










PreDigs: A Database of Context-specific Cell Type Markers and Precise Cell Subtypes for Digestive Cell Annotation

Jiayue Meng (孟佳玥) ^{1,#}, Mengyao Han (韩梦瑶) ^{2,#}, Yuwei Huang (黄俞玮) ^{1,#},
Liang Li (李梁) ¹, Yuanhu Ju (鞠元虎) ⁴, Daqing Lv (吕大庆) ⁴, Xiaoyi Chen (陈晓熠) ³,
Liyun Yuan (袁力贇) ^{1,*}, Guoqing Zhang (张国庆) ^{1,5,*}

¹National Genomics Data Center & Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai 200031, China

²Putuo People's Hospital, Shanghai Key Laboratory of Signaling and Disease Research, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China

³Ningbo Institute of Life and Health Industry, University of Chinese Academy of Sciences, Ningbo 315000, China

⁴Shanghai Southgene Technology Co., Ltd., Shanghai 201203, China

⁵Shanghai Sixth People's Hospital, Shanghai 200233, China

*Corresponding authors: gqzhang@sinh.ac.cn (Zhang G), lyyuan@sinh.ac.cn (Yuan L).

#Equal contribution.

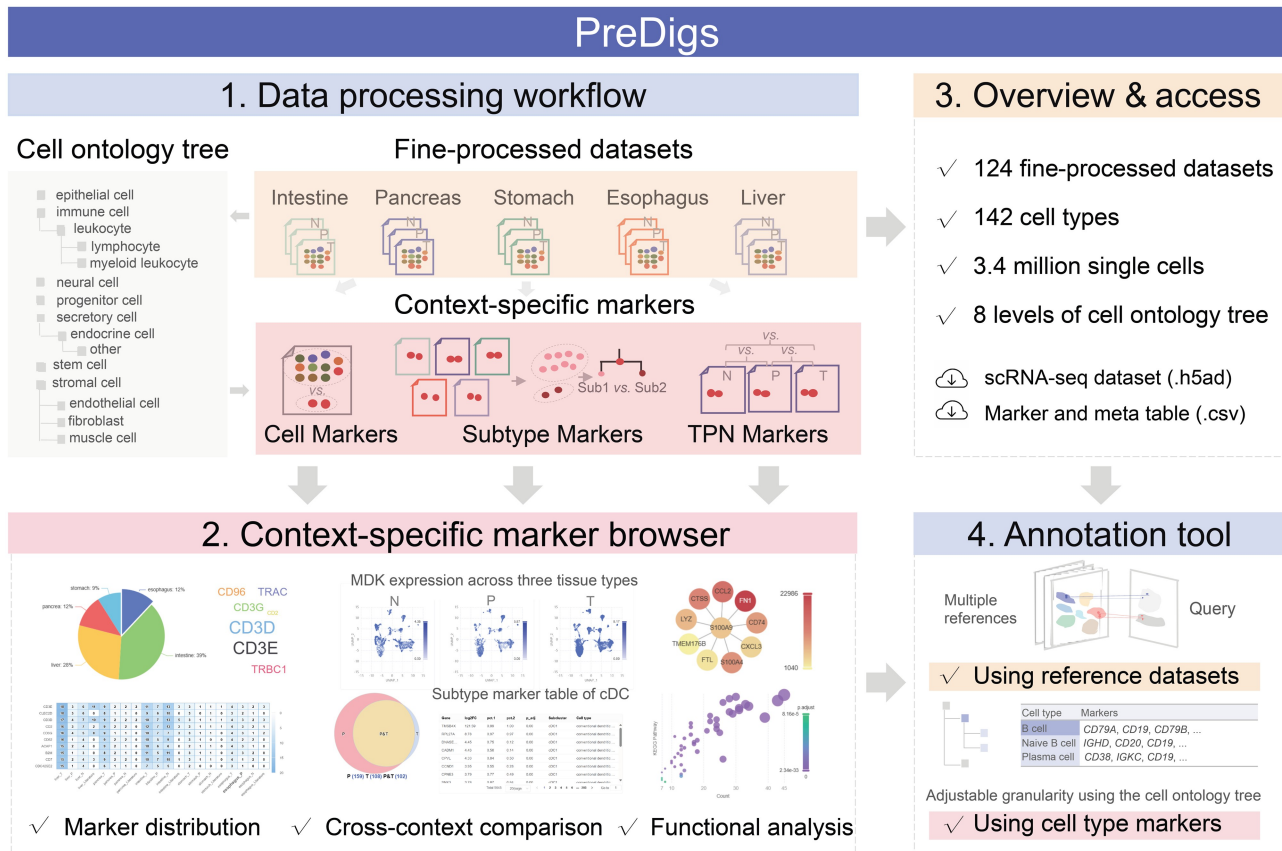
Handling Editor: Xiang Chen

Abstract

Research on cell type markers helps investigators explore the diverse cellular composition of gastrointestinal tumors, thereby enhancing our understanding of tumor heterogeneity and its impact on disease progression and treatment response. However, the integration of large-scale datasets and the standardization of cell type identification remain challenging. Here, we developed PreDigs, a user-friendly database of predicted signatures for the digestive system, which offers 124 curated single-cell RNA sequencing datasets, covering over 3.4 million cells, all available for download. After unsupervised clustering, we unified the identification and nomenclature of cell subtype labels, constructing a cell ontology tree with 142 cell types across 8 hierarchical levels. Meanwhile, we calculated three different context-specific cell type markers, including "Cell Markers", "Subtype Markers", and "TPN Markers", based on various application requirements within or across tissues. Through the integrated analysis of PreDigs data, we identified distinct cell subpopulations exclusive to tumors, one of which corresponds to tumor-specific endothelial cells. Additionally, PreDigs offers online cell annotation tools, allowing users to classify single cells with greater flexibility. PreDigs is accessible at <https://www.biosino.org/predigs/>.

Key words: Data integration; Cell type marker; Gastrointestinal tumor; Single-cell RNA sequencing; Database.

Graphical abstract



Introduction

Digestive cancers, including esophageal, gastric, liver, pancreatic, and colorectal cancers, account for more than 50% of global cancer-related morbidity and mortality [1]. As solid tumors, the complex cellular components, including malignant cells, tissue-resident cells, and tumor-infiltrating cells in the tumor microenvironment (TME), influence tumor cell growth and development [2,3]. Leveraging massive single-cell RNA sequencing (scRNA-seq) data enables comprehensive profiling of gene expression at the single-cell level, helping researchers uncover cellular diversity [4], distinguish cell lineages [5], and characterize the TME of digestive cancers [6]. Accurate cell type classification is crucial for single-cell analysis in cancer research and serves as the key to understanding intratumor heterogeneity [7–10].

Until now, great efforts have been made to develop several cell type marker data portals and resources for single-cell classification [11–20]. Notable databases include Cell Ontology (CL) [11], CellMarker [13], Panglaodb [15], Cell Taxonomy [19], and CellSTAR [20]. CL, a widely used framework for representing cell types, defines hierarchical relationships for approximately 2600 cell types, although most lack associated, assessable cell markers. CellMarker, the first database to compile a curated collection of 27,166 cell markers for 3123 cell types in human and mouse, does not include single-cell transcriptome profiles of these markers. In contrast, databases such as Panglaodb have compiled cell types, markers, and scRNA-seq

datasets; however, inconsistencies in dataset pre-processing procedures and the lack of a unified nomenclature for cell types, such as the adoption of the CL standardization, hinder direct cross-dataset integration and comparison. Cell Taxonomy and CellSTAR have addressed these issues, yet they do not support cross-tissue comparisons, such as analyzing gene expression variations in the same cell type between normal and tumor tissues. This gap limits our understanding of cell type markers in the TME and hinders precise annotation of TME cell types [21].

Here, we presented PreDigs, a comprehensive database of predicted signatures for the digestive system, which integrates extensive scRNA-seq datasets and utilizes sophisticated reference data for cell subtype annotation. PreDigs constructs a detailed cell ontology tree with 142 distinct cell types across 8 hierarchical layers. Each node within the hierarchy is linked to relevant expression profiles, enabling comparative analyses across diverse tissues and datasets. Furthermore, PreDigs affords three types of markers tailored for different research contexts: direct comparison of differentially expressed gene (DEG) profiles for a cell type across tissue types (“Cell Markers”), delineation of gene expression differences within cell subtypes via cross-tissue single-cell analysis (“Subtype Markers”), and identification of cell types specifically distributed within a tumor tissue and their gene functions in Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (“TPN Markers”). The PreDigs interface allows users to retrieve cell type markers and scRNA-seq datasets for in-depth gene expression and functional enrichment analyses.

Additionally, an online annotation tool is available to streamline single-cell classification. It supports annotation using both classical markers and customized scRNA-seq reference datasets, improving classification flexibility and reliability.

Database content and usage

Overview of data management and integration in PreDigs

In summary, we obtained 124 well-processed scRNA-seq datasets from adult digestive tissues (see Method; [Figure S1](#)). These datasets cover three types of digestive tissues (normal tissue, precancerous tissue, and tumor tissue) across five digestive organs (intestine, pancreas, esophagus, liver, and stomach), comprising approximately 3.4 million single cells ([Figure 1A](#)). Furthermore, to systematically classify digestive cell types, we developed a hierarchical cell ontology tree with up to 8 levels, incorporating 142 cell types based on the CL database (see Method; [Figure S1](#)). Of these, 74% (105 types) correspond to standardized CL identifiers (IDs). The first layer of the cell ontology tree consists of 11 major groups ([Figure 1B](#)), with immune cells, stromal cells, secretory cells, and epithelial cells being the 4 primary categories. For instance, as shown in [Figure 1B](#), immune cells are categorized into two main lineages, including lymphocyte and myeloid leukocyte. These are further divided into 64 hierarchical subtypes. PreDigs enhances cell ontology annotation by incorporating 3693 canonical markers across 74 cell types, curated from a comprehensive review. This integration enriches marker-based cell type characterization.

After obtaining the well-processed scRNA-seq datasets and cell ontology tree, we identified three context-specific types of cell markers using different strategies for different applications (see Method; [Figure 1C](#)). First, we screened DEGs with significantly higher expression in specific cell types compared to others within each dataset (“Cell Markers”). We then compared DEG sets from the same cell type across different tissue types to identify shared and unique features. Second, we performed cross-tissue integrative analysis to investigate gene expression differences among subtypes within over 140 cell types (“Subtype Markers”). Finally, we integrated multiple datasets within specific tissues to analyze expression differences of specific cell types across three tissue types and performed functional enrichment analysis (“TPN Markers”).

Through this workflow, we obtained 2,782,494 marker records ([Figure 1D](#)), including 1,994,093 “Cell Markers”, 17,245 “TPN Markers”, and 771,156 “Subtype Markers”. The number of “Cell Markers” in the intestine, liver, and pancreas was significantly higher than that in the esophagus and stomach. This trend correlates with the number of datasets. However, the number of “TPN Markers” in the intestine is significantly higher than that in other gastrointestinal organs. This could be attributed to the greater variability in TPN among intestinal samples than in other tissues ([Figure 1D](#)).

Web interface and usage

Here, we developed PreDigs, a web resource for viewing cell type markers based on different contextual needs, enabling browsing and comparison across organs, tissues, and cell types as well as downloading tailored scRNA-seq datasets. It includes six key pages ([Figure 1E](#)): Home, Cell Markers, Subtype Markers, TPN Markers, Tools, and Download. The Home

page offers a cell ontology tree, a summary of data, and marker preparation details. On the three marker pages, a marker table with detailed links to single gene expression and KEGG pathway enrichment results is provided. This setup enables users to explore gene expression features and functional implications after obtaining the marker list. Finally, these data are integrated into an online tool for accurate annotation of digestive cells.

Interface for browsing three types of cell markers

In PreDigs, a cell ontology tree is always displayed on the left side of all web pages, allowing users to browse and select cell types. Three context-specific markers, namely “Cell Markers”, “Subtype Markers”, and “TPN Markers”, are displayed, respectively, across organs, tissues, and cell types. Once a cell type is selected in the cell ontology tree on the left, detailed information including cell type name, cell ontology ID, and description will be displayed first. Meanwhile, the marker list will appear in the table below ([Figure 2A](#)), where users can click the eye icon on the right to access details for each marker gene. Here, we take “Cell Markers” page as an example to introduce how to browse marker information, while the content on the other two marker pages is relatively simpler but follows a similar structure.

The “Marker overview” tab on the “Cell Markers” page, as shown in [Figure 2B](#), presents an overview of the “Cell Markers” contained within the selected cell type, including their quantitative distribution across gastrointestinal tissues and comparisons among different tissue types. We additionally quantified the occurrences of markers in each tissue and visualized the top 10 most frequently occurring markers, helping users better understand their expression characteristics across different tissues.

The “Markers between two cell types” tab ([Figure 2C](#)) facilitates quick comparison between two distinct cell types within a particular tissue, allowing users to instantly view similarities and differences in their marker lists.

The “Dataset overview” tab ([Figure 2D](#)) focuses on individual scRNA-seq datasets, enabling users to explore cell clustering via *t*-distributed stochastic neighbor embedding (*t*-SNE), intercellular communication via CellChat, overlapping and unique markers visualized using Venn diagrams, gene expression patterns using *t*-SNE, gene interactions via protein–protein interaction (PPI) networks, prognostic impact through survival analysis, and functional insights through KEGG pathway analysis.

The “Subtype Markers” page provides a research platform for in-depth analysis of subtypes within a given cell type, as defined in the cell ontology tree on the left ([Figure S2](#)). We performed refined clustering of each cell type, allowing users to explore the distribution of subtypes across organs and tissue types. This enables a better understanding of subtype diversity and heterogeneity in digestive tissues. Additionally, users can browse detailed markers of each subtype in the table below.

The “TPN Markers” page focuses on identifying DEGs within the same cell type through pairwise comparison across distinct tissue types, namely tumor, precancerous, and normal tissues. It displays the distribution of the selected cell type within the organ, and across each individual sample ([Figure S3](#)). All pairwise comparison markers are presented in the table below, with details for each TPN marker gene accessible via the eye icon on the right. Detailed information, including gene expression [Uniform Manifold Approximation and Projection (UMAP)],

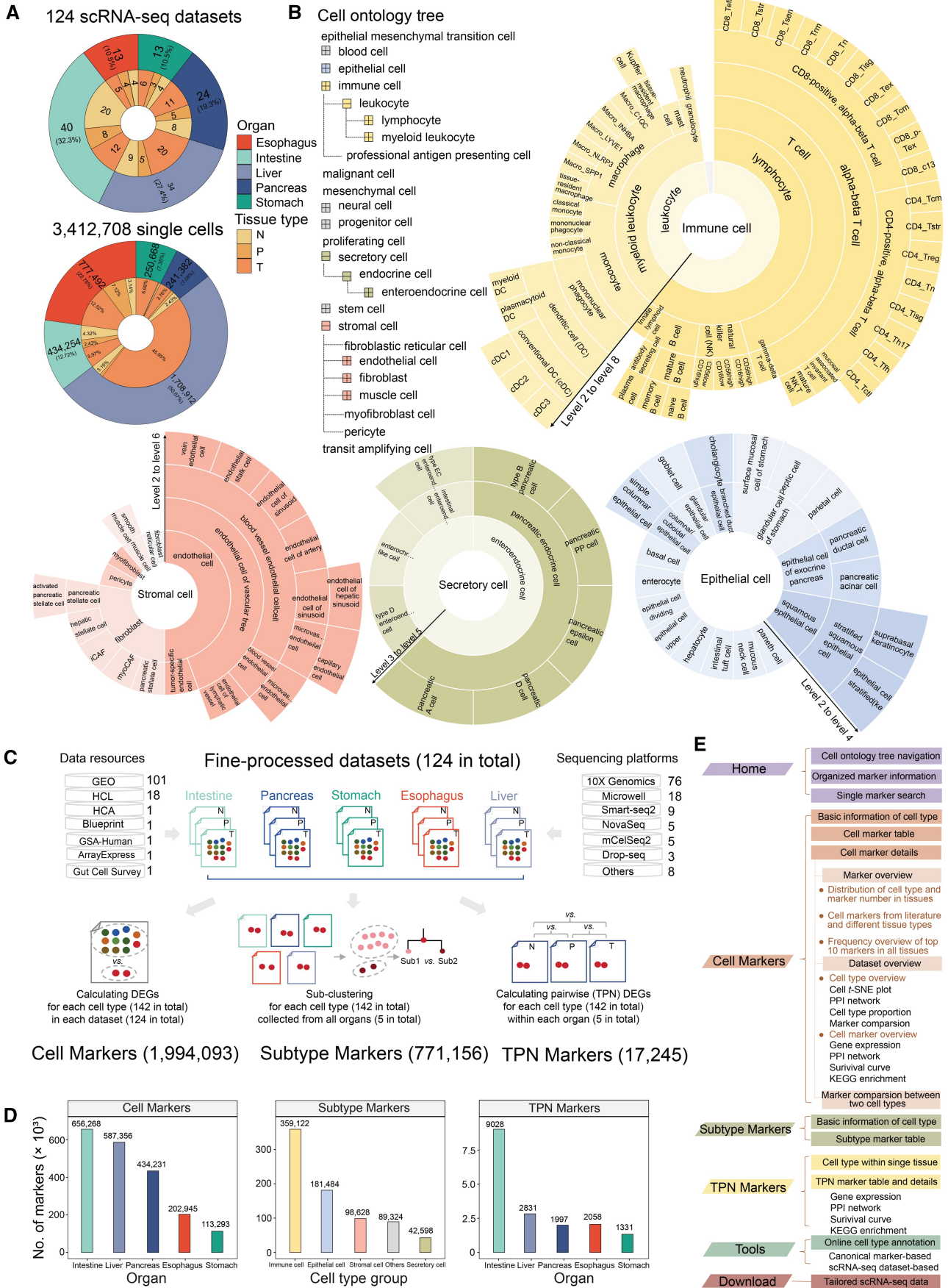


Figure 1 The summary of PreDigs data

gene network (PPI), functional analysis (survival analysis and KEGG enrichment), is displayed. These analytical results help researchers identify key signaling pathways and elucidate their biological significance for further studies.

Online cell type annotation platform

PreDigs provides an interactive platform for cell type annotation on the “Tools” page (Figure 3), offering two annotation methods: cell marker-based annotation and reference dataset-based annotation. Users can upload single-cell expression matrices and metadata tables, and customize their annotation process by selecting automated annotation tools, multiple reference datasets, and additional parameter options. For cell marker-based annotation, users receive a comprehensive list of marker genes for all cell types associated with the selected tissue, with the option to exclude any unnecessary types. For tailored reference datasets, users can select one or more datasets as references for annotation. Upon submission, the platform generates *t*-SNE plots displaying cell group names and predicted labels. Additionally, users can download prediction results and customized feature gene lists, enhancing both accessibility and efficiency when working with the PreDigs database.

Utility of PreDigs reference for revealing cell type heterogeneity

In this study, we first used endothelial cells as an example and performed a large-scale integration of 124 scRNA-seq datasets from 5 digestive organs in the PreDigs database. The integration categorized 96,429 cells into 8 distinct subtypes (Figure 4A). Notably, we identified a subtype of 20,019 single cells, predominantly in tumor tissues, designated as gastrointestinal tumor-specific endothelial cells (TECs). Furthermore, we identified 375 “Subtype Markers” highly expressed in TECs compared to other endothelial subtypes. KEGG functional enrichment results revealed that half of the top 10 enriched pathways for these markers were cancer-related, including “Pathway in cancer”, “PI3K-Akt signaling pathway”, and “Small cell lung cancer”, highlighting the pathological state of endothelial cells in tumor tissue (Figure 4B).

Using a similar approach, we integrated and clustered 6346 Macro_LYVE1 cells into 8 subclusters, and found that subcluster2 (“clu2”) was exclusively present in tumor tissues (Figure 4C). We then identified 173 “Subtype Markers” highly expressed in this cluster relative to other Macro_LYVE1 subclusters, which exhibited significant enrichment in the “IL-18 signaling pathway”, awaiting further biological investigation (Figure 4D).

Figure 1 Continued

A. PreDigs includes 124 tailored scRNA-seq datasets with 142 cell types, covering nearly 3.4 million cells across 5 digestive organs and 3 tissue types. The pie plots show the distributions of datasets (upper) and single cells (lower) across different digestive organs and tissue types. **B.** An illustration of the cell ontology tree with 8 levels. The cell type names in the hierarchical tree follow the standardized nomenclature and formatting provided by the Cell Ontology database [11]. Except for special cases, all characters are in lowercase, with words separated by spaces. **C.** Overview of data processing workflow for context-specific markers in PreDigs. **D.** Statistics on the number of three types of context-specific markers across different digestive tissues. **E.** Schematic overview of the PreDigs database interface. scRNA-seq, single-cell RNA sequencing; N, normal tissue; P, paracancerous tissue; T, tumor tissue; GEO, Gene Expression Omnibus; HCL, Human Cell Landscape; HCA, Human Cell Atlas; GSA-Human, Genome Sequence Archive for Human; PPI, protein–protein interaction; KEGG, Kyoto Encyclopedia of Genes and Genomes; TPN, tumor–paracancerous–normal pairwise comparison; DEG, differentially expressed gene.

Conclusion

One of the key strengths of PreDigs lies in its comprehensive coverage of digestive tissue types and the standardized cell ontology tree, offering unique features compared to existing resources. Each cell type node is accompanied by extensive marker calculations, which we expect to facilitate single-cell annotation across various research contexts, particularly in cross-tissue comparisons to explore cellular heterogeneity and similarities among digestive organs.

However, the construction of this cell ontology tree currently relies heavily on available marker annotations. The initial version of PreDigs focuses on gastrointestinal cancers, given the challenges of constructing a new cell type hierarchy and the substantial effort required for large-scale data integration. Moving forward, we aim to extend PreDigs to encompass a broader range of cell and tissue types, particularly by integrating datasets for digestive tumor subtypes not yet represented. To further enhance the database’s utility, we plan to incorporate multi-omics data, such as spatial transcriptomics and single-cell T cell receptor sequencing. In addition, we will introduce large language models to improve single-cell classification by leveraging their ability to process complex marker-based annotations and cross-dataset comparisons.

Combined with annual data updates, these advancements will ensure that PreDigs continues to evolve as a comprehensive resource for digestive system research.

Method

Dataset collection and pre-processing

A total of 99 original scRNA-seq datasets for the adult digestive system were collected from 7 major public data resources as of February 2024, including National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO), Human Cell Landscape (HCL) [22], Human Cell Atlas (HCA), ArrayExpress, Genome Sequence Archive for Human (GSA-Human) [23], Gut Cell Survey, and Blueprint. These datasets originate from various sequencing platforms, including 10X Genomics, Microwell, Smart-seq2, Drop-seq, STRT-seq, NovaSeq, and Fluidigm. The gene expression matrices, as well as the metadata tables containing cell type labels (if accessible in publications), were downloaded. By manually reviewing each publication and its supplemental materials, we curated the metadata for each dataset, including participant, cell number, cell type number, cancer type, technology, and dataset sources. The curated metadata was elaborated in the database.

Prior to standardized pre-processing, these 99 datasets were subdivided into 124 independent data subsets categorized by three tissue types: normal tissue, paracancerous tissue, and tumor tissue. All the scRNA-seq data were processed using

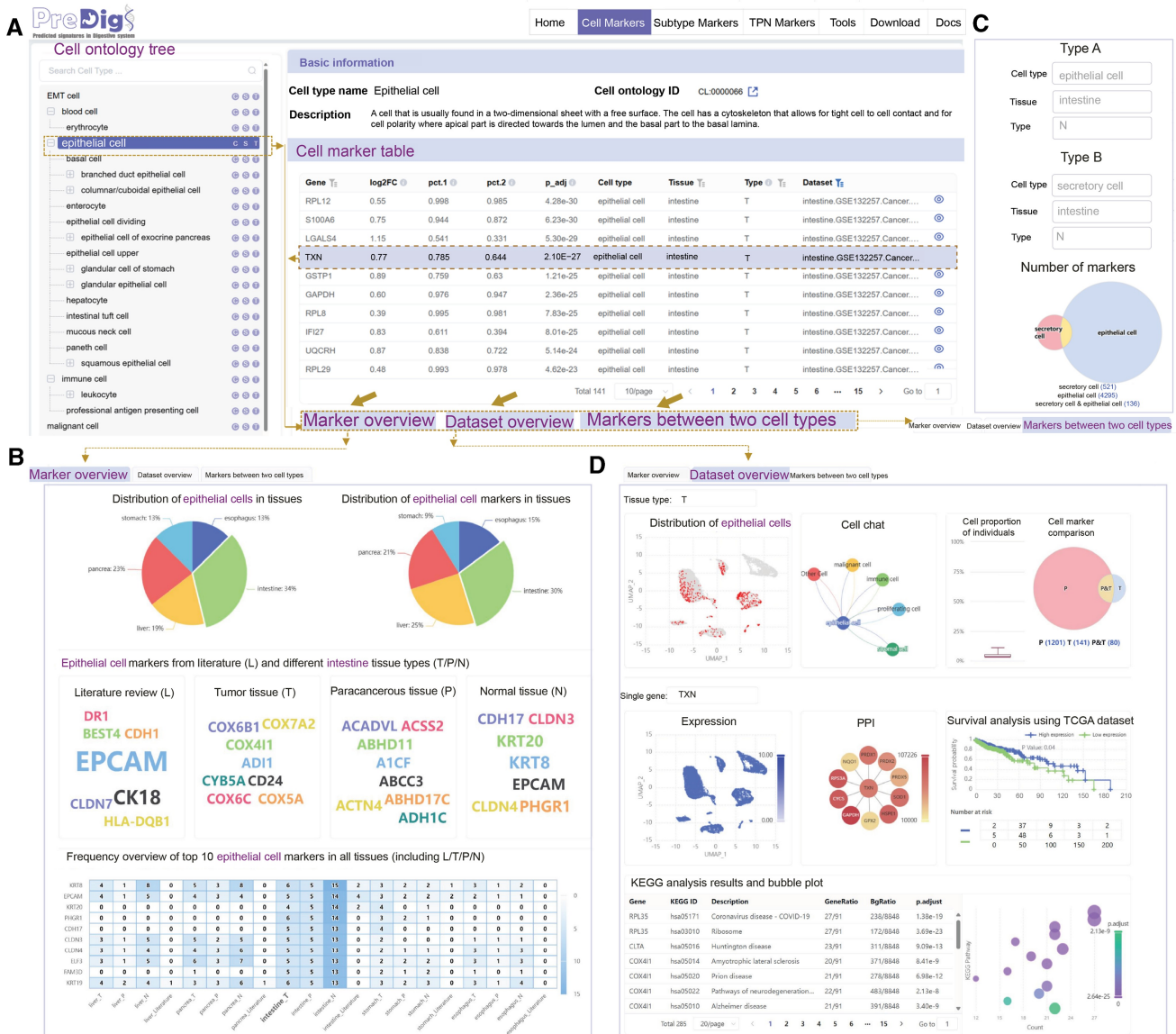


Figure 2 Screenshots of “Cell Markers” interface in PreDigs

A. “Cell Markers” interface includes a cell ontology tree, a marker table, and three tabs for different functions. **B.** “Marker overview” tab for cell marker statistics. **C.** “Markers between two cell types” tab for comparative analysis. **D.** “Dataset overview” tab for expression profiles of scRNA-seq datasets.

R (v3.6.1) with uniform methods except special explanation. Doublets were removed using DoubletFinder (v2.2.0) [24] (7% per 10000 cells) for all datasets. Seurat (v2.3.2) [25] was employed to perform a series of analyses: quality control [200 < nfeature < 5000 and mitochondrial (MT) gene percentage < 10%], normalization, principal component analysis (PCA), dimensionality reduction, cell clustering with the top 30 principal components, UMAP visualization, and differential gene expression analysis through the “FindAllMarkers” function with Wilcoxon rank-sum test (adjusted P value < $1E-04$, $|\log_2$ fold change| > 1.5). For datasets generated by full-length sequencing, transcript per million (TPM) normalization was mainly performed instead of the Seurat workflow.

Batch removal and data integration were performed by Harmony (v1.0) [26] and canonical correlation analysis algorithm in Seurat (v2.3.2) [25]. To integrate the data, we first

identified variable genes in each dataset with the “FindVariableFeatures” function. We then scaled the data using the “ScaleData” function and proceeded to integrate different datasets using Seurat’s “FindIntegrationAnchors” and “IntegrateData” functions, with parameters set as follows: k.filter = 50, resolution = 0.5, and dims = 30. The aforementioned processes were facilitated by Visualization integrated Precision Medicine Analytics Platform (ViPMAP; <https://www.biosino.org/vipmap/>).

Cell type annotation and clustering

We performed unified cell type annotations on the pre-processed gene expression matrix. The annotation process for healthy gastrointestinal tissues mainly consisted of three steps. First, we used the tool CellAssign [27] for automated cell type prediction. Next, manual annotation was conducted based on the expression of typical cell markers. Then, cell

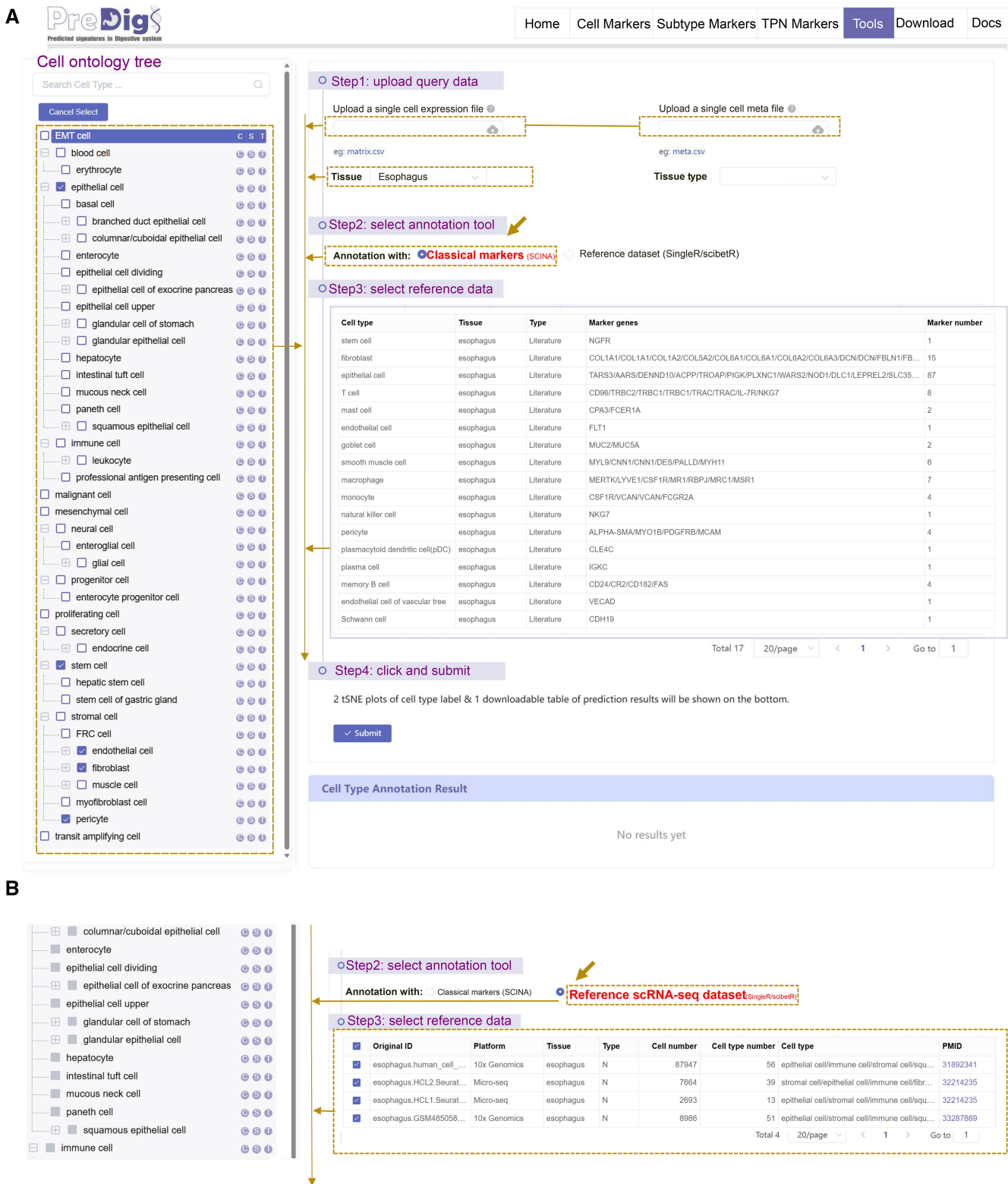


Figure 3 Screenshots of “Tools” interface in PreDigs

A. Step-by-step instructions for annotation using classical markers as reference data. **B.** Step-by-step instructions for annotation utilizing single or multiple scRNA-seq datasets as reference data.

labels were determined by comparing the re-annotated results with the original literature. Finally, cell label names were manually curated following the standardized terms in the CL

database. For malignant cell classification in tumor tissues, we further refined the annotation by incorporating the expression of typical tumor cell marker genes (*EPCAM*, *TP53*, etc.).

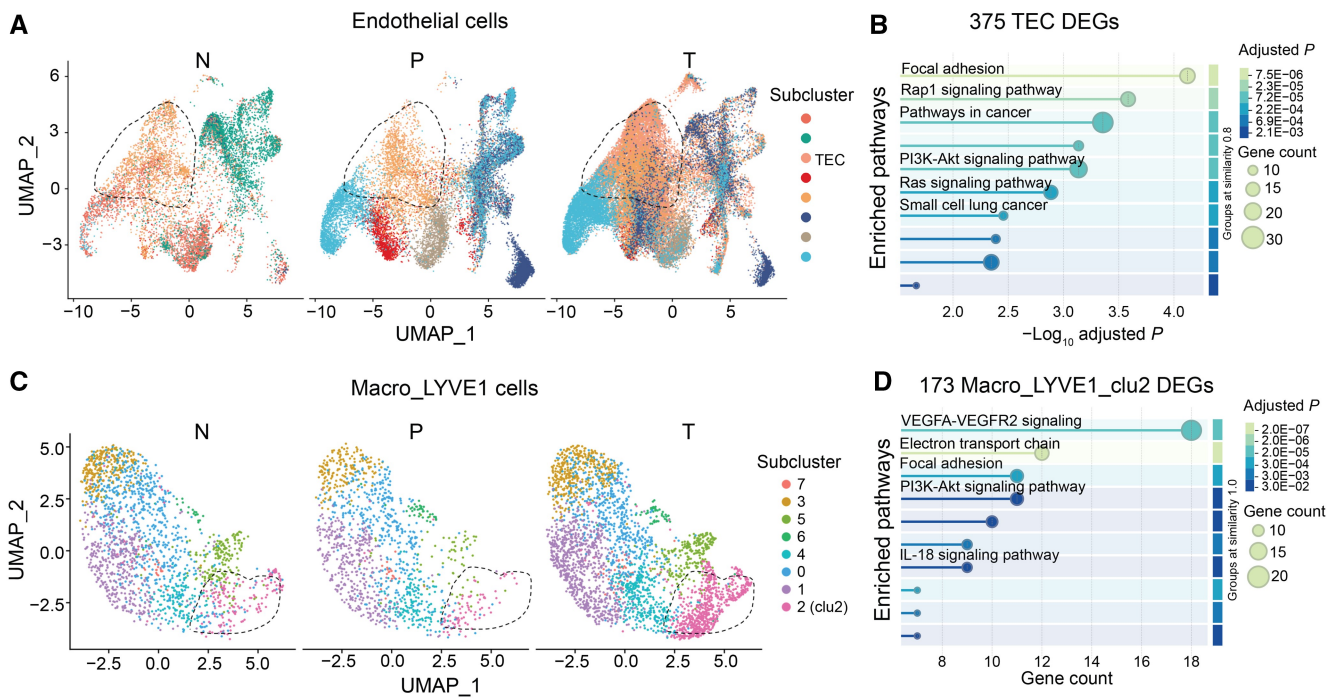


Figure 4 The utility of PreDigs reference for cell subtype annotation

A. Unsupervised clustering results of endothelial cells from the cross-tissue integration of PreDigs data. **B.** KEGG enrichment results of 375 “Subtype Markers” up-regulated in TECs compared to other endothelial subtypes, using a threshold of adjusted $P < 1E-04$ and \log_2 fold change > 1.5 . **C.** Unsupervised clustering results of Macro_LYVE1 cells resulting from the cross-tissue integration of PreDigs data. **D.** Wiki pathway enrichment analysis for 173 “Subtype Markers” up-regulated in clu2 relative to other subclusters within the Macro_LYVE1 cell population, using a threshold of adjusted $P < 1E-04$ and \log_2 fold change > 1.5 . TEC, tumor-specific endothelial cell; clu2, cluster 2.

Furthermore, to enhance cell type resolution, we conducted additional subtype annotations and classifications for six major cell types, including T cells, natural killer cells, myeloid cells, endothelial cells, B cells, and fibroblasts, using sophisticated reference data [6,28–32] (Table S1). For the first four cell types, larger-scale single-cell atlases were used as the reference dataset, combined with SingleR [33] and scibetR [34] for automatic refinement. Then, the re-annotated results were compared with the cell labels in the original literature. For the latter two cell types, due to the lack of available large-scale reference atlases, typical subtype markers were mainly referenced, combined with Semi-supervised Category Identification and Assignment (SCINA) [35] for automatic annotation. Similarly, the re-annotated results were compared with the labels in the original literature.

Additionally, cross-tissue data integration and clustering were utilized to analyze cell subtypes in the TME. First, single-cell data from various organs were integrated using Seurat (v2.3.2) [25] to classify and analyze cell subtypes, leading to the identification of tumor-specific subclusters. Secondly, Seurat’s canonical correlation analysis algorithm was applied to combine and cluster multiple datasets within the same gastrointestinal tumor, allowing differences in gene expression patterns of the same cell type across normal, paraneoplastic, and tumor tissues to be examined.

Cell ontology tree construction

Following the comprehensive annotation of digestive single cells, a standardized ontology for cell types was constructed to structurally describe these cell types in the scRNA-seq dataset of PreDigs. This cell ontology tree was manually curated and derived from the CL reference database.

A complete list of digestive cell types was obtained and matched to the existing CL hierarchy using standardized nomenclature. Subsequently, the corresponding CL numbers and parent node cell types for each cell type were compiled and organized. A unified and structured cell ontology tree was then constructed, mainly categorizing cells into five groups: immune cells, stromal cells, epithelial cells, secretory cells, and others. The text descriptions corresponding to these cell types were also documented and uniformly stored in the database.

To construct a comprehensive cell type marker set for human digestive organs and accurately profile digestive cellular characteristics, manual curation was performed based on existing resources, such as CellMarker 2.0 [13] and CellMatch [16], and canonical markers that had been experimentally validated were collected.

Approach for deriving three types of context-specific cell markers

To uncover expression differences of the same cell type across different tissue types, the Seurat package was employed to conduct three distinct differential gene expression analyses for specific cell types under various conditions, including “Cell Markers”, “Subtype Markers”, and “TPN Markers”. First, significantly up-regulated DEGs in the specific cell type, compared to other cell types, were identified within each tailored dataset (“Cell Markers”). The heterogeneity of cell type-specific DEGs across different tissues was then elucidated by comparing these DEG lists. Next, all datasets across digestive organs containing the specific cell type were integrated to construct a large-scale cell integration dataset. Expression differences between subtypes within the same cell

type were analyzed using this integrated dataset (“Subtype Markers”). Finally, to meet the research needs specific to a particular digestive organ, relevant datasets were consolidated, and an in-depth analysis of the gene expression differences for the specific cell type across different tissue types (normal tissue, paracancerous tissue, and tumor tissue) was carried out after eliminating batch effects (“TPN Markers”).

Data visualization and database construction

In addition to basic visualization of cell type composition statistics and DEG analysis, the UMAPPlot and FeaturePlot functions in Seurat (v2.3.2) [25] were employed for cell clustering visualization and feature expression. Cell–cell interaction networks were constructed by CellChat [36]. PPI networks were constructed by STRINGdb [37]. KEGG enrichment analysis on DEGs was performed for each specific cell type using the R package clusterProfiler [38]. Survival analysis was performed by the R package survival. Static figures in the database were created by ggplot2.

PreDigs web portal was constructed by standard database development techniques. Hyper Text Markup Language 5 (HTML5) and Cascading Style Sheets (CSS) were used for front-end page display and My Structured Query Language (MySQL) was served as a container for back-end data storage. ECharts and Highcharts were adopted for building interactive graphs. The following browsers were recommended for better compatibility: Google Chrome, Firefox (v64.0 and up), or Internet Explorer (v11.0 and up). The whole bioinformatic analysis results in PreDigs were generated by in-house R scripts.

Code availability

PreDigsR, a standalone version of the PreDigs database for cell type annotation, is available on GitHub (<https://github.com/BioMedBigDataCenter/predigsr>). It has also been submitted to BioCode at the National Genomics Data Center (NGDC), China National Center for Bioinformation (CNCB) (BioCode: BT007624), which is publicly accessible at <https://ngdc.cncb.ac.cn/biocode/tool/7624>.

Data availability

All data in PreDigs are freely available at <https://www.biosino.org/predigs/>, with no login or registration required. The database has been submitted to Database Commons [39] at the NGDC, CNCB, which is publicly accessible at <https://ngdc.cncb.ac.cn/databasecommons/database/id/9719>.

CRedit author statement

Jiayue Meng: Investigation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing. **Mengyao Han:** Data curation. **Yuwei Huang:** Resources. **Liang Li:** Investigation. **Yuanhu Ju:** Software. **Daqing Lv:** Software. **Xiaoyi Chen:** Investigation. **Liyun Yuan:** Projection administration, Methodology, Writing – original draft, Writing – review & editing. **Guoqing Zhang:** Conceptualization, Projection administration, Funding acquisition, Writing – original draft, Writing – review & editing. All authors have read and approved the final manuscript.

Competing interests

Yuanhu Ju and Daqing Lv are current employees of Shanghai Southgene Technology Co., Ltd. All the other authors have declared no competing interests.

Acknowledgments

This work was supported by grants from the National Major Scientific Research Instrument Development Project of NSFC (Grant No. 81827901), the Shanghai Municipal Science and Technology Major Project (GTP), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB38030100), the National Key R&D Program of China (Grant No. 2021YFF0703702), and the Key R&D Program of Shandong Province of China (Competitive Innovation Platform) (Grant No. 2023CXPT102). We gratefully acknowledge the contributors of published studies for sharing their scRNA-seq data on human digestive samples.

Supplementary material

Supplementary material is available at *Genomics, Proteomics & Bioinformatics* online (<https://doi.org/10.1093/gpbjnl/qzaf066>).

ORCID

0009-0003-5744-063X (Jiayue Meng)
 0000-0002-3751-3219 (Mengyao Han)
 0000-0003-1571-7103 (Yuwei Huang)
 0009-0007-2903-2347 (Liang Li)
 0009-0009-2812-0657 (Yuanhu Ju)
 0009-0003-5678-2486 (Daqing Lv)
 0009-0005-0349-1890 (Xiaoyi Chen)
 0000-0001-9122-6782 (Liyun Yuan)
 0000-0001-8827-7546 (Guoqing Zhang)

References

- [1] Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin* 2024;74:12–49.
- [2] Zhang Z, Wang ZX, Chen YX, Wu HX, Yin L, Zhao Q, et al. Integrated analysis of single-cell and bulk RNA sequencing data reveals a pan-cancer stemness signature predicting immunotherapy response. *Genome Med* 2022;14:45.
- [3] Maman S, Witz IP. A history of exploring cancer in context. *Nat Rev Cancer* 2018;18:359–76.
- [4] Wang Y, Navin NE. Advances and applications of single-cell sequencing technologies. *Mol Cell* 2015;58:598–609.
- [5] Kester L, van Oudenaarden A. Single-cell transcriptomics meets lineage tracing. *Cell Stem Cell* 2018;23:166–79.
- [6] Cheng S, Li Z, Gao R, Xing B, Gao Y, Yang Y, et al. A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells. *Cell* 2021;184:792–809.e23.
- [7] Ono H, Arai Y, Furukawa E, Narushima D, Matsuura T, Nakamura H, et al. Single-cell DNA and RNA sequencing reveals the dynamics of intra-tumor heterogeneity in a colorectal cancer model. *BMC Biol* 2021;19:207.
- [8] Croft W, Evans RPT, Pearce H, Elshafie M, Griffiths EA, Moss P. The single cell transcriptional landscape of esophageal adenocarcinoma and its modulation by neoadjuvant chemotherapy. *Mol Cancer* 2022;21:200.

- [9] Ferrall-Fairbanks MC, Ball M, Padron E, Altrock PM. Leveraging single-cell RNA sequencing experiments to model intratumor heterogeneity. *JCO Clin Cancer Inform* 2019;3:1–10.
- [10] Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat Biotechnol* 2021;39:599–608.
- [11] Diehl AD, Meehan TF, Bradford YM, Brush MH, Dahdul WM, Dougall DS, et al. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J Biomed Semantics* 2016;7:44.
- [12] Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, et al. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res* 2019;47:D721–8.
- [13] Hu C, Li T, Xu Y, Zhang X, Li F, Bai J, et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res* 2023;51:D870–6.
- [14] Hatano A, Chiba H, Moesa HA, Taniguchi T, Nagaie S, Yamanegi K, et al. CELLPEDIA: a repository for human cell information for cell studies and differentiation analyses. *Database (Oxford)* 2011;2011:bar046.
- [15] Franzén O, Gan LM, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)* 2019;2019:baz046.
- [16] Shao X, Liao J, Lu X, Xue R, Ai N, Fan X. scCATCH: automatic annotation on cell types of clusters from single-cell RNA sequencing data. *iScience* 2020;23:100882.
- [17] Choi JH, In Kim H, Woo HG. scTyper: a comprehensive pipeline for the cell typing analysis of single-cell RNA-seq data. *BMC Bioinformatics* 2020;21:342.
- [18] Stachelscheid H, Seltmann S, Lekschas F, Fontaine JF, Mah N, Neves M, et al. CellFinder: a cell data repository. *Nucleic Acids Res* 2014;42:D950–8.
- [19] Jiang S, Qian Q, Zhu T, Zong W, Shang Y, Jin T, et al. Cell Taxonomy: a curated repository of cell types with multifaceted characterization. *Nucleic Acids Res* 2023;51:D853–60.
- [20] Zhang Y, Sun H, Zhang W, Fu T, Huang S, Mou M, et al. CellSTAR: a comprehensive resource for single-cell transcriptomic annotation. *Nucleic Acids Res* 2024;52:D859–70.
- [21] Mayer S, Milo T, Isaacson A, Halperin C, Miyara S, Stein Y, et al. The tumor microenvironment shows a hierarchy of cell–cell interactions dominated by fibroblasts. *Nat Commun* 2023;14:5810.
- [22] Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, et al. Construction of a human cell landscape at single-cell level. *Nature* 2020;581:303–9.
- [23] Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive Family: toward explosive data growth and diverse data types. *Genomics Proteomics Bioinformatics* 2021;19:578–83.
- [24] McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst* 2019;8:329–37.e4.
- [25] Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. *Cell* 2021;184:3573–87.e29.
- [26] Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 2019;16:1289–96.
- [27] Zhang AW, O’Flanagan C, Chavez EA, Lim JLP, Ceglia N, McPherson A, et al. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat Methods* 2019;16:1007–15.
- [28] Chu Y, Dai E, Li Y, Han G, Pei G, Ingram DR, et al. Pan-cancer T cell atlas links a cellular stress response state to immunotherapy resistance. *Nat Med* 2023;29:1550–62.
- [29] Tang F, Li J, Qi L, Liu D, Bo Y, Qin S, et al. A pan-cancer single-cell panorama of human natural killer cells. *Cell* 2023;186:4235–51.e20.
- [30] Xia J, Xie Z, Niu G, Lu Z, Wang Z, Xing Y, et al. Single-cell landscape and clinical outcomes of infiltrating B cells in colorectal cancer. *Immunology* 2023;168:135–51.
- [31] Lavie D, Ben-Shmuel A, Erez N, Scherz-Shouval R. Cancer-associated fibroblasts in the single-cell era. *Nat Cancer* 2022;3:793–807.
- [32] Dai J, Xi X, Liu Z, Wu W, Zhu S, Zhang X, et al. Single-cell sequencing of multi-region resolves geospatial architecture and therapeutic target of endothelial cells in esophageal squamous cell carcinoma. *Clin Transl Med* 2023;13:e1487.
- [33] Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 2019;20:163–72.
- [34] Li C, Liu B, Kang B, Liu Z, Liu Y, Chen C, et al. SciBet as a portable and fast single cell type identifier. *Nat Commun* 2020;11:1818.
- [35] Zhang Z, Luo D, Zhong X, Choi JH, Ma Y, Wang S, et al. SCINA: a semi-supervised subtyping algorithm of single cells and bulk samples. *Genes (Basel)* 2019;10:531.
- [36] Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, et al. Inference and analysis of cell–cell communication using CellChat. *Nat Commun* 2021;12:1088.
- [37] Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* 2021;49:D605–12.
- [38] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (Camb)* 2021;2:100141.
- [39] Ma L, Zou D, Liu L, Shireen H, Abbasi AA, Bateman A, et al. Database Commons: a catalog of worldwide biological databases. *Genomics Proteomics Bioinformatics* 2023;21:1054–8.