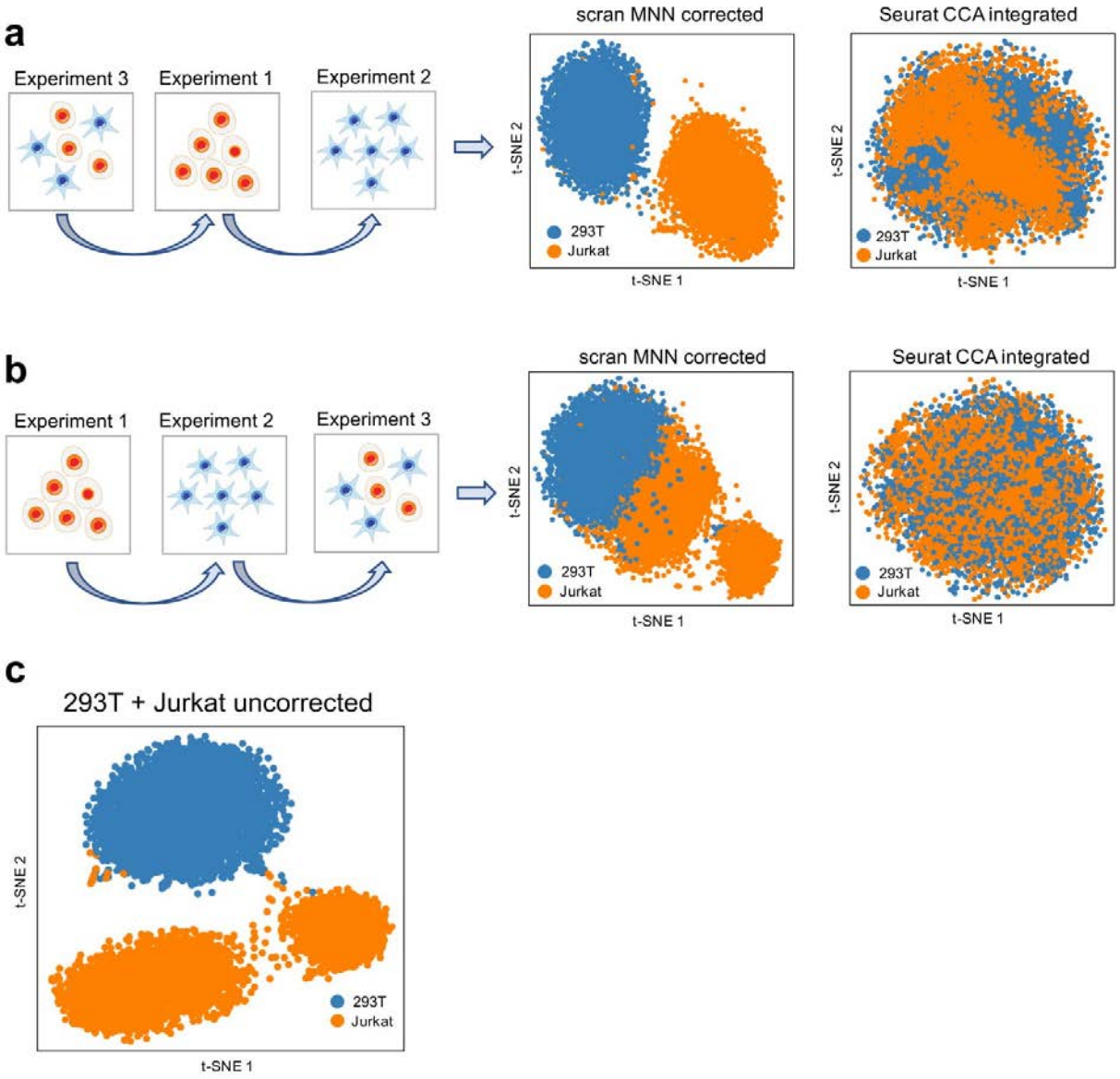
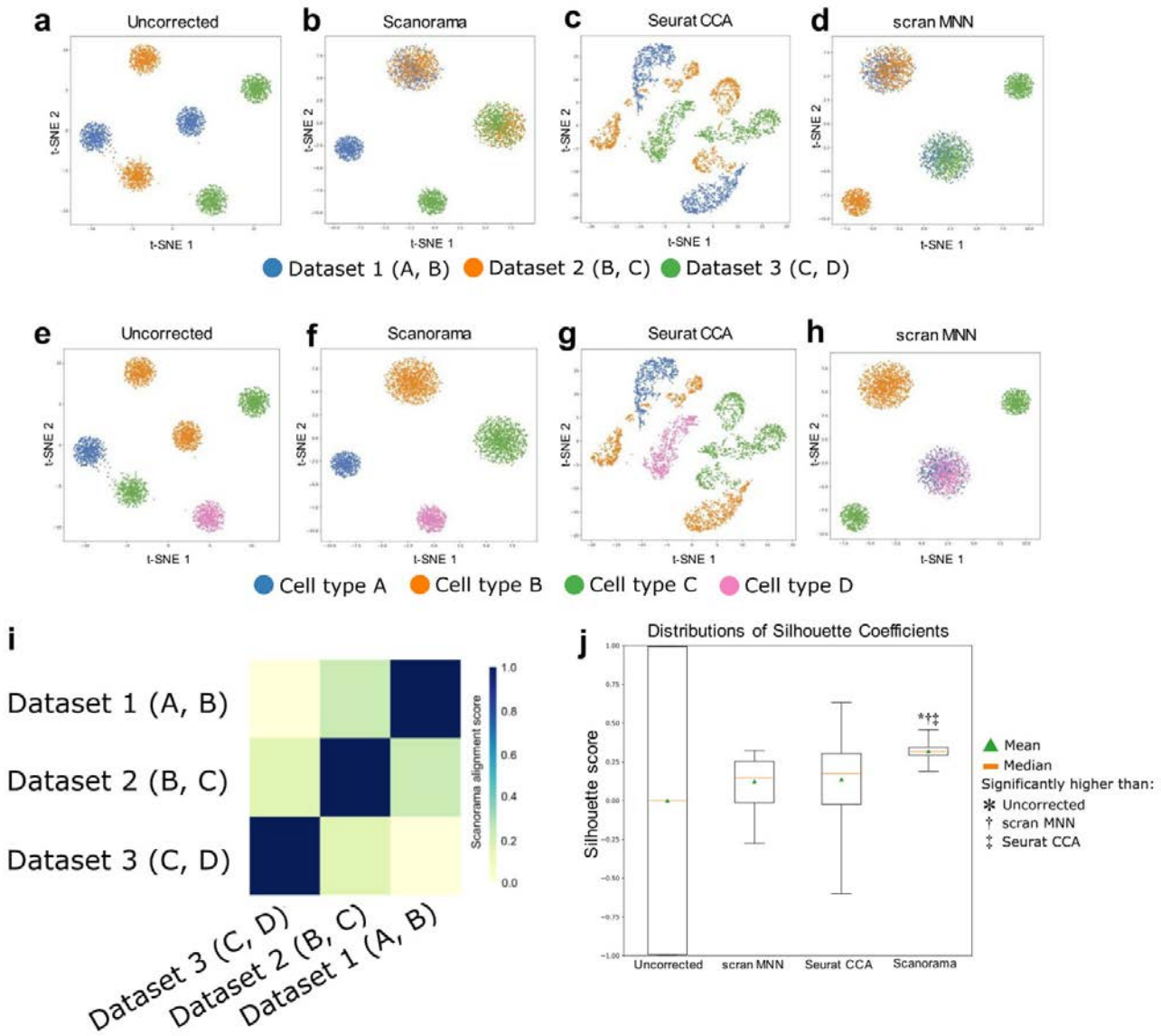


Supplementary Figure 1

Integration of a 293T/Jurkat mixture using scran MNN and Seurat CCA is sensitive to the order in which the datasets are considered.

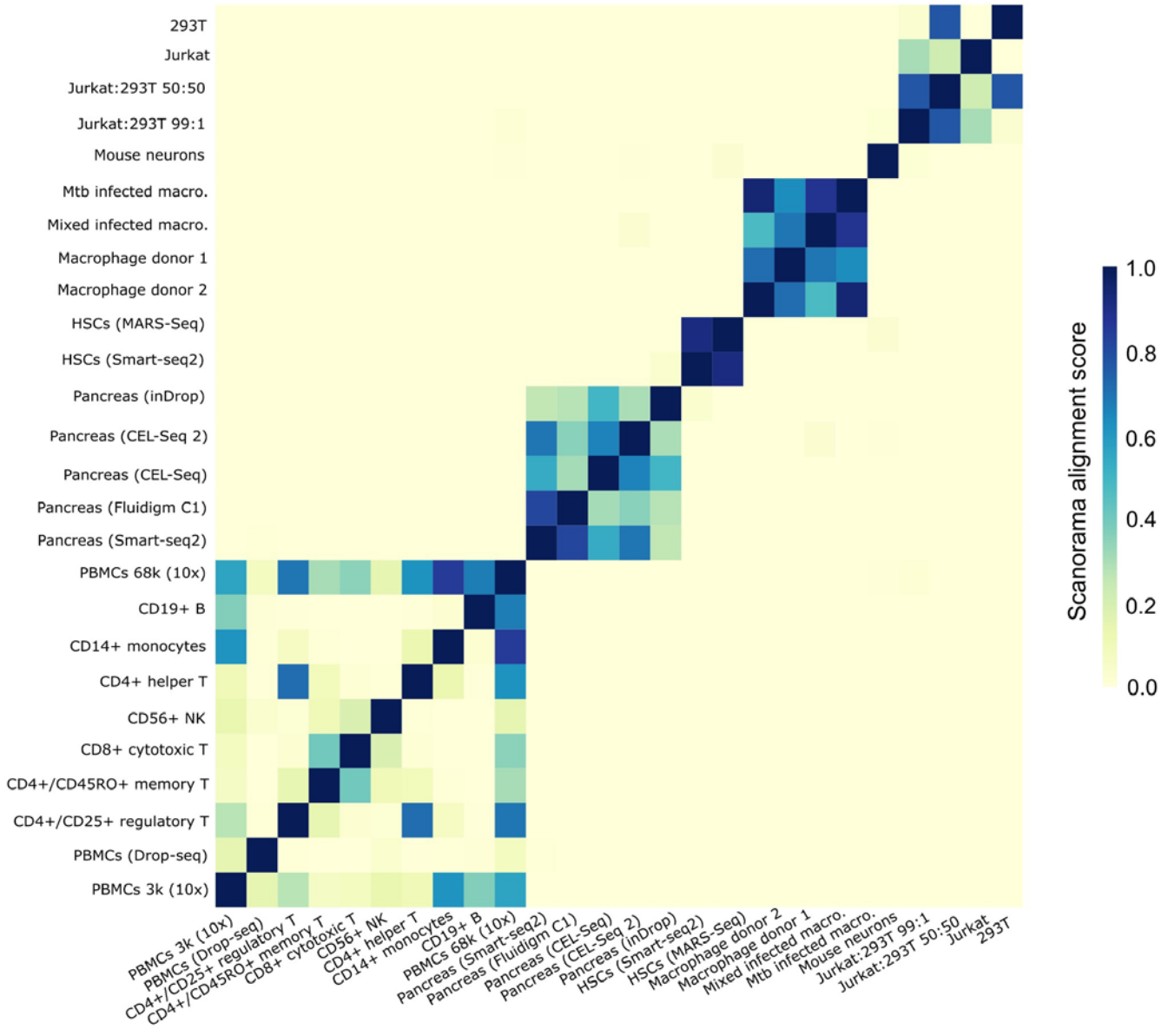
(**a**) When a mixture dataset of 293T cells and Jurkat cells is chosen as the first reference dataset (*n* = 3388 cells), scran MNN correctly integrates a second dataset of Jurkat cells (*n* = 3257) and a third dataset of 293T cells (*n* = 2885 cells). (**b**) When given the two datasets of 293T cells and Jurkat cells first, scran MNN incorrectly merges the two cell types together into a single cluster. Integration by scran MNN requires its first dataset to share at least one cell type with all other datasets that are successively integrated, which may not be a reasonable assumption. Seurat CCA was unsuccessful at integrating these three datasets in both cases (**a,b**). (**c**) Without correction, Jurkat cells cluster by batch instead of by cell type.



Supplementary Figure 2

Comparison of scRNA-seq integration methods on simulated data.

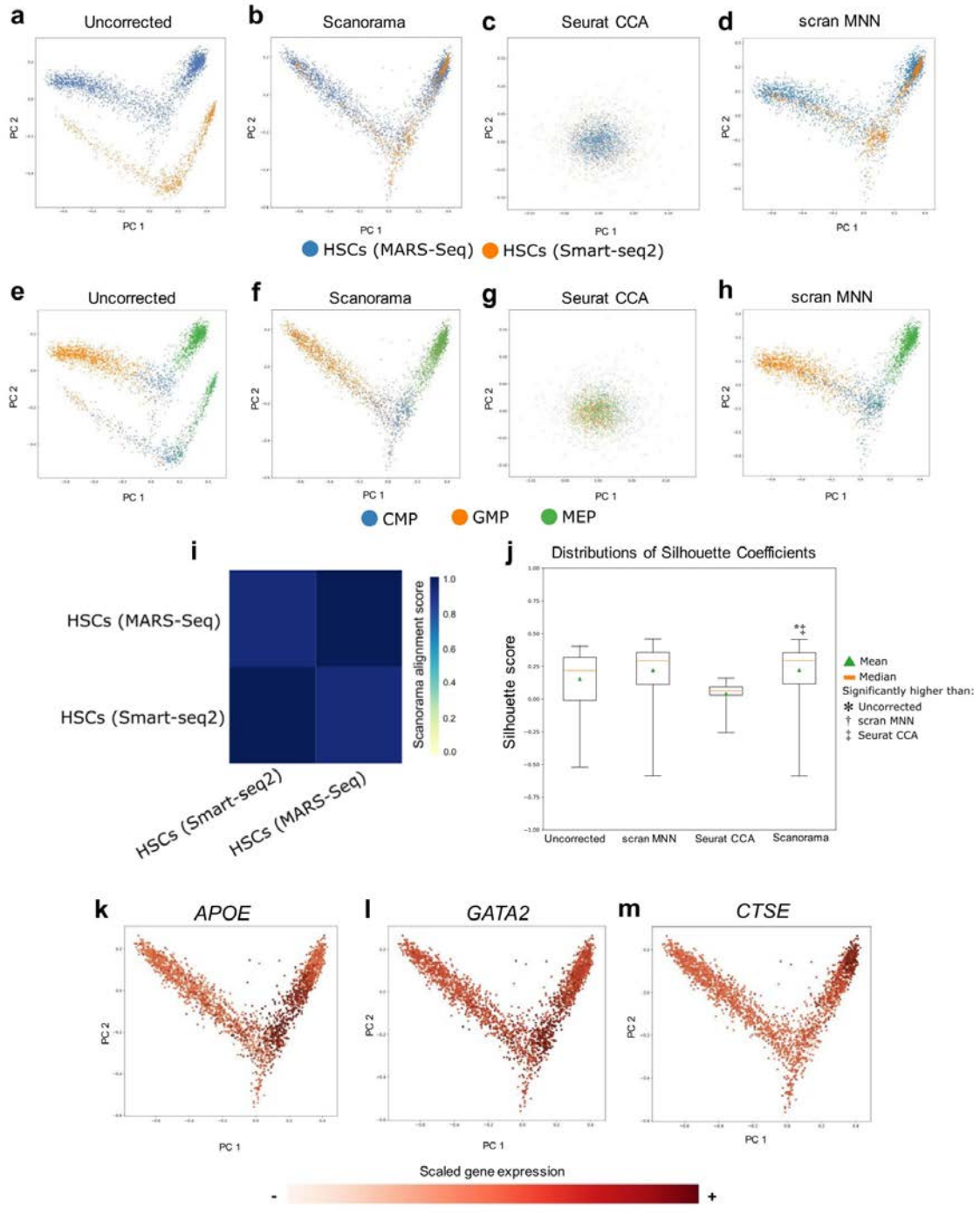
(**a-h**) We use the Splatter package to simulate three datasets with four cell types in total, where dataset 1 has cell types A and B, dataset 2 has cell types B and C, and dataset 3 has cell types C and D. In each dataset, we assign cells to a cell type with a 50/50 probability. Each dataset contains 1,000 cells. The Splatter simulation also generates batch effects between datasets such that without batch correction cells cluster by both dataset and batch (**a, e**). For Seurat CCA and scran MNN, datasets are aligned in numerical order. Scanorama correctly aligns the same cell types together (**b, f**), whereas scran MNN incorrectly merges cell types A and D and does not merge cell type C across batches (**d, h**). Seurat CCA is unable to merge the datasets together (**c, g**). (**i**) Scanorama alignment scores find the correct pairwise matches between the simulated cell types. (**j**) Scanorama has significantly improved Silhouette scores (median of 0.28) than the uncorrected data (median of 0.00; independent, two-sided t-test P < 5e-324; n = 3,000 cells), scran MNN (median of 0.16; P = 1.1e-40), and Seurat CCA (median of 0.18; P = 2.7e-37). An asterisk (*****) indicates a significantly higher Silhouette Coefficient distribution (Bonferroni corrected P < 0.05) between Scanorama and no correction, a dagger (†) indicates significance over scran MNN, and a double dagger (‡) indicates significance over Seurat CCA. t-SNE visualizations use a learning rate of 200 and a perplexity of 100. Box plot boxes extend from lower to upper quartiles with an orange line at the median and green triangle at the mean; whiskers show the range.



Supplementary Figure 3

Visualizing Scanorama alignment scores across 26 scRNA-seq datasets.

Scanorama alignment scores from aligning 26 heterogeneous scRNA-seq datasets reveal high amounts of alignment among biologically similar datasets and alignments scores close to or at zero for datasets that are not biologically similar. Heatmap rows and columns correspond to different datasets and diagonal entries are set to 1.

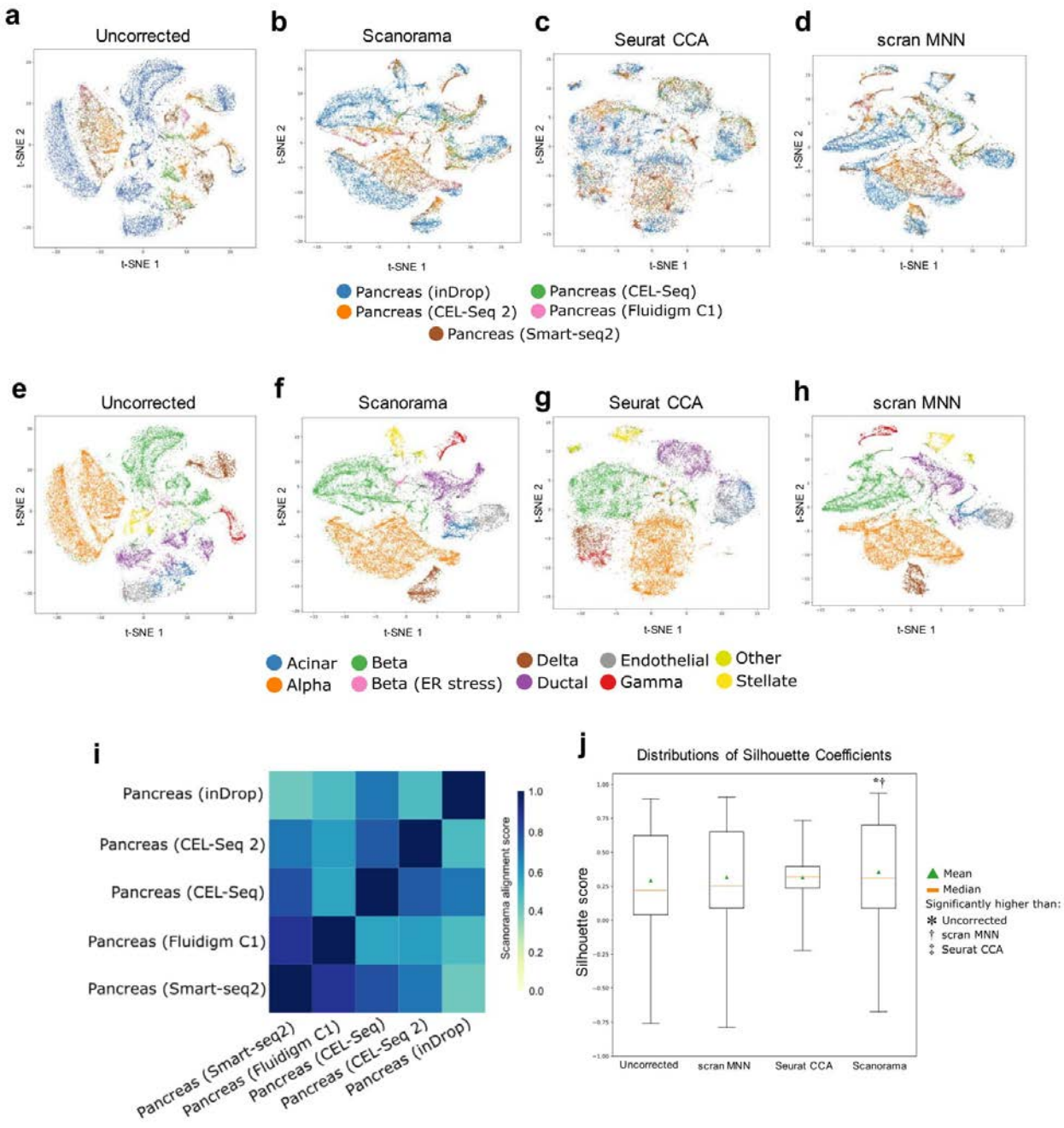


Supplementary Figure 4

Comparison of scRNA-seq integration methods on hematopoietic stem cells (HSCs).

Integration of 2401 hematopoietic stem cells (HSCs) from MARS-Seq and 774 HSCs from Smart-seq2. (**a**, **e**) Two datasets of HSCs plotted on the first two principal components (PCs) shows cell separated by batch effects along the second PC. Cells are visualized using PCs, instead of t-SNE embeddings, since they organize according to their pseudo-temporal relationships when visualized with PCA; granulocyte-macrophage progenitors (GMP) and megakaryocyte-erythrocytes (MEP) are derived from common myeloid

progenitors (CMP). (**b, f**) Scanorama removes any significant difference due to experimental batch (natural log likelihood-ratio = -902; *n* = 3,175 cells). (**c**, **g**) Seurat CCA overcorrects and places all cell types into a single cluster. (**d**, **h**) scran MNN obtains a similar result to that of Scanorama. (**i**) Scanorama alignments consists of a substantial percentage of the cells in both datasets, as expected. (**j**) Scanorama and scran MNN have similar performance and the same median Silhouette Coefficient (median of 0.28; independent, two-sided t-test *P* = 0.14; *n* = 3,175 cells), but Scanorama has significantly better performance than no correction (median of 0.22; *P* = 8e-10) and Seurat CCA (median of 0.07; *P* = 2e-132). An asterisk (*****) indicates a significantly higher Silhouette Coefficient distribution (Bonferroni corrected *P* < 0.05) between Scanorama and no correction and a double dagger (‡) indicates significance over Seurat CCA. Box plot boxes extend from lower to upper quartiles, whiskers indicate range, an orange line indicates the median, and a green triangle indicates the mean (*n* = 3,175 cells). (**k-m**) Expression of marker genes indicating different stages of erythropoiesis. *APOE* and *GATA2* are more highly expressed in the erythropoietic transition from common myeloid progenitors (CMPs) to megakaryocyte-erythrocytes (MEPs) (**k, l**) and *CTSE* is more highly expressed in MEPs (**m**).

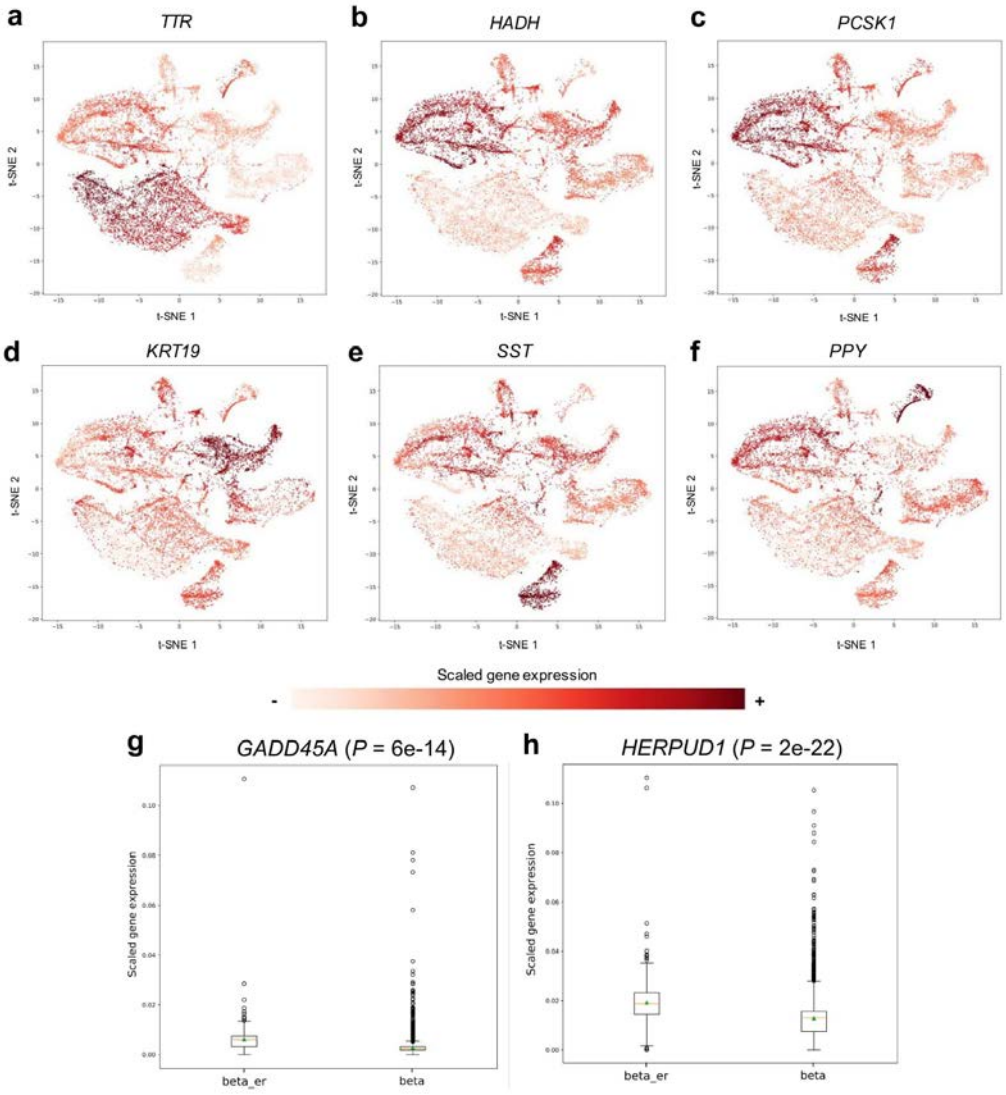


Supplementary Figure 5

Comparison of scRNA-seq integration methods on pancreatic islet cells.

Integration of 8569 pancreatic islet cells from inDrop, 2449 cells from CEL-Seq2, 1276 cells from CEL-Seq, 638 cells from Fluidigm C1, and 2989 cells from Smart-seq2. (**a**, **e**) Pancreatic islets cluster by cell type and batch in the uncorrected setting. (**b-d, f-h**) Visually, Scanorama, Seurat CCA, and scran MNN have similar performance in merging cell-type specific clusters together across datasets. (**i**) Scanorama finds substantial overlap among all five pancreatic islet datasets. (**j**) All methods have relatively similar performance, but Seurat CCA has a higher Silhouette Coefficient distribution (median of 0.30; compared to Scanorama, independent, two-sided t-test P = 4.8e-3; n = 15,921 cells) followed by Scanorama (median of 0.28), scran MNN (median of 0.25; P = 5.1e-4), and the uncorrected data (median of 0.23; P = 9.7e-5). An asterisk (*****) indicates a significantly higher Silhouette Coefficient distribution (Bonferroni corrected P <

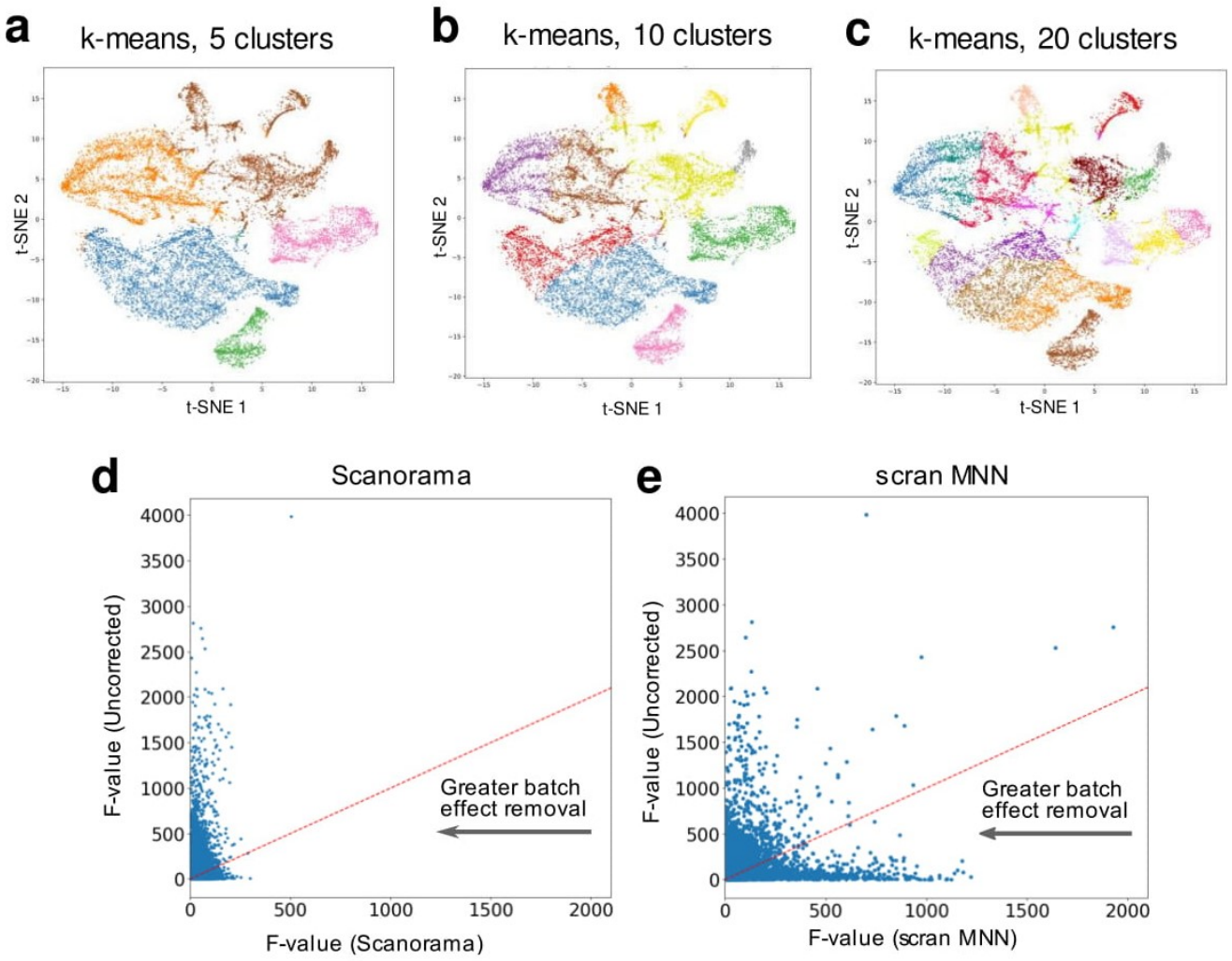
0.05) between Scanorama and no correction and a dagger (†) indicates significance over scran MNN. Box plot boxes extend from lower to upper quartiles, whiskers indicate range, an orange line indicates the median, and a green triangle indicates the mean (*n* = 15,921 cells).



Supplementary Figure 6

Marker gene expression of Scanorama integrated and batch corrected pancreatic islet datasets.

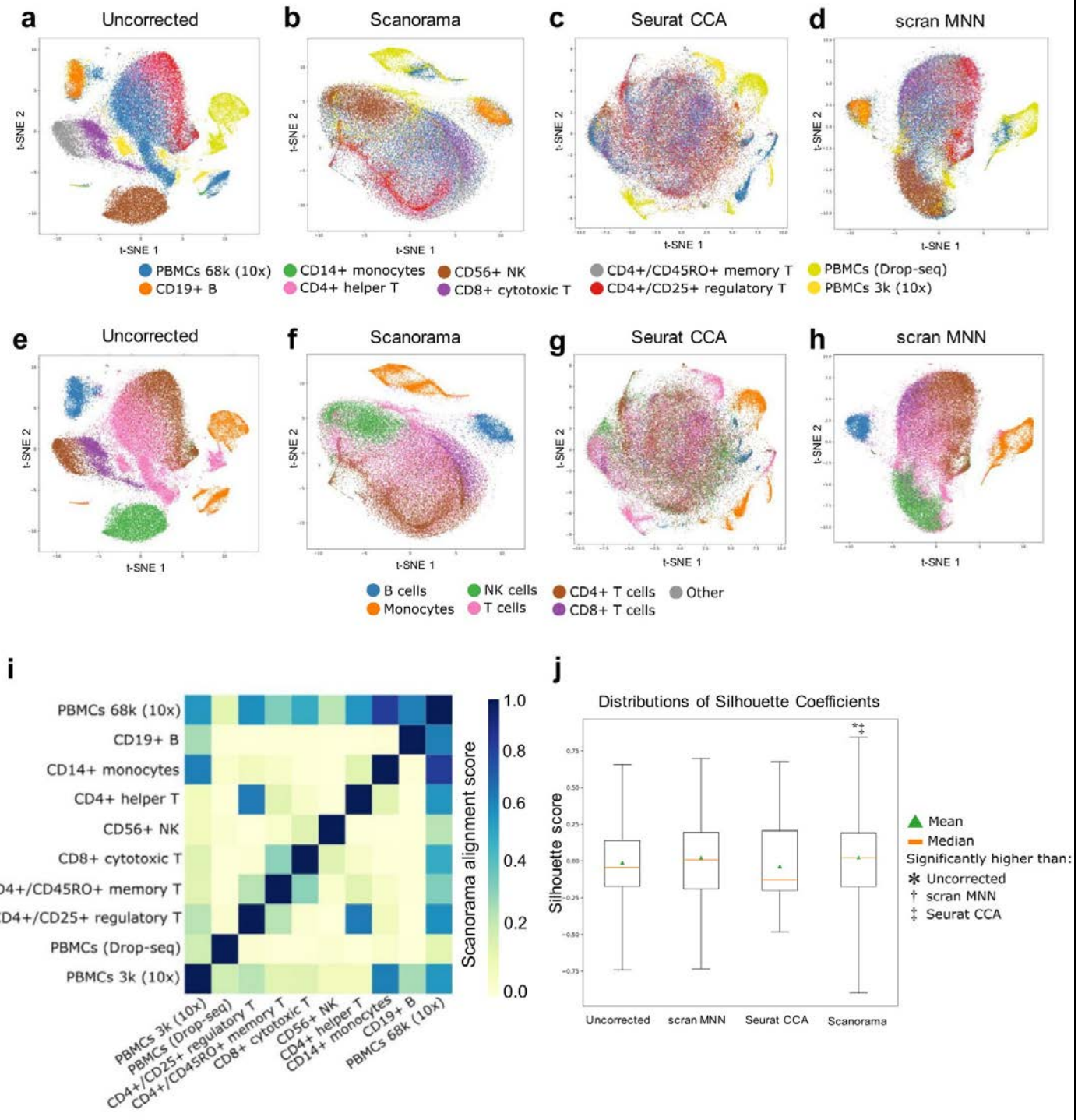
Gene expression after integration of 8569 pancreatic islet cells from inDrop, 2449 cells from CEL-Seq2, 1276 cells from CEL-Seq, 638 cells from Fluidigm C1, and 2989 cells from Smart-seq2. (**a-f**) Marker gene expression heatmaps of the t-SNE embedded panorama of pancreatic islet cells. We observe higher expression of *TTR* in alpha cells (**a**), *HADH* and *PCSK1* in beta cells (**b, c**), *KRT19* in ductal cells (**d**), *SST* in delta cells (**e**), and *PPY* in gamma cells (**f**). (**g, h**) Marker genes *GADD45A* and *HERPUD1* related to ER stress are significantly elevated among a subpopulation of beta cells (n = 320 cells) compared to other beta cells (n = 4765 cells), consistent with a rare subpopulation of beta cells marked by ER stress that was previously identified in one of the datasets. The P-values for increased expression of *GADD45A* and *HERPUD1* are also much stronger after integrating five pancreas datasets (P = 6.07e-14 for *GADD45A* and P = 2.42e-22 for *HERPUD1*) than for the initial findings in a single dataset (P = 5.21e-3 for *GADD45A* and P = 2.98e-5 for *HERPUD1*; 102 ER stress beta cells and 1,114 other beta cells). We computed P-values using a two-sided, Welch's t-test for comparing populations with unequal variances. t-SNE visualizations use a learning rate of 200 and a perplexity of 400. Box plot boxes extend from lower to upper quartiles, upper whisker extends to last point less than the third quartile plus 1.5 times the interquartile range (IQR), lower whisker extends to first point greater than the first quartile minus 1.5 times the IQR, points indicate remaining cells, an orange line indicates the median, and a green triangle indicates the mean.



Supplementary Figure 7

Clustering of Scanorama integrated pancreatic islet datasets and batch correction quality.

Batch correction performance after applying Scanorama to 8569 pancreatic islet cells from inDrop, 2449 cells from CEL-Seq2, 1276 cells from CEL-Seq, 638 cells from Fluidigm C1, and 2989 cells from Smart-seq2. (**a-c**) k-means clustering of datasets integrated with Scanorama result in clusters that are orthogonal to differences due to batch, noting that even smaller sub-clusters do not find dataset-specific structure. (**d, e**) Scanorama batch correction of five pancreas datasets results in lower one-way ANOVA *F*-values compared to scran MNN (we note that this analysis is not applicable to Seurat CCA, which finds integrated embeddings and does not modify gene expression values). Each point represents a gene; results are for 15,369 genes. Closer to the left is better, indicating more similar gene expression distributions after batch correction. The red dashed line indicates equal *F*-values between uncorrected and corrected datasets.

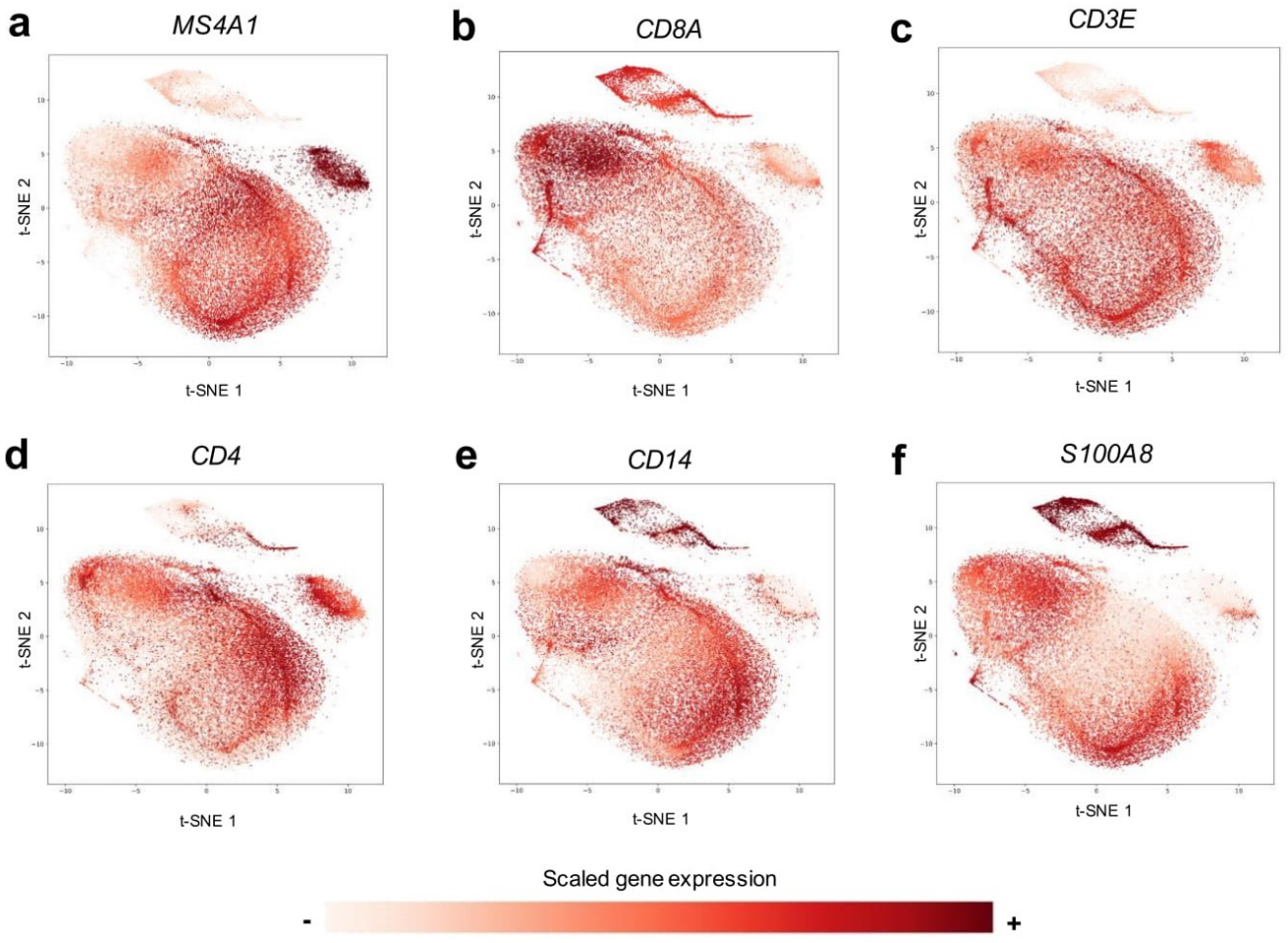


Supplementary Figure 8

Comparison of scRNA-seq integration methods on peripheral blood mononuclear cells (PBMCs).

Integration of 18018 PBMCs from 10x Genomics (donor 1), 2261 CD19+ B cells from 10x, 295 CD14+ monocytes from 10x, 3713 CD4+ helper T cells from 10x, 6657 CD56+ NK cells from 10x, 3990 CD8+ cytotoxic T cells from 10x, 3628 CD4+/CD45RO+ memory T cells from 10x, 3365 CD4+/CD25+ regulatory T cells from 10x, 3774 PBMCs using Drop-seq, and 2293 PBMCs from 10x Genomics (donor 2). (**a**, **e**) Without batch correction, PBMC datasets cluster by both cell type and dataset. (**b**, **f**) Scanorama integration results

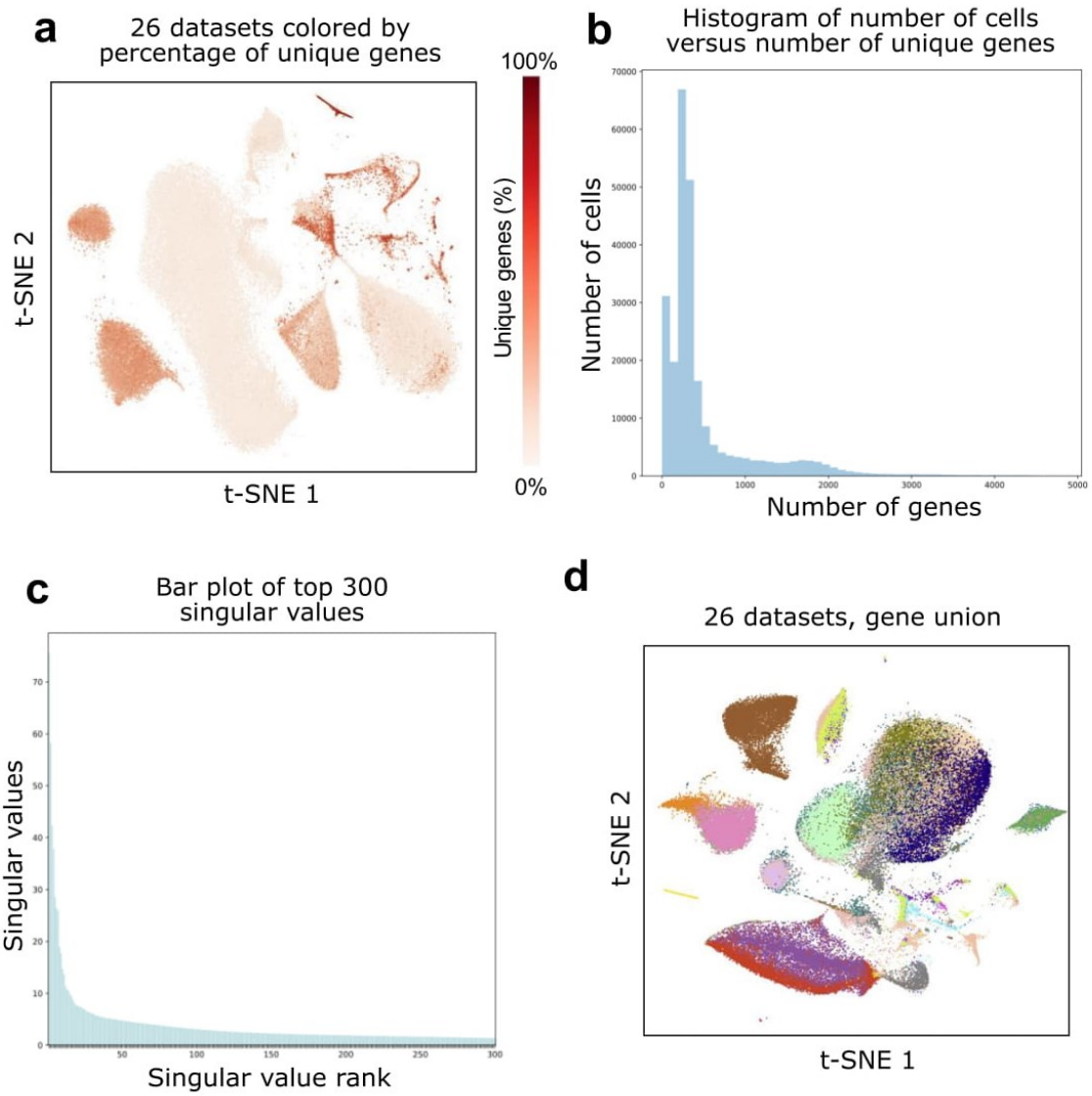
cells clustering by cell type. (**c**, **g**) Seurat CCA integration results in overcorrection. (**d**, **h**) scran MNN obtains a similar result as that of Scanorama because a large dataset of PBMCs was chosen as the first dataset. We expect performance to degrade if the large dataset were not chosen first. (**i**) Scanorama alignment scores capture relationships between the datasets. (**j**) Scanorama has the highest distribution of Silhouette Coefficients (median of 0.05) compared to scran MNN (median of 0.03; independent, two-sided t-test P = 0.0011; n = 47,994 cells), the uncorrected data (median of -0.08; P = 1e-51), and Seurat CCA (median of -0.18; P = 9e-194). An asterisk (*****) indicates a significantly higher Silhouette Coefficient distribution (Bonferroni corrected $P < 0.05$) between Scanorama and no correction and a double dagger (‡) indicates significance over Seurat CCA. Box plot boxes extend from lower to upper quartiles, whiskers indicate range, an orange line indicates the median, and a green triangle indicates the mean.



Supplementary Figure 9

Marker gene expression in Scanorama integrated and batch corrected PBMC datasets.

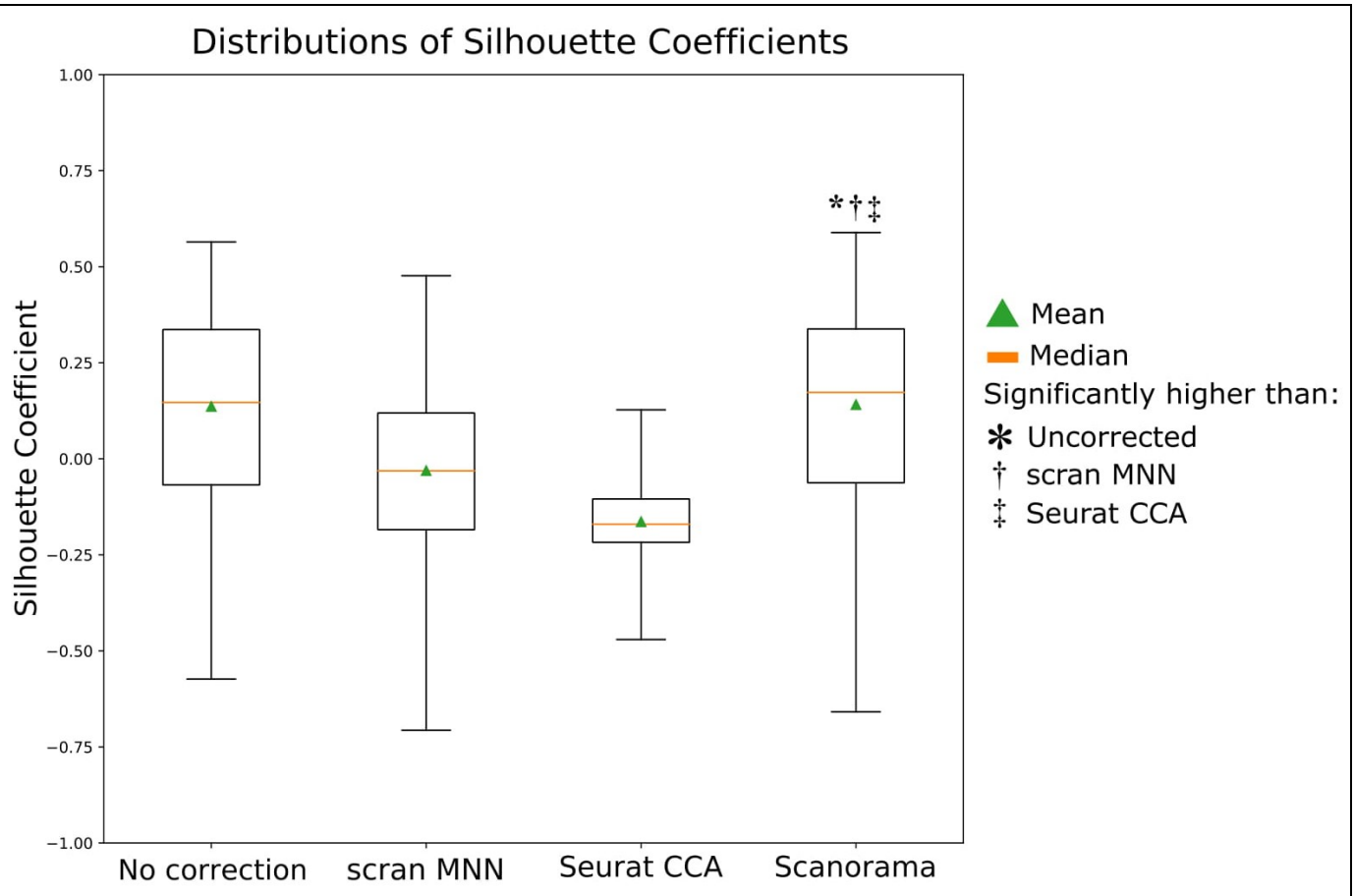
Gene expression after integration of 18018 PBMCs from 10x Genomics (donor 1), 2261 CD19+ B cells from 10x, 295 CD14+ monocytes from 10x, 3713 CD4+ helper T cells from 10x, 6657 CD56+ NK cells from 10x, 3990 CD8+ cytotoxic T cells from 10x, 3628 CD4+/CD45RO+ memory T cells from 10x, 3365 CD4+/CD25+ regulatory T cells from 10x, 3774 PBMCs using Drop-seq, and 2293 PBMCs from 10x Genomics (donor 2). (**a-f**) Marker gene expression heatmaps of the t-SNE embedded panorama of PBMCs. We observe higher expression of *MS4A1* in (**a**) B cells, (**b**) *CD8A* in NK cells, (**c, d**) *CD3E* and *CD4* in T cells, and (**e, f**) *CD14* and *S100A8* in monocytes. t-SNE visualizations use a learning rate of 200 and a perplexity of 400.



Supplementary Figure 10

26 dataset quality control.

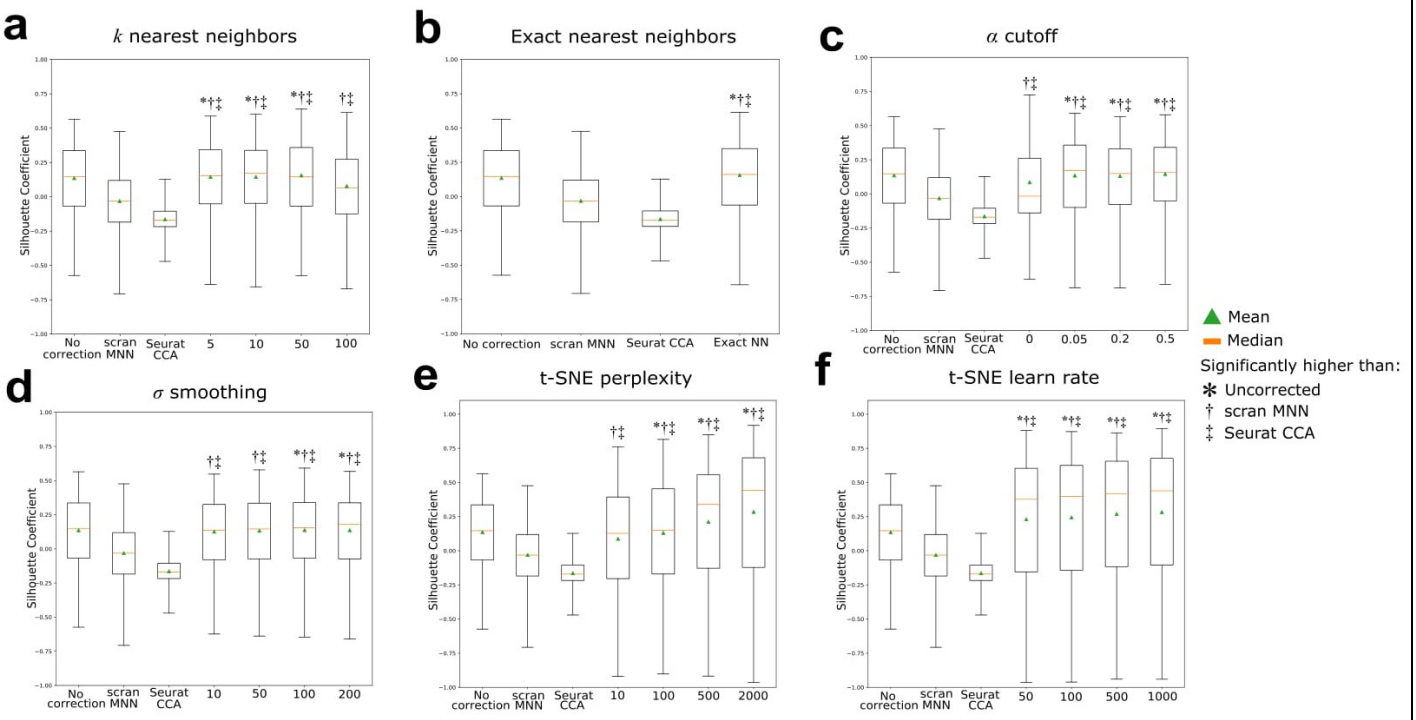
(**a**) Cells in our experiment integrating 26 diverse datasets (*n* = 105,476 cells) cluster according to cell type instead of by relative differences in the number of unique genes. E.g., the two HSC datasets are aligned despite different dataset-specific gene percentages (the MARS-Seq dataset has a relatively low average percentage of nonzero genes at 30% versus the Smart-seq2 dataset with an average of 79% nonzero genes), as are the pancreas datasets. (**b**) In our analysis of 26 datasets, cells were included if they contained greater than 600 unique genes. We observe a bimodal distribution of cells according to their number of unique genes and we filter out the mode of cells that have lower amounts of unique genes due to either transcriptional quiescence, high amounts of dropout, or other technical artefacts. (**c**) We compute the SVD of the concatenation of the 26 datasets and visualize the top 300 singular values in a bar plot. To preserve most of the variation in the data, indicated by the “elbow” in the bar plot, we use a conservative cutoff of the top 100 components from the SVD. (**d**) Integrating datasets (*n* = 105,476 cells) based on the union of all genes (setting unobserved gene expression values to zero) results in similar results as with taking the intersection (although interestingly, a small portion of CD14+ monocytes align with macrophages, which may have some biological basis); however, we caution against a union-based approach since this could introduce variability that is not reflective of the underlying biology.



Supplementary Figure 11

Silhouette Coefficient distributions across 26 scRNA-seq datasets.

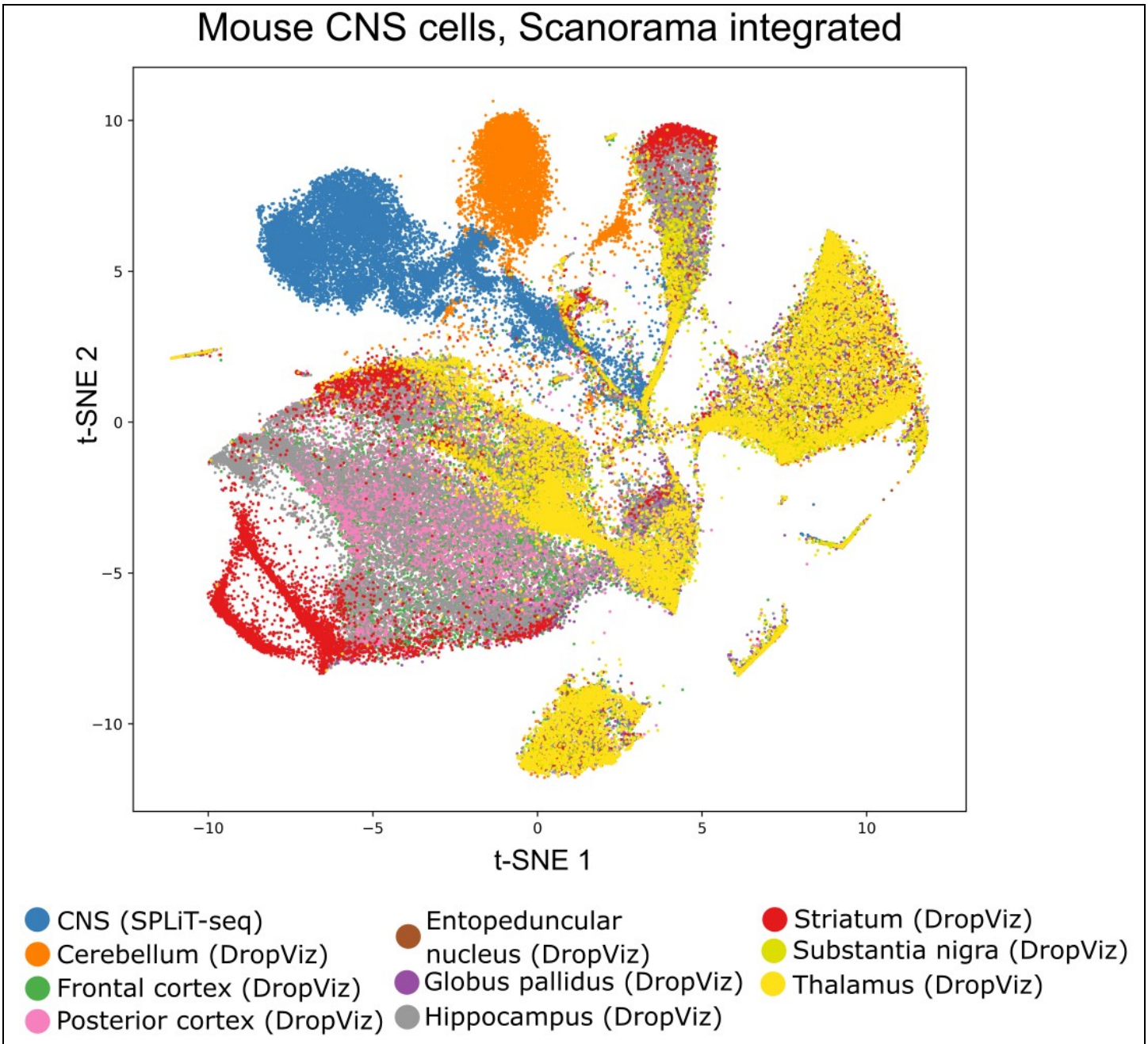
In additional to visually inspecting the clusters produced by a method like t-SNE, we can quantify the integrative performance of our method by computing a Silhouette Coefficient for each cell (**Methods**). Higher values indicate that samples from the same cell type also cluster together, indicating better clustering performance. For our experiment in which we integrate 26 diverse scRNA-seq datasets, we compute Silhouette Coefficients using low dimensional embeddings as described in **Methods**. Scanorama has a significantly higher Silhouette Coefficient distribution (median of 0.17) compared to scran MNN (median of -0.03; P < 5e-324), Seurat CCA (median of -0.18; P < 5e-324), and no correction (median of 0.14; P = 4e-6) when integrating our collection of 26 datasets containing 105,476 cells (**Figure 2a-c**). Notably, scran MNN and Seurat CCA have lower median Silhouette Coefficients than if no correction had been applied, indicating large amounts of overcorrection. Box plot boxes extend from lower to upper quartiles with an orange line at the median and green triangle at the mean; whiskers show the range. P-values are determined using an independent, two-sided t-test (n = 105,476 cells). An asterisk (*****) indicates a significantly higher Silhouette Coefficient distribution (Bonferroni corrected P < 0.05) between Scanorama and no correction, a dagger (†) indicates significance over scran MNN, and a double dagger (‡) indicates significance over Seurat CCA.



Supplementary Figure 12

Silhouette Coefficient distributions for 26 dataset integration at different parameters.

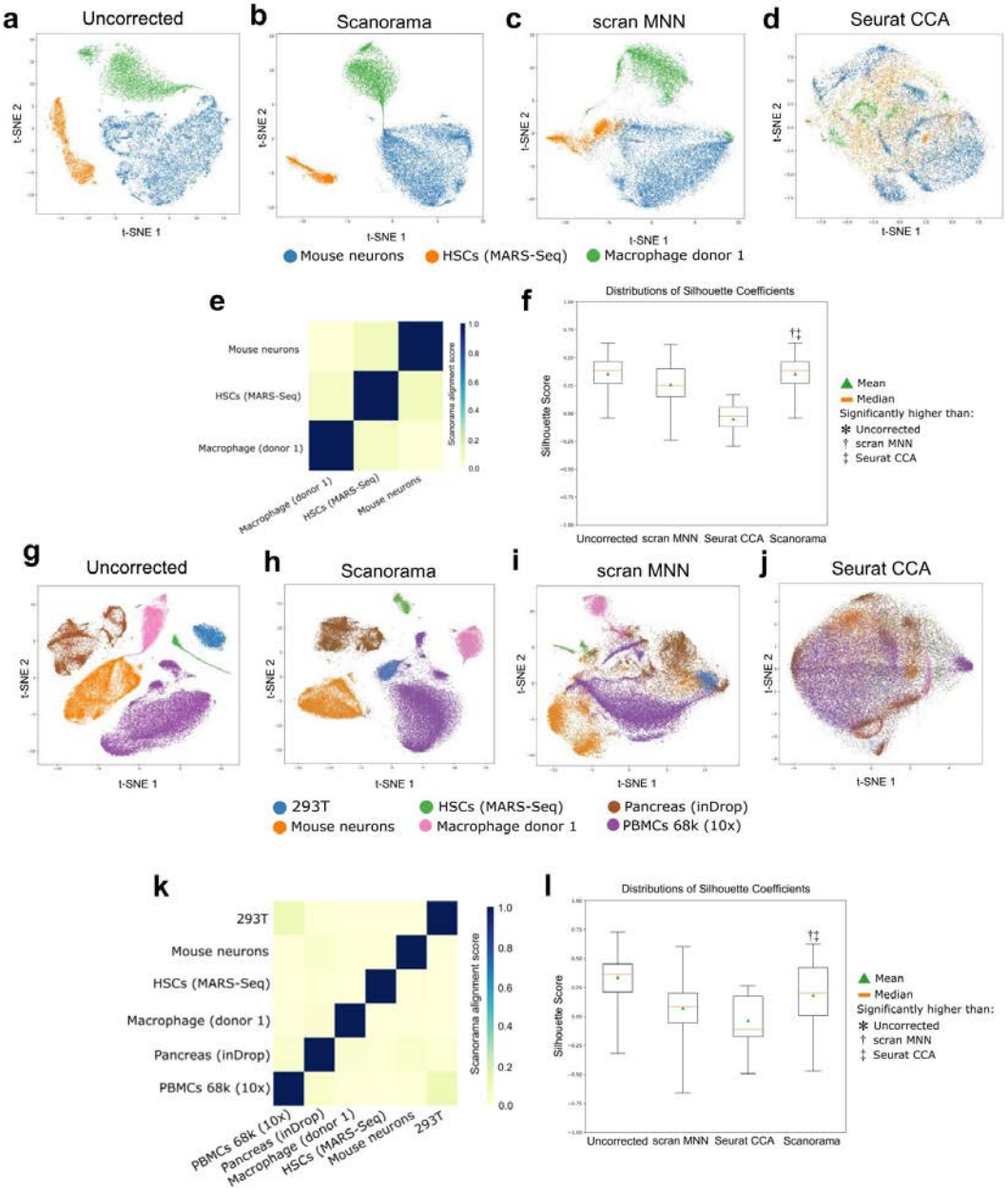
Sensitivity analysis of Scanorama alignment parameters and t-SNE visualization parameters for the integration of 26 diverse scRNA-seq datasets. Box plots show distributions of Silhouette Coefficients at different parameter settings. Box plot boxes extend from lower to upper quartiles with an orange line at the median and green triangle at the mean; whiskers show the range. All distributions are over the same 105,476 cells across 26 heterogeneous scRNA-seq datasets. An asterisk (*****) indicates a significantly higher Silhouette Coefficient distribution (two-sided independent *t*-test, Bonferroni corrected $P < 0.05$) between Scanorama and no correction, a dagger (†) indicates significance over scran MNN, and a double dagger (‡) indicates significance over Seurat CCA. Importantly, in the analysis for alignment parameters (**a-d**), Silhouette Coefficients are calculated for the integrated, low-dimensional embeddings. When assessing the sensitivity of t-SNE visualization parameters (**e**, **f**), we calculate the Silhouette Coefficients on the 2-dimensional t-SNE embeddings (which are computed off of the low dimensional embeddings). All plots also include the Silhouette Coefficient distributions for uncorrected data, Seurat CCA integration, and scran MNN correction on low dimensional embeddings as described in **Methods**. (**a**) The *k* nearest neighbor parameter is largely insensitive around the default value of 20 and can go as low as 5 without affecting performance. At larger values of *k*, the matches become more permissive and the Silhouette Coefficients start to drop, where at $k = 100$ the median Silhouette Coefficient (0.091) is below that of the uncorrected case. (**b**) There is no significant change in the distribution of Silhouette Coefficients between the approximate and exact nearest neighbors settings (independent, two-sided t-test $P = 0.39$; $n = 105{,}476$ cells), although the integration runtime increases to more than 60 minutes without the approximation algorithm. (**c**) We recommend keeping α to a low value greater than zero, which can be learned from the data if some of the cell types being integrated are known. Lower values may introduce overcorrection, while higher values approach the uncorrected case. (**d**) The median Silhouette Coefficient is largely insensitive to different values of the smoothing parameter σ for the Gaussian kernel function. (**e**) Visualizing the integration of 26 datasets requires a high perplexity (around 500 or greater) to obtain a median Silhouette Coefficient comparable to that for the low dimensional embeddings. We set the perplexity to 1,200 for visualizing the 26 datasets (**Figure 3a)**. (**f**) When visualizing the 26 datasets, a higher t-SNE learning rate improves the median Silhouette Coefficient to be comparable that for the low dimensional embeddings. The Silhouette Coefficient distributions for the t-SNE embeddings are generally wider than those for the lower dimensional embeddings since it is harder to obtain large separations between clusters in two dimensions.



Supplementary Figure 13

Scanorama integration of different regions of the mouse CNS.

Visualization of 10% (*n* = 109,553 cells) of 1,095,538 mouse CNS cells after Scanorama integration colored by dataset. Corresponding cell type labels and marker genes are given in **Figure 4**.

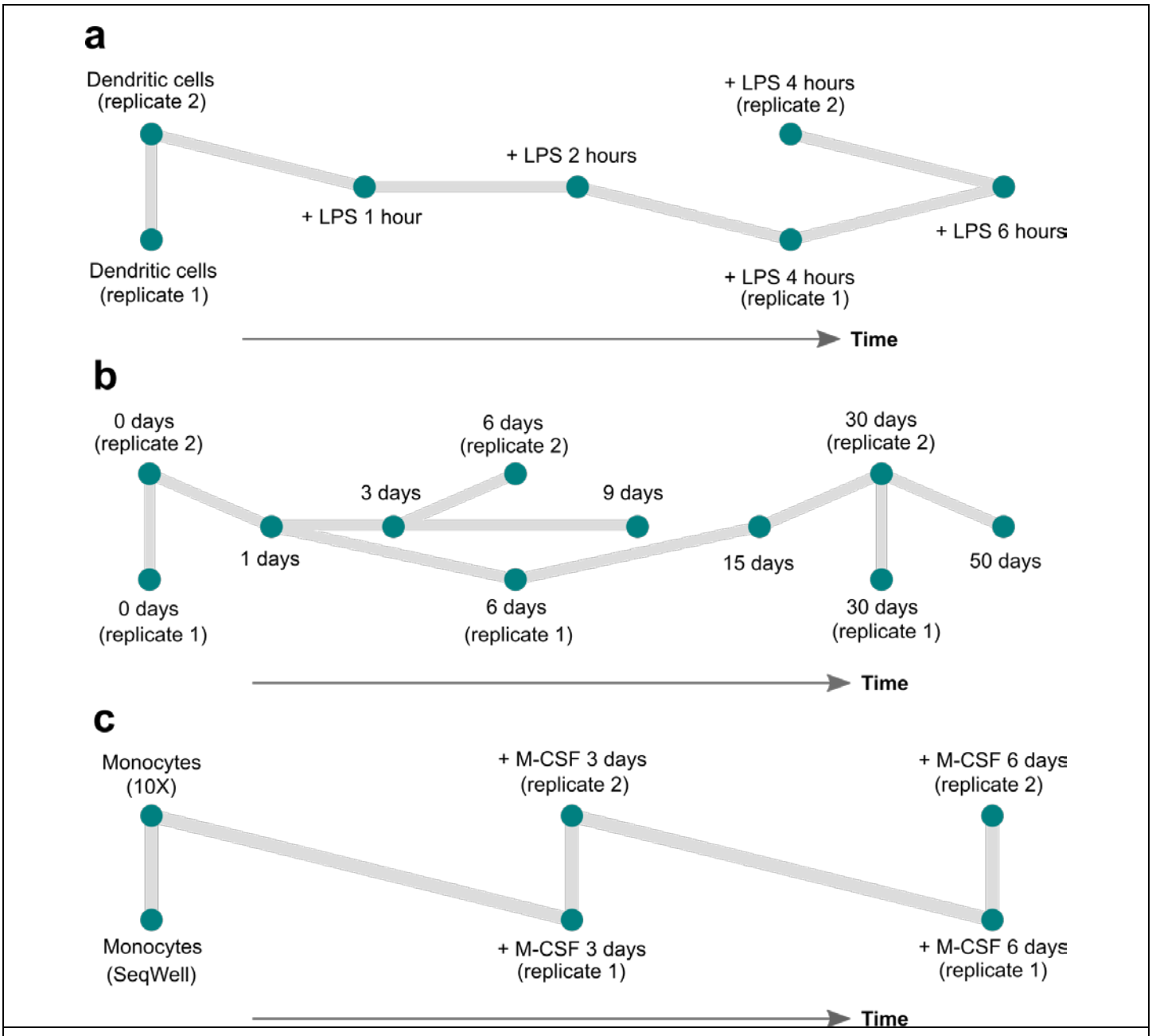


Supplementary Figure 14

Comparison of scRNA-seq integration methods on datasets with no overlapping cell types.

(**a**) A collection of three diverse datasets (9032 mouse neurons, 2401 mouse HSCs, and 4510 human macrophages) cluster separately without correction, as expected. (**b**) When given a collection of three diverse datasets with no overlapping cell types (mouse neurons, HSCs, and unstimulated macrophages), Scanorama finds a few spurious alignments between datasets, but none of the alignment scores pass the cutoff threshold of 10% (**e**). (**c**, **d**) scran MNN and Seurat CCA are more prone to overcorrection. (**f**) Both the uncorrected and Scanorama corrected data have the highest Silhouette Coefficients (both have a median of 0.37) compared to scran MNN (median of 0.20; independent, two-sided t-test P = 7e-252; n = 15,794 cells) and Seurat CCA (median of -0.12; P < 5e-324). Box plot boxes extend from lower to upper quartiles with an orange line at the median and green triangle at the mean; whiskers show the range. (**g**) A collection of six diverse datasets (2885 293T cells, 9032 mouse neurons, 2401 mouse HSCs, 4510 human macrophages, 8569 human pancreatic islet cells, and 18018 human PBMCs) cluster separately without correction, as expected. (**h**) When given the same collection of six diverse datasets with no overlapping cell types, Scanorama keeps disparate cell types separate with only a small

amount of overcorrection in matching a small portion of 293T cells with PBMCs. (**i**, **j**) Because they are not designed for heterogeneous dataset integration, both scran MNN and Seurat CCA integrate biologically disparate cell types among the same collection of datasets. (**k**) Scanorama alignment scores are at or very close to zero between the different datasets. (**l**) While the highest Silhouette Coefficient distribution belongs to the data without batch correction (median of 0.35), Scanorama has the least overcorrection among the datasets and has higher Silhouette Coefficients (median of 0.20) than scran MNN (median of 0.10; two-sided independent *t*-test $P = 5.3e\text{-}98$; $n = 36{,}755$ cells) and Seurat CCA (median of -0.18; $P < 5e\text{-}324$). A dagger (†) indicates a significantly higher Silhouette Coefficient distribution (Bonferroni corrected $P < 0.05$) between Scanorama and scran MNN, and a double dagger (‡) indicates significance over Seurat CCA. t-SNE visualizations use a learning rate of 200 and a perplexity of 400. Box plot boxes extend from lower to upper quartiles with an orange line at the median and green triangle at the mean; whiskers indicate the range.



Supplementary Figure 15

Scanorama alignment scores reconstruct temporal relationships between datasets.

Blue nodes indicate datasets and gray edges make up the maximum spanning tree (MST) on the graph with Scanorama alignment scores as the edge weights. In (**a**) mouse dendritic cells stimulated with LPS over 6 hours and (**c**) human CD14+ monocytes stimulated with M-CSF over 6 days, MST edges perfectly correspond to the temporal ordering of the datasets and only connect replicate timepoints or adjacent timepoints. In (**b**) *D. melanogaster* brain cells over 50 days, most edges connect replicate or adjacent timepoints except for edges between 3 and 9 days, between 1 and 6 days, and between 6 and 15 days, possibly indicating greater transcriptional similarity at the midpoint of the time series.