible cell-type mixing.

Keywords: bioinformatics, single-cell RNA sequencing, normalization, batch effect adjustment

Introduction

As a powerful technique to profile gene expressions of thousands to millions of cells simultaneously at a cellular resolution, single-cell RNA sequencing (scRNA-seq) has been widely used to characterize cellular heterogeneity [1–3], reconstruct developmental trajectory [4–6] and improve our understanding of human disease [7–9]. Despite recent advances in technologies, scRNA-seq data show high technical variability resulting from sequencing depth, amplification bias, RNA capture efficiency and dropout events. These technical factors introduce substantial noise, making gene-level or cell-level expression incomparable even within one individual dataset. Additionally, scRNA-seq generated in different laboratories, at different times, by different platforms have large technical variations, making datasets incomparable and difficult for integration. Those unwanted variabilities introduced by technical factors confound biological signals of interest, which complicate the downstream analysis and result in false interpretations.

Normalization and batch correction are two important procedures to remove technical noises while preserving true biological variations. Normalization methods aim to adjust the influence of technical factors on gene counts within an individual dataset. They fall into two groups. One group is to infer cell-specific normalization factors, assuming all genes in one cell are subject to the same technical biases. The simplest approach is to scale the counts by sequencing depths, i.e. library size normalization (ls). BASiCS uses spike in to infer cell-specific normalization factors [10]. Scran groups cells by their library sizes to form cell pools, estimates the pool-specific scaling factors and then obtains cell-specific scaling factors by solving a linear equation system [11]. Scran reduces the effect of dropout events by pooling cells, increases the robustness of normalization and weakens the non-DE assumption. The other group of methods suggests that genes in one cell are affected unequally by technical factors, meaning that cell-specific factor is insufficient. They model the relationship between molecular counts

Shih-Kai Chu is a postdoctoral researcher in the Department of Biostatistics at Vanderbilt University Medical Center. His research focuses on single cell transcriptomics analysis.

Shilin Zhao is an assistant professor in the Department of Biostatistics at Vanderbilt University Medical Center. His research focuses on statistical methods in high dimensional data.

Yu Shyr is a professor and director in the Department of Biostatistics at Vanderbilt University Medical Center. His research interests include statistical and bioinformatics methodology and adaptive clinical trials.

Qi Liu is an associate professor in the Department of Biostatistics at Vanderbilt University Medical Center. Her research interests include single cell omics integration, and bioinformatics approaches to study complex disease.

Received: August 16, 2021. **Revised:** November 24, 2021. **Accepted:** December 11, 2021

and sequencing depth for every gene and infer normalization factors to adjust for the count–depth relationship. For example, SCnorm uses quantile regression to estimate the count–depth relationship for every gene, groups genes with similar dependence and then estimates normalization factors within each group [12]. Sctransform, as another example, models count–depth relationship for every gene by negative binomial regression, regularizes model parameters by a kernel smoother and directly predicts normalized count from the residuals of the model [13].

Batch correction methods seek to eliminate systematic differences across scRNA-seq datasets from multiple experiments, laboratories and platforms, enabling efficient integration of heterogeneous single-cell transcriptomics. Some methods are borrowed from bulk RNA-seq analysis, such as limma [14] and ComBat [15]. They model the linear relationship between batch and gene expression based on the Gaussian-distribution assumption. To handle highly sparse and over-dispersed scRNA-seq data, ZINB-WaVE extends the linear model based on a zero-inflated negative binomial distribution [16]. These linear-based methods assume that transcriptomics differences between batches all attribute to technical factors that could be modeled and regressed out. In real practices, however, cell populational compositions contribute to transcriptomics shift as well, which are usually unknown and not identical across batches. Without modeling populational compositions as covariates, the estimated coefficient for the batch factor contains biological components, resulting in overcorrection. To account for populational composition difference across scRNA-seq studies, methods have been developed to define shared cell types across batches by nearest neighbor (NN) or mutual nearest neighbors (MNN), such as fastMNN [17], Seurat [18], scMerge [19], Scanorama [20] and BBKNN [21]. The expression differences between cells from the same cell type but different batches are then used to estimate the batch effect. To enable shared cell population identification, NN-based methods project cells into a common reduced dimensional embedding by principal component analysis (PCA), non-negative matrix factorization (NMF) or canonical correlation analysis (CCA). For example, fastMNN finds mutual nearest pairs in the low-dimensional space calculated from PCA [17]. Seurat MultiCCA employs CCA to find a common embedding, which is further used to identify MNNs as anchor points [18]. scMerge combines linear modeling and MNN search to remove unwanted variations across batches while preserving the biological signal [19]. In addition to NN-based methods, Harmony formulates an objective function to balance cell-type clustering and degree of dataset mixing in the PCA space [22], which is fast for large-scale datasets [23]. LIGER uses integrative NMF to jointly define cell types from multiple single-cell datasets by calculating shared and dataset-specific metagenes [24]. Recently, deep neural networks were applied to model library sizes and batch effect biases for single-cell RNA-seq denoising [25–27]. For example, scVI uses autoencoder to reduce the high-dimensional gene-expression matrix to a lower-dimensional representation, which can be interpreted for relevant biology and can be used for clustering and visualization [25].

Most methods target either normalization or batch correction, which need to be combined with others for a complete denoising procedure. Several methods, in contrast, handle normalization and batch correction together, such as ZINB-WaVE [16], scMerge [19] and scVI [25], returning normalized and batch-corrected data. Since each method comes with its own strengths and weaknesses and works well under certain scenarios, it is very difficult to choose appropriate denoising procedures. In this study, we performed a comprehensive evaluation of 28 denoising procedures on 55 scenarios using both simulated and real datasets. These scenarios account for a variety of technical and biological factors, including relative magnitude of batch effect, the extent of cell population imbalance, the complexity of cell group structures, the proportion and the similarity of nonoverlapping cell populations, dropout rates and variable library sizes. We evaluated the performance of methods in different scenarios by batch mixing along with cell and gene structure preserving. This study not only provides a guidance to choose appropriate methods in different scenarios but also points out the challenge scenario where most methods failed. Additionally, the result suggested that novel metrics on correct batch mixing is needed for assessing batch correction.

Results

Evaluation on 28 denoising procedures in 55 scenarios

We evaluated the performance of 28 denoising procedures in terms of their ability to remove unwanted technical variations while preserving biological signals of interest. The 28 denoising procedures consisted of 24 combinative steps of 4 normalization (ls, Scran, SCnorm and sctransform) and 6 batch correction methods (limma, ComBat, fastMNN, Scanorama, Seurat MultiCCA V3 and Harmony), 3 approaches that address both normalization and batch correction simultaneously (ZINB-WaVE, scMerge and scVI), and LIGER with its custom preprocessing (Figure 1). We created 55 scenarios involving multiple technical and biological factors, including relative magnitude of batch effects, the extent of cell population imbalance, simple or complicated cell group structures, the proportion and the similarity of nonoverlapping cell populations, dropout rates and variable library sizes. The magnitude of batch effects varied from none, mild to strong biases. The cell populational composition ranged from balance, mild to severe imbalance. In imbalanced settings, we not only designed scenarios with identical cell types, but different proportions across batches, but also those with nonoverlapping cell types. The nonoverlapping was

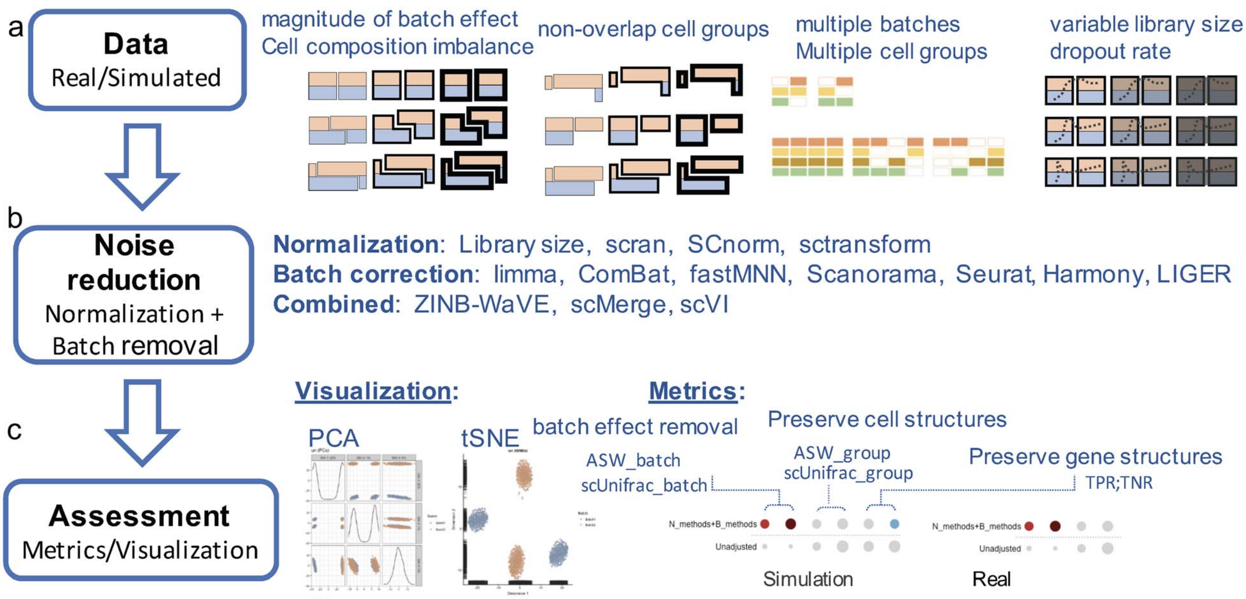

Figure 1. The evaluation workflow. (**A**) Studied scenarios and datasets. (**B**) Normalization and batch correction methods included in this study. (**C**) Adjustment performances assessed by both visualization and quantitative metrics. Quantitative metrics were summarized by a circle plot. Each circle in the plot represented the evaluation result of a procedure (row) measured by a certain metric (column). The size of the circle was determined by the metric score. The color of the circle was determined by the relative change of the score compared with the baseline (unadjusted). Red suggests an increase, while blue indicates a decrease and gray means unchanged in scores. The darker the circle, the more improved or worsened are the scores.

either one rare/dominant cell type just in one batch, or two similar/distinct cell types from two batches. Simple cell group structures comprised only two distinct cell groups, while complicated ones had multiple distinct cell groups and similar subgroups. We evaluated the performance by six quantitative metrics on batch mixing, cell and gene structure preservation in conjunction with PCA and tSNE visualizations. Average silhouette width_batch (ASW_batch) and scUnifrac_batch measure the degree of batch mixing. ASW_group and scUnifrac_group assess the cell structure preservation. True positive rate (TPR) and true negative rate (TNR) indicate the gene structure preservation by calculating the percentage of true marker (TP) and non-marker (TN) genes between cell types (details in Materials and Methods) (Figure 1).

Evaluations on 23 scenarios with overlapping cell types but different cell proportions and magnitude of batch effects between two batches

To evaluate the influence of batch effects and cell populational compositions on the denoising performance, we designed 15 scenarios with simple cell group structures (only two cell types) and 8 scenarios with complicated cell group structures (multiple cell types and subtypes) in 2 batches using both simulated and real datasets. The 15 scenarios comprised 9 simulated and 6 real studies from pancreas scRNA-seq data [28–30], including 3 levels of cell composition imbalance (balanced, mild imbalanced or severe imbalanced) and 2 or 3 levels of magnitude of batch effects, respectively (details in Materials and Methods). The eight scenarios with multiple cell types, generated from PBMC datasets [31], included two levels of batch effects (mild and strong) and four levels of cell compositions imbalance (balanced, mild, moderate and severe imbalanced) (details in Materials and Methods).

Harmony and LIGER output low-dimensional cell embeddings without a corrected-gene-expression matrix. Therefore, TPR or TNR scores were not reported for Harmony and LIGER. Although scVI and fastMNN return batch-corrected values after integration, they had poor TPR score, suggesting disrupted gene structures (Supplementary Figure S1 available online at http://bib.oxfordjournals.org/). As mentioned in their studies, the batch-corrected values from scVI and fastMNN no longer correspond to gene-expression values and they cannot be directly used in gene-based analysis [17, 25]. Although Scanorama obtained high TPR and TNR scores, the corrected expression had different ranges from the original data, resulting in low fold changes in differential analysis (Supplementary Figure S2 available online at http://bib.oxfordjournals.org/). Different ranges between corrected and original gene expression in Scanorama was also reported in a recent benchmark study [23]. In summary, Seurat, scMerge, ZINB-WaVE, limma and ComBat are recommended for gene-based downstream analysis (Table 1, gene-based analysis).

In a simple structure with only two cell types in each batch (scenarios 1–15), similar results were obtained in simulation and real datasets (Figure 2 and Supplementary Figures S1 and S3 available online at http://bib.oxfordjournals.org/). The performance of linear-based methods (limma, ComBat and ZINB-WaVE) was sensitive to cell composition imbalance. They worked well when cell compositions were balanced

Table 1. Summary of performance of batch correction methods in different scenarios

	Gene-based analysis	Two batches												Multiple batches	
		Mild batch effect						Strong batch effect						Mixed batch effects	
		Balance	Imbalance					Balance	Imbalance					Overlapping	Nonoverlapping
			Overlapping but mild/ moderate imbalance	Overlapping but severe imbalance	A rare nonoverlapping population	Very different nonoverlapping cell populations	Similar nonoverlapping cell populations		Overlapping but mild/ moderate imbalance	Overlapping but severe imbalance	A rare nonoverlapping population	Very different nonoverlapping cell populations	Similar nonoverlapping cell populations		
Combat	✓				✓						✓				
limma	✓				✓						✓				
ZINB-WaVE	✓				✓						✓				
Seurat-CCA	✓	✓	✓		✓			✓	✓		✓			✓	
scMerge	✓	✓	✓	✓		✓						✓			
fastMNNs		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		✓	⍻
Scanorama	⍻	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		⍻	
Harmony		✓	✓	✓	✓	✓	✓	⍻	⍻	⍻	⍻	⍻	⍻	⍻	⍻
scVI		✓	✓	✓		✓	✓								
LIGER		✓	✓	✓	✓	✓	✓	✓	✓		✓	✓			

⍻: Use with caution (Scanorama) or parameters need to be adjusted to achieve good performance (Harmony) or not great but rank first (fastMNN).

but resulted in data distortion in imbalanced scenarios. Compared to the original data (unadjusted), the corrected data produced by these methods in mild or severe imbalanced scenarios had lower ASW_batch and scUnifrac_batch scores, suggesting poorer batch mixing (Figure 2**B** and **C** and Supplementary Figure S1 available online at http://bib.oxfordjournals.org/). The reason was that they misrecognized the transcriptomic differences caused by population compositions as batch effects and overcorrected the data even when no technical biases existed (Supplementary Figure S3 available online at http://bib.oxfordjournals.org/). As expected, NN-based methods, fastMNN, Scanorama and scMerge, were robust to magnitude of batch effects and cell composition imbalance. They obtained better batch mixing (higher ASW_batch and scUnifrac_batch scores) than linear-based methods especially when cell compositions were imbalanced (Figure 2 and Supplementary Figures S1 and S3 available online at http://bib.oxfordjournals.org/). Seurat, one of NN-based methods, achieved good performance when there was only mild cell composition imbalance (Figure 2**B** and **E**). However, it totally changed the transcriptomic profiles when the populational imbalance was severe. Seurat mixed the two main populations from each batch even though they belonged to different cell types and no batch effect existed (low scUnifrac_group scores) (Figure 2**C** and **F** and Supplementary Figures S1 and S3 available online at http://bib.oxfordjournals.org/). The performance of scVI and Harmony was affected by the magnitude of batch effects, which had low ASW_batch and scUnifrac_batch scores when there were strong batch effects, suggesting poor batch mixing (Figure 2**D**–**F**). It should be noted that the performance of Harmony could be improved if we added more penalties on clusters with low batch-diversity and forced batch mixing (Supplementary Figure S4 available online at http://bib.oxfordjournals.org/). LIGER removed batch effects successfully in every scenario (Figure 2). However, it had low scUnifrac_group scores, indicating damaged cell group structures (see the example in PBMC datasets below). Different normalization methods had subtle impact on the performance if any in all scenarios.

In scenarios with complicated cell group structures (scenarios 16–23), linear-based methods (limma, Combat and ZINB-WaVE) and scMerge showed different results from those in settings with simple group structures. Even when cell populations were balanced and batch effects were mild, linear-based methods only had a subtle improvement on ASW_batch and scUnifrac_batch scores, suggesting incomplete batch removal (Figure 3**A**). tSNE plots showed that major cell types mixed well (T cells) but not minor groups (monocytes, megakaryocyte and cDC) (Supplementary Figure S5 available online at http://bib.oxfordjournals.org/). This was supported by higher scUnifrac_batch scores for CD4 T and CD8 T cells but lower scores for CD14 monocyte, CD16 monocyte, megakaryocyte and cDC (Supplementary Figure S6

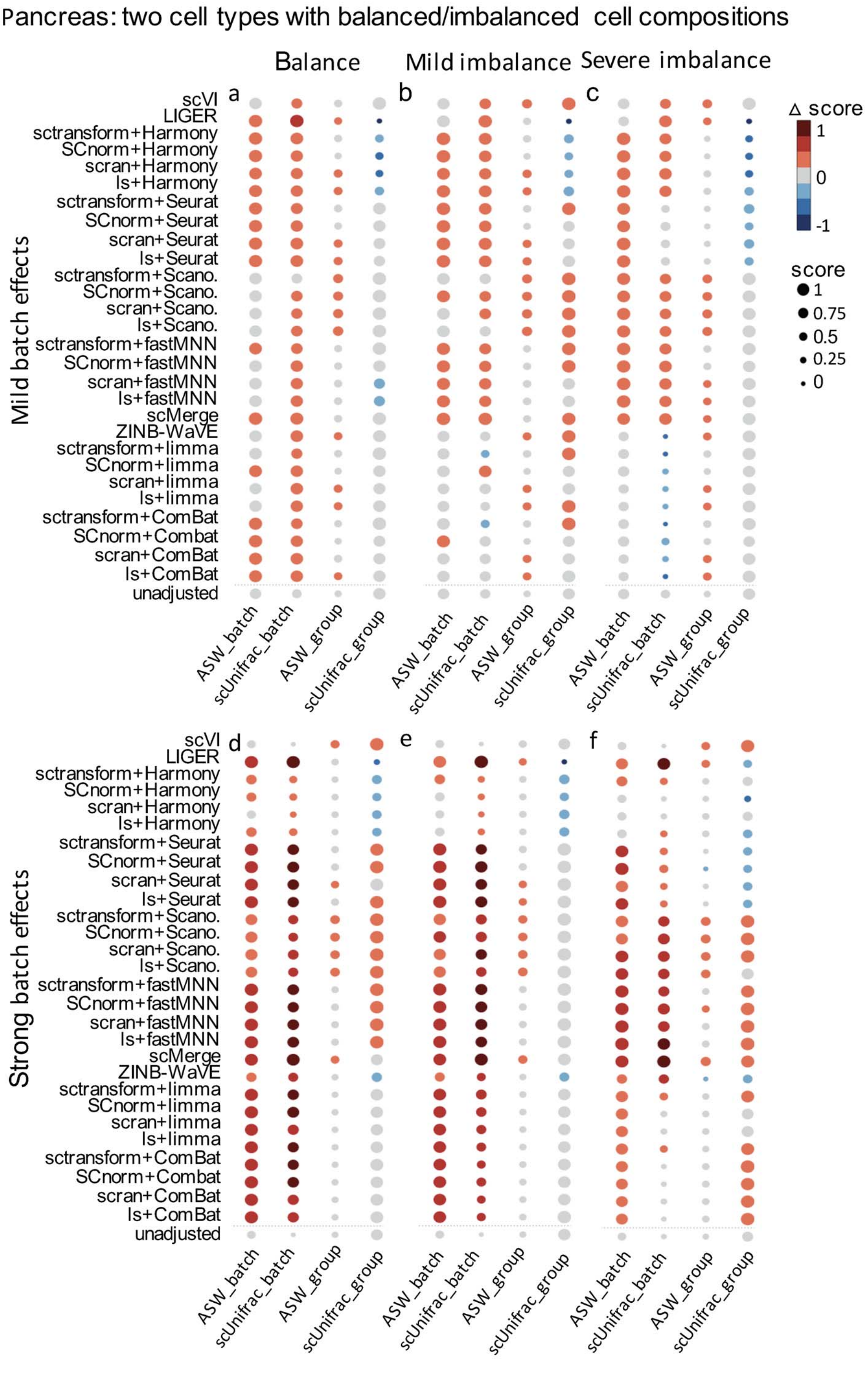

Figure 2. Evaluation of noise reduction procedures on six scenarios generated from pancreas scRNA-seq data with two overlapping cell types in two batches. Six scenarios included three levels of cell compositional difference (balanced, mild and severe imbalanced) and two levels of batch effects (mild and strong). (**A**) Balanced and mild batch effects; (**B**) Mild imbalanced and mild batch effects; (**C**) Severe imbalanced and mild batch effects; (**D**) Balanced and strong batch effects; (**E**) Mild imbalanced and strong batch effects; (**F**) Severe imbalanced and strong batch effects. Scores were represented by circle sizes and changes in scores (Δ score) after noise reduction procedures were denoted by colors: red for increase, blue for decrease and gray for unchanged.

available online at http://bib.oxfordjournals.org/). The unsuccessful batch removal by linear methods was highly likely due to the cell-type specific batch effects, which could not be removed by one single transcriptomics shift. When there were strong batch effects, batch mixing for those minor cell types got even worse with even lower scUnifrac_batch scores (Figure 3**E**, Supplementary Figures S5 and S6 available online at http://bib.oxfordjournals.org/). ComBat, modeling mean and variances simultaneously, achieved better performance than limma and ZINB-WaVE (Figure 3). In the balanced and mild effect design, Harmony mixed monocytes but not T cells, suggesting parameters need to be adjusted to increase batch mixing in those cell types (Supplementary Figures S5 and S6 available online at http://bib.oxfordjournals.org/). Surprisingly, scMerge failed when batch effects were strong, which wrongly mixed cell types together (Figure 3**E**–**H** and Supplementary Figure S5 available online at http://bib.oxfordjournals.org/). scMerge uses k-means clustering before NN search, whose performance would be greatly affected if k was chosen inappropriately. In high technical noise and complicated cell-type structures (strong batch effects and multiple cell types and subtypes), k-means clustering combined with NN search cannot align cell structures accurately between two batches. Consistent with scenarios of simple cell group structures, Seurat worked well when cell compositions were balanced, mild or modest imbalanced, no matter mild or strong batch effects (Figure 3**A**–**C** and **E**–**G**). However, it mixed different cell types in the scenarios of severe imbalance (Figure 3**D** and **H** and Supplementary Figure S7 available online at http://bib.oxfordjournals.org/). LIGER worked well except the scenario of strong batch effects and severe imbalanced cell compositions, where different cell types were mixed. In this case, the transcriptomics structure was too complicated to find shared and batch-specific metagenes, resulting in incorrect alignment of cell types (Supplementary Figure S7 available online at http://bib.oxfordjournals.org/). fastMNN achieved great performance in all scenarios (Figure 3 and Supplementary Figures S5 and S7 available online at http://bib.oxfordjournals.org/). The normalization method Scran increased batch mixing (high ASW_batch and scUnifrac_batch scores) at the cost of losing cell group structures slightly (low ASW_group and scUnifrac_group scores), especially when combined with Seurat, limma or ComBat (Figure 3). To be noted, cDC had lower scUnifrac_batch scores than other cell types, suggesting that batch removal on rare cell types were not as successful as others, especially for LIGER, Seurat, Scanorama and MNN (Supplementary Figure S6 available online at http://bib.oxfordjournals.org/).

In summary, proportional imbalance in cell groups greatly impacts the performance of linear-based methods and Seurat. Seurat multi-CCA is tolerant to populational imbalance unless the imbalance is severe. Seurat reciprocal PCA is recommended by the developer when cell populations are very different across batches. LIGER had great performance except the scenario of both strong batch effects and severe imbalanced cell compositions. The high technical noise and biological variations make it difficult to identify shared metagenes to align cells. Even LIGER removed batch effects successfully and distinct cell groups were observed in tSNE plots, and the global cell group structures were disrupted (low scUnifrac_group scores in all 23 scenarios). Using PBMC datasets as an example, global distances between two distinct cell types (T cells and monocytes) were not preserved well in low-dimensional cell embeddings of LIGER compared to other methods (Supplementary Figure S8 available online at http://bib.oxfordjournals.org/). The magnitude of batch effect affects the performance of scVI and Harmony. Harmony is flexible, and debugging parameters would help denoise the data correctly. scMerge is sensitive to cell group structures, especially when batch effects are strong. fastMNN and Scanorama have stable performance across different scenarios (Table 1; two batches; mild/strong batch effects; balance/imbalance and overlapping but mild/severe imbalance).

Evaluations on 19 scenarios with nonoverlapping cell types between two batches

In most cases, datasets do not contain the exact same cell types across batches. To test the performance when there were nonoverlapping cell types, we set up nine scenarios for simulation studies and six scenarios using pancreas scRNA-seq datasets mentioned above. The nine simulated scenarios consisted of the combination of three levels of batch effects (none, mild and strong) and three types of nonoverlapping, i.e. a rare cell type, a cell type or a dominant cell type. Here, the nonoverlapping cell type only existed in one batch but not in the other. If the nonoverlapping cell type was a rare population in the batch, it was called 'a rare nonoverlapping cell type'. If the nonoverlapping cell type was a dominant population in the batch, it was called 'a dominant nonoverlapping cell type'. If the nonoverlapping cell type was neither rare nor dominant, it was called 'a nonoverlapping cell type' (details in Materials and Methods). The six scenarios in real datasets were similar but without the no-batch effect setting (Figure 4).

The performances in real datasets were similar with those in simulation studies (Figure 4 and Supplementary Figure S9 available online at http://bib.oxfordjournals.org/). The scenarios with nonoverlapping cell types between batches were specific cases of populational composition imbalance. When one batch had a rare nonoverlapping cell type, the imbalance was subtle. Therefore, although performances of linear-based methods were greatly affected by cell composition as mentioned above, they were able to reduce batch effects (increased ASW_batch and scUnifrac_batch scores compared to unadjusted) (Figure 4**A** and **D**). ZINB-WaVE obtained lower ASW_batch and scUnifrac_batch scores than other methods even in the case of a rare nonoverlapping cell type and strong batch effects (Figure 4**D**). The tSNE plots also showed the same cell types from two batches were mixed poorly (Supplementary Figure S10 available online at http://bib.oxfordjournals.org/). LIGER, Seurat, fastMNN and Scanorama obtained improved ASW_batch and scUnifrac_batch scores, suggesting batch removal (Figure 4). Surprisingly, scMerge obtained lower ASW_batch and scUnifrac_batch scores, indicating unsuccessful batch correction in the scenario of mild batch effects (Figure 4**A**). In strong batch effects, it had lower ASW_group and scUnifrac_group scores,

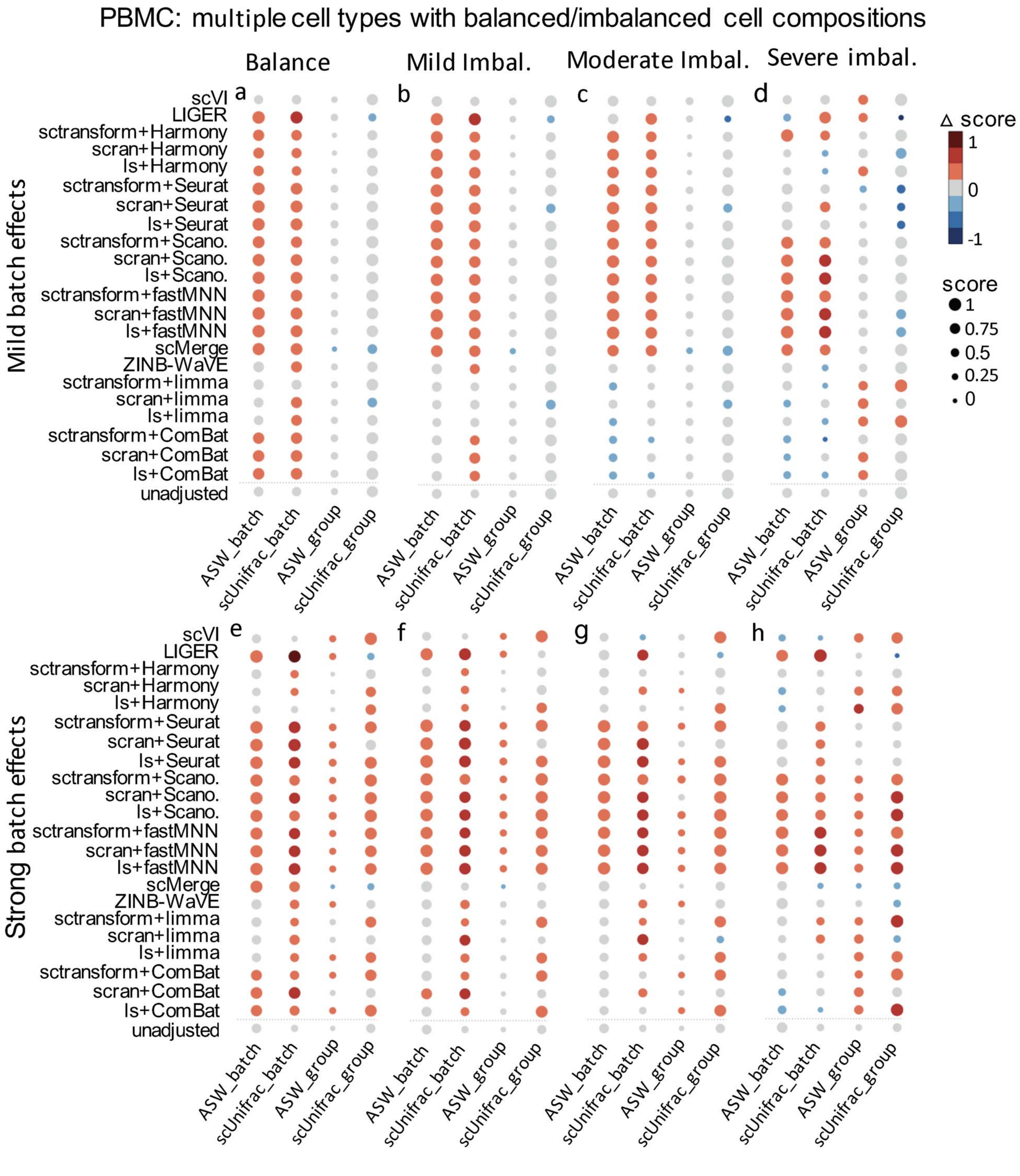

Figure 3. Evaluation of noise reduction procedures on eight scenarios generated from PBMC scRNA-seq data with multiple cell types in two batches. Eight scenarios included four levels of cell compositional difference (balanced, mild, moderate and severe imbalanced) and two levels of batch effects (mild and strong). (**A**) Balanced and mild batch effects; (**B**) Mild imbalanced and mild batch effects; (**C**) Moderate imbalanced and mild batch effects; (**D**) Severe imbalanced and mild batch effects; (**E**) Balanced and strong batch effects; (**F**) Mild imbalanced and strong batch effects; (**G**) Moderate imbalanced and strong batch effects; (**H**) Severe imbalanced and strong batch effects. Scores were represented by circle sizes and changes in scores (Δ score) after noise reduction procedures were denoted by colors: red for increase, blue for decrease and gray for unchanged.

suggesting damaged cell group structures (Figure 4D) (Supplementary Figure S10 available online at http://bib.oxfordjournals.org/). This was probably due to the reason that scMerge uses *k*-means clustering before NN search, whose performance highly depends on the predefined *k*. Additionally, the existence of a rare population violates the assumption of equal sized clusters in *k*-means clustering.

When one batch had a nonoverlapping cell type or a dominant nonoverlapping cell type, the imbalance became moderate or even severe. The performances in these cases were similar with those with identical

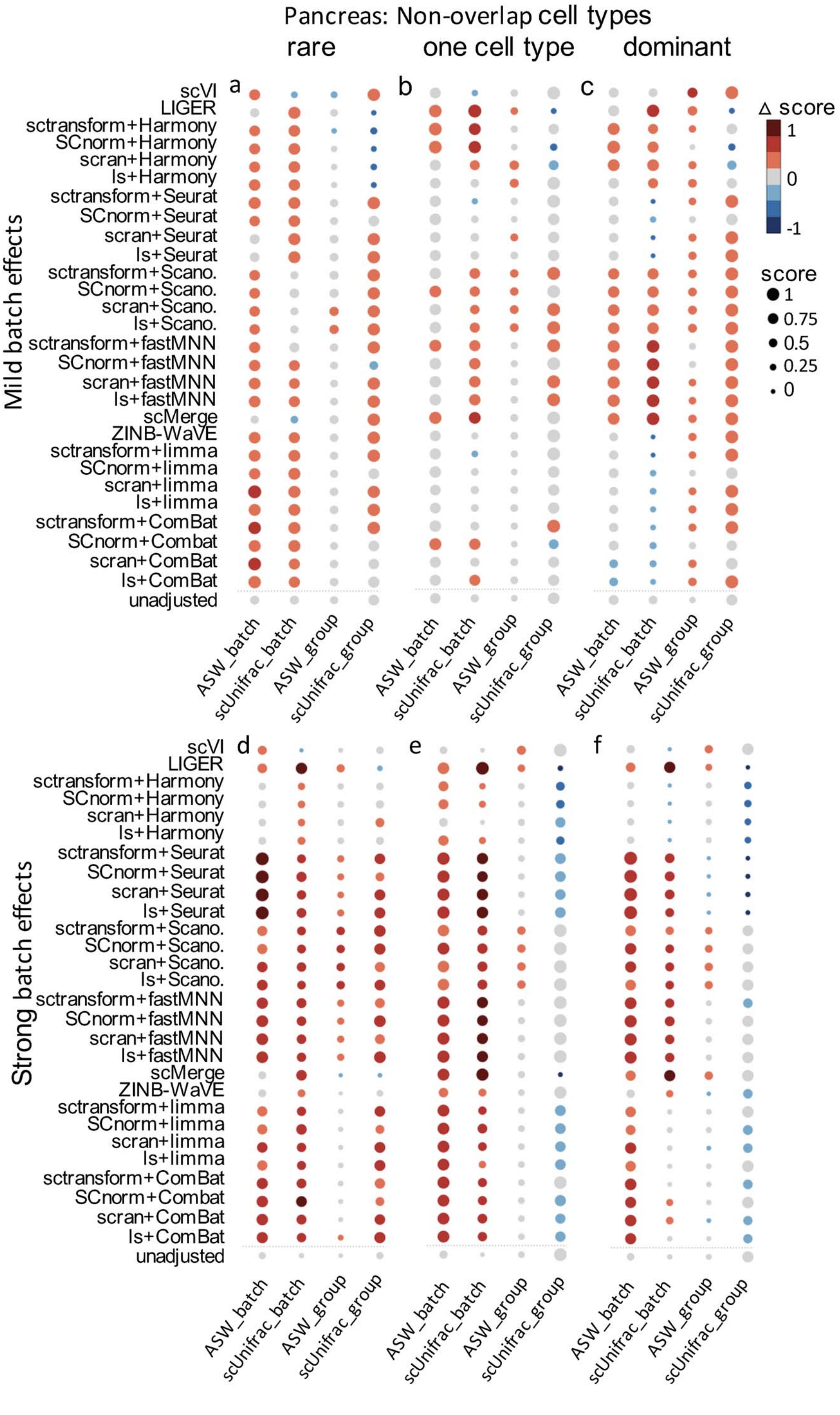

Figure 4. Evaluation of noise reduction procedures on six scenarios generated from pancreas scRNA-seq data with nonoverlapping cell types in two batches. The nonoverlapping cell type was a rare, one (neither rare nor dominant) or dominant cell type in the batch. The batch effect was either mild (top) or strong (bottom). (**A**) a rare nonoverlapping cell type and mild batch effects; (**B**) a nonoverlapping cell type and mild batch effects; (**C**) a dominant nonoverlapping cell type and mild batch effects; (**D**) a rare nonoverlapping cell type and strong batch effects; (**E**) a nonoverlapping cell type and strong batch effects; (**F**) a dominant nonoverlapping cell type and strong batch effects; Scores were represented by circle sizes and changes in scores (Δ score) after noise reduction procedures were denoted by colors: red for increase, blue for decrease and gray for unchanged.

cell types but different proportions. That is, linear-based methods, LIGER and Seurat would not work well, whereas NN-based methods reduced the batch effect successfully, including scMerge, fastMNN and Scanorama (Figure 4**B**, **C**, **E** and **F**) (Supplementary Figure S10 available online at http://bib.oxfordjournals.org/). scVI reduced mild batch effects but failed in the case of strong batch effects. Harmony failed in those cases with strong batch effects with default parameters, but performance could be improved if we adjusted parameters to penalize heavily on clusters with low batch-diversity.

Additionally, we created another four scenarios where two batches shared one common cell type while each having one batch-specific cell type using pancreas scRNA-seq data. For example, one batch consisted of alpha and ductal cells, and the other batch included beta and ductal cells. The two batch-specific cell types were either very different or similar (details in Materials and Methods). Since these scenarios were specific cases of imbalanced cell populations, linear-based methods (limma, ComBat and ZINB-WaVE) and Seurat did not work well (Figures 5 and 6). Linear-based methods resulted in poor batch mixing with a low scUnifrac_batch score. Seurat led to different cell types blending in addition to the mixture of the common cell type from two batches (Figures 5**C** and 6**C**). It should be noted that ASW_group and scUnifrac_group could not reflect the wrong cell-type mixture. For example, although alpha and beta cells were wrongly mixed (Figure 6**C**), they were still separated from ductal cells within each batch, resulting in high and misleading ASW_group and scUnifrac_group scores. The wrong cell-type mixture could be observed from tSNE plots (Figure 6**C** and Supplementary Figure S11 available online at http://bib.oxfordjournals.org/). Among NN-based methods, Scanorama achieved great performance in the mild batch effect setting, which mixed the common cell type from two batches well and also separated batch-specific cell types no matter they were distinct or similar (Figure 5). In the strong batch effect setting, however, it worked when batch-specific cell types were distinct but failed when they were similar (Figure 6 and Supplementary Figure S11 available online at http://bib.oxfordjournals.org/). fastMNN and scMerge removed batch effects successfully when batch-specific cell types were distinct (alpha and ductal cells in Figure 5). However, they mixed the batch-specific cell types when they were similar (alpha and beta cells in Figure 6). The reason was that similar batch-specific cell types were detected as MNNs since they were similar and thus mistreated as the same cell type. When batch-specific cell types were very different, they were not detected as MNNs, therefore, only the common cell type from two batches are mixed and batch-specific cell types are kept separate (Figures 5 and 6). LIGER worked well when batch-specific cell types were distinct but mixed cell types incorrectly when they were similar and batch effects were strong (Supplementary Figure S11 available online at http://bib.oxfordjournals.org/). Since LIGER aligns cells based on shared and batch-specific metagenes, batch-specific metagenes was mistakenly identified as shared metagenes if they were similar. Adding strong batch effects on top of similar batch-specific cells made the situation even worse. The performance of scVI and Harmony was not affected by the population imbalance. They worked well when batch effect was mild (Figure 5). However, scVI failed in the strong batch effect scenario as mentioned above (Supplementary Figure S11 available online at http://bib.oxfordjournals.org/). The parameters of Harmony need to be adjusted for a successful batch removal when batch effect was strong (Figure 6 and (Table 1; two batches, mild/strong batch effects, balance/imbalance and distinct/similar nonoverlapping cell populations). Each real scenario was repeated 100 times by resampling cells to evaluate the variability of the performance. Similar results were obtained from replicated scenarios.

Evaluations on three scenarios with multiple batches

Integrating multiple batches is challenging due to the high cellular heterogeneity and different levels of technical biases across datasets. We designed three scenarios from PBMC data [31], which included six batches and nine cell types. Each scenario had both complicated batch effects and cell group structures. There were both strong and mild batch effects across six datasets. Moreover, there were distinct cell types, such as T cells and monocytes, and also similar sub-groups, such as CD4 and CD8 T.

In the first scenario, each batch had the full dataset, which included all cell types in the original data. Seurat performed the best as it greatly reduced batch effects (the highest scUnifrac_batch score) and also retained cell group structures (an unchanged scUnifrac_group score) (Figure 7**A** and Supplementary Figure S12 available online at http://bib.oxfordjournals.org/). fastMNN ranked the second, which mixed six batches without altering cell group structures. Scanorama, however, only removed batch effects partially. For example, CD14 monocytes showed low scUnifrac_batch scores, suggesting insufficient batch removal (Supplementary Figure S13 available online at http://bib.oxfordjournals.org/). It also mingled some cell types (Supplementary Figure S12 available online at http://bib.oxfordjournals.org/). scMerge disrupted cell group structures (an improved scUnifrac_batch score but a decreased scUnifrac_group score), where some distinct cell types in different batches were put together wrongly (Supplementary Figure S12 available online at http://bib.oxfordjournals.org/). Linear-based methods, limma, ComBat and ZINB-WaVE, removed batch effects partially, which showed lower scUnifrac_batch scores for CD14 and CD16 monocytes (Supplementary Figure S13 available online at http://bib.oxfordjournals.org/). Combat obtained better performance than limma and ZINB-WAVE with higher scUnifrac_batch scores (Figure 7**A** and Supplementary Figures S12 and S13 available online at http://bib.oxfordjournals.org/). LIGER not only mixed batches (high scUnifrac_batch scores) but also blended distinct cell types, suggesting LIGER failed to identify shared metagenes in such complicated batch- and cell-type structures (Figure 7**A** and Supplementary Figure S12 available online at http://bib.oxfordjournals.org/). In the second scenario, each cell type has a 25% of chance to be missing in each batch. The results were consistent with those in the first scenario. fastMNN and Seurat achieved the best performance. Scanorama only removed batch effects partially. scMerge, limma, ComBat and ZINB-WaVE worked poorly to remove batch effects (Figure 7**B**

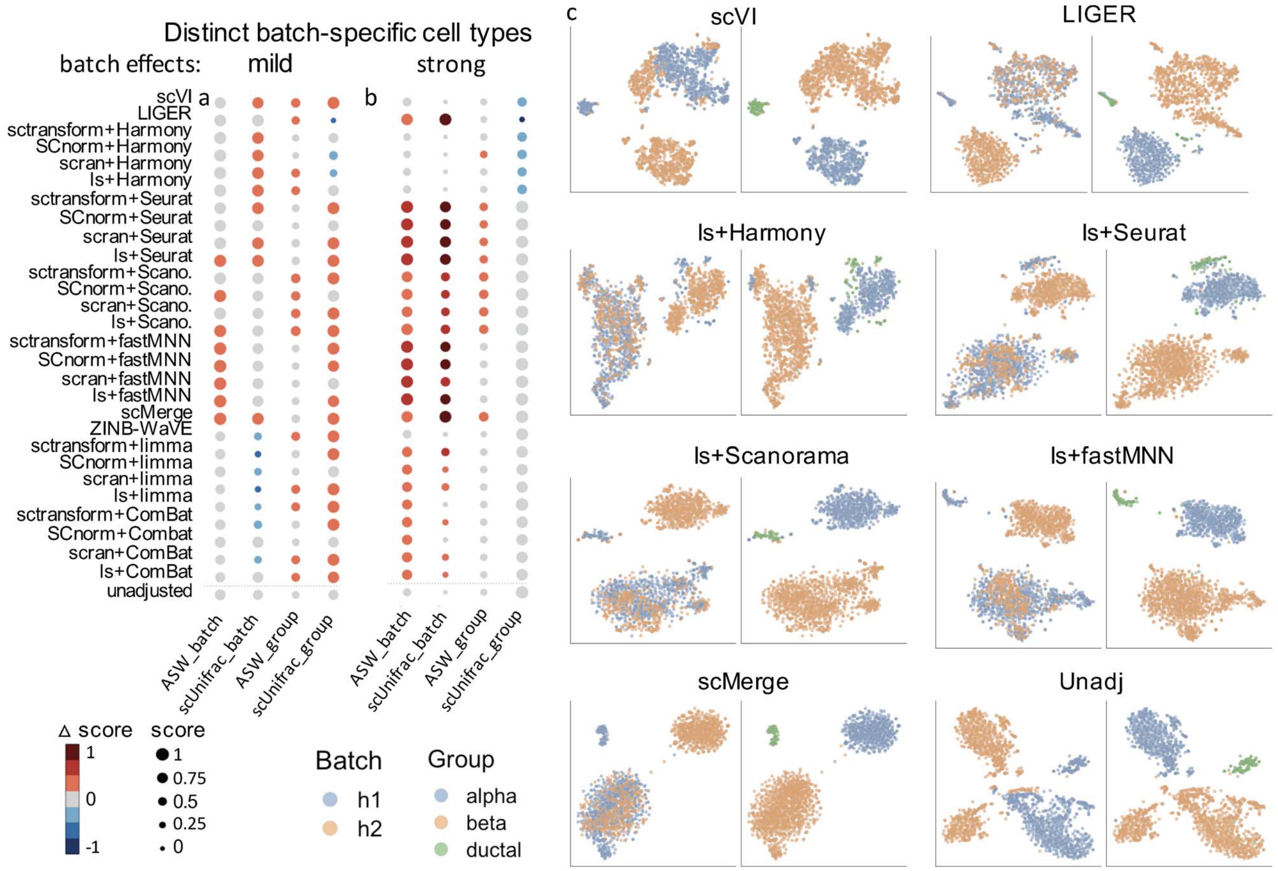

Figure 5. Evaluation of noise reduction procedures on two scenarios generated from pancreas scRNA-seq data in two batches. One batch had beta and ductal cells, and the other batch consisted of beta and alpha cells. The batch effect was either mild (**A**) or strong (**B**). tSNE visualization for corrected and uncorrected datasets in the mild batch effect setting (**C**) (left: by batch; right: by group). Scores were represented by circle sizes and changes in scores (Δ score) after noise reduction procedures were denoted by colors: red for increase, blue for decrease and gray for unchanged.

and Supplementary Figure S12 available online at http://bib.oxfordjournals.org/). Consistent with results from two batches, rare cell types (cDC and pDC) had much lower scUnifrac_batch scores than others, suggesting insufficient batch removal for those rare cell types in almost all methods (Supplementary Figure S13 available online at http://bib.oxfordjournals.org/).

In the third scenario, each cell type has 50% of chances to be missing in each batch. In this case, where a significant portion of cells were nonoverlapping across datasets, Seurat led to cell-type mixing with decreased scUnifrac_group scores than unadjusted (Supplementary Figure S12 available online at http://bib.oxfordjournals.org/). The performance of fastMNN was slightly better than Seurat, which mixed batches at the cost of some cell types blending (Figure 7**C** and Supplementary Figure S12 available online at http://bib.oxfordjournals.org/). It should be noted that Harmony and scVI did not work well due to the existence of strong batch effects. The findings were summarized in Table 1; multiple batches. In terms of normalization methods, Scran, combined with Seurat, limma or ComBat, improved batch mixing at the cost of disrupting cell group structures, which was consistent with results from two batches (Figure 7). Sctransform, in contrast, preserved cell group structures the best compared to ls and Scran, especially in the third scenario when a significant number of cell types were missing in every batch (Figure 7). The 2nd and 3rd scenarios were repeated 20 times and similar results were obtained.

Evaluations on 10 scenarios with variable dropout rate and library size

In previous scenarios, we focused on the impact of cell populational imbalance and magnitude of batch effects on denoising procedures. These two factors mostly affected the performance of batch correction methods. The normalization methods, however, showed subtle influence on the performance. To benchmark the performance of normalization methods, we created 10 scenarios using real and simulated datasets with balanced design and mild batch effect between two batches. One scenario was generated by down-sampling reads from each cell in the PBMC dataset mentioned above. Higher percentage of reads was removed in cells with smaller library sizes, leading to highly variable sizes in each batch (details in Materials and Methods). The other nine scenarios were generated by simulation, where there were three levels of variations in dropout rate and library size (low, middle and high). Dropout rate

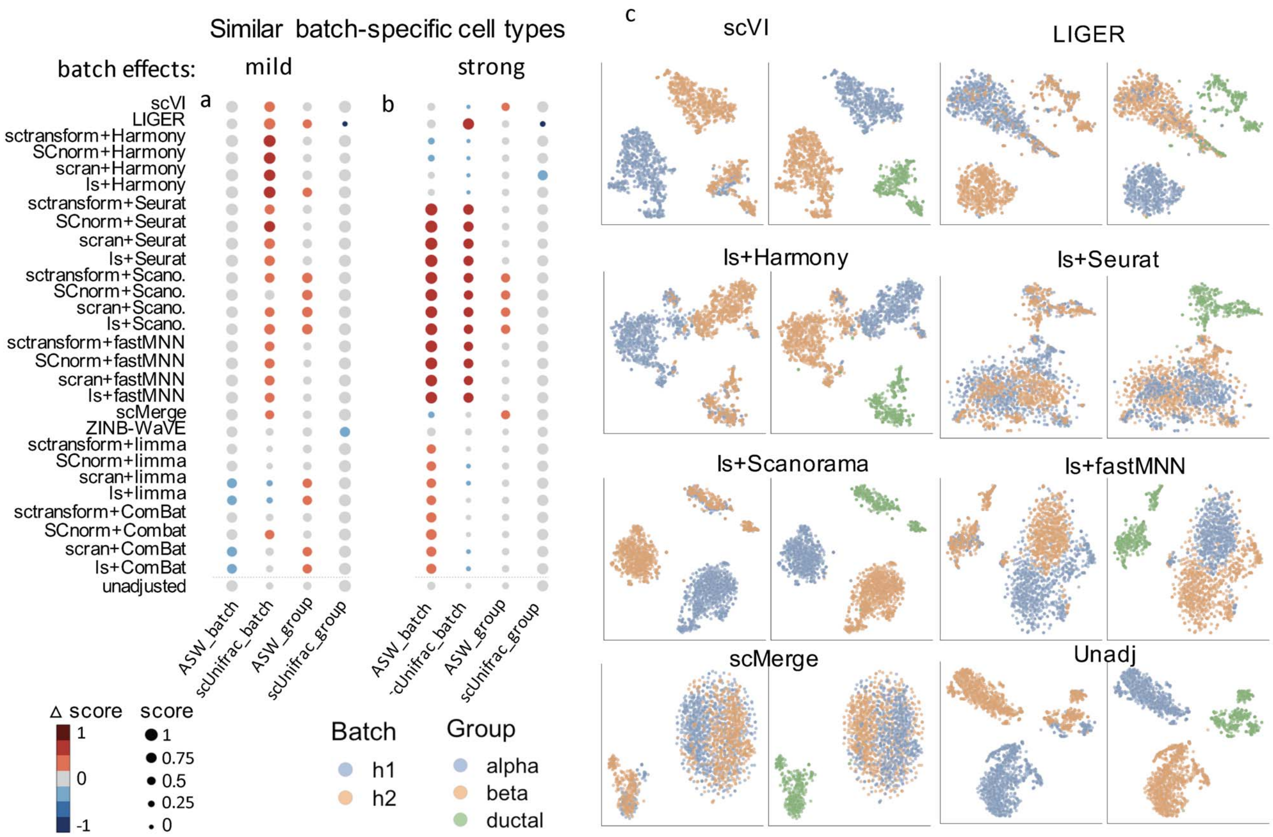

Figure 6. Evaluation of noise reduction procedures on two scenarios generated from pancreas scRNA-seq data in two batches. One batch had beta and ductal cells, and the other batch comprised alpha and ductal cells. The batch effect was either mild (**A**) or strong (**B**). tSNE visualization for corrected and uncorrected datasets in the mild batch effect setting (**C**) (left: by batch; right: by group). Scores were represented by circle sizes. Changes in scores (Δ score) after noise reduction procedures were denoted by colors: red for increase, blue for decrease and gray for unchanged.

and library size are the two common technical noises in single-cell RNA-seq datasets. The successful removal of those technical noises is critical for uncovering cell group structures.

In the down-sampled dataset, different procedures were able to remove or partially remove batch effects, whose performance was similar with those from the scenario mentioned above (two batches with multiple cell types in a balanced and mild batch effects design). Surprisingly, normalization methods had a substantial impact on preservation of cell group structures (Figure 8 and Supplementary Figure S14 available online at http://bib.oxfordjournals.org/). As mentioned above, Scran disrupted cell group structures slightly, especially combined with Seurat, limma or ComBat in the scenario of two batches and multiple batches (Figures 3 and 7). When library sizes varied a lot across cells, Scran greatly damaged cell group structures in the combination with Seurat, limma or ComBat, especially for CD4 and CD8 T cells (low scUnifrac_group scores). In comparison, Scran combined with fastMNN or Scanoroma preserved cell group structures better. ls and sctransform preserved cell group structures well in the combination of different batch correction methods (Figure 8 and Supplementary Figure S14 available online at http://bib.oxfordjournals.org/). ZINB-Wave and scMerge, which handle normalization and batch correction simultaneously, had poor performance in preservation of cell group structures when library size varied a lot (low ASW_group and scUnifrac_group scores). ZINB-WaVE brought different cell groups together and blended distinct cell types, and scMerge merged similar subgroups like CD4 and CD8 T cells (Figure 8 and Supplementary Figure S14 available online at http://bib.oxfordjournals.org/).

In simulation datasets with only two groups, when dropout rate was low and the library size across cells was similar, every normalization and batch correction methods worked well (Supplementary Figure S15 available online at http://bib.oxfordjournals.org/). When dropout rate was high, this technical noise masked the cell group structure, which was reflected by a low scUnifrac_group score before adjustment (Supplementary Figure S15 available online at http://bib.oxfordjournals.org/). scVI and ZINB-WaVE uncovered the cell group structure successfully with a high scUnifrac_group score. fastMNN achieved higher scUnifrac_group scores when combined with normalization methods ls, Scran and SCnorm than with sctransform. Harmony, however, worked better in combination with sctransform and SCnorm. Seurat, scMerge, limma and ComBat failed to recover the cell group structure when the dropout rate was high. When the library size varied a lot, sctransform

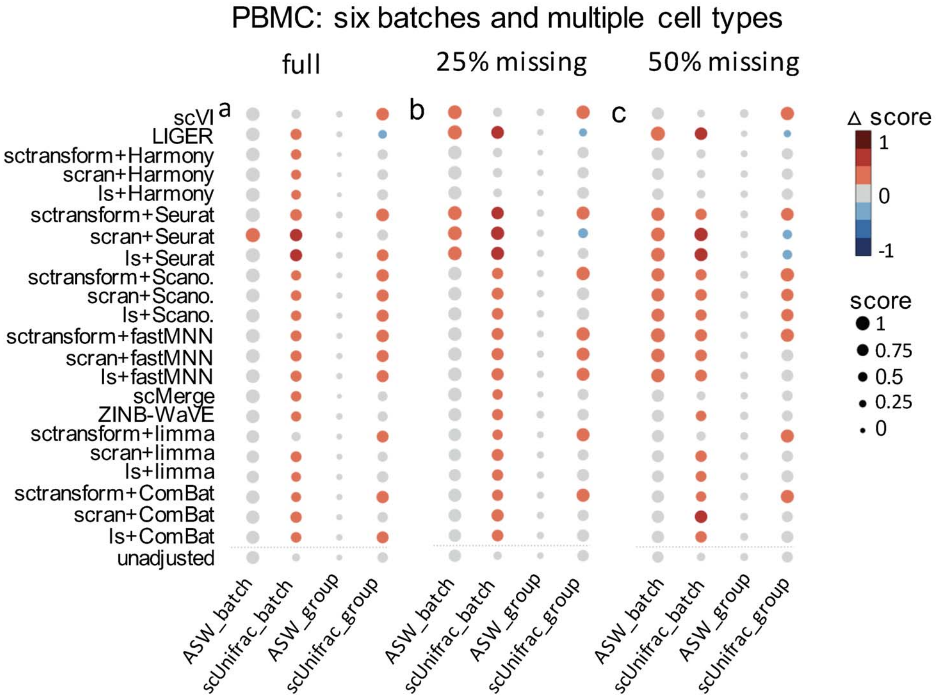

Figure 7. Evaluation of noise reduction procedures on three scenarios generated from PBMC scRNA-seq data in six batches. The three scenarios were generated from the full dataset (**A**), randomly removing cell types by a 25% of chance (**B**) or by a 50% of chance (**C**). Scores were represented by circle sizes and changes in scores (Δ score) after noise reduction procedures were denoted by colors: red for increase, blue for decrease and gray for unchanged.

performed better than other normalization methods. The combination of sctransform with Harmony, Seurat, Scanorama, limma and ComBat achieved higher scUnifrac_group scores. scVI and scMerge failed to uncover the cell group structure with low scUnifrac_group scores (Supplementary Figure S15 available online at http://bib.oxfordjournals.org/). When there were both high dropout rates and variable library sizes, Scanorama combined with sctransform and ZINB-WaVE performed the best with the highest scUnifrac_group score (Supplementary Figure S15 available online at http://bib.oxfordjournals.org/).

A guideline to select suitable procedures

When integrating multiple datasets, batch effects are generally the main source of technical noise. Selection of appropriate batch correction methods is critical for denoising since each method has its own assumption and works in certain scenarios. Under- or overcorrection will lead to false interpretation of downstream analysis. Compared to batch correction, normalization methods have subtle impact on the performance unless library sizes vary dramatically across cells. In this case, sctransform and ls are recommended rather than Scran.

The performance of batch removal methods is mostly affected by three factors, the complexity of cell group structures, cell populational imbalance and magnitude of batch effects. Linear-based methods, such as ComBat, limma and ZINB-WaVE, have relative better performance when there are no or very subtle differences (balance or a rare nonoverlapping population) in cell populational composition across datasets than imbalanced scenarios. When there are cell compositional difference across datasets, transcriptomics difference caused by cell populational compositions would be mistakenly treated as technical biases, resulting in overcorrection. The existence of cell-type specific batch effects, however, limited their performance even in balanced scenarios. Therefore, they should be used with caution for single-cell RNA-seq batch correction.

Seurat uses CCA to map cells from different batches into a common reduced dimensional space, which works well when populational composition is not severe across datasets. When cell populations are extremely imbalanced, such as substantially different composition or a unique major population in one dataset, Seurat results in erroneous cell-type mixing (Table 1). Extremely different compositions result in weakly correlated gene modules in two datasets, making it difficult for CCA to capture. In this case, reciprocal PCA is recommended by the developer.

The performance of other NN-based methods, scMerge, fastMNN and Scanorama, are generally not affected by cell populational imbalance. When there are similar batch-specific cell types, however, scMerge and fastMNN mistake them as MNNs and result in incorrect cell-type mixing (similar nonoverlapping cell populations, as shown in Table 1). scMerge finds MNNs between cell clusters instead of cells. The existence of rare nonoverlapping populations or complicated batch effects

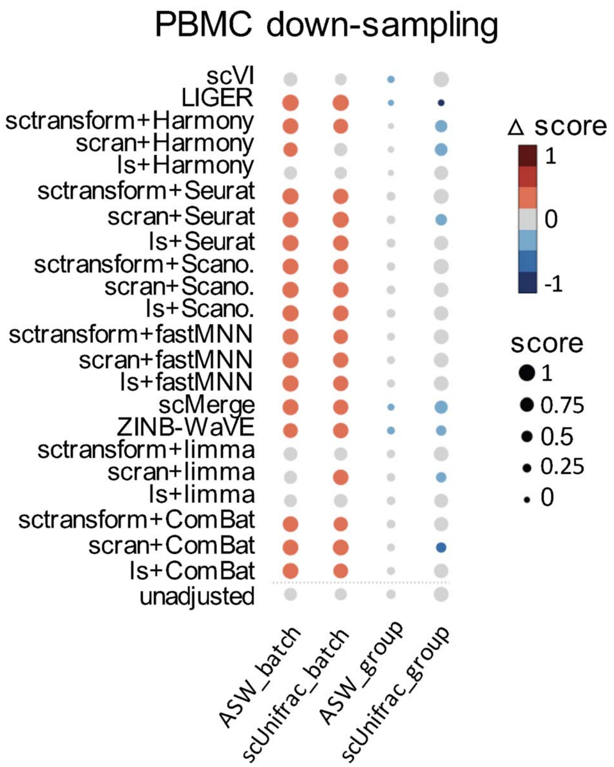

Figure 8. Evaluation of noise reduction procedures one scenario generated from PBMC scRNA-seq datasets in two batches with mild batch effects. Cells in batch were down-sampled to have highly variable library sizes. Scores were represented by circle sizes and changes in scores (Δ score) after noise reduction procedures were denoted by colors: red for increase, blue for decrease and gray for unchanged.

and cell group structures, e.g. multiple batches and coexistence of strong batch effects and complicated cell groups, make it difficult for clustering or aligning cell clusters across batches (Table 1). Scanorama performs well in scenarios with two batches, but its performance reduces in scenarios with mixed batch effects.

The performance of scVI is robust to cell populational imbalance but sensitive to magnitude of batch effects. When there is a strong batch effect, scVI leads to undercorrection. The performance of Harmony is similar to scVI, but it is flexible to balance the cluster accuracy and batch mixing. For example, we found that the default parameters led to undercorrection when the batch effect was strong (Table 1). However, the batch effect would be completely removed if we changed parameters to force dataset mixing. Those parameters should be adjusted based on the number of cells and the magnitude of batch effects. The performance of LIGER depends on the accurate identification of shared and batch-specific metagenes. When there are high technical noise and biological variations, such as the co-existence of strong batch effects and severe compositional imbalance, strong batch effects and similar batch-specific cell types and the combination of mixed batch effects and complicated cell group structures, shared metagenes are not the major source of variances and are masked by those technical and biological noises. Therefore, LIGER cannot identify shared metagenes accurately and will lead to data distortion.

In integrating multiple batches, where different magnitude of batch effects and cell population imbalance are generally involved, NN-based methods except scMerge are recommended. In the challenging case when a significant number of cell types are missing in every batch (Figure 7**C**), none of methods achieved great performance, which either have poor batch mixing or blending cell types. FastMNN performs better than other methods. Harmony is quite flexible, so it might work if cell types are pre-known and thus parameters are adjusted manually to force the same cell types from different batches' mix (Table 1).

To be noted, Harmony, scVI, LIGER and fastMNN only provide corrected cell-level embeddings, which can be used for clusters or trajectories. Cell-level embeddings generated by LIGER, however, should be treated with caution since the global structure is not preserved well. Other methods produce corrected expression matrices. The corrected gene-expression values returned by Scanorama, however, have different ranges from the original data, which should be treated with caution (Table 1).

Discussion

We performed a comprehensive evaluation of noise reduction procedures for single-cell RNA-seq data using simulated and real datasets. We summarized our findings as a guideline for selection of suitable procedures in different scenarios. Special caution should be paid on technical and biological factors, including magnitude of batch effect, the complexity of cell group structures, the extent of cell population imbalance and the proportion and the similarity of nonoverlapping cell populations, which would greatly affect the performance of each batch correction method.

Although users can apply default settings to run analysis without knowing the parameters, the knowledge of underlying algorithms and assumptions would help find right settings for better performance. For example, the default parameters of Harmony led to undercorrection in our scenarios with strong batch effect. The undercorrection suggests that cluster diversity is not penalized enough. The increase of the parameter of 'theta' forces dataset mixing and leads to successful batch effects removal. The appropriate balance between cluster accuracy and diversity is critical for avoiding over- or undercorrection, which would be adjusted by the parameters theta and sigma. As another example, Seurat and scMerge use highly variable genes as default to calculate the reduced dimension vectors, whereas fastMNN and Scanorama depend on users to select highly variable features before batch correction. The variable genes should be selected from each batch separately to get only cell-type variation. The improper choice of highly variable genes would lead to the wrong match between cell types, resulting in the removal of true biological variations.

By default, Seurat uses CCA to find a common reduced dimensional space, which would lead to wrong mixing when there is severe cell populational imbalance. In this case, choosing another option 'reciprocal PCA' would solve the problem. Additionally, scMerge and Harmony require to set number of clusters in each batch, which greatly impact the performance and may need to be adjusted in certain scenarios.

The scenario with similar nonoverlapping cell populations across batches is the most challenging. Most methods removed the difference between similar nonoverlapping cell populations and mixed them together, leading to overcorrection. It is due to the fact that similar nonoverlapping cell populations are misidentified as MNNs in the reduced dimensional space. Considering not only MNNs but also their distances might help solve this issue. Most NN methods assume that the differences between cell types are greater than the differences between batch. When batch effect is too strong to dominate over cell-type differences, special cautions should be paid to avoid erroneous cell-type mixing. Linear-based methods are very sensitive to cell populational imbalance. The performance of linear-based methods could be improved under imbalanced cases if cell groups could be included into the model as covariates. scMerge employs this strategy to improve its performance, which first uses mutual nearest clusters to define cell groups and then keeps them in the model as wanted factors.

In our evaluations, the metrics can be used to distinguish correct from wrong mixing (cell-type mixing) since the cell types are known. Incorrect mixing of two different cell types from two batches would obtain very low scUnifrac_batch and ASW_batch scores. Similarly, $\mathrm{ARI}_{\mathrm{batch}}$ and $\mathrm{LISI}_{\mathrm{batch}}$ were used to quantify batch mixing, while $\mathrm{ARI}_{\mathrm{celltype}}$ and $\mathrm{LISI}_{\mathrm{celltype}}$ were used to evaluate cell-type mixing in a latest evaluation study [23]. In real practices, however, cell types are generally unknown, which make it very challenging to determine whether batch correction work properly or not. Although there are metrics developed for assessing single-cell RNA-seq batch correction, such as kBET [32], they only focus on batch mixing. They cannot tell the correct one (the same cell type from two batches mix) from an erroneous (different cell types from two batches mix) mixture. In the similar nonoverlapping scenario, for example, if we do not know B and C are not the same cell types and should not be mixed, Seurat, fastMNN and scMerge with a better batch mixing and a highly kBET score would be wrongly chosen rather than Scanorama. One feasible way is to analyze each dataset separately to gain some idea about cell types in each batch and then explore the batch-corrected results manually to decide whether there is any suspicious mixing. But it is time-consuming and requires a strong background. To automate the process, new quantitative metrics are needed for the evaluation of batch correct methods, which not only measure batch mixing but also evaluate imperceptible cell-type mixing.

Recent efforts to benchmark batch correction methods for single-cell RNA-seq all came to the conclusion that no single method emerges as the best performer in every dataset [23, 33, 34]. They also showed the importance of the selection of highly variable genes and the limitation of different types of outputs on the downstream analysis. Instead of recommending one or several methods, our studies focused on sorting out technical and biological factors that affect the performance of each method at the view of methodological assumptions and differences. For example, Seurat distorts the data when there is severe cell composition imbalance due to dimensional reduction by CCA. In this case, reciprocal PCA instead of CCA is recommended. Harmony was reported to succeed in two datasets but to fail in one dataset [30]. Our study, however, identified the importance to adjust its parameters to balance batch mixing and cluster accuracy. LIGER aligns cells incorrectly when technical noise and biological variances are too complicated to find shared metagenes across batches. Revealing limitations of existing methods provides guidance and directions how to improve. Our study also pointed out the urgent need of new metrics for evaluation since cell types are unknown in real applications.

Materials and methods

Studied scenarios

Scenarios with identical cell populations

We designed 23 scenarios to evaluate denoising procedures on two batches with identical cell populations. In scenarios 1–9, we generated nine simulation datasets by considering three levels of batch effects (none, mild or strong batch effect) and three levels of cell compositions imbalances (balanced, moderate imbalanced or severe imbalanced) using the R package Splatter [35]. The parameters of simulation studies are summarized in Supplementary Table S1 available online at http://bib.oxfordjournals.org/. In the mild batch effect setting, the batch explained the lower or comparative proportion of variation than cell groups. In the strong batch effect setting, in contrast, the batch explained the higher proportion of variation than cell groups. In the setting of balanced cell populations, two batches had the exact same cell population compositions. In the settings of moderate or severe imbalanced, however, two batches contained different cell population compositions. The ratio of two cell populations was 7:3 in one batch but was 3:7 in the other batch when moderate imbalanced, whereas the ratio of two cell populations was 9:1 in one batch but was 1:9 in the other batch when severe imbalanced.

In scenarios 10–15, we used real datasets to create six scenarios with mild or strong batch effects and three levels of cell compositions' imbalance (balanced, moderate and severe imbalanced). The setup of cell composition imbalances was the same with simulation studies mentioned above. The datasets with mild batch effects included two scRNAseq datasets generated by inDrops from human pancreatic islets (GEO Accession ID: GSE84133) [28]. The datasets with strong batch effects comprised two scRNAseq datasets from human pancreatic islets generated by different platforms. One was produced by Fluidigm C1(GEO Assession ID: GSE86469) [29], and the other was generated by SMART-seq2 (EBI-ArrayExpress Accession ID: E-MTAB-5061) [30]. Two cell types were extracted from the datasets, alpha and beta cells (details are summarized in Supplementary Table S1 available online at http://bib.oxfordjournals.org/).

In scenarios 16–23, we used PBMC datasets generated by multiple platforms [31], including two levels of batch effects (mild and strong) and four levels of cell compositions imbalance (balanced, mild, moderate and severe imbalanced). The scenario with mild batch effects consisted of eight cell types in two scRNAseq datasets generated by 10x v2 and v3 platforms, respectively (scenarios 16–19). The scenario with strong batch effects comprised seven cell types in two scRNAseq datasets from 10x v2 and inDrops platforms, respectively (scenario 20–23). All cells were included in the balanced design. The 30% of T cells (including CD4 T, CD8 T and NK) were randomly chosen in one batch, 30% of monocyte (including CD14 and CD16) were randomly selected in the other batch and other cells (cDC, B cells and Meg) were kept in the mild imbalanced design. In the moderate imbalanced design, 10% of T cells (including CD4 T, CD8 T and NK) were randomly chosen in one batch, 10% of monocytes (including CD14 and CD16) were randomly selected in the other batch and other cells (cDC, B cells and Meg) were kept in the moderate design. In the severe imbalanced design, 10% of CD4 T were randomly chosen in one batch, 10% of CD14 monocytes were randomly selected in the other batch and all the other cells were removed.

Scenarios with nonoverlapping cell populations

We laid out two cases with nonoverlapping cell populations in two batches. One is that the nonoverlapping cell population only exists in one batch, while the other is that each batch has one nonoverlapping cell population. In the first case, we designed 15 scenarios, including 9 scenarios based on simulated datasets by Splatter (scenarios 24–32) and 6 scenarios based on real datasets (scenarios 33–38). In the second case, we created four scenarios using real datasets (scenarios 39–42).

Scenarios 24–32 consisted of the combination of three levels of batch effect (none, mild and strong) and three types of nonoverlapping, i.e. a rare cell type, a cell type or a dominant cell type. Based on real datasets, scenarios 33–38 had the similar setting with simulated datasets without the no-batch effect setting (Supplementary Table S1 available online at http://bib.oxfordjournals.org/). In scenarios 39–42, each batch has one nonoverlapping cell population. In scenarios 39–42, we used datasets with mild batch effects, i.e. two samples from human pancreatic islet in the same study (GEO ID: GSE84133). The two nonoverlapping cell populations in two batches were very different in the scenario 39 (ductal and alpha cells), where one batch had beta and ductal cells, and the other batch had beta and alpha cells. In the scenario 32, the two nonoverlapping cell populations were similar (alpha and beta cells), where one batch had ductal and beta cells, and the other batch had ductal and alpha cells. Scenarios 41–42 were similar with scenarios 39–40 but with strong batch effects, where two datasets generated by different platforms were used (one by SMART-seq2 with ID E-MTAB-5061 and the other by Fluidigm C1with ID GSE86469) (Supplementary Table S1 available online at http://bib.oxfordjournals.org/). Those scenarios 33–42 were repeated 100 times by resampling cells.

Three scenarios of multiple batches with strong batch effect and cell population imbalance

We created three scenarios (scenarios 43–45) with six batches and nine cell types using PBMC datasets [31]. Both mild and strong batch effects were involved since datasets were generated by the same or different platforms, including two datasets from 10x v2 (A and B), four datasets from 10x v3, Dropseq, inDrops and seqWell, respectively. The nine cell types were B, CD4 T, CD8 T, NK, megakaryocytes, CD14 monocytes, CD16 monocytes, cDC and pDC. Three scenarios corresponded to three levels of cell group heterogeneity, where 0%, 25% or 50% of chances were that one cell type would be missing in a batch. The second and third scenarios were regenerated 20 times.

Ten scenarios with varying dropout rates and library sizes

We designed one scenario (scenario 46) to evaluate the performance of denoising procedures on recovering cell group structures obscured by highly variable library sizes. We used the same PBMC dataset [31]. Reads were down-sampled in each batch. Down-sampling proportion of each cell was determined by its library size relative to the maximum library size with a minimum down-sampling rate of 10% (90% of reads were discarded). Higher percentage of reads was removed in cells with smaller library sizes. For example, the cell with the maximum library size was not down-sampled. The cell with half of the maximum library size was down-sampled by 50%. Cells with <10% of the maximum library size were down-sampled by 10%.

High dropout and variable library sizes might obscure original cell/gene structure and pose challenges to noise reduction procedures. We designed nine scenarios (scenarios 47–55) to assess the performance of denoising procedures on recovering original data structures using simulation. The simulation datasets with mild batch effect and two balanced cell proportion were generated by Splatter, which included three levels of dropout rates from low, modest to high dropouts and also three levels of variable library size, ranging from lowly, modestly to highly variable. The details of parameters are summarized in Supplementary Table S1 available online at http://bib.oxfordjournals.org/.

Normalization and batch effect adjustment methods

Four normalization methods, ls, Scran, SCnorm and sctransform, were included in this study. ls and Scran estimated one scaling factor per cell and used it to normalize the expression. SCnorm and sctransform, in contrast, treat each gene in a cell differently by estimating multiple scaling factors for each cell. To be noted, SCnorm was not used in PBMC datasets due to memory issues. Six batch effect adjustment methods, including limma, ComBat, fastMNN, Scanorama, Seurat and Harmony, were considered in this study. Limma and ComBat belonged to linear-based methods, while fastMNN, Scanorama and Seurat were NN-based methods. Combining normalization with batch effect adjustment methods, we had in total of 24 noise reduction procedures. In addition, three methods handling normalization and batch correction together, ZINB-WaVE, scMerge and scVI, and one method, LIGER, with its custom preprocessing were included in the study as well. The detailed settings and the version of each method are summarized in Supplementary Table S2 available online at http://bib.oxfordjournals.org/.

Metrics for assessment

To assess denoising performance, we used three different methods, ASW [36], scUnifrac [37] distance and marker genes of each cell group. To measure batch mixing, ASW_batch and scUnifrac_batch were calculated. The metrics were calculated for each cell group sharing in all batches and then the mean values were obtained [Equation (1)]. Higher ASW_batch and scUnifrac_batch scores suggest better batch mixing. To measure cell groups separation, ASW_group and scUnifrac_group were calculated [Equation (2)]. The metrics were calculated for each batch and then the mean values were obtained. Higher ASW_group and scUnifrac_group scores suggest better cell group separation. The positive, none or negative changes of the ASW_group and scUnifrac_group values after batch correction compared to those before batch correction suggests successfully recovered, retained or disrupted cell group structures, respectively.

$$\text{ASW\_batch} = 1 - \sum\nolimits_{k\in \text{group}} s(b)^k / N_{\text{group}},$$

where $s(b)^k$ is the silhouette coefficient on batches for the group k.

$$\text{scUnifrac\_batch} = 1 - \sum\nolimits_{i<j,i,j\in \text{batch},k\in \text{group}} d(i,j)^k / [C(N_{\text{batch}}, 2)^* \, N_{\text{group}}], \qquad (1)$$

where $d(i, j)^k$ is the scUnifrac distance between two batches i and j for the group k, $C(N_{\text{batch}}, 2)$ is the number of combinations to select two from N_{batch}, N_{batch} is the number of batches and N_{group} is the number of cell groups.

$$\text{ASW\_group} = \sum\nolimits_{k\in \text{batch}} s(g)^k / N_{\text{batch}},$$

where $s(g)^k$ is the silhouette coefficient on groups for the batch k.

$$\text{scUnifrac\_group} = \sum\nolimits_{i<j,i,j\in \text{ group},k\in \text{batch}} d(i,j)^k / [C(N_{\text{group}}, 2)^* \, N_{\text{batch}}], \qquad (2)$$

where $d(i, j)^k$ is the scUnifrac distance between two groups i and j for the batch k, $C(N_{\text{group}}, 2)$ is the number of combinations to select two from N_{group}, N_{group} is the number of cell groups and N_{batch} is the number of batches.

Truly differentially expressed genes between cell groups were predefined in simulated datasets. limma was used to perform differential expression analysis. Genes with FDR-adjusted P-value < 0.05 were considered to be significantly different. TPR was defined as TP/T, and TNR was defined as TN/N. The positive, none or negative changes of TPR and TNR values after batch correction compared to those before batch correction suggest successfully recovered, retained, or disrupted gene structures, respectively.

Key Points

- Normalization and batch correction are critical steps in scRNA-seq data, which remove technical effects and systematic biases to unmask biological signal of interest.
- We perform a comprehensive evaluation of noise reduction procedures for single-cell RNA-seq data using simulated and real datasets.
- Our findings show that special caution should be paid on technical and biological factors, including magnitude of batch effect, the complexity of cell group structures, the extent of cell population imbalance and the proportion and the similarity of nonoverlapping cell populations, which would greatly affect the performance of each batch correction method.
- Our studies not only provide a comprehensive guideline for selecting suitable noise reduction procedures but also point out unsolved issues in the field, especially the urgent need of developing metrics for assessing batch correction on imperceptible cell-type mixing.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Funding

National Cancer Institute grants (U2C CA233291, U54 CA217450, SPORE in Gastrointestinal Cancer 5P50CA236733-02, SPORE in Breast Cancer 5P50CA098131-18); Cancer Center Support grant (P30CA068485).

References

1. Litvinukova M, Talavera-Lopez C, Maatz H, *et al*. Cells of the adult human heart. *Nature* 2020;**588**:466–72.
2. Pepe-Mooney BJ, Dill MT, Alemany A, *et al*. Single-cell analysis of the liver epithelium reveals dynamic heterogeneity and an essential role for YAP in homeostasis and regeneration. *Cell Stem Cell* 2019;**25**(23–38):e28.
3. Han X, Zhou Z, Fei L, *et al*. Construction of a human cell landscape at single-cell level. *Nature* 2020;**581**:303–9.
4. Fawkner-Corbett D, Antanaviciute A, Parikh K, *et al*. Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell* 2021;**184**:810–26 e823.
5. Reynolds G, Vegh P, Fletcher J, *et al*. Developmental cell programs are co-opted in inflammatory skin disease. *Science* 2021;**371**:eaba6500.
6. Park JE, Botting RA, Dominguez Conde C, *et al*. A cell atlas of human thymic development defines T cell repertoire formation. *Science* 2020;**367**:eaay3224.
7. Melms JC, Biermann J, Huang H, *et al*. A molecular single-cell lung atlas of lethal COVID-19. *Nature* 2021;**595**:114–9.
8. Vieira Braga FA, Kar G, Berg M, *et al*. A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat Med* 2019;**25**:1153–63.
9. Ramachandran P, Dobie R, Wilson-Kanamori JR, *et al*. Resolving the fibrotic niche of human liver cirrhosis at single-cell level. *Nature* 2019;**575**:512–8.
10. Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol* 2015; **11**:e1004333.
11. Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* 2016;**17**:75.
12. Bacher R, Chu LF, Leng N, *et al*. SCnorm: robust normalization of single-cell RNA-seq data. *Nat Methods* 2017;**14**:584–6.
13. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* 2019;**20**:296.
14. Ritchie ME, Phipson B, Wu D, *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47.
15. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;**8**:118–27.
16. Risso D, Perraudeau F, Gribkova S, *et al*. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun* 2018;**9**:284.
17. Haghverdi L, Lun ATL, Morgan MD, *et al*. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 2018;**36**:421–7.
18. Hao Y, Hao S, Andersen-Nissen E, *et al*. Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573–87 e3529.
19. Lin Y, Ghazanfar S, Wang KYX, *et al*. scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proc Natl Acad Sci U S A* 2019;**116**:9775–84.
20. Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat Biotechnol* 2019;**37**:685–91.
21. Polanski K, Young MD, Miao Z, *et al*. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* 2020;**36**:964–5.
22. Korsunsky I, Millard N, Fan J, *et al*. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 2019;**16**:1289–96.
23. Tran HTN, Ang KS, Chevrier M, *et al*. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* 2020;**21**:12.
24. Liu J, Gao C, Sodicoff J, *et al*. Jointly defining cell types from multiple single-cell datasets using LIGER. *Nat Protoc* 2020;**15**: 3632–62.
25. Lopez R, Regier J, Cole MB, *et al*. Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**:1053–8.
26. Eraslan G, Simon LM, Mircea M, *et al*. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun* 2019;**10**: 390.
27. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. *Nat Methods* 2019;**16**:715–21.
28. Baron M, Veres A, Wolock SL, *et al*. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intracell population structure. *Cell Syst* 2016;**3**:346–60 e344.
29. Lawlor N, George J, Bolisetty M, *et al*. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res* 2017;**27**: 208–22.
30. Segerstolpe A, Palasantza A, Eliasson P, *et al*. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab* 2016;**24**:593–607.
31. Ding J, Adiconis X, Simmons SK, *et al*. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat Biotechnol* 2020;**38**:737–46.
32. Buttner M, Miao Z, Wolf FA, *et al*. A test metric for assessing single-cell RNA-seq batch correction. *Nat Methods* 2019;**16**: 43–9.
33. Chazarra-Gil R, van Dongen S, Kiselev VY, *et al*. Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench. *Nucleic Acids Res* 2021;**49**:e42.
34. Luecken MD, Büttner M, Chaichoompu K, *et al*. Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* 2021. https://doi.org/10.1038/s41592-021-01336-8. Epub ahead of print. PMID: 34949812.
35. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol* 2017;**18**:174.
36. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987;**20**: 53–65.
37. Liu Q, Herring CA, Sheng Q, *et al*. Quantitative assessment of cell population diversity in single-cell landscapes. *PLoS Biol* 2018;**16**:e2006687.