Supplementary Information for

scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets

Yingxin Lin, Shila Ghazanfar, Kevin Y. X. Wang, Johann A. Gagnon-Bartsch, Kitty K. Lo, Xianbin Su, Ze-Guang Han, John T. Ormerod, Terence P. Speed, Pengyi Yang and Jean Yee Hwa Yang

Jean Yee Hwa Yang, Pengyi Yang
E-mail: J.Y.H.Y. jean.yang@sydney.edu.au and P.Y. pengyi.yang@sydney.edu.au

This PDF file includes:

Figs. S1 to S17
Tables S1 to S2
Captions for Databases S1 to S2

Other supplementary materials for this manuscript include the following:

Databases S1 to S2

www.pnas.org/cgi/doi/10.1073/pnas.1820006116

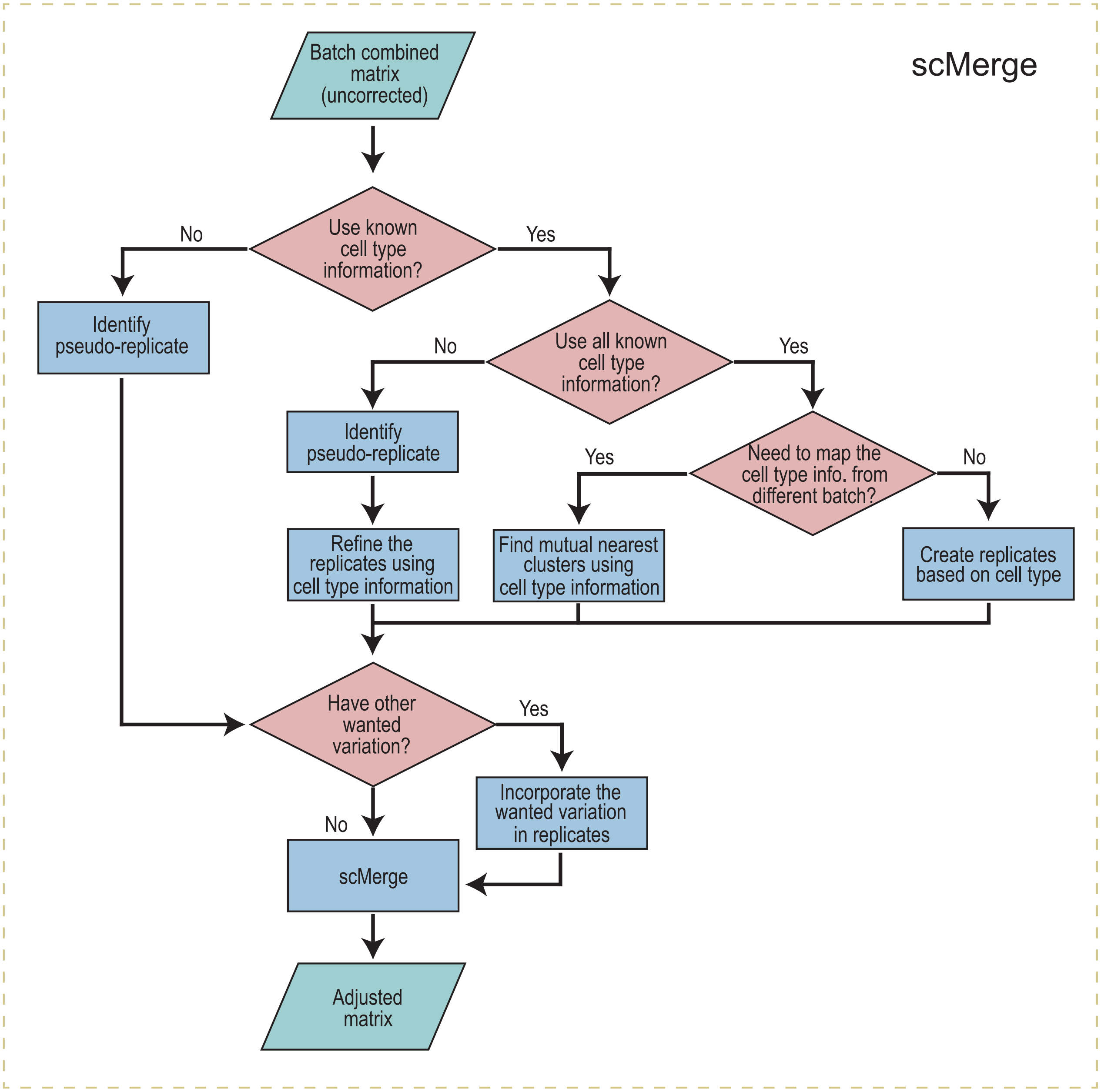

scMerge
Batch combined matrix (uncorrected)
Use known cell type information?
No
Yes
Identify pseudo-replicate
Use all known cell type information?
No
Yes
Identify pseudo-replicate
Need to map the cell type info. from different batch?
Yes
No
Refine the replicates using cell type information
Find mutual nearest clusters using cell type information
Create replicates based on cell type
Have other wanted variation?
Yes
Incorporate the wanted variation in replicates
No
scMerge
Adjusted matrix

Fig. S1. Flow chart illustrating the decision-making process associated with scMerge algorithm.

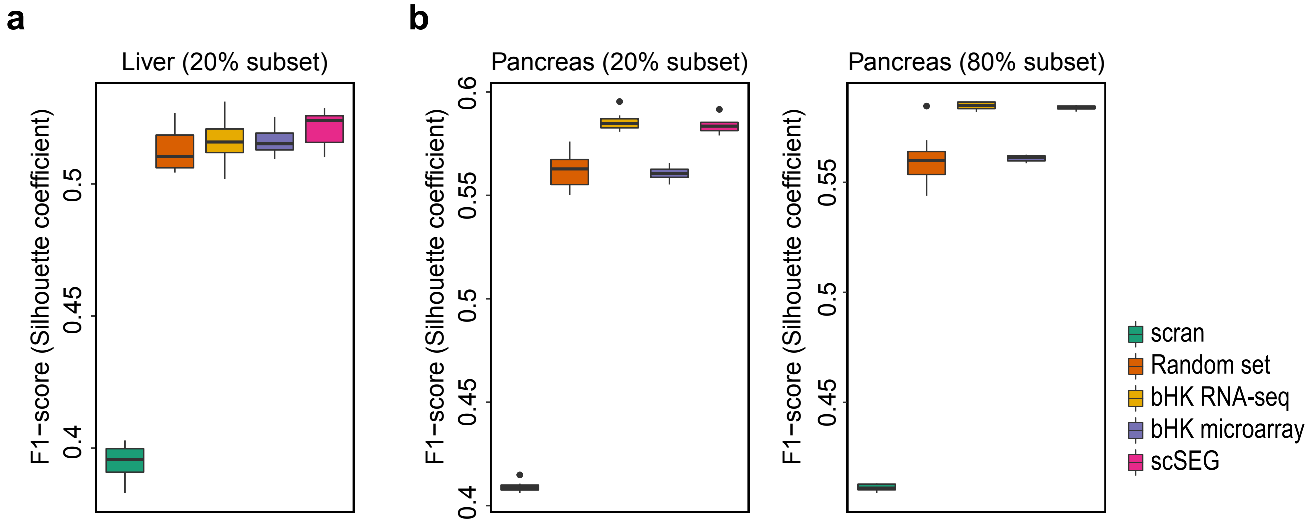


Fig. S2. A 1 by 3 panel of boxplots comparing the effect of different types of negative controls for (a) liver datasets and (b) four Pancreas datasets. The y-axis represents the F1score of Silhouette coefficients between cell type mixing and (1 - datasets mixing). Stratified sampling is performed to randomly subset 20% and 80% of cells from the datasets. This procedure is repeated 10 times. The boxplots represent the F1 score results using logcounts, RUVg using random subset of genes, bulk microarray, RNA-Seq data, and scSEG as negative control genes based on subset of the datasets.

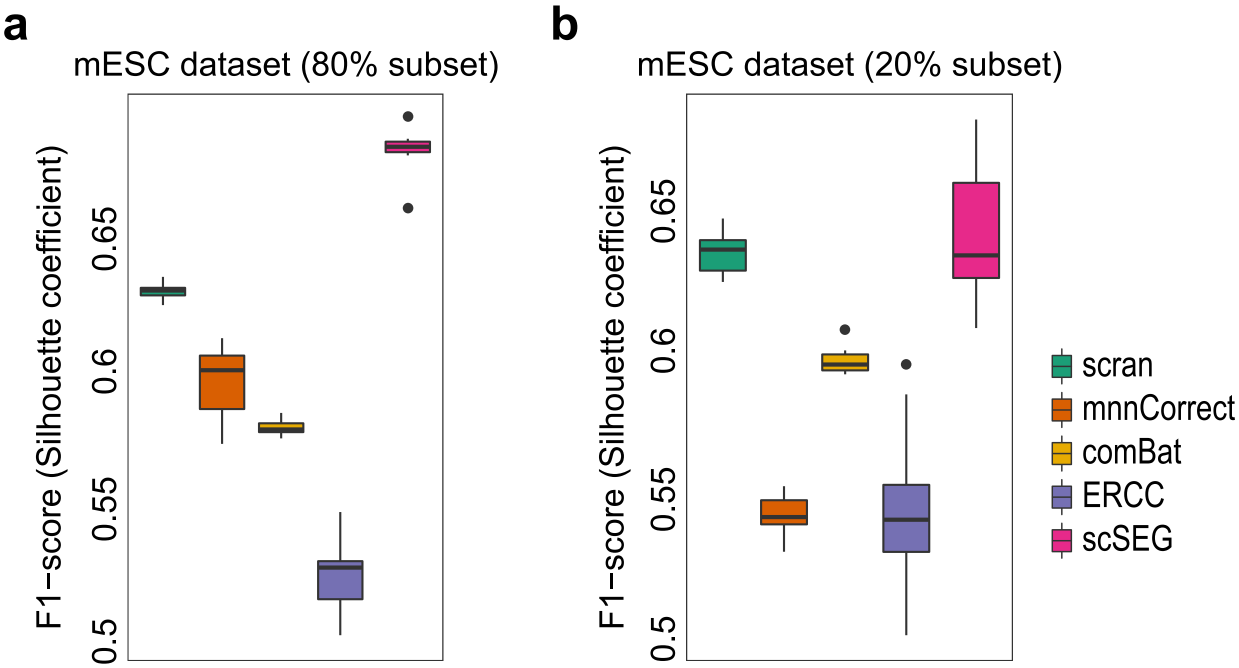


Fig. S3. A 1 by 2 panel of boxplots showing the effect of using ERCC and scSEG as negative controls with scMerge. Stratified subsampling is performed for the mESC dataset. In each stratified subsampling, we randomly selected 20% (left panel) or 80% (right panel) of the cells from the dataset and perform scMerge with ERCC spike-ins genes and mouse scSEG, and comparing the results with ComBat and mnnCorrect (default settings). This procedure is repeated 10 times.

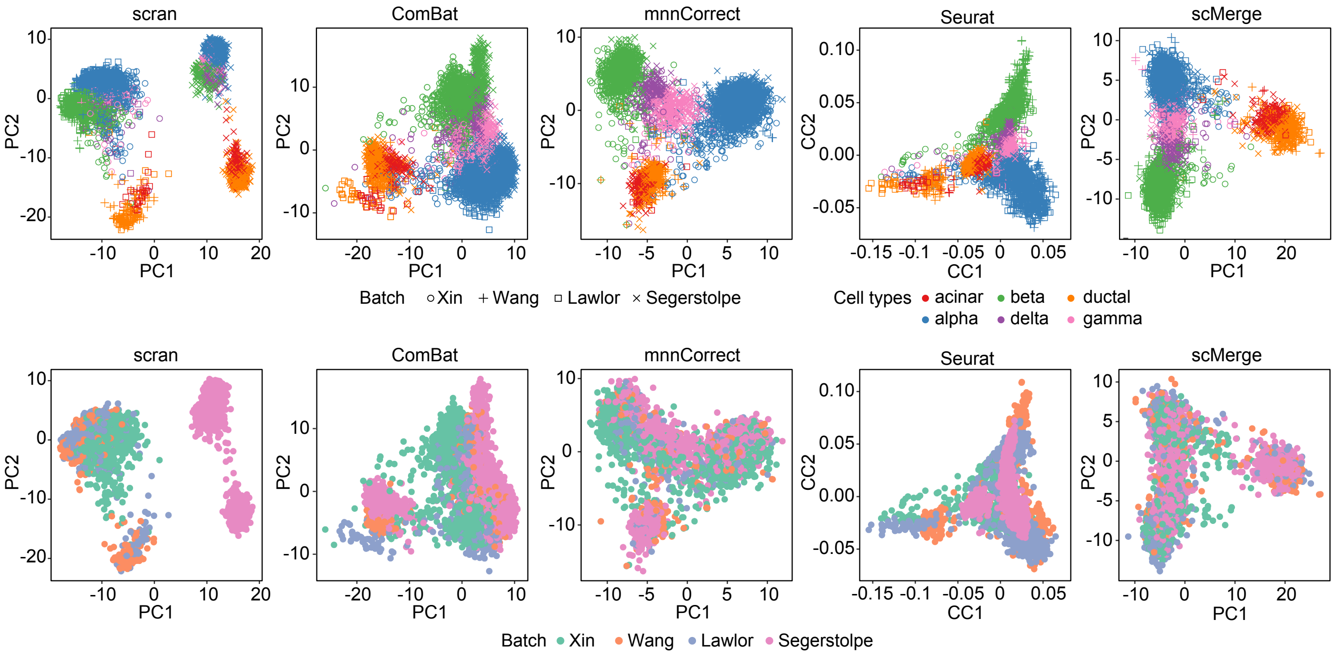


Fig. S4. A 2 by 5 panel of PCA plots of the four Pancreas datasets using the output from scran (logcounts), ComBat, mnnCorrect, Seurat, and scMerge (using scSEG as negative controls). The top row of the panel is color coded by cell types and the second row is color coded by the four different datasets.

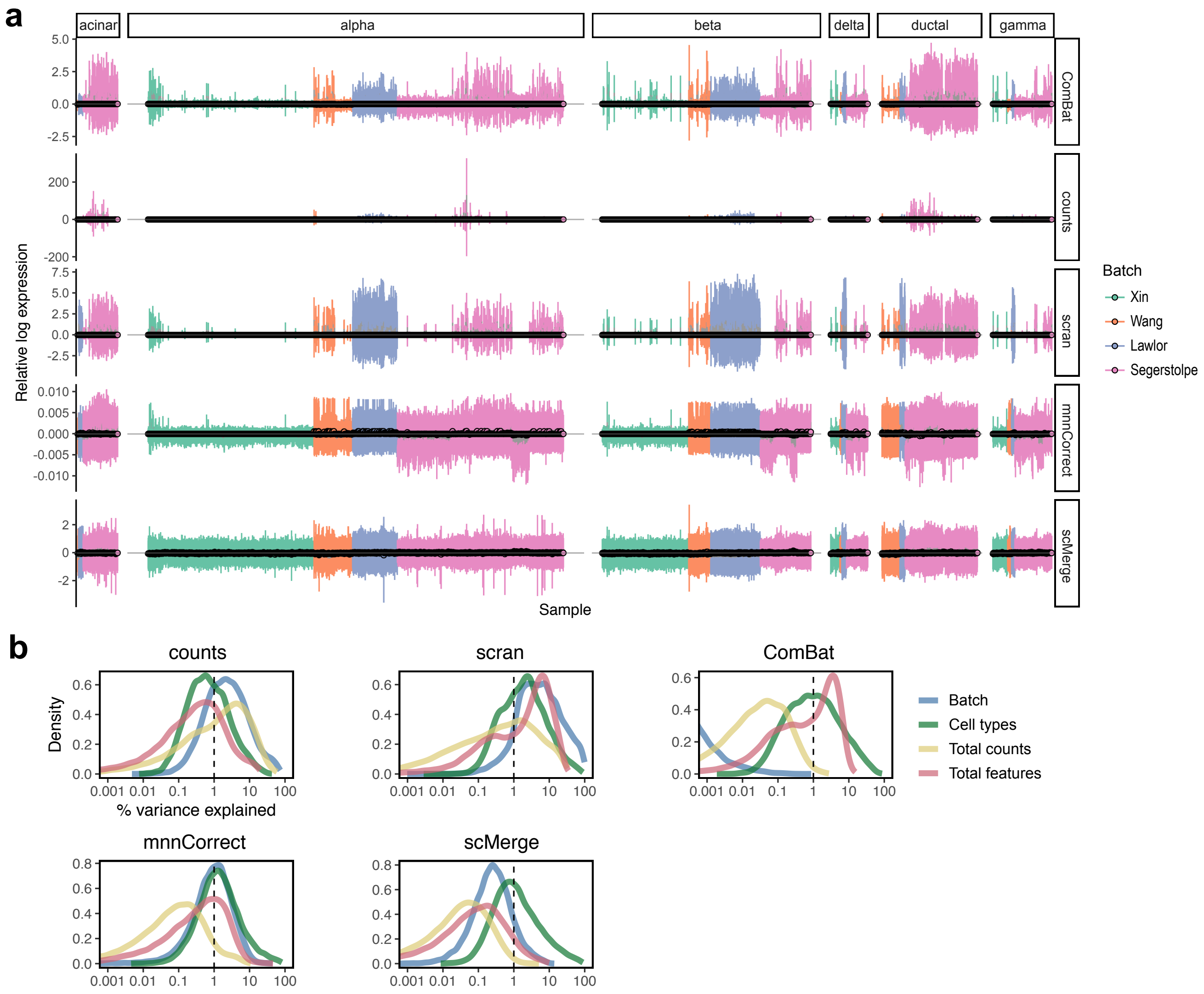


Fig. S5. Diagnostic plots from the Pancreas data collections ("Pancreas4"). A. RLE plots. The boxplot for each cell from the same cell type between different batches of scMerge shares similar inter quantile ranges. B. Percentage of variance explained for each variable. scMerge has cellType explaining the highest percentage of variation for the liver data collection.

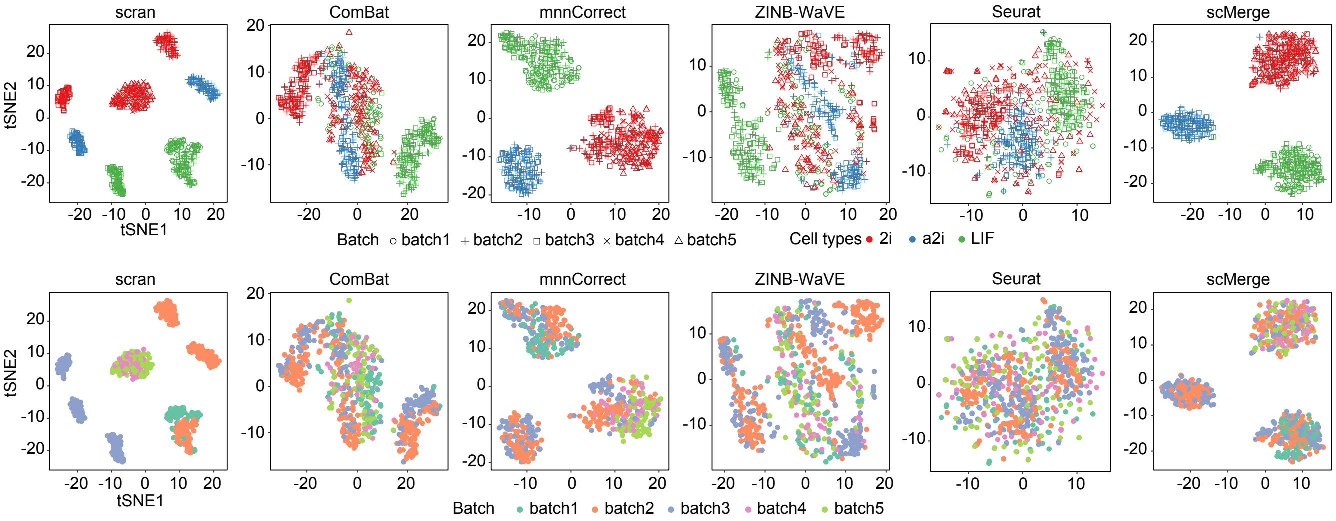


Fig. S6. A 2 by 6 panel of tSNE plots of the mESC datasets using the output from scran (logcounts), ComBat, mnnCorrect, ZINB-WaVE, Seurat, and scMerge (using scSEG as negative controls). The top row of the panel is color coded by three cell types and the second row is color coded by the three batches.

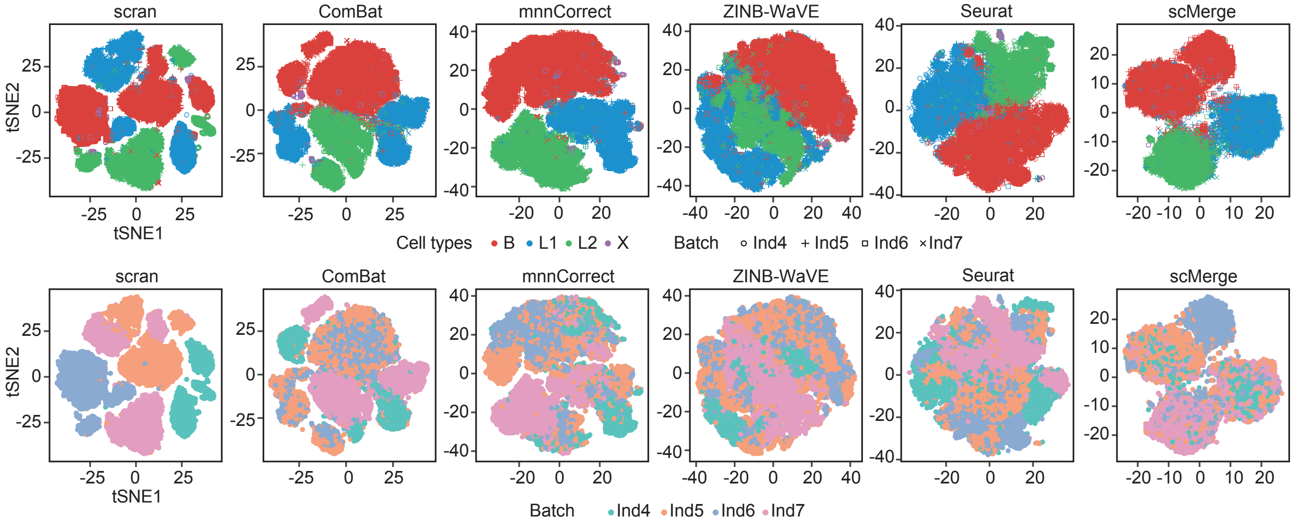


Fig. S7. A 2 by 6 panel of tSNE plots of the Breast Cancer data using the output from scran (logcounts), ComBat, mnnCorrect, ZINB-WaVE, Seurat, and scMerge (using scSEG as negative controls). The top row of the panel is color coded by cell types and the second row is color coded by the four individuals.

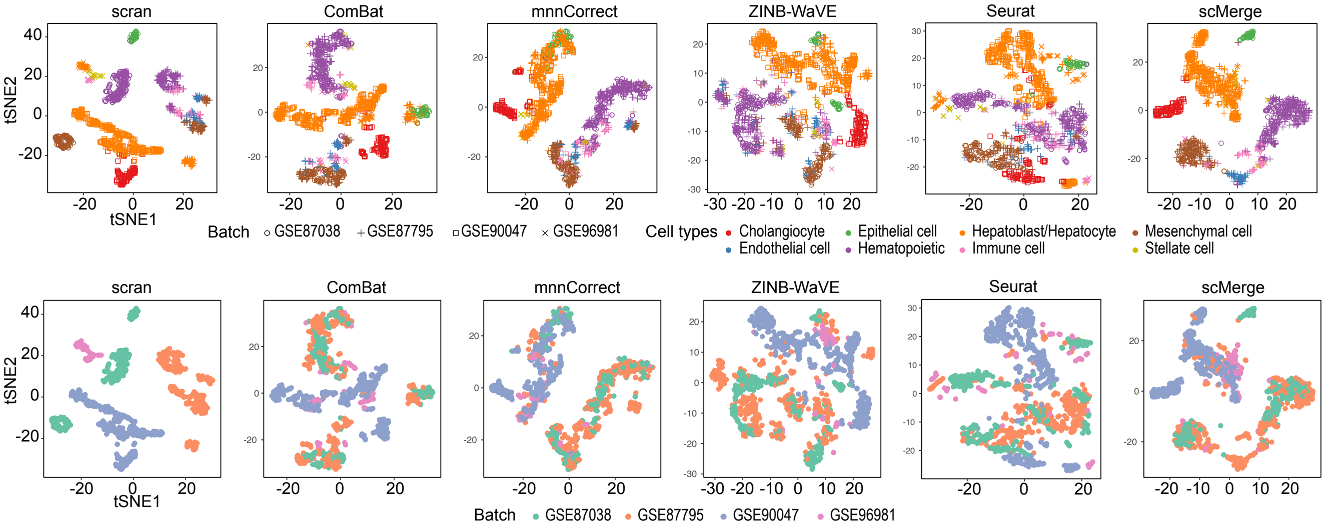


Fig. S8. A 2 by 6 panel of tSNE plots of the Liver data collection using the output from scran (logcounts), ComBat, mnnCorrect, ZINB-WaVE, Seurat, and scMerge (using scSEG as negative controls). The top row of the panel is color coded by cell types and the second row is color coded by the four datasets.

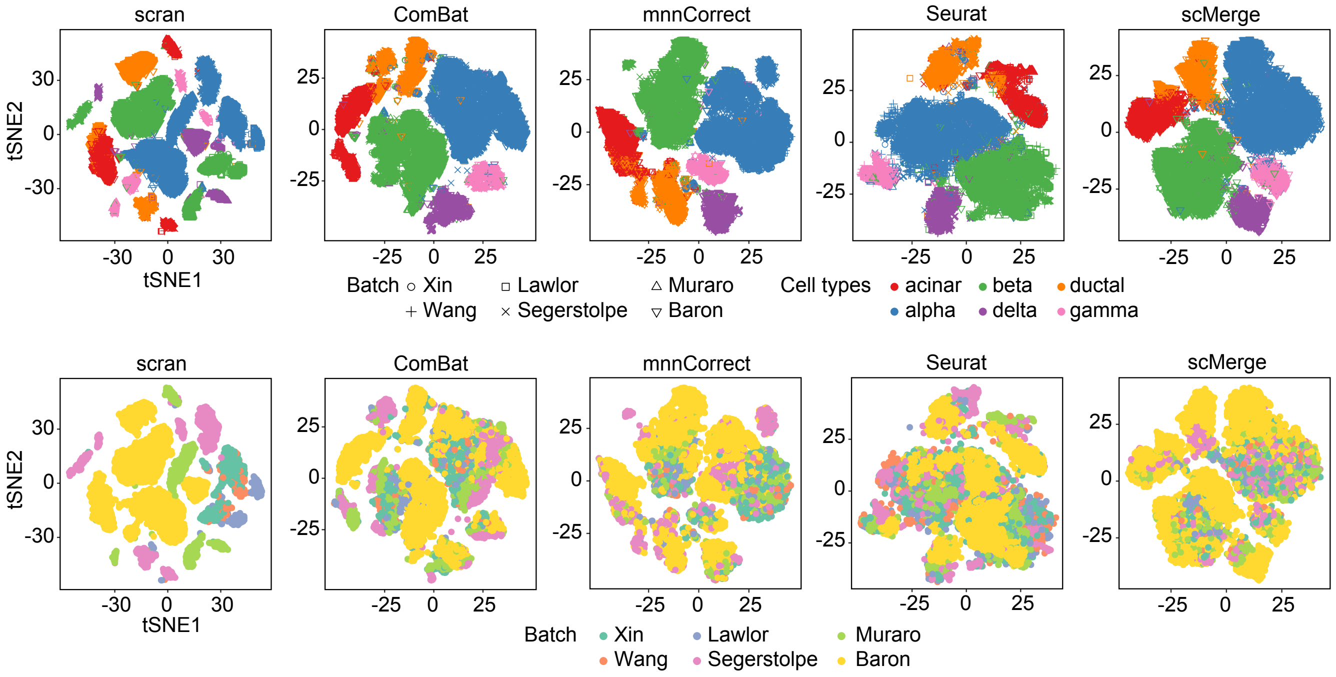


Fig. S9. A 2 by 6 panel of tSNE plots of the Olfactory data collection using the output from scran (logcounts), ComBat, mnnCorrect, ZINB-WaVE, Seurat, and scMerge (using scSEG as negative controls). The top row of the panel is color coded by cell types and the second row is color coded by the two datasets.

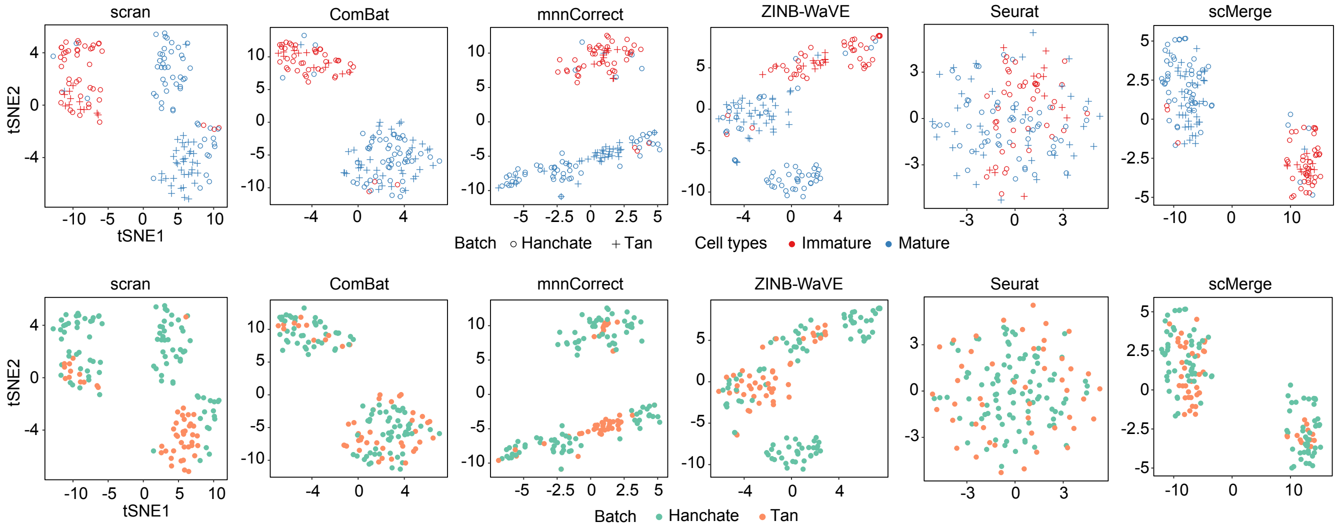


Fig. S10. A 2 by 5 panel of tSNE plots of all six Pancreas related datasets based on the output from scran (logcounts), ComBat, mnnCorrect, Seurat, and scMerge (using scSEG as negative controls). The top row of the panel is color coded by cell types and the second row is color coded by the six different datasets.

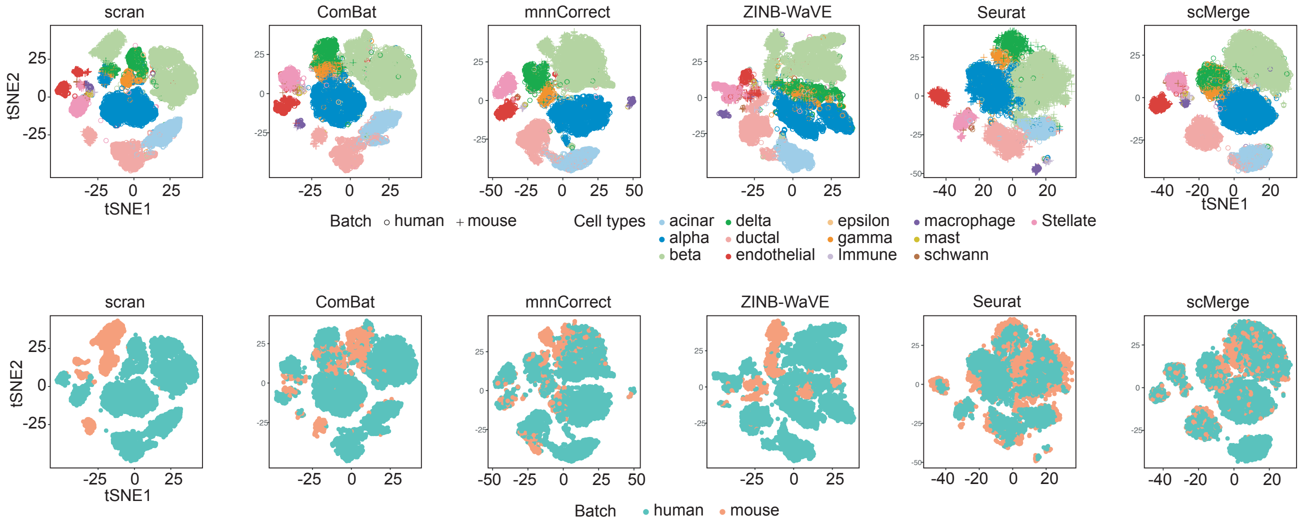


Fig. S11. A 2 by 6 panel of tSNE plots of the Pancreas Islet data collection using the output from scran (logcounts), ComBat, mnnCorrect, ZINB-WaVE, Seurat, and scMerge (using scSEG as negative controls). The top row of the panel color coded by cell types and the second row is color coded by the six datasets.

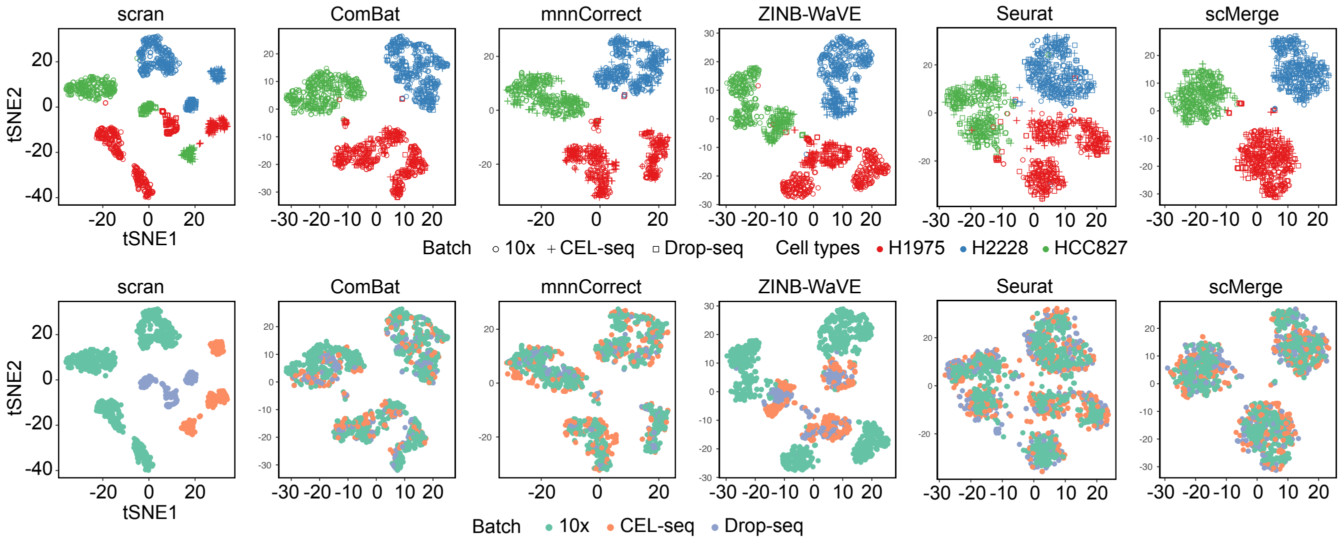


Fig. S12. A 2 by 6 panel of tSNE plots of the CellBench data using the output from scran (logcounts), ComBat, mnnCorrect, ZINB-WaVE, Seurat, and scMerge (using scSEG as negative controls). The top row of the panel is color coded by cell lines and the second row is color coded by the three different platforms.

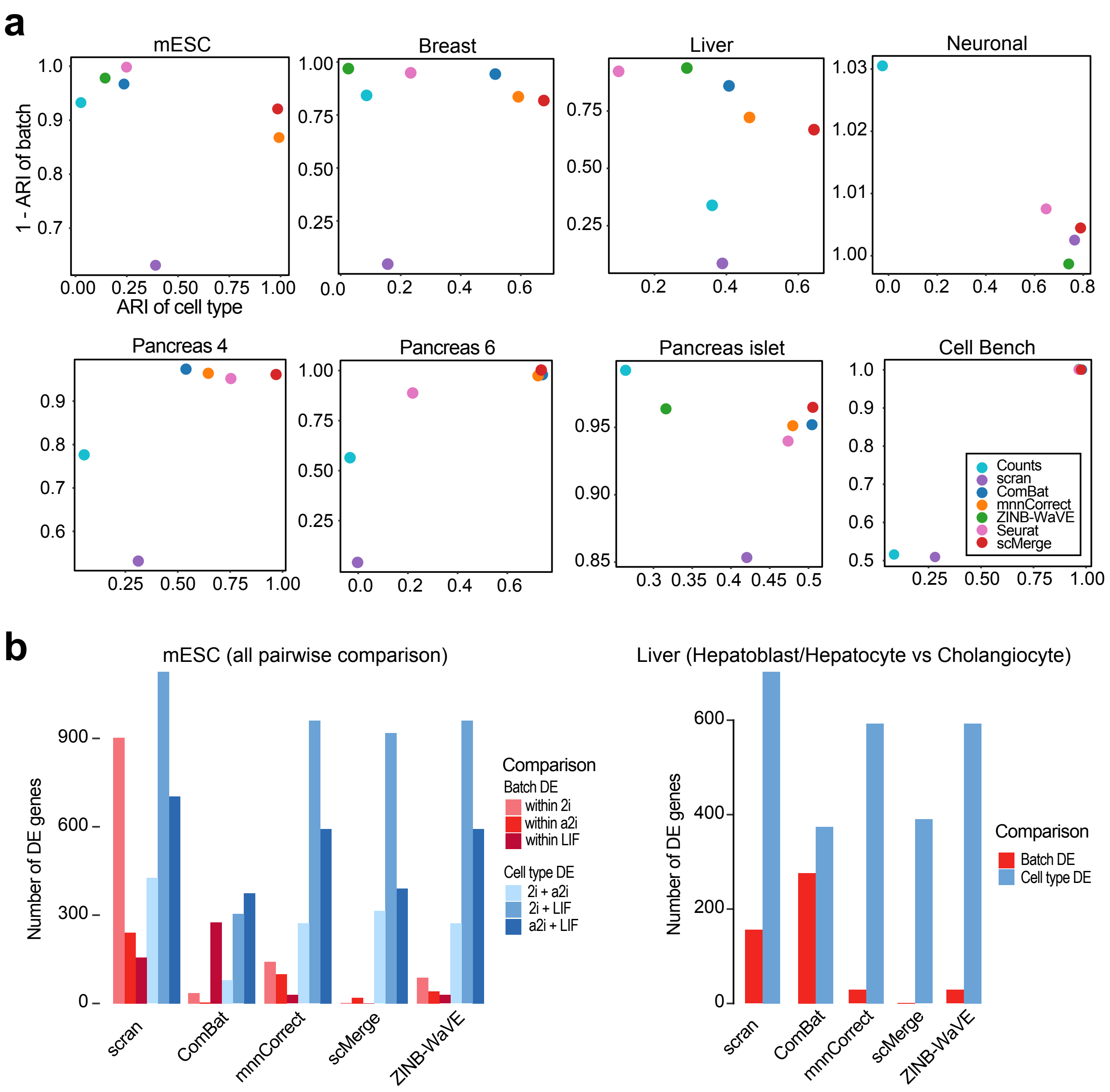


Fig. S13. (a) A 2 by 4 panel of scatter plots of ARI evaluation for no normalization, scran (logcounts), ZINB-WaVE, ComBat, mnnCorrect, Seurat, and scMerge (using scSEG as negative controls) of eight dataset collections (mESC dataset, breast). x-axes denote the ARI of cell types and y-axes denote the 1 – ARI of batch effects, where desirable outcomes are in the top-right hand corner. (b) A 1 by 2 panel of bar plots of differential expression results of mESC dataset (left panel) and liver dataset collection (right panel) on scran, ComBat, mnnCorrect, scMerge and ZINB-WaVE normalized data. The y-axis indicates the number of DE genes that are selected (adjusted p-value < 0.05 and logFC > 2). Batch DE are selected by performing the DE analysis within the same cell type, while cell type DE are selected by performing the DE analysis between two different cell types.

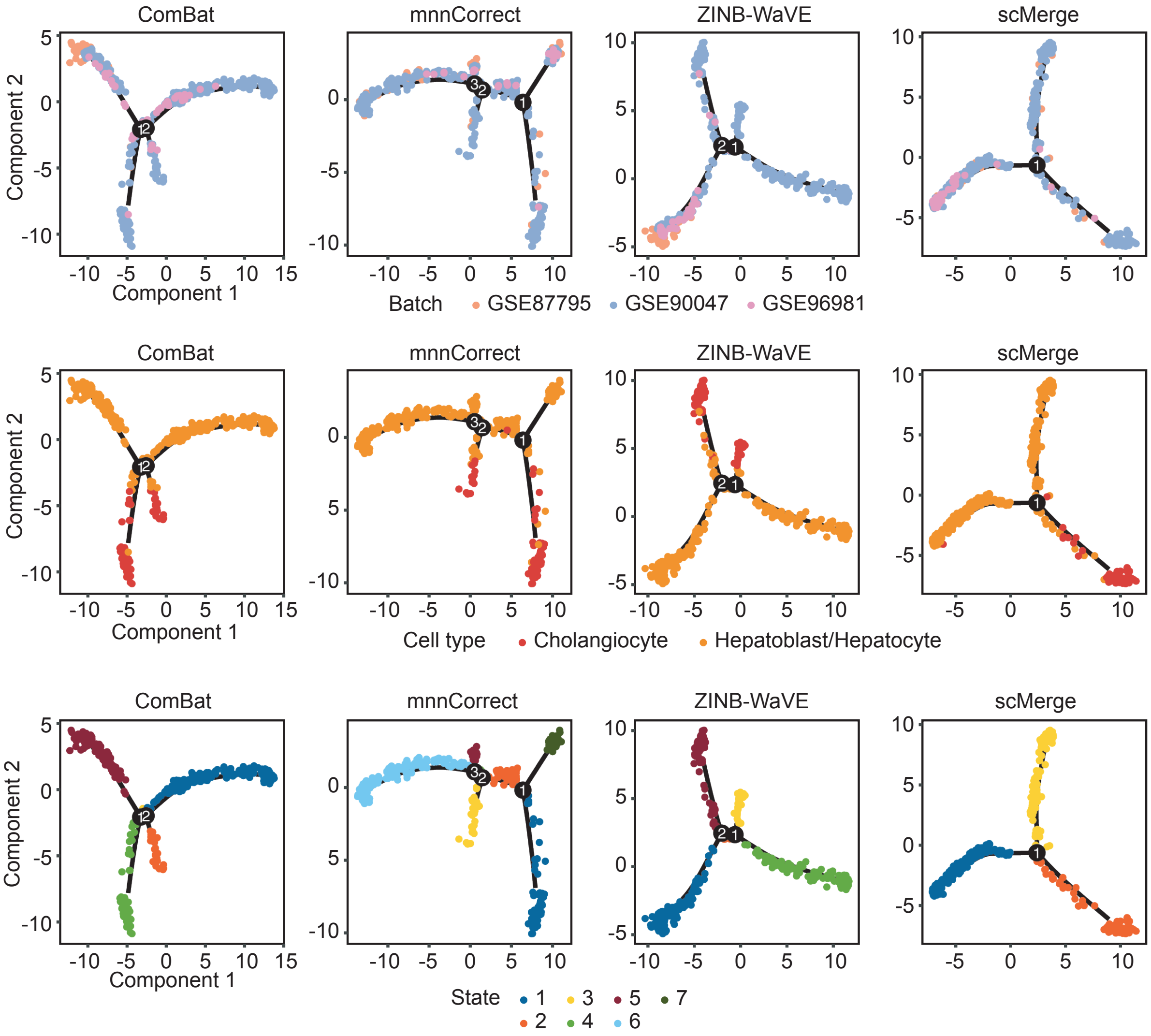


Fig. S14. A 3 by 4 panel of pseudotime trajectory from Monocle 2 with perturbed data demonstrating the stability of scMerge with output from ComBat, mnnCorrect, ZINB-WaVE and scMerge. The first row of the panel color coded by three datasets; the second row is color coded by the cell types and the third row is color coded by Monocle 2 states.

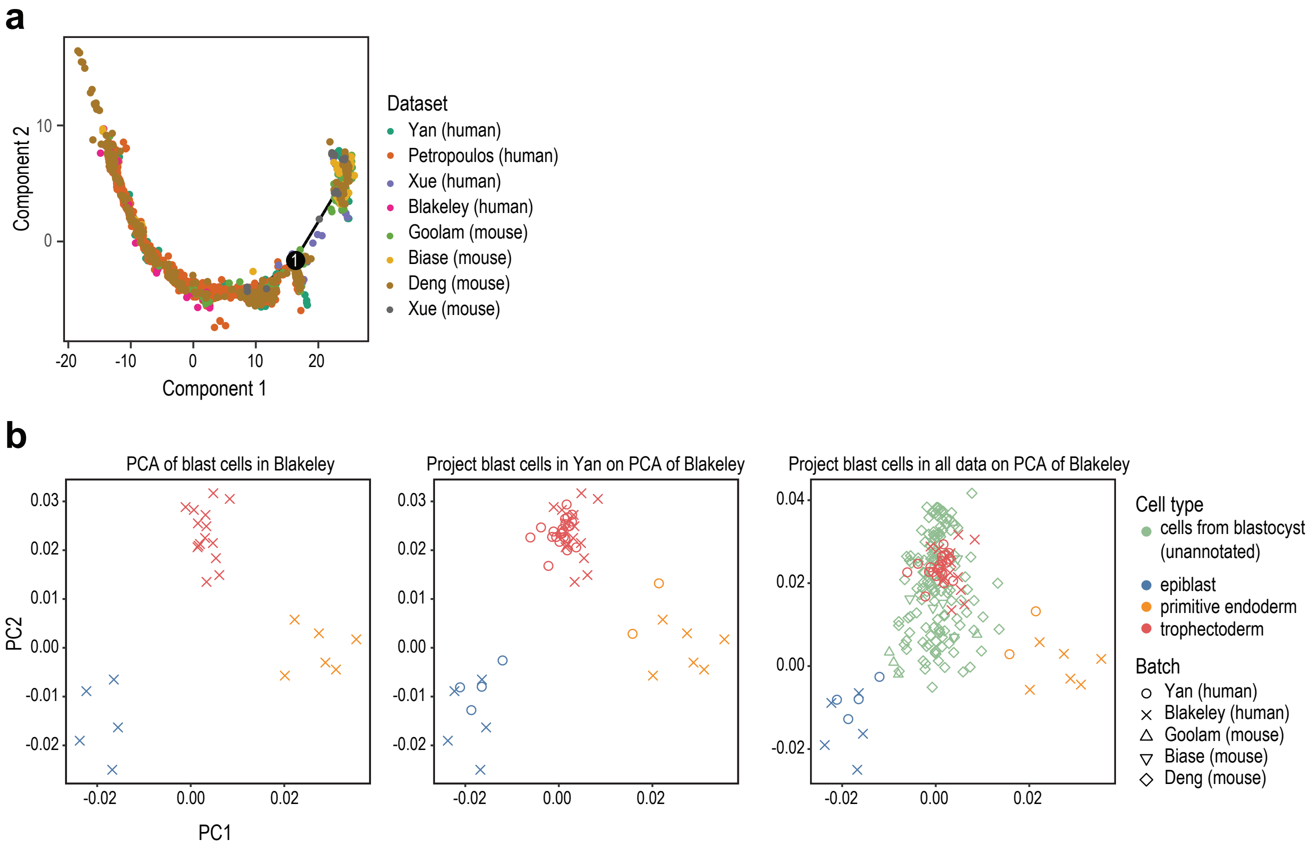


Fig. S15. (a) A pseudotime trajectory from Monocle 2 with all cells from ESC dataset collections, color coded by dataset. (b) A 1 by 3 panel of PCA plots of blast cells from ESC developmental data collections demonstrating scMerge effectively reproduce their cell type results, colored by the cell types identified from the original paper. The left panel is the PCA plot of blast cells from Blakeley et al. The medium panel is the PCA plot of blast cells from Blakeley et al. where blast cells from Yan et al. are projected. The right panel is the PCA plot of blast cells from Blakeley et al. where blast cells from mouse ESC datasets (Biase et al., Goolam et al. and Deng et al.) are projected.

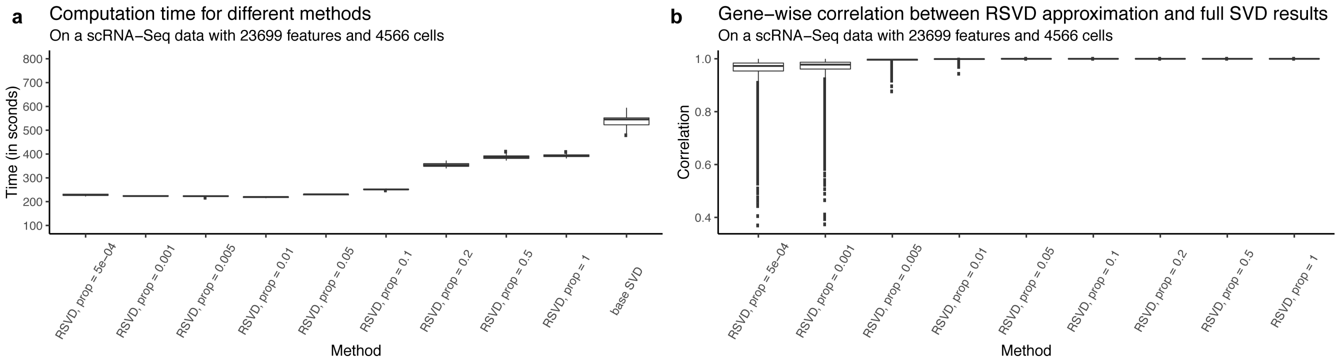


Fig. S16. (a) Computational time for different rsvd parameters with the Pancreas data collections ("Pancreas4"). This includes 23,699 features and 4566 cells. (b) Gene wise association based on Pearson correlation between rsvd approximation and full SVD calculation.

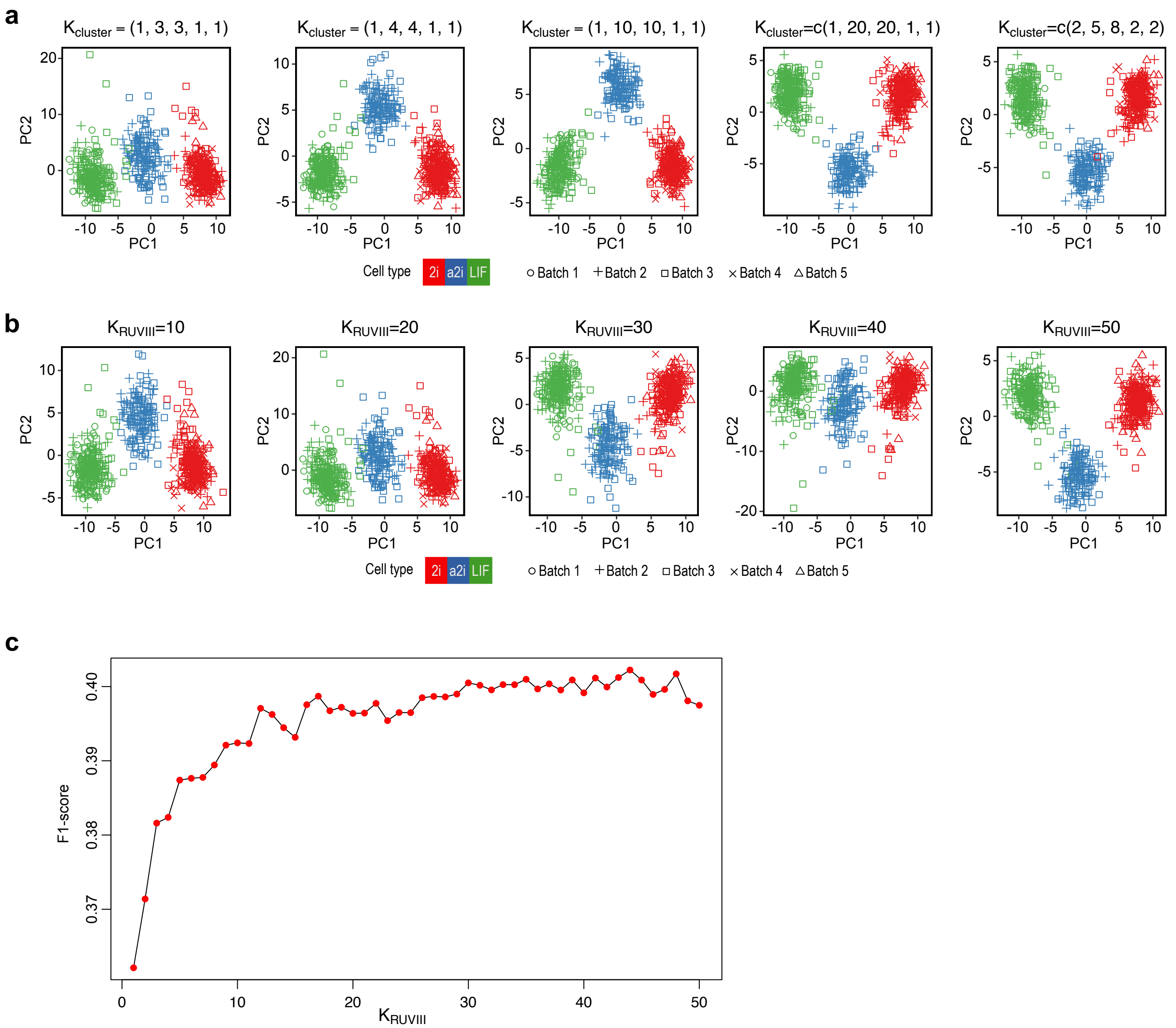
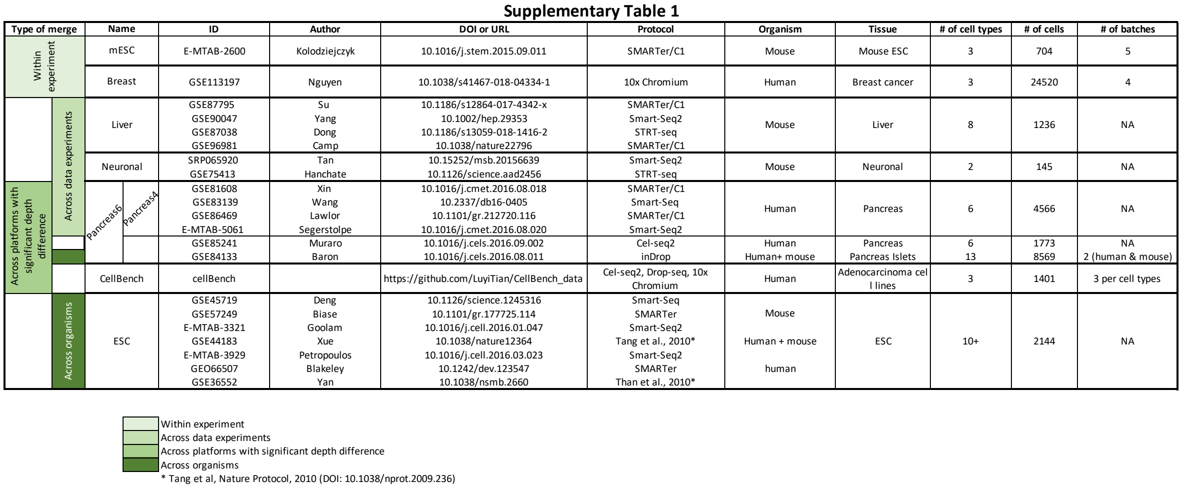


Fig. S17. (a) A 1 by 4 panel of PCA plots of mESC datasets using different $k_{Cluster}$ settings. (b) A 1 by 4 panel of PCA plots of mESC datasets using different k_{RUVIII} settings (k_{RUVIII} = 10, 20, 30, 40, 50) (c) A scatter plot showing the performance of scMerge using different k_{RUVIII} from 1 to 50, evaluated by F1 score of Silhouette coefficient. The y-axis represents the F1-score, while the x-axis represents the k_{RUVIII} values.

Table S1. More detailed summary of datasets and data collections used in this study.



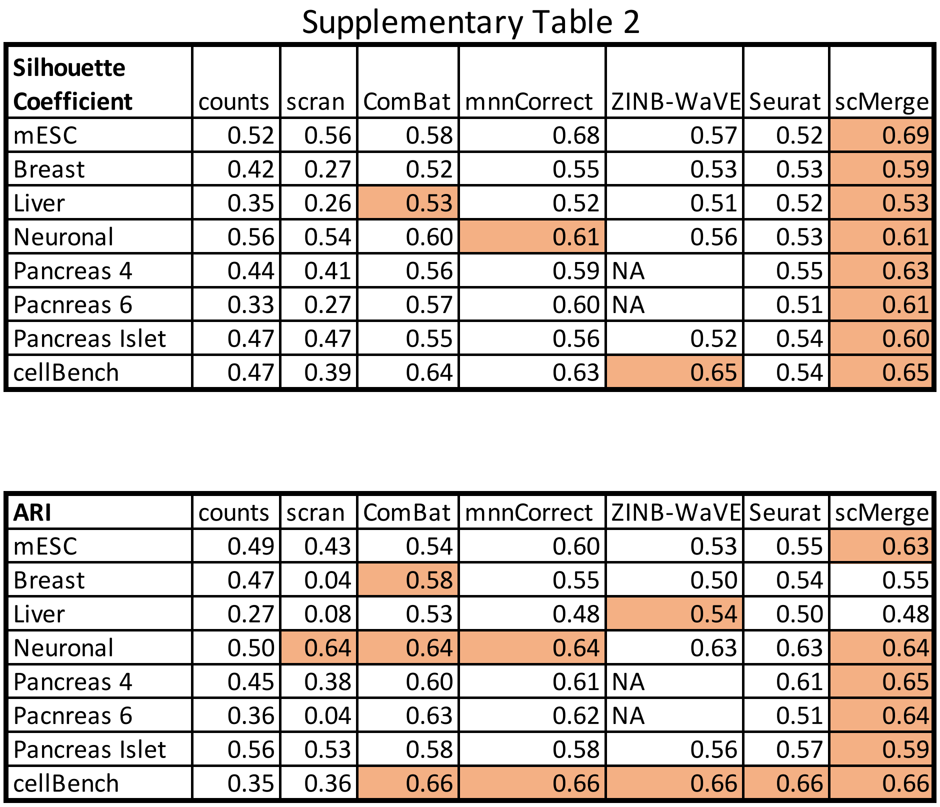
Supplementary Table 1

Type of merge		Name		ID	Author	DOI or URL	Protocol	Organism	Tissue	# of cell types	# of cells	# of batches
Within experiment		mESC		E-MTAB-2600	Kolodziejczyk	10.1016/j.stem.2015.09.011	SMARTer/C1	Mouse	Mouse ESC	3	704	5
		Breast		GSE113197	Nguyen	10.1038/s41467-018-04334-1	10x Chromium	Human	Breast cancer	3	24520	4
	Across data experiments	Liver		GSE87795 GSE90047 GSE87038 GSE96981	Su Yang Dong Camp	10.1186/s12864-017-4342-x 10.1002/hep.29353 10.1186/s13059-018-1416-2 10.1038/nature22796	SMARTer/C1 Smart-Seq2 STRT-seq SMARTer/C1	Mouse	Liver	8	1236	NA
		Neuronal		SRP065920 GSE75413	Tan Hanchate	10.15252/msb.20156639 10.1126/science.aad2456	Smart-Seq2 STRT-seq	Mouse	Neuronal	2	145	NA
Across platforms with significant depth difference		Pancreas6	Pancreas4	GSE81608 GSE83139 GSE86469 E-MTAB-5061	Xin Wang Lawlor Segerstolpe	10.1016/j.cmet.2016.08.018 10.2337/db16-0405 10.1101/gr.212720.116 10.1016/j.cmet.2016.08.020	SMARTer/C1 Smart-Seq SMARTer/C1 Smart-Seq2	Human	Pancreas	6	4566	NA
				GSE85241 GSE84133	Muraro Baron	10.1016/j.cels.2016.09.002 10.1016/j.cels.2016.08.011	Cel-seq2 inDrop	Human Human+ mouse	Pancreas Pancreas Islets	6 13	1773 8569	NA 2 (human & mouse)
		CellBench		cellBench		https://github.com/LuyiTian/CellBench_data	Cel-seq2, Drop-seq, 10x Chromium	Human	Adenocarcinoma cel l lines	3	1401	3 per cell types
	Across organisms	ESC		GSE45719 GSE57249 E-MTAB-3321 GSE44183 E-MTAB-3929 GEO66507 GSE36552	Deng Biase Goolam Xue Petropoulos Blakeley Yan	10.1126/science.1245316 10.1101/gr.177725.114 10.1016/j.cell.2016.01.047 10.1038/nature12364 10.1016/j.cell.2016.03.023 10.1242/dev.123547 10.1038/nsmb.2660	Smart-Seq SMARTer Smart-Seq2 Tang et al., 2010* Smart-Seq2 SMARTer Than et al., 2010*	Mouse Human + mouse human	ESC	10+	2144	NA

▢	Within experiment
▢	Across data experiments
▢	Across platforms with significant depth difference
■	Across organisms

* Tang et al, Nature Protocol, 2010 (DOI: 10.1038/nprot.2009.236)

Table S2. $F1_{sil}$ **and** $F1_{ARI}$ **of all methods (counts, scran (logcounts), ComBat, mnnCorrect, ZINB-WaVE, Seurat, and scMerge) across all datasets.**



Supplementary Table 2

Silhouette Coefficient	counts	scran	ComBat	mnnCorrect	ZINB-WaVE	Seurat	scMerge
mESC	0.52	0.56	0.58	0.68	0.57	0.52	0.69
Breast	0.42	0.27	0.52	0.55	0.53	0.53	0.59
Liver	0.35	0.26	0.53	0.52	0.51	0.52	0.53
Neuronal	0.56	0.54	0.60	0.61	0.56	0.53	0.61
Pancreas 4	0.44	0.41	0.56	0.59	NA	0.55	0.63
Pacnreas 6	0.33	0.27	0.57	0.60	NA	0.51	0.61
Pancreas Islet	0.47	0.47	0.55	0.56	0.52	0.54	0.60
cellBench	0.47	0.39	0.64	0.63	0.65	0.54	0.65

ARI	counts	scran	ComBat	mnnCorrect	ZINB-WaVE	Seurat	scMerge
mESC	0.49	0.43	0.54	0.60	0.53	0.55	0.63
Breast	0.47	0.04	0.58	0.55	0.50	0.54	0.55
Liver	0.27	0.08	0.53	0.48	0.54	0.50	0.48
Neuronal	0.50	0.64	0.64	0.64	0.63	0.63	0.64
Pancreas 4	0.45	0.38	0.60	0.61	NA	0.61	0.65
Pacnreas 6	0.36	0.04	0.63	0.62	NA	0.51	0.64
Pancreas Islet	0.56	0.53	0.58	0.58	0.56	0.57	0.59
cellBench	0.35	0.36	0.66	0.66	0.66	0.66	0.66

Additional data table S1 (SupplementaryFile1.xlsx)

Excel spreadsheet with six sheets listing scSEG and housekeeping genes derived from bulk microarray and bulk RNA-Seq, for human and mouse each. Mouse bulk microarray and bulk RNA-Seq derived housekeeping gene lists are homologues of the human gene lists given.

Additional data table S2 (SupplementaryFile2.xlsx)

Excel spreadsheet with pseudotime estimation for all cells from ESC data collections.