Nucleic Acids Research, 2024, **52**, D859–D870
https://doi.org/10.1093/nar/gkad874
Advance access publication date: 19 October 2023
Database issue

CellSTAR: a comprehensive resource for single-cell transcriptomic annotation

Ying Zhang [1,†], Huaicheng Sun [1,†], Wei Zhang [1,†], Tingting Fu [1,†], Shijie Huang [1], Minjie Mou [1], Jinsong Zhang [1], Jianqing Gao [1], Yichao Ge [1,2,*], Qingxia Yang [3,4,*] and Feng Zhu [1,2,*]

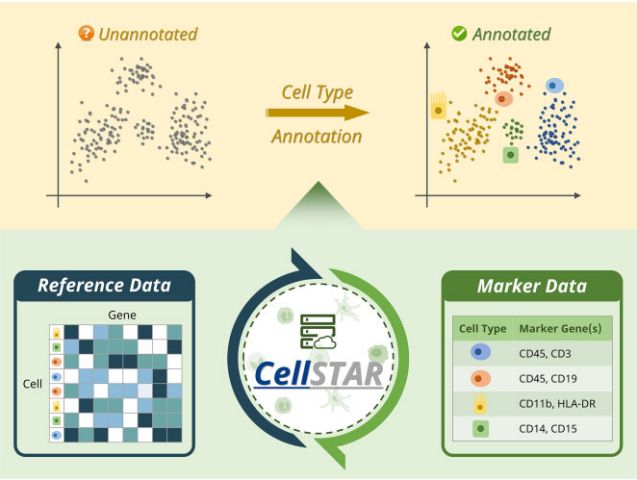
[1]College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou 310058, China
[2]Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Hangzhou 330110, China
[3]Zhejiang Provincial Key Laboratory of Precision Diagnosis and Therapy for Major Gynecological Diseases, Women's Hospital, Zhejiang University School of Medicine, Hangzhou 310058, China
[4]Department of Bioinformatics, School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

[*]To whom correspondence should be addressed. Tel: +86 189 8946 6518; Fax: +86 0571 8820 8444; Email: zhufeng@zju.edu.cn
Correspondence may also be addressed to Qingxia Yang. Email: yangqx@njupt.edu.cn
Correspondence may also be addressed to Yichao Ge. Email: geyichao@zju.edu.cn
[†]The authors wish it to be known that, in their opinion, the first four authors should be regarded as Joint First Authors.

Abstract

Large-scale studies of single-cell sequencing and biological experiments have successfully revealed expression patterns that distinguish different cell types in tissues, emphasizing the importance of studying cellular heterogeneity and accurately annotating cell types. Analysis of gene expression profiles in these experiments provides two essential types of data for cell type annotation: annotated references and canonical markers. In this study, the first comprehensive database of single-cell transcriptomic annotation resource (CellSTAR) was thus developed. It is unique in (a) offering the comprehensive expertly annotated reference data for annotating hundreds of cell types for the first time and (b) enabling the collective consideration of reference data and marker genes by incorporating tens of thousands of markers. Given its unique features, CellSTAR is expected to attract broad research interests from the technological innovations in single-cell transcriptomics, the studies of cellular heterogeneity & dynamics, and so on. It is now publicly accessible without any login requirement at: https://idrblab.org/cellstar.

Graphical abstract



Introduction

With the rapid advances in single-cell RNA sequencing (scRNA-seq), there has been a paradigm shift from 'bulk' to 'single-cell' resolution, highlighting the importance of cellular heterogeneity (1–7). This transition has led to extensive research efforts (>16 600 papers published on PubMed in the last five years) and a surge in large-scale unannotated datasets (Figure 1A), which necessitates accurate identification of cell

Received: August 15, 2023. **Revised:** September 12, 2023. **Editorial Decision:** September 27, 2023. **Accepted:** September 27, 2023

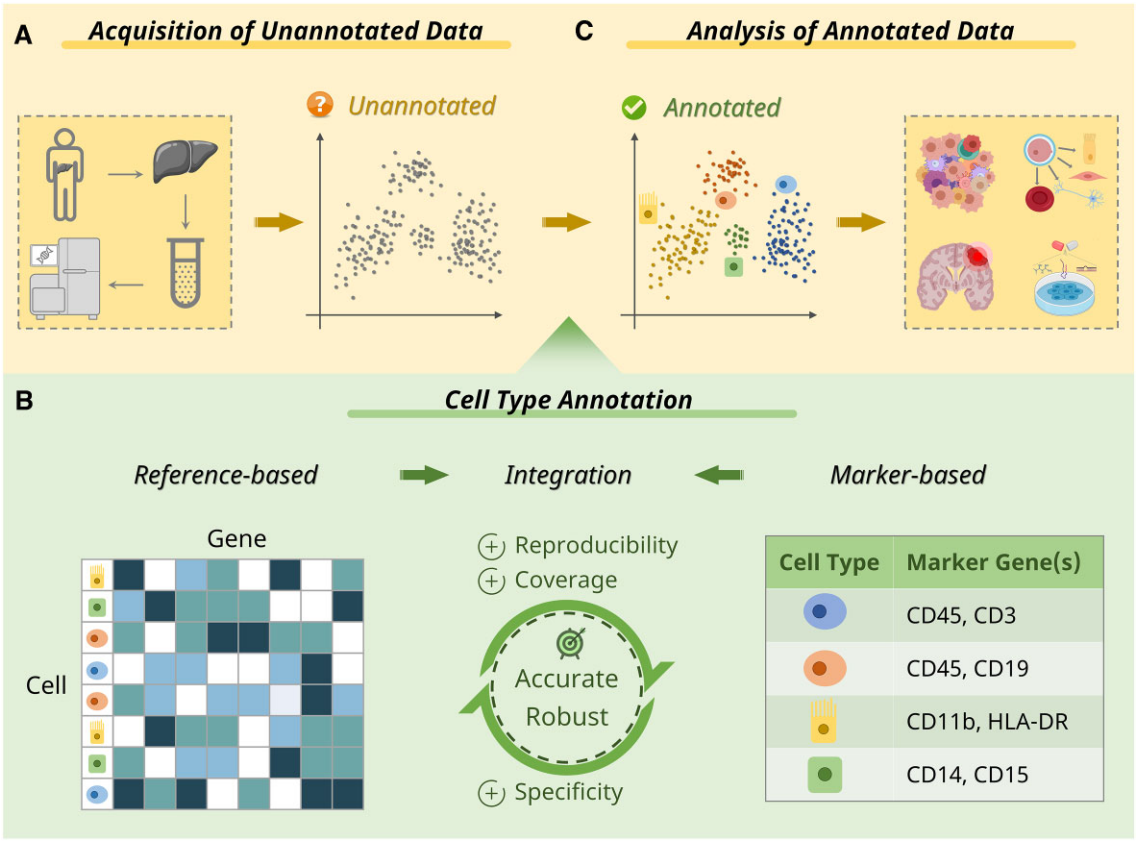


Figure 1. Schematic illustrations of the general workflow of cell type annotation and the features of annotation-related prior data (annotated reference datasets and marker genes) provided by CellSTAR. (**A**) *Acquisition of unannotated data*: the acquisition of large-scale unannotated datasets from single-cell sequencing studies necessitates accurate cell type annotation. (**B**) *Cell type annotation*: unlike the strategy that relies on traditional information of canonical marker genes that are specifically expressed in known cell types, the reference-based annotation strategy utilizes comprehensive gene expression profiles of expertly annotated reference datasets. Due to this feature, it has demonstrated superiority in capturing expression variability and coverage, exhibiting efficiency and reproducibility, and achieving high resolution (described in Supplementary Table S1). Furthermore, the accuracy, reliability and consistency of both annotation strategies heavily depend on the availability, quality and applicability of annotation data, which commonly requires a comprehensive database that integrates curated reference and marker data to achieve abundant availability, high quality, and complementary applicability. (**C**) *Analysis of annotated data*: by enabling collective considerations of both types of data, CellSTAR is expected to facilitate accurate and robust identification of cell identities and various downstream analyses, such as studies of cellular heterogeneity and dynamics, disease research, drug discovery.

identity (i.e. cell type annotation) (8,9).

Cell type annotation has become an essential step for downstream analysis in single-cell studies (10,11), which relies on two types of annotation data: 'reference datasets' (expertly annotated single-cell maps) and 'marker genes' (genes specifically expressed in known cell types) (12). Compared to the traditional marker-based annotation, the reference-based strategy can fully leverage existing expertly annotated references, which has demonstrated superior performance in identifying complex cellular compositions and deciphering cell state transitions (12–15). Furthermore, the integrated application of these two types of data has been advocated in many studies, and its accuracy, reliability, and consistency heavily depend on data comprehensiveness and quality (14,16) (Figure 1B). Therefore, the comprehensive annotation data of both references and markers are urgently needed in current single-cell transcriptomic studies.

So far, several databases related to scRNA-seq cell type annotation have been constructed (17–22). Most of them focus on describing marker genes, such as CellMarker (17), PCMDB (18), CancerSEA (19) and several others (20–22). These databases have attracted extensive interest because they bridge the gap between the availability of differentially expressed genes (DEGs) and the delivery of canonical markers to users. However, none of these databases provides reference data. In other words, there is currently no existing database that provides strictly curated scRNA-seq reference data, let alone the systematic integration of corresponding cell markers. All in all, it is highly demanded to construct a comprehensive scRNA-seq cell type annotation database that integrates both curated reference and marker data.

Herein, a comprehensive database of single-cell transcriptomic annotation resource (named 'CellSTAR') is thus introduced. The latest version of CellSTAR (a) provides expression profiles with valuable underlying annotated references collected from 515 projects and 1679 batches using 14 sequencing techniques, and includes 889 distinct cell types identified by 107 annotation approaches across 18 species and 139 tissues, (b) collects canonical cell markers, which contain >80 000 entries covering over 80% of the cell types identified in 76% of the reference data, (c) describes detailed annotation-related experimental metadata, which is valuable for revealing the impact of experimental factors on annotation and considering appropriate analytical approaches

when utilizing the data and (d) offers various interactive visualizations, enabling a holistic exploration of intricately distributed cells and potential molecular drivers. Moreover, species, tissues, cell types and corresponding markers are standardized based on the latest versions of Taxonomy (23), Uberon (24), Cell Ontology (25) and Entrez Gene (26), respectively.

All in all, CellSTAR is unique in (a) offering the comprehensive expertly annotated reference data for annotating hundreds of cell types for the first time and (b) enabling the collective consideration of reference data and marker genes by incorporating tens of thousands of markers. Given the rapid advances of AI in single-cell omics, our CellSTAR (https://idrblab.org/cellstar) is expected to have significant impacts on single-cell transcriptomic analyses (27–31), such as studies of cellular heterogeneity (32), developmental biology, disease research and drug discovery (33,34) (Figure 1C).

Factual content and data retrieval

Systematic collection, curation and integration of annotation data

In this study, a multi-step collection and curation process was proposed to obtain a diverse set of high-quality reference datasets with reliable annotation information: (i) *Comprehensive literature review*: a thorough literature review was conducted in PubMed, focusing on scRNA-seq studies that provide well-characterized gene expression patterns for specific cell types. Specific keyword combinations such as 'scRNA-seq + cell atlas', 'single cell RNA sequencing + reference data' and 'single cell transcriptomics + reference data' were used, which resulted in 10 426 relevant publications. (ii) *Public repository mining*: to complement the literature-based search strategy, additional data mining was carried out on public datasets with underlying annotations in reputable repositories such as Gene Expression Omnibus (GEO) (35) and Single Cell Expression Atlas (SCEA) (36). By employing the aforementioned keywords in GEO, 25 906 entries were retrieved. For SCEA, data were filtered based on experimental factors using the keyword 'cell type', resulting in 135 relevant experiments. (iii) *Rigorous inspection and verification*: due to factors such as data size, complexity, protection regulations and organizational barriers, sharing annotation information along with expression profiles can be challenging (37). Therefore, rigorous inspection and verification of the publications and experiments identified in the previous steps were conducted to ensure the collection of reliable data. We carefully reviewed the full texts of selected publications and cross-checked relevant supplementary materials, ensuring that only publicly accessible expression profiles associated with confidently confirmed annotations were included (such as those derived from cell sorting, cellular mixing, cell classification, and identification). Subsequently, we extracted comprehensive experimental metadata and assigned them to corresponding entries. (iv) *Data deduplication and recording*: to avoid duplication of entries, experiment entries with identical experimental conditions were merged. Detailed experimental metadata and batch information were provided for each experiment entry. As a result, CellSTAR provided 515 research projects and 1679 experimental batches sequenced by 14 techniques, containing 889 distinct cell types identified by 107 annotation approaches across 18 species and 139 tissues. Particularly, a total of 67 experiments conducted on disease samples (including 36 disease classes defined by WHO ICD-11) and 448 experiments conducted on healthy samples were included.

Given the significance of canonical markers in inferring cell identities, CellSTAR incorporates marker data as a complementary resource to establish an integrated framework that combines the strengths of both types of information. (i) *Standardization and data alignment*: the cell type annotations of our reference datasets were standardized using the latest version of Cell Ontology (25) to ensure consistent terminology across different data sources and further align the reference and marker data. (ii) *Canonical markers acquisition*: expert-validated cell markers were obtained from well-established resources such as CellMarker (17), PanglaoDB (38), CancerSEA (19) and the CD Marker Handbook (22). In some cases, the marker databases might not cover all the cell types present in previously searched reference datasets. Therefore, to ensure comprehensive mutual validation of both data types, we extensively supplemented missing marker information for uncovered cell types by conducting a thorough literature review on Google Scholar. Specifically, we used keyword combinations of marker(s) along with names of respective cell type, species, and tissue. For example, ('marker' OR 'signature') AND ('fat cell' OR 'adipocyte' OR 'adipose cell') AND ('Mus musculus' OR 'mouse') AND ('kidney'). (iii) *Expert review and validation*: moreover, biological researchers and experts participated in the review of candidate markers to extract clinically or experimentally validated marker information, ensuring their specificity, relevance, and underlying biological insights. (iv) *Combination for comprehensive information*: all records of markers derived from databases and publications were merged, resulting in >80 000 entries covering over 80% of the cell types identified in 76% of the reference data. Through these systematic and rigorous steps of data collection, curation, and integration, CellSTAR innovatively offers a comprehensive and reliable framework that combines well-characterized annotated references with well-established canonical markers to improve cell type annotation.

Data accessibility of annotated references with comprehensive metadata

The success of reference-based annotation heavily depends on the availability and selection of appropriate reference datasets that exhibit sufficient similarity to the query dataset (37,39). To address this challenge, CellSTAR provides users with systematically collected and curated reference datasets that were previously scattered across publications, along with comprehensive annotation-related metadata to enable confident utilization of the datasets (40,41).

In the online database, users can retrieve experiments of interest by searching for relevant keywords. Each query will generate a list of relevant experiments, and users can easily select appropriate references based on the provided general information and a word cloud map that illustrates the complex cellular landscape and the abundance of distinct cell populations within the reference data. Furthermore, comprehensive metadata, batch information, and reference datasets for each experiment are accessible. (a) *Experimental metadata*: this section provides a comprehensive description of the studied samples, including species and tissue names/synonyms, and specific disease under investigation. It also incorporates details about the sequencing technology, data preprocessing protocols (quality

control, normalization, transformation, data correction and integration, feature selection and dimensionality reduction), and experimental treatments such as reagent treatments, surgical procedures, gene modifications, feeding regimens and disease progression. The origin of the annotations (whether obtained manually, automatically, through cellular sorting or immunopanning) and corresponding annotation protocols are also provided. Additionally, cross-links with well-established databases such as NCBI Taxonomy (23), Uberon (24), GEO (35) and SCEA are available (36) (Figure 2A). (b) *Batch information*: this section explicitly presents the original designation or label associated with each batch, accompanied by relevant descriptions of their respective samples. By elucidating the distinctive characteristics and composition of different batches, it offers valuable insights into identifying potential batch effects, understanding the impact of batch-related variations on cell type composition, and considering appropriate analytical approaches when utilizing the dataset (42,43) (Figure 2B). (c) *Similar experiments*: although datasets are categorized into separate experiments based on distinct metadata attributes (such as sample sources and experimental conditions), it is essential to integrate and utilize these highly correlated reference datasets in order to enhance annotation accuracy and ensure comprehensive annotation coverage (44–46). Therefore, this section establishes connections among datasets associated with the same literature or common research project, which facilitates navigation through related datasets while demonstrating consistency and comparability in metadata across different experiments (Figure 2C). (d) *File download*: the online database provides two types of data that collectively serve as 'reference data', including an annotation reference file and corresponding expression profiling file(s). The 'annotation reference file' is a csv file that establishes a clear mapping between individual cells and their corresponding annotated cell identities across all batches within the expression profiling file(s) (Figure 2D). Within the annotation reference file, two key columns are crucial for indicating cell identities: 'inferred_cell_type' and 'cell_ontology_class'. The former is provided by the original data authors, which may vary in terminologies and nomenclature conventions, while the latter records standardized designations implemented by us using established Cell Ontology to ensure consistency and reliability. The 'expression profiling file' is a count matrix that captures the raw expression counts for each gene in each single cell, as speculated by the reviewer. All count matrices within CellSTAR are raw count matrices without any normalization or transformation. However, some have undergone quality control by the original data providers to remove low-quality data (e.g. experiment CSTA_000001), and we did not apply any further data operations. In summary, CellSTAR facilitates systematic connections between user queries and reference datasets based on experimental metadata while effectively organizing resulting data for convenient access and download.

Characterization of canonical markers and navigation to associated references

With the inclusion of canonical markers, the online database not only provides access to an extensive collection of annotated references but also offers valuable insights into the molecular signatures of diverse cell types. The user-friendly interface was carefully designed for intuitive navigation and efficient querying of the integrated annotation data for specific cell types, comprising the following key sections: (a) *Cell general information*: this section presents an overview of each cell type, including its name and synonyms, comprehensive descriptions of its unique morphology, function, location and other distinctive features. Additionally, the immediate superiors or higher-level cell types to which the specific cell type belongs are provided. To facilitate a deeper understanding of hierarchical relationships with various cell types, external links to the Cell Ontology Lookup Service of EMBL-EBI are provided (47) (Figure 3A); (b) *Cell-related experiment(s)*: in this section, users can access a comprehensive compilation of reference data involving the specific cell type of interest. In addition to general experimental metadata, and in-depth information can be accessed by following corresponding hyperlinks provided in the 'Details' column (Figure 3B); (c) *Cell-related canonical marker(s)*: this section lists curated canonical cell markers of various species and tissues. For each marker, essential details which include marker name, gene symbol, gene type and the protein encoded by the marker gene are carefully documented. Moreover, direct access to relevant publications through provided hyperlinks is offered for thorough investigation into these markers (Figure 3C).

Visualizations for exploring cell populations and molecular signatures

CellSTAR offers various interactive visualizations that facilitate in-depth exploration of intricately distributed cells and potential molecular drivers. Thorough analysis was conducted on each expression profile, leading to the following visualizations: (a) *Visualization of annotated cell populations*: The pie chart presents an overview of the relative abundance of diverse cell populations within each reference dataset, providing users with rapid insights into the distribution and composition of cell populations (48). Meanwhile, the tSNE map intuitively represents intricate spatial relationships among clusters based on their gene expression profiles (49) (Figure 4A). (b) *Heatmap of cell population abundance*: this heatmap empowers users to visually compare the distribution and variability of cell populations across different experimental batches, thereby enabling the discovery of potential functional significance associated with distinct cell types within each batch (Figure 4B). (c) *Heatmap of top genes*: This heatmap vividly depicts the expression patterns of top DEGs across various cell populations, facilitating the identification of potential molecular drivers underlying complex biological processes (50) (Figure 4C). Specifically, DEGs for each of the annotated cell types in a reference dataset were identified based on the Wilcoxon Rank Sum test. As recognized in the scientific community, high-quality reference datasets hold significant value in deciphering expression patterns and serve as foundational resources for identifying molecular signatures. Moreover, it is essential to cross-validate markers across multiple datasets, especially when data originate from different experimental conditions or technologies (51,52). This ensures mutual validation and enhances the robustness and reliability of annotation outcomes. Therefore, DEGs derived from reference datasets are made available for users to compare these candidate markers across datasets or with existing canonical marker information. Taking the reference dataset (CSTA_000192) sequenced by Microwell-seq in mouse liver as an example, 60% of cell types whose DEGs overlap with existing canonical markers by more than 50%, indicating its high quality and potential for

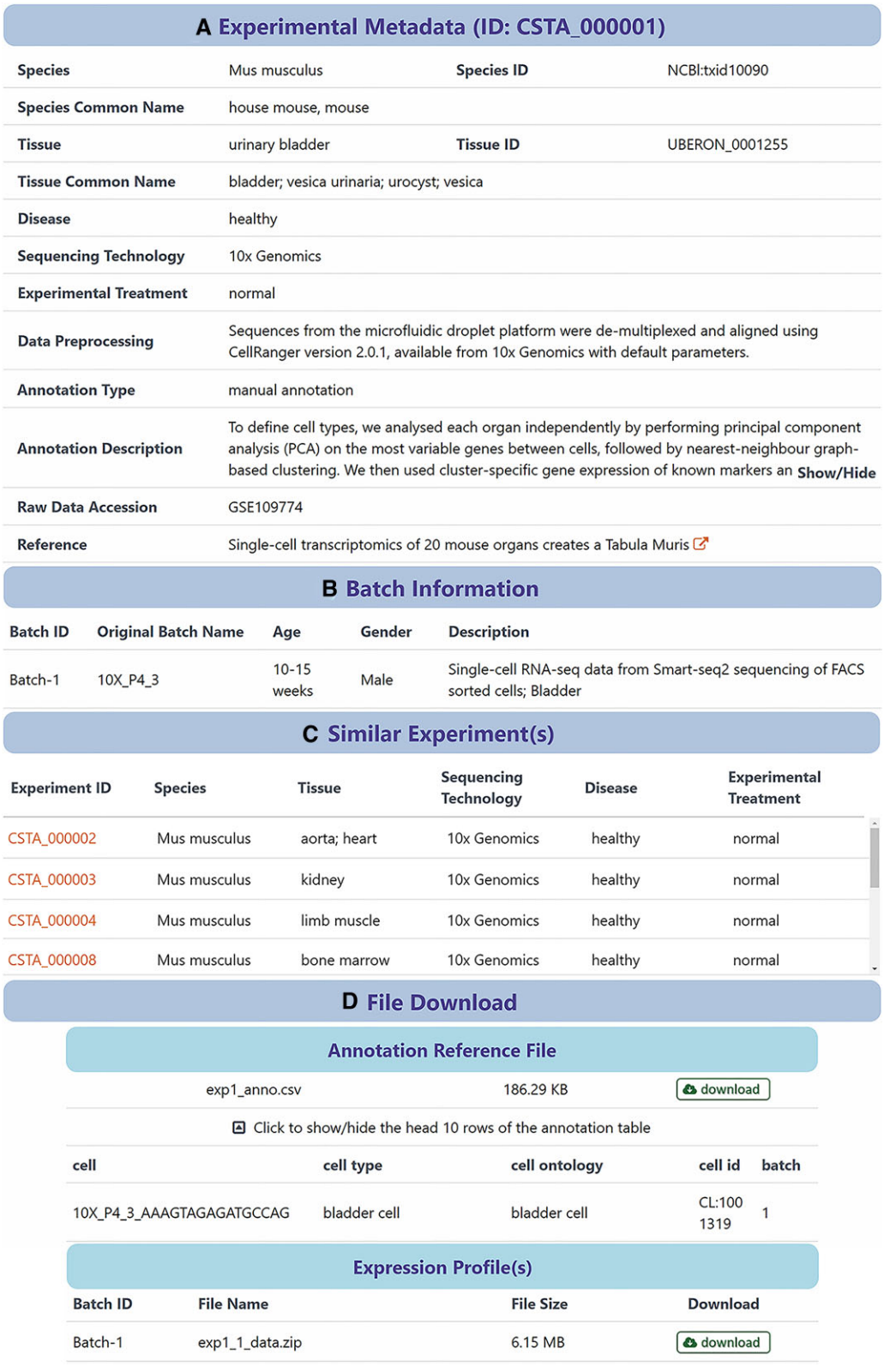


Figure 2. Detailed descriptions of annotation-related metadata and experimental batches for each curated reference dataset (using the CSTA_000001 as an example). (**A**) The comprehensive experimental metadata includes descriptions of the studied samples (species and tissue names/synonyms, and specific disease under investigation), sequencing technology, data preprocessing protocols, experimental treatments, the origin of annotations, corresponding annotation protocols, and external links to other molecular biological databases. (**B**) The batch information explicitly elucidates the distinctive characteristics and compositions of different batches within the study. (**C**) This section establishes connections among datasets associated with the same literature or common research project, which facilitates navigation through related datasets while demonstrating consistency and comparability in metadata across different experiments. (**D**) Well-organized annotated references are available for download, establishing a clear mapping between individual cells and their corresponding annotated cell identities across all batches within the expression profiling file(s). Although this figure presents only a single batch and its associated expression profiling file due to space limitations, complete information can be accessed through the online database.

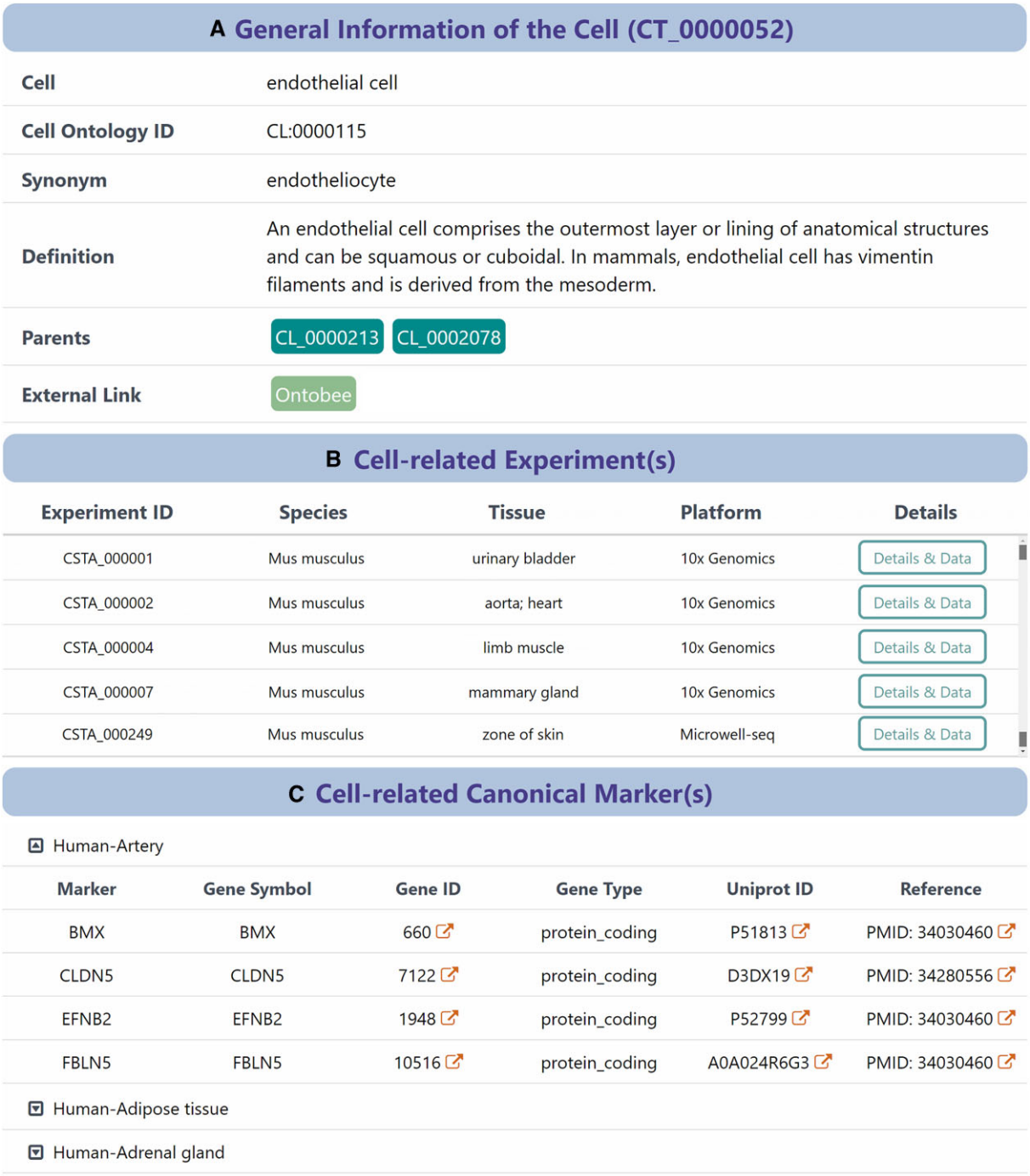


Figure 3. Characterizations of canonical cell markers and associated reference data for each cell type (using the endothelial cell as an example). (**A**) General information of the specific cell type includes cell name and synonyms, distinctive features (such as morphology, function, location), its hierarchical relationships with various cell types, and external links to the Cell Ontology Lookup Service of EMBL-EBI for through exploration. (**B**) In this section, an overview of all reference data involving the specific cell type of interest is presented, with in-depth information of each study accessible through hyperlinks provided in the 'Details' column. (**C**) A comprehensive list of canonical cell markers of various species and tissues are categorized. For each marker, essential details including marker name, gene symbol, gene type, and the protein encoded by the marker gene are carefully documented, and direct access to relevant publications are offered for further investigation.

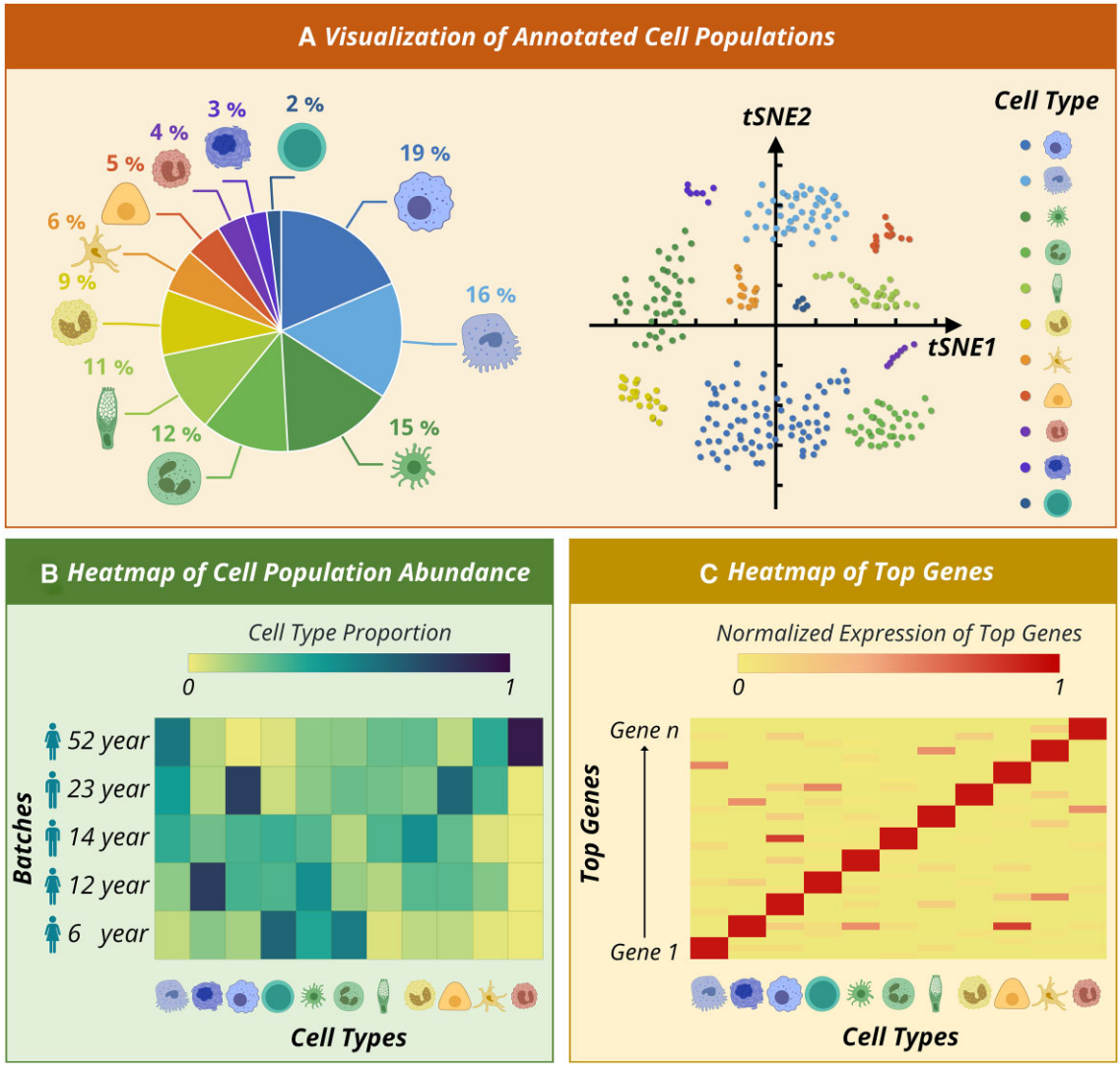


Figure 4. Diverse visualizations offered by CellSTAR for exploration of cell populations and molecular signatures. (**A**) *Visualization of annotated cell populations*: the pie chart presents an overview of the relative abundance of diverse cell populations within each reference dataset, providing users with rapid insights into the distribution and composition of cell populations. Meanwhile, the tSNE map intuitively represents intricate spatial relationships among clusters based on their gene expression profiles. (**B**) *Heatmap of cell population abundance*: this heatmap empowers users to visually compare the distribution and variability of cell populations across different experimental batches, thereby discovering potential functional significance associated with distinct cell types within each batch. (**C**) *Heatmap of top genes*: this heatmap vividly depicts the expression patterns of top DEGs across various cell populations, facilitating the identification of potential molecular drivers underlying complex biological processes.

annotation. In summary, CellSTAR's powerful visualizations not only contribute to the identification of cell types, but also hold great promise in discovering potential biomarkers and elucidating intricate molecular mechanisms.

Applications of CellSTAR for comprehensive and robust annotation

The term 'reference data' typically refers to the existing expertly annotated single-cell maps that may contain various biological factors (e.g. age, sex, and disease condition) (12). When annotating query datasets, one preferred approach is to reference annotated single-cell datasets that are relevant to the specific context (i.e. reference data, not limited to healthy or disease experiments). Therefore, CellSTAR incorporated both types of experiments in order to provide a comprehensive resource for users. Healthy experiments serve as references for understanding the typical cell types and their gene expression profiles within healthy organisms or tissues, which are essential for establishing a foundational framework against which to compare and annotate cell types across various experimental conditions. Meanwhile, there has been a shift in focus towards comparative analyses across different diseases or experimental conditions, as the number of samples per study increases. Such comparative analyses are critical for understanding disease pathogenesis, identifying biomarkers and potential therapeutic targets. Thus, in some cases, researchers may conduct experiments using samples obtained from individuals with specific diseases or conditions. These annotated datasets are valuable for annotating cell types in query datasets acquired under similar conditions, characterizing disease-specific cell types, identifying disease-specific marker genes, and investigating the impact of diseases on cellular landscape. Furthermore, these data allow for meta-

analyses across multiple single-cell maps to clarify cell type differences that correlate with disease severity and response to various therapeutic treatments.

Here, we present a comprehensive exploration of various annotation applications of CellSTAR reference datasets and its flexibility in selecting appropriate strategies for specific application scenarios (46,53–58). CellSTAR allows for the association of gene expression profiles between reference datasets and unknown cells in query datasets through similarity measurement, data integration, or (semi-)supervised classification. The most straightforward strategy is similarity measurement (Figure 5A), which has been implemented in tools such as scmap (59), scMatch (60), CHETAH (61), CIPR (62) and clustifyr (63). These tools focus on quantifying the similarity between gene expression profiles of reference and query datasets, enabling the mapping of unknown cells to known cell types. For the application scenario involving the integration of query data with reference data (Figure 5B), batch effects are an important concern as they can significantly impact cell type annotation based on integrated datasets. To deal with such interference while preserving the biological signals of interest, several tools have been proposed (42,64). For example, Harmony projects both query and reference data into a lower-dimensional space followed by soft clustering to generate multiple clusters. Correction factors are then computed based on cluster centroids and iteratively applied to minimize batch effects within the integrated dataset. This approach enables the identification of clusters spanning both datasets and facilitates transfer of reference labels to query cells within these clusters. Although this approach supports the identification of distinct cell types and gradients in cell state, it can be computationally expensive (65). Despite existing tools of either similarity measurement or data integration are based on '*a single*' reference dataset, they allow the merging of '*multiple*' references to achieve comprehensive annotation. Ultimately, the (semi-)supervised classification strategy (Figure 5C) offers a more valid alternative when '*sufficient*' reference datasets with meaningful features and cell identities are available. This approach effectively overcomes intrinsic experimental noise and variability in multiple datasets (66–70). Overall, the diverse reference data available in CellSTAR can be readily applied to these annotation tools either directly or with minor adjustments, depending on the availability of appropriate reference datasets, the specific research being conducted, and the computational and expertise resources available.

Besides the discussed flexibility in utilizing reference data, CellSTAR's marker data can be used for expert manual annotation as well as in automated annotation tools such as CellAssign (11), Garnett (70), ScType (51), scCATCH (22) and CALLR (71). Moreover, further potential applications based on the integrated considerations of reference and marker data can be explored. A number of benchmarking studies have demonstrated the complementary nature of these methods, each possessing distinct advantages (14). Therefore, it is highly recommended to combine multiple tools that integrate both reference and marker data in practical applications. This approach establishes cross-validation and enables majority voting, which significantly enhances annotation accuracy (72,73). By leveraging the synergy between reference and marker data, CellSTAR holds the potential to advance its capabilities and provide more accurate and comprehensive annotations.

Standardization, access and retrieval of CellSTAR data

Due to the inherent biases introduced during the assignment process, which involves the application of uncontrolled vocabularies for cell type labeling across different datasets, as well as the inconsistencies observed in canonical marker genes across various databases, the utilization of such resources can potentially lead to divergent annotation outcomes (74). In other words, to ensure reliable and consistent utilization of CellSTAR data, it is essential to mitigate potential variations arising from distinct terminologies and nomenclatures across different sources. Therefore, careful data cleaning and systematic standardization were conducted on the collected raw data. The standardization process relied on the latest versions of the Taxonomy (23), Uberon (24), Cell Ontology (25), Entrez Gene (26) and WHO ICD-11 for accurate alignment of species, tissues, cell types, cell markers and diseases respectively.

Moreover, a user-friendly interface was thoughtfully designed to enable intuitive navigation, efficient retrieval, and convenient downloading of the data. The browsing functionality of CellSTAR can greatly assist users in exploring reference data and identifying experiments relevant to specific species or tissues. To further expedite the data exploration process, a quick search utility is integrated into CellSTAR, enabling users to efficiently search through the entire annotation data. This can be achieved either through the main search interface or the pull-down menu options. On the 'Download' page, annotated reference datasets with experimental metadata and resulting analysis data, as well as corresponding canonical cell markers are readily available. Specifically, to enhance user convenience and improve interoperability when utilizing reference datasets, well-organized expression profiles have been converted into universally accepted formats (CSV, TSV and MTX) compatible with various single-cell annotation and analysis tools, including but not limited to SingleR (15), scDeepSort (13), and harmony (65). The CellSTAR can be freely accessed by all users without any login requirements at https://idrblab.org/cellstar, and compatibility testing has been performed across popular web browsers including Microsoft Edge, Google Chrome, Apple Safari and Mozilla Firefox.

Conclusion and prospect

In this study, we developed CellSTAR, a manually curated resource that comprehensively integrates annotated references and canonical markers across diverse species. It is the first database to provide experiment-based reference data for annotating hundreds of cell types. Such valuable data have demonstrated superior performance in identifying complex cellular compositions and deciphering cell state transitions compared to the traditional information of marker genes that are specifically expressed in known cell types. Moreover, tens of thousands of markers are also incorporated into CellSTAR to enable collective consideration of reference and marker data. Notably, various visualizations are provided to facilitate in-depth exploration of intricately distributed cells and potential molecular drivers based on thorough analysis of each expression profile. In summary, CellSTAR will be an informative and valuable resource for researchers aiming to accelerate

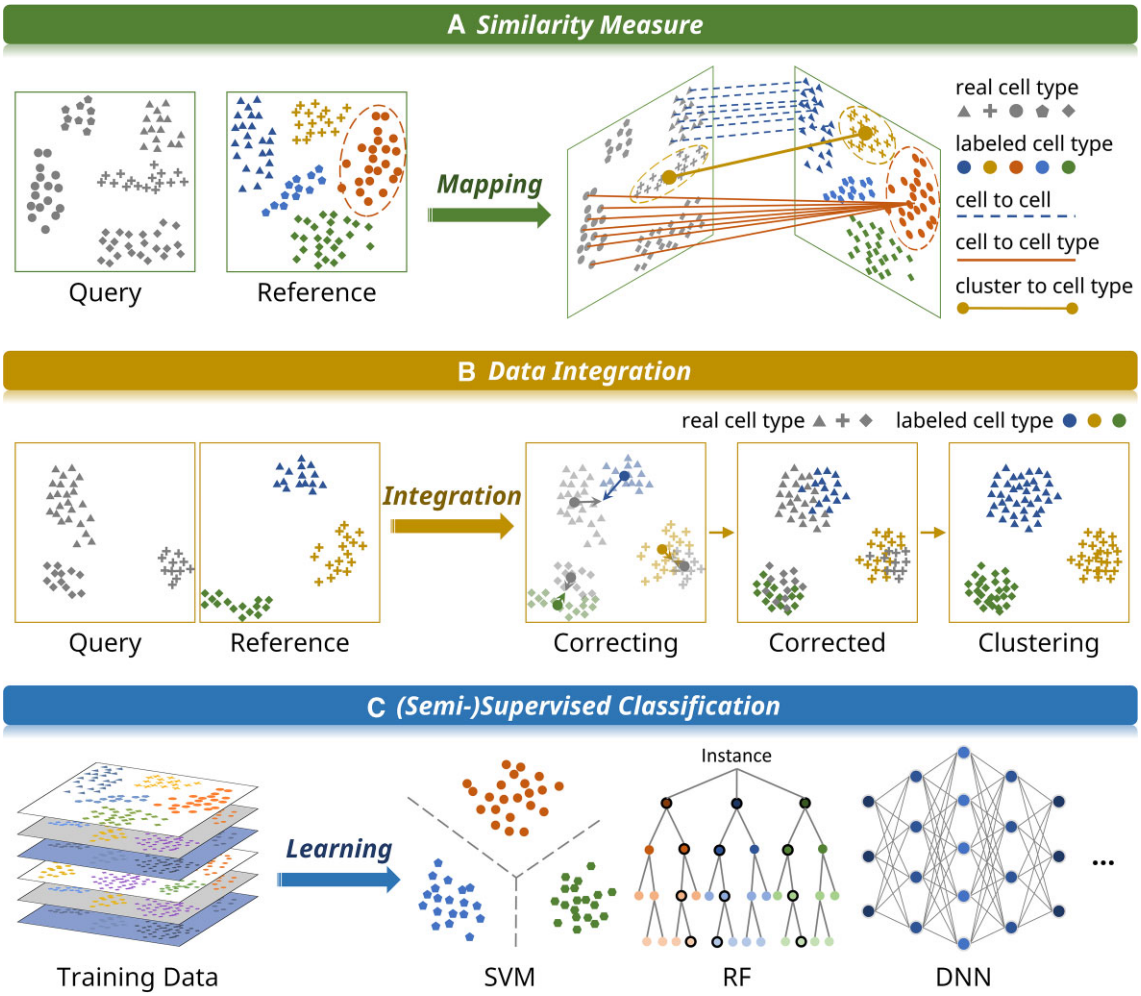


Figure 5. Schematic depictions of three widely used annotation strategies that can be employed based on CellSTAR reference data. (**A**) *Similarity measurement*: it involves quantifying the similarity between gene expression profiles of reference and query datasets, which facilitates the mapping of unknown cells or clusters to known cell types. (**B**) *Data integration*: by integrating the query dataset with a reference dataset, this strategy enables the identification of clusters that span both datasets and allows for transferring reference labels to query cells within these clusters. (***C***) *(Semi-)supervised classification*: it refers to the training of supervised or semi-supervised classifiers on the distribution of cell type labels, in terms of a defined set of features within annotated datasets. These trained models are subsequently utilized to assign labels to cells within unlabeled datasets based on their relative features.

their investigations into distinct cell types, and pave the way for groundbreaking discoveries in cell biology.

As the field of single-cell transcriptomics rapidly advances, reference atlases, new experimental technologies, annotation strategies and tools are continuously being developed to improve our ability to interpret, annotate and validate cellular landscapes. Therefore, further extensions will be conducted in the following aspects. First, reference datasets of more species and tissues, and an expanded repository of cell markers for diverse cell types (particularly rare/transitive cell types), will be collected and updated in the future version of CellSTAR to provide more comprehensive information for users. Second, analysis algorithms and tools will be integrated into CellSTAR to facilitate the extraction of meaningful insights from complex datasets. Third, contextual categorization of studies involving multiple experimental conditions will be emphasized, as the focus of single-cell map interpretation is gradually shifting to comparisons across disease, age or other conditions. Finally, it is crucial to acknowledge the inherent limitations of relying solely on reference datasets and marker genes in the current version of CellSTAR, which was initially designed to address the urgent demand for a comprehensive and curated resource for cell type annotation using scRNA-seq data. The challenge arises when attempting to annotate homogeneous or closely related cell types or states, as they often exhibit significant overlap in their expression patterns. In other words, subtle distinctions between cell subtypes may not be detectable at the transcriptional level alone and may require additional complementary genomic layers, such as epigenetic information (75,76) (e.g. single-cell assay for transposase accessible-chromatin with high-throughput sequencing (77,78) and single-cell DNA methylome sequencing (79)). Moreover, continuous advancements in experimental technologies enable the measurement of multiple modalities at the single-cell level (80), which are expected to achieve more accurate and comprehensive cell type annotations and enhance our understanding of complex multicellular systems. For instance, spatial transcriptomics combines cell imaging and scRNA-seq to capture spatial transcript patterns and cellular morphology within a single experiment (81); cellular indexing of tran-

scriptomes and epitopes by sequencing enables simultaneous immunophenotyping of cell surface proteins and scRNA-seq measurements. Overall, considering the valuable insights provided by these diverse data types, we expect expansions of these data into our database and explorations of their integration in future versions based on evolving needs of the single-cell research community (82).

Data availability

All data can be viewed, accessed and downloaded from the CellSTAR, which is freely accessible without any login requirements by all users at: https://idrblab.org/cellstar.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

Schematic illustration was created with the assistance of Medpeer.com.

Funding

National Natural Science Foundation of China [62201289, 82373790, 22220102001, U1909208, 81872798, 81971982]; Natural Science Foundation of Zhejiang Province [LR21H300001]; Leading Talents of 'Ten Thousand Plan' National High-Level Talents Special Support Plan of China; 'Double Top-Class' University Projects [181201*194232101]; Fundamental Research Funds for Central Universities [2018QNA7023]; Westlake Laboratory (Westlake Laboratory of Life Sciences and Biomedicine); Key R&D Programs of Zhejiang Province [2020C03010]; National Key R&D Program of China [2022YFC3400501]; Natural Science Foundation of Jiangsu Province [BK20210597]; Information Technology Centers of Zhejiang University; Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare; Alibaba Cloud. Funding for open access charge: Natural Science Foundation of Jiangsu Province [BK20210597].

Conflict of interest statement

None declared.

References

1. Stuart,T., Butler,A., Hoffman,P., Hafemeister,C., Papalexi,E., Mauck,W.M. 3rd, Hao,Y., Stoeckius,M., Smibert,P. and Satija,R. (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.
2. Papalexi,E. and Satija,R. (2018) Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.*, **18**, 35–45.
3. Klein,A.M., Mazutis,L., Akartuna,I., Tallapragada,N., Veres,A., Li,V., Peshkin,L., Weitz,D.A. and Kirschner,M.W. (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**, 1187–1201.
4. Bhattacharya,M. and Ramachandran,P. (2023) Immunology of human fibrosis. *Nat. Immunol.*, **9**, 1423–1433.
5. Bhattacharyya,A., Torre,P., Yadav,P., Boostanpour,K., Chen,T.Y., Tsukui,T., Sheppard,D., Muramatsu,R., Seed,R.I., Nishimura,S.L., *et al.* (2022) Macrophage Cx43 is necessary for fibroblast cytosolic calcium and lung fibrosis after injury. *Front. Immunol.*, **13**, 880887.
6. Jia,P., Hu,R., Yan,F., Dai,Y. and Zhao,Z. (2022) scGWAS: landscape of trait-cell type associations by integrating single-cell transcriptomics-wide and genome-wide association studies. *Genome Biol.*, **23**, 220.
7. Han,Q.L., Zhang,X.L., Ren,P.X., Mei,L.H., Lin,W.H., Wang,L., Cao,Y., Li,K. and Bai,F. (2023) Discovery, evaluation and mechanism study of WDR5-targeted small molecular inhibitors for neuroblastoma. *Acta Pharmacol. Sin.*, **44**, 877–887.
8. Pei,G., Yan,F., Simon,L.M., Dai,Y., Jia,P. and Zhao,Z. (2022) deCS: a tool for systematic cell type annotations of single-cell RNA sequencing data among human tissues. *Genomics Proteomics Bioinformatics*, **1**, 1–5.
9. Dai,Y., Hu,R., Manuel,A.M., Liu,A., Jia,P. and Zhao,Z. (2021) CSEA-DB: an omnibus for human complex trait and cell type associations. *Nucleic Acids Res.*, **49**, D862–D870.
10. Miao,Z., Moreno,P., Huang,N., Papatheodorou,I., Brazma,A. and Teichmann,S.A. (2020) Putative cell type discovery from single-cell gene expression data. *Nat. Methods*, **17**, 621–628.
11. Zhang,A.W., O'Flanagan,C., Chavez,E.A., Lim,J.L.P., Ceglia,N., McPherson,A., Wiens,M., Walters,P., Chan,T., Hewitson,B., *et al.* (2019) Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat. Methods*, **16**, 1007–1015.
12. Clarke,Z.A., Andrews,T.S., Atif,J., Pouyabahar,D., Innes,B.T., MacParland,S.A. and Bader,G.D. (2021) Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nat. Protoc.*, **16**, 2749–2764.
13. Shao,X., Yang,H., Zhuang,X., Liao,J., Yang,P., Cheng,J., Lu,X., Chen,H. and Fan,X. (2021) scDeepSort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. *Nucleic Acids Res.*, **49**, e122.
14. Abdelaal,T., Michielsen,L., Cats,D., Hoogduin,D., Mei,H., Reinders,M.J.T. and Mahfouz,A. (2019) A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.*, **20**, 194.
15. Aran,D., Looney,A.P., Liu,L., Wu,E., Fong,V., Hsu,A., Chak,S., Naikawadi,R.P., Wolters,P.J., Abate,A.R., *et al.* (2019) Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.*, **20**, 163–172.
16. Ma,W., Su,K. and Wu,H. (2021) Evaluation of some aspects in supervised cell type identification for single-cell RNA-seq: classifier, feature selection, and reference construction. *Genome Biol.*, **22**, 264.
17. Hu,C., Li,T., Xu,Y., Zhang,X., Li,F., Bai,J., Chen,J., Jiang,W., Yang,K., Ou,Q., *et al.* (2023) 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.*, **51**, D870–D876.
18. Jin,J., Lu,P., Xu,Y., Tao,J., Li,Z., Wang,S., Yu,S., Wang,C., Xie,X., Gao,J., *et al.* (2022) PCMDB: a curated and comprehensive resource of plant cell markers. *Nucleic Acids Res.*, **50**, D1448–D1455.
19. Yuan,H., Yan,M., Zhang,G., Liu,W., Deng,C., Liao,G., Xu,L., Luo,T., Yan,H., Long,Z., *et al.* (2019) CancerSEA: a cancer single-cell state atlas. *Nucleic Acids Res.*, **47**, D900–D908.
20. Schmitz,R., Wright,G.W., Huang,D.W., Johnson,C.A., Phelan,J.D., Wang,J.Q., Roulland,S., Kasbekar,M., Young,R.M., Shaffer,A.L., *et al.* (2018) Genetics and pathogenesis of diffuse large B-cell lymphoma. *N. Engl. J. Med.*, **378**, 1396–1407.
21. Liberzon,A., Birger,C., Thorvaldsdottir,H., Ghandi,M., Mesirov,J.P. and Tamayo,P. (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.*, **1**, 417–425.
22. Shao,X., Liao,J., Lu,X., Xue,R., Ai,N. and Fan,X. (2020) scCATCH: automatic annotation on cell types of clusters from single-cell RNA sequencing data. *iScience*, **23**, 100882.

23. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
24. Mungall,C.J., Torniai,C., Gkoutos,G.V., Lewis,S.E. and Haendel,M.A. (2012) Uberon, an integrative multi-species anatomy ontology. *Genome Biol.*, **13**, R5.
25. Osumi-Sutherland,D., Xu,C., Keays,M., Levine,A.P., Kharchenko,P.V., Regev,A., Lein,E. and Teichmann,S.A. (2021) Cell type ontologies of the Human Cell Atlas. *Nat. Cell Biol.*, **23**, 1129–1135.
26. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
27. Cortese,R., Adams,T.S., Cataldo,K.H., Hummel,J., Kaminski,N., Kheirandish-Gozal,L. and Gozal,D. (2023) Single-cell RNA-seq uncovers cellular heterogeneity and provides a signature for paediatric sleep apnoea. *Eur. Respir. J.*, **61**, 1.
28. Cheng,C., Easton,J., Rosencrance,C., Li,Y., Ju,B., Williams,J., Mulder,H.L., Pang,Y., Chen,W. and Chen,X. (2019) Latent cellular analysis robustly reveals subtle diversity in large-scale single-cell RNA-seq data. *Nucleic Acids Res.*, **47**, e143.
29. Johnson,T.S., Yu,C.Y., Huang,Z., Xu,S., Wang,T., Dong,C., Shao,W., Zaid,M.A., Huang,X., Wang,Y., *et al.* (2022) Diagnostic evidence GAuge of single cells (DEGAS): a flexible deep transfer learning framework for prioritizing cells in relation to disease. *Genome Med.*, **14**, 11.
30. Li,F., Hu,Q., Zhang,X., Sun,R., Liu,Z., Wu,S., Tian,S., Ma,X., Dai,Z., Yang,X., *et al.* (2022) DeepPROTACs is a deep learning-based targeted degradation predictor for PROTACs. *Nat. Commun.*, **13**, 7133.
31. Wang,L., Wu,Y., Yao,S., Ge,H., Zhu,Y., Chen,K., Chen,W.Z., Zhang,Y., Zhu,W., Wang,H.Y., *et al.* (2022) Discovery of potential small molecular SARS-CoV-2 entry blockers targeting the spike protein. *Acta Pharmacol. Sin.*, **43**, 788–796.
32. Zou,Q., Mao,Y., Hu,L., Wu,Y. and Ji,Z. (2014) miRClassify: an advanced web server for miRNA family classification and annotation. *Comput. Biol. Med.*, **45**, 157–160.
33. Huang,L.H., He,Q.S., Liu,K., Cheng,J., Zhong,M.D., Chen,L.S., Yao,L.X. and Ji,Z.L. (2018) ADReCS-target: target profiles for aiding drug safety research and application. *Nucleic Acids Res.*, **46**, D911–D917.
34. Cai,M.C., Xu,Q., Pan,Y.J., Pan,W., Ji,N., Li,Y.B., Jin,H.J., Liu,K. and Ji,Z.L. (2015) ADReCS: an ontology database for aiding standardization and hierarchical classification of adverse drug reaction terms. *Nucleic Acids Res.*, **43**, D907–D913.
35. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M., *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
36. Moreno,P., Fexova,S., George,N., Manning,J.R., Miao,Z., Mohammed,S., Munoz-Pomer,A., Fullgrabe,A., Bi,Y., Bush,N., *et al.* (2022) Expression Atlas update: gene and protein expression in multiple species. *Nucleic Acids Res.*, **50**, D129–D140.
37. Lotfollahi,M., Naghipourfar,M., Luecken,M.D., Khajavi,M., Buttner,M., Wagenstetter,M., Avsec,Z., Gayoso,A., Yosef,N., Interlandi,M., *et al.* (2022) Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.*, **40**, 121–130.
38. Franzen,O., Gan,L.M. and Bjorkegren,J.L.M. (2019) PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)*, **2019**, 1.
39. Heumos,L., Schaar,A.C., Lance,C., Litinetskaya,A., Drost,F., Zappia,L., Lucken,M.D., Strobl,D.C., Henao,J., Curion,F., *et al.* (2023) Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.*, **24**, 550–572.
40. Skinnider,M.A., Squair,J.W. and Courtine,G. (2021) Enabling reproducible re-analysis of single-cell data. *Genome Biol.*, **22**, 215.
41. Puntambekar,S., Hesselberth,J.R., Riemondy,K.A. and Fu,R. (2021) Cell-level metadata are indispensable for documenting single-cell sequencing datasets. *PLoS Biol.*, **19**, e3001077.
42. Haghverdi,L., Lun,A.T.L., Morgan,M.D. and Marioni,J.C. (2018) Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.*, **36**, 421–427.
43. Chen,J., Xu,H., Tao,W., Chen,Z., Zhao,Y. and Han,J.J. (2023) Transformer for one stop interpretable cell type annotation. *Nat. Commun.*, **14**, 223.
44. Liu,H., Li,H., Sharma,A., Huang,W., Pan,D., Gu,Y., Lin,L., Sun,X. and Liu,H. (2023) scAnno: a deconvolution strategy-based automatic cell type annotation tool for single-cell RNA-sequencing data sets. *Brief. Bioinform*, **24**, 1.
45. Duan,B., Chen,S., Chen,X., Zhu,C., Tang,C., Wang,S., Gao,Y., Fu,S. and Liu,Q. (2021) Integrating multiple references for single-cell assignment. *Nucleic Acids Res.*, **49**, e80.
46. Pasquini,G., Rojo Arias,J.E., Schafer,P. and Busskamp,V. (2021) Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.*, **19**, 961–969.
47. Ong,E., Xiang,Z., Zhao,B., Liu,Y., Lin,Y., Zheng,J., Mungall,C., Courtot,M., Ruttenberg,A. and He,Y. (2017) Ontobee: a linked ontology data server to support ontology term dereferencing, linkage, query and integration. *Nucleic Acids Res.*, **45**, D347–D352.
48. Wilson,N.K., Kent,D.G., Buettner,F., Shehata,M., Macaulay,I.C., Calero-Nieto,F.J., Sanchez Castillo,M., Oedekoven,C.A., Diamanti,E., Schulte,R., *et al.* (2015) Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations. *Cell Stem Cell*, **16**, 712–724.
49. Kobak,D. and Berens,P. (2019) The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.*, **10**, 5416.
50. Wang,X., Liu,C., Chen,J., Chen,L., Ren,X., Hou,M., Cui,X., Jiang,Y., Liu,E., Zong,Y., *et al.* (2022) Single-cell dissection of remodeled inflammatory ecosystem in primary and metastatic gallbladder carcinoma. *Cell Discov.*, **8**, 101.
51. Ianevski,A., Giri,A.K. and Aittokallio,T. (2022) Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat. Commun.*, **13**, 1246.
52. Li,R., Zhang,J. and Li,Z. (2023) EasyCellType: marker-based cell-type annotation by automatically querying multiple databases. *Bioinform. Adv.*, **3**, vbad029.
53. Wang,X., Xu,Z., Hu,H., Zhou,X., Zhang,Y., Lafyatis,R., Chen,K., Huang,H., Ding,Y., Duerr,R.H., *et al.* (2022) SECANT: a biology-guided semi-supervised method for clustering, classification, and annotation of single-cell multi-omics. *PNAS Nexus*, **1**, pgac165.
54. Xu,Z., Heidrich-O'Hare,E., Chen,W. and Duerr,R.H. (2022) Comprehensive benchmarking of CITE-seq versus DOGMA-seq single cell multimodal omics. *Genome Biol.*, **23**, 135.
55. Tabib,T., Huang,M., Morse,N., Papazoglou,A., Behera,R., Jia,M., Bulik,M., Monier,D.E., Benos,P.V., Chen,W., *et al.* (2021) Myofibroblast transcriptome indicates SFRP2(hi) fibroblast progenitors in systemic sclerosis skin. *Nat. Commun.*, **12**, 4384.
56. Shi,Y., Zhang,X., Yang,Y., Cai,T., Peng,C., Wu,L., Zhou,L., Han,J., Ma,M., Zhu,W., *et al.* (2023) D3CARP: a comprehensive platform with multiple-conformation based docking, ligand similarity search and deep learning approaches for target prediction and virtual screening. *Comput. Biol. Med.*, **164**, 107283.
57. Han,J., Liu,T., Zhang,X., Yang,Y., Shi,Y., Li,J., Ma,M., Zhu,W., Gong,L. and Xu,Z. (2022) D3AI-Spike: a deep learning platform for predicting binding affinity between SARS-CoV-2 spike receptor binding domain with multiple amino acid mutations and human angiotensin-converting enzyme 2. *Comput. Biol. Med.*, **151**, 106212.
58. Wu,L., Zhou,L., Mo,M., Liu,T., Wu,C., Gong,C., Lu,K., Gong,L., Zhu,W. and Xu,Z. (2022) SARS-CoV-2 Omicron RBD shows weaker binding affinity than the currently dominant Delta variant to human ACE2. *Signal Transduct Target Ther.*, **7**, 8.

59. Kiselev,V.Y., Yiu,A. and Hemberg,M. (2018) scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods*, **15**, 359–362.
60. Hou,R., Denisenko,E. and Forrest,A.R.R. (2019) scMatch: a single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics*, **35**, 4688–4695.
61. de Kanter,J.K., Lijnzaad,P., Candelli,T., Margaritis,T. and Holstege,F.C.P. (2019) CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res.*, **47**, e95.
62. Ekiz,H.A., Conley,C.J., Stephens,W.Z. and O'Connell,R.M. (2020) CIPR: a web-based R/shiny app and R package to annotate cell clusters in single cell RNA sequencing experiments. *BMC Bioinf.*, **21**, 191.
63. Fu,R., Gillen,A.E., Sheridan,R.M., Tian,C., Daya,M., Hao,Y., Hesselberth,J.R. and Riemondy,K.A. (2020) clustifyr: an R package for automated single-cell RNA sequencing cluster classification. *F1000Res*, **9**, 223.
64. Welch,J.D., Kozareva,V., Ferreira,A., Vanderburg,C., Martin,C. and Macosko,E.Z. (2019) Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, **177**, 1873–1887.
65. Korsunsky,I., Millard,N., Fan,J., Slowikowski,K., Zhang,F., Wei,K., Baglaenko,Y., Brenner,M., Loh,P.R. and Raychaudhuri,S. (2019) Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods*, **16**, 1289–1296.
66. Johnson,T.S., Wang,T., Huang,Z., Yu,C.Y., Wu,Y., Han,Y., Zhang,Y., Huang,K. and Zhang,J. (2019) LAmbDA: label ambiguous domain adaptation dataset integration reduces batch effects and improves subtype detection. *Bioinformatics*, **35**, 4696–4706.
67. Xie,P., Gao,M., Wang,C., Zhang,J., Noel,P., Yang,C., Von Hoff,D., Han,H., Zhang,M.Q. and Lin,W. (2019) SuperCT: a supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles. *Nucleic Acids Res.*, **47**, e48.
68. Ma,F. and Pellegrini,M. (2020) ACTINN: automated identification of cell types in single cell RNA sequencing. *Bioinformatics*, **36**, 533–538.
69. Alquicira-Hernandez,J., Sathe,A., Ji,H.P., Nguyen,Q. and Powell,J.E. (2019) scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol.*, **20**, 264.
70. Pliner,H.A., Shendure,J. and Trapnell,C. (2019) Supervised classification enables rapid annotation of cell atlases. *Nat. Methods*, **16**, 983–986.
71. Wei,Z. and Zhang,S. (2021) CALLR: a semi-supervised cell-type annotation method for single-cell RNA sequencing data. *Bioinformatics*, **37**, i51–i58.
72. Dominguez Conde,C., Xu,C., Jarvis,L.B., Rainbow,D.B., Wells,S.B., Gomes,T., Howlett,S.K., Suchanek,O., Polanski,K., King,H.W., *et al.* (2022) Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, **376**, eabl5197.
73. Zhao,X., Wu,S., Fang,N., Sun,X. and Fan,J. (2020) Evaluation of single-cell classifiers for single-cell RNA sequencing data sets. *Brief. Bioinform.*, **21**, 1581–1595.
74. Cao,Y., Wang,X. and Peng,G. (2020) SCSA: a cell type annotation tool for single-cell RNA-seq data. *Front. Genet.*, **11**, 490.
75. Carter,B. and Zhao,K. (2021) The epigenetic basis of cellular heterogeneity. *Nat. Rev. Genet.*, **22**, 235–250.
76. Guilhamon,P., Chesnelong,C., Kushida,M.M., Nikolic,A., Singhal,D., MacLeod,G., Madani Tonekaboni,S.A., Cavalli,F.M., Arlidge,C., Rajakulendran,N., *et al.* (2021) Single-cell chromatin accessibility profiling of glioblastoma identifies an invasive cancer stem cell population associated with lower survival. *eLife*, **10**, e64090.
77. Lin,Y., Wu,T.Y., Wan,S., Yang,J.Y.H., Wong,W.H. and Wang,Y.X.R. (2022) scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. *Nat. Biotechnol.*, **40**, 703–710.
78. Buenrostro,J.D., Wu,B., Chang,H.Y. and Greenleaf,W.J. (2015) ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.*, **109**, 21 29 21–21 29 29.
79. Angermueller,C., Clark,S.J., Lee,H.J., Macaulay,I.C., Teng,M.J., Hu,T.X., Krueger,F., Smallwood,S., Ponting,C.P., Voet,T., *et al.* (2016) Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods*, **13**, 229–232.
80. Packer,J. and Trapnell,C. (2018) Single-cell multi-omics: an engine for new quantitative models of gene regulation. *Trends Genet.*, **34**, 653–665.
81. Chen,K.H., Boettiger,A.N., Moffitt,J.R., Wang,S. and Zhuang,X. (2015) RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, **348**, aaa6090.
82. Stoeckius,M., Hafemeister,C., Stephenson,W., Houck-Loomis,B., Chattopadhyay,P.K., Swerdlow,H., Satija,R. and Smibert,P. (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods*, **14**, 865–868.

Received: August 15, 2023. **Revised:** September 12, 2023. **Editorial Decision:** September 27, 2023. **Accepted:** September 27, 2023