A human ensemble cell atlas (hECA) enables *in data* cell sorting

Sijie Chen[#1], Yanting Luo[#1], Haoxiang Gao[#1], Fanhong Li[#1], Jiaqi Li[#1], Yixin Chen[1], Renke You[2], Minsheng Hao[1], Haiyang Bian[1], Xi Xi[1], Wenrui Li[1], Weiyu Li[2], Mingli Ye[2], Qiuchen Meng[1], Ziheng Zou[1], Chen Li[1], Haochen Li[3], Yangyuan Zhang[1], Yanfei Cui[1], Lei Wei[1], Fufeng Chen[2], Xiaowo Wang[1], Hairong Lv[1,2], Kui Hua[1], Rui Jiang[1], Xuegong Zhang*[1,3,4]

1 Bioinformatics Division of BRNIST and Department of Automation, MOE Key Lab of Bioinformatics, Tsinghua University, Beijing 100084, China

2 Fuzhou Institute of Data Technology, Changle, Fuzhou 350200, China

3 School of Medicine, Tsinghua University, Beijing 100084, China

4 School of Life Sciences, Center for Synthetic and Systems Biology, Tsinghua University, Beijing 100084, China

[#] These authors contributed equally in this work and should be considered as the joint first authors.

[*] Corresponding author, email: zhangxg@tsinghua.edu.cn

SUMMARY

The significance of building atlases of human cells as references for future biological and medical studies of human in health or disease has been well recognized. Comparing to the rapidly accumulation of single-cell data, there has been fewer published work on the information structure to assemble cell atlases, or on methods for using reference atlases once they are ready. Most existing cell atlas work organize single-cell gene expression data as a collection of individual files, allowing users to download selected data sheets, or to annotate

query cells using models pretrained with the collected data. These features are useful as the basic use of cell atlases. More comprehensive uses of global cell atlases can be developed once data of cells from multiple organs across different studies can be assembled into one orchestrated data repository rather than a collection of data files. For this purpose, we presented a unified giant table or uGT to store and organize single-cell data from multiple studies into a single huge data repository, and a unified hierarchical annotation framework or uHAF to annotate cells from uncoordinated studies. Based on these technologies, we developed a system that enables users to design complex rules to recruit from the atlas cells that meet certain conditions, such as with desired expression range of a gene or multiple genes and with required organ, tissue origins or developmental stages, across multiple datasets that were otherwise unconnected. The conditions can be expressed as sophisticated logic criteria to pinpoint specific cells that cannot be easily spotted in traditional *in vivo* or *in vitro* cell sorting or in traditional searching in published data. We name this technology as *in data* cell sorting from cell atlases. With the increasing coverage of the cell atlas, this *in data* experiment paradigm will facilitate scientists to conduct investigations in the data space beyond the restrictions in traditional *in vivo* and *in vitro* experiments. In the current work, we collected scRNA-seq data of more than 1 million human cells from scattered studies and assembled them as a human Ensemble Cell Atlas or hECA using the proposed information structure, and provided comprehensive tools for *in data* experiments based on the atlas. Case examples on agile construction of atlases of particular cell types and on off-target prediction of targeted therapy showed that *in data* cell sorting is an efficient and effective way for comprehensive discoveries. hECA provides a powerful platform for assembling massive scattered single-cell data into a unified atlas, and can serve as a prototype for building future cell atlases.

INTRODUCTION

Cells are the basic unit of the human body. Cells of different types at different tissues and organs of the human body are characterized by their different molecular features, especially

transcriptomic features. Building the molecular atlas for all cell types of the human body will provide a fundamental reference for all future biological studies of the human body in health or disease. The international Human Cell Atlas (HCA) initiative and the US HuBMAP project are two major efforts for building such references (Regev *et al.*, 2017; Snyder *et al.*, 2019), among several other projects aiming at similar or related goals. These big consortiums have involved labs worldwide in generating data, but coordination of the efforts and data is challenging especially when many labs have their own research themes besides atlas building and are supported by different agencies (Arazi *et al.*, 2019; Azizi *et al.*, 2018; Bayraktar *et al.*, 2020; Chevrier *et al.*, 2017; Cillo *et al.*, 2020; Corridoni *et al.*, 2020; Fernandez *et al.*, 2019; Grubman *et al.*, 2019; Vieira Braga *et al.*, 2019; Villani *et al.*, 2017; Wang *et al.*, 2020a).

The rapid development of single-cell omics technologies also propels a huge wave of single-cell studies besides those consortium efforts. The number of independent single-cell studies is growing exponentially. As a result, massive amounts of single-cell transcriptomics data have been pouring into the public domain (e.g., Han *et al.*, 2018; Han *et al.*, 2020; Schaum *et al.*, 2018; Sunkin *et al.*, 2013). Most single-cell studies generate data for their investigations on specific scientific questions instead of for building atlases. But the single-cell data from those studies are coving all major adult human organs (e.g., Aizarani *et al.*, 2019; Bayraktar *et al.*, 2020; Guo *et al.*, 2018; Han *et al.*, 2020; Litviňuková *et al.*, 2020; Pellin *et al.*, 2019) and key developmental stages (e.g., Asp *et al.*, 2019; Cao *et al.*, 2020a; Cui *et al.*, 2019; Guo *et al.*, 2020; Kernfeld *et al.*, 2018; Park *et al.*, 2020; Zhong *et al.*, 2018). The samples are from both healthy donors and patients of certain diseases (e.g., Grubman *et al.*, 2019; Reyfman *et al.*, 2019; Wang *et al.*, 2020a; Zhang *et al.*, 2020). These studies are uncoordinated and the data are scattered, but the amount of data generated overwhelms that of the HCA or HuBMAP projects. These massive single-cell data scattered in the public domain provde great potential for gradually building up a global reference of all human cells in a “shot-gun” manner, if they can be assembled in a proper way.

The significance of building a reference of all human cells is comparable to, if not greater

than, that of building a reference of all human genes in the human genome project (HGP). But the complexity of information in cell atlases is magnitudes higher than that of the genome. A cell has multifaceted properties in its anatomical locations, physiological functions, developmental time and cellular states (Ding *et al.*, 2019; Sokolov *et al.*, 2016; Trapnell *et al.*, 2014; Zhang *et al.*, 2019). The molecular information of each cell, given the current level of scRNA-seq technologies, is already of the dimensionality in ~10,000s or higher, even before counting in other omics information. There is no "natural" framework like the 1-dimensional backbone of the genome that can allow cells to be aligned to (Regev *et al.*, 2017). This nature of complexity makes the assembly of single-cell data challenging even for data obtained from studies with well-coordinated top-down design (Han *et al.*, 2020; Pisco and Consortium, 2021; Snyder *et al.*, 2019; Weber *et al.*, 2020). There have been efforts to collect single-cell data from multiple coordinated or independent studies and to "integrate" them as unified resources (e.g., Abugessaisa *et al.*, 2018; Cao *et al.*, 2017; Cao *et al.*, 2020b; Srivastava *et al.*, 2018; Svensson *et al.*, 2020). But all their data "integrations" are at the dataset level rather than cell level: Data from different studies and sub-studies are achieved as separated files rather than in a single database; Databases are only used to manage or index the metadata of the individual datasets. The typical way to use these resources is to find specific datasets from the list and download the corresponding data files for further in-house analyses. In this sense they are only collections of data but not integrations of data. Such collections provided the convenience of finding the needed data from multiple sources in one stop, but did not provide integration or assembly. There has been no reported effort yet in assembling massive single-cell data of multiple studies into a unified repository, and the question of possible underlying information structures to organize and annotate all cells in an comprehensive atlas has not been sufficiently studied (Bard *et al.*, 2005; Börner *et al.*, 2020; Diehl *et al.*, 2016; El-Achkar *et al.*, 2021; Franzén *et al.*, 2019; Rood *et al.*, 2019; Zhang *et al.*, 2018).

To better utilize the massive amount of scattered single-cell data in the literature, and to explore a prototype for future full-scale cell atlases, we developed a system to seamlessly integrate uncoordinated data and assemble them into a unified reference atlas of human cells.

We named the system as the *human Ensemble Cell Atlas* or hECA. hECA uses of a unified giant table or uGT that facilitates the storage of millions or more cells in one single data table with each cell having tens to hundreds of thousands of features. This breaks the limitation in traditional relational databases and is the key feature for real integration of large single-cell datasets. Another key feature is the unified hierarchical annotation framework or uHAF we developed for indexing cells with multifaceted properties and unifying uncoordinated annotations across different studies. Using these two technologies, data from both well-coordinated studies and uncoordinated independent studies can be assembled in a unified data repository as if they were generated by one single study. We also designed a portraiture system for characterizing cells and genes in the atlas by sets of "quantitative portraits" of the distributions of their molecular and cellular properties. Comparing to the practice of using marker genes to characterize cell types, the portraiture representation provides more complete references of the molecular properties of cells and more complete references of cellular properties of genes. The current version of hECA (v1.0) has assembled >1 million cells from 116 published datasets, covering 38 human organs. The system is ready for much larger data sizes both in the number of cells and the number of features in each cell, limited only by the hardware of the computer server.

The ensemble atlas opened doors for comprehensive usage of the reference atlas. We developed tools that allow users to search for cells that meet complex criteria on the expression of multiple genes and on features of cells such as cell types, tissue origins, developmental stages, or disease status. We call this comprehensive searching as "*in data*" cell sorting. Eventually the ensemble atlas will converge to a digital being of virtual cells from the whole human body when the data reach sufficient coverage and depth. Users can conduct *in data* experiments by selecting cells needed to address specific scientific questions. The flexibility of *in data* cell sorting expands the scope of cell experiments to questions that cannot be investigated conveniently at the bench due to limitations in experimental technologies. Users can also quickly build their own customized cell atlases using *in data* cell sorting to collect cells with their special requirements, or integrate the selected cells with their

own in-house data. We conducted two example experiments to show the power of *in data* cell experiments based on the current version of hECA: One is the agile construction of a T cell atlas across multiple organs and the other is the searching for cells that could suffer off-target effects of targeted cancer therapy. The examples showed the power of *in data* cell sorting on the ensemble atlas even with the shallow coverage of current data, and illustrated the great potential of computational experiments in the data space in the future when more data are assembled into hECA.

RESULTS

uGT: a unified giant table for assembling cell atlases

The basic building block of cell atlases is a cell. Ideally, we wish to have all information of a cell in the atlas, including its genetic, epigenetic, transcriptomic, proteomic, metabolomic attributes as well as facts about the cell's anatomic compartment, spatial location, developmental time, physiological status, etc. We call all these information as features of the cell, which could be of the dimensionality of tens of thousands to millions. Transcriptomics data obtained with scRNA-seq are currently the most available and informative data. A typical dataset of a single-cell transcriptomic study is in the form of a data table called expression matrix, with each column representing a cell and each row representing the expression values of a gene in the cells. The expression values are the more suitable form of data to be assembled into atlases than the raw RNA-seq data from which the expression values are computed. Other types of single-cell omics data such as single-cell ATAC-seq data are in similar forms. Besides these features of cells, a dataset also includes some metadata about the donors, the study, the technical platforms and parameters. They are attached to the cell data as tags labeling the extra information. Annotations of a cell, such as the cell type it belongs to, can also be viewed as part of the metadata. Usually single-cell studies provide their data as one or several data sheet files (such as Excel files) of the expression matrix and meta data (**Figure 1A**).

Current platforms for collecting single-cell omics data organize the data from multiple studies as lists of such data files. This is a quick way to gather scattered data into the same archive, so that users can easily find the studies they need in a one-stop manner instead of surfing the internet. Examples of such platforms include the "Single-cell studies database" (Svensson *et al.*, 2020), "HCA Data Portal"(Regev *et al.*, 2017), "Cell Atlas Search"(Srivastava *et al.*, 2018), "EMBL-EBI Single Cell Expression Atlas" (https://www.ebi.ac.uk/gxa/sc/home), "Cell Blast"(Cao *et al.*, 2020b), "PanglaoDB" (Franzén *et al.*, 2019), "Single Cell Portal (beta)" (https://portals.broadinstitute.org/single_cell), and scRNASeqDB (Cao *et al.*, 2017), etc. Some of these platforms defined their own processing pipelines and provide data files after their reprocessing. Some have summarized marker genes in the collected studies and facilitated users to find datasets in which a query marker gene is expressed. Some have trained cell type classification models using the collected data to predict the cell type of a query cell. These platforms have provided primary examples for the use of cell atlases. But the collected data have not been assembled into a unified data repository for more comprehensive applications. Relational databases (SQL databases) are used to index the datasets rather than to manage the data of each cell. For projects that generated data of many organs or developmental models in the same studies (e.g., Cao *et al.*, 2020a; Han *et al.*, 2020; Yu *et al.*, 2021), their data were organized in similar manners.

Under this data management scheme, if researchers for example want to study a type of cells from multiple organs using published data, such as the work on multi-organ endothelial cells (ECs) (Kalucka *et al.*, 2020) , they would need to go to some collective data repository, search for the cell type to get a list of related datasets of different sizes, download and re-process each of them on their local computers, and try to merge them into a new unified gene expression matrix manually or with in-house software. Computational analyses can then be applied on the merged matrix. If the researchers need to study a subset of the cells that satisfy certain criteria, they need to write codes to search in the matrix. Such work can be tedious and error prone, and is a barrier for many labs to conduct comprehensive studies involving data

from multiple sources.

This type of integrative data analyses across multiple datasets can be made more efficient if the collected data are stored in databases with Structured Query Languages (SQL), which allow users to use query commands to retrieve subsets of data they need. However, traditional SQL databases have hard limitations on the "width" of each data item, although they can store limitless amount of data items. For example, the limit on the number of columns is 1000 for Oracle™ and 4096 for MySQL™, which already reached the theoretical upper limit (MySQL, 2021; Oracle, 2021). Obviously, the number of features of each cell in a cell atlas exceeds this limit by several magnitudes.

We designed the unified giant data table uGT using the NoSQL database technology (Stonebraker, 2010; Wang *et al.*, 2017) to integrate data from multiple studies into one data repository in the cloud. With this technology, the uGT can store any number of features of mixed data types for any number of cells, given sufficient storage capacity in the cloud. **Figure 1C** illustrates the current structure of the uGT. It can be further expanded when more features from other omics data are ready to be integrated. The uGT not only provides a unified storage for data from many studies, but also supports comprehensive queries such as logic combinations of conditions on the expression of multiple genes plus the conditions on the metadata.

We collected 1,093,299 cells from 116 published datasets to build the current release of hECA (v1.0) and stored them all in the uGT. The sources of the data include large projects such as the Human Cell Landscape (Han *et al.*, 2020) and Allen Brain Atlas (Sunkin *et al.*, 2013), as well as smaller datasets in many other publications. The method we used for collecting the data is described in the **Supplementary Methods** and details of the data sources are given in **Table S1**. The data cover 38 human organs and 11 systems (integumentary, endocrine, urinary, cardiovascular, lymphatic, nervous, respiratory, digestive, muscular, reproductive and

skeletal systems). They are annotated as 146 cell types using the uHAF system (see **Data S1**). **Table 1** summarizes the numbers of collected cells in each organ. uGT provides a powerful facility to search for cells using complicated conditions seamlessly among all the collected data. We developed a data access interface ECAUGT (pronounced as "e-caught") for this facility (see **Supplementary Methods**). For example, the combinatorial query of "all T cell subtypes located in heart with PTPRC positive and CD3D or CD3E positive" can be written as the following logic expression:

(organ==Heart) && (cell_type == T cell || cell_type == CD8 T cell || cell_type == Activated T cell || cell_type == NK T cell || cell_type == Proliferating T cell) && (PTPRC > 0.5) && (CD3E >= 0.5)

hECA will return all cells that satisfy these conditions in a single downloadable file to users for further analysis, no matter which study they are from. Information about the particular studies of the cells are also provided to the users. **Table S2** provides the syntax of the logic expressions in ECAUGT.

Table 1. Summary of cells collected in the organs in hECA v1.0

#	Organ	# of cells	#	Organ	# of cells
1	Adipose	1,362	20	Oesophagus	87,947
2	Adrenal gland	15,065	21	Ovary	6,927
3	Bladder	3,980	22	Pancreas	26,566
4	Blood	29,514	23	Placenta	9,926
5	Bone marrow	8,671	24	Pleura	19,695
6	Brain	214,314	25	Prostate	2,445
7	Bronchi	12,553	26	Rectum	5,718
8	Colon	22,919	27	Rib	5,907
9	Duodenum	3,743	28	Skin	6,618
10	Eye	47,275	29	Spinal cord	4,483
11	Gallbladder	14,733	30	Spleen	15,806
12	Heart	210,597	31	Stomach	22,187
13	Ileum	3,132	32	Testis	13,210
14	Intestine	41,851	33	Thymus	4,516
15	Jejunum	4,198	34	Thyroid	12,599
16	Kidney	45,368	35	Ureter	2,205
17	Liver	26,475	36	Uterine tube	6,496
18	Lung	90,521	37	Uterus	8,096
19	Muscle	26,029	38	Vessel	9,652

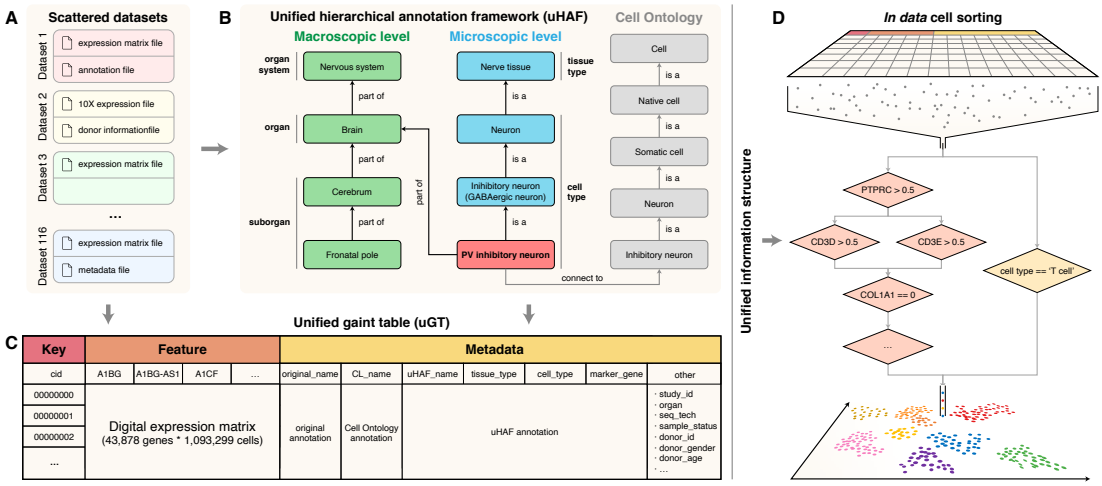

Figure 1. Overview of hECA structure and the scheme of *in data* cell sorting. (A) Collection of multiple datasets. (B) The unified hierarchical annotation framework uHAF. (C) The unified giant table uGT. (D) Illustration of the *in data* cell sorting experiments.

uHAF: a unified hierarchical annotation framework

An index or coordinate system is needed to assemble cells into an atlas so that cell annotations from different studies can be aligned. The human genome has a natural one-dimensional backbone to which all sequence segments are aligned for assembling the complete genome. The cell atlas is much more complicated. Rood et al proposed the principle to design a common coordinate framework (CCF) so that the relative "geographic" coordinates in different individuals can be aligned (Rood *et al.*, 2019). It is an important goal, but a challenging one due to the diversity and dynamic nature of the sizes, shapes and positions of organs in human bodies. Relative spatial layout of some cell types is important for the physiology and pathology in some tissues, but for many organs and tissues, the specific position of each single cell is not deterministic. Therefore, indexing cells by their 3D coordinates is not feasible or necessary. The HuBMAP project proposed an ASCT+B system to index cells, with information of anatomical structures and cell types plus the expression of biomarkers (El-Achkar *et al.*, 2021). We built an expandable unified hierarchical annotation framework uHAF with a similar idea. We unified the information of anatomical structures, source organs, and cell types into a unified hierarchical knowledge graph. uHAF can index

cells at multiple granularities, depending on the resolution provided in the original studies that generated the data. Spatial information can be added to the uHAF annotation in the future when spatial transcriptomics data are integrated into the atlas, as uHAF is designed to be flexible for extending the granularity of the tree.

We defined two types of entities using a controlled vocabulary, composing two subgraphs in uHAF. Entities in the "macroscopic subgraph" include organ, anatomical region and subregion information. Entities in the "microscopic subgraph" include annotations of cells on their histological types and cell types or subtypes determined by molecular features. **Figure 1B** shows the basic structure of uHAF. We defined two types of edges in the uHAF graph, "part of" and "is a", to represent the hierarchical relations among the entities, and an extra "connect to" type of edge to tag attributes of the entities. For example, there is a "part of" edge from the entity "left ventricle" to the entity "heart", and there is an "is a" edge from the entity "inhibitory neuron" to the entity "neuron". The entities in the macroscopic and the microscopic subgraph are organized in a hierarchical tree (details in **Supplementary Methods**) by manually surveying the canonical human anatomy structure and cell type names from classical medical textbooks (Bai and Ying, 2015; Li and Zeng, 2018; Mescher, 2016; Tang and Zhang, 2013) followed by confirmation and refinement from medical experts. The microscopic entities are attached with attributes "marker reference" consisting of the expression of marker genes by the "connect to" edges. There are "part of" connections from nodes in the microscopic subgraph to nodes in the macroscopic subgraph, indicating the cell type composition of a macroscopic entity. For example, the entity "T cell" has a "part of" connection with the entity "left ventricle", as well as "part of" connections to other anatomical units that have T cells in their tissues. Each cell in hECA is annotated with two entities of the uHAF, one macroscopic and one microscopic. The annotation can be at different levels in the two hierarchies, depending on the information provided by the original data and the specificity of the marker genes. Users of hECA can use the uHAF to annotate their query cells in the same way. **Table S3** listed the entity combinations that have been used in annotating the existing data in the current version of hECA. uHAF allows for expansion of

the knowledge graph at all levels when more data are being integrated.

The annotations of cell types in the scattered publications have not been coordinated. This is also observed even in some data generated from the same study when the scale of the study is large. We use uHAF to unify all annotations. We established unified marker references for all microscopic entities (see **Supplementary Methods**). We applied clustering analysis on each collected dataset, and compared the gene expression features of each cluster found by differential expression (DE) gene analysis with the marker references to redo annotation for each cell (details described in **Supplementary Methods**). All cells are re-annotated in this unified manner according to the uHAF before being assembled into the uGT. The uHAF annotations compose the major part of metadata for each cell in uGT (**Figure 1C**). Using uHAF as the structured indexing system, we assembled scattered data from uncoordinated studies into the ensemble cell atlas.

In data cell sorting enabling comprehensive cell experiments that cannot be conducted *in vivo* or *in vitro*

Cell sorting is a fundamental technique in cell biology. It uses flow cytometry to select cells with specific properties from biological samples. A typical cell sorting workflow is composed of multiple filtering steps. For example, to obtain the regulatory T cells (Treg) from certain type of human tissue sample, one needs to use the marker protein PTPRC (also known as CD45) to distinguish immune cells (PTPRC+) from other lineages of cells (PTPRC-), use CD3 to select the T cells (PTPRC+ and CD3+) from the PTPRC+ cells, and then use CD4, IL2RA (also known as CD25), and FoxP3 markers to filter out other T cells and get the Treg cells. The types of cells that can be selected depend on the availability and identifiability of surface markers of the cells under study, and the discriminating power of the flow cytometry technology.

The unified assembly and indexing of all cells in hECA allow for a more powerful technique for virtually searching cells with any desired conditions in their molecular features and metadata attributes. The ECAUGT allows users to define logical expressions on any features and metadata or their combinations to implement "*in data* cell sorting". Such *in data* cell sorting is no longer restricted to selection criteria on surfer markers, but can be defined in a more precise way on the expression of any genes. And the selection range is also no longer restricted with the biological samples that can be collected in one particular study, but can be extended across all cells with the desired properties in the whole human body or in a set of organs. In future versions of hECA that include epigenomic features such as chromatin openness measured by scATAC-seq, the searching will be naturally extended to allow for cells with desired epigenomic features in selected regions of the genome. Eventually hECA will become a digital being composed of cells of all human organs, alive with all their cellular and molecular characteristics and interactions. It'll be a "digit-life twin system" of a living human body. Designing cell experiments becomes a matter of designing a regular logic expression for searching hECA. This opens the new paradigm in cell biology: *in data* cell sorting from the digit-life twin system followed by *in silico* computational experiments, once the data in hECA are rich and deep enough to represent cellular and molecular features of all major cell types and major human organs.

The *in data* cell sorting facility of hECA is provided as the Python package ECAUGT for bioinformaticians at https://pypi.org/project/ECAUGT, and as an interactive web server for general users at http://eca.xglab.tech/#/cellSorting. Using the facility to search for all T cells in the heart with normalized expression of gene PTPRC greater than 0.5 and that of CD3E great than 0.5, for example, one can simply write the logic expression in python as the input of ECAUGT:

```
rows_to_get = ECAUGT.query_cells("organ==Heart && cell_type == T cell", include_children=TRUE)
gene_condition = ECAUGT.seq2filter("PTPRC > 0.5 && (CD3D>=0.5 || CD3E>=0.5)")
ECAUGT.get_columnsbycell_para(rows_to_get = rows_to_get, cols_to_get=['CD3E','PTPRC'],
col_filter=gene_condition
```

hECA will return the selection results (of around 210,000 cells in the current version) in about 190 seconds.

Case study 1: agile creation of a draft T-cell metabolic landscape

We use two cases as examples of how hECA can be used to empower comprehensive studies of human cells across the human body.

T lymphocytes or T cells are an essential type of cells in the human immune system. They must adapt to multifarious microenvironments as they circulate through or reside in different human tissues in different organs. Their differentiation, activation and quiescence are regulated by the diverse metabolites in different local tissues (Buck *et al.*, 2015; Chapman *et al.*, 2020; Shyer *et al.*, 2020; Yin *et al.*, 2019). For example, recent studies reported that microbial bile acid metabolites promoted the generation of regulatory T cells in the intestine, which is associated with inflammatory bowel disease (IBD) (Campbell *et al.*, 2020; Hang *et al.*, 2019; Song *et al.*, 2020), suggesting that targeting metabolic pathways of T cell activation and differentiation may improve therapeutic outcomes of IBD patients (Li *et al.*, 2021). Comprehensive survey of the metabolism of T cells across multiple organs is crucial for better understanding intrinsic responses of T cells to microenvironment changes, but *in vivo* or *in vitro* experiments on multiple organs are not easy. Xiao et al proposed a computational pipeline to study the metabolic landscape of cells from single cell transcriptomic data (Xiao *et al.*, 2019). The unification of cells of all types in all organs in hECA allows us to conduct study on T cell metabolism across all organs *in data*, without the need to search through datasets scattered in the literature.

Using the facilities provided by hECA, we first found all cells in uGT with label of “T cell” and its associated names (such as “CD4 T cell”, “CD8 T cell”, “Activated T cell”, etc.) across all recorded organs (**Figure S1A**). To include cells that might have been annotated to other

cell types, we also searched for cells with normalized expression values of PTPRC, CD3D or CD3E greater than 0.5 across all organs (**Figure S1B, S1C**). Then we filtered the cells by the expression of a list of negative markers such as COL1A1, CD79A (the full list provided in **Table S4**). We conducted clustering analysis on cells from the same organs, and obtained a series of candidate clusters in each organ (**Figure 2A**). We removed clusters with low expression levels of CD3D, CD3E or CD3G as they are unlikely to be T cells. With these steps, we built an agile cell atlas of T cells from 18 organs (lung, pancreas, blood, liver, muscle, thymus, jejunum, rectum, colon, kidney, gallbladder, stomach, thyroid, intestine, spleen, bone marrow, eye, and vessel).

To assign finer annotations to the cells in the T cell atlas, we performed hierarchical clustering on them using signature genes CD4, CD8A and CD8B, and divided the cells into 6 subgroups of 3 major groups (**Figure S2A**). The three major groups are CD4+, CD8+ and double-negative (CD4- and CD8-) T cells (**Figure S2B**). For the CD4+ and CD8+ groups, we further annotated the cells as resident memory T cells, central memory T cells, effector memory T cells, naïve T cells, cytotoxic T cells, etc. according to the positive markers listed in **Table S5**. **Figures 2B, 2C** show the UMAP of the CD4+ and CD8+ T cells with the subtype annotations and with the organ origin of the cells, respectively. **Figure 2D** shows the gene expression signatures of the identified T cell subtypes. For the double-negative cluster, we marked them as “T cells” without further analysis as there might be cells false negatives in CD4 or CD8 expression due to possible dropout events in scRNA-seq data.

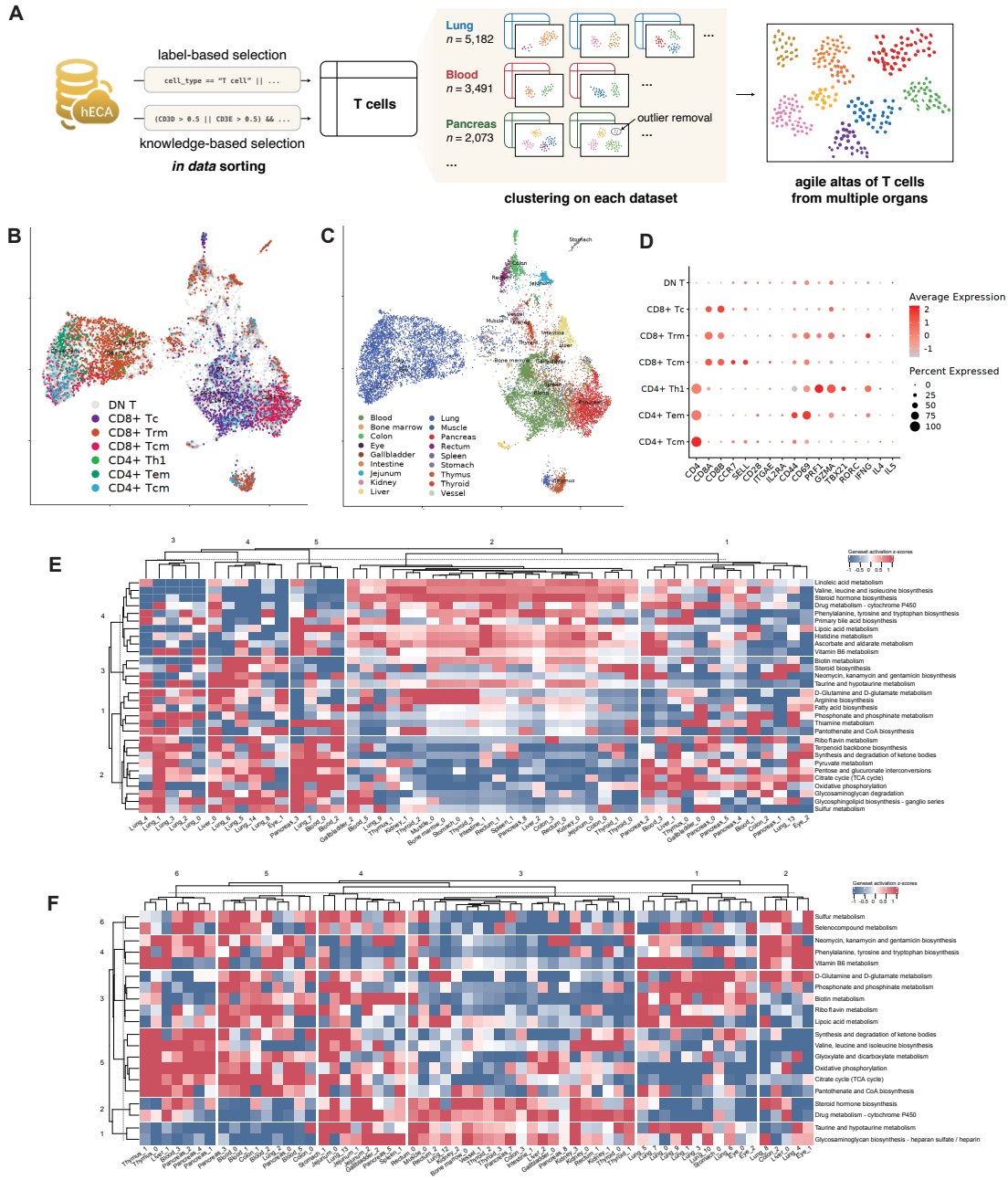

Figure 2. The agile creation of a draft T-cell metabolic landscape across multiple organs from hECA. (A) Workflow of the *in data* cell sorting from hECA to build the agile T cell atlas. (B) Subtypes of selected T cells displayed on UMAP. DN T: Double negative T cell, CD8+ Tc: CD8+ Cytotoxic T cell, CD8+ Trm: CD8+ resident memory T cell, CD4+ Th1: CD4+ T helper cell type 1, CD4+ Tem: CD4+ effector memory T cell, CD4+ Tcm: CD4+ central memory T cell. (C) Organ origins of selected T cells organ origin displayed on UMAP. (D) Gene expression signatures of the identified T cell subtypes. (E,F) Heatmaps showing z-scores of activity scores of major metabolic pathways of the T cell subtypes in multiple organs. (E) for CD4+ T cells and (F) for CD8+ T cells. Each row in the heatmap

corresponds to one selected term in the KEGG metabolism pathway database, and each column corresponds to one T cell subcluster.

For a sketchy study on the metabolic landscape of T cells across multiple organs, we evaluated each cell's metabolic activity scores with GSVA, which produced comparable values across multiple clusters or datasets and alleviated possible batch effects in the data from multiple sources (Hänzelmann *et al.*, 2013). The genes of the metabolic pathways are derived from KEGG (Kanehisa *et al.*, 2020) and Xiao et al's work (Xiao *et al.*, 2019). A heatmap of the obtained draft metabolic landscape of T cells of their activity scores of all major metabolic pathways across the human body are shown in **Figures 2E** and **2F**. Such landscapes can help to reveal different metabolic patterns across organs. For example, we found organ-level metabolic variations in lungs from the metabolic activities of organ-level CD4+ T cell clusters in **Figure 2E** and those of the organ-level CD8+ T cell clusters in **Figure 2F**. For CD4+ T cells, we observed lung-enriched metabolic pathway activations in the pathways of riboflavin metabolism, terpenoid backbone biosynthesis, TCA cycle, oxidative phosphorylation, sulfur metabolism, and D-Glutamine and D-glutamate metabolism (row blocks 1 & 2 of the lung-origin T cell clusters in **Figure 2E**). Similar enrichments can also be observed in the lung-origin CD8+ T cell clusters in **Figure 2F**. These observations deserve further investigations, but showcased the great potential of cross-organ *in data* cell experiments enabled by hECA.

Case study 2: *in data* discovery of side effects in targeted therapy

A great part (~97%) of cancer drugs tested in clinical trials failed to get approval from FDA, mainly due to their insufficient efficacy or unexpected toxicities to organs on which the drugs were not designed to take effect (Lin *et al.*, 2019). Off-target effects are usually not easy to be observed in trials on animal models. Prediction of cellular toxicities across the whole human body can significantly help reducing improper clinical trials and increasing the efficiency of discoveries of new drugs. This is a typical scenario that we need a digit-life twin system that can allow testing drugs *in data* before clinical trials.

In previous research, computational investigation of off-target effects or neurotoxicity effects of targeted therapy requires multiple steps. Usually, a group of organs must first be chosen as suspects of side effects based on existing knowledge or educated guess. For these organs, researchers need to review the literature to find and collect proper single-cell datasets. The goal is to find cell types in the suspected organs in which the target gene of the candidate drug is highly expression, and evaluate the effect of the drug on the phenotype of those cells and therefore on the phenotype of the organs. Data collected from scattered publications can be of different formats, sequencing platforms, processing and annotation methods. Researchers need to design a pipeline to convert the data and re-process them in a standardized way for the study. As an example, Parker et al found that CD19+ mural cells in the human brain were potential off-tumor targets of CAR-T therapy (Parker *et al.*, 2020). They first noticed from previous literature that CD19 CAR-T therapy could introduce neurologic adverse reactions. Then they collected 3 single cell datasets of the brain: that of the human prefrontal cortex (Zhong *et al.*, 2018), that of the human forebrain (La Manno *et al.*, 2018) and that of the human ventral forebrain (La Manno *et al.*, 2016). After applying standardized reprocessing on each dataset, cells were manually annotated by comparing highly enriched genes to known cell-type markers. They observed from the UMAP a small population of cells in the first dataset that expressed both CD19 and CD248 (a marker for mural cells). They further identified that these cells were pericytes and verified them in all three datasets.

This type of investigations can be conducted in a more efficient and effective way with the ensembled cell atlas of all major human organs and *in data* cell sorting facilitates of hECA. For any specific target gene, cells that highly express the gene can be found through *in data* cell sorting, no matter which original datasets the cells are from. A profile of cellular distribution of all major human organs that contain the found cells can be built, which highlights suspected organs that might be the off targets of the drug. Detailed analyses can be applied on the possible effects of the drug on the phenotypes of the cells by checking on the consequences of the expression change of the target gene on downstream gene expression,

signaling pathways, metabolisms, and interactions with other cells, etc. Quantitative analyses then can be applied on the cell compositions and cell-cell interactions in the suspected organs to evaluate the possible effect on the physiology or pathology.

Using this idea, we followed the example of Parker's work (Parker *et al.*, 2020) to study the possible off-target effects of CAR-T therapy using the data and facilities of hECA. CD19 is a usually target of CAR-T therapy when curing B-cell lymphoma (Wei *et al.*, 2019). Neurological toxicity is one of the main side-effects (Rubin *et al.*, 2019). To study why this toxicity occurs and what other organs might be affected by this CAR-T therapy, we used a filtering criterion on CD19 expression for *in data* cell sorting in hECA. Totally 2,566 CD19+ cells passed the filter (**Figure 3B**). This therapy aims to target malignant B cells for curing lymphoma. But B cells and plasma B cells only compose ~53% of the selected CD19+ cells (**Figure 3C, Figure S4, Table S6**). The other cells in the selected group include endothelial cells, microglia and neurons in the brain, cardiomyocytes and fibroblasts in the heart and lung, enterocytes in the rectum, etc. (**Figure 3D, Figure S4, Table S6**). They all are cells that have the potential to be off-targets of the therapy. This result explains why encephalopathy was often observed and cells constructing vessels were targeted by the drug (Parker *et al.*, 2020). Our results also suggest that there is possible toxicity on the circulatory system and digestive system, which has also been reported (Yáñez *et al.*, 2019).

CD22 is also a popular target when designing CAR-T therapy for lymphoma (Wei *et al.*, 2019). Similarly, we used *in data* cell sorting in hECA and obtained 8,724 cells with CD22 expressed (**Figure 3E**). Besides B cells (**Figure 3F, Figure S5, Table S7**), this group contains oligodendrocytes and excitatory neurons in the brain, cardiomyocytes and fibroblasts in the heard, macrophage, mast cells and monocytes in the lung, and neutrophils in the testis, etc. (**Figure 3G, Figure S5, Table S7**). These observations provide significant clues for systematic investigation on the potential side effects of targeted therapy.

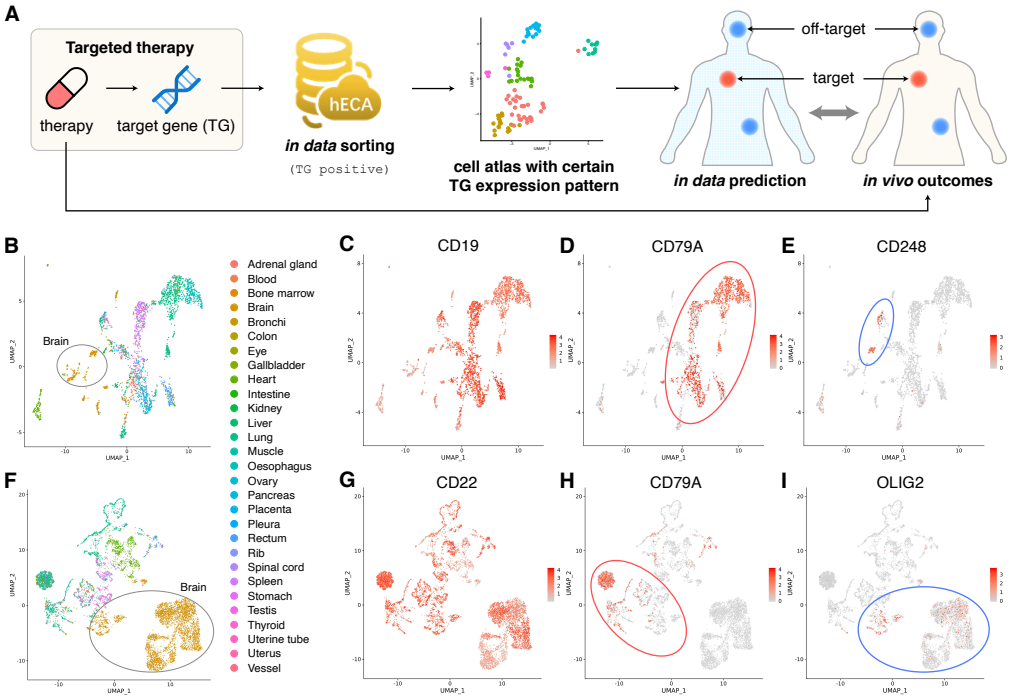

Figure 3. *In data* experiments with hECA facilitating discoveries of side effects of targeted drugs. (A) The diagram of using *in data* cell sorting to predict targets and off-targets of targeted therapy. Red dot and blue dot in the human body represent the target effected sites and side effected sites, respectively. The red and blue dot in the UMAP represent the treatment effected cells and side effected cells, respectively. (B) Visualization of CD19+ cells (expression>0.1) in UMAP, colored by organ origins of cells. CD19 is the target gene of the targeted therapy. (C) Visualization of CD19 expression levels of those CD19+ cells. (D) Visualization of CD79A expression levels of those CD19+ cells. CD79A is a marker for B cells. (E) Visualization of CD248 expression levels of those CD19+ cells. CD248 is a marker for pericytes. (F) Visualization of CD22+ (expression>0.1) cells in UMAP, colored by organ origin of cells. CD22 is the target gene of the targeted therapy. (G) Visualization of CD22 expression levels of those CD19+ cells. (H) Visualization of CD79A expression levels of those CD22+ cells. CD79A is a marker for B cells. (I) Visualization of OLIG2 expression levels of those CD22+ cells. OLIG2 is a marker for oligodendrocytes. The color bars in (C-E) represent expression levels of CD19, CD79A and CD248, and the color bars in (G-I) represent expression levels of CD22, CD79A and OLIG2, respectively, with colors grey to red indicating expression low to high. The red and blue ellipses in (D-E) and (H-I) line out the target affected cells and off-target affected cells, respectively.

A unified framework for quantitative portraiture of genes, cell types and organs

A basic usage of a global reference of molecular features of cells in the human body is to provide complete pictures of quantitative characteristics of genes, cell types and organs at all possible levels. We developed a unified quantitative portraiture framework in hECA to summarize these characteristics and to portray them in the interactive graphic user interface (GUI). The current coverage and quality of the data are still far from fully characterizing the entities in an unbiased manner. Therefore, current portraits can reflect information in the collected data rather than the complete biological picture. However, with the growing data coverage and quality in data sampling in the future, the portraiture framework will approach producing a holographic view of genes and cells in the whole human body.

The quantitative profiling of a gene's expression in different types of cells in all major organs is important for portraying the gene. hECA v1.0 provides the portraits of 43,878 genes (HGNC gene symbols) based on the currently archived data. The portrait includes basic information about the gene, links to GeneCard, NCBI, Ensemble and Wikigene pages of the gene, and profiles of its expression in cell types and organs that are covered in the current data.

A cell type is mainly characterized by two types of information: the organs that contain the cell type, and the expression patterns of genes that are specific to the cell type. hECA v1.0 portrays 146 of the 416 cell types organized by the uHAF hierarchy. On the hECA website, users can type in the name to search for a cell type or to click along the tree of cell types to display the cell type portrait. It includes the distribution of the cell type across organs, shown as the number of cells of this type collected in the organs, the list of marker genes with their characteristic expression ranges in the cell type, and a 2D PCA, UMAP or DensMAP visualization (McInnes and Healy, 2018; Narayan *et al.*, 2021; Pearson, 1901) of the cells colored by the organ of the cells or the expression of a certain gene in the cells.

An organ is usually characterized by its anatomic and physiological features, but the full portraiture of an organ should include its full cellular and molecular features at multiple resolutions. The basic cellular information is the relative composition of cell types in the organ and in its different anatomical parts. And the basic molecular information is the gene expression patterns in the organ as a whole and in its different parts, spatial locations, and at different physiological statuses. hECA v1.0 allow users to browse such information at a coarse level with the relative numbers of cells in each organ or sub-organ in the current collection of data, and the gene expression profiles in the organ.

It should be noted that most current single-cell sequencing technologies undergo some kind of cell selection, and for cells that are selected, the sampling efficiencies for different cell types are not uniform (Baran-Gale *et al.*, 2017; Phipson *et al.*, 2017; Tung *et al.*, 2017). There are also many technical reasons that may cause biases in the measured gene expression values even in the same experiment, let alone across different experiments (Chen and Zheng, 2018; Miao *et al.*, 2018; Miao and Zhang, 2016; Soneson and Robinson, 2018). Therefore, it is unrealistic to build full portraits of genes, cell types or organs with high reliability and fidelity based on the currently available data. But the quantitative portraiture system provides a unified framework for the task, and sets a goal that can lead the direction of future uniformization of cell atlas data.

DISCUSSION

Building a full reference of all human cells is an important scientific goal of this era, and will be a great breakthrough toward the dream of fully understanding the molecular and cellular mechanism of human body in health and diseases. Assembling pieces of data into a unified, well-structured data repository is essential for building the reference. We presented a solution framework for this through the development of uGT to store all data in one single repository and uHAF to provide universal indexing and annotation of cells from uncoordinated studies.

Based on these two techniques, we proposed the new paradigm of *in data* cell sorting that enables deep exploration of properties and behaviors of cells that meet the combinations of multiple logic conditions. Integrating more than the transcriptomic data of more than 1 million cells from published literatures, we built a human ensemble cell atlas hECA using this framework. Although the number of cells is still very small and the coverage of organs and cell types is very limited, case studies based on this primary version of hECA already showed the power of *in data* experiments and the feasibility of assembling single-cell data from many studies using the technologies in hECA.

Several low-level processing issues that are crucial for the construction of cell atlases, such as normalization and correction for possible batch effects. Nonuniform sampling of cells and of expressed genes is another issue that may poison any global analysis of atlas data. We followed the currently widely accepted protocols for the low-level processing of collected data. We are fully aware that these issues are far from perfectly solved yet, either in our own work or in the community. But we choose to not stuck by those issues but to work on the key questions in assembling and utilizing the data. These two types of questions are orthogonal and we should not wait till the ideal solution of the low-level processing problems to study the assembly and advanced application problems. On the other hand, the assembling and utilization of hECA can help to pinpoint what downstream analyses are more sensitive to pre-processing and what are not. From the case examples, we can see that although expressions of genes measured in separated experiments are not precisely comparable due to possible batch effects, *in data* cell sorting on the rough expression of some genes can already reveal important organ-specific patterns and can help to discover organs that are more prone to side effects of targeted therapy. If necessary, more advanced reprocessing methods can be applied on the selected data, optimized for the specific downstream scientific investigation. This is more feasible than trying to find general optimal solutions in data preprocessing without a specific aim in the downstream study.

The proposed gene, cell type and organ portraitures provide a powerful framework for

characterizing the full information of biological entities in a unified quantitative manner. Up to now, all single-cell data that have been ever generated for human cells are still only a tiny fraction of all human cells, and the data are also under the influence of multiple types of noises and biases. Therefore, the current portraits can only reveal properties of the collected data but cannot be expected of high fidelity for the underlying biology. However, keeping this reality in mind, users can already use these portraits as handy tools for exploring properties of genes, cell types and organs from a more complete view than traditional views. With the rapid advancement in data quantity, coverage and uniformity, the portraits will provide multiple-scale holographic views of all biological entities in the human body.

SUPPLENMENTAL INFORMATION

Table S1. Source datasets of hECA v1.0. Related to **Table 1**.

Table S2. Syntax of the logic expressions of ECAUGT in uGT. Related to **Figure 1**.

Table S3. uHAF annotation produces uHAF_name. Related to **Figure 1**.

Table S4. Details of case study 1: Negative markers for T cells. Related to **Figure 2**.

Table S5. Details of case study 1: Positive markers for T cells. Related to **Figure 2**.

Table S6. Details of case study 2: Organ and cell type table of CD19+ cells. Related to **Figure 3**. The number of CD19 expressed cells from each cell type and each organ. Red and blue highlights were treatment effect and side effect respectively.

Table S7. Details of case study 2: Organ and cell type table of CD22+ cells. Related to **Figure 3**. The number of CD22 expressed cells from each cell type and each organ. Red and blue highlights were treatment effect and side effect respectively.

Data S1. Organ cellular composition. Related to **Table 1**. Each page shown one organ and its cell type composition. The cell number and percentage were shown for each cell type. Database query was performed using the function “get_column_set()” of package ECAUGT to extract the cells for each cell type, and each cell type per organ was counted by “express.treemap()” function in “ploty” python package.

Figure S1. Filtering candidate T cell subpopulations. Related to **Figure 2**. (A) A UMAP showing the within-organ clustering results. (B) The general T cell markers' expressions (CD3D, CD3E, CD3G). (C) Per-cluster marker gene expressions.

Figure S2. CD4/CD8 T cell population definition. Related to **Figure 2**. (A) A heatmap showing the hierarchical clustering results based on the CD4/CD8A/CD8B genes. Other T cell signature genes are also listed on the heatmap. (B) Split view of CD4 positive T cells, CD8 positive T cells, and double-negative T cells (Note that sequencing dropouts caused some falsely recognized double-negative T cells).

Figure S3. Cell type and organ distribution of CD19 expressed cells. Related to **Figure 3**. The

number of CD19 expressed cells from different organs (A) and cell types (B). Visualization of CD19 expressed cells in UMAP labeling organs (C) and cell types (D).

Figure S4. Cell type and organ distribution of CD22 expressed cells. Related to **Figure 3**. The number of CD22 expressed cells from different organs (A) and cell types (B). Visualization of CD22 expressed cells in UMAP labeling organs (C) and cell types (D).

ACKNOWLEDGEMENTS

This work is supported in part by NSFC grants 60171003, 62050178, 32000453 and 42050101. Many students have participated in the early attempts of literature data collection and early explorations of demo systems, including Yi Liu, Yongzhen Yu, Yuehua Zhu, Zihao Jiang, Wuhan Qiu, Xinyi Ning, Qianshi Mei, Xiaosen Wei, Fangyuan Chen, Siting Li, Bin Zhang, Min Wu, Jiani Wang, Dongfang Mao and Xiaoxiao Nong. Dr. Aziz Khan helped to test some functions of the website. The work is also support by the Tsinghua-Fuzhou Institute of Data Technologies (TFIDT2021003).

AUTHOR CONTRIBUTIONS

XZ conceptualized and designed the project. XZ, SC, HG, YL and FL designed the study. SC led the design and implementation of uGT. YL and HG designed uHAF with inputs from SC, FL and JL. YL designed the methods for data collection and led the efforts of data collection and annotation. FL and SC coordinated the construction of the data system and the collaboration of all teams. JL developed the portraiture system and all portraits based on current data. HG, FL, JL, SC and YL designed the web visualization. YC, SC, HB and MH developed the ECAUGT package. SC and MH conducted case study 1. HG, SC and MH conducted case study 2. RY, Weiyu Li, MY and FC developed the web user interface system and the interface with the uGT, under the supervision of H. Lv. YC, XX, HB, MH, Wenrui Li, CL, YC, H. Li and YZ curated the data under the supervision of YL. QM and ZZ participated in developing methods for data annotation, visualization and gene symbol unification method. KH, H. Lv and RJ participated in the

conceptualization of the project and discussions on strategies of implementation. RJ, KH, LW and WW participated in many technical and strategical aspects of the study. XZ, SC, YL, FL, HG and JL wrote the manuscript, with inputs from all authors. All authors read and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Abugessaisa, I., Noguchi, S., Böttcher, M., Hasegawa, A., Kouno, T., Kato, S., Tada, Y., Ura, H., Abe, K., Shin, J.W., et al. (2018). SCPortalen: human and mouse single-cell centric database. Nucleic Acids Research *46*, D781-D787. 10.1093/nar/gkx949.

Aizarani, N., Saviano, A., Sagar, Mailly, L., Durand, S., Herman, J.S., Pessaux, P., Baumert, T.F., and Grün, D. (2019). A human liver cell atlas reveals heterogeneity and epithelial progenitors. Nature *572*, 199-204. 10.1038/s41586-019-1373-2.

Arazi, A., Rao, D.A., Berthier, C.C., Davidson, A., Liu, Y., Hoover, P.J., Chicoine, A., Eisenhaure, T.M., Jonsson, A.H., Li, S., et al. (2019). The immune cell landscape in kidneys of patients with lupus nephritis. Nat Immunol *20*, 902-914. 10.1038/s41590-019-0398-x.

Asp, M., Giacomello, S., Larsson, L., Wu, C., Fürth, D., Qian, X., Wärdell, E., Custodio, J., Reimegård, J., Salmén, F., et al. (2019). A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart. Cell *179*, 1647-1660.e1619. 10.1016/j.cell.2019.11.025.

Azizi, E., Carr, A.J., Plitas, G., Cornish, A.E., Konopacki, C., Prabhakaran, S., Nainys, J., Wu, K., Kiseliovas, V., Setty, M., et al. (2018). Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. Cell *174*, 1293-1308.e1236. 10.1016/j.cell.2018.05.060.

Bai, S., and Ying, D. (2015). Systematic Anatomy(Chinese Edition), 3 Edition (People's Medical Publishing House).

Baran-Gale, J., Chandra, T., and Kirschner, K. (2017). Experimental design for single-cell RNA sequencing. Briefings in Functional Genomics *17*, 233-239. 10.1093/bfgp/elx035.

Bard, J., Rhee, S.Y., and Ashburner, M. (2005). An ontology for cell types. Genome Biol *6*, R21. 10.1186/gb-2005-6-2-r21.

Bayraktar, O.A., Bartels, T., Holmqvist, S., Kleshchevnikov, V., Martirosyan, A., Polioudakis, D., Ben Haim, L., Young, A.M.H., Batiuk, M.Y., Prakash, K., et al. (2020). Astrocyte layers in the mammalian cerebral cortex revealed by a single-cell in situ transcriptomic map. Nat Neurosci *23*, 500-509.

10.1038/s41593-020-0602-1.

Börner, K., Ellen M. Quardokus, Bruce W. Herr II, Leonard E. Cross, Elizabeth G. Record, Yingnan Ju, Andreas D. Bueckle, James P. Sluka, Jonathan C. Silverstein, Kristen M. Browne, et al. (2020). Construction and Usage of a Human Body Common Coordinate Framework Comprising Clinical, Semantic, and Spatial Ontologies. ArXiv *abs/2007.14474*. arXiv:2007.14474.

Buck, M.D., O'Sullivan, D., and Pearce, E.L. (2015). T cell metabolism drives immunity. Journal of Experimental Medicine *212*, 1345-1360. 10.1084/jem.20151159.

Campbell, C., McKenney, P.T., Konstantinovsky, D., Isaeva, O.I., Schizas, M., Verter, J., Mai, C., Jin, W.-B., Guo, C.-J., Violante, S., et al. (2020). Bacterial metabolism of bile acids promotes generation of peripheral regulatory T cells. Nature *581*, 475-479. 10.1038/s41586-020-2193-0.

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. (2020a). A human cell atlas of fetal gene expression. Science *370*. 10.1126/science.aba7721.

Cao, Y., Zhu, J., Jia, P., and Zhao, Z. (2017). scRNASeqDB: A Database for RNA-Seq Based Gene Expression Profiles in Human Single Cells. Genes (Basel) *8*. 10.3390/genes8120368.

Cao, Z.-J., Wei, L., Lu, S., Yang, D.-C., and Gao, G. (2020b). Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST. Nature Communications *11*, 3458. 10.1038/s41467-020-17281-7.

Chapman, N.M., Boothby, M.R., and Chi, H. (2020). Metabolic coordination of T cell quiescence and activation. Nature Reviews Immunology *20*, 55-70. 10.1038/s41577-019-0203-y.

Chen, L., and Zheng, S. (2018). BCseq: accurate single cell RNA-seq quantification with bias correction. Nucleic Acids Research *46*, e82-e82. 10.1093/nar/gky308.

Chevrier, S., Levine, J.H., Zanotelli, V.R.T., Silina, K., Schulz, D., Bacac, M., Ries, C.H., Ailles, L., Jewett, M.A.S., Moch, H., et al. (2017). An Immune Atlas of Clear Cell Renal Cell Carcinoma. Cell *169*, 736-749.e718. 10.1016/j.cell.2017.04.016.

Cillo, A.R., Kürten, C.H.L., Tabib, T., Qi, Z., Onkar, S., Wang, T., Liu, A., Duvvuri, U., Kim, S., Soose, R.J., et al. (2020). Immune Landscape of Viral- and Carcinogen-Driven Head and Neck Cancer. Immunity *52*, 183-199.e189. 10.1016/j.immuni.2019.11.014.

Corridoni, D., Antanaviciute, A., Gupta, T., Fawkner-Corbett, D., Aulicino, A., Jagielowicz, M., Parikh, K., Repapi, E., Taylor, S., Ishikawa, D., et al. (2020). Single-cell atlas of colonic CD8(+) T cells in ulcerative colitis. Nat Med *26*, 1480-1490. 10.1038/s41591-020-1003-4.

Cui, Y., Zheng, Y., Liu, X., Yan, L., Fan, X., Yong, J., Hu, Y., Dong, J., Li, Q., Wu, X., et al. (2019). Single-Cell Transcriptome Analysis Maps the Developmental Track of the Human Heart. Cell Rep *26*, 1934-1950.e1935. 10.1016/j.celrep.2019.01.079.

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. 10.1186/s13326-016-0088-7.

Ding, H., Blair, A., Yang, Y., and Stuart, J.M. (2019). Biological process activity transformation of single cell gene expression for cross-species alignment. Nat Commun *10*, 4899. 10.1038/s41467-019-

12924-w.

El-Achkar, T.M., Eadon, M.T., Menon, R., Lake, B.B., Sigdel, T.K., Alexandrov, T., Parikh, S., Zhang, G., Dobi, D., Dunn, K.W., et al. (2021). A multimodal and integrated approach to interrogate human kidney biopsies with rigor and reproducibility: guidelines from the Kidney Precision Medicine Project. Physiol Genomics *53*, 1-11. 10.1152/physiolgenomics.00104.2020.

Fernandez, D.M., Rahman, A.H., Fernandez, N.F., Chudnovskiy, A., Amir, E.D., Amadori, L., Khan, N.S., Wong, C.K., Shamailova, R., Hill, C.A., et al. (2019). Single-cell immune landscape of human atherosclerotic plaques. Nat Med *25*, 1576-1588. 10.1038/s41591-019-0590-4.

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*. 10.1093/database/baz046.

Gaublomme, J.T., Li, B., McCabe, C., Knecht, A., Yang, Y., Drokhlyansky, E., Van Wittenberghe, N., Waldman, J., Dionne, D., Nguyen, L., et al. (2019). Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. Nature Communications *10*, 2907. 10.1038/s41467-019-10756-2.

Greene, D., Richardson, S., and Turro, E. (2017). ontologyX: a suite of R packages for working with ontological data. Bioinformatics *33*, 1104-1106. 10.1093/bioinformatics/btw763.

Grubman, A., Chew, G., Ouyang, J.F., Sun, G., Choo, X.Y., McLean, C., Simmons, R.K., Buckberry, S., Vargas-Landin, D.B., Poppe, D., et al. (2019). A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. Nat Neurosci *22*, 2087-2097. 10.1038/s41593-019-0539-4.

Guo, J., Grow, E.J., Mlcochova, H., Maher, G.J., Lindskog, C., Nie, X., Guo, Y., Takei, Y., Yun, J., Cai, L., et al. (2018). The adult human testis transcriptional cell atlas. Cell Res *28*, 1141-1157. 10.1038/s41422-018-0099-2.

Guo, J., Nie, X., Giebler, M., Mlcochova, H., Wang, Y., Grow, E.J., Kim, R., Tharmalingam, M., Matilionyte, G., Lindskog, C., et al. (2020). The Dynamic Transcriptional Cell Atlas of Testis Development during Human Puberty. Cell Stem Cell *26*, 262-276.e264. 10.1016/j.stem.2019.12.005.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., et al. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. Cell *172*, 1091-1107.e1017. 10.1016/j.cell.2018.02.001.

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. 10.1038/s41586-020-2157-4.

Hang, S., Paik, D., Yao, L., Kim, E., Trinath, J., Lu, J., Ha, S., Nelson, B.N., Kelly, S.P., Wu, L., et al. (2019). Bile acid metabolites control TH17 and Treg cell differentiation. Nature *576*, 143-148. 10.1038/s41586-019-1785-z.

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics *14*, 7. 10.1186/1471-2105-14-7.

Kalucka, J., de Rooij, L., Goveia, J., Rohlenova, K., Dumas, S.J., Meta, E., Conchinha, N.V., Taverna, F., Teuwen, L.A., Veys, K., et al. (2020). Single-Cell Transcriptome Atlas of Murine Endothelial Cells. Cell *180*, 764-779.e720. 10.1016/j.cell.2020.01.015.

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., and Tanabe, M. (2020). KEGG: integrating viruses and cellular organisms. Nucleic Acids Research *49*, D545-D551. 10.1093/nar/gkaa970.

Kernfeld, E.M., Genga, R.M.J., Neherin, K., Magaletta, M.E., Xu, P., and Maehr, R. (2018). A Single-Cell Transcriptomic Atlas of Thymus Organogenesis Resolves Cell Types and Developmental Maturation. Immunity *48*, 1258-1270.e1256. 10.1016/j.immuni.2018.04.015.

Kinchen, J., Chen, H.H., Parikh, K., Antanaviciute, A., Jagielowicz, M., Fawkner-Corbett, D., Ashley, N., Cubitt, L., Mellado-Gomez, E., Attar, M., et al. (2018). Structural Remodeling of the Human Colonic Mesenchyme in Inflammatory Bowel Disease. Cell *175*, 372-386.e317. 10.1016/j.cell.2018.08.067.

La Manno, G., Gyllborg, D., Codeluppi, S., Nishimura, K., Salto, C., Zeisel, A., Borm, L.E., Stott, S.R.W., Toledo, E.M., Villaescusa, J.C., et al. (2016). Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. Cell *167*, 566-580.e519. 10.1016/j.cell.2016.09.027.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494-498. 10.1038/s41586-018-0414-6.

Lake, B.B., Chen, S., Sos, B.C., Fan, J., Kaeser, G.E., Yung, Y.C., Duong, T.E., Gao, D., Chun, J., Kharchenko, P.V., and Zhang, K. (2018). Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol *36*, 70-80. 10.1038/nbt.4038.

Li, J., and Zeng, Y. (2018). Histology and Embryology (Chinese Edition), 9 Edition (People's Medical Publishing House).

Li, N., Zhan, S., Tian, Z., Liu, C., Xie, Z., Zhang, S., Chen, M., Zeng, Z., and Zhuang, X. (2021). Alterations in Bile Acid Metabolism Associated With Inflammatory Bowel Disease. Inflammatory Bowel Diseases. 10.1093/ibd/izaa342.

Lin, A., Giuliano, C.J., Palladino, A., John, K.M., Abramowicz, C., Yuan, M.L., Sausville, E.L., Lukow, D.A., Liu, L., Chait, A.R., et al. (2019). Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. Sci Transl Med *11*. 10.1126/scitranslmed.aaw8412.

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. 10.1038/s41586-020-2797-4.

Lukowski, S.W., Lo, C.Y., Sharov, A.A., Nguyen, Q., Fang, L., Hung, S.S., Zhu, L., Zhang, T., Grünert, U., Nguyen, T., et al. (2019). A single-cell transcriptome atlas of the adult human retina. Embo j *38*, e100811. 10.15252/embj.2018100811.

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. 10.1186/s13059-019-1906-x.

McInnes, L., and Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv *abs/1802.03426*. arXiv:1802.03426.

Menon, M., Mohammadi, S., Davila-Velderrain, J., Goods, B.A., Cadwell, T.D., Xing, Y., Stemmer-Rachamimov, A., Shalek, A.K., Love, J.C., Kellis, M., and Hafler, B.P. (2019). Single-cell transcriptomic atlas of the human retina identifies cell types associated with age-related macular degeneration. Nat Commun *10*, 4902. 10.1038/s41467-019-12780-8.

Mescher, A.L. (2016). Junqueira's Basic Histology Text and Atlas, 13 Edition (McGraw-Hill).

Miao, Z., Deng, K., Wang, X., and Zhang, X. (2018). DEsingle for detecting three types of differential expression in single-cell RNA-seq data. Bioinformatics *34*, 3223-3224. 10.1093/bioinformatics/bty332.

Miao, Z., and Zhang, X. (2016). Differential expression analyses for single-cell RNA-Seq: old questions on new data. Quantitative Biology *4*, 243-260. 10.1007/s40484-016-0089-7.

MySQL (2021). Limits on Table Column Count and Row Size. Available at: https://dev.mysql.com/doc/refman/8.0/en/column-count-limit.html (Accessed: 21 July 2021).

Narayan, A., Berger, B., and Cho, H. (2021). Assessing single-cell transcriptomic variability through density-preserving data visualization. Nature Biotechnology *39*, 765-774. 10.1038/s41587-020-00801-7.

Oracle (2021). Datatype Limits. Available at: https://docs.oracle.com/en/database/oracle/oracle-database/21/refrn/datatype-limits.html (Accessed: 21 July 2021).

Parikh, K., Antanaviciute, A., Fawkner-Corbett, D., Jagielowicz, M., Aulicino, A., Lagerholm, C., Davis, S., Kinchen, J., Chen, H.H., Alham, N.K., et al. (2019). Colonic epithelial cell diversity in health and inflammatory bowel disease. Nature *567*, 49-55. 10.1038/s41586-019-0992-y.

Park, J.-E., Botting, R.A., Domínguez Conde, C., Popescu, D.-M., Lavaert, M., Kunz, D.J., Goh, I., Stephenson, E., Ragazzini, R., Tuck, E., et al. (2020). A cell atlas of human thymic development defines T cell repertoire formation. Science *367*. 10.1126/science.aay3224.

Parker, K.R., Migliorini, D., Perkey, E., Yost, K.E., Bhaduri, A., Bagga, P., Haris, M., Wilson, N.E., Liu, F., Gabunia, K., et al. (2020). Single-Cell Analyses Identify Brain Mural Cells Expressing CD19 as Potential Off-Tumor Targets for CAR-T Immunotherapies. Cell *183*, 126-142.e117. 10.1016/j.cell.2020.08.022.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science *2*, 559-572. 10.1080/14786440109462720.

Pellin, D., Loperfido, M., Baricordi, C., Wolock, S.L., Montepeloso, A., Weinberg, O.K., Biffi, A., Klein, A.M., and Biasco, L. (2019). A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. Nature Communications *10*, 2395. 10.1038/s41467-019-10291-0.

Phipson, B., Zappia, L., and Oshlack, A. (2017). Gene length and detection bias in single cell RNA sequencing protocols. F1000Res *6*, 595. 10.12688/f1000research.11290.1.

Pisco, A., and Consortium, T.S. (2021). Tabula Sapiens release 1.0. figshare, doi.org/10.6084/m9.figshare.14267219.v1.

Plasschaert, L.W., Žilionis, R., Choo-Wing, R., Savova, V., Knehr, J., Roma, G., Klein, A.M., and Jaffe, A.B. (2018). A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. Nature *560*, 377-381. 10.1038/s41586-018-0394-6.

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*. 10.7554/eLife.27041.

Renthal, W., Boxer, L.D., Hrvatin, S., Li, E., Silberfeld, A., Nagy, M.A., Griffith, E.C., Vierbuchen, T., and Greenberg, M.E. (2018). Characterization of human mosaic Rett syndrome brain tissue by single-nucleus RNA sequencing. Nat Neurosci *21*, 1670-1679. 10.1038/s41593-018-0270-6.

Reyfman, P.A., Walter, J.M., Joshi, N., Anekalla, K.R., McQuattie-Pimentel, A.C., Chiu, S., Fernandez, R., Akbarpour, M., Chen, C.-I., Ren, Z., et al. (2019). Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. American Journal of Respiratory and Critical Care Medicine *199*, 1517-1536. 10.1164/rccm.201712-2410OC.

Rood, J.E., Stuart, T., Ghazanfar, S., Biancalani, T., Fisher, E., Butler, A., Hupalowska, A., Gaffney, L., Mauck, W., Eraslan, G., et al. (2019). Toward a Common Coordinate Framework for the Human Body. Cell *179*, 1455-1467. 10.1016/j.cell.2019.11.019.

Rubin, D.B., Danish, H.H., Ali, A.B., Li, K., LaRose, S., Monk, A.D., Cote, D.J., Spendley, L., Kim, A.H., Robertson, M.S., et al. (2019). Neurological toxicities associated with chimeric antigen receptor T-cell therapy. Brain *142*, 1334-1348. 10.1093/brain/awz053.

Schaum, N., Karkanias, J., Neff, N.F., May, A.P., Quake, S.R., Wyss-Coray, T., Darmanis, S., Batson, J., Botvinnik, O., Chen, M.B., et al. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature *562*, 367-372. 10.1038/s41586-018-0590-4.

Shyer, J.A., Flavell, R.A., and Bailis, W. (2020). Metabolic signaling in T cells. Cell Research *30*, 649-659. 10.1038/s41422-020-0379-5.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., and Kasprzyk, A. (2009). BioMart--biological queries made easy. BMC Genomics *10*, 22. 10.1186/1471-2164-10-22.

Snyder, M.P., Lin, S., Posgai, A., Atkinson, M., Regev, A., Rood, J., Rozenblatt-Rosen, O., Gaffney, L., Hupalowska, A., Satija, R., et al. (2019). The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. Nature *574*, 187-192. 10.1038/s41586-019-1629-x.

Sokolov, A., Paull, E.O., and Stuart, J.M. (2016). ONE-CLASS DETECTION OF CELL STATES IN TUMOR SUBTYPES. Pac Symp Biocomput *21*, 405-416. 10.1142/9789814749411_0037.

Soneson, C., and Robinson, M.D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. Nature Methods *15*, 255-261. 10.1038/nmeth.4612.

Song, X., Sun, X., Oh, S.F., Wu, M., Zhang, Y., Zheng, W., Geva-Zatorsky, N., Jupp, R., Mathis, D., Benoist, C., and Kasper, D.L. (2020). Microbial bile acid metabolites modulate gut RORγ+ regulatory T cell homeostasis. Nature *577*, 410-415. 10.1038/s41586-019-1865-0.

Srivastava, D., Iyer, A., Kumar, V., and Sengupta, D. (2018). CellAtlasSearch: a scalable search engine for single cells. Nucleic Acids Research *46*, W141-W147. 10.1093/nar/gky421.

Stonebraker, M. (2010). SQL databases v. NoSQL databases. Commun. ACM *53*, 10–11. 10.1145/1721654.1721659.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., III, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1-15. 10.1016/j.cell.2019.05.031.

Sunkin, S.M., Ng, L., Lau, C., Dolbeare, T., Gilbert, T.L., Thompson, C.L., Hawrylycz, M., and Dang, C. (2013). Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic Acids Res *41*, D996-d1008. 10.1093/nar/gks1042.

Svensson, V., da Veiga Beltrame, E., and Pachter, L. (2020). A curated database reveals trends in single-cell transcriptomics. Database *2020*. 10.1093/database/baaa073.

Tang, J., and Zhang, L. (2013). Histology and Embryology(Chinese Edition), 3 Edition (Peking University Medica).

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol *32*, 381-386. 10.1038/nbt.2859.

Tung, P.-Y., Blischak, J.D., Hsiao, C.J., Knowles, D.A., Burnett, J.E., Pritchard, J.K., and Gilad, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. Scientific Reports *7*, 39921. 10.1038/srep39921.

Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419. 10.1126/science.1260419.

Venteicher, A.S., Tirosh, I., Hebert, C., Yizhak, K., Neftel, C., Filbin, M.G., Hovestadt, V., Escalante, L.E., Shaw, M.L., Rodman, C., et al. (2017). Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science *355*. 10.1126/science.aai8478.

Vieira Braga, F.A., Kar, G., Berg, M., Carpaij, O.A., Polanski, K., Simon, L.M., Brouwer, S., Gomes, T., Hesse, L., Jiang, J., et al. (2019). A cellular census of human lungs identifies novel cell states in health and in asthma. Nat Med *25*, 1153-1163. 10.1038/s41591-019-0468-5.

Villani, A.C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., et al. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science *356*. 10.1126/science.aah4573.

Voigt, A.P., Mulfaul, K., Mullin, N.K., Flamme-Wiese, M.J., Giacalone, J.C., Stone, E.M., Tucker, B.A., Scheetz, T.E., and Mullins, R.F. (2019). Single-cell transcriptomics of the human retinal pigment epithelium and choroid in health and macular degeneration. Proc Natl Acad Sci U S A *116*, 24100-24107. 10.1073/pnas.1914143116.

Wang, L., Yu, P., Zhou, B., Song, J., Li, Z., Zhang, M., Guo, G., Wang, Y., Chen, X., Han, L., and Hu, S. (2020a). Single-cell reconstruction of the adult human heart during heart failure and recovery reveals the cellular landscape underlying cardiac function. Nat Cell Biol *22*, 108-119. 10.1038/s41556-019-0446-7.

Wang, X., Williams, C., Liu, Z.H., and Croghan, J. (2017). Big data management challenges in health research—a literature review. Briefings in Bioinformatics *20*, 156-167. 10.1093/bib/bbx086.

Wang, Y., Song, W., Wang, J., Wang, T., Xiong, X., Qi, Z., Fu, W., Yang, X., and Chen, Y.G. (2020b). Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine. J Exp Med *217*. 10.1084/jem.20191130.

Weber, G.M., Ju, Y., and Börner, K. (2020). Considerations for Using the Vasculature as a Coordinate

System to Map All the Cells in the Human Body. Front Cardiovasc Med *7*, 29. 10.3389/fcvm.2020.00029.

Wei, J., Han, X., Bo, J., and Han, W. (2019). Target selection for CAR-T therapy. Journal of Hematology & Oncology *12*, 62. 10.1186/s13045-019-0758-x.

Xiao, Z., Dai, Z., and Locasale, J.W. (2019). Metabolic landscape of the tumor microenvironment at single cell resolution. Nature Communications *10*, 3763. 10.1038/s41467-019-11738-0.

Yáñez, L., Sánchez-Escamilla, M., and Perales, M.-A. (2019). CAR T Cell Toxicity: Current Management and Future Directions. HemaSphere *3*, e186. 10.1097/hs9.0000000000000186.

Yin, Z., Bai, L., Li, W., Zeng, T., Tian, H., and Cui, J. (2019). Targeting T cell metabolism in the tumor microenvironment: an anti-cancer therapeutic strategy. Journal of Experimental & Clinical Cancer Research *38*, 403. 10.1186/s13046-019-1409-3.

Yu, Q., Kilik, U., Holloway, E.M., Tsai, Y.H., Harmel, C., Wu, A., Wu, J.H., Czerwinski, M., Childs, C.J., He, Z., et al. (2021). Charting human development using a multi-endodermal organ atlas and organoid models. Cell *184*, 3281-3298.e3222. 10.1016/j.cell.2021.04.028.

Zhang, J.Y., Wang, X.M., Xing, X., Xu, Z., Zhang, C., Song, J.W., Fan, X., Xia, P., Fu, J.L., Wang, S.Y., et al. (2020). Single-cell landscape of immunological responses in patients with COVID-19. Nat Immunol *21*, 1107-1118. 10.1038/s41590-020-0762-x.

Zhang, P., Yang, M., Zhang, Y., Xiao, S., Lai, X., Tan, A., Du, S., and Li, S. (2019). Dissecting the Single-Cell Transcriptome Network Underlying Gastric Premalignant Lesions and Early Gastric Cancer. Cell Rep *27*, 1934-1947.e1935. 10.1016/j.celrep.2019.04.052.

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M., et al. (2018). CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Research *47*, D721-D728. 10.1093/nar/gky900.

Zhong, S., Ding, W., Sun, L., Lu, Y., Dong, H., Fan, X., Liu, Z., Chen, R., Zhang, S., Ma, Q., et al. (2020). Decoding the development of the human hippocampus. Nature *577*, 531-536. 10.1038/s41586-019-1917-5.

Zhong, S., Zhang, S., Fan, X., Wu, Q., Yan, L., Dong, J., Zhang, H., Li, L., Sun, L., Pan, N., et al. (2018). A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. Nature *555*, 524-528. 10.1038/nature25980.

SUPPLEMENTARY METHODS

Dataset collection

In the first version of hECA (v1.0), we present a collection of 1,093,299 cells from 116 datasets belonging to 20 published studies (Asp *et al.*, 2019; Cui *et al.*, 2019; Gaublomme *et al.*, 2019; Han *et al.*, 2020; Kinchen *et al.*, 2018; Lake *et al.*, 2018; Lukowski *et al.*, 2019; Madissoon *et al.*, 2019; Menon *et al.*, 2019; Parikh *et al.*, 2019; Plasschaert *et al.*, 2018; Renthal *et al.*, 2018; Sunkin *et al.*, 2013; Venteicher *et al.*, 2017; Vieira Braga *et al.*, 2019; Voigt *et al.*, 2019; Wang *et al.*, 2020a; Wang *et al.*, 2020b; Zhong *et al.*, 2020; Zhong *et al.*, 2018), details provided in **Table S1**. We designed hECA as a framework for collection of all accessible human single cell data, regardless of the technology, platform, researcher, study design or other factors in data generation. Toward this goal, we selected 20 peer-reviewed studies in hECA v1.0, preferably studies with high throughput in cell numbers and coverage of multiple healthy organs. These studies covered 38 organs and spanned the developmental stages from fetal to adult. In hECA v1.0 we only include transcriptomic data of healthy doners, but future versions will cover multi-omics data as well as data of disease samples.

In each of the studies, we collected the expression matrix of every dataset in the study. In addition, we collected all the descriptive information in study level and dataset level, and analysis results of the cells in the original papers. They are referred to as metadata and original annotations in hECA, respectively. Metadata includes the following information if available: sample organ, sample tissue, anatomical region, subregion, donor ID, donor gender, donor age or developmental stage, sequencing technology. Annotations are the assigned cell type of each cell in the original study. The completeness of metadata and annotations vary among datasets according to original studies.

Processing of data matrixes

The collected datasets were processed for integration into uGT. The 116 datasets we collected in hECA v1.0 are all expression matrices, every matrix was transformed into gene by cell matrix, with each row representing a gene and each column representing a cell. For those expression value in log scale, we performed the value transformation back to raw values.

For the purpose of integrating data into uGT, we unified the gene names for all datasets. For datasets identifying gene with Ensembl ID, we used the R package biomaRt (Smedley *et al.*, 2009) to convert Ensembl ID into gene symbol. Datasets with different number of gene symbols were unified with an in-house built toolkit: we compared gene symbols in the datasets to the list of 43,878 HUGO Gene Nomenclature Committee (HGNC) approved symbols (see "HGNC gene symbol list" in https://github.com/XuegongLab/hECA), all previous, withdrawn and alias symbols were converted into HGNC approved symbols. Genes that are in the list but not sequenced in any dataset were filled with zeros. After processing, every dataset was with 43,878 genes as rows.

For datasets that have cell type annotations in the original study, the original annotations were kept and stored in column "original_name" in the uGT. Regardless of the original annotations, we performed clustering analysis and annotation in each dataset with Seurat v3.2 (Stuart *et al.*, 2019). We implemented a standard processing procedure for each dataset: We created a Seurat object from the expression matrix, conducted quality control to filter out genes and cells, selected variable genes, conducted normalization, scaling, dimensional reduction and cell clustering. The parameters for quality control and cell filtering were determined specifically for each dataset following the original studies or following the tutorial of Seurat. The parameter for cell clustering is determined based on the consistency with original cell clustering results. Then the analysis pipeline of Seurat was performed to get cell cluster-specific expressed genes. After the quality control, we got a total of 1,093,299 cells from 116 datasets.

Integrating data into uGT

After the processing, the datasets were ready to be integrated into a uGT. The data were depth-adjusted and log-normalized. We stored all data in an instance of the multi-model NoSQL data storage service Tablestore on Aliyun Cloud. It utilized distributed storage and index-based search engine to store a large amount of structured data and support fast query. We use the Wide Column model in Tablestore as the framework of uGT. We integrated all 1,093,299 cells into the uGT. In the wide-column table. In the current version, every cell is a row with a unique identifier (column "cid"), followed by 43,878 columns of genes expression values and 17 columns of metadata (columns "user_id", "study_id",

“cell_id”, “organ”, “region”, “subregion”, “seq_tech”, “sample_status”, “donor_id”, “donor_gender”, “donor_age”, “original_name”, “cl_name”, “uHAF_name”, “tissue_type”, “cell_type”, and “marker_gene”). Another 6 columns of metadata are about annotation including the original annotation (column “original_name”) and annotations assigned in our analysis (column “cl_name”, “uHAF_name”, “tissue_type”, “cell_type”, and “marker_gene”).

ECAUGT: the data access interface on uGT

Based on the Tablestore python SDK, we developed a command line tool “ECAUGT” (pronounced as “e-caught”) to query data from hECA for advanced users to implement *in data* cell sorting. Users can query the cells with the provided query conditions and download the selected data of these cells. Function “query_cells()” will query cells with conditions on the columns of metadata and provide a user-friendly interface, with which users can combine multiple conditions into a logical expression in a structured string with logical operators ‘&&’ (for logical operation AND), ‘||’ (for logical operation OR), and ‘!’ (for logical operation NOT). Then “query_cells()” will return the cid list of the queried cells. Function “get_columnsbycell ()” will allow users to download data with this id list. Users can select interested columns and add gene conditions in this function with the similar interface by “query_cells()”. The “get_columnsbycell ()” can provide downloaded data in two forms: a python list, where each element represents a cell, or a pandas.DataFrame object. User can choose the form they want with the parameter “do_transform”. We also provide the parallel acceleration version with similar interface by “get_columnsbycell_para()”. Function “get_all_rows()” will provide the cid list of all cells in uGT and can be convenient when user require information of the whole hECA. Function “get_column_set()” receives a cid list and will provide all unique values in the selected column of these cells. “ECAUGT” is available on https://pypi.org/project/ECAUGT and can be installed with PyPI.

For users without much programming background, we provided a lightweight command line tool “Cell_Download” to download data from hECA. Users first query cells in the website interface of hECA and download a cell id list file. Then “Cell_Download” only need one-line command to assign the input and output path and will automatically download all columns of the selected cells in the id list and save the result in four files: a .csv file “metadata.csv” for columns of metadata, a .npz file for

sparse expression matrix, and two .csv files for the row names and column names of this matrix.

Constructing the unified hierarchical cell annotation framework (uHAF)

The uHAF is a structured framework we designed for the representation and hierarchical annotation of cell types in hECA. uHAF records the knowledge and relationships among different anatomical levels and cell types.

There are two major structured ontologies in uHAF: the macroscopic level and the microscopic level. The macroscopic level mainly contains anatomical entities including system, organ, suborgan (see "uHAF macroscopic ontology" in https://github.com/XuegongLab/hECA). The microscopic level consists of histological information with four tissue types (epithelial tissue, connective tissue, muscle tissue and nerve tissue) and a series of hierarchical entities of cell types (see "uHAF microscopic ontology" in https://github.com/XuegongLab/hECA). Both the macroscopic level and microscopic level were represented from general to detail by nodes from parent to child, and organized into ontology by the protégé tool (https://protege.stanford.edu/). The macroscopic and microscopic ontologies can be connected if the cell type composes one organ or anatomical region by relationship of "part of". We listed all the connection observed in our collected data in hECA v1.0 in "uHAF macro-micro map" (https://github.com/XuegongLab/hECA).

We established the marker references for the microscopic entities in uHAF (see "uHAF marker reference" in https://github.com/XuegongLab/hECA). We collected the markers of cell types that were investigated in the original studies and toke the intersection from multiple studies as the marker references of the cell types. For cell types that marker genes were not given in the original studies, we surveyed for markers from multiple sources including PanglaoDB (https://panglaodb.se/), the Human Protein Atlas (https://www.proteinatlas.org/), and CellMarker (http://biocc.hrbmu.edu.cn/CellMarker/) to replenish the marker references. Such processes were implemented iteratively to curate the final marker references. The references will be continuously updated along with the release of new versions of hECA.

The macroscopic part of uHAF was built based on several standard textbooks including *Junqueira's Basic Histology: Text & Atlas* (Mescher, 2016), *Histology and Embryology* (in Chinese)

(Tang and Zhang, 2013), *Systematic Anatomy* (in Chinese) (Bai and Ying, 2015), *Histology and Embryology* (in Chinese) (Li and Zeng, 2018)) as well as several public studies and databases (Franzén *et al.*, 2019).

We provided the uHAF related files in https://github.com/XuegongLab/hECA.

Cell identify assignment

Annotation of all collected cells was performed under the framework of uHAF.

uHAF name assignment

For each Seurat cluster, we identified the cluster-specific differentially expressed genes (DEGs) by FindAllMarkers function. We referred to the marker reference to determine the cell type labels, and use the top ranked DEGs to further annotate the subtypes. We first determined the most general labels among the four tissue types (epithelial tissue, connective tissue, muscle tissue, nerve tissue), and then chose the deepest child cell type on which markers can be used to support the cell type assignment in the uHAF. In this way, we annotate each cluster in both macroscopic and microscopic level as "organ-tissue_type-cell_type-markers". For cells that cannot be annotated based on available information, we named them as "Unclassified". This label produced from uHAF is called "uHAF_name" (see **Table S3**).

Mapping uHAF names to Cell Ontology terms

We downloaded the basic Cell Ontology terms from CL website (Cell Ontology - Summary | NCBO BioPortal (http://bioontology.org)), retained "Preferred Label", "Definitions" and "Parents" (Table), and used the "Preferred Label" for CL term assignment. We converted the "uHAF_name" to "cl_name" by a combined strategy: We preferably used the Cell Ontology terms with the exact matching of the whole string of "cell_type". For the "cell_type" that did not appear in the Cell Ontology terms, we further searched their parent "cell_type" in our uHAF until the Cell Ontology term is matched completely. For the remaining "cell_type"s, we manually determined the most similar Cell Ontology terms by ontologyIndex R package (Greene *et al.*, 2017). If none term was found, we labeled them as "none" (see **Table S3**).

Generation of quantitative portraits

We designed a portraiture system as a systematic way to characterize the full properties of biological entities of all levels in hECA. There are three major types of biological entities in hECA: organs (including sub-organs), cell types (including subtypes) and genes. A full quantitative portrait of a biological entity is its holographic picture of the entity at anatomical, cellular and molecular levels. However, both the quality and quantity of the currently available data in hECA are far from constructing such full portraits. Therefore, the quantitative portraits in hECA v1.0 only illustrated the idea of the portraiture system using the available information. They reflect more about the characteristics of the collected data of and related to each entity, rather than about the biological truth of the entity.

Organ portraits: a portrait of an organ is composed of 3 major parts: the cell composition viewer, the cell embedding viewer, and the organ hierarchy viewer. The cell composition viewer shows the counts and fractions of cell types observed in one organ's datasets. It is notable that statistics in the organ portraits only reflect the counts/fractions of the collected cells, not the true counts/percentages of cell types in an organ. The embedding viewer visualizes cells of an organ with a 2-dimensional scattergram (UMAP/PCA/DensMAP for users to choose). This viewer supports coloring embedded cells by their cell types, sequencing technologies, sample origins, and any given gene's expression level. The organ hierarchy viewer shows the position of the organ in the uHAF macroscopic annotation system.

Cell type portraits: The cell type portrait depicts cells belong to the same cell types/subtypes across all collected organs, and it is composed of 4 major parts: cell distribution, marker genes, 2D visualization and taxonomy relationship with other cell types. The cell distribution part describes the relationship of this cell type with organs, with bar plots showed the organ origin of this cell type in numbers and proportions. The marker gene part provides a table with genes highly expressed in this cell type, which were defined by comparing gene expression level with all other cell types using Seurat v3.2. We filtered out genes with adjusted p-value larger than 0.05 or expressed in fewer than 25% cells in this cell type, and showed top 50 genes with highest log fold-changes. In 2D visualization part, we

plotted an interactive scatter plot showing the distribution and landscape of cells in this cell type. Like organ portraits, we also showed the cell type's hierarchical relationship with other uHAF tree cell types.

Gene portraits: The portrait of a gene is composed of 2 major parts: basic gene information and gene expression distribution. In the basic gene information part, for each gene, we collected the full name of gene, the position where the gene is on the genome, commonly used aliases of the gene, and description that introduce the basic function of the gene. The "known as marker of" section denote cell types that highly express this gene, which is calculated by comparing the expression level of the gene in a cell type with it in other cells. For the gene expression distribution part, we first performed data normalization of all cells in uGT using function NormalizeData in Seurat v3.2. For each gene, we present its distribution in an organ or in a cell type by drawing a ridge plot. The ridge plot is fitted by expression value of the gene in the organ or cell type, while zero-value are truncated before fitting. The median expression level and non-zero percentage are also provided on the ridge plot.

The hECA website

We provided two portals for users to access hECA. One is a computer programing portal for users to access the data and do *in data* cell experiment using the ECAUGT package. The portal is at https://pypi.org/project/ECAUGT/. It is powerful but requires users to be comfortable with some programming skills. The other portal is a website at http://eca.xglab.tech/ with graphic user interface (GUI) that enables both browsing hECA at all levels and searching the data for *in data* cell experiments. ECAUGT can also be accessed from the website portal.

The interactive functions of the hECA website (http://eca.xglab.tech/) are divided into four parts: "*in data* cell sorting", "uHAF cells", "uHAF organs" and "gene portraits", plus a link to the "ECAUGT" portal. Users can browse these functions anonymously, but should sign in to get the full service of the results.

"Cell sorting" is the graphical interface for *in data* cell sorting in hECA v1.0. It supports flexible multi-step cell selection with all kinds of filters regarding to cell features (gene, cell type in uHAF, organ in uHAF and other metadata). Filters can be combined with basic logic operators (AND, OR,

NOT) to form complex logic expressions. Users can have a quick view of the selected data with real-time statistical analysis and visualization of the organ origin and cell type composition, and can adjust the sorting criteria accordingly if necessary. For more in-depth analysis, we provide the organ-wise cell type composition and gene expressions across cell types or organs and pseudo-FACS visualization of expression correlation between any two genes. Cell sorting processes can be saved to users' collections for future reference. After users selected their interested cell groups, a cid list can be downloaded for further data query with ECAUGT. Examples of *in data* cell sorting and vignettes are provided.

Cell types and organs are organized in uHAF trees in hECA. The "uHAF cells" entry provides an interactive visualization of the hierarchical cell type tree, which is the microscopic subgraph of uHAF. The "uHAF organ" entry provides the view of the macroscopic subgraph of uHAF. Each cell type is assigned with a unique uHAF ID with a brief description. We provide portraits of the cell types on nodes with data available in the current version. Users can click "view details" to check the cell type portraits which include information of original organs, marker genes and embedding view of the cell types. The embedding view is drawn as UMAP by default, and users can switch to PCA or DensMAP view. The plots can be colored by the organs, expression level of selected gene, sequencing platform or the original study. The organ portraits provide information of cell type composition (as reflected by the current data), similar embedding views, anatomy relationships and position in the uHAF.

The "gene portraits" entry allows users to select any particular gene and visualize the distribution of the gene in all organs and cell types (as reflected by the currently available data). The basic information includes the distribution of non-zero expression values in the organs and cell types, and the proportion of non-zero values (%Expr). Users should keep in mind the fact that the current scRNA-seq data are quite noisy and suffer from dropout events when using the information. The gene portraits also provide basic information of the gene collected from public databases and links to the corresponding pages at Genecard, NCBI, Ensembl and Wikigenes.