Nucleic Acids Research, 2025, gkaf1153
https://doi.org/10.1093/nar/gkaf1153
Database issue

scMOVIR: a single-cell multi-omics database for human viral infections and immune responses

Xue Zhang[1,2], Shounan Yang[1], Xiaobin Xu[1,2], Yitao Lin[1,2], Huaicheng Sun[1], Bangyu Zhu[1], Wenyi Zhao[1,2], Binbin Zhou[3], Yan Lou[4], Xinyu Wang[4], Shuqing Chen[1], Qiaojun He[1,2], Feng Zhu[1,5,*], Zhan Zhou[1,2,6,*]

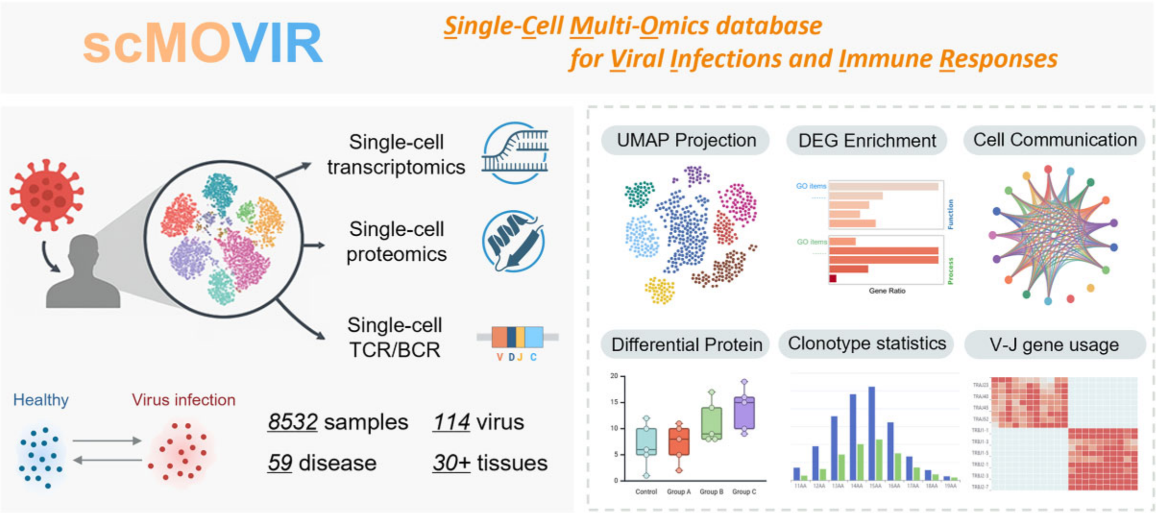
[1]State Key Laboratory of Advanced Drug Delivery and Release Systems & Innovation Institute for Artificial Intelligence in Medicine, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China
[2]Zhejiang Provincial Key Laboratory of Anti-Cancer Drug Research & MOE Engineering Research Center of Innovative Anticancer Drugs, Zhejiang University, Hangzhou 310018, China
[3]School of Computer and Computing Science, Hangzhou City University, Hangzhou 310015, China
[4]The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310003, China
[5]The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310009, China
[6]The Fourth Affiliated Hospital, Zhejiang University School of Medicine, Yiwu 322000, China

*To whom correspondence should be addressed. Email: zhanzhou@zju.edu.cn
Correspondence may also be addressed to Feng Zhu. Email: zhufeng@zju.edu.cn

Abstract

Viral infections impose a substantial threat to human health, characterized by a wide range of pathogens, clinical manifestations, and complex immune responses. Single-cell multi-omics technologies have revolutionized the study of antiviral immunity by resolving cellular heterogeneity, transcriptional reprogramming, and clonal dynamics. However, no resource has yet comprehensively integrated such datasets in the context of viral infections. Here, we present scMOVIR, a single-cell multi-omics database for human viral infections and immune responses. The database systematically integrates transcriptomic, proteomic, and immune receptor repertoire profiles at single-cell resolution, compiling 8532 human samples across 114 viral species, subtypes, and strains. These datasets span 59 well-defined disease states, including acute and chronic infections, virus-associated malignancies, and immune-related disorders, and extend across >30 tissue types. In addition, these datasets incorporate vaccination cohorts, therapeutic interventions, and time-course models of infection. To ensure consistency and interoperability, all datasets undergo standardized preprocessing, including batch correction and unified cell-type annotation, with metadata harmonized using authoritative ontologies. scMOVIR provides user-friendly interfaces for dataset exploration and interactive visualization of cellular dynamics and molecular profiles, offering a high-resolution reference for investigating virus–host immune interactions and supporting antiviral research. The database is freely accessible at https://pgx.zju.edu.cn/scmovir.

Graphical abstract



Received: August 15, 2025. **Revised:** October 9, 2025. **Accepted:** October 11, 2025

Introduction

Viral infections represent a major threat to human health, encompassing diverse pathogens and heterogeneous clinical manifestations ranging from acute self-limiting infections to persistent or latent infections [1–3]. Epidemiological studies estimate that over 12% of newly diagnosed cancers are attributable to oncogenic viral infections, with an even higher proportion in resource-limited regions [4–6]. For example, human immunodeficiency virus (HIV) establishes a persistent infection characterized by chronic immune activation, progressive CD4$^+$ T cell depletion, and compromised host defense [7]. Hepatitis B virus (HBV) and human papillomavirus are strongly associated with malignancies such as hepatocellular carcinoma and cervical cancer [8–10]. Epstein–Barr virus (EBV) is linked to multiple lymphoid malignancies and exhibits complex latency and reactivation cycles [11]. Additionally, the outbreak of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has led to a global pandemic, resulting in millions of deaths worldwide [12, 13]. Therefore, elucidating virus–host immune interactions is critical to advancing our understanding of viral pathogenesis and its clinical implications.

Viral infections elicit dynamic and spatially compartmentalized immune responses characterized by activation of distinct cellular subsets, phenotypic remodeling, and antigen-driven clonal expansion [14–16]. These responses are accompanied by transcriptional reprogramming and signaling network rewiring, collectively reshaping cellular functional states [16, 17]. Conventional bulk omics approaches, which average signals across heterogeneous cell populations, are insufficient to resolve these complexities [18–20]. Single-cell multi-omics technologies provide the necessary resolution to capture cellular heterogeneity and dynamic molecular responses during viral infection [21, 22]. For example, paired single-cell RNA sequencing (scRNA-seq) and T-cell receptor sequencing (scTCR-seq) in SARS-CoV-2 infection have revealed rapid antigen-specific T-cell expansion and interferon-driven mucosal immune responses during early infection [23]. In HBV-associated hepatocellular carcinoma, multi-omics profiling identified exhausted and cytotoxic CD8$^+$ T cell subsets associated with distinct tumor recurrence patterns [24]. Similarly, studies of HIV and EBV infections applied single-cell multi-omics to delineate virus-specific clonal hierarchies and immune dysregulation pathways [25, 26]. These advances underscore the critical role of single-cell technologies in elucidating the cellular and molecular basis of antiviral immunity and pathogenesis.

Several publicly available databases have been developed. For example, MVIP offers extensive bulk multi-omics datasets alongside a subset of scRNA-seq data on virus–host interactions [27], and VThunter provides single-cell resolution profiles of viral receptor expression across human tissues to assess tissue susceptibility [28]. Other platforms, such as TISCH2 [29] and SC2Disease [30], emphasize the immune landscape in cancer and other diseases based on single-cell transcriptomic profiles. More recently, scImmomics has integrated T-cell receptor and B-cell receptor (TCR/BCR) repertoires with scRNA-seq data to reconstruct immune cell lineage relationships [31], and additional resources further expand specific aspects of viral single-cell immunology [32, 33]. While these resources have greatly contributed to the field, the inherent complexity of antiviral immune responses, such as rapid clonal dynamics, systemic immune activation, and pathogen-specific signatures, remains challenging to fully elucidate through bulk data. Moreover, existing single-cell resources primarily focus on transcriptomic modalities within oncological contexts, with limited coverage of diverse viral species and infection stages. In addition, inconsistencies in data preprocessing and annotation standards across studies hinder seamless integration and comprehensive cross-study analysis. These limitations underscore the need for a unified single-cell multi-omics platform specifically designed for human viral infections and immune responses.

Therefore, we developed scMOVIR, a multi-omics database for human viral infections and immune responses, which integrates transcriptomic, proteomic, and immune receptor repertoire data at single-cell resolution. scMOVIR incorporates 8532 samples covering 114 viral species, subtypes, and strains, 59 disease states, and over 30 tissue types, spanning acute and chronic infection contexts, virus-associated cancers, vaccination cohorts, and therapeutic intervention studies. All datasets are subjected to standardized processing, quality control (QC), and unified annotation to ensure consistency and interoperability. We performed multi-layered analyses, including differential gene and protein expression profiling, functional enrichment, cell–cell communication inference, and clonality characterization. These results are systematically integrated into the database and accessible through flexible, interactive visualization tools. Compared with existing resources, scMOVIR provides a comprehensive, virus-focused single-cell multi-omics platform with standardized datasets and comparative analysis, enabling in-depth investigation of virus–host interactions and immune response heterogeneity across diverse infection contexts (Supplementary Table S1).

Materials and methods

Systematic collection and rigorous curation of datasets

To establish a comprehensive and high-quality collection of single-cell datasets related to viral infections, we implemented a multi-step data acquisition and management strategy: (i) Comprehensive literature review: We conducted exhaustive searches of public virus repositories, including ViPR [34], ICTV [35], and NCBI Virus [36], to identify viruses known to infect humans. Subsequently, a systematic literature search was conducted on PubMed using Entrez E-utility, employing keyword combinations such as "single-cell RNA sequencing + virus name," "single-cell proteomics + virus name," and "single-cell BCR/TCR + virus name." The search, completed in December 2024, yielded 3316 relevant publications. (ii) Public database mining: To complement the literature-based search, we conducted targeted mining of public omics repositories, including the Gene Expression Omnibus (GEO) [37], ArrayExpress [38], FlowRepository [39], SingPro [40], and SPDB [41], among others. This effort identified an additional 58 experiments covering various omics modalities. (iii) Verification and validation: All retrieved publications and datasets undergo rigorous validation through manual inspection of original articles and supplementary materials. Only publicly available single-cell datasets with clearly confirmed viral infection status were included. Comprehensive metadata were manually curated to ensure consistency and completeness. (iv) Data deduplication and integration: To eliminate redundancy, duplicated datasets were manually reviewed and consolidated. In cases where samples overlap across multiple studies, the dataset with more complete metadata is retained, and annotations are added to indicate the source of the original data to ensure data traceability.

Data quality control and preprocessing

scMOVIR employs a modality-specific standardized preprocessing pipeline to ensure data quality and enable cross-study comparability [42]. For single-cell transcriptomic data, we implemented a two-step QC strategy that combines empirical thresholds with sample-level median absolute deviation (MAD)-based outlier detection for all datasets. QC was performed using Scanpy [43]. First, low-quality cells were explicitly checked and removed based on established empirical thresholds: (i) cells with low gene counts (n_genes <200), (ii) cells with low total counts (total_counts <500), and (iii) cells with high mitochondrial content (pct_counts_mt $>10\%$). Then, sample-level outlier detection was performed based on the MAD of four metrics [44]: log-transformed gene counts, log-transformed total transcript counts, proportion of transcripts from the top 20 most highly expressed genes, and mitochondrial transcript fraction. MAD thresholds of 5 were applied to the first three metrics and 3 to the mitochondrial proportion.

For droplet-based single-cell sequencing, Scrublet [45] was applied to each sample to detect and remove potential doublets. The algorithm assigns a doublet score to each cell by simulating synthetic doublets and comparing them to observed profiles. Cells exceeding the score threshold were excluded using the following parameters: min_counts = 2, min_cells = 3, and min_gene_variability_pctl = 85. For each dataset, genes detected in fewer than three cells were filtered out. Expression counts were then normalized by total counts per cell (TCM), scaled to 10,000 counts, and log-transformed using log1p. Highly variable genes (HVGs) were identified using a dispersion-based selection method with the following parameters: min_disp = 0.5, min_mean = 0.0125, max_mean = 3, span = 0.3, and n_bins = 20. Principal component analysis (PCA) was subsequently performed on the selected HVGs using the ARPACK solver to compute 50 principal components. Batch effects were corrected using Harmony [46], with sample ID specified as the batch key. Nearest-neighbor graphs were constructed from the Harmony-corrected PCA embeddings. UMAP was applied for dimensionality reduction and visualization, using Euclidean distance with n_neighbors = 15, min_dist = 0.5, and spread = 1.0. Cell clustering was performed using the Leiden algorithm with an initial resolution of 1, which was further adjusted during manual cell-type annotation to resolve subpopulation structures more accurately.

For single-cell proteomic data from cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) experiments, cells with low counts, determined based on the distribution of captured antibody-derived tags, were removed. Sample-level QC was performed using MAD-based filtering on detected gene counts and total UMI counts with a 5-MAD cutoff. Normalization was conducted via centered log-ratio transformation. For flow cytometry-based proteomic datasets, raw data were imported into FlowJo (v 10.8) for QC, including removal of anomalous events based on scatter and signal intensity. Dead cells and atypical populations were manually gated out, and signal intensities were normalized using arcsinh transformation.

For scTCR/BCR-seq data, processing was performed using Scirpy [47]. Only cells with paired productive receptor sequences were retained: T cells with both α-chain (TRA) and β-chain (TRB), and B cells with both heavy (IGH) and light (IGK or IGL) chains. Clonotypes were defined as unique paired TRA–TRB or IGH–IGK/IGL sequence combinations, with clonal events required to be detected in at least two cells. For each clonotype, clonal frequency per cell, the proportion of clonotype-positive cells within each cell type, and clonotype distribution across cell types were calculated.

Cell-type annotation

We have implemented a combined automated and manual curation strategy. Initially, CellTypist was used to automatically annotate each dataset [48], employing the pretrained Immune_All_Low.pkl model. Predictions were first made at the single-cell level, followed by a majority-vote-based consensus to assign robust cell type labels and mitigate local overclustering artifacts. For clusters with low prediction confidence scores, we performed manual inspection using a curated list of marker genes validated in the literature, and manually corrected annotations for clusters that were inconsistent with expected marker gene expression. Supplementary Table S2 provides a comprehensive list of cell types, subtypes, and the marker genes collected from literatures [49–52]. For datasets that were derived from cell lines or specific cell types, manual curation was conducted based on literature and associated metadata.

Differentially expressed genes and proteins

Differentially expressed genes (DEGs) and proteins were identified using Welch's *t*-test on normalized, log-transformed expression data. DEGs were selected using false discovery rate <0.05, absolute $\log_2$ fold change >1, and gene expression in $>20\%$ of cells within the cluster. Cluster signature genes were defined as those significantly upregulated relative to all other clusters and expressed in the majority of cells within the cluster.

Functional enrichment analysis

Functional enrichment analysis was conducted using overrepresentation analysis. DEGs with adjusted P-values $<.05$ were assessed for enrichment of Gene Ontology (GO) biological processes [53, 54], Reactome [55], and KEGG pathways using Fisher's exact test with Benjamini–Hochberg correction [56, 57]. Only pathways meeting the corrected P-value threshold ($<.05$) were considered significant. The entire workflow was implemented via GSEApy [58].

Cell–cell communication

Cell–cell communication networks were inferred using CellphoneDB [59], which predicts ligand–receptor interactions between cell types based on scRNA-seq expression data. For each condition group, 1000 empirical permutations were performed by randomly shuffling cell cluster labels to estimate the null distribution of average ligand–receptor expression within interacting clusters. Ligand–receptor pairs with adjusted P-values $<.05$ were considered significant. To identify differential interactions, the $\log_2$ fold change ($\log_2$FC)

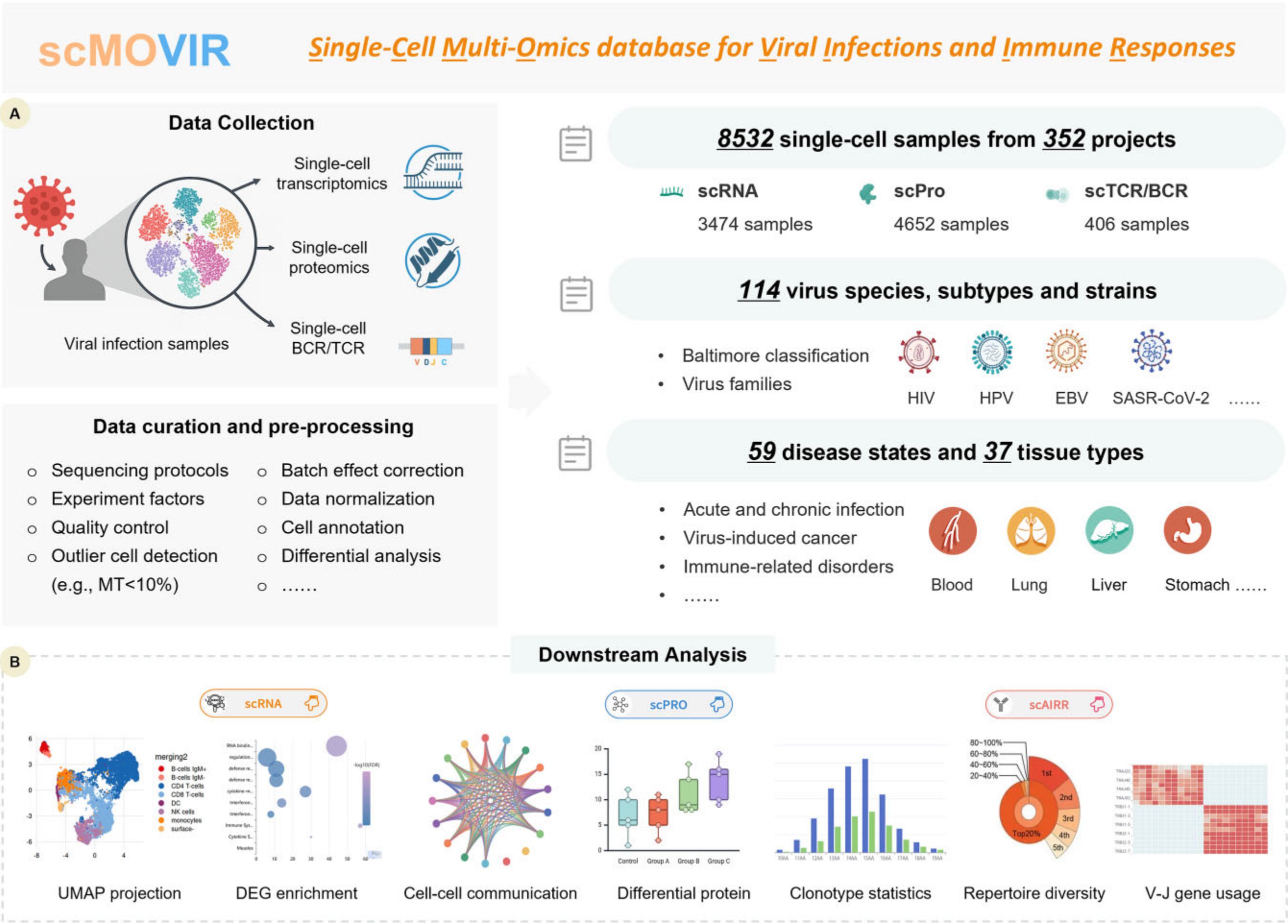


Figure 1. Overview of the scMOVIR workflow. (**A**) Systematic collection of single-cell multi-omics datasets related to viral infections from literature and multiple public repositories. All datasets undergo stringent QC and preprocessing, supplemented by manual curation and comprehensive metadata harmonization to ensure consistency and accuracy. (**B**) Downstream analyses and visualization are performed for each omics modality. Transcriptomic analyses encompass cell UMAP projection, cell-type composition, differential gene expression, functional enrichment, and cell–cell communication. Proteomic analyses involve the identification of differentially expressed proteins. Immune repertoire analyses include quantification of clonal abundance, repertoire diversity, and VJ gene usage patterns.

of ligand–receptor communication scores between the disease/virus group and the control group was calculated for each interacting cell pair, enabling the systematic detection of intercellular interactions that are specifically altered under experimental or infection conditions.

Clonal diversity scoring

Clonal diversity was assessed by computing the Shannon entropy of clonotype frequencies within each sample. To account for variability in sequencing depth, the observed entropy was normalized by dividing by its theoretical maximum, yielding a normalized diversity index ranging from 0, where a single clonotype dominates, to 1, representing maximal clonotype diversity. This approach enables quantitative comparison of clonal heterogeneity across samples.

Database implementation

scMOVIR is built on a modular architecture, with the backend developed using Django (v3.2.16) and MySQL (v5.7.34) for data management. The frontend utilizes Vue.js (v3.5.1) combined with Element Plus components to deliver an interactive and user-friendly interface. Visualization functionalities are implemented using ECharts (v5.5.1) [60] and Highcharts (v11.4.8), enabling dynamic rendering of gene expression profiles, high-dimensional single-cell data distributions, and immune receptor repertoires, thereby supporting comprehensive multi-omics data exploration.

Factual content and data retrieval

Overview of scMOVIR

The overall framework of scMOVIR is illustrated in Fig. 1. Through systematic collection from GEO [37], PubMed, and other related repositories, scMOVIR is a rigorously curated, systematically organized database that integrates high-quality single-cell multi-omics datasets focused on human viral infections and immune responses. The database comprises 8532 samples from 352 independent experiments, encompassing single-cell transcriptomics, single-cell proteomics (including scCITE-seq and fluorescence-based flow cytometry), and single-cell immune repertoire data, which includes paired TCR and BCR sequencing data (Fig. 1A). These datasets comprehensively capture the cellular and molecular landscape of virus–host interactions, encompassing 114 viral species, subtypes, and strains, including major pathogens such as HIV, HBV, and SARS-CoV-2. This collection covers 59 disease states, spanning acute and chronic infections, virus-associated malignancies (such as cervical cancer and hepatocellular carcinoma), and virus-associated immune dysregulation (such as lymphocytopenia and systemic lupus erythematosus), across >30 tissue types, including lung, liver, blood, lymph nodes, and brain. In addition, these datasets incorporate vaccination cohorts, clinical treatment interventions, longitudinal infection dynamics, and *in vitro* viral perturbation experiments.

All datasets undergo stringent QC, standardized preprocessing, and harmonized downstream processing, ensuring analytical robustness and cross-cohort comparability. The scMOVIR platform provides a unified framework for systematic interrogation of virus–host interactions, delineation of immune response dynamics, and mechanistic insights into viral pathogenesis across diverse clinical and experimental contexts.

Intuitive and structured user interface

The scMOVIR platform offers a user-friendly and structured interface to facilitate efficient exploration of single-cell datasets related to viral infections and associated diseases. Users can access the data through different modules, including Browser, Search, Analysis, and Download (Fig. 2). Within the "Browse" module, users can access an interactive portal for data navigation (Fig. 2A). The left panel categorizes viruses and diseases, incorporating a hierarchical classification system. Viruses are organized according to the Baltimore classification, while diseases follow ICD-11 standards, including acute infection, chronic disease, cancer, immune dysfunction, vaccination, sequelae, and others. Selecting a subcategory displays card summaries on the right of relevant entries, each summarizing the scientific name, identifier, encompassed omics types, and experiment counts. Users can click the "Detail" button to navigate to the detail page, which presents the relationship between the selected virus and associated diseases, tissue involvements, and all relevant datasets.

scMOVIR offers a search module with autocompletion, enabling queries across various omics datasets based on "virus name," "disease name," and "tissue name" (Fig. 2B). All search terms are standardized and linked to authoritative databases to ensure consistency. In the context of single-cell transcriptomics, datasets derived from *in vitro* viral infections can also be retrieved using viral strain names and experimental models (e.g. cell lines, organoids). For proteomics data, queries can be performed based on specific protein markers. Search results are presented in both list and table formats, summarizing key information for each single-cell experiment. Users can access detailed views of individual experiments, supporting in-depth analysis and data download.

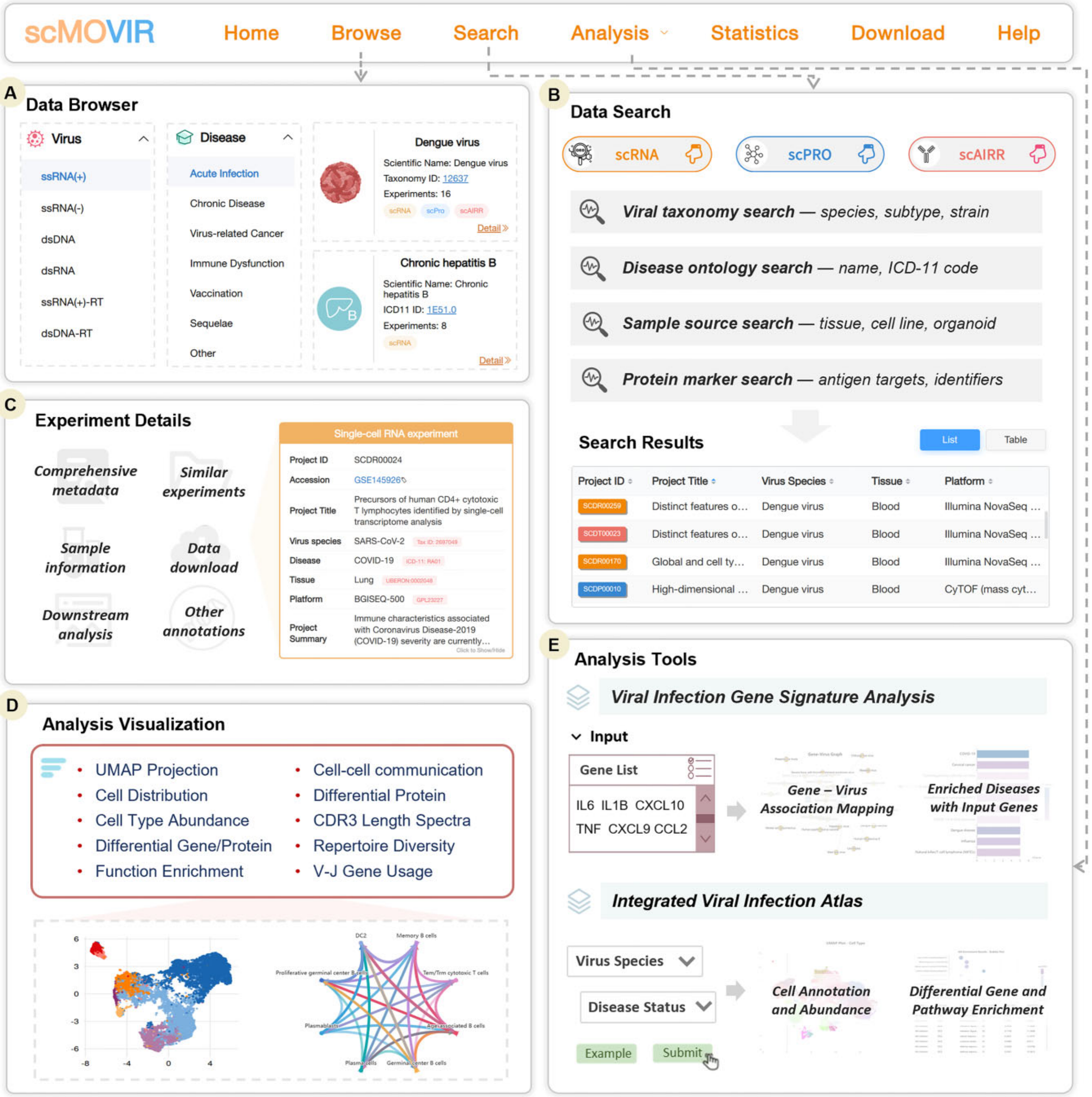
Comprehensive experimental metadata and data accessibility

Comprehensive and standardized metadata are critical for ensuring transparency, reproducibility, and cross-study integration in large-scale single-cell research. Therefore, scMOVIR systematically curates all experiments, providing detailed information, including experimental metadata, sample list of characteristics, similar experiments, and downloadable files (Fig. 2C). The upper section of each experiment page presents metadata including project identifiers, study titles, and reference links to original publications (Fig. 3A). Project descriptions and experimental design details, such as study objectives, grouping strategies, library preparation (e.g. 10× Chromium 3′, 10× Chromium 5′, SMART-seq2), sequencing platforms (e.g. Illumina NovaSeq 6000), number of samples, and processed single-cells, are fully documented. Standardized biological context information is provided, such as viral species/subtypes, infection status, disease classification, and tissue origin. Furthermore, the complete data processing pipeline is also recorded, specifying QC procedures, normalization methods, batch correction strategies, clustering algorithms, and all software versions and parameter settings. Each single-cell experiment is provided with a detailed sample characteristics list containing sample identifier, title, source tissue, patient information, and experimental variables (e.g. infection status), along with technical batch annotations to support robust stratification and reproducibility (Fig. 3B). For proteomic datasets, antibody panel specifications, including target antigens, labeling strategies (fluorescent or isotopic), and detection modalities, are fully documented (Fig. 4A). Immune receptor datasets are annotated with paired TCR/BCR chain sequences, clonotype assignments, CDR3 characteristics, and VJ gene usage profiles, enabling precise delineation of antigen-specific immune repertoires (Fig. 4B).

Furthermore, scMOVIR establishes systematic links between related experiments associated with the same viral infections, enabling comprehensive navigation and interrogation of infection-specific data within a unified metadata framework. For each experiment, the database provides downloadable files, including standardized annotations for cell types, proteomic profiles, and immune receptor repertoires, supporting cell classification, immune profiling, and multi-omics integration. Analytical results, including differential expression, cell-type abundance, and functional enrichment, are also made available to facilitate further data mining and validation.

Interactive visualization of cellular and molecular features

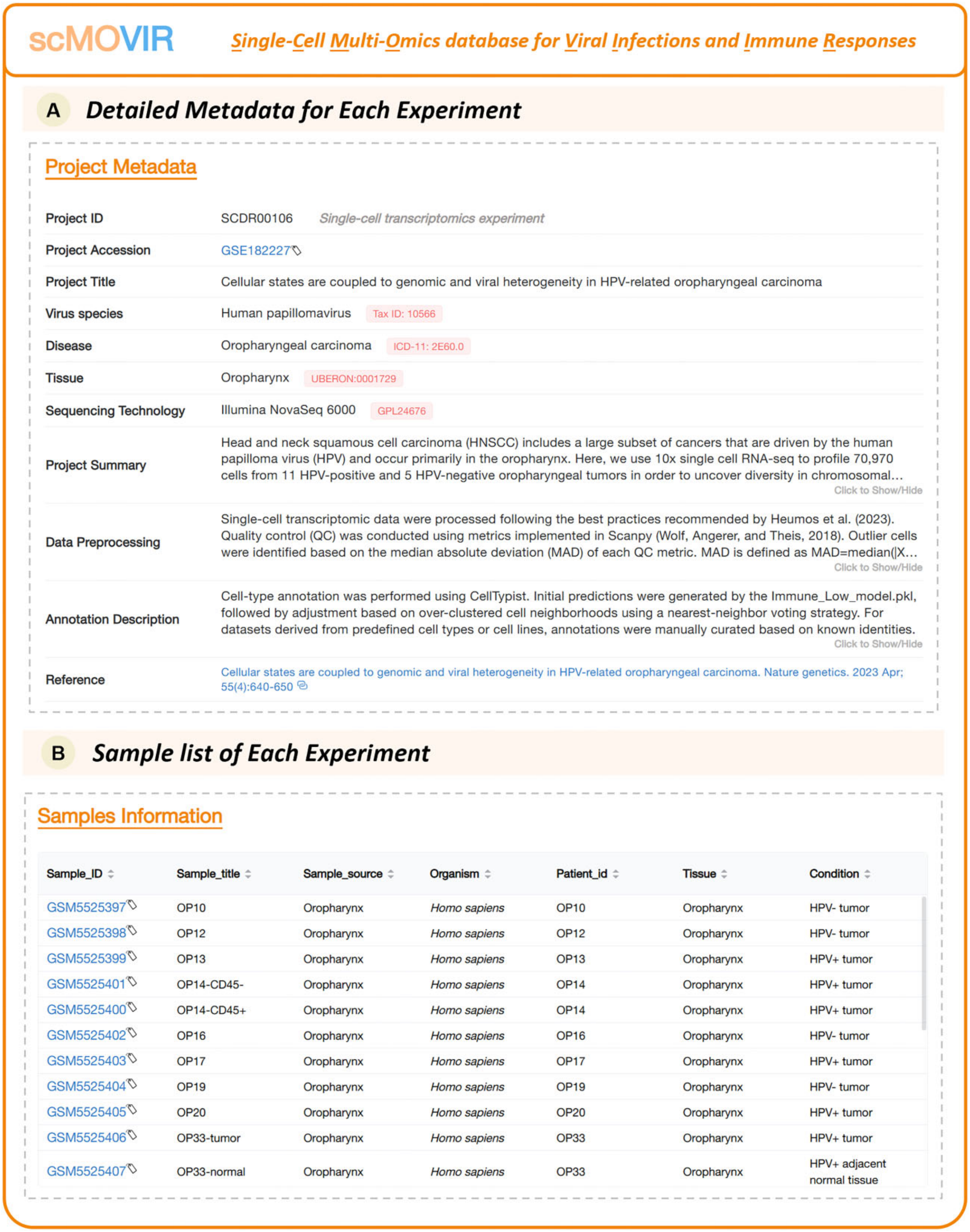
scMOVIR provides a variety of interactive visualizations to facilitate in-depth exploration of cellular heterogeneity, molecular signatures, intercellular communication, and immune repertoire diversity in the context of viral infection (Figs 1B and 2D). For single-cell transcriptomics, annotated cell populations are visualized using UMAP projections based on clustered gene expression profiles, which intuitively represent the complex spatial relationships between clusters and cell types. For datasets exceeding 5000 cells, a downsampled subset is displayed to ensure efficient web rendering while preserving the overall data structure and cell type proportions. Pie charts summarize the relative abundance of different cell populations within each reference dataset, offering an overview of cellular composition. Boxplots combined with statistical tables enable quantitative comparisons of cell population distributions across experimental groups, facilitating the identification of infection- or phenotype-specific cellular remodeling. The heatmaps display the most significantly differentially expressed genes across cell types, providing an intuitive visualization of expression patterns and cell type-specific features. Differential gene expression analysis identifies genes significantly up- or downregulated in defined cell types and experimental conditions, with functional interpretation supported by enrichment analysis using GO, KEGG, and Reactome. Enrichment results are presented as bubble plots integrating gene counts and statistical significance. Intercellular communica-



Project ID	Project Title	Virus Species	Tissue	Platform
SCDR00259	Distinct features o…	Dengue virus	Blood	Illumina NovaSeq …
SCDT00023	Distinct features o…	Dengue virus	Blood	Illumina NovaSeq …
SCDR00170	Global and cell ty…	Dengue virus	Blood	Illumina NovaSeq …
SCDP00010	High-dimensional …	Dengue virus	Blood	CyTOF (mass cyt…

Project ID	SCDR00024
Accession	GSE145926
Project Title	Precursors of human CD4+ cytotoxic T lymphocytes identified by single-cell transcriptome analysis
Virus species	SARS-CoV-2 Tax ID: 2697049
Disease	COVID-19 ICD-11: RA01
Tissue	Lung UBERON:0002048
Platform	BGISEQ-500 GPL23227
Project Summary	Immune characteristics associated with Coronavirus Disease-2019 (COVID-19) severity are currently…

Figure 2. Main modules and functions of scMOVIR. (**A**) The Browse page provides an interactive portal to explore viruses and diseases, organized by the Baltimore classification and the ICD-11 system, respectively. Users can select categories to access summary cards with key information and navigate to detailed pages that present virus–disease relationships and associated datasets. (**B**) The Search page supports keyword queries across multiple omics datasets, including virus names, diseases, tissues, viral strains, experimental models, and specific proteins for proteomics data. Search results are displayed in both list and table formats, summarizing key information of single-cell experiments. (**C**) The experiment details page provides comprehensive metadata, including experimental design, sample and batch information, related experiments and datasets, analytical outcomes, interactive visualizations, downloadable files, and others. (**D**) The analysis and visualization module comprises a suite of interactive, high-resolution visualizations tailored to each omics layer. All plots are dynamically linked to metadata and results, enabling multi-scale interrogation of viral infection-associated cell states, molecular profiles, and intercellular communication. (**E**) The database incorporates two analytical tools: the Gene Signature Analysis module, which evaluates enrichment of user-defined gene sets in viral infection-related conditions, and the Integrated Viral Infection Atlas module, which enables integrative analysis of single-cell transcriptomic landscapes across various viral infections.



Single-Cell Multi-Omics database for Viral Infections and Immune Responses

Detailed Metadata for Each Experiment

Project Metadata

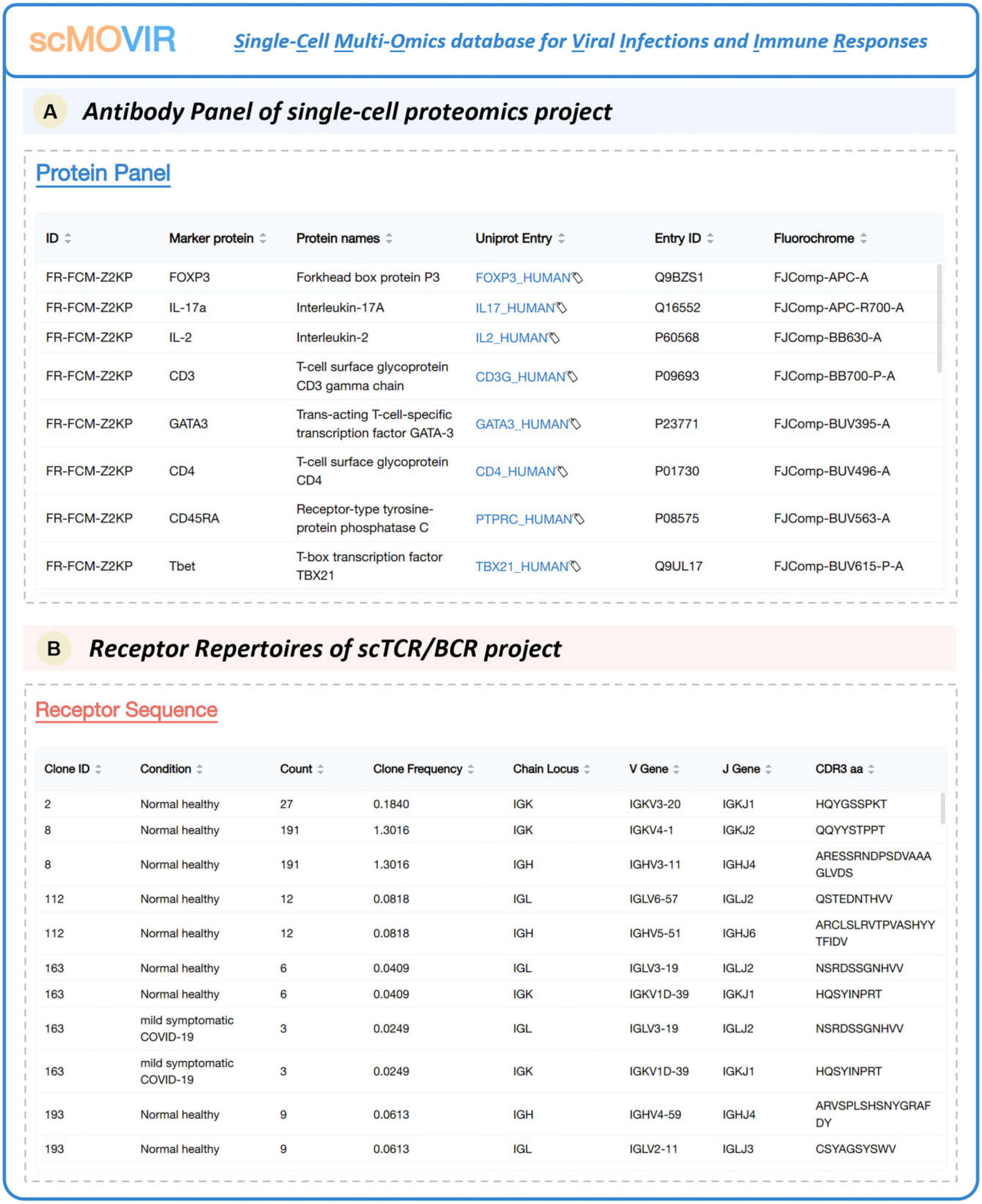
Project ID	SCDR00106 *Single-cell transcriptomics experiment*
Project Accession	GSE182227
Project Title	Cellular states are coupled to genomic and viral heterogeneity in HPV-related oropharyngeal carcinoma
Virus species	Human papillomavirus Tax ID: 10566
Disease	Oropharyngeal carcinoma ICD-11: 2E60.0
Tissue	Oropharynx UBERON:0001729
Sequencing Technology	Illumina NovaSeq 6000 GPL24676
Project Summary	Head and neck squamous cell carcinoma (HNSCC) includes a large subset of cancers that are driven by the human papilloma virus (HPV) and occur primarily in the oropharynx. Here, we use 10x single cell RNA-seq to profile 70,970 cells from 11 HPV-positive and 5 HPV-negative oropharyngeal tumors in order to uncover diversity in chromosomal… Click to Show/Hide
Data Preprocessing	Single-cell transcriptomic data were processed following the best practices recommended by Heumos et al. (2023). Quality control (QC) was conducted using metrics implemented in Scanpy (Wolf, Angerer, and Theis, 2018). Outlier cells were identified based on the median absolute deviation (MAD) of each QC metric. MAD is defined as MAD=median(|X… Click to Show/Hide
Annotation Description	Cell-type annotation was performed using CellTypist. Initial predictions were generated by the Immune_Low_model.pkl, followed by adjustment based on over-clustered cell neighborhoods using a nearest-neighbor voting strategy. For datasets derived from predefined cell types or cell lines, annotations were manually curated based on known identities. Click to Show/Hide
Reference	Cellular states are coupled to genomic and viral heterogeneity in HPV-related oropharyngeal carcinoma. Nature genetics. 2023 Apr; 55(4):640-650

Sample list of Each Experiment

Samples Information

Sample_ID	Sample_title	Sample_source	Organism	Patient_id	Tissue	Condition
GSM5525397	OP10	Oropharynx	*Homo sapiens*	OP10	Oropharynx	HPV- tumor
GSM5525398	OP12	Oropharynx	*Homo sapiens*	OP12	Oropharynx	HPV- tumor
GSM5525399	OP13	Oropharynx	*Homo sapiens*	OP13	Oropharynx	HPV+ tumor
GSM5525401	OP14-CD45-	Oropharynx	*Homo sapiens*	OP14	Oropharynx	HPV+ tumor
GSM5525400	OP14-CD45+	Oropharynx	*Homo sapiens*	OP14	Oropharynx	HPV+ tumor
GSM5525402	OP16	Oropharynx	*Homo sapiens*	OP16	Oropharynx	HPV- tumor
GSM5525403	OP17	Oropharynx	*Homo sapiens*	OP17	Oropharynx	HPV+ tumor
GSM5525404	OP19	Oropharynx	*Homo sapiens*	OP19	Oropharynx	HPV- tumor
GSM5525405	OP20	Oropharynx	*Homo sapiens*	OP20	Oropharynx	HPV+ tumor
GSM5525406	OP33-tumor	Oropharynx	*Homo sapiens*	OP33	Oropharynx	HPV+ tumor
GSM5525407	OP33-normal	Oropharynx	*Homo sapiens*	OP33	Oropharynx	HPV+ adjacent normal tissue

Figure 3. Representative scMOVIR detail page illustrating general information for a single-cell experimental dataset. (**A**) The upper section of each project page displays metadata, including project identifier, project title, virus type, disease classification, tissue origin, study description, experimental design details, data processing, and reference to the original publication. (**B**) The sample list displays sample identifiers, sample titles, source tissues, patient information, experimental variables, and batch information, supporting reliable stratified analysis and ensuring experimental reproducibility.



scMOVIR
Single-Cell Multi-Omics database for Viral Infections and Immune Responses
A
B

Antibody Panel of single-cell proteomics project

Protein Panel

ID	Marker protein	Protein names	Uniprot Entry	Entry ID	Fluorochrome
FR-FCM-Z2KP	FOXP3	Forkhead box protein P3	FOXP3_HUMAN	Q9BZS1	FJComp-APC-A
FR-FCM-Z2KP	IL-17a	Interleukin-17A	IL17_HUMAN	Q16552	FJComp-APC-R700-A
FR-FCM-Z2KP	IL-2	Interleukin-2	IL2_HUMAN	P60568	FJComp-BB630-A
FR-FCM-Z2KP	CD3	T-cell surface glycoprotein CD3 gamma chain	CD3G_HUMAN	P09693	FJComp-BB700-P-A
FR-FCM-Z2KP	GATA3	Trans-acting T-cell-specific transcription factor GATA-3	GATA3_HUMAN	P23771	FJComp-BUV395-A
FR-FCM-Z2KP	CD4	T-cell surface glycoprotein CD4	CD4_HUMAN	P01730	FJComp-BUV496-A
FR-FCM-Z2KP	CD45RA	Receptor-type tyrosine-protein phosphatase C	PTPRC_HUMAN	P08575	FJComp-BUV563-A
FR-FCM-Z2KP	Tbet	T-box transcription factor TBX21	TBX21_HUMAN	Q9UL17	FJComp-BUV615-P-A

Receptor Repertoires of scTCR/BCR project

Receptor Sequence

Clone ID	Condition	Count	Clone Frequency	Chain Locus	V Gene	J Gene	CDR3 aa
2	Normal healthy	27	0.1840	IGK	IGKV3-20	IGKJ1	HQYGSSPKT
8	Normal healthy	191	1.3016	IGK	IGKV4-1	IGKJ2	QQYYSTPPT
8	Normal healthy	191	1.3016	IGH	IGHV3-11	IGHJ4	ARESSRNDPSDVAAAGLVDS
112	Normal healthy	12	0.0818	IGL	IGLV6-57	IGLJ2	QSTEDNTHVV
112	Normal healthy	12	0.0818	IGH	IGHV5-51	IGHJ6	ARCLSLRVTPVASHYYTFIDV
163	Normal healthy	6	0.0409	IGL	IGLV3-19	IGLJ2	NSRDSSGNHVV
163	Normal healthy	6	0.0409	IGK	IGKV1D-39	IGKJ1	HQSYINPRT
163	mild symptomatic COVID-19	3	0.0249	IGL	IGLV3-19	IGLJ2	NSRDSSGNHVV
163	mild symptomatic COVID-19	3	0.0249	IGK	IGKV1D-39	IGKJ1	HQSYINPRT
193	Normal healthy	9	0.0613	IGH	IGHV4-59	IGHJ4	ARVSPLSHSNYGRAFDY
193	Normal healthy	9	0.0613	IGL	IGLV2-11	IGLJ3	CSYAGSYSWV

Figure 4. Representative scMOVIR page describing the protein panel and immune receptor information. (**A**) Proteomics datasets provide detailed antibody panel specifications, encompassing target antigens, labeling strategies (fluorescent or isotopic), and detection modalities. (**B**) Immune receptor datasets are annotated with paired TCR/BCR chain sequences, clonotype assignments, VJ gene usage, and CDR3 sequences, enabling precise characterization of antigen-specific immune repertoires.

tion is inferred through ligand–receptor interaction networks, visualized as circular plots of intracellular and intercellular signaling pathways, allowing users to interrogate major signal sources and targets, communication probabilities, and top-ranked ligand–receptor pairs that reflect infection-induced immune regulatory circuits. Differential signaling results are further highlighted with a scatter plot, which depicts the most significant interactions between cell types and illustrates both the magnitude and specificity of cell–cell communication changes across disease groups.

For single-cell proteomics, differential protein analysis is visualized using boxplots and numerical summaries for specific markers, capturing infection-associated changes in protein abundance. For immune receptor analysis, the normalized Shannon entropy index quantifies clonal diversity, repertoire evenness, and oligoclonal expansion. Histograms of CDR3 length distributions reveal clonal expansion patterns and selection pressures, reflecting structural biases in antigen recognition. Heatmaps depict V(D)J gene rearrangements and usage preferences, highlighting remodeling of antigen-specific immune receptors at the gene segment level. Collectively, these interactive visualizations enable comprehensive characterization of cellular states, molecular remodeling, intercellular signaling, and adaptive immune receptor dynamics during viral infection.

Integrated analytical functions in scMOVIR

On the analysis page, scMOVIR offers two analytical modules: Viral Infection Gene Signature Analysis and the Integrated Viral Infection Atlas, designed to facilitate systematic investigation of host–virus interactions at single-cell resolution (Fig. 2E). The first tool assesses whether user-defined genes are associated with viral infection-related diseases or phenotypes and systematically evaluates their enrichment patterns. For example (Fig. 5A), upon submission of a gene list comprising inflammatory cytokines and chemokines (e.g. IL6, IL1B, TNF, CXCL10, CXCL9, and CCL2), the system calculates enrichment significance across viral infection-related datasets using Fisher's exact test. The outputs include a gene–virus association map illustrating differential expression of the submitted genes across viral infection contexts, an enrichment overview across disease states highlighting infection-specific patterns, and differential expression statistics presented as boxplots with accompanying tables. Collectively, these results enable identification of infection-associated gene signatures and provide a robust reference for cross-dataset comparisons and functional validation.

The second module integrates large-scale, patient-derived viral infection scRNA-seq datasets to characterize immune cell state remodeling and transcriptional reprogramming during infection and disease progression. As an example, for HIV infection (Fig. 5B), scMOVIR aggregates all HIV-related scRNA-seq datasets, harmonizes disease state annotations across studies, and employs the probabilistic scVI model to construct an integrated reference anchored on healthy controls. This analysis generates a comprehensive patient sample list (131 samples from 12 independent studies) and provides harmonized cell type annotations spanning immune and non-immune populations. Based on these annotations, cell composition profiling quantifies the relative abundance of each cell type, revealing dynamic shifts in immune cell proportions during infection, with dominant populations including Tcm/Naïve helper T cells, Tem/Temra cytotoxic T cells, Tem/Effector helper T cells, and classical monocytes. At the molecular level, differential gene expression analysis delineates transcriptional programs significantly up- or downregulated in specific cell types during infection, uncovering HIV-induced immune perturbations and transcriptional remodeling. Functional pathway enrichment analysis further contextualizes these changes within biological processes and signaling pathways derived from GO, KEGG, and Reactome, providing mechanistic insights into infection-driven immune responses and inter-patient molecular heterogeneity. Together, this module enables high-resolution single-cell characterization of virus–host interactions and provides an extensible, clinically interpretable platform to support mechanistic and translational research in viral immunology.

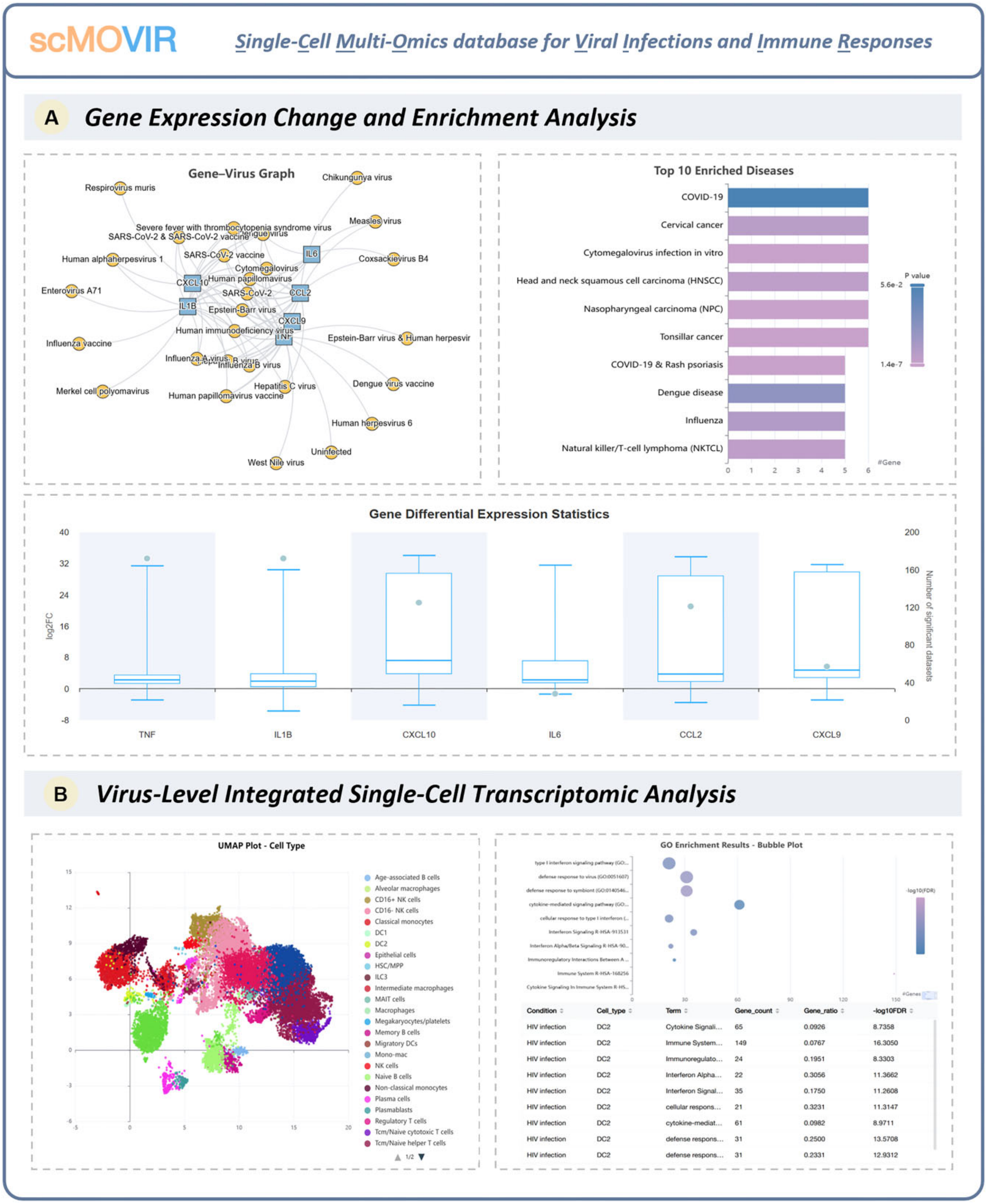
Standardization, access, and download of data

The integration of single-cell multi-omics datasets from heterogeneous sources is often challenged by inconsistencies in experimental metadata, annotation standards, and the nomenclature of diseases or viral taxa, which can compromise data comparability and downstream analyses. To address these issues, scMOVIR implements a rigorous standardization pipeline grounded in authoritative biomedical ontologies. All viruses, diseases, tissues, and proteins were standardized using the authoritative databases such as NCBI Taxonomy [36], WHO ICD-11 [61], Uberon [62], and UniProt [63]. Based on this standardization, scMOVIR offers a user-friendly interface for efficient data exploration and retrieval. Users can explore the database either by performing targeted queries using keyword-based searches and structured filters or by navigating through the Browse interface, which organizes viruses and diseases in a hierarchical manner. Flexible download options include access to individual files as well as batch archival downloads using built-in platform tools. All database functionalities are freely accessible without login at https://pgx.zju.edu.cn/scmovir, facilitating seamless access for the research community.

Conclusion and perspectives

This study presents scMOVIR, a comprehensive single-cell multi-omics database dedicated to human viral infection and immune response. Although numerous single-cell datasets have been generated, there remains a lack of integrated multi-omics data and broad coverage across different virus types, infection states, and tissue sources. scMOVIR systematically integrates large-scale, high-quality single-cell datasets spanning transcriptomics, proteomics, and immune receptor repertoires, encompassing a wide spectrum of viral pathogens, disease states, and tissue sources. All data undergo rigorous standardization, annotation, and QC to ensure interoperability and consistency. The platform provides a comprehensive suite of interactive visualization and analytical tools, coupled with a user-friendly interface, to facilitate detailed exploration of multi-omics features and cellular and molecular heterogeneity in the context of viral infection.

scMOVIR will be continuously updated to incorporate emerging single-cell datasets and novel multi-omics modalities. Future expansions will encompass additional viral species, diverse clinical sample cohorts, and advanced data types such as single-cell ATAC-seq and spatial multi-omics.


scMOVIR
Single-Cell Multi-Omics database for Viral Infections and Immune Responses
A
Gene Expression Change and Enrichment Analysis
Gene–Virus Graph
Respirovirus muris
Chikungunya virus
Severe fever with thrombocytopenia syndrome virus
SARS-CoV-2 & SARS-CoV-2 vaccine
Dengue virus
Measles virus
Human alphaherpesvirus 1
SARS-CoV-2 vaccine
IL6
Coxsackievirus B4
Cytomegalovirus
CXCL10
Human papillomavirus
Enterovirus A71
SARS-CoV-2
CCL2
IL1B
Epstein-Barr virus
CXCL9
Human immunodeficiency virus
TNF
Epstein-Barr virus & Human herpesvir
Influenza vaccine
Influenza A virus
Hepatitis B virus
Influenza B virus
Hepatitis C virus
Dengue virus vaccine
Merkel cell polyomavirus
Human papillomavirus vaccine
Human herpesvirus 6
Uninfected
West Nile virus
Top 10 Enriched Diseases
COVID-19
Cervical cancer
Cytomegalovirus infection in vitro
Head and neck squamous cell carcinoma (HNSCC)
Nasopharyngeal carcinoma (NPC)
Tonsillar cancer
COVID-19 & Rash psoriasis
Dengue disease
Influenza
Natural killer/T-cell lymphoma (NKTCL)
P value
5.6e-2
1.4e-7
#Gene
0
1
2
3
4
5
6
Gene Differential Expression Statistics
log2FC
40
32
24
16
8
0
-8
Number of significant datasets
200
160
120
80
40
0
TNF
IL1B
CXCL10
IL6
CCL2
CXCL9
B
Virus-Level Integrated Single-Cell Transcriptomic Analysis
UMAP Plot - Cell Type
15
12
9
6
3
0
-3
-6
-5
0
5
10
15
20
Age-associated B cells
Alveolar macrophages
CD16+ NK cells
CD16- NK cells
Classical monocytes
DC1
DC2
Epithelial cells
HSC/MPP
ILC3
Intermediate macrophages
MAIT cells
Macrophages
Megakaryocytes/platelets
Memory B cells
Migratory DCs
Mono-mac
NK cells
Naive B cells
Non-classical monocytes
Plasma cells
Plasmablasts
Regulatory T cells
Tcm/Naive cytotoxic T cells
Tcm/Naive helper T cells
▲ 1/2 ▼
GO Enrichment Results - Bubble Plot
type I interferon signaling pathway (GO:...
defense response to virus (GO:0051607)
defense response to symbiont (GO:0140546...
cytokine-mediated signaling pathway (GO:...
cellular response to type I interferon (...
Interferon Signaling R-HSA-913531
Interferon Alpha/Beta Signaling R-HSA-90...
Immunoregulatory Interactions Between A ...
Immune System R-HSA-168256
Cytokine Signaling In Immune System R-HS...
0
30
60
90
120
150
-log10(FDR)
#Genes

Condition ⇕	Cell_type ⇕	Term ⇕	Gene_count ⇕	Gene_ratio ⇕	-log10FDR ⇕
HIV infection	DC2	Cytokine Signali...	65	0.0926	8.7358
HIV infection	DC2	Immune System...	149	0.0767	16.3050
HIV infection	DC2	Immunoregulato...	24	0.1951	8.3303
HIV infection	DC2	Interferon Alpha...	22	0.3056	11.3662
HIV infection	DC2	Interferon Signal...	35	0.1750	11.2608
HIV infection	DC2	cellular respons...	21	0.3231	11.3147
HIV infection	DC2	cytokine-mediat...	61	0.0982	8.9711
HIV infection	DC2	defense respons...	31	0.2500	13.5708
HIV infection	DC2	defense respons...	31	0.2331	12.9312

Figure 5. Analysis tools available in scMOVIR. (**A**) Viral Infection Gene Signature Analysis assesses associations between user-defined gene sets and viral infection-related diseases or phenotypes. Given an input gene list (e.g. inflammatory cytokines IL6, IL1B, TNF, CXCL10, CXCL9, CCL2), enrichment analysis across viral datasets is performed using Fisher's exact test. The output comprises a gene–virus association map, an enrichment overview stratified by disease state, and differential expression statistics, thereby enabling the identification of infection-associated signatures and facilitating cross-dataset comparisons. (**B**) Integrated Viral Infection Atlas aggregates large-scale, patient-derived single-cell RNA sequencing datasets to delineate immune cell remodeling and transcriptional dynamics during viral infection. Using HIV as an example, disease annotations from all related experiments are harmonized, and the scVI model is applied to construct a reference framework anchored to healthy controls. The output provides cell type annotations, cell composition profiles, differential gene expression, and pathway enrichment analyses, offering mechanistic insights into specific virus–host interactions.

Furthermore, systematic integration of perturbation datasets collected before and after viral infection will establish a valuable resource for elucidating causal mechanisms of virus–host interactions and for training predictive computational models. With the rapid advancement of artificial intelligence, scMOVIR is also envisioned to interface with large language model-based frameworks, enabling the development of intelligent agents capable of automated data interpretation, hypothesis generation, and interactive knowledge retrieval. Collectively, scMOVIR provides a standardized, scalable, and openly accessible platform, offering critical support for systematic characterization of virus–host immune interactions at single-cell resolution.

Acknowledgements

We thank the Information Technology Center and State Key Lab of CAD&CG, the Innovation Institute for Artificial Intelligence in Medicine, Zhejiang University for the support of computing resources. We also express our gratitude to all data contributors for providing the metadata that form the basis of this study. Additionally, we would like to acknowledge the group members and the numerous users who reported bugs and provided constructive suggestions.

Author contributions: Xue Zhang (Data curation [lead], Formal analysis [lead], Investigation [lead], Methodology [equal], Visualization [lead], Writing—original draft [lead]), Shounan Yang (Data curation [supporting], Formal analysis [supporting], Investigation [supporting], Methodology [supporting], Visualization [supporting]), Xiaobin Xu (Data curation [supporting], Formal analysis [supporting], Investigation [supporting]), Yitao Lin (Data curation [supporting], Formal analysis [supporting], Methodology [supporting]), Huaicheng Sun (Investigation [supporting], Methodology [supporting]), Bangyu Zhu (Data curation [supporting], Investigation [supporting]), Wenyi Zhao (Investigation [supporting], Methodology [supporting]), Binbin Zhou (Methodology [supporting], Resources [supporting]), Yan Lou (Funding acquisition [supporting], Resources [supporting]), Xinyu Wang (Funding acquisition [supporting], Resources [supporting]), Shuqing Chen (Methodology [supporting], Resources [supporting]), Qiaojun He (Funding acquisition [equal], Resources [supporting]), Feng Zhu (Conceptualization [equal], Methodology [equal], Supervision [equal], Writing—review & editing [equal]), and Zhan Zhou (Conceptualization [equal], Funding acquisition [lead], Methodology [equal], Resources [lead], Supervision [lead], Writing—review & editing [lead]).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the Noncommunicable Chronic Diseases—National Science and Technology Major Project (2024ZD0524900), the Zhejiang Provincial Natural Science Foundation of China (LDT23H1901), the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2024C03003, 2025C02068), the Key Project under the Zhejiang Provincial and Ministry Co-construction Plan (WK-ZJ-2425), the National Natural Science Foundation of China (92369107, 32370712), the International Cooperation and Exchange Programs of the National Natural Science Foundation of China (W2412017, 2024VTJP1003), and the Fundamental Research Funds for the Central Universities (226-2025-00065). Funding to pay the Open Access publication charges for this article was provided by Noncommunicable Chronic Diseases—National Science and Technology Major Project (2024ZD0524900).

Data availability

All data can be viewed, accessed, and downloaded from the scMOVIR, which is freely accessible without any login requirements by all users at https://pgx.zju.edu.cn/scmovir.

References

1. Liang G, Bushman FD. The human virome: assembly, composition and host interactions. *Nat Rev Microbiol* 2021;19:514–27. https://doi.org/10.1038/s41579-021-00536-5
2. Koyuncu OO, Hogue IB, Enquist LW. Virus infections in the nervous system. *Cell Host Microbe* 2013;13:379–93. https://doi.org/10.1016/j.chom.2013.03.010
3. He M, He CQ, Ding NZ. Human viruses: an ever-increasing list. *Virology* 2025;604:110445. https://doi.org/10.1016/j.virol.2025.110445
4. Mesri EA, Feitelson MA, Munger K. Human viral oncogenesis: a cancer hallmarks analysis. *Cell Host Microbe* 2014;15:266–82. https://doi.org/10.1016/j.chom.2014.02.011
5. Bouvard V, Baan R, Straif K *et al.* A review of human carcinogens—Part B: biological agents. *Lancet Oncol* 2009;10:321–2. https://doi.org/10.1016/S1470-2045(09)70096-8
6. de Martel C, Georges D, Bray F *et al.* Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. *Lancet Glob Health* 2020;8:e180–90. https://doi.org/10.1016/S2214-109X(19)30488-7
7. Wei S, Evans PC, Strijdom H *et al.* HIV infection, antiretroviral therapy and vascular dysfunction: effects, mechanisms and treatments. *Pharmacol Res* 2025;217:107812. https://doi.org/10.1016/j.phrs.2025.107812
8. Shen C, Jiang X, Li M *et al.* Hepatitis virus and hepatocellular carcinoma: recent advances. *Cancers* 2023;15:533. https://doi.org/10.3390/cancers15020533
9. Perkins RB, Wentzensen N, Guido RS *et al.* Cervical cancer screening: a review. *JAMA* 2023;330:547–58. https://doi.org/10.1001/jama.2023.13174
10. Rahangdale L, Mungo C, O'Connor S *et al.* Human papillomavirus vaccination and cervical cancer risk. *BMJ* 2022;379:e070115. https://doi.org/10.1136/bmj-2022-070115
11. Sausen DG, Basith A, Muqeemuddin S. EBV and lymphomagenesis. *Cancers* 2023;15:2133. https://doi.org/10.3390/cancers15072133
12. Banerjee AK, Blanco MR, Bruce EA *et al.* SARS-CoV-2 disrupts splicing, translation, and protein trafficking to suppress host defenses. *Cell* 2020;183:1325–39.e21. https://doi.org/10.1016/j.cell.2020.10.004
13. V'kovski P, Kratzel A, Steiner S *et al.* Coronavirus biology and replication: implications for SARS-CoV-2. *Nat Rev Microbiol* 2021;19:155–70.
14. Gambadauro A, Galletta F, Li Pomi A *et al.* Immune response to respiratory viral infections. *Int J Mol Sci* 2024;25:6178. https://doi.org/10.3390/ijms25116178

15. Jia Z, Ren Z, Ye D *et al.* Immune-ageing evaluation of peripheral T and NK lymphocyte subsets in chinese healthy adults. *Phenomics* 2023;3:360–74. https://doi.org/10.1007/s43657-023-00106-0
16. Walters KA, Blatti CA, Zhu R *et al.* Nasal and systemic immune responses correlate with viral shedding after influenza challenge in people with complex preexisting immunity. *Sci Transl Med* 2025;17:eadt1452. https://doi.org/10.1126/scitranslmed.adt1452
17. Traxler P, Reichl S, Folkman L *et al.* Integrated time-series analysis and high-content CRISPR screening delineate the dynamics of macrophage immune regulation. *Cell Syst* 2025;16:101346. https://doi.org/10.1016/j.cels.2025.101346
18. Stubbington MJT, Rozenblatt-Rosen O, Regev A *et al.* Single-cell transcriptomics to explore the immune system in health and disease. *Science* 2017;358:58–63. https://doi.org/10.1126/science.aan6828
19. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol* 2018;18:35–45. https://doi.org/10.1038/nri.2017.76
20. Cui A, Huang T, Li S *et al.* Dictionary of immune responses to cytokines at single-cell resolution. *Nature* 2024;625:377–84. https://doi.org/10.1038/s41586-023-06816-9
21. Liu X, Zhang L, Li X *et al.* Single-cell multi-omics profiling uncovers the immune heterogeneity in HIV-infected immunological non-responders. *eBioMedicine* 2025;115:105667. https://doi.org/10.1016/j.ebiom.2025.105667
22. Zhang B, Upadhyay R, Hao Y *et al.* Multimodal single-cell datasets characterize antigen-specific CD8$^+$ T cells across SARS-CoV-2 vaccination and infection. *Nat Immunol* 2023;24:1725–34. https://doi.org/10.1038/s41590-023-01608-9
23. Lindeboom RGH, Worlock KB, Dratva LM *et al.* Human SARS-CoV-2 challenge uncovers local and systemic response dynamics. *Nature* 2024;631:189–98. https://doi.org/10.1038/s41586-024-07575-x
24. Chen S, Huang C, Liao G *et al.* Distinct single-cell immune ecosystems distinguish true and *de novo* HBV-related hepatocellular carcinoma recurrences. *Gut* 2023;72:1196–210. https://doi.org/10.1136/gutjnl-2022-328428
25. Satija N, Patel F, Schmidt G *et al.* Tracking HIV persistence across T cell lineages during early ART-treated HIV-1-infection using a reservoir-marking humanized mouse model. *Nat Commun* 2025;16:2233. https://doi.org/10.1038/s41467-025-57368-7
26. Qiu MZ, Wang C, Wu Z *et al.* Dynamic single-cell mapping unveils Epstein−Barr virus-imprinted T-cell exhaustion and on-treatment response. *Signal Transduct Target Ther* 2023;8:370. https://doi.org/10.1038/s41392-023-01622-1
27. Tang Z, Fan W, Li Q *et al.* MVIP: multi-omics portal of viral infection. *Nucleic Acids Res* 2021;50:D817–27. https://doi.org/10.1093/nar/gkab958
28. 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
29. Han Y, Wang Y, Dong X *et al.* TISCH2: expanded datasets and new tools for single-cell transcriptome analyses of the tumor microenvironment. *Nucleic Acids Res* 2022;51:D1425–31. https://doi.org/10.1093/nar/gkac959
30. Zhao T, Lyu S, Lu G *et al.* SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Res* 2021;49:D1413–9. https://doi.org/10.1093/nar/gkaa838
31. Li YY, Zhou LW, Qian FC *et al.* scImmOmics: a manually curated resource of single-cell multi-omics immune data. *Nucleic Acids Res* 2024;53:D1162–72. https://doi.org/10.1093/nar/gkae985
32. Zhang X, Wu J, Luo Y *et al.* CovEpiAb: a comprehensive database and analysis resource for immune epitopes and antibodies of human coronaviruses. *Brief Bioinform* 2024;25:bbae183. https://doi.org/10.1093/bib/bbae183
33. Wei M, Wu J, Bai S *et al.* TRAIT: a comprehensive database for T-cell receptor–antigen interactions. *Genomics Proteomics Bioinf* 2025;23:qzaf033. https://doi.org/10.1093/gpbjnl/qzaf033
34. Pickett BE, Sadat EL, Zhang Y *et al.* ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res* 2012;40:D593–8. https://doi.org/10.1093/nar/gkr859
35. Lefkowitz EJ, Dempsey DM, Hendrickson RC *et al.* Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res* 2018;46:D708–17. https://doi.org/10.1093/nar/gkx932
36. 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
37. Barrett T, Wilhite SE, Ledoux P *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 2013;41:D991–5. https://doi.org/10.1093/nar/gks1193
38. Athar A, Füllgrabe A, George N *et al.* ArrayExpress update—from bulk to single-cell expression data. *Nucleic Acids Res* 2019;47:D711–5. https://doi.org/10.1093/nar/gky964
39. Spidlen J, Breuer K, Rosenberg C *et al.* FlowRepository: a resource of annotated flow cytometry datasets associated with peer-reviewed publications. *Cytometry A* 2012;81:727–31. https://doi.org/10.1002/cyto.a.22106
40. Lian X, Zhang Y, Zhou Y *et al.* SingPro: a knowledge base providing single-cell proteomic data. *Nucleic Acids Res* 2023;52:D552–61. https://doi.org/10.1093/nar/gkad830
41. Wang F, Liu C, Li J *et al.* SPDB: a comprehensive resource and knowledgebase for proteomic data at the single-cell resolution. *Nucleic Acids Res* 2024;52:D562–71. https://doi.org/10.1093/nar/gkad1018
42. Ren L, Shi L, Zheng Y. Reference materials for improving reliability of multiomics profiling. *Phenomics* 2024;4:487–521. https://doi.org/10.1007/s43657-023-00153-7
43. 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
44. Heumos L, Schaar AC, Lance C *et al.* Best practices for single-cell analysis across modalities. *Nat Rev Genet* 2023;24:550–72. https://doi.org/10.1038/s41576-023-00586-w
45. Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst* 2019;8:281–91. https://doi.org/10.1016/j.cels.2018.11.005
46. 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
47. Sturm G, Szabo T, Fotakis G *et al.* Scirpy: a Scanpy extension for analyzing single-cell T-cell receptor-sequencing data. *Bioinformatics* 2020;36:4817–8. https://doi.org/10.1093/bioinformatics/btaa611
48. Domínguez Conde C, Xu C, Jarvis LB *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* 2022;376:eabl5197. https://doi.org/10.1126/science.abl5197
49. Hu C, Li T, Xu Y *et al.* CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res* 2023;51:D870–6. https://doi.org/10.1093/nar/gkac947
50. Cheng S, Li Z, Gao R *et al.* A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells. *Cell* 2021;184:792–809. https://doi.org/10.1016/j.cell.2021.01.010
51. Kock KH, Tan LM, Han KY *et al.* Asian diversity in human immune cells. *Cell* 2025;188:2288–306. https://doi.org/10.1016/j.cell.2025.02.017
52. Zhang L, Yu X, Zheng L *et al.* Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. *Nature* 2018;564:268–72. https://doi.org/10.1038/s41586-018-0694-x
53. Ontology Consortium G, Aleksander SA, Balhoff J *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* 2023;224:iyad031. https://doi.org/10.1093/genetics/iyad031
54. Ashburner M, Ball CA, Blake JA *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* 2000;25:25–9. https://doi.org/10.1038/75556

55. Milacic M, Beavers D, Conley P *et al.* The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res* 2024;52:D672–8. https://doi.org/10.1093/nar/gkad1025
56. Cai G, Zhao W, Zhou Z *et al.* MATTE: a pipeline of transcriptome module alignment for anti-noise phenotype-gene-related analysis. *Brief Bioinform* 2023;24:bbad207. https://doi.org/10.1093/bib/bbad207
57. Kanehisa M, Furumichi M, Sato Y *et al.* KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* 2021;49:D545–51. https://doi.org/10.1093/nar/gkaa970
58. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* 2023;39:btac757. https://doi.org/10.1093/bioinformatics/btac757
59. Efremova M, Vento-Tormo M, Teichmann SA *et al.* CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat Protoc* 2020;15:1484–506. https://doi.org/10.1038/s41596-020-0292-x
60. Li D, Mei H, Shen Y *et al.* ECharts: a declarative framework for rapid construction of web-based visualization. *Visual Inf* 2018;2:136–46. https://doi.org/10.1016/j.visinf.2018.04.011
61. The Lancet. ICD-11. *Lancet* 2019;393:2275. https://doi.org/10.1016/S0140-6736(19)31205-X
62. Mungall CJ, Torniai C, Gkoutos GV *et al.* Uberon, an integrative multi-species anatomy ontology. *Genome Biol* 2012;13:R5. https://doi.org/10.1186/gb-2012-13-1-r5
63. Ahmad S, Jose da Costa Gonzales L, Bowler-Barnett EH *et al.* The UniProt website API: facilitating programmatic access to protein knowledge. *Nucleic Acids Res* 2025;53:W547–53. https://doi.org/10.1093/nar/gkaf394

Received: August 15, 2025. **Revised:** October 9, 2025. **Accepted:** October 11, 2025