https://doi.org/10.1093/nar/gkac930

SPEED: Single-cell Pan-species atlas in the light of Ecology and Evolution for Development and Diseases

Yangfeng Chen[1,2,†], Xingliang Zhang[3,4,†], Xi Peng[1,2,5,†], Yicheng Jin[1,2,†], Peiwen Ding[1,2,†], Jiedan Xiao[1,2], Changxiao Li[1,2], Fei Wang[6], Ashley Chang[1,2], Qizhen Yue[7], Mingyi Pu[8], Peixin Chen[9], Jiayi Shen[10], Mengrou Li[11], Tengfei Jia[11], Haoyu Wang[5], Li Huang[12], Guoji Guo[13], Wensheng Zhang[9,10], Hebin Liu[11], Xiangdong Wang[14,*] and Dongsheng Chen[1,2,*]

[1]Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005, China, [2]Suzhou Institute of Systems Medicine, Suzhou 215123, China, [3]Department of Respiratory Diseases, Institute of Pediatrics, Shenzhen Children's Hospital, Shenzhen 518038, China, [4]Department of Pediatrics, the Affiliated Hospital of Guangdong Medical University, Zhanjiang 524001, China, [5]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China, [6]Department of Biomedicine, Aarhus University, Aarhus 8000, Denmark, [7]State Key Laboratory of Cell Biology, CAS Center for Excellence in Molecular Cell Science, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai 200031, China, [8]Department of Medicine, Sun Yat-sen University, Shenzhen 518106, China, [9]Cam-Su Genomic Resource Center, Medical College of Soochow University, Suzhou 215123, China, [10]Peninsula Cancer Research Center, School of Basic Medical Sciences, Binzhou Medical University, Yantai 264003, China, [11]Institutes of Biology and Medical Sciences (IBMS), Soochow University, Suzhou 215123, China, [12]The Future Laboratory, Tsinghua University, Beijing 100084, China, [13]Center for Stem Cell and Regenerative Medicine, Zhejiang University School of Medicine, Hangzhou 310058, China and [14]Zhongshan Hospital, Department of Pulmonary and Critical Care Medicine, Institute for Clinical Science, Shanghai Institute of Clinical Bioinformatics, Shanghai 200032, China

Received August 14, 2022; Revised September 30, 2022; Editorial Decision October 03, 2022; Accepted October 12, 2022

ABSTRACT

It is a challenge to efficiently integrate and present the tremendous amounts of single-cell data generated from multiple tissues of various species. Here, we create a new database named SPEED for single-cell pan-species atlas in the light of ecology and evolution for development and diseases (freely accessible at http://8.142.154.29 or http://speedatlas.net). SPEED is an online platform with 4 data modules, 7 function modules and 2 display modules. The 'Pan' module is applied for the interactive analysis of single cell sequencing datasets from 127 species, and the 'Evo', 'Devo', and 'Diz' modules provide comprehensive analysis of single-cell atlases on 18 evolution datasets, 28 development datasets, and 85 disease datasets. The 'C2C', 'G2G' and 'S2S' modules explore intercellular communications, genetic regulatory networks, and cross-species molecular evolution. The 'sSearch', 'sMarker', 'sUp', and 'sDown' modules allow users to retrieve specific data information, obtain common marker genes for cell types, freely upload, and download single-cell datasets, respectively. Two display modules ('HOME' and 'HELP') offer easier access to the SPEED database with informative statistics and detailed guidelines. All in all, SPEED is an integrated platform for single-cell RNA sequencing (scRNA-seq) and single-cell whole-genome sequencing (scWGS) datasets to assist the deep-mining and understanding of heterogeneity among cells, tissues, and species at multi-levels, angles, and orientations, as well as provide new insights into molecular mechanisms of biological development and pathogenesis.

[*]To whom correspondence should be addressed. Tel: +86 512 62873780; Fax: +86 512 62873779; Email: cds@ism.pumc.edu.cn
Correspondence may also be addressed to Xiangdong Wang. Email: xdwang@fuccb.com

[†]The authors wish it to be known that, in their opinion, the first five authors should be regarded as Joint First Authors.

INTRODUCTION

Heterogeneity is ubiquitous among cells, tissues, organs, and species. High cellular diversity exists in the mammalian lung (1,2). Early embryonic cells and pluripotent stem cells are differentiated into function-specific cells within tissues and organs by switching on/off diverse gene expression patterns during the development (3–5). The interactions between gene expression and stimuli play decisive roles in the maintenance of health status or disease development (6). Each cell type demonstrates different transcriptome changes in response to microbial and environmental insults (7,8). It is critical to compare gene expression profiles of cell types between normal and pathological tissues and clarify the biological significance and pathogenesis mechanisms.

Methods of cell classification by cell morphology or specific gene/protein expression panels characterize well-known cell types, while lacking the systematicity and correlation. The transcriptome or proteome at bulk cells and tissue fails to define the gene expression profile at single cell resolution. The scRNA-seq is an advanced high-throughput sequencing technology to capture the transcriptome of single cells on a large scale and to unbiasedly classify cell types in complex tissues and dissect the gene expression in individual cells (9). scRNA-seq provides opportunities to reliably identify novel or rare cell types (10), discriminate closely related cell populations (11), and define cell heterogeneity, differentiation, and development (12). With rapid development, scRNA-seq offers an unprecedented impetus to the initiation of an ambitious international consortium, the Human Cell Atlas (HCA) (13). In addition, the cellular atlas in multiple tissues or organs of other species has been accomplished by scRNA-seq in recent years, e.g. tissue cells from 11 non-model animals (14), circulating immune cells among 12 species (15), eight organs and tissues of mouse embryos at different stages (12), mice lung cells across aging stages (16,17), and pig cerebral cortex, hypothalamus (18), and lung (2). Single-cell transcriptomic data from pathologic tissues of patients and animal models of diseases is generated rapidly, e.g. cells from lower airways of asthmatic patients (19), fibrotic lungs of mice (20), and three brain regions of the Parkinson's disease model mouse (21).

Fast-increasing amounts of single-cell data on multiple tissues among different species are generated worldwide. The challenges are the deposition of raw data in scRNA-seq datasets and the difficulty of deep-mining the needed information for non-bioinformatics scientists. Therefore, effective integration and presentation of scRNA-seq datasets will facilitate and maximize the utilization by the research community. Raw data and expression matrix datasets from single-cell transcriptomes are submitted to several free academic websites such as Gene Expression Omnibus (22) and ArrayExpress (23). Several websites were established for collecting single-cell datasets across multiple species, such as Single Cell Portal with 11 species (https://singlecell.broadinstitute.org/single_cell), Single Cell Expression Atlas with 20 species (24), and UCSC Cell Browser with 15 species (25). Some portals were created specifically to collect scRNA-seq data for certain types of tissues or organs. LungMAP was established for illustrating the human and mouse lung molecular atlas (26), and ScdbLung for exploring single-cell data from the lung of human, mouse, rat, and pig (2). Of diseases-related scRNA-seq databases, CancerSEA provides distinct functional states of single-cell expression signatures for 25 cancer types (27), and CSEA-DB for 598 GWAS traits associated with the underlying cell types on basis of scRNA-seq (28). SC2disease is an accurate resource of gene expression profiles for 25 diseases (29). TISCH integrates scRNA-seq profiles from 76 tumor datasets (30). PBatlas was recently established as a versatile website to access the pig brain atlas (18). The CellMarker database records cell markers for human tissues and mouse tissues (31). VThunter presents the cell expression pattern of 107 virus receptors from 47 animal species (32).

Although some databases are focused on specific applications, there is still a practical need to construct databases with more species and with a special focus on the heterogeneities among cells, tissues, organs, and species. Thus, it is urgent to build a systematic, hierarchical, and comprehensive gene expression profiling database from public scRNA-seq data. To address those issues, we centralized multiple scRNA-seq datasets and established a comprehensive and freely accessible online database called SPEED.

DATA COLLECTION AND DATABASE CONTENT

A total of 5,770,427 cells from 634 datasets across 127 species were collected into the SPEED database. The scRNA-seq datasets were downloaded from multiple sources, including NCBI/GEO (22), EMBL-EBI/SCEA (24), ArrayExpress (23), HCL (33), UCSC Cell Browser (25), MCA (34), and Single Cell Portal (https://singlecell.broadinstitute.org/single_cell), prioritizing studies with RDS files or gene expression matrix available. The navigation menu contains 13 modules, such as 'HOME', 'Pan', 'Evo', 'Devo', 'Diz', 'C2C', 'G2G', 'S2S', 'sSearch', 'sUp', 'sDown', 'sMarker', and 'HELP' (Figure 1A). All detailed information on scRNA-seq datasets of 122 species, is found on the 'sDown' module of SPEED, including data sources, timing, technologies, common and Latin names of species, sample tissues, treatments, and cell numbers. In addition, we collected the single-cell WGS (sc-WGS) datasets from 16 species in the 'Pan' module (35). The whole-genome sequencing (WGS) technology provides an in-depth description of single nucleotide polymorphisms (SNPs), small insertions, and deletions (INDELs).

The scRNA-seq datasets were checked manually and processed using Seurat v4.1.1 (36). In brief, quality control was conducted for the downstream analysis by retaining cells with nFeature_RNA between 200 and 6000. The 'NormalizeData' function in Seurat v4.1.1 was then utilized to normalize gene expression matrix of the sparse single cell. The 'FindVariableFeatures' function was used to identify highly variable genes. The top 2000 highly variable genes were then dimensionally reduced using the principal component analysis (PCA). The top 30 principal components were picked for clustering. The visualization of single cell data sets was achieved using ShinyCell package (37). Cellular communication was analyzed in the 'C2C' module using CellChat (38). In the 'G2G' module, putative interactions between

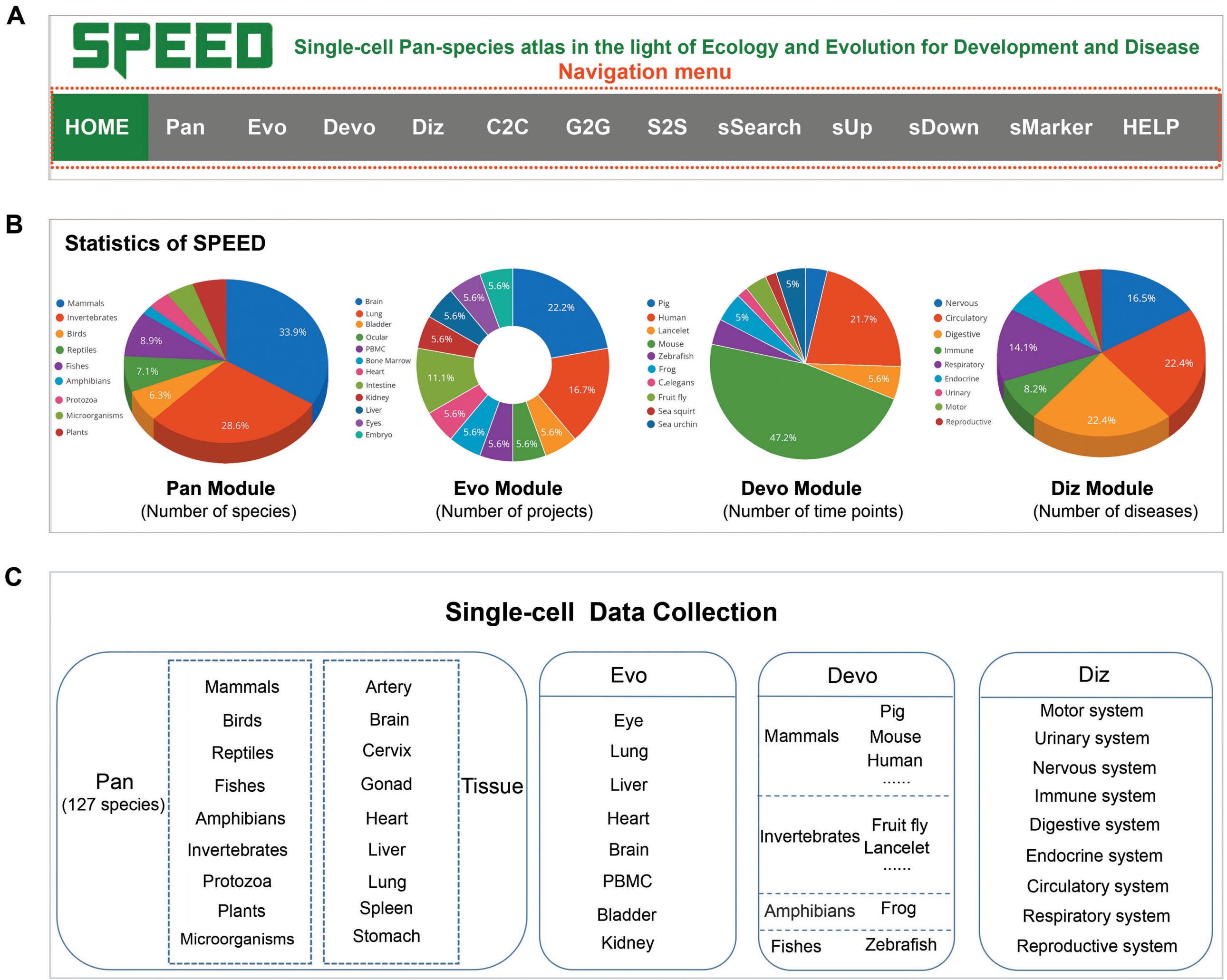


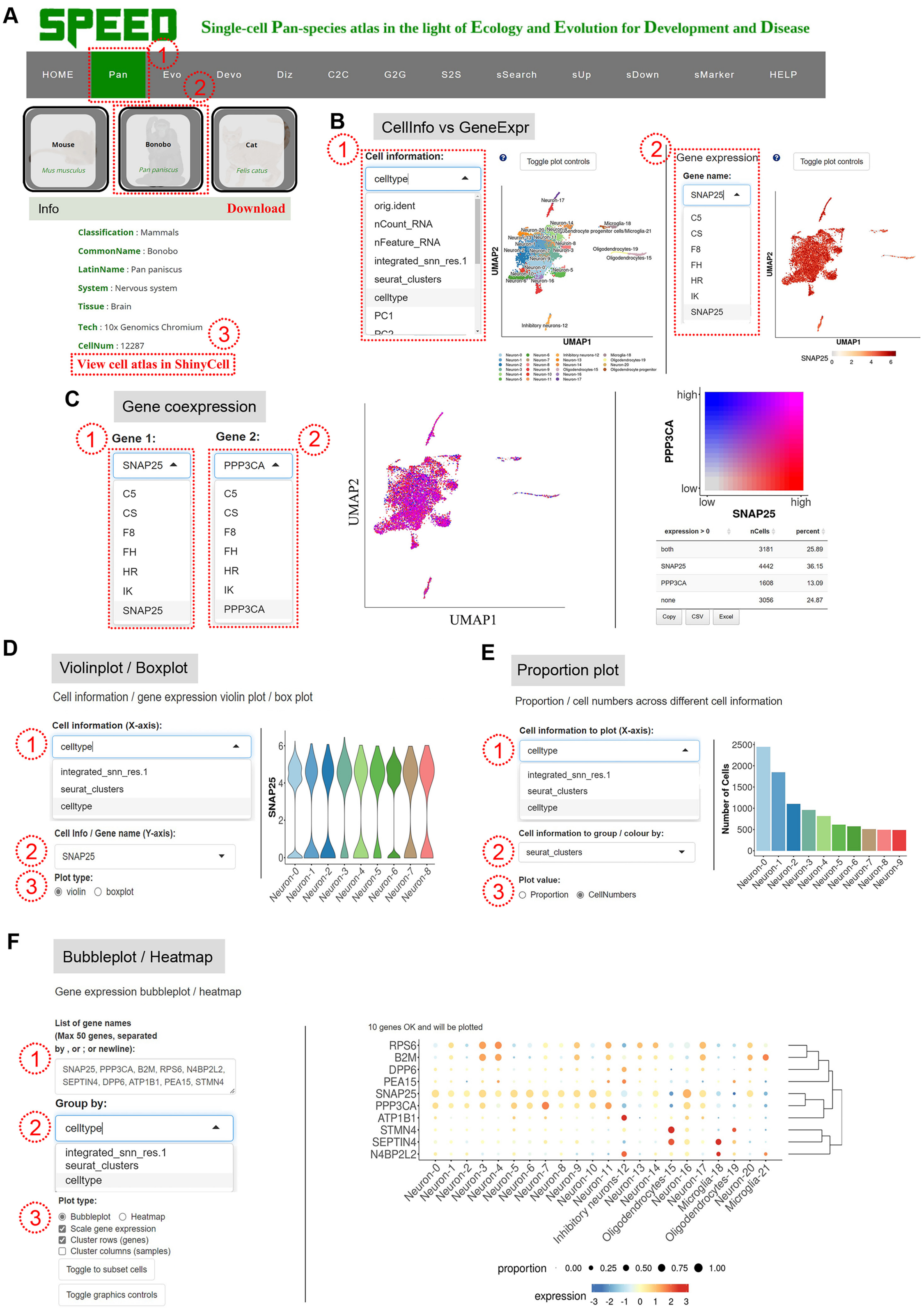
Figure 1. Overview of SPEED database. (**A**) Navigation menu of the SPEED website. (**B**) Statistics of Pan, Evo, Devo, and Diz modules in SPEED. (**C**) Single-cell data collection of the four data modules.

transcriptional factors (TFs) and downstream targets were predicted using GENIE3 (39) and GO enrichment was performed using clusterProfiler (40).

USER INTERFACE

The web application and search engine in SPEED are currently mounted on a high-performance Linux server. Users worldwide can freely access database for in-depth data visualization and custom analysis of 127 species and function modules for further personalized analysis. There are 13 modules in the navigation menu of SPEED (Figure 1A), which are grouped into three categories: display, data, and function. The display category includes 'HOME' and 'HELP' modules. The 'HOME' module introduces the statistics of datasets (Figure 1B). The 'HELP' module guides users to easily understand the SPEED user interface. A brief description and a frame-by-frame animation demonstration are provided for each module to help users to easily catch key points of the module usage.

SPEED is designed mainly for the data categories: 'Pan', 'Evo', 'Devo', and 'Diz' modules in the navigation menu (Figure 1C). The 'Pan' module is one of the most core modules of SPEED, containing scRNA-seq datasets from 122 species and sc-WGS data for 16 mammalian species. On the 'Pan' page, the 122 species scRNA-seq datasets were grouped into 9 classes (i.e. mammals, birds, reptiles, fishes, amphibians, invertebrates, protozoa, plants, and microorganisms), according to different kingdoms and phyla (Figure 1C). The 'Pan' module allows the interactive exploration of each individual cell-type gene expression profile. If users wonder the cell-types-specific expression profile of a certain gene of interest in a certain tissue of a certain species, they can perform the following operations to obtain the desired information as shown in Figure 2A and B. When users move the cursor onto one species' image (e.g. 'Bonobo') (41), a sliding tab will pop up with the common species name in black font and its Latin name in green italics (Figure 2A). By clicking the sliding tab on the species image and the link in the column of 'Rds', users can browse the basic information of the species and the detailed information of single-cell atlas on a new tab page. Clicking the 'View cell atlas in ShinyCell' in red font guides users to a new page with



Classification : Mammals

CommonName : Bonobo

LatinName : Pan paniscus

System : Nervous system

Tissue : Brain

Tech : 10x Genomics Chromium

CellNum : 12287

Cell information:

celltype

orig.ident
nCount_RNA
nFeature_RNA
integrated_snn_res.1
seurat_clusters
celltype
PC1
PC2

Gene expression

Gene name:

SNAP25

C5
CS
F8
FH
HR
IK
SNAP25

Gene 1:

SNAP25

C5
CS
F8
FH
HR
IK
SNAP25

Gene 2:

PPP3CA

C5
CS
F8
FH
HR
IK
PPP3CA

expression > 0	nCells	percent
both	3181	25.89
SNAP25	4442	36.15
PPP3CA	1608	13.09
none	3056	24.87

Cell information / gene expression violin plot / box plot

Cell information (X-axis):

celltype

integrated_snn_res.1
seurat_clusters
celltype

Cell Info / Gene name (Y-axis):

SNAP25

Plot type:

● violin ○ boxplot

Proportion / cell numbers across different cell information

Cell information to plot (X-axis):

celltype

integrated_snn_res.1
seurat_clusters
celltype

Cell information to group / colour by:

seurat_clusters

Plot value:

○ Proportion ● CellNumbers

Gene expression bubbleplot / heatmap

List of gene names (Max 50 genes, separated by , or ; or newline):

SNAP25, PPP3CA, B2M, RPS6, N4BP2L2, SEPTIN4, DPP6, ATP1B1, PEA15, STMN4

Group by:

celltype

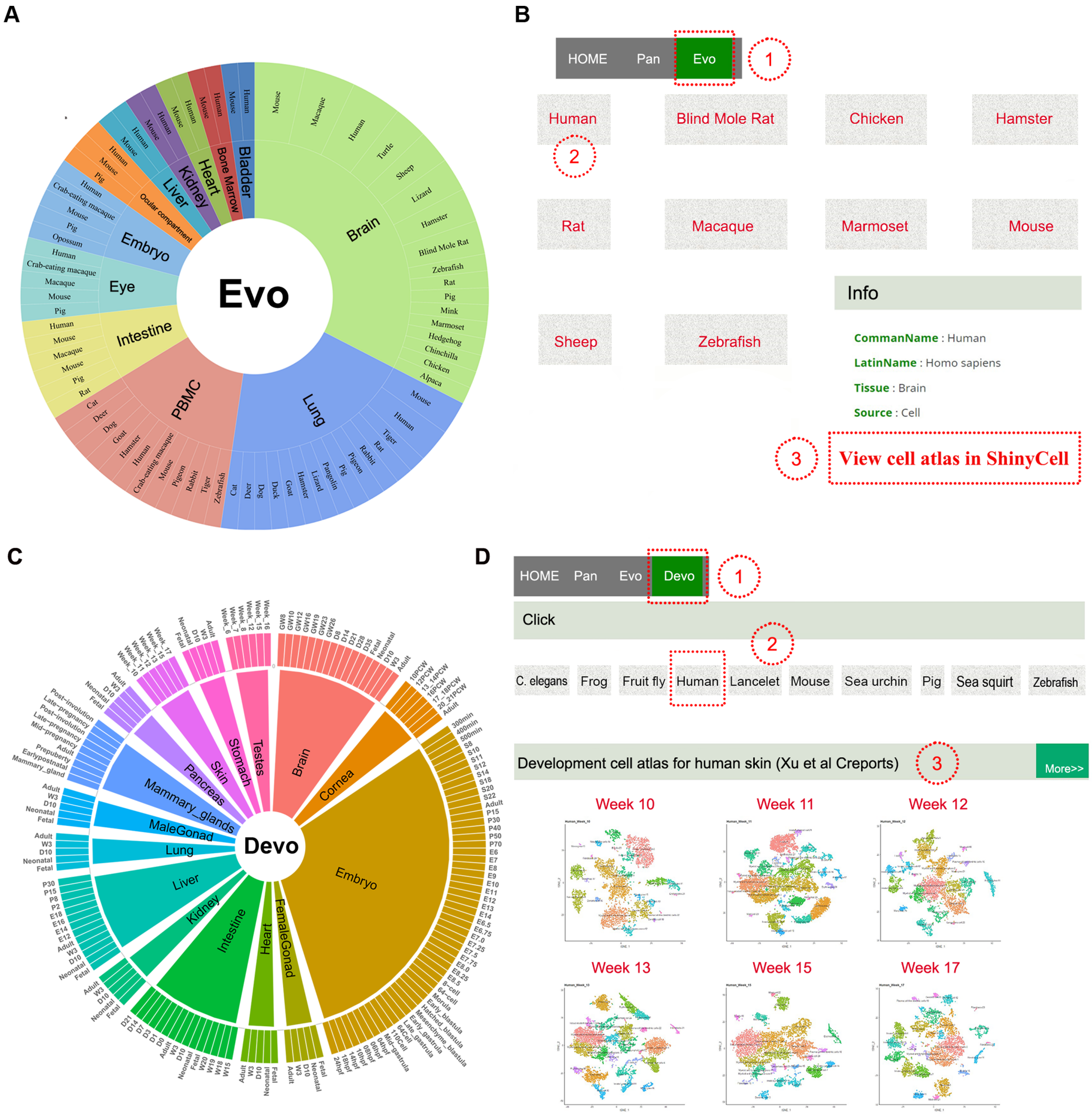
integrated_snn_res.1
seurat_clusters
celltype

Plot type:

● Bubbleplot ○ Heatmap
☑ Scale gene expression
☑ Cluster rows (genes)
□ Cluster columns (samples)

10 genes OK and will be plotted

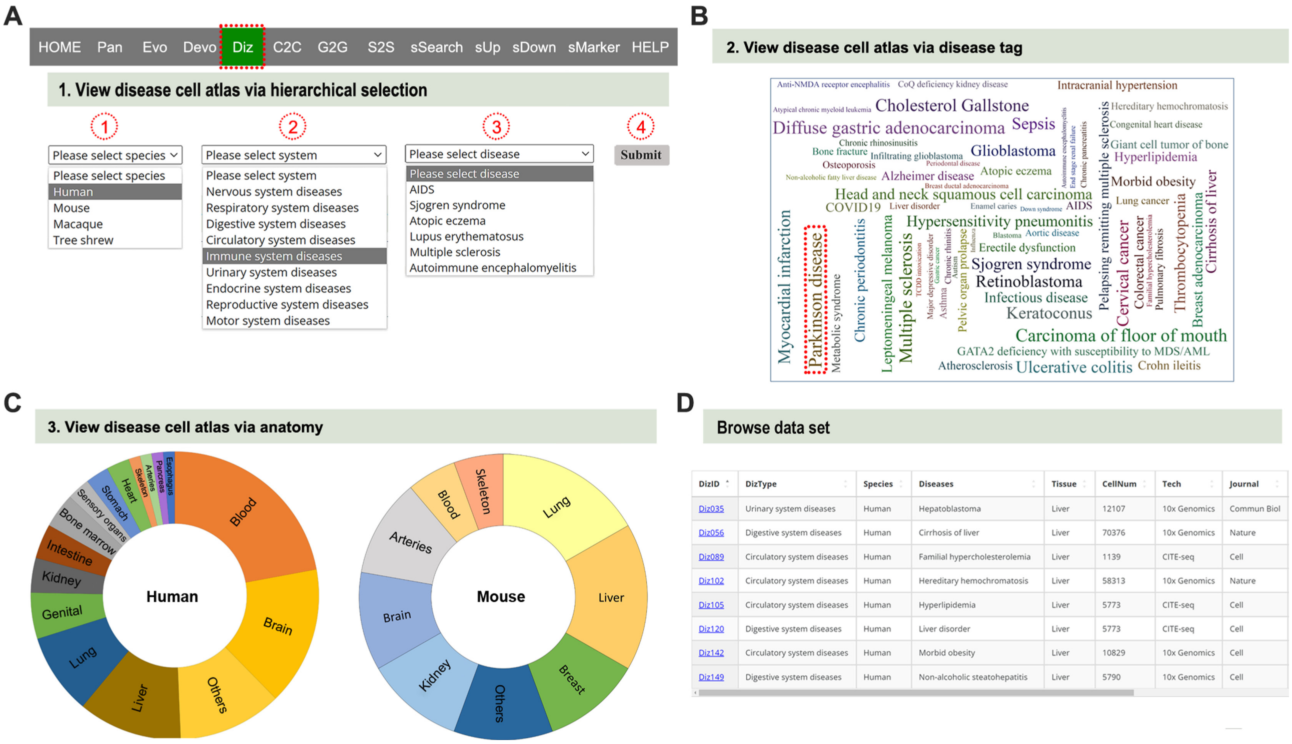
Figure 2. Demonstration of Pan module. (**A**) Operation to access cell atlas in ShinyCell on the species profile page and download RDS files. (**B**) Illustration of viewing multi-classified cell and gene expression information in ShinyCell. (**C**) Illustration of viewing gene co-expression information. (**D**) Operation to view cell or gene information in Violinplot or Boxplot. (**E**) Operation to view the single cell composition.(**F**) Operation to view the gene expression pattern in Bubbleplot/Heatmap.



CommanName : Human
LatinName : Homo sapiens
Tissue : Brain
Source : Cell

Figure 3. Overview of Evo and Devo modules. (**A**) Tissue types of each species in Evo module. Tissue types and species names are shown from the inner circle to the outer circle. (**B**) Operation to access the datasets in Evo module. (**C**) Distribution of tissue types in development time points in Devo module. (**D**) Operation to view the datasets in Devo module.

seven tabs on the navigation menu (Figure 2A). In doing so, users may continue to visualize gene expression profiles and cell-type information for genes of interest. The 'CellInfo vs GenExpr' tab simultaneously visualizes cell information and gene expression side-by-side on low-dimensional representations. There are multiple options (e.g. 'tSNE1' for X-axis and 'tSNE2' for Y-axis) to choose from in the drop-down box under the X-axis and Y-axis of 'Dimension Reduction'. Different dimensions of cell information are presented by switching the options (e.g. 'celltype') in the drop-down box under 'Cell Information' (Figure 2B, left). The cell-type gene expression profiles are visualized by choosing gene name of interest (e.g. '*SNAP25*') in the drop-down box under 'Gene Information' (Figure 2B, right). The 'CellInfo vs CellInfo' and 'GenExpr vs GenExpr' tabs visualize two cell information and two gene expressions side-by-side on low-dimensional representations, respectively (Figure S1). Researchers can analyze the co-expression relationships between two genes on each cell-type of a certain tissue of interest, which may be realized through the following steps. The 'Gene coexpression' tab visualizes the cell-type co-expression of two genes (e.g. '*SNAP25*' and '*PPP3CA*') on low-dimensional representations when users choose or input two genes of interest in the boxes under 'Gene Expression' button (Figure 2C). The 'Violinplot/Boxplot' tab visualizes the gene expression or continuous cell information



Browse data set

DizID	DizType	Species	Diseases	Tissue	CellNum	Tech	Journal
Diz035	Urinary system diseases	Human	Hepatoblastoma	Liver	12107	10x Genomics	Commun Biol
Diz056	Digestive system diseases	Human	Cirrhosis of liver	Liver	70376	10x Genomics	Nature
Diz089	Circulatory system diseases	Human	Familial hypercholesterolemia	Liver	1139	CITE-seq	Cell
Diz102	Circulatory system diseases	Human	Hereditary hemochromatosis	Liver	58313	10x Genomics	Nature
Diz105	Circulatory system diseases	Human	Hyperlipidemia	Liver	5773	CITE-seq	Cell
Diz120	Digestive system diseases	Human	Liver disorder	Liver	5773	CITE-seq	Cell
Diz142	Circulatory system diseases	Human	Morbid obesity	Liver	10829	10x Genomics	Cell
Diz149	Digestive system diseases	Human	Non-alcoholic steatohepatitis	Liver	5790	10x Genomics	Cell

Figure 4. Demonstration of Diz module. (**A**) Operation to search for the disease atlas by hierarchical selection. (**B**) Illustration of obtaining disease cell atlas by tissue name. (**C**) Illustration of viewing disease cell atlas by anatomical label. (**D**) Illustration of obtaining disease cell atlas in the summary table.

across groups of cells when users choose the option or input gene of interest in the boxes under 'Cell information (X-axis)' and 'Cell Info/Gene name (Y-axis)' (Figure 2D). The 'Proportion plot' tab shows the proportion and number of single cells (Figure 2E). The 'Bubbleplot/Heatmap' tab visualizes the gene expression patterns of multiple genes which are grouped by categorical cell information (e.g. celltype, RNA_snn_res.1, and seurat_clusters) (Figure 2F).

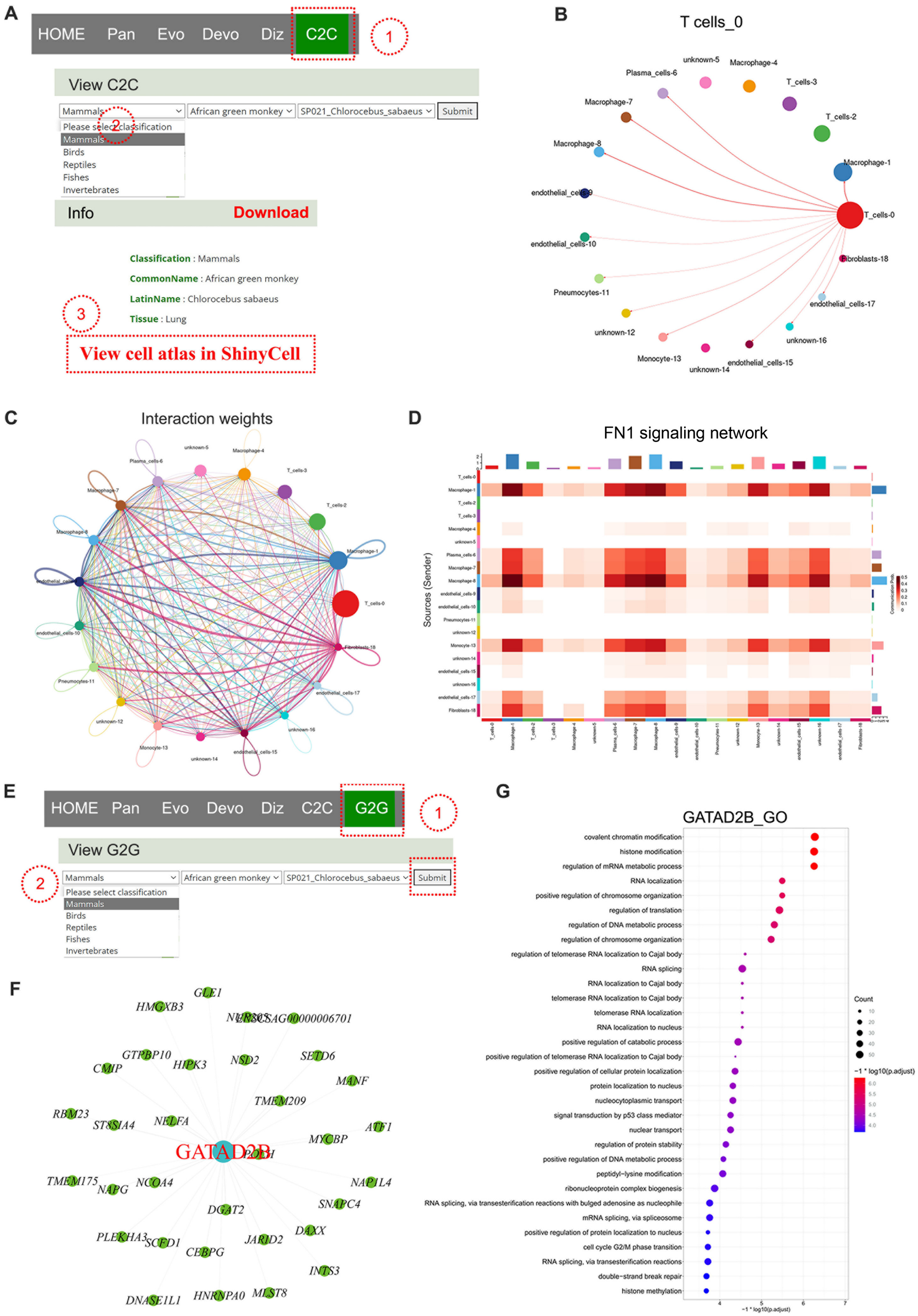
In addition to scRNA-seq datasets, sc-WGS datasets of 16 mammalian species are available in the tenth class on the 'Pan' module, which provide detailed information of SNPs and INDELs (35) (Figure S2A). By clicking the sliding tab on the species image (e.g. 'King colobus'), a new page pops up and shows basic information of the dataset and two link buttons, 'View SNP information' and 'View Indel information' (Figure S2B). When users click the two links buttons, a new tab page appears with detailed information of the SNPs or INDELs from sc-WGS datasets (Figure S2C-D), including chromosomes, start and end sites, intergenic, intronic or exonic, sequencing methods, and others.

The 'Evo' module displays 18 datasets of multiple tissues from 28 species, including brain, lung, heart, bladder, ocular compartment, eye, bone marrow, intestine, kidney, liver, embryo, and peripheral blood mononuclear cell (PBMC) (Figure 3A). Cross-species comparisons of single-cell atlas of mammals, reptiles, and other species provide a reference evolutionary reservoir of developmental programs to explore potential networks among evolutionarily distant species. When users click the dataset picture (e.g. 'Brain data sets (Geirsdottir et al Cell)') on the 'Evo' module (Figure 3B), the current page switches to a new page to show all species pictures and their common names linked with the corresponding literature (42). By clicking the species picture (e.g. 'Human'), users get more detailed information on the scRNA-seq atlas via the link button 'View cell atlas in ShinyCell' in a new tab page (Figure 3B).

The developmental single-cell atlas is critical for understanding of stem cell biology. The 'Devo' module contains 28 single-cell datasets from 10 species (Figure 3C). On the 'Devo' page, the scRNA-seq dataset picture at each developmental stage (e.g. 'Week 10' under 'Homo sapiens') (43) (Figure 3D) was linked with a new tab page to show more details of the dataset by clicking the button 'View cell atlas in ShinyCell'.

The 'Diz' module encompasses scRNA-seq datasets of 85 diseases (Figure 4A), including neurological, respiratory, digestive, cardiovascular, immunological, urinary, endocrinological, reproductive, and motor system diseases (Figure 4A). The disease-related scRNA-seq atlas (e.g. 'Parkinson disease') (44) can be viewed via the species-system-disease hierarchical select (Figure 4A), the disease tag (Figure 4B), or the tissue name (Figure 4C), to direct users to a new tab page with the basic information of the disease and dataset. By clicking the link button (e.g. 'Diz035) (45) in the column of 'DizlD' (Figure 4D), users can obtain more detailed information on disease-related scRNA-seq atlas via the link button 'View cell atlas in ShinyCell' in a new tab page.

Seven function modules in the navigation menu include 'C2C', 'G2G', 'S2S', 'sSearch', 'sUp', 'sDown', and 'sMarker' (Figure 1A). The 'C2C' module provides potential cell interaction networks and signaling pathways between cell types. The potential intercellular communication


A
HOME
Pan
Evo
Devo
Diz
C2C
1
View C2C
Mammals
African green monkey
SP021_Chlorocebus_sabaeus
Submit
Please select classification
Mammals
Birds
Reptiles
Fishes
Invertebrates
2
Info
Download
3
View cell atlas in ShinyCell
B
T cells_0
unknown-5
Macrophage-4
Plasma_cells-6
T_cells-3
Macrophage-7
T_cells-2
Macrophage-8
Macrophage-1
endothelial_cells-9
T_cells-0
endothelial_cells-10
Fibroblasts-18
Pneumocytes-11
endothelial_cells-17
unknown-12
unknown-16
Monocyte-13
endothelial_cells-15
unknown-14
C
Interaction weights
D
FN1 signaling network
Sources (Sender)
Communication Prob.
E
HOME
Pan
Evo
Devo
Diz
C2C
G2G
1
View G2G
2
Mammals
African green monkey
SP021_Chlorocebus_sabaeus
Submit
Please select classification
Mammals
Birds
Reptiles
Fishes
Invertebrates
F
GATAD2B
GLE1
HMGXB3
NUP205
ENSCSAG00000006701
GTPBP10
CMIP
HIPK3
NSD2
SETD6
MANF
TMEM209
RBM23
ST8SIA4
NELFA
ATF1
MYCBP
POGK
TMEM175
NAPG
NCOA4
NAP1L4
SNAPC4
DGAT2
PLEKHA3
SCFD1
CEBPG
JARID2
DAXX
INTS3
DNASE1L1
HNRNPA0
MEST8
G
GATAD2B_GO
covalent chromatin modification
histone modification
regulation of mRNA metabolic process
RNA localization
positive regulation of chromosome organization
regulation of translation
regulation of DNA metabolic process
regulation of chromosome organization
regulation of telomerase RNA localization to Cajal body
RNA splicing
RNA localization to Cajal body
telomerase RNA localization to Cajal body
telomerase RNA localization
RNA localization to nucleus
positive regulation of catabolic process
positive regulation of telomerase RNA localization to Cajal body
positive regulation of cellular protein localization
protein localization to nucleus
nucleocytoplasmic transport
signal transduction by p53 class mediator
nuclear transport
regulation of protein stability
positive regulation of DNA metabolic process
peptidyl-lysine modification
ribonucleoprotein complex biogenesis
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
mRNA splicing, via spliceosome
positive regulation of protein localization to nucleus
cell cycle G2/M phase transition
RNA splicing, via transesterification reactions
double-strand break repair
histone methylation
Count
10
20
30
40
50
−1 * log10(p.adjust)
6.0
5.5
5.0
4.5
4.0
4
5
6
7
−1 * log10(p.adjust)

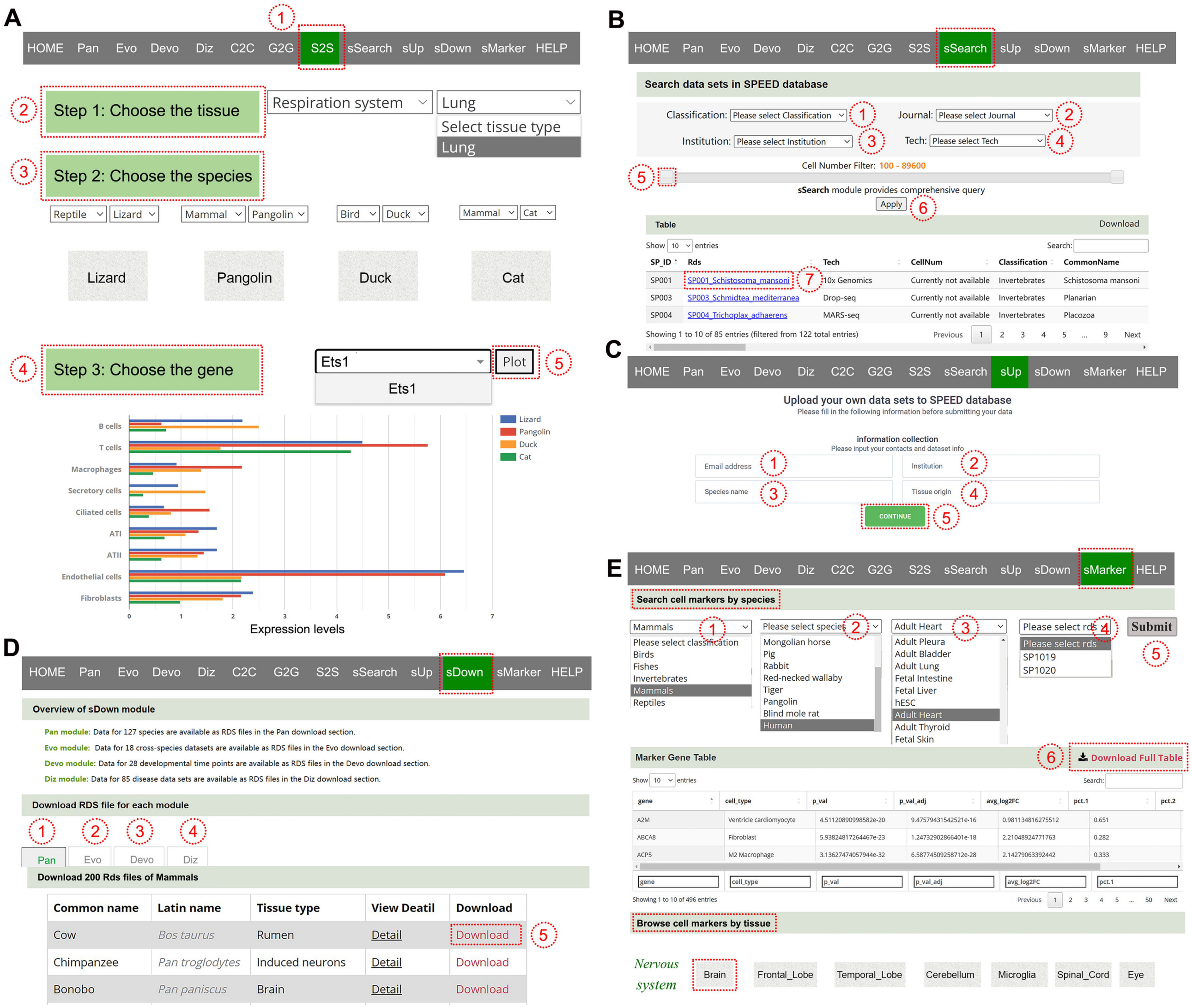
Classification : Mammals

CommonName : African green monkey

LatinName : Chlorocebus sabaeus

Tissue : Lung

Figure 5. Demonstration of C2C and G2G modules. (**A**) Operation to search for the dataset in C2C module. (**B**) Intercellular communications between one cell type and other cell types. (**C**) Intercellular communications for all cell types. (**D**) Heatmap showing the FN1 signalling pathway network among cell types. (**E**) Operation to search for the dataset in G2G module. (**F**) Genetic regulatory networks (GRNs) of *GATAD2B* and its candidate target genes. (**G**) Gene Ontology (GO) term enrichment analysist of *GATAD2B*.



SP_ID	Rds	Tech	CellNum	Classification	CommonName
SP001	SP001_Schistosoma_mansoni	10x Genomics	Currently not available	Invertebrates	Schistosoma mansoni
SP003	SP003_Schmidtea_mediterranea	Drop-seq	Currently not available	Invertebrates	Planarian
SP004	SP004_Trichoplax_adhaerens	MARS-seq	Currently not available	Invertebrates	Placozoa

Showing 1 to 10 of 85 entries (filtered from 122 total entries) Previous 1 2 3 4 5 ... 9 Next

gene	cell_type	p_val	p_val_adj	avg_log2FC	pct.1	pct.2
A2M	Ventricle cardiomyocyte	4.51120890998582e-20	9.47579431542521e-16	0.981134816275512	0.651	
ABCA8	Fibroblast	5.93824817264467e-23	1.24732902866401e-18	2.21048924771763	0.282	
ACP5	M2 Macrophage	3.13627474057944e-32	6.58774509258712e-28	2.14279063392442	0.333	

Pan module: Data for 127 species are available as RDS files in the Pan download section.

Evo module: Data for 18 cross-species datasets are available as RDS files in the Evo download section.

Devo module: Data for 28 developmental time points are available as RDS files in the Devo download section.

Diz module: Data for 85 disease data sets are available as RDS files in the Diz download section.

Common name	Latin name	Tissue type	View Deatil	Download
Cow	*Bos taurus*	Rumen	Detail	Download
Chimpanzee	*Pan troglodytes*	Induced neurons	Detail	Download
Bonobo	*Pan paniscus*	Brain	Detail	Download

Figure 6. Demonstration of S2S, sSearch, sUp, sDown, and sMarker modules. (**A**) Steps to view gene expression level for cell types of four species in S2S module. (**B**) Operation to search for the dataset in sSearch module. (**C**) Operation to upload the dataset to SPEED in sUp module. (**D**) Operation to download the dataset of four data modules in sDown module. (**E**) Operation to obtain marker genes of organs in sMarker module.

mediated by ligand-receptor pairs analysis is performed using R package CellChat (38). The ligand-receptor pairs are assigned to different signaling pathways. On the ‘C2C’ page, when users choose one species (e.g. ‘African green monkey’) (46) under the ‘View C2C’ drop-down box via hierarchical selection and then click ‘Submit’ button (Figure 5A), a new tab page appears and presents the basic information of dataset, intercellular communication networks, and signaling pathways. The intercellular communications of one cell type with other cell types are shown in one image (Figure 5B), and all intercellular communications among cell types are merged in the last image (Figure 5C). The thickness of lines between cell types represents the strength of intercellular communication. Some signaling pathways are mainly enriched in intercellular communications between certain cell types. Each heatmap in Figure 5D shows one signaling pathway network among particular cell types. The table at the bottom shows the detailed information on every ligand-receptor interaction and related signaling pathway between any two cell types. Users can get customized cellular information by inputting the keywords in the ‘Search’ box.

TFs and candidate target genes in each cell type are determined based on the scRNA-seq data. The ‘G2G’ module predicts the putative GRNs between TFs and candidate targets, to promote the understanding of the transcriptional regulation on the cell type level. Users view GRNs of each TF via hierarchical selection at the drop-down box of ‘View G2G’ (46) (Figure 5E). The TF name in red font is located in the hub of GRNs, around which candidate target genes are scattered (Figure 5F). The ‘View GO enrichment for each regulon’ catalogue performs GO (Gene Ontology) term enrichment analysis on candidate target genes of each regulon. The top GO terms for each regulon are shown under the ‘View GO enrichment for each regulon’ item (Figure 5G).

The ‘S2S’ module allows cross-species comparison of expression patterns of transcription factor encoding gene.

Users query the cell-type expressions of one gene (Figure 6A) by selecting the tissue, choosing four species from four drop-down boxes, and selecting one gene and then clicking the 'Plot' button. Afterwards, a horizontal bar chart is generated to show the gene expression on cell types across species (14) (Figure 6A).

The 'sSearch' module provides an optional query mode combination for detailed information on scRNA-seq datasets. On the 'sSearch' page, users can query individually or in combination of multiple criteria (Figure 6B). After users click the 'Apply' button, the searched results are shown in the table below (Figure 6B). The RDS file of interest in the 'Table' could be downloaded via clicking the link in blue font of Rds (Figure 6B).

The 'sUp' module allows users to upload their own datasets to SPEED database. Users need to fill in email address, institution, species name, and tissue origin in the columns of 'information collection' before uploading the datasets (Figure 6C).

The 'sDown' module allows users to download all datasets in Rds file format. On the 'sDown' page, users choose one (e.g. 'Pan') of the 'Pan', 'Evo', 'Devo', and 'Diz' modules under the 'Download RDS file for each module', followed by the appearance of a key information table with details and download links for all Rds files in 'Pan' module (Figure 6D). Users can obtain the single-cell sequencing dataset of interest by clicking the 'Download' button.

Cell types in scRNA-seq data are identified by cell clustering with known marker genes. The 'sMarker' module gathers marker genes of different species in nine systems. On the 'sMarker' page, users query and download marker gene tables by clicking 'Search cell markers by species' or 'Browse cell markers by tissue'. When users choose the classification/species/tissue/rds options in the drop-down box of 'Search cell markers by species' and press the 'Submit' button (Figure 6E), a new page appears with the species image and the basic information, as well as a link 'View Marker Gene Table' in red font. By clicking this link, users are directed to a new tab page with the table of marker genes. Users can search for any marker gene and download marker gene tables (Figure 6E). The entire datasets with marker gene tables are accessible via the link 'View cell atlas in ShinyCell'. Alternatively, users can obtain the marker gene by choosing the tissue or organ image of interest under the 'Browse cell markers by tissue' item (Figure 6E). Once the tissue or organ image (e.g. 'Brain') is clicked, users will see a dataset table with the links of Rds files on a new page. Users can click the link to access the marker gene table.

SUMMARY AND FUTURE PERSPECTIVES

With rapid development of scRNA-seq technology, the number of scRNA-seq datasets is growing. The valuable data are systematically summarized and presented in a freely and publicly accessible database for global researchers to optimize the utility, explore scRNA-seq datasets, and facilitate scientific research, especially for scientists who lack bioinformatics experiences. Here, we collected publicly available single-cell sequencing datasets to create the freely accessible website SPEED, which enables researchers to easily interpret high-quality data resources. SPEED is a powerful tool to deeply mine and define the heterogeneity among cells, tissues, and species. We sorted relevant scRNA-seq datasets to establish the 'Evo', 'Devo', and 'Diz' modules for the convenience of researchers. Seven function modules were built in SPEED to conveniently perform personalized analysis and mining of these scRNA-seq datasets.

SPEED, with its dynamic integrations, will increase in value and importance as more large-scale scRNA-seq studies are performed and the amount of scRNA-seq data is growing exponentially. SPEED pays a special attention to multidimensional cell information from genomics, DNA methylation, chromatin accessibility sequencing, multi-omics (e.g. metabolome and proteome), and trans-omics at single-cell resolution. Data generated by spatial transcriptomics will be integrated into SPEED to interpret the original location of cells in the tissue. The feedbacks and suggestions from researchers will improve the user experience of SPEED.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

The work was supported by Natural Science Foundation of Jiangsu Province (Grants No BK20220279), the CAMS Innovation Fund for Medical Sciences (CIFMS) (2021-I2M-1-061), the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (2022-RC416-01), Gusu Innovation and Entrepreneurship Leading Talents Program (ZXL2022475) and the Shenzhen Science and Technology Program (JCYJ20210324125413036).

Conflict of interest statement. None declared.

REFERENCES

1. Tata,P.R. and Rajagopal,J. (2017) Plasticity in the lung: making and breaking cell identity. *Development*, **144**, 755–766.
2. Zhang,L., Zhu,J., Wang,H., Xia,J., Liu,P., Chen,F., Jiang,H., Miao,Q., Wu,W., Zhang,L. *et al.* (2021) A high-resolution cell atlas of the domestic pig lung and an online platform for exploring lung single-cell data. *J Genet Genomics*, **48**, 411–425.
3. Chan,M.M., Smith,Z.D., Grosswendt,S., Kretzmer,H., Norman,T.M., Adamson,B., Jost,M., Quinn,J.J., Yang,D., Jones,M.G. *et al.* (2019) Molecular recording of mammalian embryogenesis. *Nature*, **570**, 77–82.
4. Chen,D., Jiang,S., Ma,X. and Li,F. (2017) TFBSbank: a platform to dissect the big data of protein-DNA interaction in human and model species. *Nucleic Acids Res.*, **45**, D151–D157.
5. Ma,P., Liu,X., Xu,Z., Liu,H., Ding,X., Huang,Z., Shi,C., Liang,L., Xu,L., Li,X. *et al.* (2022) Joint profiling of gene expression and chromatin accessibility during amphioxus development at single-cell resolution. *Cell Rep.*, **39**, 110979.
6. Hollman,A.L., Tchounwou,P.B. and Huang,H.C. (2016) The association between gene-environment interactions and diseases involving the human GST superfamily with SNP variants. *Int. J. Environ. Res. Public Health*, **13**, 379.
7. Cohen,M., Giladi,A., Gorki,A.D., Solodkin,D.G., Zada,M., Hladik,A., Miklosi,A., Salame,T.M., Halpern,K.B., David,E. *et al.* (2018) Lung single-cell signaling interaction map reveals basophil role in macrophage imprinting. *Cell*, **175**, 1031–1044.
8. 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. *Am. J. Respir. Crit. Care Med.*, **199**, 1517–1536.

9. Tang,F., Barbacioru,C., Wang,Y., Nordman,E., Lee,C., Xu,N., Wang,X., Bodeau,J., Tuch,B.B., Siddiqui,A. *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, **6**, 377–382.
10. Farber,D.L. and Sims,P.A. (2019) Dissecting lung development and fibrosis at single-cell resolution. *Genome Med*, **11**, 33.
11. Paul,F., Arkin,Y., Giladi,A., Jaitin,D.A., Kenigsberg,E., Keren-Shaul,H., Winter,D., Lara-Astiaso,D., Gury,M., Weiner,A. *et al.* (2015) Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, **163**, 1663–1677.
12. Dong,J., Hu,Y., Fan,X., Wu,X., Mao,Y., Hu,B., Guo,H., Wen,L. and Tang,F. (2018) Single-cell RNA-seq analysis unveils a prevalent epithelial/mesenchymal hybrid state during mouse organogenesis. *Genome Biol.*, **19**, 31.
13. Regev,A., Teichmann,S.A., Lander,E.S., Amit,I., Benoist,C., Birney,E., Bodenmiller,B., Campbell,P., Carninci,P., Clatworthy,M. *et al.* (2017) The human cell atlas. *Elife*, **6**, e27041.
14. Chen,D., Sun,J., Zhu,J., Ding,X., Lan,T., Wang,X., Wu,W., Ou,Z., Zhu,L., Ding,P. *et al.* (2021) Single cell atlas for 11 non-model mammals, reptiles and birds. *Nat. Commun.*, **12**, 7083.
15. Li,Z., Sun,C., Wang,F., Wang,X., Zhu,J., Luo,L., Ding,X., Zhang,Y., Ding,P., Wang,H. *et al.* (2022) Molecular mechanisms governing circulating immune cell heterogeneity across different species revealed by single-cell sequencing. *Clin. Transl. Med.*, **12**, e689.
16. Angelidis,I., Simon,L.M., Fernandez,I.E., Strunz,M., Mayr,C.H., Greiffo,F.R., Tsitsiridis,G., Ansari,M., Graf,E., Strom,T.M. *et al.* (2019) An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nat. Commun.*, **10**, 963.
17. Lee,J.H., Tammela,T., Hofree,M., Choi,J., Marjanovic,N.D., Han,S., Canner,D., Wu,K., Paschini,M., Bhang,D.H. *et al.* (2017) Anatomically and functionally distinct lung mesenchymal populations marked by lgr5 and lgr6. *Cell*, **170**, 1149–1163.
18. Zhu,J., Chen,F., Luo,L., Wu,W., Dai,J., Zhong,J., Lin,X., Chai,C., Ding,P., Liang,L. *et al.* (2021) Single-cell atlas of domestic pig cerebral cortex and hypothalamus. *Science Bulletin*, **66**, 1448–1461.
19. 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.
20. 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.
21. Zhong,J., Tang,G., Zhu,J., Wu,W., Li,G., Lin,X., Liang,L., Chai,C., Zeng,Y., Wang,F. *et al.* (2021) Single-cell brain atlas of parkinson's disease mouse model. *J Genet Genomics*, **48**, 277–288.
22. 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.
23. Athar,A., Fullgrabe,A., George,N., Iqbal,H., Huerta,L., Ali,A., Snow,C., Fonseca,N.A., Petryszak,R., Papatheodorou,I. *et al.* (2019) ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res.*, **47**, D711–D715.
24. Papatheodorou,I., Moreno,P., Manning,J., Fuentes,A.M., George,N., Fexova,S., Fonseca,N.A., Fullgrabe,A., Green,M., Huang,N. *et al.* (2020) Expression atlas update: from tissues to single cells. *Nucleic Acids Res.*, **48**, D77–D83.
25. Speir,M.L., Bhaduri,A., Markov,N.S., Moreno,P., Nowakowski,T.J., Papatheodorou,I., Pollen,A.A., Raney,B.J., Seninge,L., Kent,W.J. *et al.* (2021) UCSC cell browser: visualize your single-cell data. *Bioinformatics*, **37**, 4578–4580.
26. Ardini-Poleske,M.E., Clark,R.F., Ansong,C., Carson,J.P., Corley,R.A., Deutsch,G.H., Hagood,J.S., Kaminski,N., Mariani,T.J., Potter,S.S. *et al.* (2017) LungMAP: the molecular atlas of lung development program. *Am. J. Physiol. Lung Cell. Mol. Physiol.*, **313**, L733–L740.
27. Yuan,H., Yan,M., Zhang,G., Liu,W., Deng,C., Liao,G., Xu,L., Luo,T., Yan,H., Long,Z. *et al.* (2019) CancerSEA: a cancer single-cell state atlas. *Nucleic Acids Res.*, **47**, D900–D908.
28. Dai,Y., Hu,R., Manuel,A.M., Liu,A., Jia,P. and Zhao,Z. (2021) CSEA-DB: an omnibus for human complex trait and cell type associations. *Nucleic Acids Res.*, **49**, D862–D870.
29. Zhao,T., Lyu,S., Lu,G., Juan,L., Zeng,X., Wei,Z., Hao,J. and Peng,J. (2021) SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Res.*, **49**, D1413–D1419.
30. Sun,D., Wang,J., Han,Y., Dong,X., Ge,J., Zheng,R., Shi,X., Wang,B., Li,Z., Ren,P. *et al.* (2021) TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. *Nucleic Acids Res.*, **49**, D1420–D1430.
31. Zhang,X., Lan,Y., Xu,J., Quan,F., Zhao,E., Deng,C., Luo,T., Xu,L., Liao,G., Yan,M. *et al.* (2019) CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.*, **47**, D721–D728.
32. Chen,D., Tan,C., Ding,P., Luo,L., Zhu,J., Jiang,X., Ou,Z., Ding,X., Lan,T., Zhu,Y. *et al.* (2022) VThunter: a database for single-cell screening of virus target cells in the animal kingdom. *Nucleic Acids Res.*, **50**, D934–D942.
33. Han,X., Zhou,Z., Fei,L., Sun,H., Wang,R., Chen,Y., Chen,H., Wang,J., Tang,H., Ge,W. *et al.* (2020) Construction of a human cell landscape at single-cell level. *Nature*, **581**, 303–309.
34. 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.
35. Cagan,A., Baez-Ortega,A., Brzozowska,N., Abascal,F., Coorens,T.H.H., Sanders,M.A., Lawson,A.R.J., Harvey,L.M.R., Bhosle,S., Jones,D. *et al.* (2022) Somatic mutation rates scale with lifespan across mammals. *Nature*, **604**, 517-524.
36. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck,W.M., 3rd, Zheng,S., Butler,A., Lee,M.J., Wilk,A.J., Darby,C., Zager,M. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.
37. Ouyang,J.F., Kamaraj,U.S., Cao,E.Y. and Rackham,O.J.L. (2021) ShinyCell: Simple and sharable visualisation of single-cell gene expression data. *Bioinformatics*, **39**, 3374–3376.
38. Jin,S., Guerrero-Juarez,C.F., Zhang,L., Chang,I., Ramos,R., Kuan,C.H., Myung,P., Plikus,M.V. and Nie,Q. (2021) Inference and analysis of cell-cell communication using CellChat. *Nat Commun*, **12**,1088.
39. Huynh-Thu,V.A., Irrthum,A., Wehenkel,L. and Geurts,P. (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **5**,e12776.
40. Wu,T., Hu,E., Xu,S., Chen,M., Guo,P., Dai,Z., Feng,T., Zhou,L., Tang,W., Zhan,L. *et al.* (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)*, **2**,100141.
41. Khrameeva,E., Kurochkin,I., Han, D., Guijarro,P., Kanton,S., Santel,M., Qian,Z., Rong,S., Mazin,P., Sabirov,M. *et al.* (2020) Single-cell-resolution transcriptome map of human, chimpanzee, bonobo, and macaque brains. *Genome Res.*, **30**, 776–789.
42. Geirsdottir,L., David,E., Keren-Shaul,H., Weiner,A., Bohlen,S.C., Neuber,J., Balic,A., Giladi,A., Sheban,F., Dutertre,C.-A. *et al.* (2019) Cross-Species Single-Cell Analysis Reveals Divergence of the Primate Microglia Program. *Cell*, **179**, 1609–1622.
43. Xu,Y., Zhang,J., Hu,Y., Li,X., Sun,L., Peng,Y., Sun,Y., Liu,B., Bian,Z. and Rong,Z. (2021) Single-cell transcriptome analysis reveals the dynamics of human immune cells during early fetal skin development. *Cell Reports*, **36**,109524.
44. Kamath,T., Abdulraouf,A., Burris,S.J., Langlieb,J., Gazestani,V., Nadaf,N.M., Balderrama,K., Vanderburg,C. and Macosko,E.Z. (2022) Single-cell genomic profiling of human dopamine neurons identifies a population that selectively degenerates in Parkinson's disease. *Nat Neurosci*, **25**, 588–595.
45. Bondoc,A., Glaser,K., Jin,K., Lake,C., Cairo,S., Geller,J., Tiao,G. and Aronow,B. (2021) Identification of distinct tumor cell populations and key genetic mechanisms through single cell sequencing in hepatoblastoma. *Commun Biol*, **4**, 1–14.
46. Speranza,E., Williamson,B.N., Feldmann,F., Sturdevant,G.L., Pérez-Pérez,L., Meade-White,K., Smith,B.J., Lovaglio,J., Martens,C., Munster,V.J. *et al.* (2021) Single-cell RNA sequencing reveals SARS-CoV-2 infection dynamics in lungs of African green monkeys. *Sci Transl Med*, **13**,eabe8146.