D596–D602 *Nucleic Acids Research, 2022, Vol. 50, Database issue* *Published online 17 November 2021*
https://doi.org/10.1093/nar/gkab1020

DISCO: a database of Deeply Integrated human Single-Cell Omics data

Mengwei Li[1], Xiaomeng Zhang[1], Kok Siong Ang[1], Jingjing Ling[1], Raman Sethi[1], Nicole Yee Shin Lee[1], Florent Ginhoux[1,3,4] and Jinmiao Chen[1,2,*]

[1]Singapore Immunology Network (SIgN), Agency for Science, Technology and Research (A*STAR), 8A Biomedical Grove, Immunos Building, Level 3, Singapore 138648, Singapore, [2]Immunology Translational Research Program, Yong Loo Lin School of Medicine, Department of Microbiology and Immunology, National University of Singapore (NUS), 5 Science Drive 2, Blk MD4, Level 3, Singapore 117545, Singapore, [3]Shanghai Institute of Immunology, Shanghai JiaoTong University School of Medicine, Shanghai 200025, China and [4]Translational Immunology Institute, SingHealth Duke-NUS Academic Medical Centre, Singapore 169856, Singapore

Received August 15, 2021; Revised October 04, 2021; Editorial Decision October 11, 2021; Accepted October 13, 2021

ABSTRACT

The ability to study cellular heterogeneity at single cell resolution is making single-cell sequencing increasingly popular. However, there is no publicly available resource that offers an integrated cell atlas with harmonized metadata that users can integrate new data with. Here, we present DISCO (https://www.immunesinglecell.org/), a database of Deeply Integrated Single-Cell Omics data. The current release of DISCO integrates more than 18 million cells from 4593 samples, covering 107 tissues/cell lines/organoids, 158 diseases, and 20 platforms. We standardized the associated metadata with a controlled vocabulary and ontology system. To allow large scale integration of single-cell data, we developed FastIntegration, a fast and high-capacity version of Seurat Integration. We also developed CELLiD, an atlas guided automatic cell type identification tool. Employing these two tools on the assembled data, we constructed one global atlas and 27 sub-atlases for different tissues, diseases, and cell types. DISCO provides three online tools, namely Online FastIntegration, Online CELLiD, and CellMapper, for users to integrate, annotate, and project uploaded single-cell RNA-seq data onto a selected atlas. Collectively, DISCO is a versatile platform for users to explore published single-cell data and efficiently perform integrated analysis with their own data.

INTRODUCTION

Single-cell RNA sequencing has emerged as a powerful tool for dissecting cellular heterogeneity to discover rare cell types and study gene regulation at the cellular level. In the past decade, there has been an exponential growth in single-cell transcriptome studies, covering a wide range of tissue types and diseases (1–4). Advances in technology have not only slashed sequencing costs, but also increased the number of cells sequenced per experiment, with a coverage of more than a million cells being reported (5–7). The growing availability of single-cell data offers opportunities for data integration to create comprehensive cell maps and enhance the power of downstream analyses.

These developments also bring about challenges in the management and integration of single-cell data. Presently, there are >400 single-cell RNA-seq datasets available in public databases (Figure 1A). Several resource websites have sought to organize the plethora of published datasets. For example, the Human Cell Atlas Data Portal (https://data.humancellatlas.org/explore/projects) integrates community generated single-cell data from 151 projects and provides uniformly processed data. PanglaoDB (8) collects 305 and 1063 single-cell datasets of human and mouse, respectively. Single Cell Portal (Broad Institute, https://singlecell.broadinstitute.org/single_cell) is a data portal that hosts 366 user uploaded single-cell RNA-seq datasets as of July 2021. TISCH focuses on the tumor microenvironment, hosting an integrated single-cell atlas constructed from 79 cancer studies (9). The current single-cell databases have three key drawbacks. Firstly, most of these databases only provide processed data within their respective studies; no data integration was done or no batch-corrected values are provided. Having integrated atlases for specific tissues or diseases are highly useful as consensus reference maps and for enhancing downstream analyses. Secondly, the associated metadata is not harmonized, with non-standard formatting and naming conventions. In particular, the cell type labels do not follow any standardized cell type ontology. Thirdly, they provide limited analysis functionalities and visualiza-

*To whom correspondence should be addressed. Tel: +65 64070395; Fax: +65 64642056; Email: Chen_Jinmiao@immunol.a-star.edu.sg

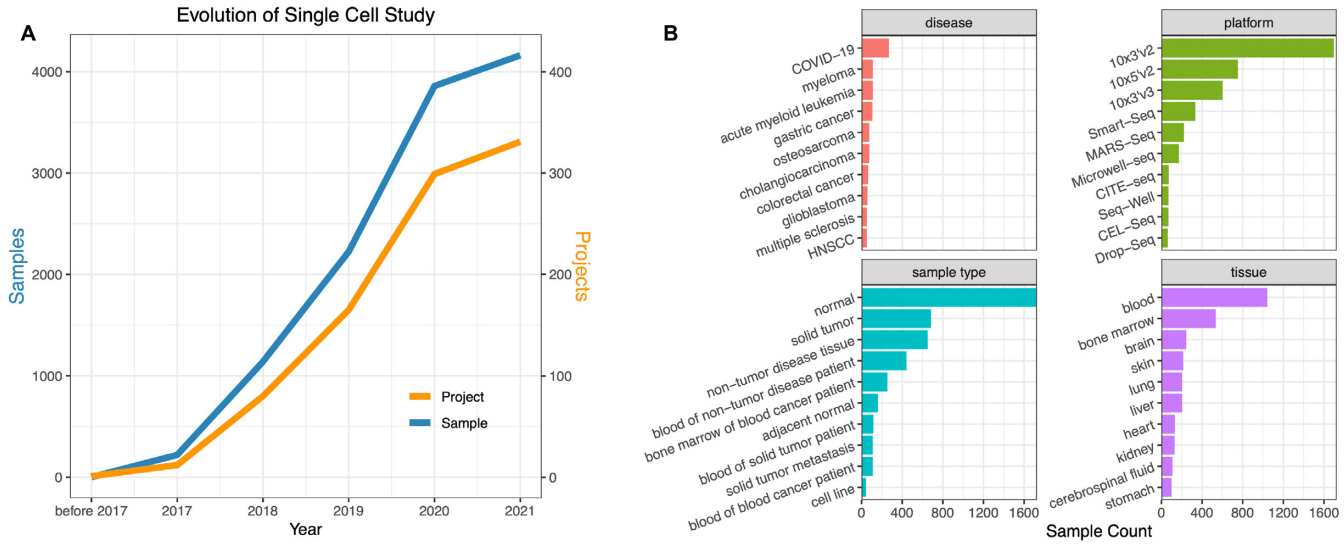

Figure 1. Statistics of single-cell studies. (**A**) The growth in number of single-cell projects and samples. (**B**) Top diseases, platforms, sample types, and tissues in the single-cell field.

tion capabilities. For example, none of them allow users to project their own data onto the hosted data.

We present DISCO, a database of Deeply Integrated Single-Cell Omics data. DISCO integrates more than 18 million cells from 4593 samples in 351 projects, covering 107 tissues/cell lines/organoids, 158 diseases, and 20 platforms. All the data hosted on DISCO were processed from raw fastq files using a standardized pipeline (Figure 2). Leveraging on the large number of public cell type annotation, we developed CELLiD and applied it to annotate the cell types in an automatic and standardized way. To integrate the single-cell data and create consensus reference maps, we developed FastIntegration, which can integrate more than four million cells. Currently, DISCO provides one global atlas and 27 sub-atlases for 23 tissues, 3 diseases (COVID-19, breast cancer, and colorectal cancer), and B/plasma cells. DISCO is equipped with three online tools, FastIntegration for online data integration, CELLiD for online cell type identification, and CellMapper for online cell projection. These tools enable users to perform custom data integration, and to upload their own data for cell type annotation and mapping onto the available atlases. The integrated atlases and all sample data used to construct the atlases are also available for downloading. In summary, we believe that DISCO is a valuable data resource for exploring cell types and gene expressions across different healthy and diseased human tissues, and can help accelerate the discovery of novel cell types and their associated functions.

MATERIALS AND METHODS

Data pre-processing

We first retrieved human tissue, organoid, or cell line acquired single-cell RNA-seq datasets with publicly available raw reads from GEO, ArrayExpress, GSA, and other public resources (10–12). By only considering datasets with raw reads (fastq, SRA, or bam file) we can perform re-alignment to a single reference genome. This ensures consistent gene identifiers across different studies to facilitate data integration and help minimize batch effects. For data in SRA and bam format, we converted them to fastq using SRA-toolkit (https://github.com/ncbi/sra-tools) and bamtofastq (https://github.com/10XGenomics/bamtofastq), respectively. We then employed demultiplexing methods appropriate for the different sequencing technologies. For the 10x Genomics platform, cell debarcoding and unique molecular identifiers (UMIs) identification was performed using Cell Ranger (version 3.1) with the RNA read offset increased from 26 to 39 in the SC5P-PE mode to remove the TSO sequence. For the Drop-seq platform, demultiplexing was done using Drop-seq tools (version 2.4.0, https://github.com/broadinstitute/Drop-seq). Finally, we employed UMI-tools (13) for demultiplexing data acquired with other technologies like BD Rhapsody, CEL-seq2, and Seq-Well. After debarcoding, reads tagged with a cell barcode and UMI were mapped to the human reference genome assembly hg38 using STAR (version 020201) (14), and then assigned to the corresponding gene (Ensembl 93) with featureCounts (version 2.0.2) (15). We then performed cell level quality control (QC) by filtering on mitochondrial mRNA counts and unique feature counts. We fitted the normal distribution to both count distributions and applied cutoffs based on Z-scores. Cells with high mitochondrial mRNA counts were removed (Z-score $>$ 1.64, P-value 0.05) (16), as were cells with too high or too low unique feature counts (|Z-score| $>$ 1.64). We further filtered out samples with less than 500 cells remaining. Though excluded from the data integration, these samples are marked as low quality and remain downloadable on DISCO. The QC script is available at (https://github.com/JinmiaoChenLab/DISCO_manuscript/blob/master/QC.R).

We also retrieved and manually curated the accompanying metadata for harmonization across datasets. Achieving metadata consistency facilitates efficient data retrieval and downstream data integration. The harmonized metadata is organized into eight common fields (sample ID, project ID,

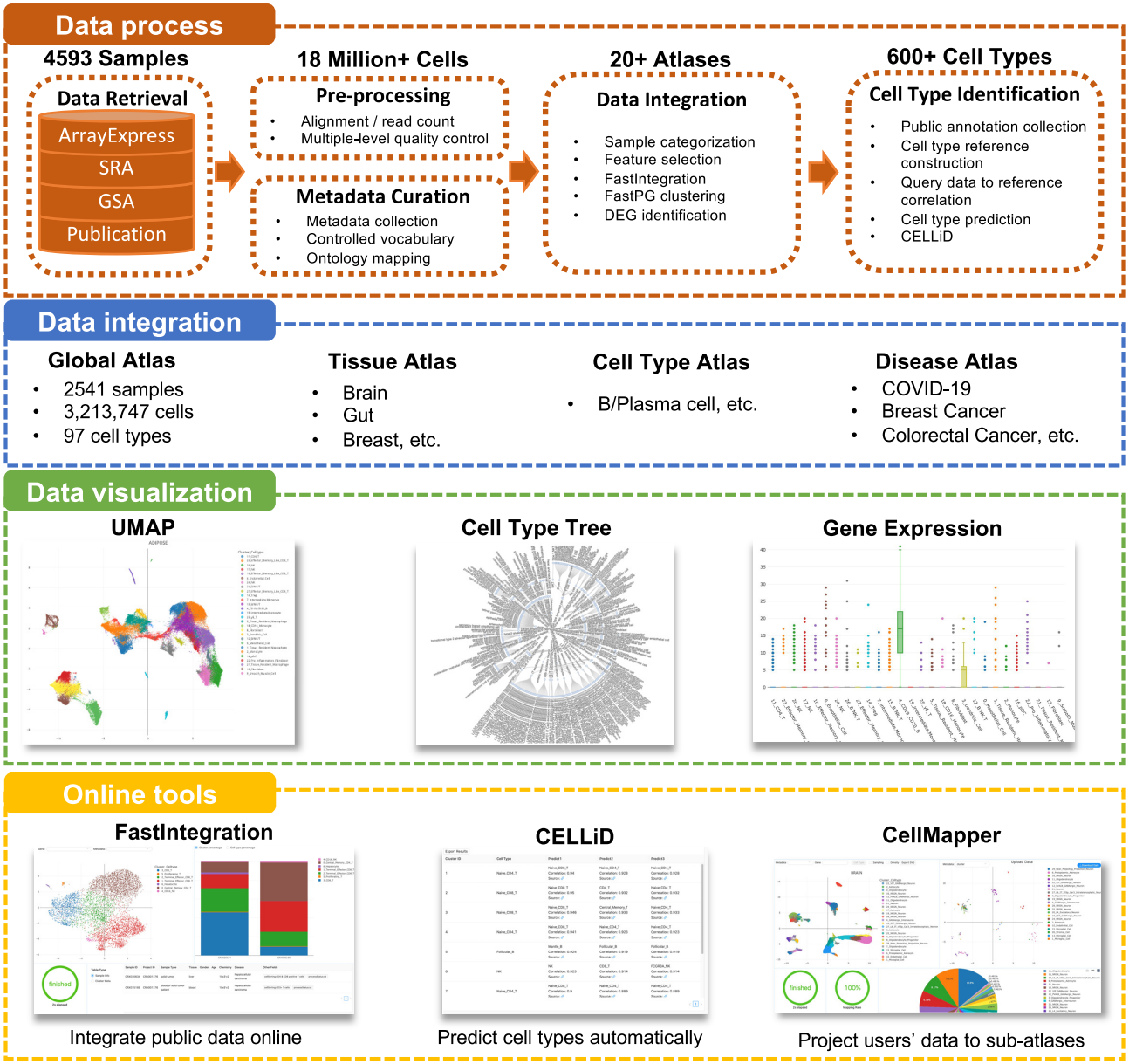

Figure 2. The data processing pipeline, and the database content and tools of DISCO.

sample type, disease status, tissue, platform, gender, and cell number) and 26 study-specific fields. For each field, we set up a controlled vocabulary or applied Experimental Factor Ontology (EFO) on the entries therein (17). For example, in the sample type field, samples were categorized into 18 types according to tissue type and disease status.

CELLiD: cell type annotation

As cell type annotation varies wildly across the datasets used in our atlas construction, we harmonized the cell type labels with our own automatic cell type annotation tool. For successful automatic cell type annotation, a high quality cell type reference database is necessary. Therefore, we first constructed a cell type specific gene expression reference by separately collecting 84 single-cell datasets that contain detailed cell type annotations. We extracted the original cell type names and manually standardized them for the same cell type. For compatibility with the studies used in atlas construction, we mapped the original cell type names to their closest Cell Ontology types (18). We also constructed a hierarchical tree that illustrates the relationships between cell types.

With this curated cell type reference, we developed CELLiD (CELL type iDentification) to annotate single-cell data at the cluster level (Figure 3A). CELLiD takes in the average gene expression of a cell cluster and performs two rounds of predictions. In the first round (coarse grain stage), it computes the Spearman correlation between each input cell cluster and each reference cell type with all overlap-

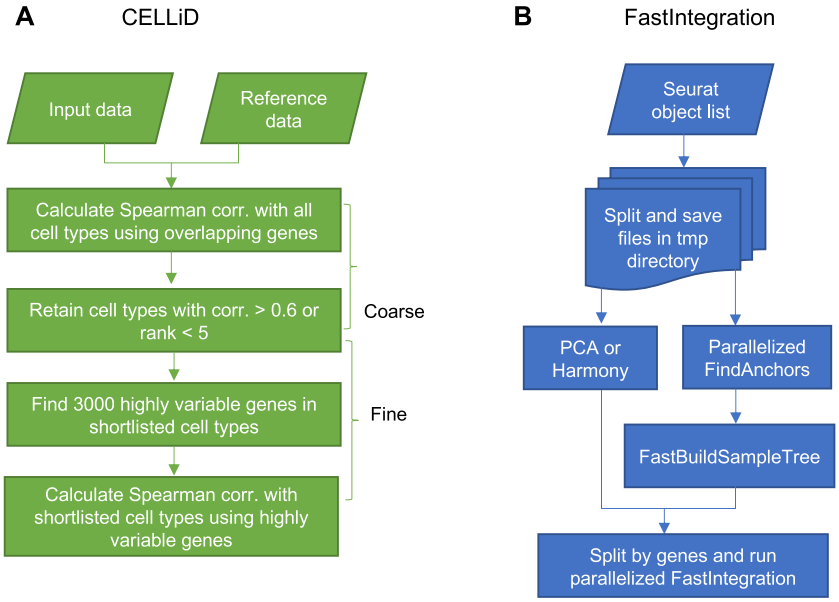

Figure 3. The workflow diagrams of (**A**) CELLiD and (**B**) FastIntegration.

ping genes. Cell types with correlation coefficients >0.6 or ranked within the top 5 will be retained. In the second round (fine grain stage), it selects the top 3000 highly variable genes in the retained cell types and calculates the Spearman correlation again. The cell type with the highest correlation coefficient will then be reported. The full report contains the original cell type name, standardized name, associated cell ontology, hierarchical tree, and information on originating dataset.

FastIntegration: data integration

In our previous benchmark study, we found Seurat and Harmony to be among the best performing data integration algorithms (19). While Harmony was ranked higher overall than Seurat, it does not provide batch corrected gene expression values (20), which hinders downstream analysis, including the identification of differentially expressed genes and inference of regulatory networks. Conversely, Seurat returns corrected gene expression values, but does not scale well with large numbers of cells and batches. This renders Seurat unusable for constructing large atlases with millions of cells. To overcome Seurat's performance limitations, we created FastIntegration (https://github.com/JinmiaoChenLab/FastIntegration) by modifying the original Seurat batch integration functions to enhance parallelization and employ matrix operations. Moreover, we also compiled R with OpenBLAS (https://www.openblas.net/) to speed up all matrix operations. These changes enabled us to integrate more than four million cells in one week. We improved the original Seurat functions in these aspects (Figure 3B):

1) In the FindIntegrationAnchors stage, Seurat uses the future package for parallelization, which is memory intensive. We switched to the pbmclappy package (https://cran.r-project.org/web/packages/pbmcapply/) for parallelization and also save each RNA expression matrix in a temporary file to be read when needed. These changes improved parallelization performance while consuming less memory.
2) In the BuildSampleTree stage, Seurat calculates the sample similarity matrix in a loop which is very slow when the number of samples is large. We converted it to matrix operations where computation for thousands of samples only requires several minutes.
3) When integrating multiple datasets in the IntegrateData, Seurat constructs a hierarchical tree based on the similarity between all pairs of datasets and conducts recursive pairwise correction up the tree. Principal Component Analysis (PCA) is performed on the query datasets during each round of integration, and the nearest anchors are identified in this PCA space. In FastIntegration, we execute PCA or Harmony only once on all unintegrated datasets and identify the nearest anchors in this PCA space. In doing so, we can parallelize the calculation of integrated expression values for different genes to speed up the computation.

With FastIntegration, we constructed a global cell atlas and respective sub-atlases. For the tissue and disease specific sub-atlases, all cells from the respective tissue or disease samples were integrated. For the cell type specific sub-atlases, we selected cells by their predicted cell type from all samples for integration. To construct the global atlas, we integrated 1000 cells from each sample. Within a sample, the same number of cells were randomly selected from each cluster, subject to the maximum number of cells in the cluster. To visualize the integrated data, we performed dimension reduction by PCA and used the top 50 PCs for Uniform Manifold Approximation and Projection (UMAP) (21). FastPG, a fast phenograph-like clustering method, was used to cluster the cells (22). Finally, we performed cell type identification of clusters using CELLiD.

CellMapper: Projection of query data to DISCO atlases

In DISCO, users can upload their own dataset and project it onto a desired atlas. This projection functionality is provided by our modified version of Seurat's reference mapping method. For mapping purposes, each sub-atlas is subsampled such that only 500–1000 cells in each cluster are used for the data projection. By default, Seurat computes the reference anchors in PCA reduced space, which is more efficient for big data. Here, we adopt Canonical-Correlation Analysis (CCA) dimension reduction for higher accuracy.

Implementation

The front end of DISCO is implemented using React (https://reactjs.org/), a JavaScript library for building user interfaces. We employ Redux (https://redux.js.org/), an open-source JavaScript library, to manage the front-end data. Most of the visualization functions are implemented with PlotyJS. We use WebGL (https://get.webgl.org/) to enable GPU accelerated plot rendering. The backend of DISCO is written in Java with the application framework based on Spring Boot. MySQL is used to store data while Redis serves as the middle layer to store hot data. Data exchange between the client and server uses Axios (https://github.com/axios/axios). FastIntegration, CELLiD, and CellMapper are implemented in R.

DATABASE CONTENT AND ONLINE TOOLS

DISCO hosts a comprehensive collection of single-cell RNA-seq datasets that cover a wide range of tissues, diseases, and platforms. Currently, it integrates 4593 samples from 351 projects, involving 107 tissues/cell lines/organoids, 158 diseases, and 20 platforms. It is accompanied by harmonized cell-level metadata and curated cell type annotation for consistency. We used this collection to construct a global cell atlas, as well as tissue, disease, and cell type specific sub-atlases that users can explore online. DISCO is organized into five main pages, sub-atlas, repository, cell types, gene, and download. We also provide a global search function on the home page to allow users to find gene specific information or tissue sub-atlases.

DISCO provides one global atlas, 23 tissue, 3 disease, and 1 cell type sub-atlases. Each sub-atlas has its own page where users can interactively explore the data. Each atlas page has three main panels which provide different information and visualization capabilities. In the UMAP panel, users can color individual cells by cell type, metadata, or gene expression values. For sub-atlases with more than one million cells, users can downsample the data to accelerate plot rendering. DISCO also provides density plotting, which is useful when too many cells overlap. The information panel allows users to view a variety of information on the cells clusters, genes, and samples. The cell type table lists the clusters found in the atlas and their predicted cell type information. The feature genes of the clusters are plotted as a heatmap in the feature gene tab, while the full list of markers by cluster are found in the marker tab. Users can also visualize the expression of a selected gene as a box plot in different clusters in the gene expression tab. The abundance of different cell types in the contributing samples are given as box plots in the cell type frequency tab and detailed sample information such as data source are listed in the sample tab. All plots generated onsite can be exported as high-quality vector graphics, while the cell types, marker, and sample information can be downloaded as formatted tables.

The repository page (https://www.immunesinglecell.org/repository) lists all the processed data samples and associated information, including tissue, disease, sample type, and gender. Users can browse and filter them for downloading. Here, users can also employ FastIntegration for online data integration with up to five samples. Upon successful integration, users can view the UMAP plot of the integrated data and the breakdown of cluster percentage per sample and cell type percentage. Similar to the UMAP plots of the constructed atlases, users can highlight cells by the metadata or the expression levels of selected genes.

Detailed information on cell types and genes can be found on the cell type browser (https://www.immunesinglecell.org/ct_pub) and gene browser (https://www.immunesinglecell.org/gene) page, respectively. On the cell type page, we provide detailed information on each collected cell type including their original name, standardized name, mapped ontology, and source publication. We also provide a hierarchical tree to visualize the relationships between cell types. The gene browser page lists all the genes found in the data samples. Users can narrow down the list by filtering for cell type specific genes. Each gene has a detailed information page accessible by clicking on the gene symbol. Here, users can view its cluster expression in a dot plot for clusters and on the UMAP plot of a selected atlas. We also provide the list of other genes in the atlas that are highly correlated.

In addition to FastIntegration at the repository page, DISCO also provides CELLiD and CellMapper online. CELLiD (http://www.immunesinglecell.org/cellpredictor) is the same tool employed for annotating the DISCO datasets. From our experience, CELLiD is able to give accurate cell type predictions comparable to manual annotations. Users can upload their own data to CELLiD and export the cell type predictions in a tabular format. The second tool is CellMapper (http://www.immunesinglecell.org/ct_mapper) which enables projection of uploaded data onto the constructed atlases. Taking advantage of the numerous sub-atlases and global atlas, CellMapper provides a powerful way to investigate single-cell RNA-seq data. On the result page of CellMapper, users can view the UMAP of the selected atlas and mapped data, and highlight individual cells according to their metadata. Although CellMapper also provides cell type annotating functionality via label transfer, we recommend CELLiD for cell type prediction tasks, as the cell type labels in the reference atlas were also predicted by CELLiD.

Finally, the download page (https://www.immunesinglecell.org/download) allows users to download all integrated data, including gene expression level, metadata of sample, and metadata of cells in RDS format. These files contain all the necessary data for downstream analysis. The page also contains links to download the R scripts for CELLiD, CellMapper, and QC, as well as the CELLiD's reference data.

DISCUSSION

Capturing cellular behavior at single cell resolution is essential for dissecting the heterogeneity of healthy and diseased tissues. This has enabled us to discover new cell types, investigate cell-cell interactions and consequently understand how different cells coordinate to give rise to overall tissue behavior. Integrating data from multiple studies constructs comprehensive cellular maps of human tissues that can enhance the power of analyses. We present DISCO, a comprehensive single-cell RNA-seq database. To construct this resource, we remapped all single-cell studies with available raw reads to standardize the gene annotation and minimize batch effects. We also curated the associated metadata to conform to a standardized vocabulary, and re-annotated the cell types with a newly developed tool, CELLiD. To integrate the reprocessed data, we developed FastIntegration that can integrate more than four million cells within a week. The final integrated data is available as a global cell atlas or as thematic sub-atlases specific to tissue, cell type, or disease. Users can visualize a desired atlas, the breakdown of cell types, and the expression patterns of genes. The tools, CELLiD, FastIntegration, and CellMapper, are available to users. Users can upload their own data for cell type annotation with CELLiD, and map their data onto the available atlases with CellMapper. Users can also select specific datasets within DISCO for custom data integration with FastIntegration. Finally, the reprocessed data can be freely downloaded for constructing custom atlases with other datasets, targeting various tissues and diseases. We intend DISCO to be a comprehensive and high quality repository for processed single-cell omics data.

As single-cell technological developments continue apace, we intend to continuously update and upgrade DISCO. We will update DISCO as new studies are published. New sub-atlases will also be constructed and annotation updated as needed to reflect any new developments. We also plan to expand the scope of DISCO to encompass other single-cell omics data, such as scATAC-seq, scTCR-seq, scBCR-seq, and spatial transcriptomics. The different omics data will be integrated to create a single-cell multi-omics reference atlas.

DATA AVAILABILITY

All data and scripts in DISCO are free for academic research. The codes of FastIntegration are maintained in the GitHub: https://github.com/JinmiaoChenLab/FastIntegration Scripts of CELLiD, CellMapper are available at https://github.com/JinmiaoChenLab/DISCO_manuscript.

ACKNOWLEDGEMENTS

We thank Vishuo Biomedical Pte Ltd for providing servers and technical support. We also thank a number of users for reporting bugs and providing suggestions, as well as the anonymous reviewers for their valuable comments on this work.

FUNDING

Open Fund Individual Research Grant [MOH-OFIRG18nov-2013]. Funding for open access charge: Open Fund Individual Research Grant [MOH-OFIRG18nov-2013].
Conflict of interest statement. None declared.

REFERENCES

1. Han,X., Zhou,Z., Fei,L., Sun,H., Wang,R., Chen,Y., Chen,H., Wang,J., Tang,H., Ge,W. *et al.* (2020) Construction of a human cell landscape at single-cell level. *Nature*, **581**, 303–309.
2. Cao,J., O'Day,D.R., Pliner,H.A., Kingsley,P.D., Deng,M., Daza,R.M., Zager,M.A., Aldinger,K.A., Blecher-Gonen,R., Zhang,F. *et al.* (2020) A human cell atlas of fetal gene expression. *Science*, **370**, eaba7721.
3. Stephenson,E., Reynolds,G., Botting,R.A., Calero-Nieto,F.J., Morgan,M.D., Tuong,Z.K., Bach,K., Sungnak,W., Worlock,K.B., Yoshida,M. *et al.* (2021) Single-cell multi-omics analysis of the immune response in COVID-19. *Nat. Med.*, **27**, 904–916.
4. Eze,U.C., Bhaduri,A., Haeussler,M., Nowakowski,T.J. and Kriegstein,A.R. (2021) Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. *Nat. Neurosci.*, **24**, 584–594.
5. Hwang,B., Lee,J.H. and Bang,D. (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.*, **50**, 96.
6. Zheng,G.X., Terry,J.M., Belgrader,P., Ryvkin,P., Bent,Z.W., Wilson,R., Ziraldo,S.B., Wheeler,T.D., McDermott,G.P., Zhu,J. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.
7. Picelli,S., Faridani,O.R., Bjorklund,A.K., Winberg,G., Sagasser,S. and Sandberg,R. (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.*, **9**, 171–181.
8. 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**, baz046.
9. Sun,D., Wang,J., Han,Y., Dong,X., Ge,J., Zheng,R., Shi,X., Wang,B., Li,Z., Ren,P. *et al.* (2021) TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. *Nucleic Acids Res.*, **49**, D1420–D1430.
10. 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.
11. Kolesnikov,N., Hastings,E., Keays,M., Melnichuk,O., Tang,Y.A., Williams,E., Dylag,M., Kurbatova,N., Brandizi,M., Burdett,T. *et al.* (2015) ArrayExpress update–simplifying data submissions. *Nucleic Acids Res.*, **43**, D1113–D1116.
12. Wang,Y., Song,F., Zhu,J., Zhang,S., Yang,Y., Chen,T., Tang,B., Dong,L., Ding,N., Zhang,Q. *et al.* (2017) GSA: Genome Sequence Archive. *Genomics Proteomics Bioinformatics*, **15**, 14–18.
13. Smith,T., Heger,A. and Sudbery,I. (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.*, **27**, 491–499.
14. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
15. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
16. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck,W.M. 3rd, Zheng,S., Butler,A., Lee,M.J., Wilk,A.J., Darby,C., Zager,M. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.
17. Malone,J., Holloway,E., Adamusiak,T., Kapushesky,M., Zheng,J., Kolesnikov,N., Zhukova,A., Brazma,A. and Parkinson,H. (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, **26**, 1112–1118.
18. Diehl,A.D., Meehan,T.F., Bradford,Y.M., Brush,M.H., Dahdul,W.M., Dougall,D.S., He,Y., Osumi-Sutherland,D., Ruttenberg,A., Sarntivijai,S. *et al.* (2016) The Cell Ontology 2016:

enhanced content, modularization, and ontology interoperability. *J Biomed Semantics*, **7**, 44.
19. Tran,H.T.N., Ang,K.S., Chevrier,M., Zhang,X., Lee,N.Y.S., Goh,M. and Chen,J. (2020) A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.*, **21**, 12.
20. 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.
21. McInnes,L., Healy,J. and Melville,J. (2020) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv doi: https://arxiv.org/abs/1802.03426v3, 18 September 2020, preprint: not peer reviewed.
22. Bodenheimer,T., Halappanavar,M., Jefferys,S., Gibson,R., Liu,S., Mucha,P.J., Stanley,N., Parker,J.S. and Selitsky,S.R. (2020) FastPG: fast clustering of millions of single cells. bioRxiv doi: https://doi.org/10.1101/2020.06.19.159749, 20 June 2020, preprint: not peer reviewed.