Hierarchical cell-type identifier accurately distinguishes immune-cell subtypes enabling precise profiling of tissue microenvironment with single-cell RNA-sequencing

Joongho Lee [a], Minsoo Kim [a], Keunsoo Kang[b], Chul-Su Yang[c] and Seokhyun Yoon [d*]

[a] Dept. of Computer Science, College of SW Convergence, Dankook University, Yongin-si, Korea, 16890
[b] Dept. of Microbiology, College of Natural Sciences, Dankook University, Cheonan-si, Korea, 31116
[c] Dept. of Molecular and Life Science, Center for Bionano Intelligence Education and Research, Hanyang University, Ansan, Korea, 15588
[d] Dept. of Electronics & Electrical Eng., College of Engineering, Dankook University, Yongin-si Korea, 16890

[*] Corresponding: syoon@dku.edu

ABSTRACT

Single-cell RNA-seq enabled in-depth study on tissue micro-environment and immune profiling, where a crucial step is to annotate cell identity. Immune cells play key roles in many diseases while their activities are hard to track due to their diverse and highly variable nature. Existing cell-type identifiers had limited performance for this purpose.

We present HiCAT, a hierarchical, marker-based cell-type identifier utilizing gene set analysis for statistical scoring for given markers. It features successive identification of major-type, minor-type and subsets utilizing subset markers structured in a three-level taxonomy tree.

Comparison with manual annotation and pairwise match test showed HiCAT performs better than others in major and minor-type identification. For subset identification, since no dataset provides canonical subset annotation, we qualitatively evaluated marker expression profile, where HiCAT was the best in the sense that, with its annotation, marker genes were expressed mostly in the designated subset while rarely expressed in non-designated subsets.

Key words: Cell-type identification, Marker-based, hierarchical identification, using cell-type taxonomy tree, single-cell RNA-seq.

ABBREVIATIONS

AUC: Area under ROC curve, DC: dendritic cell, DB: database, EM: Expectation maximization, GEP: gene expression profile, GMM: Gaussian mixture model, GSA: Gene set analysis, ROC: Receiver operating characteristic, scRNA-seq: single-cell RNA-seq, tSNE: t-distributed stochastic neighbor embedding, UMAP: Uniform Manifold Approximation and Projection, Th: T helper cell, Treg: regulatory T cell.

INTRODUCTION

Since the introduction of scRNA-seq technologies[1,2,3], it is becoming one of standardized methods for many biological researches. Unlike bulk RNA-seq technologies, scRNA-seq enabled microscopic studies at single-cell resolution on intra- and inter-cellular activities in heterogenous tissue. Studies on tumor microenvironment and immune-profiling of many disease[4,5,6,7] are being conducted by many groups. An essential step to scRNA-seq analysis is cell-type identification[8,9] since a tissue is heterogeneous complex of cells with diverse phenotypic features[10] interacting with each other. Specifically, identification of immune cells and inferring their activation state are important issue in those studies. Due to its significance, there were vast amount of research for cell-type identification, either marker- or reference-based approaches. The former utilizes existing cell markers for given GEP. Garnett[11], SCINA[12], scSorter[13], scType[14], scCatch[15], DigitalCellSorter[16,17], CellAsign[18] and MarkerCount[19] are of this class. While, the latter utilizes prior annotation to obtain cell-type profiles or markers to be used for identification[20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37]. Although manual annotation using Seurat[38], SCRAN[39], SC3[40] or SCANPY[41] are also usable, it is time-consuming and not scalable for the explosive growth of scRNA-seq data generation. Among these cell-type identifiers, the majority was the reference-based[9]. In comparisons of the two approaches[19] with the same datasets, the reference-based approaches were slightly better than the marker-based one. It can be due to the lack of good markers uniformly applicable to different tissues and/or its usage was not suitable. Compared to the marker-based approach, the reference-based identifiers search for data/tissue-specific markers from the prior annotation possibly leading to better results, even though one need to prepare the reference first.

Despite these tendencies, we focus on the marker-based approach since it is easy to use requiring only markers. It can be a better option if we have good markers and utilize them appropriately. Then, how one can improve the marker-based approach? Our answer is to use the markers in a structured manner. There were studies on transcriptome-based cell-type taxonomy for scRNA-seq experiment[42, 43]. Based on transcriptomic profile, one can define different levels of cell-type taxonomies, for example, major-type, minor-type, and subset, for each of which one can devise identifiers separately and organize them in a hierarchy. In fact, distinguishing T cells and myeloid cells is not the same as classification of myeloid cell minor-types (e.g., distinguishing DC, monocyte, and macrophage). Classification of subsets, such as Tregs, Th1,2,9,17 and 22, might be even different from minor-type classification since they typically represent different polarization states of the same minor type. Common markers in many subsets make subset identification harder, and one must not rely only on a few markers to annotate subset. For example, *SELL*(*CD62L*) is a common marker of monocyte, naïve CD4 T and Tregs. For such subsets, a better decision can be made if their minor-type is known *a priori*.

With these observations, we present HiCAT, a hierarchical, marker-based cell-type annotation tool, featuring improved performance for subsets, particularly of immune cells. The method utilizes subset markers structured in a taxonomy tree. Although it roughly corresponds to the lineage tree, we rather define them by their transcriptomic profiles for classification purposes. The brief procedure and the taxonomy tree used are depicted in Figure 1, where the annotation is performed successively for major-type, minor-type and subsets along the tree. Although marker lists are defined only for subsets, those for minor- or major-types is obtained from subset markers by taking union for the subsets belonging to a specific minor or major-type. There are previous works utilizing hierarchy in reference-based approaches[20,37], while we applied to marker-based method in a more structured way utilizing the taxonomy tree.

RESULTS

The identification performances were evaluated for major-type, minor-type and subsets, separately.

Major-type identification performance

***HiCAT shows the best performance in most datasets*:** Figure 2(a) shows comparisons of major-type

identification performance with the existing marker-based identifiers for 8 datasets with manual annotation available, where 2 datasets (BRCA 100K and Melanoma 5K) contain tumor cells. In Figure 2(a), we used five performance criteria used in previous works[19, 37], i.e., correct(C), error(E), correctly-unassigned(CUA), erroneously-assigned(EA), erroneously-unassigned(EUA). HiCAT was the best in terms of correct criterion, except for PBMC 68K, for which scSorter was slightly better than HiCAT. Considering only valid major-types, i.e., not including tumor cells, HiCAT showed mostly greater than 95% accuracy, except for PBMC 68K for which it was 92.7%. Major-types include T cell, B cell, myeloid cell, epithelial cell, fibroblast, endothelial cell, smooth muscle cell and mast cell. These cells are typically well clustered with sufficient separation from other major-types. Even with these characteristics, some identifiers didn't perform good enough. For datasets with tumors, HiCAT almost perfectly identified tumor cells in Melanoma 5K by correctly unassigning any labels, while, in BRCA 100K, most of the identifiers, including HiCAT, annotated them as normal epithelial cell, except for scCatch, which, however, the performance was the worst.

Primary statistics and ROC/AUC show good identifiability of major-types: To give insight into how HiCAT works, we showed in Figure 2(b) and (c) the distribution of major-type GSA scores along with ROC and AUC. Figure 2(b) shows that major-types are largely identifiable using their scores while there are small overlaps between the two distributions of target and non-target indicating the existence of erroneously identified cells. They can be corrected based on their similarity in GEP (proximity in feature space), for example, by using the GMM-based corrections. GMM identifies the region of population for each major-type and by using the likelihood for given GEP, one can improve the performance as shown by the ROC/AUC before and after the GMM-based correction. ROC/AUC shows how well they are separable from other major-types.

Minor-type identification

Performance for minor-types were worse than those for major-types in all identifiers raising question if manual annotation of minor-types is reliable enough: Compared to major-types, minor-types are not well clustered so that the performance is expected to be worse than those for major-types. Figure 3 shows similar plots to Figure 2, where, as expected, the overall performances are much lower than those for major-type. Comparing identifiers, HiCAT was the best in four datasets, Lung 114K, BALF 43K, Colon 365K and BRCA 100K, scType was the best in three datasets, PBMC 68K, Lung 57K and Melanoma 5K, and DigitalCellSorter was the best in CBMC 8K. PBMC 68K was the worst, for which all identifiers showed less than 50% of correct assignment. Manual annotation in CBMC 8K also couldn't be well predicted by most identifiers when compared to other datasets. This raises a question whether manual annotation is reliable enough to evaluate performance against this. In Figure 2(a) and 3(a), the performance was evaluated assuming the manual annotations are perfect. It is plausible assumption for major-types since major-types are well clustered with enough separation between them so that one can easily determine major-types for each separable cluster by checking expression of a few well-known markers, for example *CD3* for T cells, *CD79B* for B cells. Unfortunately, however, it is not the case for minor-types as well as subsets. In many cases, minor-types of a major-type do not clearly separable even though they may form a cluster localized within a larger cluster of a major-type. In cluster-basis annotation, all cells within a cluster (obtained blindly by a clustering algorithm) are assigned by the same cell-type so that the annotation precision is limited by clustering resolution. In major-type annotation, it isn't a big problem since they are largely separable from others. While, in minor-type annotation, a cluster can be spanning over two or more minor-types, with which annotation error may occur.

Pairwise match test shows higher match rate than those with manual annotation and HiCAT showed the best match: Accounting possibly erroneous annotation in minor-types, it would be interesting to check how well prediction results from different identifiers match to each other. Surprisingly, some identifier pairs had better match than pairs with manual annotation. In Table 1(a), we summarized the pairwise

prediction match for minor-type identification where we showed identifier pair and their match rate for the best, 2nd best and the best with manual annotation. The average of the best match rates with manual annotation was 0.71 across all datasets, while the one for the best pairs was 0.82. Of note, in all datasets, HiCAT was one of the best pair. The higher match rate can happen if specific processes and/or algorithms used in two identifiers in a pair are similar to each other so that they produce similar bias and outcomes, regardless of their actual precision. If it is the case, there must exist one majority of pairs in the best pair list. In Table 1(a), however, there exist different pairs, three HiCAT–scSorter, two HiCAT–scType, two HiCAT–Manual annotation and one HiCAT–SCINA pair. Although HiCAT was shown in the best pairs for all the datasets, it is paired with different identifiers and, therefore, the argument above cannot be justified. Even for the best match with manual annotation, HiCAT appeared 5 times among 8 datasets. The best match rate with manual annotation for major-type in Figure 2(a) was higher than 95% while it was 0.71 for minor-type (the average of the last column in Table 1(a)), which reflects difficulty of identifying minor-types compared to major-type. On the other hand, higher average of the best match rates than those with manual annotation reflects the manual annotation of minor-type may contain errors and it is risky to use them as ground truth for performance evaluation and/or as references for reference-based identifiers. The difficulty of minor-type identification is also clear from their GSA score. Figure 3(b) and (c) show the minor-type score distribution and their ROC and AUC for Colon 365K and BRCA 100K, respectively. Compared to those for major-type, two distributions of target and non-target considerably overlap and the AUCs are much smaller than those in Figure 2(c). Even if HiCAT seems to be the best among those datasets, we cannot say it is completely reliable too.

Subset identification

Identification of subsets, for example, M1 and M2 polarization of macrophages or Th1, Th2 and Th17 in CD4+ T cells, is even more challenging since they mostly represent activation states of the same minor-type and, sometimes, the states are reversible. Although some datasets provide subset annotation for some immune cell subsets using some well-recognized markers, they are mostly non-canonical so that they are not reusable for various downstream analysis and, most of all, it is impossible to evaluate performance for subset. Therefore, pairwise match rates were evaluated as did for minor-type.

HiCAT-scType pair shows the best match in all datasets: Table 1(b) summarizes the best, 2nd and 3rd best match pair and their respective match rates. The average match rate over all datasets were 0.67, 0.53 and 0.50 for the best, 2nd and 3rd best pair, respectively, which are all even lower than those for minor-type identification, reflecting it is more difficult to precisely identify cell-type subsets. Regardless of this tendency, HiCAT-scType pair showed the best match rate in all datasets. As mentioned before, it may be because similar processing and algorithms were used in the two identifiers. However, from the results for minor-type identification, it seems not the case. Without the ground truth available, we cannot say which one is better than the other, even though HiCAT was better in minor-type identification.

Qualitative evaluation of marker expression profile: Although quantitative evaluation is impossible, we can give insight into HiCAT performance by showing "marker expression profile" within each identified subset. Figure 4 shows the marker expression profile under HiCAT annotation in Colon 365K for subsets of T cells, ILCs/NK cells, myeloid cells and B cells, where marker genes were grouped for each subset and ordered according to the fraction of cells expressing the marker (dot size). Although this type of analysis is used in manual annotation to determine cell-type for each cluster, here we used it to check how well the identifier determined cell-types. Ideally, the marker genes, except for common markers, must be expressed only in their designated subset so that their expressions in non-designated subsets indicate annotation errors. One thing to note, however, is that subsets are mostly not clearly separable and we need to allow some "expression leakage" in other non-designated subsets. Now, looking at T cell subsets in Figure 4(a), we observe the followings:

(a) The fractions of marker expressing cells in its designated subset were mostly much larger than those in non-designated subset, even though some of the markers appeared not expressed, e.g., *MME* in Tfh, *NRP1* and *TGFB3* in Tregs.
(b) Some markers have non-negligible fraction in non-designated subsets, e.g., *ZFP36* for CD8 T, *CXCR4* for Tfh and Th2. Their mean expression levels, however, are less than those in the designated subset.
(c) *BATF* is common markers of Th17 and Th22, and *STAT3* is common marker of Tfh, Th17 and Th22, so that they were shown to be comparable in all their designated subsets.

The rest marker expression profiles in Figure 4(b)-(d) for other major-types show similar tendencies even though they were not as clear as T cell subsets.

HiCAT annotation shows the best marker expression profile in immune cell subsets: Now, it would be interesting to compare the marker expression profiles for immune cell subsets under the annotations from other methods. Figure 5 shows the marker expression profiles for T cell subsets under the annotations from different tools including HiCAT, SCINA, scSorter, scType, and DigitalCellSorter. The markers with the dot size less than 0.15 were removed for better comparison. Note that rows of a dot plot are the subsets identified by the corresponding tool so that, if there are subsets not showing in the rows, it means the identifier decided those subsets not existing in the datasets. Comparing Figure 5(a) to (e), we observe the followings:

(a) HiCAT, SCINA and scSorter were identified all 9 T cell subsets, while scType identified 5 and DigitalCellSorter did 6 only. Although one can argue that there indeed exist only 5 or 6 T cell subsets, it is not likely because other 3 tools identified all 9 subsets.
(b) The number of markers with its expressing fraction greater than 0.15 was highest under HiCAT annotation, i.e., 7, 3, 8, 5, 3, 3, 4, 5, 12 makers respectively of 9 T cell subsets have their fraction of marker expressing cells greater than 0.15. Compared to HiCAT, the numbers with SCINA annotation were only 7, 3, 7, 4, 2, 2, 4, 3, 3 and those for others were even less.
(c) Compared with HiCAT annotation, other methods were inaccurate in terms of dot sizes. For example, the fraction of *STAT3* expressing cells of Tfh, Th17 and Th22 were much smaller in (b) SCINA and (c) scSorter, respectively, and was shown only in Tfh in (d) and Th22 in (e).

Note that one must not assess the quality of annotation relying only on a few markers or separately for each subset and many markers must be considered for the qualitative evaluation. Overall, the results in Figure 5 implicitly indicate better identification of T cell subsets by using HiCAT.

DISCUSSION

The successive identification gives us a better result in minor-type and subset identification. Specifically, immune cell subsets typically do not have clear boundaries between them. Rather, they form "a continuum beginning with hard and ending with soft distinctions among cell-types with an ambiguous transition between these extremes" [43]. In this case, simultaneous identification of all minor-types or a mix of minor-types and subsets will not give us a good result. MarkerCounter_GSA seems to work well by successively determining cell's identity given their broader class as a condition for statistical scoring of lower classes.

Another aspect of HiCAT is that cluster-basis annotation is used only for major-type identification while kNN is applied for minor-types and subsets. In many identifier, cell-types are assigned in cluster-basis, i.e., the same cell-type is assigned to all cells in a cluster. In this case, annotation accuracy is limited by clustering resolution and it is desired to perform clustering with high resolution setting. However, if it is too high, the statistics on marker expression become inaccurate as the number of cells in a cluster gets smaller. One the other hand, if the resolution is too small, two or more subsets may be put into one cluster leading to incorrect assignment. Therefore, cluster-basis assignment is not a good option for subset annotation, while it is fine for major or possibly minor-types. HiCAT applies cluster-basis assignment only for major-type. While, in minor-type and subset, the annotation is initially made based only on the GSA

score. Cell-type correction is then applied using *k*-nearest neighbor (*k*NN) rule, with different number *k* for minor-type and subset identification.

Another thing to account when using markers for cell-type annotation is the fact that transcript expression does not match to the protein expression. CD4 and CD8 surface markers in T cells are always present on their surface so that one can make sure they are CD4+ or CD8+ using antibodies for physical cell sorting. However, this does not hold when counting RNAs since they are not necessarily present in their transcriptome at the time of measurement. Therefore, annotation using only a few markers defined in protein level is risky to use confidently. If a minor-type is well clustered, one can identify them cluster-by-cluster basis since some portion of the cells in the cluster is expressing the markers in transcriptome level. However, this may not be the case for subsets, and it is desirable to use as many markers as possible simultaneously to make better decisions using their statistics. From these unique features, HiCAT could reliably be used for identification of immune cell subsets. Although we cannot say the subset annotation of HiCAT is completely reliable, one can select confident cells using GSA scores provided in the HiCAT output.

METHOD

Datasets

For performance evaluation, we collected 8 scRNA-seq datasets as summarized in Table 1. The 8 datasets are PBMC 68K(GSE93421)[44], CBMC 8K(GSE100866)[45], Melanoma 5K(GSE72056)[46], Lung 57K(ERP114453)[47], Lung 114K(SRP218543)[48], BALF 43K(GSE158055)[7], Colon 365K(SCP259)[5], and BRCA 100K(GSE176078)[6], where the first three were used many times in previous works for cell-type identification[19,37]. For all the datasets, manual annotations are available for specific cell-type so that one can easily reannotate their minor and major-types for the purpose of performance evaluation both in major-type and minor-type level. Supplemental Table 1 summarizes the basic specification of the datasets. Note that Melanoma 5K and BRCA 100K contain tumor cell population, with which one can evaluate rejection function for unknown cell population. Most of all, the 8 datasets contain lots of immune cells, including T cells, B cells and myeloid cells, which are our main focuses in the evaluation of the proposed method and other state of the art tools. Supplemental Figure 1 and 2 show UMAP plots for BRCA 100K and Colon 365K and comparisons of the manual and HiCAT annotation as examples. As shown, major-types are clearly separable from others so that they are easily identifiable, while minor-types are not clearly separable even though they are mostly localized within a major-type cluster.

HiCAT processing

The building blocks of HiCAT includes (A) PCA and clustering, (B) Marker Counting and GSA scoring, (C) Unknown cluster detection, (D) GMM based correction, (E) Rejecting unclear cells, and (F) kNN based correction. Using these building blocks, HiCAT successively determine major-type, minor-type and subset utilizing the subset markers structured in three-level taxonomy tree. The rationale is that one can possibly make better decision on subset identity knowing their broader type, especially when there exist common markers in several minor-types or subsets from different major-types.

Three-level taxonomy tree: The taxonomy tree was shown Figure 1 (b). Although the three levels of cell-type roughly correspond to the lineage tree, we rather define them by their transcriptomic profiles for classification and clustering purposes. First, major-types are defined as cell populations that are clearly separable from others so that they are easily identifiable in a low-dimensional landscape, e.g., in tSNE or UMAP plots, while minor-types are a subpopulation of a major-type and not clearly separable even though they can be well localized within a major-type cluster. In our taxonomy, CD4, CD8 T cells, ILC and NK cells were set as minor-types of T cells. Subsets are a subpopulation of a minor-type and typically represents specific activation/polarization states of a minor population, e.g., M1 or M2 polarization of macrophage. They are not clearly separable from other subsets and also not so-well clustered such that

blind clustering is not applicable to catch a specific subset. As mentioned, these three-level taxonomies are defined only for the classification purposes and development of hierarchical identification, and do not necessarily correspond to lineage tree or specific activation state. We implemented hierarchical identification based on the taxonomy tree in Figure 1 (b), where, since our primary interest was on immune cells, other non-immune cells were not divided into minor-types or subsets, and this can be considered as that there is only one minor-type or one subset.

Structured marker DB: To implement hierarchical identification, we need list of markers for each subset in the taxonomy tree, which are fed to the identifier as a tap-separated values file having 5 columns (major-type, minor-type, subset, marker type and comma-separate list of markers (gene symbols). The taxonomy tree can be easily recovered from the tuples of (major-type, minor-type and subset). The marker type is one of positive (pos), negative (neg) or secretory (sec), which were adapted from the type of markers in R&D systems (https://www.rndsystems.com/resources/cell-markers) where we obtained the subset markers. Positive markers are either cell surface markers or intracellular markers that are typically expressed in a subset, regardless of its expression level. Negative markers are those not expressed. The negative markers are considered as a hit when evaluating GSA score if the cell does not express the negative markers. Although it provides useful information in minor-type or subset identification within a major-type (e.g., CD4+ and CD8+ T cell identification within a T cell cluster), it may not work in major-type identification if other major-types expresses those negative markers defined only for a few specific minor-type or subset. The last one is the secretory factors, of which coded proteins are secreted by some subsets presenting their respective activation state. Although they are also an important information when identifying subsets, they are highly variable according to the biological state of the tissue they reside. Therefore, we used only positive markers in this work even though the usage of the 3 marker types can be user-specified. Using the subset marker defined in the marker DB, the markers of minor-types and major-types can be defined as the union of markers for the subsets belonging to that minor or major-type.

Input cell-by-gene matrix: Another input to the HiCAT is the cell-by-gene matrix, of which each row is a cell, and each column is a gene. The matrix should be formatted as a data frame, for which the row names are cell barcodes and column names are gene symbols as in typical in a scRNA-seq data. The column names are used to identify the expression state of a specific marker.

Preprocessing: In the preprocessing step, we perform count normalization, log transformation and principal component analysis (PCA) by choosing highly variable genes, similar to those in Seurat[38] or SCANPY[41]. Clustering is then performed to be used for the cluster-basis major-type identification and unknown cluster detection. HiCAT support two clustering algorithms, Louvain's algorithm or GMM-based clustering. User can select either one even though we used Louvain's algorithm in this work.

Major-type identification and unknown cluster detection: This process is performed in 5 steps; (1) computing GSA scores, (2) rejection of unclear cells, (2) detection of "unknown cell-type" clusters, and (4) Gaussian mixture model (GMM) based correction.

Marker counts and gene set analysis (GSA) score: For the first step in the cell-type identification, we need cell-type markers, where, for the major-types, they are given by a union of subset markers as mentioned before. Given the major-type markers, M_i, of size $|M_i|$ for the ith cell-type and the gene expression matrix of size $N_c \times N_g$, the marker counts is given by $n_{ij} = |M_i \cap G_j|$ for the ith cell-type of the jth cell, where G_j is the set of genes expressed in the jth cell, regardless of their expression level. The GSA score is defined by

$$s_{ij} = -\log_{10}(p_{ij}) \tag{1}$$

with the p-value, p_{ij}, given by

$$p_{ij} = 1 - \sum_{k=0}^{n_{ij}-1} \frac{\binom{|M_i|}{k}\binom{N_g-|M_i|}{|G_j|-\mathrm{k}}}{\binom{N_g}{|G_j|}} \tag{2}$$

which is the probability of the marker counts n_{ij} is obtained by chance [49]. Once the scores are obtained, the initial major-type of the jth cell is given by $\mathrm{argmax}_i\, s_{ij}$ and its score is set to $\max_i s_{ij}$. But this process is performed along with the rejection of unclear cells in the next.

Rejection of unclear cells: Next step is to set threshold for rejection. Rejecting unclear cell is performed per-cell basis. Taking the batch effect into account, we need to use cell-type specific threshold that can be obtained from the dataset itself. To this end, we first determine the best and the 2nd best type for each cell given s_{ij} for $i \in T$, $j \in C$, where T is the set of major-types to identify and C is the set of cells. Let $C_i^{(1)}$ and $C_i^{(2)}$ be the set of cells of which their best and the 2nd best type is i. Apparently, they are disjoint to each other. As cells in $C_i^{(2)}$ might be some other types than i with high probability, their score s_{ij} for $j \in C_i^{(2)}$ may give us a clue for the threshold of the ith cell-type. To be specific, define two set of cells for given rejection threshold t, $C_i^{(1)}(t)$ and $C_i^{(2)}(t)$. They are the set of cells from $C_i^{(1)}$ and $C_i^{(2)}$, respectively, of which the GSA score is greater than the threshold t. We obtained the threshold t_i for the ith cell-type as the minimum of t satisfying the false positive rate defined below is less than a certain threshold, $FPR_{\mathrm{threshold}}$, i.e.,

$$FPR(t) = 1 - \frac{\left|C_i^{(2)}(t)\right|}{\left|C_i^{(1)}(t)\right|} < FPR_{\mathrm{threshold}} \tag{4}$$

With the threshold t_i obtained, the cells with its score, $\max_i s_{ij}$ being less than t_i are replaced with 'unassigned'.

Detecting unknown cell-type clusters: It is an important feature of cell-type identifiers to detect unclear cell-type cluster that can be possibly unidentified-so-far or tumor cells. The average score within a cluster can be used as an indicator of unknown cell-type. If there exist some unknown (major) type of cells that are not contained in the marker DB entries, the cell cluster might have much less scores than others. The problem is that it is hard to find a universal threshold that applies to every dataset as they might be obtained by different platforms utilizing different reference genomes and, sometimes, they were preprocessed in different ways. Moreover, different cell-types have different number of markers and different average scores. Therefore, it is not a good idea to use a universal threshold, but to use dataset specific threshold that can be obtained from the dataset itself. To identify those unknown clusters, we devised the following steps.

(a) For each cluster c, the percentage, q_c, of 'usable' cells with its score being greater than or equal to a certain threshold, s_{th}, is computed. Note that $0 < q_c \leq 1$.
(b) We sort q_c's to obtain a sequence $q_1, q_2, \ldots, q_{N_{clusters}}$, to which a linear fit is performed for high ranked q_c's , where the fitting cost J is defined as the sum of squared error weighted by q_k, i.e., $J(a,b) = \sum_{i=n+1}^{N_{clusters}} q_k \cdot (ak + b - q_k)^2$, where n is given by a specific proportion of $N_{clusters}$, e.g., round($\alpha N_{clusters}$) with α =0.3.
(c) Once the fitted line $a^*k + b^*$ is obtained, the decision whether cluster k is kept or discarded is made by a test if $q_k < a^*k + b^* - t$, where $t = (1-\alpha) \cdot \max_j \left|a^*j + b^* - q_j\right|$.

Gaussian mixture model (GMM) based correction: This is to reassign major-type to a cluster unintentionally excluded and or erroneously identified. Suppose that a cluster is marked 'unknown' even if it is not "clearly separable" from some nearby clusters with proper cell-type assignment. If it is not clearly separable from them, it might not be unknown, especially their major-type. To identify those

clusters and reassign a major-type if applicable, we proceed as follows.

(a) For each major-type, say the ith major-type identified in the previous steps, we model their distribution as a mixture of multivariate normal on the dimension reduced space, i.e.,

$$f_i(\boldsymbol{x}) = \sum_{k=1}^{M_i} \pi_{i,k} N\left(\boldsymbol{x}; \boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}\right) \quad (3)$$

where $\pi_{i,k}$, $\boldsymbol{\mu}_{i,k}$ and $\boldsymbol{\Sigma}_{i,k}$ are the size, mean and covariance matrix, respectively, of the kth component of the ith major-type, M_i is the number of mixture components and $\boldsymbol{x}$ is the (PCA transformed) feature vector in the dimension reduced space. Given the major-type assignment performed using GSA score, the model parameters can be obtained by using the expectation maximization algorithm, for which various software tools are available.

(b) With the GMM parameters for all the major-type, we compute the log-likelihood, $l_{ij} = \log f_i(\boldsymbol{x}_j)$, for each cell using its feature vector $\boldsymbol{x}_j$ and the major-type are revised by finding the maximum of l_{ij}, i.e., the major-type of the jth cell is given by $\operatorname{argmax}_i l_{ij}$ and its score by $\max_i l_{ij}$.

(c) *Cluster-basis major-type assignment*: Once unclear clusters were excluded, major cell-types are assigned per-cluster basis majority voting, i.e., for each cluster, simply the majority is identified and all the cells in that cluster are assigned by the same major-type regardless of their respective percentage. This is the final major-type assignment, where all the cells in each cluster have the same major-type, either valid or invalid, i.e., "unassigned".

The last step is from the assumption that major cell-types are well clustered and clearly separable from other major-types. The 'unassigned' cell rejected previously can also be reassigned through this procedure.

Minor-type identification: Once major-type is assigned to all the clusters, minor-types are assigned separately to each major-type utilizing the taxonomy hierarchy. The assignment procedure is as follows.

(a) For each major-type, say the mth major-type, select the relevant minor-types and their markers as the union of the markers for the subsets belonging to the minor-type.

(b) Count expressed markers and compute GSA score, s_{ij} for $j \in C_m, j \in T_m$, in the same way as for major-type assignment, where C_m is the set of cells assigned to the mth major-type and T_m is the set of minor-types belonging to the mth major-type.

(c) Once the scores are obtained, the initial minor-type of the jth cell is given by $\operatorname{argmax}_i s_{ij}$ and its score is set to $\max_i s_{ij}$.

(d) *Application of k nearest neighbor (kNN) rule for minor-type correction*: Since minor-types are assumed to be roughly clustered, but not clearly separable to each other, we apply kNN rule to correct minor-type in PCA transformed space, i.e., for each cell in C_m, select k nearest neighbors and take majority among k to correct its minor-type. By doing so, each minor-type can be roughly divided within the clusters of the corresponding major-type. For minor type correction, we set k to 31 in all experiments, even though it can be specified by the user.

Subset identification and minor-type correction: Finally, subset assignment is performed similar to the minor-type assignment, with $k = 11$ for kNN rule for correction, where, instead of identifying subsets within a minor type, it is performed for each major-type, but rather than using only subset GSA scores, the initial subset identification is based on the sum of scores for the subset and the corresponding minor-type obtained previously. Then, kNN rule is applied for correction to obtain the final subset identify. Minor-type is finally corrected using its subset.

Running other marker-based methods

We compared HiCAT with 6 existing marker-based methods, Garnett[11], SCINA[12], scSorter[13], scType[14], scCatch[15], DigitalCellSorter[16,17]. For CellAssign, we couldn't install the package due to error. Although some of them provide their own list of markers, we used R&D systems markers

(https://www.rndsystems.com/resources/cell-markers) in all methods for fair comparison and their default parameter settings. The usage of each method was reproduced from their Github page, except for Garnett and scSorter. We referred https://cole-trapnell-lab.github.io/garnett/ for Garnett and https://cran.r-project.org/web/packages/scSorter/vignettes/scSorter.html for scSorter. All the existing marker-based methods were run 3 times, each for major-type, minor-type and subset identification, respectively, while HiCAT was run once for each dataset since it provides the 3 levels of cell-type, simultaneously.

Performance criteria

***Match test with manual annotation*:** To evaluate performance, we used five performance measures used in previous works[19,37], correct (C), error (E), erroneously assigned (EA), correctly unassigned (CUA), and erroneously unassigned (EUA). To define these criteria, we divide the cells based on their prediction results into 'unassigned' and others with valid predicted label (their cell-types exist in the marker DB). The unassigned is counted as CUA if its original label (i.e., the manual annotation) is not in the cell type list in the marker DB. Otherwise, they are counted as EUA. A cell with valid predicted label is counted as EA if its original label is not in the cell type list in the marker DB, or, counted as C if its original label exists in the cell type list and is equal to the predicted label. Otherwise, it is counted as E, i.e., if its original label exists in the cell type list but is not equal to the predicted label.

***Pairwise match test*:** In pairwise match test, instead of using the 5 criteria below, we simply compared the two results from two different identifiers, where we counted the fraction of cells for which the two results are equal, regardless of whether they are 'unassigned' or other valid cell-type, i.e., it is counted as a hit if the two results are both 'unassigned'.

AVAILABILITY OF SOFTWARE AND DATASETS

The python code and the example in Jupyter notebook were deposited to Github and can be obtained at https://github.com/combio-dku/HiCAT (Project name: HiCAT, license: GPL 3.0, Operating system(s): Platform independent, Programming language: python 3, other requirement: None). The datasets used in this work can be freely downloaded from the internet: two lung datasets from the human cell atlas, https://www.humancellatlas.org/, and others from the gene expression omnibus https://www.ncbi.nlm.nih.gov/geo/ (GEO) using their accession number.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no competing interest.

FUNDING

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2020R1F1A1066320 and NRF-2021R1A4A5032463).

AUTHORS' CONTRIBUTION

SY and JL devised the concept and key idea. SY developed the code of HiCAT. JL and MK performed experiments. SY guided experiments and data analysis. KK and CY gave interpretation and feedback on the results. All authors wrote, read, and approved the final manuscript.

REFERENCES

1. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377-382 (2009).
2. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10**, 1096-1098 (2013).

3. Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C. & Teichmann, S.A. The technology and biology of single-cell RNA sequencing. *Mol Cell* **58**, 610-620 (2015).
4. Chung, W. et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat Commun* **8**, 15081 (2017).
5. Smillie, C.S. et al. Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **178**, 714-730 e722 (2019).
6. Wu, S.Z. et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet* **53**, 1334-1347 (2021).
7. Ren, X. et al. COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* **184**, 1895-1913 e1819 (2021).
8. Abdelaal, T. et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol* **20**, 194 (2019).
9. Xie, B., Jiang, Q., Mora, A. & Li, X. Automatic cell type identification methods for single-cell RNA sequencing. *Comput Struct Biotechnol J* **19**, 5874-5887 (2021).
10. Balkwill, F.R., Capasso, M. & Hagemann, T. The tumor microenvironment at a glance. *J Cell Sci* **125**, 5591-5596 (2012).
11. Pliner, H.A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat Methods* **16**, 983-986 (2019).
12. Zhang, Z. et al. SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples. *Genes (Basel)* **10** (2019).
13. Guo, H. & Li, J. scSorter: assigning cells to known cell types according to marker genes. *Genome Biol* **22**, 69 (2021).
14. Ianevski, A., Giri, A.K. & Aittokallio, T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun* **13**, 1246 (2022).
15. Shao, X. et al. scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. *iScience* **23**, 100882 (2020).
16. Domanskyi, S., Hakansson, A., Bertus, T.J., Paternostro, G. & Piermarocchi, C. Digital Cell Sorter (DCS): a cell type identification, anomaly detection, and Hopfield landscapes toolkit for single-cell transcriptomics. *PeerJ* **9**, e10670 (2021).
17. Domanskyi, S. et al. Polled Digital Cell Sorter (p-DCS): Automatic identification of hematological cell types from single cell RNA-sequencing clusters. *BMC Bioinformatics* **20**, 369 (2019).
18. Zhang, A.W. et al. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nature Methods* **16**, 1007-+ (2019).
19. Kim, H., Lee, J., Kang, K. & Yoon, S. MarkerCount: A stable, count-based cell type identifier for single-cell RNA-seq experiments. *Comput Struct Biotechnol J* **20**, 3120-3132 (2022).
20. Michielsen, L., Reinders, M.J.T. & Mahfouz, A. Hierarchical progressive learning of cell identities in single-cell data. *Nat Commun* **12**, 2799 (2021).
21. Fu, R. et al. clustifyr: an R package for automated single-cell RNA sequencing cluster classification. *F1000Res* **9**, 223 (2020).
22. Brbic, M. et al. MARS: discovering novel cell types across heterogeneous single-cell experiments. *Nat Methods* **17**, 1200-1206 (2020).
23. Zhang, R., Luo, Y., Ma, J., Zhang, M. & Wang, S. scPretrain: Multi-task self-supervised learning for cell type classification. *Bioinformatics* (2022).
24. Duan, B. et al. Learning for single-cell assignment. *Sci Adv* **6** (2020).
25. Wang, L.F. et al. An interpretable deep-learning architecture of capsule networks for identifying cell-type gene expression programs from single-cell RNA-sequencing data. *Nat Mach Intell* **2**, 693-+ (2020).
26. Ma, F. & Pellegrini, M. ACTINN: automated identification of cell types in single cell RNA sequencing.

Bioinformatics **36**, 533-538 (2020).
27. Li, C.W. et al. SciBet as a portable and fast single cell type identifier. *Nature Communications* **11** (2020).
28. Boufea, K., Seth, S. & Batada, N.N. scID Uses Discriminant Analysis to Identify Transcriptionally Equivalent Cell Types across Single-Cell RNA-Seq Data with Batch Effect. *Iscience* **23** (2020).
29. Tan, Y.Q. & Cahan, P. SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. *Cell Systems* **9**, 207-+ (2019).
30. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods* **15**, 1053-+ (2018).
31. Hou, R., Denisenko, E. & Forrest, A.R.R. scMatch: a single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics* **35**, 4688-4695 (2019).
32. Nguyen, V. & Griss, J. scAnnotatR: framework to accurately classify cell types in single-cell RNA-sequencing data. *Bmc Bioinformatics* **23** (2022).
33. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* **20**, 163-172 (2019).
34. Alquicira-Hernandez, J., Sathe, A., Ji, H.P., Nguyen, Q. & Powell, J.E. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol* **20**, 264 (2019).
35. Kiselev, V.Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat Methods* **15**, 359-362 (2018).
36. Lieberman, Y., Rokach, L. & Shay, T. CaSTLe - Classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLoS One* **13**, e0205499 (2018).
37. de Kanter, J.K., Lijnzaad, P., Candelli, T., Margaritis, T. & Holstege, F.C.P. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res* **47**, e95 (2019).
38. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502 (2015).
39. Lun, A.T., McCarthy, D.J. & Marioni, J.C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* **5**, 2122 (2016).
40. Kiselev, V.Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* **14**, 483-486 (2017).
41. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018).
42. Miller, J.A. et al. Common cell type nomenclature for the mammalian brain. *Elife* **9** (2020).
43. Yuste, R. et al. A community-based transcriptomics classification and nomenclature of neocortical cell types (vol 23, pg 1456, 2020). *Nat Neurosci* **24**, 612-612 (2021).
44. Zheng, G.X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017).
45. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**, 865-868 (2017).
46. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196 (2016).
47. Madissoon, E. et al. scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome Biol* **21**, 1 (2019).
48. Reyfman, P.A. et al. Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. *Am J Respir Crit Care Med* **199**, 1517-1536 (2019).
49. Maleki, F., Ovens, K., Hogan, D.J. & Kusalik, A.J. Gene Set Analysis: Challenges, Opportunities, and Future Research. *Frontiers in Genetics* **11** (2020).

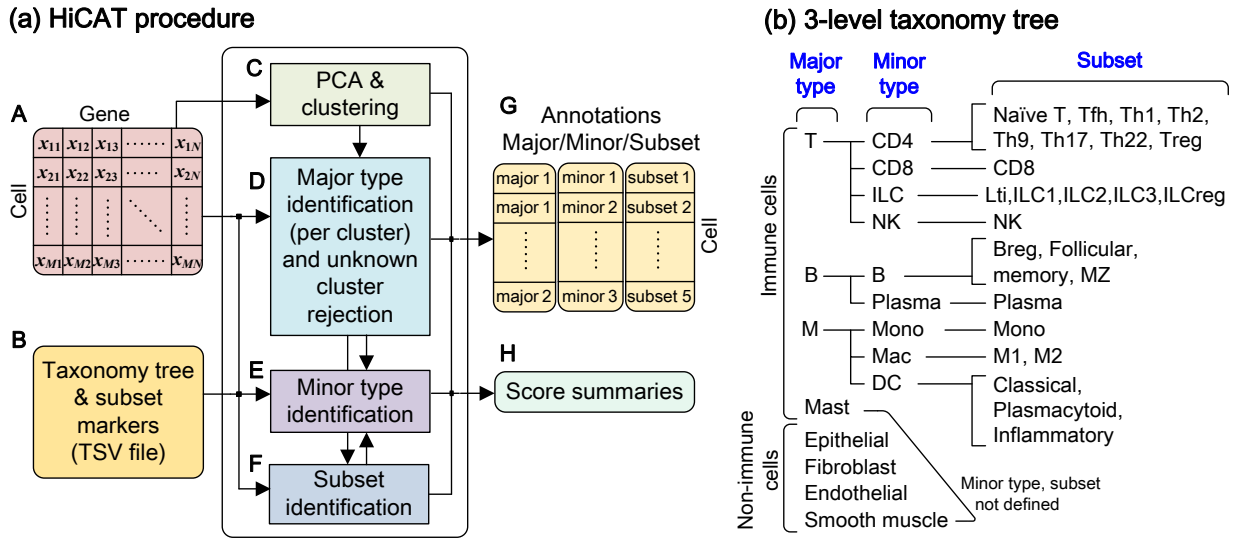


Figure 1. HiCAT processing flow (a) and taxonomy tree used for hierarchical identification (b). In (a), the proposed identifier takes two inputs; a cell-by-gene matrix (A) and the marker DB (B). The cell-by-gene matrix is first PCA transformed and cells are clustered in dimension-reduced space (C). In major-type identification step (D), unclear clusters are rejected by using the best and 2nd best major-type scores. Minor-type (E) and subset identification (F) follow to finally outputs 3-level annotation results (G) and score summaries (H). Note that markers are specified only for subsets and, for major or minor-type identification, the union of makers of its subsets was used. In the taxonomy tree (b), as our focus is on immune cell subsets, other non-immune cells were not further divided into minor-types or subsets. T: T cells, B: B cells, M: Myeloid cells, MZ: marginal zone B cell, Treg: regulatory T cells, Breg: regulatory B cells.

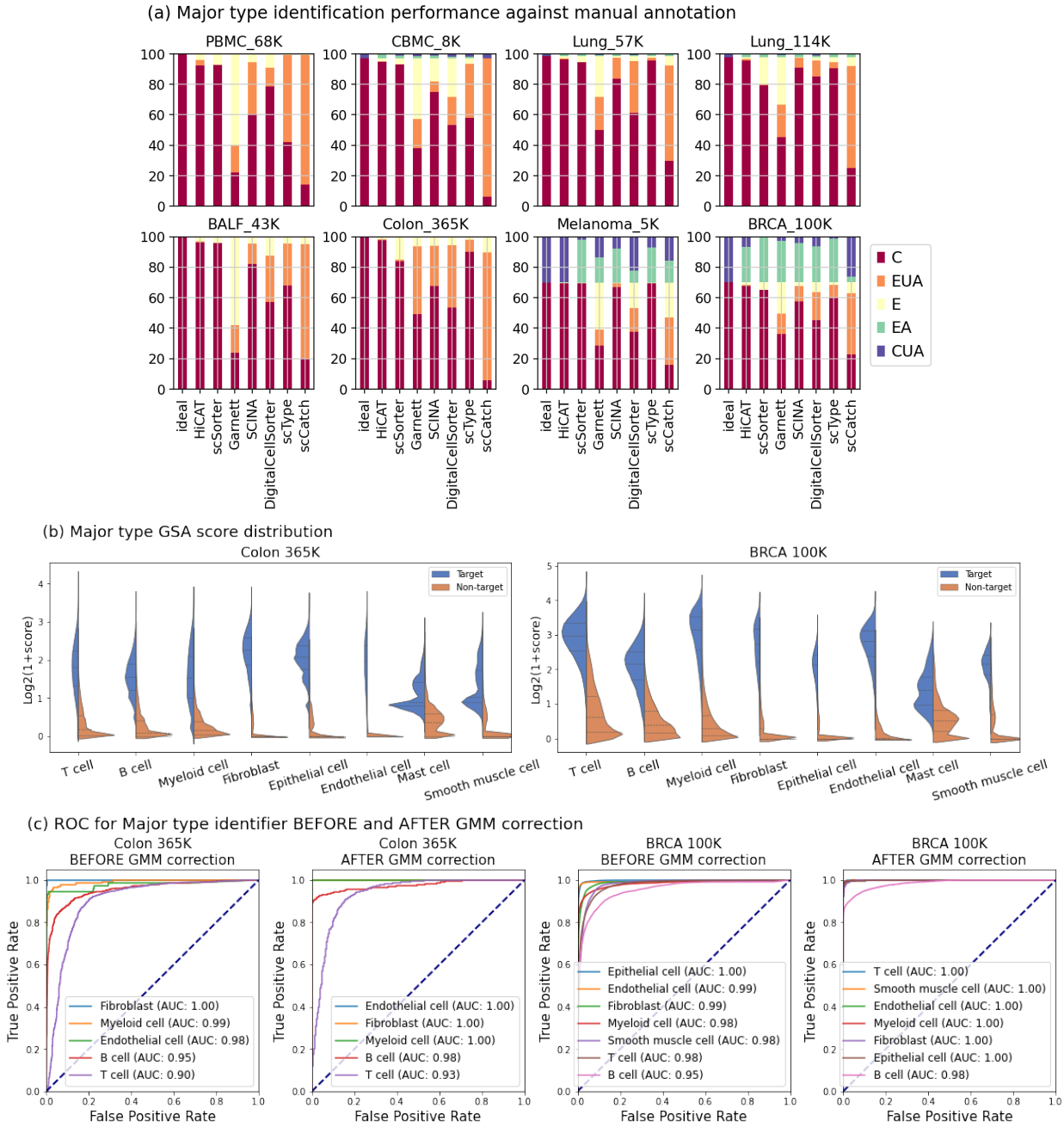


Figure 2 Comparisons of major-type identification performances (a), major-type GSA scores distribution of target cells versus non-target for Colon 365K and BRCA 100K (b), and ROC curves before and after GMM-based correction for the two datasets (c). C: correct, EUA: erroneously unassigned, E: error, EA: erroneously assigned, CUA: correctly unassigned. In (a), the first bar marked as 'ideal' shows the ideal performance, where CUA account for tumor cells or those cell-types not in the marker DB entry, which should ideally be 'unassigned'. Excluding CUA portion in Ideal performance, HiCAT showed 95% or higher accuracy in most datasets, except for PBMC 68K. In (b), GSA score distribution for target is the distribution of 'target' cell-type scores for the cells identified as 'target', i.e., for those cells of which the cell-type with the highest score was the 'target', while that for non-target is the score distribution of the 'target' for the cells of which the 2nd best cell type was the 'target', i.e., the best one was something else. The violin plots in (b) show that major-types are largely separable from others using only the GSA scores, even though there are small overlaps between the scores of targets and non-targets. Rejection thresholds for each major-type are set using the two distributions such that false positive rate is less than or equal to a pre-defined value, say 0.05. In (c), comparing ROCs before and after GMM based correction clearly shows the improvement in the identification performance.

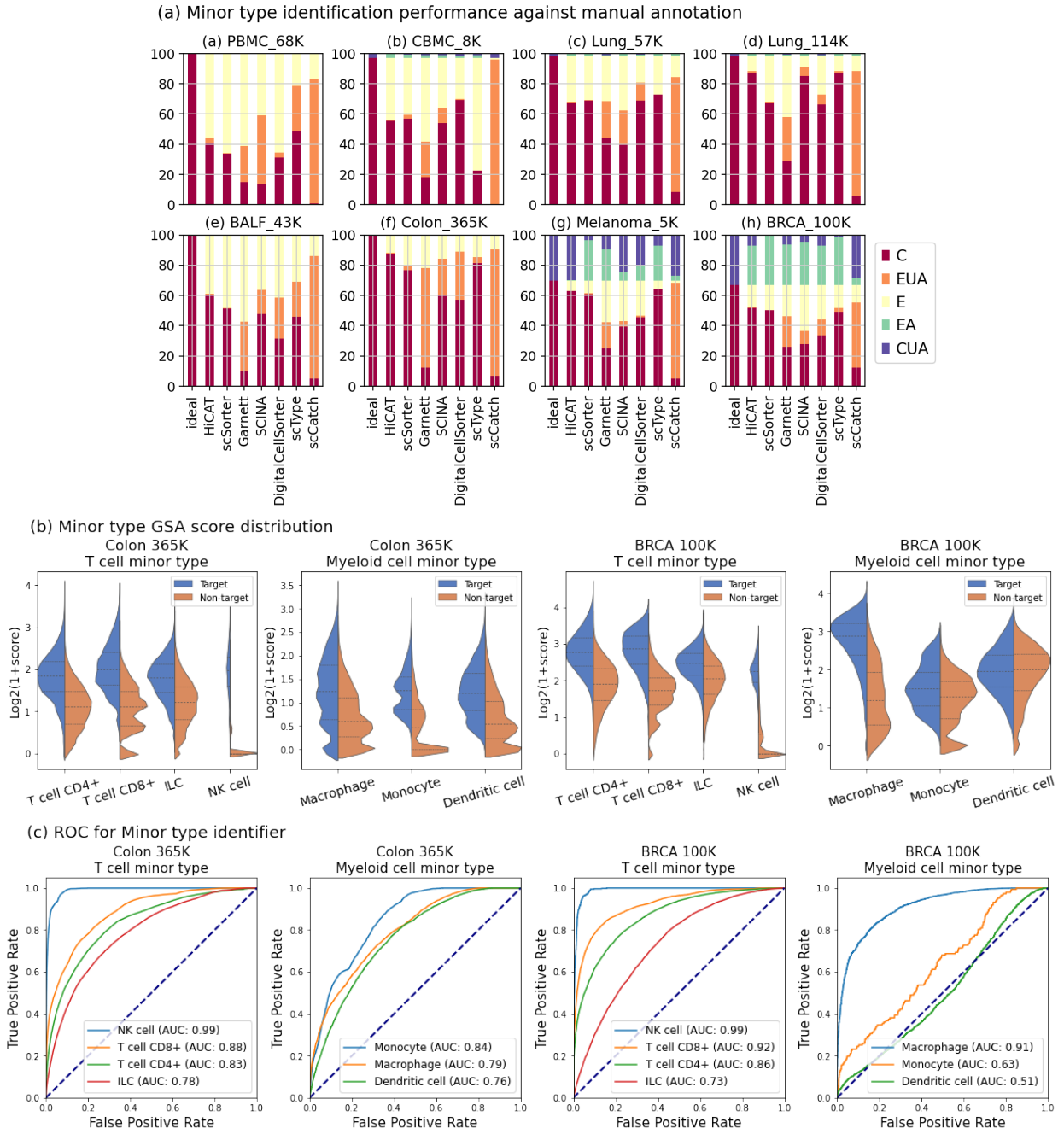


Figure 3 Comparisons of minor-type identification performances (a), T cell and myeloid cell minor-type GSA scores distribution for Colon 365K and BRCA 100K (b), and ROC curves for T cell and myeloid cell minor-type identification for the two datasets (c). C: correct, EUA: erroneously unassigned, E: error, EA: erroneously assigned, CUA: correctly unassigned. In (a), the first bar marked as 'ideal' shows the ideal performance, where CUA account for tumor cells or those cell-types not in the marker DB entry, which should ideally be 'unassigned'. Identification performances for minor-type are much less than those for major-types in all identifiers raising question whether manual annotations are reliable enough to be used as ground truth for performance evaluation. In (b), comparing with those for major-types, there exist considerable overlap between the two distributions for target and non-target, respectively, indicating that they may not be well separable using only the GSA scores. In minor-type identification, instead of cluster basis reassignment, we used *k*-nearest neighbor (*k*NN) rules to correct initial decisions. The number *k* can be pre-specified by the user and we set it 31 in all experiments in this work. In (c), the AUCs are shown to be much smaller than those for major-types.

Table 1. A summary of pairwise prediction match (a) for minor-type identification showing the best match, 2nd best match and the best match with manual annotations and (b) for subset identification showing the best, 2nd best and 3rd best match. Tumor cells are considered to be 'unknown' and counted as match if they were 'unassigned' by identifier. In (a), the average best match rate (0.819) is higher than the best match rate with manual annotation (0.713), raising question whether manual annotation is reliable enough for performance evaluation against it. Of note, in both (a) and (b), HiCAT was one of the best match pair.

(a)	Dataset	Best match pair	Best match rate	2nd Best match pair	2nd Best match rate	Best tool with manual annot.	Best match rate with manual annot.
Minor-type	PBMC_68K	HiCAT - scSorter	0.715	scSorter - DCS	0.642	scType	0.488
	CBMC_8K	HiCAT - scSorter	0.806	HiCAT - SCINA	0.754	DCS	0.704
	Lung_57K	HiCAT - scType	0.788	Manual - scType	0.72	scType	0.72
	Lung_114K	HiCAT - SCINA	0.878	HiCAT - scType	0.868	HiCAT	0.819
	BALF_43K	HiCAT - scSorter	0.784	HiCAT - SCINA	0.69	HiCAT	0.603
	Colon_365K	HiCAT - Manual	0.873	Manual - scType	0.812	HiCAT	0.873
	Melanoma_5K	HiCAT - Manual	0.93	Manual - scType	0.713	HiCAT	0.93
	BRCA_100K	HiCAT - scType	0.774	HiCAT - scSorter	0.713	HiCAT	0.51
		Average rate	0.819	Average rate	0.739	Average rate	0.713

(b)	Dataset	Best match pair	Best match rate	2nd Best match pair	2nd Best match rate	3rd Best match pair	3rd Best match rate
Subset	PBMC_68K	HiCAT - scType	0.544	HiCAT - scSorter	0.53	scSorter - scType	0.44
	CBMC_8K	HiCAT - scType	0.787	scSorter - scType	0.603	HiCAT - scSorter	0.601
	Lung_57K	HiCAT - scType	0.555	Garnett - scCatch	0.404	scSorter - SCINA	0.382
	Lung_114K	HiCAT - scType	0.68	scSorter - scType	0.648	HiCAT - scSorter	0.621
	BALF_43K	HiCAT - scType	0.751	HiCAT - DCS	0.209	DCS - scType	0.204
	Colon_365K	HiCAT - scType	0.763	HiCAT - SCINA	0.667	SCINA - scType	0.612
	Melanoma_5K	HiCAT - scType	0.578	scSorter - SCINA	0.54	scSorter - scType	0.525
	BRCA_100K	HiCAT - scType	0.714	HiCAT - scSorter	0.7	scSorter - scType	0.676
		Average rate	0.672	Average rate	0.538	Average rate	0.508

* DCS: DigitalCellSorter

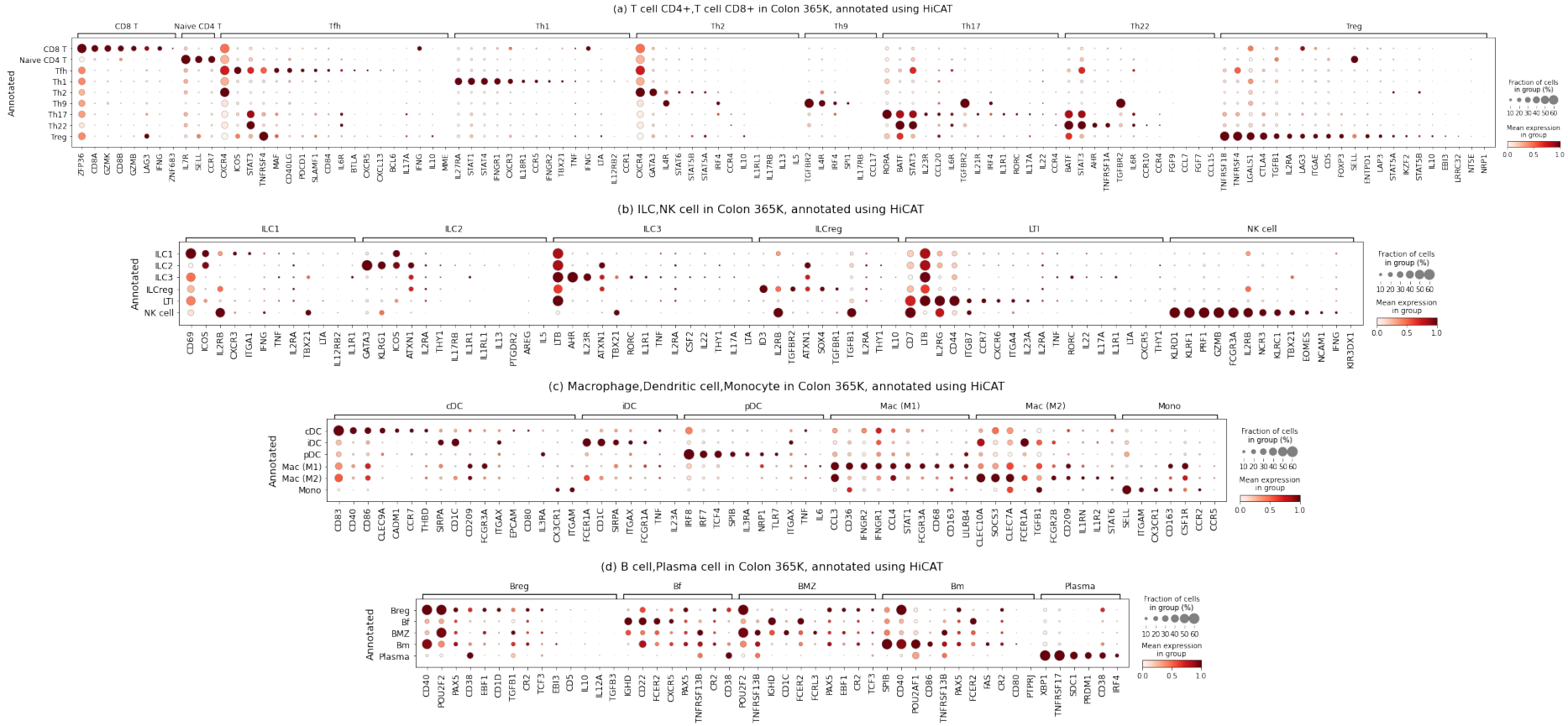


Figure 4. Dot plots showing marker expression profile under HiCAT annotation in Colon 365K. (a) CD4 and CD8 T cell subsets, (b) ILC and NK cell subsets, (c) myeloid cell subsets, and (d) B cell subsets. The dot size is the fraction of cells that expresses the marker gene in the identified subsets, regardless of their expression level, while the color corresponds to the average expression level of the marker scaled between 0 and 1, where 0 and 1 represent respectively the minimum and the maximum expression level of that marker among all subsets. The fraction was clipped at 0.6. For better visualization, we grouped the marker genes for each subset and ordered according to the fraction of marker expressing cells in the designated subset. We removed some common marker that frequently occurs (in 3 or more subsets), e.g., CD3 and CD4 in T cell subset, CD14 in myeloid cell subsets, PTPRC, IL7R and KLRB1 in ILC subsets, IGHM, CD19, CD27, MS4A1 in B cell subsets.

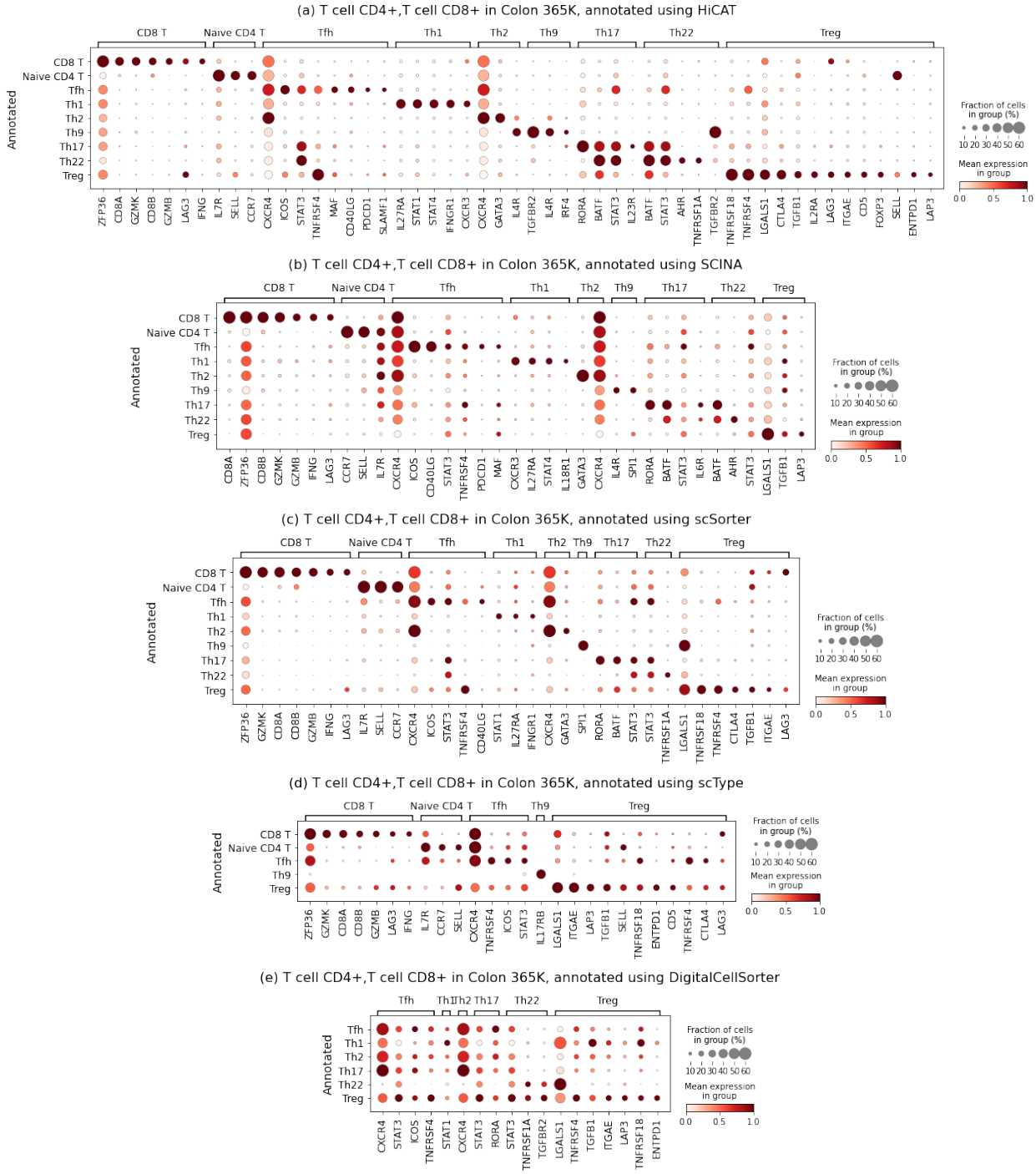


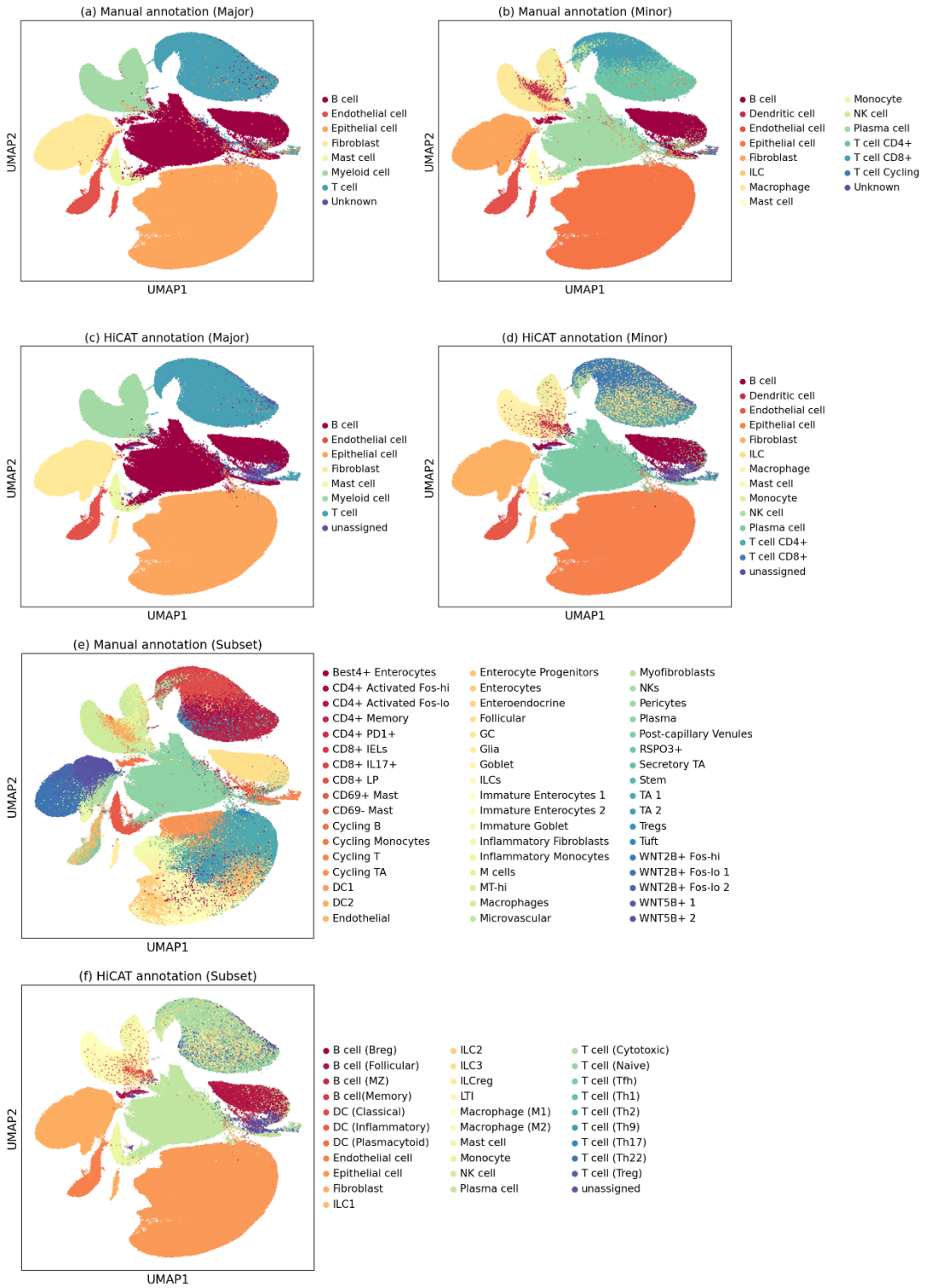
Figure 5. Dot plots showing marker expression of T cell subsets in Colon 365K identified by different tools. (a) HiCAT, (b) SCINA, (c) scSorter, (d) scType, and (e) DigitalCellSorter. The dot size is the fraction of cells in an identified subset that expresses the marker, regardless of their expression level, while the color corresponds to the expression level scaled between 0 and 1, where 0 and 1 represent respectively the minimum and the maximum expression level of that marker among all subsets. The fraction was clipped at 0.6, i.e., fractions greater than 0.6 were set to 0.6, for better visualization. For better visualization, we grouped the marker genes for each subset and ordered according to the fraction of marker expressing cells in the target subsets. Each row are subsets identified by each method. HiCAT, SCINA and scSorter identified all 9 T cell subsets, while scType and DigitalCellSorter identified only 5 and 6 subsets, respectively.

Supplemental Figures and Tables

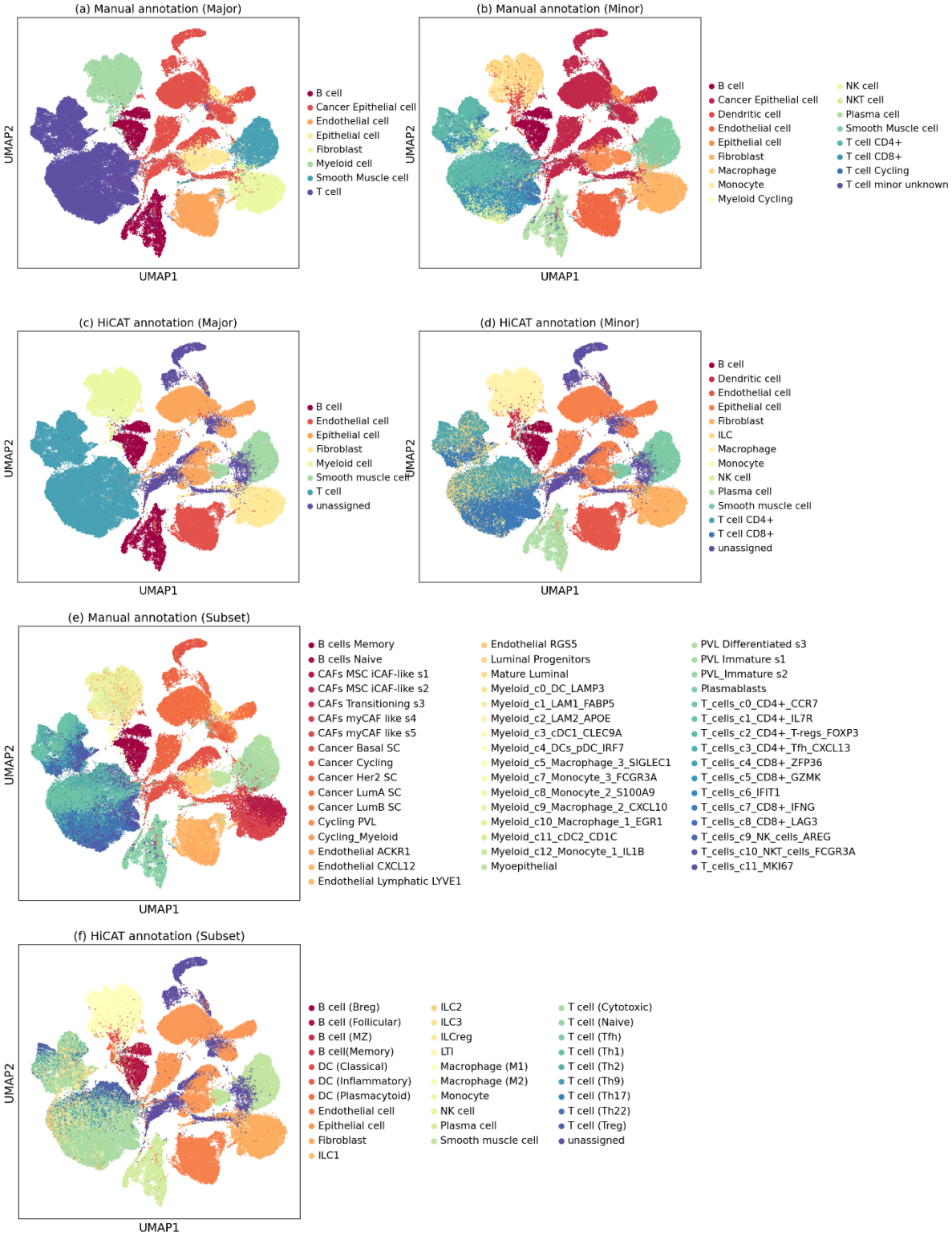
Supplemental Table 1. A summary of datasets used for the analysis and performance comparisons.

Dataset	Accession	Protocol	Num. of genes	Num. of cells total	Num. of Tumor cells	Num of major /minor-types
Colon 365K	SCP259	10X genomics	18,151	365,492	None	8/14
BRCA 100K	GSE176078	10X genomics	29,733	100,064	24,489	8/17
PBMC 68K	GSE93421	10X genomics	32,643	68,579	None	4/7
CBMC 8K	GSE100866	Drop-seq	20,501	8617	None	8/12
Lung 57K	ERP114453	10X v2	25,204	57,202	None	8/15
Lung 114K	SRP218543	10X v2	33,694	114,396	None	9/14
BALF 43K*	GSE158055	10X genomics	27,943	42,723	None	6/11
Melanoma 5K	GSE72056	Smart-seq2	23,684	4,513	1,251	6/8

* Selected only bronchoalveolar lavage fluid (BALF) samples



Supplemental Figure 1. UMAP plot and comparison of manual annotation and HiCAT results for Colon 365K datasets. (a), (b), and (e) are manual annotation of major-type, minor-type and subset (non-canonical), respectively. (c), (d), and (f) are HiCAT annotation of major-type, minor-type and subsets (canonical), respectively. Comparing (a) and (b), major-types are clearly separable from others with sufficient separation between them, while minor-types, for example T cell and myeloid cell minor types, are not clearly separable even though they are mostly well localized within a major-type cluster. Nevertheless, it is questionable if manual annotation of minor types is reliable enough to evaluate identifier performance against it.



Supplemental Figure 2. UMAP plot and comparison of manual annotation and HiCAT results for BRCA 100K datasets. (a), (b), and (e) are manual annotation of major-type, minor-type and subset (non-canonical), respectively. (c), (d), and (f) are HiCAT annotation of major-type, minor-type and subsets (canonical), respectively. Major-types in (a) are clearly separable from others, while minor-types in (b), for example T cell and myeloid cell minor types, are not clearly separable even though they are mostly well localized within a major-type cluster.