Method details

Detailed description of HiCAT processing

Preprocessing: In the preprocessing step, we perform count normalization, log transformation and principal component analysis (PCA) by choosing highly variable genes, similar to those with Seurat[1] or SCANPY[2]. Clustering is then performed to be used for the cluster-basis major-type identification and unknown cluster detection. HiCAT support three clustering algorithms, including *k*-means clustering, 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 4 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 count and gene set analysis (GSA) scoring: For the first step in the cell-type identification, we need cell-type markers, where, for the major-types, they are given by the union of subset markers as mentioned before. Given the set of major-type markers, M_i, of size $|M_i|$ for the ith cell-type and the gene expression matrix of size $N_c \times N_g$, with N_c being the number of cells and N_g the number of genes, the marker counts is given by $n_{ij} = \left|M_i \cap G_j\right|$ 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

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

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[3]. Once the scores are obtained, the initial major-type of the jth cell is given by

$$i_j^* = \mathrm{argmax}_i\, s_{ij}$$

and its score is set to $\max_i s_{ij}$. But this process is performed in conjunction 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.,

$$FPR(t) = 1 - \frac{\left|C_i^{(2)}(t)\right|}{\left|C_i^{(1)}(t)\right|} < FPR_{\mathrm{threshold}} \tag{3}$$

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 cluster. 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 this end, we first made the following assumption on the average scores of known and unknown cell-type clusters, i.e., the average scores of known and unknown cell-type clusters are somewhat varied, while the distribution of their average scores of the two are largely separable from each other such that the average scores for unknown cell-type cluster is far below from the mean of average scores of known cell-type cluster, e.g., below 2 standard deviation from the mean of known cell-type clusters. Since we do not know which are known and unknown, we first sort average scores, where the scores of known clusters are assumed to be ranked in higher portion. Using only high ranked scores, we obtain a fitted line, by which we obtain the threshold. To be specific, we took the following steps to identify unknown cell-type clusters.

(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 $q_1, q_2, \dots, q_{N_{clusters}}$, 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 $t = (1-\alpha) \cdot \max_j |a^*j + b^* - q_j|$.

Supplemental Fig. 4 shows two examples of unknown cluster detection, one for Melanoma 5K and the other for BRCA 100K. The two datasets contain lots of tumor cells forming multiple clusters and these must be detected as unknown since we do not have markers of tumor cells in the database. Once the average scores are sorted, one can find the threshold to divide clusters into known or unknown, using the steps described above. Although the threshold is not always clearly determined, it looks quite clear especially in Melanoma 5K dataset and it can give us useful information which clusters are likely to be unknown.

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 those with valid cell-type assignment, it might not likely be unknown, but it was classified as unknown because the average score was just below the threshold due, for example, to some tissue/sample specific perturbation. 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.,

$$f_i(\boldsymbol{x}) = \sum_{k=1}^{M_i} \pi_{i,k} N(\boldsymbol{x}; \boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}) \quad (4)$$

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) gene expression vector. Given the major-type assignment performed using GSA score, the model parameters can be obtained by using the EM algorithm, for which various software tools are available.

(b) With the GMM parameters for all the major-type, we compute the log-likelihood, $l_{ij} = \lg f_i(\boldsymbol{x}_j)$, 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 $\mathrm{argmax}_i\, l_{ij}$ 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.

(d) *Correction using aggregated network topology*: When using graph-based clustering, one can identify aggregated network of clusters using the interconnection strength between clusters. Based on this interconnection network of clusters, one can further identify tightly connected group of clusters to correct major-type identified in step (c). For example, let's say 4 nearby clusters were tightly interconnected, where 3 of them were T cells, while one was unassigned (or B cell with relatively low average score). Then, the unassigned cluster is highly likely to be T cell and we can correct it to the majority in the tightly connected group. 'Unassigned' cell cluster rejected previously can be reassigned through this procedure.

Note that the last two steps are from the assumption that major cell-types are well clustered and clearly separable from other major-types and, therefore, not applicable to minor-type identification.

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 subsets and their markers.

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

(c) Minor type score is the maximum subset score among those of subsets belonging to that minor type, i.e., $s'_{kj} = \max_{i \in S'_k} s_{ij}$ where S'_k is the set of subsets belonging to the kth minor type.

(d) Once the minor type scores are obtained, the initial minor-type of the jth cell is given by $\mathrm{argmax}_k\, s'_{kj}$ and its score is set to $\max_k s'_{kj}$.

(e) *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 *k*NN 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. Although *k*NN-rule based correction is supported for subsets as well, we set $k = 1$ for kNN rule (no correction) since there is no evidence reported so far if subsets are well localized within minor-type cluster.

Running other marker-based methods

We compared HiCAT with 6 existing marker-based methods, Garnett[4], SCINA[5], scSorter[6], scType[7], scCatch[8], DigitalCellSorter[9, 10]. 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-types simultaneously. Since some of identifiers could not be run for big-sized datasets, such as BRCA 100K, Lung 114K, and Colon 365K, we ran all the identifiers separately for pair of samples. For fair comparison, we used the default hyper parameter setting for HiCAT too, the same for all the datasets. It is important since an identifier can perform better for a dataset with a certain setting of hyper parameter, while the setting gives us a worse result for other datasets.

Construction of marker database

We adopted most subset markers from R&D systems (https://www.rndsystems.com/resources/cell-markers), where three types of markers are defined, i.e., cell surface markers, intracellular markers and secreted factors, where, for the first two, the markers are specifically marked as positive(+), negative(−), 'high' or 'low'. Since the 'high' and 'low' cannot be clearly distinguished by absolute gene expression level, we simply divided only into positive and negative by classifying both 'high' and 'low' as positive. Also, since R&D systems does not provide markers for some important subsets, such as naïve CD4 T cell and cytotoxic T cell, we added their markers by literature search. The list of cell-type markers we used in this work can be found in Supplemental Table 1 in Supplemental File 2. Although we used markers from R&D systems, users of HiCAT may easily edit the marker database file for their own study objective.

Performance criteria

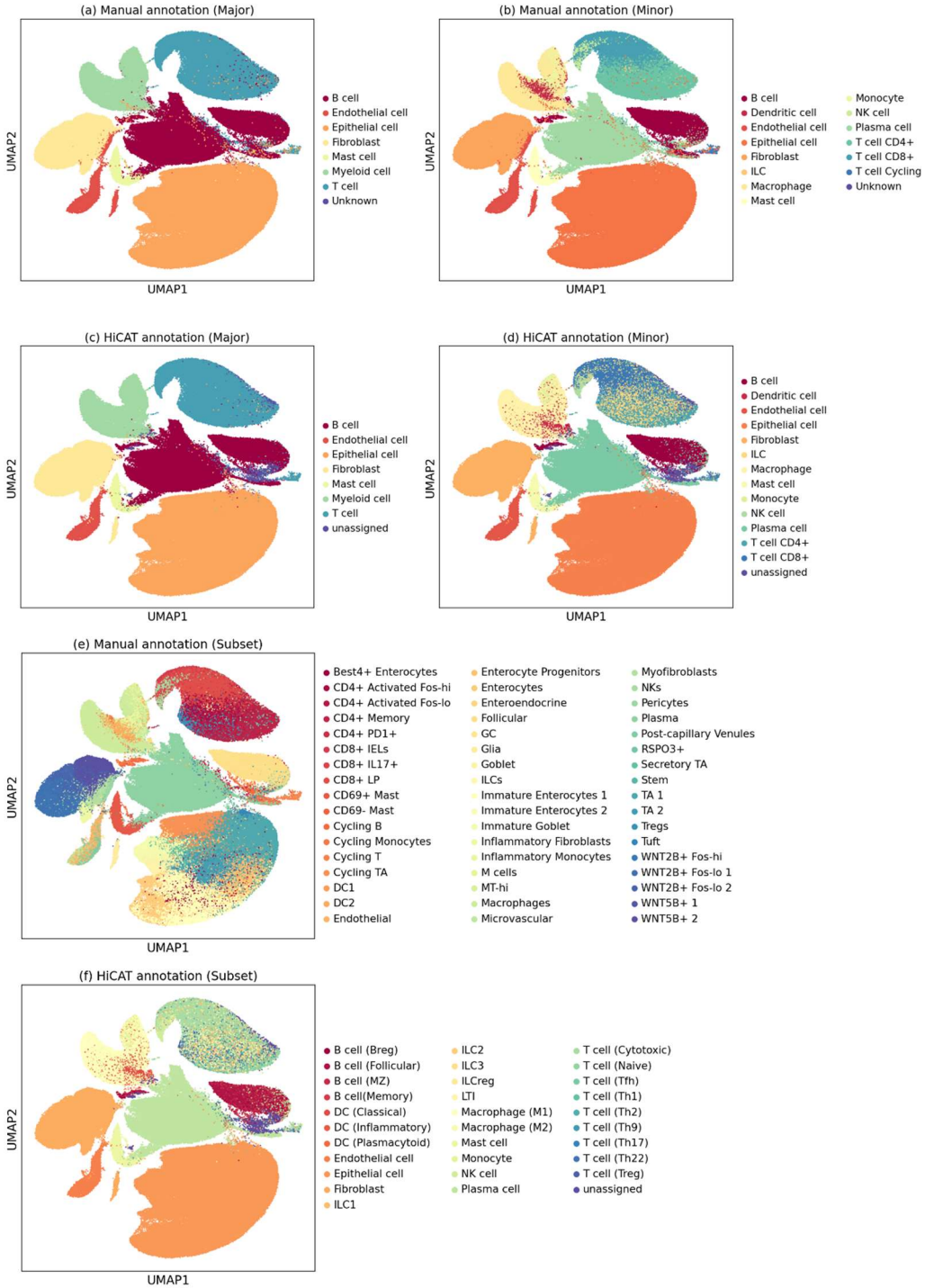
***Match test with manual annotation*:** To evaluate performance for major- and minor-type, we used five performance measures used in previous works[11, 12], i.e., correct (C), error (E), erroneously assigned (EA), correctly unassigned (CUA), and erroneously unassigned (EUA). To define these criteria, we divide the cells based on their prediction results into 'unassigned' and others with valid predicted label (the cell-types existing in the marker DB). The unassigned is counted as CUA if its original label (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. The graphical explanation was depicted as reference in Figure 2 and 3.

***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 valid cell-types, i.e., it is counted as a hit even if the two results are both 'unassigned'.

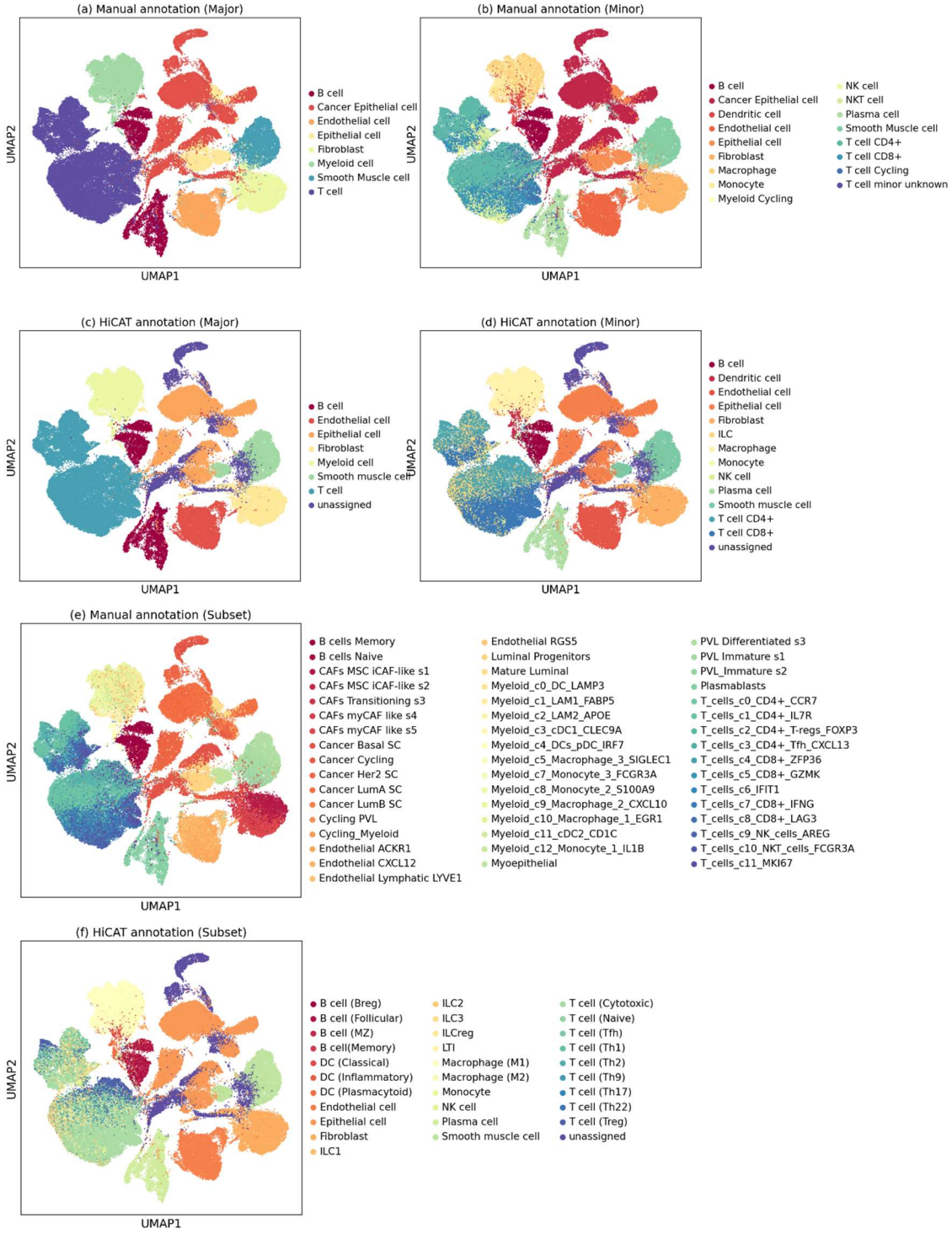
DEG and GSE analysis

Differentially expressed gene (DEG) and gene set enrichment (GSE) analysis were performed for immune cell profiling in ulcerative colitis. Seurat package was used for DEG and FGSEA package for GSE analysis. The analyses were performed separately for the two datasets, Colon 365K and Colon 32K. First, DEG analysis was performed in inflamed (test) versus healthy (control) condition setting, separately for all myeloid cell subsets and T cell subsets. Then, DEGs were obtained with adjusted p-value cutoff 0.05. GSE analysis was then performed using DEGs. Additionally, DEGs for M2a/M2b/M2c versus M1 was obtained to run GSE analysis to check possible M1-skewed polarization of the three M2 macrophage subtypes.

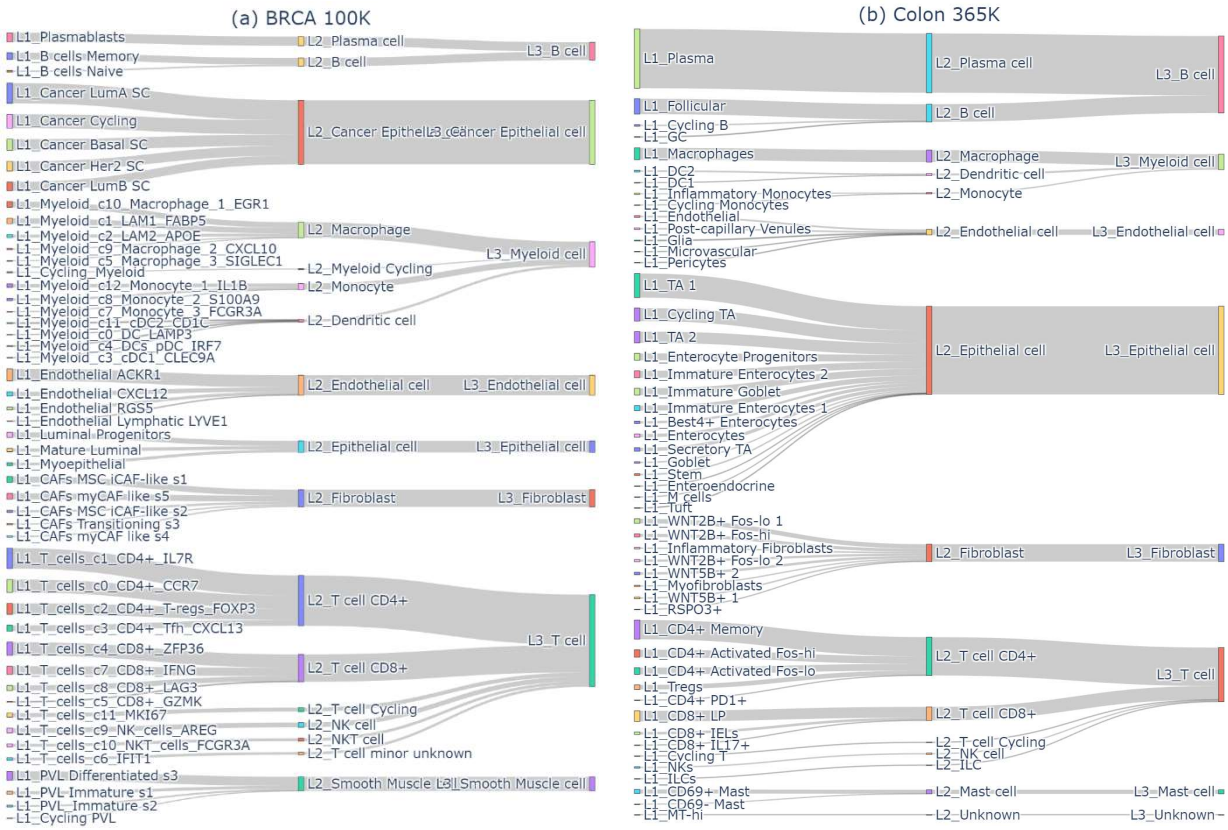
Supplemental Figures



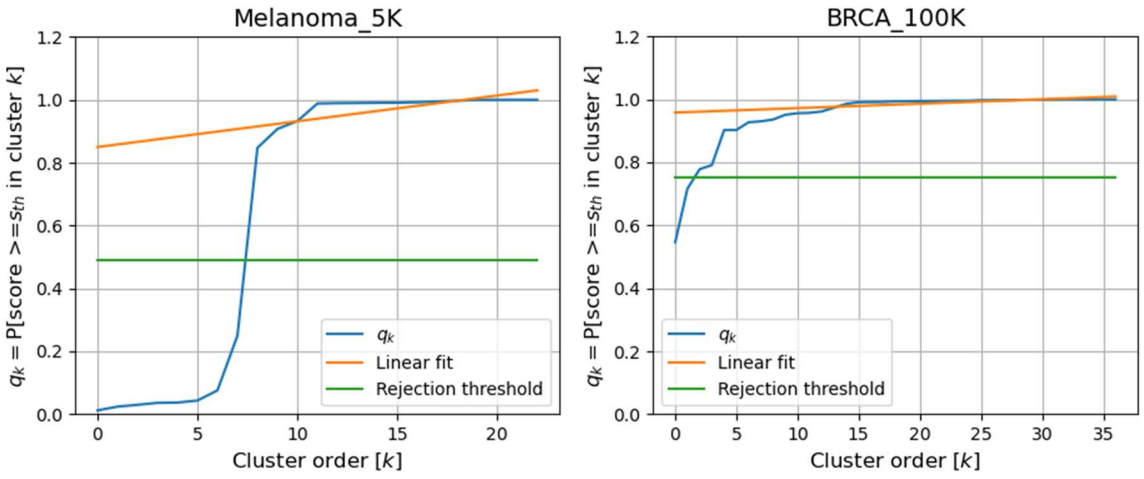
Supplemental Fig. 1. UMAP plot and comparison of manual annotation and HiCAT results for Colon 365K datasets. (a), (b), and (e) are manual annotation of major-type, minor-type and subset (non-canonical), respectively. (c), (d), and (f) are HiCAT annotation of major-type, minor-type and subsets (canonical), 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.



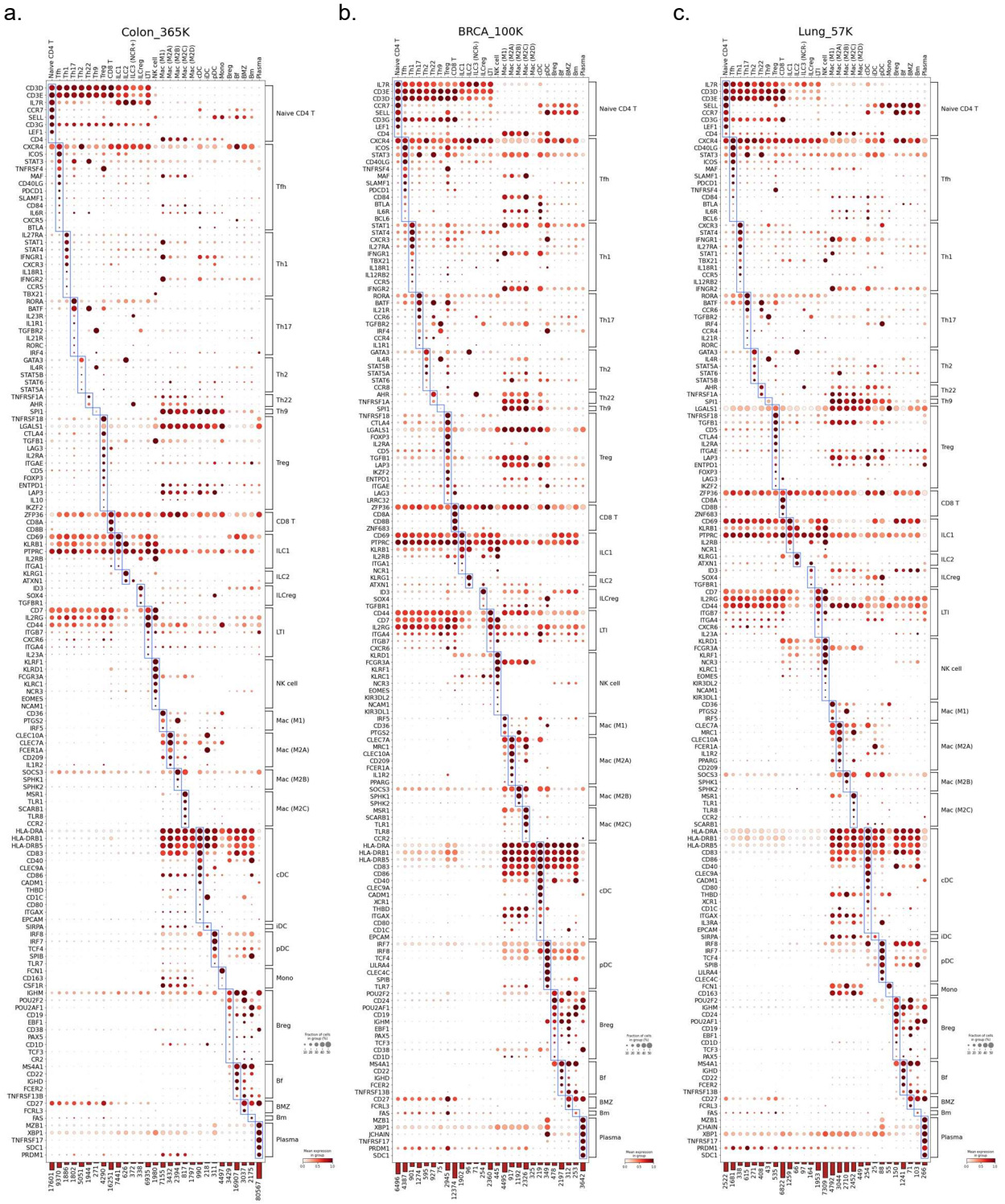
Supplemental Fig. 2. UMAP plot and comparison of manual annotation and HiCAT results for BRCA 100K datasets. (a), (b), and (e) are manual annotation of major-type, minor-type and subset (non-canonical), respectively. (c), (d), and (f) are HiCAT annotation of major-type, minor-type and subsets (canonical), respectively. Like Supplemental Figure 10, 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.



Supplemental Fig. 3. Examples of setting minor-type and major-type given the specific type annotation contained in the datasets for BRCA 100K (a) and Colon 365K (b). Left: manual annotations provided along with the dataset, Center: reannotated minor-types, Right: reannotated major-types. We used these minor-types and major-types to compared with the decisions of HiCAT and other existing identifiers, excluding 'Unknown's and cycling cells.

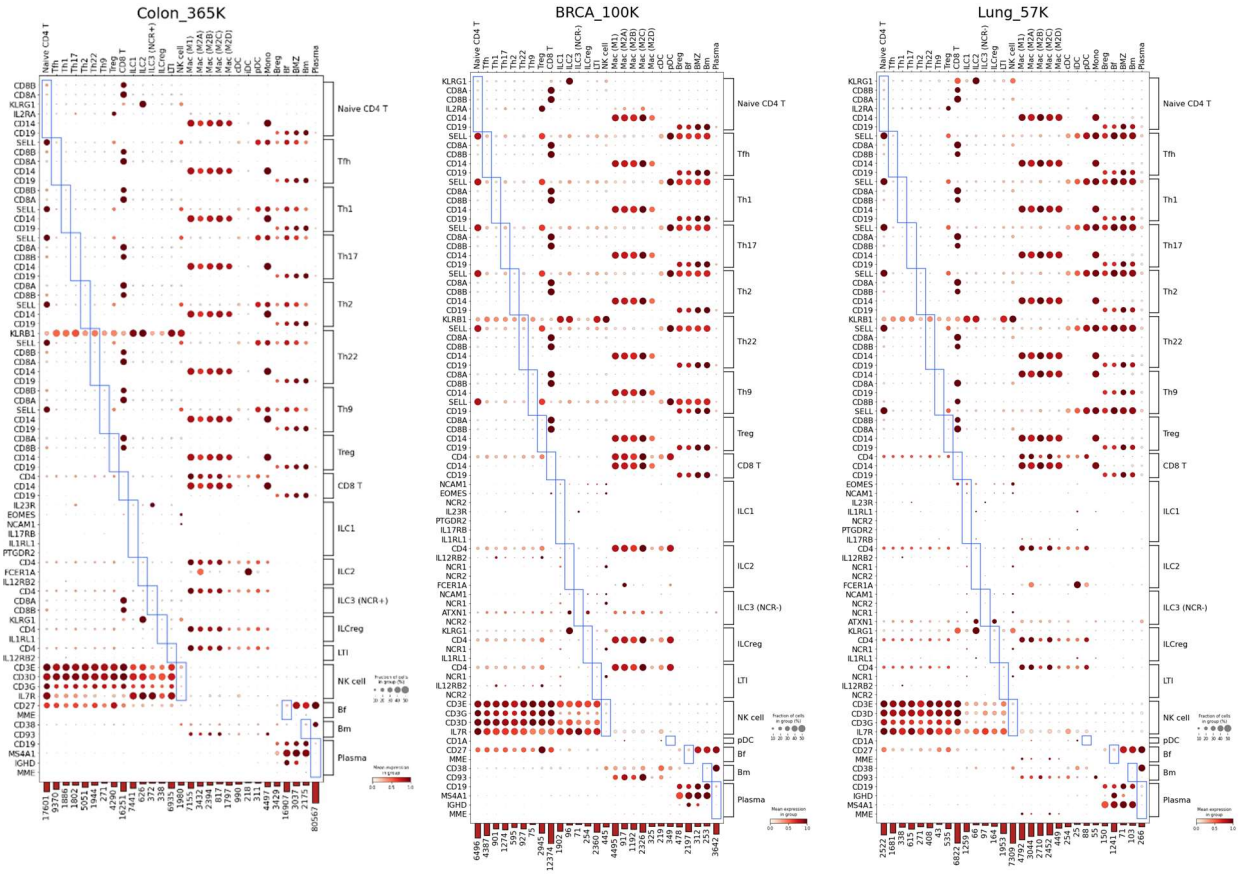


Supplemental Fig. 4. Statistics for unknown cluster rejection for (a) Melanoma 5K and (b) BRCA 100K. The cluster index is ordered according to the percentage of 'usable' cells, q_k, defined by $Pr\{\text{GSA score} \geq \text{threshold}\}$ for cluster k. Linear fitting was applied with those q_k's for upper 60% (reconfigurable) and the rejection threshold is obtained from the maximum difference between the fitted line and those q_k's for lower 40% of clusters.

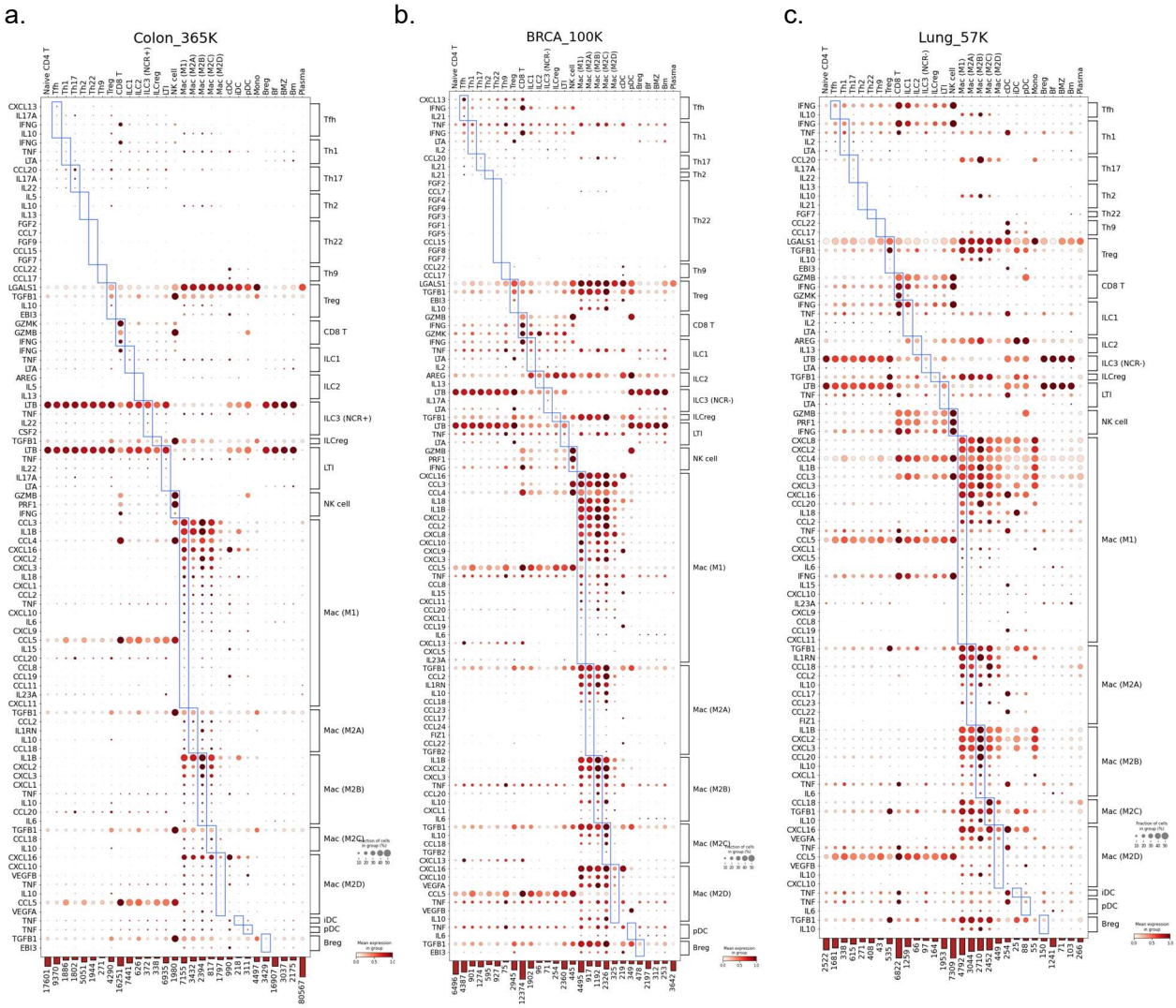


Supplemental Fig. 5. Expression patterns of POSITIVE markers, including cell surface and intracellular markers for 3 datasets. For better clarity, (1) we showed common markers only once in its first occurrence among all cell-type subsets, e.g., the common markers of all T cells, *CD3D*/*E*/*G* were shown only in gene group of Naïve CD4 T cell and (2) showed only markers for which the percentage of marker expressing cell in a subset is higher than or equal to 5%. As shown in the figure, *CD3D*/*E*/*G* were highly expressed in CD4+ and CD8+ T cells, while it is also expressed in ILCs, except for NK cell. Note that CD4+ T cells express *CD4* marker in all CD4+ T cell subsets. But the percentage were shown to be very small. It might be due to the difference of transcript expression and protein expression as discussed in Discussion section.

a. b. c.



Supplemental Fig. 6. Expression patterns of NEGATIVE markers for 3 datasets. The figure shows that, compared to the expressions of positive markers, the negative markers were mostly not or very rarely expressed in all immune cell subsets. Although rare or low expression of negative markers can be due to misclassification of cell-type subset, they are mostly very small, except for *KLRB1* (*CD161*) in Th22. In these results for 3 datasets, *KLRB1* was expressed in non-negligible portion of Th22 even though the expression level is quite low. Although it can also be due to misclassification of Th22, it might not be likely because its positive markers, *AHR* and *TNFRSF1A*, were clearly expressed as shown in Supplemental Figure 4.



Supplemental Fig. 7. Expression patterns of SECRETORY markers for 3 datasets. Compared to the expression of positive markers, many secretory factors were not shown to be uniquely expressed only in the designated immune cell subsets. And this is why we did not use them for identification of cell-type subset.

References

1. Satija, R., et al., *Spatial reconstruction of single-cell gene expression data.* Nat Biotechnol, 2015. **33**(5): p. 495-502.
2. Wolf, F.A., P. Angerer, and F.J. Theis, *SCANPY: large-scale single-cell gene expression data analysis.* Genome Biol, 2018. **19**(1): p. 15.
3. Maleki, F., et al., *Gene Set Analysis: Challenges, Opportunities, and Future Research.* Frontiers in Genetics, 2020. **11**.
4. Pliner, H.A., J. Shendure, and C. Trapnell, *Supervised classification enables rapid annotation of cell atlases.* Nat Methods, 2019. **16**(10): p. 983-986.
5. Zhang, Z., et al., *SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples.* Genes (Basel), 2019. **10**(7).
6. Guo, H. and J. Li, *scSorter: assigning cells to known cell types according to marker genes.* Genome Biol, 2021. **22**(1): p. 69.
7. Ianevski, A., A.K. Giri, and T. Aittokallio, *Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data.* Nat Commun, 2022. **13**(1): p. 1246.
8. Shao, X., et al., *scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data.* iScience, 2020. **23**(3): p. 100882.
9. Domanskyi, S., et al., *Digital Cell Sorter (DCS): a cell type identification, anomaly detection, and Hopfield landscapes toolkit for single-cell transcriptomics.* PeerJ, 2021. **9**: p. e10670.
10. Domanskyi, S., et al., *Polled Digital Cell Sorter (p-DCS): Automatic identification of hematological cell types from single cell RNA-sequencing clusters.* BMC Bioinformatics, 2019. **20**(1): p. 369.
11. Kim, H., et al., *MarkerCount: A stable, count-based cell type identifier for single-cell RNA-seq experiments.* Comput Struct Biotechnol J, 2022. **20**: p. 3120-3132.
12. de Kanter, J.K., et al., *CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing.* Nucleic Acids Res, 2019. **47**(16): p. e95.