Benchmarking single cell transcriptome matching methods for incremental growth of reference atlases

Joyce Hu[1,#], Beverly Peng[1,#], Ajith V. Pankajam[2], Bingfang Xu[2], Vikrant Anil Deshpande[3], Andreas Bueckle[3], Bruce W. Herr II[3], Katy Börner[3], Christopher Dupont[4], Richard H. Scheuermann[2,*], Yun Zhang[1,2,#,*]

Affiliations: [1]Department of Informatics, J. Craig Venter Institute, La Jolla, CA, USA. [2]Division of Intramural Research, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. [3]Department of Intelligent Systems Engineering, Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA. [4]Departments of Genomic Medicine, Human Health and Disease, and Human Microbiome, J. Craig Venter Institute, La Jolla, CA, USA.

[#]These authors contributed equally to this manuscript.

*Corresponding authors: Richard H. Scheuermann (richard.scheuermann@nih.gov), Yun Zhang (yun.zhang@nih.gov)

Abstract

Background: The advancement of single cell technologies has driven significant progress in constructing a multiscale, pan-organ Human Reference Atlas (HRA) for healthy human cells, though challenges remain in harmonizing cell types and unifying nomenclature. Multiple machine learning and artificial intelligence methods, including pre-trained and fine-tuned models on large-scale atlas data, are publicly available for the single cell community users to computationally annotate and match their cell clusters to the reference atlas.

Results: This study benchmarks four computational tools for cell type annotation and matching – Azimuth, CellTypist, scArches, and FR-Match – using two lung atlas datasets, the Human Lung Cell Atlas (HLCA) and the LungMAP single-cell reference (CellRef). Despite achieving high overall performance while comparing algorithmic cell type annotations to expert annotated data, variations in accuracy were observed, especially in annotating rare cell types, underlining the need for improved consistency across cell type prediction methods. The benchmarked methods were used to cross-compare and incrementally integrate 61 cell types from HLCA and 48 cell types from CellRef, resulting in a meta-atlas of 41 matched cell types, 20 HLCA-specific cell types, and 7 CellRef-specific cell types.

Conclusion: This study reveals complementing strengths of the benchmarked methods and presents a framework for incremental growth of the cell type inventory in the reference atlases, leading to 68 unique cell types in the meta-atlas across CellRef and HLCA. The benchmarking analysis contributes to improving the coverage and quality of HRA construction by assessing the reliability and performance of cell type annotation approaches for single cell transcriptomics datasets.

Keywords

Single cell transcriptomics, cell type, machine learning, meta-analysis, lung biomarkers, human reference atlas

Introduction

The advancement of single cell transcriptomic technologies has led to the rapid growth of single cell/nucleus RNA-seq (sc/snRNA-seq) datasets. Several large international consortia have made these data publicly available to accelerate the cellular characterization of complex tissues. For example, the CZ CellxGene Discover platform [1] (https://cellxgene.cziscience.com) is a data corpus of ~108 million unique cells from 1,765 datasets; the Human BioMolecular Atlas Program (HuBMAP) [2] data portal (https://portal.hubmapconsortium.org) contains 3,355 datasets for 31 organs; the Single Cell Portal [3] (https://singlecell.broadinstitute.org) features 820 studies of ~57 million cells. These efforts have contributed to the ongoing endeavors to build complete reference atlases for human tissues (e.g., the Human Cell Atlas (HCA) [4] and Human Reference Atlas (HRA) [5, 6]) at single cell resolution. One of the goals of these atlasing projects is to achieve a community-driven consensus of the diversity of cell phenotypes (cell types and cell states) in the human body. While there is a growing consensus on gene names and anatomical structures, there is less agreement on the nomenclature of cell phenotypes and their defining characteristics, making the harmonization of cell types between datasets, technologies, and atlases challenging.

Since the beginning of single cell biology, numerous computational methods have been developed for batch correction and data integration, towards building a comprehensive atlas. Batch correction methods, such as Harmony [7] and scVI [8], aim to eliminate experimental batch effects and technical noises while preserving true biological variation. Moreover, data integration methods, such as Seurat [9, 10] and LIGER [11], aim to integrate heterogeneous cells across datasets, modalities, and conditions to form an integrated data matrix with a shared latent space for downstream analysis. Luecken et al. [12] benchmarked 16 batch correction and data integration methods and revealed complementing properties of these methods. Combining the strengths of the integration methods, the first versions of single cell atlases were constructed by integrating data from selected high-quality studies using computational methods and re-assigning cell types in the integrated dataset. Due to the initial analyses in the individual studies involved varying levels of cell type annotations, most atlas projects would re-analyze the entire integrated dataset (i.e., principal component analysis, low-dimensional embedding, clustering, and annotation) to produce consistent and high-quality cell type annotations. Atlases that followed these strategies include the integrated Human Lung Cell Atlas (HLCA) v1.0 [13] (https://data.humancellatlas.org/hca-bio-networks/lung/atlases/lung-v1-0) from the HCA Lung Network, and the Human CellCards Multi-Study CellRef 1.0 Atlas [14] (https://www.lungmap.net/omics/?experiment_id=LMEX0000004396) from the LungMAP Consortium; both serve as reference catalogs of healthy human lung cell types for the research community. These data-driven reference atlases form the necessary foundation for building a trustworthy knowledgebase that covers all healthy cell types and generalizes to data from future studies.

Currently, there is no good strategy to update a reference atlas when new data become available or to merge two overlapping reference atlases. The current strategy that requires data integration and joint clustering is problematic for two reasons: (1) Each update will require a re-analysis of all cells, including the cells that are already in the atlas. Re-analysis will introduce slightly different grouping of cells in each update, so that studies which used an old version of the atlas will not be reproducible using an updated version, contradicting the sustainability and stability desired for the reference atlas and knowledgebase. (2) The joint clustering and manual annotation procedure is not efficient for the growing size of the atlas, as those computationally

joined clusters will need to be annotated again using marker genes and domain knowledge and require manual curation to identify rare cell types given larger datasets.

To support the sustainable growth of reference atlases and to construct an evolving cell type knowledgebase, we introduce a strategy for "incremental" growth of the cell type collection in the knowledgebase. The idea of the "incremental" growth is to compare cell clusters in the new data with cell types in the reference atlas so that common cell types will be merged, and novel cell types will be added during an update of the reference atlas and knowledgebase. This strategy would require computational methods to perform comparative pattern matching at the cell type/cluster level instead of searching for the nearest neighbors at the individual cell level. As an alternative to the cell integration methods, a cell type pattern matching method, such as FR-Match [15, 16], that not only matches the common cell types but also statistically identifies distinct or novel cell types between two datasets, will complement and potentially improve the long-term atlas building.

In this study, we present a framework to construct a meta-atlas of healthy human lung cell types using benchmarked computational methods for incremental cell type growth (**Figure 1**). We harmonize cell types identified in two highly-recognized human lung cell atlases – the Human Lung Cell Atlas (HLCA) [14] and the LungMAP single-cell reference (CellRef) [13]. Four computational methods, Azimuth [10] (https://azimuth.hubmapconsortium.org), CellTypist [17, 18] (https://www.celltypist.org), scArches [19] (https://docs.scarches.org), and FR-Match [15, 16] (https://github.com/JCVenterInstitute/FRmatch), were used to integrate the 61 cell types identified in the HLCA atlas with the 48 cell types identified in the CellRef atlas. To use Azimuth, CellTypist, and scArches, models pre-trained on the same reference (i.e., HLCA) are publicly available for users. With these three methods, one potential shortcoming is that the models do not explicitly recognize new cell types that are not in the reference during the pre-training process. To complement these pre-trained models, FR-Match based on the HLCA reference was also used for benchmarking the cell type matching performance with potential candidates for new cell types. All methods were used to match cell types from CellRef as the query dataset to HLCA as the reference dataset (i.e., CellRef → HLCA). In addition, FR-Match has more flexible matching options: (1) it implements two types of matching – cell-to-cluster (all methods) and cluster-to-cluster (FR-Match only); and (2) it can take any well-annotated dataset as the reference, i.e., allowing the reciprocal matching of HLCA as query and CellRef as reference (HLCA → CellRef). Combining these options, the two-way (CellRef ←→HLCA) cluster-to-cluster matching by FR-Match only was performed, in addition to the benchmarkable one-way (CellRef → HLCA) cell-to-cluster matching by all four methods. To construct a meta-atlas of lung cell types from the two datasets, the cell-level matching results were summarized into the cluster-level, and the final catalog of cell types (i.e., matched cell types, HLCA-specific cell types, and CellRef-specific cell types) were curated by examining all five matching results and their data visualizations (see **Results**).

Results

Characteristics of datasets and methods

The Human Lung Cell Atlas (HLCA) is an integrated cell atlas consisting of 2.4 million cells from 486 individuals spanning 49 datasets of the human respiratory system in health and disease [14]. A filtered subset of healthy cells forms the HLCA core dataset, which contains more than 0.5 million cells. Hereinafter, we will use "HLCA dataset" to mean the HLCA core dataset. The HLCA atlas was constructed by adopting a supervised data integration strategy that builds a

five-level hierarchical cell identity reference framework. At the finest level, the HLCA core defines 61 annotated cell types ("ann_finest_level") with well-distributed metadata of donors' demographics, anatomical, and technical variables, suggesting that any batch effects have been effectively normalized. Note that the HLCA includes nasal brushing in addition to lung structures.

The LungMAP single-cell reference (CellRef) is a reference atlas for both normal human and mouse lungs developed using a guided approach for atlas construction [13], where a pre-defined cell type dictionary (i.e., LungMAP CellCards - https://www.lungmap.net/research/cell-cards/) was used to guide the annotation of cell identities. The CellRef human dataset contains more than 0.3 million cells/nuclei consolidated from 10 datasets and 104 donors using a computational pipeline designed to account for both biological and technical variations. The atlas defines 48 cell types ("celltype_level3_fullname") in normal human lung. Key dataset characteristics are summarized in **Table 1**.

Table 1. Summary of healthy adult human lung scRNA-seq datasets used in this study.

Dataset	Year	Consortium	# cells	# genes	# cell types	Platform	Sample type
HLCA core	2023	Human Cell Atlas	584,944	28,024	61	10x 3' v1/v2/v3, 10x 5' v1/v2	Cells (in core)
CellRef human	2023	LungMap	347,970	32,284	48	10x 3' v2/v3, 10x 5' v1	Cells/ nuclei

In this study, we introduce the notions of unsupervised and supervised cell type matching, and cell-based and cluster-based matching for describing and categorizing the matching method properties. **Figure 1A** illustrates the unsupervised and supervised cell type matching scenarios. The setup of the cell type matching problem is to compare two datasets, namely a query dataset and a reference dataset. The reference dataset is an annotated dataset with the clusters labeled by cell type names (visualized as colored UMAP or tSNE plots). The query dataset is often not annotated (visualized as uncolored UMAP or tSNE plots). If the query dataset is not annotated, we refer to the matching from query to reference as unsupervised cell type matching. If the query dataset is annotated, we refer to the matching as the supervised cell type matching. One advantage of the supervised matching is that matching direction can be two-way, i.e., query → reference and reference → query. Supervised matching is useful when matching cell types between two well-annotated cell atlases, so that the cell type annotations in the original atlases can be preserved and assessed in whole in the cross-atlas matching results.

Figure 1B categorizes the selected methods used in this study based on their matching properties. Three pre-trained models were selected for comparison: Azimuth, CellTypist, and scArches. These three models have all been pre-trained on the HLCA reference. The pre-trained part of these models are the latent spaces and pre-tuned algorithms trained using the chosen reference, so that the input query data will be projected to the latent space and annotated to the reference cell types by the pre-tuned algorithm. Despite the different underlying algorithms, these methods all perform cell-based matching (i.e., assigning each query cell to a reference cell type cluster). These models require minimal user input as no tuning of the algorithm is needed and no clustering is needed to prepare the query data. The downside is that it is difficult for the user to assess the annotation quality of the query cells.

In this study, we benchmark these pre-trained models with each other and with an alternative category of method – FR-Match – with different matching design features. FR-Match allows for the choice of any user-defined reference atlas, making it more flexible than the other methods. Instead of a latent space, FR-Match utilizes a reduced feature marker gene space determined by the NS-Forest marker gene selection algorithm [20, 21], as a data-driven way to represent reference cell types in a reduced informative dimensional space. For the query dataset, FR-Match utilizes the upstream clustering results that group similar query cells together, making use of aggregated properties of similar cells. The FR-Match algorithm iteratively compares pairs of query and reference clusters with two options for matching: cell-to-cluster and cluster-to-cluster. FR-Match cell-to-cluster can be used to assign each query cell to a reference cell type cluster (i.e., cell-based matching), allowing refinement to the preliminary grouping of the query cells. When the query dataset is optimally clustered, FR-Match cluster-to-cluster can be used to perform two-way supervised matching between the query cell types and the reference cell types (i.e., cluster-based matching).

Using the selected methods and datasets, we present a lung reference meta-atlas after integrating the cell types between the HLCA and CellRef datasets and propose a framework for incrementally growing the meta-atlas by adding new cell types from future studies while preserving the knowledge and cell types in the original atlases (**Figure 1C**). Selected methods were benchmarked first in a cross-validation setting using the HLCA dataset. Then, all methods were used to perform cell type matching between HLCA and CellRef. All results are cross-compared at the cluster-level after the cell-based matching results were aggregated to the cluster-level. Discrepancies in the matching results were closely examined in the HLCA and CellRef UMAPs to resolve the discrepancies. Following this procedure, a collection of 68 cell types (41 matched cell types, 20 HLCA-specific cell types, and 7 CellRef-specific cell types) was collated into the lung reference meta-atlas.

Benchmarking of the cell-based methods on the HLCA dataset

The performance of Azimuth, CellTypist, scArches, and FR-Match cell-to-cluster matching methods were directly benchmarked on the HLCA reference. The benchmarking comparison was conducted in a cross-validation setting, where the HLCA cells were randomly split into 10 folds within each cluster. Each fold was then used as a query dataset to be assigned with the predicted annotations. After matching all folds, the predicted annotations of all cells were pooled and compared with the true annotations in the HLCA dataset (**Supplementary Figure 1**).

In the HLCA core dataset, the finest level of annotation consists of 61 cell types distributed among half a million cells, with the cell abundances varying depending on the cell type. For example, Alveolar macrophages is the most abundant cell type with 68,487 cells, and Lymphatic EC proliferating is the least abundant with only 28 cells (**Figure 2A**). Notably, 55% of cells are contained within the top 7 most abundant cell types, highlighting the large variation in cell abundances (cluster sizes) among the cell types. Potential effects of unequal cluster sizes when evaluating the matching performance are worth noting.

For benchmarking, each of the 10 folds of data was considered as the query set iteratively for predicting cell type annotations using the four computational methods. After pooling results for all cells by each method, the overall metrics in **Figure 2B(i)**, indicate that all four methods – Azimuth, CellTypist, scArches, and FR-Match – achieved similar performance across the average precision, recall, and F1-score, with all metrics above 0.8. This high performance suggests that all four methods can computationally annotate collections of cells at similar

performance levels. However, while the overall results can provide a good baseline for analysis, it is also important to examine whether the models are able to provide equally accurate predictions for each of the different clusters. When the results were evaluated per cluster (i.e. grouped by each true HLCA cell type), the performance metrics showed varying distribution among the methods in **Figure 2B(ii)**. In contrast to the consistently high overall F1-scores, the median F1-score when calculated per cluster for Azimuth (0.815), CellTypist (0.758), and scArches (0.841) were much lower than FR-Match (0.952); similar patterns were also observed in precision and recall. The box plots of the per-cluster metrics reveal that Azimuth, CellTypist, and scArches have lower median cluster-level scores and wider ranges, which suggests more inconsistency in the cluster-level performance. In particular, CellTypist had several per-cluster F1-scores of 0 but an overall cell-based F1-score of 0.859. The outlier clusters from all four methods demonstrate that while the overall F1-score is high, certain clusters are low performing.

The variation of the cluster-level performance is further illustrated by the bar plots in **Figure 2C** which depict the proportion of correct-to-incorrect annotations made for each cell type, ordered by cluster size. FR-Match achieved the highest average accuracy (96.7%), followed by scArches (71.7%), Azimuth (69.3%), and CellTypist (60.9%). Looking at the cluster-level results, Azimuth, CellTypist, and scArches follow similar patterns where larger clusters tended to have higher proportions of correct annotations and smaller clusters tended to have higher proportion of incorrect annotations. In contrast, FR-Match follows a reverse trend where larger clusters tended to have slightly higher proportion of incorrect annotations than smaller clusters, while the ratio of correct annotations for smaller clusters were remarkably improved compared to the other three methods. The relationship observed suggests that the discrepancy between cell-level and cluster-level results in **Figure 2B** may be attributed to the high performance of larger clusters masking the low performance of smaller clusters. Therefore, when selecting the most appropriate metrics to evaluate the cell type annotation methods, it is important to not only evaluate overall performance but also to consider the method's effectiveness in accurately identifying rare cell types.

Furthermore, utilizing the confidence scores produced by each of the methods, receiver operating characteristic (ROC) analysis was conducted in this cross-validation setting. Each of the four methods employs their own methodology for producing a confidence score to quantify a prediction's certainty. Details about these methodologies are provided in the **Methods** section. The underlying assumption of a confidence score is that a method would assign high confidence to the correct predictions and low confidence to the incorrect ones, which can be used to construct ROC curves to assess the trade-off between the rate of true positive and false positive prediction pairs at various confidence score thresholds. FR-Match exhibits the highest performance with an area under the curve (AUC) score of 0.934, followed by Azimuth (AUC = 0.905), scArches (AUC = 0.877), and CellTypist (AUC = 0.799) (**Figure 2D**), consistent with the observations in **Figure 2B-C**.

To understand further the ROC results, colorized ROC curves are shown in **Supplementary Figure 2** and heatmaps of confidence scores are shown in **Supplementary Figure 3**. The colorized ROC curves (**Supplementary Figure 2**), shaded by confidence score, help visualize the relationship between the prediction confidence with the true positive and false positive rates. In an ideal case, the colorized plot would display a 1 to 0 score gradient as the curve transitions from the start point, i.e., (0, 0) to the end point, i.e., (1,1), which would indicate high confidence predictions associated with correct predictions, i.e., true positives (red-orange portion of the curve) and medium to low confidence predictions associated with incorrect predictions, i.e., false positives (cyan-blue portion of the curve). The heatmaps (**Supplementary Figure 3**)

further elaborate on the results of the colorized ROC curves by depicting the mean confidence for each true label and predicted label pairing generated by the corresponding method. A perfect heatmap would display a clean distinct diagonal line, signifying a perfect alignment between the prediction and true labels at a high confidence level (dark color). Darker squares in the heatmap correspond to the red portion of the colorized ROC curve; and the lighter regions correspond to the blue portion. This visualization illustrates how cells from specific clusters that were mislabeled with high confidence would influence the shape of the ROC curves and the quality of predictions output by each method.

For example, FR-Match, which has the highest AUC score, has a nearly clean diagonal in its heatmap. There are a few dark squares just off the diagonal, which means that high confidences were assigned to similar cell types but not exactly the true label (note that clusters in the heatmap are ordered by their similarities to each other based on the HLCA taxonomy dendrogram; see **Methods**). Azimuth, CellTypist, and scArches have noisier heatmaps and lower AUC scores. In particular, the CellTypist heatmap has more dark squares scattered off the diagonal, which corresponds to the shallower slope in its ROC curve.

To conclude this cross-validation evaluation, by iteratively validating each of the 10 folds of the HLCA dataset, we found that 1) although all methods achieved high overall cell-based performance metrics values, the cluster-level performance varied from method to method, which can be obscured by the imbalance in cluster sizes in the overall evaluation; 2) the ROC analysis demonstrated that method performance was driven by the confidence scores calculated in each method; and 3) the confidence score heatmaps suggest a strong need to eliminate false negatives (i.e., off-diagonal high confidence scores) in the score calculation.

Benchmarking of the cell-based methods across datasets

After cross-validation on the HLCA dataset, the four methods were used for cell-based matching of the CellRef dataset as query to the HLCA reference (CellRef → HLCA). Each query cell was assigned to a reference cell type or “unassigned” (see **Methods**), resulting in a result table of cell-level predictions for the 347,970 CellRef cells with corresponding confidence scores for each method. The combined results table for all four cell-based methods is in **Supplementary Table 1**.

As noted, cluster sizes had an impact on the matching results. In CellRef, there are 48 cell types, whose cluster sizes are shown in **Figure 3A**. The most abundant cell type is alveolar macrophage with 101,380 cells and the least abundant cell type is chondrocyte with only 6 cells. It is also worth noting that despite many synergies between the HLCA and CellRef efforts, a major issue that arose when attempting to compare cell types in both datasets was the inconsistent naming conventions. As the respective datasets use differing nomenclatures to annotate cell types, after using the computational methods to match the CellRef cells to HLCA, many of the predicted HLCA annotations did not have a direct match with the CellRef original labels.

Unlike the cross-validation evaluation, instead of validating the predicted HLCA labels with the original CellRef labels, each cell’s prediction by one method is cross-compared with its corresponding predictions from the other three methods. An agreement score of 4 is given to a cell where all four methods gave the same prediction, 3 where three methods agreed upon the prediction, and so forth. In **Figure 3B**, 76.7% of cell predictions were agreed upon by all four methods, and an additional 14.7% showed a majority 3 out of 4 agreement. Interestingly, 8.4%

of annotations were agreed between only two methods, and only 0.2% had no agreement at all. While agreement among methods does not necessarily confirm the accuracy of the predictions, the high level of agreement across methods suggests that the predictions for most cells were similar.

As observed above, it would be informative to look at the prediction agreement at the cluster-level. As expected, when looking at the agreement levels for individual clusters, there was a trend where larger clusters had higher agreement, and smaller clusters had lower agreement (**Figure 3C**). In this figure, cells are grouped by their original CellRef labels. Looking at the most abundant cluster, 91.56% of alveolar macrophage cells' predictions had majority agreement among all four methods (score of 4 or 3). As this cluster consists of 29.13% of the dataset, alveolar macrophage and other large, high agreement clusters may have biased the overall cell-level results. Similarly, all four methods assigned the same predictions for almost all cells in abundant alveolar type 2 cell and alveolar type 1 cell clusters (98.38% and 97.51% of cells had score of 4, respectively). In contrast, there was no majority agreement (score of 4 or 3) in clusters such as megakaryocyte/platelet and chondrocyte. Although the only 0.2% of cell-level predictions that had no method agreement (score of 1) is a low number, most of those predictions are found within the megakaryocyte/platelet and chondrocyte clusters (containing 41 and 6 cells respectively). Both clusters, although small, have high proportions of little-to-no method agreement (score of 2 or 1) for its predictions (97.56% and 100% respectively), which further points out the difficulties in accurately predicting rare cell types.

Assessing the pairwise agreement between methods revealed additional insights into the performance for individual methods (**Figure 3D**). Overall, a general trend between cluster size and prediction agreement was observed for each pair, where the larger clusters tended to have more agreements. While there was a higher proportion of disagreement (red) towards the right of these plots, there was also low agreement for some moderate sized clusters. Some disagreement could be attributed to certain methods. For example, Azimuth showed a high disagreement with all other methods about predictions for respiratory airway secretory cell and secretory cell; in contrast, the other method pairs have high agreement on the prediction for these two cell types. Thus, there is no agreement score of 4 in **Figure 3C** for these two types. While we cannot tell from these plots whether Azimuth correctly predicted for these two cell types while other methods were incorrect, or vice versa, this example demonstrates how even among similarly performing algorithms, methods may still vary in their ability to predict specific cell types, regardless of how well the clusters are represented in the dataset. In the case of basal cell and intestinal macrophage types, which have sufficient cells (12,989 and 6,437 cells respectively), relatively low and variable agreement was observed between method pairs, especially in the pairs of FR-Match and others, resulting in an agreement score of 4 with only 17.92% and 7.49% of cells (**Figure 3C**). The matching results of basal cell and intestinal macrophage types will be further explored in the next section. Interestingly, the highest overall agreement achieved in this pairwise comparison are the pair of Azimuth and scArches (0.745), and the pair of CellTypist and FR-Match (0.745), suggesting some similarity in these pairs of methods.

Cell-based matching results of the lung cell atlases

Cell-based matching of CellRef as query and HLCA as reference results in each CellRef cell being assigned with an HLCA cell type label prediction. **Figure 4A** shows UMAP plots in which CellRef cells in the CellRef UMAP coordinates are colored by the HLCA label predictions by each method. Overall, Azimuth predicted 54 cell types out of the 61 HLCA labels; CellTypist

predicted 47; scArches predicted 61; and FR-Match predicted 49. While the original CellRef contains 48 cell type labels, the prediction to all 61 HLCA labels by scArches may suggest that the scArches model is overfitted as it was trained by the HLCA team. While a global evaluation of these UMAPs (i)-(iv) suggest that the matching results are very similar (**Figure 4A**), since 76.7% of cells agreed on the predictions by all four methods (**Figure 3B**), the differences between the methods require more careful examinations.

Two regions where major differences were observed are highlighted in **Figure 4A** and shown in detail in **Figure 4B-C**, where the original CellRef labels and the predictions are compared. One region is the interstitial macrophage ("IM") region in CellRef (**Figure 4B**), where the well-segregated IM cluster is matched to a mixture of HLCA cell types by Azimuth and scArches, two HLCA cell types split into roughly equal sizes by CellTypist, and one HLCA cell type (interstitial Mph perivascular) uniformly colored by FR-Match. The FR-Match result reflects the algorithm's property that it borrows information among the clustered query cells, so that its matching results are more likely to preserve the shape of a well-segregated query cluster. The CellTypist result here is also interesting, where the beige color in the predictions is the monocyte-derived Mph cell type in HLCA, suggesting a transitional state of cell types from the inflammatory monocyte (iMON) to interstitial macrophage (IM) in CellRef.

The other region showing major difference is the "basal" region in CellRef (**Figure 4C**), where there is a separation of suprabasal and basal in the CellRef original labels, but spillover of suprabasal (cyan color) into the "basal" region (yellow color) in the predictions. The FR-Match result had the least spillover of the cyan color because those spillover cells were likely of low confidence, thus were "unassigned" in FR-Match (**Supplementary Figure 4 and 5**). All predicted labels as well as the original CellRef and HLCA labels for the "IM" and "basal" regions are in **Supplementary Figure 4**, where the "unassigned" cells in FR-Match are colored white, implying a clean-up of the low confidence cells.

In **Supplementary Figure 5A**, cells in the UMAP plots are colored by their confidence scores by each method. While most of the cells had high confidence scores in each plot, all methods produced weakly matched cells spread over many clusters. FR-Match produced some regions packed with very low confidence cells (i.e., dark islands in the plot), including the spillover "basal" region, where many of the spillover cells are "unassigned" (**Supplementary Figure 5B**). Labeling the spillover cells as "unassigned", reflecting the low matching confidence, is more informative than a fuzzy matching to a similar label, e.g., suprabasal. Another "unassigned" pattern in the FR-Match plot (**Supplementary Figure 5B**) is highlighted by the black arrows where the "unassigned" cells form an entire cluster that is isolated from others, suggesting CellRef cell types that are not in the HLCA reference. Lastly, due to a different version of HLCA annotations being used in the Azimuth training model, the Azimuth-only labels had decent confidence scores (**Supplementary Figure 5A-B**), suggesting an update is needed for the Azimuth lung model to be consistent with the cell type annotation in the HLCA literature.

Cluster-level matching of cell types based on the cell-based matching results

As previously shown, the cluster-level pairwise prediction agreement varied a lot among clusters (**Figure 3D**). By aggregating the cell-level predictions per query cluster, **Figure 5** shows the proportion of CellRef cells matched between a pair of the original CellRef label (horizontal axis) and the predicted HLCA label (vertical axis) for each method. The HLCA labels are ordered by the hierarchical clustering such that similar clusters are grouped next to each other according to their gene expression profiles (**Methods**). The CellRef labels are ordered accordingly so that

the highest matched proportions would ideally align along the diagonal. FR-Match showed the cleanest aggregated cluster-level matching pattern (**Figure 5**), with the lowest off-diagonal proportions. Two examples are highlighted to reveal the different matching patterns among the methods. In the highlighted box (A), the three dendritic cell (DC) types of CellRef are matched differently by all methods. Azimuth matched all three types to DC2 in HLCA; CellTypist clearly matched classical DC subset 2 to DC2 but not the other two types; scArches showed relatively smaller proportions of matched cells, with some matched to other DC types and some matched to non-DC types; and FR-Match identified unique one-to-one matches for each of the three types.

The highlighted box (B) illustrates the matching pattern for a "unassigned" cell type by FR-Match. For the megakaryoctye/platelet type in CellRef, it was matched poorly to many types in HLCA by Azimuth and scArches and was matched to the mast cells in HLCA by CellTypist. Only FR-Match determined that those poorly matched cells should be "unassigned". The disagreement among Azimuth, CellTypist and scArches also indicates that there is no consistent match to the megakaryoctye/platelet type and should be "unassigned". These examples suggest that further refinements are needed to fine tune the pre-trained models of Azimuth, CellTypist, and scArches for more accurate matching (e.g., for the DC types) and to address the scenario where weakly matched cells are forced to match to some cell type in the reference.

These benchmarking results further reveal considerations and challenges for meaningfully integrating cell types in the atlases using the cell-based matching methods. First, three out of the four methods tend to introduce quite a few off-diagonal matching even though the query cluster is optimally clustered, in this case, the well-annotated CellRef cell types. Only FR-Match preserves most of the optimally-clustered cell types in CellRef. Preserving the existing cell type membership in well-annotated reference atlases is important for the systematic regular update of a reference atlas. Second, the ability to identify not matched cells (i.e., "unassigned") is important for incorporating new cell types into the reference atlas; however, this feature is missing in the benchmarked methods other than FR-Match. Currently, novel cell types were commonly identified by re-annotating the jointly clustered data after integrating all the cells when constructing an updated reference atlas. Instead, a mechanism to identify novel cell types, annotate them using human expertise, then add them to an updated reference atlas, is desired.

FR-Match cluster-to-cluster matching results of the lung cell atlases

FR-Match can perform cluster-to-cluster matching, which allows two-way matching (i.e., from CellRef as query to HLCA as reference and vice versa, denoted as CellRef ←→ HLCA), by utilizing the cell type annotations in both datasets. The two-way matching results are shown in **Figure 6**, and each of the one-way matching results are shown in **Supplementary Figure 6**. The two-way matches indicate that the pair of the matched CellRef and HLCA clusters are found by using both the CellRef as reference and the HLCA as reference. In **Figure 6**, the two-way matches found by the FR-Match cluster-to-cluster matching strategy exhibit a very clean matching pattern, where most of the two-way matches are one-to-one matches and only a few are 1-to-2 or 2-to-1 matches. The cluster-to-cluster results are easier to interpret and not impacted by the varying cluster sizes as each cluster is assessed as a whole in the cluster-based method. Although it is possible to have one-to-many or many-to-one matches, there are fewer of these in the cluster-based results than in the cell-based results. It is worth noting that the "unassigned" row at the bottom also shows two-way matches, suggesting that these cell types are only found in the CellRef atlas. Similarly, the empty rows in the figure indicate the cell types that are only found in the HLCA atlas.

The cluster-to-cluster matching strategy, in addition to the cell-based matching strategies, provides the foundation for incrementally introducing new cell types into the reference atlases while preserving the existing cell types in the atlases, which is a more sustainable approach for the regular update of reference atlases as new datasets become available.

Recommended matching for the lung cell type meta-atlas across CellRef and HLCA

For the lung reference atlases, a comprehensive meta-comparison of all the benchmarked cell-based matching results (**Figure 5**) and the cluster-based FR-Match results (**Figure 6**) was conducted to construct a meta-atlas of lung cell types across CellRef and HLCA. For the cell-based results, a computational strategy was applied to determine high confidence matches based on the matching proportions (**Methods**), so that all cell-based results could be uniformly processed into a binary "match" (Y) or "no match" (N) format between the cell types. Then, the two-way cluster-to-cluster matches from FR-Match were combined with the high confidence matches from the cell-based matching results to provide evidence for determining the recommended matching in the meta-atlas. **Figure 7** shows the cross-comparison table for the recommended matching of the cell types between CellRef and HLCA. Candidates of "HLCA2CellRef_Match" are suggested by the agreement of the cluster-based matching method and at least one cell-based matching method; the remaining are candidates for atlas-specific cell types. All candidate pairs and any match to the atlas-specific candidates by a single method were carefully examined by looking at their marker gene expression patterns using the marker gene selection algorithm NS-Forest [20, 21] (**Methods**) before determining the final matching recommendation. 29 matches were identified by all methods with all Y's and 16 HLCA-specific types were identified by all methods with all N's. The remaining were determined by majority voting and examination on the marker gene expression pattern.

Figure 8 shows the use of NS-Forest marker genes to disambiguate potential matches. The NS-Forest marker gene combinations form unique "barcode" (i.e., on-and-off) patterns (see next section) to characterize different cell types in the single cell atlas. **Figure 8A** are the "barcode" plots of the HLCA AT0 and pre-TB secretory types and the CellRef respiratory airway secretory cell and secretory cell types. The four cell types form three pairs of two-way matches in the cluster-based matching results, and both CellRef types were matched to the HLCA pre-TB secretory type by 3 out of 4 cell-based methods. Using the "barcode" plots, the expression combination of the three marker genes – *SCGB3A2*, *SCGB1A1*, and *SFTPB* can clearly delineate the recommended matching pair – HLCA pre-TB secretory and CellRef respiratory airway secretory cell – from the other two types. The matched pair have high expression in all three markers (i.e., $SCGB3A2^{high}SCGB1A1^{high}SFTPB^{high}$), whereas the HLCA AT0 is $SCGB3A2^{high}SCGB1A1^{low}SFTPB^{high}$, and the CellRef secretory cell is $SCGB3A2^{low}SCGB1A1^{high}SFTPB^{low}$ in the corresponding datasets. Another example is the CellRef regulatory T cells, which is recommended as a CellRef-specific type, despite the potential matching to the HLCA CD4 T cells by all methods. Although these are very similar cell types, markers like *FOXP3* and *CTLA4* can distinguish the T regulatory cells (**Figure 8B**). A special case of "HLCA2CellRef_broadMatch" is recommended between the HLCA club (non-nasal) and CellRef goblet cell types, because the CellRef goblet cell type is likely under-partitioned, containing both club cells expressing *SCGB1A1*, the secretoglobin 1A1 (also known as club cell secretory protein) [22], and goblet cells expressing the well-known marker *MUC5AC* [23] (**Figure 8C**). In these two atlases, there are three location-specific goblet types (nasal, bronchial, and subsegmental) in HLCA, thus none of which were matched to the under-partitioned goblet cell type in CellRef. There is no club cell annotated in CellRef, which is likely grouped in the goblet cluster in the CellRef atlas. It is also worth noting that none of the four

“nasal” cell types in HLCA are matched to anything in CellRef in this meta-comparison (**Figure 7**), which is consistent with the fact that nasal specimens were not included in the CellRef experiments. However, that was not the case in the individual matching results of Azimuth, CellTypist, and scArches (**Figure 5**), where nasal cell types were weakly matched to similar cell types in different tissue locations. In contrast, cell-based and cluster-based FR-Match approaches did not match “nasal” cell types (**Figure 5-6**).

In summary, out of the 61 HLCA cell types and 48 CellRef cell types, the meta-comparison recommends 41 matched cell types, 20 HLCA-specific cell types, and 7 CellRef-specific cell types. Taken these results together, it suggests that an integrated version of these two atlases – a meta-atlas – would serve as a higher-quality and more comprehensive reference for normal human lung cell types. This workflow demonstrates a strategy that supports for the incremental update and growth of reference atlases.

Cell type “barcode” of marker gene expression

The notion of a gene expression “barcode” for characterizing cell types from scRNA-seq data was independently proposed by two groups [15, 16, 20, 21, 24, 25]. “Barcode” is used to describe the on-and-off modes of cell type-specific marker genes (i.e., “on” means a gene is highly expressed and “off” means a gene has little-to-no expression). Heuristically, the combinations of on-and-off modes of all marker genes identified for all clusters in a given experiment form unique “barcode”-like patterns for distinguishing different cell types. Grabski et al. used the gene expression “barcode” for unsupervised clustering with scRNA-seq data [24, 25]. Zhang et al. used the cell type “barcode” produced by NS-Forest marker genes [20, 21] to perform cell type matching in FR-Match [15, 16]. In the previous section, we showed the use of NS-Forest cell type “barcode” plots to disambiguate potential matches or non-matches. The “barcode” plots also showed convincing evidence of highly similar expression patterns for the cell type pairs of recommended matching in the reciprocal atlases (**Figure 9**). The NS-Forest v4 algorithm was applied to each dataset, which selected 115 unique marker genes for the 61 HLCA cell types and 89 unique marker genes for the 48 CellRef cell types. In each sub-figure, gene expression of 30 randomly selected cells from each of the matched cell types are plotted in the HLCA marker gene space (left two plots) and the CellRef marker gene space (right two plots). Though different marker gene combinations were selected for HLCA and CellRef due to their different composition of cell types, the “barcode” plots of matched cell types across the two atlases showed very similar expression patterns in both marker gene spaces. The “barcode”-like patterns of NS-Forest marker genes are robust across datasets, which also supports for the pairs in the recommended matching being truly matched cell types across atlases based on their gene expression patterns in each dataset. The NS-Forest “barcode” plots for all recommended matching pairs and atlas-specific cell types are in **Supplementary Figure 7-11**. The recommended matching results with NS-Forest marker genes are in **Supplementary Table 2**.

Discussion

In this study, cell type matching performance of four methods – Azimuth, CellTypist, scArches, and FR-Match – were systematically benchmarked using two well-annotated scRNA-seq human lung atlas datasets serving as the ground truth. The results highlight the strengths and limitations of current cell type matching approaches when applied to large-scale datasets. (1) While all four methods demonstrated high accuracy in the overall results, particularly in well-represented cell types, significant variation was observed in their ability to accurately predict

rare cell types. The discrepancy between overall performance and cluster-specific results suggests that many of the current approaches are biased towards larger and more common cell type populations. (2) The impact of imbalances in cluster sizes among these data is clearly elaborated in this study, as well as Maan et al. (2023) [26]. While many of the current tools provide the user community with convenient cell-based predictions, more evidence has emerged that suggests that the cell-based approaches may produce misleading results if cluster-level variation is not adequately taken into account. (3) One common limitation to many of the methods is the lack of mechanism to identify novel cell types. With the goal of completing a reference atlas that contains all healthy cell types in the human body, the ability to recognize novel cell types not identified in initial atlases will be essential as new datasets are evaluated. Among the benchmarked methods, only FR-Match has a mechanism to determine “unassigned” novel cell types, while all other methods seem to insufficiently address weakly matched cells in their algorithms. These limitations highlight the importance of future efforts to further refine matching algorithms to better capture rare and novel cell types.

Earlier in this project, some attempts were made to manually match the CellRef annotated cell type labels with their HLCA counterparts, and vice versa. Even with domain knowledge, these attempts failed to find one-to-one matches for several cell types. This issue is not unique to lung. In fact, this portion of the study reveals a common struggle in current computational cell type annotation development efforts. Alternative methods based on nature language processing (NLP) and generative pretrained model (GPT), such as OnClass [27] and scGPT [28], have also been proposed to perform cell type annotation for single cell data. However, their performances have yet to be benchmarked.

To circumvent these issues, this study assumes that high-quality predictions are more likely when there is greater consensus among the methods. After obtaining results from the benchmarked cell-based and cluster-based methods, a majority voting strategy was adopted to recommend the consensus matching results. The resulting meta-comparison table deciphers commonalities and complementary results that the constituent methods produce, leading to an evidence-based meta-atlas. Moreover, further exploration into how cell type definitions and marker genes could be standardized across datasets could lead to improved consensus in cell type nomenclature and enhance the construction of reference atlases for the single cell community.

Last but not least, this study sheds light on how to reliably grow the reference atlas when new datasets are ingested into the data corpus. Currently, the common approach is to update the reference atlas by re-running the data integration pipeline and updating the cluster labels by co-clustering of all cells when new data is added to the reference atlas. This approach will likely produce slightly different cluster membership every time the reference is updated, thereby changing the original cell annotations in the reference. It will make future studies hard to reproduce results or leverage findings that are derived using a previous version of the reference. A more stable reference would be obtained if using a mechanism that allows organic growth of cell types in an incremental fashion by adding in new cell types when there is enough evidence from new data while preserving the cluster membership used in the original atlas. This would require the existing and new methods to have more enhanced and consistent performance for cluster-level matching and the ability to identify novel cell types, distinguishing similar, but not exact, matches of cell types.

Conclusion

The meta-analysis presented in this study highlights the challenges of identifying rare and novel cell types. Combining the benchmarked cell type matching results largely improves the accuracy of cell type annotation. Based on the benchmarking results, 41 cell types are matched between HLCA and CellRef, 20 cell types are HLCA-specific, and 7 cell types are CellRef-specific, leading to a growing meta-atlas for healthy adult human lung. This computational framework supports the incremental growth of cell types and the systematic update of reference atlases.

Methods

Algorithms

Table 2 summarizes the four cell type annotation algorithms benchmarked in this study along with notable parameters and features for replication. “Model specification” indicates the name of the pre-trained models used in Azimuth, CellTypist, and scArches, and the dimensionality-reduced gene space used in FR-Match (see below). “Reference” specifies the reference dataset that the pre-trained models were trained on; FR-Match uses any dataset provided as an input reference to the algorithm. “Code env” states the code environment that the method is available in. “Confidence score” describes the process used to assign a confidence score to each prediction. “Preprocessing” details the notable pre-processing steps taken in this study to prepare the query data for cell type matching. The preprocessing steps are followed as described in the original documentations of each algorithm.

Table 2: Summary of algorithms

Algorithm	Model specification	Reference	Code env	Confidence score	Preprocessing
Azimuth	“Human - Lung v2 (HLCA)”	HLCA	R	Prediction score from multiple anchors at cell level, ranging from 0-1.	count data + SCTransform for normalization and variance stabilization
CellTypist	“Human_Lung_Atlas”	HLCA	Python	Decision score that is a linear combination of the scaled gene expression and the model coefficients, ranging 0-1.	count data + scanpy.pp.normalize_total() + scanpy.pp.log1p()
scArches	“HLCA_reference_model" and "surgery_model”	HLCA	Python	Uncertainty score i.e., the probability of the predicted label is incorrect given the set of k-nearest neighbors, ranging from 0-1.	count data + sparse matrix + ensembl ID for features + perform surgery to integrate query with reference
FR-Match	“marker_selection = ‘binary_genes’”	User input	Python / R	Adjusted p-value based on the Friedman-Rafsky test	log-transformed data + normalization()

Azimuth

Azimuth employs anchors in lower-dimensional space to transfer cell-type labels from reference to query data clusters. This method first projects the query data onto a lower-dimensional space (UMAP) by applying sparse principal component analysis (sPCA) to the reference data. Subsequently, each query cell is compared to reference-based anchor cells to construct a k-nearest neighbor graph. Finally, Azimuth utilizes weighted voting to classify each query cell to a

corresponding reference cluster. Several levels of annotation granularity are available for this method.

An automated web application is available as cell annotation; however, for large-scale query datasets and the purposes of this study, Azimuth was executed locally in R using the "RunAzimuth()" function from the Azimuth library. "RunAzimuth()" accepts several forms of input for query data, including path names to AnnData files and Seurat objects. The pre-trained reference model is readily downloadable from the Azimuth website through Zenodo (https://zenodo.org/records/6342228).

To perform the HLCA cross-validation, the query data, was subset into 10 folds according to pre-split indices. As no pre-processing was required for the query data, each subset was used as input into "RunAzimuth()". The CellRef dataset required additional steps to be taken before running Azimuth. As the CellRef dataset is available as a H5AD file, the file path was provided as input for "RunAzimuth()". Internally, the algorithm uses Seurat's "LoadH5AD()" function to convert the H5AD file into a Seurat object for annotation. However, due to some differences in file formatting, both the function and H5AD object were manually modified to successfully convert the data into a Seurat object. A more comprehensive tutorial detailing this process is available in the GitHub repository. Once the H5AD file was converted to a Seurat object, the same procedure was applied to subset the query data into 10 folds for annotation.

CellTypist

CellTypist utilizes a logistic regression framework with stochastic gradient descent (SGD) optimization to train its reference models. To annotate each cell, CellTypist scales each gene according to parameters in the reference model and calculates decision scores for each query cell based on the linear combination of its scaled gene expression and the model coefficients. It assigns the cell type with the largest score to the query cell. CellTypist also provides an additional refinement step, majority voting, that was enabled in this study. Majority voting uses Leiden clustering to over-cluster query cells and re-assigns each subcluster to the majority predicted label for that cluster.

Same as in the Azimuth process, the query data was first subset into 10 folds using the same fold indices. Each subset was then preprocessed using "sc.pp.normalize_total(query_adata, target_sum = 1e4)" and "sc.pp.log1p(query_adata)" from Scanpy. Then, each fold was annotated using the annotate method from CellTypist with "majority_voting=True".

scArches

scArches uses architectural surgery to incorporate new input nodes from query data with trainable weights into a pre-trained neural-net reference model. The method first maps query data to the pre-trained neural-net reference model and trains new input nodes from the query data. Then, it uses a weighted k-nearest-neighbor classifier trained on the latent space representation of the reference dataset. Lastly, it calculates the uncertainty of each of the k labels for each query cell and classifies the query cell with the least uncertain label.

To use scArches, the query data was first checked for whether 1) it is a sparse matrix, 2) whether the matrix contains raw counts, and 3) whether the data uses the same features as the input features for its reference model. If the criterion were met, the query data was subset into 10 folds in the same process as Azimuth and CellTypist. Each subset was trained using

“surgery_model.train(accelerator='cpu', devices=1, max_epochs=surgery_epochs, **early_stopping_kwargs_surgery)” to incorporate the subset into the reference model. The specific parameters used for surgery model training can be found in the tutorial notebook. After surgery, the new surgery model is used to create a latent representation of the reference model to be used in label transfer.

FR-Match

Originally developed in R, FR-Match is now implemented in Python 3 (https://github.com/BeverlyPeng/frmatch) to streamline the matching workflow in the same code environment as Scanpy’s AnnData object [29] and NS-Forest algorithm [20, 21]. AnnData is a Python package that manages annotated single cell experiment data with computational efficient features. NS-Forest is a Python package for machine learning-based selection of cell type marker genes, which is compatible with the AnnData object and the Scanpy workflow.

The inputs to FR-Match are two AnnData objects (one for query and one for reference) and user’s choice of “marker_selection”. The choices for “marker_selection” can be “NSForest_markers”, “binary_genes”, or a user-provided list of marker genes. The reference AnnData object has the NS-Forest results saved in the unstructured metadata slot (.uns). If “NSForest_markers” or “binary_genes” is chosen, the corresponding list of genes will be extracted directly from .uns. Both the query and reference objects will be reduced to the chosen list of genes as the dimensionality reduction step in FR-Match. In this study, “marker_selection=‘binary_genes’” was used.

Before matching, the “normalization()” function is used to align the expression distributions of the two datasets, which performs a weighted per-gene min-max scaling to normalize the gene expression values to the range of [0, 1]. There are two main functions for conducting the matching: FRmatch() for cluster-based matching, and FRmatch_cell2cluster() for cell-based matching. As the other methods, the cell-based matching was conducted through subsetting the query data into 10 folds and matching each fold to the reference respectively. The cluster-based FR-Match does not need to subset the query data as it leverages the query clusters and assigns matching to each of the query clusters as a whole. One of the outputs from FR-Match algorithm is the adjusted p-values based on the non-parametric Friedman-Rafsky statistical test, which serves as an objective confidence score for the matching quality. Based on the adjusted p-value, a matching assignment is determined by adjusted p-value >= 0.05, and “unassigned” is determined by adjusted p-value < 0.05.

For the two-way matching, FRmatch() was run twice with the query and reference datasets switched. The plot_bi_FRmatch() function was used to combine the two sets of results and produce the two-way matching plot. Note that, since FR-Match utilizes the query cluster information, a query cluster with too few cells are removed in the algorithm. Hence, the query lymphatic EC proliferating in HLCA and the query chondrocyte in CellRef were removed in the FR-Match analyses.

Cell type dendrogram

Dendrogram is an informative visualization of the cell type taxonomy. The HLCA reference dendrogram is generated using “scanpy.tl.dendrogram()” [30], which computes a hierarchical clustering for the given cell type categories. Hierarchical clustering is an unsupervised machine learning technique that groups cell types based on the similarity of their expression profiles in a

tree structure, so that similar cell types are grouped together in the same branch. In the HLCA dendrogram, the cell types form major branches of immune, epithelial, endothelial and stromal cells, reflecting the hierarchy of cell classes and types. The cell type order in the dendrogram reflects the similarity of these cell types.

High confidence match

For the cell-based methods, the matching results were saved in a table (**Supplementary Table 1**) where rows are the CellRef cells, and columns are the true CellRef label ("cellref_true") and the predicted HLCA label and corresponding prediction score by each method ("azimuth_pred", "azimuth_score", "celltypist_pred", "celltypist_score", "frmatch_pred", "frmatch_score", "scarches_pred", "scarches_score"). To summarize the cell-based matching results into cluster-level matching, a contingency table of matching proportions p_{ij} for predicted HLCA label = i and true CellRef label = j were calculated, where

$$p_{ij} = \frac{number\ of\ cells\ matched\ between\ i\ and\ j}{total\ number\ of\ cells\ of\ j}, \text{ so that } \sum_{i \in \{all\ HLCA\ labels\}} p_{ij} = 1.$$

For each method, these values were plotted in Figure 5A. In Figure 5A, the column sums should be one; however, the row sums of these proportions may be larger than 1, which means that, at the cluster-level, many CellRef clusters were matched to the same HLCA cluster. To normalize the row sums, we define $q_{ij} = \frac{p_{ij}}{\sum_j p_{ij}}$, which can be interpreted as the cluster-level precision, ranging from 0 to 1. In the most ideal situation where the CellRef cluster and the HLCA cluster is a one-to-one match, $p_{ij} = q_{ij} = 1$. However, this is very hard to achieve. Practically, the high confidence matches are determined by requiring $p_{ij} > 0.5$ and $q_{ij} > 0.3$, which means the majority of the CellRef cells of label j were matched to HLCA label i, and the cluster-level precision is at least 30%. The values for p_{ij} and q_{ij} for the recommended matches are in **Supplementary Table 3**; and all high confidence matches for each method are in **Supplementary Table 4-7**.

NS-Forest marker gene selection

NS-Forest is a marker gene selection algorithm based on random forest machine learning. The NS-Forest v4 algorithm is used in this study, which prioritizes the gene pre-selection of binary-expressed genes prior to the random forest model by using "gene_selection=BinaryFirst_high". The NS-Forest algorithm reports the F-beta score, Positive Predictive Value (PPV, a.k.a. precision), and recall as metrics to quantify the strength of cell type classification using the selected marker genes. NS-Forest marker genes are the minimum combination among the top gene features ranked by random forest and binary score that achieves the highest F-beta score in the decision tree evaluation. The algorithm also outputs the top 10 binary genes after the random forest feature selection, which represents a more comprehensive list of genes that are highly expressed in the given cell type. The NS-Forest marker genes were visualized in the "barcode" plots; and the binary genes were used in FR-Match.

Data Availability

The HLCA dataset was download from CZ CELLxGENE: https://cellxgene.cziscience.com/collections/6f6d381a-7701-4781-935c-db10d30de293, namely "An integrated cell atlas of the human lung in health and disease (core)".

The CellRef dataset was downloaded from LungMAP: https://www.lungmap.net/omics/?experiment_id=LMEX0000004396, namely "LungMAP_HumanLung_CellRef.v1.1.h5ad".

Code Availability

All codes used in this study are documented in the tutorial notebooks, as a benchmark and protocols to compare and optimize current and future cell type matching tools, available via GitHub (https://github.com/jjoycehu/manuscript).

Azimuth in R: https://github.com/satijalab/azimuth

CellTypist in Python: https://github.com/Teichlab/celltypist

scArches in Python: https://github.com/theislab/scarches

FR-Match in R: https://github.com/JCVenterInstitute/FRmatch

FR-Match in Python: https://github.com/BeverlyPeng/frmatch

Acknowledgement

This work was supported by the U.S. National Institutes of Health under award 1R03OD036499, the NIH Common Fund through the Office of Strategic Coordination/Office of the NIH Director under awards OT2OD033756 and OT2OD026671, and the Division of Intramural Research of the National Library of Medicine (NLM), National Institutes of Health. The funding bodies had no role in the design or conclusions of this study.

Author contributions

YZ, RHS and KB conceived the project. YZ and RHS designed the study. JH and BP implemented the workflows and conducted the analyses. JH, BP, AVP, BX, and VAD contributed to the data analyses. JH, RHS and YZ interpreted the results. JH and YZ managed the datasets and results. BP built the software package. AB, BWH, and KB coordinate cell type annotation usage in the Human Reference Atlas resource. YZ and CD contributed to the project management. JH, BP, KB, RHS and YZ wrote the manuscript. All authors agreed on the contents of the manuscript.

References

1. Abdulla S, Aevermann B, Assis P, Badajoz S, Bell SM, Bezzi E, Cakir B, Chaffer J, Chambers S, Cherry JM, Chi T, Chien J, Dorman L, Garcia-Nieto P, Gloria N, Hastie M, Hegeman D, Hilton J, Huang T, Infeld A, Istrate AM, Jelic I, Katsuya K, Kim YJ, Liang K, Lin M, Lombardo M, Marshall B, Martin B, McDade F, Megill C, Patel N, Predeus A, Raymor B, Robatmili B, Rogers D, Rutherford E, Sadgat D, Shin A, Small C, Smith T, Sridharan P, Tarashansky A, Tavares N, Thomas H, Tolopko A, Urisko M, Yan J, Yeretssian G, Zamanian J, Mani A, Cool J, Carr A. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. Nucleic Acids Res. 2025;53(D1):D886-d900. Epub 2024/11/28. doi: 10.1093/nar/gkae1142. PubMed PMID: 39607691; PMCID: PMC11701654.
2. Jain S, Pei L, Spraggins JM, Angelo M, Carson JP, Gehlenborg N, Ginty F, Gonçalves JP, Hagood JS, Hickey JW, Kelleher NL, Laurent LC, Lin S, Lin Y, Liu H, Naba A, Nakayasu ES, Qian WJ, Radtke A, Robson P, Stockwell BR, Van de Plas R, Vlachos IS, Zhou M, Börner K, Snyder MP. Advances and prospects for the Human BioMolecular Atlas Program (HuBMAP). Nat Cell Biol. 2023;25(8):1089-100. Epub 2023/07/20. doi: 10.1038/s41556-023-01194-w. PubMed PMID: 37468756; PMCID: PMC10681365.
3. Tarhan L, Bistline J, Chang J, Galloway B, Hanna E, Weitz E. Single Cell Portal: an interactive home for single-cell genomics data. bioRxiv. 2023. Epub 2023/07/28. doi: 10.1101/2023.07.13.548886. PubMed PMID: 37502904; PMCID: PMC10370058.
4. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, Clevers H, Deplancke B, Dunham I, Eberwine J, Eils R, Enard W, Farmer A, Fugger L, Göttgens B, Hacohen N, Haniffa M, Hemberg M, Kim S, Klenerman P, Kriegstein A, Lein E, Linnarsson S, Lundberg E, Lundeberg J, Majumder P, Marioni JC, Merad M, Mhlanga M, Nawijn M, Netea M, Nolan G, Pe'er D, Phillipakis A, Ponting CP, Quake S, Reik W, Rozenblatt-Rosen O, Sanes J, Satija R, Schumacher TN, Shalek A, Shapiro E, Sharma P, Shin JW, Stegle O, Stratton M, Stubbington MJT, Theis FJ, Uhlen M, van Oudenaarden A, Wagner A, Watt F, Weissman J, Wold B, Xavier R, Yosef N. The Human Cell Atlas. Elife. 2017;6. Epub 2017/12/06. doi: 10.7554/eLife.27041. PubMed PMID: 29206104; PMCID: PMC5762154.
5. Börner K, Teichmann SA, Quardokus EM, Gee JC, Browne K, Osumi-Sutherland D, Herr BW, 2nd, Bueckle A, Paul H, Haniffa M, Jardine L, Bernard A, Ding SL, Miller JA, Lin S, Halushka MK, Boppana A, Longacre TA, Hickey J, Lin Y, Valerius MT, He Y, Pryhuber G, Sun X, Jorgensen M, Radtke AJ, Wasserfall C, Ginty F, Ho J, Sunshine J, Beuschel RT, Brusko M, Lee S, Malhotra R, Jain S, Weber G. Anatomical structures, cell types and biomarkers of the Human Reference Atlas. Nat Cell Biol. 2021;23(11):1117-28. Epub 2021/11/10. doi: 10.1038/s41556-021-00788-6. PubMed PMID: 34750582; PMCID: PMC10079270.
6. Börner K, Blood PD, Silverstein JC, Ruffalo M, Satija R, Teichmann SA, Pryhuber GJ, Misra RS, Purkerson JM, Fan J, Hickey JW, Molla G, Xu C, Zhang Y, Weber GM, Jain Y, Qaurooni D, Kong Y, Bueckle A, Herr BW, 2nd. Human BioMolecular Atlas Program (HuBMAP): 3D Human Reference Atlas construction and usage. Nat Methods. 2025. Epub 2025/03/14. doi: 10.1038/s41592-024-02563-5. PubMed PMID: 40082611.
7. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh PR, Raychaudhuri S. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16(12):1289-96. Epub 2019/11/20. doi: 10.1038/s41592-019-0619-0. PubMed PMID: 31740819; PMCID: PMC6884693.
8. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053-8. Epub 2018/12/07. doi: 10.1038/s41592-018-0229-2. PubMed PMID: 30504886; PMCID: PMC6289068.

9. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888-902.e21. Epub 2019/06/11. doi: 10.1016/j.cell.2019.05.031. PubMed PMID: 31178118; PMCID: PMC6687398.
10. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87.e29. Epub 2021/06/02. doi: 10.1016/j.cell.2021.04.048. PubMed PMID: 34062119; PMCID: PMC8238499.
11. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell. 2019;177(7):1873-87.e17. Epub 2019/06/11. doi: 10.1016/j.cell.2019.05.006. PubMed PMID: 31178122; PMCID: PMC6716797.
12. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colomé-Tatché M, Theis FJ. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022;19(1):41-50. Epub 2021/12/25. doi: 10.1038/s41592-021-01336-8. PubMed PMID: 34949812; PMCID: PMC8748196 Dermagnostix GmbH and Cellarity. The remaining authors declare no competing interests.
13. Guo M, Morley MP, Jiang C, Wu Y, Li G, Du Y, Zhao S, Wagner A, Cakar AC, Kouril M, Jin K, Gaddis N, Kitzmiller JA, Stewart K, Basil MC, Lin SM, Ying Y, Babu A, Wikenheiser-Brokamp KA, Mun KS, Naren AP, Clair G, Adkins JN, Pryhuber GS, Misra RS, Aronow BJ, Tickle TL, Salomonis N, Sun X, Morrisey EE, Whitsett JA, Xu Y. Guided construction of single cell reference for human and mouse lung. Nat Commun. 2023;14(1):4566. Epub 2023/07/30. doi: 10.1038/s41467-023-40173-5. PubMed PMID: 37516747; PMCID: PMC10387117.
14. Sikkema L, Ramírez-Suástegui C, Strobl DC, Gillett TE, Zappia L, Madissoon E, Markov NS, Zaragosi LE, Ji Y, Ansari M, Arguel MJ, Apperloo L, Banchero M, Bécavin C, Berg M, Chichelnitskiy E, Chung MI, Collin A, Gay ACA, Gote-Schniering J, Hooshiar Kashani B, Inecik K, Jain M, Kapellos TS, Kole TM, Leroy S, Mayr CH, Oliver AJ, von Papen M, Peter L, Taylor CJ, Walzthoeni T, Xu C, Bui LT, De Donno C, Dony L, Faiz A, Guo M, Gutierrez AJ, Heumos L, Huang N, Ibarra IL, Jackson ND, Kadur Lakshminarasimha Murthy P, Lotfollahi M, Tabib T, Talavera-López C, Travaglini KJ, Wilbrey-Clark A, Worlock KB, Yoshida M, van den Berge M, Bossé Y, Desai TJ, Eickelberg O, Kaminski N, Krasnow MA, Lafyatis R, Nikolic MZ, Powell JE, Rajagopal J, Rojas M, Rozenblatt-Rosen O, Seibold MA, Sheppard D, Shepherd DP, Sin DD, Timens W, Tsankov AM, Whitsett J, Xu Y, Banovich NE, Barbry P, Duong TE, Falk CS, Meyer KB, Kropski JA, Pe'er D, Schiller HB, Tata PR, Schultze JL, Teichmann SA, Misharin AV, Nawijn MC, Luecken MD, Theis FJ. An integrated cell atlas of the lung in health and disease. Nat Med. 2023;29(6):1563-77. Epub 2023/06/09. doi: 10.1038/s41591-023-02327-2. PubMed PMID: 37291214; PMCID: PMC10287567.
15. Zhang Y, Aevermann BD, Bakken TE, Miller JA, Hodge RD, Lein ES, Scheuermann RH. FR-Match: robust matching of cell type clusters from single cell RNA sequencing data using the Friedman-Rafsky non-parametric test. Brief Bioinform. 2021;22(4). Epub 2020/11/30. doi: 10.1093/bib/bbaa339. PubMed PMID: 33249453; PMCID: PMC8294536.
16. Zhang Y, Aevermann B, Gala R, Scheuermann RH. Cell type matching in single-cell RNA-sequencing data using FR-Match. Sci Rep. 2022;12(1):9996. Epub 2022/06/16. doi: 10.1038/s41598-022-14192-z. PubMed PMID: 35705694; PMCID: PMC9200772.
17. Domínguez Conde C, Xu C, Jarvis LB, Rainbow DB, Wells SB, Gomes T, Howlett SK, Suchanek O, Polanski K, King HW, Mamanova L, Huang N, Szabo PA, Richardson L, Bolt L, Fasouli ES, Mahbubani KT, Prete M, Tuck L, Richoz N, Tuong ZK, Campos L, Mousa HS, Needham EJ, Pritchard S, Li T, Elmentaite R, Park J, Rahmani E, Chen D, Menon DK, Bayraktar OA, James LK, Meyer KB, Yosef N, Clatworthy MR, Sims PA, Farber DL, Saeb-Parsy

K, Jones JL, Teichmann SA. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science. 2022;376(6594):eabl5197. Epub 2022/05/14. doi: 10.1126/science.abl5197. PubMed PMID: 35549406; PMCID: PMC7612735.
18. Xu C, Prete M, Webb S, Jardine L, Stewart BJ, Hoo R, He P, Meyer KB, Teichmann SA. Automatic cell-type harmonization and integration across Human Cell Atlas datasets. Cell. 2023;186(26):5876-91.e20. Epub 2023/12/23. doi: 10.1016/j.cell.2023.11.026. PubMed PMID: 38134877.
19. Lotfollahi M, Naghipourfar M, Luecken MD, Khajavi M, Büttner M, Wagenstetter M, Avsec Ž, Gayoso A, Yosef N, Interlandi M, Rybakov S, Misharin AV, Theis FJ. Mapping single-cell data to reference atlases by transfer learning. Nat Biotechnol. 2022;40(1):121-30. Epub 2021/09/01. doi: 10.1038/s41587-021-01001-7. PubMed PMID: 34462589; PMCID: PMC8763644 equity in Celsius Therapeutics and Rheos Medicines. The remaining authors declare no competing interests.
20. Aevermann B, Zhang Y, Novotny M, Keshk M, Bakken T, Miller J, Hodge R, Lelieveldt B, Lein E, Scheuermann RH. A machine learning method for the discovery of minimum marker gene combinations for cell type identification from single-cell RNA sequencing. Genome Res. 2021;31(10):1767-80. Epub 2021/06/06. doi: 10.1101/gr.275569.121. PubMed PMID: 34088715; PMCID: PMC8494219.
21. Liu A, Peng B, Pankajam Ajith V, Duong TE, Pryhuber G, Scheuermann RH, Zhang Y. Discovery of optimal cell type classification marker genes from single cell RNA sequencing data. BMC Methods. 2024;1(1):15. doi: 10.1186/s44330-024-00015-2.
22. Martinu T, Todd JL, Gelman AE, Guerra S, Palmer SM. Club Cell Secretory Protein in Lung Disease: Emerging Concepts and Potential Therapeutics. Annu Rev Med. 2023;74:427-41. Epub 2022/12/01. doi: 10.1146/annurev-med-042921-123443. PubMed PMID: 36450281; PMCID: PMC10472444.
23. Thai P, Chen Y, Dolganov G, Wu R. Differential regulation of MUC5AC/Muc5ac and hCLCA-1/mGob-5 expression in airway epithelium. Am J Respir Cell Mol Biol. 2005;33(6):523-30. Epub 2005/09/10. doi: 10.1165/rcmb.2004-0220RC. PubMed PMID: 16151054; PMCID: PMC2715330.
24. Grabski IN, Irizarry RA. A probabilistic gene expression barcode for annotation of cell types from single-cell RNA-seq data. Biostatistics. 2022;23(4):1150-64. Epub 2022/07/01. doi: 10.1093/biostatistics/kxac021. PubMed PMID: 35770795; PMCID: PMC9802389.
25. Grabski IN, Street K, Irizarry RA. Significance analysis for clustering with single-cell RNA-sequencing data. Nat Methods. 2023;20(8):1196-202. Epub 2023/07/11. doi: 10.1038/s41592-023-01933-9. PubMed PMID: 37429993; PMCID: PMC11282907.
26. Maan H, Zhang L, Yu C, Geuenich MJ, Campbell KR, Wang B. Characterizing the impacts of dataset imbalance on single-cell data integration. Nat Biotechnol. 2024;42(12):1899-908. Epub 2024/03/02. doi: 10.1038/s41587-023-02097-9. PubMed PMID: 38429430.
27. Wang S, Pisco AO, McGeever A, Brbic M, Zitnik M, Darmanis S, Leskovec J, Karkanias J, Altman RB. Leveraging the Cell Ontology to classify unseen cell types. Nat Commun. 2021;12(1):5556. Epub 2021/09/23. doi: 10.1038/s41467-021-25725-x. PubMed PMID: 34548483; PMCID: PMC8455606 (Personalis, 23andme); consulting or advisory role (United Health, Second Genome, Karius, UK Biobank, Swiss Personalized Health Network, Danish National Genome Center, All of Us Project, and Bridge Bio).
28. Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, Wang B. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods. 2024;21(8):1470-80. Epub 2024/02/27. doi: 10.1038/s41592-024-02201-0. PubMed PMID: 38409223.
29. Virshup I, Rybakov S, Theis FJ, Angerer P, Wolf FA. anndata: Access and store annotated data matrices. Journal of Open Source Software. 2024;9(101):4371.

30. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15. Epub 2018/02/08. doi: 10.1186/s13059-017-1382-0. PubMed PMID: 29409532; PMCID: PMC5802054.

A

B

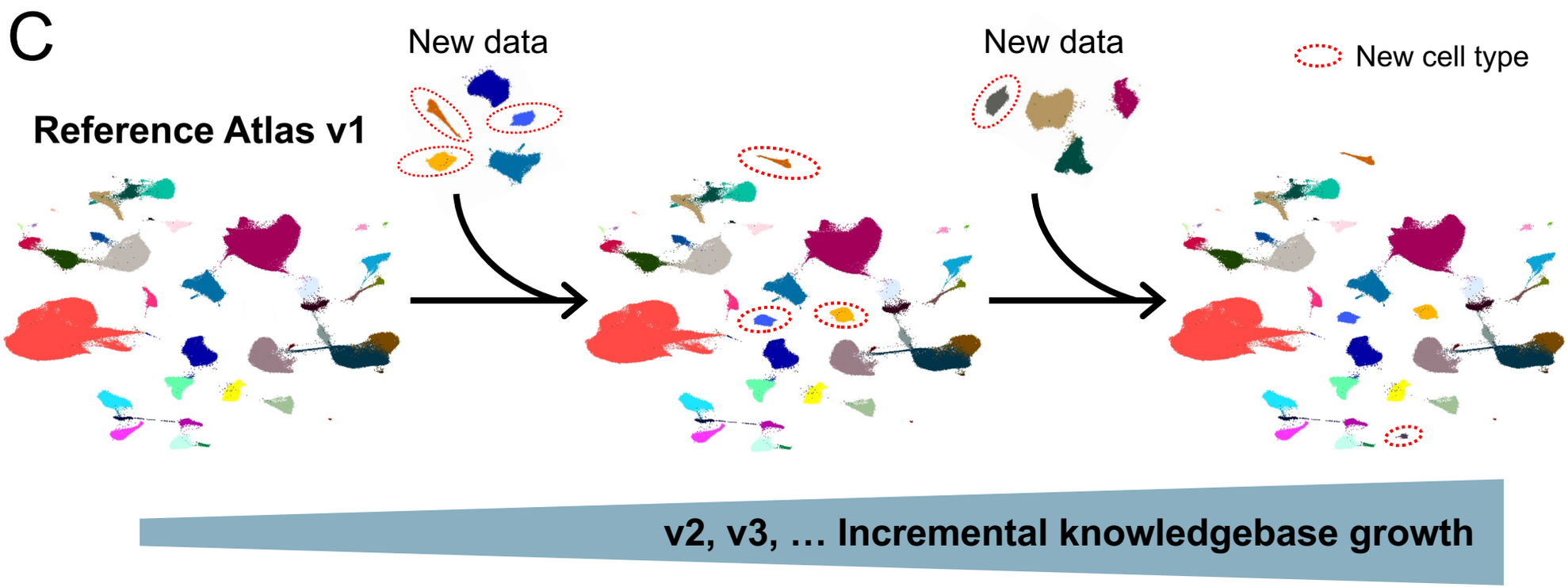
Cell type matching

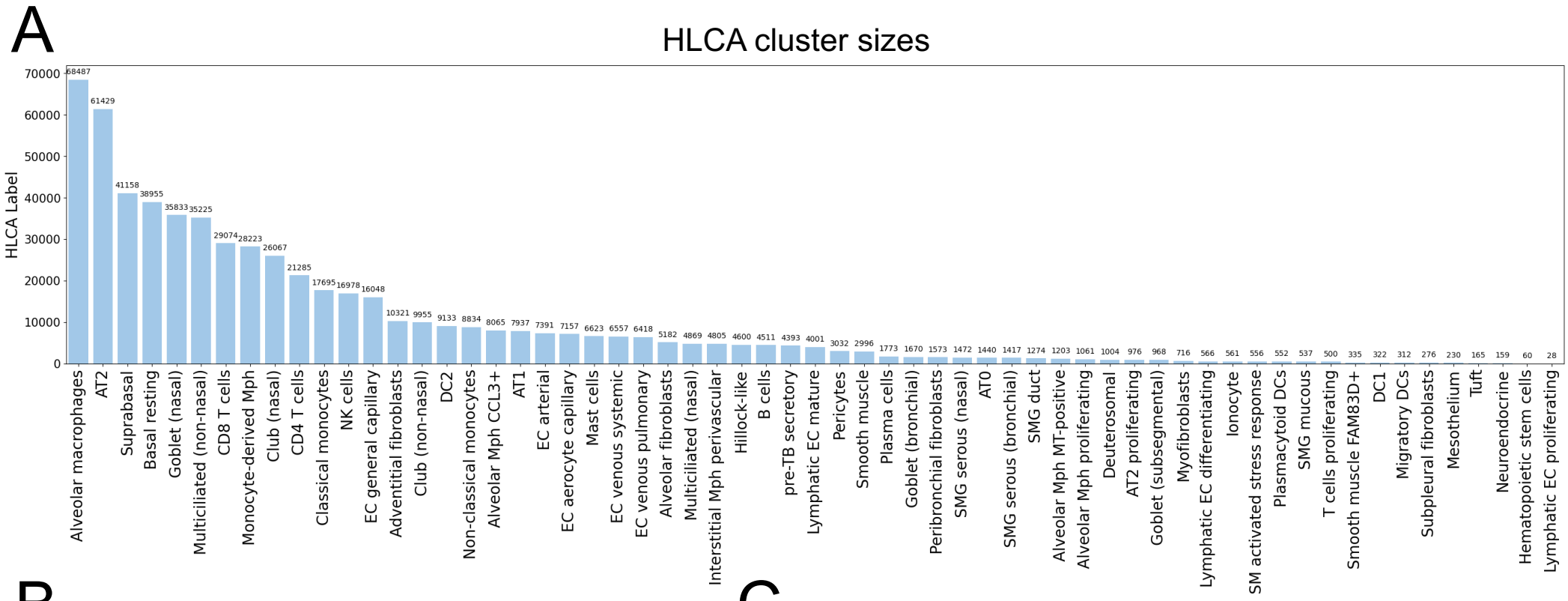
Cell-based matching

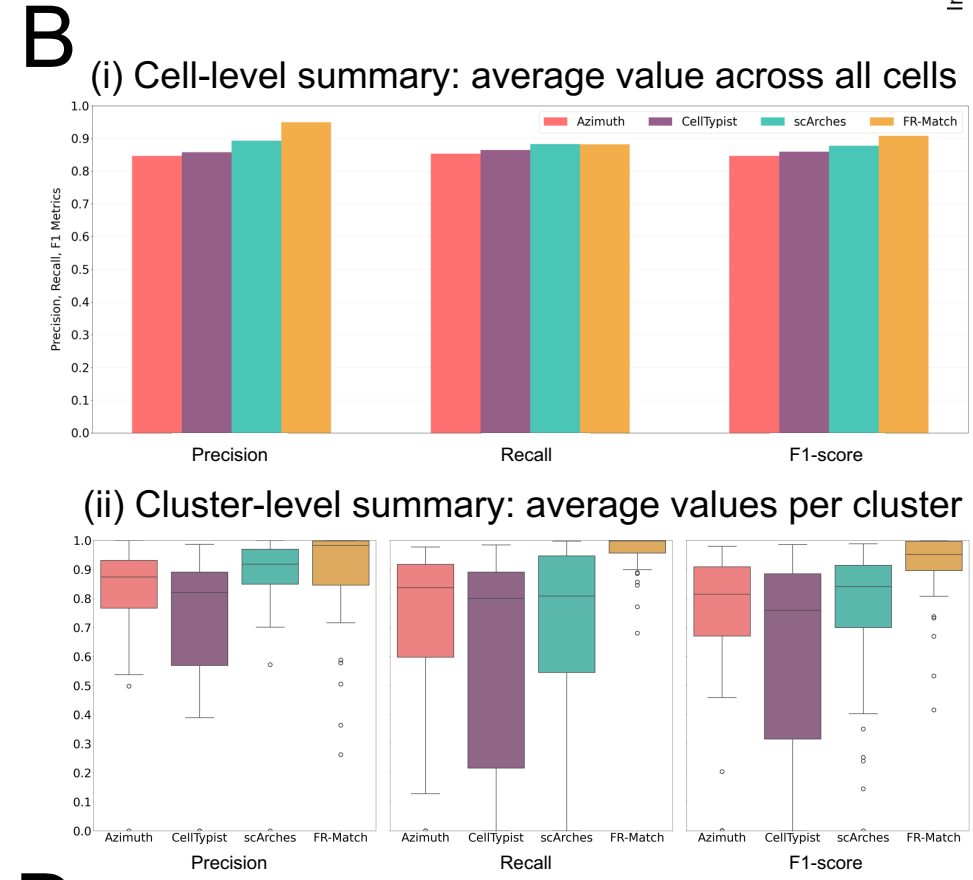
Cluster-based matching

Method	Reference	Query
Azimuth (pre-trained)	HLCA	Unsupervised
CellTypist (pre-trained)	HLCA	Unsupervised
scArches (pre-trained)	HLCA	Unsupervised
FR-Match cell-to-cluster	Flexible	Unsupervised
FR-Match cluster-to-cluster	Flexible	Supervised

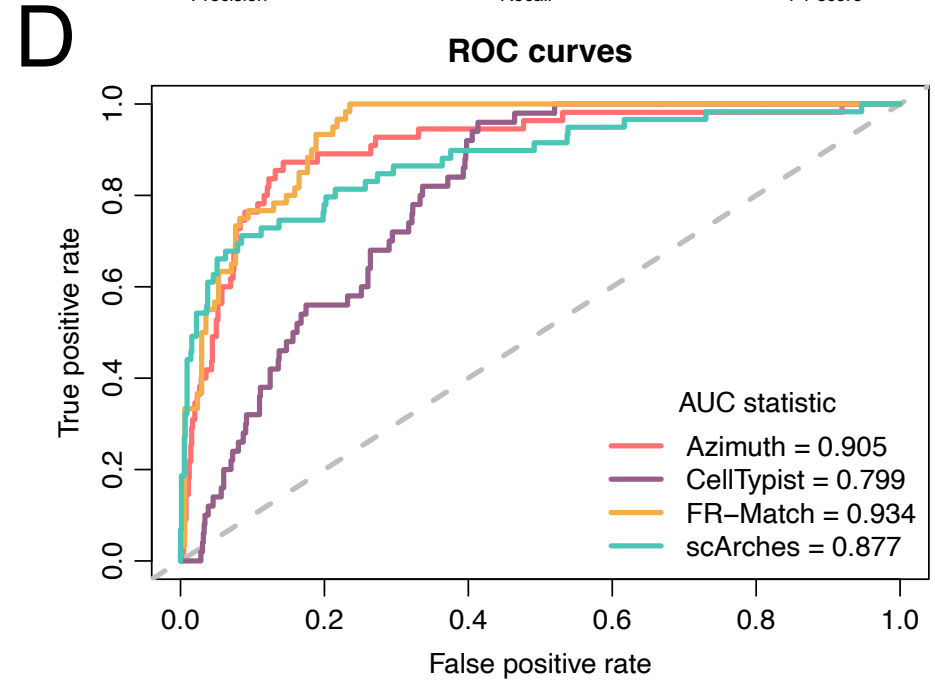
C

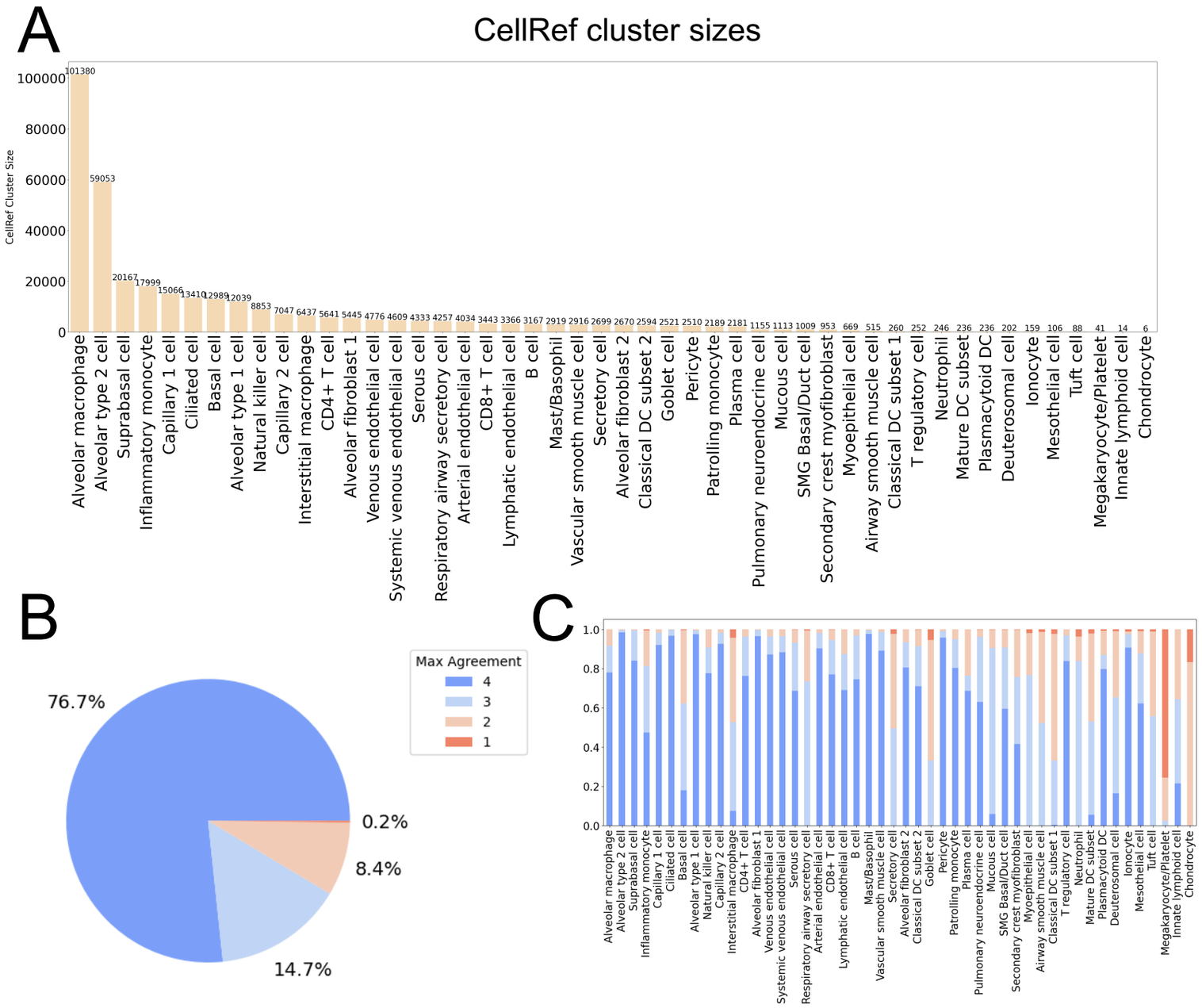


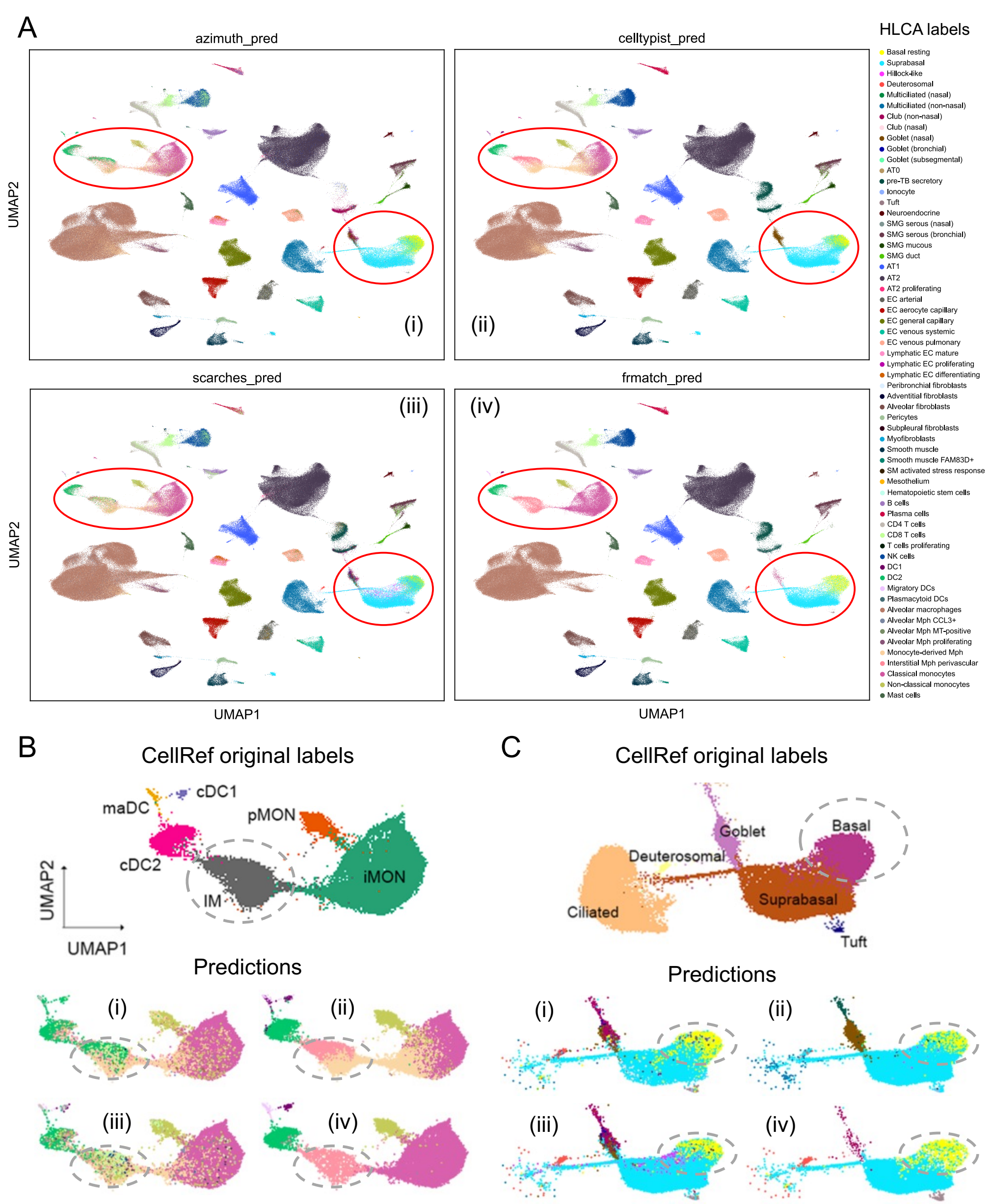

A
HLCA cluster sizes
HLCA Label
68487
61429
41158
38955
35833
35225
29074
28223
26067
21285
17695
16978
16048
10321
9955
9133
8834
8065
7937
7391
7157
6623
6557
6418
5182
4869
4805
4600
4511
4393
4001
3032
2996
1773
1670
1573
1472
1440
1417
1274
1203
1061
1004
976
968
716
566
561
556
552
537
500
335
322
312
276
230
165
159
60
28
Alveolar macrophages
AT2
Suprabasal
Basal resting
Goblet (nasal)
Multiciliated (non-nasal)
CD8 T cells
Monocyte-derived Mph
Club (nasal)
CD4 T cells
Classical monocytes
NK cells
EC general capillary
Adventitial fibroblasts
Club (non-nasal)
DC2
Non-classical monocytes
Alveolar Mph CCL3+
AT1
EC arterial
EC aerocyte capillary
Mast cells
EC venous systemic
EC venous pulmonary
Alveolar fibroblasts
Multiciliated (nasal)
Interstitial Mph perivascular
Hillock-like
B cells
pre-TB secretory
Lymphatic EC mature
Pericytes
Smooth muscle
Plasma cells
Goblet (bronchial)
Peribronchial fibroblasts
SMG serous (nasal)
AT0
SMG serous (bronchial)
SMG duct
Alveolar Mph MT-positive
Alveolar Mph proliferating
Deuterosomal
AT2 proliferating
Goblet (subsegmental)
Myofibroblasts
Lymphatic EC differentiating
Ionocyte
SM activated stress response
Plasmacytoid DCs
SMG mucous
T cells proliferating
Smooth muscle FAM83D+
DC1
Migratory DCs
Subpleural fibroblasts
Mesothelium
Tuft
Neuroendocrine
Hematopoietic stem cells
Lymphatic EC proliferating

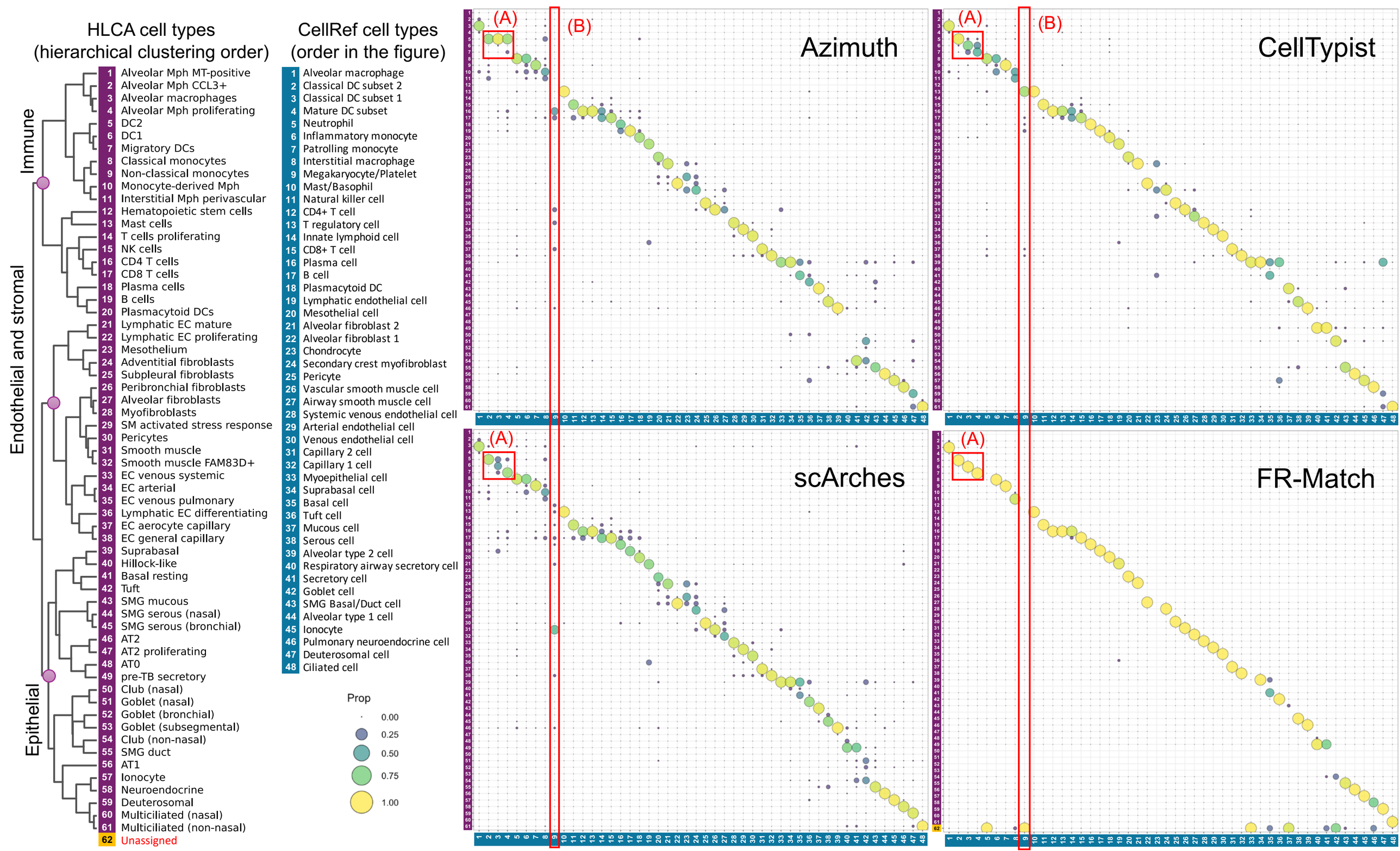

B
(i) Cell-level summary: average value across all cells
Precision, Recall, F1 Metrics
Azimuth
CellTypist
scArches
FR-Match
Precision
Recall
F1-score
(ii) Cluster-level summary: average values per cluster
Azimuth
CellTypist
scArches
FR-Match
Precision
Recall
F1-score

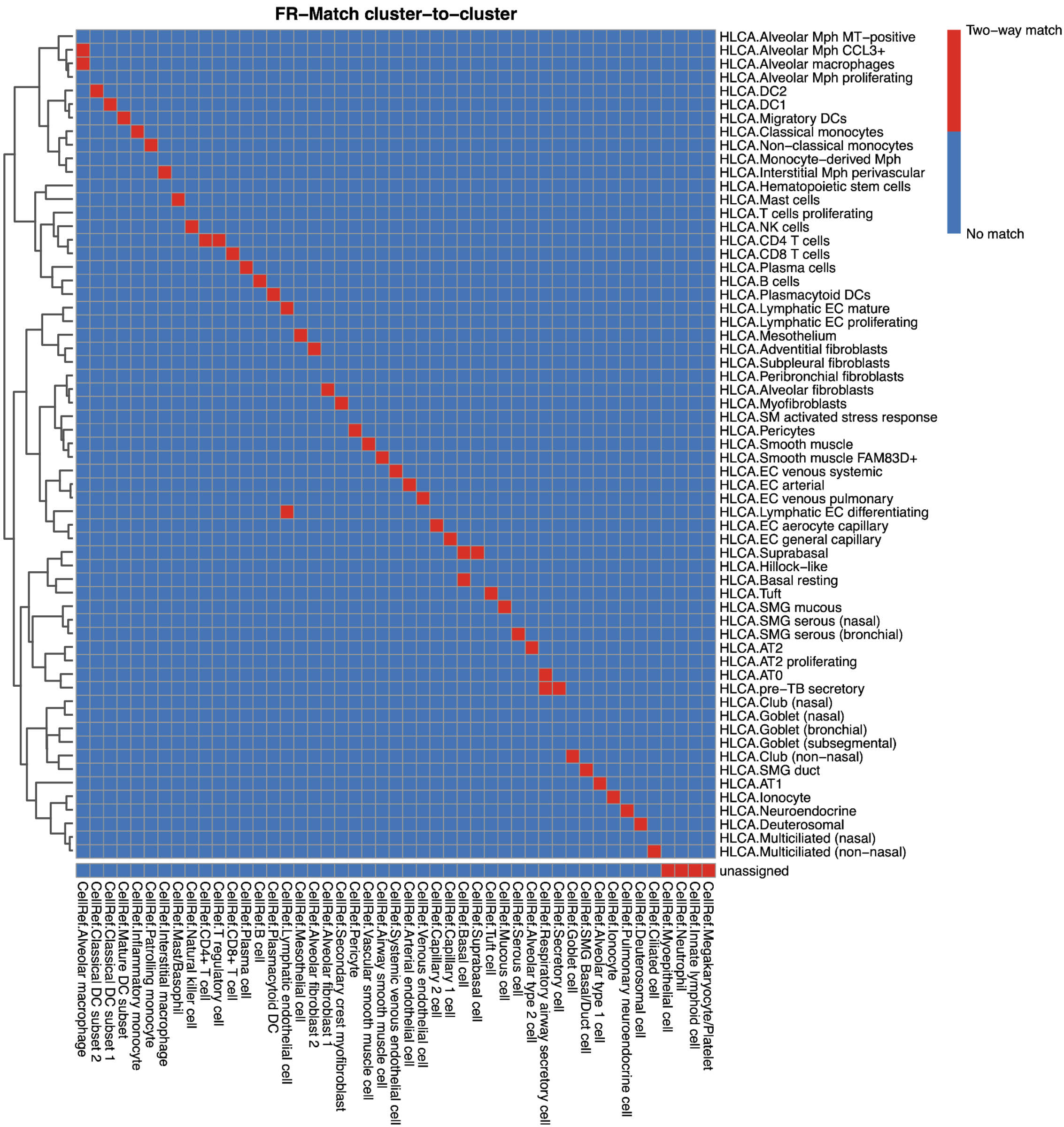
C
Cluster-level annotation accuracy (ordered by cluster size)
Azimuth - Overall: 0.693
CellTypist - Overall: 0.609
scArches - Overall: 0.717
FR-Match - Overall: 0.967
Proportion of Cells
Correct
Mislabeled
NA
Overall
HLCA Label


D
ROC curves
True positive rate
False positive rate
AUC statistic
Azimuth = 0.905
CellTypist = 0.799
FR–Match = 0.934
scArches = 0.877

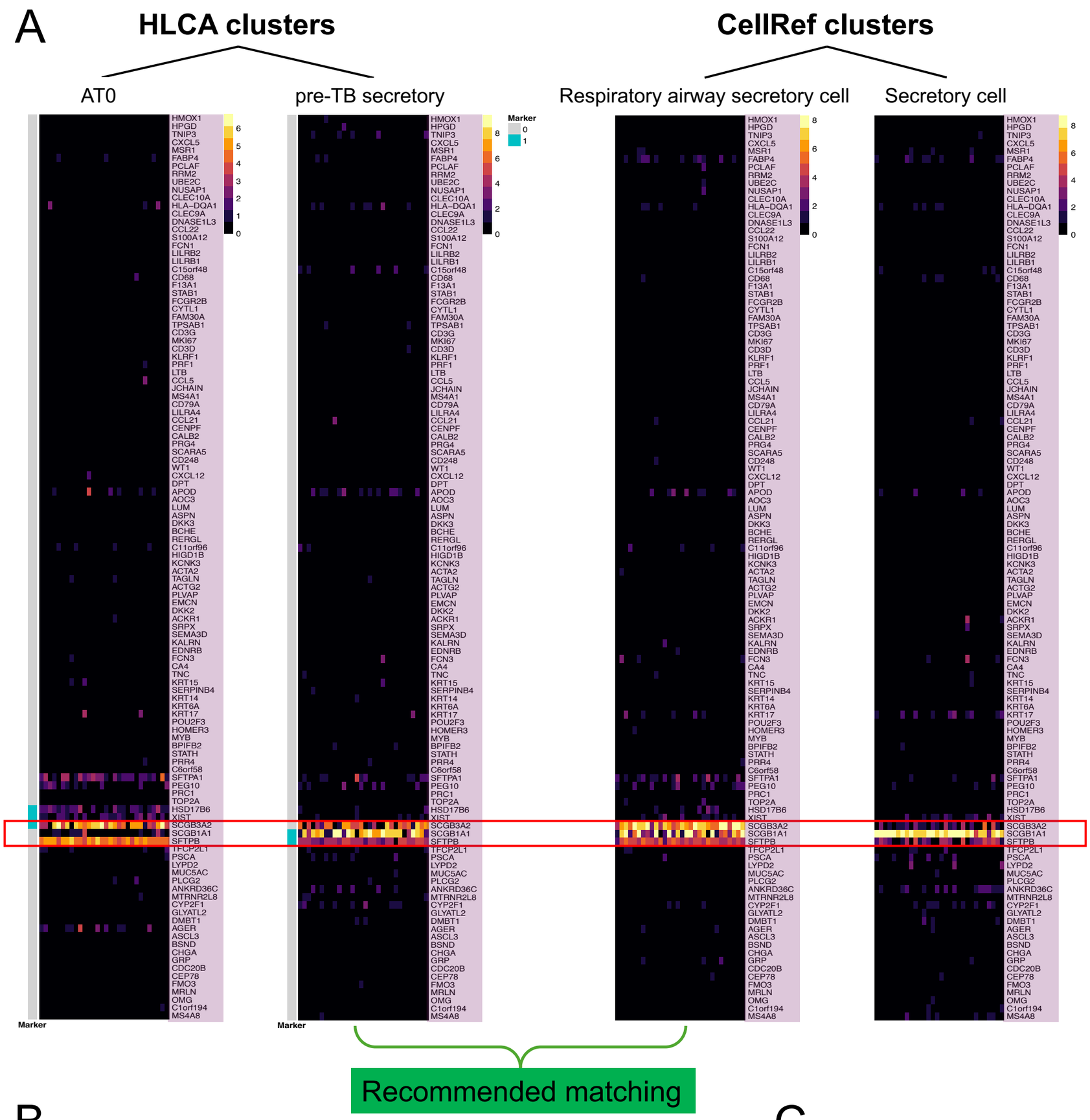

A
CellRef cluster sizes
CellRef Cluster Size
101380
59053
20167
17999
15066
13410
12989
12039
8853
7047
6437
5641
5445
4776
4609
4333
4257
4034
3443
3366
3167
2919
2916
2699
2670
2594
2521
2510
2189
2181
1155
1113
1009
953
669
515
260
252
246
236
236
202
159
106
88
41
14
6
Alveolar macrophage
Alveolar type 2 cell
Suprabasal cell
Inflammatory monocyte
Capillary 1 cell
Ciliated cell
Basal cell
Alveolar type 1 cell
Natural killer cell
Capillary 2 cell
Interstitial macrophage
CD4+ T cell
Alveolar fibroblast 1
Venous endothelial cell
Systemic venous endothelial cell
Serous cell
Respiratory airway secretory cell
Arterial endothelial cell
CD8+ T cell
Lymphatic endothelial cell
B cell
Mast/Basophil
Vascular smooth muscle cell
Secretory cell
Alveolar fibroblast 2
Classical DC subset 2
Goblet cell
Pericyte
Patrolling monocyte
Plasma cell
Pulmonary neuroendocrine cell
Mucous cell
SMG Basal/Duct cell
Secondary crest myofibroblast
Myoepithelial cell
Airway smooth muscle cell
Classical DC subset 1
T regulatory cell
Neutrophil
Mature DC subset
Plasmacytoid DC
Deuterosomal cell
Ionocyte
Mesothelial cell
Tuft cell
Megakaryocyte/Platelet
Innate lymphoid cell
Chondrocyte
B
Max Agreement
4
3
2
1
76.7%
0.2%
8.4%
14.7%
C
D
Cluster-level pairwise prediction agreement (ordered by cluster sizes)
Azimuth vs. CellTypist - Overall: 0.703
Azimuth vs. FR-Match - Overall: 0.64
Azimuth vs. scArches - Overall: 0.745
CellTypist vs. FR-Match - Overall: 0.745
CellTypist vs. scArches - Overall: 0.741
FR-Match vs. scArches - Overall: 0.702
Proportion of Cells
Agree
Disagree
NA
Overall
CellRef Original Labels

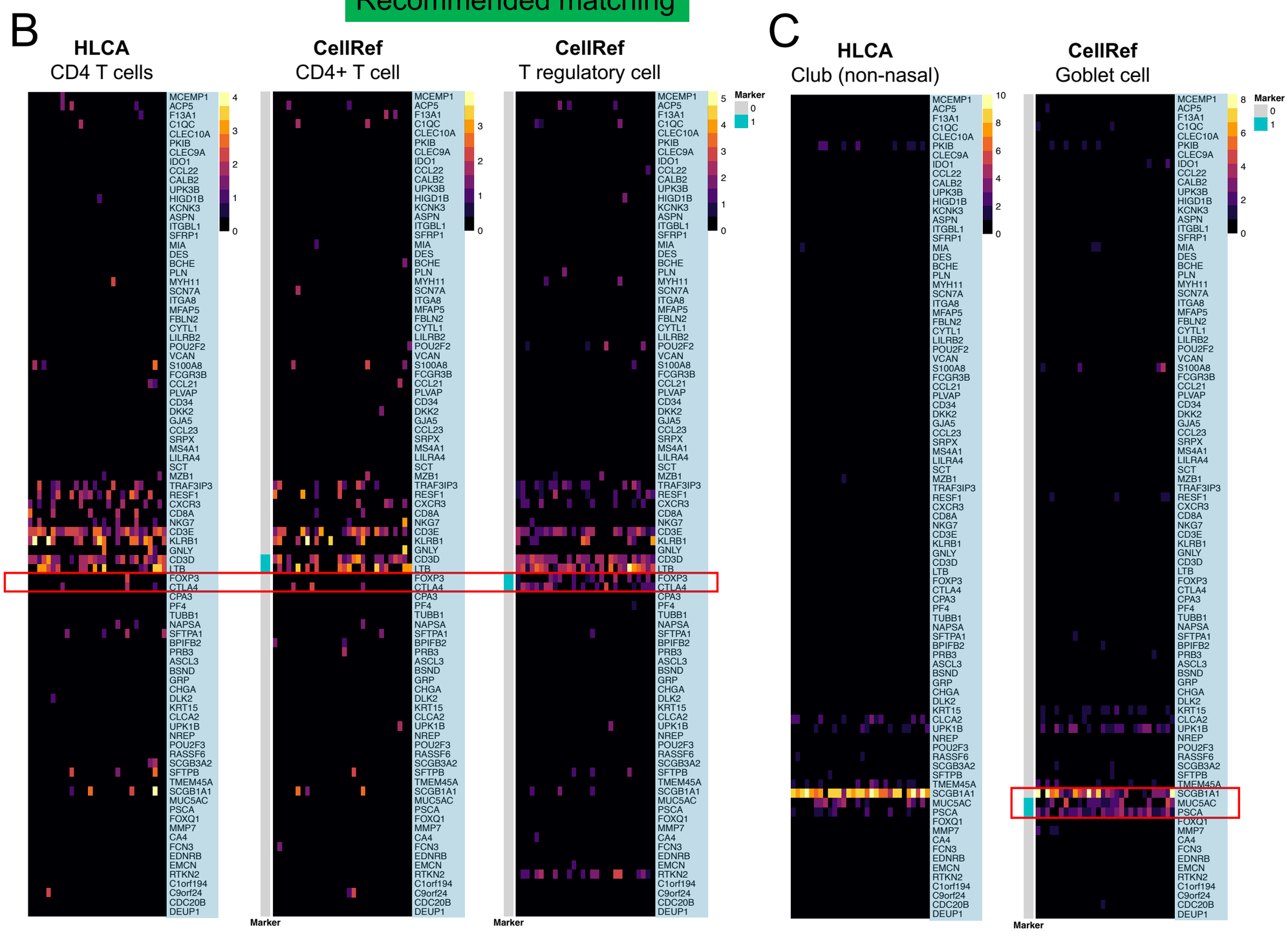

A
azimuth_pred
celltypist_pred
scarches_pred
frmatch_pred
(i)
(ii)
(iii)
(iv)
UMAP2
UMAP1
HLCA labels
Basal resting
Suprabasal
Hillock-like
Deuterosomal
Multiciliated (nasal)
Multiciliated (non-nasal)
Club (non-nasal)
Club (nasal)
Goblet (nasal)
Goblet (bronchial)
Goblet (subsegmental)
AT0
pre-TB secretory
Ionocyte
Tuft
Neuroendocrine
SMG serous (nasal)
SMG serous (bronchial)
SMG mucous
SMG duct
AT1
AT2
AT2 proliferating
EC arterial
EC aerocyte capillary
EC general capillary
EC venous systemic
EC venous pulmonary
Lymphatic EC mature
Lymphatic EC proliferating
Lymphatic EC differentiating
Peribronchial fibroblasts
Adventitial fibroblasts
Alveolar fibroblasts
Pericytes
Subpleural fibroblasts
Myofibroblasts
Smooth muscle
Smooth muscle FAM83D+
SM activated stress response
Mesothelium
Hematopoietic stem cells
B cells
Plasma cells
CD4 T cells
CD8 T cells
T cells proliferating
NK cells
DC1
DC2
Migratory DCs
Plasmacytoid DCs
Alveolar macrophages
Alveolar Mph CCL3+
Alveolar Mph MT-positive
Alveolar Mph proliferating
Monocyte-derived Mph
Interstitial Mph perivascular
Classical monocytes
Non-classical monocytes
Mast cells
B
CellRef original labels
maDC
cDC1
pMON
cDC2
iMON
IM
UMAP2
UMAP1
Predictions
(i)
(ii)
(iii)
(iv)
C
CellRef original labels
Goblet
Basal
Deuterosomal
Ciliated
Suprabasal
Tuft
Predictions
(i)
(ii)
(iii)
(iv)


HLCA cell types
(hierarchical clustering order)
CellRef cell types
(order in the figure)
Immune
Endothelial and stromal
Epithelial
1 Alveolar Mph MT-positive
2 Alveolar Mph CCL3+
3 Alveolar macrophages
4 Alveolar Mph proliferating
5 DC2
6 DC1
7 Migratory DCs
8 Classical monocytes
9 Non-classical monocytes
10 Monocyte-derived Mph
11 Interstitial Mph perivascular
12 Hematopoietic stem cells
13 Mast cells
14 T cells proliferating
15 NK cells
16 CD4 T cells
17 CD8 T cells
18 Plasma cells
19 B cells
20 Plasmacytoid DCs
21 Lymphatic EC mature
22 Lymphatic EC proliferating
23 Mesothelium
24 Adventitial fibroblasts
25 Subpleural fibroblasts
26 Peribronchial fibroblasts
27 Alveolar fibroblasts
28 Myofibroblasts
29 SM activated stress response
30 Pericytes
31 Smooth muscle
32 Smooth muscle FAM83D+
33 EC venous systemic
34 EC arterial
35 EC venous pulmonary
36 Lymphatic EC differentiating
37 EC aerocyte capillary
38 EC general capillary
39 Suprabasal
40 Hillock-like
41 Basal resting
42 Tuft
43 SMG mucous
44 SMG serous (nasal)
45 SMG serous (bronchial)
46 AT2
47 AT2 proliferating
48 AT0
49 pre-TB secretory
50 Club (nasal)
51 Goblet (nasal)
52 Goblet (bronchial)
53 Goblet (subsegmental)
54 Club (non-nasal)
55 SMG duct
56 AT1
57 Ionocyte
58 Neuroendocrine
59 Deuterosomal
60 Multiciliated (nasal)
61 Multiciliated (non-nasal)
62 Unassigned
1 Alveolar macrophage
2 Classical DC subset 2
3 Classical DC subset 1
4 Mature DC subset
5 Neutrophil
6 Inflammatory monocyte
7 Patrolling monocyte
8 Interstitial macrophage
9 Megakaryocyte/Platelet
10 Mast/Basophil
11 Natural killer cell
12 CD4+ T cell
13 T regulatory cell
14 Innate lymphoid cell
15 CD8+ T cell
16 Plasma cell
17 B cell
18 Plasmacytoid DC
19 Lymphatic endothelial cell
20 Mesothelial cell
21 Alveolar fibroblast 2
22 Alveolar fibroblast 1
23 Chondrocyte
24 Secondary crest myofibroblast
25 Pericyte
26 Vascular smooth muscle cell
27 Airway smooth muscle cell
28 Systemic venous endothelial cell
29 Arterial endothelial cell
30 Venous endothelial cell
31 Capillary 2 cell
32 Capillary 1 cell
33 Myoepithelial cell
34 Suprabasal cell
35 Basal cell
36 Tuft cell
37 Mucous cell
38 Serous cell
39 Alveolar type 2 cell
40 Respiratory airway secretory cell
41 Secretory cell
42 Goblet cell
43 SMG Basal/Duct cell
44 Alveolar type 1 cell
45 Ionocyte
46 Pulmonary neuroendocrine cell
47 Deuterosomal cell
48 Ciliated cell
Prop
0.00
0.25
0.50
0.75
1.00
(A)
(B)
Azimuth
CellTypist
scArches
FR-Match


FR−Match cluster−to−cluster
Two−way match
No match
HLCA.Alveolar Mph MT−positive
HLCA.Alveolar Mph CCL3+
HLCA.Alveolar macrophages
HLCA.Alveolar Mph proliferating
HLCA.DC2
HLCA.DC1
HLCA.Migratory DCs
HLCA.Classical monocytes
HLCA.Non−classical monocytes
HLCA.Monocyte−derived Mph
HLCA.Interstitial Mph perivascular
HLCA.Hematopoietic stem cells
HLCA.Mast cells
HLCA.T cells proliferating
HLCA.NK cells
HLCA.CD4 T cells
HLCA.CD8 T cells
HLCA.Plasma cells
HLCA.B cells
HLCA.Plasmacytoid DCs
HLCA.Lymphatic EC mature
HLCA.Lymphatic EC proliferating
HLCA.Mesothelium
HLCA.Adventitial fibroblasts
HLCA.Subpleural fibroblasts
HLCA.Peribronchial fibroblasts
HLCA.Alveolar fibroblasts
HLCA.Myofibroblasts
HLCA.SM activated stress response
HLCA.Pericytes
HLCA.Smooth muscle
HLCA.Smooth muscle FAM83D+
HLCA.EC venous systemic
HLCA.EC arterial
HLCA.EC venous pulmonary
HLCA.Lymphatic EC differentiating
HLCA.EC aerocyte capillary
HLCA.EC general capillary
HLCA.Suprabasal
HLCA.Hillock−like
HLCA.Basal resting
HLCA.Tuft
HLCA.SMG mucous
HLCA.SMG serous (nasal)
HLCA.SMG serous (bronchial)
HLCA.AT2
HLCA.AT2 proliferating
HLCA.AT0
HLCA.pre−TB secretory
HLCA.Club (nasal)
HLCA.Goblet (nasal)
HLCA.Goblet (bronchial)
HLCA.Goblet (subsegmental)
HLCA.Club (non−nasal)
HLCA.SMG duct
HLCA.AT1
HLCA.Ionocyte
HLCA.Neuroendocrine
HLCA.Deuterosomal
HLCA.Multiciliated (nasal)
HLCA.Multiciliated (non−nasal)
unassigned
CellRef.Megakaryocyte/Platelet
CellRef.Innate lymphoid cell
CellRef.Neutrophil
CellRef.Myoepithelial cell
CellRef.Ciliated cell
CellRef.Deuterosomal cell
CellRef.Pulmonary neuroendocrine cell
CellRef.Ionocyte
CellRef.Alveolar type 1 cell
CellRef.SMG Basal/Duct cell
CellRef.Goblet cell
CellRef.Secretory cell
CellRef.Serous cell
CellRef.Alveolar type 2 cell
CellRef.Respiratory airway secretory cell
CellRef.Mucous cell
CellRef.Tuft cell
CellRef.Suprabasal cell
CellRef.Basal cell
CellRef.Capillary 1 cell
CellRef.Capillary 2 cell
CellRef.Venous endothelial cell
CellRef.Arterial endothelial cell
CellRef.Systemic venous endothelial cell
CellRef.Airway smooth muscle cell
CellRef.Vascular smooth muscle cell
CellRef.Pericyte
CellRef.Secondary crest myofibroblast
CellRef.Alveolar fibroblast 1
CellRef.Alveolar fibroblast 2
CellRef.Mesothelial cell
CellRef.Lymphatic endothelial cell
CellRef.Plasmacytoid DC
CellRef.B cell
CellRef.Plasma cell
CellRef.CD8+ T cell
CellRef.T regulatory cell
CellRef.CD4+ T cell
CellRef.Natural killer cell
CellRef.Mast/Basophil
CellRef.Interstitial macrophage
CellRef.Patrolling monocyte
CellRef.Inflammatory monocyte
CellRef.Mature DC subset
CellRef.Classical DC subset 1
CellRef.Classical DC subset 2
CellRef.Alveolar macrophage

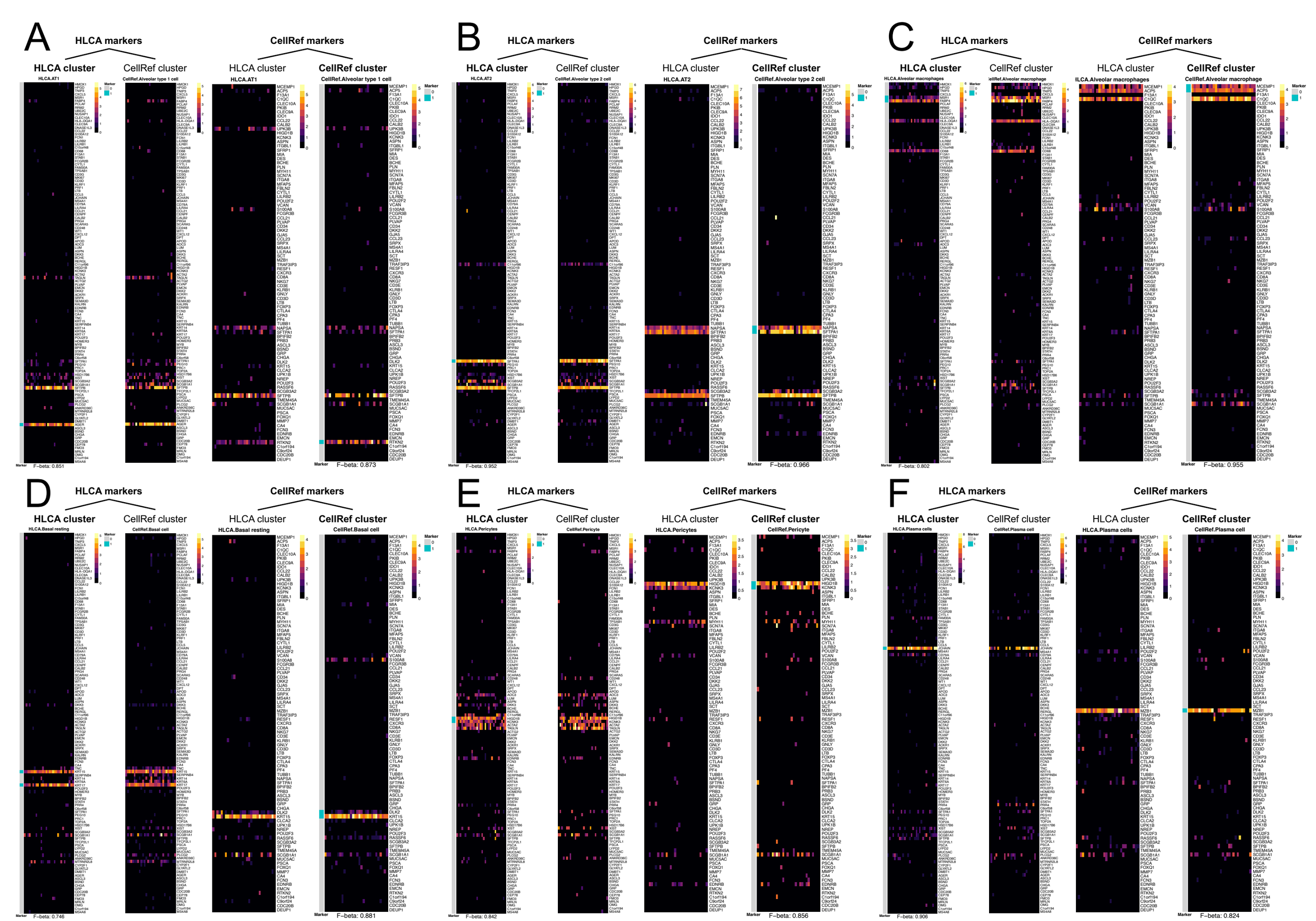
HLCA	CellRef	Match	HLCA_ClusterName	CellRef_ClusterName	Cluster-based matching	Cell-based matching				RecommendedMatching
					TwoWayMatch_FRMatch	HiConfMatch_Azimuth	HiConfMatch_CellTypist	HiConfMatch_scArches	HiConfMatch_FRMatch	
3	1	1	Alveolar macrophages	Alveolar macrophage	Y	Y	Y	Y	Y	HLCA2CellRef_Match
8	6	2	Classical monocytes	Inflammatory monocyte	Y	Y	Y	Y	Y	HLCA2CellRef_Match
9	7	3	Non-classical monocytes	Patrolling monocyte	Y	Y	Y	Y	Y	HLCA2CellRef_Match
13	10	4	Mast cells	Mast/Basophil	Y	Y	Y	Y	Y	HLCA2CellRef_Match
15	11	5	NK cells	Natural killer cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
16	12	6	CD4 T cells	CD4+ T cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
17	15	7	CD8 T cells	CD8+ T cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
18	16	8	Plasma cells	Plasma cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
19	17	9	B cells	B cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
20	18	10	Plasmacytoid DCs	Plasmacytoid DC	Y	Y	Y	Y	Y	HLCA2CellRef_Match
21	19	11	Lymphatic EC mature	Lymphatic endothelial cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
23	20	12	Mesothelium	Mesothelial cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
24	21	13	Adventitial fibroblasts	Alveolar fibroblast 2	Y	Y	Y	Y	Y	HLCA2CellRef_Match
27	22	14	Alveolar fibroblasts	Alveolar fibroblast 1	Y	Y	Y	Y	Y	HLCA2CellRef_Match
28	24	15	Myofibroblasts	Secondary crest myofibroblast	Y	Y	Y	Y	Y	HLCA2CellRef_Match
30	25	16	Pericytes	Pericyte	Y	Y	Y	Y	Y	HLCA2CellRef_Match
31	26	17	Smooth muscle	Vascular smooth muscle cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
33	28	18	EC venous systemic	Systemic venous endothelial cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
34	29	19	EC arterial	Arterial endothelial cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
35	30	20	EC venous pulmonary	Venous endothelial cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
37	31	21	EC aerocyte capillary	Capillary 2 cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
38	32	22	EC general capillary	Capillary 1 cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
45	38	23	SMG serous (bronchial)	Serous cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
46	39	24	AT2	Alveolar type 2 cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
55	43	25	SMG duct	SMG Basal/Duct cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
56	44	26	AT1	Alveolar type 1 cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
57	45	27	Ionocyte	Ionocyte	Y	Y	Y	Y	Y	HLCA2CellRef_Match
58	46	28	Neuroendocrine	Pulmonary neuroendocrine cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
61	48	29	Multiciliated (non-nasal)	Ciliated cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
5	2	30	DC2	Classical DC subset 2	Y	N	Y	Y	Y	HLCA2CellRef_Match
6	3	31	DC1	Classical DC subset 1	Y	N	Y	N	Y	HLCA2CellRef_Match
7	4	32	Migratory DCs	Mature DC subset	Y	N	Y	Y	Y	HLCA2CellRef_Match
11	8	33	Interstitial Mph perivascular	Interstitial macrophage	Y	N	N	N	Y	HLCA2CellRef_Match
32	27	34	Smooth muscle FAM83D+	Airway smooth muscle cell	Y	N	Y	Y	Y	HLCA2CellRef_Match
39	34	35	Suprabasal	Suprabasal cell	Y	Y	N	Y	Y	HLCA2CellRef_Match
41	35	36	Basal resting	Basal cell	Y	Y	Y	N	Y	HLCA2CellRef_Match
42	36	37	Tuft	Tuft cell	Y	Y	N	Y	Y	HLCA2CellRef_Match
43	37	38	SMG mucous	Mucous cell	Y	Y	Y	Y	N	HLCA2CellRef_Match
49	40	39	pre-TB secretory	Respiratory airway secretory cell	Y	N	Y	Y	Y	HLCA2CellRef_Match
54	42	40	Club (non-nasal)	Goblet cell	Y	N	N	N	N	HLCA2CellRef_boardMatch
59	47	41	Deuterosomal	Deuterosomal cell	Y	Y	N	Y	Y	HLCA2CellRef_Match
1		42	Alveolar Mph MT-positive		N	N	N	N	N	HLCA-specific
2		43	Alveolar Mph CCL3+		Y (Alveolar macrophage)	N	N	N	N	HLCA-specific
4		44	Alveolar Mph proliferating		N	N	N	N	N	HLCA-specific
10		45	Monocyte-derived Mph		N	N	N	N	N	HLCA-specific
12		46	Hematopoietic stem cells		N	N	N	N	N	HLCA-specific
14		47	T cells proliferating		N	N	N	N	N	HLCA-specific
22		48	Lymphatic EC proliferating		N	N	N	N	N	HLCA-specific
25		49	Subpleural fibroblasts		N	N	N	N	N	HLCA-specific
26		50	Peribronchial fibroblasts		N	N	N	N	N	HLCA-specific
29		51	SM activated stress response		N	N	N	N	N	HLCA-specific
36		52	Lymphatic EC differentiating		Y (Lymphatic endothelial cell)	N	N	N	N	HLCA-specific
40		53	Hillock-like		N	N	N	N	N	HLCA-specific
44		54	SMG serous (nasal)		N	N	N	N	N	HLCA-specific
47		55	AT2 proliferating		N	N	N	N	N	HLCA-specific
48		56	AT0		Y (Respiratory airway secretory cell)	N	N	N	N	HLCA-specific
51		57	Goblet (nasal)		N	N	Y (Goblet)	N	N	HLCA-specific
50		58	Club (nasal)		N	N	N	N	N	HLCA-specific
52		59	Goblet (bronchial)		N	N	N	N	N	HLCA-specific
53		60	Goblet (subsegmental)		N	N	N	N	N	HLCA-specific
60		61	Multiciliated (nasal)		N	N	N	N	N	HLCA-specific
	5	62		Neutrophil	N	Y (Classical monocytes)	Y (Classical monocytes)	Y (Classical monocytes)	N	CellRef-specific
	9	63		Megakaryocyte/Platelet	N	N	Y (Mast cells)	Y (Smooth muscle)	N	CellRef-specific
	13	64		T regulatory cell	Y (CD4 T cells)	Y (CD4 T cells)	Y (CD4 T cells)	Y (CD4 T cells)	Y (CD4 T cells)	CellRef-specific
	14	65		Innate lymphoid cell	N	N	N	N	Y (CD4 T cells)	CellRef-specific
	23	66		Chondrocyte	NA	N	N	N	NA	CellRef-specific
	33	67		Myoepithelial cell	N	Y (Suprabasal)	N	Y (Suprabasal)	N	CellRef-specific
	41	68		Secretory cell	Y (pre-TB secretory)	Y (Club (non-nasal))	Y (pre-TB secretory)	Y (pre-TB secretory)	Y (pre-TB secretory)	CellRef-specific

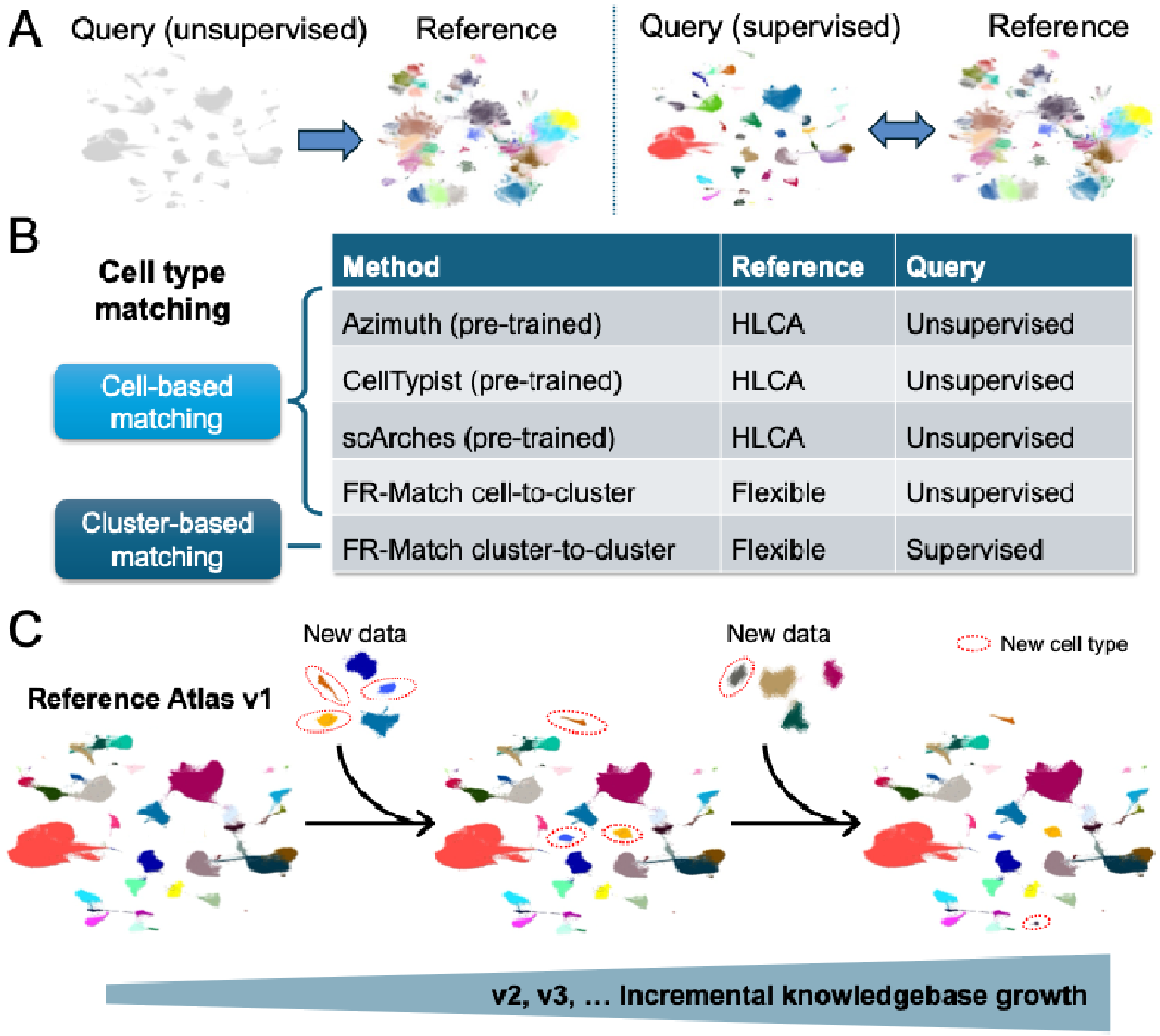




D

HLCA markers	CellRef markers
HMOX1	MCEMP1
HPGD	ACP5
TNIP3	F13A1
CXCL5	C1QC
MSR1	CLEC10A
FABP4	PKIB
PCLAF	CLEC9A
RRM2	IDO1
UBE2C	CCL22
NUSAP1	CALB2
CLEC10A	UPK3B
HLA–DQA1	HIGD1B
CLEC9A	KCNK3
DNASE1L3	ASPN
CCL22	ITGBL1
S100A12	SFRP1
FCN1	MIA
LILRB2	DES
LILRB1	BCHE
C15orf48	PLN
CD68	MYH11
F13A1	SCN7A
STAB1	ITGA8
FCGR2B	MFAP5
CYTL1	FBLN2
FAM30A	CYTL1
TPSAB1	LILRB2
CD3G	POU2F2
MKI67	VCAN
CD3D	S100A8
KLRF1	FCGR3B
PRF1	CCL21
LTB	PLVAP
CCL5	CD34
JCHAIN	DKK2
MS4A1	GJA5
CD79A	CCL23
LILRA4	SRPX
CCL21	MS4A1
CENPF	LILRA4
CALB2	SCT
PRG4	MZB1
SCARA5	TRAF3IP3
CD248	RESF1
WT1	CXCR3
CXCL12	CD8A
DPT	NKG7
APOD	CD3E
AOC3	KLRB1
LUM	GNLY
ASPN	CD3D
DKK3	LTB
BCHE	FOXP3
RERGL	CTLA4
C11orf96	CPA3
HIGD1B	PF4
KCNK3	TUBB1
ACTA2	NAPSA
TAGLN	SFTPA1
ACTG2	BPIFB2
PLVAP	PRB3
EMCN	ASCL3
DKK2	BSND
ACKR1	GRP
SRPX	CHGA
SEMA3D	DLK2
KALRN	KRT15
EDNRB	CLCA2
FCN3	UPK1B
CA4	NREP
TNC	POU2F3
KRT15	RASSF6
SERPINB4	SCGB3A2
KRT14	SFTPB
KRT6A	TMEM45A
KRT17	SCGB1A1
POU2F3	MUC5AC
HOMER3	PSCA
MYB	FOXQ1
BPIFB2	MMP7
STATH	CA4
PRR4	FCN3
C6orf58	EDNRB
SFTPA1	EMCN
PEG10	RTKN2
PRC1	C1orf194
TOP2A	C9orf24
HSD17B6	CDC20B
XIST	DEUP1
SCGB3A2	
SCGB1A1	
SFTPB	
TFCP2L1	
PSCA	
LYPD2	
MUC5AC	
PLCG2	
ANKRD36C	
MTRNR2L8	
CYP2F1	
GLYATL2	
DMBT1	
AGER	
ASCL3	
BSND	
CHGA	
GRP	
CDC20B	
CEP78	
FMO3	
MRLN	
OMG	
C1orf194	
MS4A8	


A
HLCA markers
CellRef markers
HLCA cluster
CellRef cluster
HLCA cluster
CellRef cluster
HLCA.AT1
CellRef.Alveolar type 1 cell
HLCA.AT1
CellRef.Alveolar type 1 cell
Marker
F−beta: 0.851
F−beta: 0.873
B
HLCA markers
CellRef markers
HLCA cluster
CellRef cluster
HLCA cluster
CellRef cluster
HLCA.AT2
CellRef.Alveolar type 2 cell
HLCA.AT2
CellRef.Alveolar type 2 cell
Marker
F−beta: 0.952
F−beta: 0.966
C
HLCA markers
CellRef markers
HLCA cluster
CellRef cluster
HLCA cluster
CellRef cluster
HLCA.Alveolar macrophages
CellRef.Alveolar macrophage
ILCA.Alveolar macrophages
CellRef.Alveolar macrophage
Marker
F−beta: 0.802
F−beta: 0.955
D
HLCA markers
CellRef markers
HLCA cluster
CellRef cluster
HLCA cluster
CellRef cluster
HLCA.Basal resting
CellRef.Basal cell
HLCA.Basal resting
CellRef.Basal cell
Marker
F−beta: 0.746
F−beta: 0.881
E
HLCA markers
CellRef markers
HLCA cluster
CellRef cluster
HLCA cluster
CellRef cluster
HLCA.Pericytes
CellRef.Pericyte
HLCA.Pericytes
CellRef.Pericyte
Marker
F−beta: 0.842
F−beta: 0.856
F
HLCA markers
CellRef markers
HLCA cluster
CellRef cluster
HLCA cluster
CellRef cluster
HLCA.Plasma cells
CellRef.Plasma cell
HLCA.Plasma cells
CellRef.Plasma cell
Marker
F−beta: 0.906
F−beta: 0.824



Method	Reference	Query
Azimuth (pre-trained)	HLCA	Unsupervised
CellTypist (pre-trained)	HLCA	Unsupervised
scArches (pre-trained)	HLCA	Unsupervised
FR-Match cell-to-cluster	Flexible	Unsupervised
FR-Match cluster-to-cluster	Flexible	Supervised

Figure 1. Study overview. Constructing an integrated human lung reference atlas using computational cell type matching between lung atlas datasets from single cell transcriptomic profiling, in support of incremental growth of a cell type knowledgebase. **(A)** Unsupervised and supervised cell type matching between query and reference datasets. The unsupervised cell type matching is when the query dataset has no cluster / cell type labels, so that the matching direction can only be in one direction from the query to the reference. The supervised cell type matching is when both datasets have annotated cell types, so that two-way cell type matching is allowed to crossmatch the cell types in the two datasets. **(B)** Cell-based and cluster-based cell type matching methods. Azimuth, CellTypist, and scArches perform cell-based cell type matching, i.e., from query cell to reference cell, and assign the reference cell label to the query cell, using models pre-trained on a pre-defined reference dataset. The lung models in these three methods were pre-trained using the HLCA dataset as the reference. FR-Match has both cell-based and cluster-based matching options: cell-to-cluster is from query cell to reference cluster and cluster-to-cluster is from query cluster to reference cluster. The FR-Match approach has a flexible choice of reference dataset. If both datasets are optimally-clustered, FR-Match

cluster-to-cluster can perform supervised two-way matching (**Methods**). **(C)** A sustainable workflow for continued integration of reference atlases requires a strategy that identifies new cell types from new datasets to be integrated to the reference atlas while preserving the cell type memberships of existing cells, to support the incremental growth of the knowledgebase.

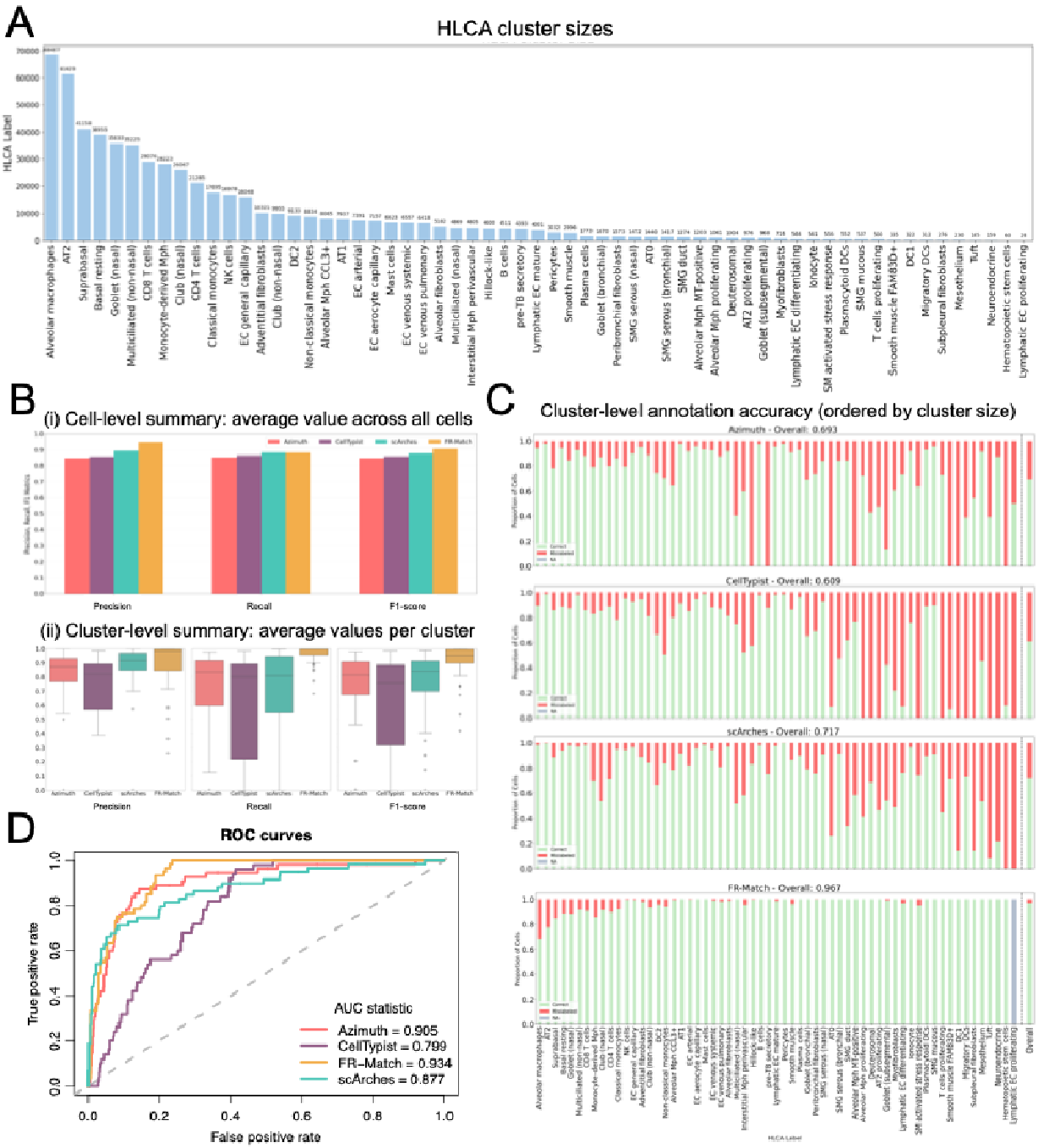


Figure 2. Benchmarking performance of cell-based annotation/matching methods in the HLCA 10-fold cross-validation. (A) Distribution of cluster sizes of the cell types represented in the HLCA dataset. **(B)** Precision, recall, and F1-score for each method (Azimuth, CellTypist, scArches, and FR-Match cell-to-cluster): (i) overall performance averaged by all cells and (ii) performance by each cluster. **(C)** Bar plots of cluster-level annotation accuracy for each method, sorted by cluster size. Note that due to a lack of cells in the smallest cluster after 10-fold split (less than 3 cells in each fold), FR-Match filtered out such cluster as having too few cells (see **Methods**). **(D)** ROC curve and AUC values for each method.

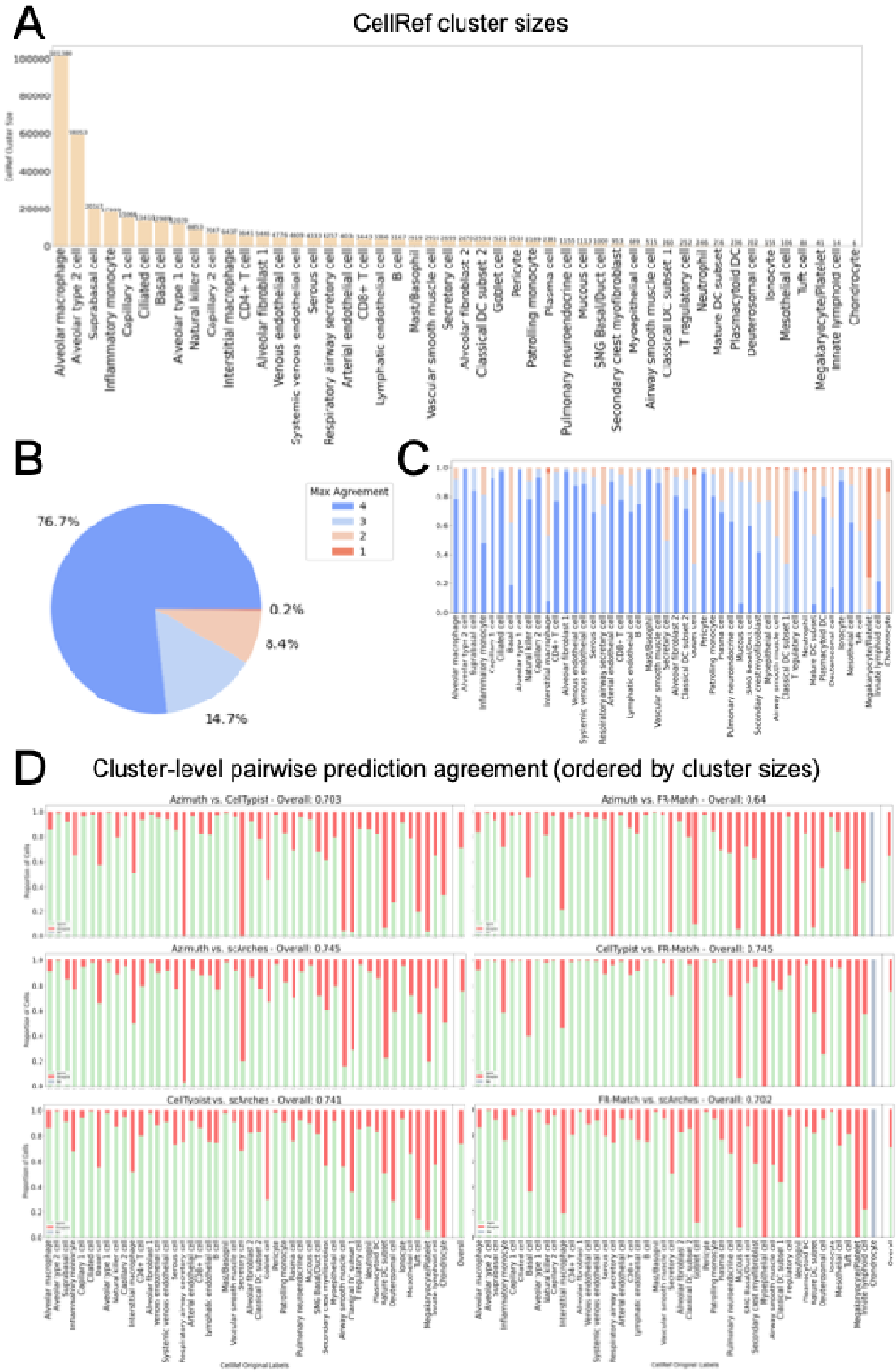

A
CellRef cluster sizes
CellRef Cluster Size
Alveolar macrophage
Alveolar type 2 cell
Suprabasal cell
Inflammatory monocyte
Capillary 1 cell
Ciliated cell
Basal cell
Alveolar type 1 cell
Natural killer cell
Capillary 2 cell
Interstitial macrophage
CD4+ T cell
Alveolar fibroblast 1
Venous endothelial cell
Systemic venous endothelial cell
Serous cell
Respiratory airway secretory cell
Arterial endothelial cell
CD8+ T cell
Lymphatic endothelial cell
B cell
Mast/Basophil
Vascular smooth muscle cell
Secretory cell
Alveolar fibroblast 2
Classical DC subset 2
Goblet cell
Pericyte
Patrolling monocyte
Plasma cell
Pulmonary neuroendocrine cell
Mucous cell
SMG Basal/Duct cell
Secondary crest myofibroblast
Myoepithelial cell
Airway smooth muscle cell
Classical DC subset 1
T regulatory cell
Neutrophil
Mature DC subset
Plasmacytoid DC
Deuterosomal cell
Ionocyte
Mesothelial cell
Tuft cell
Megakaryocyte/Platelet
Innate lymphoid cell
Chondrocyte
B
Max Agreement
4
3
2
1
76.7%
0.2%
8.4%
14.7%
C
D
Cluster-level pairwise prediction agreement (ordered by cluster sizes)
Azimuth vs. CellTypist - Overall: 0.703
Azimuth vs. FR-Match - Overall: 0.64
Azimuth vs. scArches - Overall: 0.745
CellTypist vs. FR-Match - Overall: 0.745
CellTypist vs. scArches - Overall: 0.741
FR-Match vs. scArches - Overall: 0.702
Proportion of Cells
CellRef Original Labels
Overall

Figure 3. Benchmarking of the cell-based methods across the CellRef and HLCA datasets. The cell type annotation/matching is from CellRef as the query dataset to HLCA as the reference dataset (CellRef → HLCA). **(A)** Distribution of cluster sizes of cell types represented in the CellRef dataset. **(B)** Summary of level of prediction agreement across all methods (Azimuth, CellTypist, scArches, and FR-Match cell-to-cluster) at the cell-level. **(C)** Stacked bar plot depicting level of prediction agreement across all methods at the cluster-level, ordered by cluster size. **(D)** Proportion of agreement between two methods' predictions for each cluster, ordered by cluster size. Note that chondrocyte was filtered out in FR-Match due to its small cluster size.

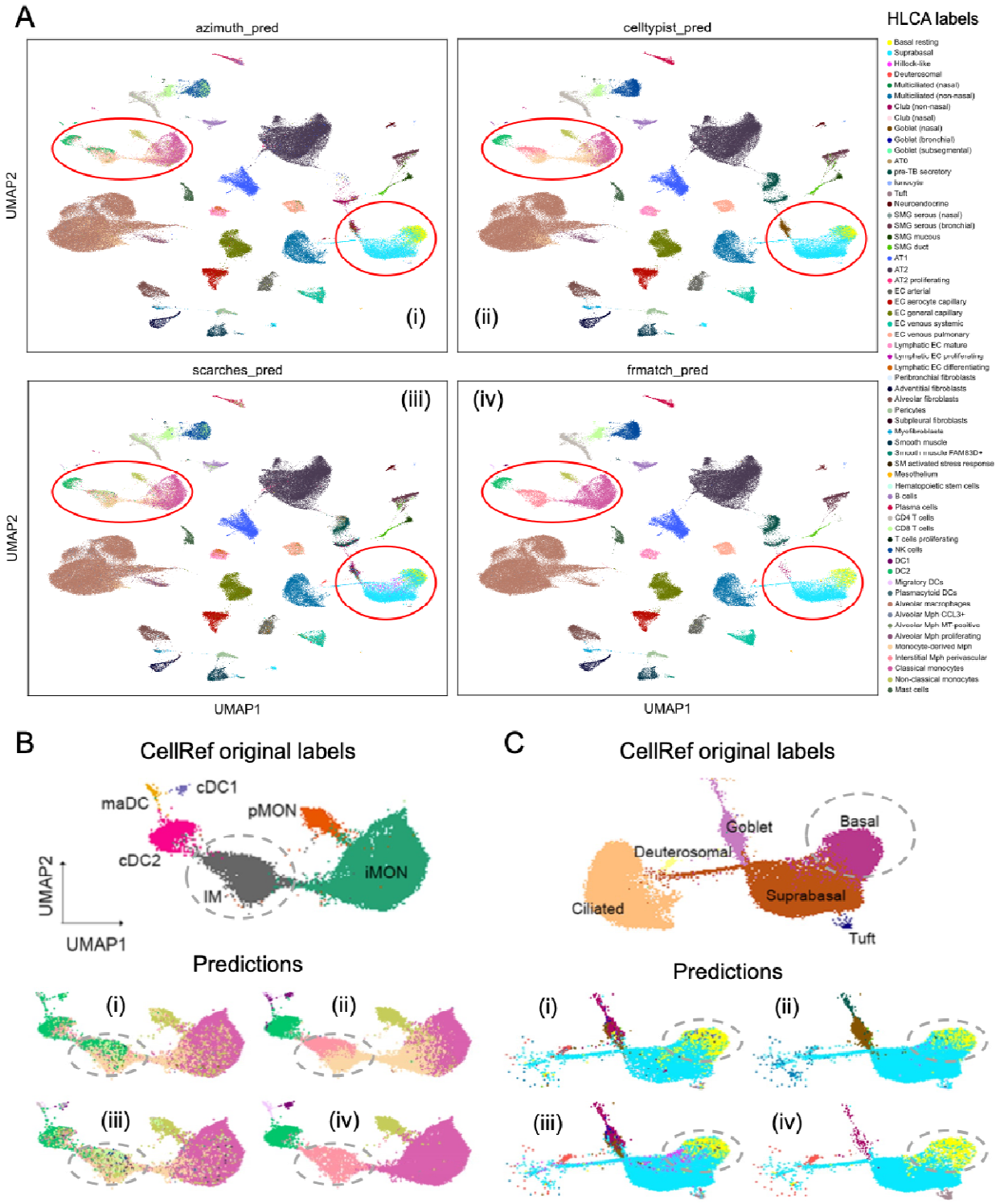


Figure 4. UMAP of cell-based matching results of CellRef cells to the HLCA reference. (A) CellRef cells in the CellRef UMAP embeddings are colored by HLCA labels predicted by Azimuth (i), CellTypist (ii), scArches (iii), and FR-Match cell-to-cluster (iv). Red circles highlight regions with major differences. **(B-C)** Detailed views of the highlighted regions in (A) for

comparison of the CellRef original labels and the HLCA predictions for these cells. Methods used in (i)-(iv) are the same as in (A). The dotted circles indicate the major differences in interstitial macrophage (IM) and basal cell. Cells with the original label of cDC1, maDC, cDC2, IM, iMON, pMON are plotted with predictions in the “IM” region plot in (B). Cells with the original label of deuterosomal, goblet, suprabasal, tuft, and basal are plotted with predictions in the “basal” region plot in (C).

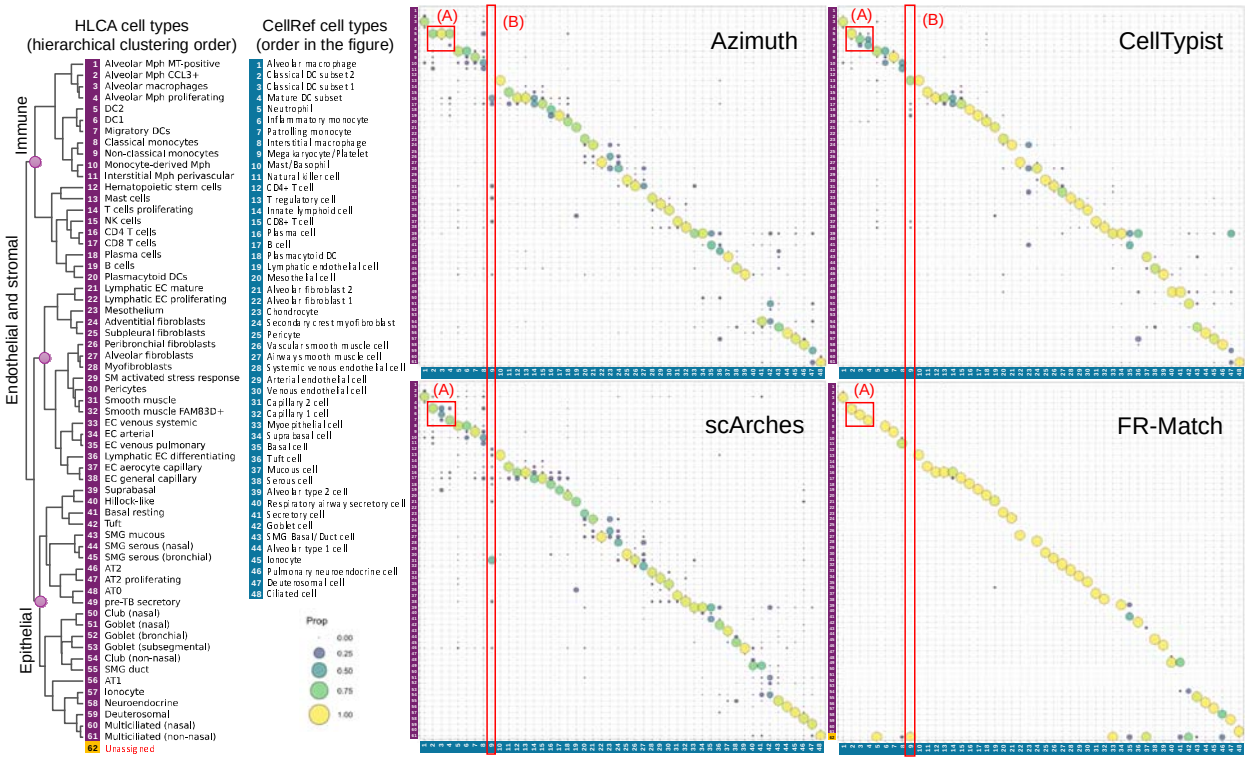


Figure 5. Matching between the 48 cell types in CellRef and the 61 cell types in HLCA. Cell-based matching results are summarized at the cluster-level. The 48 CellRef cell types are on the horizontal axis; the 61 HLCA cell types are on the vertical axis. Each circle in the plot is the proportion of cells in the original CellRef cluster matched to each of the corresponding HLCA clusters. Since the cell-based matching is one directional (CellRef as query to HLCA as reference), the proportions in each column sum up to 1. The dendrogram to the left shows the hierarchical clustering order of the HLCA cell types, which groups similar cell types to each other based on their transcriptional profiles (**Methods**). The CellRef cell types are ordered according to the HLCA dendrogram order so that the matching proportions in the figure align close to the diagonal; an off-diagonal circle suggests a mismatch. In the FR-Match figure, the "unassigned" cells are collected in the bottom row, indicating the statistically weak matching of these cells. The highlighted boxes are examples of different matching patterns by the methods for dendritic cell types (A) and for "unassigned" cells (B).

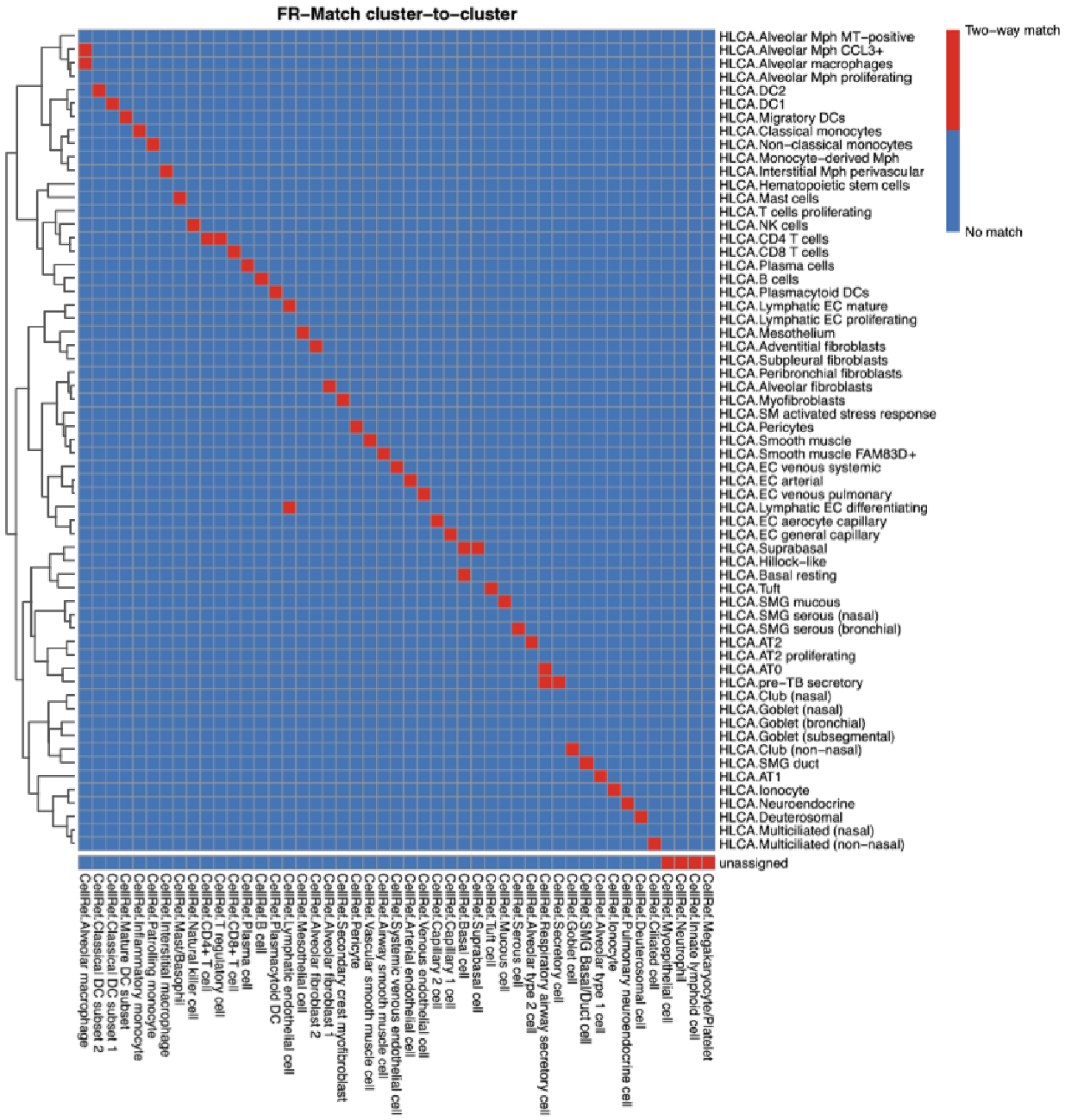
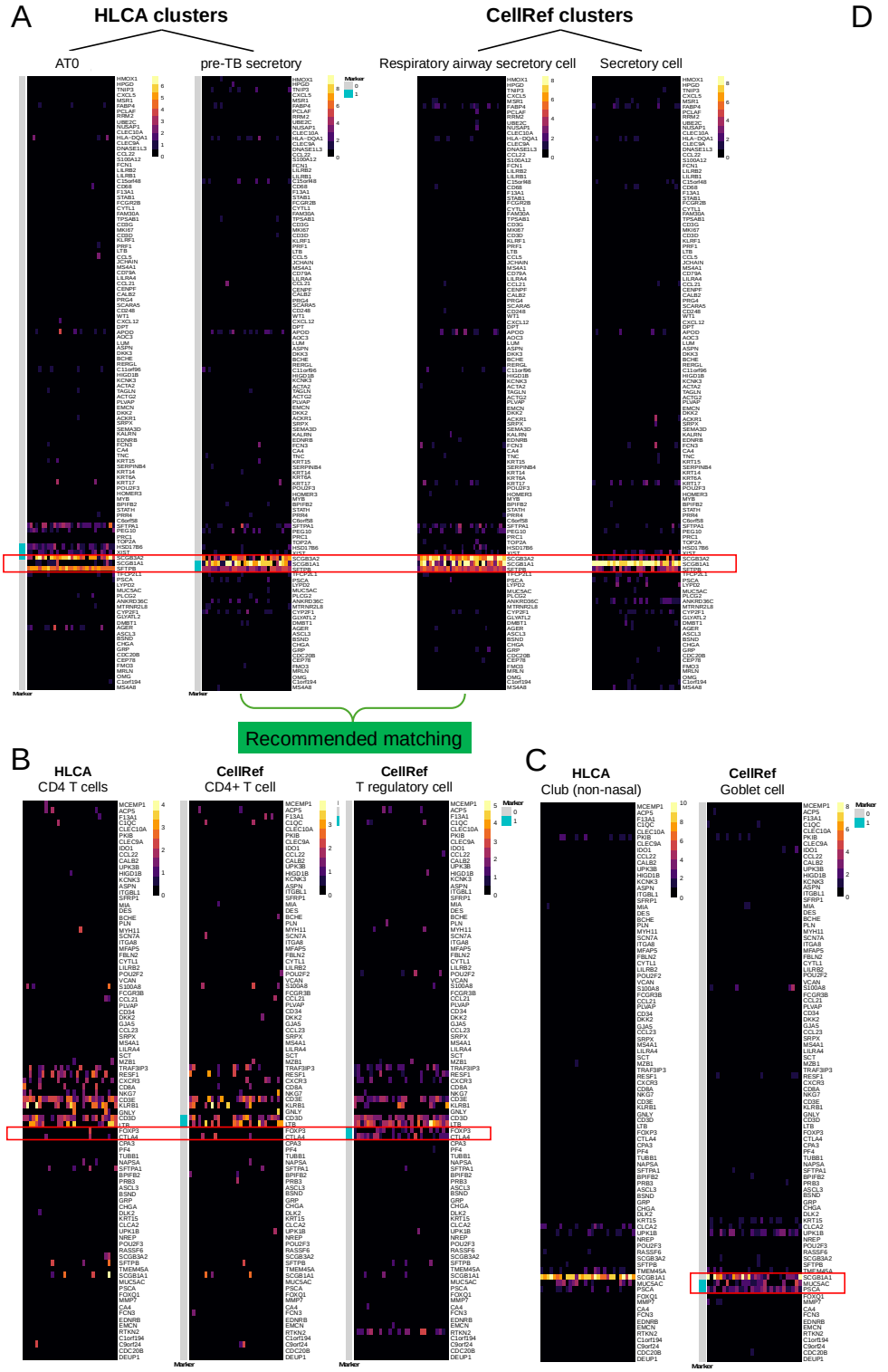


Figure 6. FR-Match cluster-to-cluster results of the CellRef ←→ HLCA cell type matching. Two-way matches of the CellRef and HLCA cell types. CellRef as query to HLCA as reference and HLCA as query to CellRef as reference were both performed by FR-Match. The red blocks indicate two-way matches where the matched/unassigned cell types are agreed by both matching directions.

HLCA	CellRef	Match	HLCA_ClusterName	CellRef_ClusterName	Cluster-based matching	Cell-based matching				RecommendedMatching
					TwoWayMatch_FRMatch	HiConfMatch_Azimuth	HiConfMatch_CellTypist	HiConfMatch_scArches	HiConfMatch_FRMatch	
3	1	1	Alveolar macrophages	Alveolar macrophage	Y	Y	Y	Y	Y	HLCA2CellRef_Match
8	6	2	Classical monocytes	Inflammatory monocyte	Y	Y	Y	Y	Y	HLCA2CellRef_Match
9	7	3	Non-classical monocytes	Patrolling monocyte	Y	Y	Y	Y	Y	HLCA2CellRef_Match
13	10	4	Mast cells	Mast/Basophil	Y	Y	Y	Y	Y	HLCA2CellRef_Match
15	11	5	NK cells	Natural killer cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
16	12	6	CD4 T cells	CD4+ T cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
17	15	7	CD8 T cells	CD8+ T cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
18	16	8	Plasma cells	Plasma cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
19	17	9	B cells	B cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
20	18	10	Plasmacytoid DCs	Plasmacytoid DC	Y	Y	Y	Y	Y	HLCA2CellRef_Match
21	19	11	Lymphatic EC mature	Lymphatic endothelial cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
23	20	12	Mesothelium	Mesothelial cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
24	21	13	Adventitial fibroblasts	Alveolar fibroblast 2	Y	Y	Y	Y	Y	HLCA2CellRef_Match
27	22	14	Alveolar fibroblasts	Alveolar fibroblast 1	Y	Y	Y	Y	Y	HLCA2CellRef_Match
28	24	15	Myofibroblasts	Secondary crest myofibroblast	Y	Y	Y	Y	Y	HLCA2CellRef_Match
30	25	16	Pericytes	Pericyte	Y	Y	Y	Y	Y	HLCA2CellRef_Match
31	26	17	Smooth muscle	Vascular smooth muscle cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
33	28	18	EC venous systemic	Systemic venous endothelial cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
34	29	19	EC arterial	Arterial endothelial cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
35	30	20	EC venous pulmonary	Venous endothelial cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
37	31	21	EC aerocyte capillary	Capillary 2 cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
38	32	22	EC general capillary	Capillary 1 cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
45	38	23	SMG serous (bronchial)	Serous cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
46	39	24	AT2	Alveolar type 2 cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
55	43	25	SMG duct	SMG Basal/Duct cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
56	44	26	AT1	Alveolar type 1 cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
57	45	27	Ionocyte	Ionocyte	Y	Y	Y	Y	Y	HLCA2CellRef_Match
58	46	28	Neuroendocrine	Pulmonary neuroendocrine cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
61	48	29	Multiciliated (non-nasal)	Ciliated cell	Y	Y	Y	Y	Y	HLCA2CellRef_Match
5	2	30	DC2	Classical DC subset 2	Y	N	Y	Y	Y	HLCA2CellRef_Match
6	3	31	DC1	Classical DC subset 1	Y	N	Y	N	Y	HLCA2CellRef_Match
7	4	32	Migratory DCs	Mature DC subset	Y	N	Y	Y	Y	HLCA2CellRef_Match
11	8	33	Interstitial Mph perivascular	Interstitial macrophage	Y	N	N	N	Y	HLCA2CellRef_Match
32	27	34	Smooth muscle FAM83D+	Airway smooth muscle cell	Y	N	Y	Y	Y	HLCA2CellRef_Match
39	34	35	Suprabasal	Suprabasal cell	Y	Y	N	Y	Y	HLCA2CellRef_Match
41	35	36	Basal resting	Basal cell	Y	Y	Y	N	Y	HLCA2CellRef_Match
42	36	37	Tuft	Tuft cell	Y	Y	N	Y	Y	HLCA2CellRef_Match
43	37	38	SMG mucous	Mucous cell	Y	Y	Y	Y	N	HLCA2CellRef_Match
49	40	39	pre-TB secretory	Respiratory airway secretory cell	Y	N	Y	Y	Y	HLCA2CellRef_Match
54	42	40	Club (non-nasal)	Goblet cell	Y	N	N	N	N	HLCA2CellRef_boardMatch
59	47	41	Deuterosomal	Deuterosomal cell	Y	Y	N	Y	Y	HLCA2CellRef_Match
1		42	Alveolar Mph MT-positive		N	N	N	N	N	HLCA-specific
2		43	Alveolar Mph CCL3+		Y (Alveolar macrophage)	N	N	N	N	HLCA-specific
4		44	Alveolar Mph proliferating		N	N	N	N	N	HLCA-specific
10		45	Monocyte-derived Mph		N	N	N	N	N	HLCA-specific
12		46	Hematopoietic stem cells		N	N	N	N	N	HLCA-specific
14		47	T cells proliferating		N	N	N	N	N	HLCA-specific
22		48	Lymphatic EC proliferating		N	N	N	N	N	HLCA-specific
25		49	Subpleural fibroblasts		N	N	N	N	N	HLCA-specific
26		50	Peribronchial fibroblasts		N	N	N	N	N	HLCA-specific
29		51	SM activated stress response		N	N	N	N	N	HLCA-specific
36		52	Lymphatic EC differentiating		Y (Lymphatic endothelial cell)	N	N	N	N	HLCA-specific
40		53	Hillock-like		N	N	N	N	N	HLCA-specific
44		54	SMG serous (nasal)		N	N	N	N	N	HLCA-specific
47		55	AT2 proliferating		N	N	N	N	N	HLCA-specific
48		56	AT0		Y (Respiratory airway secretory cell)	N	N	N	N	HLCA-specific
51		57	Goblet (nasal)		N	N	Y (Goblet)	N	N	HLCA-specific
50		58	Club (nasal)		N	N	N	N	N	HLCA-specific
52		59	Goblet (bronchial)		N	N	N	N	N	HLCA-specific
53		60	Goblet (subsegmental)		N	N	N	N	N	HLCA-specific
60		61	Multiciliated (nasal)		N	N	N	N	N	HLCA-specific
	5	62		Neutrophil	N	Y (Classical monocytes)	Y (Classical monocytes)	Y (Classical monocytes)	N	CellRef-specific
	9	63		Megakaryocyte/Platelet	N	N	Y (Mast cells)	Y (Smooth muscle)	N	CellRef-specific
	13	64		T regulatory cell	Y (CD4 T cells)	Y (CD4 T cells)	Y (CD4 T cells)	Y (CD4 T cells)	Y (CD4 T cells)	CellRef-specific
	14	65		Innate lymphoid cell	N	N	N	N	Y (CD4 T cells)	CellRef-specific
	23	66		Chondrocyte	NA	N	N	N	NA	CellRef-specific
	33	67		Myoepithelial cell	N	Y (Suprabasal)	N	Y (Suprabasal)	N	CellRef-specific
	41	68		Secretory cell	Y (pre-TB secretory)	Y (Club (non-nasal))	Y (pre-TB secretory)	Y (pre-TB secretory)	Y (pre-TB secretory)	CellRef-specific

Figure 7. Recommended matching of the lung cell types between CellRef and HLCA. 41 recommended cell type matches (cluster-based matching + cell-based matching) out of 61 HLCA cell types and 48 CellRef cell types. An additional 20 cell types are unique to HLCA and 7 cell types are unique to CellRef. The complete table with characterizing marker genes is in **Supplementary Table 2**.



D

HLCA markers	CellRef markers
HMOX1	MCEMP1
HPGD	ACP5
TNIP3	F13A1
CXCL5	C1QC
MSR1	CLEC10A
FABP4	PKIB
PCLAF	CLEC9A
RRM2	IDO1
UBE2C	CCL22
NUSAP1	CALB2
CLEC10A	UPK3B
HLA−DQA1	HIGD1B
CLEC9A	KCNK3
DNASE1L3	ASPN
CCL22	ITGBL1
S100A12	SFRP1
FCN1	MIA
LILRB2	DES
LILRB1	BCHE
C15orf48	PLN
CD68	MYH11
F13A1	SCN7A
STAB1	ITGA8
FCGR2B	MFAP5
CYTL1	FBLN2
FAM30A	CYTL1
TPSAB1	LILRB2
CD3G	POU2F2
MKI67	VCAN
CD3D	S100A8
KLRF1	FCGR3B
PRF1	CCL21
LTB	PLVAP
CCL5	CD34
JCHAIN	DKK2
MS4A1	GJA5
CD79A	CCL23
LILRA4	SRPX
CCL21	MS4A1
CENPF	LILRA4
CALB2	SCT
PRG4	MZB1
SCARA5	TRAF3IP3
CD248	RESF1
WT1	CXCR3
CXCL12	CD8A
DPT	NKG7
APOD	CD3E
AOC3	KLRB1
LUM	GNLY
ASPN	CD3D
DKK3	LTB
BCHE	FOXP3
RERGL	CTLA4
C11orf96	CPA3
HIGD1B	PF4
KCNK3	TUBB1
ACTA2	NAPSA
TAGLN	SFTPA1
ACTG2	BPIFB2
PLVAP	PRB3
EMCN	ASCL3
DKK2	BSND
ACKR1	GRP
SRPX	CHGA
SEMA3D	DLK2
KALRN	KRT15
EDNRB	CLCA2
FCN3	UPK1B
CA4	NREP
TNC	POU2F3
KRT15	RASSF6
SERPINB4	SCGB3A2
KRT14	SFTPB
KRT6A	TMEM45A
KRT17	SCGB1A1
POU2F3	MUC5AC
HOMER3	PSCA
MYB	FOXQ1
BPIFB2	MMP7
STATH	CA4
PRR4	FCN3
C6orf58	EDNRB
SFTPA1	EMCN
PEG10	RTKN2
PRC1	C1orf194
TOP2A	C9orf24
HSD17B6	CDC20B
XIST	DEUP1
SCGB3A2	
SCGB1A1	
MUC5AC	
PSCA	
TFCP2L1	
LYPD2	
MUC5AC	
PLCG2	
ANKRD36C	
MTRNR2L8	
CYP2F1	
GLYATL2	
DMBT1	
AGER	
ASCL3	
BSND	
CHGA	
GRP	
CDC20B	
CEP78	
FMO3	
MRLN	
OMG	
C1orf194	
MS4A8	

Figure 8. Use of NS-Forest marker genes to disambiguate potential matches. Each heatmap shows the gene expression pattern of 30 randomly selected cells (columns) from the plotting cell type in the NS-Forest marker gene space (rows). **(A)** The recommended matching of HLCA pre-TB secretory and CellRef respiratory airway secretory cell are

SCGB3A2^{high}*SCGB1A1*^{high}*SFTPB*^{high}, whereas the HLCA AT0 is *SCGB3A2*^{high}*SCGB1A1*^{low}*SFTPB*^{high}, and the CellRef secretory cell is *SCGB3A2*^{low}*SCGB1A1*^{high}*SFTPB*^{low} **(B)** Distinction of T regulatory cells from CD4+ T cells by markers *FOXP3* and *CTLA4*. **(C)** The CellRef goblet cell type is likely under-partitioned, containing both club cells expressing *SCGB1A1* and goblet cells expressing *MUC5AC.* **(D)** NS-Forest marker gene spaces of HLCA and CellRef.

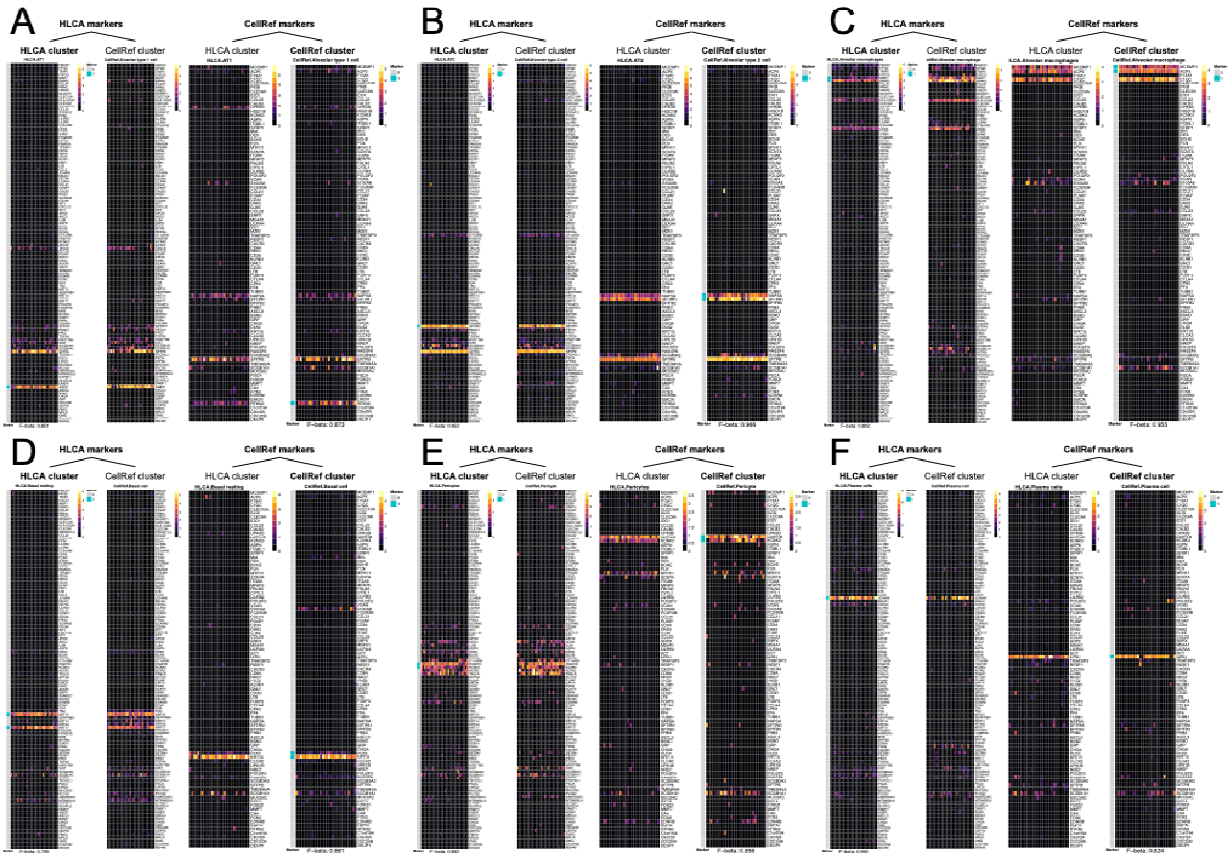


Figure 9. Examples of NS-Forest cell type "barcode" plots of recommended matching pairs in the reciprocal atlases. Cell type "barcode" plots of AT1 (A), AT2 (B), alveolar macrophages (C), basal resting (D), pericytes (E), and plasma cells (F) in HLCA, and the corresponding recommended matching in CellRef show highly similar expression patterns in both marker gene spaces of HLCA and CellRef.