Nucleic Acids Research, 2026, **54**, D1251–D1269
https://doi.org/10.1093/nar/gkaf1137
Advance access publication date: 18 November 2025
Database issue

scGeneBank: cross-species screening of functional gene sets at single-cell resolution

Siying Qin[1,†], Yuexi Yang[1,†], Lingmin Peng[1,†], Linlin Xu[1,†], Yushan Deng[1], Haoyu Wang[1], Xiwen Qiu[1,2], Yuanqing He[1], Xiangdong Wang[3,*], Yuting Ma[4,*], Li Huang[1,*], Dongsheng Chen[1,*]

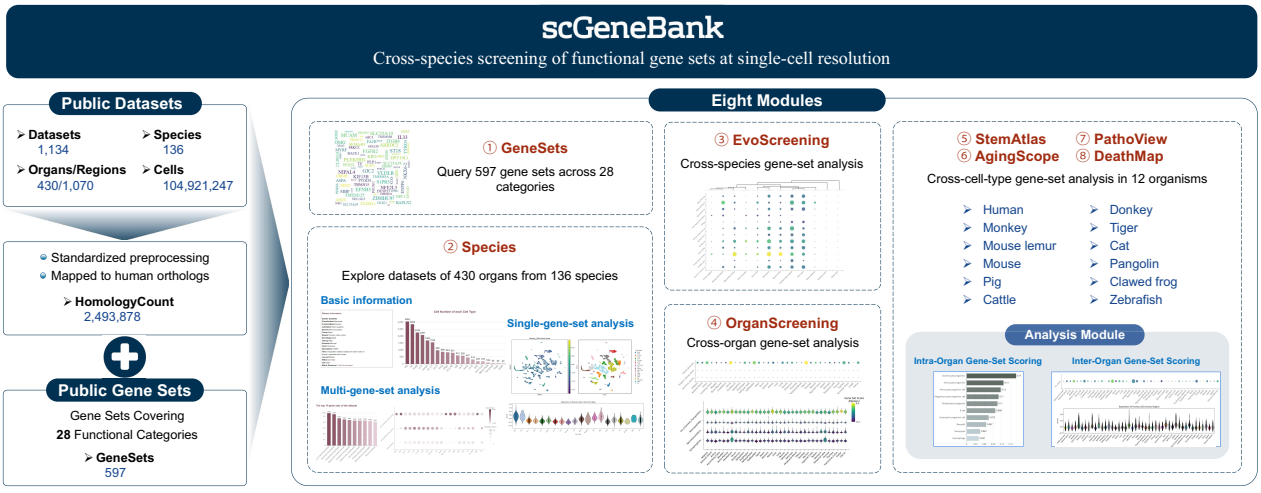
[1]State Key Laboratory of Common Mechanism Research for Major Diseases, Suzhou Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, Suzhou 215123, Jiangsu, China
[2]Department of Obstetrics, The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital, Gusu School of Nanjing Medical University, Suzhou 215000, Jiangsu, China
[3]Zhongshan Hospital, Department of Pulmonary and Critical Care Medicine, Institute for Clinical Science, Shanghai Institute of Clinical Bioinformatics, Shanghai 200032, China
[4]National Key Laboratory of Immunity and Inflammation, Suzhou Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, Suzhou 215123, Jiangsu, China

*To whom correspondence should be addressed. Email: cds@ism.pumc.edu.cn
Correspondence may also be addressed to Li Huang. Email: hl@ism.cams.cn
Correspondence may also be addressed to Yuting Ma. Email: yuting_ma1984@163.com
Correspondence may also be addressed to Xiangdong Wang. Email: xdwang@fuccb.com
†The first four authors should be regarded as Joint First Authors.

Abstract

Single-cell transcriptomics is reshaping the study of cellular physiology and pathology by revealing fine-grained cell functions and states. Interpreting these signals across cell types, organs, and species benefits from curated gene sets and tools for comparative analysis. Here, we introduce scGeneBank (http://8.142.154.29/scGeneBank or http://www.genebank.info), a database and analysis portal for gene set-driven exploration at single-cell resolution. The resource integrates 1134 carefully selected single-cell transcriptome datasets, totaling 104.9 million cells from 136 species and covering 430 organs and 1070 anatomical areas. It supports interactive analysis and visualization of 597 expert-curated gene sets organized into 28 biologically meaningful categories. Eight linked modules address common use cases: GeneSets (curated gene-set navigation), Species (dataset-level exploration), EvoScreening (cross-species comparisons within organs), OrganScreening (gene-set activity analysis across 35 human organs at cell-type resolution), StemAtlas (stemness profiling across species and organs), AgingScope (aging-related pathway analysis), PathoView (disease-associated gene expression analysis), and DeathMap (cell-death pathway activity visualization). Together, these modules provide a unified framework for comparing gene-set activity across species, organs, and cell types, enabling systematic studies of development, evolution, aging, disease, and cell death.

Graphical abstract



Received: August 11, 2025. **Revised:** September 12, 2025. **Accepted:** September 29, 2025

Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized our ability to resolve heterogeneity and characterize gene expression dynamics at unprecedented resolution. It has deepened our understanding of cellular behaviors [1–6], tissue architecture and complexity [1, 6–9], individual development [4, 5, 10–12], disease [3, 13–18], and evolution [4, 6, 19–23]. However, despite substantial efforts to collect and organize single-cell datasets, platforms that enable functional interpretation of gene lists remain limited. Gene sets—collections of genes grouped by shared biological functions, regulatory mechanisms, or disease associations—are widely used to interpret omics data [24]. Gene set analysis (GSA) is critical for revealing heterogeneity among cell populations: by aggregating signals from biologically related genes, GSA reduces the stochastic noise inherent to single-cell data and increases power to detect coordinated functional programs that may be missed at the single-gene level [25–28]. It also facilitates cross-cell-type and cross-condition comparisons while linking observed expression patterns to curated biological pathways [29]. Systematic evaluation of gene-set activity across diverse biological contexts is therefore critical for deciphering complex biological processes.

Several databases have been developed to support downstream analysis of transcriptomic data. For example, CancerSEA [30] and GSCA [31] primarily target cancer-related signatures in humans. A growing number of databases now integrate single-cell datasets. Single Cell Portal [32] provides interactive visualization of gene expression and, for some datasets, pathway enrichment tools, making it one of the most widely used platforms for data hosting and exploration. Other databases, such as scMMO-atlas [33], OrganoidDB [34], and TEDD [35], have expanded to include multi-organ and multi-omics single-cell data integration. DRscDB [36] supports cell-type-specific gene exploration in model organisms, while SPathDB [37] provides specialized pathway annotations. EDomics [38] collects datasets from 12 species but lacks gene-set-level scoring. Although these platforms have greatly accelerated the interpretation of results from fundamental and clinical research, few existing resources support large-scale, gene set-based functional analysis at single-cell resolution.

To meet this need, we developed scGeneBank, a comprehensive platform for profiling functional gene-set activity at single-cell resolution across species, organs, and cell types. The current release integrates 1134 scRNA-seq datasets, covering 104.9 million cells in 136 species, spanning 430 organs and 1070 anatomical regions. scGeneBank compiles 597 expert-curated gene sets organized into 28 biologically meaningful categories, including key life processes such as stemness, aging, disease, and cell death. All data sets are processed with a unified, standardized pipeline to compute gene-set activity, ensuring comparability and reproducibility. Users can explore the resource through eight interactive modules: GeneSets, Species, EvoScreening, OrganScreening, StemAtlas, AgingScope, PathoView, and DeathMap. Together, these modules establish scGeneBank as a versatile, powerful resource for investigating gene-set activity across development, evolution, and disease.

Materials and methods

Single-cell transcriptomics data collection and curation

We systematically searched PubMed [39] using keywords such as "scRNA-seq," "single-cell transcriptomic sequencing," "whole-organism single-cell atlas," "cross-species," and "evolutionary transcriptomics" to identify relevant single-cell transcriptomic datasets. Only publicly accessible datasets were included. Eligibility criteria required the availability of raw count matrices with cell annotations, or data that could be converted into expression matrices. Studies limited to cell lines, bulk sequencing, or lacking cell-type annotations were excluded. In addition, we prioritized datasets from non-human and non-mouse species, as these organisms are underrepresented in existing databases yet are essential for cross-species comparative analyses. In addition, we collected cross-species and whole-organism datasets from large-scale resources, including CZ CELLxGENE Discover [40], Allen Brain Atlas [41], Tabula Sapiens [42], SPEED [43], SCAN [44], and VThunter [45]. Together, we compiled a comprehensive single-cell transcriptomic compendium spanning 136 species, 430 organs, and 1070 anatomical regions. Twelve representative species were selected for cross-species screening of gene set activity in EvoScreening module: human (*Homo sapiens*) [6], crab-eating macaque (*Macaca fascicularis*) [46], mouse lemur (*Microcebus murinus*) [47], mouse (*Mus musculus*) [48], pig (*Sus scrofa*) [49], cattle (*Bos taurus*) [50], donkey (*Equus asinus*) [51], tiger (*Panthera tigris altaica*) [52], domestic cat (*Felis catus*) [52], pangolin (*Manis javanica*) [52], African clawed frog (*Xenopus laevis*) [53], and zebrafish (*Danio rerio*) [54].

For metadata management, we retrieved author-provided information from supplementary tables, GEO records, and related sources, and matched these to the processed expression matrices through cell barcode alignment. To standardize meta across studies, we added two unified fields:

i. "Tissue": a harmonized tissue label summarizing fine-grained anatomical regions into higher-level categories (e.g. cortex and white matter & primary motor cortex → brain).
ii. "LatinName": the standardized species name based on NCBI taxonomy.

Finally, for whole-organism atlases, we harmonized cell-type annotations while retaining the integrity of the original labels. Synonymous or ambiguous terms [e.g. "B cell" and "B cell (Plasmocyte)"] were merged into unified categories, and inconsistent formats (e.g. "stromal cell" and "Stromal_cell") were standardized. Very small subpopulations (<10 cells) were removed to ensure consistency and comparability across organs and species.

Gene-set collection and classification

We organized key biological themes into 28 major categories, comprising 597 curated gene sets in total (Table 1). Categories such as Stem, Death (cell death), CellType, Development, BrainRegion, NeuralGenes, Immune, and CellCycle include fundamental genes regulating cell fate, lineage specificity, and organ-level organization. These sets are critical for deciphering cellular heterogeneity and reconstructing developmental trajectories [55–58]. Aging, Longevity, StressResponse, CircadianGene, SleepDisorders, HypoxiaGenes, and Pain capture organismal adaptation to intrinsic and extrinsic stress, and are highly relevant to age-associated diseases, metabolic dysregulation, and mechanisms of homeostatic regulation [59–64]. We also included human pathologies- and therapy-related categories such as Diz (disease), Obesity, and DrugTarget, which enable systematic identification of disease biomarkers and potential drug targets [65–68]. Gene sets including TFsig,

Table 1. Summary of the 28 curated gene-set categories in scGeneBank

ID	Gene set category	Description	Gene set count
A001	Stem	Stemness-related gene sets	29
A002	Aging	Aging-related gene sets	12
A003	Diz	Disease-related gene sets	35
A004	Death	Cell death pathway-related gene sets	26
A005	Pathway	Signaling pathway-related gene sets	14
A006	Metabolism	Metabolism-related gene sets	86
A007	DrugTargets	Drug targets-related gene sets	37
A008	VirusReceptors	Virus receptor gene sets	94
A009	Development	Development-related gene sets	15
A010	BrainRegion	Brain region-related gene sets	6
A011	Immune	Immune-related gene sets	14
A012	TFsig	Human transcription factor family gene sets	73
A013	HominoidSpecificGenes	Hominoid lineage-specific gene sets	5
A014	Pain	Pain-related gene sets	5
A015	HypoxiaGenes	Hypoxia-related gene sets	5
A016	SpeciesEvo	Species evolution-related gene sets	5
A017	NeuralGenes	Neural-related gene sets	4
A018	CellType	Cell type marker gene sets	44
A019	UltraConserved	Ultra-conserved gene sets	3
A020	Obesity	Obesity-related gene sets	9
A021	CircadianGene	Circadian rhythm-related gene sets	5
A022	CellCycle	Cell cycle-related gene sets	13
A023	SleepDisorders	Sleep disorder-related gene sets	8
A024	Imprinting	Imprinted gene sets	5
A025	Longevity	Longevity-related gene sets	6
A026	FishGenes	Fish-related gene sets	4
A027	StressResponse	Stress-related gene sets	3
A028	GlycoGene	Glycosylation-related gene sets	32

For each category, the table lists the category ID, name, brief description, and the number of gene sets it contains.

Imprinting, Ultraconserved, HominoidSpecificGenes, SpeciesEvo, and FishGenes highlight the evolutionary and epigenetic dimensions of gene regulation [69, 70], supporting analysis of conserved and species-specific mechanisms that shape gene expression programs and evolutionary adaptation. Finally, Pathway, Metabolism, VirusReceptor, and GlycoGene represent major signaling cascades, metabolic networks [71], host–pathogen interactions [45], and glycosylation-related genes essential for studying cell–environment interactions and infectious disease mechanisms.

We followed a systematic workflow for the collection and curation of gene sets. First, for each thematic category, we performed targeted PubMed searches (e.g. "circadian gene signature," "hypoxia response genes") to identify relevant publications, from which candidate gene sets were extracted from the main text and supplementary tables. Each dataset was manually reviewed to ensure a clear biological context. In addition to literature-based curation, we integrated gene sets from domain resources, including AnimalTFDB [72–74], CircaKB [62], GeneCards [75], and PanglaoDB [76]. For gene sets derived from human studies, gene identifiers (e.g. Ensembl ID and Entrez IDs) were standardized by mapping to official gene symbols using BioMart (v2.56.1) [77].

For gene sets originating from non-human species, all genes were converted to their human gene-symbol equivalents to ensure consistency and enable cross-species comparisons. In cases where datasets provided only gene IDs rather than official symbols, we used species-specific homology tables to map both human and non-human IDs to standardized human symbols. To resolve potential duplication issues arising from many-to-one mappings (e.g. multiple IDs mapping to a single symbol), we summed the corresponding expression values into a single aggregated count.

Finally, all gene sets were manually curated to remove redundancy and resolve inconsistencies, while adopting a unified naming convention. This process ensured that each gene set is biologically well-defined, interpretable, and standardized for direct use in downstream analyses and comparative studies.

Extraction of homologous gene protein sequences

To place gene-set scores on a common scale across species, we constructed ortholog tables linking genes from 135 non-human species to their human counterparts. These ortholog tables are downloadable from the Home and Species modules (Fig. 2C and E). Using these mappings, all non-human gene expression matrices were converted to human-ortholog space before gene-set scoring.

To generate the ortholog tables, we first recorded the reference genome/annotation version used for quantification in each dataset. For datasets aligned to Ensembl-derived genomes, human ortholog information was retrieved via Ensembl BioMart [77], using the full set of 70 711 human genes as the reference query set. For datasets aligned to other references, we obtained genome annotation (GTF) files and protein sequence (FAA) files from NCBI RefSeq [78] or the publications' supplementary materials. For each species, we retained the longest unique protein isoform per gene and inferred pairwise orthology to the human protein-coding gene set (23 820 genes) using OrthoFinder (v2.5.5) [79].

For gene expression matrix conversion, one-to-one orthologs were directly mapped. For many-to-one orthologs, expression values of all corresponding non-human genes were aggregated by summing across rows in the adata.X matrix, with the result assigned to the matched human gene. Finally,

the converted matrix was intersected with the human protein-coding gene set to produce ortholog-converted count matrices.

Data processing and quality control

We obtained single-cell transcriptomic data via download links cited in the source publications. Inputs varied by study: for datasets with raw FASTQ files, we retrieved the latest reference genome from NCBI, built the appropriate reference, and then performed read quantification with Cell Ranger [80] (v7.1.0; 10x Genomics) to generate count matrices; for datasets provided as raw matrices, RDS files, or H5AD files, we extracted the expression matrix and standardized all inputs into a unified H5AD format using SCANPY (v1.9.3) [81] in Python (3.10.13). Metadata for each dataset was manually curated and matched to the processed matrices.

We retained single-cell transcriptomic datasets from 136 species and used the original cell-type annotations provided by the source publications for each dataset; these annotations served as the basis for downstream gene-set scoring. For whole-body single-cell atlases used in our database, including human (*Homo sapiens*) [6], crab-eating macaque (*Macaca fascicularis*) [46], mouse lemur (*Microcebus murinus*) [47], mouse (*Mus musculus*) [48], pig (*Sus scrofa*) [49], cattle (*Bos taurus*) [50], donkey (*Equus asinus*) [51], tiger (*Panthera tigris altaica*) [52], domestic cat (*Felis catus*) [52], pangolin (*Manis javanica*) [52], African clawed frog (*Xenopus laevis*) [53], and zebrafish (*Danio rerio*) [54], we manually curated and standardized cell-type annotations. To improve the robustness of cell-type-specific statistics, we excluded any cell type represented by fewer than 10 cells within a given organ, minimizing the impact of rarely represented cell types on gene-set scoring.

Cells with fewer than 200 or >6000 detected genes were excluded to remove low-quality cells using the function sc.pp.filter.cells from SCANPY (v1.9.3). We employed Scrublet (v0.2.3) [82] with default settings to identify and exclude doublets. Expression matrices were normalized to a total count of 10 000 per cell using sc.pp.normalize_total, followed by log-transformation using sc.pp.log1p.

Dataset integration

For six modules—EvoScreening, OrganScreening, StemAtlas, AgingScope, PathoView, and DeathMap—gene-set scoring was performed at the species level based on integrated multi-organ single-cell transcriptomic datasets. Depending on the integration status reported in the original publication, we adopted different preprocessing strategies. If a study provided a pre-integrated expression matrix or low-dimensional embeddings across organs, we used those directly for downstream gene-set scoring. For datasets lacking such integration, we performed batch correction with Harmony (v0.0.10) [83]. Harmony is efficient and accurate for large-scale single-cell transcriptomic integration [83, 84] and operates on principal component (PC) embeddings, facilitating downstream dimensionality reduction and clustering without extensive parameter tuning. Within each species, we used Tissue or Library as the batch label to account for organ-specific effects. To assess the performance of this integration, we evaluated batch effect removal using scIB [85], a benchmarking framework that evaluates both batch mixing and biological conservation. After Harmony integration, the overall scIB scores improved consistently across multiple levels: cross-tissue, cross-dataset within tissue, and cross-library within dataset. These results demonstrate that Harmony effectively reduces batch effects and improves the comparability and reliability of cross-organ and cross-dataset analyses (see Supplementary Note 1).

Importantly, the integration step described earlier refers only to the preprocessing and harmonization of multi-organ data for visualization and clustering. All subsequent gene-set scoring was conducted exclusively on normalized expression matrices using the scanpy.tl.score_genes function. For cross-species comparisons, we further applied ComBat [86] correction (sva v3.46.0) [87] at the gene-set score matrix level, ensuring comparability while minimizing non-biological variation.

Gene set scoring and expression ratio calculation

We deployed different strategies for scoring at the individual dataset level and cross-dataset analyses. For individual datasets, we used the scanpy.tl.score_genes function from SCANPY (v1.9.3) [81] to compute per-cell activity scores for 597 gene sets. This function estimates the relative enrichment of a gene set for each cell by subtracting the average expression of a randomly selected control gene set from that of the target gene set. This provides a proxy for gene-set-specific function activity at single-cell resolution. To ensure scoring stability and comparability across datasets with varying gene coverage, we scaled the ctrl_size parameter by the number of genes from the target set detected in the dataset: if the overlap was ≥50 genes, we set ctrl_size = overlap; otherwise, we used the default ctrl_size = 50.

For cross-dataset or cross-species comparisons, gene-set scoring was performed independently for each dataset. Then, the resulting activity scores were batch-corrected using ComBat from the sva package (v.3.46.0), with species specified as the batch factor. ComBat applies an empirical Bayes framework to adjust for both mean and variance across batches, thereby reducing non-biological, species-specific effects and improving comparability of gene-set activity across evolutionary lineages.

To quantify gene-set expression within a given organ or cell type, we computed for each cell the fraction of genes in the set that were expressed. Cells with >10% of set genes detected were designated as gene-set-positive. The expression ratio was then defined as the percentage of gene-set-positive cells among all cells in the selected organ or cell type.

Visualization

Across the Species, EvoScreening, OrganScreening, StemAtlas, AgingScope, PathoView, and DeathMap modules, results are shown with interactive bubble plots as well as static violin plots. In these plots, the *y*-axis lists the user-selected gene sets, and the *x*-axis shows either multiple species (for cross-species comparisons) or multiple organs within a single species (for cross-organ comparisons). In bubble plots, color intensity represents the average per-cell gene set activity score, while bubble size reflects the proportion of cells expressing the gene set. In violin plots, color intensity denotes the median activity score across cells.

To facilitate within-species comparisons, the StemAtlas, AgingScope, PathView, and DeathMap modules also provide horizontal bar plots. Here, the *y*-axis lists all annotated cell types in the user-selected organ of the chosen species, and bar length represents the average per-cell score of the selected gene sets for each cell type.

Database Info		scRNA-seq Data Statistics						Gene-Set Analysis Module			
Name	Publication Year	Species	Organs/ Regions	Datasets	Cells	Gene Sets	Homology Count	Gene-Set Scoring	Cross-Cell-Type Analyses	Cross-Organ Analyses	Cross-Species Analyses
scGeneBank	-	136	430/1,070	1,134	104.9 million	597	2.4 million	Yes	Yes	Yes	Yes
CancerSEA	2019	1	27 (Cancer types)	74	93 thousand	14	-	Yes	No	Yes	No
GSCA	2023	1	33 (Cancer types)	4 (Data sources)	Bulk	0 (User-provided)	-	Yes	Yes	Yes	No
SPathDB	2025	2	36	84	1.6 million spots	114,998	0	Yes	Yes	No	No
DRscDB	2021	5	24	90	1 million	0	< 0.5 million	No	No	Yes	Yes
scMMO-atlas	2024	2	27	70	3 million	0	0	No	No	No	No
OrganoidDB	2023	2	46	145	1.5 million	0	0	No	No	No	No
TEDD	2023	4	103	177	3.8 million	0	0	No	No	No	No
EDomics	2023	12	-	-	0.5 million	0	0	No	No	No	No

Figure 1. Comparative overview of scGeneBank and existing gene set-related databases by content and functionality. scGeneBank integrates 1134 curated single-cell transcriptomic datasets, encompassing 104.9 million cells from 136 species, 430 organs, and 1070 regions. It includes 597 manually curated gene sets and supports ortholog mapping comprising 2.4 million gene pairs. In comparison to other resources such as CancerSEA [30], GSCA [31], SPathDB [37], DRscDB [36], scMMO-atlas [33], OrganoidDB [34], TEDD [35], and EDomics [38], scGeneBank is the first and only platform that enables large-scale gene set scoring with standardized processing pipelines and supports cross-cell-type, cross-organ, and cross-species functional analyses with unified visualization tools.

Results

Overview of the scGeneBank database

scGeneBank applies curated functional gene sets at scale to single-cell transcriptomic data. We curated 597 gene sets organized into 28 categories and scored them across 1134 single-cell transcriptomic datasets comprising 104 921 247 cells from 136 species, spanning 430 organs and 1070 regions. The resource encompasses major evolutionary lineages, including vertebrates (mammals, birds, reptiles, amphibians, and fishes) and invertebrates (e.g. arthropoda). The wide coverage of species enables us to examine conserved and divergent biological processes across the tree of life. Gene set activity scores were computed across all datasets using a standardized method. To enable cross-dataset and cross-species comparison, genes from non-human species were mapped to their human orthologs. A total of 2 493 878 orthologous gene relationships are available to date (freely accessible at http://8.142.154.29/scGeneBank/HomoGene.php).

Compared with existing resources, scGeneBank uniquely enables functional gene set profiling across species, organs, and cell types at single-cell resolution (Fig. 1). Our website features eight core interactive modules—GeneSets, Species, EvoScreening, OrganScreening, StemAtlas, AgingScope, PathoView, and DeathMap—that facilitate exploration of evolutionary patterns, stemness, aging, disease, and cell-death signatures. GeneSets module provides access to all curated gene sets and their classification into 28 biological categories. Species module allows users to browse and search across 136 species by taxonomy and organ, and explore dataset-level gene set activity with associated metadata. EvoScreening module and OrganScreening module enable users to assess gene set activity across species or across human organs, respectively. Results are visualized through bubble plots and violin plots that reflect both activity levels and expression prevalence. StemAtlas, AgingScope, PathoView, and DeathMap focus on distinct functional themes and application scenarios (stemness, aging, disease, and cell death), allowing users to examine gene set activity across organs and cell types in 12 representative species.

scGeneBank provides multiple visualization tools, including Uniform Manifold Approximation and Projection (UMAP) projections, bubble plots, bar charts, and violin plots, to help present results clearly. All outputs, including gene-set activity scores and generated figures, are available for download to facilitate further analysis. Overall, scGeneBank is a user-friendly platform for functional gene set profiling and cross-context analysis.

The Home page displays all gene-set categories and species provided by scGeneBank (Fig. 2A). The Home page gives an overview of our database and enables users to initiate exploration through intuitive visual elements. Clicking on the gene set category images directs users to the GeneSets module (Fig. 2B). Hovering over the species image on the Home page brings up an information card displaying details such as the species ID (SPID), taxonomic classification, common name, Latin name, a download button for the homology table, and dataset statistics related to the selected species (Fig. 2C). Clicking on the images redirects users to the Species module, where they can select an organ of the selected species to view all associated single-cell transcriptomic datasets.

GeneSets module: curated functional gene sets across 28 biological categories

The GeneSets module organizes 597 curated gene sets into 28 biological categories (Table 1), compiled from publicly available studies. Users can enter any category by clicking its image to view all associated gene sets. Selecting a specific gene set displays a detailed table listing each gene symbol, Ensembl ID, and direct links to external resources like

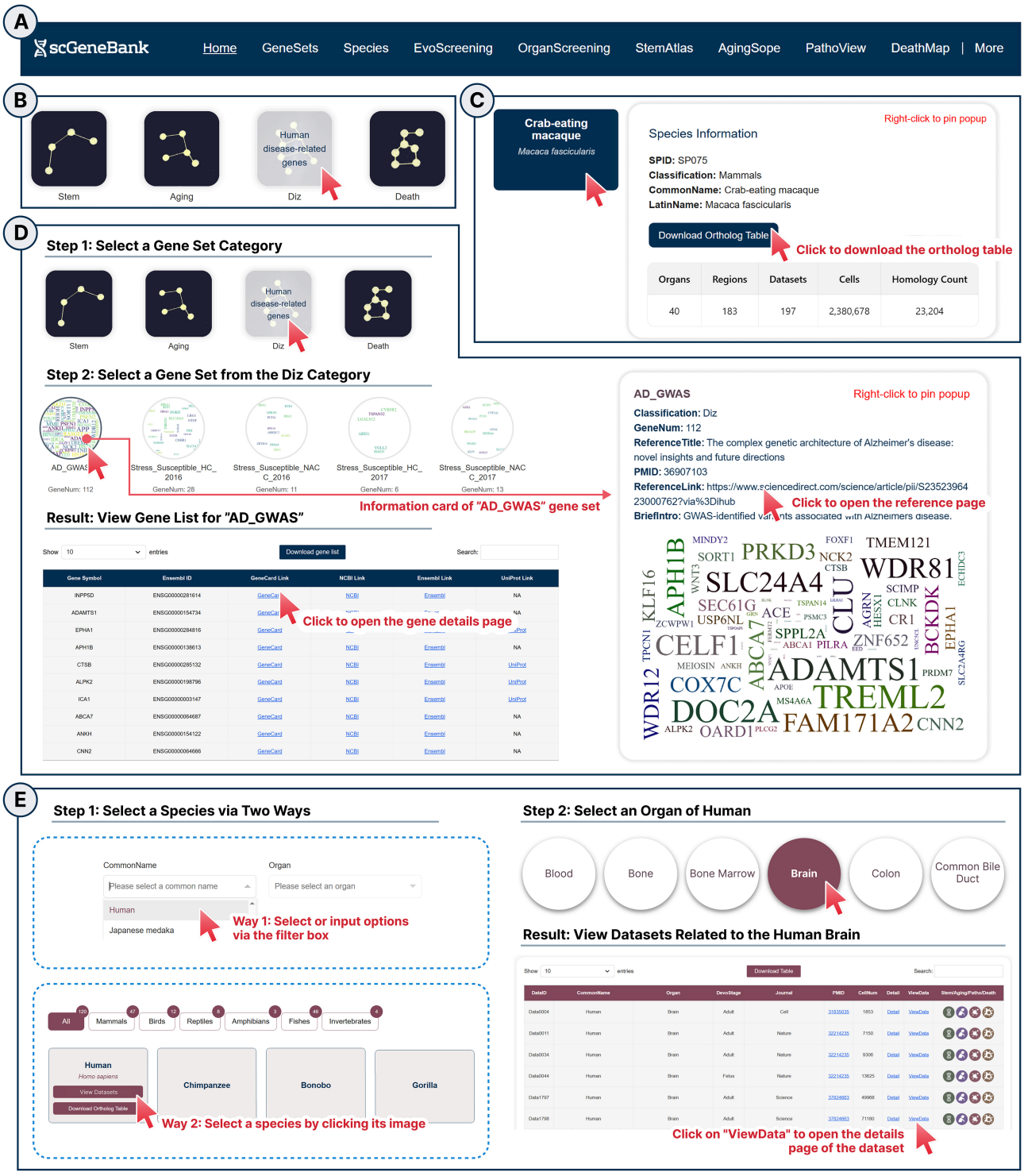


Figure 2. Overview of the Home, GeneSets, and Species modules in scGeneBank. (**A**) Home page displays all gene-set categories and available species. (**B**) Selecting a gene-set category opens the GeneSets module. (**C**) Species information card with taxonomic details, dataset statistics, and ortholog table download option. (**D**) GeneSets module workflow: Step 1, select a category; Step 2, select a gene set; results include a gene list with links to external resources. (**E**) Species module workflow: Step 1, select a species by the filter box or image; Step 2, select an organ; results list associated datasets with metadata and access to detailed dataset pages.

GeneCards [75], NCBI [88], Ensembl [89], and UniProt [90] for further exploration. As an example, choosing "Diz" category (Fig. 2D, Step 1) reveals all gene sets within that category under the Step 1 panel. Users can browse the 35 available gene sets on the Step 2 panel and select one of interest, such as AD_GWAS [91] (Fig. 2D, Step 2). Right-clicking on the AD_GWAS image opens metadata (e.g. source and number of genes) for AD_GWAS and a ReferenceLink to the original publication. Left-clicking the image loads the full table of the 112 genes included in the set. To investigate a particular gene (e.g. *INPP5D*), users can follow the embedded links to GeneCards, NCBI, and Ensembl.

Species module: exploration of single-cell transcriptomics across 136 species

The Species module aggregates 1134 single-cell transcriptomic datasets from 136 species and provides cross-species exploration of gene expression and gene-set activity. After selecting a species and an organ, the relevant datasets are displayed in a table. Each row lists the species' common name, organ, developmental stage, PMID, cell count, and links to detailed dataset pages for interactive exploration.

To explore human-related datasets, users have two entry points. First, enter the common name "human" in the filter box beneath the header and select a human organ (e.g. brain) to begin exploration (Fig. 2E, Step 1, Way 1). Alternatively, left-click the "human" image in the Step 1 panel (Fig. 2E, Step 1, Way 2) and then choose an organ in the Step 2 panel (Fig. 2E, Step 2). Hovering over the "human" image reveals two buttons: "View Datasets" and "Download Ortholog Table." Clicking "View Datasets" displays all available human organs in the Step 2 panel, while "Download Ortholog Table" provides a file of orthologous gene relationships between humans and the selected species. After selecting the human brain, 117 associated datasets and their metadata are displayed in a table (Fig. 2E, Result). For example, Data1798 (PMID: 37824663) [92] contains 71 180 cells from the adult human brain. Detailed metadata, accessible via the "Detail" link, indicate that samples originated from the human cerebrovascular (CBV) region and were obtained from donors aged 29, 42, and 50 years under normal physiological conditions. Single-cell sequencing libraries were constructed using the Chromium v3 platform, and the corresponding expression matrix can be downloaded via the "Click to download" button (Fig. 3A).

A dataset-specific exploration page is available via the "ViewData" button. For example, clicking "ViewData" for Data1798 opens its in-depth exploration page (Fig. 3). Section 1 presents dataset metadata (Fig. 3A). To the right, a ranked bar chart displays the top 10 enriched gene sets (Fig. 3B) based on activity scores calculated with Scanpy (scanpy.tl.scores_genes). The highest-scoring gene sets highlight evolutionary conservation, for example, Intelligence_Evo_Genes (from the SpeciesEvo category), CTF_NFI (from the TFsig category), and Genes_Overlapping_UCEs (UltraConserved). Strong enrichment of development-, stemness-, and neuronal/neuroendocrine-related sets implies the presence of developmentally plastic cell populations in the region. Signals from drug-target sets (e.g. Topiramate, Primidone) and obesity/metabolism-related sets (Obesity-risk genes) suggest potential involvement of CBV-resident cells in neurological and metabolic disease pathways. Section 2 shows clustering results via UMAP, accompanied by bar plots summarizing cell-type distributions and relative abundances. The UMAP and bar plots (Fig. 3C) indicate that Data1798 captures the characteristic cellular composition of the CBV region, with enrichment of vascular and glial populations alongside other supportive cell types.

To explore functional patterns within a selected dataset, users choose a gene-set category and then a specific gene set; Section 3 displays the results (Fig. 3D). For example, selecting the "Diz" category reveals all disease-related gene sets. Choosing AD_GWAS [91], then evaluates its activity pattern in Data1798. Three visualizations are generated: a UMAP plot colored by gene-set activity across cells, an annotated UMAP plot with cell-type labels, and violin plots comparing gene-set scores across clusters. In Data1798, microglia exhibit the highest AD_GWAS scores, consistent with a previous study identifying microglia as a key cell type associated with Alzheimer's disease genetic risk [93].

Section 4 visualizes gene-set activity for a selected category. A bubble plot summarizes the activity of all gene sets within that category, for example, the "Diz" set across cell types in Data1798 (Fig. 3E). Color intensity reflects the average gene-set score, and bubble size indicates prevalence (the fraction of cells in each cell type exhibiting the signal). All visualizations and result tables can be downloaded for further analysis.

EvoScreening module: cross-species profiling of functional gene set activity

EvoScreening enables cross-species profiling of gene-set activity within a selected organ. The module incorporates whole-body atlases from 12 species: human (*Homo sapiens*) [6], crab-eating macaque (*Macaca fascicularis*) [46], mouse lemur (*Microcebus murinus*) [47], mouse (*Mus musculus*) [48], pig (*Sus scrofa*) [49], cattle (*Bos taurus*) [50], donkey (*Equus asinus*) [51], tiger (*Panthera tigris altaica*) [52], domestic cat (*Felis catus*) [52], pangolin (*Manis javanica*) [52], African clawed frog (*Xenopus laevis*) [53], and zebrafish (*Danio rerio*) [54] (Fig. 4A). The distribution and overlap of available organ datasets across these species are summarized in an UpSet plot. The plot highlights both shared and species-specific organ coverage (Fig. 4B). Once an organ and a functional gene category are selected, gene set activity is visualized in a bubble plot. The color intensity represents the gene set score and bubble size indicates the proportion of cells expressing the set within the organ.

To examine pathway-related gene-set activity across liver datasets from multiple species, users could select the liver icon and then choose the "Pathway" category (Fig. 4C). The resulting bubble plot indicates relatively high activity for ERK/MAPK (BioCarta_Erk1/Erk2_signaling and MAPK_Signaling), STING (Ishikawa_Signaling), NF-κB (NFKkapaB_Signaling), and PI3K/Akt (PI3K/Akt_Signaling) [94] pathways in liver, consistent with the organ's homeostatic roles in metabolic sensing and stress/innate immune regulation. ERK/MAPK and PI3K/Akt mediate insulin- and growth factor-driven metabolic and survival signaling [95, 96], whereas the STING–IRF3/NF-κB axis provides immune surveillance against pathogens and cellular stress [97, 98]. By contrast, Wnt/β-catenin and Hedgehog show lower activity in the adult homeostatic liver and are typically induced during pathological states [99–101]. Species-specific differences are also apparent: Murine livers show elevated Erk1/Erk2 and

Figure 3. Functional gene set visualization and analysis interface in the Species module. (**A**) Dataset metadata card showing species, anatomical origin, developmental stage, sequencing platform, and download links [92]. (**B**) Ranked bar chart of the top 10 enriched gene sets by activity score; both the figure and the full ranking table are downloadable. (**C**) UMAP embedding and bar chart summarizing cell-type clusters and their relative abundances. (**D**) Single gene-set analysis workflow: after selecting a category (e.g. Diz) and a set (e.g. AD_GWAS [91]), three views are shown: UMAP colored by gene-set score, cell-type–annotated UMAP, and violin plots of score distributions across clusters. (**E**) Multi-set activity overview: Step 1, select a category; a bubble plot summarizes activity for all gene sets in the category, with color indicating average activity and bubble size indicating prevalence (fraction of cells exhibiting the signal). All visualizations and underlying tables can be downloaded for further analysis.

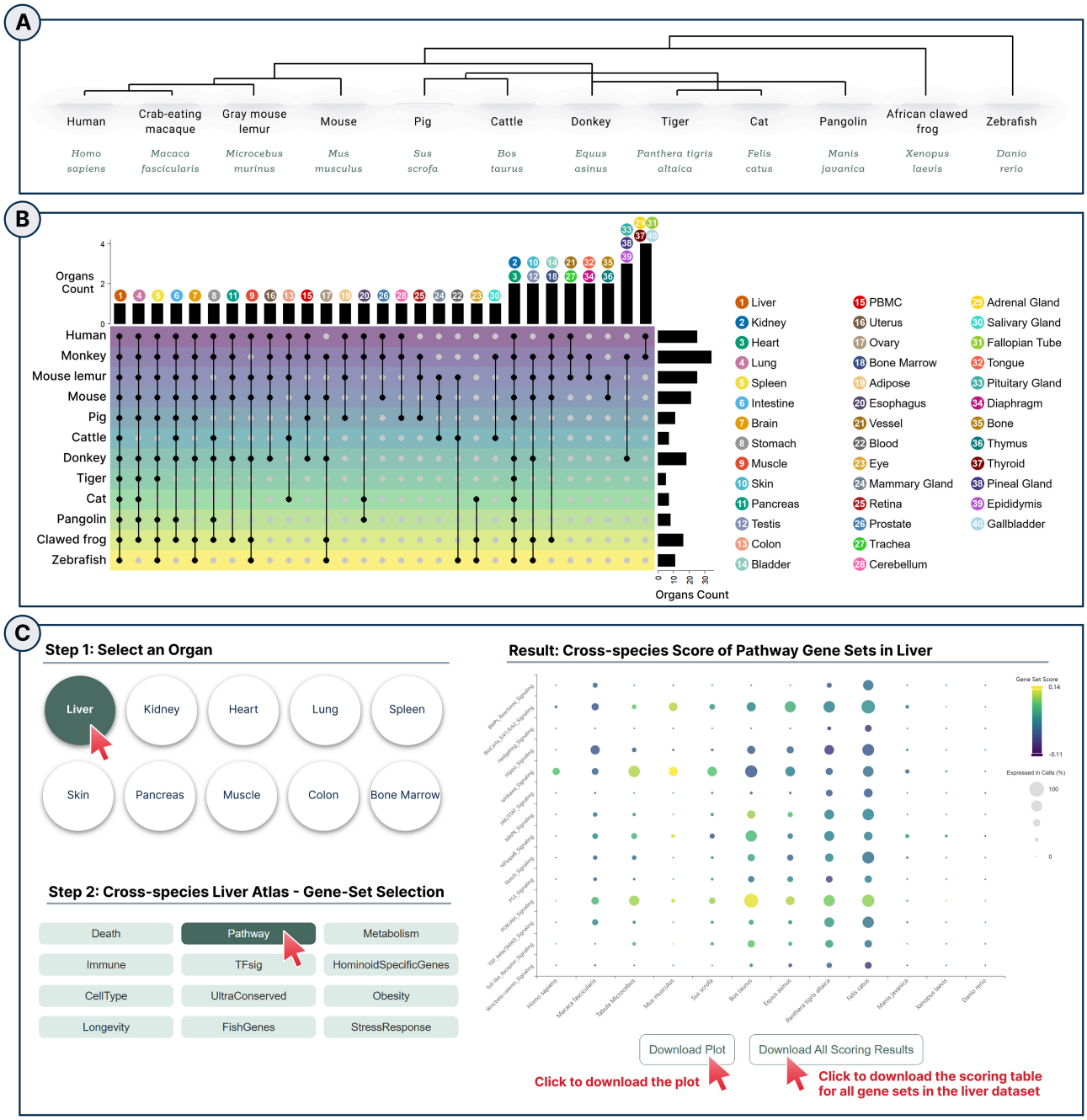


Figure 4. EvoScreening module for cross-species functional GSA. (**A**) Phylogenetic tree of 12 representative species included in the EvoScreening module. (**B**) UpSet plot showing the distribution and overlap of organ datasets across the 12 species. (**C**) Workflow for cross-species analysis: Step 1, select an organ (e.g. liver); Step 2, select a functional gene set category (e.g. Pathway); the resulting bubble plot visualizes gene-set activity across species, where color intensity denotes the average activity score and bubble size indicates the proportion of cells expressing the signal.

Ishikawa_STING activity. In *Bos taurus*, PI3K/Akt activity is particularly pronounced, which may reflect adaptations to ruminant-specific insulin and nutrient metabolism [102, 103].

OrganScreening module: cross-organ profiling of functional gene-set activity

OrganScreening allows users to explore gene set expression across 35 human organs [6] (Fig. 5A). By clicking on the cell type image, users can select a cell type of interest and then choose a functional gene category to perform the results. The result shows gene activity across organs for the selected cell type. It provides insight into organ-specific functional patterns.

For example, users can apply OrganScreening to cross-organ analysis of cell cycle gene activity in human macrophages (Fig. 5B). In the resulting bubble plot, 12 of 13 cell cycle-related gene sets show consistently low scores across organs (Fig. 5C). This suggests that, under normal homeostatic conditions, most organ-resident and circulating macrophages remain quiescent and non-proliferative. By contrast, the "Cancer-Essential_Cell-Cycle_Genes" [57] gene set shows markedly elevated activity, most notably in the fallop-

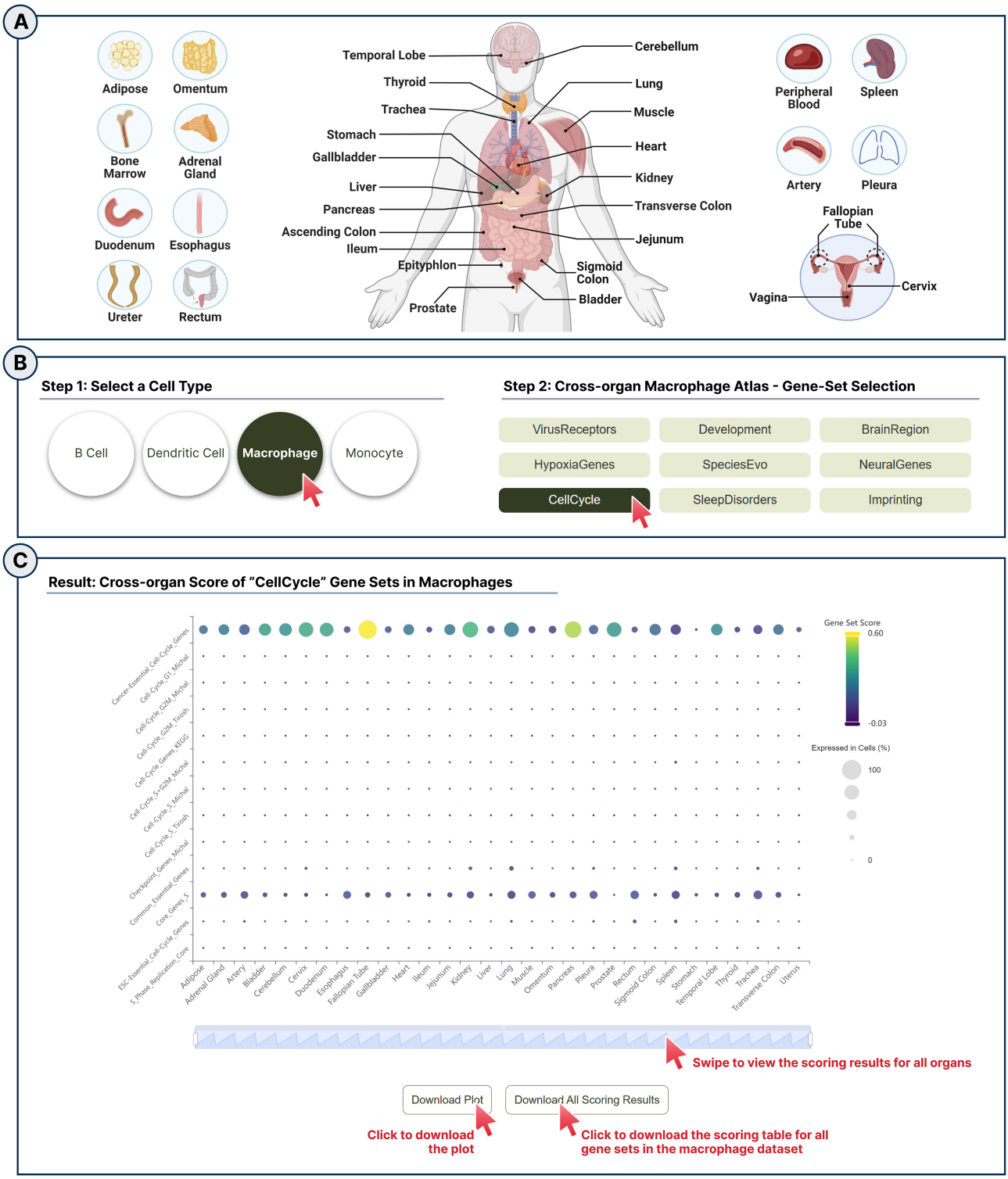


Figure 5. OrganScreening module for cross-organ functional gene-set analysis in humans. (**A**) Overview of 35 human organs available for analysis [6], covering multiple systems (digestive, circulatory, endocrine, and reproductive). (**B**) Analysis workflow: Step 1, select a cell type of interest (e.g. macrophage); Step 2, choose a functional gene-set category for evaluation (e.g. CellCycle). (**C**) Example output showing a bubble plot of CellCycle gene-set activity in macrophages across organs. Color intensity represents the average activity score, and bubble size indicates prevalence (fraction of cells exhibiting the signal). All plots and scoring tables are downloadable for further analysis. Created in BioRender. Qin, S. (2025) https://BioRender.com/0yigvyo.

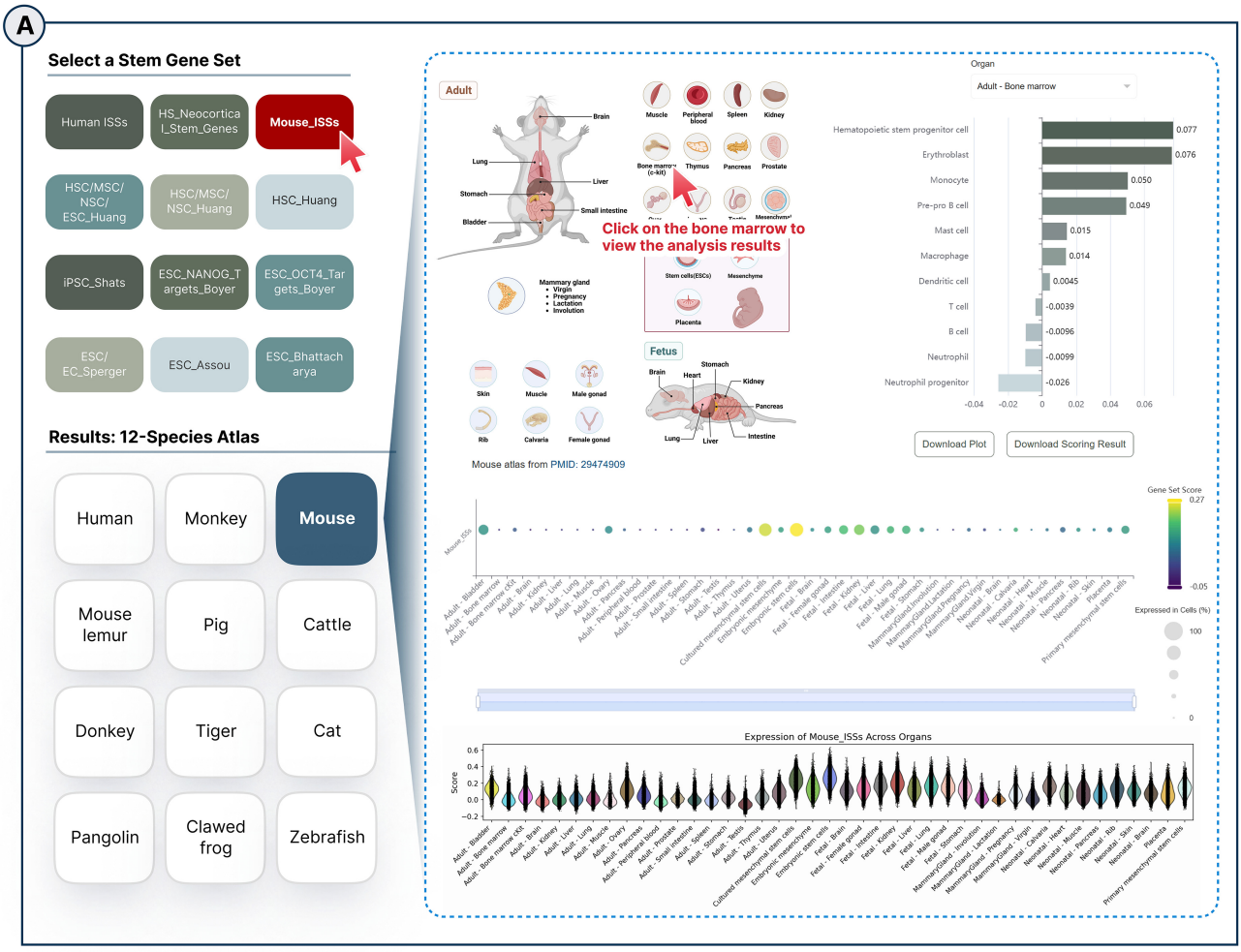


Figure 6. StemAtlas module for cross-organ and cross-species exploration of stemness-related gene sets. (**A**) Example workflow using the "Mouse_ISSs" [105] stemness-related gene set. Users select an organ (bone marrow) within a chosen species (mouse) [48] to view cell-type-resolved activity scores. Results are presented as bar plots of gene set scores across annotated cell types and as bubble plots and violin plots summarizing cross-organ activity. Color intensity of bubble plots reflects the average gene-set score, and bubble size indicates prevalence (fraction of cells exhibiting the signal). All visualizations and result tables are downloadable. (**B**, **C**) Corresponding outputs for human (Homo sapiens) [6] and crab-eating monkey (Macaca fascicularis) [46], respectively. Created in BioRender. Qin, S. (2025) https://BioRender.com/2te2o34.

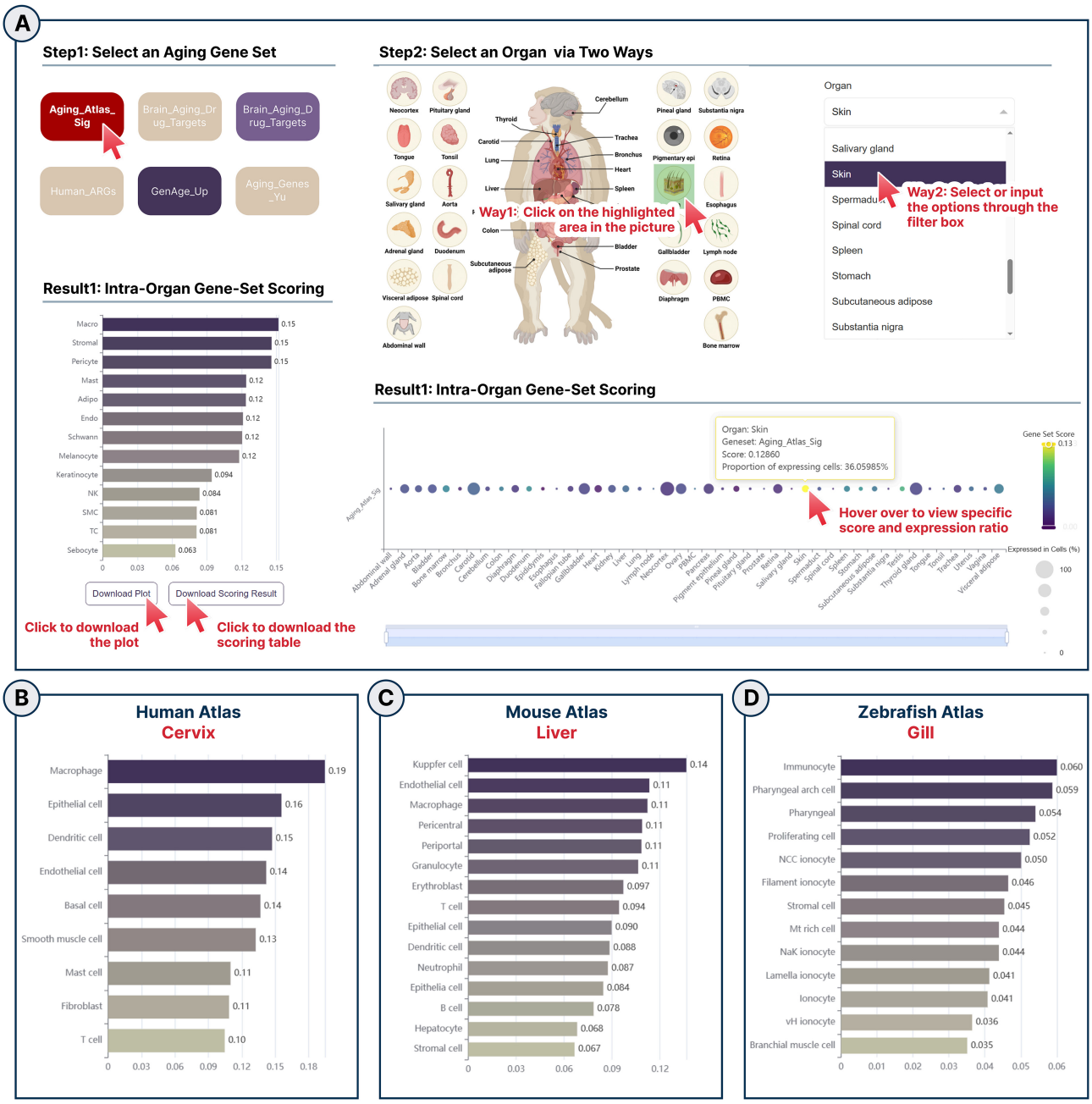


Figure 7. AgingScope module for cross-organ and cross-species analysis of aging-related gene sets. (**A**) Module workflow: users select an aging-related gene set (e.g. Aging_Atlas_Sig [106]) and an organ of interest via direct atlas interaction or filter bar. Results include intra-organ bar plots showing average gene set scores across annotated cell types and cross-organ bubble plots summarizing activity patterns across all organs within the species [46]. (**B**–**D**) Example results for Aging_Atlas_Sig in representative organs from three species. Bar plots display the top-ranked cell types contributing to aging-related gene set activity in each organ: cervix (human) [6], liver (mouse) [48], and gill (zebrafish) [54]. Created in BioRender. Qin, S. (2025) https://BioRender.com/ge9xrlv.

ian tube. This enrichment suggests the presence of a distinct microenvironment within the fallopian tube that may enhance macrophage transcriptional programs associated with proliferation. Previous studies report that the fallopian tube epithelium and nearby immune cells activate proliferative signaling during ovulation and inflammatory episodes [104]. Cancer-essential cell cycle genes are also known to be persistently active in cancers and highly regenerative tissues [57]. Their elevated expression in the fallopian tube may indicate increased proliferative potential or heightened susceptibility to transformation.

StemAtlas module: atlas of stemness-related gene set activity

StemAtlas, AgingScope, PathoView, and DeathMap are four thematic modules that analyze gene-set activity related to stemness, aging, disease, and cell death.

The StemAtlas module profiles stemness-associated gene sets. It features a curated panel of 29 genes implicated in stem cell function and regulation. Users can select a species, choose target organs, and query specific gene sets. For each query, gene-set scores are computed for all annotated cell

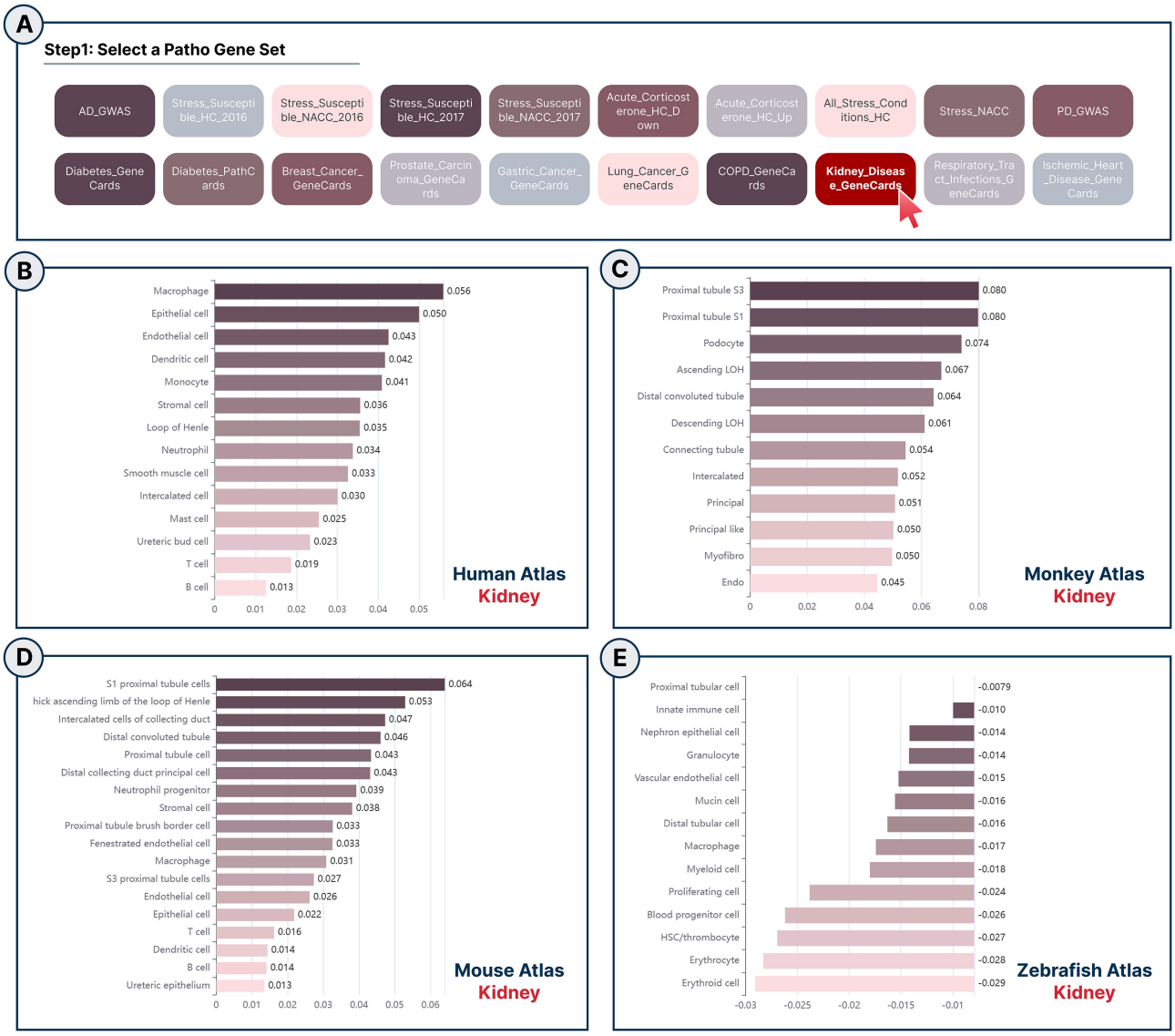


Figure 8. PathoView module for cross-organ and cross-species analysis of disease-related gene sets. (**A**) Module interface for selecting a disease-associated gene set (e.g. Kidney_Disease_GeneCards [113]). Intra-organ bar plots of gene-set scores across annotated kidney cell types for human (**B**) [6], crab-eating macaque (**C**) [46], mouse (**D**) [48], and zebrafish (**E**) [54]. Bars represent average gene-set activity; greater height and color intensity indicate higher activity.

types within the chosen organ and visualized as bar plots to highlight relative activity across cell types. The module also provides cross-organ activity profiles within each species, enabling systematic comparisons of stemness characteristics across tissues.

As an example, to explore the activity of the mouse integrated stemness signatures (*Mouse_ISSs*) [105], click the gene-set icon, navigate to the mouse panel, and either select an organ (e.g. bone marrow) or examine the cross-organ activity panel at the bottom (Fig. 6A). In adult mouse bone marrow, Mouse_ISSs activity is highest in hematopoietic stem cells, erythroblasts, and early B cells—populations that remain undifferentiated or possess high proliferative potential (Fig. 6A). By contrast, mature immune cells such as neutrophils, T cells, and B cells score lower, consistent with downregulation of stemness programs upon terminal differentiation. Similar patterns are observed in other species. In human bone marrow, erythroid progenitors, cord-blood $CD34^+$ cells, and erythroid cells rank highest, indicating relatively active stem cell programs (Fig. 6B). In crab-eating macaques bone marrow, megakaryocyte-erythroid progenitors, late-stage erythroid cells, and pre-B cells show the strongest activity (Fig. 6C). These cross-species trends provide a useful reference for conserved stemness programs.

AgingScope module: atlas of aging-related gene set activity

AgingScope focuses on aging-related transcriptional programs and pathways. The module currently includes 12 curated gene sets, encompassing aging biomarkers from Aging Atlas [106], brain aging drug target genes [107], genes overexpressed or underexpressed with age [108], and signatures associated with cellular senescence [109]. Users can examine gene-set activity across organs and cell types in 12 supported organisms. After selecting a gene set (e.g. Aging_Atlas_Sig [106]), users choose an organ via the filter bar or by clicking the corresponding

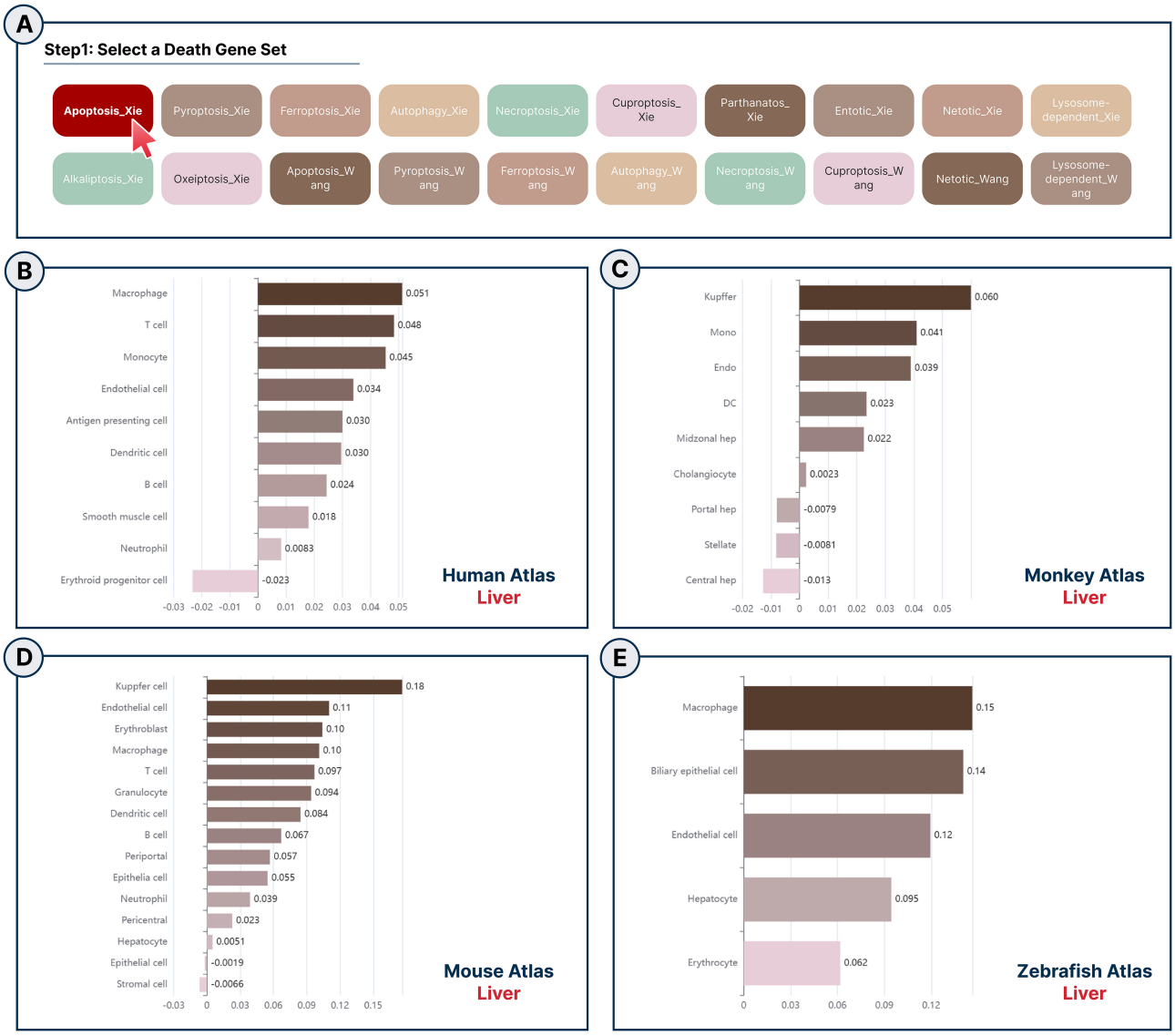


Figure 9. DeathMap module for organ- and cell type–resolved analysis of cell-death–related gene sets across four species. (**A**) Module interface for selecting a cell-death–related gene set (e.g. Apoptosis_Xie [116]). (**B**–**E**) Bar plots showing intra-organ average gene set scores for the liver (human [6], crab-eating macaque [46], mouse [48], and zebrafish [54]).

region in the interactive anatomical diagram (Fig. 7A). Bar plots then visualize gene-set scores results across annotated cell types within the selected organ, revealing organ-specific or conserved aging patterns. A cross-organ panel at the bottom summarizes activity for the chosen gene set across organs within the selected species.

Analysis of *Aging_Atlas_Sig* reveals a consistent trend: aging-related signals accumulate in organs and cell types that function as barriers, metabolic hubs, or immune surveillance sites. In humans [6], the highest activity appears in the cervix, driven by epithelial and immune cells, including macrophages, T cells, and dendritic cells. This aligns with the cervix's role as a mucosal barrier and its exposure to microbe-driven chronic inflammation [110] (Fig. 7B). In the crab-eating macaques [46], the skin shows the highest signal (Fig. 7A), dominated by macrophages, stromal cells, and pericytes, aligning with hallmarks of skin aging, such as matrix remodeling, immune activation, and chronic inflammation [111]. In mice [48], the liver ranks highest, with Kupffer cells and liver endothelial cells contributing most of the activity (Fig. 7C). This supports the previous study that hepatic aging starts with vascular and immune interface dysfunction, followed by pro-inflammatory activation [112]. In zebrafish [54], the gills lead the ranking, enriched in immune cells and multiple ionocyte subtypes (Fig. 7D), reflecting their dual roles in barrier immunity and ion balance. Together, these results point to a conserved cross-species principle: aging signatures tend to concentrate in barrier-immune-metabolic niches.

PathoView module: atlas of disease-related gene set activity

PathoView focuses on 35 disease-associated gene sets, covering conditions such as Alzheimer's disease, Parkinson's disease, diabetes, breast cancer, and stress susceptibility. It enables exploration of disease-relevant signatures across the whole-body single-cell atlas of 12 representative species. By selecting a gene set (e.g. *Kidney_Disease_GeneCards* [113]),

users can visualize how disease-related transcriptional programs are distributed across cell types and organs within a selected organ and across organs within the investigated species (Fig. 8A).

Across species, the *Kidney_Disease_GeneCards* gene set shows cell-type specificity and species divergence, with high scores in macrophages (human), proximal tubules (monkey, mouse), loop of Henle (human, monkey), and podocytes (monkey) (Fig. 8B–D). These patterns are accompanied by evidence of innate immune activation and stromal responses. The pattern matches the well-established kidney disease pathway: epithelial injury → immune/stromal activation → fibrotic progression [114, 115]. While in zebrafish, scores are generally low (Fig. 8E). Overall, this cross-species trend underscores the gene set's ability to capture core pathological axes of mammalian kidney disease.

DeathMap module: atlas of cell death-related gene set activity

DeathMap focuses on cell death-related transcriptional programs and includes 28 curated gene sets representing key pathways such as apoptosis, pyroptosis, ferroptosis, autophagy, and others. Users begin by selecting a gene set (e.g. *Apoptosis_Xie* [116]), then choose an organ of interest from 12 supported species. Upon organ selection (e.g. liver), gene-set activity scores are computed across all annotated cell types of that organ and visualized as bar plots. The gene set activity across organs within each display species is also provided in this module (Fig. 9A).

Results show that the *Apoptosis_Xie* gene set is primarily enriched in immune and vascular compartments of the liver across mammals. In humans, macrophages, T cells, and monocytes show the highest scores (Fig. 9B). In crab-eating macaques, Kupffer cells rank first, followed by monocytes and endothelial/dendritic cells, while hepatocytes and hepatic stellate cells display low or negative scores (Fig. 9C). Mice exhibit a similar hierarchy, with Kupffer cells and endothelial cells leading (Fig. 9D). In zebrafish, apoptotic activity is prominent in macrophages, biliary epithelial cells, and endothelial cells, moderate in hepatocytes, and low in erythrocytes (Fig. 9E).

Discussion

Here, we provide scGeneBank, which, to our knowledge, is the first resource to support large-scale, gene-set–based functional analysis across a wide taxonomic range at single-cell resolution. The current release covers 1134 scRNA-seq datasets from 136 species, totaling 104.9 million cells. All datasets are processed with a unified pipeline: non-human genes are mapped to human orthologs, and gene-set activity scores are computed using a standardized method. We manually curated 597 gene sets, grouped into 28 biological categories. Together, these elements enable comparative interrogation of key biological processes such as stemness, aging, disease progression, and cell death across species, organs, and cell types.

scGeneBank enables cross-dataset and cross-species comparisons at single-cell resolution. The platform offers multiple interactive modules to satisfy diverse analytical needs: GeneSets for browsing 597 gene sets across 28 categories with standardized annotations and links to external databases; EvoScreening for cross-species comparisons of gene-set activity within selected organs; OrganScreening for cross-organ analyses across 35 human organs; and four thematic modules—StemAtlas, AgingScope, PathoView, and DeathMap—that focus on stemness, aging, disease, and cell death, respectively, leveraging whole-body atlases for 12 supported species. At the dataset level, the ViewData page within the Species module supports interactive exploration of gene-set activity and cell-type composition with multiple visualization options. All figures and result tables are downloadable to facilitate reuse and publication.

In future updates, we will incorporate newly published single-cell datasets and expand the catalog of functional gene sets. Furthermore, we aim to integrate single-cell multi-omics datasets to enrich functional analyses and facilitate more comprehensive interpretation of cellular states.

Acknowledgement

We acknowledge the high-throughput sequencing and high-performance computing platform of Institute of Systems Medicine, Chinese Academy of Medical Sciences/Suzhou Institute of Systems Medicine (ISM) for technical support.

Author contributions: D.C., L.H., Y.M., and X.W. conceived and supervised the project. S.Q., Y.Y., L.P., L.X., Y.D., H.W., X.Q., and Y.H. collected datasets and developed bioinformatics pipelines. S.Q., Y.Y., L.P., and L.X. curated and organized the gene sets. D.C., L.H., S.Q., Y.Y., L.P., and L.X. developed the scGeneBank website. D.C., L.H., S.Q., and Y.Y. wrote the manuscript. All authors reviewed and approved the final version.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

The work was supported by National Key Research and Development Program of China (2024YFC3607500), Science and Technology Innovation 2030 Major Project (STI2030-Major Projects 2022ZD0205700), Non communicable Chronic Diseases-National Science and Technology Major Project (2024ZD0519600), National Natural Science Foundation of China under Grant (32300560), the Natural Science Foundation of Jiangsu Province (Grants No BK20220279), the CAMS Innovation Fund for Medical Sciences (CIFMS) (2021-I2M-1-061, 2021-I2M-1-074, 2022-I2M-2-004), the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (2022-RC416-01), the NCTIB Fund for R&D Platform for Cell and Gene Therapy, the Suzhou Municipal Key Laboratory (SZS2022005, SZS2023005), Gusu Innovation and Entrepreneurship Leading Talents Program (ZXL2022475), Beijing Research Ward Excellence Program (BRWEP2024W114060100), and 2025 PHC Cai Yuanpei project, the Special Research Fund for Central Universities, Peking Union Medical College (3332024089).

Data availability

All data for scGeneBank are freely accessible at http://8.142.154.29/scGeneBank.

References

1. Jin K, Yao Z, van Velthoven CTJ *et al.* Brain-wide cell-type-specific transcriptomic signatures of healthy ageing in mice. *Nature* 2025;638:182–96. https://doi.org/10.1038/s41586-024-08350-8
2. Qian Z, Qin J, Lai Y *et al.* Large-scale integration of single-cell RNA-seq data reveals astrocyte diversity and transcriptomic modules across six central nervous system disorders. *Biomolecules* 2023;13:692. https://doi.org/10.3390/biom13040692.
3. Kamath T, Abdulraouf A, Burris SJ *et al.* Single-cell genomic profiling of human dopamine neurons identifies a population that selectively degenerates in Parkinson's disease. *Nat Neurosci* 2022;25:588–95. https://doi.org/10.1038/s41593-022-01061-1
4. Sepp M, Leiss K, Murat F *et al.* Cellular development and evolution of the mammalian cerebellum. *Nature* 2024;625:788–96. https://doi.org/10.1038/s41586-023-06884-x
5. Feng X, Gao Y, Chu F *et al.* Cortical arealization of interneurons defines shared and distinct molecular programs in developing human and macaque brains. *Nat Commun* 2025;16:672. https://doi.org/10.1038/s41467-025-56058-8
6. Han X, Zhou Z, Fei L *et al.* Construction of a human cell landscape at single-cell level. *Nature* 2020;581:303–9. https://doi.org/10.1038/s41586-020-2157-4
7. Longo SK, Guo MG, Ji AL *et al.* Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat Rev Genet* 2021;22:627–44. https://doi.org/10.1038/s41576-021-00370-8
8. Guilliams M, Bonnardel J, Haest B *et al.* Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. *Cell* 2022;185:379–96. https://doi.org/10.1016/j.cell.2021.12.018
9. Ammons DT, Harris RA, Hopkins LS *et al.* A single-cell RNA sequencing atlas of circulating leukocytes from healthy and osteosarcoma affected dogs. *Front Immunol* 2023;14:1162700. https://doi.org/10.3389/fimmu.2023.1162700
10. Wang L, Wang C, Moriano JA *et al.* Molecular and cellular dynamics of the developing human neocortex. *Nature* 2025. https://doi.org/10.1038/s41586-024-08351-7
11. Mayshar Y, Raz O, Cheng S *et al.* Time-aligned hourglass gastrulation models in rabbit and mouse. *Cell* 2023;186:2610–27. https://doi.org/10.1016/j.cell.2023.04.037
12. Zhao C, Plaza Reyes A, Schell JP *et al.* A comprehensive human embryo reference tool using single-cell RNA-sequencing data. *Nat Methods* 2025;22:193–206. https://doi.org/10.1038/s41592-024-02493-2
13. Adams TS, Schupp JC, Poli S *et al.* Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci Adv* 2020;6:eaba1983. https://doi.org/10.1126/sciadv.aba1983
14. Reyfman PA, Walter JM, Joshi N *et al.* Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *Am J Respir Crit Care Med* 2019;199:1517–36. https://doi.org/10.1164/rccm.201712-2410OC
15. Mathys H, Davila-Velderrain J, Peng Z *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* 2019;570:332–7. https://doi.org/10.1038/s41586-019-1195-2
16. Almeida MC, Eger SJ, He C *et al.* Single-nucleus RNA sequencing demonstrates an autosomal dominant Alzheimer's disease profile and possible mechanisms of disease protection. *Neuron* 2024;112:1778–94. https://doi.org/10.1016/j.neuron.2024.02.009
17. Yang AC, Kern F, Losada PM *et al.* Dysregulation of brain and choroid plexus cell types in severe COVID-19. *Nature* 2021;595:565–71. https://doi.org/10.1038/s41586-021-03710-0
18. Wang Q, Wang M, Choi I *et al.* Molecular profiling of human substantia nigra identifies diverse neuron types associated with vulnerability in Parkinson's disease. *Sci Adv* 2024;10:eadi8287. https://doi.org/10.1126/sciadv.adi8287
19. Hahn J, Monavarfeshani A, Qiao M *et al.* Evolution of neuronal cell classes and types in the vertebrate retina. *Nature* 2023;624:415–24. https://doi.org/10.1038/s41586-023-06638-9
20. Tibi M, Biton Hayun S, Hochgerner H *et al.* A telencephalon cell type atlas for goldfish reveals diversity in the evolution of spatial structure and cell types. *Sci Adv* 2023;9:eadh7693. https://doi.org/10.1126/sciadv.adh7693
21. Geirsdottir L, David E, Keren-Shaul H *et al.* Cross-species single-cell analysis reveals divergence of the primate microglia program. *Cell* 2019;179:1609–22. https://doi.org/10.1016/j.cell.2019.11.010
22. Lamanna F, Hervas-Sotomayor F, Oel AP *et al.* A lamprey neural cell type atlas illuminates the origins of the vertebrate brain. *Nat Ecol Evol* 2023;7:1714–28. https://doi.org/10.1038/s41559-023-02170-1
23. Wu B, Xu W, Wu K *et al.* Single-cell analysis of the amphioxus hepatic caecum and vertebrate liver reveals genetic mechanisms of vertebrate liver evolution. *Nat Ecol Evol* 2024;8:1972–90. https://doi.org/10.1038/s41559-024-02510-9
24. Liberzon A, Birger C, Thorvaldsdóttir H *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Systems* 2015;1:417–25. https://doi.org/10.1016/j.cels.2015.12.004
25. Franchini M, Pellecchia S, Viscido G *et al.* Single-cell gene set enrichment analysis and transfer learning for functional annotation of scRNA-seq data. *NAR Genom Bioinform* 2023;5:lqad024. https://doi.org/10.1093/nargab/lqad024
26. Gibbs DL, Strasser MK, Huang S. Single-cell gene set scoring with nearest neighbor graph smoothed data (gssnng). *Bioinform Adv* 2023;3:vbad150. https://doi.org/10.1093/bioadv/vbad150
27. Frost HR. Leveraging cell type-specificity for gene set analysis of single cell transcriptomics. bioRxiv, https://doi.org/10.1101/2024.09.25.615040, 27 September 2024, preprint: not peer reviewed
28. Mooney MA, Wilmot B. Gene set analysis: a step-by-step guide. *Am J Med Genet B Neuropsychiatr Genet* 2015;168:517–27. https://doi.org/10.1002/ajmg.b.32328
29. Figueiredo RQ, Del Ser SD, Raschka T *et al.* Elucidating gene expression patterns across multiple biological contexts through a large-scale investigation of transcriptomic datasets. *BMC Bioinformatics* 2022;23:231. https://doi.org/10.1186/s12859-022-04765-0
30. Yuan H, Yan M, Zhang G *et al.* CancerSEA: a cancer single-cell state atlas. *Nucleic Acids Res* 2018;47:D900–8. https://doi.org/10.1093/nar/gky939
31. Liu CJ, Hu FF, Xie GY *et al.* GSCA: an integrated platform for gene set cancer analysis at genomic, pharmacogenomic and immunogenomic levels. *Brief Bioinform* 2023;24:bbac558.
32. Tarhan L, Bistline J, Chang J *et al.* Single Cell Portal: an interactive home for single-cell genomics data. bioRxiv, https://doi.org/10.1101/2023.07.13.548886, 17 July 2023, preprint: not peer reviewed.
33. Cheng W, Yin C, Yu S *et al.* scMMO-atlas: a single cell multimodal omics atlas and portal for exploring fine cell heterogeneity and cell dynamics. *Nucleic Acids Res* 2024;53:D1186–94. https://doi.org/10.1093/nar/gkae821
34. Ma Q, Tao H, Li Q *et al.* OrganoidDB: a comprehensive organoid database for the multi-perspective exploration of bulk and single-cell transcriptomic profiles of organoids. *Nucleic Acids Res* 2023;51:D1086–93. https://doi.org/10.1093/nar/gkac942
35. Zhou Z, Tan C, Chau MHK *et al.* TEDD: a database of temporal gene expression patterns during multiple developmental periods

in human and model organisms. *Nucleic Acids Res* 2023;51:D1168–78. https://doi.org/10.1093/nar/gkac978
36. Hu Y, Tattikota SG, Liu Y *et al.* DRscDB: a single-cell RNA-seq resource for data mining and data comparison across species. *Comput Struct Biotechnol J* 2021;19:2018–26. https://doi.org/10.1016/j.csbj.2021.04.021
37. Li F, Song X, Fan W *et al.* SPathDB: a comprehensive database of spatial pathway activity atlas. *Nucleic Acids Res* 2024;53:D1205–14. https://doi.org/10.1093/nar/gkae1041
38. Wei J, Liu P, Liu F *et al.* EDomics: a comprehensive and comparative multi-omics database for animal evo-devo. *Nucleic Acids Res* 2022;51:D913–23. https://doi.org/10.1093/nar/gkac944
39. National Library of Medicine (US); National Center for Biotechnology Information. PubMed [Internet]. Bethesda (MD): Available from: https://pubmed.ncbi.nlm.nih.gov (27 October 2024, date last accessed).
40. Program CCS, Abdulla S, Aevermann B *et al.* CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res* 2024;53:D886–900. https://doi.org/10.1093/nar/gkae1142
41. Sunkin SM, Ng L, Lau C *et al.* Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res* 2013;41:D996–1008. https://doi.org/10.1093/nar/gks1042
42. Jones RC, Karkanias J, Krasnow MA *et al.* The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science* 2022;376:eabl4896.
43. Chen Y, Zhang X, Peng X *et al.* SPEED: single-cell pan-species atlas in the light of ecology and evolution for development and diseases. *Nucleic Acids Res* 2022;51:D1150–9. https://doi.org/10.1093/nar/gkac930
44. Deng Y, Lu Y, Li M *et al.* SCAN: spatiotemporal cloud atlas for neural cells. *Nucleic Acids Res* 2023;52:D998–1009. https://doi.org/10.1093/nar/gkad895
45. Chen D, Tan C, Ding P *et al.* VThunter: a database for single-cell screening of virus target cells in the animal kingdom. *Nucleic Acids Res* 2021;50:D934–42. https://doi.org/10.1093/nar/gkab894
46. Han L, Wei X, Liu C *et al.* Cell transcriptomic atlas of the non-human primate *Macaca fascicularis*. *Nature* 2022;604:723–31. https://doi.org/10.1038/s41586-022-04587-3
47. de Morree A, De Vlaminck I, Shapiro L *et al.* A molecular cell atlas of mouse lemur, an emerging model primate. *Nature* 2025;644:173–84.
48. Han X, Wang R, Zhou Y *et al.* Mapping the mouse cell atlas by Microwell-seq. *Cell* 2018;172:1091–107. https://doi.org/10.1016/j.cell.2018.02.001
49. Wang F, Ding P, Liang X *et al.* Endothelial cell heterogeneity and microglia regulons revealed by a pig cell landscape at single-cell level. *Nat Commun* 2022;13:3620. https://doi.org/10.1038/s41467-022-31388-z
50. Wu J-J, Zhu S, Gu F *et al.* Cross-tissue single-cell transcriptomic landscape reveals the key cell subtypes and their potential roles in the nutrient absorption and metabolism in dairy cattle. *J Adv Res* 2022;37:1–18. https://doi.org/10.1016/j.jare.2021.11.009
51. Tian Y, Liu S, Shi H *et al.* Revealing the transcriptional and metabolic characteristics of sebocytes based on the donkey cell transcriptome atlas. *Adv Sci* 2025;12:e2413819. https://doi.org/10.1002/advs.202413819
52. Chen D, Sun J, Zhu J *et al.* Single cell atlas for 11 non-model mammals, reptiles and birds. *Nat Commun* 2021;12:7083. https://doi.org/10.1038/s41467-021-27162-2
53. Liao Y, Ma L, Guo Q *et al.* Cell landscape of larval and adult *Xenopus laevis* at single-cell resolution. *Nat Commun* 2022;13:4306. https://doi.org/10.1038/s41467-022-31949-2
54. Jiang M, Xiao Y, E W *et al.* Characterization of the zebrafish cell landscape at single-cell resolution. *Front Cell Dev Biol* 2021;9:743421. https://doi.org/10.3389/fcell.2021.743421
55. Florio M, Heide M, Pinson A *et al.* Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. *eLife* 2018;7:e32332.
56. Aibar S, González-Blas CB, Moerman T *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 2017;14:1083–6. https://doi.org/10.1038/nmeth.4463
57. Viner-Breuer R, Yilmaz A, Benvenisty N *et al.* The essentiality landscape of cell cycle related genes in human pluripotent and cancer cells. *Cell Div* 2019;14:15. https://doi.org/10.1186/s13008-019-0058-4
58. Lucas CL, Lenardo MJ. Identifying genetic determinants of autoimmunity and immune dysregulation. *Curr Opin Immunol* 2015;37:28–33. https://doi.org/10.1016/j.coi.2015.09.001
59. Saul D, Kosinsky RL, Atkinson EJ *et al.* A new gene set identifies senescent cells and predicts senescence-associated pathways across tissues. *Nat Commun* 2022;13:4827. https://doi.org/10.1038/s41467-022-32552-1
60. Yu Z, Seim I, Yin M *et al.* Comparative analyses of aging-related genes in long-lived mammals provide insights into natural longevity. *The Innovation* 2021;2:100108. https://doi.org/10.1016/j.xinn.2021.100108
61. Zhang Y, Kiryu H. Identification of oxidative stress-related genes differentially expressed in Alzheimer's disease and construction of a hub gene-based diagnostic model. *Sci Rep* 2023;13:6817. https://doi.org/10.1038/s41598-023-34021-1
62. Zhu X, Han X, Li Z *et al.* CircaKB: a comprehensive knowledgebase of circadian genes across multiple species. *Nucleic Acids Res* 2024;53:D67–78. https://doi.org/10.1093/nar/gkae817
63. Zou X, Ptáček LJ, Fu Y-H. The genetics of human sleep and sleep disorders. *Annu Rev Genomics Hum Genet* 2024;25:259–85. https://doi.org/10.1146/annurev-genom-121222-120306
64. Tao G, Jiao C, Wang Y *et al.* Comprehensive analysis of hypoxia-related genes for prognosis, immune features, and drugs treatment strategy in gastric cancer using bulk and single-cell RNA-sequencing. *Sci Rep* 2022;12:21739. https://doi.org/10.1038/s41598-022-26395-5
65. Gandal MJ, Haney JR, Parikshak NN *et al.* Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* 2018;359:693–7. https://doi.org/10.1126/science.aad6469
66. Loos RJF, Yeo GSH. The genetics of obesity: from discovery to biology. *Nat Rev Genet* 2022;23:120–33. https://doi.org/10.1038/s41576-021-00414-z
67. Zhou Y, Arribas GH, Turku A *et al.* Rare genetic variability in human drug target genes modulates drug response and can guide precision medicine. *Sci Adv* 2021;7:eabi6856. https://doi.org/10.1126/sciadv.abi6856
68. Thul PJ, Lindskog C. The human protein atlas: a spatial map of the human proteome. *Protein Sci* 2018;27:233–44. https://doi.org/10.1002/pro.3307
69. Tay S-K, Blythe J, Lipovich L. Global discovery of primate-specific genes in the human genome. *Proc Natl Acad Sci USA* 2009;106:12019–24. https://doi.org/10.1073/pnas.0904569106
70. Li Y, Hu M, Zhang Z *et al.* Origin and stepwise evolution of vertebrate lungs. *Nat Ecol Evol* 2025;9:672–91. https://doi.org/10.1038/s41559-025-02642-6
71. Huang YF, Aoki K, Akase S *et al.* Global mapping of glycosylation pathways in human-derived cells. *Dev Cell* 2021;56:1195–209. https://doi.org/10.1016/j.devcel.2021.02.023
72. Shen WK, Chen SY, Gan ZQ *et al.* AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. *Nucleic Acids Res* 2023;51:D39–45. https://doi.org/10.1093/nar/gkac907
73. Zhang H-M, Chen H, Liu W *et al.* AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res* 2012;40:D144–9. https://doi.org/10.1093/nar/gkr965
74. Hu H, Miao Y-R, Jia L-H *et al.* AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal

transcription factors. *Nucleic Acids Res* 2018;47:D33–8. https://doi.org/10.1093/nar/gky822
75. Safran M, Dalah I, Alexander J *et al.* GeneCards Version 3: the human gene integrator. *Database* 2010;2010:baq020. https://doi.org/10.1093/database/baq020, baq020.
76. Franzén O, Gan LM, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)* 2019;2019:baz046.
77. Durinck S, Moreau Y, Kasprzyk A *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 2005;21:3439–40. https://doi.org/10.1093/bioinformatics/bti525
78. O'Leary NA, Wright MW, Brister JR *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;44:D733–45. https://doi.org/10.1093/nar/gkv1189
79. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 2019;20:238. https://doi.org/10.1186/s13059-019-1832-y
80. Zheng GXY, Terry JM, Belgrader P *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;8:14049. https://doi.org/10.1038/ncomms14049
81. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;19:15. https://doi.org/10.1186/s13059-017-1382-0
82. Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Systems* 2019;8:281–91. https://doi.org/10.1016/j.cels.2018.11.005
83. Korsunsky I, Millard N, Fan J *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 2019;16:1289–96. https://doi.org/10.1038/s41592-019-0619-0
84. Arevalo J, Su E, Ewald JD *et al.* Evaluating batch correction methods for image-based cell profiling. *Nat Commun* 2024;15:6516. https://doi.org/10.1038/s41467-024-50613-5
85. Luecken MD, Büttner M, Chaichoompu K *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* 2022;19:41–50. https://doi.org/10.1038/s41592-021-01336-8
86. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2006;8:118–27. https://doi.org/10.1093/biostatistics/kxj037
87. Leek JT, Johnson WE, Parker HS *et al.* The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 2012;28:882–3. https://doi.org/10.1093/bioinformatics/bts034
88. Sayers EW, Beck J, Bolton EE *et al.* Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res* 2025;53:D20–9. https://doi.org/10.1093/nar/gkae979
89. Dyer SC, Austine-Orimoloye O, Azov AG *et al.* Ensembl 2025. *Nucleic Acids Res* 2024;53:D948–57. https://doi.org/10.1093/nar/gkae1071
90. Consortium TU. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* 2024;53:D609–17. https://doi.org/10.1093/nar/gkae1010
91. Andrews SJ, Renton AE, Fulton-Howard B *et al.* The complex genetic architecture of Alzheimer's disease: novel insights and future directions. *EBioMedicine* 2023;90:104511. https://doi.org/10.1016/j.ebiom.2023.104511
92. Siletti K, Hodge R, Mossi Albiach A *et al.* Transcriptomic diversity of cell types across the adult human brain. *Science* 2023;382:eadd7046. https://doi.org/10.1126/science.add7046
93. Leng F, Edison P. Neuroinflammation and microglial activation in Alzheimer disease: where do we go from here? *Nat Rev Neurol* 2021;17:157–72. https://doi.org/10.1038/s41582-020-00435-y
94. Subramanian A, Tamayo P, Mootha VK *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;102:15545–50. https://doi.org/10.1073/pnas.0506580102
95. Moon H, Ro SW. MAPK/ERK signaling pathway in hepatocellular carcinoma. *Cancers* 2021;13:3026. https://doi.org/10.3390/cancers13123026.
96. Shamsan E, Almezgagi M, Gamah M *et al.* The role of PI3k/AKT signaling pathway in attenuating liver fibrosis: a comprehensive review. *Front Med* 2024;11:1389329. https://doi.org/10.3389/fmed.2024.1389329
97. Zhang Z, Zhou H, Ouyang X *et al.* Multifaceted functions of STING in human health and disease: from molecular mechanism to targeted strategy. *Signal Transduct Target Ther* 2022;7:394. https://doi.org/10.1038/s41392-022-01252-z
98. Li X, Chen X, Zheng L *et al.* Non-canonical STING–PERK pathway dependent epigenetic regulation of vascular endothelial dysfunction via integrating IRF3 and NF-κB in inflammatory response. *Acta Pharmaceutica Sinica B* 2023;13:4765–84. https://doi.org/10.1016/j.apsb.2023.08.015
99. Liu J, Xiao Q, Xiao J *et al.* Wnt/β-catenin signalling: function, biological mechanisms, and therapeutic opportunities. *Signal Transduct Target Ther* 2022;7:3. https://doi.org/10.1038/s41392-021-00762-6
100. Jing J, Wu Z, Wang J *et al.* Hedgehog signaling in tissue homeostasis, cancers and targeted therapies. *Signal Transduct Target Ther* 2023;8:315. https://doi.org/10.1038/s41392-023-01559-5
101. Fu M, Hu Y, Lan T *et al.* The Hippo signalling pathway and its implications in human health and diseases. *Signal Transduction and Targeted Therapy* 2022;7:376.
102. Hill MM, Clark SF, Tucker DF *et al.* A role for protein kinase Bbeta/Akt2 in insulin-stimulated GLUT4 translocation in adipocytes. *Mol Cell Biol* 1999;19:7771–81. https://doi.org/10.1128/MCB.19.11.7771
103. Kra G, Daddam JR, Moallem U *et al.* Alpha-linolenic acid modulates systemic and adipose tissue-specific insulin sensitivity, inflammation, and the endocannabinoid system in dairy cows. *Sci Rep* 2023;13:5280. https://doi.org/10.1038/s41598-023-32433-7
104. King SM, Hilliard TS, Wu LY *et al.* The impact of ovulation on fallopian tube epithelial cells: evaluating three hypotheses connecting ovulation and serous ovarian cancer. *Endocr Relat Cancer* 2011;18:627–42. https://doi.org/10.1530/ERC-11-0107
105. Barata T, Duarte I, Futschik ME. Integration of stemness gene signatures reveals core functional modules of stem cells and potential novel stemness genes. *Genes* 2023;14:745. https://doi.org/10.3390/genes14030745.
106. Consortium AA. Aging Atlas: a multi-omics database for aging biology. *Nucleic Acids Res* 2020;49:D825–30. https://doi.org/10.1093/nar/gkaa894
107. Yi F, Yuan J, Somekh J *et al.* Genetically supported targets and drug repurposing for brain aging: a systematic study in the UK Biobank. *Sci Adv* 2025;11:eadr3757. https://doi.org/10.1126/sciadv.adr3757
108. de Magalhães JP, Abidi Z, Dos Santos GA *et al.* Human Ageing Genomic Resources: updates on key databases in ageing research. *Nucleic Acids Res* 2024;52:D900–8. https://doi.org/10.1093/nar/gkad927
109. Liu S, Meng Y, Zhang Y *et al.* Integrative analysis of senescence-related genes identifies robust prognostic clusters with distinct features in hepatocellular carcinoma. *J Adv Res* 2025;69:107–23. https://doi.org/10.1016/j.jare.2024.04.007
110. Dong M, Dong Y, Bai J *et al.* Interactions between microbiota and cervical epithelial, immune, and mucus barrier. *Front Cell Infect Microbiol* 2023;13:1124591. https://doi.org/10.3389/fcimb.2023.1124591
111. Jin S, Li K, Zong X *et al.* Hallmarks of skin aging: update. *Aging and Disease* 2023;14:2167–76. https://doi.org/10.14336/AD.2023.0321

112. Kim IH, Kisseleva T, Brenner DA. Aging and liver disease. *Curr Opin Gastroenterol* 2015;31:184–91. https://doi.org/10.1097/MOG.0000000000000176
113. Stelzer G, Rosen N, Plaschkes I *et al*. The GeneCards Suite: from gene data mining to disease genome sequence analyses. *Curr Protoc Bioinformatics* 2016;54:1.30.31–31.30.33. https://doi.org/10.1002/cpbi.5
114. Yuan Q, Tang B, Zhang C. Signaling pathways of chronic kidney diseases, implications for therapeutics. *Signal Transduct Target Ther* 2022;7:182. https://doi.org/10.1038/s41392-022-01036-5
115. Reiss AB, Jacob B, Zubair A *et al*. Fibrosis in chronic kidney disease: pathophysiology and therapeutic targets. *J Clin Med* 2024;13:1881. https://doi.org/10.3390/jcm13071881.
116. Zou Y, Xie J, Zheng S *et al*. Leveraging diverse cell-death patterns to predict the prognosis and drug sensitivity of triple-negative breast cancer patients after surgery. *Int J Surg* 2022;107:106936. https://doi.org/10.1016/j.ijsu.2022.106936

Received: August 11, 2025. **Revised:** September 12, 2025. **Accepted:** September 29, 2025