iScience, Volume *23*

Supplemental Information

scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data

Xin Shao, Jie Liao, Xiaoyan Lu, Rui Xue, Ni Ai, and Xiaohui Fan

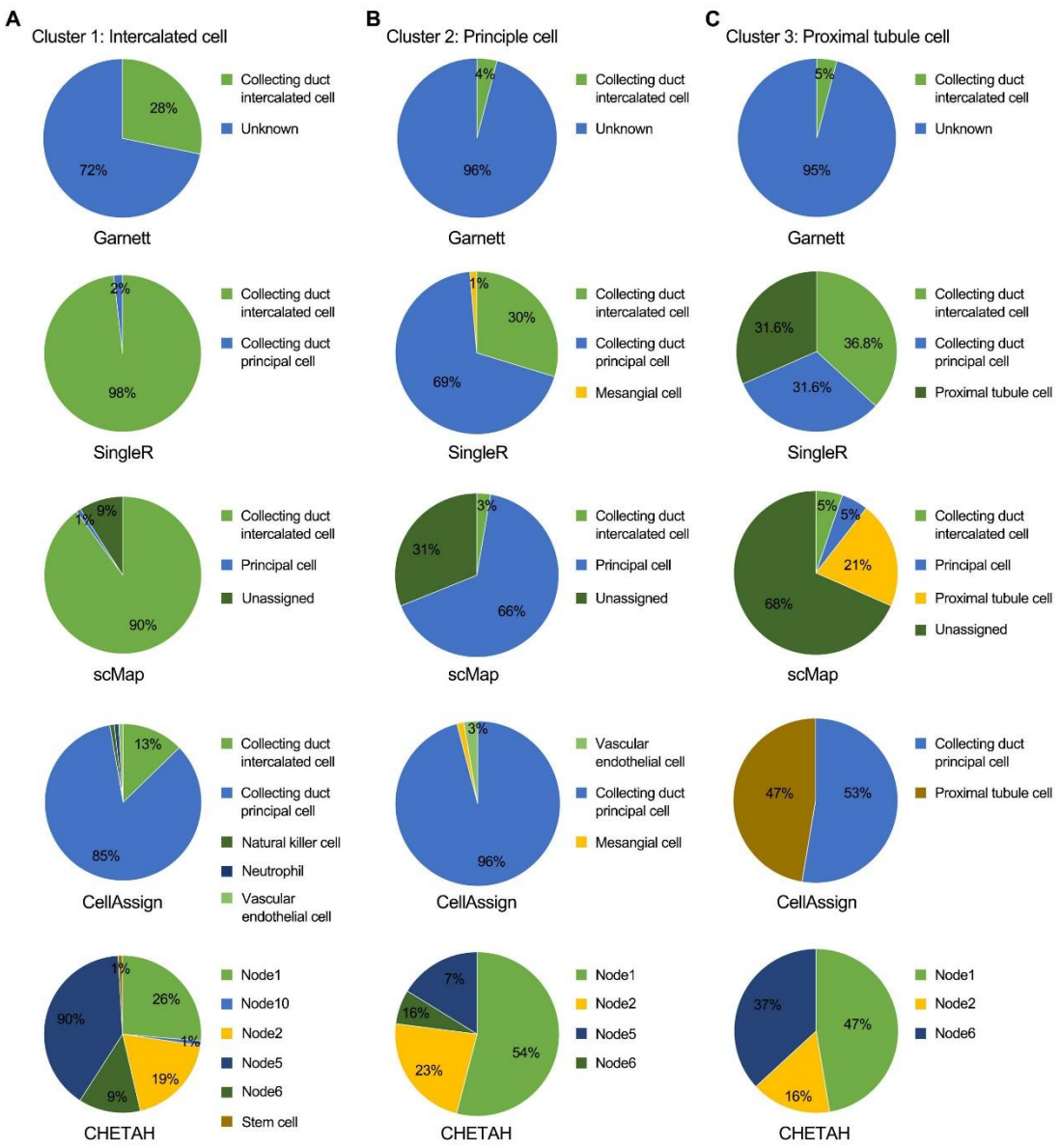

Figure S1.Cell composition in each cluster of Chen datasets, Related to Figure 3 and Table 1. Cell composition in cluster 1 **(A)**, cluster 2 **(B)** and cluster 3 **(C)**, annotated by Garnett, SingleR, scMap, CellAssign and CHETAH.

Table S1. Cell annotation by SingleR and CHETAH with various referred databases on Chen datasets, Related to Figure 3 and Table 1.

Dataset	Cluster	Cell type	Consistent rate				
			SingleR referred database			CHETAH referred database	
			Immgen	Mouse.RNAseq	CellMatch	Headneck	CellMatch
Chen	1	Intercalated cell	0%	0%	98%	NA	0%
	2	Principle cell	0%	0%	69%	NA	0%
	3	Proximal tubule cell	0%	0%	32%	NA	0%
	All	-	0%	0%	81%	NA	0%

Table S2. Cell annotation by SingleR and CHETAH with various referred databases on Xin and Gierahn datasets, Related to Figure 3 and Table 1.

Dataset	Cluster	Cell type	Consistent rate				
			SingleR referred database			CHETAH referred database	
			HPCA	Blueprint.encode	CellMatch	Headneck	CellMatch
Xin	1	Beta cell	0%	0%	95%	0%	0%
	2	Alpha cell	0%	0%	99%	0%	0%
	3	Delta cell	0%	0%	67%	0%	0%
	4	PP cell	0%	0%	72%	0%	0%
	All	-	0%	0%	95%	0%	0%
Gierahn	1	B cell	62%	56%	1%	0%	0%
	2	T cell	90%	95%	9%	20%	0%
	3	DC	74%	88%	0%	4%	0%
	4	NK cell	84%	75%	55%	0%	0%
	5	Monocyte	89%	52%	1%	0%	0%
	All	-	85%	67%	10%	5%	0%

Table S3. The clustering method and initialization platform for all datasets, Related to Table 1 and Table 2.

Dataset	Clustering method	Initialization platform	Compatibility
Chen	Seurat package	C1 Fluidigm system	√
Xin	Seurat package	C1 Fluidigm system	√
Gierahn	Seurat package	Seq-Well	√
Enge	NA	Smart-seq2	√
Wu	Louvain-Jaccard graph clustering	Drop-Seq	√
Lindsey	SPRING	Droplet microfluidics	√
Zheng	Seurat package	10X Genomics	√
Zeisel	BackSPIN	C1 Fluidigm system	√
Heng	K-means clustering	10X Genomics	√

NA, not available. √ represent the compatibility with the pipeline of scCATCH.

Transparent Methods

Datasets

scRNA-seq datasets were retrieved from several high-quality reports and Gene Expression Omnibus (GEO), including human and mouse primary tissues such as peripheral blood, brain, lung, kidney, and pancreas, wherein unannotated cells were excluded. The Zheng dataset (2,700 peripheral blood mononuclear cells [PBMCs]) was directly downloaded from Satija Lab (https://satijalab.org/seurat/). Validation datasets included the Chen, Xin, and Gierahn datasets, wherein cell types were experimentally validated via FACS or in situ hybridization and IHC. Test datasets included three internal datasets of Enge, Wu, and Lindsey and three external datasets of Zheng, Zeisel, and Heng, wherein cell types were annotated using known marker genes.

Construction of the CellMatch reference database

Human and mouse cell markers from CellMarker (http://biocc.hrbmu.edu.cn/CellMarker), MCA (https://figshare.com/articles/MCA_DGE_Data/5435866), CancerSEA (http://biocc.hrbmu.edu.cn/CancerSEA), and the CD Marker Handbook (http://static.bdbiosciences.com/documents/cd_marker_handbook.pdf) were retrieved.

For CellMarker database, cell markers derived from undefined tissue were excluded in this study and only cell markers with at least one supporting article were included. After integrating cell markers from the same tissue of origin, it led to 45,090 records (28,636 for human and 16,454 for mouse) involving 175 tissue types, 635 cell types, 20,213 cell marker genes and 2,085 references.

For MCA database, raw counts and the meta information of cell types were downloaded and processed. For each tissue, cells with the same annotation were merged and cell marker genes for each cell type were identified with FindAllMarkers of Seurat by using Normalized data via LogNormalize, in which the percentage of expressed cells was set to 75%, P value from Wilcoxon Rank-Sum (WRS) test to 0.01, and log10 fold change to 0.5. After the integration of marker gene from same tissue origin, it led to a total of 4,014 records related with 31 tissue

types, 139 cell types, and 1,486 cell marker genes.

For CancerSEA database, cell marker genes were curated from cancer related single-cell studies, wherein studies sourced from tissue were included. A total of 1,196 records were obtained from 11 references related with 6 tissue types, 27 cell types as well as 997 cell marker genes.

For CD Marker Handbook database, human and mouse key cell markers were collected as blood cell markers. A total of 54 records were obtained from CD Marker Handbook database related with 15 cell types and 50 cell marker genes.

Cell marker gene symbols and gene IDs were revised in accordance with NCBI gene data (https://www.ncbi.nlm.nih.gov/gene/) updated on July 1, 2019, wherein unmatched genes were removed from the CellMatch. Repeated records were combined, and cell types and subtypes were extracted from the names of annotated cells in accordance with histological origin, expression of specific markers or degrees of differentiation. To ensure the accuracy of CellMatch, manual confirmation were performed via independently examining the marker genes and reference by three reviewers. Lastly, cell marker genes curated from CellMarker, MCA, CancerSEA and CD Marker Handbook database were integrated to establish species-specific and tissue-specific reference database CellMatch, which includes 49,635 records (29,836 for human and 19,799 for mouse) involving 184 tissue types, 353 cell types and related 686 subtypes, 20,792 cell marker genes and 2,097 references.

Data pre-processing

All scRNA-seq data were processed using R (version 3.6.1). For Zheng datasets, the raw count was processed in accordance with the pipeline of the Satija Lab tutorial, using Seurat 3.0, wherein cells with unique feature counts of >2,500 or <200 and >5% mitochondrial counts were filtered out. For other datasets, all cells in the datasets were included in the filtered matrices and the meta information of cell clusters and cell types for all cells were obtained from the literature. Cells with same annotation were merged into the same cluster, and duplicated genes were combined through summation of raw counts for each cell. All datasets

were then saved as the CellDataSet class prepared for running scCATCH and other methods.

Data preparation for scCATCH

All datasets were transferred as Seurat objects from CellDataSet objects by extracting raw count and meta information of cell clusters and cell types. Then the raw counts were normalized via the global-scaling normalization method *LogNormalize*. Principal component analyses (PCA) were performed followed by uniform manifold approximation and projection (UMAP) analysis for dimensional reduction and visualization. All datasets were stored as Seurat objects prepared for running scCATCH.

Identification of cluster potential marker genes with scCATCH

For clusters i and j among n clusters ($i \neq j$, $i \,\&\, j \leq n$), $G_{i,j}$ was defined as the gene set in which every gene's average expression in cluster i is significantly greater than that in cluster j with the percentage of expressed cells (≥25%), using WRS test (P<0.05) and a $\log_{10}$ fold change of ≥0.25. For each cluster i, the cluster potential marker gene set M_i was obtained using the following equation:

$$M_i = G_{i,1} \cap G_{i,2} \cap G_{i,\ldots} \cap G_{i,j}$$

Cluster annotating process with scCATCH using cluster potential marker genes

Evidence-based scoring (*ES*) protocol in scCATCH involved two steps. The first step was to determine the cell type, and the second step was to determine the subtype of the corresponding cell type. For each cluster i, the cluster marker gene set M_i was matched with species-specific (human or mouse) and tissue-specific (blood, brain, kidney, etc.) cell markers from CellMatch database on the basis of revised gene symbols. c_i was considered as the matched unique cell type candidates. For each candidate k among c_i, the ES_k was determined as follows:

$$ES_k = \sqrt{\frac{l_k}{l_k+1} \times \frac{g_k}{g_k+1}} \quad (1)$$

In equation (1), l_k represents the unique number of related studies, while g_k is the number of associated cell marker genes, referring to the intersection of set M_i and the cell markers in

CellMatch database. Candidate k with the maximal ES_k was determined as the cell type for cluster i. Furthermore, s_i was considered the matched unique subtypes belonging to candidate k. For each subtype m among s_i, the $ES_{k,m}$ was determined as follows:

$$ES_{k,m} = \sqrt{\frac{l_{k,m}}{l_{k,m}+1} \times \frac{g_{k,m}}{g_{k,m}+1}} \quad (2)$$

In equation (2), $l_{k,m}$ represents the unique number of related evidence/reference while $g_{k,m}$ is the number of associated cell marker genes. Subtype m with the maximal $ES_{k,m}$ (> 0.5) was determined as the cell subtype for cluster i.

For scCATCH annotation, the mouse kidney cell markers was selected from CellMatch database for Chen dataset. The human pancreas and pancreatic islet cell markers were used for Xin and Enge datasets. The human blood, peripheral blood, and bone marrow cell markers were picked for Gierahn and Zheng datasets; mouse brain cell markers for Wu, Zeisel, and Heng datasets; human lung cell markers for Lindsey dataset.

Annotation with cluster potential marker genes identified by Seurat and scCATCH

For Seurat, cluster potential marker genes of three validation datasets were identified with *FindAllMarkers*, wherein Normalized data via "LogNormalize" were processed to determine the positive cluster potential marker genes with default parameters with the percentage of expressed cells (>10%), using WRS test (P<0.01) and a $\log_{10}$ fold change >0.25. For scCATCH, cluster potential marker genes of three validation datasets were identified with *findmarkergenes*, wherein Normalized data via *LogNormalize* were processed to determine the positive cluster potential marker genes with default parameters with the percentage of expressed cells (≥25%), using WRS test (P<0.05) and a $\log_{10}$ fold change ≥0.25. Both cluster potential marker genes generated from Seurat and scCATCH were used to annotate the cell types of three validation datasets via the function of *scCATCH* on the basis of *ES*s. As previously described, the tissue type for the Chen dataset, Xin dataset and Gierahn dataset was set to kidney, pancreas and pancreatic islet and blood, peripheral blood, and bone marrow, respectively, when annotating.

Performance comparison on the validation datasets with various methods

For CellAssign, three validation datasets were first transformed as SingleCellExperiment objects from CellDataSet object by extracting raw count and meta information of cell types. CellMatch database was then used as reference while all other parameter in CellAssign were kept as default (learning rate, 0.01).

For Garnett, the marker genes of mouse kidney cells, human pancreas and pancreatic islet cells, and human blood, peripheral blood and bone marrow cells from CellMatch database were first extracted to train corresponding classifiers of three validation datasets. The parameter of the number of unknown type cells was set as an outgroup during classification with a value of 50. Then the trained classifiers were used to classify the cells of Chen, Xin and Gierahn datasets.

For SingleR, three validation datasets were first transformed as a SingleR object from CellDataSet object by extracting raw count and meta information of cell types. To annotate Chen dataset by SingleR, the default databases of the Immunological Genome Project (ImmGen) and the mouse RNA-seq were used as the reference list, wherein a new reference list generated from CellMatch database by extracting the marker genes of mouse kidney cells was incorporated. To annotate Xin and Gierahn dataset by SingleR, the default databases of HPCA as well as Encode and Blueprint Epigenomics transcriptomes were used as the reference list, wherein a new reference list generated from CellMatch database by extracting the marker genes of human pancreas and pancreatic islet cells was incorporated for Xin dataset, and another new reference list by extracting the marker genes of human blood, peripheral blood and bone marrow cells was incorporated for Gierahn dataset.

For scMap, three validation datasets were transformed as SingleCellExperiment objects from CellDataSet object by extracting raw count and meta information of cell types. The three validation datasets were normalized with $\log_2$ raw counts and duplicated genes were removed from the normalized matrices. The individual cells in each dataset were used as the reference to projects cells of the corresponding dataset by scMap-cell.

For CHETAH, three validation datasets were transformed as SingleCellExperiment objects from CellDataSet object by extracting raw count and meta information of cell types. To annotate Chen, Xin and Gierahn datasets by CHETAH, the default reference of head and neck

were used and three new references were created and added by extracting the corresponding marker genes from CellMatch database, as previously described in SingleR.

Consistent rate evaluation

Consistent rate for scCATCH was defined as the percentage of consistent clusters annotated with the same cell type as in the literature, while consistent rate for CellAssign, Garnett, SingleR, scMap and CHETAH was defined as the percentage of consistent cells with the same cell type, as in the literature.

Code and data availability

The source code of scCATCH is implemented in R and is freely available at https://github.com/ZJUFanLab/scCATCH. The source code and results of performance comparison on the detail of the process among scCATCH, CellAssign, Garnett, SingleR, scMap and CHETAH, and CellMatch database are implemented in R and is freely available at https://github.com/ZJUFanLab/scCATCH_performance_comparison. All data are accessible in GEO with the following accession codes: (a) for Chen dataset, the accession code is GSE99701; (b) Xin dataset, GSE81608; (c) Gierahn dataset, GSM2486333; (d) Enge dataset, GSE81547; (f) Wu dataset, GSE103976; (g) Lindsey dataset, GSE102580; (h) Zeisel dataset, GSE60361; (i) Heng dataset, GSE125708.