Knowledge-based classification of fine-grained immune cell types in single-cell RNA-Seq data with ImmClassifier

Authors

Xuan Liu¹, Sara J.C. Gosline², Lance T. Pflieger³, Pierre Wallet³, Archana Iyer⁴, Justin Guinney², Andrea H. Bild³, Jeffrey T. Chang^{1*}

Affiliations

¹ Department of Integrative Biology & Pharmacology, University of Texas Health Science Center at Houston, Houston, TX 77030, USA

² Sage Bionetworks, Seattle, WA 98109, USA

³ Department of Medical Oncology & Therapeutics Research, City of Hope National Medical Center, Duarte, CA 91010, USA

⁴ Center for Cancer Systems Immunology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

* Corresponding author

Corresponding author

Jeffrey T. Chang, Ph.D.
Associate Professor

Department of Integrative Biology & Pharmacology,
University of Texas Health Science Center at Houston,
Houston, TX 77030, USA.

Email: Jeffrey.T.Chang@uth.tmc.edu

Abstract

Single-cell RNA sequencing is an emerging strategy for characterizing the immune cell population in diverse environments including blood, tumor or healthy tissues. While this has traditionally been done with flow or mass cytometry targeting protein expression, scRNA-Seq has several established and potential advantages in that it can profile immune cells and non-immune cells (e.g. cancer cells) in the same sample, identify cell types that lack precise markers for flow cytometry, or identify a potentially larger number of immune cell types and activation states than is achievable in a single flow assay. However, scRNA-Seq is currently limited due to the need to identify the types of each immune cell from its transcriptional profile, which is not only time-consuming but also requires a significant knowledge of immunology. While recently developed algorithms accurately annotate coarse cell types (e.g. T cells vs macrophages), making fine distinctions has turned out to be a difficult challenge. To address this, we developed a machine learning classifier called ImmClassifier that leverages a hierarchical ontology of cell type. We demonstrate that ImmClassifier outperforms other tools (+20% recall, +14% precision) in distinguishing fine-grained cell types (e.g. CD8+ effector memory T cells) with comparable performance on coarse ones. Thus, ImmClassifier can be used to explore more deeply the heterogeneity of the immune system in scRNA-Seq experiments.

Introduction

Single-cell RNA-Seq (scRNA-Seq) has emerged as a powerful technique to catalog cell types [1, 2], including immune cells, which play critical roles in a wide range of diseases. In cancer, they have been shown to impact survival, drug resistance and evolution [3, 4]. However, annotation of the cells based on their transcriptional profiles remains a challenge [5, 6] due both to the diversity of cell types as well as ambiguous divisions along the developmental lineage and activation states [7]. That is, while myeloid and lymphoid cells have drastically different transcriptional profiles and are trivial to distinguish, the differences within each lineage are more subtle—specific types of T cells are difficult to identify [8].

Currently, the most commonly used approach to annotate cell types is to start with an unsupervised clustering algorithm (like t-SNE [9] or UMAP [10]) to group cells with similar profiles, and then to manually inspect each cluster for the expression of marker genes that distinguish specific cell types [11]. While conceptually straightforward, using these markers is challenging in practice due to poor expression or dropout [12, 13], low conservation of markers across studies [14], ambiguity of markers [15], lack of reliable markers [16] and transcriptional similarity of cell types [17]. Further, cell type

annotations are not yet easily transferred between different datasets, and therefore, each data set needs to be manually annotated by experts with an understanding of both immunology and the idiosyncrasies of scRNA-Seq data. As a final complication, cell type annotation is an iterative procedure where the clustering influences the immune cell classifications, which then reveals discrepancies (e.g. a memory T cell cluster that contains both CD4+ and CD8+ T cells) that need to be resolved by refining the clusters (e.g. by clustering with different genes) [18].

To simplify and automate the process of identifying cell types, several bioinformatics methods have recently been developed. Correlation-based methods such as scmap [19] and SingleR [20] correlate the query cells to a pre-defined set of reference cell types and assign the label of the type with maximum correlation. Hierarchy-based methods such as Garnett [21] and CHETAH [22] construct a reference cell type hierarchy and search for the optimal cell type from a generic root node to increasingly more specific types. CellAssign [23] utilizes a Bayesian statistical framework to model cell types, marker gene expression, and other covariates such as processing batches. While these methods have been successful and quickly adopted into scRNA-Seq pipelines, approaches with improved ability in identifying fine-grained cell types are still needed.

To increase prediction accuracy, and in particular, on more fine-grained cell types, we have developed a method that includes an explicit knowledge-based and hierarchical model of immune cell types. The immune cell types are naturally organized into class

hierarchy. ImmClassifier (Immune cell classifier) integrates the biology of immune cells from a hierarchical ontology, and synthesizes heterogeneous reference datasets using a two-step machine learning and deep learning process. Within each reference dataset, the first step trains a random forest classifier to assign probabilities according to the cell types of the reference dataset, which preserves the intra-dataset cell type relations and avoids batch effects when pooling the cells from different reference datasets. To resolve the differences in reference annotations, cell types from all reference datasets were mapped to unified and non-redundant cell ontology hierarchy. The second step employs a deep learning approach to integrate numerous reference datasets and directly learn the cell ontology hierarchy, assigning the optimal annotation based on the distribution of probabilities across the hierarchy. This enables ImmClassifier to synthesize cell type assignments across reference datasets. Evaluating on a number of independent scRNA-Seq datasets, ImmClassifier accurately classified and outperformed existing methods over a variety of immune cells collected from different tissues, especially at coarser annotation granularity. ImmClassifier is available as a Docker container at <https://github.com/xliu-uth/ImmClassifier>.

Results

Evaluating ImmClassifier on tumor-infiltrating immune cells

We developed a computational tool ImmClassifier to annotate immune cells in gene expression data (Methods, Fig 1). We evaluated the performance of ImmClassifier on

three test datasets (not included in the training) [17, 24, 25]. We began by analyzing the microarray profiles of purified cell populations, which lacks some of the technical challenges, such as drop-out, that would be seen in the scRNA-Seq data. In this situation, ImmClassifier recovered 13 out of 15 original cell types and distinguished CD4 from CD8 T cells accurately, recovering 100% of CD4 T cells and 83% of CD8 T cells. 95% of the cells identified under the CD4 T hierarchy were CD4 T cells in the original annotation (Fig. 2a). A more challenging situation was seen in the myeloid lineage, where there was less overall coverage in the reference data sets. None of the five eosinophil samples were identified exactly as eosinophils, although three were correctly placed in the myeloid lineage (macrophage, monocytes, dendritic cells). It is likely that greater coverage of this cell type will be needed to improve its accuracy.

To assess its ability to classify scRNA-Seq data, we ran ImmClassifier on tumor-infiltrating immune cells from two distinct cancer types. ImmClassifier achieved an overall precision of 72% and 77% (Fig. 2b, 2c). Again, the highest rate of misclassification was made across closely related lineages. 55% of the B cells were predicted to be plasma cells, which are terminally differentiated B cells (Fig. 2b), which highlight the challenges in distinguishing related closely cell types.

Comparison to existing methods

To determine whether the performance we observe is comparable to related methods, we compared ImmClassifier against Garnett (extended mode) [21] and SingleR [20], which are commonly used representatives for the hierarchy- and correlation-based

methods. We applied each method to four scRNA-Seq datasets from three independent publications and visualized the predictions using UMAP clustering. These four datasets cover a variety of sequencing specifications including InDrop 3' sequencing, 10X 5' sequencing, 10X 3' sequencing, Smart-Seq2 Full-length (Table S1). In addition, those datasets covered a wide range of immune populations, collected from cancer patients and healthy donors.

We first quantified the performance of those methods by recall and precision for four different depths across our cell hierarchy (Fig. 3a). (Greater depth is finer cell types). Across all conditions, recall and precision decrease with increasing depth, reflecting the difficulty in distinguishing closely related cell types. For coarse cell types (e.g. myeloid vs lymphoid cells) (depth 1), all methods performed well, with over 95% performance overall. However, the performance dropped with finer types (depth 4) (e.g. central vs effector memory T cells). ImmClassifier achieved an average precision of 97%, 76%, 68%, and 27% for depths 1-4, respectively. At depths 1-2, the recall was comparable to that of SingleR, with only a +3% difference (Fig. 3b). However, at depths 3-4, the recall was improved by +20%. For difference in precision was similar, with a -1% difference at depths 1-2, and +14% at depths 3-4. Thus, while all methods performed well with coarse cell types, ImmClassifier was considerably more accurate at higher depths, although challenges still remain.

We also compared the methods with respect to (1) spatial concordance with original annotation, (2) frequency of matched and mis-matched cell types to the original annotation, and (3) the distance between the centroids of the predicted and original cell types. We used a data set comprised of a complex mixture of 12 immune cell types, including lymphoid and myeloid cells collected from four tissues [7]. Here, ImmClassifier achieved a higher mean recall and precision across all cell types (45.5% and 43.2%) than SingleR (35.5% and 37.9%) and Garnett (12.5% and 10.7%), with high visual concordance to original annotations in a UMAP plot (Fig. 4a). To assess the overall similarity of the expression profiles of each predicted cell type, we averaged the expression profiles of all cells of each type and computed the Euclidean distance between the averaged profiles with those of the original cells (Fig. 4b). This revealed that the overall profiles of the cell types from ImmClassifier was most similar to those of the original data type. Furthermore, ImmClassifier could recover 11 of the 12 cell types, in contrast to SingleR (8 cell types) and Garnett (6 cell types) (Fig. 4c, Fig.S1). Notably, ImmClassifier was able to distinguish a mast cell population (73% recall, 88% precision), which was completely missed by the other methods (Fig. 4a,4c). However, all methods failed to identify the cluster of NKT cells. This cell type was missing in the ImmClassifier training set, and both ImmClassifier and SingleR annotated them as a mix of NK cells and T cells (Fig. 4c). Nevertheless, this demonstrates ImmClassifier's capacity to accurately recapitulate the complexity of the immune cell types.

At the deepest level (e.g. distinguishing between central memory and effector memory T cells), the accuracy for ImmClassifier drops to 21.2% recall and 27.1% precision.

At this level, the differences in the transcriptional profiles of these cells are subtle, and biological distinctions ambiguous and under heavy investigation [8]. We noticed that ImmClassifier frequently predicted effector memory T cells to be other types of T cells, and in particular, tissue-resident memory (TRM) T cells (Fig. 5a). TRM is a newly identified type of memory T cell that resides in tissue without recirculating, yet is transcriptionally, phenotypically and functionally distinct from recirculating central and memory T cells [26, 27]. It has been shown that TRM cells exhibit high expression of *CXCR6* and low expression *SELL* [28]. To determine whether these cells may be mis-annotated in the gold standard data set, we examined the ratio of *CXCR6* to *SELL* expression, and see that it is higher in the cells that ImmClassifier predicted to be TRM (Fig. 5b). This suggests that the TRMs were not accurately annotated in the original data set. It is likely that, due to the difficulties in manual annotation, as well as the high degree of expertise required, immune cell subtypes, especially those deep in the hierarchy, are not yet accurately annotated in the existing data sets.

Discussion

Despite the need for improved annotations of immune cells in scRNA-Seq data, it remains a challenging problem, in particular for cells closely related in the developmental lineage. To address this challenge, the use of a cell type hierarchy has emerged as a critical component in the latest cell type annotation tools. For example,

while Garnett is hierarchical, it uses a pairwise classification strategy that does not consider information across the overall ontology, which has been shown in other contexts to improve accuracy [29, 30]. This is in contrast to ImmClassifier, which uses a deep learning framework to model the whole hierarchy and assign the probabilities considering information across all cell types simultaneously, which appears to improve the performance at deeper levels of granularity.

Currently, ImmClassifier cannot detect new and intermediate cell types. ImmClassifier is trained on known cell types and will assign a cell of an unobserved type to its closest known reference cell type. The limit will need to be addressed in future work. To allow identification of intermediate cell types, the continuous distribution of probabilities that ImmClassifier generates in its intermediate step may be informative. But perhaps most importantly, as well annotated scRNA-Seq data sets grow, new reference cell types may be integrated in a straightforward manner within the architecture of ImmClassifier.

We anticipate that the performance of ImmClassifier will continue to increase as additional high-quality data sets become available. Indeed, our results demonstrate the need for a greater quantity and quality of annotated immune cell data set, especially with fine-grained cell types, to train this and other classifiers. The ImmClassifier framework is scalable to new reference data sets, with the most difficult step in mapping the original annotations to the cell hierarchy. We anticipate that ImmClassifier will be

most useful in experimental settings that require accurate, comprehensive and robust immune cell annotation.

Acknowledgement

Research reported in this publication is supported by the National Cancer Institute of the National Institutes of Health under grant number U54CA209978 and the Cancer Prevention and Research Institute of Texas (RP170668). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests

The authors declare no competing interests.

References

1. Svensson, V., R. Vento-Tormo, and S.A. Teichmann, *Exponential scaling of single-cell RNA-seq in the past decade*. Nature protocols, 2018. **13**(4): p. 599.
2. Heath, J.R., A. Ribas, and P.S. Mischel, *Single-cell analysis tools for drug discovery and development*. Nature reviews Drug discovery, 2016. **15**(3): p. 204.
3. Gajewski, T.F., H. Schreiber, and Y.-X. Fu, *Innate and adaptive immune cells in the tumor microenvironment*. Nature immunology, 2013. **14**(10): p. 1014.
4. Taube, J.M., et al., *Association of PD-1, PD-1 ligands, and other features of the tumor immune microenvironment with response to anti-PD-1 therapy*. Clinical cancer research, 2014. **20**(19): p. 5064-5074.
5. Papalexi, E. and R. Satija, *Single-cell RNA sequencing to explore immune cell heterogeneity*. Nature Reviews Immunology, 2018. **18**(1): p. 35.
6. Tirosh, I., et al., *Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq*. Science, 2016. **352**(6282): p. 189-196.
7. Azizi, E., et al., *Single-cell map of diverse immune phenotypes in the breast tumor microenvironment*. Cell, 2018. **174**(5): p. 1293-1308. e36.

8. Jameson, S.C. and D. Masopust, *Understanding subset diversity in T cell memory*. Immunity, 2018. **48**(2): p. 214-226.
9. Maaten, L.v.d. and G. Hinton, *Visualizing data using t-SNE*. Journal of machine learning research, 2008. **9**(Nov): p. 2579-2605.
10. McInnes, L., J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction*. arXiv preprint arXiv:1802.03426, 2018.
11. Kiselev, V.Y., T.S. Andrews, and M. Hemberg, *Challenges in unsupervised clustering of single-cell RNA-seq data*. Nature Reviews Genetics, 2019. **20**(5): p. 273-282.
12. Kharchenko, P.V., L. Silberstein, and D.T. Scadden, *Bayesian approach to single-cell differential expression analysis*. Nature methods, 2014. **11**(7): p. 740.
13. Stoeckius, M., et al., *Simultaneous epitope and transcriptome measurement in single cells*. Nature methods, 2017. **14**(9): p. 865.
14. Nirmal, A.J., et al., *Immune cell gene signatures for profiling the microenvironment of solid tumors*. Cancer immunology research, 2018. **6**(11): p. 1388-1400.
15. Boesch, M., A. Cosma, and S. Sopper, *Flow Cytometry: To Dump or Not To Dump*. The Journal of Immunology, 2018. **201**(7): p. 1813-1815.
16. Godfrey, D.I., et al., *NKT cells: facts, functions and fallacies*. Immunology today, 2000. **21**(11): p. 573-583.
17. Novershtern, N., et al., *Densely interconnected transcriptional circuits control cell states in human hematopoiesis*. Cell, 2011. **144**(2): p. 296-309.
18. Zilionis, R., et al., *Single-cell transcriptomics of human and mouse lung cancers reveals conserved myeloid populations across individuals and species*. Immunity, 2019. **50**(5): p. 1317-1334. e10.
19. Kiselev, V.Y., A. Yiu, and M. Hemberg, *scmap: projection of single-cell RNA-seq data across data sets*. Nature methods, 2018. **15**(5): p. 359.
20. Aran, D., et al., *Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage*. Nature immunology, 2019. **20**(2): p. 163.
21. Pliner, H.A., J. Shendure, and C. Trapnell, *Supervised classification enables rapid annotation of cell atlases*. Nature Methods, 2019.
22. de Kanter, J.K., et al., *CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing*. bioRxiv, 2019: p. 558908.
23. Zhang, A.W., et al., *Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling*. Nature methods, 2019: p. 1-9.
24. Sade-Feldman, M., et al., *Defining T cell states associated with response to checkpoint immunotherapy in melanoma*. Cell, 2018. **175**(4): p. 998-1013. e20.
25. Puram, S.V., et al., *Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer*. Cell, 2017. **171**(7): p. 1611-1624. e24.
26. Schenkel, J.M. and D. Masopust, *Tissue-resident memory T cells*. Immunity, 2014. **41**(6): p. 886-897.
27. Zhang, L. and Z. Zhang, *Recharacterizing Tumor-Infiltrating Lymphocytes by Single-Cell RNA Sequencing*. Cancer immunology research, 2019. **7**(7): p. 1040-1046.
28. Savas, P., et al., *Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis*. Nature medicine, 2018. **24**(7): p. 986-993.
29. Silla, C.N. and A.A. Freitas, *A survey of hierarchical classification across different application domains*. Data Mining and Knowledge Discovery, 2011. **22**(1-2): p. 31-72.

30. Silla Jr, C.N. and A.A. Freitas. *A global-model naive bayes approach to the hierarchical prediction of protein functions*. in *2009 Ninth IEEE International Conference on Data Mining*. 2009. IEEE.

Figure Legends

Figure 1. ImmClassifier Architecture. ImmClassifier takes a matrix containing the gene expression of immune cells from a single-cell RNA-Seq experiment (far left). Using feature genes pre-calculated from multiple training datasets, ImmClassifier predicts the probability that each cell corresponds to a cell type annotated in the original training data sets using random forest models (mid left). Here, “DS #1” refers to the first training data set. This probability matrix is further converted to a hierarchical cell type probability matrix using deep neural network classifiers (DNNs). The mean and standard deviation of probabilities from the DNN models is incorporated into a cell ontology hierarchy. Traversing the cell ontology hierarchy, the cell type with maximum information gain is assigned. **(b)** Depiction of the cell types derived from the EBI Cell Ontology to enable machine learning probabilities. Types of immune cells are represented according to granularity, from coarse cell types at the root, to fine-grained ones in the leaf nodes. There are 35 cell types in all. **(c)** Each flat cell type is converted to hierarchical cell type based on the path from root to the flat cell type. The hierarchical cell type is encoded in a 35-bit binary vector by marking the flat cell types on the path 1

and 0 otherwise. The binary vector is used as a target in the DNN training and output in prediction.

Figure 2. ImmClassifier predicts immune cell types. These heatmaps compare the cell types from the original publication (rows) to those inferred by ImmClassifier (columns). In the recall plot (left), the color represents the recall (as a percent) of each original cell type predicted by ImmClassifier. In the precision plot (right), the color represents the precision of each original cell type predicted by ImmClassifier. Recall or precision scores no less than 20 are labeled. Three datasets tested are **(a)** purified immune populations sequenced by microarray platform [17]. **(b)** SKCM [24]. The *Other* row includes a small number of Megakaryocytes, Mast cells, Eosinophils and Neutrophils. **(c)** HNSCC [25]. The *Other* row includes a minimal number of NK cells, Erythrocytes, Megakaryocytes, Monocytes and Neutrophils. Since ImmClassifier has finer annotation granularity than the original annotation, for the purpose of comparison, the annotation terms of equivalent granularity to the original annotation were used.

Figure 3. Classification accuracy at different annotation granularities. **(a)** These boxplots show the recall and precision across cell types, organized by depth. Depth 1 includes CD34+, M and L; depth 2 includes HSC, Dendritic, Macrophage, Monocyte, Neutrophil, Mast, T, NK, B; depth 3 includes CD4+T, CD8+ T, mDC, pDC and depth 4 includes CD4+Tn, CD4+Tcm, CD4+Tem, CD4+Tex, CD4+ Treg, CD4+ Tfh, CD8+Tn, CD8+Tcm, CD8+Tem, CD8+Tex and CD8+ MAIT. **(b)** This table shows the average of

performance ImmClassifier compared to Garnett and SingleR. The median value of recall/precision across different cell types at each depth for each method was calculated. The difference of median recall/precision (ΔSR , ΔGN) between ImmClassifier and Garnett and SingleR, respectively, was calculated. Comparisons in which ImmClassifier has higher accuracy are shown in red, and those with lower accuracy are blue.

Figure 4. Visualization of brca3p dataset by different annotation methods. For visualization, abundant cell populations (cell number >200) were sub-sampled to 200 per cell type. The mapping IDs of cell types between different annotations is listed in Table S2. If the annotation method produces a finer annotation than the original one, the annotation terms of equivalent granularity were used. **(a)** UMAP plots colored by annotation method. In the first row (left to right), the panels are colored by ImmClassifier, SingleR and Garnett. In the second row (left to right), panels are colored based on the annotations in the original publication, or after random shuffling. **(b)** Boxplots of the Euclidean distance between centroids of predicted annotations and original annotations. P-values are by Wilcoxon test. The significance of P-values is shown as ** (<0.01), *** (<0.001) and **** (<0.0001). **(c)** Alluvial plots connect the original annotated cell types to cell types annotated by ImmClassifier (left) and SingleR (right).

Figure 5. ImmClassifier infers tissue-resident memory T cells in brca5p dataset.

(a) This shows the T cell clusters from Fig. S1a in greater detail. The UMAP plots are colored by original annotation, ImmClassifier and SingleR. **(b)** This boxplot shows the ratio of the gene expression of TRM marker genes in three cell populations. The *EM/CM* (effector-memory and central memory T) cells are predicted by ImmClassifier, and consistent with the original annotation. The *Naïve* (naïve T) cells are predicted by ImmClassifier and consistent with the original annotation. The *TRM* (effector-memory and central memory T) cells are predicted by ImmClassifier, and inconsistent with the original annotation.

Methods

ImmClassifier

An outline of ImmClassifier is shown in (Fig. 1a). ImmClassifier employs a machine learning paradigm that takes as input a vector of the expression of n genes of a cell and returns an m length vector of the probabilities that the cell is from one of the m cell types. However, it also deviates from a classical machine learning setup to handle the complexity of the immune cell types and resultant idiosyncrasies in the training sets. Because no single training set includes all desired cell types, multiple ones must be integrated. Furthermore, the cell type labels are frequently inconsistent, using not only different labels, but also labeling at different granularities of the cell type hierarchy. For example, [1] classified T cells into $\alpha\beta$ and $\gamma\delta$ T cells, [2-4] classified T cells as CD4+ and CD8+ T cells and [5] labeled T cells by numeric cluster ID, without an explicit cell type. To resolve these differences, ImmClassifier takes a stepwise approach where independent classifiers are developed for each training set, and the outputs for each training set are resolved by a final classifier that determines the ultimate cell type assignment. Specifically, the input goes through a random forest classifier for each training set (currently, seven). The output matrices are concatenated and processed by ten independently trained deep neural network classifiers so that a robust estimate of the average performance can be estimated. The mean and standard deviation of the scores for each query cell in each cell type were calculated and projected to a cell ontology hierarchy. The cell type with maximum information gain, relative to its parent node, was assigned to each query cell.

Reference data sets

To address the heterogeneity of immune cell types and sequencing platforms, seven independent scRNA-Seq datasets covering a wide range of cell types, sources, and sequencing platforms, were used for training (Table S3). Seven independent test data sets were used. Each reference dataset was divided into a training and a test set. The reference datasets are normalized to logarithm of counts per million reads ($\log_2$ CPM). Non-immune cells were excluded.

Dataset-specific classification

Feature genes are obtained for each reference dataset. If the cluster-associated marker genes are provided by the original publication, the top 20 marker genes are used. Otherwise, the cell type markers are called by Seurat [6] using the gene expression count and provided cell type annotations. To account for drop-out in scRNA-Seq data, only genes expressed in the query dataset are used.

To reduce the bias introduced by cluster size, a balanced training set is generated by randomly selecting 500 cells per cell type (without replacement for abundant cell types (>1000 cell) and with replacement for rare cell types (<1000 cell)). Batch correction is performed between the query dataset and a reference dataset on feature genes using ComBat [7]. The original cell type labels from the reference dataset were used as the target variable. A random forest classifier is trained using default parameters and evaluated for each reference dataset using MLR [8].

Hierarchical immune cell annotations

To integrate annotations and resolve differences across data sets, we developed a hierarchical set of cell types (Fig. 1b and Table S4). The cell types are found from Ontobee [9] (matching based on names and marker genes) and organized using the EBI Cell Ontology [10]. In this hierarchy, the top or root node represents the coarsest cell type, and successive levels describe increasingly finer ones. More precisely, child nodes are related to their parents via “is-a” relationships. This provides a framework to synthesize annotations from the reference data sets provided at different levels of granularity. Using this hierarchy, the 171 cell types (including redundancies) from the seven reference datasets are unified to a set of 35 non-redundant and hierarchical cell types (Fig. S2).

Integrating annotations using a Deep Neural Network (DNN)

We predict the types in the hierarchy as a multi-class classification problem, leveraging the ability of DNNs to learn complex structure [11]. Cells were subsampled per cell type to avoid the dominance by cell number (Table S5). For each cell, the input layer takes the concatenation of the probability vectors from the dataset-specific classifiers and the output layer returns a vector of probabilities associated with hierarchical cell types (Fig. 1c). The position of each cell type in the hierarchy is encoded as a 35-bit vector that traces the path from the cell type to the root. There are three hidden layers with 200,

400 and 200 nodes, respectively, a topology that worked well in the training set, and the dropout rate is 0.2. The DNN is implemented as a multi-label multi-class classification model using keras [12] and tensorflow [13]. We used 10 trained DNN classifiers with identical hyper-parameters to produce a robust estimate of the mean and standard deviation of the probabilities for the subtypes. 10 DNN classifiers were trained using the balance dataset with epochs = 5 and batch_size = 4096.

Hierarchical cell type assignment

For each query cell, we assigned its cell type based on the cell type hierarchy that exhibited the maximum entropy gain [14]. For each node (cell type) c on the cell type hierarchy, the entropy is calculated

$$Entropy(c) = -p(c) * \log(p(c)) - (1 - p(c)) * \log(1 - p(c))$$

where p is the mean probability of that the query cell belongs to cell type c from the 10 DNN models above. Entropy is 0 when p equals to 0, by definition.

We adapted the well-known decision tree algorithm and revised the information gain from [15]

$$Information\ gain(c) = entropy(c) - \sum_{v \in desc(c)} entropy(v)$$

where c is the cell type and $desc(c)$ is the set of child nodes of c .

This will select the cell type that can be assigned most confidently. To break ties in ranking, the cell type with the largest ratio of standard variation to mean is chosen.

SingleR and Garnett predictions

SingleR annotation was generated by running SingleR using default parameters except for `normalize.gene.length=T`. Garnett annotation was generated by Garnett extended mode using the pre-trained classifier `hsPBMC`. The ImmClassifier annotation was generated using default parameters. Annotations were mapped across algorithms, as shown in Table S5.

UMAP clustering and alluvial plots

The spatial coordinates of the cells were obtained using UMAP for each of the four datasets. The UMAP clustering of test datasets was performed using BETSY [16]. For dataset `brca3p`, `brca5p` and `hcc`, the top 500 most variable genes were used for PCA. Top 1000 most variable genes were chosen for dataset `pbmc68k`. Top 10 principle components, the nearest 50 neighbors and `resolution=0.8` were used to generate the tSNE-plots. Alluvial plots were generated using `ggalluvial` [17].

References (Methods)

1. Zheng, C., et al., *Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing*. Cell, 2017. **169**(7): p. 1342-1356. e16.
2. Hay, S.B., et al., *The Human Cell Atlas bone marrow single-cell interactive web portal*. Experimental hematology, 2018. **68**: p. 51-61.
3. Oetjen, K.A., et al., *Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry*. JCI insight, 2018. **3**(23).
4. Azizi, E., et al., *Single-cell map of diverse immune phenotypes in the breast tumor microenvironment*. Cell, 2018. **174**(5): p. 1293-1308. e36.
5. Zilionis, R., et al., *Single-cell transcriptomics of human and mouse lung cancers reveals conserved myeloid populations across individuals and species*. Immunity, 2019. **50**(5): p. 1317-1334. e10.
6. Butler, A., et al., *Integrating single-cell transcriptomic data across different conditions, technologies, and species*. Nature biotechnology, 2018. **36**(5): p. 411.

7. Leek, J.T., et al., *The sva package for removing batch effects and other unwanted variation in high-throughput experiments*. Bioinformatics, 2012. **28**(6): p. 882-883.
8. Bischl, B., et al., *mlr: Machine Learning in R*. The Journal of Machine Learning Research, 2016. **17**(1): p. 5938-5942.
9. Xiang, Z., et al. *Ontobee: A linked data server and browser for ontology terms*. in *ICBO*. 2011.
10. Jupp, S., et al. *A new Ontology Lookup Service at EMBL-EBI*. in *SWAT4LS*. 2015.
11. Webb, S., *Deep learning for biology*. Nature, 2018. **554**(7693).
12. Gulli, A. and S. Pal, *Deep Learning with Keras*. 2017: Packt Publishing Ltd.
13. Abadi, M., et al. *Tensorflow: A system for large-scale machine learning*. in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 2016.
14. Shannon, C.E., *A mathematical theory of communication*. Bell system technical journal, 1948. **27**(3): p. 379-423.
15. Clare, A., *Machine learning and data mining for yeast functional genomics*. 2003, University of Wales, Aberystwyth.
16. Chen, X. and J.T. Chang, *Planning bioinformatics workflows using an expert system*. Bioinformatics, 2017. **33**(8): p. 1210-1215.
17. Brunson, J.C., *ggalluvial: Alluvial Plots in 'ggplot2'. R package version 0.11.1* <https://CRAN.R-project.org/package=ggalluvial>. 2019.

Figure 1

Figure 2

Figure 3

b

	Δ Recall								Δ Precision							
	Δ GN	Δ SR	Δ GN	Δ SR	Δ GN	Δ SR	Δ GN	Δ SR	Δ GN	Δ SR	Δ GN	Δ SR	Δ GN	Δ SR	Δ GN	Δ SR
pbmc68k	28	3	20	6	41	6	28	28	1	-3	4	-3	9	13	20	20
brca3p	25	0	43	0	54	43	14	6	3	-1	35	7	26	18	10	-6
brca5p					22	13	9	7					6	13	33	6
hcc					85	17	34	34					53	6	46	46
average	26	2	32	3	50	20	21	19	2	-2	20	2	23	12	27	16
	1		2		3		4		1		2		3		4	

Depth

Δ GN: ImmClassifier-Garnett
 Δ SR: ImmClassifier-SingleR

Figure 5