Automatic cell type harmonization and integration across Human Cell Atlas datasets

Chuan Xu[1], Martin Prete[1], Simone Webb[2,1], Laura Jardine[2], Benjamin Stewart[1,3,4], Regina Hoo[1], Peng He[1,5], Sarah A. Teichmann[1,6,7,*]

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK
[2]Biosciences Institute, Newcastle University, NE2 4HH, UK
[3]Molecular Immunity Unit, Department of Medicine, University of Cambridge, Cambridge, CB2 0QQ, UK
[4]Cambridge University Hospitals NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge CB2 0QQ, UK
[5]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge, UK
[6]Department of Physics, Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, UK
[7]Lead contact
*Correspondence: st9@sanger.ac.uk (S.A.T)

Summary

Harmonizing cell types with respect to transcriptomics identity and nomenclature across the single-cell community and assembling them into a common framework is an essential step towards building a standardized Human Cell Atlas. Here we present CellTypist v2.0 (a new version of our automated annotation tool, *https://github.com/Teichlab/celltypist*), where we develop a predictive clustering tree-based approach to resolve cell type differences across datasets that have different naming conventions, annotation resolution, and technical biases. CellTypist v2.0 accurately quantifies cell-cell transcriptomic similarities and enables robust and efficient cross-dataset meta-analyses. Cell types are placed into a relationship graph that hierarchically defines shared and novel cell subtypes. Application to multiple immune datasets confirms its sensitivity and specificity by recapitulating expert-curated cell annotations. We also apply CellTypist to datasets from eight diseases which all lead to pulmonary fibrosis, and reveal underexplored relationships between healthy cell types and a variety of diseased cell states. Furthermore, we present a workflow for fast cross-dataset integration guided by the harmonized cell types at different levels of annotation granularity. Finally, we apply CellTypist to 12 tissue atlases from 38 datasets, and provide a deeply curated cross-tissue database (*www.celltypist.org/organs*) with more than 3.6 million cells and 250 cell types. This is complemented by a large collection of machine learning models for automatic cell type annotation across human tissues.

Introduction

The past decade has seen an accumulation of single-cell genomics datasets profiling a variety of tissues at different developmental stages. By integrating these resources, international consortia such as the Human Cell Atlas (HCA) have begun to chart a standardized reference map in the human body[1,2]. A challenge in assembling distinct cell atlas datasets is that different laboratories use their own definitions of cell types, often resulting in inconsistencies in the naming schema. This challenge is being partially met by the ongoing Cell Ontology database effort, which aims at unifying cell type labels across communities[3]. Other contributions come from the NIH Human BioMolecular Atlas Program (HuBMAP)[4] project and its ASCT+B tables that name and interlink major anatomical structures, cell types and biomarkers[5], and the Cell Annotation Platform (a collaborative effort funded by Schmidt Futures, https://celltype.info). However, standardizing cell types across datasets, denoted here as the process of cell type harmonization, has intrinsic challenges beyond nomenclature. These include the variability in cell quality caused by various technical confounders, presence of shared versus novel cell types across datasets, and the discrepancy in annotation resolution due to differences in cell sampling depth and in laboratory-specific annotation strategies.

A number of label transfer methods have been developed to match cell types across datasets[6–13]. A limitation of these methods is that they are only applicable to two datasets (reference and query). Aligning multiple datasets simultaneously is still arduous as it requires multiple pairwise comparisons and brings challenges in information integration and consistent visualization. A further limitation lies in the nature of machine learning-based models that, without a prebuilt cell type hierarchy, neglect the relationships among cell types both within and across datasets and usually lead to “flat” classifiers.

Furthermore, current efforts for single-cell data integration largely focus on correcting batch effects across datasets[9,14–18]. Information relating to cell types annotated by individual studies is not fully utilized in batch effect regression owing to naming inconsistencies. Once cell types are successfully harmonized across datasets, data integration can be supervised in an annotation-aware manner. Although deep learning-based methods such as scANVI[19] and scGen[20] are able to include cell type information in the integration pipeline, further efforts are needed to control for the degree of intervention from cell annotations, as well as to incorporate cell type relationships and cell hierarchy into the data integration process.

To meet these challenges, we develop an automated workflow for cell type harmonization connected with cross-dataset integration. Building on the infrastructure in CellTypist v1.0[8], we develop a predictive clustering tree (PCT)-based tool in CellTypist v2.0 to efficiently align multiple datasets by assessing their cell-cell similarities and harmonizing cell annotations. Based on this, CellTypist defines semantic relationships among cell types and captures their underlying biological hierarchies, which are further leveraged to guide the downstream data integration at different levels of annotation granularities. We have applied this pipeline to 43 datasets, confirmed its effectiveness in data harmonization and integration, and provided a collection of organ atlases and machine learning models to the wider community for automatic cell annotation.

Results

An automated workflow for cell type harmonization and integration

To infer cell type relationships across datasets, we use PCT, a multi-target regression tree algorithm. Different from classical multi-output decision trees, with PCT the dependency among cell types in the reference datasets is considered, and dissimilarities between query cells and reference cell types are predicted to be the prototypes of the matched tree leaves (**Figure 1**). To reduce runtime and avoid overfitting, during the top-down induction of PCT in each dataset, the clustering tree is pruned at the cell taxonomy (i.e., node) where further splitting will result in two homogeneous cell states based on an F-test. By iteratively building such a tree using each dataset as a reference and further extrapolating it to cells from the remaining datasets, CellTypist derives a global matrix representing the inferred relationships between all cells and cell types. In other words, once a PCT is built for the reference dataset, each query cell will fall into a given leaf of the reference tree, and this information is used to predict its dissimilarity profile across reference cell types, thereby effectively scaling the query datasets to the reference. Since this algorithm has no assumption about gene expression distribution and cell-cell distances are predicted within the span of the clustering trees, CellTypist is able to produce batch-insensitive dissimilarity measures, enabling a robust cross-dataset meta-analysis.

The resulting relationships between cells and cell types are integrated into a harmonization graph by considering all possible pairs of reference-to-query predictions (**Figure 1**). Specifically, we define a dataset core which summarizes original cell types into

semantically connected ones including matched and divisible cell types. Through iterative addition of a new dataset, this core is gradually enlarged to re-define an expanded cell type repertoire, namely, the CellTypist harmonization table (**Figure 1; Methods**). The advantage of this procedure is that the order of datasets added has a minimal impact on the structure of the harmonization graph. As a result, cell types in the graph can be reorganized to approximate the underlying biological hierarchies including the high- and low-hierarchies (**Figure 1**). This also represents an avenue to integrate knowledge across studies by connecting their manual (and biologically meaningful) groupings. Moreover, to improve sensitivity in discovering dataset-specific cell types, CellTypist defines two levels of novelties for cell types in a given dataset: unmatched cell types ("NONE") which cannot align with any cell type from the other datasets, and unharmonized cell types ("UNRESOLVED") which fail to integrate into the harmonization graph after the final iteration.

CellTypist further utilizes the cell type relationships to supervise the single-cell data integration (**Figure 1**). Given the harmonized cell types across datasets, the neighborhood graph is built by searching the neighbors of each cell in a restricted manner within the same cell type group which is determined by combining the cell type hierarchy with the inferred transcriptomic similarities (**Methods**). By controlling the size of cell type groups, CellTypist can adjust the extent of supervision from cell annotations. Data scalability and efficiency are also anticipated to be improved as the neighborhood search space is dramatically narrowed down.

CellTypist harmonization recapitulates expert-curated annotations

To validate the harmonization pipeline in CellTypist, we selected five immune datasets which were initially included in the CellTypist v1.0 database and deeply curated by experts into consistent cell types[8]. These datasets had varying cell type compositions, developmental stages, tissue coverages, and transcriptome sparsities, presenting a real-world scenario for cell type harmonization (**Figure S1A**). As expected, we observed substantial batch effects when visualizing the 417,866 cells onto the uniform manifold approximation and projection (UMAP)[21] (**Figure S1B**). We thus asked whether CellTypist can harmonize the cell types of the datasets to recapitulate the ground-truth annotations (**Figure S1C**).

Despite the various confounding factors present in the five datasets, CellTypist was able to reconstruct their cell type relationships from two complementary angles. Firstly, CellTypist predicted the transcriptomic distances among the 129 original cell types, and based on this, revealed a high degree of cross-dataset cell type replicability (**Figure S2**). Specifically, unsupervised hierarchical clustering using the inferred distances not only grouped the same cell types from different datasets regardless of their batch confounders, but also demonstrated a biological structure comprising compartments of hematopoietic progenitors, myeloid, and lymphoid cells. This meta-analysis also uncovered finer structures underlying these datasets such as the innate lymphoid cells and T cells, the latter of which were further subdivided into CD4+ αβ, CD8+ αβ, and γδ T cells (**Figure S2**). Secondly, CellTypist harmonized the 129 cell types into 66 at the low-hierarchy level, identifying both dataset-shared and -specific cell types (**Figures 2A and S3A**). Reorganization of the harmonization graph further categorized them into 52 high-hierarchy cell groups (**Figures 2B and S3B**).

As well as the harmonizations that could be confirmed at the nomenclature level (e.g., "DC_2", "DC2", and "cDC2"), CellTypist resolved complex cell type alignments not obvious based on their nomenclatures. For example, CellTypist harmonized the eight annotations of CD8+ αβ T cells into three cell states (**Figure 2C**). These included a reassuring match between "Tnaive/CM_CD8" and "naive CD8 T cell", as well as two matches not easily resolvable using nomenclature only: between "T_CD8_CTL" and "Tem/emra_CD8"; and among "T_CD8_activated", "Trm/em_CD8", "CD8 T", and "mature CD8 T cell". Examination of marker gene expression validated these cell identities aligned (**Figure 2C**). Similarly, for natural killer (NK) cells, CellTypist harmonized seven NK annotations into two subtypes as supported by their respective gene signatures, that is, the CD16+ ("NK_FCGR3Apos", "NK_CD16+", and "mature NK") and CD16- ("NK_CD160pos", "NK_CD56bright_CD16-", "NK", and "CD56 bright NK") NK cells (**Figure S4A**).

A number of cell types were hierarchically harmonized, leading to the identification of cell subtypes not annotated as such in the original publications (**Figures 2A and S3A**). In one of the datasets, we previously described a specialized and rare memory B cell population, age-associated B cells (ABCs), and extrapolated its presence to other datasets by means of manual inspection[8]. Using the automated harmonization algorithm in CellTypist, we reproduced these findings by corroborating the existence of ABCs in two other datasets profiling the adult spleen and bone marrow, respectively[22] (**Figure 2D**). The other example was the "B cell" population identified in the fetal liver[23], which we demonstrated to consist of a spectrum of transitional

subpopulations from immature to naive B cells (**Figure S4B**). Moreover, the harmonization module in CellTypist is able to detect homogeneous subtypes which may not readily be separated using clustering-based approaches. In line with our manually curated results, we found three transcriptionally similar myeloid populations ("monocyte-DC", "CD14 monocyte", and "CD16 monocyte") collectively forming the "Macrophage" annotated in the gut dataset[24] (**Figure 2E**), as well as three distinct CD4+ T cell populations ("Tnaive/CM_CD4", "Tfh", and "Teffector/EM_CD4") constituting a single heterogeneous "T_CD4_fh" cell type in a spleen dataset[22] (**Figure S4C**).

Interestingly, CellTypist discovered new cell types that were previously overlooked. For instance, the γδ T cells ("gd T") initially annotated in the gut dataset[24] were revealed by CellTypist to contain a hidden population of epithelium-resident CD8+ αβ T cells ("Trm_gut_CD8") which closely resembled the γδ T cells (**Figure S4D**). This was also the case for plasma cells defined in two datasets[8,22] where CellTypist hierarchically divided them into the IgA ("B cell IgA Plasma") and IgG ("B cell IgG Plasma") plasma subpopulations based on another dataset[24] (**Figure S4E**).

To further test the stability of our algorithm in multi-dataset alignment, we permuted the order of datasets incorporated during harmonization and confirmed the robustness of the aforementioned observations (**Figure S5**). Taken together, through comparisons with expert-graded annotations across multiple immune datasets, we showcase the capacity of CellTypist to perform meta-analysis, harmonize cell types, define cell hierarchies, and inform potential new biology.

CellTypist disentangles disease-enriched cell states

Cell diversity is often increased in pathological conditions due to the emergence of novel cell types or molecular reorganization of the same cell types[25,26]. Resolving the cell state complexity in disease and ascertaining its association with the underlying molecular alterations are important for understanding pathogenesis. To test the utility of CellTypist in these contexts, we applied it to four single-cell datasets[27–30] profiling the adult human lung, collectively covering 479,135 cells from 130 individuals in health plus eight diseases resulting in pulmonary fibrosis: idiopathic pulmonary fibrosis (IPF), chronic obstructive pulmonary disease (COPD), nonspecific interstitial pneumonia (NSIP), chronic hypersensitivity

pneumonitis (cHP), interstitial lung disease (ILD), sarcoidosis (SA), systemic sclerosis (SSc), and polymyositis (PM) (**Figure S6A**).

Prominent technical batch effects existed across these datasets, making the discovery of disease patterns challenging (**Figures S6B-S6D**). Despite these batch effects, with CellTypist, epithelial and stromal cell types in the four datasets were consistently harmonized, in a manner that discriminated the transcriptomic differences not only among cell subtypes but also among disease statuses (**Figures 3A, 3B, and S7**). For example, the basal, ciliated, and alveolar epithelial cell types were clearly separated into healthy and diseased cell states, with most of the disease statuses further distinguishable in the harmonization graph. Conversely, in the immune cell compartment, many cell types were harmonized irrespective of their disease contexts, such as the partial alignment between normal and COPD mast cells, as well as among NK cells in normal, COPD, NSIP and IPF conditions (**Figure S8**). These observations were in line with the clinical manifestations of pulmonary fibrosis by which epithelial and stromal cells were most affected[31], while the relatively weaker transcriptomic changes in immune cell types were insufficient to drive a cell identity transition.

The harmonization graph also allowed us to locate cell types and subtypes missed by individual studies and to interpret pathological cell states by extending their contexts across studies. We observed that the aberrant basaloid cells ("Aberrant_Basaloid"), a pathogenic population lining the myofibroblast foci in the IPF lung[27], were aligned with the distal alveolar epithelial population ("KRT5-/KRT17+") enriched in pulmonary fibrosis[29] (**Figure 4A**), suggestive of a single aberrant population named differently between studies. This was further corroborated by an independent analysis using logistic regression-based label transfer. In addition, the harmonization graph indicated mixing of this population into the "AT1" cell type of a third dataset[30]. *In silico*, we were able to resolve these populations and verified their presence by combinatorial expression of multiple marker genes (**Figure 4A**). Similarly, goblet cells in the diseased state from two of the datasets[27,28] were demonstrated to be heterogeneous, divisible into *MUC5B+* and *SCGB3A2+SCGB1A1+* populations (**Figure 4B**). CellTypist can further refine the original cell annotations to match their underlying biology. For instance, the IPF "Fibroblast" in *Adams et al*[27] was shown to align with the lipofibroblast-like subpopulation, as supported by its overexpression of *HAS1* and *PLIN2* (**Figure S9A**). The IPF "Fibroblast" in *Morse et al*[28], on the other hand, was *ACTA2+* and corresponded to the activated myofibroblasts during pulmonary fibrosis (**Figure 4C**). Other examples included the macrophage (**Figure 4D**) and T cell (**Figure S9B**) subpopulations harmonized by CellTypist and confirmed at the

molecular level with respect to markers and expression programs. All these highlight the power of CellTypist in enhancing the specificity and interpretability of cell types across diseases and studies.

Furthermore, harmonization of all cell types into one unified representation allows for systematic investigation of cellular pathodevelopmental trajectories. We used a list of pulmonary fibrosis-defining biomarkers[27,29] as signature genes to quantify the degree of epithelial remodeling in the lung. Among alveolar epithelial cell types, this signature was particularly prominent in the aberrant basaloid cells, as expected. In addition to this, the alveolar type 1 (AT1) and 2 (AT2) cells were scored higher in fibrotic states of the eight diseases as compared to the healthy controls, potentially accounting for their disease-specific alignments (**Figure 3C**). Moreover, consistent with the CellTypist alignment where a proportion of transitional AT2 cells were intermixed within the AT1 cells in the IPF lung, these two cell types exhibited higher fibrosis scores than healthy and diseased AT2 cells. This indicated an impaired alveolar epithelial regeneration process from AT2, with transitional AT2 cells developing into AT1 cells[32] (**Figure 3C**).

For airway (as opposed to alveolar) epithelial cells, overrepresentation of the pulmonary fibrosis signature was observed in the secretory cell compartment. These included diseased club cells that aligned across four datasets, and IPF-related goblet cells, in particular the *MUC5B+* subtype harmonized by CellTypist in three of the datasets (**Figure S10A**). Among the collagen-producing stromal cells, an excessive accumulation of extracellular matrix (ECM) was detected in pathological fibroblasts and myofibroblasts (**Figure 3D**). Between them, myofibroblasts in five diseases (ILD, SA, cHP, NSIP, and IPF) had activation signatures, and showed a much stronger enrichment of ECM genes than healthy and diseased fibroblasts, including the IPF "Fibroblast" in *Morse et al*[28] which was harmonized into the myofibroblast population (**Figure 4C**). In the lung endothelium, we located a match between the peribronchial endothelial cells from *Adams et al*[27], and all IPF endothelial cells in the other three datasets that were *COL15A1+*, a marker suggestive of a widespread bronchiolization occurring in endothelial cells of the distal lung in IPF (**Figure S10B**). Lastly, we assessed the distribution of the profibrotic gene signature[27] across harmonized immune cells, and found its enrichment in the *SPP1+* macrophages, and to a lesser extent, the *F13A1+* macrophages, in comparison to the *FCN1+* macrophages (**Figure S10C**). This was consistent with the reported role of this *SPP1+* macrophage population in activating myofibroblasts during fibrosis[33].

We next grouped cell types and disease conditions according to their alignments in the cell type hierarchy. Through this, duplicated information can be consolidated and molecular signals can be systematically interrogated. In alveolar epithelial cells, the AT2-transitional AT2-AT1 molecular axis was maintained across the various disease states; at the same time, specific disease variations in this trajectory were detected (**Figure 3E**). We found activation of genes such as *BCL3* and *SOCS2* in AT1 and AT2 cells from cHP and SSc, as compared to cells from PM and IPF. Interestingly, *BCL3* has been reported to alleviate pulmonary injury in both mice and humans[34,35], and the observed downregulation in the most severe form of pulmonary fibrosis (i.e., IPF) may reflect its association with disease severity. In contrast, compared to other diseases, PM and IPF induced overexpression of several other genes in AT1 and AT2 cells, including *IGFBP2*, a biomarker of IPF[36], *CLIC6*, a key gene upregulated in IPF[37], and *TRIM2*, part of the microRNA-369-3p/TRIM2 axis relevant for regulating pulmonary fibrosis[38].

Transcriptional alterations were also detectable in a more cell type- and disease-restricted manner (**Figure 3E**). For instance, in AT2 cells, expression of *PLPP3* and *GALNT2* was largely specific to cells in COPD, whereas *CYP1B1* was detected as uniquely expressed in SSc, a chronic autoimmune disease affecting multiple organs including the kidney and lung (and inhibition of *CYP1B1* has been shown to prevent fibrosis in the kidney[39]). We also observed a partial loss of the AT1 cell identity specifically in PM, illustrated by the deficiency in expression of AT1 markers *AGER* and *RTKN2*, as well as a gain of expression in known pulmonary fibrosis marker *MMP7*[40]. Further extending this analysis to the secretory cells (**Figure S11A**) and fibroblasts (**Figure S11B**), we characterized both disease-shared and -specific alterations in transcriptomes of multiple cell types and subtypes as harmonized by CellTypist.

In summary, application of CellTypist to four diseased lung datasets successfully harmonized the cell types in different compartments, disentangled the aberrant cell states in multiple diseases, revealed their underlying pathological signatures, and facilitated a comprehensive investigation of transcriptomic changes.

Cell re-annotation and data integration by CellTypist

Given the harmonized cell types across datasets, CellTypist can further refine the identity of each cell by reannotating it within the context of multiple datasets. Specifically, on the basis of

cell harmonization information, including cell identities annotated in the dataset of origin and inferred from the other datasets, each cell can be reannotated to one of the harmonized cell types in the graph corresponding to a given row in the harmonization table (**Figure 1**). CellTypist reannotated the cells from the five immune datasets into 66 low-hierarchy and 52 high-hierarchy cell types, providing precise classifications with different granularities for individual cells (**Figure 5A**). For instance, the previously annotated γδ T cells were grouped as a single cell type at the high hierarchy and were further classified into two low-hierarchy cell types: 88% as γδ T cells ("NONE = Trm_Tgd ∈ gd T = NONE = NONE") and 12% as epithelium-resident CD8+ T cells ("NONE = Trm_gut_CD8 ∈ gd T = NONE = NONE").

Through reannotation, all cells were placed into a common naming schema, allowing data integration to be supervised by annotations. In particular, the cell hierarchies built under this schema, in combination with the transcriptional distances, offered an accurate and efficient approach to construct the neighborhood graph. This confined neighbor search to cells from the same branches or similar cell types, thereby purifying cell neighbors at a minimal search cost. Notably, the extent to which the cell hierarchies guide data integration can be adjusted through controlling the number of neighboring cell types (i.e., meta neighbors) that are incorporated into the search space. This approach also applies to partially annotated datasets where local transcriptomic structures can be tuned using annotated cells only. We employed this strategy to integrate cells from the five immune datasets, and based on this, revealed the expected transcriptomic organization of immune cells (**Figures 5B and S12**).

To benchmark CellTypist against widely used data integration methods, we compared the performance of CellTypist on the five immune datasets with that of other methods, including six unsupervised (FastMNN, Harmony, Seurat v4, BBKNN, Scanorama, and scVI) and two supervised (scANVI and scGen) ones[9,14–20] (**Figure S13**). A set of integration metrics were used to assess the performance of these tools (**Methods**). With respect to the biological conservation, annotation-aware methods (scANVI, scGen, and CellTypist) showed better preservation of cell types across datasets than unsupervised ones, regardless of use of low- or high-hierarchy annotations (**Figure 5C**). The degrees of batch removal, on the other hand, were comparable across most methods, with the top performers being Seurat v4, BBKNN, and CellTypist. The performance of CellTypist was further confirmed by benchmarking on the four diseased lung datasets (**Figures S14 and S15**).

Since only part of the dataset is exploited when anchoring neighbors to query cells, we also expect an improved efficiency of data integration by using CellTypist. To test this, we benchmarked the runtime of all nine methods based on different subsets of the five immune datasets generated by downsampling cells from 10,000 to 400,000 (**Figure 5D**). Five methods were able to integrate these cells within 15 minutes: scVI and scGen (using graphics processing unit [GPU]), Scanorama, CellTypist, and BBKNN. Among them, CellTypist and BBKNN were the only two methods that integrated more than 400,000 cells in two minutes.

In conclusion, after reannotating cells through the CellTypist harmonization module, CellTypist further provides an integration module to tune the data structure towards harmonized cell types whilst mitigating the effects from batch confounders in an efficient manner.

A multi-organ reference map for cell harmonization, integration, and classification

Integrating single-cell resources across the community, a current focus of the HCA consortium, is often laborious. Taking advantage of its highly automated workflow from harmonization to integration, we anticipate that CellTypist could accelerate this process. We thus employed CellTypist to build multi-organ reference maps following three steps: i) manual collection of public organ-focused datasets and their associated meta-data, ii) automatic cell type harmonization and integration by CellTypist, iii) curation and finalization of the derived cell atlases (**Figure 6A**).

We compiled a total of 38 single-cell and single-nucleus transcriptomics datasets profiling 3,694,864 cells from 369 adult human donors in 12 tissues and organs, including the blood, bone marrow, heart, hippocampus, intestine, kidney, liver, lung, lymph node, pancreas, skeletal muscle, and spleen (**Table S1**). After applying the automated pipeline in CellTypist, we generated 12 standardized organ reference atlases, available at *https://www.celltypist.org/organs*. All cells in these atlases were reannotated into the harmonized groups inferred by CellTypist, which after further manual inspection and categorization were finally curated into biologically meaningful cell type labels. Based on this resource, we offered three complementary functionalities for the single-cell community (**Figure 6B**). Firstly, we provided a two-dimensional space (UMAP) onto which cells integrated across datasets were projected, enabling inspection of cell types and gene expression.

Secondly, to demonstrate how cell types were harmonized and transcriptomic connections were established across datasets, we visualized the cell type alignments in the form of a harmonization graph or cell type hierarchy, referred to as the "tree plot" in CellTypist. Thirdly, we leveraged the label transfer framework in CellTypist v1.0, a light-weight cell annotation algorithm based on logistic regression[8], to train a suite of machine learning classifiers. Thus, we provided 48 dataset-specific and 12 organ-specific models to achieve automatic cell type classifications in a broad tissue context.

Discussion

Although the single-cell community has been using different batch correction methods to integrate cells, information about independently annotated cell types across datasets, a valuable part of single-cell transcriptomics, has not been fully exploited. To benefit from annotations for cross-dataset integration, we developed workflows for cell type harmonization and integration, combined them into the CellTypist infrastructure, and employed this pipeline to address challenges in developing cell type standards. Combining with the cell type classification module offered by CellTypist v1.0, CellTypist v2.0 is now a harmonization-integration-classification triad contributing to assembling HCA reference data in health and disease.

The PCT algorithm endows CellTypist with scalability, speed, and accuracy. We examined the performance of CellTypist on five heterogeneous immune datasets, and found consistency between cell types harmonized by CellTypist and those curated manually. In CellTypist, we adopt the harmonization graph to iteratively redefine and reannotate cell types without losing relationships between any pairs of datasets. This approach overcomes the challenge in multi-dataset alignment and dataset order instability, and generates consistent cell type hierarchies. Due to constraints on representation in two dimensions, cell types with complex one-to-one and one-to-many alignments may be overlaid in the resulting tree plot (see one example from the blood atlas at https://www.celltypist.org/organs/tree/Blood). Though these cross-connections are partially resolvable through visualization with a different order of datasets, three-dimensional tree plot or other forms of visualizations may be further avenues to pursue.

Reconstructing cell type hierarchies is a long-standing theme in cell biology. Whereas cell type relationships are imprinted in their molecular traits, transcriptome-based cell groupings (e.g., via unsupervised hierarchical clustering) do not necessarily mirror developmental history or biology. Expert knowledge is needed to group or segregate cell types regardless of their transcriptomic similarities. We speculate that this hard-earned legacy knowledge is often already incorporated into manual annotations, and our approach now allows this information to be integrated across multiple datasets to approximate the underlying biological hierarchies. CellTypist represents an effort towards this direction and lays the foundation for future work aiming to achieve a data- and knowledge-driven cell type hierarchy.

Application of CellTypist to additional datasets further demonstrates its potential to unveil molecular and cellular variations in both physiology and pathology. Since cells from independent laboratories are harmonized, the resulting information can be exploited to put each cell into different contexts beyond its original study. This leads to consolidation of well-defined cell types and identification of novel cell states, as well as alignment of the underlying gene expression signatures and their altered characteristics in disease. For instance, CellTypist was able to uncover pathological cell types and states that could only be identified when multiple datasets were harmonized and greater resolution was achieved. Of note, detecting rare cell types and reflecting subtle molecular changes poses a challenge in many single-cell studies[41–43]. CellTypist defines two levels of novelties to capture these signals: cells either unmatched or unharmonized across datasets, which are identified during the workflow. A limitation of this approach is that when ambient RNAs are highly abundant in a given dataset, not all cell types belonging to the same category can be well matched across datasets (**Figure S4F**). How to distinguish these artifacts from real transcriptomic deviations is a topic of interest for further developing CellTypist.

For data integration, CellTypist uses an annotation-aware algorithm to supervise the construction of neighborhood graphs. Influence of cell annotation on the data structure can range from forcibly merging the same cell types to a more lenient cell grouping. Adjusting the stringency of cell type annotations provides an effective way of integrating the cell type hierarchy into the integration. We show that CellTypist is able to integrate diverse datasets to alleviate their batch effects and maintain the cell type relationships across datasets. Having a consistent database and search architecture to quantitatively compare cell types is critical to the single-cell field[44,45]. The seamless integration of the two functionalities (harmonization and integration) in CellTypist allows us to semi-automate the atlas building effort. We harmonized

~250 high-quality cell types from more than 3.6 million cells for 12 human tissues, providing, to our knowledge, the first multi-organ resource that allows information of cell type harmonization, integration, and classification to be simultaneously inspected and queried. We also anticipate that this highly structured CellTypist architecture, together with the multi-organ landmark resource, will aid in the assembly of diverse single-cell resources, facilitate the propagation of cell knowledge across communities, and improve the interpretability of organ-to-organ and health-to-disease cell relationships.

Acknowledgements

We thank D. Sumanaweera and K. Polanski for feedback on the CellTypist algorithm; K. Meyer, A. Cujba, E. Madissoon, and other members of the Teichmann lab for helpful discussions; A. Maartens and R. Vento-Tormo for editing and commenting on the manuscript. This work was funded in part, by the Wellcome Trust (grants 206194 and 108413/A/15/D to S.A.T.). This work is also funded by Technology Vision Fund under Turing project code R-TEC-001. This publication is part of the Human Cell Atlas (www.humancellatlas.org/publications/).

Author contributions

C.X. and S.A.T. conceived the project. C.X. developed algorithms, collected data and performed all analyses. C.X., S.W., L.J., B.S., R.H., and P.H. contributed to the immune cell annotation in the CellTypist v1.0 database. C.X. and M.P. structured the database and web portal. C.X. and S.A.T. wrote the manuscript.

Declaration of interests

S.A.T. is a remunerated member of the Scientific Advisory Board, Foresite Labs, and Qiagen. She is co-founder and equity holder of Transition Bio.

Methods

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Dr. Sarah A. Teichmann (st9@sanger.ac.uk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All datasets used in this study are publicly available. The five immune datasets were downloaded from the tissue stability cell atlas (https://www.tissuestabilitycellatlas.org), cross-tissue immune cell atlas (https://www.tissueimmunecellatlas.org), gut cell atlas (https://www.gutcellatlas.org), HCA data portal (https://data.humancellatlas.org/explore/projects/cc95ff89-2e68-4a08-a234-480eca21ce79), and developmental cell atlas (https://developmental.cellatlas.io/fetal-liver). The four lung datasets were downloaded from Gene Expression Omnibus: GSE136831, GSE128033, GSE135893, and GSE122960. The 38 datasets used for building the multi-organ atlases are also publicly available, with their detailed information listed in Table S1.
CellTypist source code is available at https://github.com/Teichlab/celltypist, and the Python package can be found at https://pypi.org/project/celltypist. Detailed usage of CellTypist is also available at https://github.com/Teichlab/celltypist, with step-by-step tutorials at https://celltypist.readthedocs.io/en/latest. Notebooks and scripts for reproducing the results in this study are available at https://github.com/Teichlab/CellTypist2_notebook.

METHOD DETAILS

Induction of the PCT-based tree

A direct calculation of the transcriptomic distance between two cells (or cell types) usually suffers from substantial batch effects when they originate from two different studies. To overcome this, CellTypist leverages the PCT algorithm to predict the transcriptional dissimilarities instead of calculating them. Since there is no ready-made PCT algorithm for use in the single-cell field, we adopt the commonly used tree structure, that is, parallel arrays from

the Python scikit-learn package[46] (version 0.24.1), and implement the PCT framework adapted for single-cell transcriptomics.

A distance matrix $d_{c,ct}$ is first constructed for each dataset independently, where c and ct represent the cells and cell types (defined as the cell centroids) in a given dataset, respectively. The general assumption here is that cells within one dataset are less impacted by batch effects in comparison to those across datasets. The metric of correlation distance (1 - Pearson's r) or Euclidean distance is used to calculate this matrix based on the gene expression space or low-dimensional latent representations (e.g., PCA), respectively. For each cell c (each row), this matrix records its dissimilarity profile across cell types.

We next define two key components of the clustering tree: distance measure and prototype[47]. In CellTypist, the dissimilarity between two cells in the matrix d is assessed using the Euclidean distance, and the prototype $p(n)$ is defined as the dissimilarity profile averaged across cells that are involved during the tree induction of a given node n. Accordingly, the intra-node variance $\sigma^2(n)$ is defined as the sum of squared distances between each cell c and the prototype $p(n)$:

$$\sigma^2(n) = \sum_{i=1}^{|n|} \sum_{j=1}^{|ct|} \left(c(i,j) - p_j(n) \right)^2$$

where i enumerates each cell contained in the node n, j enumerates each cell type in the reference dataset of interest, and $c(i,j)$ is the dissimilarity value for cell i and cell type j. The benefit of such definition is that minimizing the intra-node variance during PCT is equivalent to minimizing the squared error when considering each ct independently and further summing them up across all cell types as defined in a classical multi-output regression tree (CART):

$$\sigma^2(n') = \sum_{j=1}^{|ct|} \sum_{i=1}^{|n'|} \left(c(i,j) - p_j(n') \right)^2$$

Also through this definition, during the top-down induction of the clustering tree, the node impurity (i.e., the degree of difference among cells in a given node) will be calculated in a similar way as CART. Specifically, since the impurity $I(n)$ for the node n in CART is calculated as the mean squared error further averaged across targets (cell types), the node impurity for a PCT will be $|ct| \times I(n)$ by treating the dissimilarity profile of each cell as a

single response variable. The proxy impurity improvement is then quantified by summing the impurities of the two daughters of the node n and adding a negative sign.

The clustering tree is split at each possible node by a "test" corresponding to an expression cutoff in a given gene that mostly separates cells within this node. By incorporating all above optimizations, a clustering tree is induced for each dataset, with a stringent requirement for node sizes (i.e., the number of cells to be greater than 20 which is adjustable in CellTypist). However, the clustering tree is not exhaustively split to avoid overfitting. In the meantime, the resulting tree is not pruned by cost complexity pruning (CCP) as used by many tree-related algorithms, but by a more intuitive F-test as detailed below.

Impurity reduction test and tree pruning

CellTypist uses an F-test to determine whether the clustering tree needs to be further induced to a lower hierarchy. Compared to CCP, this approach underscores the clustering tendency of similar cells which likely correspond to a homogeneous cell state but places less emphasis on the tree complexity (i.e., number of leaves generated by possible splits). The level of significance for a possible split at the node n in the reference dataset can be assessed by an F-test:

$$F_{|ct|\times(|n|-1),|ct|\times(|n|-2)} = \frac{\dfrac{\sigma^2(n)}{|ct| \times (|n| - 1)}}{\dfrac{\sigma^2(l) + \sigma^2(r)}{|ct| \times (|n| - 2)}}$$

where $\sigma^2(n)$, $\sigma^2(l)$, and $\sigma^2(r)$ are the intra-cluster variances in the node, left daughter, and right daughter, respectively. Replacing all the three variances in this equation with the impurity values ($I(n)$, $I(l)$, $I(r)$) in CART allows us to infer the F-statistics in PCT as:

$$F_{|ct|\times(|n|-1),|ct|\times(|n|-2)} = \frac{\dfrac{I(n) \times |n|}{|n| - 1}}{\dfrac{I(l) \times |nl| + I(r) \times |nr|}{|n| - 2}}$$

A p-value cutoff which defaults to 0.05 in CellTypist is used to prune the node (and its recursive descendent groups) which has a small F-statistics.

Prediction of dissimilarity profiles across datasets

After a PCT is built for each dataset separately, it will be used to predict the cell-by-cell-type distance matrix d for cells in the remaining datasets. Each query cell will be, according to its gene expression profile, assigned to one of the leaves in the clustering tree and predicted to be the prototype. The resulting distance matrices are then concatenated across cells from all datasets to form an expanded matrix $\widehat{D}_{c,ct}$ where c covers cells from all datasets and ct represents the cell types from the reference dataset of interest. Prediction accuracy is calculated as the coefficient of determination (R^2) between the actual ($D_{c,ct}$) and predicted ($\widehat{D}_{c,ct}$) dissimilarity profiles:

$$R^2 = 1 - \frac{\sum_{i=1}^{|n|}\sum_{j=1}^{|ct|}(D_{i,j} - \widehat{D}_{i,j})^2}{\sum_{i=1}^{|n|}\sum_{j=1}^{|ct|}(D_{i,j} - \overline{D}_j)^2}$$

$\widehat{D}_{c,ct}$ is next turned into a rank matrix starting from 1 to the number of elements in this matrix, and is further maximum-normalized to lie between 0 and 1. This is also the strategy adopted by MetaNeighbor[48]. Finally, all $\widehat{D}_{c,ct}$ are concatenated across datasets, forming a global matrix $G_{c,ct}$ of dissimilarities among all cells (c) and all cell types (ct). Of note, because a cell type is usually defined at the cluster level and theoretically no cluster is pure, CellTypist can optionally filter out cells whose gene expression profiles do not correlate most with the cell type they belong to. This will speed up the run as only a subset of cells are used, but will render these filtered cells unharmonized in the downstream workflow (see the section below).

Meta-analysis is performed by averaging cells (i.e., rows) from the same cell type in the matrix G, leading to a cell-type-by-cell-type distance matrix among all cell types studied. This matrix is further transposed and averaged with the original matrix to form the final matrix that enables a consistent visualization given its symmetry.

Cell type harmonization

Before undergoing the following stages, all cells are predicted into the most likely cell types in each dataset, on the basis of the global matrix G mentioned above. Specifically, for each dataset, except for the cells which originate from this dataset, all other cells will be assigned one of the cell types that has the smallest distance with the query cell. We denote the result as the assignment matrix $A \in R^{c \times d}$ in CellTypist depicting how each cell c is designated in each dataset d.

Stage I: initialization of the dataset core and definition of cell type relations

Two datasets are first selected, named as D1 and D2. Using their cross-prediction information in *A*, CellTypist calculates two confusion matrices *M12* and *M21*, representing how cell types in D1 are assigned across cell types of D2, and vice versa. CellTypist performs a pairwise alignment to form the dataset core for the two datasets with an iterative cell type removal procedure:

1. Initialize an empty harmonization table *HT* with three columns (D1, relation, D2)
2. Normalize each row of *M12* and *M21* to a sum of 1
3. Check whether novel cell types exist in D1 or D2. If yes, append them to *HT*, remove them in both *M12* and *M21*, and go back to 2; if no, go to 4
4. Check whether one-to-one alignments exist. If yes, append them to *HT*, remove them in *M12* and *M21*, row-normalize *M12* and *M21*, and go back to 4; if no, go to 5
5. Check whether one-to-many alignments exist. If yes, append them to *HT*, remove them in *M12* and *M21*, row-normalize *M12* and *M21*, and go back to 5; if no, go to 6
6. Check whether many-to-one alignments exist. If yes, append them to *HT*, remove them in *M12* and *M21*, row-normalize *M12* and *M21*, and go back to 6; if no, go to 7
7. Check whether any cell types are still maintained in *M12* and *M21* after 1-6. If yes, append them to *HT* as unharmonized, done; if no, done

From step 3 onwards, if *M12* or *M21* becomes an empty matrix in any step, the workflow will immediately end. A novel cell type C in D1 is appended as “C = NONE”; a novel cell type C in D2 is appended as “NONE = C”; an 1:1 alignment between C1 from D1 and C2 from D2 is appended as “C1 = C2”; an 1:N alignment between C1 from D1 and C2 and C3 from D2 is appended as “C1 ∋ C2” and “C1 ∋ C3”; a N:1 alignment between C2 and C3 from D1 and C1 from D2 is appended as “C2 ∈ C1” and “C3 ∈ C1”; an unharmonized cell type C from D1 is appended as “C = UNRESOLVED”; an unharmonized cell type C from D2 is appended as “UNRESOLVED = C”.

Definitions of the above seven categories are as follows: 1) Based on *M21*, if no cell types in D2 can align with a cell type C in D1 with more than *maximum_novel_percent* (default to 0.05 in CellTypist), this is a novel cell type in D1. 2) Based on *M12*, if no cell types in D1 can align with a cell type C in D2 with more than *maximum_novel_percent* (default to 0.05 in CellTypist), this is a novel cell type in D2. 3) In both *M12* and *M21*, if the alignments between C1 from D1 and C2 from D2 are greater than *minimum_unique_percent* (default to 0.5), plus that these alignments are not 1:N or N:1 (see 4 and 5), this will be an 1:1 match between C1

and C2. 4) If based on *M12*, one cell type C in D1 has more than two cell types aligned in D2 with a match proportion greater than *minimum_divide_percent* (default to 0.1), and these matched cell types in D2 have a back-match proportion greater than *minimum_unique_percent* (the same parameter used in 3), this will be an 1:N match. 5) If based on *M21*, one cell type C in D2 has more than two cell types aligned in D1 with a match proportion greater than *minimum_divide_percent*, and these matched cell types in D1 have a back-match proportion greater than *minimum_unique_percent*, this will be a N:1 match. 6) Based on *M12*, if after all these categorizations, a cell type in D1 remains unharmonized (at this time, *M12* should have no columns available), this will be an unharmonized cell type in D1. 7) Based on *M21*, if after all these categorizations, a cell type in D2 remains unharmonized, this will be an unharmonized cell type in D2. Of note, during multi-dataset (>=3) alignments, all three cutoffs used above are dynamically chosen to find the best alignments during iterations (see stage IV).

Stage II: combination of datasets and cell type renaming

The aim of this stage is to combine D1 and D2 into a new D1, allowing addition of a new D2 in stage III. As a first step, cells from D1 and D2 are combined, and the combined set of cells are called a new dataset D1. Next, CellTypist reannotates the identities of these cells into cell types in the harmonization table *HT* obtained from stage I. Specifically, based on what cell type each cell is predicted to be in each dataset using the information from the assignment matrix *A*, each cell can be assigned one of the rows of *HT*, or "UNASSIGNED" if this cell does not have a consistent alignment pattern existing in *HT*. These "UNASSIGNED" cells are not considered in the following stages.

Stage III: expansion of the dataset core and harmonization table

After the new D1 and associated annotations of cells in D1 are obtained, an additional dataset D2 (which is still named D2 because the earlier D1 and D2 have been combined and renamed to D1) is incorporated. The dataset core is expanded by aligning the D2 with D1 using a procedure below:

1. Form the *M12* matrix
2. Form the *M21* matrix
3. Perform the procedure in stage I with *M12* and *M21*, and obtain a new *HT*
4. Disentangle *HT* to have multiple columns (one column per original dataset)
5. Perform the procedure in stage II with the new *HT*, and form a new D1

6. Incorporate an additional dataset D2, and repeat stage III with the new D1

For step 1, $M12$ is formed by examination of the predicted cell types of D1 cells in D2, on the basis of the information from the assignment matrix A. For step 2, $M21$ illustrates what cell types D2 cells are predicted to be in D1. Notably, since cells in D2 are predicted into cell types in D1 which correspond to rows in HT, not all cells in D2 can be annotated and are thus assigned "UNASSIGNED", which are not used when generating $M21$.

After step 3, a new HT with three columns (D1, relation, and D2) is formed. The third column (D2) contains cell types from D2, whereas the first column (D1) contains cell types (rows) as defined in the earlier HT. This new HT will be expanded to have multiple columns with one dataset per column joined by a "relation" column, serving as input for step 5. When all datasets are iteratively incorporated, the final output will be a HT covering all cell types across all datasets.

Stage IV: tuning the optimal dataset order and cell reannotation

Although all pairwise predictions between datasets are considered by the above stages (information from the assignment matrix A), an optimal order of datasets is still useful with respect to grouping of similar datasets and visualization. CellTypist uses a greedy strategy to determine the optimal order of datasets. CellTypist first calculates the similarity between any pair of datasets. Next, two datasets with the highest similarity are added in the first iteration of the harmonization pipeline, followed by addition of another dataset which has the highest total similarity with the datasets already added in the previous round. Here, the inter-dataset similarity is quantified by the numbers of alignments between two datasets (with adjustable weights of 2, 1, -1, and -2 for one-to-one, one-to-many, novel, and unharmonized matches, respectively) divided by the total number of cell types, denoting the degree of cell type share between the two given datasets.

During stage I to III, combinations of five *minimum_unique_percent* (0.4, 0.5, 0.6, 0.7, 0.8) and three *minimum_divide_percent* (0.1, 0.15, 0.2) are exhaustively tried to find the best combo, by choosing the parameters that result in the least alignments when a new dataset is added. This procedure is repeatedly conducted in each iteration. After the final HT is obtained, each cell will be reannotated to one of the cell types (rows) in HT based on the assignment matrix A.

Stage V: reorganization and visualization of HT and its cell type hierarchy

An interconnected (i.e., connections spanning multiple rows of HT) group of cell types is further defined at the high hierarchy for HT, with each group being independent of other groups. For each cell type group, if it contains novel ("NONE") or unharmonized ("UNRESOLVED") representations, these blank labels will be renamed in order to distinguish the many blank labels in the HT. Specifically, a blank element in the first column (dataset) will be searched against cell types from the other columns across the given row until a non-blank one is found and renamed to it. A blank element in other columns (except the first column) will be searched in the opposite direction and renamed to the first non-blank element in the given row. The newly formed HT forms the basis for the Sankey plot and the flat tree plot in CellTypist.

Next, columns (datasets) of each cell type group will be reordered according to how many unique values each column has, including the renamed blank cell types. To disentangle the cross connections in HT for visualizations, rows of each group are further reordered to enable consistent alignments to lie at bottom of each group. This reordered HT is used to plot the cell type hierarchy by cross-linking unique cell types including the renamed blank cell types across datasets.

Annotation-aware data integration

CellTypist employs the approach of restricted neighborhood search to integrate cells from different datasets (or batches), a concept similar to that used by BBKNN[16]. After cell type harmonization (or other procedures that can result in consistent cell type labels across datasets), all cells are placed into a common naming schema. In CellTypist, for each cell belonging to a specific cell type, its nearest neighbors can only be found out of cells that belong to the same cell type. Since this strategy cannot produce a fully connected neighborhood graph and each cell type is strongly separated with other cell types, CellTypist uses the cell type hierarchy and meta-neighbors to further expand the search space and preserve cell type spectrum. Specifically, the cell type hierarchy, especially the cell types grouped at the high hierarchical level, contains information about what cell types are transcriptionally similar. During the neighborhood search, each cell will search, in addition to the cell type it originates from, other cell types at the high hierarchy. Moreover, CellTypist leverages the distance matrix among cell types (see the section "*Prediction of dissimilarity profiles across datasets*") to find the nearest neighbors of cell types. For each cell type in each dataset, two nearest cell types will be located (the default parameter

of n_meta_neighbors is three, meaning that two other cell types will be incorporated). These cell types will be next combined across datasets to form a final cell type group, against which neighbors of each cell will be searched.

Due to the presence of dataset-specific cell types, some cell type groups may not cover all datasets. For these groups, the neighbors of a query cell will be contributed only by datasets (or batches) that qualify, through providing more cell neighbors per dataset. The resulting neighborhood graph is further trimmed to keep the top largest connectivities for each cell. To tackle the problem of partial annotation such as "UNASSIGNED" cells derived from the harmonization pipeline, CellTypist incorporates these cells into every cell type group and thus expands the size of each group. Accordingly, for these unannotated cells, their search space will be among all cell type groups available.

Analysis of five immune datasets

The five immune datasets were downloaded from the CellTypist database[8] and combined into one cell-by-gene count matrix. After normalization and log-transformation (with a pseudocount of 1), the highly variable genes (HVGs) were selected using the function scanpy.pp.highly_variable_genes (batch_key is set to the dataset of origin) from Scanpy[49] (version 1.7.2), followed by data scaling (scanpy.pp.scale) and principal component analysis (PCA) (scanpy.tl.pca). Cell type harmonization and cell reannotation were conducted by following the aforementioned pipeline in the Methods section "*Cell type harmonization*", collectively wrapped as the function celltypist.harmonize in CellTypist. The result was visualized as a tree plot or a cell type hierarchy using the function celltypist.treeplot, details of which can be found in the stage V of the Methods section "*Cell type harmonization*". Meta-analyses of cell types across datasets were performed by averaging the cell-by-cell-type matrix G across rows (cells) that belong to the same cell types. To confirm the existence of cell subtypes (e.g., ABCs) harmonized and detected by CellTypist, associated cells (e.g., memory B cells) were extracted and re-clustered using a canonical PCA-based clustering approach. Marker genes used for inspection and validation of cell subtypes were from the CellTypist immune cell encyclopedia (https://www.celltypist.org/encyclopedia/Immune/v2).

Analysis of four diseased lung datasets

The four lung datasets were downloaded from respective studies[27–30]. For each dataset, cells with fewer than 200 expressed genes were discarded. We next combined cell type annotations

with disease conditions to define the statuses of cells in the four datasets, and further removed the cell status with a size fewer than 20 cells. Normalization, log-transformation, HVGs selection, scaling, and PCA were performed as in the Methods section "*Analysis of five immune datasets*". The harmonization graph and cell type hierarchy were obtained using the function celltypist.harmonize in CellTypist, and visualized with celltypist.treeplot. Label transfer analysis (e.g., between "Aberrant_Basaloid" and "KRT5-/KRT17+") was conducted using the logistic regression framework in CellTypist v1.0 based on the model training (celltypist.train) followed by cell type prediction (celltypist.annotate) with the default parameters. Existence of cell subtypes harmonized and revealed by CellTypist was confirmed in a similar manner as that for the five immune datasets.

The following gene signatures were used. Pulmonary fibrosis: *CDKN2A*, *COL1A1*, *FN1*, *MMP7*, *MUC5B*, *SMAD3*, *ITGB6*, *GDF15*, and *EPHB2*; ECM: from the Gene Ontology (GO) term of "extracellular matrix structural constituent"; profibrotic signature: *SPP1*, *LIPA*, *SPARC*, *GPC4*, *PALLD*, *CTSK*, *MMP9*, and *CSF1*. Scores of gene signatures were calculated using the Scanpy function scanpy.tl.score_genes to account for background gene expression levels. For the analysis of molecular alterations, differential expression analysis was performed using the Scanpy function scanpy.tl.rank_genes_groups (with the parameter pts turned on to enable quantification of percentage of gene expression in a given cell type), based on the grouped cell types and diseases from the harmonization graph. Genes were selected with three criteria: a *p*-value less than 0.05, expression percent in the cell type of interest greater than 0.4, and expression percent in other cell types less than 0.2. A subset of genes with pathological relevance were selected and displayed ultimately.

Integration of datasets with different tools

The five immune datasets or four diseased lung datasets were integrated in an annotation-aware manner using CellTypist, with the algorithmic details in the Methods section "*Annotation-aware data integration*". The dataset of origin was considered as a batch confounder and the cell reannotation information from the harmonization pipeline was treated as a biological label key. The default parameters (e.g., n_meta_neighbors = 3) were used. The Scanpy function scanpy.tl.umap was further utilized to generate UMAP embeddings. We also used the following eight methods to integrate the datasets.

1) FastMNN: normalized expression data (HVGs only) was used. The data was split by the dataset of origin, and served as input for the wrapper function RunFastMNN in the R package

Seurat[9] (version 4.3.0). After this, neighbors were located using the function FindNeighbors (tag-value pairs: dims = 1:50, reduction = "mnn") and UMAP was generated with the function RunUMAP (tag-value pairs: dims = 1:50, reduction = "mnn").

2) Harmony: normalized and scaled expression data (HVGs only) with pre-calculated PCA coordinates were used. The wrapper function RunHarmony in Seurat was used with the dataset of origin being a batch confounder. After this, neighbors were located using the function FindNeighbors (tag-value pairs: dims = 1:50, reduction = "harmony") and UMAP was generated with the function RunUMAP (tag-value pairs: dims = 1:50, reduction = "harmony").

3) Seurat V4: normalized expression data (HVGs only) was used. The data was split by the dataset of origin, and for each dataset gene expression was scaled separately. Next, the function FindIntegrationAnchors (tag-value pairs: normalization.method = 'LogNormalize', scale = FALSE, dims = 1:50) followed by IntegrateData (tag-value pairs: normalization.method = 'LogNormalize', dims = 1:50) were used to generate a batch-corrected expression matrix. This matrix underwent a canonical pipeline of data scaling (ScaleData), PCA (RunPCA), neighborhood search (FindNeighbors), and UMAP generation (RunUMAP).

4) BBKNN: pre-calculated PCA coordinates were used. The function bbknn.bbknn was applied to the data with the dataset of origin being a batch confounder. The resulting neighborhood graph was used to generate UMAP embeddings with the function scanpy.tl.umap in Scanpy.

5) Scanorama: scaled expression data (HVGs only) was used. The wrapper function scanpy.external.pp.scanorama_integrate in Scanpy was used to generate the latent representations with the dataset of origin being a batch confounder, followed by neighborhood construction (scanpy.pp.neighbors) using the derived latent space and UMAP generation (scanpy.tl.umap).

6) scVI: raw count expression (HVGs only) was used with the batch covariate being the dataset of origin. A scVI model was built using the function scvi.model.SCVI (tag-value pairs: n_layers = 2, n_latent = 30, gene_likelihood = "nb") and further trained to generate the latent representations, followed by neighborhood construction (scanpy.pp.neighbors) using the derived latent space and UMAP generation (scanpy.tl.umap).

7) scANVI: raw count expression (HVGs only) was used with the batch covariate being the dataset of origin. The function scvi.model.SCVI was first applied as in scVI, and a scANVI model was next built using the function scvi.model.SCANVI.from_scvi_model (the biological

key is set as the cell annotations derived from the CellTypist harmonization procedure; both high- and low-hierarchy annotations were used). This model was trained with 20 epochs to generate the latent representations, followed by neighborhood construction (scanpy.pp.neighbors) using the derived latent space and UMAP generation (scanpy.tl.umap).

8) scGen: normalized expression (HVGs only) was used with the batch covariate being the dataset of origin and the biological covariate (high- or low-hierarchy annotations) being the cell annotations derived from the CellTypist harmonization procedure. The function scgen.SCGEN in scGen was used to build the model, which after training, resulted in the latent representations, followed by neighborhood construction (scanpy.pp.neighbors) using the derived latent space and UMAP generation (scanpy.tl.umap).

Benchmarking analysis of CellTypist

We used the metrics summarized in scIB[50] (version 1.1.1) to quantify the degree of biological conservation (average silhouette width [ASW], isolated label score ASW, and isolated label score F1) and batch removal (batch ASW, graph connectivity, and kBET score). Local inverse Simpson's index metrics were not selected, and other metrics requiring a cluster key were also not used here. Due to an elongated runtime, only 50,000 cells were randomly sampled for assessing the six metrics. Considering that not all tools produce batch-corrected low-dimensional latent representations, we used the ultimate UMAP embeddings (see the previous section), together with the neighborhood graph constructed from the UMAP, to directly assess biological conservation and batch removal at the level of UMAP. Cells from the five immune datasets were next downsampled to 10,000, 50,000, 100,000, 200,000, 300,000 or all cells (417,866). The integration procedures as detailed in the Methods section "*Integration of datasets with different tools*" were followed for each cell subset. For scVI, scANVI and scGen, both GPU (NVIDIA Tesla T4) and CPU (Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz) were utilized.

Generation of multiple organ atlases

A total of 38 scRNA-seq and snRNA-seq datasets were collected. Detailed information of these datasets can be found in **Table S1**. Each tissue or organ consisted of four datasets. Because some datasets contained multiple organs, these datasets were repeatedly used in different tissue atlases. For each dataset, raw count expression data was downloaded, accompanied by the

meta-information of original annotations, experimental assays, donor identities, ages, genders, and suspension types. These metadata were next unified across datasets.

Using the procedure in the Methods section "*Cell type harmonization*", we performed automatic cell type harmonization for cells from each organ that comprised four datasets. After harmonization, we injected the cell reannotation information into the combined dataset for supervised data integration. For data integration, we used the annotation-aware procedure as detailed in the Methods section "*Annotation-aware data integration*", with the batch key being the dataset of origin (for organs of heart, lung, and pancreas) or the donor ID (for organs of blood, bone marrow, hippocampus, intestine, kidney, liver, lymph node, skeletal muscle and spleen) depending on how these batches manifested in the organ studied. We next renamed the harmonized cell types into biologically meaningful labels, followed by further manual curations of cell annotations including missed cell subtypes. These cell types were also linked with the cell type labels from the Cell Ontology[51] for future community-wide query.

Machine learning-based CellTypist models were next built for each tissue and each dataset from the given tissue. The function celltypist.train in CellTypist v1.0 was used to train logistic regression-based classifiers, with the manual annotation being the label key and feature selection turned on (tag-value pairs: feature_selection = True, n_top_genes = 300). Ultimately, five models were trained for each tissue (four dataset-specific models and one combined model), resulting in a total of 60 models that could be used for automatic cell type annotation for cells querying our database in different tissue contexts. The integrated datasets with harmonized and curated cell annotations in 12 organs, as well as the CellTypist models, are all available at https://www.celltypist.org/organs.

References

1. Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). The Human Cell Atlas. Elife *6*. 10.7554/eLife.27041.

2. Rozenblatt-Rosen, O., Stubbington, M.J.T., Regev, A., and Teichmann, S.A. (2017). The Human Cell Atlas: from vision to reality. Nature *550*, 451–453.

3. Osumi-Sutherland, D., Xu, C., Keays, M., Levine, A.P., Kharchenko, P.V., Regev, A., Lein, E., and Teichmann, S.A. (2021). Cell type ontologies of the Human Cell Atlas. Nat. Cell Biol. *23*, 1129–1135.

4. HuBMAP Consortium (2019). The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. Nature *574*, 187–192.

5. Börner, K., Teichmann, S.A., Quardokus, E.M., Gee, J.C., Browne, K., Osumi-Sutherland, D., Herr, B.W., 2nd, Bueckle, A., Paul, H., Haniffa, M., et al. (2021). Anatomical structures, cell types and biomarkers of the Human Reference Atlas. Nat. Cell Biol. *23*, 1117–1128.

6. Kiselev, V.Y., Yiu, A., and Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. Nat. Methods *15*, 359–362.

7. Kimmel, J.C., and Kelley, D.R. (2021). Semisupervised adversarial neural networks for single-cell classification. Genome Res. *31*, 1781–1793.

8. Domínguez Conde, C., Xu, C., Jarvis, L.B., Rainbow, D.B., Wells, S.B., Gomes, T., Howlett, S.K., Suchanek, O., Polanski, K., King, H.W., et al. (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science *376*, eabl5197.

9. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29.

10. Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. *20*, 163–172.

11. Bernstein, M.N., Ma, Z., Gleicher, M., and Dewey, C.N. (2021). CellO: comprehensive and hierarchical cell type classification of human cells with the Cell Ontology. iScience *24*, 101913.

12. Hou, R., Denisenko, E., and Forrest, A.R.R. (2019). scMatch: a single-cell gene expression profile annotation tool using reference datasets. Bioinformatics *35*, 4688–4695.

13. Kang, J.B., Nathan, A., Weinand, K., Zhang, F., Millard, N., Rumker, L., Moody, D.B., Korsunsky, I., and Raychaudhuri, S. (2021). Efficient and precise single-cell reference atlas mapping with Symphony. Nat. Commun. *12*, 5890.

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

15. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296.

16. Polański, K., Young, M.D., Miao, Z., Meyer, K.B., Teichmann, S.A., and Park, J.-E. (2020). BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics *36*, 964–965.

17. Hie, B., Bryson, B., and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat. Biotechnol. *37*, 685–691.

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

19. Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M.I., and Yosef, N. (2021). Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Mol. Syst. Biol. *17*, e9620.

20. Lotfollahi, M., Wolf, F.A., and Theis, F.J. (2019). scGen predicts single-cell perturbation responses. Nat. Methods *16*, 715–721.

21. Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. 10.1038/nbt.4314.

22. Madissoon, E., Wilbrey-Clark, A., Miragaia, R.J., Saeb-Parsy, K., Mahbubani, K.T., Georgakopoulos, N., Harding, P., Polanski, K., Huang, N., Nowicki-Osuch, K., et al. (2019). scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. Genome Biol. *21*, 1.

23. Popescu, D.-M., Botting, R.A., Stephenson, E., Green, K., Webb, S., Jardine, L., Calderbank, E.F., Polanski, K., Goh, I., Efremova, M., et al. (2019). Decoding human fetal liver haematopoiesis. Nature *574*, 365–371.

24. James, K.R., Gomes, T., Elmentaite, R., Kumar, N., Gulliver, E.L., King, H.W., Stares, M.D., Bareham, B.R., Ferdinand, J.R., Petrova, V.N., et al. (2020). Distinct microbial and immune niches of the human colon. Nat. Immunol. *21*, 343–353.

25. Rood, J.E., Maartens, A., Hupalowska, A., Teichmann, S.A., and Regev, A. (2022). Impact of the Human Cell Atlas on medicine. Nat. Med. *28*, 2486–2496.

26. Lindeboom, R.G.H., Regev, A., and Teichmann, S.A. (2021). Towards a Human Cell Atlas: Taking Notes from the Past. Trends Genet. *37*, 625–630.

27. Adams, T.S., Schupp, J.C., Poli, S., Ayaub, E.A., Neumark, N., Ahangari, F., Chu, S.G., Raby, B.A., DeIuliis, G., Januszyk, M., et al. (2020). Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. Sci Adv *6*, eaba1983.

28. Morse, C., Tabib, T., Sembrat, J., Buschur, K.L., Bittar, H.T., Valenzi, E., Jiang, Y., Kass, D.J., Gibson, K., Chen, W., et al. (2019). Proliferating SPP1/MERTK-expressing macrophages in idiopathic pulmonary fibrosis. Eur. Respir. J. *54*. 10.1183/13993003.02441-2018.

29. Habermann, A.C., Gutierrez, A.J., Bui, L.T., Yahn, S.L., Winters, N.I., Calvi, C.L., Peter, L., Chung, M.-I., Taylor, C.J., Jetter, C., et al. (2020). Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. Sci Adv *6*, eaba1972.

30. Reyfman, P.A., Walter, J.M., Joshi, N., Anekalla, K.R., McQuattie-Pimentel, A.C., Chiu, S., Fernandez, R., Akbarpour, M., Chen, C.-I., Ren, Z., et al. (2019). Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. Am. J. Respir. Crit. Care Med. *199*, 1517–1536.

31. Cottin, V., Hirani, N.A., Hotchkin, D.L., Nambiar, A.M., Ogura, T., Otaola, M., Skowasch, D., Park, J.S., Poonyagariyagorn, H.K., Wuyts, W., et al. (2018). Presentation, diagnosis and clinical course of the spectrum of progressive-fibrosing interstitial lung diseases. Eur. Respir. Rev. *27*. 10.1183/16000617.0076-2018.

32. Choi, J., Park, J.-E., Tsagkogeorga, G., Yanagita, M., Koo, B.-K., Han, N., and Lee, J.-H. (2020). Inflammatory Signals Induce AT2 Cell-Derived Damage-Associated Transient Progenitors that Mediate Alveolar Regeneration. Cell Stem Cell *27*, 366–382.e7.

33. Ogawa, T., Shichino, S., Ueha, S., and Matsushima, K. (2021). Macrophages in lung fibrosis. Int. Immunol. *33*, 665–671.

34. Kreisel, D., Sugimoto, S., Tietjens, J., Zhu, J., Yamamoto, S., Krupnick, A.S., Carmody, R.J., and Gelman, A.E. (2011). Bcl3 prevents acute inflammatory lung injury in mice by restraining emergency granulopoiesis. J. Clin. Invest. *121*, 265–276.

35. Song, L., Wörmann, S., Ai, J., Neuhöfer, P., Lesina, M., Diakopoulos, K.N., Ruess, D., Treiber, M., Witt, H., Bassermann, F., et al. (2016). BCL3 Reduces the Sterile Inflammatory Response in Pancreatic and Biliary Tissues. Gastroenterology *150*, 499–512.e20.

36. Guiot, J., Bondue, B., Henket, M., Corhay, J.L., and Louis, R. (2016). Raised serum levels of IGFBP-1 and IGFBP-2 in idiopathic pulmonary fibrosis. BMC Pulm. Med. *16*, 86.

37. Wan, H., Huang, X., Cong, P., He, M., Chen, A., Wu, T., Dai, D., Li, W., Gao, X., Tian, L., et al. (2021). Identification of Hub Genes and Pathways Associated With Idiopathic Pulmonary Fibrosis Bioinformatics Analysis. Front Mol Biosci *8*, 711239.

38. Yi, H., Luo, D., Xiao, Y., and Jiang, D. (2021). Knockdown of long non-coding RNA DLEU2 suppresses idiopathic pulmonary fibrosis by regulating the microRNA-369-3p/TRIM2 axis. Int. J. Mol. Med. *47*. 10.3892/ijmm.2021.4913.

39. Li, F., Zhu, W., and Gonzalez, F.J. (2017). Potential role of CYP1B1 in the development and treatment of metabolic diseases. Pharmacol. Ther. *178*, 18–30.

40. Rosas, I.O., Richards, T.J., Konishi, K., Zhang, Y., Gibson, K., Lokshin, A.E., Lindell, K.O., Cisneros, J., Macdonald, S.D., Pardo, A., et al. (2008). MMP1 and MMP7 as

potential peripheral blood biomarkers in idiopathic pulmonary fibrosis. PLoS Med. *5*, e93.

41. Wegmann, R., Neri, M., Schuierer, S., Bilican, B., Hartkopf, H., Nigsch, F., Mapa, F., Waldt, A., Cuttat, R., Salick, M.R., et al. (2019). CellSIUS provides sensitive and specific detection of rare cell populations from complex single-cell RNA-seq data. Genome Biol. *20*, 142.

42. Jiang, L., Chen, H., Pinello, L., and Yuan, G.-C. (2016). GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol. *17*, 144.

43. Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature *525*, 251–255.

44. Li, M., Zhang, X., Ang, K.S., Ling, J., Sethi, R., Lee, N.Y.S., Ginhoux, F., and Chen, J. (2022). DISCO: a database of Deeply Integrated human Single-Cell Omics data. Nucleic Acids Res. *50*, D596–D602.

45. Franzén, O., Gan, L.-M., and Björkegren, J.L.M. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database *2019*. 10.1093/database/baz046.

46. Garreta, R., and Moncecchi, G. (2013). Learning scikit-learn: Machine Learning in Python (Packt Publishing Ltd).

47. Blockeel, H., Dzeroski, S., Struyf, J., and Zenko, B. (2012). Predictive Clustering (Springer).

48. Crow, M., Paul, A., Ballouz, S., Huang, Z.J., and Gillis, J. (2018). Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. Nat. Commun. *9*, 884.

49. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

50. Luecken, M.D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colomé-Tatché, M., et al. (2022). Benchmarking atlas-level data integration in single-cell genomics. Nat. Methods *19*, 41–50.

51. Diehl, A.D., Meehan, T.F., Bradford, Y.M., Brush, M.H., Dahdul, W.M., Dougall, D.S., He, Y., Osumi-Sutherland, D., Ruttenberg, A., Sarntivijai, S., et al. (2016). The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. J. Biomed. Semantics *7*, 44.

Figures and figure legends

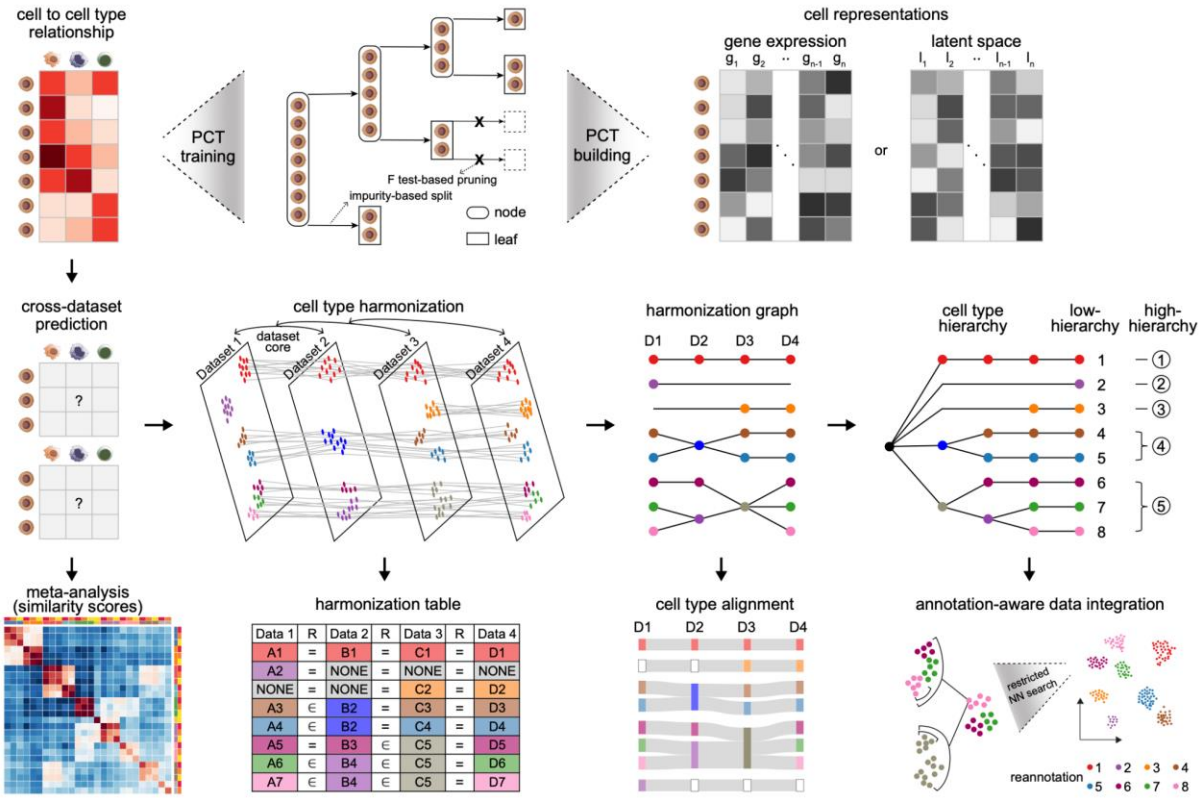


Figure 1. Framework of CellTypist v2.0

CellTypist builds a clustering tree at the gene expression or low-dimensional latent space, and based on this, predicts transcriptomic dissimilarities among cells and cell types to enable a batch-insensitive meta-analysis. A harmonization graph is then constructed by iteratively adding datasets and inferring their cell type relationships, leading to a semantically defined harmonization table. Two levels of cell type hierarchies are obtained through reorganization of the harmonization graph, which serve as input to restrict the neighborhood search space during an annotation-aware data integration. PCT, predictive clustering tree; NN, nearest neighbors.

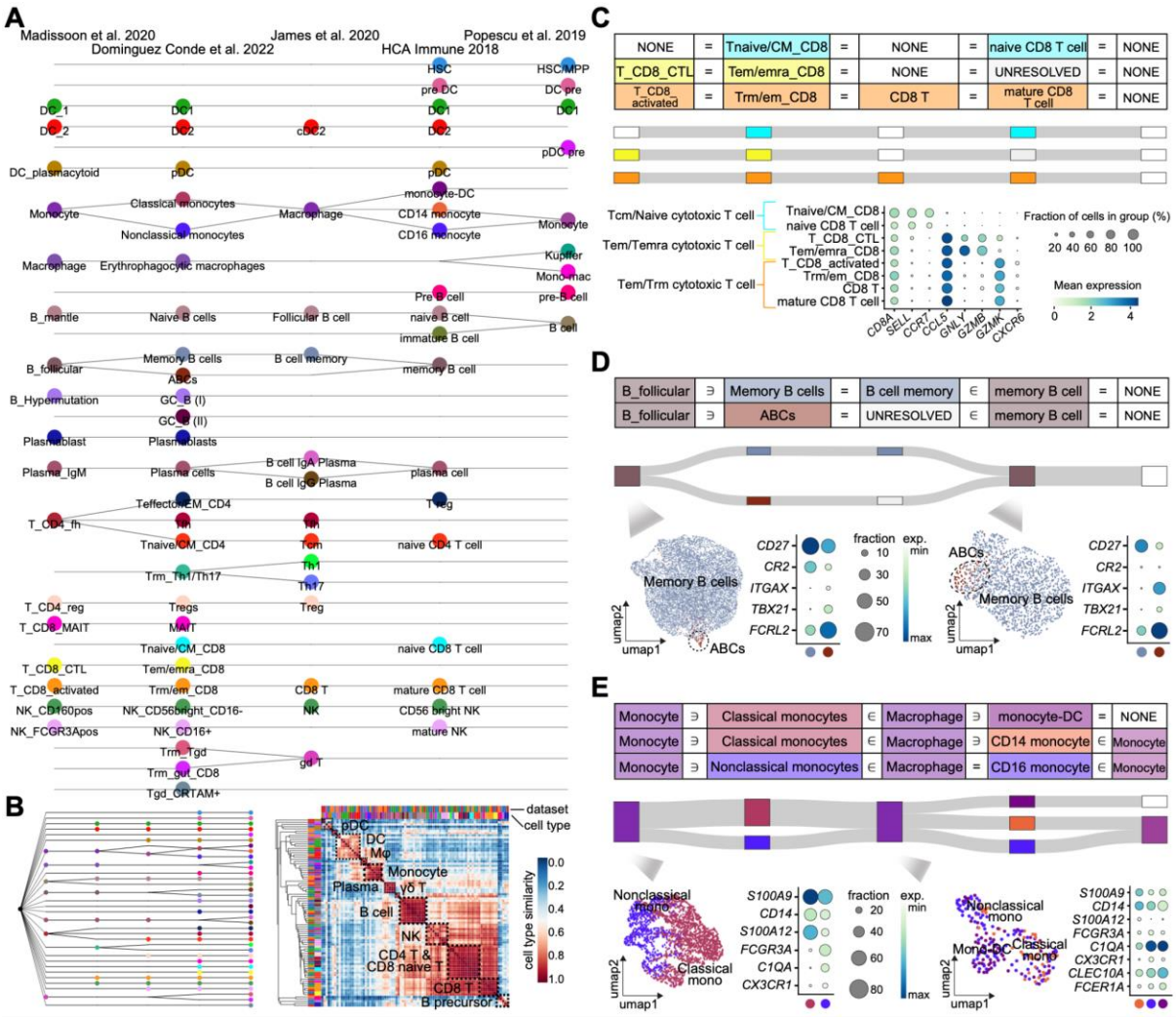


Figure 2. CellTypist harmonization recapitulates manual annotations across five immune datasets

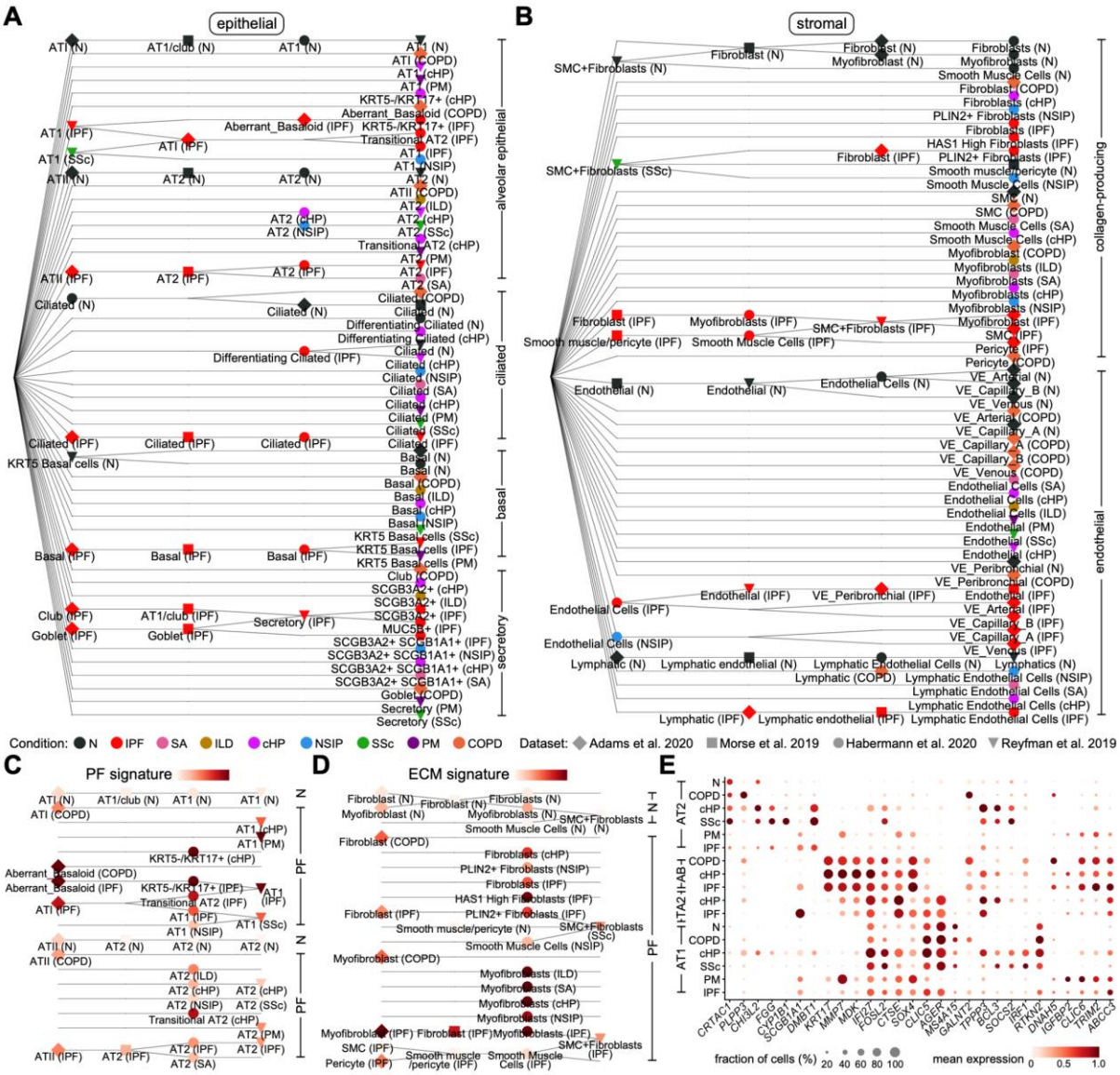
(A) Harmonization graph showing selected cell type relationships across five immune datasets shown on top. Cell type labels are colored according to their low-hierarchy alignments. The order of datasets applies to (C), (D), and (E).

(B) Left: cell type hierarchy reorganized from (A). Right: heat map displaying the unsupervised hierarchical clustering of cell types based on their inferred similarities by CellTypist. Cell types are colored as in (A), and clustered into cell compartments marked in the plot.

(C) Upper: part of the harmonization table containing eight CD8+ T cell populations, as well as their alignments visualized as a Sankey plot. Lower: dot plot showing expression of marker genes across the eight cell populations corresponding to three subtypes harmonized by CellTypist. Color of the dot represents normalized gene expression, and size represents the percentage of cells expressing a given gene.

(D) Upper: as in (C), but for the five memory B cell-related populations. Lower: UMAP visualizations of memory B cells in two datasets that contain a hidden population of age-associated B cells (ABCs). Dot plots show expression of marker genes for memory B cells and ABCs, with the color and size representing normalized gene expression and the percentage of cells expressing a given gene, respectively.

(E) Upper: as in (D), but for the eight monocyte-related cell types. Lower: as in (D), but for the two cell populations consisting of classical and nonclassical monocytes (left), and of three monocyte subtypes (right), respectively.



997

998 **Figure 3. CellTypist harmonizes diseased cell states for discovery of molecular alterations**
999 (A) Cell type hierarchy visualizing the harmonized epithelial cell types across the four lung
1000 datasets. Color and shape of the dot represent the disease condition and dataset of origin,
1001 respectively. N, normal; IPF, idiopathic pulmonary fibrosis; COPD, chronic obstructive
1002 pulmonary disease; NSIP, nonspecific interstitial pneumonia; cHP, chronic hypersensitivity
1003 pneumonitis; ILD, interstitial lung disease; SA, sarcoidosis; SSc, systemic sclerosis; PM,
1004 polymyositis.
1005 (B) As in (A), but for harmonized stromal cell types across the four lung datasets.
1006 (C) Distribution of the pulmonary fibrosis (PF)-defining gene signature in the harmonization
1007 graph focusing on the alveolar epithelial cell types. Color and shape of the dot represent the
1008 degree of PF signature and dataset of origin, respectively.
1009 (D) As in (C), but for distribution of the extracellular matrix (ECM) signature in collagen-
1010 producing cell types.
1011 (E) Dot plot showing expression of differentially expressed genes among alveolar epithelial
1012 cell types and diseases, with the color and size representing normalized gene expression and
1013 the percentage of cells expressing a given gene, respectively. Cell types and diseases are

grouped across datasets based on their alignments inferred by CellTypist. AB, aberrant basaloid cells; TA2, transitional AT2 cells.

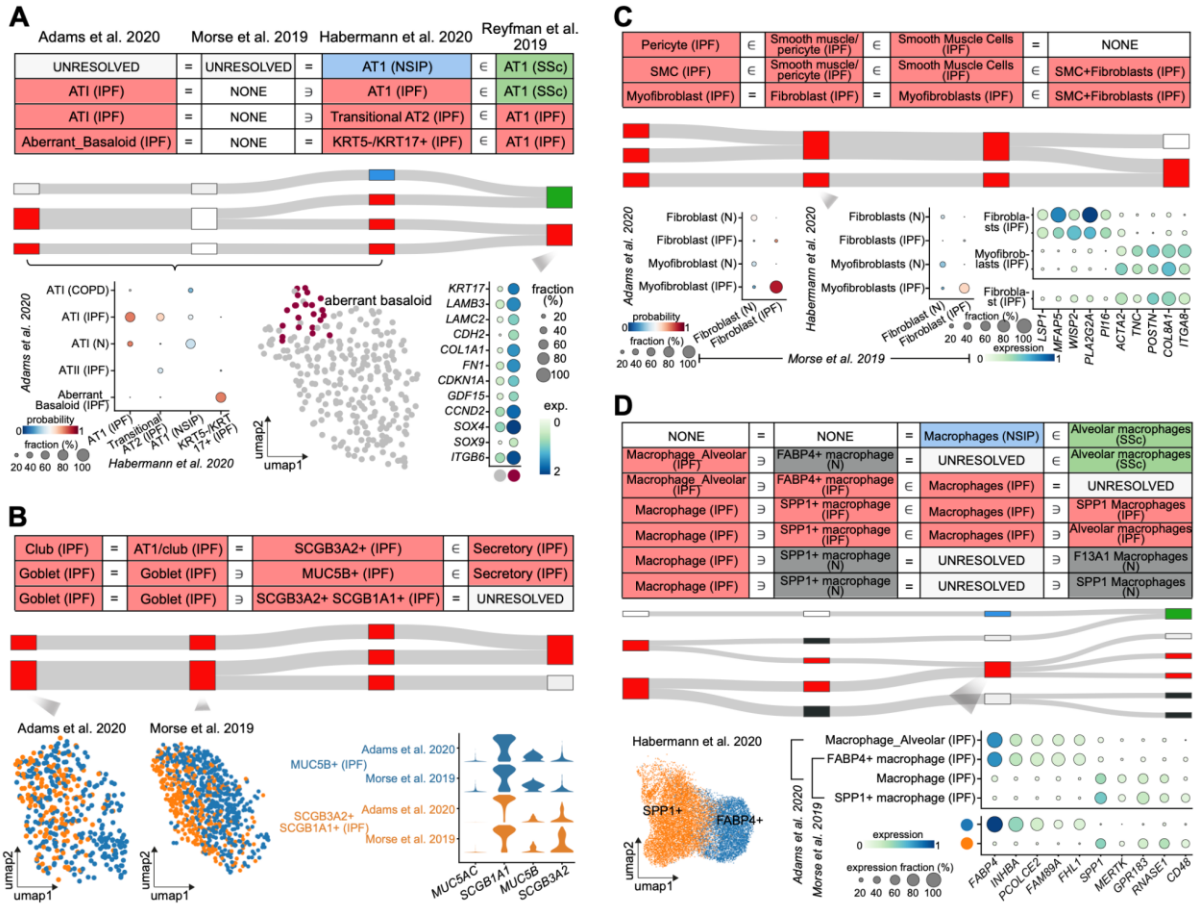


Figure 4. Examples of pathological cell states harmonized by CellTypist

(A) Upper: part of the harmonization table containing alveolar epithelial populations, as well as their alignments visualized as a Sankey plot. Lower: scatter plot showing label transfers from cell types in the first dataset to those in the third dataset, highlighting a correspondence between "Aberrant_Basaloid" and "KRT5-/KRT17+". Color and size of the dot denote the mean prediction probability and cell type alignment fraction, respectively. Dot plot to the right shows expression of marker genes of the aberrant basaloid cells separated from the "AT1" cells of the fourth dataset, with color and size of the dot representing normalized gene expression and percentage of cells expressing a given gene, respectively. The order of datasets applies to (B), (C), and (D).

(B) Upper: as in (A), but for the pathological secretory cells. Lower: UMAP plots showing goblet cells in two datasets that are harmonized into two subtypes based on the third dataset. Violin plot to the right displays marker gene expression across the two goblet subtypes.

(C) Upper: as in (A), but for all collagen-producing cell types. Lower: as in (A), but for the IPF "Fibroblast" in the second dataset that is harmonized into myofibroblasts based on the first and third datasets. Dot plot to the right shows expression of marker genes for the "Fibroblast" identified (bottom) as compared to those from the other two datasets (top).

(D) Upper: as in (A), but for macrophage-related populations in the four datasets. Lower: as in (B), but for the pathological "Macrophages" in the third dataset that are harmonized into two macrophage subtypes. Dot plot to the right shows expression of marker genes for the two subtypes identified (bottom) as compared to those from the other two datasets (top).

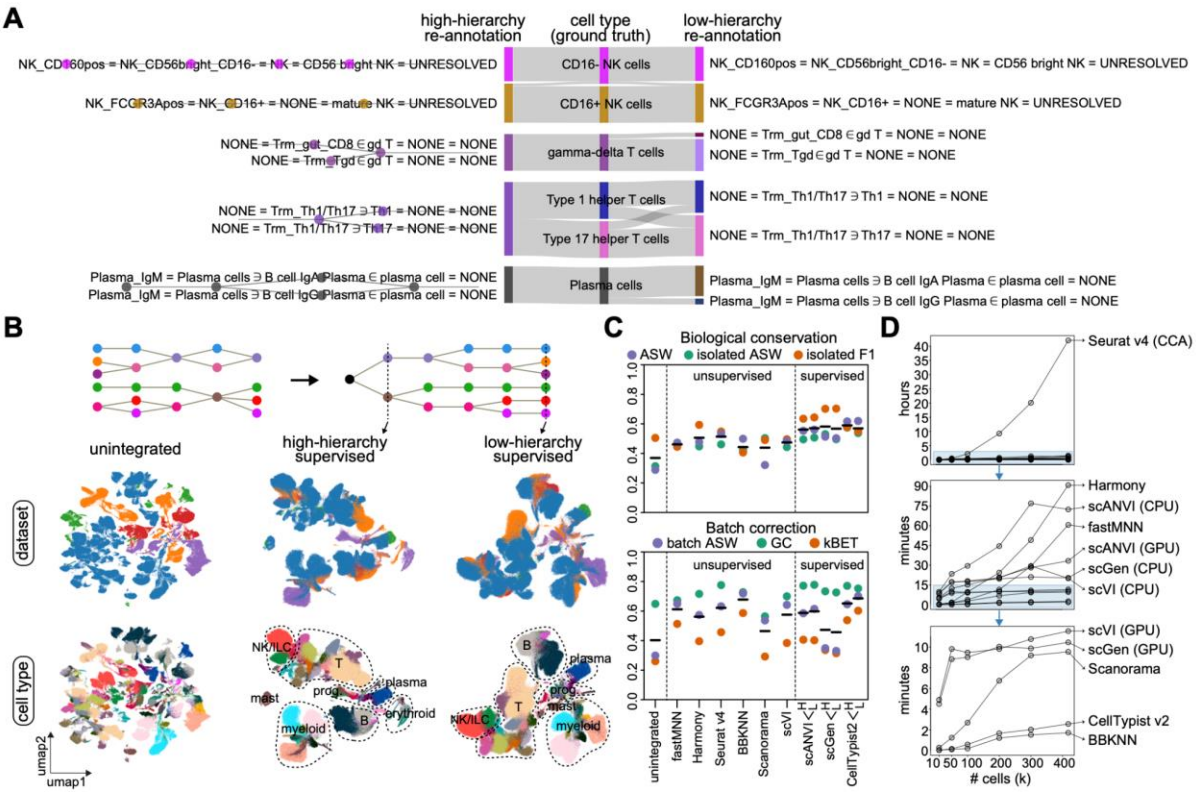


Figure 5. CellTypist enables cell reannotation and supervised data integration

(A) Example Sankey diagram depicting the correspondence among the high-hierarchy cell reannotation (left), the ground-truth annotation (middle), and the low-hierarchy cell reannotation (right) from five immune datasets. Cells are reannotated by concatenating original cell type labels across datasets with predefined symbols, corresponding to rows designated by the harmonization table.

(B) Upper: schematic of the harmonization graph and the cell type hierarchy used as input for CellTypist-based data integration. Lower: UMAP representations of cells from the five immune datasets overlaid by the dataset of origin (top) and cell type (bottom). Transcriptomic structures are shown for unintegrated cells (left), and cells integrated with supervision from high-hierarchy (middle) and low-hierarchy (right) cell types. Color codes of datasets and cell types can be found in Figure S12.

(C) Assessment of biological conservation (upper) and batch removal (lower) after data integration using the six unsupervised and three supervised methods shown on bottom. For supervised methods, both high- and low-hierarchy annotations are used. Horizontal black lines denote the averages of the integration metrics. ASW, average silhouette width; GC, graph connectivity; kBET, k-nearest-neighbor batch-effect test.

(D) Comparisons of runtime (y axis) among the nine integration methods, based on cell subsets ranging from 10,000 to 400,000 (x axis). For data integration using scVI, scANVI and scGen, both CPU (Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz) and GPU (NVIDIA Tesla T4) are tested.

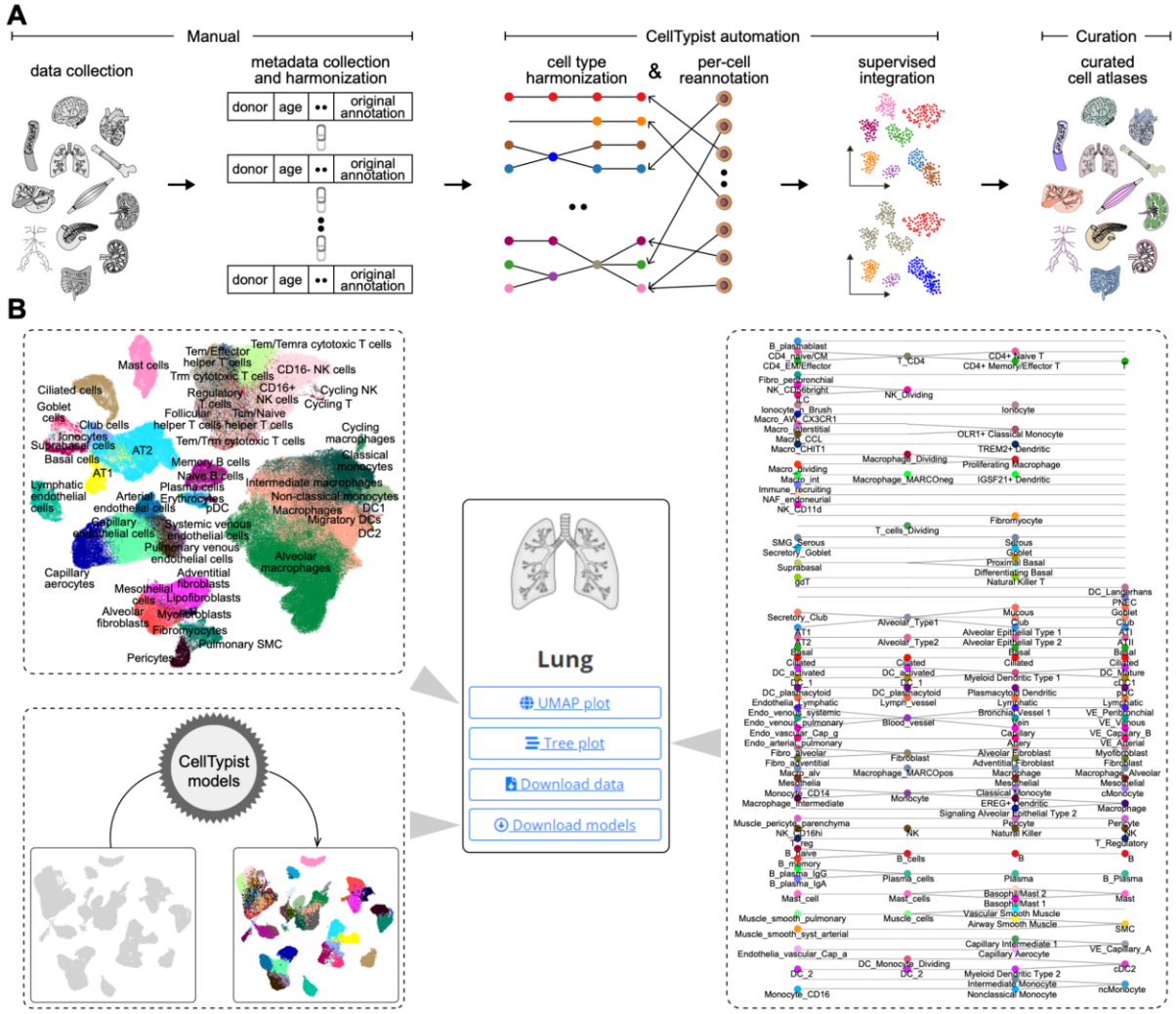


Figure 6. A standardized multi-organ resource of cell type harmonization, integration, and classification

(A) Workflow of building standardized organ atlases using CellTypist, from collection of data and metadata, cell type harmonization, cell reannotation, supervised data integration, to the final manual curations.

(B) Illustration of the three modules in our multi-organ resource, including integrated atlases with harmonized and curated cell types (top left), harmonization graphs with cross-dataset relationships among cell types (right), and light-weight machine learning models for automated cell classifications (bottom left).

A

Dataset	Tissue	Stage	# Cells	# Cell types	# Expressed genes
Madissoon et al. 2020	spleen, oesophagus	adult	37,810	25	1,050
Dominguez Conde et al. 2022	multi-organs (11)	adult	233,834	34	1,991
James et al. 2020	intestine, lymph node	adult	24,900	16	1,389
HCA Immune 2018	blood, bone marrow	fetal, adult	78,171	32	1,059
Popescu et al. 2019	liver	fetal	43,151	22	2,750

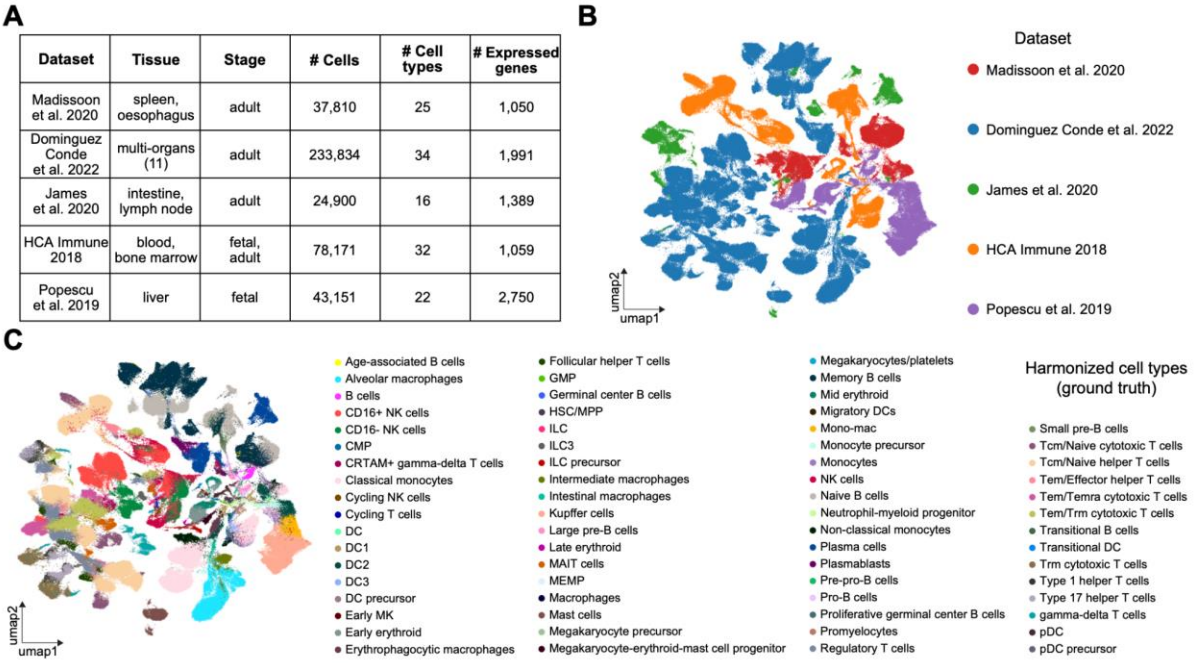


Figure S1. Overview of the five immune datasets used for validation

(A) Table showing the summary information (tissue, developmental stage, numbers of cells, cell types and expressed genes) of each dataset.

(B) UMAP visualization of all cells colored by the dataset of origin, on the basis of a canonical PCA-based single-cell pipeline.

(C) As in (B), but colored by the manually curated cell types.

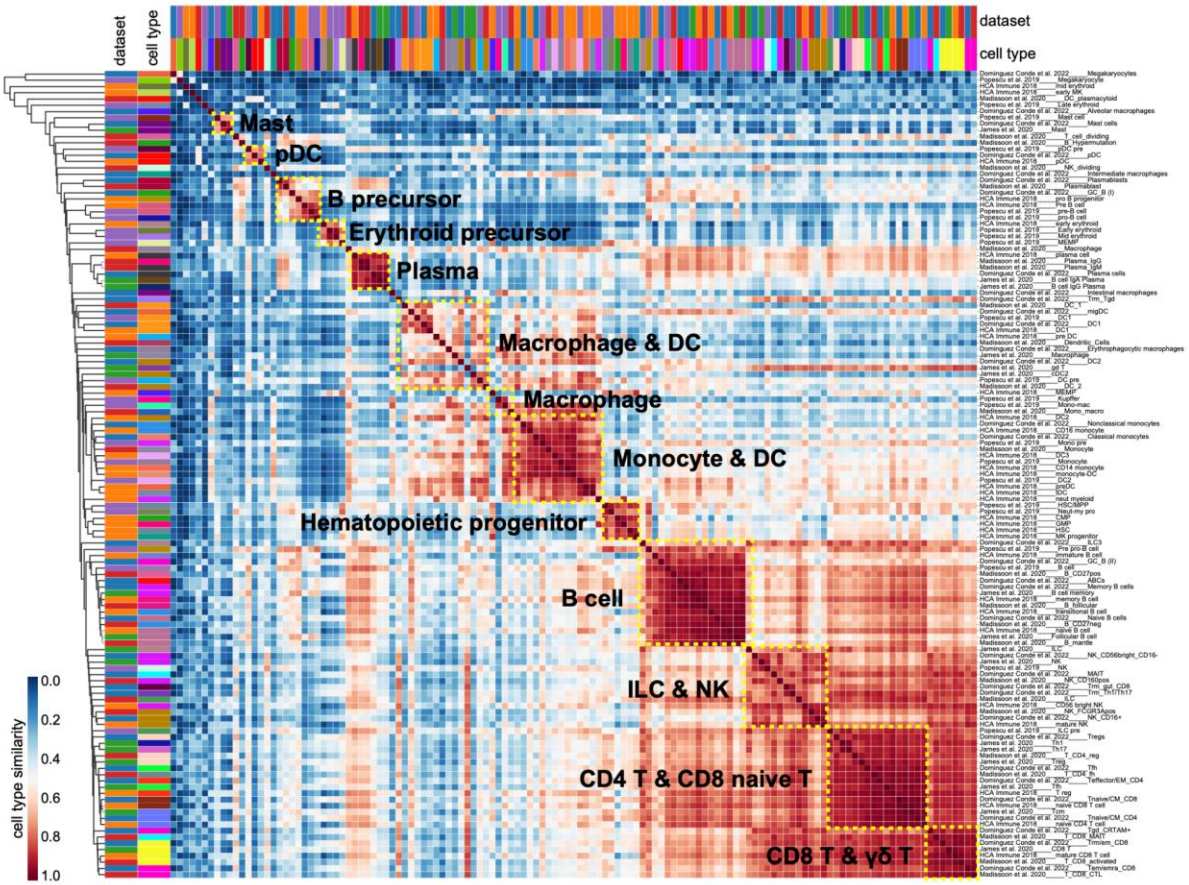
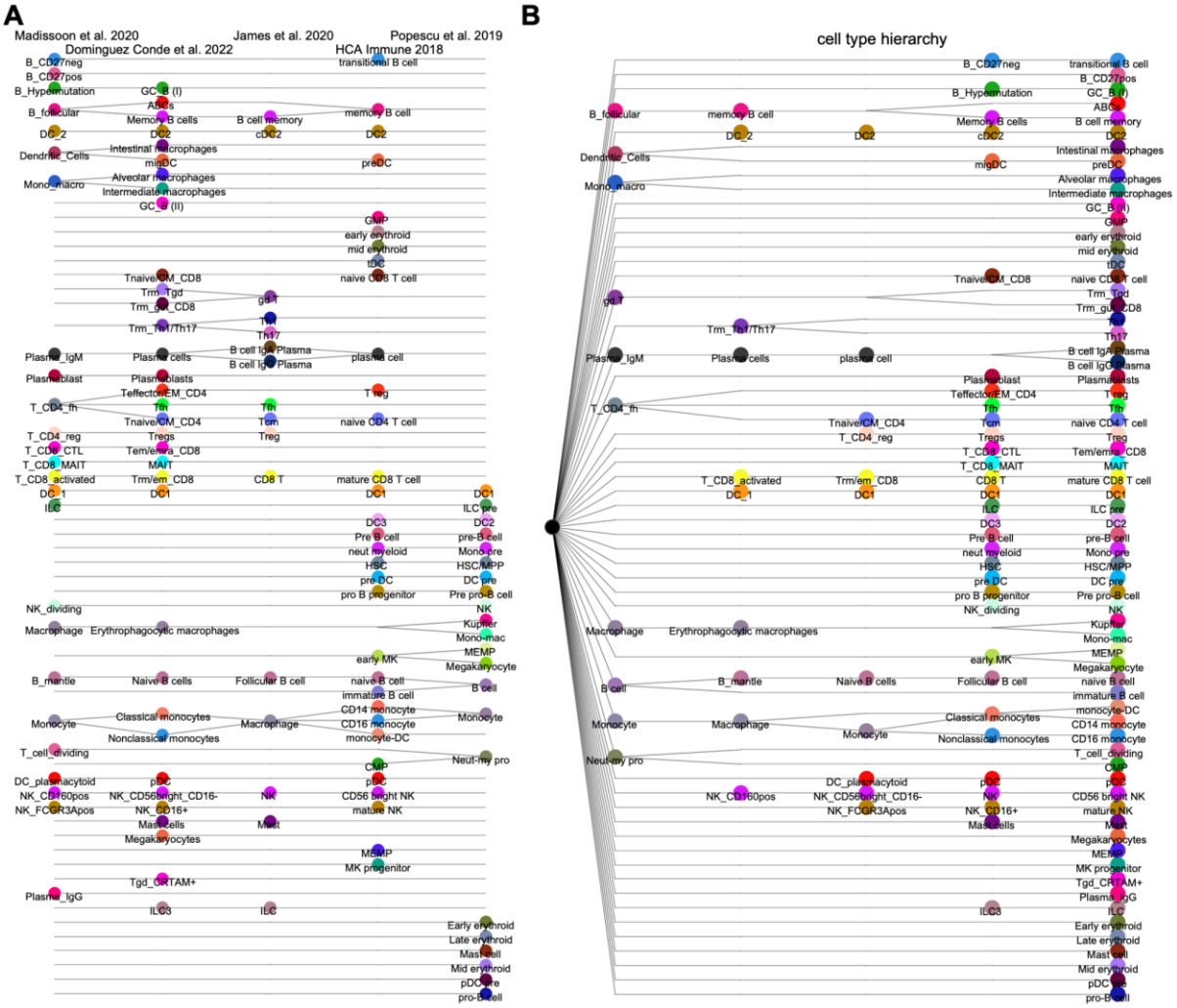


Figure S2. Heat map displaying the unsupervised hierarchical clustering of cell types based on their inferred similarities by CellTypist

Cell types are colored as in Figure 2A and Figure S3, and clustered into cell compartments marked in the plot.



1082

1083 **Figure S3. Harmonization of all cell types in the five immune datasets**

1084 (A) Harmonization graph showing cell type relationships across five immune datasets shown
1085 on top. Cell type labels are colored according to their low-hierarchy alignments.
1086 (B) Cell type hierarchy reorganized from (A).

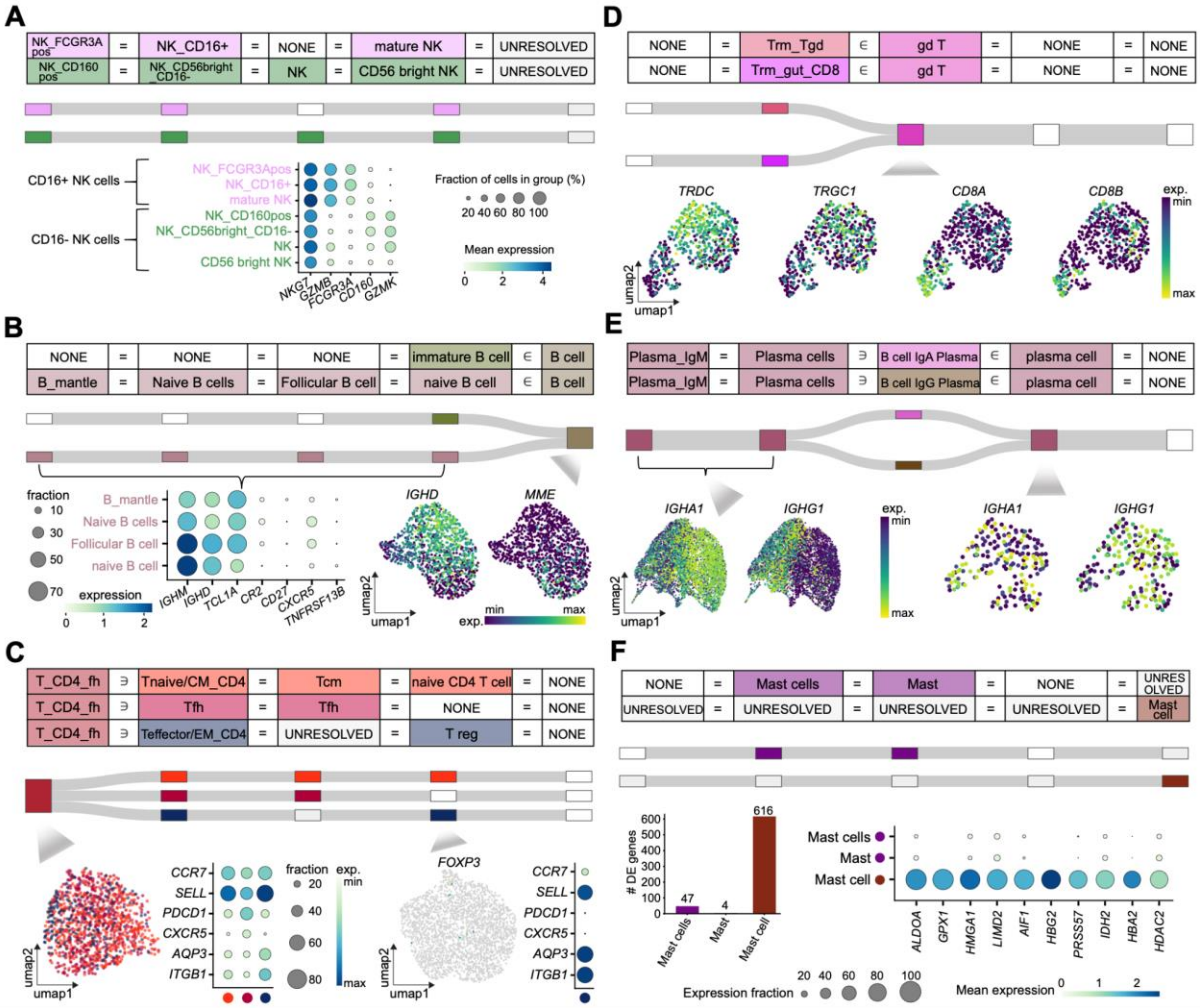


Figure S4. Examples of immune cell types and subtypes harmonized by CellTypist

(A) Upper: part of the harmonization table containing seven NK cell populations, as well as their alignments visualized as a Sankey plot. The order of datasets is, from left to right, Madissoon et al. 2020, Dominguez Conde et al. 2022, James et al. 2020, HCA Immune 2018, to Popescu et al. 2019, and applies to the rest of panels. Lower: dot plot showing expression of marker genes across the seven cell populations corresponding to two subtypes harmonized by CellTypist. Color of the dot represents normalized gene expression, and size represents the percentage of cells expressing a given gene.

(B) As in (A), but for the six B cell-related populations, including a "B cell" population that is harmonized into a spectrum of immature and naive B cells.

(C) As in (A), but for the eight CD4+ T cell-related populations, including a "T_CD4_fh" population that is harmonized into three homogeneous subtypes and a "T reg" population that was mis-annotated as demonstrated by the lack of *FOXP3* expression and is actually an effector memory CD4+ T cell subtype.

(D) As in (A), but for the three cell populations transcriptomically similar with γδ T cells in the gut, including the "gd T" population that contains a hidden gut-resident CD8+ T cell subtype.

(E) As in (A), but for the five plasma cell-related populations, including plasma cells from three datasets that are harmonized into IgA and IgG plasma subtypes.

(F) As in (A), but for the three mast cell-related populations, including a “Mast cell” population highly abundant in ambient RNAs (and also with more differentially expressed genes) and failing to match other mast cells.

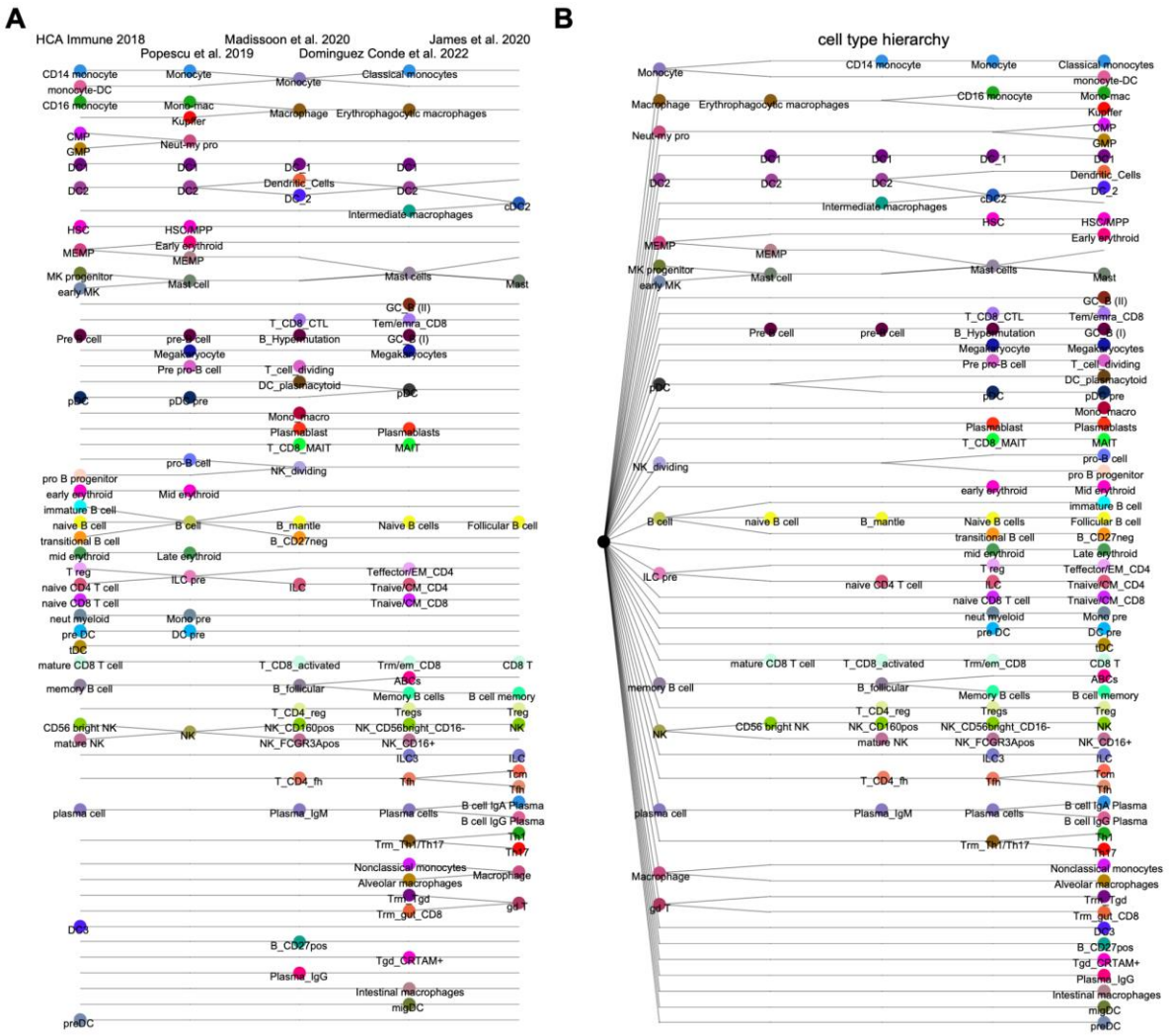
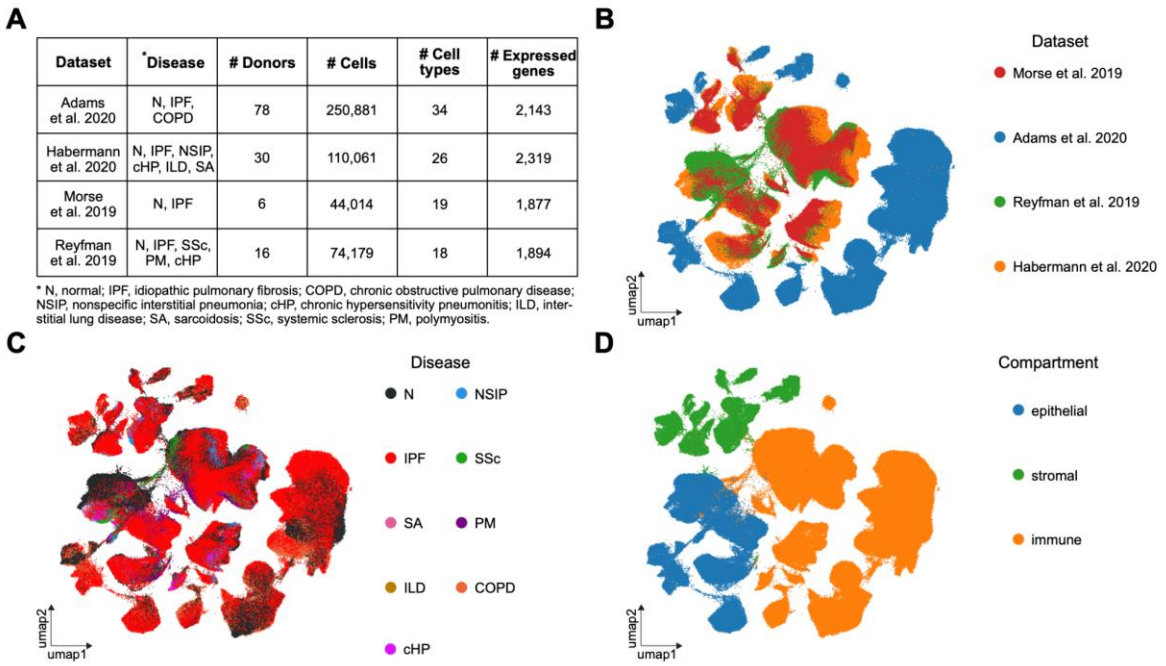


Figure S5. Robustness of CellTypist harmonizations to the order of datasets

(A) Harmonization graph showing cell type relationships across five immune datasets shown on top, with a different order of datasets as compared to Figure 2A and Figure S3A. Cell type labels are colored according to their low-hierarchy alignments.

(B) Cell type hierarchy reorganized from (A).



A

Dataset	*Disease	# Donors	# Cells	# Cell types	# Expressed genes
Adams et al. 2020	N, IPF, COPD	78	250,881	34	2,143
Habermann et al. 2020	N, IPF, NSIP, cHP, ILD, SA	30	110,061	26	2,319
Morse et al. 2019	N, IPF	6	44,014	19	1,877
Reyfman et al. 2019	N, IPF, SSc, PM, cHP	16	74,179	18	1,894

* N, normal; IPF, idiopathic pulmonary fibrosis; COPD, chronic obstructive pulmonary disease; NSIP, nonspecific interstitial pneumonia; cHP, chronic hypersensitivity pneumonitis; ILD, interstitial lung disease; SA, sarcoidosis; SSc, systemic sclerosis; PM, polymyositis.

Figure S6. Overview of the four lung datasets

(A) Table showing the summary information (disease, donor, numbers of cells, cell types and expressed genes) of each dataset. Abbreviations of diseases are summarized at the bottom of the table.

(B) UMAP visualization of all cells colored by the dataset of origin, on the basis of a canonical PCA-based single-cell pipeline.

(C) As in (B), but colored by the normal and eight disease conditions.

(D) As in (B), but colored by the three cell compartments.

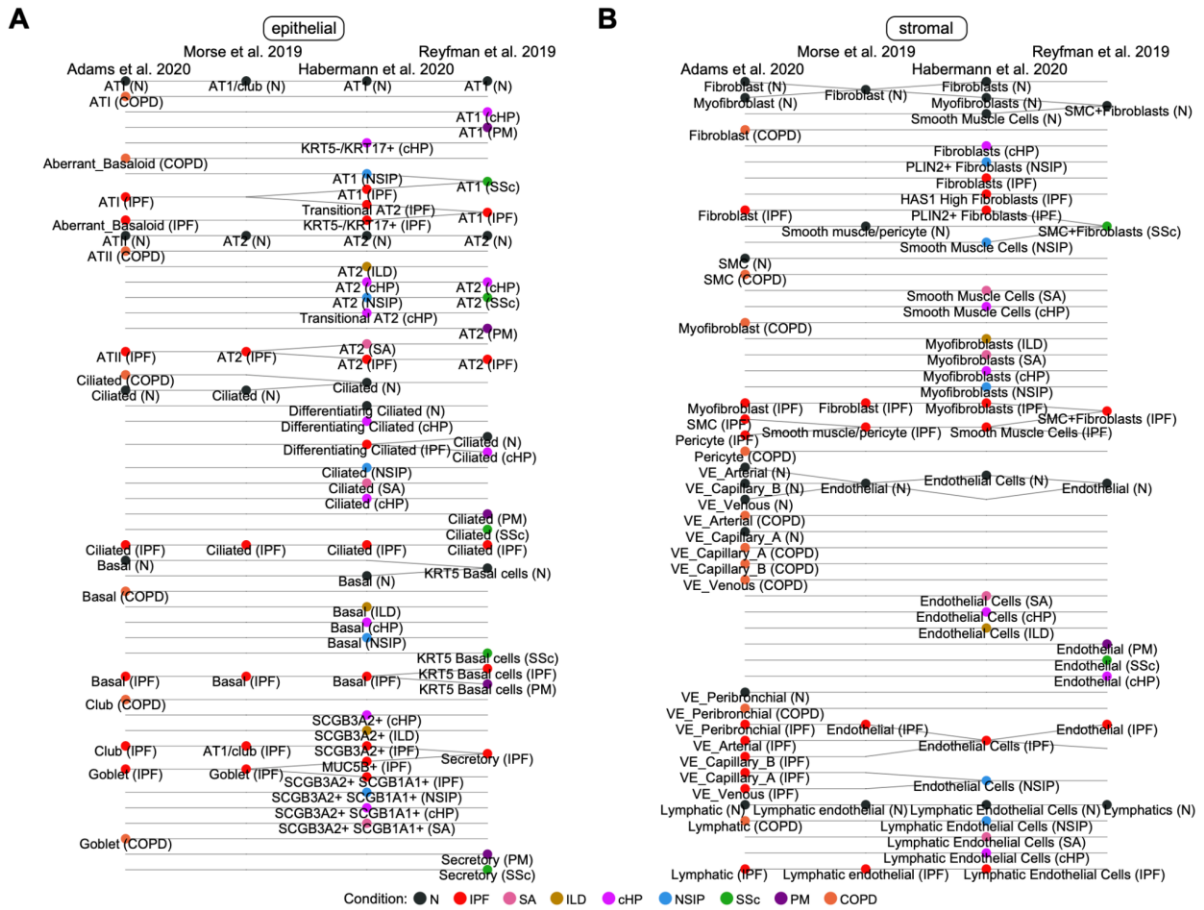


Figure S7. Harmonization graph showing cell type relationships across four lung datasets (A and B) Dataset names are shown on top. Color of the dot represents the normal or eight disease conditions for cells in the epithelial (A) and stromal (B) cell compartments. N, normal; IPF, idiopathic pulmonary fibrosis; COPD, chronic obstructive pulmonary disease; NSIP, nonspecific interstitial pneumonia; cHP, chronic hypersensitivity pneumonitis; ILD, interstitial lung disease; SA, sarcoidosis; SSc, systemic sclerosis; PM, polymyositis.

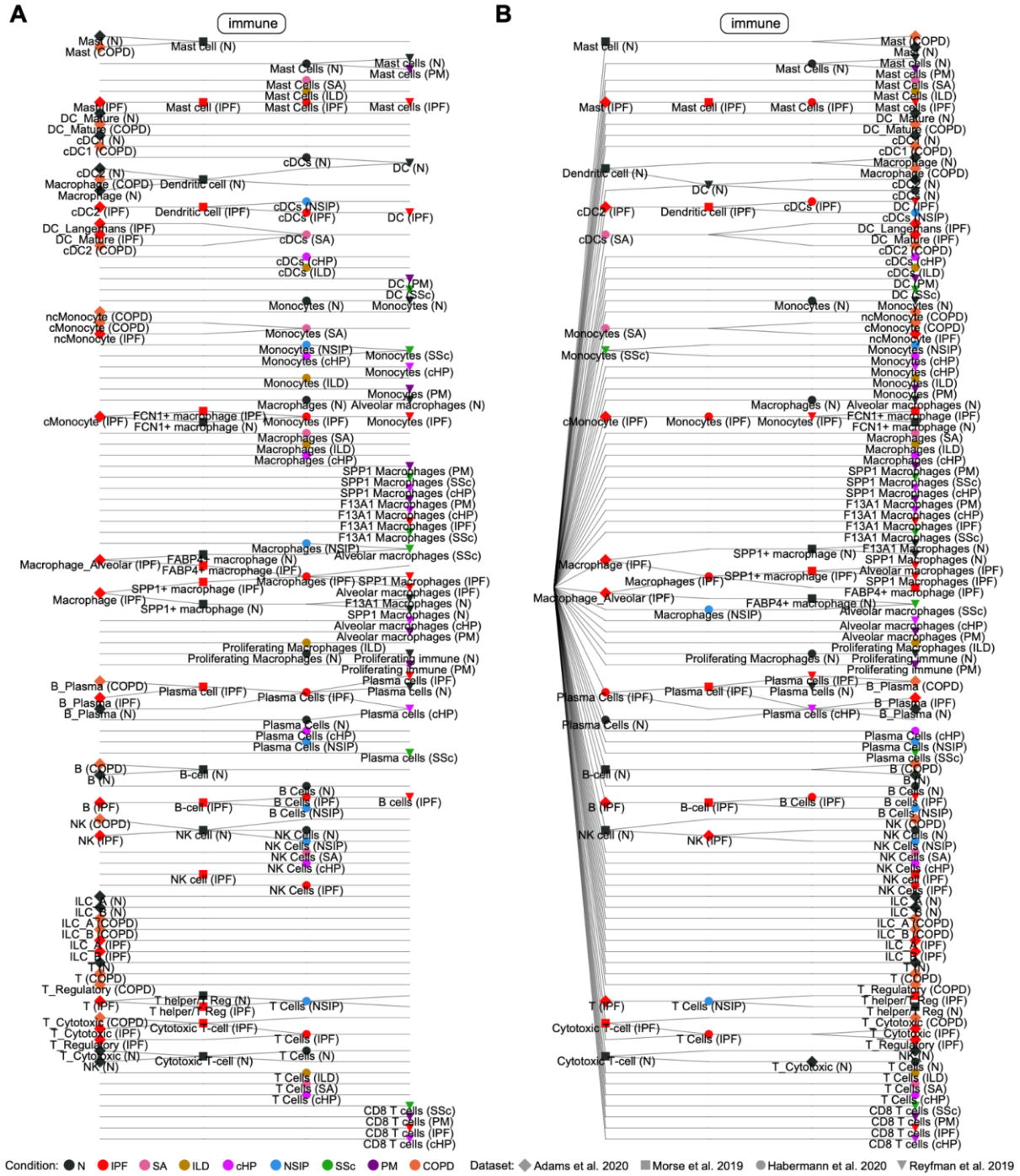


Figure S8. Harmonization of all immune cell types in the four lung datasets

(A) Harmonization graph showing cell type relationships across four lung datasets. Color and shape of the dot represent the disease condition and dataset of origin, respectively. N, normal; IPF, idiopathic pulmonary fibrosis; COPD, chronic obstructive pulmonary disease; NSIP, nonspecific interstitial pneumonia; cHP, chronic hypersensitivity pneumonitis; ILD, interstitial lung disease; SA, sarcoidosis; SSc, systemic sclerosis; PM, polymyositis.

(B) Cell type hierarchy reorganized from (A).

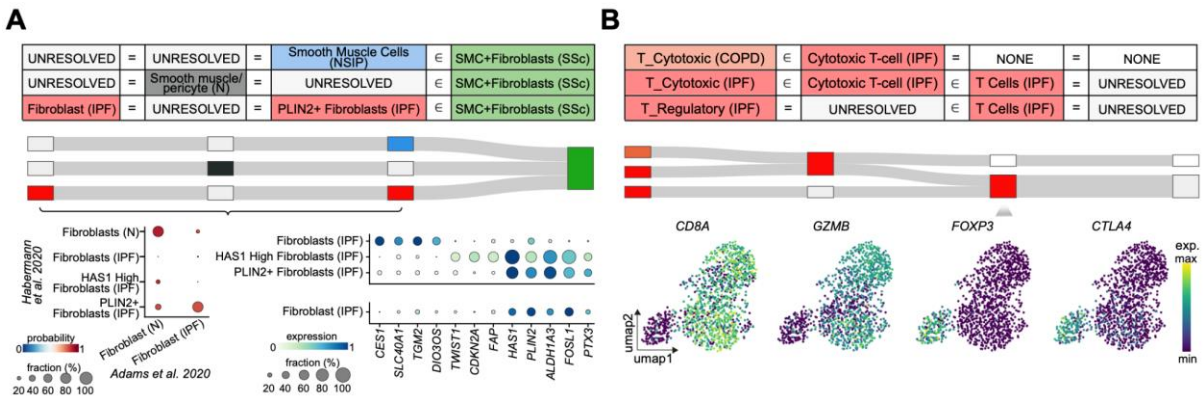


Figure S9. Examples of pathological cell states harmonized by CellTypist

(A) Upper: part of the harmonization table containing fibroblasts and smooth muscle cells, as well as their alignments visualized as a Sankey plot. The order of datasets is, from left to right, Adams et al. 2020, Morse et al. 2019, Habermann et al. 2020, to Reyfman et al. 2019, and applies to (B). Lower: scatter plot showing label transfers from cell types in the third dataset to those in the first dataset, highlighting a correspondence between "PLIN2+ Fibroblasts" and "Fibroblast". Color and size of the dot denote the mean prediction probability and cell type alignment fraction, respectively. Dot plot to the right shows expression of marker genes for the "Fibroblast" identified (bottom) as compared to those from the third dataset (top), with color and size of the dot representing normalized gene expression and percentage of cells expressing a given gene, respectively.

(B) Upper: as in (A), but for a subset of T cell populations. Lower: UMAP representations overlaid by expression of marker genes of cytotoxic (*CD8A*+, *GZMB*+) and regulatory (*FOXP3*+, *CTLA4*+) T cell populations which together constitute the "T Cells" of the third dataset.

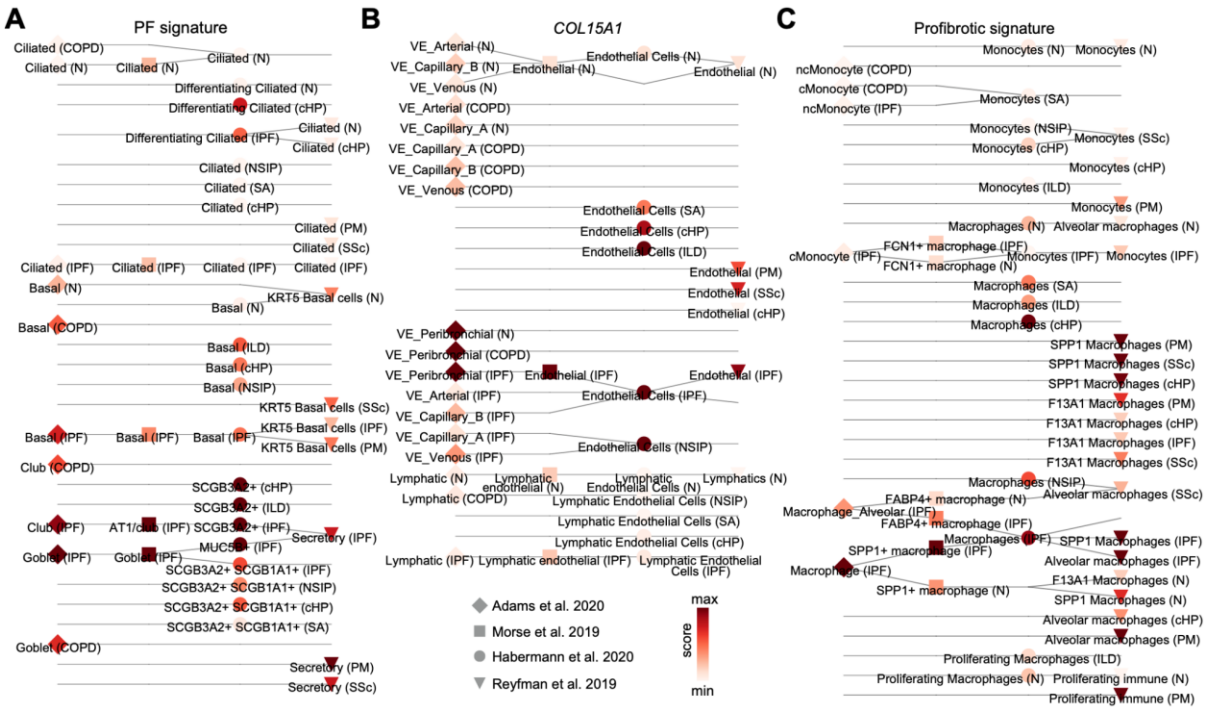


Figure S10. Distribution of different gene signatures across harmonized lung cell types
(A-C) Part of the harmonization graph that focuses on the pulmonary fibrosis (PF)-defining signature in airway epithelial cells (A), expression of *COL15A1* in endothelial cells (B), and the profibrotic signature in monocytes and macrophages (C). Color and shape of the dot represent the degree of a given signature and dataset of origin, respectively.

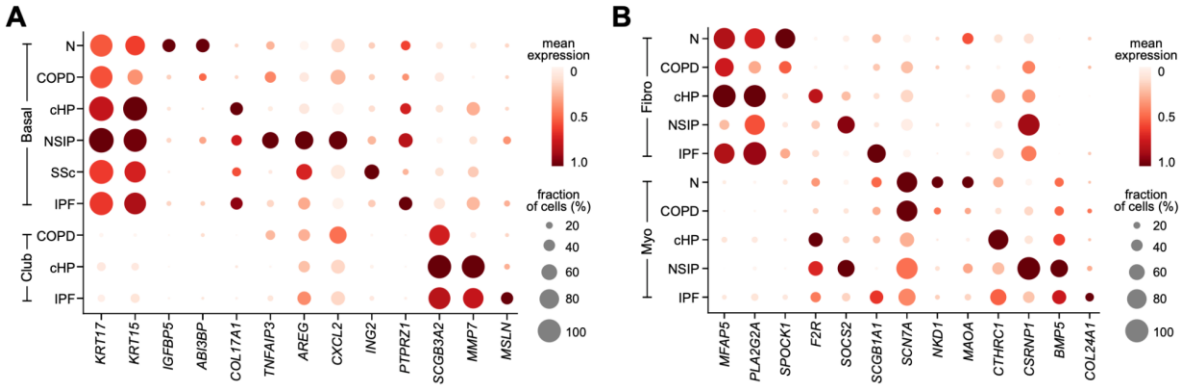


Figure S11. Molecular alterations in lung cell types across disease conditions

(A) Dot plot showing expression of differentially expressed genes among basal and club cell types and among disease conditions, with the color and size representing normalized gene expression and the percentage of cells expressing a given gene, respectively. Cell types and diseases are grouped across datasets based on their alignments inferred by CellTypist. N, normal; IPF, idiopathic pulmonary fibrosis; COPD, chronic obstructive pulmonary disease; NSIP, nonspecific interstitial pneumonia; cHP, chronic hypersensitivity pneumonitis; ILD, interstitial lung disease; SA, sarcoidosis; SSc, systemic sclerosis; PM, polymyositis.

(B) As in (A), but for the fibroblasts and myofibroblasts in the four datasets.

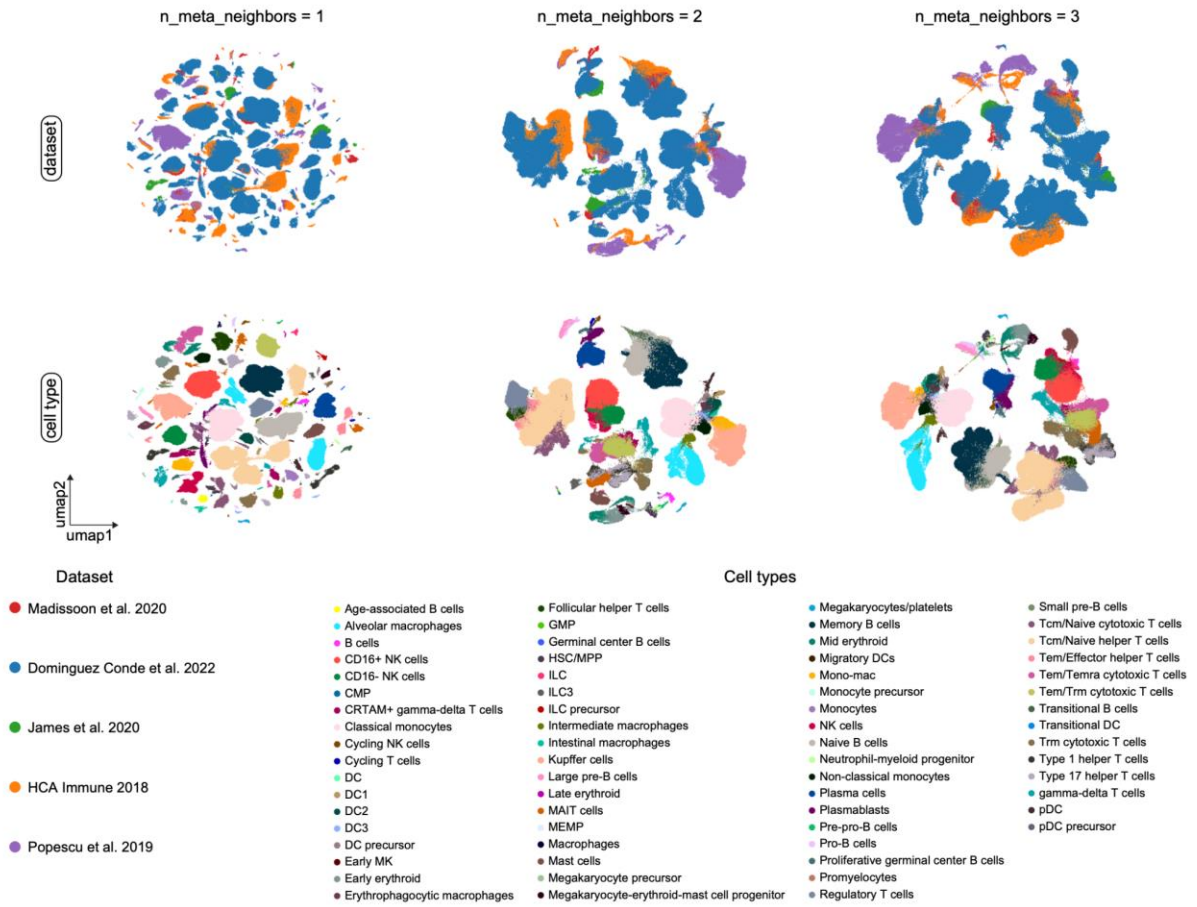


Figure S12. Impact of the number of cell type neighbors on transcriptomic structures UMAP representations of cells from the five immune datasets overlaid by the dataset of origin (top) and cell type (bottom). Transcriptomic structures are shown with different numbers of neighbors to anchor for each cell type when determining the group size of cell types: 1 (left), 2 (middle), and 3 (right).

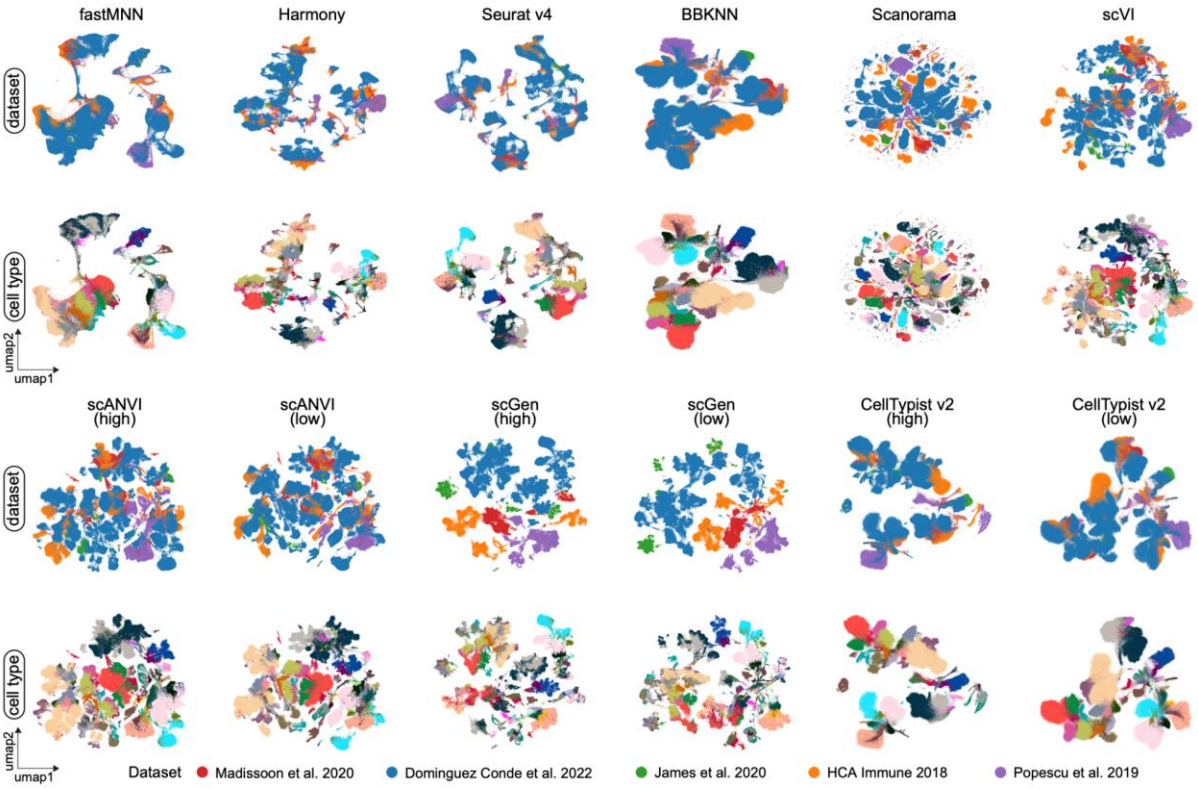


Figure S13. Transcriptome organizations of immune cells integrated using different methods

UMAP representations of cells from the five immune datasets integrated using six unsupervised (top two rows) and three supervised (bottom two rows) methods. For each method, cells are colored by the dataset of origin (upper) and the cell type (lower). For supervised methods, both high- and low-hierarchy annotations are used. Color codes of datasets and cell types can be found in Figure S12.

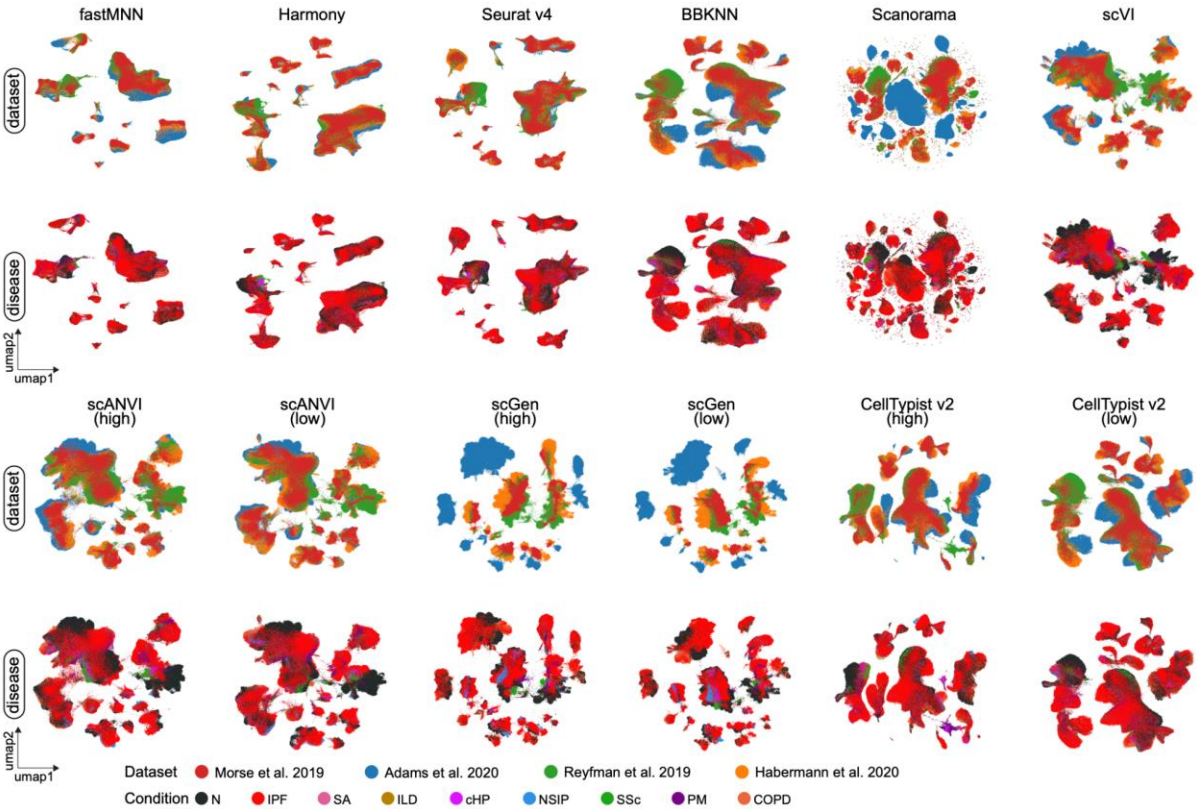


Figure S14. Transcriptome organizations of lung cells integrated using different methods UMAP representations of cells from the four lung datasets integrated using six unsupervised (top two rows) and three supervised (bottom two rows) methods. For each method, cells are colored by the dataset of origin (upper) and the disease condition (lower). For supervised methods, both high- and low-hierarchy annotations are used. N, normal; IPF, idiopathic pulmonary fibrosis; COPD, chronic obstructive pulmonary disease; NSIP, nonspecific interstitial pneumonia; cHP, chronic hypersensitivity pneumonitis; ILD, interstitial lung disease; SA, sarcoidosis; SSc, systemic sclerosis; PM, polymyositis.

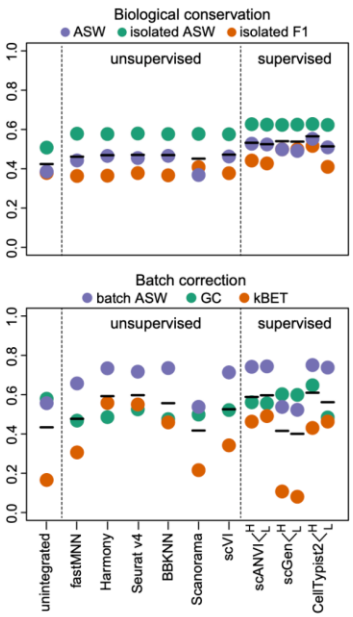


Figure S15. Assessment of biological conservation (upper) and batch removal (lower) after lung data integration using the six unsupervised and three supervised methods
Names of tools are shown on the bottom. Both high- and low-hierarchy annotations are used for supervised methods. Horizontal black lines denote the averages of the integration metrics. ASW, average silhouette width; GC, graph connectivity; kBET, k-nearest-neighbor batch-effect test.