D1244–D1254 *Nucleic Acids Research, 2022, Vol. 50, Database issue* *Published online 4 October 2021*
https://doi.org/10.1093/nar/gkab857

huARdb: human Antigen Receptor database for interactive clonotype-transcriptome analysis at the single-cell level

Lize Wu[1,2,†], Ziwei Xue[3,†], Siqian Jin[3], Jinchun Zhang[3], Yixin Guo[3], Yadan Bai[3], Xuexiao Jin[1], Chaochen Wang[3], Lie Wang[4], Zuozhu Liu[5], James Q. Wang[3], Linrong Lu[1,2,3,6,*] and Wanlu Liu[2,3,6,7,8,*]

[1]Institute of Immunology and Department of Rheumatology at Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang 310058, China, [2]Liangzhu Laboratory, Zhejiang University Medical Center, 1369 West Wenyi Road, Hangzhou, Zhejiang 311121, China, [3]Zhejiang University-University of Edinburgh Institute (ZJU-UoE Institute), Zhejiang University School of Medicine, International Campus, Zhejiang University, Haining, Zhejiang 314400, China, [4]Department of Immunology, Zhejiang University School of Medicine, Hangzhou, Zhejiang 310058, China, [5]Zhejiang University-University of Illinois at Urbana-Champaign Institute (ZJU-UIUC Institute), International Campus, Zhejiang University, Haining, Zhejiang 314400, China, [6]Dr. Li Dak Sum & Yip Yio Chin Center for Stem Cell and Regenerative Medicine, Zhejiang University, Hangzhou, Zhejiang 310058, China, [7]Department of Orthopedic Surgery of the Second Affiliated Hospital of Zhejiang University School of Medicine, Zhejiang University, Hangzhou, Zhejiang 310003, China and [8]Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Zhejiang University, Hangzhou, Zhejiang 310058, China

Received July 26, 2021; Revised August 31, 2021; Editorial Decision September 08, 2021; Accepted September 14, 2021

ABSTRACT

T-cell receptors (TCRs) and B-cell receptors (BCRs) are critical in recognizing antigens and activating the adaptive immune response. Stochastic V(D)J recombination generates massive TCR/BCR repertoire diversity. Single-cell immune profiling with transcriptome analysis allows the high-throughput study of individual TCR/BCR clonotypes and functions under both normal and pathological settings. However, a comprehensive database linking these data is not yet readily available. Here, we present the human Antigen Receptor database (huARdb), a large-scale human single-cell immune profiling database that contains 444 794 high confidence T or B cells (hcT/B cells) with full-length TCR/BCR sequence and transcriptomes from 215 datasets. All datasets were processed in a uniform workflow, including sequence alignment, cell subtype prediction, unsupervised cell clustering, and clonotype definition. We also developed a multi-functional and user-friendly web interface that provides interactive visualization modules for biologists to analyze the transcriptome and TCR/BCR features at the single-cell level. HuARdb is freely available at https://huarc.net/database with functions for data querying, browsing, downloading, and depositing. In conclusion, huARdb is a comprehensive and multi-perspective atlas for human antigen receptors.

INTRODUCTION

The human adaptive immune system is a branch of the immune system that is responsible for specific antigen recognition and clearance (1). Through interacting with specific antigens, the adaptive immune system is activated and can store long-term immunological memories for targeted antigens (2). Long-term immunological memory with high antigen-specificity can therefore generate a more robust response during subsequent exposure to the antigens (2). Adaptive immune response activation requires antigen recognition by receptors expressed on T or B cells, known as T cell receptors (TCRs) or B cell receptors (BCRs), respectively (3).

TCRs are composed of paired α and β peptide chains and BCRs are composed of heavy and light chains, with each chain consisting of a variable region (V region) and constant region (C region) (3,4). The V region of each TCR/BCR peptide chain is encoded by the stochastic

*To whom correspondence should be addressed. Tel: +86 571 87572818; Fax: +86 571 88981979; Email: wanluliu@intl.zju.edu.cn
Correspondence may also be addressed to Linrong Lu. Tel: +86 571 88981173; Fax: +86 571 88208022; Email: lu_linrong@zju.edu.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

recombination of the variable (V) gene, diversity (D) gene, and joining (J) gene (5). During T or B cell development, V(D)J gene recombination produces unique complementarity-determining regions (CDRs) for TCRs/BCRs on each T or B cell, which endows their specificity (6,7). The V regions of each TCR/BCR peptide chain incorporate CDR1, CDR2 and CDR3, with CDR3 playing the predominant role in antigen recognition (8–11). Diverse V(D)J gene recombination, junctional diversity, and chain combination together have been estimated to generate up to 10^{20} possible different TCRs and BCRs, allowing the adaptive immune system to recognize almost an infinite number of antigens (12,13). Characterization of TCRs/BCRs is critical for biologists to better understand how the adaptive immune system exerts its effector function in a pathogen- or antigen-specific manner (14).

Bulk TCR/BCR sequencing has been used for years to investigate clonally expanded T/B cells during normal or pathogenic immunological responses (15–17). However, bulk TCR/BCR sequencing does not provide necessary information for receptor pairing, which is essential to reconstruct functional TCRs/BCRs for experimental validation. The recent development of single-cell immune profiling techniques allows the detection of paired full-length TCR/BCR sequences simultaneously with genome-wide transcriptomes at the single-cell level. The revolutionary advance in immune receptor profiling provides a unique opportunity for immunologists to study the functional relevance of clonally expanded T/B cells under different settings. With single-cell immune profiling techniques, tumor-specific TCRs can be identified from individual patients to guide the development of TCR-T-therapy for eradicating tumors with high specificity and few side effects (18). Single-cell immune profiling performed in ulcerative colitis (UC) patients revealed the clonal evolution of T cells in UC pathogenesis (19). These studies demonstrated that single-cell immune profiling provides new insights into disease pathogenesis and may shed light on targeted therapy of immune-related diseases. These large depositions of publicly available single-cell immune profiling data call for their in-depth analysis to yield new findings on their immunological significance. Such analysis would be aided by a publicly available single-cell immune profiling database.

Here, we present the human Antigen Receptor database (huARdb), a single-cell immune profiling database with a multi-functional and user-friendly web interface. HuARdb has collected 215 single-cell immune profiling datasets from fourteen projects and features over 440 000 high confidence T or B cells (hcT/B cells) containing over 880 000 paired TCR/BCR chains. The interactive web interface of huARdb allows the coupled clonotype-transcriptome analysis of collected datasets to functionally characterize the clonally expanded T or B cells.

MATERIALS AND METHODS

Data retrieval and pre-processing

Fourteen human single-cell immune profiling datasets were collected from various publications and resources including Gene Expression Omnibus (GEO) database, Sequence Read Archive (SRA) database, and Genome Sequence Archive (20–22) (Supplementary Table S1). Pre-processing for coupled single-cell (sc) RNA-seq and scV(D)J-seq data were performed using Cell Ranger (v6.1.0) with default parameters except for sample C30_R from Boland *et al.* which was processed using Cell Ranger (v4.0.0) (Figure 1) (23). Briefly, *mkgtf* function was used to retain the protein-coding sequence (23). Reference index based on human genome assembly GRCh38 (hg38, http://ensembl.org/Homo_sapiens/) was built via *mkref* and *mkvdjref* functions (23). Raw data in FASTQ format for scRNA-seq were then processed by the *count* function to generate a unique molecular identifier (UMI) count matrix for protein-coding genes (23). Meanwhile, raw data in FASTQ format for scV(D)J-seq were processed by the *vdj* function to produce V, (D), J, C gene usage, CDR3 sequences, and UMI counts of TCR/BCR chains (23).

Quality control and cell filtering based on transcriptome information

For each sample, gene expression matrix containing UMI counts was loaded in R (v4.0.4) with Seurat (v4.0.2) R-package (24). Cells with unique feature counts <200 or with >20% mitochondrial features were filtered. Potential doublets produced in the library construction step was filtered out with DoubletFinder (v2.0.3) R-package, since it was scored with the highest usability in the computational doublet-detection methods benchmarking study (25,26). In brief, DoubletFinder R-package defining doublet based on transcriptome similarity between test cell and the simulated artificial doublet.

Cell subtype prediction and clustering

Cell subtype prediction. After doublet filtering, original gene expression matrix was transformed into Single Cell Experiment (SCE) Objects by Seurat (v4.0.2) R-package (24). UMI counts were log-normalized with the *LogNormCounts* function within scater (v1.18.6) R-package (27). Cell subtype prediction was performed with SingleR (v1.4.1) R-package using previously published human T or B cell reference datasets (Figure 1) (28). Predicted effector memory $CD8^+$ T cells, central memory $CD8^+$ T cells, terminal effector $CD8^+$ T cells and naïve $CD8^+$ T cells was classified as 'unpredicted' cells if they expressed *CD4*. Predicted helper T cells (Th cells) and follicular helper T cells were classified as 'unpredicted' cells if they expressed *CD8A*. Data from Corridoni *et al.* were generated from FACS (Fluorescence-activated cell sorting) sorted $CD8^+$ T cells, cell subtype prediction were limited to effector memory $CD8^+$ T cell, central memory $CD8^+$ T cell, terminal effector $CD8^+$ T cell, naïve $CD8^+$ T cell, and MAIT (mucosal-associated invariant T) cell subtypes. Top 10 marker genes for each predicted cell subtype were defined and visualized with *pl.rank_genes_groups_matrixplot* function within Scanpy Python-package (v1.6.1) (29).

Unsupervised cell clustering. To perform unsupervised cell clustering analysis, the gene expression matrix containing UMI counts was loaded into Python3 (v3.8.7) with Scanpy Python-package (v1.7.2) (29). The UMI

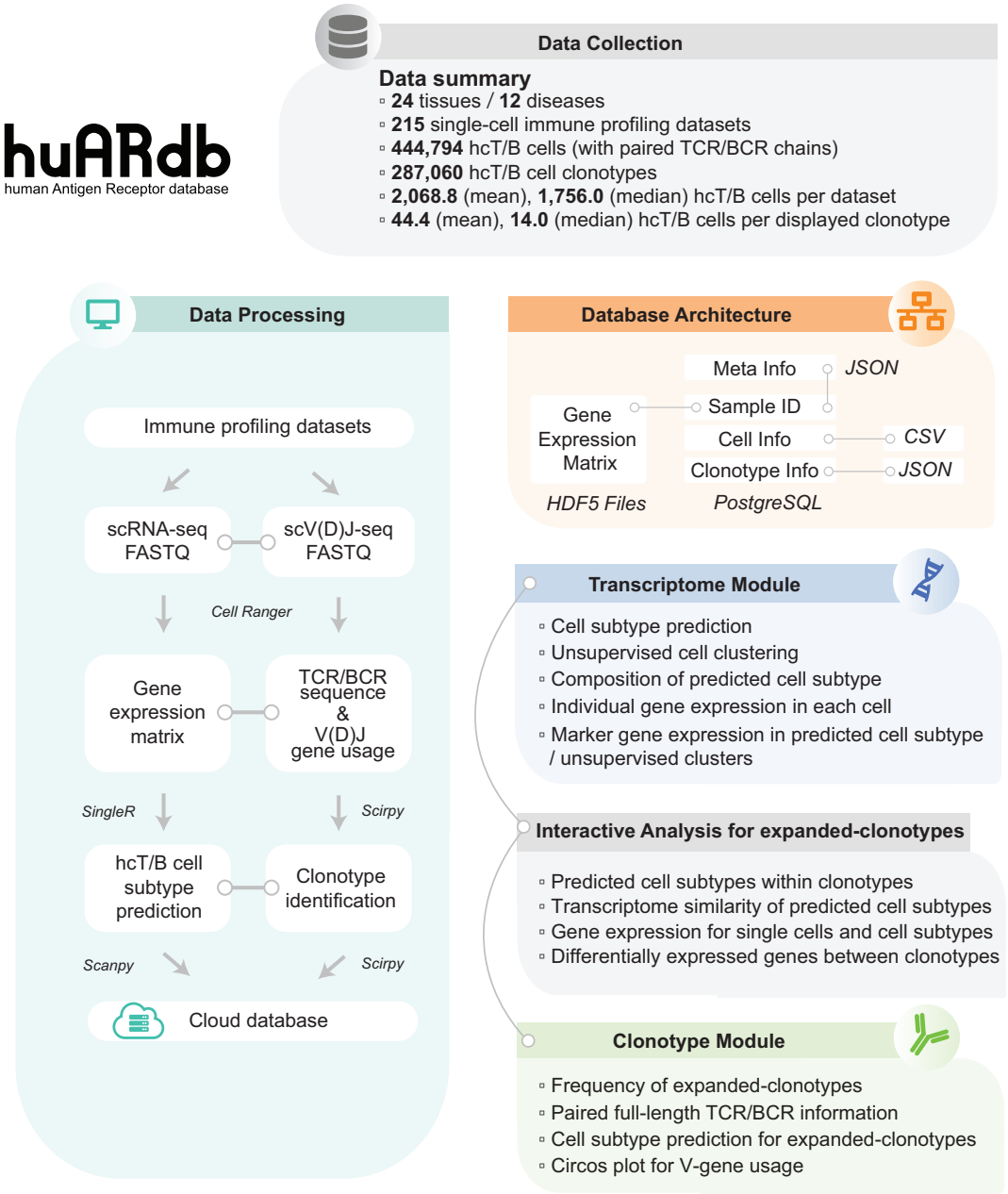

Figure 1. Overview of huARdb workflow and modules. All single-cell immune profiling datasets were retrieved from public databases (SRA, GEO, and GSA). All datasets were uniformly processed with our workflow. All processed data were stored in PostgreSQL database and HDF5 files, using sample ID as primary keys. For the web user interface, various interactive visualization and data analysis modules were provided for analyzing transcriptome and clonotype features.

counts were normalized to counts per million (CPM) with *scanpy.pp.normalize_total* function, followed by log-transformation and PCA (principal component analysis) using *scanpy.pp.log1p* and *scanpy.tl.pca* functions (29). The neighborhood graph was calculated based on the PCA results (*scanpy.pp.neighbors* function) and the Leiden algorithm was used to perform unsupervised cell clustering (*scanpy.tl.leiden*) (29,30). T-Distributed Stochastic Neighbor Embedding (tSNE) was used for the data visualization of predicted cell subtypes and unsupervised cell clustering. Top 10 marker genes for cell clusters were defined and visualized with *pl.rank_genes_groups_matrixplot* function within *Scanpy* Python-package (v1.6.1) (29).

Characterization of high confidence T or B cells (hcT/B cells)

The V(D)J gene annotation and CDR3 sequences of each TCR/BCR were loaded in Python3 (v3.8.7) with Scirpy (v0.7.0) Python-package (31). To define high confidence T or B cells (hcT/B cells), we implemented a stringent two-

step quality control strategy. In the first step, only cells with both transcriptome and TCR/BCR information were retained for downstream analysis. Next, only cells with paired light/heavy chains (BCR) or α/β chains (TCR) were considered as valid T or B cells in our database. Cells with extra, orphan, or unpaired (two chains with the same type) TCR/BCR chains were filtered out. In the end, all the single-cell immune profiling data displayed in our database contained transcriptome information and strictly paired TCR/BCR chains for each cell. Datasets with less than 30 cells were excluded from our database. With this stringent filtering strategy, 753 385 out of 1 198 179 cells were discarded and 444 794 hcT/B cells were displayed in huARdb.

Clonotype identification

The clonotype of hcT/B cells in our database was defined via the *pp.ir_dist* and *tl.define_clonotypes* functions within Scirpy (v0.7.0) Python-package (Figure 1) (31). Briefly, the nucleotide sequences of CDR3 were used to define clonotypes. Only cells with the same CDR3 nucleotide sequences in both VJ and VDJ chains (e.g. the same CDR3 on both light and heavy BCR chains, or both α and β TCR chains) were characterized as an identical clonotype.

Data analysis functions

To analyze transcriptome similarity for predicted cell subtypes in top 10 expanded-clonotypes, pairwise Pearson correlation coefficient for the average gene expressions for cells within a predicted cell subtype for certain clonotypes were calculated Pandas (v1.2.3) and Numpy (v1.19.5) Python-package (32). The pairwise Pearson correlation coefficient can be visualized in a heatmap on the webpage.

To identify pairwise differentially expressed genes (DEG) for top 10 expanded-clonotypes, gene expression fold change was first calculated based on the average CPM over certain clonotype using Pandas (v1.2.3) and Numpy (v1.19.5) Python-package (32). P-value were calculated using Student's t-test followed by FDR (False Discovery Rate) multiple testing correction using Scipy (v1.7.0) and Statsmodels (v0.10.1) Python-package (33). The DEG of any two expanded-clonotypes can be visualized in a volcano plot on the webpage.

To analyze the V gene usage for all hcT/B cells in a sample, V genes usage frequency on α (or light) chains and β (or heavy) chains were calculated and the V gene pairing information was stored. The top 60%, 80%, 90% ranking percentile for the most used V genes can be visualized in a Circos plot on the webpage.

Web implementation for the database

We developed the huARdb, a multi-functional and user-friendly database with advanced interactivity and visualization to present our uniformly analyzed single-cell immune profiling datasets. The front-end interface was developed with HTML5 and CSS3 languages, and all data visualizations were developed through Javascript using D3.js framework (34). The back-end data containing cell, clonotype, and metadata information were maintained and could be queried through the PostgreSQL database management system (v2.6.0) (Figure 1 and Supplementary Figure S1). Python3 (v3.7.9) and Javascript were used to communicate between the back-end and the front-end. The huARdb database is deployed with an Nginx web server (v1.14.1) on a Linux CentOS (v8.3.2011) cloud server system and is freely available at https://huarc.net/database without any registration or login. All features of huARdb were thoroughly tested on Google Chrome and Apple Safari browsers and were also accessible and legible on phone and tablet screens.

RESULTS

Data summary

Currently, our database contains cells from 24 tissue types and 12 diseases (Figure 1). We collected 231 coupled scRNA-seq and scV(D)J-seq datasets. After quality control and data filtering (see Methods), 215 datasets consisting of 444 794 hcT/B cells with paired TCR/BCR chains were retained. On average, 2 069 hcT/B cells and 13 493 genes were captured in each dataset (Figure 1 and Supplementary Figure S2A, B). In total, we characterized 287 060 hcT/B cell clonotypes with 1 335 clonotypes on average in each dataset (Figure 1 and Supplementary Figure S2C). To discover clonotypes with potential biological significance, we displayed the top 10 expanded-clonotypes for each dataset in huARdb. On average, we obtained 44 hcT/B cells per expanded-clonotype (Figure 1 and Supplementary Figure S2D).

With cell subtype prediction (see Methods), we classified 402 557 hcT cells (90.5%) and 42 237 hcB cells (9.5%) in huARdb (Supplementary Figure S2E). T cells were further classified into 13 different subtypes including effector memory $CD8^+$ T cells, Th1/Th17 cells, T regulatory cells, etc., while B cells were further classified into naïve B cells, exhausted B cells, non-switched memory B cells, switched memory B cells, and plasmablasts (Supplementary Figure S2F, G). For the top 10 expanded-clonotypes, we observed the enrichment of plasmablasts and effector memory T cells (Supplementary Figure S2H-J).

Since the CDR3 amino acid sequences play critical roles in antigen receptor recognition, we also analyzed the amino acid length and usage for CDR3 on both VJ chains (α and light chains) and VDJ chains (β and heavy chains). For hcT cells, the CDR3 amino acid length peaked around 13 for α chains and around 15 for β chains (Supplementary Figure S3A, B). For hcB cells, the CDR3 amino acid length peaked around 11 for light chains and around 16 for heavy chains (Supplementary Figure S3C, D). We observed an enriched usage for Glycine (G), Alanine (A), Phenylalanine (F), leucine (L) and Serine (S) for TCR CDR3 amino acid usage in both α and β chains (Supplementary Figure S3E, F). For BCR CDR3 usage, Glutamine (Q), Threonine (T), and S had higher usage rates in the light chain, while Tyrosine (Y), G, and Aspartate (D) were enriched in the heavy chain (Supplementary Figure S3G, H). In summary, huARdb collected amounts of single-cell immune profiling datasets and formed the human antigen receptor map with diverse features.

Utility of huARdb

All samples included in huARdb illustrated coupled transcriptome-clonotype information for each hcT/B cell. For the transcriptome feature analysis, we provided users predicted cell subtypes as well as unsupervised cell clustering. Users may visualize the composition of predicted cell subtypes as well as individual gene expression in each cell. To uncover clonotypes with potential biological significance, we displayed the top 10 expanded-clonotypes for each dataset. The huARdb web interface also enables users to explore clonotype-transcriptome modules interactively. For example, a user may explore cell numbers, predicted cell subtype composition, and individual gene expression for certain clonotypes concurrently (Figure 1).

Home page. The home page displays the metadata information for datasets collected in huARdb. To provide a better user experience, we generated a tutorial video for huARdb (Figure 2 and Supplementary Video). Users may query specific datasets performed in tissue types, diseases, projects and specific samples. The detailed publication and sample information will be displayed on the home page. Uniform Resource Locator (URL) links to the raw data are provided (Figure 2A). After users submit the query, detailed analysis results of the dataset will be exhibited on a new webpage. Users may submit no more than four samples to enable cross-sample comparison function. A link to a detailed tutorial on how to use the database is included on the home page.

Transcriptome features of the dataset. In the transcriptome module of huARdb, we provide cell subtype prediction and unsupervised cell clustering analysis along with interactive visualization of gene expression level at the single-cell and cell subtype level. Figures used for demonstration (Figures 2 and 3) were based on Luoma *et al.* published datasets, which were the immune profile of colon $CD3^+$ cells in checkpoint inhibitor (CPI)-treated melanoma patients with colitis (sample name: 'CPIc_C2') (35).

Users may choose to visualize predicted cell subtypes, or unsupervised cell clustering results via dropdown options on top of the tSNE plot. In the mode of unsupervised clustering, cells were clustered by the Leiden algorithm with variable options for parameters controlling clustering resolution (30)(see Methods). In the mode of cell subtype prediction, cell subtypes were calculated automatically based on the transcriptome similarity compared to the reference dataset (28,36) for hcT/B cells (see Methods). Users may select certain cell subtypes via the radio buttons on the top of the webpage, and cells predicted as the selected subtype will be spotlighted on the tSNE plot (Figure 2B). The predicted cell subtype composition of the dataset will be displayed on a bar plot, and the sample 'CPIc_C2' revealed enriched effector memory $CD8^+$ T cells in the colon (Figure 2C).

To visualize gene expression levels at the single-cell level, a colored tSNE plot was employed (Figure 2D). Meanwhile, the gene expression level in each predicted cell subtype could also be analyzed via the violin plot, in which the short red line represents the average expression level in each predicted cell subtype and the grey dots represent the gene expression level in each cell (Figure 2E). The expression level of the marker genes for the predicted cell subtypes were visualized with heatmap on the webpage (Figure 2F).

To illustrate the use of the transcriptome module of huARdb, we queried 'Colon' for tissue, 'Ulcerative Colitis (UC)' for disease, 'Daniele Corridoni *et al.*' for paper, and 'S33' for sample. This is a single-cell immune profiling dataset of UC patient colon $CD8^+$ T cell collected in the huARdb database. Interleukin (IL)-26 expression was elevated in inflammatory bowel disease (IBD) and UC, according to Corridoni *et al.* and other investigations (19,37). To confirm the previous findings, we queried 'IL26' in the text input option on top of the tSNE plot to evaluate *IL26* expression patterns across different predicted cell subtypes. Our analysis indicated that *IL26* was highly expressed in UC patient T cells, particularly in predicted MAIT cells (Supplementary Figure S4A). As a control, we queried *IL26* expression in a healthy control dataset (Sample 'S22' from the same study) and observed a low level of *IL26* expression in all cell subtypes (Supplementary Figure S4B). Furthermore, IL17A expression was upregulated in the MAIT cells of UC patients compared to healthy control (Supplementary Figure S4C, D) (19,38). Altogether, the transcriptome module of huARdb displays cell subtype prediction, unsupervised cell clustering and individual gene expression information for each collected dataset.

Clonotype features of the dataset. We reasoned that the top ten largest clonotypes in diseases likely exhibit biological significance in disease settings. Thus, we allow users to explore the features of the top ten largest clonotypes in huARdb. For each dataset, we displayed the frequency of the top 10 expanded-clonotypes with a network plot and a bar plot (Figure 3A, C). At the same time, users may explore and download the paired full-length TCR/BCR information including receptor type, V, D, J and C gene usage, CDR3 sequences, pairing information for VJ (α/light) and VDJ (β/heavy) chains in the top 10 expanded-clonotypes (Figure 3B).

Interactive clonotype-transcriptome analysis. In the clonotype module of huARdb, users can perform interactive clonotype-transcriptome analysis at the single-cell level.

(i) *predicted cell subtypes in each expanded-clonotype*.

We allow users to analyze the composition of cell subtypes in each expanded-clonotype in a network and bar plot (Figure 3A, F). With this function, biologists may infer the potential biological significance for certain expanded-clonotypes.

(ii) *gene expression level in each expanded-clonotype*.

If users identified an expanded-clonotype of interest, by clicking on specific bars in the clonotype frequency bar plot (Figure 3C), users may explore and visualize the gene expression level and cell subtype information on the tSNE and violin plot (Figure 3C–E). With this function, users may analyze the expression level of a specific gene for certain cell subtypes in a specifically expanded-clonotype.

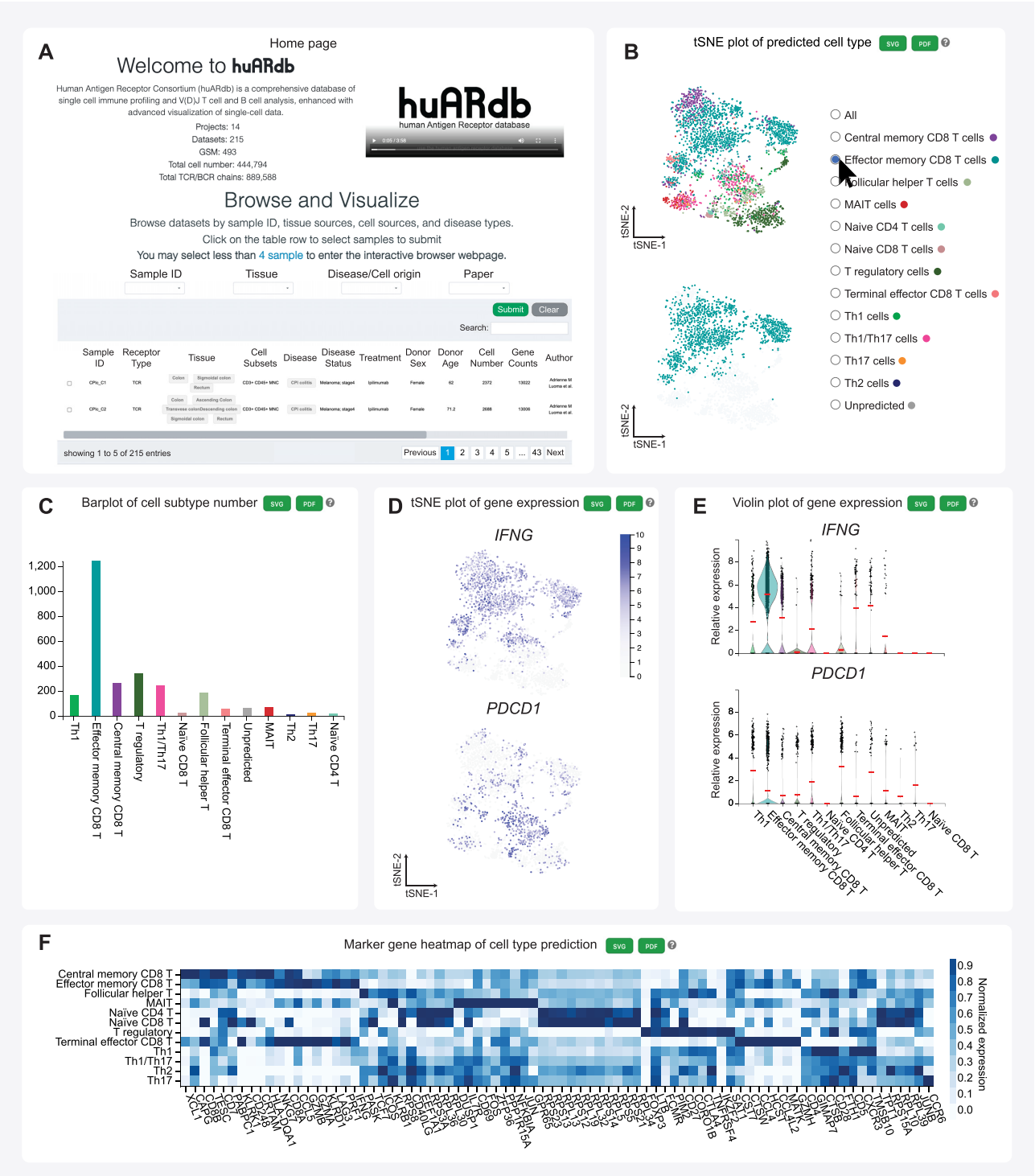

Figure 2. The transcriptome module of huARdb. (**A**) Screenshot for the huARdb home page. Users can select samples from various tissues, diseases and papers. (**B**) The tSNE visualization module for predicted cell subtype. Users can select a cell subtype on the side menu and obtain a zoomed-in view. (**C**) The bar plot shows the number of predicted cell subtype in the dataset. (**D**, **E**) The gene expression modules. The web interface provides a feature plot (D) and violin plot (E) for visualizing individual gene expression level. The red bar in the violin plot represents the average expression level for certain predicted cell subtype. *IFNG* and *PDCD1* were used as example genes for gene expression modules. (**F**) Heatmap of marker genes for predicted cell types. The average expression values are normalized between 0 to 1. Sample 'CPIc_C2' from Luoma *et al.* was used as examples in this figure (35). All panels were exported and downloaded from the web interface.

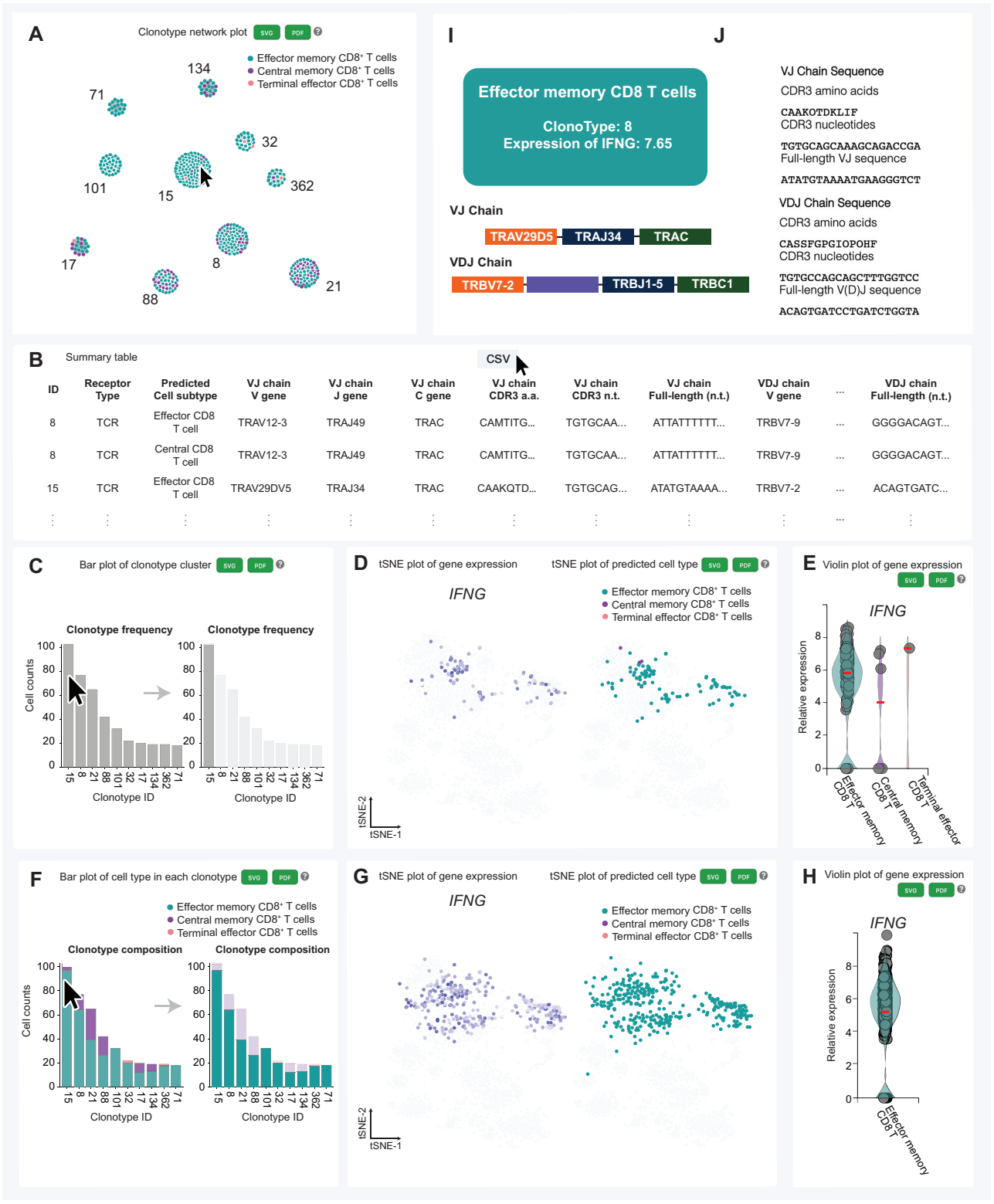

Figure 3. The clonotype module of huARdb. (**A**) The web interface provides clonotype network plot of the top 10 expanded-clonotypes. (**B**) Summary table of single cell information. (**C**) Bar plots showing number of cells of the top 10 expanded-clonotypes. After clicking the bar in the left panel, the bar will be highlighted, and the tSNE plot (**D**) and violin plot (**E**) exhibit the cells in the selected clonotype. (**F**) Bar plots showing cell subtype composition of the top 10 expanded-clonotypes. After clicking the bar, the tSNE plot (**G**) and violin plot (**H**) exhibit effector memory CD8$^+$ T cells in all top 10 expanded-clonotypes. (**I**, **J**) After clicking a node in (**A**), the clonotype information including predicted cell subtype (**I**, upper panel), V(D)J gene usage (**I**, lower panel), amino acid/nucleotide sequences of CDR3 (**J**), and full-length V(D)J sequence (**J**) of the selected cell will be displayed in the menu bar on the web interface (**I**). Sample 'CPIc_C2' from Luoma *et al.* was used as examples in this figure (35). *IFNG* were used as example genes for gene expression modules. All panels were exported and downloaded from the web interface.

(iii) *gene expression level in cell subtypes of interest for all top 10 expanded-clonotypes*.

If users are interested in specific cell subtypes in all top 10 expanded-clonotypes, by clicking on the colored bars in the clonotype composition bar plot (Figure 3F), users may explore and visualize the gene expression level and distribution for certain cell subtypes on the tSNE and violin plot (Figure 3F–H). With this function, users may analyze the expression level of a specific gene for a specific cell subtype in all top 10 expanded-clonotypes.

To exemplify the interactive clonotype-transcriptome analysis with huARdb, we analyzed the clonotype and clonotype-related transcriptome features from Corridoni *et al.* published sample 'S33' dataset (19). For clonotype features of the dataset, in the UC patient, the cell count of the largest clonotype was about 2-fold more than the second-largest clonotype, while the top two clonotypes contained a similar count of cells in the healthy control from the same study (sample 'S22') (Supplementary Figure S5A, B). In the largest expanded-clonotype in the UC patient (sample 'S33', clonotype 19), most T-cells were predicted as MAIT-cells with highly expressed *IL26* and *IL17A* (Supplementary Figure S5C–E). By contrast, in the largest expanded-clonotype in the healthy control (sample 'S22', clonotype cluster 73), most cells were predicted as effector memory $CD8^+$ T cells with no expression of *IL26* and *IL17A* (Supplementary Figure S5F–H). Given its enrichment and the highly expressed *IL26* and *IL17A*, clonotype 21 in sample 'S33' may play a critical role in recognizing and targeting self-antigens, leading to an autoimmune response in the UC patient.

In summary, huARdb provides the coupled clonotype-transcriptome analysis. In addition to the overall clonotype characteristics of the dataset, huARdb provides the transcriptome features of the enriched clonotypes in each dataset through variable interactive functions, including the cell subtype composition and the individual gene expression pattern for top 10 expanded-clonotypes.

Single cell information. By double-clicking the dots on the tSNE plot, the violin plot, or clonotype network plot, users may view the individual cell information displayed on the side menu, including (i) the predicted cell subtype, (ii) the V, D, J and (C) gene usage of TCR/BCR, (iii) the full-length amino acid and nucleotide sequence of V(D)J gene used by the TCR/BCR, (iv) the amino acid and nucleotide sequence of TCR/BCR CDR3 and (v) the expression level of a specific gene (Figure 3I, J).

Data analysis functions. HuARdb also provide users various analysis function for transcriptome and clonotype features. Pairwise Pearson correlation coefficient measuring the transcriptome similarity for predicted cell subtypes within the top 10 expanded-clonotypes could be visualized via a heatmap on the webpage (Figure 4A). We also allow users to explore the differentially expressed genes (DEG) for between any clonotypes of interest within the top 10 expanded-clonotypes. The fold change and statistics significance for the DEGs could be visualized via a volcano plot on the webpage (Figure 4B). In addition, the V gene usage for all hcT/B cells and pairing information could be visualized via a Circos plot on the webpage (Figure 4C).

Cross-sample comparison function. To enable users to compare multiple samples simultaneously, we allow user to submit no more than four samples on the homepage to enter the cross-sample comparison mode. In this mode, most the plots and functions mentioned above could be visualized and analyzed side by side (Supplementary Figure S6).

Data download. HuARdb enables the user to download detailed information including predicted cell subtypes, V, D, J and C gene usage, and CDR3 sequences for each cell in the top 10 expanded-clonotypes.

Data deposit and customized URL generation for publication. As part of the mission of huARdb is to build a holistic atlas of human antigen receptors, other scientists are encouraged to deposit their single-cell immune profiling datasets to huARdb. We will generate an interactive visualization with unique URL subpaths for each dataset. Thus, scientists could incorporate the URL in their publications and provide readers with a non-static, interactive way to explore the high-dimensional single-cell immune profiling data.

DISCUSSION AND FUTURE EXTENSIONS

The expanding use of single-cell sequencing technologies has generated a vast number of publicly available datasets, including single-cell immune profiling. Our huARdb database collects, analyses, and provides a comprehensive user-friendly web-based resource for biologists to interrogate the potential biological significance of individual TCR/BCR sequences under different settings. Through the multi-functional interactive web interface, huARdb allows coupled clonotype-transcriptome analysis. The full-length TCR/BCR sequence information at the single-cell level enables the pairing of TCR/BCR chains and provides sufficient information for biologists to clone the receptor subunit genes and perform functional analysis of TCRs/BCRs of interest. The coupled transcriptome information may further assist biologists to infer the cell subtype and function with specific TCRs/BCRs in different disease settings.

In the future, we plan to extend huARdb as follows. We will continue to collect publicly available human single-cell immune profiling data to expand the tissues and diseases covered in huARdb. Together with our collaborators, we are also generating novel single-cell immune profiling data for various immune-related diseases, and we will include our data on huARdb. Meanwhile, the algorithm for cell subtype prediction could be further improved in huARdb. At this stage, huARdb uses SingleR to predict T or B cell subtypes, based on the reference dataset published by Monaco *et al.* (28,36). However, some T or B cell subtypes may be absent in the references, such as exhausted T cells (36). To avoid cell subtype mis-annotation, huARdb also provides unsupervised clustering results based on the Leiden algorithm (30). Users may manually annotate cell subtypes based on signature gene expression. In the future, with more comprehensive and accurate immune cell reference datasets available, we would update our reference dataset for cell subtype

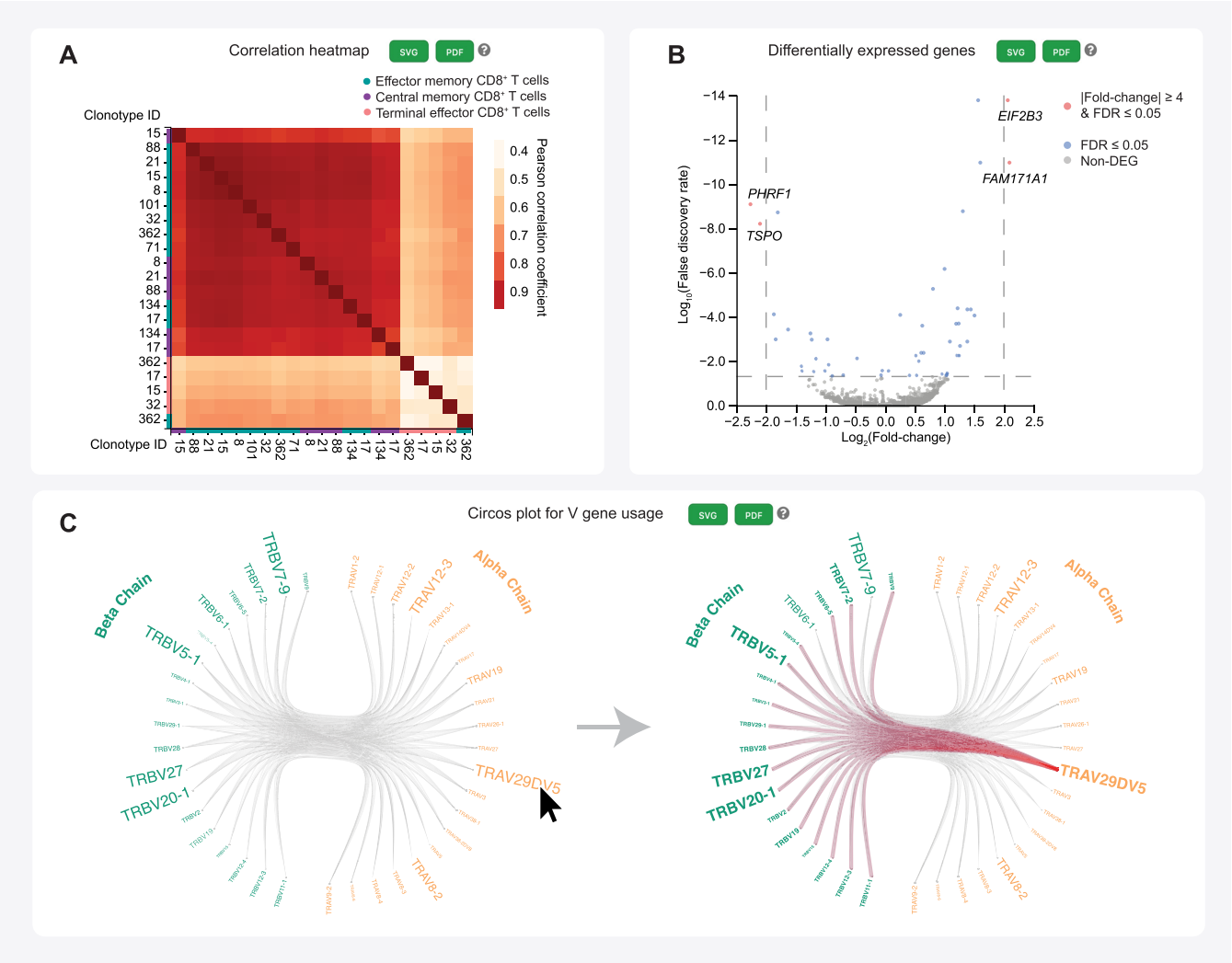

Figure 4. The transcriptome-clonotype analysis module of huARdb. (**A**) The correlation heatmap of the transcriptome for different predicted cell types in the top 10 expanded-clonotypes. Pearson correlation coefficient are used to reflect the similarity of the transcriptome. (**B**) The differentially expressed genes for any two clonotypes among the top 10 expanded-clonotypes. Users may select any two clonotypes in the web interface for comparison. (**C**) Circos plot for V gene usage for all hcT/B cells in the sample. Sample 'CPIc_C2' from Luoma *et al.* was used as examples in this figure (35). All panels were exported and downloaded from the web interface.

prediction. In addition, to define hcT/B cells, huARdb only includes T and B cells with paired α/β or light/heavy chains. However, some evidence suggests that a considerable proportion of T cells express two distinct α chains (39). Similarly, some B cells expressed dual surface immunoglobulin light chains (40). T cells with dual TCRs and B cells expressing dual immunoglobins might play important roles in promoting autoimmunity (40–42). Therefore, we will include T or B cells that express dual α or dual light chains in future versions of huARdb.

In conclusion, our huARdb will be a rapidly growing resource available to facilitate the in-depth study of TCRs/BCRs under different immunological settings. As more single-cell immune profiling data are generated, we invite more scientists to deposit their data on huARdb. Through our collective efforts, huARdb will be a unique resource that provides the most up-to-date information on functional immune receptors.

DATA AVAILABILITY

HuARdb is a database with online and open access, available at https://huarc.net/database.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank all researchers who generated the single-cell immune profiling datasets that are collected, analyzed, and displayed in huARdb. We thank Dr Chen, Di and Dr Liu, Jian from Zhejiang University for helpful discussions and lab members of Lab Liu and Lu for their valuable discussions and suggestions. We thank the Life Science Editors Foundation for editing services.

FUNDING

National Natural Science Foundation of China [31930038, 31770954, 31530019 to L.L.]; Fundamental Research Funds for the Central Universities [2021QN81016 to L.W.]; Innovative Institute of Basic Medical Sciences of Zhejiang University [to L.L.]; Alibaba Cloud [to L.W.]. Funding for open access charge: National Natural Science Foundation of China [31930038].

Conflict of interest statement. The authors have filed a patent related to this work.

REFERENCES

1. Lee,S.W. and Whelan,R.L. (2006) Immunologic and oncologic implications of laparoscopic surgery: what is the latest? *Clin. Colon Rectal. Surg.*, **19**, 5–12.
2. Natoli,G. and Ostuni,R. (2019) Adaptation and memory in immune responses. *Nat. Immunol.*, **20**, 783–792.
3. Dong,D., Zheng,L., Lin,J., Zhang,B., Zhu,Y., Li,N., Xie,S., Wang,Y., Gao,N. and Huang,Z. (2019) Structural basis of assembly of the human T cell receptor-CD3 complex. *Nature*, **573**, 546–552.
4. Treanor,B. (2012) B-cell receptor: from resting state to activate. *Immunology*, **136**, 21–27.
5. Roth,D.B. (2014) V(D)J recombination: mechanism, errors, and fidelity. *Microbiol. Spectr.*, **2**, 18.
6. Stadinski,B.D., Trenh,P., Duke,B., Huseby,P.G., Li,G., Stern,L.J. and Huseby,E.S. (2014) Effect of CDR3 sequences and distal V gene residues in regulating TCR-MHC contacts and ligand specificity. *J. Immunol.*, **192**, 6071–6082.
7. Chi,X., Li,Y. and Qiu,X. (2020) V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. *Immunology*, **160**, 233–247.
8. D'Angelo,S., Ferrara,F., Naranjo,L., Erasmus,M.F., Hraber,P. and Bradbury,A.R.M. (2018) Many routes to an antibody heavy-chain CDR3: necessary, yet insufficient, for specific binding. *Front. Immunol.*, **9**, 395.
9. Davis,M.M., Boniface,J.J., Reich,Z., Lyons,D., Hampl,J., Arden,B. and Chien,Y. (1998) Ligand recognition by alpha beta T cell receptors. *Annu. Rev. Immunol.*, **16**, 523–544.
10. Zheng,M., Zhang,X., Zhou,Y., Tang,J., Han,Q., Zhang,Y., Ni,Q., Chen,G., Jia,Q., Yu,H. *et al.* (2019) TCR repertoire and CDR3 motif analyses depict the role of alphabeta T cells in Ankylosing spondylitis. *EBioMedicine*, **47**, 414–426.
11. Dondelinger,M., Filee,P., Sauvage,E., Quinting,B., Muyldermans,S., Galleni,M. and Vandevenne,M.S. (2018) Understanding the significance and implications of antibody numbering and antigen-binding surface/residue definition. *Front. Immunol.*, **9**, 2278.
12. Qi,Q., Liu,Y., Cheng,Y., Glanville,J., Zhang,D., Lee,J.Y., Olshen,R.A., Weyand,C.M., Boyd,S.D. and Goronzy,J.J. (2014) Diversity and clonal selection in the human T-cell repertoire. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 13139–13144.
13. Laydon,D.J., Bangham,C.R. and Asquith,B. (2015) Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **370**, 20140291.
14. Mayer,A., Zhang,Y., Perelson,A.S. and Wingreen,N.S. (2019) Regulation of T cell expansion by antigen presentation dynamics. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 5914–5919.
15. Rosati,E., Dowds,C.M., Liaskou,E., Henriksen,E.K.K., Karlsen,T.H. and Franke,A. (2017) Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnol.*, **17**, 61.
16. Allez,M., Auzolle,C., Ngollo,M., Bottois,H., Chardiny,V., Corraliza,A.M., Salas,A., Perez,K., Stefanescu,C., Nancey,S. *et al.* (2019) T cell clonal expansions in ileal Crohn's disease are associated with smoking behaviour and postoperative recurrence. *Gut*, **68**, 1961–1970.
17. Matos,T.R., O'Malley,J.T., Lowry,E.L., Hamm,D., Kirsch,I.R., Robins,H.S., Kupper,T.S., Krueger,J.G. and Clark,R.A. (2017) Clinically resolved psoriatic lesions contain psoriasis-specific IL-17-producing alphabeta T cell clones. *J. Clin. Invest.*, **127**, 4031–4041.
18. Ping,Y., Liu,C. and Zhang,Y. (2018) T-cell receptor-engineered T cells for cancer treatment: current status and future directions. *Protein Cell*, **9**, 254–266.
19. 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.
20. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
21. Leinonen,R., Sugawara,H., Shumway,M. and International Nucleotide Sequence Database, C. (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–D21.
22. Wang,Y., Song,F., Zhu,J., Zhang,S., Yang,Y., Chen,T., Tang,B., Dong,L., Ding,N., Zhang,Q. *et al.* (2017) GSA: Genome Sequence Archive. *Genomics Proteomics Bioinformatics*, **15**, 14–18.
23. Zheng,G.X., Terry,J.M., Belgrader,P., Ryvkin,P., Bent,Z.W., Wilson,R., Ziraldo,S.B., Wheeler,T.D., McDermott,G.P., Zhu,J. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.
24. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck,W.M. III, Zheng,S., Butler,A., Lee,M.J., Wilk,A.J., Darby,C., Zager,M. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.
25. McGinnis,C.S., Murrow,L.M. and Gartner,Z.J. (2019) DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.*, **8**, 329–337.
26. Xi,N.M. and Li,J.J. (2021) Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. *Cell Syst.*, **12**, 176–194.
27. McCarthy,D.J., Campbell,K.R., Lun,A.T. and Wills,Q.F. (2017) Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, **33**, 1179–1186.
28. Aran,D., Looney,A.P., Liu,L., Wu,E., Fong,V., Hsu,A., Chak,S., Naikawadi,R.P., Wolters,P.J., Abate,A.R. *et al.* (2019) Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.*, **20**, 163–172.
29. Wolf,F.A., Angerer,P. and Theis,F.J. (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 15.
30. Traag,V.A., Waltman,L. and van Eck,N.J. (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.*, **9**, 5233.
31. Sturm,G., Szabo,T., Fotakis,G., Haider,M., Rieder,D., Trajanoski,Z. and Finotello,F. (2020) Scirpy: a Scanpy extension for analyzing single-cell T-cell receptor-sequencing data. *Bioinformatics*, **36**, 4817–4818.
32. Harris,C.R., Millman,K.J., van der Walt,S.J., Gommers,R., Virtanen,P., Cournapeau,D., Wieser,E., Taylor,J., Berg,S., Smith,N.J. *et al.* (2020) Array programming with NumPy. *Nature*, **585**, 357–362.
33. Virtanen,P., Gommers,R., Oliphant,T.E., Haberland,M., Reddy,T., Cournapeau,D., Burovski,E., Peterson,P., Weckesser,W., Bright,J. *et al.* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261–272.
34. Bostock,M., Ogievetsky,V. and Heer,J. (2011) D-3: data-driven documents. *IEEE. T. Vis. Comput. Gr.*, **17**, 2301–2309.
35. Luoma,A.M., Suo,S., Williams,H.L., Sharova,T., Sullivan,K., Manos,M., Bowling,P., Hodi,F.S., Rahma,O., Sullivan,R.J. *et al.* (2020) Molecular pathways of colon inflammation induced by cancer immunotherapy. *Cell*, **182**, 655–671.
36. Monaco,G., Lee,B., Xu,W., Mustafah,S., Hwang,Y.Y., Carre,C., Burdin,N., Visan,L., Ceccarelli,M., Poidinger,M. *et al.* (2019) RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell Rep.*, **26**, 1627–1640.
37. Fujii,M., Nishida,A., Imaeda,H., Ohno,M., Nishino,K., Sakai,S., Inatomi,O., Bamba,S., Kawahara,M., Shimizu,T. *et al.* (2017) Expression of Interleukin-26 is upregulated in inflammatory bowel disease. *World J. Gastroenterol*, **23**, 5519–5529.
38. Haga,K., Chiba,A., Shibuya,T., Osada,T., Ishikawa,D., Kodani,T., Nomura,O., Watanabe,S. and Miyake,S. (2016) MAIT cells are activated and accumulated in the inflamed mucosa of ulcerative colitis. *J. Gastroenterol. Hepatol.*, **31**, 965–972.

39. Schuldt,N.J. and Binstadt,B.A. (2019) Dual TCR T cells: identity crisis or multitaskers? *J. Immunol.*, **202**, 637–644.
40. Fraser,L.D., Zhao,Y., Lutalo,P.M., D'Cruz,D.P., Cason,J., Silva,J.S., Dunn-Walters,D.K., Nayar,S., Cope,A.P. and Spencer,J. (2015) Immunoglobulin light chain allelic inclusion in systemic lupus erythematosus. *Eur. J. Immunol.*, **45**, 2409–2419.
41. Yamaguchi,S., Hamana,H., Shitaoka,K., Sukegawa,K., Nagata,T., Hayee,A., Kobayashi,E., Ozawa,T., Fujii,T., Muraguchi,A. *et al.* (2021) TCR function analysis using a novel system reveals the multiple unconventional tumor-reactive T cells in human breast cancer-infiltrating lymphocytes. *Eur. J. Immunol.*, **51**, 2306–2316.
42. Alli,R., Nguyen,P., Boyd,K., Sundberg,J.P. and Geiger,T.L. (2012) A mouse model of clonal CD8+ T lymphocyte-mediated alopecia areata progressing to alopecia universalis. *J. Immunol.*, **188**, 477–486.