FIRM: Fast Integration of single-cell RNA-sequencing data across Multiple platforms

Jingsi Ming[1,2*], Zhixiang Lin[3,*], Xiang Wan[4], Can Yang[2,†], Angela Ruohao Wu[5,†]

[1] Academy for Statistics and Interdisciplinary Sciences, Faculty of Economics and Management, East China Normal University, Shanghai, China

[2] Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong SAR, China

[3] Department of Statistics, The Chinese University of Hong Kong, Hong Kong SAR, China

[4] Shenzhen Research Institute of Big Data, Shenzhen, China

[5] Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China

[†] Corresponding authors: macyang@ust.hk, angelawu@ust.hk

[*] These authors contributed equally to this work.

Abstract:

Single-cell RNA-sequencing (scRNA-seq) has now been used extensively to discover novel cell types and reconstruct developmental trajectories by measuring mRNA expression patterns of individual cells. However, datasets collected using different scRNA-seq technology platforms, including the popular SMART-Seq2 (SS2) and 10X platforms, are difficult to compare because of their heterogeneity. Each platform has unique advantages, and integration of these datasets would provide deeper insights into cell biology and gene regulation. Through comprehensive data exploration, we found that accurate integration is often hampered by differences in cell-type compositions. Herein we describe FIRM, an algorithm that addresses this problem and achieves efficient and accurate integration of heterogeneous scRNA-seq datasets across multiple platforms. We applied FIRM to numerous scRNA-seq datasets generated using SS2 and 10X from mouse, mouse lemur, and human, comparing its performance

in dataset integration with other state-of-the-art methods. The integrated datasets generated using FIRM show accurate mixing of shared cell type identities and superior preservation of original structure for each dataset. FIRM not only generates robust integrated datasets for downstream analysis, but is also a facile way to transfer cell type labels and annotations from one dataset to another, making it a versatile and indispensable tool for scRNA-seq analysis.

The advent of single-cell RNA-sequencing (scRNA-seq) technology has enabled discovery of new cell types[1], understanding of dynamic biological processes[2,3] and spatial reconstruction of tissues[4]. Ongoing advancement in scRNA-seq technology has led to vast improvements in the scale and cost of the experiments[5–8], providing unprecedented opportunities for deep biological insight. As scRNA-seq becomes more widely accessible both in availability and cost of the technology, many single cell transcriptomic datasets have now been generated for the same tissue types in various organisms, but often using different techniques and technology platforms. Prominent examples of this are recent efforts to generate cell atlases for whole organisms, including the Tabula Muris[9] and Tabula Microcebus consortium projects. These cell atlas projects have provided scRNA-seq datasets encompassing a comprehensive set of tissues from the organism of interest, generated using two different single-cell profiling technology platforms, SMART-seq2 (SS2) and 10X Chromium (10X). Integrating datasets from different platforms not only enables the transfer of information from one dataset to another, such as the transfer of cell-type labels and annotations, but also makes the atlases more comprehensive and cohesive, which then benefits downstream biological analyses. However, complex technical variations and heterogeneities that exist between datasets makes integration challenging.

SS2 and 10X are two frequently used scRNA-seq platforms with their unique strengths and weaknesses. SS2 is a plate-based full-length approach with high transcriptome coverage per cell and greater sensitivity[10], whereas the microfluidic droplet-based method, 10X, generally has lower coverage per cell and a higher

dropout rate[11] . But 10X is able to profile hundreds of thousands of cells per study with low per cell costs[8], which enables more reliable detection of rare cell types, and the inclusion of unique molecular identifiers (UMIs) in 10X allows removal of amplification bias and in turn enables more accurate transcript abundance quantification[12]. Harmonizing datasets across multiple platforms for integrative analysis can take advantage of the strengths of each technology and improve the robustness, as well as achieve higher accuracy for visualization; better comparison across datasets and studies; and higher statistical power for differentially expression analysis. Furthermore, integration would make it possible to use 10X for discovery of new cell types, while taking advantage of the greater depth and sensitivity of SS2 to investigate their biology, including enabling analysis of transcript isoforms, splicing[13–15], and allelic expression[14,16].

Existing methods have been designed for integration of scRNA-seq datasets across different samples, experiments, species or types of measurement. For example, batch correction methods based on mutual nearest neighbor in the original high-dimensional space, such as MNN[17] and Scanorama[18]; methods by identification of shared low-dimensional space, such as CCA[19], ZINB-WaVE[20] and scVI[21]; and methods which combine the previous two to identify shared subpopulations in the low-dimensional space, such as Seurat[22] and LIGER[23]. All these methods are applicable but not specifically designed for integration of datasets across multiple platforms. Due to the characteristic differences in SS2 and 10X datasets, technical variations preventing accurate integration often emerge under different scenarios: some result in poor alignment of cells from the same cell type; others may mix cells from different cell types inappropriately, giving rise to the overcorrection problem. The ideal method requires identification of the main technical variation for integration and designing a specific approach to address it.

Through comprehensive data exploration, we found that the heterogeneity in cell type composition between SS2 and 10X datasets accounts for the main technical variation preventing accurate integration. Datasets with different cell type compositions have different directions of maximum variance chosen by principal

component analysis (PCA) and perform differently after preprocessing. We have developed an efficient algorithm, FIRM, to specifically account for this composition effect thereby harmonizing SS2 and 10X datasets. Authors of other methods such as MNN and Scanorama have also observed the influence of cell type composition on integration and tried to reduce this effect by modifying the underlying expression data to align cells with high similarity. However, using this approach, other problems such as overcorrection can emerge, especially when there are dataset-specific cell types. In overcorrecting, close but not identical cell types may be merged inappropriately. In contrast, FIRM applies a re-scaling procedure based on subsampling for both datasets in a unified workflow. With this approach, overcorrection can be avoided and the original structure for each dataset can be largely preserved in the integrated dataset, thus generating a reliable input for downstream analysis. We applied our FIRM approach to integrate numerous scRNA-seq datasets generated using SS2 and 10X platform, and in comparison with existing state-of-the-art methods, FIRM not only demonstrates great integration performance but also effectively transfers cell type identity labels in all tested datasets.

Results

Differences in cell type composition is a major factor preventing accurate integration of scRNA-seq data generated by different technology platforms

To specifically investigate the influence of cell type composition on scRNA-seq dataset integration outcomes, we consider two scenarios using hypothetical datasets in which all technical variations have been removed except the difference in their cell type compositions. In the first scenario, the cell type proportions are consistent across different platforms (SS2: 50% cell type 1 + 50% cell type 2, 10x: 50% cell type 1 + 50% cell type 2); in the second scenario, the cell type proportions are different (SS2: 50% cell type 1 + 50% cell type 2, 10x: 80% cell type 1 + 20% cell type 2). We scaled the expression value for each gene to unit variance for each dataset, which is the standard preprocessing procedure applied to prevent the dominance of highly

expressed genes and is also necessary to reduce the difference in sequencing depth for dataset integration across platforms. In the first scenario, where cell type composition in SS2 and 10X datasets are the same, cells belonging to the same cell type have similar gene expression levels after scaling and are well mixed across platforms (Fig. 1a-c). This is the ideal case for downstream analysis, in which only the biological variations among cell types are reserved. However, in the second scenario, the scaled expression values for cells in SS2 and 10X datasets show large differences because of their heterogeneity in cell type composition (Fig. 1d, e), resulting in poor integration of these two datasets (Fig. 1f). This demonstrates that when cell-type composition is skewed between the two datasets being integrated, it impacts the integration outcome and can result in inaccurate cell merging.

To verify our hypothesis using real scRNA-seq datasets across multiple platforms, we analyzed the Tabula Muris mouse mammary gland scRNA-seq data that was generated using SS2 and 10X[9]. We extracted the basal cells and stromal cells in this dataset to create a simple example resembling the previously illustrated case two scenario, in which cell type proportions across platforms are vastly different (SS2: 75% basal cells + 25% stromal cells; 10X: 35% basal cells + 65% stromal cells). After preprocessing each dataset, we compare the gene expression of two marker genes for stromal cells (Vim and Fn1) and another two for basal cells (Krt5 and Krt14) between the SS2 and 10X dataset. We found that the expression levels for the same cell type marker across platforms are different in expression modes and dispersions (Fig. 2a-c). We then integrated the dataset by concatenating the scaled SS2 and 10X expression matrices, and in visualizing the outcome we found that in this cell type composition scenario (SS2: 75% basal cells + 25% stromal cells; 10X: 35% basal cells + 65% stromal cells), basal cells across platforms did not correctly merge into one single cluster (Fig. 2d, e, leftmost panel).

In order to confirm whether this poor alignment is caused by the difference in cell type proportions, we performed subsampling to gradually reduce the proportion of basal cells in SS2 dataset from 75% to 35%, to match that of the 10X dataset. Then we integrated the 10X dataset with these subsets of SS2 dataset, and evaluated the

performance. In addition to the UMAP plot for visualization, we also calculated the mixing metric (see Methods) to measure how well the datasets mixed after integration, where a lower score typically indicates better mixing performance. We indeed observed that more consistent cell type proportions gave rise to better alignments (Fig. 2d-f). Therefore, we concluded that the effects of heterogeneity in cell type composition between SS2 and 10X datasets accounts for one of the main technical variation preventing accurate integration of scRNA-seq data across platforms.

FIRM can provide accurate mixing of shared cell type identities and preserve local structure for each dataset.

FIRM harmonizes SS2 and 10X datasets while accounting for cell type composition. The alignment workflow takes two scRNA-seq expression matrices as the input, typically one SS2 and one 10X dataset, and performs the following steps: (i) For each dataset, we conduct the standard pre-processing procedure which includes normalization, scaling and feature selection; (ii) Then we perform dimension reduction for each dataset using PCA and cluster cells; (iii) In order to align clusters in 10X with clusters in SS2 of the same cell type, we check the alignment via subsampling to avoid overcorrection; (iv) For each paired cluster, we subsample the cells to ensure that cell-type proportions are the same in SS2 and 10X, then based on these cells, we calculate the standard deviation to perform re-scaling on each of the full datasets; (v) Finally, we merge the scaled data to obtain the integrated dataset. (Methods)

We applied FIRM to numerous SS2 and 10X scRNA-seq datasets, both published and unpublished, to show its superior performance in integration compared with existing state-of-the-art methods (Supplementary Figs. 1-40; SI unpublished data is embargoed until publication).

To demonstrate accurate integration using FIRM, we again used the mammary gland scRNA-seq datasets from Tabula Muris. Cells in SS2 dataset include basal cells, endothelial cells, luminal epithelial cells of mammary gland and stromal cells.

Three additional cell types, B cells, macrophages and T cells, are detected in 10X dataset. We performed UMAP on the integrated or corrected data for the highly variable genes using FIRM and six other methods (Seurat[22], LIGER[23], Scanorama[18], MNN[17], scVI[21] and ZINB-WaVE[20]) (Fig. 3a, b). To evaluate integration performance, we use of four metrics (Methods): mixing metric, local structure metric, average silhouette width (ASW) and adjusted rand index (ARI). The local structure metric is designed to measure how well the original structure of each dataset was preserved after integration, where a lower score indicates worse preservation and higher probability of overcorrection. ASW and ARI are calculated based on cell identities given in the Tabula Muris cell annotation. Higher values of ASW indicate that cells of the same type are closer to each other and are father from cells of other types. Higher values of ARI indicate higher similarities between the clustering of integrated data with the predefined cell types. As these metrics show different aspects of performance, joint consideration is required for effective comparison. For example, there is a trade-off between the mixing metric and the local structure metric – a low mixing metric does not always mean accurate integration, since overcorrection would be characterized by a low mixing metric and a low local structure metric.

Among all the methods compared, FIRM demonstrated the best integration performance with relatively low mixing metric, high ASW, and the highest local structure metric and ARI (Fig. 3c). Based on the UMAP plots and high-level mixing metrics, scVI and ZINB-WaVE have lower mixing performance. LIGER had a low ARI and inappropriately mixed B cells, macrophages, and T cells together, which are disparate cell types that only exist in the 10X dataset. Although Seurat scored a mixing metric comparable to that of FIRM (FIRM 68.84, Seurat 68.07), and a slightly lower ASW and ARI compared to FIRM (ASW: FIRM 3.16, Seurat 3.01; ARI: FIRM 0.68, Seurat 0.65), its local structure metric is much lower than that of FIRM (FIRM 0.58, Seurat 0.41).

We went on to test FIRM extensively on several other scRNA-seq datasets, including 12 pairs of data from Tabula Muris[9], 27 pairs from Tabula Microcebus (unpublished), and one pair in Human Lung Cell Atlas[24] (Methods). In these tests,

FIRM outperformed or was comparable in performance to all other benchmarked methods for integration of SS2 and 10X datasets (Supplementary Figs. 1-40; SI embargoed until publication). Specifically, scVI and ZINB-WaVE are the two methods with the highest mixing metrics for the most cases, indicating their poor mixing performance. LIGER overcorrected the datasets for some cases and inappropriately merged different cell types, resulting in low ARI. For example, LIGER incorrectly grouped together the kidney collecting duct epithelial cells with the kidney loop of Henle ascending limb epithelial cells in the Tabula Muris kidney dataset (Supplementary Fig. 4; SI embargoed until publication); the natural killer cells with the T-cells in the Human Lung Atlas (Supplementary Fig. 41; SI embargoed until publication); and the subtypes of natural killer cells with the T-cells in the Tabula Microcebus spleen data from lemur 2 (Supplementary Fig. 40; SI embargoed until publication). MNN is also prone to overcorrecting the data and showed low ARIs for colon (Supplementary Fig. 17; SI embargoed until publication), eye retina (Supplementary Fig. 18; SI embargoed until publication), and pancreas (Supplementary Fig. 22; SI embargoed until publication) from lemur 4 in Tabula Microcebus. Scanorama was observed to have poor mixing performance when applied to the Tabula Muris spleen data (Supplementary Fig. 10; SI embargoed until publication), as well as to the Tabula Microcebus bone marrow data from lemur 2 (Supplementary Fig. 33; SI embargoed until publication). It also incorrectly mixed cells from different cell types, such as in the subtypes of pancreatic acinar cells in pancreas from lemur 2 in Tabula Microcebus (Supplementary Fig. 39; SI embargoed until publication). Seurat is the method with the closest performance to FIRM (Supplementary Fig. 42; SI embargoed until publication). Seurat and FIRM have comparable performance in terms of ASW, but FIRM is superior in terms of ARI. Although Seurat usually has lower mixing metrics, FIRM does not show any obvious deficiency for mixing based on the UMAP plots of the integrated dataset. Considering the trade-off between the mixing metric and local structure metric, FIRM's higher local structure metric suggests that it is more robust than Seurat in avoiding overcorrection.

One of the greatest advantages of FIRM is that it preserves the local structure for each dataset. This is because FIRM harmonizes datasets using a re-scaling procedure without modifying the underlying expression data for each cell separately, so that the relative expression patterns across cells within each dataset can be largely preserved. For all the integrated datasets, we found that FIRM achieved the highest local structure metric comparing with all the methods (Supplementary Fig. 43; SI embargoed until publication), indicating the minimal distortion of cell-type relationship, thus providing more credible integrated data for downstream analysis.

FIRM is robust against overcorrection

Other methods of integration do so by directly adjusting the data matrices so that neighboring cells across different datasets have similar adjusted expression profiles. This process may be vulnerable to overcorrection because the cells that are close in distance across datasets may not always be biologically similar. We have shown that even with the prerequisite that at least one cell type is shared across datasets, other methods cannot completely avoid overcorrection. Different from other methods, which project reference dataset onto query dataset based on neighboring cells across datasets, FIRM harmonizes datasets by incorporating scaling factors that account for differences in cell type compositions across datasets. As a result, FIRM can avoid overcorrection even if there are no overlapping cell types across datasets. To demonstrate the robustness of FIRM to overcorrection, we performed integration of two datasets with no shared cell types: SS2 dataset of kidney, and 10X dataset of brain cortex of lemur 2 in Tabula Microcebus. We manually removed several cell types in each dataset to ensure that they have no cell types in common. We applied FIRM, Seurat, LIGER, Scanorama, and MNN to integrate these two datasets (Fig. 4); we excluded scVI and ZINB-WaVE from this assessment as these two methods did not work well even when there were shared cell types across datasets. Of all the methods assessed, only FIRM perfectly separated the cell types from each dataset. Seurat, LIGER, and MNN incorrectly mixed neurons with B-cells and T-cells, indicating overcorrection. The advantage of local structure preservation is especially

beneficial under this scenario where datasets are divergent with few or no shared cell populations. The local structure metric of FIRM (0.52) is much greater than Seurat (0.33), LIGER (0.28), Scanorama (0.33) and MNN (0.34).

FIRM can provide better clustering and transfer cell type identity labels across datasets.

By harmonizing SS2 and 10X datasets, we are able to perform clustering on the integrated dataset which can provide more reliable and comparable cluster labels for each dataset. Since 10X datasets have higher throughput, more cell types are expected to be captured. However, in the SS2 dataset, some cell types may contain very few cells and fail to be identified. The clustering can be refined by analyzing the SS2-10X integrated dataset. Cell type which is unseparated or weakly separated with other cell types when analyzing SS2 dataset alone may form a nicer cluster in the integrated dataset. For example, in the Tabula Microcebus testes SS2 dataset, we are not able to identify spermatogonia as there are only a few of them. By incorporating the information from 10X dataset, spermatogonia are well clustered including three cells from the SS2 dataset that show marker expression patterns (KIT, SOHLH1, PHOXF1, ZBTB16) consistent with the spermatogonia in the 10X dataset (Supplementary Fig. 44; SI embargoed until publication). We also noted that some markers (OVOL1, SPO11, TEX101) show clearer patterns in the SS2 dataset compared with the 10X dataset, indicating the benefit of detecting low abundance transcripts using SS2.

We are also able to transfer information between datasets. One way to effectively label cell populations in SS2 data is by transferring the manually annotated 10X cell type identity labels to SS2 cells by detecting nearest neighbors for each SS2 cell in 10X dataset (Methods). For cases where the SS2 dataset contains more cell types than 10X dataset, we designed match scores such that cells with low scores can be labeled as unknown (Methods). We compared FIRM with Seurat, which also has a function to transfer cell labels. The prediction scores designed in Seurat measures whether the cell in SS2 is present in 10X data, and a lower score indicates greater likelihood of being an unknown cell type. In the Tabula Muris heart dataset, there are three cell

types that are only found using SS2 and not 10X: leukocytes, myofibroblast cells, and smooth muscle cells. For cells that are found in both SS2 and 10X datasets, both FIRM and Seurat showed effective transformation (Fig. 5a-g). However, Seurat misclassified myofibroblast cells and smooth muscle cells as fibroblasts. In contrast, FIRM assigned these cells low match scores, indicating that they belong to unknown cell types.

We conducted label transfer from 10X to SS2 for 27 tissues in Tabula Microcebus and calculated the accuracy of label transfer based on the cell type identities manually annotated by experts (Fig. 5h). To quantify the performance of label transfer, we consider two types of accuracy, cell-type accuracy and cell-subtype accuracy (Methods). The cell-type accuracy measures the proportion of cells where the transferred label is consistent with the manual expert-annotated label at a high level of cell-type classification, an example would be “T-cell”; the cell-subtype accuracy measures the proportion of cells where the transferred label is consistent with the manual expert-annotated label down to the subpopulation level, such as “CD4 T-cell”. Compared with Seurat, FIRM was shown to achieve higher accuracy in both the cell-type and the cell-subtype levels in most tissues, indicating its better performance in label transfer.

Discussion

FIRM provides an efficient method for integration of scRNA-seq datasets across multiple platforms. The integrated dataset can help to answer relevant biological questions and increase the confidence of analytical conclusions. For downstream analysis to be biologically meaningful, it is important to minimize technical variations such as batch effects while preserving biological variations of interest. Generally, it is very difficult to distinguish technical from biological variation, and overcorrection can occur when attempting to remove technical variation, resulting in loss of critical underlying biological variations. The best way to avoid overcorrection is to design methods that target minimization of specific types of confounding variation. FIRM

successfully does so by specifically accounting for the heterogeneity in cell type composition between datasets. In addition to datasets across different platforms, FIRM can also be applied to integrate datasets across different individuals (Supplementary Figs. 45-46; SI embargoed until publication). As FIRM only removes the effect from cell type composition differences, the integrated dataset can be used to study other differences across samples, such as the effect of age, sex, and disease pathology. Other existing integration methods that use a general approach to account for variation between datasets do so by aligning cells with high similarity, and as such they are prone to inadvertently removing the biological differences across individuals as well. In contrast with existing methods, FIRM requires no assumption about shared cell populations between datasets, and is therefore applicable even without prior knowledge about the dataset composition.

FIRM integrates datasets based on the intersection of highly variable genes in each dataset. This preserves the most important expression patterns with less noise for alignment. However, we cannot ensure that the expressions of all genes are harmonized between SS2 cells and 10X cells, especially for low abundance transcripts. This is a remaining challenge to be addressed by future method development.

Through analysis of a diverse collection of human, mouse, and mouse lemur datasets, we show that FIRM outperforms or performs comparably to existing methods in terms of accuracy of integration and superior preservation of original structure for each dataset. Ultimately, our data integration tool enables new biological insights, and provides efficiency and utility for large scale projects, as demonstrated by its use in the Tabula Microcebus study (unpublished).

Methods

Data preprocessing. For all scRNA-seq datasets, we performed the standard pre-processing workflow in Seurat[22], which includes normalization, scaling and feature selection.

For each dataset, we employed the log-normalization which is the default normalization method in Seurat. We used the gene expression matrix $\boldsymbol{X}$, where X_{ij} is the number of reads (for SS2) or unique molecular identified (UMI, for 10X) for gene i that are detected in cell j. For each cell, the feature counts are divided by the total counts for that cell, multiplied by a scale factor M and then transformed using log1p: $Y_{ij} = log\left(1 + M\frac{X_{ij}}{\sum_{j\prime} X_{ij\prime}}\right)$, where M=10,000, the log is base e.

Then we scaled the expression values for each gene across all cell in each dataset so that the variance across cells is 1. Different from the default scaling method in Seurat, we didn't center the expression to have zero mean across cells for the convenience of the following scaling adjustment steps.

In order to highlight the biological signal in scRNA-seq datasets, we aimed to identify genes with high variability across cells. For each dataset, we performed the FindVariableFeatures function in Seurat to select highly variable genes, where it ranked genes according to the dispersion after controlling the mean expression. In default, we selected the top 4,000 genes. For integrated analysis of two datasets across platforms, we were required to select genes which were detected as highly variable genes for both datasets in order to capture the connection between the datasets. Therefore, we chose the overlapped highly variable genes for each dataset for the following dataset integration.

PCA and cell clustering for each data set. To reduce the extensive technical noise and make the data more tractable, we performed the traditional dimensionality reduction method, PCA, for each dataset on the scaled data using only the highly variable genes selected for integrated analysis. We aimed to obtain the low-dimensional embedding of the expression values for each cell. We chose the number of PCs according to its relationship with the variance explained. For integrated analysis of two datasets across platforms, the number of PCs need to be the same. Typically, larger dimensionality is needed for larger dataset. For the scRNA-seq datasets analyzed in this paper, we found the dimensionality is not very sensitive. We

chose dimensionality as the larger dimensionality in separately analysis for SS2 and 10X dataset in the original paper.

The analysis in this paper is unsupervised, where we supposed the cell annotations were unknown. For each dataset, we cluster cells based on their PCA scores using the clustering approach in Seurat which identify clusters of cells by a shared nearest neighbor (SNN) modularity optimization-based clustering algorithm. A resolution parameter can be used to control the number of clusters, which can be tuned for better integration.

Cluster alignment via subsampling. We aimed to align clusters in 10X with clusters in SS2 of the same cell type. We checked the alignment via subsampling to avoid overcorrection. First, we obtained the center for each cluster based on cell embeddings. Next, for each 10X cluster, we considered its five nearest SS2 cluster for whether cells in any SS2 cluster are from the same biological state as those in this 10X cluster. We calculated the distance between the center of the 10X cluster to each of the five SS2 clusters and checked whether it is less than the 75% quantile for the current SS2 cluster. We considered the nearest SS2 cluster that satisfied this criterion to be aligned with the 10X cluster. However, clusters may not be aligned although they had the same biological state because of their difference in proportions. Therefore, when the proportions of the 10X cluster and SS2 cluster being considered are different, we first performed subsampling to reduce the number of cells in the cluster with larger proportion in order to achieve the same cell-type proportion. Then we calculated the standard deviation of each gene across the current cells and re-scaled the expression values for each dataset. Based on the re-scaled data, we checked again based the criterion described above.

Re-scaling via subsampling and generation of integrated data. Based on the cluster pairs identified using the above procedure, we performed subsampling for cells in either the clusters of SS2 or 10X dataset to ensure their cell-type compositions were the same. The subsampled datasets not only contained the same type of cells but

also had same cell-type proportion. We computed the standard deviation for each gene across cells for each of the subsampled datasets and using this standard deviation to rescale the gene expression values. We merged the rescaled data directly to obtain the integrated dataset for downstream analysis.

Integration Metrics

Mixing metric. We used the mixing metric designed in Seurat[22] to evaluate how well the datasets mixed after integration. If the local neighborhood for a cell is well mixed across datasets, at least a small number (k=5) of cells from each dataset is assumed to be its neighbors. For each cell, we obtain its (k.max=300) ranked nearest neighbors in the integrated dataset and record the rank of the 5th nearest neighbor in each dataset (with a max of 300). The average of the ranks across all cells is defined as the mixing metric. As a result, smaller mixing metric typically indicates better mixing.

Local structure metric. We used the local structure metric designed in Seurat[22] to determine how well the original structure of each dataset was reserved after integration. For each cell, we compare its k=100 nearest neighbors in the original dataset and the integrated dataset. The average value of the fraction of overlapped neighbor across all cells is defined as the local structure metric. A large local structure metric indicates good preservation.

Average silhouette width (ASW). The silhouette width for cell i from cell type c is $s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$, where a_i is the average distance from cell i to all cells in cell type c, b_i is lowest value of average distances from cell i to all cells for each cell type other than c. ASW is the mean of silhouette widths across all cells, where a higher score indicates cells are closer to cells of the same cell type and are more far from cells of different cell types. We calculated ASW based on the predefined cell identities and low-dimensional embedding space of the integrated dataset using PCA.

Adjusted rand index (ARI). ARI measures the similarity between two clustering. The ARI is defined as $\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$, where n_{ij}, a_i, b_j are

values from the contingency table. For the integrated dataset, we clustered cells based on their PCA scores using the clustering approach in Seurat with default settings. Then we calculated ARI to compare the clustering of integrated data with the predefined cell types, where higher values indicate higher similarities.

Resolution in clustering. The appropriateness of cluster alignment is the key for effective integration. Unsuitable resolutions for clustering of SS2 and 10X cluster may lead to bad integration results. We searched the best resolution pairs for SS2 and 10X datasets in a default range of 0.1 to 2. Under each resolution pair, we integrated the datasets using our method and calculated the mixing metric. As our method would not suffer overcorrection, smaller mixing metric always indicates better integration. Therefore, we chose the pair yielding the smallest mixing metric.

Baseline model. We considered the special case without re-scaling to be the baseline model. We directly merged the scaled expression matrix for the overlapped highly variable genes after data processing to obtain the integrated dataset. If the mixing metric of integrated dataset after re-scaling does not decrease, we chose the baseline model.

Label transfer and match scores. The integration of datasets enables efficient label transfer between datasets. Suppose we only have the annotations for cells in 10X dataset and would like to use this information to annotate cells in SS2 dataset. For each SS2 cell, we found its 10 nearest 10X cells in the integrated dataset and summarized the cell types they were belonging to. We chose the cell type with the highest frequency to annotate the SS2 cell.

In case that some cell types do not exist in 10X dataset, we defined the match score to measure whether the cell in SS2 is present in 10X data. For each SS2 cell, we divided its averaged distance to its 10 nearest neighbors in SS2 dataset by that in 10X dataset. Lower score means less likely to be present in 10X data.

Cell-type accuracy and cell-subtype accuracy. We compared the transferred label with the predefined cell type identities manually performed by experts. The cell-subtype accuracy is the proportion of cells with the exact same transferred cell subtype with the predefined cell type. To account for potential differences across 10X and SS2 platforms in the classification of cell-subtypes within a broader cell-type category, we also defined a cell-type accuracy measure, where cells which is transferred to subtypes of the predefined cell type would also be classified as effective transfer.

Integration for Tabula Muris. We selected the tissues which have been sequenced using both SS2 and 10X. We integrated SS2 and 10X datasets for each tissue. We only considered the cells with annotations in the original study.

Integration of Tabula Microcebus. We selected the tissues which have been sequenced using both SS2 and 10X in each individual. We integrated SS2 and 10X datasets for each tissue in each individual. We only considered the cells with annotations in the original study.

Data availability

The datasets in Tabula Muris used in this manuscript are available at http://tabula-muris.ds.czbiohub.org/.

Code availability

FIRM code is available as Supplementary Code and at https://github.com/mingjingsi/FIRM.

1. Villani, A. C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science 356, (2017).

2. Treutlein, B. et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509, 371–375 (2014).

3. Enge, M. et al. Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. Cell 171, 321-330.e14 (2017).

4. Halpern, K. B. et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. Nature 542, 1–5 (2017).

5. Zilionis, R. et al. Single-cell barcoding and sequencing using droplet microfluidics. Nature Protocols 12, 44–73 (2017).

6. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161, 1187–1201 (2015).

7. Gierahn, T. M. et al. Seq-Well: Portable, low-cost rna sequencing of single cells at high throughput. Nature Methods 14, 395–398 (2017).

8. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. Nature Protocols vol. 13 599–604 (2018).

9. Schaum, N. et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature 562, 367–372 (2018).

10. Baran-Gale, J., Chandra, T. & Kirschner, K. Experimental design for single-cell RNA sequencing. Briefings in Functional Genomics 17, 233–239 (2017).

11. Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. Nature Communications 11, 1169 (2020).

12. Islam, S. et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nature Methods 11, 163–166 (2014).

13. Shalek, A. K. et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 498, 236–240 (2013).

14. Marinov, G. K. et al. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. Genome Research 24, 496–510 (2014).

15. Song, Y. et al. Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron Differentiation. Molecular Cell 67, 148-161.e5 (2017).

16. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 343, 193–196 (2014).

17. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nature Biotechnology 36, 421–427 (2018).

18. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nature Biotechnology 37, 685–691 (2019).

19. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology 36, 411–420 (2018).

20. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J. P. A general and flexible method for signal extraction from single-cell RNA-seq data. Nature Communications 9, 1–17 (2018).

21. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. Nature Methods 15, 1053–1058 (2018).

22. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. Cell 177, 1888-1902.e21 (2019).

23. Welch, J. D. et al. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell 177, 1873-1887.e17 (2019).

24. Travaglini, K. J. et al. A molecular cell atlas of the human lung from single cell RNA sequencing. bioRxiv 7191, 742320 (2019).

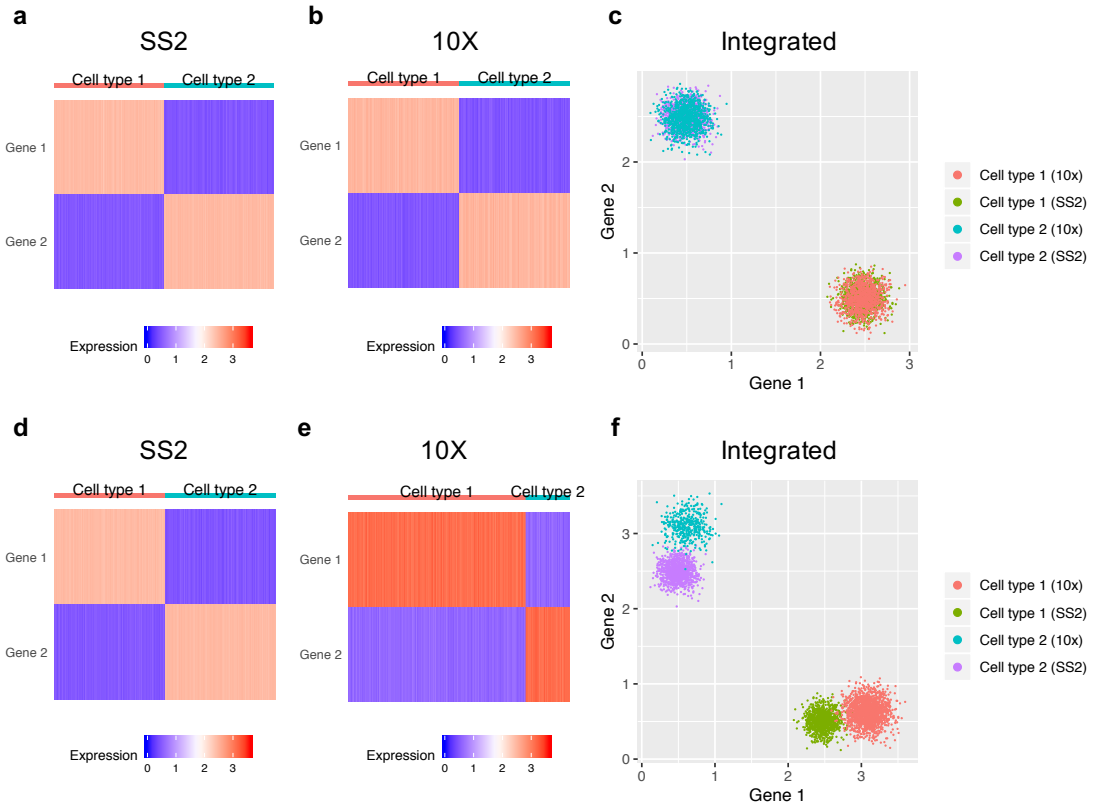

Fig. 1 Illustration of the influence of cell type composition for scRNA-seq datasets integration. a, **b**, **c**, Gene expressions for cells in SS2 dataset (**a**), 10X dataset (**b**) and integrated dataset (**c**) after scaling to unit variance for each gene, when the cell type compositions are the same across datasets (SS2: 50% cell type 1 + 50% cell type 2; 10X: 50% cell type 1 + 50% cell type 2). **d**, **e**, **f**, Gene expressions for cells in SS2 dataset (**d**), 10X dataset (**e**) and integrated dataset (**f**) after scaling to unit variance for each gene, when the cell type compositions are different across datasets (SS2: 50% cell type 1 + 50% cell type 2; 10X: 80% cell type 1 + 20% cell type 2).

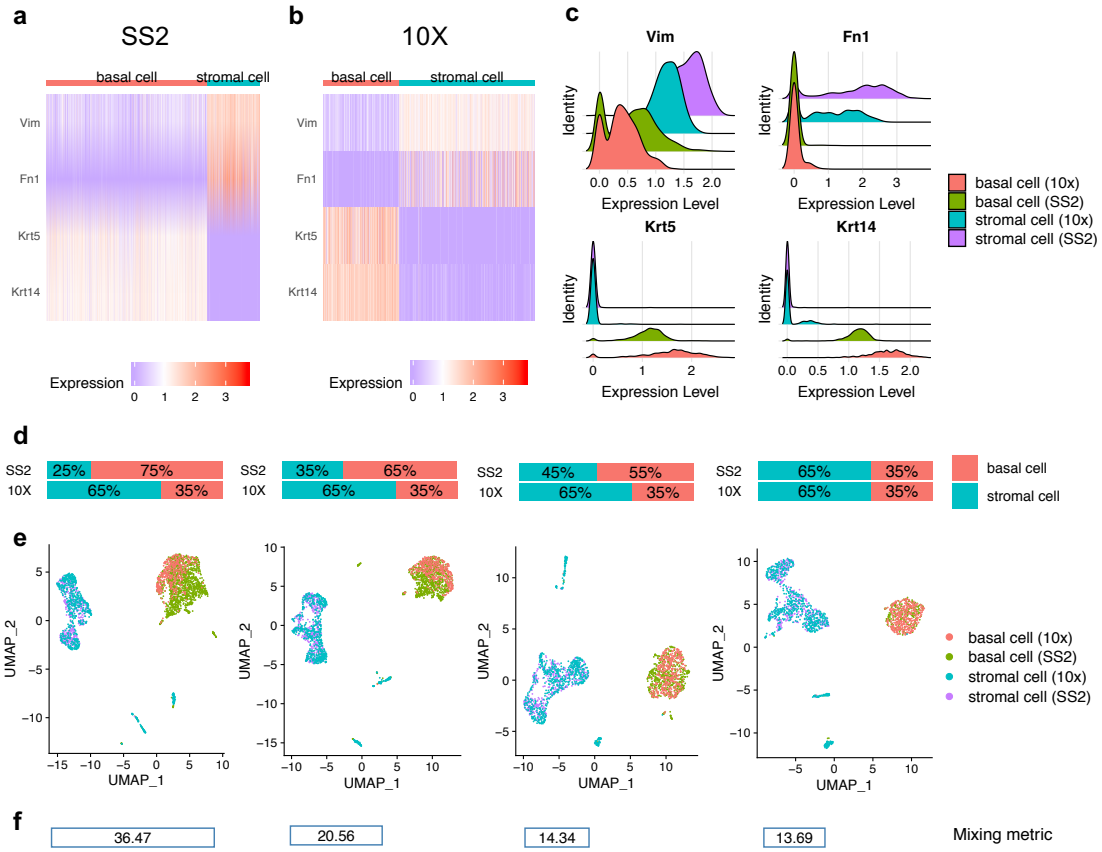

Fig. 2 Illustration of the key problem for integration based on the mammary gland scRNA-seq datasets generated by SS2 and 10X from Tabula Muris, withholding only the basal cells and stromal cells. a, **b**, Marker expressions for basal cells and stromal cells in SS2 dataset (**a**) and 10X dataset (**b**) after scaling to unit variance for each gene, where the cell type compositions are different across datasets (SS2: 75% basal cells + 25% stromal cells; 10X: 35% basal cells + 50% stromal cells). **c**, Ridge plots for marker expressions for each cell type in each dataset. **d**, **e**, **f**, Uniform manifold approximation and projection (UMAP) visualization (**e**) and mixing metric (**f**) for the integrated dataset with different cell type composition (**d**) by subsampling basal cells in SS2 dataset.

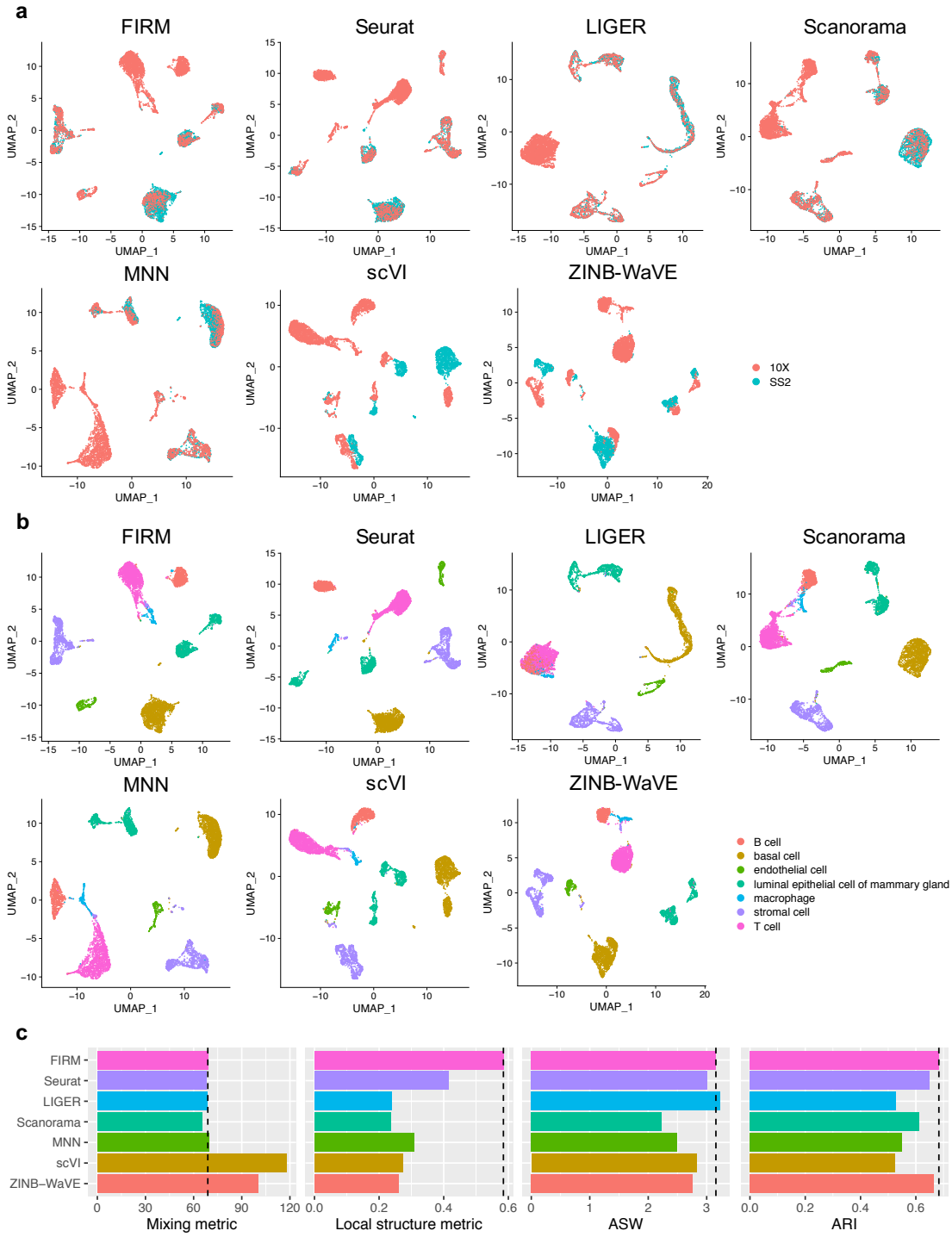

Fig. 3 Comparison of integration methods based on the mammary gland scRNA-seq datasets generated by SS2 and 10X from Tabula Muris. a, **b**, UMAP plots of mammary gland scRNA-seq datasets colored by platform (**a**) and by cell type (**b**) after integration using FIRM, Seurat, LIGER, Scanorama, MNN, scVI and ZINB-WaVE. **c**, Metrics for evaluating performance across the seven methods on four properties: cell mixing across platforms (Mixing metric), the preservation of within-dataset local structure (Local structure metric),

average silhouette width of annotated subpopulations (ASW) and adjusted rand index (ARI). The dashed lines were set at the values for FIRM as reference lines.

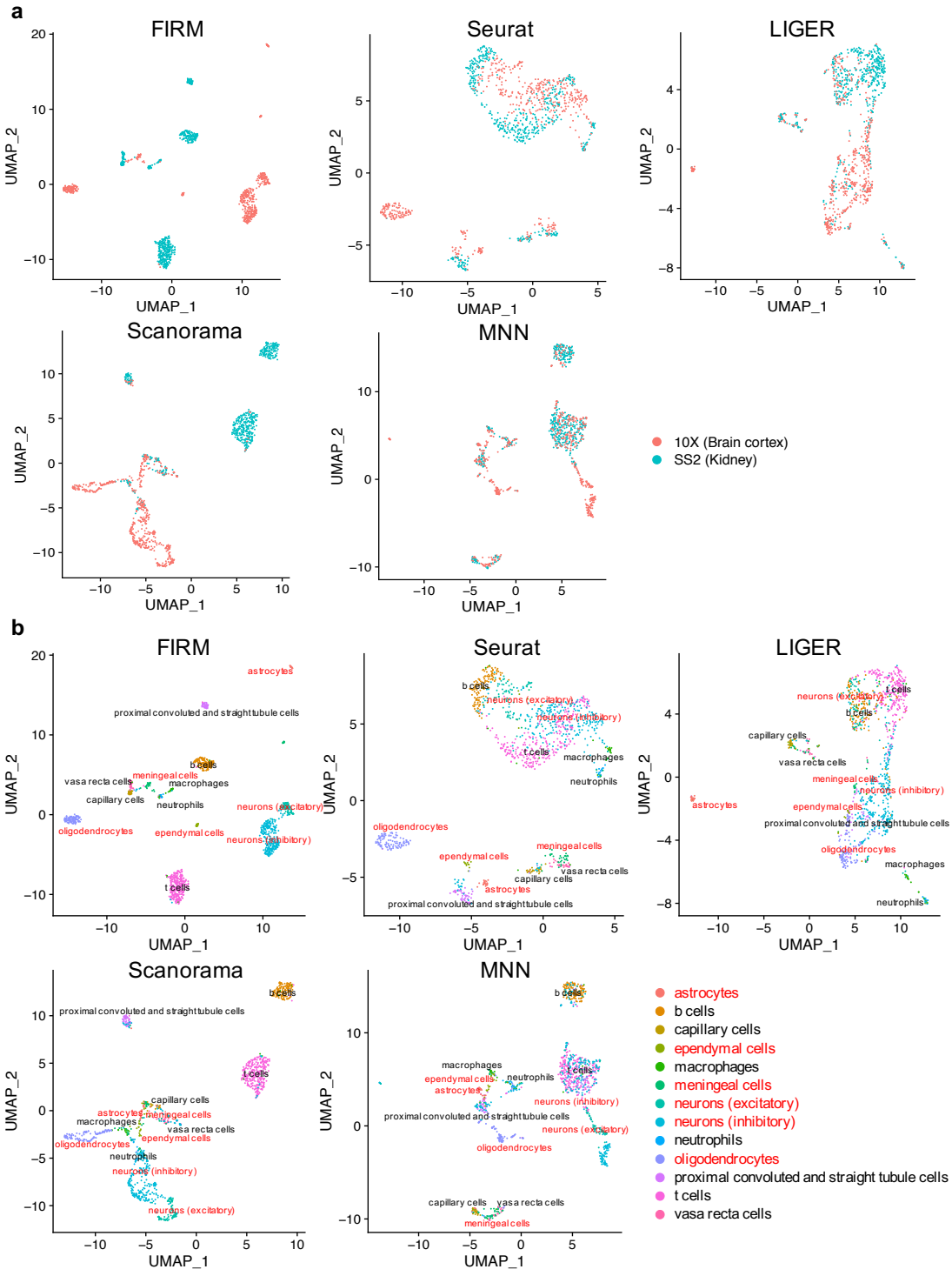

Fig. 4 Comparison of integration methods for scRNA-seq datasets from two tissues in Tabula Microcebus (lemur 2) generated by different platforms (Kidney: SS2, Brain cortex: 10X. For clear illustration, we withheld several cell types in each of the dataset to make the cell types non-overlapped across datasets. **a**, **b**, UMAP plots of scRNA-seq datasets colored by platform (**a**) and by cell type (**b**) after integration using FIRM, Seurat, LIGER, Scanorama and MNN. The labels for cell types in Brain cortex (10X) are colored by red.

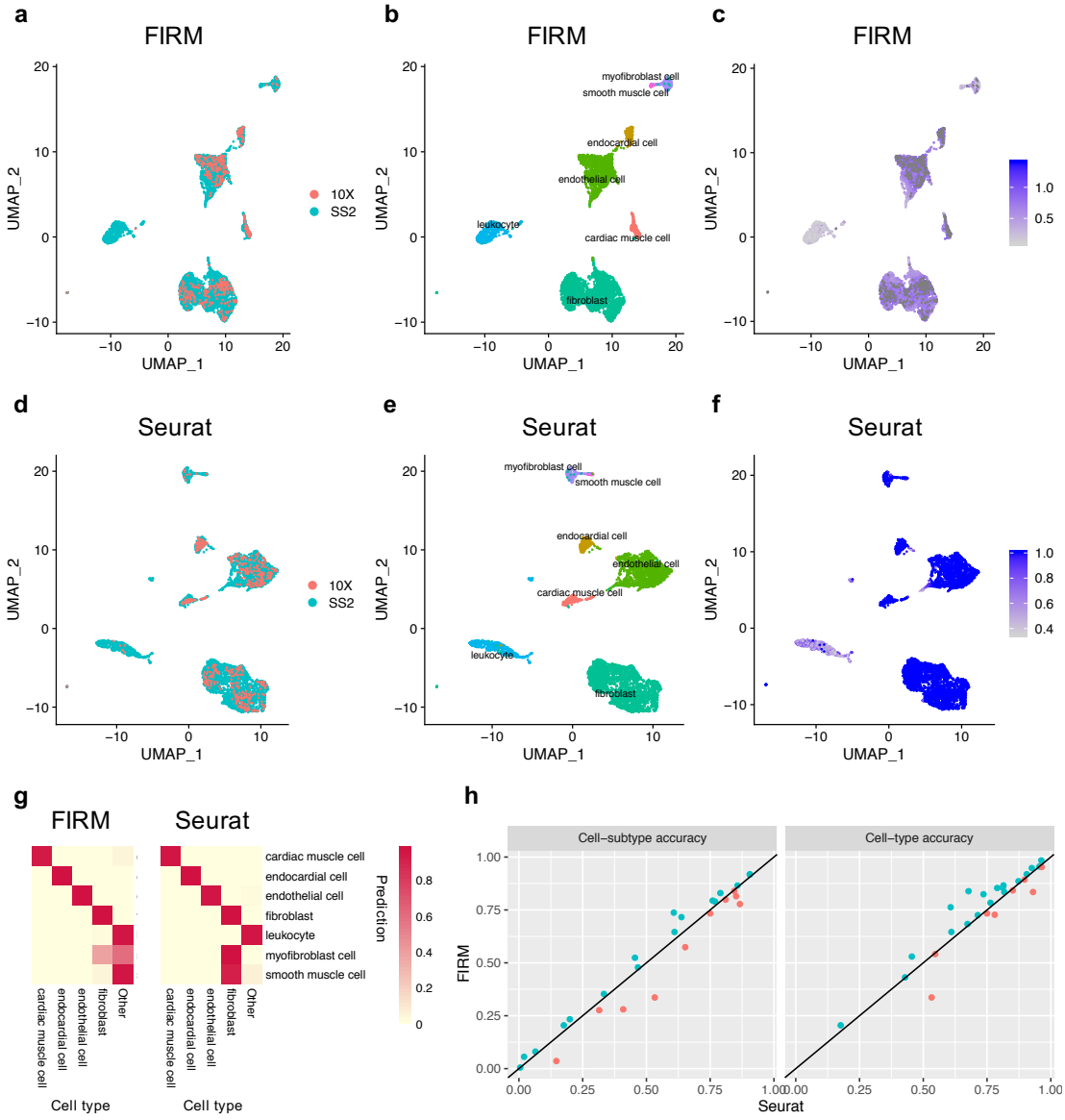

Fig. 5 Comparison of FIRM and Seurat for label transfer from 10X dataset to SS2 dataset. a-c, UMAP plots of the Heart scRNA-seq datasets from Tabula Muris after integration using FIRM colored by platform (**a**), by cell type (**b**) and by match scores (**c**, cells with low scores are more likely to belong to unknown cell types). **d-f**, UMAP plots of the Heart scRNA-seq datasets from Tabula Muris after integration using Seurat colored by platform (**d**), by cell type (**e**) and by prediction scores (**f**, cells with low scores are more likely to belong to unknown cell types). **g**, Confusion matrix for evaluation of label transfer using FIRM and Seurat based on the Heart scRNA-seq datasets from Tabula Muris. **h**, Accuracy of label transfer from 10X dataset to SS2 dataset for 27 tissues in Tabula Microcebus using

FIRM and Seurat. The tissues with higher or equivalent accuracy using FIRM are colored by blue.