biomolecules

Article

Automatic Cell Type Annotation Using Marker Genes for Single-Cell RNA Sequencing Data

Yu Chen [1] and Shuqin Zhang [1,2,3,*]

[1] School of Mathematical Sciences, Fudan University, Shanghai 200433, China
[2] Key Laboratory of Mathematics for Nonlinear Science, Ministry of Education, Fudan University, Shanghai 200433, China
[3] Shanghai Key Laboratory for Contemporary Applied Mathematics, Fudan University, Shanghai 200433, China
* Correspondence: zhangs@fudan.edu.cn

Abstract: Recent advancement in single-cell RNA sequencing (scRNA-seq) technology is gaining more and more attention. Cell type annotation plays an essential role in scRNA-seq data analysis. Several computational methods have been proposed for automatic annotation. Traditional cell type annotation is to first cluster the cells using unsupervised learning methods based on the gene expression profiles, then to label the clusters using the aggregated cluster-level expression profiles and the marker genes' information. Such procedure relies heavily on the clustering results. As the purity of clusters cannot be guaranteed, false detection of cluster features may lead to wrong annotations. In this paper, we improve this procedure and propose an Automatic Cell type Annotation Method (ACAM). ACAM delineates a clear framework to conduct automatic cell annotation through representative cluster identification, representative cluster annotation using marker genes, and the remaining cells' classification. Experiments on seven real datasets show the better performance of ACAM compared to six well-known cell type annotation methods.

Keywords: cell type annotation; marker genes; scRNA-seq

Citation: Chen, Y.; Zhang, S. Automatic Cell Type Annotation Using Marker Genes for Single-Cell RNA Sequencing Data. *Biomolecules* **2022**, *12*, 1539. https://doi.org/10.3390/biom12101539

Academic Editors: Vladimir Babenko, Gang Hu and Kui Wang

Received: 6 August 2022
Accepted: 11 October 2022
Published: 21 October 2022

1. Introduction

The development of single-cell RNA sequencing (scRNA-seq) technology has provided the opportunity for studying genes' expression at each single-cell level [1]. It has greatly advanced the understanding of biology and medicine in many aspects by analyzing the transcriptome-wide cell-to-cell variations. For example, investigation of the heterogeneity of different cell types in cancer ecosystems contributes to studying the disease progression and response to therapy [2–5], and exploration of the cell type transitions benefits studying the cell-state progression in the developing embryos [6,7]. With the wide applications of scRNA-seq technology, more and more scRNA-seq data from different platforms are being generated.

Annotation of the cell types plays an essential role in scRNA-seq data analysis. Several computational methods have been proposed for automatic annotation [8–20]. According to the databases used for conducting annotation, such methods can be divided into two categories. One is to take the previously annotated scRNA-seq database as reference for labelling the unannotated cells (reference scRNA-seq-data-based) [8,9,11,13,14,18,21]. Additionally, the other category is to directly use the marker genes to annotate the cells (marker-gene-based) [16,17,19,22].

The reference scRNA-seq-data-based cell type annotation methods can be divided into several modelling frameworks. Some of these methods map the unannotated cells to the previously annotated reference datasets using selected features, and then assign them the cell types according to their nearest neighbors based on some similarity measures. Such methods include SingleR [9], scmap [14], scMatch [11], cellHarmony [23], SeuratTransfer [21], and so on. Some other methods belonging to this category directly train a supervised learning model in the annotated reference database, and then predict the cell types of those unannotated, for example, scPred [8], CHETAH [13], and scDeepsort [18]. Deep learning methods have also been proposed on the basis of the annotated scRNA-seq database, such as MARS [10], and ItClust [12]. Since heterogeneity exists between different datasets, this category of methods puts forward high requirements for the cell type matching across different datasets. Some methods are developed using the annotated cells from the same dataset to infer the cell types of the remaining, for example, CASSL [24]. For such methods, to obtain the labels of part of the cells is still the cell type annotation problem.

The marker-gene-based cell type annotation methods also fall into different types. CellAssign takes into account the prior knowledge of cell type specific marker genes into the proposed probabilistic model to infer the type of each cell [19], which is unstable due to the large noise in scRNA-seq data. Garnett first labels a number of representative cells by scoring the marker genes, and then uses logistic regression with elastic net to classify the remaining cells [16]. CALLR improves Garnett by proposing a semi-supervised model for classifying the cells [22]. The performance of these two methods is greatly dependent on the representative cells selected using TD-IDF, which does not work stably for scRNA-seq data from different platforms. SCSA calculates cell type scores of each cluster by adding up re-scaled log2-based fold change values of differentially expressed marker genes. Clusters are then annotated as the cell type with the highest cell type score [25]. scCATCH first obtains the meta information of cell clusters, then by paired comparison of the groups, the potential marker genes for each cluster are identified. The cell types are determined by matching them with the validated marker genes [17]. Similar to scCATCH, deCS [20] annotates the cells using Fisher's exact test to choose the maximum overlap between the differentially expressed genes found in different clusters and the marker genes, though it can also annotate the cells with the annotated reference scRNA-seq dataset. SCSA, scCATCH, and deCS all assume the clusters are well defined, which may not be the truth in real data analysis. Current clustering methods are still far from sufficient for accurate annotation.

In this work, to overcome the problems existing in the marker gene based annotation methods, we propose an Automatic Cell type Annotation Method (ACAM) based on marker genes' information with no annotated cells needed. This method first finds the representative clusters by searching for the consistent subgroups across the results of several popular clustering methods, such as the method in Seurat [26], SC3 [27], CIDR [28], t-SNE+k-means [29], and SIMLR [30]. Such a technique guarantees that the cells in the same cluster have very high probabilities of being from the same cell type. Then, by selecting the features that discriminate one cluster from all the remaining cells, the potential marker genes are identified. The cell types are determined by defining a cell type importance score to match these potential marker genes with the validated ones. For those cells that do not belong to any of these clusters, we use k-nearest neighbors to determine their cell type. We did experiments on seven real-world datasets, and compared the results with six well-known methods. Results show the better performance of ACAM. ACAM fits well with our intuition for cell type annotation, takes advantage of the properties of scRNA-seq data, and is easily implementable.

2. Materials and Methods

2.1. Datasets

Seven real-world datasets were selected for comparison and testing. Information of the datasets is given in Table 1. All the cells in these datasets have known annotated labels. These datasets were chosen from various platforms. Dataset Chen [31] and Xin [32] were generated using Fluidigm C1 system. Dataset Kidney, Mammary [33], and PBMC [26] were generated using 10× Genomics. Other datasets were chosen from platforms, such as SeqWell and DropSeq [34,35]. The selected datasets have various magnitudes, ranging from 203 cells to 20679 cells. Several tissues from both human and mouse were selected to demonstrate the overall performance of the methods.

Table 1. Summary of datasets.

Dataset	Platform	Samples	Cell Types	Species	Tissue
Chen [31]	Fluidigm C1 system	203	3	Mouse	Kidney
Xin [32]	Fluidigm C1 system	1600	4	Human	Pancreatic Islet
Gierahn [34]	Seqwell	3694	5	Human	Peripheral Blood
Wu [35]	DropSeq	20,679	7	Mouse	Brain
PBMC [26]	10×	2638	4	Human	Peripheral Blood
Kidney [33]	10×	2781	8	Mouse	Kidney
Mammary [33]	10×	4481	7	Mouse	Mammary Gland

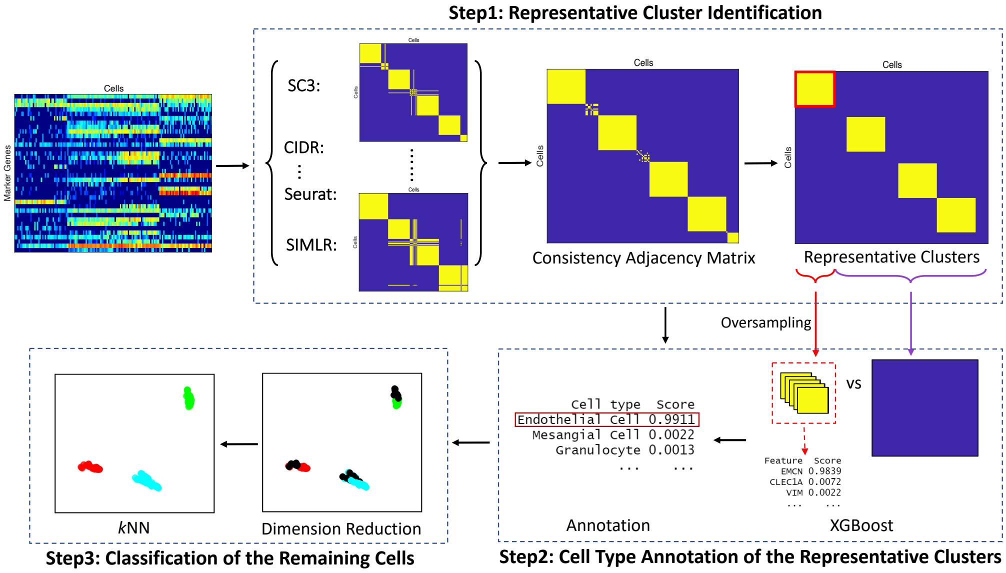
According to the cell type annotation method Garnett [16], consensus cell types were merged together. To be specific, 'AT1 cells', 'AT2 cells', and 'alveolar bipotent progenitors' were merged into 'alveolars'. 'Ciliated cells', 'clara cells', and 'dividing cells' were merged into 'ciliated cells'. 'Stromal cells' and 'fibroblasts' were merged into 'fibroblasts'. 'Neutrophils', 'eosinophils', 'basophils', and 'granulocytes' were merged into 'granulocytes'. 'Nuocytes' and 'T cells' were merged into 'T cells'. 'Dentritic cells', 'monocyte progenitor cells', 'monocytes', and 'macrophages' were merged into 'monocytes'. Consensus cell types 'Cajal-Retzius cells' and 'GABAergic cells' were merged into 'neurons' in dataset Wu [36].

In our study, we use the marker gene database CellMatch [17], which is derived from several popular database, such as CellMarker [37], MCA [38], CancerSEA [39], and the CD Marker Handbook [40]. The corresponding species and tissue of the dataset are selected in the subjects 'speciesType' and 'tissueType', and the 'Single-cell sequencing' entry is chosen in 'markerResource'. Then, cell types and their markers are chosen from the subjects 'cellMarker' and 'shortname', respectively. Markers for each cell type are then collected as input of the proposed method.

2.2. Methods

In this subsection, we present the proposed automatic cell type annotation method ACAM. The workflow of ACAM is shown in Figure 1.

Let $\tilde{X}_{p\times n}$ be the scRNA-seq gene expression matrix with p genes and n cells, which is firstly log-normalized after size factor adjustment for read depth [16]. We denote $\mathcal{U}$ as the cell set with $|\mathcal{U}| = n$. Let $\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_T\}$ be the list of marker genes for the considered species and tissue retrieved from the known database, where $\mathcal{G}_t$ denotes the list of markers for cell type t. We keep the related marker genes' expression in $\tilde{X}$ only, and remove those with zero expression across all the cells. Cells with zero expression across all marker genes are annotated as 'unknown', and are removed directly. Without confusion, we still use $\mathcal{U}$ and n to denote the remaining cell set and the remaining number of cells. The resulted data matrix is denoted as X, which is of size $M \times n$, where M is the number of marker genes for all considered cell types.


Step1: Representative Cluster Identification
Cells
Marker Genes
SC3:
CIDR:
⋮
Seurat:
SIMLR:
Cells
Cells
Consistency Adjacency Matrix
Representative Clusters
Oversampling
vs
Annotation
XGBoost
kNN
Dimension Reduction
Step3: Classification of the Remaining Cells
Step2: Cell Type Annotation of the Representative Clusters

```
       Cell type  Score
Endothelial Cell 0.9911
  Mesangial Cell 0.0022
     Granulocyte 0.0013
             ...    ...
```

```
Feature  Score
   EMCN 0.9839
 CLEC1A 0.0072
    VIM 0.0022
    ...    ...
```

Figure 1. The workflow of ACAM. The input is a log-normalized expression matrix for marker genes only. Markers are selected from the database CellMatch. Step 1. Apply different clustering methods, such as SC3 [27], CIDR [28], t-SNE+k-means [29], and SIMLR [30] to conduct clustering independently, and define the consistency adjacency matrix. Louvain algorithm is applied to identify the representative clusters. Step 2. Apply XGBoost to each representative cluster versus all the remaining cells to obtain each feature's importance score. Clusters are annotated by the maximum cell type score, which is defined as the sum of the importance score for all the features in each cell type. Step 3. Classify the remaining cells using k-nearest neighbors (kNN) after dimension reduction.

2.2.1. Representative Cluster Identification

Annotation accuracy usually heavily depends on clustering results. Each existing clustering method is insufficient for accurate annotation. Thus, to guarantee that cells from the same cluster are of the same cell type with high probability, we implement several state-of-the-art clustering methods independently, and the consensus subgroups are identified as the representative clusters. We note that any clustering method can be chosen here.

In this work, we choose five clustering methods, which include SC3 [27], CIDR [28], Seurat [41], t-SNE [29]+k-means, and SIMLR [30] according to [42]. After applying these methods, we obtain five different partitions of the cells: $\mathcal{C}_i = \{\mathcal{C}_{i1}, \ldots, \mathcal{C}_{ik_i}\}, i = 1, 2, \ldots, 5$ corresponding to the five clustering methods. $\mathcal{C}_{il}$ denotes the l-th cluster for the i-th clustering method, and k_i is the corresponding number of clusters. A brief description of five clustering methods is put in Appendix A. We then choose four of the five clustering results having the largest difference according to the variations in the pairwise Adjusted Rand Index (ARI) between any two different clustering methods [43]. To be specific, a 5 × 5 ARI matrix R is constructed by calculating

$$R(i,j) = \mathrm{ARI}(\mathcal{C}_i, \mathcal{C}_j).$$

For each row of R, we calculate the variance, and remove the clustering of the minimum variance. Without confusion, we use 1, 2, 3, and 4 to denote the four remaining methods.

To figure out the consistent clusters of the four methods, we construct graphs corresponding to the clustering results, and apply community detection methods. Let A_i ($i = 1, 2, 3, 4$) be the adjacency matrix of the graph corresponding to the results of clustering method i, where

$$A_i(u,v) = \begin{cases} 1, & \text{cell } u, v \text{ from the same cluster,} \\ 0, & \text{otherwise.} \end{cases}$$

The information from the four clustering methods is combined by adding A_i's up and the consistency adjacency matrix A^{con} is defined as follows:

$$\tilde{A} = \sum_{i=1}^{4} A_i, \qquad A^{con}(u,v) = \begin{cases} 1, & \tilde{A}(u,v) = 4, \\ 0, & \text{otherwise.} \end{cases}$$

We apply Louvain algorithm [44] to A^{con} to identify the communities, which are taken as the consistent clusters. Clusters with size larger than a threshold are finally set as the representative clusters, which are denoted as $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_C\}$, where C is the number of representative clusters. In our experiments, we set the threshold to be 10.

2.2.2. Cell Type Annotation of the Representative Clusters

In our setting, no annotated cells are given, thus supervised learning methods cannot be directly applied to label the unannotated cells. We assign each representative cluster a temporary label, and apply supervised learning methods to extract the features that discriminate it from the remaining cells. Then we match the extracted features to the known cell type associated marker genes, and assign the most probable cell type to the cluster.

Since marker genes of one particular cell type are more likely to be highly expressed in the cells of the type, while comparatively merely expressed in other cell types, extreme gradient boosting (XGBoost) [45] is a good choice for extracting the important features. XGBoost is known as a fast, flexible, and efficient gradient boosting tree skilled in tackling highly sparse data, and performs very well in many classification problems. It constructs the tree by splitting the features into two nodes according to each feature's value. This fits well with the property of marker genes. Here, we apply XGBoost to extract the features that discriminate each representative cluster $\mathcal{P}_c$ (target group) from all the remaining cells $\mathcal{U} - \mathcal{P}_c$ (adversarial group), namely the whole set with the subset $\mathcal{P}_c$ removed. According to the property of marker genes, to make sure the features having high feature importance score are the marker genes of $\mathcal{P}_c$, each gene's mean expression level is compared between the target group and the adversarial group, and those having lower mean expression values in the target group are removed before putting into the XGBoost model. Since normally the size of $\mathcal{P}_c$ is much smaller than that of $\mathcal{U} - \mathcal{P}_c$, to balance the size of the two groups, we adopt the oversampling technique to make the target group have a similar size to the adversarial group. Specifically, we randomly select the cells belonging to $\mathcal{P}_c$, until the size of the target group is the same as that of the adversarial group. Hinge loss is chosen as the objective and the tree depth is set to 1 in the XGBoost model. We implement XGBoost from R package `xgboost`. Feature importance w_m of each marker gene m can be obtained after running XGBoost. Then the feature importance for each cell type t is calculated by

$$Score_t = \sum_{m \in \mathcal{G}_t} w_m,$$

The representative cluster is annotated as the cell type t_0, where t_0 is the cell type that maximizes $Score_t$ for all t.

2.2.3. Classification of the Remaining Cells

We apply k-nearest neighbors (kNN) to annotate the cells that do not belong to any representative cluster. Before doing kNN, we apply uniform manifold approximation and projection (UMAP) [46], which efficiently conducts dimension reduction and preserves the high dimensional structure, to project the cells into two-dimensional space for the following classification and visualization. kNN is then applied to assign the remaining cells to the annotated representative clusters. We simply set k to be 1.

We put the overall procedure in Algorithm 1.

Algorithm 1 ACAM: Automatic Cell type Annotation Method.

```
Input: The pre-processed data matrix $X$, marker gene set $\mathcal{G}$
Initialize $thresh = 10, k = 1$
$\mathcal{C}_1 \leftarrow SC3(X); \mathcal{C}_2 \leftarrow CIDR(X); \mathcal{C}_3 \leftarrow Seurat(X);$
  $\mathcal{C}_4 \leftarrow t\text{-}SNE + k\text{-}means(X); \mathcal{C}_5 \leftarrow SIMLR(X)$
$R_{i,j} \leftarrow ARI(\mathcal{C}_i, \mathcal{C}_j), \quad i = 1, \dots, 5$
Remove the method of $\arg\min_i var(R_{i,\cdot})$
$A_i$: $A_i(u,v) \leftarrow \begin{cases} 1, & u,v \text{ from the same cluster,} \\ 0, & \text{otherwise.} \end{cases}$
$\tilde{A} \leftarrow \sum_{i=1}^{4} A_i$
$A^{con}$: $A^{con}(u,v) \leftarrow \begin{cases} 1, & \tilde{A}(u,v) = 4, \\ 0, & \text{else.} \end{cases}$
$\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_C\} \leftarrow Louvain(A^{con}, thresh)$
for c = 1,...,C do
  $\tilde{\mathcal{P}}_c \leftarrow Oversample(\mathcal{P}_c)$
  Select $m$ with $mean(X[m, \tilde{\mathcal{P}}_c]) > mean(X[m, \mathcal{U} - \mathcal{P}_c])$
  $w_m \leftarrow XGBoost(\tilde{\mathcal{P}}_c, \mathcal{U} - \mathcal{P}_c), m \in \mathcal{G}_t, t = 1, \dots T$
  $Score_t = \sum_{m \in \mathcal{G}_t} w_m, \quad t = 1, \dots, T$
  $y_u \leftarrow$ cell type $t_0$: $t_0 = \arg\max_t(Score_t), \quad u \in \mathcal{P}_c$
end for
$y_u \leftarrow kNN(X[, \mathcal{P}], k), \quad u \in \mathcal{U} - \mathcal{P}$
Output: Cell labels $y$
```

2.3. Results Evaluation Metrics

We choose four metrics: accuracy, balanced accuracy, macro F1-score, and Matthews correlation coefficient (MCC) to measure the performance. Let the total number of cell types be T and the total number of cells be n. Let TP_t, FP_t, and FN_t denote the true positive, false positive, and false negative for the cell type t in the confusion matrix constructed for the underlying true labels and the inferred labels. Row_t and Col_t denote the t-th row and column of the confusion matrix.

- Accuracy: It is defined as the percentage of true positives of the annotations:

$$Accuracy = \frac{\sum_{t=1}^{T} TP_t}{n}.$$

- Balanced Accuracy: It is defined as the average Recall of each cell type,

$$Balanced\ Accuracy = \frac{\sum_{t=1}^{T} Recall_t}{T},$$

where

$$Recall_t = \frac{TP_t}{TP_t + FN_t}.$$

- Macro F1-Score: It is defined as the harmonic mean of average Precision and average Recall:

$$Macro\ F1\text{-}Score = 2 \times \frac{Average\ Precision \times Average\ Recall}{Average\ Precision + Average\ Recall}$$

where

$$Precision_t = \frac{TP_t}{TP_t + FP_t}.$$

- Matthews Correlation Coefficient (MCC): It takes into account the true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes:

$$MCC = \frac{\sum_{t=1}^{T} TP_t \times n - \sum_{t=1}^{T} Col_t \times Row_t}{\sqrt{(n^2 - \sum_{t=1}^{T} Col_t^2)(n^2 - \sum_{t=1}^{T} Row_t^2)}}.$$

The detailed definitions of these metrics can be found in [47].

3. Results

We evaluated ACAM using seven real-world datasets, and compared with six well-known cell type annotation methods, especially the marker-gene-based methods.

3.1. Comparison Methods

ACAM was compared with four marker-gene-based methods: CellAssign [19], deCS [20], Garnett [16], SCSA (SCSA_Scran [48], and SCSA_Seurat [25,26]). To give a more general picture of the annotation methods, we also added two well-known reference scRNA-seq-data-based methods: SeuratTransfer [21] and SingleR [9] into comparisons.

- CellAssign It takes into account the prior knowledge of marker genes into a probabilistic model to estimate cell types with parameters selected by the maximum a posteriori probability, and google tensorflow is used in EM step.
- deCS It first conducts clustering by Seurat [26]. Differentially expressed genes of clusters are then extracted using function `FindAllMarkers` in R package `Seurat`. It then annotates clusters as the cell type with the maximum overlap between cell type markers and the differentially expressed genes.
- Garnett It first chooses representative cells by aggregating marker scores from the TF-IDF matrix, and then trains the logistic regression model with elastic net to classify the remaining cells, regarding the representative cells as training set.
- SCSA Similar to deCS, SCSA first conducts clustering by Seurat [26]. Differentially expressed genes of clusters are extracted using the function `FindAllMarkers` in R package `Seurat` (SCSA_Seurat) and the function `findMarkers` in R package `Scran` (SCSA_Scran). SCSA calculates cell type scores of each cluster by adding up re-scaled log2-based fold change values (LFC) of differentially expressed marker genes. Clusters are then annotated by the cell type with the highest cell type score.
- SeuratTransfer It uses the function `TransferData` in the R package `Seurat`. It is a strategy to 'anchor' datasets together. By placing both the annotated reference scRNA-seq dataset and the unannotated dataset in a shared low-dimensional space using canonical correlation analysis (CCA), pairwise correspondences between cells from both datasets are identified as anchors by mutual nearest neighbors (MNN). For each cell in the unannotated dataset, it is scored and annotated depending on the distances to anchors.
- SingleR It first calculates the Spearman coefficients on variable genes between each unannotated cell and the annotated ones of each type in the reference scRNA-seq data. The same procedure is iteratively performed using the cell types with top correlations in the previous step. The cell is annotated as the type that is left till the last round.

3.2. Methods' Implementation Details

In the representative clustering identification step of ACAM, we implemented the five clustering methods using their corresponding R package. We accelerated the clustering procedure for the datasets of sample size larger than 4000. In SC3 and SIMLR method, the number of cluster was calculated by the function `sc3_estimate_k` in R package `SC3` and function `SIMLR_Estimate_Number_of_Clusters` in R package `SIMLR`, respectively. The threshold of the size of representative clusters was normally set to 10, apart from the dataset

Chen, which was set to 5 independently due to its small sample size. This parameter can be set manually according to the prior knowledge.

For the marker-gene-based methods: Garnett, CellAssign, deCS and SCSA, CellMatch is selected as the input marker database. For the annotated reference scRNA-seq-data-based methods: SingleR and SeuratTransfer, we constructed the reference dataset by setting the expression of each marker gene in CellMatch database to be 1 in its cell type, and 0 otherwise to fairly compare the annotation capability. The parameters for all the methods were set to default. In the method SeuratTransfer, we changed *k.weight* ranging from 10 to the maximum, and the one with the highest accuracy was chosen in our comparisons. Note that all the input expression matrices were in the log-normalized form. For the dataset Wu, due to its time and memory cost in the five clustering procedures and the tensorflow procedure in the method CellAssign, we randomly split it into five subsets with equal size to complete the annotation independently, and then summarized the results.

3.3. Results

Figure 2 shows the cell type annotation results for the compared methods using four measures. ACAM performed stably in all datasets. Most scores of ACAM were ranked first and second. CellAssign did not perform as good as the other methods in our comparison. It failed to find more than 10 true labels in dataset Chen, Xin, Gierahn, and PBMC. Garnett had good performance only in the dataset PBMC and Gierahn, both of which are datasets of human peripheral blood. deCS and SCSA did not have stable annotation scores. deCS reached the best and second best scores in dataset Kidney and Wu, and two SCSA methods reached the best accuracy in dataset Chen. However, deCS obtained the top three worst performance in dataset Chen, Gierahn, and Mammary, which shows the instability of deCS. SingleR also reached high accuracy in all datasets. Though not as good as ACAM, two annotated reference-data-based methods, SeuratTransfer and SingleR reached a stable accuracy in all datasets.

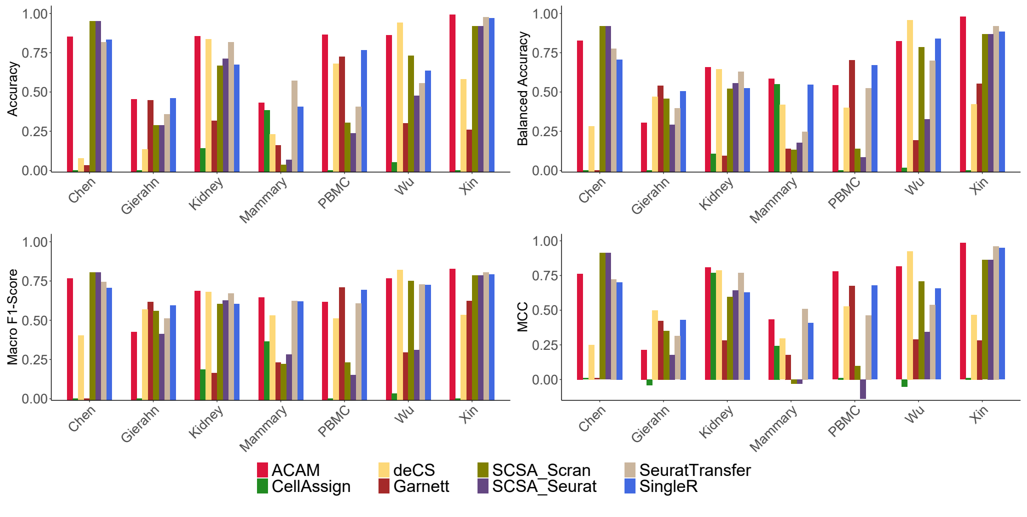


Figure 2. Annotation results comparison. Results of the compared methods using four evaluation metrics: accuracy, balanced accuracy, macro F1-score, and MCC on seven real-world datasets are shown.

To give an overall evaluation of all the seven methods, we ranked the methods according to the four metrics in each dataset ranging from one to eight. There are a total of 28 ranks for each method. Lower rank represents better performance (one is the best and eight is the worst). Figure 3 shows the boxplot of all methods. The overall ranks of ACAM are much lower than the other methods, especially the marker-gene-based methods.

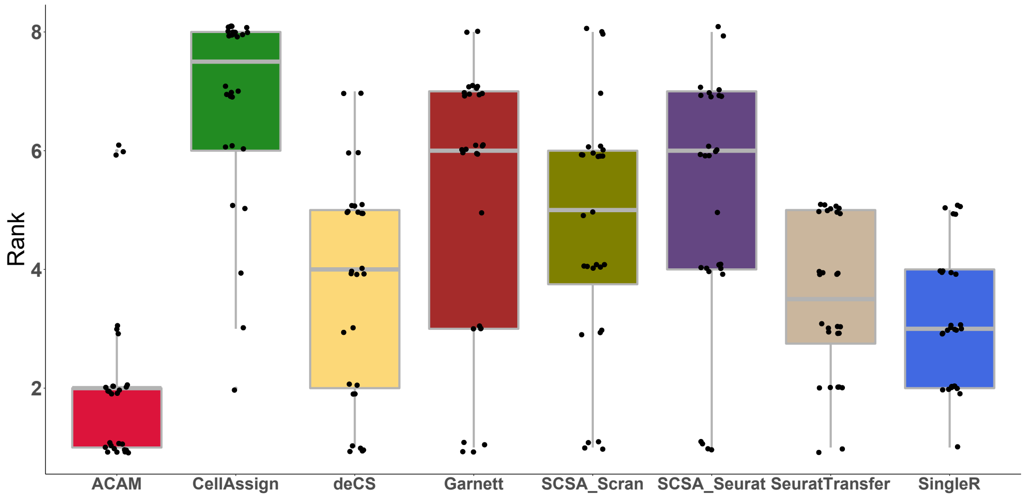


Figure 3. Rank of the compared methods. Boxplot of the rank of each method according to four evaluation metrics: accuracy, balanced accuracy, macro F1-score, and MCC on seven datasets is shown. A lower rank represents better performance (one is the best and eight is the worst).

To take a deeper look into the performance of all the compared methods, visualization of three datasets: Kidney, Mammary, and Wu is shown in Figures 4–6. ACAM managed to annotate the main clusters stably and correctly, while other marker-gene-based methods did not. CellAssign, which involves a more complicated model and uses iteration technique, failed to tell difference across some cell types of large size. It wrongly annotated most cells into the cell type ‘thick ascending limb of the loop of Henles’ in dataset Kidney (Figure 4) and ‘Martinotti cells’ and ‘neurons’ in dataset Wu (Figure 6). Garnett was able to annotate only a small part of cell types correctly. Most cells were left unassigned in all three datasets (Figures 4–6). This should be due to the TD-IDF scoring system used for constructing the training set for the following supervised learning. It may not work well for datasets from various platforms and tissues. deCS and SCSA annotated cells based on the clustering results. deCS wrongly annotated part basel cells of the dataset Mammary (Figure 5). Two SCSA methods failed to correctly annotate lots of cells in dataset Mammary and Wu (Figures 5 and 6). This should be due to the clustering results, which will strongly affect the accuracy of annotations. SeuratTransfer did not perform as stable as SingleR and ACAM. It performed well in dataset Kidney and Mammary, but for dataset Wu, it failed to discriminate the combination of subgroups, and hardly annotated cells in the bottom right corner in Figure 6. In addition, some of the microglial cells was wrongly annotated as monocytes. Though SingleR reached high accuracy in most datasets, it did not perform as well as ACAM. As shown in Figure 4, ACAM correctly labeled ‘thick ascending limb of the loop of Henles’ in dataset Kidney, SingleR, however, failed to annotated them correctly. The same happened for ‘basel cells’ in dataset Mammary and ‘neurons’ in dataset Wu, as shown in Figures 5 and 6, respectively.

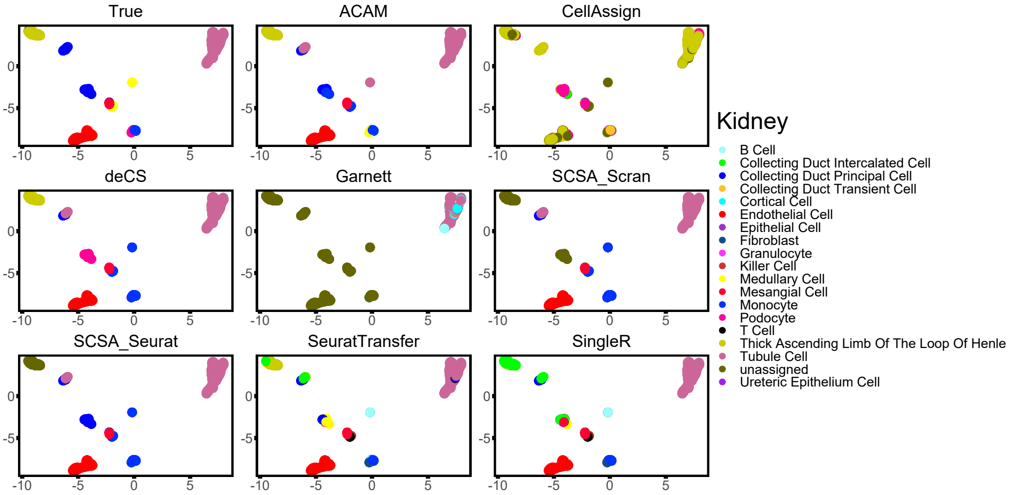


Figure 4. Two-dimensional visualization of the annotation results for dataset Kidney using UMAP.

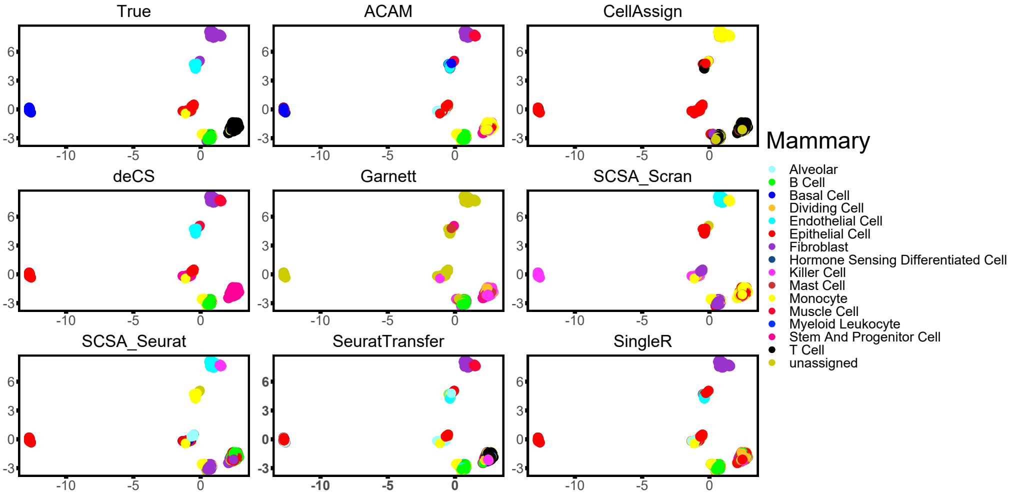


Figure 5. Two-dimensional visualization of the annotation results for dataset Mammary using UMAP.

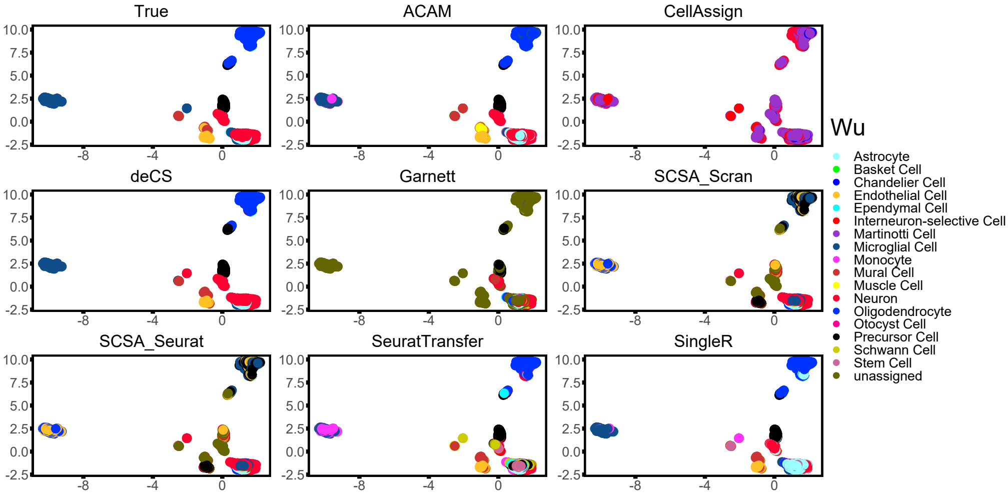


Figure 6. Two-dimensional visualization of the annotation results for dataset Wu using UMAP.

4. Conclusions and Discussion

In this study, we present an automatic marker-gene-based cell type annotation method: ACAM. It has a clear framework composed of three steps. First, trustworthy representative clusters are identified. Then, a marker-gene-based annotation strategy is designed to perform cell type assignment according to the importance score of the marker genes that discriminate one specific representative cluster from the remaining cells. After all the representative clusters are labeled, *k*NN is applied to annotate the remaining cells outside the representative clusters.

The comparison of ACAM with other methods including both marker-gene-based methods and reference scRNA-seq-data-based methods shows the superiority of ACAM in cell type annotation. ACAM performed better in datasets with different attributes, such as various sample sizes, different species and tissues, and several data generation platforms. The better performance of ACAM against the marker-gene-based methods that conduct annotation based on clustering results indicates that clustering is still an important problem for accurate cell type annotation. In our current setting, though the consistent clusters across several clustering results in ACAM give better annotation, it is at the cost of more computational time. The better performance of ACAM over the marker-gene-based methods that annotate each single cell individually suggests that cluster-level information is more stable, especially when there exists severe noise in the data.

In our current study, each cell is assigned to a known cell type of size greater than a given threshold (10 as default), which may mis-classify the rare cells, and the cells of unknown cell type. How to define the unknown cell types and find the rare cell types according to the marker genes' expression is still worth further exploration.

Author Contributions: Conceptualization, S.Z.; methodology, Y.C. and S.Z.; numerical experiments, Y.C.; writing, Y.C. and S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: S. Zhang's research is supported in part by Science and Technology Commission of Shanghai Municipality No. 20ZR1407700.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: ACAM is available as an R package at https://github.com/yuc0824/ACAM (accessed on 30 September 2022). No new data were generated for this study. All data used in this study are publicly available. Datasets below are accessible in GEO with the following accession codes: (a) Chen dataset, GSE99701. (b) Xin dataset, GSE81608. (c) Gierahn dataset, GSM2486333. (d) Wu dataset, GSE103976. (e) Kidney and Mammary dataset, GSE109774. For PBMC dataset, data is available at https://www.10xgenomics.com/ (accessed on 19 October 2021).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

We give a brief description of the five clustering methods used in ACAM.

SC3 [27]: Distance between the cells are calculated using the Euclidean, Pearson, and Spearman metrics to construct distance matrices. Dimensions of three distance matrices are then reduced using either principal component analysis (PCA) or associated graph Laplacian. It does *k*-means clustering on these six metrics. A consensus matrix is then calculated using the Cluster-based Similarity Partitioning Algorithm (CSPA) on these six matrices and it is clustered using hierarchical clustering with complete agglomeration.

CIDR [28]: It imputes dropout candidates, constructs the dissimilarity matrix and maps it into a low-dimensional space by the principal coordinate analysis (PCoA) method. It finally clusters cells by the hierarchical clustering.

Seurat [41]: It identifies clusters of cells by a shared nearest neighbor (SNN) modularity optimization based clustering algorithm. It first calculate *k*-nearest neighbors and construct the SNN graph. It then optimize the modularity function to determine clusters.

t-SNE [29]+*k*-means: It conducts t-SNE to reduce the dimension and then conducts clustering by *k*-means method.

SIMLR [30]: It conducts the multi-kernel learning framework to obtain a sparse similarity matrix. A spectral clustering algorithm is then applied, which is very effective for clustering sparse similarities and scaling cells.

References

1. Kolodziejczyk, A.A.; Kim, J.K.; Svensson, V.; Marioni, J.C.; Teichmann, S.A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **2015**, *58*, 610–620. [CrossRef] [PubMed]
2. Friebel, E.; Kapolou, K.; Unger, S.; Núñez, N.G.; Utz, S.; Rushing, E.J.; Regli, L.; Weller, M.; Greter, M.; Tugues, S.; et al. Single-cell mapping of human brain cancer reveals tumor-specific instruction of tissue-invading leukocytes. *Cell* **2020**, *181*, 1626–1642. [CrossRef]
3. Tirosh, I.; Izar, B.; Prakadan, S.M.; Wadsworth, M.H.; Treacy, D.; Trombetta, J.J.; Rotem, A.; Rodman, C.; Lian, C.; Murphy, G.; et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **2016**, *352*, 189–196. [CrossRef] [PubMed]
4. Wagner, J.; Rapsomaniki, M.A.; Chevrier, S.; Anzeneder, T.; Langwieder, C.; Dykgers, A.; Rees, M.; Ramaswamy, A.; Muenst, S.; Soysal, S.D.; et al. A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell* **2019**, *177*, 1330–1345. [CrossRef] [PubMed]
5. Zheng, H.; Pomyen, Y.; Hernandez, M.O.; Li, C.; Livak, F.; Tang, W.; Dang, H.; Greten, T.F.; Davis, J.L.; Zhao, Y.; et al. Single-cell analysis reveals cancer stem cell heterogeneity in hepatocellular carcinoma. *Hepatology* **2018**, *68*, 127–140. [CrossRef] [PubMed]
6. Li, L.; Guo, F.; Gao, Y.; Ren, Y.; Yuan, P.; Yan, L.; Li, R.; Lian, Y.; Li, J.; Hu, B.; et al. Single-cell multi-omics sequencing of human early embryos. *Nat. Cell Biol.* **2018**, *20*, 847–858. [CrossRef]
7. Wagner, D.E.; Weinreb, C.; Collins, Z.M.; Briggs, J.A.; Megason, S.G.; Klein, A.M. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **2018**, *360*, 981–987. [CrossRef]
8. Alquicira-Hernandez, J.; Sathe, A.; Ji, H.P.; Nguyen, Q.; Powell, J.E. scPred: Accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol.* **2019**, *20*, 1–17. [CrossRef]
9. Aran, D.; Looney, A.P.; Liu, L.; Wu, E.; Fong, V.; Hsu, A.; Chak, S.; Naikawadi, R.P.; Wolters, P.J.; Abate, A.R.; et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **2019**, *20*, 163–172. [CrossRef]
10. Brbić, M.; Zitnik, M.; Wang, S.; Pisco, A.O.; Altman, R.B.; Darmanis, S.; Leskovec, J. MARS: Discovering novel cell types across heterogeneous single-cell experiments. *Nat. Methods* **2020**, *12*, 1200–1206. [CrossRef]
11. Hou, R.; Denisenko, E.; Forrest, A.R. scMatch: A single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics* **2019**, *35*, 4688–4695. [CrossRef] [PubMed]
12. Hu, J.; Li, X.; Hu, G.; Lyu, Y.; Susztak, K.; Li, M. Iterative transfer learning with neural network for clustering and cell type classification in single-cell RNA-seq analysis. *Nat. Mach. Intell.* **2020**, *2*, 607–618. [CrossRef] [PubMed]
13. de Kanter, J.K.; Lijnzaad, P.; Candelli, T.; Margaritis, T.; Holstege, F.C. CHETAH: A selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res.* **2019**, *47*, e95. [CrossRef] [PubMed]
14. Kiselev, V.Y.; Yiu, A.; Hemberg, M. scmap: Projection of single-cell RNA-seq data across data sets. *Nat. Methods* **2018**, *15*, 359–362. [CrossRef] [PubMed]
15. Pasquini, G.; Arias, J.E.R.; Schäfer, P.; Busskamp, V. Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 961–969. [CrossRef]
16. Pliner, H.A.; Shendure, J.; Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* **2019**, *16*, 983–986. [CrossRef]
17. Shao, X.; Liao, J.; Lu, X.; Xue, R.; Ai, N.; Fan, X. scCATCH: Automatic annotation on cell types of clusters from single-cell RNA sequencing data. *Iscience* **2020**, *23*, 100882. [CrossRef]
18. Shao, X.; Yang, H.; Zhuang, X.; Liao, J.; Yang, P.; Cheng, J.; Lu, X.; Chen, H.; Fan, X. scDeepSort: A pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. *Nucleic Acids Res.* **2021**. [CrossRef]
19. Zhang, A.W.; Oflanagan, C.H.; Chavez, E.A.; Lim, J.L.P.; Ceglia, N.; Mcpherson, A.; Wiens, M.; Walters, P.; Chan, T.M.; Hewitson, B.; et al. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat. Methods* **2019**, *16*, 1007–1015. [CrossRef]
20. Pei, G.; Yan, F.; Simon, L.M.; Dai, Y.; Jia, P.; Zhao, Z. deCS: A tool for systematic cell type annotations of single-cell RNA sequencing data among human tissues. *Genom. Proteom. Bioinform.* **2022**, *22*. [CrossRef]
21. Stuart, T.; Butler, A.; Hoffman, P.; Hafemeister, C.; Papalexi, E.; Mauck III, W.M.; Hao, Y.; Stoeckius, M.; Smibert, P.; Satija, R. Comprehensive integration of single-cell data. *Cell* **2019**, *177*, 1888–1902. [CrossRef]

22. Wei, Z.; Zhang, S. CALLR: A semi-supervised cell-type annotation method for single-cell RNA sequencing data. *Bioinformatics* **2021**, *37*, i51–i58. [CrossRef]
23. DePasquale, E.A.; Schnell, D.; Dexheimer, P.; Ferchen, K.; Hay, S.; Chetal, K.; Valiente-Alandí, Í.; Blaxall, B.C.; Grimes, H.L.; Salomonis, N. cellHarmony: Cell-level matching and holistic comparison of single-cell transcriptomes. *Nucleic Acids Res.* **2019**, *47*, e138. [CrossRef]
24. Seal, D.B.; Das, V.; De, R.K. CASSL: A cell-type annotation method for single cell transcriptomics data using semi-supervised learning. *Appl. Intell.* **2022**. [CrossRef]
25. Cao, Y.; Wang, X.; Peng, G. SCSA: A cell type annotation tool for single-cell RNA-seq data. *Front. Genet.* **2020**, *11*, 490. [CrossRef]
26. Butler, A.; Hoffman, P.J.; Smibert, P.; Papalexi, E.; Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **2018**, *36*, 411–420. [CrossRef]
27. Kiselev, V.Y.; Kirschner, K.; Schaub, M.T.; Andrews, T.S.; Yiu, A.; Chandra, T.; Natarajan, K.N.; Reik, W.; Barahona, M.; Green, A.R.; et al. SC3: Consensus clustering of single-cell RNA-seq data. *Nat. Methods* **2017**, *14*, 483–486. [CrossRef]
28. Lin, P.; Troup, M.; Ho, J.W.K. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* **2017**, *18*, 59. [CrossRef]
29. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
30. Wang, B.; Zhu, J.; Pierson, E.; Ramazzotti, D.; Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* **2017**, *14*, 414–416. [CrossRef]
31. Chen, L.; Lee, J.W.; Chou, C.L.; Nair, A.V.; Battistone, M.A.; Păunescu, T.G.; Merkulova, M.; Breton, S.; Verlander, J.W.; Wall, S.M.; et al. Transcriptomes of major renal collecting duct cell types in mouse identified by single-cell RNA-seq. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E9989–E9998. [CrossRef]
32. Xin, Y.; Kim, J.; Okamoto, H.; Ni, M.; Wei, Y.; Adler, C.; Murphy, A.J.; Yancopoulos, G.D.; Lin, C.; Gromada, J. RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab.* **2016**, *24*, 608–615. [CrossRef]
33. Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **2018**, *562*, 367–372. [CrossRef]
34. Gierahn, T.M.; Wadsworth, M.H.; Hughes, T.K.; Bryson, B.D.; Butler, A.; Satija, R.; Fortune, S.; Love, J.C.; Shalek, A.K. Erratum: Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **2017**, *14*, 752. [CrossRef]
35. Wu, Y.E.; Pan, L.; Zuo, Y.; Li, X.; Hong, W. Detecting activated cell populations using single-cell RNA-seq. *Neuron* **2017**, *96*, 313–329. [CrossRef] [PubMed]
36. Zeisel, A.; Hochgerner, H.; Lönnerberg, P.; Johnsson, A.; Memic, F.; Van Der Zwan, J.; Häring, M.; Braun, E.; Borm, L.E.; La Manno, G.; et al. Molecular architecture of the mouse nervous system. *Cell* **2018**, *174*, 999–1014. [CrossRef] [PubMed]
37. Zhang, X.; Lan, Y.; Xu, J.; Quan, F.; Zhao, E.; Deng, C.; Luo, T.; Xu, L.; Liao, G.; Yan, M.; et al. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* **2019**, *47*, D721–D728. [CrossRef] [PubMed]
38. Han, X.; Wang, R.; Zhou, Y.; Fei, L.; Sun, H.; Lai, S.; Saadatpour, A.; Zhou, Z.; Chen, H.; Ye, F.; et al. Mapping the mouse cell atlas by microwell-seq. *Cell* **2018**, *172*, 1091–1107. [CrossRef]
39. Yuan, H.; Yan, M.; Zhang, G.; Liu, W.; Deng, C.; Liao, G.; Xu, L.; Luo, T.; Yan, H.; Long, Z.; et al. CancerSEA: A cancer single-cell state atlas. *Nucleic Acids Res.* **2019**, *47*, D900–D908. [CrossRef]
40. BD Biosciences. CD Marker Handbook. Available online: http://static.bdbiosciences.com/documents/cd_marker_handbook.pdf (accessed on 15 August 2022).
41. Hao, Y.; Hao, S.; Andersen-Nissen, E.; Mauck, W.M., III; Zheng, S.; Butler, A.; Lee, M.J.; Wilk, A.J.; Darby, C.; Zager, M.; et al. Integrated analysis of multimodal single-cell data. *Cell* **2021**, *184*, 3573–3587. [CrossRef]
42. Huh, R.; Yang, Y.; Jiang, Y.; Shen, Y.; Li, Y. SAME-clustering: Single-cell Aggregated Clustering via Mixture Model Ensemble. *Nucleic Acids Res.* **2020**, *48*, 86–95. [CrossRef]
43. Hubert, L.; Arabie, P. Comparing partitions. *J. Classif.* **1985**, *2*, 193–218. [CrossRef]
44. Blondel, V.D.; Guillaume, J.L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [CrossRef]
45. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K. Xgboost: Extreme gradient boosting. *R Package Version 0.4-2* **2015**, *1*, 1–4.
46. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *Stat* **2020**, *1050*, 18.
47. Grandini, M.; Bagli, E.; Visani, G. Metrics for multi-class classification: An overview. *arXiv* **2020**, arXiv:2008.05756.
48. Lun, A.T.; McCarthy, D.J.; Marioni, J.C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **2016**, *5*, 2122. [CrossRef]