Computational and Structural Biotechnology Journal 23 (2024) 3270–3280

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Method article

VICTOR: Validation and inspection of cell type annotation through optimal regression

Chia-Jung Chang [a,b,c], Chih-Yuan Hsu [a,b], Qi Liu [a,b,*], Yu Shyr [a,b,*]

[a] Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN 37203, USA
[b] Center for Quantitative Sciences, Vanderbilt University Medical Center, Nashville, TN 37203, USA
[c] Department of Biomedical Engineering, National Cheng Kung University, Tainan, Taiwan

ARTICLE INFO

Keywords:
Single-cell RNA sequencing analysis
Assessment of annotation quality
Cell type annotation
Elastic-net regularized regression

ABSTRACT

Single-cell RNA sequencing provides unprecedent opportunities to explore the heterogeneity and dynamics inherent in cellular biology. An essential step in the data analysis involves the automatic annotation of cells. Despite development of numerous tools for automated cell annotation, assessing the reliability of predicted annotations remains challenging, particularly for rare and unknown cell types. Here, we introduce VICTOR: Validation and inspection of cell type annotation through optimal regression. VICTOR aims to gauge the confidence of cell annotations by an elastic-net regularized regression with optimal thresholds. We demonstrated that VICTOR performed well in identifying inaccurate annotations, surpassing existing methods in diagnostic ability across various single-cell datasets, including within-platform, cross-platform, cross-studies, and cross-omics settings.

1. Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized biological research, offering unprecedented opportunities to explore cellular heterogeneity, identify rare cell types and states, characterize cellular microenvironments, and reconstruct developmental trajectories [1–6]. An essential step in scRNA-seq analysis is assignment of cell identities, which is essential for interpretability. Manual annotation, while effective, is time-consuming, requires expert input, and becomes increasingly challenging as dataset size and complexity grow. Therefore, there is a rising demand for tools that leverage existing annotated datasets to automate cell annotation, as well as methodologies that assess the quality of these automatically assigned identities.

Numerous tools have been developed for the automatic assignment of cell identities. Some tools employ a cluster-and-annotate strategy, where cells are first clustered based on gene expression similarities, and each cluster is then assigned an identity as a whole [7–12]. In contrast, other tools directly annotate individual cells without prior clustering [13–32]. Most annotation tools utilize a reference expression profile with known labels and automatically assign cell identity by either finding the best match in the reference [7–23] or by using a supervised classifier trained on the reference [24–32]. Some tools, such as scmap [13], SCINA [14], CHETAH [15], scClassify [16], Seurat [17], and scPred [24], label cells with predictive probability scores lower than a predetermined threshold as "unknown" or "unassigned". Alternatively, methods like singleR [18] evaluate the confidence of labels based on the distribution of predicted scores. However, these assessment strategies have limited efficacy when cells are rare or absent in the reference, or when the reference contains highly similar cell types.

To overcome this limitation, we introduce VICTOR: Validation and Inspection of Cell Type Annotation through Optimal Regression. VICTOR utilizes an elastic-net regularized regression along with cell type-specific optimal thresholds selection, maximizing the sum of sensitivity and specificity based on Youden's J statistic [33]. VICTOR effectively identified unreliable annotations from seven widely-used automated annotation methods, including singleR, scmap, scPred, SCINA, CHETAH, scClassify, and Seurat. Additionally, it outperformed these tools in diagnostic accuracy across different studies, platforms, tissues, and even cross-omics scenarios. Notably, VICTOR effectively pinpointed rare and unknown cells that were misclassified by these methods when such cells were underrepresented in the reference.

* Corresponding authors at: Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN 37203, USA.
E-mail addresses: qi.liu@vumc.org (Q. Liu), yu.shyr@vumc.org (Y. Shyr).

https://doi.org/10.1016/j.csbj.2024.08.028
Received 8 May 2024; Received in revised form 30 August 2024; Accepted 31 August 2024
Available online 2 September 2024

2. Results

2.1. Challenges in existing cell type annotations

We assessed the performance of seven widely-used cell annotation approaches: singleR, scmap, SCINA, scPred, CHETAH, scClassify, and Seurat. After assigning each cell to a label, singleR evaluates the prediction's reliability using a diagnostic metric. If the difference between the score for the assigned label and the median score across all labels falls below a certain threshold, the prediction is designated as unreliable, and the corresponding cell is labeled as 'unknown'. In contrast, scmap, SCINA, scPred, CHETAH, scClassify, and Seurat directly reject unreliable annotations with predictive scores or probabilities below a certain threshold, labeling the corresponding cells as 'unknown'. We defined true positives (TP) as correct annotations diagnosed as reliable, true negatives (TN) as incorrect annotations diagnosed as unreliable, false negatives (FN) as correct annotations diagnosed as unreliable, and false positives (FP) as incorrect annotations diagnosed as reliable.

We utilized a PBMC dataset as the reference and a second dataset generated from the same platform (10xV2) as the query [34]. The reference comprised 1611 cells representing nine cell types: 141 B cells, 273 CD4 + T cells, 325 CD14 + monocytes, 53 CD16 + monocytes, 568 cytotoxic T cells, 26 dendritic cells, 120 megakaryocytes, 87 natural killer cells, 18 plasmacytoid dendritic cells. The query comprised 1681 cells covering the same nine cell types, including 426 B cells, 479 CD4 + T cells, 208 CD14 + monocytes, 24 CD16 + monocytes, 348 cytotoxic T cells, 47 dendritic cells, 13 megakaryocytes, 117 natural killer cells, and 19 plasmacytoid dendritic cells. To simulate scenarios involving unknown cells, we deliberately excluded all B cells from the reference. Consequently, in this case, all B cells in the query were expected to be assigned as an unreliable and categorized as 'unknown'.

Surprisingly, singleR, scmap, CHETAH, and scClassify misclassified most queried B cells as other cell types and mistakenly deemed those incorrect annotations as reliable, leading to high false positive rates and low accuracies of 1 %, 2 %, 15 %, and 4 %, respectively (Fig. 1A). In contrast, SCINA, Seurat, and scPred correctly identified most queried B cells as unknown, resulting in only 8, 4, and 1 FPs, respectively and achieved accuracies of > 98 % (Fig. 1A).

In addition to the challenge of annotating unknown cells, current methods have difficulty in determining the identities of rare cells or closely related cells. For example, scmap correctly annotated 13 rare megakaryocytes (Fig. S1A), but it mischaracterized these annotations as unreliable (Fig. S1B), resulting in 13 false negatives and 0 % of accuracy (Fig. 1A). scPred obtained 8 false negatives out of 19 rare plasmacytoid dendritic cells–only 58 % accuracy (Fig. 1A). All seven methods struggled with cytotoxic T cells, showing elevated numbers of false positives and/or false negatives. This compromised performance is likely caused by similar expression profiles in CD4 + and cytotoxic T cells, making it challenging to distinguish between these two types of cells (Fig. 1A).

In summary, existing cell type annotation methods generated unreliable annotation assessments for cells that were underrepresented and among similar cell types included in the reference.

2.2. VICTOR improved diagnostic ability

VICTOR utilizes an elastic-net regularized regression to train a

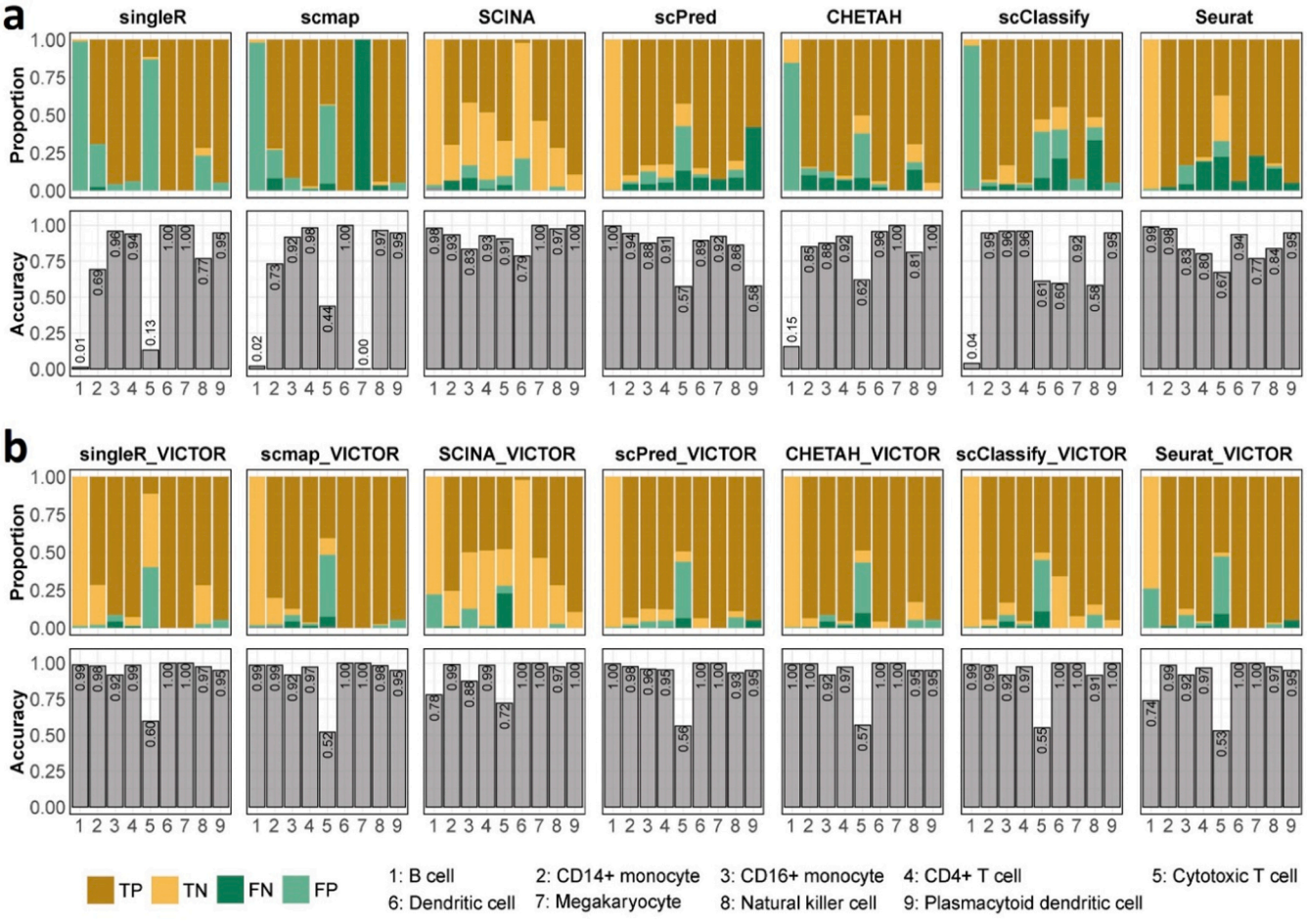

Fig. 1. One example in diagnosing the reliability of cell annotations. a) diagnostic performance of singleR, scmap, SCINA, scPred, CHETAH, scClassify, and Seurat. b) diagnostic performance of VICTOR when applied to annotations from singleR, scmap, SCINA, scPred, CHETAH, scClassify, and Seurat.

classifier. Instead of selecting one threshold across all cell types to assign annotation reliability, VICTOR chooses an optimal threshold for each cell type by maximizing the sum of specificity and sensitivity (Fig. 2) (details in Methods).

In the previous scenario where both the PBMC reference and query were generated from 10xV2 and B cells were absent from the reference, VICTOR significantly enhanced the diagnostic performance of all seven methods. Specifically, VICTOR successfully recognized that nearly all incorrectly annotated B cells from singleR, scmap, CHETAH, and scClassify were unreliable, resulting in accuracies exceeding 99 % (Fig. 1B). Additionally, VICTOR identified 13 false negatives as true positives in megakaryocytes from scmap, improving accuracy from 0 % to 100 % (Fig. 1B). The most significant improvement by VICTOR for SCINA annotations was observed in dendritic cells; SCINA misclassified 10 out of 47 dendritic cells as reliable CD14 + monocytes (false positives), but VICTOR recognized these annotations as incorrect (true negatives), resulting in an accuracy improvement from 79 % to 100 % (Fig. 1B). In the case of scPred annotations, VICTOR significantly reduced the number of false negatives across all cell types except cytotoxic T cells. For instance, scPred mischaracterized the annotations of 8 plasmacytoid dendritic cells as unreliable (false negatives), while VICTOR correctly identified 7 of these as reliable (true positives), improving accuracy from 58 % to 95 % (Fig. 1B). VICTOR also improved the accuracy of Seurat annotations for megakaryocytes from 77 % to 100 % and natural killer cells from 84 % to 97 % (Fig. 1B). Additionally, VICTOR's ROC curves displayed AUC values close to 1 for every cell type, demonstrating exceptional performance (Fig. S2).

We further evaluated the minimum number of B cells required in the reference to achieve optimal performance with VICTOR. As previously demonstrated, VICTOR maintained strong performance even without B cells in the reference. With as few as 10 B cells, VICTOR achieved nearly 100 % accuracy, with performance remaining stable as the number of B cells increased. An exception was observed with scPred annotations, where VICTOR consistently achieved near-perfect accuracy with at least 30 B cells. In contrast, other methods performed poorly without B cells and/or showed significant fluctuations in accuracy even as the number of B cells increased (Fig. S3).

2.3. *VICTOR improved diagnostic ability in within-platform annotations*

We first conducted a comprehensive evaluation of VICTOR's performance using PBMC datasets [34] within a single-platform setting. Each PBMC dataset was randomly divided into halves, with one designated as the reference and the other as the query. This strategy ensured that both the reference and the query originated from the same platform, minimizing any technical or platform biases. Additionally, we excluded individual cell types from the reference to simulate scenarios involving unknown cell types. Since PMBC datasets include nine cell types, this process yielded a total of 10 scenarios, including one where no cell types were excluded ('none' in Fig. 3), and nine scenarios where one cell type was omitted from the reference. The PBMC datasets originate from seven distinct platforms, including three samples from the 10X V2 platform, and one sample each from six other platforms. Resulting in a total of fifteen sets, this includes nine sets from the 10X V2 platform (3 samples combined in pairs) and six additional sets from the other platforms. In each set, both the reference and the query originated from the same platform.

In the scenario where the reference included all cell types (the 'none' scenario), VICTOR exhibited improved diagnostic performance, with a median accuracy increase of 23.8 % compared to singleR, 22.7 % compared to CHETAH, 11.5 % compared to scmap, 7.1 % compared to SCINA, 3.9 % compared to scClassify, 2.8 % compared to Seurat, and 2.7 % compared to scPred (Fig. 3).

In scenarios where the reference lacked certain cell types, VICTOR demonstrated even greater improvements in performance (Fig. 3). When cells were annotated by singleR, VICTOR enhanced diagnostic accuracy with median increases ranging from 19.7 % to 40.7 %. It is important to note that VICTOR performed especially better in scenarios where B cells were absent, resulting in a median accuracy increase of 40.7 % (Fig. 3). For instance, in the scenario lacking B cells in the 10xV2 platform, singleR misidentified 140 B cells as CD4 + T cells, 7 as dendritic cells, and 48 as plasmacytoid dendritic cells, erroneously considering all 195 annotations as reliable (FP). In contrast, VICTOR correctly identified all 195 annotations as unreliable (TN) (Fig. S4A). When cells were annotated by scmap, VICTOR exhibited median accuracy increases ranging from 10.3 % to 29.1 %, with the most substantial improvement occurring when megakaryocytes were absent from the reference (Fig. 3). In the scenario without megakaryocytes in the inDrop platform, scmap correctly annotated 74 natural killer cells but mischaracterized 71 cells as unreliable (FN); VICTOR reduced the number of FNs to 4 (Fig. S4B). For SCINA annotations, VICTOR showed median accuracy increases ranging from 0.2 % to 6.8 %, with a particularly notable enhancement

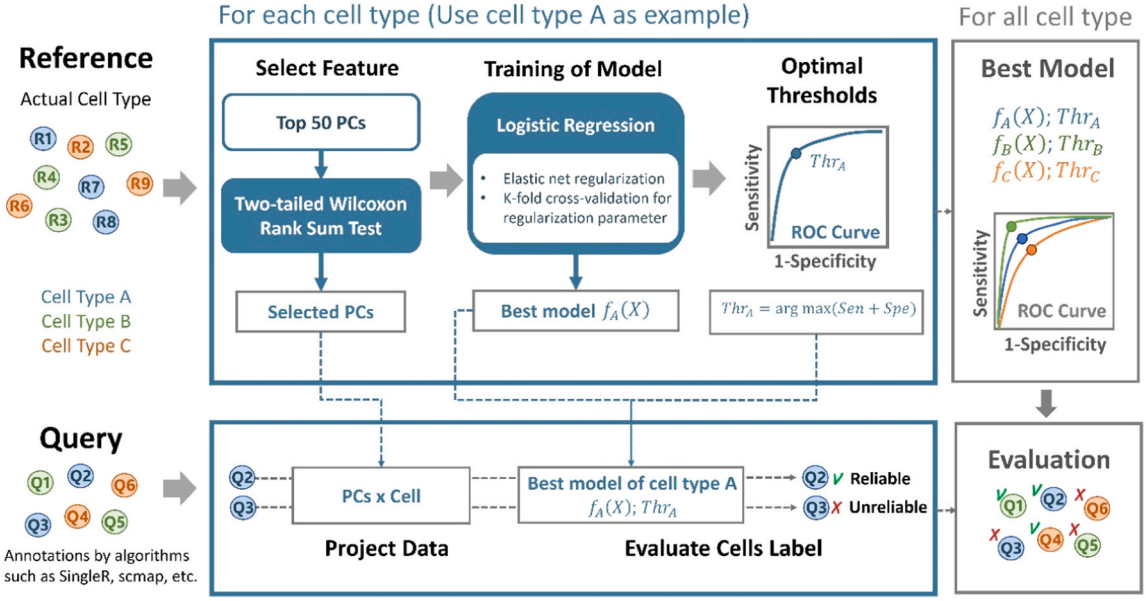

Fig. 2. Overview of VICTOR.

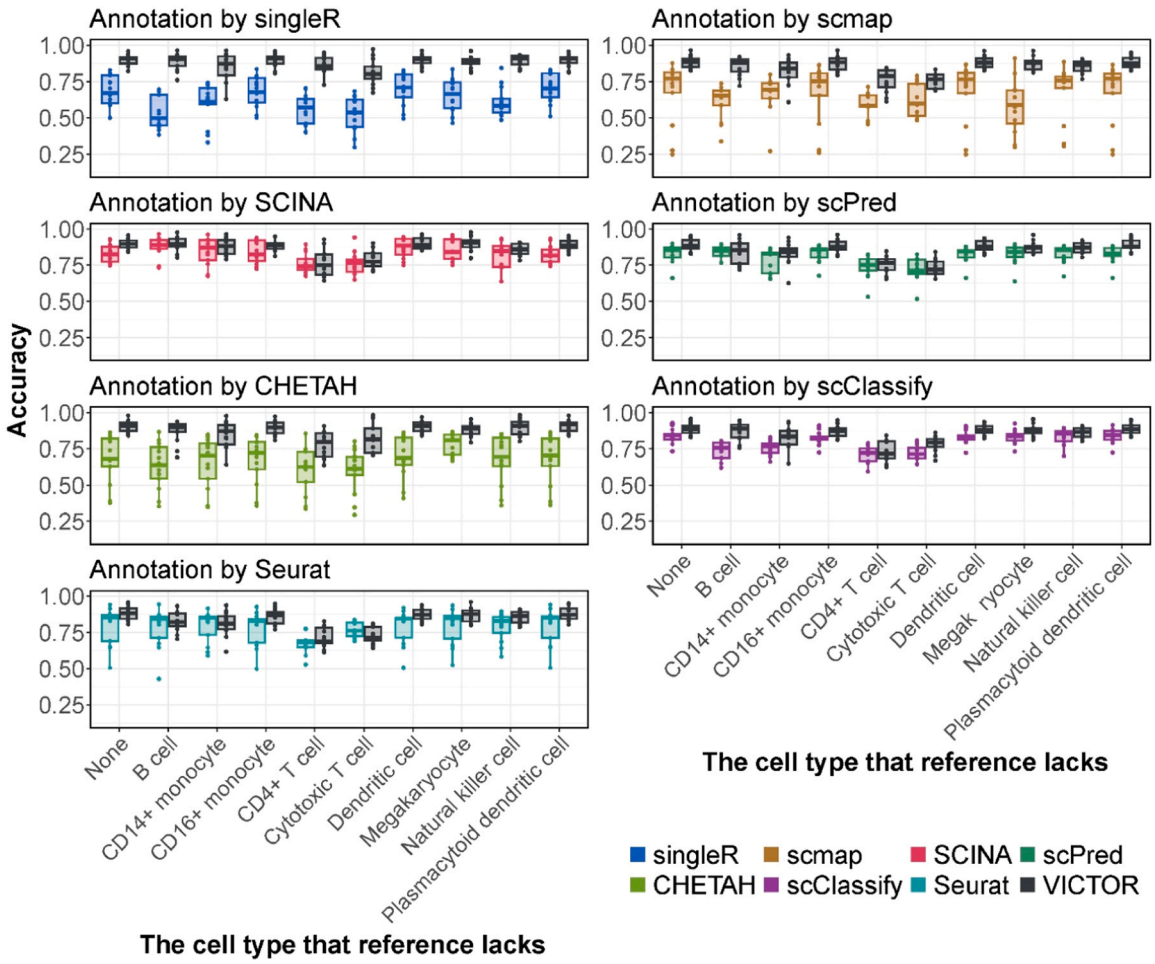

Fig. 3. Diagnostic performance in within-platform annotations using PBMC datasets generated from the same platform.

in the scenario where plasmacytoid dendritic cells were missing, increasing median accuracy from 81.7 % to 88.5 %. For instance, in the 10xV2 platform, SCINA accurately identified 390 CD4 + T cells but incorrectly labeled 198 as unreliable (FN). VICTOR correctly recognized 378 out of the 390 as reliable (Fig. S4C). scPred consistently performed well across various scenarios, with VICTOR slightly enhancing diagnostic accuracy by 0.6 % to 4.9 % (Fig. 3). Notably, in the scenario lacking plasmacytoid dendritic cells in the 10xV2 platform, scPred incorrectly annotated 19 plasmacytoid dendritic cells as CD14 + monocytes and considered them reliable (FP). VICTOR correctly identified these annotations as unreliable (TN) (Fig. S4D). When cells were annotated by CHETAH, VICTOR demonstrated median accuracy increases ranging from 7.9 % to 26.2 %, with the most significant improvement observed when B cells were absent from the reference (Fig. 3). For example, in the scenario where B cells were missing in the 10xV3 platform, CHETAH incorrectly identified 159 B cells as dendritic cells, 3 as cytotoxic T cells, and 3 as CD4 + T cells, mistakenly classifying them as reliable (FP). In contrast, VICTOR correctly flagged 169 out of 171 annotations as unreliable (TN) (Fig. S4E). For scClassify annotations, VICTOR's median accuracy gains ranged from 1.3 % to 13.3 %, except in cases where CD4 + T cells were missing, with the most notable enhancement occurring when B cells were missing from the reference (Fig. 3). For instance, scClassify misidentified 160 B cells as dendritic cells, considering 158 reliable (FP), 1 as a cytotoxic T cell and marked it as reliable, and 10 as CD4 + T cells, with 6 marked as reliable. VICTOR accurately recognized 169 out of those 171 annotations as unreliable (TN) (Fig. S4F). With Seurat annotations, VICTOR showed median accuracy improvements between 0.5 % and 3.9 %, except in cases where B cells, CD14 + monocytes, and cytotoxic T cells were missing. The most substantial boost was noted when CD16 + monocytes were absent from the reference (Fig. 3). For instance, in the absence of CD16 + monocytes in the 10xV2 platform, Seurat misclassified all 31 CD16 + monocytes as CD14 + monocytes, incorrectly deeming them reliable (FP), whereas VICTOR correctly identified 16 out of the 31 annotations as unreliable (TN) (Fig. S4G).

2.4. *VICTOR improved diagnostic ability in cross-platform annotations*

We assessed VICTOR's performance using PBMC datasets [34] in a cross-platform setting. For these comparisons, the reference and query datasets were generated on different sequencing platforms. Each platform possesses specific features or biases, rendering the annotation of one to another to be challenging. As in previous comparisons, we excluded individual cell types from the reference to simulate scenarios with unknown cell types. This process yielded a total of 10 scenarios. In the baseline scenario, no cell types were excluded ('none' in Fig. 4), whereas 9 scenarios were created to represent cell type that were excluded from reference. The PBMC datasets include three samples from the 10X V2 platform and each of six samples from six other platforms, resulting in a total of 66 sets. This comprised 18 sets created by pairing three queries from the 10X V2 platform with six references from non-10X V2 platforms, and an additional 48 sets formed by pairing six queries from non-10X V2 platforms with eight references from platforms other than themselves.

VICTOR consistently outperformed singleR, scmap, CHETAH, Seurat, and scClassify across all scenarios, while demonstrating slight

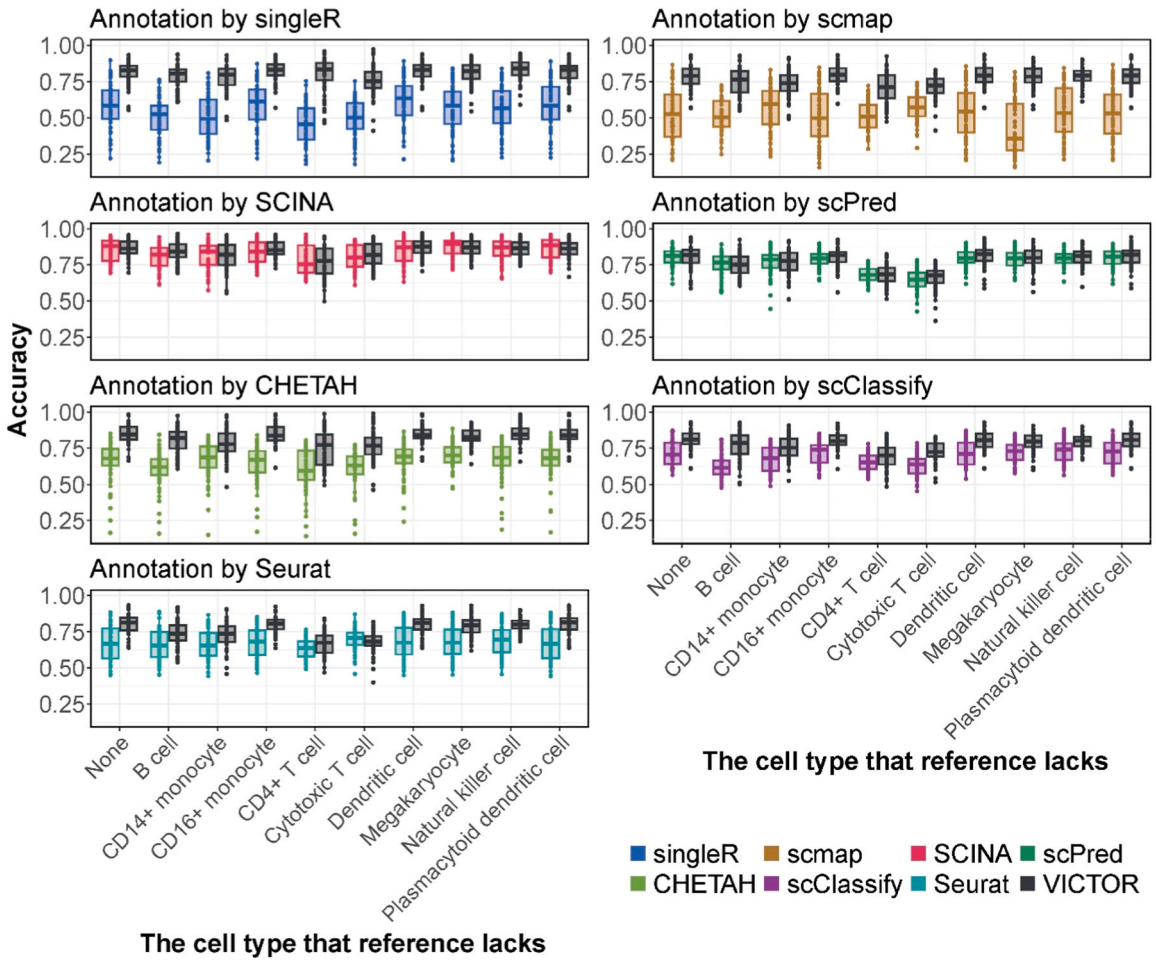

Fig. 4. Diagnostic performance in cross-platform annotations using PBMC datasets generated from different platforms.

improvements over SCINA and scPred. In the scenario where the reference included all cell types (the 'none' scenario), VICTOR achieved a median accuracy increase of 26.5 % compared to scmap, 24.6 % compared to singleR, 16.5 % compared to CHETAH, 14.9 % compared to Seurat, 10.4 % compared to scClassify, 0.5 % compared to scPred, but a slight decrease of 1.6 % compared to SCINA (Fig. 4).

In scenarios where the reference excluded specific cell types, VICTOR significantly enhanced the diagnostic performance of singleR, resulting in median accuracy increases ranging from 19.7 % to 37.7 %. The most notable improvement was observed when VICTOR raised the median accuracy of singleR annotations from 45.6 % to 83.4 % in the scenario lacking CD4 + T cells in the reference (Fig. 4). For instance, with the reference from inDrop and the query from 10xV2, singleR misclassified 465 CD4 + T cells as B cells, along with 14 cells of other types, leading to 479 erroneous diagnoses as reliable (false positives). VICTOR correctly identified 471 of these as unreliable (true negatives) (Fig. S5A). Similarly, VICTOR consistently outperformed scmap, with median accuracy increases ranging from 14.6 % to 42.8 % (Fig. 4). The most significant improvement occurred when megakaryocytes were absent from the reference, with median accuracy increasing from 35.7 % to 78.5 % (Fig. 4). For example, with the query from 10xV2 and the reference from Smart-seq2, scmap correctly identified 117 megakaryocytes but misdiagnosed them as unreliable (false negatives). VICTOR effectively recognized 113 of these as reliable (true positives) (Fig. S5B). VICTOR also showed slight improvements with SCINA and scPred. For SCINA annotations, the most significant increase was observed in the scenario lacking CD4 + T cells in the reference, with median accuracy rising from 75.5 % to 77.7 %. In this case, with the query from the 10xV2 platform and the reference from the SeqWell platform, SCINA correctly annotated 631 cytotoxic T cells but erroneously labeled 531 as unreliable (false negatives). VICTOR reduced the number of false negatives to 38 (Fig. S5C). For scPred annotations, the largest improvement was observed when cytotoxic T cells were absent from the reference, with median accuracy increasing from 64.7 % to 67.7 %. For instance, with the query from 10xV2 and the reference from CELSeq2, scPred misclassified 337 cytotoxic T cells as CD4 + T cells, 110 as natural killer cells, and 16 as B cells, resulting in a total of 463 erroneous diagnoses as reliable (false positives). VICTOR successfully reduced these false positives to 296 (Fig. S5D). VICTOR consistently outperformed CHETAH, with median accuracy improvements ranging from 9.1 % to 19.9 % (Fig. 4). The most striking enhancement occurred in the scenario where B cells were absent from the reference, with median accuracy increasing from 62.0 % to 81.9 % (Fig. 4). For instance, with the query dataset from 10xV3 and the reference from 10xV2, CHETAH misclassified 478 cytotoxic T cells as megakaryocytes and incorrectly identified them as reliable (false positives). VICTOR correctly identified all these as true positives (Fig. S5E). VICTOR demonstrated superior performance compared to scClassify, with median accuracy improvements ranging from 4.9 % to 17.2 % (Fig. 4). The most notable improvement was observed when B cells were absent from the reference, boosting median accuracy from 61.5 % to 78.7 % (Fig. 4). For example, with the query dataset from 10xV2 and the reference from 10xV3, scClassify incorrectly identified 177 B cells as dendritic cells and mistakenly classified them as reliable (false positives). VICTOR accurately recognized all these dendritic cells as unreliable (true negatives) (Fig. S5F). With Seurat annotations, VICTOR consistently outperformed except in the scenario where

cytotoxic T cells were absent from the reference, achieving median accuracy gains ranging from 3.6 % to 14.8 % (Fig. 4). The most significant improvement was observed when plasmacytoid dendritic cells were missing from the reference, with median accuracy increasing from 66.7 % to 81.5 % (Fig. 4). For instance, with the query dataset from 10xV2 and the reference from Smart-seq2, Seurat correctly annotated 457 CD4 + T cells but mistakenly labeled 425 of them as unreliable (false negatives). VICTOR impressively reduced the number of false negatives to just 8 (Fig. S5G).

2.5. *VICTOR improved diagnostic ability in cross-platform and cross-study annotations*

Moreover, we evaluated VICTOR's performance in a cross-platform and cross-study framework using pancreas datasets from multiple studies, including Baron [35] (inDrop), Muraro [36] (CEL-seq), and Segerstolpe [37] (Smart-seq2). Cross-platform and -study designs introduce biases and batch effects, further complicating automated annotation. We selected seven common cell types that are present in every pancreas dataset: acinar, alpha, beta, delta, ductal, endothelial, and epsilon cells. Similarly, we excluded individual cell types from the reference to simulate scenarios involving unknown cell types. This process resulted in eight scenarios, including one scenario in which no cell types were excluded ('none' in Fig. 5) and seven in which one cell type was omitted from the reference.

In the 'none' scenario where the reference included all cell types, VICTOR exhibited improved diagnostic performance overall all seven methods, with a median accuracy increase of 15.2 % compared to CHETAH, 13.4 % compared to singleR, 10.8 % compared to scmap, 8.3 % compared to scClassify, 3.5 % compared to Seurat, 1.8 % compared scPred, and 0.1 % compared to SCINA (Fig. 5).

When certain cell types were excluded from the reference, VICTOR markedly enhanced the diagnostic performance of singleR, resulting in median accuracy increases ranging from 8.6 % to 38.6 %. The most notable improvement occurred in the scenario where alpha cells were absent from the reference, with VICTOR boosting the median accuracy for singleR annotations from 54.8 % to 93.4 % (Fig. 5). For instance, with the query from Nanodrop and the reference from Smart-seq2, singleR misclassified 206 alpha cells as delta cells and 3 as epsilon cells, erroneously classifying all 209 as reliable (false positives). VICTOR successfully reduced the number of false positives to 0 (Fig. S6A). VICTOR also demonstrated superior performance to scmap, with median accuracy increases of 0.3 % to 21.9 %. The most significant improvement occurred in the scenario lacking alpha cells, where VICTOR increased the median accuracy from 73.7 % (scmap) to 95.6 % (Fig. 5). In this case, with the query from Nanodrop and the reference from Smart-seq2, scmap incorrectly labeled 243 out of 248 cells as reliable (false positives), including 241 alpha cells misclassified as delta cells, 1 as an epsilon cell, and 1 as a beta cell. VICTOR accurately identified 247 of these cells as unreliable (true negatives) (Fig. S6B). For SCINA annotations, VICTOR improved performance in most scenarios except when beta cells were missing. The most pronounced enhancement was in the scenario lacking alpha cells, where VICTOR increased the diagnostic accuracy from 87.4 % to 98.8 % (Fig. 5). Specifically, with the query from Nanodrop and the reference from Smart-seq2, SCINA misclassified 248 alpha cells as epsilon cells and wrongly deemed 129 of

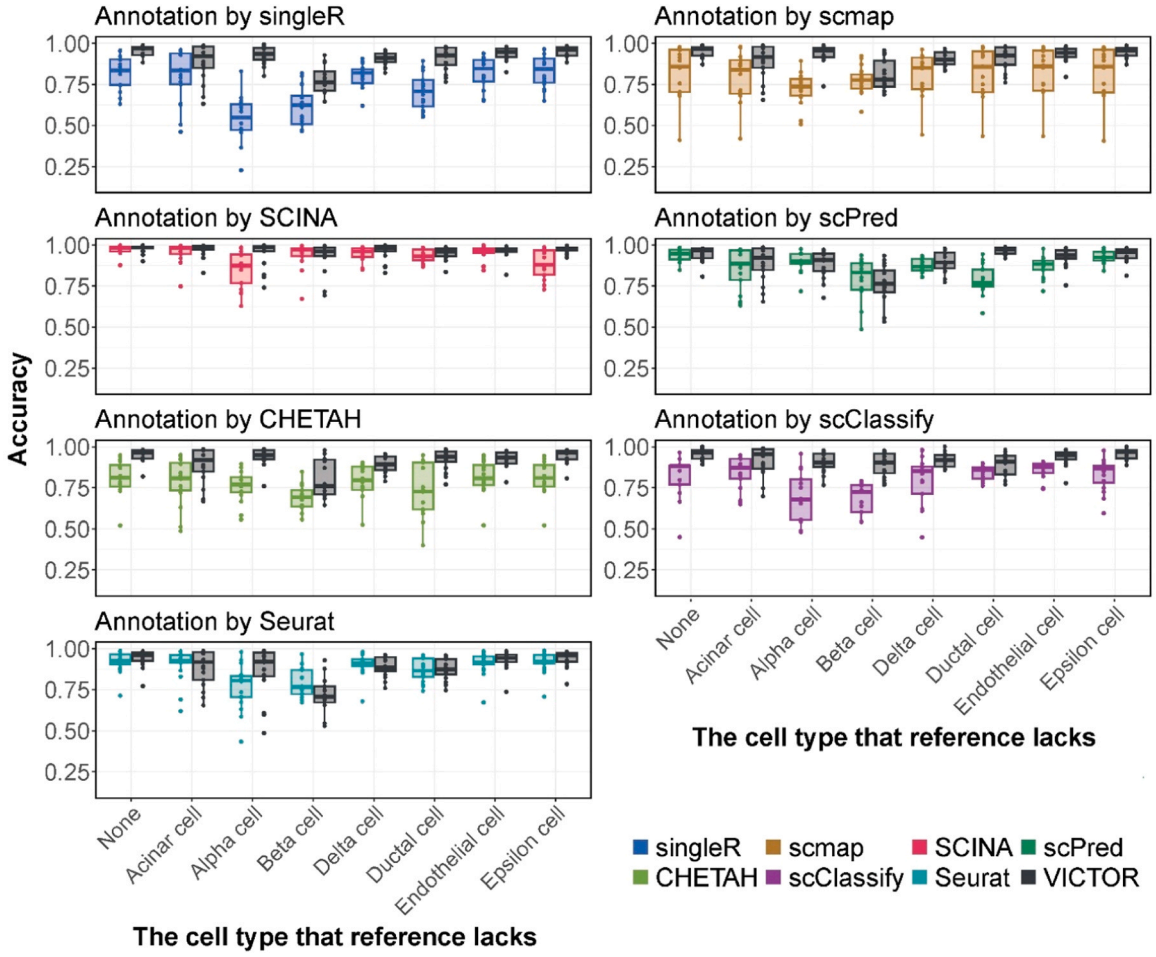

Fig. 5. Diagnostic performance in cross-platform and cross-study annotations using pancreas datasets from multiple studies.

these as reliable (false positives). VICTOR correctly identified all 248 alpha cells as unreliable (true negatives) (Fig. S6C). VICTOR also outperformed scPred in most scenarios, except when beta cells were missing from the reference. The most significant improvement was in the absence of ductal cells, with median accuracy rising from 76.8 % to 97.2 % (Fig. 5). For instance, with the query from Nanodrop and the reference from Smart-seq2, scPred misclassified 79 ductal cells as endothelial cells, erroneously marking all as reliable (false positives). VICTOR accurately identified all these cells as unreliable (true negatives) (Fig. S6D). VICTOR consistently surpassed CHETAH, with median accuracy improvements ranging from 6.7 % to 21.1 % (Fig. 5). The most notable enhancement occurred when ductal cells were excluded from the reference, increasing median accuracy from 72.7 % to 93.8 % (Fig. 5). For example, with the query from Nanodrop and the reference from Smart-seq2, CHETAH correctly identified 43 delta cells but misclassified 37 of them as unreliable (false negatives). VICTOR accurately recognized all 43 delta cells as reliable (true positives) (Fig. S6E). With scClassify annotations, VICTOR demonstrated superior performance, with median accuracy improvements ranging from 4.9 % to 22.6 % (Fig. 5). The most significant improvement occurred in the scenario lacking alpha cells, where median accuracy increased from 67.9 % to 90.5 % (Fig. 5). For instance, with the query from Nanodrop and the reference from inDrop, scClassify misclassified 203 alpha cells as beta cells, incorrectly marking 192 of these as reliable (false positives). VICTOR effectively reduced the number of false positives to just 1 (Fig. S6F). VICTOR consistently outperformed Seurat, except in scenarios where acinar, beta, or delta cells were missing (Fig. 5). The most substantial improvement was in the absence of alpha cells, with median accuracy rising from 80.4 % to 92.2 % (Fig. 5). For instance, with the query from Nanodrop and the reference from Smart-seq2, Seurat misidentified 204 alpha cells as delta cells, mistakenly classifying 165 of these as reliable (false positives). VICTOR successfully identified all as unreliable (true negatives) (Fig. S6G).

2.6. *VICTOR improved diagnostic ability in large-scale annotations*

To evaluate VICTOR's performance on large-scale datasets, we utilized an integrated cell atlas of the human lung in health and disease from the Human Cell Atlas (HCA HLCA core), comprising 584,944 cells [38]. This dataset includes 11 different studies from the 10X platform. We designed the queries by splitting each of the 11 studies into units of 20,000 cells, resulting in a total of 36 queries. Each query was then annotated against references formed by randomly selecting 5000 cells from the remaining 10 studies, excluding the study from which the query originated. Additionally, we excluded specific cell types from the references to simulate scenarios involving unknown cell types. Since the HLCA dataset includes 61 cell types, we randomly selected 13 cell types to be excluded, along with one scenario where no cell types were excluded ('none' in Fig. 6), resulting in a total of 14 scenarios.

VICTOR significantly outperformed singleR, scmap, SCINA, CHETAH, and Seurat across all scenarios, while demonstrating slight improvements over scClassify and scPred. In the 'none' scenario where the reference included all cell types, VICTOR exhibited improved diagnostic performance over all seven methods, with a median accuracy increase of 23.4 % compared to Seurat, 19.4 % compared to scmap, 13.3 % compared to SCINA, 11.8 % compared to CHETAH, 11.1 % compared to

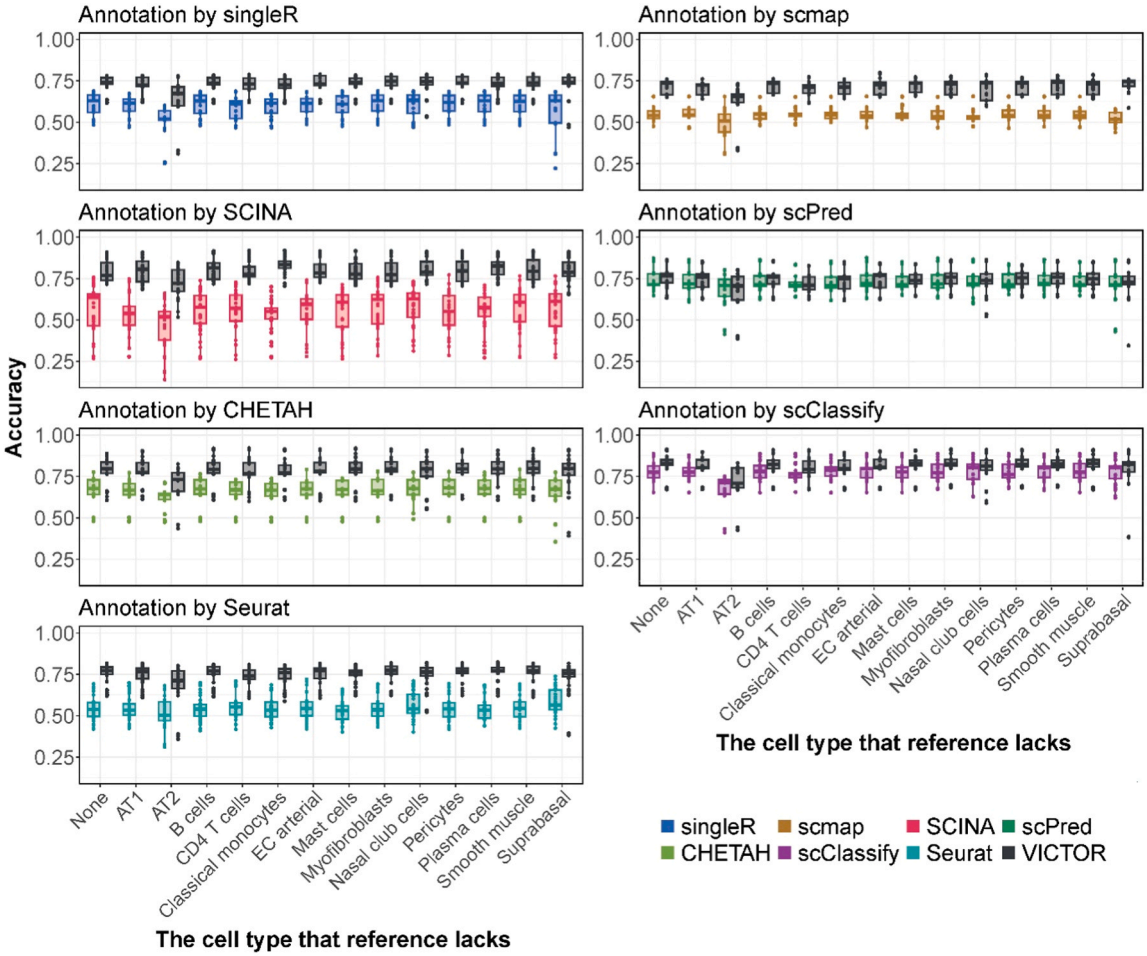

Fig. 6. Diagnostic performance in annotating large-scale datasets from the HLCA core.

singleR, 5.3 % compared to scClassify, and 5.0 % compared scPred (Fig. 6).

In scenarios where the reference lacked certain cell types, VICTOR consistently enhanced the diagnostic performance of singleR, resulting in median accuracy increases ranging from 10.5 % to 15.2 %. The most notable improvement occurred in the scenario where Alveolar Type II (AT2) cells were absent from the reference, with VICTOR increasing the median accuracy for singleR annotations from 52.0 % to 67.3 % (Fig. 6). In one example, 99 SMG mucous cells were mistakenly identified as Bronchial serous SMG cells and were incorrectly classified as reliable (FP) by singleR. VICTOR successfully identified all of these cells as unreliable (TN) (Fig. S7A). Moreover, VICTOR outperformed scmap, with median accuracy improvements ranging from 14.8 % to 22.8 %. The greatest improvement was observed in the scenario lacking suprabasal cells, where VICTOR increased the median accuracy from 51.7 % to 74.4 % (Fig. 6). For instance, although scmap correctly identified 606 mast cells, it erroneously diagnosed them as unreliable (FN). In contrast, VICTOR correctly identified 605 of these as reliable (TP) (Fig. S7B). Regarding SCINA annotations, VICTOR enhanced performance across all scenarios, with median accuracy increases ranging from 14.9 % to 28.1 %. The most pronounced improvement was observed in the scenario where classical monocytes were absent from the reference, with diagnostic accuracy increasing from 55.3 % to 83.4 % (Fig. 6). In one example, SCINA incorrectly categorized 21 alveolar Mph proliferating cells as interstitial Mph perivascular cells and 20 as T cells proliferating cells, all of which were mistakenly considered reliable (FP). In contrast, VICTOR accurately classified all 41 alveolar Mph proliferating cells as unreliable (TN) (Fig. S7C). Furthermore, VICTOR demonstrated higher median accuracy than scPred across most scenarios, except when the reference lacked AT2 cells. The most notable improvement occurred in the scenario where EC arterial cells were absent from the reference, with median accuracy increasing from 72.0 % to 76.9 % (Fig. 6). In one example, 377 nasal serous SMG cells were mislabeled as bronchial serous SMG cells, 373 of which were mistakenly considered reliable by scPred (FP). VICTOR successfully reduced the number of FPs to 231 (Fig. S7D). When cells were annotated by CHETAH, VICTOR consistently outperformed it, leading to median accuracy improvements ranging from 9.2 % to 12.9 % (Fig. 6). The most notable enhancement occurred in the scenario where myofibroblasts were excluded from the reference, resulting in a median accuracy increase from 6.6 % to 7.9 % (Fig. 6). For example, CHETAH incorrectly identified 212 smooth muscle cells as Pericytes, and mistakenly classifying 199 of them as reliable (FP). VICTOR reduced the number of FPs to 14 (Fig. S7E). When cells were annotated using scClassify, VICTOR demonstrated higher median accuracy across most scenarios except when the reference lacked AT2 cells. The most significant improvement was observed when pericytes were absent from the reference, resulting in a median accuracy increase from 76.1 % to 82.4 % (Fig. 6). For instance, although scClassify correctly identified 427 Non-basal club cells, it erroneously classified 288 of them as unreliable (FN). In contrast, VICTOR correctly recognized 425 of these as reliable (TP) (Fig. S7F). VICTOR consistently outperformed Seurat, with median accuracy improvements ranging from 18.8 % to 23.8 % (Fig. 6). The most substantial improvement occurred when plasma cells were missing from the reference, leading to a median accuracy increase from 53.3 % to 77.0 % (Fig. 6). For example, although Seurat correctly identified 92 smooth muscle cells, it erroneously classified 88 of them as unreliable (FN). In contrast, VICTOR correctly recognized 85 of these as reliable (TP) (Fig. S7G).

2.7. VICTOR improved diagnostic ability in cross-omics annotations

Finally, we explored VICTOR's ability in cross-omics annotations using a PBMC dataset, where both scRNA-seq and scATAC-seq profiles were generated for ~12,000 cells. By using scRNA-seq as the reference and scATAC-seq as the query, VICTOR demonstrated superior diagnostic ability compared to all seven methods (Fig. S8). Specifically, VICTOR significantly enhanced the performance of singleR, CHETAH, and Seurat, with accuracy increases of 39.3 %, 34.0 %, and 19.5 %, respectively. It also achieved slight improvements over SCINA, scmap, scClassify, and scPred, with accuracy increases of 6.5 %, 2.7 %, 1.9 %, and 0.2 %, respectively (Fig. S8). These results underscore VICTOR's effectiveness in improving annotation evaluation in cross-omics applications.

3. Discussion

Single-cell transcriptomics has become widely utilized for characterizing cellular heterogeneity and uncovering novel cell types and states in multicellular systems. Determination of cell identities is essential to effectively interpret single-cell data. Despite the development of numerous methods for automating cell annotation, assessing which annotations are truly correct versus those that are unreliable remains a significant challenge. Here, we presented VICTOR, an approach designed to assess the reliability of cell annotations through elastic-net regularized regression with optimized thresholds. VICTOR effectively quantified cell annotation reliability in a wide variety of scenarios as demonstrated by our study. VICTOR demonstrated superior identification of inaccurate labeling across within-platform, cross-platform, cross-study, and even cross-omics designs. VICTOR effectively pinpointed problematic annotations, thereby enhancing cell labeling. This advance holds significant promise to improve downstream analyses, such as marker gene identification and enrichment analysis, and improves the value of insights about cellular function and heterogeneity.

VICTOR employs elastic-net regularized logistic regression to build optimal models for individual cell types. Elastic net is the hybrid of ridge and lasso regularization, striking a balance between variable selection and model accuracy. Moreover, VICTOR enhances its efficacy by calibrating optimized thresholds tailored to each cell type. These cell-type-specific thresholds ensure both specificity and sensitivity, particularly for rare, unknown, or subtly distinct cell types.

VICTOR is a reference-based method, therefore its performance depends on the quality and the completeness of the reference utilized. Performance would be compromised if the reference is incomplete, contains inaccurate annotations, or poorly aligns with the query. Despite VICTOR's superior evaluation of cell annotation, the selection of a reference encompassing a broad spectrum of meticulously characterized cell types is paramount for labeling accuracies in single-cell data analysis. With single-cell data expanding exponentially, one potential solution is to construct a comprehensive reference atlas by aggregating multiple datasets from diverse platforms and conditions with standardized annotations. Cell types frequently undergo continuous change, therefore defining and characterizing them poses significant challenges [39]. Establishing a knowledge base of cell types to facilitate a thorough understanding of dynamic changes in cellular states would greatly enhance single-cell interpretation.

Although VICTOR is specifically designed as a tool for assessing the reliability of cell-type annotations, it can be extended to function as a cell-type annotation tool. The extension would involve training the model on the reference for each cell type, then scoring queried cells based on the trained model, and assigning them to the cell type with the highest score that passes specific thresholds for final annotation. Cells that do not meet any thresholds would be labeled as unassigned.

Currently, natural language processing (NLP) and large language models (LLM) offer advanced capabilities for cell type annotations [40–42]. However, these approaches require extensive computational resources. In contrast, VICTOR is much less computationally intensive, making it particularly effective for minor cell types and smaller datasets. It demonstrates robustness in handling limited observations, whereas NLP and LLM models often need substantial data to perform optimally and may suffer from performance issues, bias, or overfitting when certain cell types are underrepresented [42–44]. Furthermore, VICTOR can complement NLP- and LLM-based methods by preprocessing and

cleaning data before training those models, thereby enhancing performance and offering an assessment of annotation reliability generated by such models.

4. Materials and methods

4.1. VICTOR

VICTOR comprises two primary steps (Fig. 2). The first step involves training the model using the reference profile with labeled cells, encompassing data processing, the selection of informative PCs, elastic-net regularized logistic regression, and determination of cell type-specific optimal thresholds. The second step involves projecting the query profiles to the selected PCs and assessing the reliability of cell annotations predicted by any methods. For instance, when a query cell is annotated as cell type A, VICTOR applies the cell profile to the trained model specific to cell type A, thereby obtaining a reliability score. If the score surpasses the optimal threshold for cell type A, the annotation is deemed reliable; otherwise, it is considered unreliable.

4.2. Data preprocessing and feature Selection

VICTOR utilizes Seurat [17] for data preprocessing, which includes normalization, the selection of highly variable genes, and dimensional reduction. VICTOR normalizes data by the total count of each cell, and selects the top 2000 highly variable genes, followed by principal component analysis.

After preprocessing, VICTOR identifies the most informative principal components (PCs) to train the model. Following a strategy similar to scPred [24], VICTOR employs a two-tailed Wilcoxon rank-sum test on the top 50 PCs to select those with projected scores showing significant differences between cell types. VICTOR adjusts the p-values using the Benjamini-Hochberg false discovery rate correction (FDR) and selects PCs with an FDR < 0.01.

4.3. Model training using elastic-net regularized logistic regression

VICTOR trains one model for each cell type using elastic-net regularized logistic regression [45].

X represents the projected score matrix for the selected principal components (PCs), where each column denotes the score for each PC, and each row corresponds to a cell. Specifically, X_{ij} denotes the score of cell i projected to the PC j, where $i = 1,2\ldots n$ and $j = 1,2\ldots p$. Here, n is the number of cells and p is the number of selected PCs. Y_A is a binary vector, where $Y_{Ai} = 1$ if cell i belongs to cell type A, and 0 otherwise. The elastic net-regularized logistic regression model is formulated as below:

$$f_A(X) = \log\left(\frac{P(Y_{Ai}=1|X_i)}{1-P(Y_{Ai}=1|X_i)}\right) = \beta_0^A + \beta_1^A X_{i1} + \beta_2^A X_{i2} + \ldots + \beta_p^A X_{ip}$$

In this model, $P(Y_{Ai} = 1|X_i)$ is the probability that a given cell i belongs to cell type A, given its PC scores X_i. The model parameters β_0^A, β_1^A, …, β_p^A are estimated through the objective function for logistic regression by the penalized negative binomial log-likelihood:

$$\min_{(\beta_0^A,\beta^A)\in\mathbb{R}^{p+1}} -\left[\frac{1}{N}\sum_{i=1}^{N} Y_{Ai}\left(\beta_0^A + X_i^T\beta^A\right) - \log\left(1+e^{\left(\beta_0^A + X_i^T\beta^A\right)}\right)\right] + \lambda\sum_{j=0}^{p}\left[\frac{1}{2}(1-\alpha)\left(\beta_j^A\right)^2 + \alpha|\beta_j^A|\right]$$

where N is the total number of cells in the dataset. The balance between L1 and L2 penalties, denoted as α, is set to 0.5. The optimal regularization parameter λ is determined through cross-validation, employing the 'cv.glmnet' function in the 'glmnet' R package with default settings. This process involves computing a series of models across a predefined grid of λ values and selecting the λ that minimizes the cross-validated error.

4.4. Cell type-specific optimal thresholds

Instead of selecting one threshold for every cell type, VICTOR determines the optimal threshold for each cell type to maximize the sum of sensitivity and specificity, following Youden's J statistic [33].

$$Thr_A = \underset{\text{cutoff}}{\text{argmax}}\left(Sen_{A,cutoff} + Spe_{A,cutoff}\right)$$

Here, Thr_A represents the optimal threshold for cell type A, $Sen_{A,cutoff}$ denotes the sensitivity for cell type A at the cutoff, and $Spe_{A,cutoff}$ signifies the specificity for cell type A at the cutoff.

4.5. Assessment the reliability of cell annotations

After training the model for each cell type, VICTOR employs the trained model to evaluate the reliability of cell annotations predicted by any methods. VICTOR takes the gene expression profile and the predicted cell type of the query cell as input. Initially, VICTOR projects the query cell, denoted as *i'*, to the selected PCs, obtaining the projected score $X_{i'}$.

If cell *i'* is assigned to cell type A, VICTOR utilizes the trained model specific to cell type A to assess the reliability of this annotation.

$$f_A(X_{i'}) = P\left(Y_{A_i} = 1|X_{i'}\right) = \frac{1}{1-e^{-\left(\beta_0^A + \beta^A \bullet X_{i'}\right)}}$$

The reliability of the annotation of query cell *i'* belonging to cell type A is determined by comparing its predicted probability $f_A(X_{i'})$ against the optimal threshold Thr_A:

$$\begin{cases} reliable & \text{if } f_A(X_{i'}) \geq Thr_A \\ unreliable & \text{Otherwise} \end{cases}$$

4.6. Single-cell datasets

4.6.1. PBMC scRNA-seq datasets

The PBMC dataset comes from a comprehensive study aimed at systematically benchmarking various single-cell RNA-sequencing platforms [34]. We focused on seven platforms, including two low-throughput plate-based techniques (Smart-seq2 and CEL-Seq2) and five high-throughput methods (10x Chromium - v2, v3, Drop-seq, Seq-Well, and inDrops). Single-cell expression profiles were downloaded from GEO (GSE132044), and cell annotations were obtained from the Single Cell Portal.

4.6.2. Pancreas scRNA-seq datasets

The pancreas datasets were obtained from three studies generated by different platforms. These include Baron [35] (GSE84133) from the inDrop platform, Muraro [36] (GSE85241) from the CEL-seq platform, and Segerstolpe [37] (E-MTAB-5061) from the Smart-seq2 platform. Single-cell expression profiles and cell annotations were obtained from the scRNAseq package.

4.6.3. HCA HLCA core scRNA-seq datasets

The integrated Human Lung Cell Atlas (HLCA) is part of the Human Cell Atlas (HCA) initiative, representing the first large-scale, integrated single-cell reference atlas of the human lung [38]. The HLCA is divided into the HLCA core and the extended HLCA full. For this study, we used the HLCA core, which includes data from healthy lung tissue of 107 individuals, sourced from 14 different datasets, totaling 584,944 cells. This core dataset features manual cell type annotations based on consensus across six independent experts, as well as demographic, biological, and technical metadata.

4.6.4. PBMC multi-omics datasets

The PBMC multi-omics dataset is publicly available from 10x

Genomics, featuring approximately 12,000 human PBMCs that were granulocyte-depleted and sorted. In this dataset, both scRNA-seq and scATAC-seq profiles were simultaneously collected from the same cells, enabling comprehensive multi-omics analysis of gene expression and chromatin accessibility within individual cells.

4.7. Performance evaluation

4.7.1. Evaluation metric

We utilizes the accuracy to evaluate the performance of diagnosing cell type annotations.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Where true positives (TP) denote correct annotations that are further diagnosed as reliable, true negatives (TN) represent incorrect annotations that are further diagnosed as unreliable, false positives (FP) indicates incorrect annotations that are mistakenly diagnosed as reliable, and false negatives represents correct annotations that are erroneously diagnosed as unreliable.

4.8. Methods comparison

We compared VICTOR with seven commonly-used cell annotation methods: singleR[18], scPred[24], scmap[13], SCINA[14], CHETAH [15], scClassify [16], and Seurat [17]. These methods not only predict cell identities, but also determine the reliability of their predictions. singleR evaluates the prediction reliability using a diagnostic metric, designating the annotation as unreliable and the corresponding cell as "unknown" when the difference between its score for the assigned label and the median score across all labels is below a certain threshold. In contrast, scmap, SCINA, scPred, CHETAH, scClassify, and Seurat directly label annotations with predictive scores or probabilities below a certain threshold as “unreliable” and designate corresponding cells as "unknown.".

CRediT authorship contribution statement

Qi Liu: Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Chia-Jung Chang:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Chih-Yuan Hsu:** Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Yu Shyr:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The VICTOR Package is freely available at https://github.com/Charlene717/VICTOR. The curated and cell type annotated GSE132044 PBMC dataset is available at the Single Cell Portal (https://singlecell.broadinstitute.org/single_cell/study/SCP424). The curated and cell-type annotated Pancreas datasets, including GSE84133, GSE85241, and E-MTAB-5061, can be obtained from the scRNAseq package. The integrated Human Lung Cell Atlas, including the HLCA core, is publicly available through the CellxGene platform (https://cellxgene.cziscience.com/collections/6f6d381a-7701-4781-935c-db10d30de293). The PBMC multiomics dataset can be accessed and downloaded from the following URL: https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k. Additionally, the PBMC multiomics dataset can also be loaded using the SeuratData package, where it is available under the dataset names "pbmc.rna" for the RNA-seq data and "pbmc.atac" for the ATAC-seq data.

Acknowledgements

We would like to thank Bryan R. Helm for proofreading the manuscript and the anonymous reviewers for their valuable comments and suggestions, which have significantly improved our work. This work is supported by National Cancer Institute grants (U2C CA233291, P01CA229123 and U54 CA274367), National Institutes of Health (P01 AI139449), and Cancer Center Support Grant (P30CA068485). Additionally, Chia-Jung Chang is supported by a visiting scholarship (Overseas Project for Post Graduate Research, Grant ID: 112 - 2917 - I - 006 - 010) from the National Science and Technology Council, Taiwan. This scholarship facilitated her visit to the Department of Biostatistics at Vanderbilt University Medical Center.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.08.028.

References

[1] Werba G, et al. Single-cell RNA sequencing reveals the effects of chemotherapy on human pancreatic adenocarcinoma and its tumor microenvironment. Nat Commun 2023;14(1):797.
[2] Xu X, et al. Single-cell RNA sequencing to identify cellular heterogeneity and targets in cardiovascular diseases: from bench to bedside. Basic Res Cardiol 2023;118(1):7.
[3] Van de Sande B, et al. Applications of single-cell RNA sequencing in drug discovery and development. Nat Rev Drug Discov 2023:1–25.
[4] Jovic D, et al. Single-cell RNA sequencing technologies and applications: A brief overview. Clin Transl Med 2022;12(3):e694.
[5] Bukhari S, et al. Single-cell RNA sequencing reveals distinct T cell populations in immune-related adverse events of checkpoint inhibitors. Cell Rep Med 2023;4(1).
[6] Qiu C, et al. A single-cell time-lapse of mouse prenatal development from gastrula to birth. Nature 2024:1–10.
[7] Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat Commun 2022;13(1):1246.
[8] Shao X, et al. scCATCH: automatic annotation on cell types of clusters from single-cell RNA sequencing data. Iscience 2020;23(3).
[9] Franzén O, Björkegren JL. alona: a web server for single-cell RNA-seq analysis. Bioinformatics 2020;36(12):3910–2.
[10] Yang L, et al. Single-cell Mayo Map (scMayoMap): an easy-to-use tool for cell type annotation in single-cell RNA-sequencing data analysis. BMC Biol 2023;21(1):223.
[11] Cao Y, Wang X, Peng G. SCSA: a cell type annotation tool for single-cell RNA-seq data. Front Genet 2020;11:490.
[12] Ekiz HA, et al. CIPR: a web-based R/shiny app and R package to annotate cell clusters in single cell RNA sequencing experiments. BMC Bioinforma 2020;21(1):1–15.
[13] Kiselev VY, Yiu A, Hemberg M. scmap: projection of single-cell RNA-seq data across data sets. Nat Methods 2018;15(5):359–62.
[14] Zhang Z, et al. SCINA: a semi-supervised subtyping algorithm of single cells and bulk samples. Genes 2019;10(7):531.
[15] De Kanter JK, et al. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. Nucleic Acids Res 2019;47(16). p. e95-e95.
[16] Lin Y, et al. scClassify: sample size estimation and multiscale classification of cells using single and multiple reference. Mol Syst Biol 2020;16(6):e9389.
[17] Hao Y, et al. Integrated analysis of multimodal single-cell data. Cell 2021;184(13):3573–87. e29.
[18] Aran D, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol 2019;20(2):163–72.
[19] Hou R, Denisenko E, Forrest AR. scMatch: a single-cell gene expression profile annotation tool using reference datasets. Bioinformatics 2019;35(22):4688–95.
[20] Cortal A, et al. Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. Nat Biotechnol 2021;39(9):1095–102.
[21] Kang JB, et al. Efficient and precise single-cell reference atlas mapping with Symphony. Nat Commun 2021;12(1):5890.
[22] Zhang AW, et al. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. Nat Methods 2019;16(10):1007–15.
[23] Pliner HA, Shendure J, Trapnell C. Supervised classification enables rapid annotation of cell atlases. Nat Methods 2019;16(10):983–6.
[24] Alquicira-Hernandez J, et al. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome Biol 2019;20(1):1–17.

[25] Shao X, et al. 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;49(21). p. e122-e122.
[26] Chen J, et al. Transformer for one stop interpretable cell type annotation. Nat Commun 2023;14(1):223.
[27] Ma F, Pellegrini M. ACTINN: automated identification of cell types in single cell RNA sequencing. Bioinformatics 2020;36(2):533–8.
[28] Xu J, et al. CIForm as a Transformer-based model for cell-type annotation of large-scale single-cell RNA-seq data. Brief Bioinforma 2023:bbad195.
[29] Xie P, et al. SuperCT: a supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles. Nucleic Acids Res 2019;47 (8). p. e48-e48.
[30] Domínguez Conde C, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science 2022;376(6594)). p. eabl5197.
[31] Tan Y, Cahan P. SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species. Cell Syst 2019;9(2):207–13. e2.
[32] Mereu E, et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nat Biotechnol 2020;38(6):747–55.
[33] Youden WJ. Index for rating diagnostic tests. Cancer 1950;3(1):32–5.
[34] Ding J, et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat Biotechnol 2020;38(6):737–46.
[35] Baron M, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. Cell Syst 2016;3(4):346–60. e4.
[36] Muraro MJ, et al. A single-cell transcriptome atlas of the human pancreas. Cell Syst 2016;3(4):385–94. e3.
[37] Segerstolpe Å, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. Cell Metab 2016;24(4):593–607.
[38] Sikkema L, et al. An integrated cell atlas of the lung in health and disease. Nat Med 2023;29(6):1563–77.
[39] Zeng H. What is a cell type and how to define it? Cell 2022;185(15):2739–55.
[40] Cui H, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods 2024:1–11.
[41] Yang F, et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nat Mach Intell 2022;4(10):852–66.
[42] Szałata A, et al. Transformers in single-cell omics: a review and new perspectives. Nat Methods 2024;21(8):1430–43.
[43] Alsabbagh AR, et al. Foundation models meet imbalanced single-cell data when learning cell type annotations. bioRxiv 2023:2023. 10. 24.563625.
[44] Zhao, H., et al., Evaluating the Utilities of Large Language Models in Single-cell Data Analysis. 2023.
[45] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33(1):1.