Toward highly accurate reference and method selection for universal cross-dataset cell type annotation

Qunlun Shen[1,2], Shuqin Zhang[1,3*] and Shihua Zhang[2,4,5*]

[1]School of Mathematical Sciences, Fudan University, Shanghai, 200433, China.
[2]NCMIS, CEMS, RCSDS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China;
[3]Center for Applied Mathematics, Research Institute of Intelligent Complex Systems, and Shanghai Key Laboratory for Contemporary Applied Mathematics, Fudan University, Shanghai 200433, China;
[4]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China;
[5]Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou 310024, China;

*To whom correspondence should be addressed. Shuqin Zhang, Tel: 86-21-65647484; Email: zhangs@fudan.edu.cn. Shihua Zhang, Tel: 86-10-82541360; Email: zsh@amss.ac.cn.

Abstract

Cell type annotation is a critical and essential task in single-cell data analysis. Various reference-based methods have provided rapid annotation for diverse single-cell data. However, how to select the optimal references and methods is often overlooked. To this end, we present a cross-dataset cell-type annotation methodology with a universal reference data and method selection strategy (CAMUS) to achieve highly accurate and efficient annotations. We demonstrate the advantages of CAMUS by conducting comprehensive analyses on 672 pairs of cross-species scRNA-seq datasets. The annotation results with references selected by CAMUS achieved substantial accuracy gains (25.0-124.7%) across five reference-based methods. CAMUS achieved high accuracy in choosing the best reference-method pair among 3360 pairs (49.1%). Combined with reference-integrated strategies, they improved further on 3522 dataset triplets (9.3-39.3%). Moreover, CAMUS showed high accuracy in selecting the best methods on the 80 scST datasets (82.5%) and five scATAC-seq datasets (100.0%), illustrating its universal applicability. In addition, we utilized the CAMUS score with other metrics to predict the annotation accuracy, providing direct guidance on whether to accept current annotation results.

Introduction

Single-cell technologies such as single-cell RNA sequencing (scRNA-seq)[1], single-cell spatial transcriptomics (scST)[2], and single-cell Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq)[3] have revolutionized our understanding of cellular diversity and complexity. These technologies enable researchers to investigate the transcriptomic, spatial, and epigenomic landscapes at single-cell resolution, offering unprecedented insights into understanding organismal development, physiological responses, and disease progression[2, 4, 5].

One critical task in single-cell data analysis is cell type annotation (or cell typing)[6, 7], allowing researchers to decipher cell heterogeneity and their functions within the tissue microenvironment. The prevalent manual annotation approach groups single cells into several clusters and annotates them based on unique molecular signatures[6]. This process needs prior biological knowledge and is time-consuming. However, as scRNA-seq can profile all genes (typically numbering in the tens of thousands), the clustering results from well-profiled scRNA-seq data are relatively distinct, resulting in high accuracy of cell type annotation. Recently, as extensive scRNA-seq datasets are generated and annotated, reference-based annotation methods have gained prominence[6]. These methods effectively transfer cell labels from the reference dataset to the query one by leveraging the well-annotated scRNA-seq datasets. Reference-based methods such as CAME[8], Seurat[9], SciBet[10], scmap[11], and SCN[12] have demonstrated rapid cell annotation capabilities without needing prior biological knowledge. These methods potentially enable the cell type identification that may be overlooked or difficult to ascertain through manual annotation and have been successfully extended to various scenarios, encompassing different conditions, technologies, and species[8, 9].

Cell type annotation for scST data presents even more challenges, as spatial technologies such as seqFISH[13], STARmap[14], and MERFISH[15] typically profile hundreds of genes only. This limitation leads to blurred cluster boundaries and difficulties in manual annotation[16]. Therefore, most scST studies utilize scRNA-seq datasets as a reference to help annotate scST datasets. For example, scANVI was employed to annotate MERFISH data derived from the kidney and liver[17, 18]; RCTD was used onto Slide-seq v2 data collected from the cerebellum and hippocampus[19]; Tangram was applied to STARmap and MERFISH data originated from the cortex[20]; and STAMapper was evaluated under scST datasets collected from eight scST technologies and five tissues[16].

Regarding scATAC-seq data, its extreme sparsity often constrains the performance of cell type annotation[21]. This limitation necessitates the development of more sophisticated reference-based computational methods to assist cell typing. For instance, GLUE[22], scJoint[21], and scDART[23] learn a

joint cell embedding from scRNA-seq and scATAC-seq data, which aids in annotating the scATAC-seq data.

Although we have well-designed reference-based cell type annotation methods for various omics datasets, no single method consistently achieves optimal performance across all omics types. Furthermore, for any given tissue or disease, numerous well-annotated scRNA-seq datasets are available as references. The choice of reference can significantly impact the annotation since single-cell assays may miss certain cell types, or these cell types may be underrepresented or wrongly annotated[24]. Consequently, reference and method selection is a critical consideration. Additionally, without a solid understanding of biological context, users could not assess the accuracy of algorithmic annotations. Estimating the accuracy of these annotations can also guide users directly in deciding whether to accept a particular annotation.

To this end, we present a cross-dataset annotation methodology with a universal reference data and method selection strategy (CAMUS) for reference-based single-cell annotation methods. CAMUS prioritizes the annotation performance of reference-based methods by comparing the concordance between the annotation results and the per-clustered labels. This strategy offers four valuable benefits. First, it directly compares the clustering information with the algorithmic annotations. Since clustering typically approximates the gold standard of cell typing, this comparison is biologically meaningful and ensures high accuracy. Second, CAMUS offers a framework for prioritizing annotation accuracy across various structured biological data without imposing constraints on the original data type. Third, this strategy utilizes the low-dimensional clustering information and annotation results rather than directly modeling the similarities in high-dimensional data, thereby enhancing the computational efficiency. Finally, the CAMUS score highly correlates with annotation accuracy, making it a powerful predictor of annotation performance. Extensive tests on diverse scenarios (e.g., cross-species, cross-modalities, and cross-omics) served as benchmarks to demonstrate its superiorities for reference and method selection.

Results

Overview of CAMUS.

CAMUS aims to prioritize reference-method pairs and select the optimal one, ensuring the highest accuracy for a reference-based single-cell annotation task. CAMUS utilizes the single-cell expression profiles from the query data and the annotation results of various reference-method pairs as inputs (**Fig. 1a**, **Methods**). It initially pre-clusters cells from query data into several distinct clusters (**Fig. 1a**, **Methods**) to approximate the cell type composition of the query data. It compares the concordance score between the pre-clustering and annotation results with various methods and references. A higher CAMUS score indicates a potential higher annotation accuracy (**Methods**). CAMUS

can be applied to three distinct cell type annotation scenarios: (1) Cross-species (including species from human, mouse, zebrafish, chick, lizard, turtle, and macaque); (2) Cross modalities (from scRNA-seq dataset to scST dataset); (3) Cross omics (from scRNA-seq dataset to scATAC-seq dataset). CAMUS effectively prioritizes reference-method pairs with rankings highly correlated with the actual annotation accuracy. Furthermore, CAMUS precisely estimates annotation accuracy by utilizing CAMUS score and additional cellular indices as predictors (**Methods**), offering users reliable guidance (**Fig. 1b**).

CAMUS enables highly accurate reference selection for cross-species annotation.

To demonstrate the effectiveness of CAMUS in prioritizing the reference datasets, we gathered 672 pairs of cross-species datasets from seven species and five tissues, each pair accompanied by manual annotations serving as the ground truth (**Methods**, **Table S1)**. There are 53 distinct query datasets, 47 of which have three or more reference datasets. We applied five reference-based annotation methods (i.e., CAME[8], Seurat[9], SciBet[10], scmap[11], and SCN[12]) to all the data pairs.

The CAMUS score and the annotation accuracy of the above five methods indeed demonstrated significant positive correlations ranging from 0.6286 to 0.8290 across all data pairs (**Fig. 2a**). In approximately 99% of all cases, the annotation accuracy using different references exhibited a positive correlation with the CAMUS score (**Supplementary Fig. 1a**). Detailed analysis of retinal data from four different species (with approximately thirty reference datasets for each) reveals that across all query datasets and methods, the references prioritized by CAMUS consistently align with the accuracy of their corresponding annotations. More importantly, in nearly all datasets, the reference with the highest CAMUS score consistently proves optimal (or suboptimal) (**Fig. 2b-d**). Furthermore, the reference selection has a substantial impact on the annotation accuracy. The best reference can improve it by more than 150% compared to the worst (**Fig. 2c**), highlighting the importance of reference selection.

Next, we quantitatively evaluated the improvements by selecting references using CAMUS. For five distinct methods, the annotation accuracy improves ranging from 25.0% to 124.7% (**Fig. 2f**) using the best reference selected by CAMUS compared to random selection for all 672 data pairs. Substantial gains were also observed in both macro F1 score (12.1-99.0%) and weighted F1 score (24.3-127.8%) (**Supplementary Fig. 1b, c**). This implies reference data selection can enhance annotation accuracy and rare cell type identification. Furthermore, we observed significant variation in the dependence on references among different annotation methods. For instance, under the random selection strategy, CAME achieved significantly higher accuracy than Seurat (median 66.71% vs 50.21%, p=2.2e-12). However, this

gap was greatly reduced when reference selection was guided by the CAMUS score (median 83.41% vs 81.01%, *p*=7.5e-04), and the annotation accuracy improved significantly (**Fig. 2f**). This phenomenon was similarly reflected in other methods. That is to say, prioritizing the reference datasets is essential for more accurate annotation of a given query data. CAMUS significantly outperforms random selection in consistently identifying the top 1-3 references with much higher accuracy (**Fig. 2g**). Moreover, CAMUS demonstrates strong robustness in selecting the best reference across different clustering resolutions and various clustering methods, where we achieved stable top-1 accuracy in selecting the best reference (~55% for CAME, Seurat, and SCN, ~80% for SciBet and scmap) (**Supplementary Fig. 1d, e**).

CAMUS enables highly accurate method selection for cross-species annotation.

For all the 672 pairs of cross-species datasets considered, the average correlation between the CAMUS score and the annotation accuracy across five methods stands at 0.8807, suggesting that higher CAMUS scores are typically associated with more accurate annotations by the methods. Specifically, for the retina data (**Table S1**), 595 out of the 598 data pairs demonstrated a positive correlation between the CAMUS score and the accuracy of the five annotation methods. It showed an average correlation of over 0.74 across all query datasets (**Fig. 3a**). Moreover, for all the 672 pairs of cross-species datasets, CAMUS achieved an accuracy of over 0.82 in selecting the best methods (**Fig. 3b**). Using methods selected by CAMUS, we observed a median of 51.3% increase in accuracy, a 49.4% increase in macro f1 score, and a 57.3% increase in weighted f1 score compared to the random selection (**Fig. 3c**).

Having confirmed that CAMUS can accurately select reference data and methods separately, we next examined the relationship between annotation accuracy and CAMUS scores for all reference-method pairs. For all the 3360 reference-method pairs, we achieved high accuracy in selecting the optimal pair for each query data. Specifically, we attained 49.06% for top-1 accuracy, 63.78% for top-2 accuracy, and 71.70% for top-3 accuracy, respectively. These results represented a significant improvement compared to random selection (**Supplementary Fig. 2a**). We next utilized the Chick P10 dataset[25] as the query data and utilized a total of 31 datasets from mice, humans, and zebrafish as reference data (**Fig. 3a**). When we aggregated all reference-method pairs, we observed a more complex relationship between actual accuracy and the CAMUS score, but the correlation remains strongly positive overall (r=0.9446, p=6.7e-76). Moreover, the highest-ranking reference-method pairs according to CAMUS were still those that best annotated this dataset (**Fig. 3d**). Meanwhile, in annotating the Macosko dataset using Zebrafish TNFa Adult as the reference, the annotation accuracy of the five

methods exhibited a slight negative correlation with the CAMUS scores (**Fig. 3a**). Additionally, all methods demonstrated relatively poor annotation accuracy, ranging from 16.42% to 32.37% (**Table S2**). To evaluate the cause, we carefully examined the cell proportions and the annotation quality between the reference and query datasets. We identified issues with the annotation of rod cells in the reference data, where some bipolar neurons also expressed gene *rho,* which is the marker of the rod cell[26] (**Supplementary Fig. 2b, c**). Rod cells happen to be the most prevalent cell type in the Macosko dataset (**Fig. 3e**). This indicates that reference quality is a critical factor in the process of reference-based annotation. We found that our clustering of the Mascosko dataset resulted in 17 groups, significantly more than the 11 existing cell types (**Fig. 3e**, **Supplementary Fig. 2d**). However, the prioritization of methods based on CAMUS scores still effectively reflected the performance of the annotations. We also discovered that a closer match in cell type proportions between the reference and query data does not necessarily lead to better annotation (**Fig. 3e**). Furthermore, the reference data Chick P10, which performed the best with CAME, was not the optimal choice for Seurat (**Supplementary Fig. 2e**). This indicates that different methods have different preferences for the reference data. Therefore, the optimal dataset should be re-selected for each.

CAMUS achieved additional enhancements when incorporated with reference-integrated strategies.

Due to limited reference datasets, we may encounter situations where the annotation may be unsatisfactory even with the best reference-method pair selection. This issue often arises because some references may lack distinctive cell types or are underrepresented or because some cells are mislabeled. Integrating multiple reference datasets can cover a broader range of cell types and mitigate the impacts of annotation errors in some datasets. How to use multiple references collectively to enhance the annotation quality is a critical issue. However, it remains unclear whether using multiple reference datasets as input can improve query annotation. Here, we collected 3522 pairs of cross-species datasets, where each query data has two distinct references (named reference 1 and 2, respectively) (**Methods**).

We employed the ensemble and multi-reference (multi-ref) strategies to integrate these references. However, we didn't apply the ensemble strategy with SciBet and scmap because they do not provide probability-based results (**Methods**). For most methods, employing a multi-ref strategy yields less than a 50% probability of outperforming both single references, i.e., CAME (n=1496, 42.5%), Seurat (n=1384, 39.3%), SCN (n=1333, 37.8%), SciBet (n=1896, 53.8%), and scmap (n=934, 26.5%). We can observe analogous results under an ensemble strategy, i.e., CAME (n=1602, 45.5%), Seurat (n=1499, 42.6%), SCN (n=911, 25.9%) (**Supplementary Fig. 3a**). Moreover, these two strategies ensure that in the majority of instances (~80% for multi-

ref and ~95% for ensemble), the accuracy will surpass the inferior single reference. In other words, these two strategies enhance annotation accuracy by moderately raising the upper limit and significantly ensuring the lower limit. That's why multi-ref and ensemble strategies likely yield superior performance compared to randomly selecting a reference (**Fig. 4a, b**). Besides integrating the reference data, another simple approach is to achieve the best-performing reference selection using CAMUS. CAMUS ensures a higher probability that the selected result will surpass any single reference, i.e., CAME (n=2613, 74.2%), Seurat (n=2689, 76.3%), SCN (n=2617, 74.2%), SciBet (n=2550, 72.4%), and scmap (n=2604, 73.9%). CAMUS significantly raises the lower limit of annotation results (**Supplementary Fig. 3a**). We also combined the above three strategies, i.e., using CAMUS to select the best among multi-ref, ensemble, and two single references. We found that this combined approach could yield improvements ranging from 9.3% to 39.3%, compared to randomly selecting from the two individual references (**Fig. 4a, b, Supplementary Fig. 3b**).

Next, we annotated the mouse NMDA P60 retinal data[27] using the zebrafish NMDA 20hr retinal data[27] (reference 1), zebrafish TNFa 36hr retinal data[27] (reference 2), and the two integrating strategies (**Fig. 4c, d**). Using references 1 and 2 individually led to annotation errors for bipolar neurons and rod cells, respectively. The ensemble and multi-ref strategies leveraged the advantages of both reference datasets, thereby correcting the misannotated cell types. Moreover, CAMUS perfectly aligned the ranking of these four scenarios (i.e., reference 1 only, reference 2 only, ensemble, and multi-ref) with their actual annotation performance (**Fig. 4d**). Additionally, when the two integrating strategies do not yield improvements, CAMUS can also assist us in rejecting them (**Supplementary Fig. 3c**).

In short, CAMUS can help us determine whether reference integration strategies offer improvements over single references and select the better-performing one (**Fig. 4a, b**). When reference data is limited, CAMUS can incorporate the outcomes of approaches for integrating the references and ensembles of different methods to achieve better annotation.

CAMUS enables highly accurate method selection for scST and scATAC-seq data.

We further applied CAMUS to 80 scST and five scATAC-seq datasets as the query data to assess its applicability to other complex scenarios (**Methods**). The scST datasets were derived from brain[15, 28-33], embryo[34], retina[35], kidney[18], and liver[18] profiled by six distinct spatial technologies (**Methods, Table S3**) and the five scATAC-seq datasets were obtained from PBMC[36], BMMC[37], Cortex[38], Skin[39], and Mop[40]. Each query dataset pairs with a well-annotated scRNA-seq dataset as a reference (**Methods, Table S4**). The CAMUS score and annotation accuracy in 78 scST datasets showed distinct positive correlations. The two exceptions come from the smFISH embryo 1

and 2 data due to their limited 33 genes (**Supplementary Fig. 4a, Table S3**). CAMUS achieved a median probability of 0.8250 to select the best methods (**Fig. 5a**). Even with significant adjustments in clustering resolution (from 0.2 to 0.6), CAMUS consistently maintained robust top-1 accuracy ranging from 82.50% to 86.25% (**Supplementary Fig. 4b**), highlighting its robustness to clustering parameters and underscoring its versatility and reliability in various biological scenarios. It is worth noting that, for the clustering of scST data, the limited number of genes measured may lead to blurry clustering boundaries or different cell types wrongly mixed into a single cluster. Specifically, we focused on the mouse hypothalamic region scST dataset containing 11 cell types[28]. With only 161 profiled genes, the clustering method cannot clearly distinguish excitatory and inhibitory neurons (**Supplementary Fig. 4c, d**). However, this does not affect the efficacy of CAMUS, which still provides method prioritizations that align perfectly with annotation accuracy (**Fig. 5b**).

For the scATAC-seq datasets, we applied a higher resolution for the clustering method because their annotations were at the subtype level. Notably, we did not employ scDART for the Mop data due to its runtime exceeding a week (**Methods**). We found that, in most cases, GLUE outperformed other methods, which was consistent with its CAMUS score. scJoint achieved slightly higher accuracy on the cortex data. CAMUS successfully captured this and selected scJoint as the optimal method (**Fig. 5c**). CAMUS exhibited perfect top-1 accuracy and stable performance with varied clustering resolution (**Supplementary Fig. 4f**). To further assess CAMUS's performance at the cellular subtype resolution, we carefully analyzed the PBMC data which comprises 19 subpopulations. Although CAMUS pre-clustered the data into only 12 clusters, showing a clear mismatch with the original number of subpopulations, it still achieved perfect consistency with annotation accuracy (**Fig. 5d**, **Supplementary Fig. 4e**).

CAMUS is a powerful predictor for the accuracy of cell type annotation. Although CAMUS facilitates the reference and method selection, users may not know whether the chosen reference-method pair can produce reliable annotations. For instance, if the best reference-method pair achieves a CAMUS score of 0.5, users can not predict the corresponding annotation accuracy, which may range from ~20% to ~90% (**Fig. 2a**). When users lack relevant biological knowledge, or when annotating data from a new species, it is exceedingly difficult for them to estimate the annotation accuracy. Even with relevant biological background knowledge, such as verifying the cell type annotations by examining the marker genes' expression, a significant amount of time is required. Therefore, it becomes critical to provide an estimate of annotation accuracy rapidly.

Given that the CAMUS score significantly correlates with accuracy, it could serve as a powerful predictor. Other predictors such as the cell number, the cluster number, and cell entropy under different clustering resolutions in

the query data, also contribute to predicting the cell annotation accuracy (**Methods**). We adopted the recently published machine learning method AutoGluon[41] and the median absolute error (MAE) between the estimated accuracy and the actual accuracy as the evaluation metric (**Methods**). We achieved an MAE of 0.0453 for the ten-fold cross-validation on 672 pairs of cross-species data encompassing all the cell annotation methods (**Fig. 6a, Methods**). Methods that performed better in annotation achieved even lower MAE values: CAME (0.0375), Seurat (0.0448), and SciBet (0.0344) (**Fig. 6a, Methods**). Moreover, we validated the trained model on scST data, achieving an MAE of 0.0560 (**Fig. 6b**). Our model exhibits low error across all transcriptomic data, potentially attributable to the similarity in their data types and the same normalization procedures. Despite the preprocessing and normalization methods for scATAC-seq data being entirely different from those of scRNA-seq, our model still achieved a low MAE (0.0557) (**Fig. 6c**). Consistent with the conclusions drawn from cross-species data, for the best-performing method, GLUE, we achieved lower MAE (0.0428). This indicated that for the best results obtained through CAMUS selection, we can expect even lower MAEs.

Conclusion

Precise and rapid annotation of single-cell data with reference-based methods is essential for deciphering cell types and their complex interactions[42]. The selection of references and methods distinctly influences the outcomes of annotations (**Fig. 2a, 3c**). However, the critical selection process has not been seriously explored. In this study, we develop a universal reference and method selection method CAMUS. CAMUS provides a score significantly correlated with the annotation accuracy. It works by directly compares the clustering results of data with its annotations, thus enabling a smooth expansion to any biologically structured single-cell data.

In real-world datasets, CAMUS demonstrated high accuracy to select reference and methods for annotating scRNA-seq, scST and scATAC-seq data. Extensive tests indicate that the clustering resolution has a minimal impact on the performance of CAMUS. Despite unclear boundaries or some errors in clustering, CAMUS consistently achieved high top-1 accuracy. Even when there was a significant discrepancy between the number of clusters and the actual cell types, The performance of CAMUS was not compromised. CAMUS can also be employed for method selection when dealing with annotations of subtypes within the data.

The reference data quality is critical for the reference-based annotation of single cells. Moreover, incorrect cell annotations can adversely affect the outcomes of annotation methods. Therefore, users should carefully evaluate the reference data quality and its annotations before employing reference-based methods. CAMUS is a powerful tool for assessing the effectiveness of

the reference data for annotating query data. When the reference data is of high quality, the differences among methods diminish significantly. Thus, when sufficient reference datasets are available, users should prioritize selecting the appropriate references. When a specific new single-cell technique lacks mature annotation methods, collecting and selecting proper references through CAMUS can be highly beneficial. This also poses new demands for establishing databases to collect more high-quality datasets (e.g., SpatialDB[43] and PanglaoDB[44]).

We also discovered that the best reference for one method might not be the best for another when given a specific query dataset. Therefore, merely comparing the similarity between query and reference data is insufficient. A more viable approach would involve comparing the annotation results with the biological structure of the query data itself. One potential limitation of CAMUS is that it does not consider the prior biological knowledge inherent in the data, such as marker genes for its potential cell types, which could serve as an additional measure of annotation accuracy.

Methods

Data preprocessing

We processed the count matrices for scRNA-seq and scST datasets via Scanpy[45] (https://scanpy.readthedocs.io/en/stable/index.html). We first normalized the library size for each cell, followed by a logarithmic transformation of the expression data with a pseudo-count. Subsequently, we selected the top 3000 highly variable genes (HVGs). In cases where the scST datasets comprised fewer than 3000 genes, we used the entire gene set. We clustered the data by the Leiden[46] algorithm with a default resolution parameter of 0.4. We used SnapATAC2[47] to preprocess the scATAC-seq datasets. Following its documentation (https://kzhang.org/SnapATAC2/index.html), we first imported data fragment files and the reference genome (hg38 for human and mm10 for mouse). We generated a cell-by-bin matrix containing insertion counts across genome-wide 500-bp bins. We then employed spectral embedding calculated on the top 250k features and clustered the data by the Leiden algorithm with a default resolution parameter of 0.7.

Benchmarking cell-type annotation

Cross-species scRNA-seq datasets

We performed (1) CAME[8] with its Python package, referred to its documentation (https://xingyanliu.github.io/CAME/tut_notebooks/getting_started_pipeline_un.html); (2) Seurat[9] with its R package detailed in its documentation (https://satijalab.org/seurat/articles/integration_mapping); (3) SciBet[10] with its R package detailed in its documentation (http://scibet.cancer-pku.cn/); (4) scmap-cluster[11] with the R package scmap, following its documentation (https://bioconductor.org/packages/release/bioc/vignettes/scmap/inst/doc/scmap.html); and (5) SCN[12] with its R package singleCellNet, following its documentation (https://github.com/CahanLab/singleCellNet?tab=readme-ov-file#introduction).

scST datasets

We performed (1) STAMapper with its Python package, referred to its documentation (https://github.com/zhanglabtools/STAMapper/tree/main/Tutorials); (2) scANVI[17] with the scVI Python package, referencing the “Integration and label transfer with Tabula Muris” section from its documentation (https://docs.scvi-tools.org/en/stable/tutorials/notebooks/scrna/tabula_muris.html); (3) RCTD[19] with the R package spacexr, utilizing its default workflow available at (https://raw.githack.com/dmcable/spacexr/master/vignettes/spatial-transcriptomics.html); and (4) Tangram[20] with its Python package, employing the cluster mode configuration detailed in the tutorial (https://tangram-sc.readthedocs.io/en/latest/tutorial_sq_link.html).

scATAC-seq datasets

We implemented (1) GLUE[22] with its Python package scglue, referring to its documentation (https://scglue.readthedocs.io/en/latest/); (2) scJoint with its Python package detailed in its GitHub page (https://github.com/SydneyBioX/scJoint); (3) Seurat[9] with its R package, referring to its documentation (https://satijalab.org/seurat/articles/seurat5_atacseq_integration_vignette); (4)

scANVI[17], referring to the documentation from SnapATAC2[47] (https://kzhang.org/SnapATAC2/tutorials/annotation.html); (5) scDART[23] with its Python package, following its documentation (https://github.com/PeterZZQ/scDART). Since GLUE and scDART only perform the data integration of scRNA-seq and scATAC-seq data, we trained a k-nearest neighbor (kNN) classifier from the sklearn[48] Python package with k=10 on the integrated embeddings of scRNA-seq data to predict the cell labels from scATAC-seq data.

Calculation of CAMUS score

CAMUS takes the count matrix of query data and the annotation results generated by different references and methods as input. CAMUS consists of two parts, we first pre-cluster the query data (section "Data preprocessing"). For query data with m cells: $C = \{c_1, \dots, c_m\}$, suppose we have the clustering results $U = \{U_1, \dots, U_s\}$ with s clusters, where U_i denotes ith cluster set and the annotation results $V = \{V_1, \dots, V_t\}$ with t cell types, where V_i denotes the cell set of type i. It should be noted that for all $i \neq j$:

$$U_i \cap U_j = V_i \cap V_j = \emptyset,$$

and

$$\cup_{i=1}^{s} U_i = \cup_{j=1}^{t} V_j = C.$$

In other words, we can view U and V as two partitions of C.

We then define the concordance score between the two partitions U and V as the adjusted mutual information[49]. Let matrix $M = [n_{ij}]$, $i = 1, \dots, s$, $j = 1, \dots, t$, where n_{ij} denotes the number of shared cells for U_i and V_j. We calculate the probability of randomly picking a cell that belongs to U_i:

$$P_U(i) = \frac{|U_i|}{m}.$$

The entropy of the partition U is:

$$H(U) = -\sum_{i=1}^{s} P_U(i) log P_U(i).$$

Similarly, we calculate the entropy for partition V:

$$H(V) = -\sum_{j=1}^{t} P_V(j) log P_V(j),$$

where $P_V(j) = \frac{|V_j|}{m}$. The mutual information between the two partitions is:

$$MI(U,V) = \sum_{i=1}^{s} \sum_{j=1}^{t} P_{UV}(i,j) log \frac{P_{UV}(i,j)}{P_U(i) P_V(j)},$$

where $P_{UV}(i,j)$ denotes the probability of randomly picking a cell that belongs to $U_i \cap V_j$, i.e., $P_{UV}(i,j) = \frac{|U_i \cap V_j|}{m}$. We also calculated the expected mutual information:

$$E\big(MI(U,V)\big) = \sum_{i=1}^{s}\sum_{j=1}^{t}\sum_{n_{ij}=(a_i+b_j-N)^+}^{\min(a_i,b_j)} \frac{n_{ij}}{m}\log\left(\frac{m\times n_{ij}}{a_i b_j}\right)$$

$$\times\frac{a_i!\,b_j!\,(m-a_i)!\,\big(m-b_j\big)!}{m!\,n_{ij}!\,\big(a_i-n_{ij}\big)!\,\big(b_i-n_{ij}\big)!\,\big(m-a_i-b_j+n_{ij}\big)!},$$

where $\big(a_i+b_j-N\big)^+$ denotes $\max(0, a_i+b_j-N)$ and $a_i=\sum_{j=1}^{t} n_{ij},\ b_j=\sum_{i=1}^{s} n_{ij}$.

Finaly, we defined the CAMUS concordance score between U and V as:

$$\frac{MI(U,V)-E(MI(U,V))}{\max\big(H(U),H(V)\big)-E(MI(U,V))}.$$

The score is assigned a value of one when the two partitions are identical, and it takes a value of zero when the mutual information between the two partitions is equal to the value expected by chance alone (**Supplementary Notes**). The CAMUS score approaches one as the alignment between the pre-cluster and the annotation results improves. Conversely, the CAMUS score tends toward zero when there is a marked discrepancy between these results.

Estimate the annotation accuracy

We used the number of total cells as the first predictor. We also used the number of clusters, cell entropy, and URAS concordance score under the clustering resolution of 0.2, 0.4, 0.6, and 0.8 as other predictors. Suppose we have a dataset of m cells, and clustering results $U=\{U_1,\dots,U_s\}$, where the number of cells in U_i is m_i. To calculate the cell entropy, we first calculate the cell ratio for each cluster: $\frac{m_i}{m}, i=1,\dots,s$, and then the cell entropy as:

$$H=-\sum_{i=1}^{s}\frac{m_i}{m}log\frac{m_i}{m}.$$

We applied AutoGuon to 13 predictors to estimate the annotation accuracy, we set eval_metric='median_absolute_error' for the "TabularPredictor" function, and set time_limit=1800, presets='best_quality' for the "fit" function.

Reference-integrated strategies

Suppose we have a query dataset with m cells and two reference datasets: reference 1 with n_1 cells and p_1 genes, and reference 2 with n_2 cells and p_2 genes. Here, we propose two strategies to integrate these two references. For the references of the same species and tissue type, we extract all possible combinations of them, ultimately obtaining 3,522 data triplets (reference 1, reference 2, query).

Ensemble

Suppose we have q_1 distinct cell types for reference 1 and q_2 distinct cell types for reference 2. Both references share q common cell types. Then, using reference 1 and query datasets as the input of the cell annotation method, we will have a

probability matrix $P_1 \in R^{m \times q_1}$, where each the row i and column j of P_1 denotes the probability of the cell i belongs to cell type j. And we will have $P_2 \in R^{m \times q_2}$ when using reference 2 and query datasets as the input. We denote P_1', $P_2' \in R^{m \times q}$ as the probability matrix for the q shared cell types. Also, utilizing the annotation results from references 1 and 2, we can calculate the CAMUS scores u_1 and u_2, respectively. We then ensemble the two probability matrices:

$$P = \frac{u_1}{u_1+u_2} \times P_1' + \frac{u_2}{u_1+u_2} \times P_2'.$$

We next incorporate the unique cell type probabilities to matrix P and assign the cell type with the highest probability to the corresponding cells.

Multi-ref

Here, we concatenate the two reference datasets. Assuming that reference 1 and 2 datasets share p genes, the resulting combined matrix consists of $n_1 + n_2$ cells and p genes. We used the combined matrix as a new reference to annotate query data.

Data availability

All datasets analyzed in this study are publicly available from the links provided.

Cross-species scRNA-seq datasets

Dataset 1-2: (Pancreas, Human, Mouse), inDrop, GSE84133 in the GEO database.
Dataset 3: (Pancreas, Human), CelSeq, CelSeq, Fluidigm C1, and SMART-Seq2, a mixed pancreas dataset produced across four technologies, distributed in SeuratData package, https://satijalab.org/seurat/articles/integration_mapping.
Dataset 4: (Pancreas, Mouse), Smart-seq2, https://figshare.com/articles/dataset/Single-cell_RNA-seq_data_from_Smart-seq2_sequencing_of_FACS_sorted_cells_v2_/5829687/7.
Dataset 5-7: (Testis, Human, Monkey, Mouse), Drop-seq, GSE142585 in the GEO database.
Dataset 8: (Brain, Human), snDrop-seq, GSE97942 in the GEO database.
Dataset 9: (Brain, Mouse), Smart-seq2, GSE115746 in the GEO database.
Dataset 10: (Brain, Mouse), Drop-seq, GSE93374 in the GEO database.
Dataset 11: (Brain, Mouse), Smart-seq2, https://figshare.com/articles/dataset/Single-cell_RNA-seq_data_from_Smart-seq2_sequencing_of_FACS_sorted_cells_v2_/5829687/7.
Dataset 12-13: (Brain, Lizard, Turtle), Drop-seq, https://brain.mpg.de/research/laurent-department/software-techniques.html.
Dataset 14: (Kidney, Mouse), Drop-seq, GSE94333 in the GEO database.
Dataset 15: (Kidney, Mouse), Drop-seq, GSE111107 in the GEO database.
Dataset 16: (Kidney, Mouse), 10x, GSE107585 in the GEO database.
Dataset 17: (Kidney, Mouse), 10x, https://figshare.com/articles/dataset/Single-cell_RNA-seq_data_from_microfluidic_emulsion_v2_/5968960/2.
Dataset 18: (Kidney, Human), 10x snRNA-seq, GSE118184 in the GEO database.
Dataset 19: (Kidney, Human), 10x, EGAS00001002171, EGAS00001002486, EGAS00001002325, and EGAS00001002553 in European Genome-phenome Archive.

Dataset 20: (Kidney, Human), 10x, GSE114530 in the GEO database.
Dataset 21: (Kidney, Human), 10x, GSE109488 in the GEO database.
Dataset 22: (Retina, Human), microfluidics, GSE137847 in the GEO database.
Dataset 23: (Retina, Human), Seq-Well, GSE137847 in the GEO database.
Dataset 24: (Retina, Mouse), Drop-seq, GSE63473 in the GEO database.
Dataset 25-37: (Retina, Mouse), 10x, GSE135406 in the GEO database.
Dataset 38-53: (Retina, Zebrafish), 10x, GSE13540 in the GEO database.
Dataset 54: (Retina, Chick), 10x, GSE13540 in the GEO database.
A detailed summary of the 672 reference-query pairs can be found in **Table S1**.

Paris of scRNA-seq and scST datasets

We collected 80 pairs of scRNA-seq and scST datasets from https://github.com/zhanglabtools/STAMapper/. A human liver cancer dataset profiled by NanoStrring was excluded due to the absence of manual annotation.

Paris of scRNA-seq and scATAC-seq datasets

Dataset 1 (PBMC, Human), 10x Multiome, https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k?.
Dataset 2 (BMMC, Human), 10x Multiome, GSE194122 in the GEO database.
Dataset 3 (Cortex, Mouse), SHARE-seq, GSE126074 in the GEO database.
Dataset 4 (Skin, Mouse), SHARE-seq, GSE140203 in the GEO database.
Dataset 5 (Mop, Mouse), The 10x scRNA-seq and scATAC-seq datasets can be accessed from NeMO archive https://assets.nemoarchive.org/dat-ch1nqb7. We used samples with id: CEMBA171206_3C, CEMBA171207_3C, CEMBA171213_4B, CEMBA180104_4B, CEMBA180409_2C, CEMBA180410_2C, CEMBA180612_5D, CEMBA180618_5D.

Code availability

CAMUS is implemented in Python and is available on GitHub [https://github.com/zhanglabtools/CAMUS].

Reference

1. Kharchenko, P.V. The triumphs and limitations of computational methods for scRNA-seq. *Nature methods* **18**, 723-732 (2021).
2. Longo, S.K., Guo, M.G., Ji, A.L. & Khavari, P.A. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nature Reviews Genetics* **22**, 627-644 (2021).
3. Fang, R. et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nature communications* **12**, 1337 (2021).
4. Jia, G. et al. Single cell RNA-seq and ATAC-seq analysis of cardiac progenitor cell transition states and lineage settlement. *Nature communications* **9**, 4877 (2018).
5. Srivatsan, S.R. et al. Embryo-scale, single-cell spatial transcriptomics. *Science* **373**, 111-117 (2021).
6. Clarke, Z.A. et al. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nature protocols* **16**, 2749-2764 (2021).
7. Brbić, M. et al. Annotation of spatially resolved single-cell data with STELLAR. *Nature Methods* **19**, 1411-1418 (2022).
8. Liu, X., Shen, Q. & Zhang, S. Cross-species cell-type assignment from single-cell RNA-

seq data by a heterogeneous graph neural network. *Genome Research* **33**, 96-111 (2023).

9. Hao, Y. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature biotechnology* **42**, 293-304 (2024).
10. Li, C. et al. SciBet as a portable and fast single cell type identifier. *Nature Communications* **11**, 1818 (2020).
11. Kiselev, V.Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nature Methods* **15**, 359-362 (2018).
12. Tan, Y. & Cahan, P. SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species. *Cell systems* **9**, 207-213. e202 (2019).
13. Shah, S. et al. Dynamics and spatial genomics of the nascent transcriptome by intron seqFISH. *Cell* **174**, 363-376. e316 (2018).
14. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
15. Zhang, M. et al. Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature* **598**, 137-143 (2021).
16. Shen, Q., Dong, K., Zhang, S. & Zhang, S. High-precision cell-type mapping and annotation of single-cell spatial transcriptomics with STAMapper. *bioRxiv*, 631859 (2025).
17. Xu, C. et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular systems biology* **17**, e9620 (2021).
18. Liu, J. et al. Concordance of MERFISH spatial transcriptomics with bulk and single-cell RNA sequencing. *Life science alliance* **6** (2023).
19. Cable, D.M. et al. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature biotechnology* **40**, 517-526 (2022).
20. Biancalani, T. et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature methods* **18**, 1352-1362 (2021).
21. Lin, Y. et al. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. *Nature biotechnology* **40**, 703-710 (2022).
22. Cao, Z.-J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology* **40**, 1458-1466 (2022).
23. Zhang, Z., Yang, C. & Zhang, X. scDART: integrating unmatched scRNA-seq and scATAC-seq data and learning cross-modality relationship simultaneously. *Genome biology* **23**, 139 (2022).
24. Garmire, L.X. et al. Challenges and perspectives in computational deconvolution of genomics data. *Nature Methods* **21**, 391-400 (2024).
25. Macosko, E.Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202-1214 (2015).
26. Hoang, T. et al. Gene regulatory networks controlling vertebrate retinal regeneration. *Science* **370**, eabb8598 (2020).
27. Baron, M. et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems* **3**, 346-360. e344 (2016).
28. Moffitt, J.R. et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018).
29. Russell, A.J. et al. Slide-tags enables single-nucleus barcoding for multimodal spatial genomics. *Nature* **625**, 101-109 (2024).
30. Codeluppi, S. et al. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature methods* **15**, 932-935 (2018).

31. Zeng, H. et al. Integrative in situ mapping of single-cell transcriptional states and tissue histopathology in a mouse model of Alzheimer's disease. *Nature Neuroscience* **26**, 430-446 (2023).

32. Shah, S., Lubeck, E., Zhou, W. & Cai, L. seqFISH accurately detects transcripts in single cells and reveals robust spatial organization in the hippocampus. *Neuron* **94**, 752-758. e751 (2017).

33. Eng, C.-H.L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568**, 235-239 (2019).

34. Lohoff, T. et al. Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. *Nature biotechnology* **40**, 74-85 (2022).

35. Choi, J. et al. Spatial organization of the mouse retina at single cell resolution by MERFISH. *nature communications* **14**, 4929 (2023).

36. PBMC from a healthy donor, single cell multiome ATAC gene expression demonstration data by Cell Ranger ARC 1.0.0. *10X Genomics* (2020).

37. Luecken, M.D. et al. in Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2) (2021).

38. Chen, S., Lake, B.B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology* **37**, 1452-1457 (2019).

39. Ma, S. et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103-1116. e1120 (2020).

40. Yao, Z. et al. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature* **598**, 103-110 (2021).

41. Erickson, N. et al. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505* (2020).

42. Dimitrov, D. et al. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nature communications* **13**, 3224 (2022).

43. Fan, Z., Chen, R. & Chen, X. SpatialDB: a database for spatially resolved transcriptomes. *Nucleic acids research* **48**, D233-D237 (2020).

44. Franzén, O., Gan, L.-M. & Björkegren, J.L. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019**, baz046 (2019).

45. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology* **19**, 1-5 (2018).

46. Traag, V.A., Waltman, L. & Van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports* **9**, 1-12 (2019).

47. Zhang, K., Zemke, N.R., Armand, E.J. & Ren, B. A fast, scalable and versatile tool for analysis of single-cell omics data. *Nature methods* **21**, 217-227 (2024).

48. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825-2830 (2011).

49. Vinh, N.X., Epps, J. & Bailey, J. in Proceedings of the 26th annual international conference on machine learning 1073-1080 (2009).

Figure Caption

Figure 1. Overview of CAMUS. **a**, CAMUS first pre-clusters single-cell data into several distinct clusters and then combines diverse reference datasets and reference-based annotation methods to annotate the query dataset. CAMUS compares the concordance score between the pre-clustering and the various annotation results. A higher CAMUS score indicates a potential higher annotation accuracy. **b**, CAMUS is applied to three cell type annotation scenarios: (1) Cross-species (including species from human, mouse, zebrafish, chick, lizard, turtle, and macaque); (2) Cross-modalities (from scRNA-seq dataset to scST dataset); (3) Cross-omics (from scRNA-seq dataset to scATAC-seq dataset). CAMUS helps to define the rank of reference-method pairs which is positively correlated with the corresponding real accuracy. CAMUS serves as a predictor to estimate annotation accuracy for users.

Figure 2. Application of CAMUS for reference selection in cross-species cell type annotation. **a-e**, Scatter plot showing the significant positive correlation between the CAMUS score and annotation accuracy in all data pairs (n=672) (**a**), and specific query datasets from humans, mice, chicks, and zebrafishes (**b-e**) for five cell annotation methods, i.e., CAME, Seurat, SciBet, scmap, and SCN. **f**, Boxplot indicating the accuracy of cell type annotation where references are selected by CAMUS versus random. The p-value was calculated using the paired t-test. The data following the red arrow represents the percentage improvement for the median value relative to random selection. **g**, Dot plot showing the top-1, top-2, and top-3 accuracy of selecting the best reference by CAMUS or randomly.

Figure 3. Application of CAMUS for method selection in 672 pairs of cross-species cell type annotation. **a**, Dot plot showing the correlation between the accuracy of five cell annotation methods (i.e. CAME, Seurat, SciBet, scmap, and SCN) and the corresponding CAMUS for retina data collected from zebrafishes, mice, humans, and chicks. For each dot, color represents a positive correlation, and gray represents a negative correlation. **b**, Bar plot showing the top-1, top-2, and top-3 accuracy of selecting the best methods by CAMUS or randomly. **c**, Boxplot indicating the accuracy, macro F1 score, and weighted F1 score of cell type annotation for methods selected by CAMUS or random. The data following the red arrow represents the percentage improvement for the median value relative to random selection. **d**, Pearson correlation between the CAMUS score and cell type annotation accuracy, where each point represents a reference-method pair. **e**, UMAP plot of mouse retina cells from the macosko dataset, cells are colored according to the

manual annotation, and the annotations of CAME using the menon seqwell, menon microfluidics, and chick P10 as references (upper panel). Bar plot showing the cell ratio of the corresponding dataset (lower panel).

Figure 4. Combined CAMUS with reference integrating strategies in cross-species cell type annotation. **a**, Boxplot indicating the accuracy of random selection from two given references (ref 1 and ref 2), using multi-reference strategy, using ensemble strategy, using CAMUS to select references for the given two references (ref 1, ref 2), using CAMUS to select references for the given two references and the ensemble results (ref 1, ref 2, ensemble), using CAMUS to select references for given two references alongside the multi-refs results (ref 1, ref 2, multi-ref), using CAMUS to select references for given two references, the multi-refs results, and the ensemble results (total). The p-value was calculated using the paired t-test. Blue p-values indicate that the average value of the boxplot on the right is smaller than that on the left, while black indicates the opposite. **b**, Boxplot indicating the accuracy of random selection from two given references, using multi-refs strategy, using CAMUS to select references for the given two references, and using CAMUS to select references for the given two references alongside the multi-refs results. **c**, UMAP plot of mouse NMDA P60 dataset from the retina, cells are colored by the manual annotation. AC, amacrine cells; EC, endothelial cells; PC, pericytes; RET, reticulocyte; HC, retinal horizontal cells; Cones, cone cells; RGC, retinal ganglion cells; RPE, retinal pigment epithelium; Rods, Rod cells. **d**, UMAP plot of mouse NMDA P60 dataset from the retina, cells are colored by annotation of CAME where using single reference zebrafish NMDA 20hr (the first panel), zebrafish TNFa 36hr (the second panel), the ensemble annotation results of the two references (the third panel), the annotation results which used two references as the input of CAME (the fourth panel).

Figure 5. Application of CAMUS is to select annotation methods in cross-modalities and cross-omics cell annotation scenarios. **a**, Bar plot showing the top-1 accuracy of selecting the best methods by CAMUS for annotating scST datasets. **b**, UMAP plots of mouse hypothalamic region scST dataset from the retina, cells are colored by manual annotation, the annotation of STAMapper, scANVI, RCTD, and Tangram, respectively. Astro, astrocytes; Endo, endothelial cells; Ep, ependymal cells; Exc, excitatory neurons; Inh, inhibitory neurons; OD, oligodendrocyte. **c**, Dot plot showing the positive relationship between CAMUS and ACC. PBMC, Peripheral Blood Mononuclear Cells; BMMC, Bone Marrow Mononuclear Cells; Mop, primary motor cortex. **d**, UMAP plots of PBMC scATAC-seq dataset, cells are colored according to manual annotations, and the annotations by GLUE, Seurat, scJoint, scANVI, and scDART, respectively.

Figure 6. Applications of CAMUS to estimate annotation accuracy. **a**, 10-fold cross-validation on 672 pairs of cross-species datasets using CAMUS and other features to estimate annotation accuracy of five different methods. The dashed line shows the median value of absolute error. **b**, Validation on 80 scST datasets, using the model trained on 672 pairs of cross-species datasets annotated by CAME, Seurat, and SciBet, respectively. **c**, Validation on five scATAC-seq datasets using the model trained on 672 pairs of cross-species datasets annotated by CAME, Seurat, and SciBet, respectively.

Figure 1

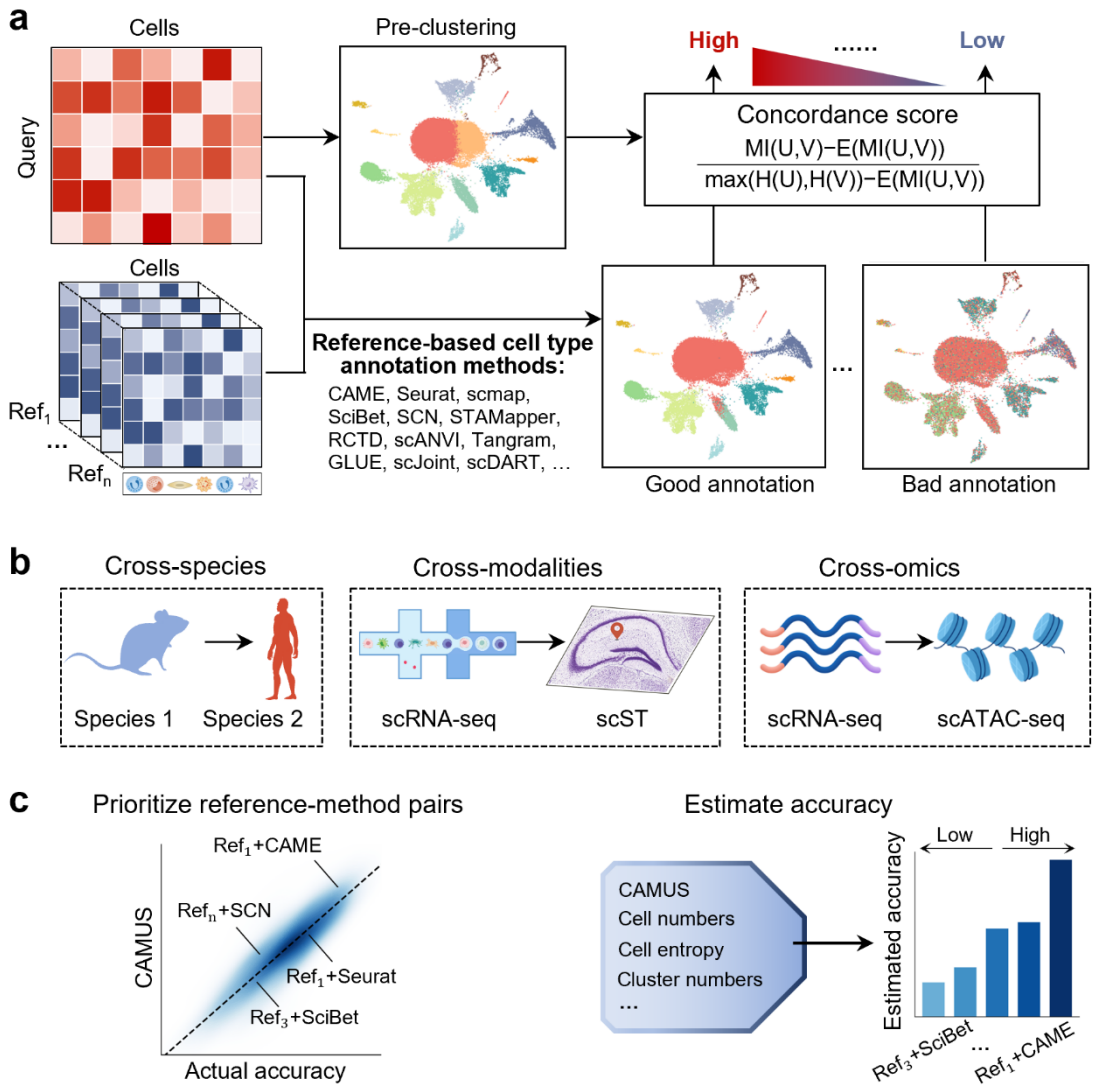

a
Cells
Query
Pre-clustering
High
.....
Low
Concordance score
MI(U,V)−E(MI(U,V))
max(H(U),H(V))−E(MI(U,V))
Cells
Ref1
...
Refn
Reference-based cell type annotation methods:
CAME, Seurat, scmap, SciBet, SCN, STAMapper, RCTD, scANVI, Tangram, GLUE, scJoint, scDART, …
Good annotation
Bad annotation
b
Cross-species
Species 1
Species 2
Cross-modalities
scRNA-seq
scST
Cross-omics
scRNA-seq
scATAC-seq
c
Prioritize reference-method pairs
Ref1+CAME
Refn+SCN
Ref1+Seurat
Ref3+SciBet
CAMUS
Actual accuracy
Estimate accuracy
CAMUS
Cell numbers
Cell entropy
Cluster numbers
...
Estimated accuracy
Low
High
Ref3+SciBet
...
Ref1+CAME

Figure 2

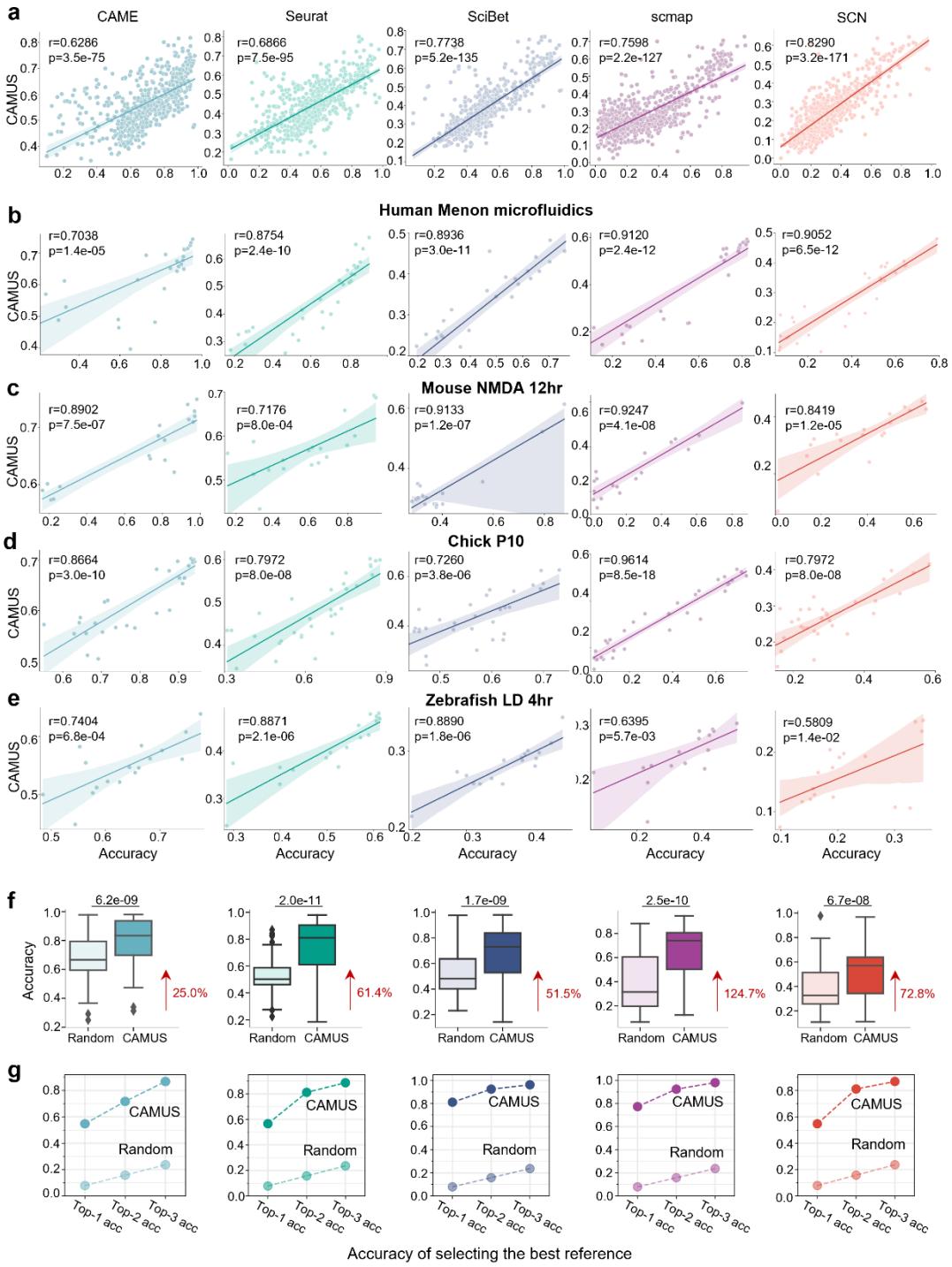

a
CAME
Seurat
SciBet
scmap
SCN
CAMUS
r=0.6286
p=3.5e-75
r=0.6866
p=7.5e-95
r=0.7738
p=5.2e-135
r=0.7598
p=2.2e-127
r=0.8290
p=3.2e-171
b
Human Menon microfluidics
r=0.7038
p=1.4e-05
r=0.8754
p=2.4e-10
r=0.8936
p=3.0e-11
r=0.9120
p=2.4e-12
r=0.9052
p=6.5e-12
c
Mouse NMDA 12hr
r=0.8902
p=7.5e-07
r=0.7176
p=8.0e-04
r=0.9133
p=1.2e-07
r=0.9247
p=4.1e-08
r=0.8419
p=1.2e-05
d
Chick P10
r=0.8664
p=3.0e-10
r=0.7972
p=8.0e-08
r=0.7260
p=3.8e-06
r=0.9614
p=8.5e-18
r=0.7972
p=8.0e-08
e
Zebrafish LD 4hr
r=0.7404
p=6.8e-04
r=0.8871
p=2.1e-06
r=0.8890
p=1.8e-06
r=0.6395
p=5.7e-03
r=0.5809
p=1.4e-02
Accuracy
f
Accuracy
6.2e-09
2.0e-11
1.7e-09
2.5e-10
6.7e-08
25.0%
61.4%
51.5%
124.7%
72.8%
Random
CAMUS
g
CAMUS
Random
Top-1 acc
Top-2 acc
Top-3 acc
Accuracy of selecting the best reference

Figure 3

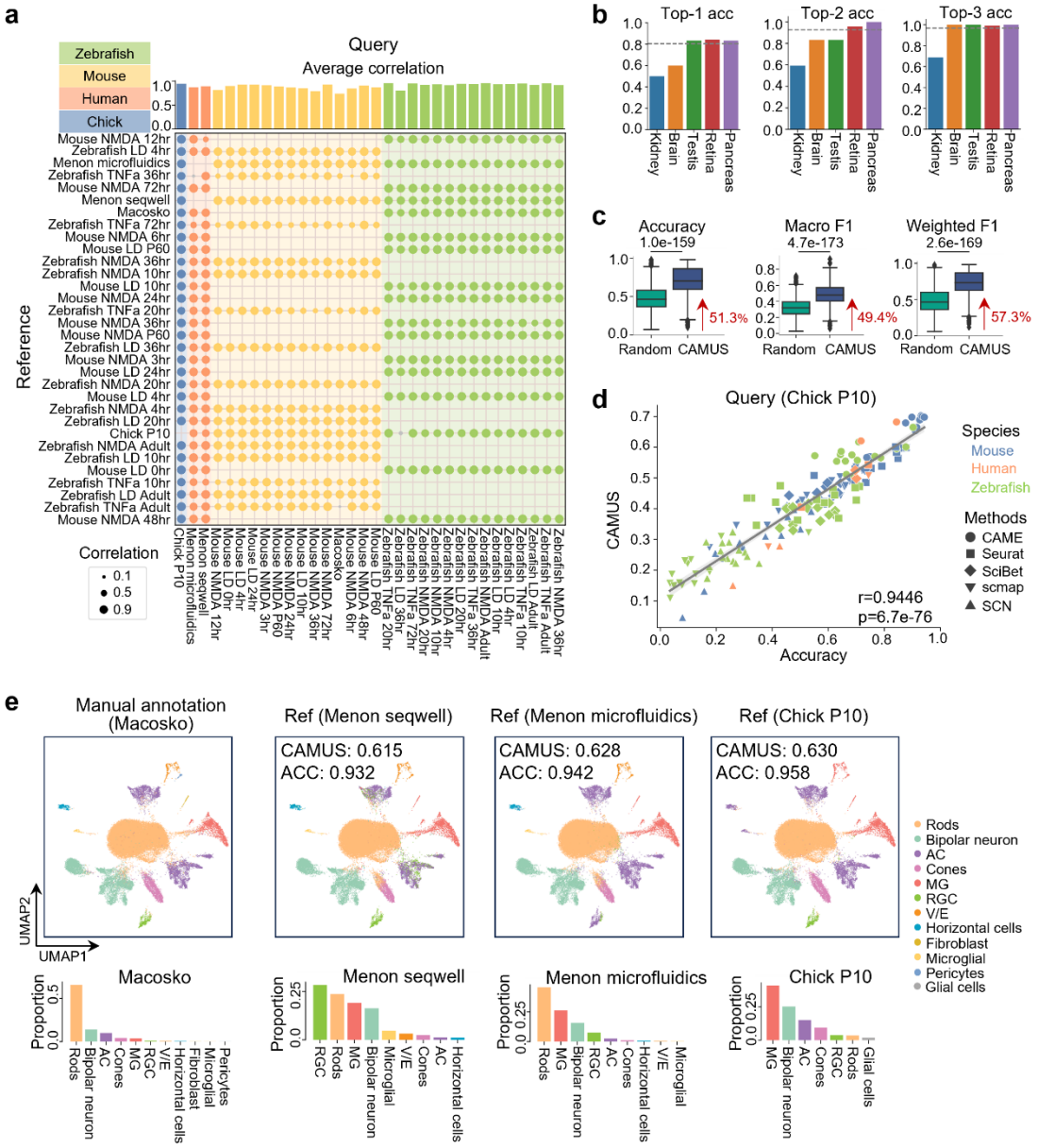

a
Zebrafish
Mouse
Human
Chick
Query
Average correlation
Reference
Correlation
b
Top-1 acc
Top-2 acc
Top-3 acc
Kidney
Brain
Testis
Retina
Pancreas
c
Accuracy
1.0e-159
Macro F1
4.7e-173
Weighted F1
2.6e-169
51.3%
49.4%
57.3%
Random
CAMUS
d
Query (Chick P10)
CAMUS
Accuracy
r=0.9446
p=6.7e-76
Species
Mouse
Human
Zebrafish
Methods
CAME
Seurat
SciBet
scmap
SCN
e
Manual annotation (Macosko)
Ref (Menon seqwell)
Ref (Menon microfluidics)
Ref (Chick P10)
CAMUS: 0.615
ACC: 0.932
CAMUS: 0.628
ACC: 0.942
CAMUS: 0.630
ACC: 0.958
UMAP2
UMAP1
Rods
Bipolar neuron
AC
Cones
MG
RGC
V/E
Horizontal cells
Fibroblast
Microglial
Pericytes
Glial cells
Macosko
Menon seqwell
Menon microfluidics
Chick P10
Proportion

Figure 4

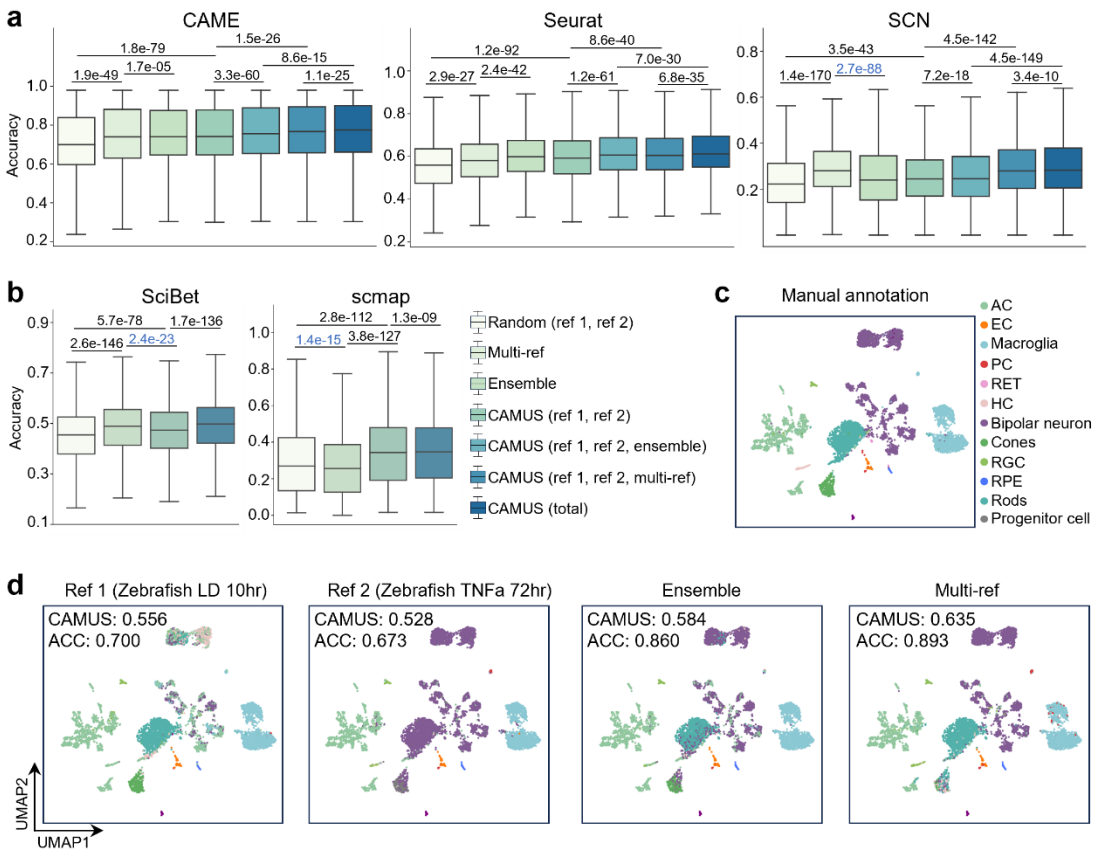

a
CAME
Seurat
SCN
Accuracy
1.9e-49
1.8e-79
1.7e-05
1.5e-26
3.3e-60
8.6e-15
1.1e-25
2.9e-27
1.2e-92
2.4e-42
8.6e-40
1.2e-61
7.0e-30
6.8e-35
1.4e-170
3.5e-43
2.7e-88
4.5e-142
7.2e-18
4.5e-149
3.4e-10
b
SciBet
scmap
2.6e-146
5.7e-78
2.4e-23
1.7e-136
1.4e-15
2.8e-112
3.8e-127
1.3e-09
Random (ref 1, ref 2)
Multi-ref
Ensemble
CAMUS (ref 1, ref 2)
CAMUS (ref 1, ref 2, ensemble)
CAMUS (ref 1, ref 2, multi-ref)
CAMUS (total)
c
Manual annotation
AC
EC
Macroglia
PC
RET
HC
Bipolar neuron
Cones
RGC
RPE
Rods
Progenitor cell
d
Ref 1 (Zebrafish LD 10hr)
CAMUS: 0.556
ACC: 0.700
Ref 2 (Zebrafish TNFa 72hr)
CAMUS: 0.528
ACC: 0.673
Ensemble
CAMUS: 0.584
ACC: 0.860
Multi-ref
CAMUS: 0.635
ACC: 0.893
UMAP2
UMAP1

Figure 5

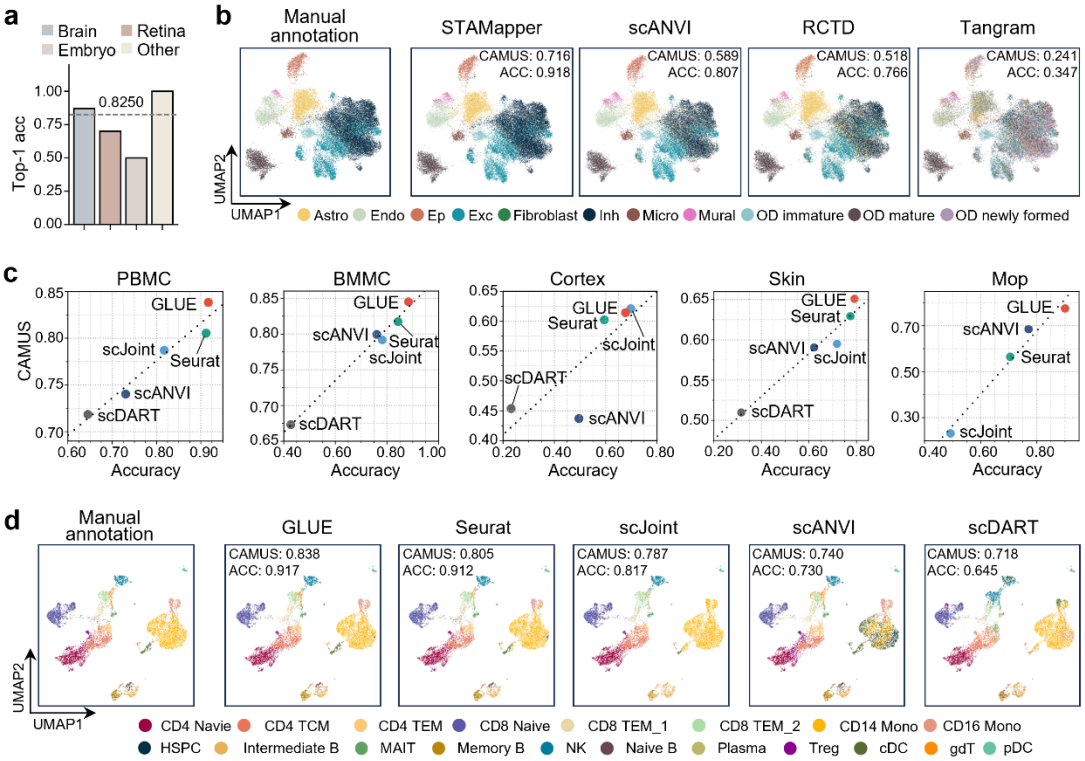
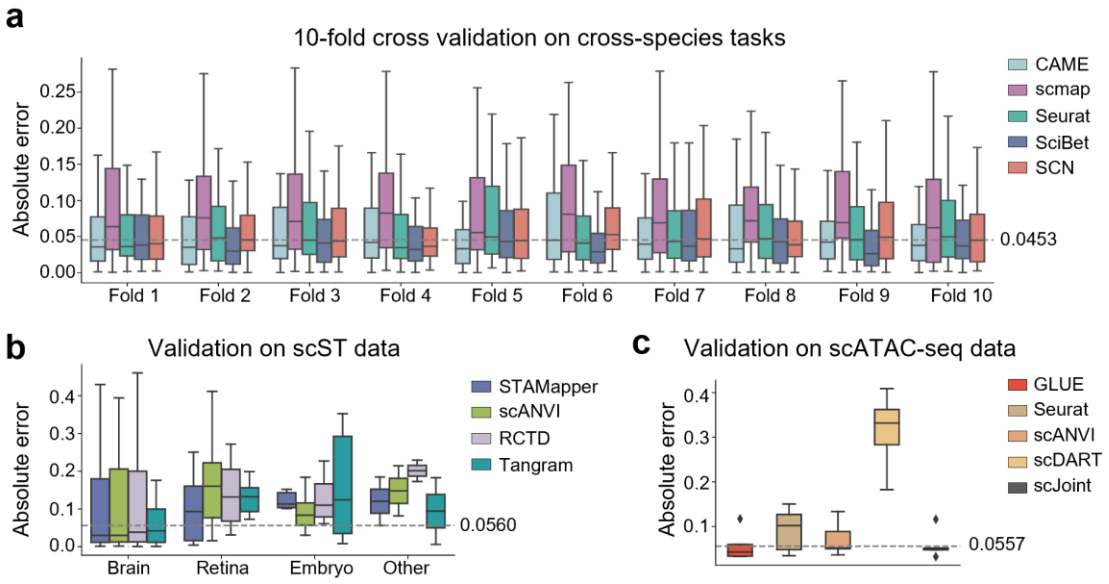

a
Brain
Retina
Embryo
Other
Top-1 acc
0.8250
b
Manual annotation
STAMapper
scANVI
RCTD
Tangram
CAMUS: 0.716
ACC: 0.918
CAMUS: 0.589
ACC: 0.807
CAMUS: 0.518
ACC: 0.766
CAMUS: 0.241
ACC: 0.347
UMAP2
UMAP1
Astro
Endo
Ep
Exc
Fibroblast
Inh
Micro
Mural
OD immature
OD mature
OD newly formed
c
PBMC
BMMC
Cortex
Skin
Mop
CAMUS
Accuracy
GLUE
Seurat
scJoint
scANVI
scDART
d
Manual annotation
GLUE
Seurat
scJoint
scANVI
scDART
CAMUS: 0.838
ACC: 0.917
CAMUS: 0.805
ACC: 0.912
CAMUS: 0.787
ACC: 0.817
CAMUS: 0.740
ACC: 0.730
CAMUS: 0.718
ACC: 0.645
UMAP2
UMAP1
CD4 Navie
CD4 TCM
CD4 TEM
CD8 Naive
CD8 TEM_1
CD8 TEM_2
CD14 Mono
CD16 Mono
HSPC
Intermediate B
MAIT
Memory B
NK
Naive B
Plasma
Treg
cDC
gdT
pDC

Figure 6


a
10-fold cross validation on cross-species tasks
Absolute error
0.25
0.20
0.15
0.10
0.05
0.00
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Fold 6
Fold 7
Fold 8
Fold 9
Fold 10
0.0453
CAME
scmap
Seurat
SciBet
SCN
b
Validation on scST data
Absolute error
0.4
0.3
0.2
0.1
0.0
Brain
Retina
Embryo
Other
0.0560
STAMapper
scANVI
RCTD
Tangram
c
Validation on scATAC-seq data
Absolute error
0.4
0.3
0.2
0.1
0.0557
GLUE
Seurat
scANVI
scDART
scJoint