Published online 21 September 2022 *Nucleic Acids Research, 2023, Vol. 51, Database issue* ***D1019–D1028***
https://doi.org/10.1093/nar/gkac791

HTCA: a database with an in-depth characterization of the single-cell human transcriptome

Lu Pan [1,2,†], Shaobo Shan [3,†], Roman Tremmel [4,5], Weiyuan Li [6], Zehuan Liao [7,8], Hangyu Shi [9], Qishuang Chen [10], Xiaolu Zhang [11,*] and Xuexin Li [12,*]

[1]Institute of Environmental Medicine, Karolinska Institutet, Solna 17165, Sweden, [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Solna 17165, Sweden, [3]Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, Beijing 100050, China, [4]Dr. Margarete Fischer-Bosch Institute of Clinical Pharmacology, Stuttgart 70376, Germany, [5]University of Tuebingen, Tuebingen 72076, Germany, [6]School of Medicine, Yunnan University, Kunnan, Yunnan 650091, China, [7]Department of Microbiology, Tumor, and Cell Biology, Karolinska Institute, Solna 17177, Sweden, [8]School of Biological Sciences, Nanyang Technological University, 637 551, Singapore, [9]Department of Acupuncture, Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing 100029, China, [10]Graduate School, Beijing University of Chinese Medicine, Beijing 100029, China, [11]Department of Physiology and Pathophysiology, School of Basic Medical Sciences, Cheeloo College of Medicine, Shandong University, Jinan, Shandong 250012, China and [12]Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Solna 17165, Sweden

Received June 23, 2022; Revised August 12, 2022; Editorial Decision August 29, 2022; Accepted September 02, 2022

ABSTRACT

Single-cell RNA-sequencing (scRNA-seq) is one of the most used single-cell omics in recent decades. The exponential growth of single-cell data has immense potential for large-scale integration and in-depth explorations that are more representative of the study population. Efforts have been made to consolidate published data, yet extensive characterization is still lacking. Many focused on raw-data database constructions while others concentrate mainly on gene expression queries. Hereby, we present HTCA (www.htcatlas.org), an interactive database constructed based on ∼2.3 million high-quality cells from ∼3000 scRNA-seq samples and comprised in-depth phenotype profiles of 19 healthy adult and matching fetal tissues. HTCA provides a one-stop interactive query to gene signatures, transcription factor (TF) activities, TF motifs, receptor–ligand interactions, enriched gene ontology (GO) terms, etc. across cell types in adult and fetal tissues. At the same time, HTCA encompasses single-cell splicing variant profiles of 16 adult and fetal tissues, spatial transcriptomics profiles of 11 adult and fetal tissues, and single-cell ATAC-sequencing (scATAC-seq) profiles of 27 adult and fetal tissues. Besides, HTCA provides online analysis tools to perform major steps in a typical scRNA-seq analysis. Altogether, HTCA allows real-time explorations of multi-omics adult and fetal phenotypic profiles and provides tools for a flexible scRNA-seq analysis.

INTRODUCTION

The rapid advancement of biotechnologies in recent decades has leveraged the measurement resolutions and dimensions to enable observations of biological events at the single cellular level. Many new discoveries have been made at the single-cell level (1–6), which led to a quick dominance of single-cell omics in the current bioscience research. However, due to the high costs of single-cell omics, the number of samples used in many studies was far less representative of their study populations. In the area of scRNA-seq, the exponential increase in the number of single-cell studies in the recent decades with a dispersed focus in many areas of biology fosters opportunities for the research community to consolidate datasets and carry out large-sample analyses to increase study statistical power and decrease the number of false positives introduced by small sample studies. To date, for single-cell transcriptomics, scientists have made substantial efforts in transforming online and/or their in-house sequencing data into online databases (7–22) and they mainly fall into three categories. (i) Databases containing raw sequencing or raw/processed data gene-cell matrix files, e.g. Gene Expression Omnibus (GEO) (7), Human Cell Atlas (8) and European Nucleotide Archive (9);

*To whom correspondence should be addressed. Tel: +46 0704998515; Email: xuexin.li@ki.se
Correspondence may also be addressed to Xiaolu Zhang. Email: xiaolu.zhang@sdu.edu.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

(ii) individual tissue atlases for the depiction of study results, e.g. Heart Cell Atlas (10), Kidney Cell Atlas (11), Covid19 Cell Atlas (12), Tabula Sapiens (13) and Descartes atlas (14) and (iii) databases summarizing published studies, e.g. DISCO (15), CellMarker (16), JingleBells (17) and PanglaoDB (18). Databases in category (i) are repository sites to store raw/processed study data for the ease of raw data retrieval and they do not serve as databases to present any in-depth analysis of the data present. Data portal of the tissue-specific atlases in category (ii) served to showcase their independent study results and are very often overviews of the cell type compositions from the study data. An in-depth exploration or cross-comparison to other studies would require time and effort to download and analyze the relevant data. In particular, Tabula Sapiens atlas platform (13) contained extensive phenotypic characterizations of tissues, yet many other phenotypic characterizations such as receptor–ligand interactions, enriched motifs, TF networks, etc., are still lacking. Descartes atlas (14), on the other hand, is an atlas hosting a spectrum of their subsequent study results (14) and provided easy access to data downloads. Yet, Descartes showed less extensive features compared to Tabula Sapiens and in addition, in terms of human gene expression, Descartes consisted of only datasets from a single study. Databases in category (iii) generally consolidated datasets from various studies and provide gene signature or expression queries to users of their platforms. For example, DISCO is a multi-tissue scRNA-seq database integrating diseased and healthy human tissues, cell lines, and organoids to showcase cell type- or gene-specific signatures. Mixing healthy, diseased cells, cell lines, and organoids together for users to retrieve cell type-specific phenotypes for each tissue type, especially when each tissue consisted of mixtures of cells from various disease types, might cause falsified interpretations and discoveries that might jeopardize studies or future studies concerned. Furthermore, for the database, DISCO only provided cell type constitutions, gene signatures, differentially expressed genes (DEGs), gene expressions, and cell-type frequencies. Some other databases from category (iii) gathered published results and provide direct queries to study papers and study results, with no/less integrative insights into transcriptome profiles other than cell-type-specific DEGs signatures, which is a common characteristic of the databases from this category. Some served as databases only to provide external links to the studies they have consolidated. So far, in terms of integrating and showcasing phenotypic profiles of data from various studies, i.e. category (iii), these databases did not make extensive use of the data they have acquired to carry out vigorous assessments from various aspects of the scRNA-seq data.

To address the current limitations of category (iii) and utilize resources from categories (i) and (ii), we constructed the database HTCA. HTCA was built based on a collection of scRNA-seq data from ∼3000 samples with a total of ∼25 million cells from 19 healthy adult tissues and their matching fetal tissues (Figure 1). The final dataset contained ∼2.3 million cells after quality control (QC). We carried out in-depth assessments of the data we have consolidated to provide extensive phenotypic landscape overviews and data queries, including cell type constitutions, DEGs, gene expressions queries, cell type frequencies queries, correlations between adult and fetal tissue-specific cell types, transcription factor activities, top TFs specific to each cell types, enriched TF motifs (23), enriched GO terms (24,25) (biological pathways, cellular components and molecular functions), cell–cell and receptor–ligand interactions (Figure 1). In addition, HTCA is also a multi-omics atlas that provides phenotypic queries to single-cell isoform expressions of 16 adult and fetal tissues; gene expressions of spatial transcriptomics in 11 adult and fetal tissues; and chromatin co-accessibilities and TF motifs of scATAC-seq in 27 adult and fetal tissues.

HTCA also provides easy-to-use online analysis tools to allow users to process and analyze direct post-quantification outputs, including QC assessments and filtering, data imputation, data integration, dimension reduction, clustering, differential expression (DE) analysis, cell type prediction, manual annotation, data splicing, and cell–cell communication using various methods and database repositories (Figure 1), which are all parameter-adjustable. For example, the extent of filtering can be freely controlled by the user based on the QC plots HTCA provided. Users could compare and contrast their dataset with the data from HTCA using the tools we provided to enable fast comparisons with datasets across multiple studies. All in all, HTCA would serve as a one-stop solution to carry out quick and in-depth assessments of multi-omics single-cell data across tissues and cell types while enabling fast analysis of their own data.

MATERIALS AND METHODS

Data screened and data cleaning

For scRNA-seq data, we screened and downloaded raw data counts from various raw data resources (Figure 2A) such as the GEO, the Human Cell Atlas, Kidney Cell Atlas, Heart Cell Atlas, etc. (7,9–12,26–36). A total of 24 652 615 cells, comprising 6 584 880 fetal cells and 18 067 735 adult cells across 19 adult and fetal tissues were consolidated (Figure 2B). We excluded diseased cells, and cells from cell lines or organoids across the projects to give a clean database with only healthy human single-cell transcriptomes. We carried out stringent quality controls using Seurat (37), to remove low-quality cells with $>$ 5% mitochondrial counts, empty droplets with a low number of genes detected ($\leq$200), and doublets or multiplets with an abnormally high number of genes detected ($>$20 000). To further identify multiplets, DoubletDecon (38) was used with *rhop* value set to 0.6. Gene annotation was standardized to gene symbols, and samples with Ensembl ID (39) annotated were transformed using org.Hs.eg.db (version 3.8.2, https://bioconductor.org/packages/org.Hs.eg.db). After the quality control steps, we obtained 2 265 015 high-quality cells, consisting of 1 641 102 adult and 623 913 fetal cells across 19 adult tissues and their matching fetal tissues (Figure 2B). To obtain tissue-specific and cell-type-specific splicing variants, raw sequencing data from 16 adult and fetal tissues of the Human Cell Landscape (HCL) (19) was used. Early processing was done using the customized script from HCL (19). Alevin was used for alignment with UMI and cell

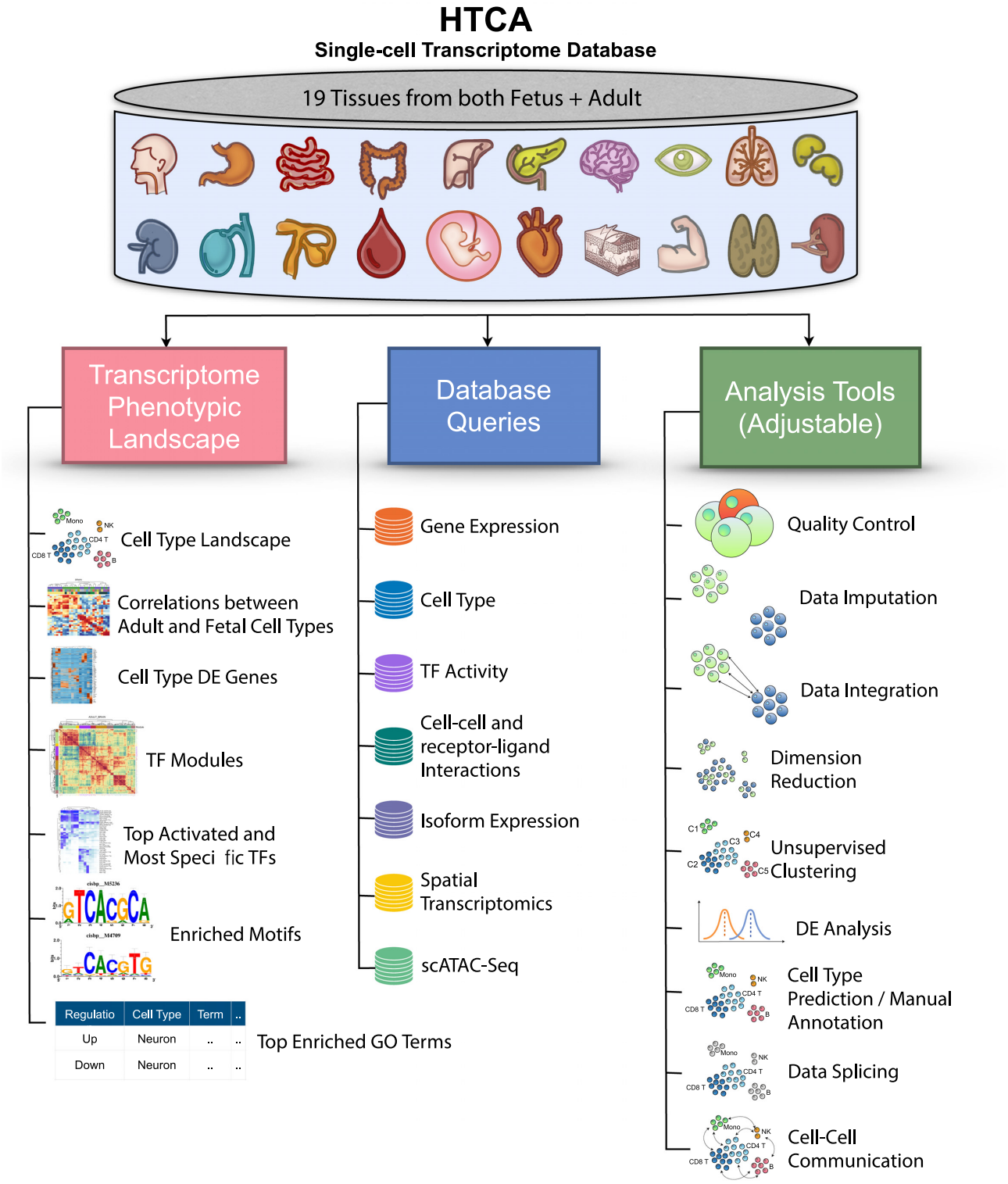

Figure 1. Schematic diagram of the HTCA database portal. HTCA incorporated phenotypic assessments from 19 adult tissues and their matching fetal tissues to form a single-cell transcriptome database. The database consisted of three main categories, namely tissue-wise phenotypic landscapes, database queries, and analysis tools.

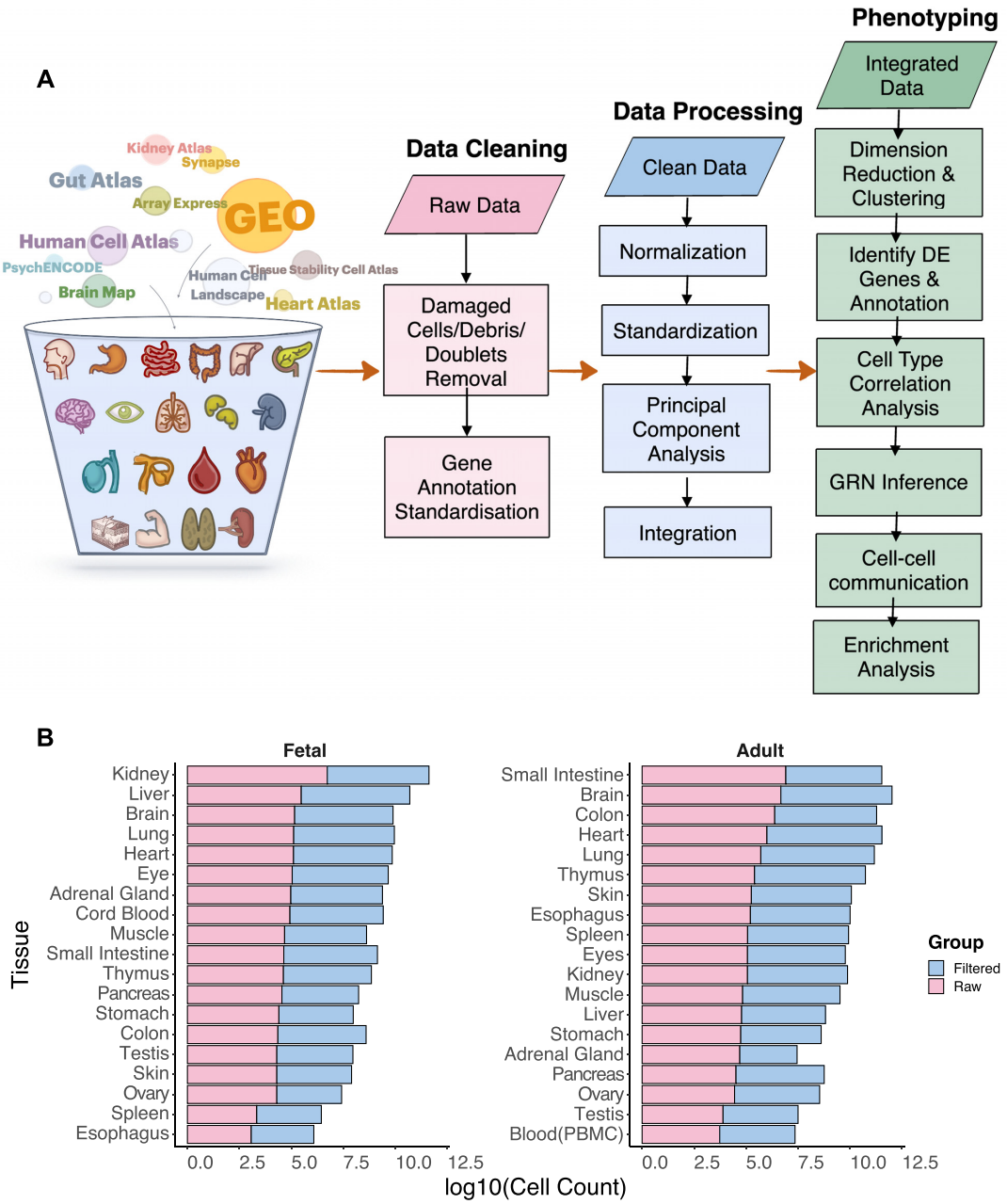

Figure 2. Workflow and cell summary of the HTCA database. (**A**) Diagram showing the workflow to construct the HTCA database, which consists of data cleaning, processing, and phenotyping. (**B**) Cell count of each tissue in the HTCA atlas in $\log_{10}$ scale.

barcode adjusted according to HCL Microwell-seq protocol (19,40,41), followed by quantification using Scasa (42). For spatial transcriptomics, post-quantification counts and histology images were obtained from various sources (43–50), and processed phenotypic profiles were obtained for scATAC-seq (51,52).

Data processing

For gene and isoform counts of scRNA-seq data, we normalized the expression counts for each cell by the total expression count of the cell and multiplies by a normalization factor of 10 000 (37). This was followed by standardization of expression values to a mean of 0 and variance of 1. To determine the number of principal components enough to cover most variance in the data, we performed principal component analysis (PCA) analysis (37) prior to data integration for each tissue (Figure 2A). To account for technical and technological differences introduced by different projects and technologies, data integration is needed to alleviate such variabilities across studies. Among 14 data integration methods (53), Harmony (54) was chosen for data integration as it was the most competent integration method considering its performance on datasets with different tech-

nologies as well as its low time complexity (53). Integration was done using Harmony based on PCA embeddings, via soft clustering and iterative embedding corrections (54) to correct for data source differences. A corrected set of PCA embeddings, known as Harmony embeddings, were used for downstream analyses. For spatial transcriptomics data, similar processing was done. Spot counts were normalized using a high-variance detection method, sctransform (37,55), by fitting the expression counts to a regularized binomial model to address for technical variations. For scATAC-seq, processed data underwent data cleaning to extract information on cell type, 2-dimension projection coordinates, enriched motifs, and co-accessibility regions.

Phenotyping: gene and isoform expression

For gene expression data of each tissue in scRNA-seq, nonlinear dimension reductions, tSNE and UMAP, were then performed based on the first 30 batch-corrected Harmony embeddings (Figure 2A), followed by unsupervised clustering via a shared nearest neighbor (SNN) modularity optimization clustering procedure (54) with default resolution. DE analysis was performed using Wilcoxon rank-sum test (54) by comparing each cell cluster to the rest of the cells present in the data in order to determine a list of DEGs uniquely and significantly expressed in each cluster. Correction for multiple testing was done using Bonferroni correction (56). At absolute average $\log_2$-fold-change ($\log_2$FC) values of > 0.5 and Bonferroni corrected $P < 0.01$, significant DEGs in each cluster were used for cell-type annotation. Cell types were annotated with reference to Human Primary Cell Atlas (57) using SingleR (58), and manually verified and corrected the final cell type identity for each cluster by cross-checking the DEGs of each cluster in each tissue across literature. We identified DEGs of each cell type by comparing the gene expressions of each cell type with all other cells in each tissue, using Wilcoxon rank-sum test and Bonferroni correction similar to the previously described cluster-wise DE analysis method. To measure cell type similarities and differences between adult and its matching fetal tissue, for each adult–fetal tissue pair, at Bonferroni corrected $P < 0.01$, average $\log_2$FC > 0, and in decreasing $\log_2$FC, we took the top 100 up-regulated genes of each cell type in the tissue pair to perform correlation analysis (Figure 2A) using neighbor voting (59). The correlation values were hierarchically clustered within each tissue pair to obtain the final clustering patterns between adult and fetal cell types. For isoform expression data, similar procedures were followed. Harmony was used for data integration of the same tissue. For both isoform and spatial transcriptomics data, similar downstream procedures were followed. Dimension reduction was done using tSNE and UMAP based on Harmony embeddings for integrated data in each tissue, and PCA embeddings for spatial transcriptomics data or tissues with a single sample in isoform data. Clustering was performed using SNN with default resolution. This was followed by DE analysis of clusters using the Wilcoxon rank-sum test followed by multiple testing corrections using the Bonferroni method. The same threshold was used compared to gene expression data, to identify significant DEGs in each cluster. For isoform expression data, cell type annotations from HCL were used, and the same method for DE analysis of cell types was used to identify DEGs in each cell type in each tissue.

Phenotyping: gene regulatory network (GRN) inference

The activities of TFs across cell types in the tissues were evaluated based on GRN inference and cell state identification method (23). For each tissue, we carried out soft gene filtering by retaining genes with expression counts >3 in at least 1% of the total cell and also genes that were detected in at least 1% of the cells. Subsampling was done to reduce the computational burden. Genes co-expressed with TFs were first identified and re-evaluated based on *cis*-regulatory motif analysis using RcisTarget (23), to eliminate false positives and retained upstream TF bundles with significantly enriched motifs (23). TFs were then assigned with cell-type-specific activity scores in terms of AUROC values using AUCell (23). We retained the final active TFs signatures for each cell type based on AUROC >0.1 and tissue-specific TFs based on regulon specificity scores (23) (RSS) >0.1 and incorporated them as part of the HTCA database (Figures 1 and 2A), together with the enriched motifs obtained from *cis*-regulatory motif analysis for each cell type.

Phenotyping: cell–cell communication

To investigate communication patterns between cell types in each tissue, CellPhoneDB ligand–receptor repository (60) was used to predict their interactions based on receptor–ligand interactions.

For each tissue, pairwise receptor–ligand expression comparisons were made between every two cell types to obtain a co-expression mean value for each receptor–ligand pair. Repeated permutation of cell type labels was then performed to disrupt the biological significance between cells to form a background null distribution of expression values specific to each interaction pair of the two cell types. Interaction pairs with co-expression values significantly higher than the background ($P < 0.05$) were retained and receptor–ligand profiles for each cell type in every tissue were incorporated into the HTCA database (Figures 1 and 2A).

Phenotyping: test for over-representation of GO Terms

Based on absolute $\log_2$FC >0.25 and Bonferroni corrected P <0.05, top 1000 up/down-regulated DEGs (ranked by decreasing log2FC for up-regulated gene sets and ranked by increasing $\log_2$FC for down-regulated gene sets) were served as inputs to GO functional analysis (Figure 2A). This was done separately for each tissue, cell type, and regulation group (i.e. up or down-regulation) using Limma (61). For each cell type in each tissue, multiple testing was corrected using Benjamini-Hochberg (BH) false discovery rate (FDR) controlling procedure (62). We retained enriched GO terms with BH FDR <0.05.

Methods used in the analysis tools

The main purpose of the analysis tools is to facilitate quick analysis of data from the user or to carry out integrative

analysis of the user data with the data from HTCA. In the QC step, the percentage of mitochondrial genes expressed in each cell, number of genes, and RNA molecules detected per cell will be calculated to remove possible empty droplets, doublets, artifacts, and damaged cells present in the sample as a result of experimental procedures (37). In the data imputation step, ALRA (63) was used for true expression signal amplification based on the low-rank matrix approximation (63) computed using singular vector decomposition (SVD) (64). In the data integration step, if more than one sample is submitted, Seurat or Harmony integration will be performed based on the filtered data and batch information provided by the user. Samples will be normalized and standardized prior to integration. Depending on the integration method, dimension reduction steps will be performed based on the pre-calculated PCA embeddings if Seurat integration was performed. Otherwise, dimension reduction steps will be performed based on Harmony embeddings. For unsupervised clustering, *k*-NN and SNN modular optimization will be performed (37) on the integrated data with user-defined clustering resolution. For each cluster, DE analysis using Wilcoxon rank-sum test will be performed and multiple-testing will be corrected using Bonferroni correction. Cell type identity will be predicted based on the clustering results (58) using Human Primary Cell Atlas (57) as the default annotation reference, or alternative reference provided by the user. For cell–cell communication analysis, LIANA was used to run the analysis using different methods and data repositories (65,66).

Database construction

Database queries were written in R and hosted by RShiny. The database of HTCA was stored in the RShiny server and the interactive part of tissue-wise phenotypic landscapes (Figure 1) was hosted using Rshiny. Analysis tools were implemented using R and hosted by RShiny. The backbone of HTCA was supported by HTML and Javascript.

RESULTS

Overview of HTCA and its analysis tools

HTCA is a database comprising in-depth phenotypic profiles of single-cell transcriptomes across 19 healthy human adult tissues and their matching fetal tissues. It also serves as a database query for cell-type-specific and tissue-specific splicing variants. In addition, HTCA is also a multi-omics database containing spatial transcriptomics and scATAC-seq phenotypic profiles in adult and fetal tissues. HTCA consisted of three major components, individual scRNA-seq tissue atlases depicting tissue-wise phenotypic landscapes, interactive database queries for different cellular profiles across tissues and omics, and online analysis tools to facilitate instantaneous analysis and visualizations of single-cell transcriptomics data (Figure 1).

Within each tissue atlas, HTCA provides interactive visualization of cell type constitutions of the tissue for users to click and zoom on different cell clusters. Other phenotype profiles present in the atlas include visualizations of the correlations of cell types of each adult tissue with cell types from the matching fetal tissue, DEGs signatures of each cell type, TF modules, TF signatures, and topmost specific TFs for each cell type, as well as enriched motifs and GO terms with search functions.

For database queries, seven sub-databases were created, including gene expression, cell type, TF activity, cell–cell interactions with receptor–ligand interactions, isoform expression, spatial transcriptomics profiles, and scATAC-seq profiles. For each of these query options, the gene expression query allows a user to search for a gene of interest (GOI) and the expression of the GOI in each cell type across tissues will be interactively visualized. To aid fast reference, a description of the gene from GeneCards (67) and genomic locations, and distribution of exons across transcripts of the GOI from the UCSC genome browser (68) will also be shown to the user. For the cell type query, the search of any cell type present in the HTCA database will interactively display the distribution of this cell type and its related cell type of lower granularity across tissues (only tissues with this cell type), as well as both up and down-regulated DEGs of this cell type across tissues, including fold-change, the significance level of post-DE testing for each DE gene, and the GRCh38 (69) genomic location of the gene. Users could rearrange, search or filter in the DEGs list based on a GOI or other filtering criteria. TF activity query allows users to interactively visualize the activity of a TF of interest (activity score in AUROC or in short AUC value) in the cells of a particular tissue on a tSNE plot, an interactive cell type tSNE will also be displayed. In the cell–cell and receptor–ligand query, users could search for a tissue of interest to observe the interaction patterns between cell types from the same tissue. Top receptor–ligand interactions between cell types will also be shown and users could filter, search and sort within the list of interactions. For isoform or splicing variant query option, UMAP constructed based on isoform expressions for each tissue will be displayed and users could select to view cell clusters, cell types, or expression of particular isoform across cells in each adult or fetal tissue. DE isoforms in table form will be shown for clusters or cell types depending on user selection. For spatial transcriptomics, the clusters mapped onto the histology image of each tissue will be shown, together with a volcano plot illustrating the fold-change (in $\log_2$ scale) of each gene in each cluster. DEGs in table form will be shown simultaneously. For querying scATAC-seq, the UMAP of each tissue with color coding indicating cell types will be shown, together with chromatin coaccesibility (if available in processed data) in each cell type, as well as cell-type-specific enriched motifs across tissues.

Analysis tools available on the HTCA include QC, data integration, data imputation, dimension reduction, clustering, DE analysis, cell type prediction, manual annotation, data splicing, and cell–cell communication, basically covering all the major steps in a typical scRNA-seq analysis workflow. All steps come with adjustable parameters and plots to visualize the analysis steps and examples are available for users in each tool to carry out fast exploration with sample files (and/or meta files) available for download. To provide users with publish-ready figures, the height and width of the plots can be expanded or shrank, the colors of the plots, size of the points in the scattered plots, and size of the labels can be changed. All plots are available for down-

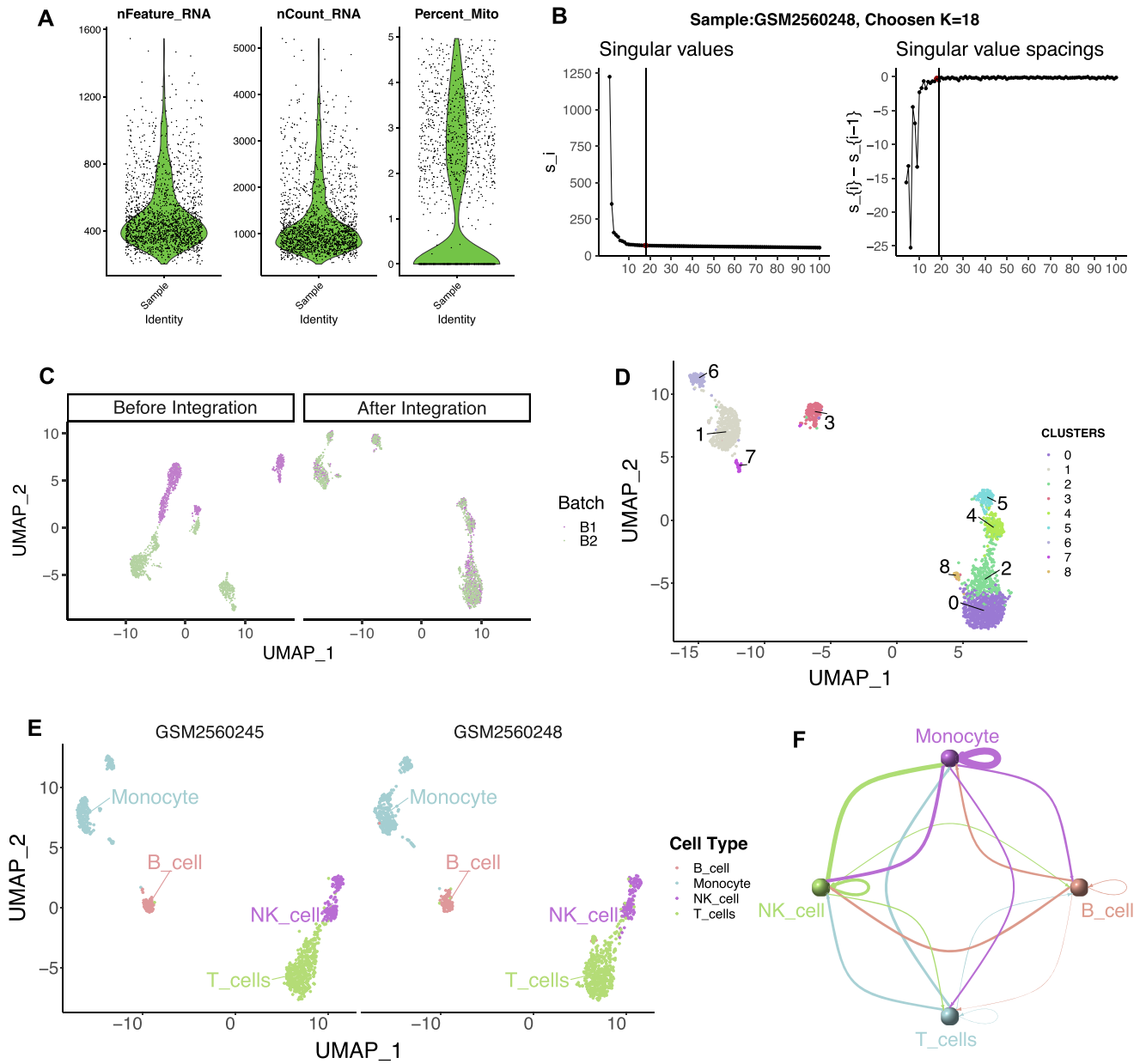

Figure 3. A walkthrough of the step-by-step online analysis tools. (**A**) Quality control step to enable manual filtering of cells based on their number of genes (i.e. nFeature_RNA), number of RNA molecules (i.e. nCount_RNA), and the percentage of mitochondria (Percent_Mito). (**B**) Data imputation step based on SVD, to predict missing values due to inadequate sequencing depths. (**C**) Data integration step to eliminate potential batch effect present in the data. (**D**) Dimension reduction, clustering, and DE analysis steps based on post-integration results. Cluster resolution can be freely adjustable to facilitate the aim of a project. (**E**) Cell type annotation step to allow automated or manual annotation of each cluster based on their DEGs. (**F**) Cell-cell communication step to predict possible interactions between defined cell types or clusters. All steps come with visualizations and download of intermediate files to enhance the flexibility and practicability of the tools for the users.

load in PNG or editable PDF formats and output files in *.RDS* format. In the QC tool, a user could upload the direct input of 10X Genomics (50,70) *.h5* filtered matrix files, or gene-to-cell matrix files in .csv/.txt format from other technologies, or Seurat object in *.RDS* format (one sample in one *.RDS* file). For a multiple-samples project, a meta file containing the batch and group information of each sample is required and the user could follow the format indicated in the sample meta file. Once uploaded, the user could visualize the number of genes, number of mRNA molecules, and percentage of mitochondrial genes in each cell (Figure 3A). The user will decide on the QC filtering cut-offs based on the plots to remove any cells that are most probable damaged cells, doublets or debris. Once they have decided, they could click 'create filtered dataset' to trigger the download of the post-filtering *.RDS* format Seurat object. The data imputation tool is an optional procedure for users to impute sparse single-cell data to enhance true biological sig-

nals. A single post-QC Seurat object in *.RDS* format can be uploaded to the tool and post-imputation plots showing singular values in each rank and singular value spacings between ranks will be displayed (Figure 3B). An optimal rank will be chosen and shown in the plots as vertical lines. Post-imputation data is downloadable and is compatible with subsequent tools. For a multiple-samples project, the user could proceed with the data integration step in the analysis tools by simply uploading the post-filtering *.RDS* Seurat object from the previous step and choose an integration method of interest. For faster runtime, Harmony is recommended. Once the file has been uploaded, the integration will be done automatically and the download of the post-integration *.RDS* Seurat object will start right after the completion of the integration process. Dimension reduction plots of before and after integration will be shown in tSNE and UMAP formats. For dimension reduction, users could upload the post-integration (for multiple samples) or post-filtering (for a single-sample project) .RDS Seurat object, and select the type(s) of dimension reductions to carry out (PCA, and/or UMAP, and/or tSNE) and the selected dimension reduction plots displaying batch information before and after integration will be shown (Figure 3C). In the clustering step, unsupervised clustering will be carried out. Users are required to upload the post-dimension reduction *.RDS* file from the previous step and proceed with the clustering analysis. The user could decide the number of clusters to set using the resolution option based on the tSNE or UMAP visualizations shown (Figure 3D). Depending on their final clustering degree (i.e. the resolution or number of clusters), DE analysis will be carried out to identify DEGs of each cluster. For the cell type prediction tool, the user will upload his post-clustering *.RDS* Seurat object, and the identity of each cell will be predicted based on its gene expression profile. Post-prediction tSNE or UMAP visualizations will also be shown to the user. For a multiple-groups or multiple-samples project, group-wise dimension reduction plots can also be shown, to allow users to compare between the groups via direct visualization (Figure 3E). To carry out a manual check on the cell type identity of each cluster, the user could refer to the DEGs list to validate the cell type annotations or proceed on with the manual annotation tool to annotate cells on their own. Users could produce data subsets using the data splicing tool to further analyze their cells of interest. For the cell–cell communication tool, a list of cell interaction methods and database repositories made available by LIANA will be provided to the user. Analysis can be done across different methods using different database resources at the same time and a cell–cell interaction network plot (Figure 3F) and a table of top receptor–ligand interaction pairs will be shown.

HTCA also consisted of a source page for viewing the demographics of HTCA interactively. A forum page is available to allow users to post questions, suggestions, or problems they have encountered while using the database so that HTCA could continue to expand, improve and include more features for the greater benefit of the research community. Many database portals or atlases do not provide such a function. We also provide a download page for users to download tissue-wise scRNA-seq datasets in HTCA.

DISCUSSION

Over the past decades, we have seen how single-cell technologies gained their dominance in biomedical science research, leading to new insights and putting forward new methods and applications driven by the new forms of high-dimension data. However, the high cost of sequencing data at the single-cell level caused a phenomenon that, the number of study samples in the majority of single-cell studies were not representative of their study populations. Therefore, data generated by studies could be consolidated and analyzed together to increase the statistical power based on the increase in sample/cell number, this is especially important for cell types that are present in minute amounts naturally in organism bodies, and the consolidation of the datasets may help to reveal their presence. To this end, we constructed HTCA, a comprehensive healthy human cell database. The platform hosts in-depth phenotypic profiles for 19 adult tissues and matching fetal tissues from scRNA-seq datasets to allow users to interactively explore these features. Users could query directly from the database depending on the type of phenotypes (i.e. gene expression, receptor–ligand interactions) they are interested in. HTCA also provided phenotypic queries of splicing variations, spatial transcriptomics profiles, and chromatin accessibility profiles across adult and fetal tissues. Analysis tools on the HTCA portal will enable users to carry out quick analyses on their data sets or to compare the data sets with the ones provided by HTCA. We hope that our one-stop single-cell multi-omics database and analytic tools could help researchers to save time and effort digging into various data sources just to observe specific phenotypic features. HTCA will continue to expand through the incorporation of more tissue types, analytic tools, and omics types, to piece up a more complete and diverse landscape of multi-omics healthy human landscape at the single-cell level.

DATA AVAILABILITY

All data including data sources, as well as analysis tools, are available on www.htcatlas.org. Codes to the analysis tools are maintained at the GitHub repository https://github.com/eudoraleer/HTCA.

ACKNOWLEDGEMENTS

The computations and data handling were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Rackham, partially funded by the Swedish Research Council through grant agreement no. 2018-05973. We would like to thank Liming Zhang for his help in the construction of the website. R.T. was supported by the Robert Bosch Stiftung, Stuttgart, Germany.

FUNDING

Karolinska Institute Network Medicine Global Alliance Collaborative Grant [C24401073 to X.L. and L.P.]; National Natural Science Foundation of China [8210100902 to X.Z.]; Nature Science Foundation of Shandong Province [ZR2021MH393 to X.Z.]. Funding for open access charge:

Karolinska Institute Network Medicine Global Alliance Collaborative Grant [C24401073 to X.L. and L.P.].
Conflict of interest statement. None declared.

REFERENCES

1. Ginhoux,F., Yalin,A., Dutertre,C.A. and Amit,I. (2022) Single-cell immunology: past, present, and future. *Immunity*, **55**, 393–404.
2. Mogilenko,D.A., Shchukina,I. and Artyomov,M.N. (2021) Immune ageing at single-cell resolution. *Nat. Rev. Immunol.*, **22**, 484–498.
3. Aldridge,S. and Teichmann,S.A. (2020) Single cell transcriptomics comes of age. *Nat. Commun.*, **11**, 4307.
4. Suvà,M.L. and Tirosh,I. (2019) Single-Cell RNA sequencing in cancer: lessons learned and emerging challenges. *Mol. Cell*, **75**, 7–12.
5. Lawson,D.A., Kessenbrock,K., Davis,R.T., Pervolarakis,N. and Werb,Z. (2018) Tumour heterogeneity and metastasis at single-cell resolution. *Nat. Cell Biol.*, **20**, 1349–1360.
6. Armand,E.J., Li,J., Xie,F., Luo,C. and Mukamel,E.A. (2021) Single-Cell sequencing of brain cell transcriptomes and epigenomes. *Neuron*, **109**, 11–26.
7. 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.
8. Regev,A., Teichmann,S.A., Lander,E.S., Amit,I., Benoist,C., Birney,E., Bodenmiller,B., Campbell,P., Carninci,P., Clatworthy,M. *et al.* (2017) The human cell atlas. *Elife*, **6**, e27041.
9. Harrison,P.W., Ahamed,A., Aslam,R., Alako,B.T.F., Burgin,J., Buso,N., Courtot,M., Fan,J., Gupta,D., Haseeb,M. *et al.* (2020) The european nucleotide archive in 2020. *Nucleic Acids Res.*, **49**, D82–D85.
10. Litviňuková,M., Talavera-López,C., Maatz,H., Reichart,D., Worth,C.L., Lindberg,E.L., Kanda,M., Polanski,K., Heinig,M., Lee,M. *et al.* (2020) Cells of the adult human heart. *Nature*, **588**, 466–472.
11. Stewart,B.J., Ferdinand,J.R., Young,M.D., Mitchell,T.J., Loudon,K.W., Riding,A.M., Richoz,N., Frazer,G.L., Staniforth,J.U.L., Braga,F.A.V. *et al.* (2019) Spatiotemporal immune zonation of the human kidney. *Science*, **365**, 1461–1466.
12. Delorey,T.M., Ziegler,C.G.K., Heimberg,G., Normand,R., Yang,Y., Segerstolpe,Å., Abbondanza,D., Fleming,S.J., Subramanian,A., Montoro,D.T. *et al.* (2021) COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. *Nature*, **595**, 107–113.
13. Jones,R.C., Karkanias,J., Krasnow,M.A., Pisco,A.O., Quake,S.R., Salzman,J., Yosef,N., Bulthaup,B., Brown,P., Harper,W. *et al.* (2022) The tabula sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science*, **376**, eabl4896.
14. 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.
15. Li,M., Zhang,X., Ang,K.S., Ling,J., Sethi,R., Lee,N.Y.S., Ginhoux,F. and Chen,J. (2022) DISCO: a database of deeply integrated human single-cell omics data. *Nucleic Acids Res.*, **50**, D596–D602.
16. Zhang,X., Lan,Y., Xu,J., Quan,F., Zhao,E., Deng,C., Luo,T., Xu,L., Liao,G., Yan,M. *et al.* (2019) CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.*, **47**, D721–D728.
17. Ner-Gaon,H., Melchior,A., Golan,N., Ben-Haim,Y. and Shay,T. (2017) JingleBells: a repository of immune-related single-cell RNA-Sequencing datasets. *J. Immunol.*, **198**, 3375–3379.
18. Franzén,O., Gan,L.-M. and Björkegren,J.L.M. (2019) PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database*, **2019**, baz046.
19. 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.
20. Karlsson,M., Zhang,C., Méar,L., Zhong,W., Digre,A., Katona,B., Sjöstedt,E., Butler,L., Odeberg,J., Dusart,P. *et al.* (2021) A single-cell type transcriptomics map of human tissues. *Sci. Adv.*, **7**, eabh2169.
21. 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.
22. Zhao,T., Lyu,S., Lu,G., Juan,L., Zeng,X., Wei,Z., Hao,J. and Peng,J. (2020) SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Res.*, **49**, D1413–D1419.
23. Aibar,S., González-Blas,C.B., Moerman,T., Huynh-Thu,V.A., Imrichova,H., Hulselmans,G., Rambow,F., Marine,J.-C., Geurts,P., Aerts,J. *et al.* (2017) SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods*, **14**, 1083–1086.
24. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.
25. Gene Ontology Consortium (2020) The gene ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
26. Eraslan,G., Drokhlyansky,E., Anand,S., Fiskin,E., Subramanian,A., Slyper,M., Wang,J., Wittenberghe,N.V., Rouhana,J.M., Waldman,J. *et al.* (2022) Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science*, **376**, eabl4290.
27. Perez-Riverol,Y., Zorin,A., Dass,G., Vu,M.-T., Xu,P., Glont,M., Vizcaíno,J.A., Jarnuczak,A.F., Petryszak,R., Ping,P. *et al.* (2019) Quantifying the impact of public omics data. *Nat. Commun.*, **10**, 3512.
28. Bakken,T.E., Jorstad,N.L., Hu,Q., Lake,B.B., Tian,W., Kalmbach,B.E., Crow,M., Hodge,R.D., Krienen,F.M., Sorensen,S.A. *et al.* (2021) Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature*, **598**, 111–119.
29. Hodge,R.D., Bakken,T.E., Miller,J.A., Smith,K.A., Barkan,E.R., Graybuck,L.T., Close,J.L., Long,B., Johansen,N., Penn,O. *et al.* (2019) Conserved cell types with divergent features in human versus mouse cortex. *Nature*, **573**, 61–68.
30. Tasic,B., Yao,Z., Graybuck,L.T., Smith,K.A., Nguyen,T.N., Bertagnolli,D., Goldy,J., Garren,E., Economo,M.N., Viswanathan,S. *et al.* (2018) Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, **563**, 72–78.
31. Li,M., Santpere,G., Kawasawa,Y.I., Evgrafov,O.V., Gulden,F.O., Pochareddy,S., Sunkin,S.M., Li,Z., Shin,Y., Zhu,Y. *et al.* (2018) Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science*, **362**, eaat7615.
32. Madissoon,E., Wilbrey-Clark,A., Miragaia,R.J., Saeb-Parsy,K., Mahbubani,K.T., Georgakopoulos,N., Harding,P., Polanski,K., Huang,N., Nowicki-Osuch,K. *et al.* (2019) scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome Biol.*, **21**, 1.
33. Kanz,C., Aldebert,P., Althorpe,N., Baker,W., Baldwin,A., Bates,K., Browne,P., van den Broek,A., Castro,M., Cochrane,G. *et al.* (2005) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **33**, D29–D33.
34. Young,M.D., Mitchell,T.J., Braga,F.A.V., Tran,M.G.B., Stewart,B.J., Ferdinand,J.R., Collord,G., Botting,R.A., Popescu,D.-M., Loudon,K.W. *et al.* (2018) Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science*, **361**, 594–599.
35. Elmentaite,R., Kumasaka,N., Roberts,K., Fleming,A., Dann,E., King,H.W., Kleshchevnikov,V., Dabrowska,M., Pritchard,S., Bolt,L. *et al.* (2021) Cells of the human intestinal tract mapped across space and time. *Nature*, **597**, 250–255.
36. Travaglini,K.J., Nabhan,A.N., Penland,L., Sinha,R., Gillich,A., Sit,R.V., Chang,S., Conley,S.D., Mori,Y., Seita,J. *et al.* (2020) A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature*, **587**, 619–625.
37. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck,W.M., 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.
38. DePasquale,E.A.K., Schnell,D.J., Van Camp,P.-J., Valiente-Alandí,Í., Blaxall,B.C., Grimes,H.L., Singh,H. and Salomonis,N. (2019) DoubletDecon: deconvoluting doublets from single-cell RNA-Sequencing data. *Cell Rep.*, **29**, 1718–1727.
39. Cunningham,F., Allen,J.E., Allen,J., Alvarez-Jarreta,J., Amode,MR., Armean,IrinaM., Austine-Orimoloye,O., Azov,AndreyG., Barnes,I., Bennett,R. *et al.* (2021) Ensembl 2022. *Nucleic Acids Res.*, **50**, D988–D995.

40. Patro,R., Duggal,G., Love,M.I., Irizarry,R.A. and Kingsford,C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.
41. Srivastava,A., Malik,L., Smith,T., Sudbery,I. and Patro,R. (2019) Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biol.*, **20**, 65.
42. Pan,L., Dinh,H.Q., Pawitan,Y. and Vu,T.N. (2021) Isoform-level quantification for single-cell RNA sequencing. *Bioinformatics*, **38**, 1287–1294.
43. Garcia-Alonso,L., Lorenzi,V., Mazzeo,C.I., Alves-Lopes,J.P., Roberts,K., Sancho-Serra,C., Engelbert,J., Marečková,M., Gruhn,W.H., Botting,R.A. *et al.* (2022) Single-cell roadmap of human gonadal development. *Nature*, **607**, 540–547.
44. Garcia-Alonso,L., Handfield,L.-F., Roberts,K., Nikolakopoulou,K., Fernando,R.C., Gardner,L., Woodhams,B., Arutyunyan,A., Polanski,K., Hoo,R. *et al.* (2021) Mapping the temporal and spatial dynamics of the human endometrium in vivo and in vitro. *Nat. Genet.*, **53**, 1698–1711.
45. Kadur Lakshminarasimha Murthy,P., Sontake,V., Tata,A., Kobayashi,Y., Macadlo,L., Okuda,K., Conchola,A.S., Nakano,S., Gregory,S., Miller,L.A. *et al.* (2022) Human distal lung maps and lineage hierarchies reveal a bipotent progenitor. *Nature*, **604**, 111–119.
46. Guilliams,M., Bonnardel,J., Haest,B., Vanderborght,B., Wagner,C., Remmerie,A., Bujko,A., Martens,L., Thoné,T., Browaeys,R. *et al.* (2022) Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. *Cell*, **185**, 379–396.
47. Wu,R., Guo,W., Qiu,X., Wang,S., Sui,C., Lian,Q., Wu,J., Shan,Y., Yang,Z., Yang,S. *et al.* (2021) Comprehensive analysis of spatial architecture in primary liver cancer. *Sci. Adv.*, **7**, eabg3750.
48. Bäckdahl,J., Franzén,L., Massier,L., Li,Q., Jalkanen,J., Gao,H., Andersson,A., Bhalla,N., Thorell,A., Rydén,M. *et al.* (2021) Spatial mapping reveals human adipocyte subpopulations with distinct sensitivities to insulin. *Cell Metab.*, **33**, 1869–1882.
49. Fawkner-Corbett,D., Antanaviciute,A., Parikh,K., Jagielowicz,M., Gerós,A.S., Gupta,T., Ashley,N., Khamis,D., Fowler,D., Morrissey,E. *et al.* (2021) Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell*, **184**, 810–826.
50. Weisenfeld,N.I., Kumar,V., Shah,P., Church,D.M. and Jaffe,D.B. (2017) Direct determination of diploid genome sequences. *Genome Res.*, **27**, 757–767.
51. Domcke,S., Hill,A.J., Daza,R.M., Cao,J., O'Day,D.R., Pliner,H.A., Aldinger,K.A., Pokholok,D., Zhang,F., Milbank,J.H. *et al.* (2020) A human cell atlas of fetal chromatin accessibility. *Science*, **370**, eaba7612.
52. Zhang,K., Hocker,J.D., Miller,M., Hou,X., Chiou,J., Poirion,O.B., Qiu,Y., Li,Y.E., Gaulton,K.J., Wang,A. *et al.* (2021) A single-cell atlas of chromatin accessibility in the human genome. *Cell*, **184**, 5985–6001.
53. 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.
54. 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.
55. Hafemeister,C. and Satija,R. (2019) Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.*, **20**, 296.
56. Bland,J.M. and Altman,D.G. (1995) Multiple significance tests: the bonferroni method. *BMJ*, **310**, 170.
57. Mabbott,N.A., Baillie,J.K., Brown,H., Freeman,T.C. and Hume,D.A. (2013) An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC Genomics*, **14**, 632.
58. 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.
59. Crow,M., Paul,A., Ballouz,S., Huang,Z.J. and Gillis,J. (2018) Characterizing the replicability of cell types defined by single cell RNA-sequencing data using metaneighbor. *Nat. Commun.*, **9**, 884.
60. Efremova,M., Vento-Tormo,M., Teichmann,S.A. and Vento-Tormo,R. (2020) CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.*, **15**, 1484–1506.
61. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
62. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B (Methodological)*, **57**, 289–300.
63. Linderman,G.C., Zhao,J., Roulis,M., Bielecki,P., Flavell,R.A., Nadler,B. and Kluger,Y. (2022) Zero-preserving imputation of single-cell RNA-seq data. *Nat. Commun.*, **13**, 192.
64. Klema,V. and Laub,A. (1980) The singular value decomposition: its computation and some applications. *IEEE Trans. Autom. Control*, **25**, 164–176.
65. Dimitrov,D., Türei,D., Garrido-Rodriguez,M., Burmedi,P.L., Nagai,J.S., Boys,C., Ramirez Flores,R.O., Kim,H., Szalai,B., Costa,I.G. *et al.* (2022) Comparison of methods and resources for cell–cell communication inference from single-cell RNA-Seq data. *Nat. Commun.*, **13**, 3224.
66. Türei,D., Valdeolivas,A., Gul,L., Palacio-Escat,N., Klein,M., Ivanova,O., Ölbei,M., Gábor,A., Theis,F., Módos,D. *et al.* (2021) Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol. Syst. Biol.*, **17**, e9923.
67. Safran,M., Rosen,N., Twik,M., BarShir,R., Stein,T.I., Dahary,D., Fishilevich,S. and Lancet,D. (2021) In: Abugessaisa,I. and Kasukawa,T. (eds). *Practical Guide to Life Science Databases*. Springer Nature Singapore, Singapore, pp. 27–56.
68. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
69. Schneider,V.A., Graves-Lindsay,T., Howe,K., Bouk,N., Chen,H.-C., Kitts,P.A., Murphy,T.D., Pruitt,K.D., Thibaud-Nissen,F., Albracht,D. *et al.* (2017) Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.*, **27**, 849–864.
70. Zheng,G.X.Y., 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.