**scAEQN: A Batch Correction Joint Dimension Reduction Method on scRNA-seq Data**

CONTENT

[Notes 2](#_Toc126011363)

[Supplementary note S1: human and mouse brain RNA dataset (brain) 2](#_Toc126011364)

[Supplementary note S2: mouse embryonic stem cell dataset (MESC) 2](#_Toc126011365)

[Supplementary note S3: mouse neurons dataset 2](#_Toc126011366)

[Supplementary note S4: human pancreas datasets from different donors (human pancreas1) 2](#_Toc126011367)

[Supplementary note S5: human pancreas dataset from different protocols (human pancreas2) 3](#_Toc126011368)

[Supplementary note S6: mouse pancreas dataset (mouse pancreas) 3](#_Toc126011369)

[Supplementary note S7: Experimental design 3](#_Toc126011370)

[Supplementary note S7: Adjusted Rand Index 5](#_Toc126011371)

[Supplementary note S8: Average Silhouette Width 5](#_Toc126011372)

[Supplementary note S9: Local Inverse Simpson’s Index 6](#_Toc126011373)

[Tables 6](#_Toc126011374)

[Supplementary Table S1. Summary of different batch correction methods 6](#_Toc126011375)

[Supplementary Table S2: parameter selection of scAEQN on mouse embryonic stem cells. 7](#_Toc126011376)

[Supplementary Table S3: parameter selection of scAEQN on mouse neuron dataset. 8](#_Toc126011377)

[Supplementary Table S4: parameter selection of scAEQN on human pancreas (different donors) dataset. 8](#_Toc126011378)

[Supplementary Table S5: parameter selection of scAEQN on mouse and human brain dataset. 9](#_Toc126011379)

[Supplementary Table S6: parameter selection of scAEQN on mouse pancreas dataset. 9](#_Toc126011380)

[Supplementary Table S7: parameter selection of scAEQN on human pancreas (different protocols) dataset. 10](#_Toc126011381)

[Supplementary Table S8: Statistics of the identified DEGs between alpha and beta cells on human pancreas1 dataset corrected by scAEQN based on 20 different sampling groups. 10](#_Toc126011382)

[Supplementary note S9: Co-expressed genes of uncorrected and different batch correction methods (MNN, Limma, Combat, scBatch, scAEQN) 11](#_Toc126011383)

[Figures 12](#_Toc126011384)

[Supplementary Figure S1: batch integration on mouse pancreas dataset. 12](#_Toc126011385)

[Supplementary Figure S2: batch integration on human pancreas2 dataset. 12](#_Toc126011386)

[Supplementary Figure S3: batch integration on brain dataset. 13](#_Toc126011387)

[Supplementary Figure S4: batch integration on mouse neuron dataset. 14](#_Toc126011388)

[Supplementary Figure S5: batch integration on human pancreas1 dataset. 14](#_Toc126011389)

[Supplementary Figure S6: the number of DEGs identified by datasets of different correct methods. 15](#_Toc126011390)

[Addition analysis 15](#_Toc126011391)

[Running time 15](#_Toc126011392)

# Notes

## Supplementary note S1: human and mouse brain RNA dataset (brain)

This dataset is the FPKM matrix data set of human and mouse brain cells generated by Zhang et al. It contains 15041 genes and 62 cell samples, including 41 human brain cells and 21 mouse brain cells. We regard different populations as batches of this dataset, i.e., human and mouse batches. Batch information has been provided, and cell types of all samples are also known, including many types of astrocytes, neurons, oligodendrocytes, endothelial cells and microglia cells.

## Supplementary note S2: mouse embryonic stem cell dataset (MESC)

This dataset was generated by Kim et al. The original dataset can be obtained in the Array Express database with accession number E-MTAB-2600. In this paper, we use the preprocessed dataset obtained from the public database of Hemberg group (https://hemberglab.github.io/scRNA

.seq.data-sets/), which contains 469 cell samples and 38616 genes. The dataset was generated by three different cell culture conditions, divided two batches to experimentation. Thus, cell types are marked by three different culture conditions set by the experimental author, and the batch information comes from cell culture carried out by the experimental author in two batches. The distribution of between batches and cell types is almost balanced.

## Supplementary note S3: mouse neurons dataset

This dataset was obtained by Usoskin et al. The original dataset can be obtained in the NCBI Gene Expression Omnibus (GEO) with accession ID GSE59739. Similarly, we obtained the preprocessed dataset from the public database of Hemberg group. The preprocessed dataset contains 622 cells and 25334 genes. The batch of the dataset comes from different preparation libraries. We also removed two batches with too small samples, and the final batches are 8 different preparation libraries. In addition, the true labels of the cells are determined according to the marker gene in the data, which is convenient for our model verification.

## Supplementary note S4: human pancreas datasets from different donors (human pancreas1)

This dataset was obtained by Xin et al, and the original dataset can also be obtained in NCBI GEO with accession ID GSE81608. The data we use is the preprocessed dataset obtained from the public database of Hemberg group. We also use the gene filtering method mentioned by Xin in the article. Genes with RPKM count >100 were retained in no less than 10 sample species. The dataset contains samples of healthy and diseased donors. Here, we only take the cells of healthy donors and regard different donors as batch effects. The final dataset contains 651 cells and 6797 genes. Cell types and real labels are also provided in the dataset.

## Supplementary note S5: human pancreas dataset from different protocols (human pancreas2)

The dataset contains two datasets measured by different sequencing technologies: one obtained by CEL-Seq sequencing technology (GSE81076) and the one obtained by CEL-Seq2 sequencing technology (GSE85241). These datasets can be obtained in NCBI GEO with accession number GSE81076 and GSE85241. By integrating two datasets from different protocols and marking sequencing tools of samples as batches, and filtering low expressions by Scanpy [23], a Python library for performing single-cell analysis, the final data contained 3289 cells and 12088 genes. Real cell labels are also available in the dataset.

## Supplementary note S6: mouse pancreas dataset (mouse pancreas)

The dataset is obtained by Baron et al, which can be obtained by NCBI GEO with accession number GSE84133. The dataset contains 4 human pancreas cells and 2 mouse pancreas cells. Here, we selected only pancreas cells from two different mouse populations, which represent different batches. It includes 1886 cells and 14878 genes, where are 13 real cell types that are also available in the dataset.

## Supplementary note S7: Experimental design

### Simulate experimental design

We applied Splatter [37] package to simulate scRNA-seq data. In dataset, gene expression counts for 10000 genes and 2000 cells were divided into four cell types in four batches. In order to make the characteristics of cell proportions of simulated data close to real scRNA-seq data, we varied the cell type proportions across the batches. Specifically, we set the proportion of four cell types as 0.4:0.3:0.2:0.1. For batch, we set the values of batch location as 0.1:0.2:0.05:0.15 and the values of batch scale as 0.05:0.1:0.25:0.3 in every cell type. Besides, the probability of outliers is set to 0.05 and dropout type is global variable.

### Batch correction and integration

In order to evaluate the performance of scAEQN, we select six public scRNA-seq datasets where the number of cells varies from tens to thousands, and different batch correction methods (Supplementary Note S3). Specific description and preprocessing work of datasets are in Supplementary Note S1-S6. Firstly, we evaluate the results of batch correction by low dimensional visualization method t-distributed stochastic neighbor embedding (t-SNE) [27]. We visualize the distribution of batches and cell types of the raw data, and corrected data obtained by MNN [8], Limma [3], Combat [2], QuantNorm [12], scBatch [15], Harmony [11], scVI [14], Scanorama [10], scETM [35] and scAEQN. To help to evaluate the effectiveness of scAEQN, we utilize some metrics that measure the level of batch integration quantitatively, such as adjusted rand index (ARI) [28], average silhouette width (ASW) [29], local inverse Simpson’s index (LISI) [30]. After obtaining the batch-corrected outputs, we compute the PCA vectors and use the top 10 PCs as inputs to calculate the LISI, ASW, and ARI scores respectively for assessing cell type purity and batch mixing separately. To obtain stable results, we repeat 20 times to obtain 20 metric scores, normalize their median score to range from 0 to 1, and then combine the assessments into an F1 score, as described in Supplementary Note S7-S9.

### Cluster analysis

Cluster analysis is performed on the batch corrected data. We repeat k-means clustering on the corrected count matrix or similarity matrix 10 times. Hyperparameter k of k-means selects the real number of cell types in each dataset. Then, we use ARI [28] to evaluate the consistency between real cell types and clustering labels by k-means. The closer the ARI value is to 1, the better the clustering result is. In this way, we can assess the performance of different batch correction methods through cluster analysis. For cluster analysis, we select six popular single-cell RNA-seq datasets (Table 1) and perform batch correction utilizing various batch correction methods. These methods are scBatch, scVI, scETM, BBKNN, Combat, Harmony, Limma, MNN, QuantNorm, and Scanorama. We utilize k-means to cluster the corrected expression data or similarity matrix on six datasets and to calculate the ARI scores. In addition, to ensure that the batch correction is effective, we also cluster the uncorrected raw matrix using the same clustering algorithm as a control group. Considering k-means clustering produces unstable clustering results, we run 10 times to obtain the average ARI value, which serves as the final evaluation metric

### DEGs analysis

DEGs analysis plays an important role in downstream analyses of scRNA-seq data. To perform DEG analysis, we use FindAllMarkers function from Seurat package [31], selecting a log2FC value greater than 0.05 as the threshold. We perform DEGs analysis on the gene expression matrix of raw data with batch effects and on the batch-corrected expression matrix by scAEQN, and then visualize the top 5 differential expression genes by heatmap function in ComplexHeatmap [32]. We assess the stability of DEGs identified by scAEQN corrected data and raw data through experimental design [37]. In addition, we also evaluate the feasibility of each batch correction method by comparing the running time on different number of cells. To further assess scAEQN, DEGs analysis is performed from the corrected data by scAEQN compared with the uncorrected original data. We utilize FindAllMarkers function in Seurat3 package to obtain DEGs on six datasets, where lilog2FC threshold is set to 0.05 because of the corrected data is normalized and log transformed in scAEQN. The top 5 DEGs are selected by ranking the value of log2FC in every cell type on six datasets, and then plotted heatmap by ComplexHeatmap [28] method (Figure 4B).

## Supplementary note S7: Adjusted Rand Index

The adjusted rand index (ARI) [26] can be employed to evaluate batch correction methods in terms of cell type purity and batch mixing. To assess cell type purity using ARI, the cell type labels were compared against the k-means clustering results using ARI as follows:

(1)

where represents the numbers of samples in which the true labels and the cluster labels overlapped, is the total number of samples,  is the total number of the class in the true labels, and  is the total number of the class in the cluster labels. A higher ARI corresponds to cluster cell block better. For batch mixing assessment, the ARI is calculated by their respective batch labels were compared to the k-means clustering labels, the low ARI represents that superior batch integration,

(2)

where is the total number of the batch in the batch information. To produce stable results, the computations for batch and cell type assessments were repeated 20 times each with random subsampling. The median ARI scores were then normalized to range between 0 and 1, and a combined F1 score was obtained for each batch correction method by computing the harmonic mean of the ARI scores:

(3)

A higher ARI F1 score will result from a lower ARI batch mixing score, and a higher ARI cell type mixing score.

## Supplementary note S8: Average Silhouette Width

The silhouette score of a data point is computed by subtracting its average distance to other members in the same cluster from its average distance to all members of the neighboring clusters, and then dividing by the larger of the two values. The resulting score ranges from − 1 to 1, where a high score denotes that the data point fits well in the current cluster, while a low score denotes a poor fit. The average score of all data points is used to measure overall cell type purity or batch mixing through the choice of labels. The average silhouette width (ASW) of cell type purity and batch mixing defines as follows respectively:

(4)

(5)

where denotes the number of samples,  represents the average distance between sample point and other samples of its same cluster class, represents the minimum average distance between sample point and the samples of other clusters, represents the average distance between sample point and other samples of its same batch class, represents the minimum average distance between sample point and the samples of other batches. To ensure the stable result, the median values of ASW were calculated and normalized for 20 times, and then calculated the harmonic mean of batch and cell type ASW scores to obtain the F1 score for assessing cell type purity and batch mixing integrally:

. (6)

A higher F1 score (that is, smaller ASW score for batch and higher ASW score for cell type) indicates superior batch correction.

## Supplementary note S9: Local Inverse Simpson’s Index

LISI is a local level evaluate metric, which can be used to assess batch and cell type mixing. LISI selects the nearest neighbors based on the local distribution of distances with a fixed perplexity. The selected neighbors are then used to compute the inverse Simpson’s index for diversity, which is the effective number of types present in this neighborhood. The Simpson’s index for diversity as follows:

(7)

where is the number of clusters, is the number of selected the nearest neighbors, and represents the proportion of the sample number of cluster to the total sample numbers. To assess cell type and batch mixing, we calculate LISI for cell type (cLISI) and batch (bLISI). For batch mixing, the bLISI score closer to the number of batches denotes good mixing, and cLISI close to 1 denotes that the clusters contain pure cell types. For combined assessment of cell type purity and batch mixing, the harmonic mean of cLISI and bLISI was computed to obtain the F1 score:

(8)

The F1 score is higher, and batch correction is better, likewise.

# Tables

## Supplementary Table S1. Summary of different batch correction methods

| Method | Language | Batch labels | Method-based | Batch-corrected output | Reference |
| --- | --- | --- | --- | --- | --- |
| MNN | R | Yes | Mutual nearest neighbor in gene expression space | gene expression matrix | [8] |
| Harmony | R | Yes | Iterative clustering in dimen-sionally reduced space | dimension reduction ma-trix | [11] |
| Combat | R | Yes | Empirical Bayesian frame-work | gene expression matrix | [2] |
| limma | R | Yes | Linear/empirical Bayes model | gene expression matrix | [3] |
| Seurat V3 | R | Yes | Canonical correlation analysis and MNN-anchors | gene expression matrix | [31] |
| BBKNN | Python/R | Yes | Batch balanced k-nearest neighbors | dimension reduction ma-trix and connective graph | [9] |
| QuantNorm | R | Yes | Interpolation quantile normali-zation | Similarity matrix | [12] |
| Scanorama | Python/R | Yes | MNN and panoramic stitching | gene expression matrix | [10] |
| scbatch | R | Yes | Iterative interpolation quantile normalization and matrix de-composition | gene expression matrix | [15] |
| scVI | R | Yes | Variational autoencoder | Dimension reduction ma-trix | [14] |
| scETM | R | Yes | Embedded topic model | Dimension reduction ma-trix | [35] |
| scAEQN | Python+R | Yes | QuantNorm and autoencoder | Gene expression matrix and dimension reduction matrix |  |

## Supplementary Table S2: parameter selection of scAEQN on mouse embryonic stem cells.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Loss function error | Learning rate | epoch | ARI | Loss function  error | Learning rate | epoch | ARI |
| 2.00E-03 | 1.00E-07 | 1847 | 0.4954 | 1.00E-01 | 5.00E-04 | 0 | 0.0279 |
| 2.00E-03 | 1.00E-06 | 1075 | 0.4553 | 5.00E-02 | 5.00E-04 | 0 | 0.0166 |
| 2.00E-03 | 8.00E-05 | 711 | 0.8380 | 1.00E-02 | 5.00E-04 | 30 | 0.5232 |
| 2.00E-03 | 6.00E-05 | 534 | 0.8059 | 8.00E-03 | 5.00E-04 | 31 | 0.4808 |
| 2.00E-03 | 4.00E-05 | 482 | 0.7585 | 6.00E-03 | 5.00E-04 | 43 | 0.9274 |
| 2.00E-03 | 2.00E-05 | 437 | 0.5328 | 4.00E-03 | 5.00E-04 | 47 | 0.5015 |
| 2.00E-03 | 1.00E-05 | 298 | 0.6574 | 2.00E-03 | 5.00E-04 | 98 | 0.7343 |
| 2.00E-03 | 8.00E-04 | 127 | 0.6116 | 1.00E-03 | 5.00E-04 | 345 | 0.6265 |
| 2.00E-03 | 6.00E-04 | 92 | 0.6125 | 8.00E-04 | 5.00E-04 | 482 | 0.3145 |
| 2.00E-03 | 4.00E-04 | 87 | 0.9293 | 5.00E-04 | 5.00E-04 | 5520 | 0.3910 |
| 2.00E-03 | 2.00E-04 | 139 | 0.5420 | 1.00E-04 | 5.00E-04 | 10000 | / |
| 2.00E-03 | 1.00E-04 | 254 | 0.5064 |  |  |  |  |
| 2.00E-03 | 1.00E-03 | 413 | 0.4958 |  |  |  |  |
| 2.00E-03 | 1.00E-02 | 699 | 0.5476 |  |  |  |  |

## Supplementary Table S3: parameter selection of scAEQN on mouse neuron dataset.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Loss function error | Learning rate | epoch | ARI | Loss function  error | Learning rate | epoch | ARI |
| 1.00E-02 | 1.00E-07 | 10000 | / | 1.00E-01 | 1.00E-04 | 9 | 0.2349 |
| 1.00E-02 | 1.00E-06 | 5428 | 0.4481 | 5.00E-02 | 1.00E-04 | 17 | 0.5724 |
| 1.00E-02 | 8.00E-05 | 161 | 0.4670 | 1.00E-02 | 1.00E-04 | 174 | 0.6533 |
| 1.00E-02 | 6.00E-05 | 216 | 0.5073 | 8.00E-03 | 1.00E-04 | 221 | 0.5340 |
| 1.00E-02 | 4.00E-05 | 274 | 0.4613 | 6.00E-03 | 1.00E-04 | 372 | 0.4462 |
| 1.00E-02 | 2.00E-05 | 505 | 0.6955 | 4.00E-03 | 1.00E-04 | 672 | 0.5096 |
| 1.00E-02 | 1.00E-05 | 880 | 0.6826 | 2.00E-03 | 1.00E-04 | 10000 | / |
| 1.00E-02 | 8.00E-04 | 131 | 0.4696 | 1.00E-03 | 1.00E-04 | 10000 | / |
| 1.00E-02 | 6.00E-04 | 108 | 0.4136 | 8.00E-04 | 1.00E-04 |  |  |
| 1.00E-02 | 4.00E-04 | 136 | 0.4890 | 5.00E-04 | 1.00E-04 |  |  |
| 1.00E-02 | 2.00E-04 | 146 | 0.5579 | 1.00E-04 | 1.00E-04 |  |  |
| 1.00E-02 | 1.00E-04 | 174 | 0.4863 |  |  |  |  |
| 1.00E-02 | 1.00E-03 | 321 | 0.4079 |  |  |  |  |
| 1.00E-02 | 1.00E-02 | 10000 | / |  |  |  |  |

## Supplementary Table S4: parameter selection of scAEQN on human pancreas (different donors) dataset.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Loss function error | Learning rate | epoch | ARI | Loss function  error | Learning rate | epoch | ARI |
| 1.00E-03 | 1.00E-07 | 10000 | / | 1.00E-01 | 5.00E-04 | 4 | 0.3732 |
| 1.00E-03 | 1.00E-06 | 10000 | / | 5.00E-02 | 5.00E-04 | 15 | 0.4092 |
| 1.00E-03 | 8.00E-05 | 1338 | 0.2572 | 1.00E-02 | 5.00E-04 | 56 | 0.3990 |
| 1.00E-03 | 6.00E-05 | 1603 | 0.3046 | 8.00E-03 | 5.00E-04 | 94 | 0.4769 |
| 1.00E-03 | 4.00E-05 | 1937 | 0.3103 | 6.00E-03 | 5.00E-04 | 110 | 0.4116 |
| 1.00E-03 | 2.00E-05 | 3631 | 0.3581 | 4.00E-03 | 5.00E-04 | 149 | 0.4092 |
| 1.00E-03 | 1.00E-05 | 2043 | 0.4347 | 2.00E-03 | 5.00E-04 | 240 | 0.4660 |
| 1.00E-03 | 8.00E-04 | 822 | 0.4276 | 1.00E-03 | 5.00E-04 | 945 | 0.4234 |
| 1.00E-03 | 6.00E-04 | 824 | 0.4661 | 8.00E-04 | 5.00E-04 | 1999 | 0.3067 |
| 1.00E-03 | 4.00E-04 | 863 | 0.4082 | 5.00E-04 | 5.00E-04 | / | / |
| 1.00E-03 | 2.00E-04 | 960 | 0.4091 | 1.00E-04 | 5.00E-04 | / | / |
| 1.00E-03 | 1.00E-04 | 1441 | 0.4464 |  |  |  |  |
| 1.00E-03 | 1.00E-03 | 865 | 0.4018 |  |  |  |  |

## Supplementary Table S5: parameter selection of scAEQN on mouse and human brain dataset.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| loss function error | learning rate | epoch | ARI | loss function error | learning rate | epoch | ARI |
| 8.00E-03 | 1.00E-07 | 5446 | 0.1883 | 1.00E-01 | 5.00E-04 | 0 | 0.1769 |
| 8.00E-03 | 1.00E-06 | 1157 | 0.2366 | 5.00E-02 | 5.00E-04 | 2 | 0.3625 |
| 8.00E-03 | 1.00E-05 | 212 | 0.3804 | 1.00E-02 | 5.00E-04 | 59 | 0.3567 |
| 8.00E-03 | 8.00E-04 | 45 | 0.4533 | 8.00E-03 | 5.00E-04 | 89 | 0.4196 |
| 8.00E-03 | 6.00E-04 | 53 | 0.4340 | 6.00E-03 | 5.00E-04 | 250 | 0.3917 |
| 8.00E-03 | 4.00E-04 | 65 | 0.4124 | 4.00E-03 | 5.00E-04 | / | / |
| 8.00E-03 | 2.00E-04 | 48 | 0.3567 | 2.00E-03 | 5.00E-04 | / | / |
| 8.00E-03 | 1.00E-04 | 45 | 0.2430 | 1.00E-03 | 5.00E-04 | / | / |
| 8.00E-03 | 8.00E-03 | 87 | 0.3034 | 8.00E-04 | 5.00E-04 | / | / |
| 8.00E-03 | 5.00E-03 | 83 | 0.4348 | 5.00E-04 | 5.00E-04 | / | / |
| 8.00E-03 | 2.00E-03 | 66 | 0.4417 | 1.00E-04 | 5.00E-04 | / | / |
| 8.00E-03 | 1.00E-03 | 48 | 0.3070 |  |  |  |  |
| 8.00E-03 | 5.00E-02 | 549 | 0.2666 |  |  |  |  |
| 8.00E-03 | 1.00E-02 | 159 | 0.3052 |  |  |  |  |

## Supplementary Table S6: parameter selection of scAEQN on mouse pancreas dataset.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| loss function error | learning rate | epoch | ARI | loss function error | learning rate | epoch | ARI |
| 5.00E-03 | 1.00E-07 | 10000 | / | 1.00E-01 | 5.00E-04 | 8 | 0.5677 |
| 5.00E-03 | 1.00E-06 | 7824 | 0.5368 | 5.00E-02 | 5.00E-04 | 13 | 0.4822 |
| 5.00E-03 | 1.00E-05 | 1633 | 0.5732 | 1.00E-02 | 5.00E-04 | 45 | 0.6006 |
| 5.00E-03 | 8.00E-04 | 74 | 0.7357 | 8.00E-03 | 5.00E-04 | 55 | 0.8797 |
| 5.00E-03 | 6.00E-04 | 85 | 0.7391 | 6.00E-03 | 5.00E-04 | 76 | 0.8498 |
| 5.00E-03 | 4.00E-04 | 121 | 0.7490 | 4.00E-03 | 5.00E-04 | 149 | 0.5493 |
| 5.00E-03 | 2.00E-04 | 137 | 0.8737 | 2.00E-03 | 5.00E-04 | 4786 | 0.5742 |
| 5.00E-03 | 1.00E-04 | 235 | 0.8111 | 1.00E-03 | 5.00E-04 | 7638 | 0.5477 |
| 5.00E-03 | 5.00E-03 | 159 | 0.5597 | 5.00E-04 | 5.00E-04 | 10000 | / |
| 5.00E-03 | 1.00E-03 | 67 | 0.6458 | 1.00E-04 | 5.00E-04 | 10000 | / |
| 5.00E-03 | 1.00E-02 | 2492 | 0.5497 |  |  |  |  |

## Supplementary Table S7: parameter selection of scAEQN on human pancreas (different protocols) dataset.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| loss function error | learning rate | epoch | ARI | loss function error | learning rate | epoch | ARI |
| 5.00E-03 | 1.00E-06 | 5000 | / | 1.00E-01 | 5.00E-04 | 3 | 0.4546 |
| 5.00E-03 | 8.00E-05 | 223 | 0.5329 | 5.00E-02 | 5.00E-04 | 6 | 0.4900 |
| 5.00E-03 | 6.00E-05 | 255 | 0.5522 | 1.00E-02 | 5.00E-04 | 40 | 0.5307 |
| 5.00E-03 | 4.00E-05 | 324 | 0.5375 | 8.00E-03 | 5.00E-04 | 47 | 0.5817 |
| 5.00E-03 | 2.00E-05 | 574 | 0.5607 | 6.00E-03 | 5.00E-04 | 59 | 0.5626 |
| 5.00E-03 | 1.00E-05 | 952 | 0.5532 | 4.00E-03 | 5.00E-04 | 110 | 0.5411 |
| 5.00E-03 | 8.00E-04 | 75 | 0.5860 | 2.00E-03 | 5.00E-04 | 517 | 0.5062 |
| 5.00E-03 | 6.00E-04 | 67 | 0.6025 | 1.00E-03 | 5.00E-04 | 10000 | / |
| 5.00E-03 | 4.00E-04 | 87 | 0.5320 | 5.00E-04 | 5.00E-04 | 10000 | / |
| 5.00E-03 | 2.00E-04 | 117 | 0.5073 | 1.00E-04 | 5.00E-04 | 10000 | / |
| 5.00E-03 | 1.00E-04 | 170 | 0.4917 |  |  |  |  |
| 5.00E-03 | 1.00E-03 | 70 | 0.5504 |  |  |  |  |
| 5.00E-03 | 1.00E-02 | 5000 | / |  |  |  |  |

We suggest that loss function threshold default is set 5e-3 and learning rate default is set 5e-4, then adjust the parameters for the specific dataset.

## Supplementary Table S8: Statistics of the identified DEGs between alpha and beta cells on human pancreas1 dataset corrected by scAEQN based on 20 different sampling groups.

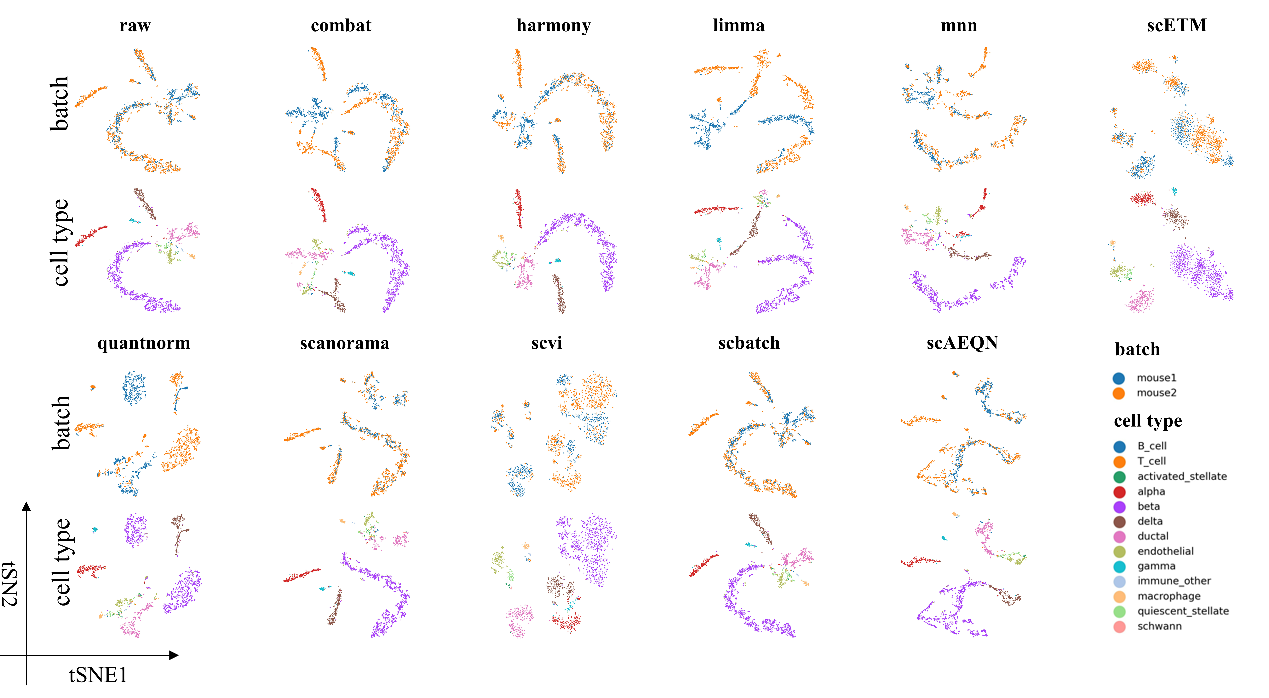
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Group** | **No. of DEGs** | **No. of DEGs also identified by other N sampling groups (N=)** | | | | | | | | | | | | | | | | | | | |
| **19** | **18** | **17** | **16** | **15** | **14** | **13** | **12** | **11** | **10** | **9** | **8** | **7** | **6** | **5** | **4** | **3** | **2** | **1** | **0** |
| **1** | 242 | **126** | 11 | 9 | 6 | 6 | 7 | 5 | 7 | 5 | 7 | 7 | 7 | 2 | 8 | 14 | 6 | 3 | 6 | 0 | 0 |
| **2** | 257 | **126** | 12 | 9 | 7 | 6 | 7 | 5 | 7 | 5 | 7 | 7 | 8 | 2 | 8 | 16 | 8 | 3 | 8 | 5 | 1 |
| **3** | 166 | **126** | 10 | 7 | 3 | 4 | 7 | 0 | 1 | 0 | 1 | 2 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| **4** | 192 | **126** | 11 | 8 | 7 | 5 | 4 | 5 | 4 | 4 | 7 | 5 | 4 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| **5** | 154 | **126** | 10 | 3 | 2 | 3 | 5 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| **6** | 232 | **126** | 12 | 9 | 7 | 6 | 11 | 5 | 7 | 5 | 6 | 7 | 7 | 3 | 6 | 9 | 3 | 0 | 3 | 0 | 0 |
| **7** | 157 | **126** | 9 | 6 | 3 | 3 | 3 | 4 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **8** | 246 | **126** | 12 | 9 | 7 | 7 | 11 | 5 | 7 | 5 | 8 | 10 | 8 | 2 | 7 | 9 | 7 | 3 | 1 | 1 | 1 |
| **9** | 245 | **126** | 12 | 9 | 7 | 7 | 11 | 5 | 7 | 5 | 4 | 6 | 5 | 3 | 3 | 6 | 4 | 3 | 10 | 3 | 9 |
| **10** | 174 | **126** | 10 | 7 | 3 | 5 | 6 | 5 | 5 | 2 | 2 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **11** | 294 | **126** | 12 | 9 | 7 | 7 | 11 | 5 | 7 | 5 | 8 | 10 | 9 | 3 | 8 | 17 | 10 | 5 | 17 | 11 | 7 |
| **12** | 279 | **126** | 12 | 9 | 7 | 7 | 11 | 5 | 7 | 5 | 8 | 8 | 8 | 3 | 8 | 17 | 9 | 5 | 14 | 7 | 3 |
| **13** | 191 | **126** | 12 | 8 | 7 | 5 | 5 | 5 | 7 | 3 | 7 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **14** | 238 | **126** | 12 | 9 | 7 | 6 | 10 | 5 | 7 | 5 | 7 | 7 | 7 | 2 | 6 | 10 | 5 | 3 | 2 | 1 | 1 |
| **15** | 204 | **126** | 12 | 9 | 7 | 7 | 11 | 5 | 7 | 4 | 4 | 7 | 2 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| **16** | 174 | **126** | 12 | 9 | 7 | 6 | 8 | 0 | 1 | 0 | 1 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **17** | 173 | **126** | 11 | 7 | 4 | 3 | 7 | 0 | 0 | 0 | 1 | 3 | 1 | 0 | 1 | 0 | 0 | 2 | 5 | 1 | 1 |
| **18** | 178 | **126** | 12 | 8 | 7 | 6 | 9 | 1 | 2 | 0 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **19** | 215 | **126** | 12 | 9 | 7 | 7 | 11 | 5 | 5 | 3 | 6 | 6 | 4 | 1 | 3 | 2 | 2 | 2 | 3 | 1 | 0 |
| **20** | 192 | **126** | 12 | 9 | 7 | 6 | 10 | 0 | 2 | 2 | 1 | 4 | 2 | 1 | 1 | 0 | 1 | 2 | 3 | 2 | 1 |

## Supplementary note S9: Co-expressed genes of uncorrected and different batch correction methods (MNN, Limma, Combat, scBatch, scAEQN)

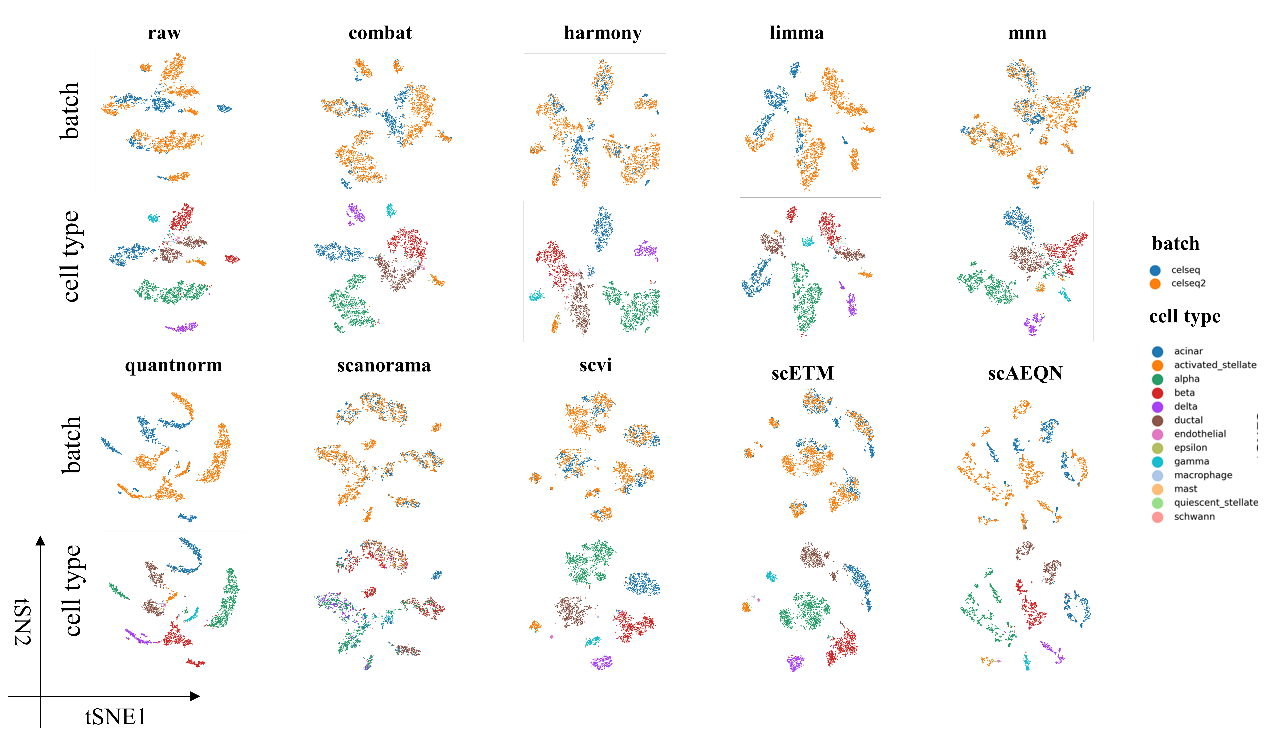
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **44261** | **ABCC8** | **ABHD15** | **AES** | **ARL6IP5** | **ARX** |
| **ASAP1** | **ATP6** | **BBX** | **BCHE** | **BTG3** | **C1orf127** |
| **CACNB2** | **CALD1** | **CD9** | **CFC1** | **CHGB** | **COTL1** |
| **COX2** | **CPNE8** | **CRH** | **CSGALNACT1** | **DLG5** | **DLK1** |
| **DNAJB9** | **DNAJC3** | **DPP4** | **EIF2S3** | **EIF3E** | **EIF4A2** |
| **ENAH** | **ENO1P1** | **ETNK1** | **EZR** | **F10** | **FABP5P7** |
| **FAM110B** | **FAM159B** | **FAM162A** | **FAM84A** | **FAP** | **FSTL5** |
| **FXYD5** | **G3BP1** | **GAD2** | **GNB2L1** | **GPX2** | **GRIA3** |
| **GSN** | **HIST1H4C** | **HLA-B** | **HPS1** | **HSP90AA1** | **HSPA8** |
| **IGF1R** | **IGFBP7** | **IGFBPL1** | **INS** | **INS-IGF2** | **KIAA1377** |
| **LINC00657** | **LOXL4** | **MAP1B** | **MEIS2** | **MGAT4A** | **MRC1** |
| **MT1X** | **MTND2P28** | **NACA** | **ND3** | **ND4L** | **ND5** |
| **NPNT** | **NUCB1** | **NUDT19** | **PALLD** | **PAPPA2** | **PDX1** |
| **PDZK1** | **PLCE1** | **PPM1H** | **PPP1CB** | **PPY** | **PRDX1** |
| **PRMT2** | **PRSS23** | **PRUNE2** | **PTPN3** | **RAB12** | **RAPH1** |
| **RASGEF1B** | **RBP4** | **RBP4** | **RFX6** | **RGS16** | **RGS4** |
| **RMND5A** | **RNASE4** | **ROBO2** | **RPL10** | **RPL12** | **RPL13A** |
| **RPL14** | **RPL15** | **RPL21P16** | **RPL21P75** | **RPL3** | **RPL31** |
| **RPL34** | **RPL35A** | **RPL36AP37** | **RPL5** | **RPL7** | **RPL7A** |
| **RPL7P** | **RPL7P19** | **RPL7P23** | **RPS10-NUDT3** | **RPS12** | **RPS18** |
| **RPS20** | **RPS27** | **RPS28** | **RPS4X** | **RPS6** | **RRAGD** |
| **S100A10** | **SCG3** | **SEPW1** | **SERPINI1** | **SH3RF1** | **SLC38A4** |
| **SLC6A6** | **SLC7A2** | **SNORA11E** | **SOD1** | **SORL1** | **SSR3** |
| **SSX2IP** | **ST13** | **SUMF2** | **TBC1D24** | **TBL1XR1** | **TGFBR3** |
| **THAP5** | **TIMP2** | **TMBIM4** | **TMED6** | **TMEM176B** | **TMEM33** |
| **TOB1** | **TPM3** | **TSPAN1** | **TSPYL1** | **TTR** | **TUBA1A** |
| **TUBA1C** | **UCHL1** | **VIM** | **VPS35** | **WLS** | **WSCD2** |
| **YWHAZ** | **ZBTB20** | **ZNF652** |  |  |  |

# Figures

## Supplementary Figure S1: batch integration on mouse pancreas dataset.



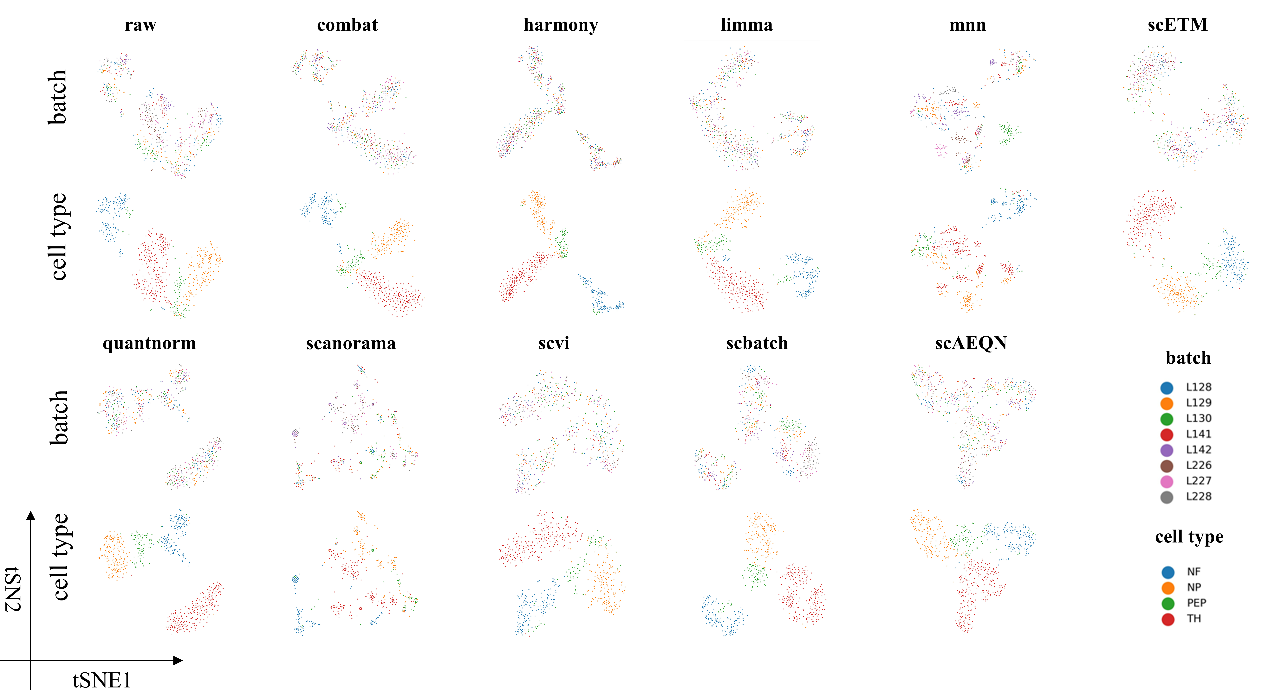
## **Supplementary Figure S2:** batch integration on human pancreas2 dataset.

****

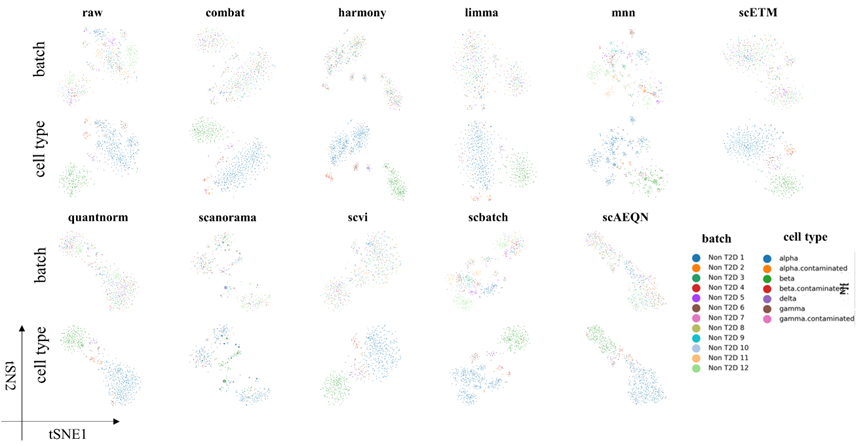
## Supplementary Figure S3: batch integration on brain dataset.

****

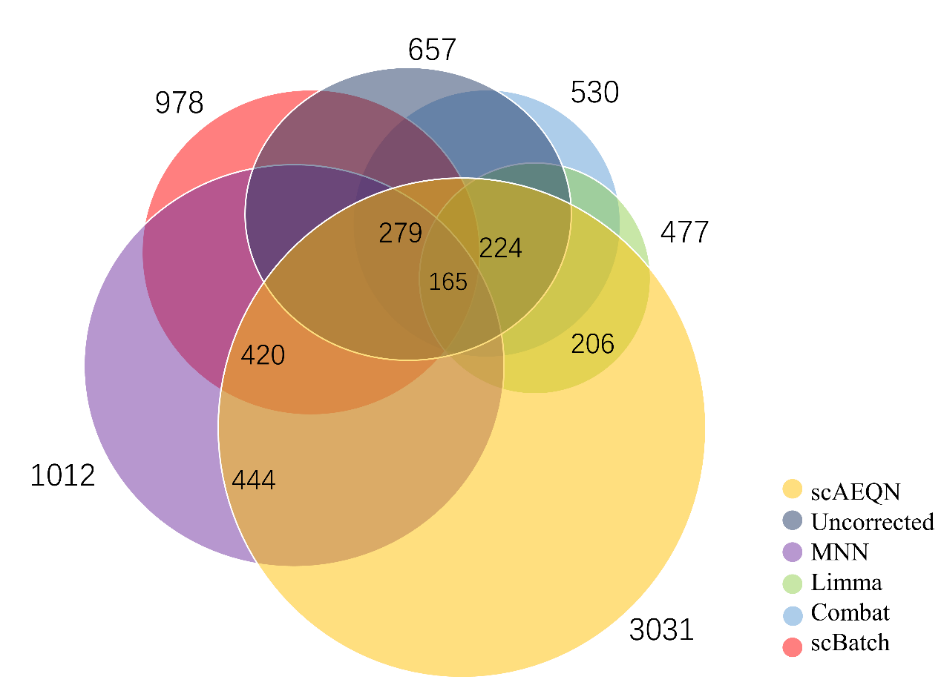
## Supplementary Figure S4: batch integration on mouse neuron dataset.

****

## Supplementary Figure S5: batch integration on human pancreas1 dataset.

****

## Supplementary Figure S6: the number of DEGs identified by datasets of different correct methods.



# Addition analysis

## Running time

To verify the operability of scAEQN, we record the running time in seconds of different batch correction methods under five datasets with different cell numbers varied from tens to thousands, and parameters setting are described in section 2.2.

As can be seen from Figure S7, the abscissa represents the number of cells arranged from small to large, the ordinate represents running time in seconds, and the lines of different colors represent the running time of different batch corrections methods. Figure S7A(b) is an enlarged plot of the line marked by the red circle in Figure S7A(a). There is a dotted line between 1886 and 3289 in Figure S7A(a), because the scBatch method takes 12 hours on a dataset of 3289 cells and does not get results. The running time of the scBatch method increases exponentially with the increase of the number of cells. MNN method also takes more time with the increase in the number of cells similarly, but the running time is still acceptable. Combat, Limma and BBKNN are stable in running time. The running time of scAEQN is related to the parameters of the selected QuantNorm. When the parameters of each dataset in scAEQN are set to “method=vectorization”, the running time of our method will increase with the growth of the number of cells, but the longest time is only a few minutes. For instance, on the dataset with 651 cells, when the parameter is set to “method =row/column”, the running time of scAEQN is only six minutes.

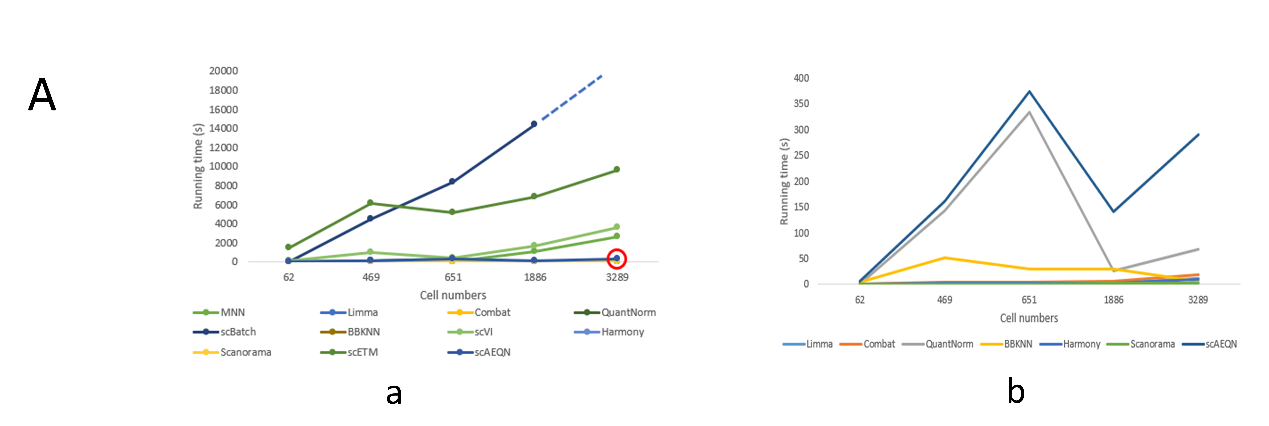


Figure S7: A: running time of different batch correction methods on different numbers of cells. (a): running time of eleven batch correction methods. (b): seven methods of overlapped part marked by the red circle in C(a).