scientific reports

OPEN

MOI is a comprehensive database collecting processed multi-omics data associated with viral infection

Xuefei Guo[1,2]✉, Yang Zhao[1,2] & Fuping You[1]

Viral infections pose significant public health challenges, exemplified by the global impact of COVID-19 caused by SARS-CoV-2. Understanding the intricate molecular mechanisms governing virus-host interactions is pivotal for effective intervention strategies. Despite the burgeoning multi-omics data on viral infections, a centralized database elucidating host responses to viruses remains lacking. In response, we have developed a comprehensive database named 'MOI' (available at http://www.fynn-guo.cn/), specifically designed to aggregate processed Multi-Omics data related to viral Infections. This meticulously curated database serves as a valuable resource for conducting detailed investigations into virus-host interactions. Leveraging high-throughput sequencing data and metadata from PubMed and Gene Expression Omnibus (GEO), MOI comprises over 3200 viral-infected samples, encompassing human and murine infections. Standardized processing pipelines ensure data integrity, including bulk RNA sequencing (RNA-seq), single-cell RNA-seq (scRNA-seq), Chromatin Immunoprecipitation sequencing (ChIP-seq), and Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq). MOI offers user-friendly interfaces presenting comprehensive cell marker tables, gene expression data, and epigenetic landscape charts. Analytical tools for DNA sequence conversion, FPKM calculation, differential gene expression, and Gene Ontology (GO)/ Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment enhance data interpretation. Additionally, MOI provides 16 visualization plots for intuitive data exploration. In summary, MOI serves as a valuable repository for researchers investigating virus-host interactions. By centralizing and facilitating access to multi-omics data, MOI aims to advance our understanding of viral pathogenesis and expedite the development of therapeutic interventions.

Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has posed an ongoing global health threat since its emergence in late 2019, infecting over 6.3 million individuals worldwide as of July 15, 2022[1–3]. Understanding the intricate molecular mechanisms of SARS-CoV-2 infection is paramount for controlling its spread and managing severe cases. Host cells mount defensive responses upon viral invasion to combat the pathogen[4–7]. Recent advancements in next-generation sequencing technology have facilitated the exploration of host cell gene expression profiles, offering invaluable insights into virus-host interactions in vitro and in vivo[8–11].

The GEO database is a cornerstone resource for viral-related NGS data, housing extensive gene expression and genomics datasets. From this wealth of information, specialized databases have emerged to delve into specific aspects of viral biology. VISDB focuses on viral integration sites[12], while VirHost facilitates research on viral pathogenesis[13]. ViPR integrates genomic and proteomic data for comprehensive analysis[14], while PhEVER delves into viral infection pathology and evolution[15]. HPIDB catalogs interactions between host proteins and pathogens[16], while ViMIC highlights viral mutation accumulation[17]. ViralZone offers curated information on viral protein families[18], and PHI-base specializes in pathogen-host interactions[18]. MVIP integrates multi-omics data to investigate host responses to viruses, providing insights into viral pathogenesis and host-virus interactions[19]. However, a cavity looms large in databases tailored to sorting and cataloging gene expression profiles, epigenetic landscapes, and cell markers for scRNA-seq during viral infection.

ScRNA-seq is a powerful tool for dissecting gene expression dynamics at the single-cell level, revolutionizing research in oncology, immunology, neurobiology, and developmental biology[20–23]. However, analyzing scRNA-seq

[1]Institute of Systems Biomedicine, Department of Immunology, School of Basic Medical Sciences, Beijing Key Laboratory of Tumor Systems Biology, NHC Key Laboratory of Medical Immunology, Peking University Health Science Center, Beijing, China. [2]These authors contributed equally: Xuefei Guo and Yang Zhao. ✉email: guoxf2025@163.com

data involves intricate processes such as alignment, quality control, normalization, dimensionality reduction, clustering, and cell annotation, which are crucial for drawing accurate conclusions[24–29]. While software packages like SingleR and scMatch aid in cell annotation through machine learning algorithms trained on existing scRNA-seq data[30,31], there is a pressing need for comprehensive databases that streamline this process.

To address this gap, we introduce the MOI database (http://www.fynn-guo.cn/), an intuitive platform providing comprehensive expression profiles and abundant cell markers from multi-omics data under various virus infection conditions. MOI enables users to browse, retrieve seamlessly, and download information about viral infections, including cell markers, gene expression profiles, and epigenetic landscapes. Moreover, MOI integrates a suite of online tools for downstream multi-omics data analysis and visualization, such as FPKM calculation, detection of Differentially Expressed Genes (DEGs), Gene Ontology (GO) annotation, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment, and visualization through 16 plots. By offering these resources, MOI aims to alleviate the burdens associated with omics data analysis, ultimately facilitating more profound insights into molecular events during infection and empowering downstream applications. We anticipate that MOI will be a valuable resource for researchers engaged in virus-related multi-omics studies, fostering a deeper understanding of viral pathogenesis and developing innovative therapeutic strategies.

Data collection and database content

Data collection

We rigorously adhered to a series of standardized procedures to systematically collect virus-related multi-omics data, ensuring consistency and reliability throughout the data collection process[19,32]. Initially, we conducted a comprehensive search in the PubMed database, retrieving 2500 abstracts containing the keyword 'virus' as of May 2021. These candidate articles underwent meticulous filtration based on the availability of multi-omics data and the methodology utilized for library construction. Specifically, we selected articles employing high-throughput sequencing or wet experimental techniques such as bulk RNA-seq, single-cell RNA-seq (scRNA-seq), ATAC-seq, and ChIP-seq assays to extract validated information (Fig. 1).

The current database release comprises data extracted from 1757 articles related explicitly to NGS technology. Each article's full text underwent a thorough manual review by at least two independent researchers to

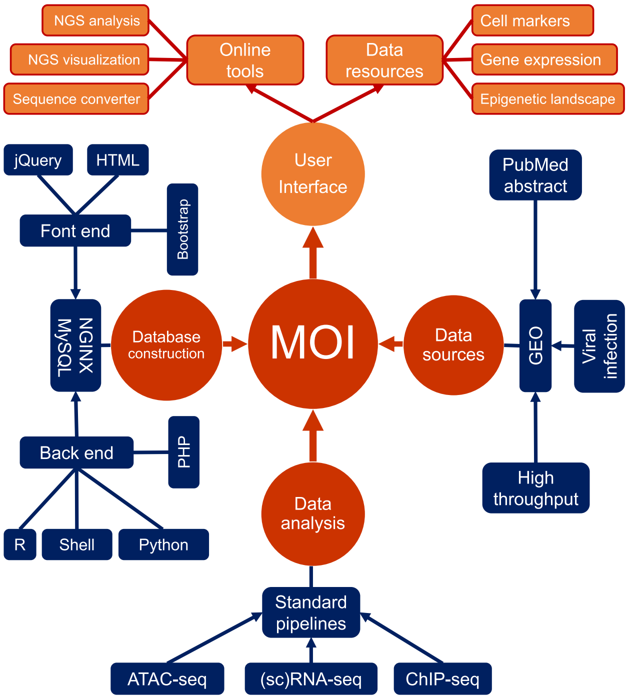


Figure 1. The design and structure of MOI. This schematic diagram comprehensively portrays MOI's design and structure, elucidating its fundamental constituents and organizational schema. Data Acquisition: The raw data utilized in this study were sourced from the Gene Expression Omnibus (GEO) database, adhering to stringent inclusion criteria to ensure data quality and relevance. Data Integration: At the core of MOI lies a robust data integration pipeline that harmonizes disparate multi-omics datasets from diverse origins into a cohesive and standardized repository. Database Construction: The integrated multi-omics data are housed within a meticulously structured database architecture developed with MySQL, PHP, NGINX, Bootstrap, and so on, facilitating efficient data management, retrieval, and manipulation. User Interface: MOI boasts an intuitive user interface designed to streamline user interaction and enable seamless navigation across its functionalities. The user interface provides intuitive access to many tools and features, including data querying, analysis tools, and visualization modules.

extract comprehensive information concerning virus infection. Each database entry includes essential metadata, including species, cell line, infection duration, library type, reference genome, PubMed ID of the publication, and details regarding the experimental or computational methodology employed for detection. This stringent methodology ensures the inclusion of high-quality, curated data within the database, facilitating robust analysis and interpretation for researchers in viral infections and multi-omics research.

Data processing and annotation

The sequencing data generated in FASTQ file format was filtered by FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (V0.11.9) and Trim-Galore(https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) (V0.6.4) software for quality control. The parameters of Trim-Galore are as follows: `trim_galore --gzip --trim-n --phred33 -j 7 --paired ${var}_1.fq.gz ${var}_2.fq.gz -o $wrk_dir/clean_result/`. Then, the mouse genome version mm10 was used as a reference genome to align clean data with Subread (V2.0.0) software[33]. The gene count matrix was acquired using the featureCounts (V2.0.0) program[34]. The specific arguments for Subread and featureCounts are as shown below: `subread-align -i $idx_dir -r $cle_dir/${var}_1_val_1.fq.gz -R $cle_dir/${var}_2_val_2.fq.gz -o $aln_dir/${var}.bam -T 30 -t 0`; `featureCounts -p -t exon -g gene_id -a $gtf_dir -o $cnt_dir/count_refGene $aln_dir/*.bam -T 29`. Then, the gene count data was normalized using the Fragments Per Kilobase Million (FPKM) formula[35]. R package DESeq2 (V1.38.3) was used to identify the DEGs (with the criteria: abs(log2(FoldChange)) > 2 and p-value < 0.05)[36]. The GO annotation and KEGG pathway enrichment analysis were conducted by the DAVID (https://david.ncifcrf.gov/summary.jsp) database (V6.8)[37–39].

Single-cell sequencing data were aligned and quantified with Cell Ranger (v6.0.1) against the GRCh38 human and mm10 reference genome downloaded from 10× Genomics official website (https://www.10xgenomics.com/). The parameters of Cell Ranger are as follows: `cellranger count --id=$var --transcriptome = $idx_dir --fastqs=$raw_dir --sample=$var --expect-cells = 1000 --localcores 28 --localmem 88`. Preliminary counts were then used for downstream analysis. Quality control was applied to cells based on three metrics: the total UMI counts, the number of detected genes, and the proportion of mitochondrial gene count per cell. Specifically, cells with less than 1000 UMI counts and 500 detected genes and cells with more than 10% mitochondrial gene counts were filtered. To integrate cells into a shared space from different datasets for unsupervised clustering, we used the harmony algorithm (V1.2.0) to do batch effect correction[26]. We performed variable gene selection separately for each sample to detect the most variable genes used for the harmony algorithm. Then, we use the R package Seurat (V5.0.3) to conduct the normalization, dimensional reduction clustering, etc. (https://satijalab.org/seurat/index.html)[25]. The annotation of the prepared single cell was completed by one shared R package named SingleR (V2.0.0) (https://github.com/dviraran/SingleR)[29]. The DEGs (with the criteria: abs(log2(FoldChange)) > 0.5 and p-value < 0.05) of scRNA-seq were detected by MAST (V1.24.1) (https://github.com/RGLab/MAST)[40].

As for the ChIP-seq and ATAC-seq data downloaded from the available database, we first conducted the quality control analysis for the raw data in fastq format with Trim-Galore (V0.6.4). Then, the filtered reads were aligned and quantified to their corresponding genome, such as hg38 or mm10, using aligner bowtie2 (V2.3.5.1)[41,42], which would output the SAM or BAM files. The parameters of bowtie2 are as follows: `bowtie2 --very-sensitive -X 2000 -x $Bowtie2Index -1 $cln_res/${sample}_1_val_1.fq.gz -2 $cln_res/${sample}_2_val_2.fq.gz -p $PPN | samtools view -buSh -@ $PPN | samtools sort -@ $PPN -O BAM -o $aln_res/${sample}.sorted.bam`. Then, the samtools (V1.10) were used to sort and index the SAM/BAM files with the following parameter: `samtools index -@ $PPN $aln_res/${sample}.sorted.bam`[43]. Peak calling was completed with MACS3 (V3.0.0a5)[44] with the default parameters, which produced the peak files with bed format for each sample. In the downstream of ATAC-seq and ChIP-seq, we adopted different strategies. The ATAC-seq data were first merged by the bedtools (V2.31.1) merge tool for all samples[45]. Then, they were normalized by the bamCoverage (V3.3.2) tool in deepTools (V3.3.2) suites[46], using the parameter `bamCoverage --bam $var -o ${var%.*}.bw --binSize 100 --normalizeUsing RPKM --effectiveGenomeSize 2864785220 --ignoreForNormalization chrM –extendReads`. Then, the normalized data could be used for differential analysis with DESeq2 (V1.38.3)[36]. For ChIP-seq data: The reads that fell in the peak gap were normalized by RPKM first using bamCoverage (V3.3.2) with the above parameters. The treatment samples were compared to their Input data using the bamCompare (V3.3.2) tool with the parameter: `bamCompare -b1 idx-MUTMSCLACH.bam -b2 idx-MUTMSCLAIN.bam -o bwResIpInput/MUTIPratioInput --binSize 145 --normalizeUsing RPKM --effectiveGenomeSize 2864785220 --ignoreForNormalization chrM --extendReads -p 22 --scaleFactorsMethod None`. Then, the R package ChIPseeker (V1.34.1) was used to annotate the identified peak files[47]. Last, the Integrative Genomics Viewer (IGV) (V2.17.4) was used to visualize these peaks from ChIP-seq or ATAC-seq[48].

Database statistics

The data sources of the MOI database primarily originate from original data from the GEO database and published literature. Following retrieval, the data are centralized onto a local server where a standardized analysis and annotation process is employed. Presently, the database comprises 1578 bulk RNA-seq samples, 787 single-cell RNA-seq (scRNA-seq) samples, 285 ChIP-seq samples, and 484 ATAC-seq samples (Fig. 2a).

The compiled data are predominantly categorized into three main classes: (1) specific cell markers of mouse and human cells across various tissues and experimental treatments; (2) gene expression profiles of cells post-viral infection at diverse time points; and (3) dynamics of cell transcription factors and epigenetic landscapes of human and mouse cells. As of December 2022, MOI encompasses 1597 human cell marker records, 897 mouse cell marker records, and 524 cell marker annotations derived from single-cell sequencing (Fig. 2b). The database includes 1638 human and 1543 mouse transcription factors (Supplementary Fig. 1b and 1d). Furthermore, MOI incorporates a comprehensive array of innate immune cells, commonly utilized laboratory cell lines, and gene expression profiles spanning diverse time points and viral infections (Supplementary Fig. 1a and 1c). Moving

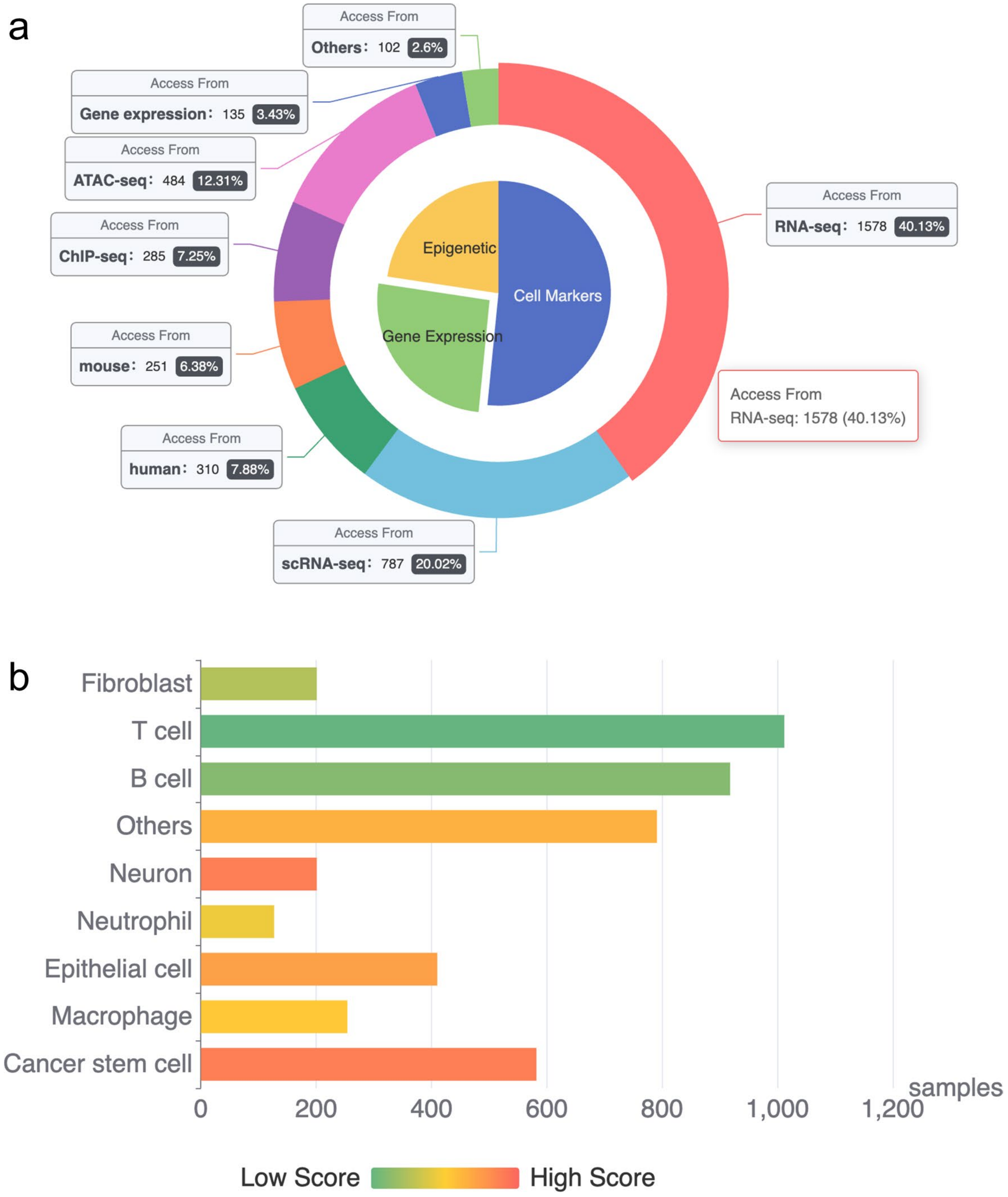


Figure 2. The composition and statistics of MOI. (**a**) Comprehensive statistics of multi-omics samples, which delineates detailed statistics about the multi-omics samples employed within the MOI study. The encompassed methodologies include bulk RNA-seq, single-cell RNA-seq (scRNA-seq), Chromatin Immunoprecipitation sequencing (ChIP-seq), and Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq). (**b**) The total cell markers enumeration across immune cell types.

forward, the database is poised for continuous expansion and enrichment. The repository will be augmented with additional scRNA-seq, ChIP-seq, and ATAC-seq data to supplement the growing repertoire of cell types and treatment time points.

Database features and applications

User-friendly browsing and searching

An intelligible web interface has been meticulously crafted to facilitate scholarly inquiry into information pertinent to viral infection. Access to this comprehensive repository is afforded through a freely accessible database website, accessible at http://www.fynn-guo.cn/home.php. The website's interface is structured into ten distinct sections, comprising 'Home', 'Cell Marker', 'Gene Expression', 'Epigenetic Landscape', 'Analysis', 'Statistics', 'Download', 'Submit', 'Help', and 'Search' (Fig. 3a).

The Home page serves as the principal gateway, providing users with fundamental insights into the MOI database and a concise overview of its contents. Upon further navigation, users are directed to sub-sections such as 'Get Started', 'Explore', and 'Analyze', each seamlessly linked to corresponding pages, namely 'Help', 'Markers', and 'Analysis'. This structured approach facilitates efficient navigation and access to relevant information. Furthermore, interactive visual aids are seamlessly integrated to elucidate the features and functionalities of MOI, thereby enhancing user engagement and comprehension. Notably, a dynamic world map located at the bottom of the Home page meticulously logs the access history of MOI, offering valuable insights into user interactions and usage patterns.

A user-friendly graphical navigation system has been implemented to expedite the exploration of MOI's extensive data, empowering users to navigate to their specific areas of interest effortlessly. For example, the Marker page facilitates the exploration of cell markers associated with a diverse array of cells derived from human or mouse tissue under various conditions, meticulously categorized by species. Additionally, users can utilize the 'Filterable Table' feature to input keywords, such as 'T cell' or 'CD14', thereby streamlining the process of retrieving relevant information tailored to their research needs. This functionality enhances the efficiency and precision of data exploration within the MOI database.

Similarly, the Expression and Transcription Factors (TFs) pages replicate the functionality observed in the Marker page, allowing users to explore gene expression profiles and glean insights into transcription factor activity. This consistent functionality ensures coherence and uniformity in the user experience across distinct sections of the MOI database. Moreover, users can expedite their search endeavors by leveraging the 'Search' button at the interface's apex. Upon activation, this feature directs users to the search page, where they can execute finely tailored queries within MOI's extensive database. This refined approach enhances the efficiency and efficacy of information retrieval, empowering users to access pertinent data to their research inquiries swiftly.

Data download and submission

On the one hand, users are directed to the Download page to access specific information within the MOI database. Here, data is systematically categorized based on species, tissue, cell line, and identification method. Datasets are available in various standard-compliant formats, including JSON, TSV, and XML (Fig. 3b). To ensure the integrity of downloaded files, users are provided with corresponding MD5 checksum files for verification purposes. Users can also directly access specific MOI data through Python packages such as urllib and JSON.

On the other hand, the Submit page facilitates user contributions of new virus-related data by soliciting the submission of the PubMed ID, accessible data URL, and contributor information (Fig. 3c). This process undergoes a standardized internal collection and verification procedure. Subsequently, additional data is curated and slated for publication in the subsequent stable release within a maximum timeframe of 10 months. Simultaneously, MOI offers a 'Help' page containing documentation detailing the platform's user guide information to enhance user experience and usability. Users are encouraged to engage with the platform, explore its features, and provide feedback for ongoing improvements.

Online data analysis and visualization tools

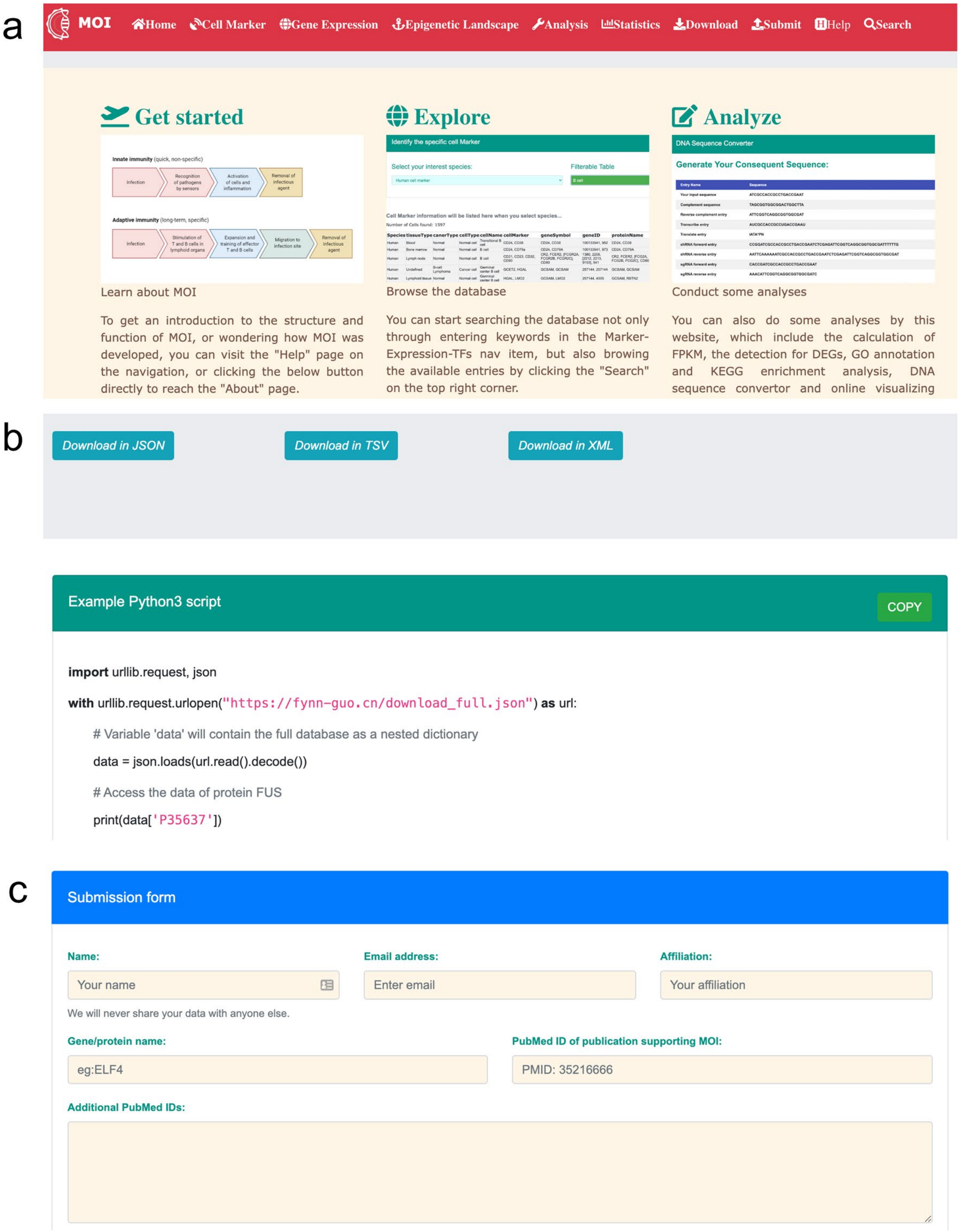
In addition to the aforementioned data sources, MOI presents a suite of online data analysis tools (Fig. 4a), primarily tailored for downstream analysis of bulk RNA-seq data. These tools encompass a range of functionalities, including the computation of FPKM values within expression matrices containing biological replicates. Users can leverage these tools for tasks such as detecting DEGs, elucidating validated gene functions, and conducting GO and KEGG annotations on DEGs. Notably, the gene list length for annotation should be at most 2000 characters.

Since its inception, molecular biology has emerged as a cornerstone of fundamental research in life sciences. Primer design has thus become an indispensable skill for researchers engaged in wet laboratory experiments. Consequently, MOI offers a user-friendly DNA sequence processing tool to aid researchers in primer sequence design. This tool facilitates various tasks, including identifying reverse complementary, transcription, translation, and shRNA or sgRNA sequences designed based on plasmid characteristics and restriction endonuclease sites (Fig. 4b).

Furthermore, MOI has developed a data visualization toolbox to provide users with customizable online representations of 16 typical scientific graphs and charts. These visualizations encompass volcano plots, heatmaps, bubble plots, Venn diagrams, boxplots, violin plots, chord diagrams, and more (Fig. 4c). By offering these comprehensive and freely accessible online data analysis and visualization tools, MOI empowers researchers lacking bioinformatic expertise and programming skills to conduct complex analyses and generate publication-ready plots independently and conveniently.

System design and implementation

As shown in Fig. 1, the MOI website is hosted on a Linux-based NGINX (V1.21.0) web server, accessed at https://www.nginx.com/. Its user interface is meticulously designed with optimization techniques utilizing the Bootstrap (V4.6.1) framework, accessible at https://getbootstrap.com/docs/4.6/getting-started/introduction/. To facilitate sophisticated data presentation, we have integrated plugins for prominent JavaScript libraries, such as jQuery (V3.6.1), incorporating DataTables (V1.10.19) (https://datatables.net), Biodalliance (v0.13.8) (http://www.biodalliance.org/about.html), and morris.js (v0.5.0) (https://morrisjs.github.io/morris.js/index.html). In the server-side architecture, we deploy PHP (V7.4.24) (http://www.php.net), Python (V3.9.7) (https://www.python.org/), and R (V4.1.3) (https://www.r-project.org/). The MOI database is securely housed within a MySQL (V8.0.24) (http://www.mysql.com) database infrastructure. A meticulous testing regime has been conducted across diverse


a
MOI
Home
Cell Marker
Gene Expression
Epigenetic Landscape
Analysis
Statistics
Download
Submit
Help
Search
Get started
Innate immunity (quick, non-specific)
Infection
Recognition of pathogens by sensors
Activation of cells and inflammation
Removal of infectious agent
Adaptive immunity (long-term, specific)
Infection
Stimulation of T and B cells in lymphoid organs
Expansion and training of effector T and B cells
Migration to infection site
Removal of infectious agent
Explore
Analyze
b
Download in JSON
Download in TSV
Download in XML
COPY
c

Learn about MOI

To get an introduction to the structure and function of MOI, or wondering how MOI was developed, you can visit the "Help" page on the navigation, or clicking the below button directly to reach the "About" page.

Identify the specific cell Marker

Select your interest species: Human cell marker

Filterable Table: B cell

Cell Marker information will be listed here when you select species...

Number of Cells found: 1597

Species	tissueType	canerType	cellType	cellName	cellMarker	geneSymbol	geneID	proteinName
Human	Blood	Normal	Normal cell	Transitional B cell	CD24, CD38	CD24, CD38	100133941, 952	CD24, CD38
Human	Bone marrow	Normal	Normal cell	B cell	CD24, CD79a	CD24, CD79A	100133941, 973	CD24, CD79A
Human	Lymph node	Normal	Normal cell	B cell	CD21, CD23, CD32, CD80	CR2, FCER2, [FCGR2A, FCGR2B, FCGR2C], CD80	1380, 2208, [2212, 2213, 9103], 941	CR2, FCER2, [FCG2A, FCG2B, FCG2C], CD80
Human	Undefined	B-cell Lymphoma	Cancer cell	Germinal center B cell	GCET2, HGAL	GCSAM, GCSAM	257144, 257144	GCSAM, GCSAM
Human	Lymphoid tissue	Normal	Normal cell	Germinal center B cell	HGAL, LMO2	GCSAM, LMO2	257144, 4005	GCSAM, RBTN2

Browse the database

You can start searching the database not only through entering keywords in the Marker-Expression-TFs nav item, but also browing the available entries by clicking the "Search" on the top right corner.

DNA Sequence Converter

Generate Your Consequent Sequence:

Entry Name	Sequence
Your input sequence	ATCGCCACCGCCTGACCGAAT
Complement sequence	TAGCGGTGGCGGACTGGCTTA
Reverse complement entry	ATTCGGTCAGGCGGTGGCGAT
Transcribe entry	AUCGCCACCGCCUGACCGAAU
Translate entry	IATA*PN
shRNA forward entry	CCGGATCGCCACCGCCTGACCGAATCTCGAGATTCGGTCAGGCGGTGGCGATTTTTTG
shRNA reverse entry	AATTCAAAAAATCGCCACCGCCTGACCGAATCTCGAGATTCGGTCAGGCGGTGGCGAT
sgRNA forward entry	CACCGATCGCCACCGCCTGACCGAAT
sgRNA reverse entry	AAACATTCGGTCAGGCGGTGGCGATC

Conduct some analyses

You can also do some analyses by this website, which include the calculation of FPKM, the detection for DEGs, GO annotation and KEGG enrichment analysis, DNA sequence convertor and online visualizing

Example Python3 script

```
import urllib.request, json
with urllib.request.urlopen("https://fynn-guo.cn/download_full.json") as url:
    # Variable 'data' will contain the full database as a nested dictionary
    data = json.loads(url.read().decode())
    # Access the data of protein FUS
    print(data['P35637'])
```

Submission form

Name: Your name

Email address: Enter email

Affiliation: Your affiliation

We will never share your data with anyone else.

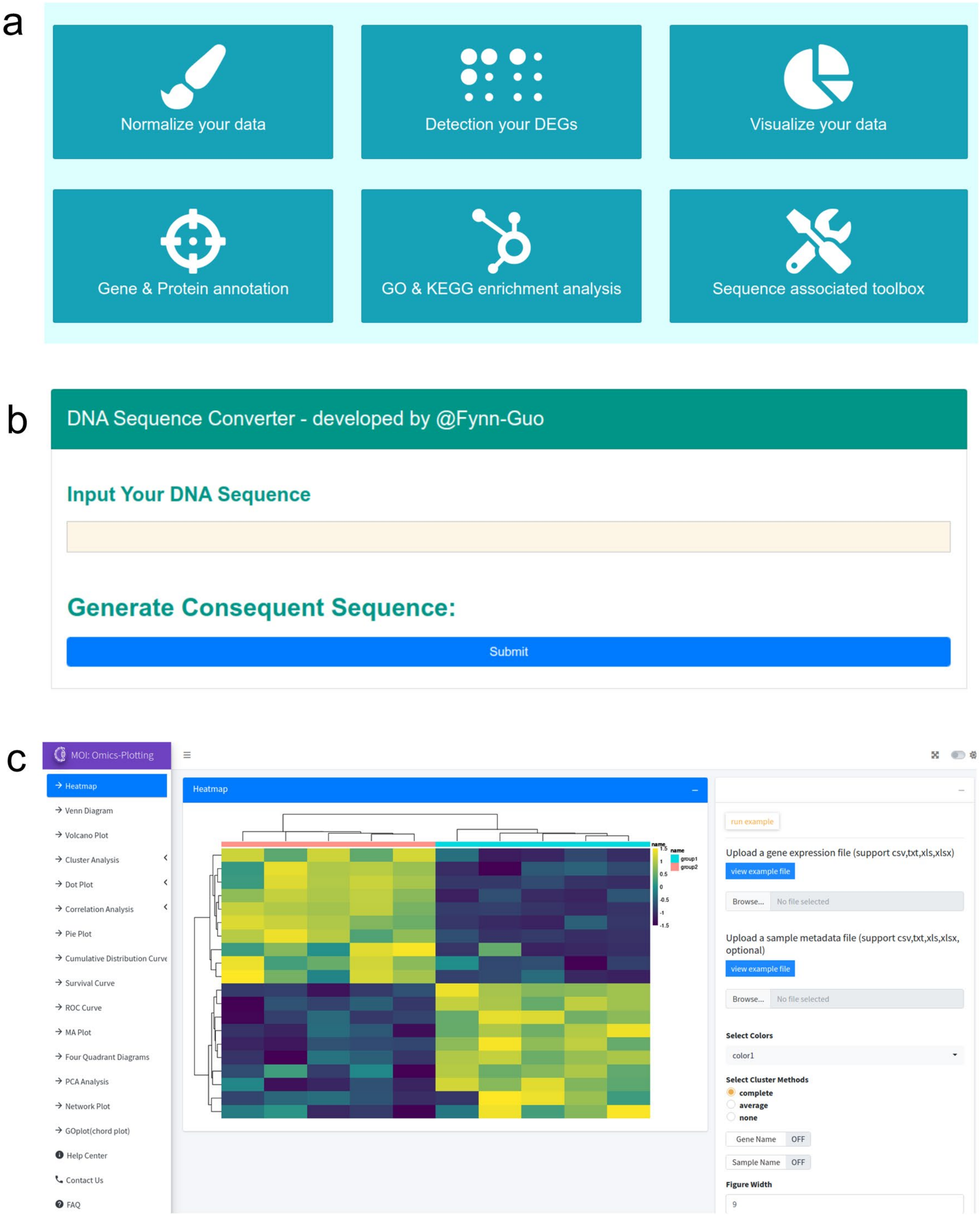
Gene/protein name: eg:ELF4

PubMed ID of publication supporting MOI: PMID: 35216666

Additional PubMed IDs:

Figure 3. The user interface and navigation of MOI. (**a**) The home page interface of MOI. This section illustrates the interface of the MOI home page, providing an overview of its layout and features. (**b**) The data download options of MOI. Various methods for downloading data from MOI are showcased, offering users convenient access to integrated multi-omics datasets. (**c**) The data submission form of MOI. A form enabling users to submit their data to MOI is displayed, facilitating the expansion and enrichment of the integrated database.

mainstream web browsers, including Google Chrome, Firefox, Opera, Apple Safari, and Microsoft Edge to ascertain optimal performance and cross-browser compatibility. Access to the MOI database is readily extended to the scholarly community through the online portal at http://www.fynn-guo.cn/. This repository is a pivotal resource for conducting molecular immunology investigations focusing on viral infections.



Upload a gene expression file (support csv,txt,xls,xlsx)

No file selected

Upload a sample metadata file (support csv,txt,xls,xlsx, optional)

No file selected

Select Colors: color1

Select Cluster Methods: ● complete ○ average ○ none

Gene Name OFF

Sample Name OFF

Figure Width: 9

Figure 4. The online analysis and visualization tools of MOI. (**a**) An overview of downstream analysis and visualization tools. (**b**) An interface of a DNA sequence converter that could convert DNA sequences into various formats. (**c**) The interface of Omics-Plotting. An overview of the Omics-Plotting tool is presented, showcasing its capability to generate 16 standard plots relevant to multi-omics analysis, enabling effective visualization and interpretation of complex data sets.

Case study: exploring host responses to viral infections utilizing MOI

Introduction Viral infections present substantial global health challenges, epitomized by the widespread ramifications of diseases like COVID-19 stemming from SARS-CoV-2. Comprehending the intricate molecular interplay between viruses and their hosts is a cornerstone for devising efficacious intervention strategies. Despite the proliferation of multi-omics data pertinent to viral infections, a conspicuous absence of a centralized repository still facilitates exhaustive exploration into host responses. We introduce the MOI database to address this lacuna. MOI is conceived to furnish researchers with a meticulously curated compendium of multi-omics data delineating virus-host interactions.

Accessing MOI Researchers gain entry to MOI via the web interface accessible at http://www.fynn-guo.cn/. Upon ingress, users are greeted with an intuitive, user-friendly interface meticulously crafted to expedite seamless navigation and data retrieval.

Dataset overview MOI encompasses a vast repository of over 3200 samples afflicted with viral infections sourced from human and mouse models. These samples undergo scrupulous curation and are collated from eminent repositories such as PubMed and the GEO. Spanning an array of high-throughput sequencing modalities, including (single-cell) RNA-seq, ChIP-seq, and ATAC-seq, MOI offers a comprehensive panorama of multi-omics data pertinent to viral infections.

Data processing and integrity Upholding data integrity, MOI implements standardized processing pipelines tailored for data normalization and stringent quality control measures. This rigorous protocol ensures the reliability and suitability of the database contents for downstream analytical endeavors.

Features and tools MOI boasts an array of analytical tools and features meticulously designed to facilitate profound exploration and interpretation of multi-omics data. These encompass:

Cell Marker Tables: Exhaustive tabulations delineating cell markers attributed to diverse cell types implicated in viral infections.

Gene Expression Data: Access to gene expression profiles spanning various viral infections allows users to scrutinize gene expression signature alterations.

Epigenetic Landscape Charts: These are visual representations of epigenetic landscapes. They elucidate chromatin accessibility and transcription factor binding motifs, unraveling intricate regulatory mechanisms.

Analytical Tools: A repertoire encompassing DNA sequence conversion utilities, FPKM calculation modules, differential gene expression analysis platforms, and enrichment analysis tools for elucidating GO/KEGG pathways.

Visualization Plots: Sixteen bespoke visualization plots, meticulously crafted to facilitate intuitive exploration and comprehension of multi-omics data landscapes.

Discussion

Investigating viral infections is fundamental to biological inquiry and is pivotal for understanding disease mechanisms. However, until recently, comprehensive elucidation of the complex interplay among gene expression, cell markers, and regulatory elements has posed significant challenges. The emergence of state-of-the-art experimental techniques and computational methodologies has heralded a transformative era, shedding light on the intricate dynamics of viral-host interactions. Recent advancements in NGS research have propelled numerous datasets to the forefront, offering unparalleled insights into the molecular terrain of viral infections[1–5].

Furthermore, the landscape of databases and web portals dedicated to viruses, viral infections, and host organisms is rich and diverse, each offering unique features and focusing on distinct research areas. For instance, VISDB is a dedicated repository for curating information on viral integration sites within the human genome[12]. In contrast, VirHost is tailored to facilitate research on viral pathogenesis by providing comprehensive insights into host-virus interactions, including identifying host factors implicated in viral infection and targeting host factors by viral proteins[13]. Similarly, ViPR stands out as a comprehensive database that consolidates genomic, proteomic, and functional data spanning various viruses. Equipped with data analysis and visualization tools, ViPR empowers researchers to delve into multiple aspects of viral evolution, epidemiology, and pathogenesis[14]. In a parallel vein, PhEVER fills a niche by focusing on phylogenetic analysis of viral evolution and its ramifications for disease dynamics. Through its suite of tools for sequence analysis and evolutionary prediction, PhEVER enables researchers to discern the evolutionary trajectories of viral sequences[15]. Likewise, HPIDB serves as a repository cataloging human protein interactions, including those pivotal in viral infections. By furnishing information on protein–protein interactions relevant to viral pathogenesis and host response, HPIDB aids in unraveling the intricate interplay between viruses and their hosts[16]. Additionally, resources like ViMIC cater to viral microbiology, providing comprehensive insights into viral genomes, proteins, and interactions[17]. Similarly, ViralZone is a meticulously curated database offering exhaustive information on viral protein families and their functions. This resource is indispensable for researchers seeking to elucidate the molecular mechanisms underpinning viral infections[18]. Moreover, PHI-base is a valuable resource by focusing on interactions between pathogens and their hosts, including viruses. By providing curated information on genes implicated in pathogen-host interactions, PHI-base facilitates the study of host defense mechanisms and pathogen virulence factors[49]. On a broader scale, MVIP represents a pioneering initiative to integrate multi-omics data to unravel the intricate landscape of host responses to viral infections. By amalgamating RNA-seq, ChIP-seq, and CLIP-seq data, MVIP endeavors to shed light on the molecular mechanisms underlying viral infections and host-virus interactions[19]. However, despite the resources available, a notable gap persists in databases explicitly dedicated to organizing and cataloging gene expression profiles, epigenetic landscapes, and cell markers for single-cell RNA sequencing (scRNA-seq) during viral infections.

In response to this notable void, we have developed what we believe to be the first comprehensive database dedicated to the intricacies of gene expression, epigenetic landscape, cell markers, and transcription factors. Through meticulously crafted systematic data collection protocols and standardized datasets, alongside designing a user-friendly web platform adhering to rigorous research standards, our aspiration for MOI is to serve as a cornerstone resource enriching the endeavors of biologists and data scientists alike. MOI is envisioned as a tool to foster a deeper comprehension of the intricate interplay between viruses and their hosts, elucidating gene expression alterations and the regulatory mechanisms underlying these interactions. Additionally, we aim to provide researchers with the means to construct more exhaustive gene regulatory networks by amalgamating the wealth of data accessible through MOI.

Illustrating the breadth of MOI's scope, the database boasts an extensive repository encompassing thousands of cell marker records for human and mouse cells, along with comprehensive catalogs comprising 1638 human transcription factors and 1543 mouse transcription factors. Moreover, MOI includes diverse innate immune cells and standard laboratory cell lines, facilitating the exploration of gene expression profiles across different treatment time points and in response to various viral strains. To further enhance utility, MOI seamlessly integrates a suite of bioinformatic analysis and visualization tools within its framework, offering researchers resources to dissect and interpret the wealth of data.

Following the inaugural release of MOI, our immediate focus is on enhancing its capabilities to encompass more comprehensive epigenomic annotations. This includes delving into diverse facets such as transcription factor binding, histone modification, and chromatin accessibility across various cell lines in response to viral infections[50]. One notable area of emphasis is integrating information about transposable elements (TEs) and endogenous retroviruses (ERVs) into MOI[51–53]. Emerging research suggests that TEs undergo up-regulation shortly after viral infection, preceding significant increases in virus replication and interferon β expression[54]. This underscores their potential role in modulating host innate immune responses, a phenomenon we aim to elucidate within MOI's curated datasets. By incorporating detailed information on TEs and ERVs, we strive to provide researchers with a nuanced understanding of their contributions to viral pathogenesis. Furthermore, we remain committed to refining MOI's user experience by enhancing visualization tools for scRNA-seq gene expression comparisons and optimizing search query speeds. These improvements aim to bolster MOI's utility as a comprehensive data resource, catering to the needs of both traditional molecular biology research and burgeoning computational methodologies.

In summary, MOI is a testament to our dedication to advancing scientific inquiry. By continuously evolving and expanding its capabilities, we endeavor to empower researchers with the tools and resources needed to unravel the complexities of virus-host interactions, driving forward both traditional molecular biology research and the burgeoning field of computational approaches.

Data availability

Users can access any feature available in the MOI database without registering or logging in. All data is freely accessible to the research community at http://www.fynn-guo.cn/json_data.json or http://www.fynn-guo.cn/download_infor.tsv. In addition to various options for downloading data on the *Download* page, users can customize and download the filtered data on the *Search* page.

Received: 26 January 2024; Accepted: 21 June 2024
Published online: 26 June 2024

References

1. Zhao, Y. *et al.* SARS-CoV-2 spike protein interacts with and activates TLR41. *Cell Res.* **31**, 818–820. https://doi.org/10.1038/s41422-021-00495-9 (2021).
2. Su, Y. *et al.* Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19. *Cell* **183**, 1479-1495.e20 (2020).
3. Wang, D. *et al.* The SARS-CoV-2 subgenome landscape and its novel regulatory features. *Mol. Cell* **81**, 2135-2147.e5 (2021).
4. Andres-Terre, M. *et al.* Integrated, multi-cohort analysis identifies conserved transcriptional signatures across multiple respiratory viruses. *Immunity* **43**, 1199–1211 (2015).
5. Wu, A. *et al.* One year of SARS-CoV-2 evolution. *Cell Host Microbe* **29**, 503–507 (2021).
6. Zou, L. *et al.* SARS-CoV-2 viral load in upper respiratory specimens of infected patients. *New Engl. J. Med.* **382**, 1177 (2020).
7. Ruff, W. E., Greiling, T. M. & Kriegel, M. A. Host–microbiota interactions in immune-mediated diseases. *Nat. Rev. Microbiol.* **18**, 521–538. https://doi.org/10.1038/s41579-020-0367-2 (2020).
8. Mahalingam, S. *et al.* Landscape of humoral immune responses against SARS-CoV-2 in patients with COVID-19 disease and the value of antibody testing. *Heliyon* **7**, 4. https://doi.org/10.1016/j.heliyon.2021.e06836 (2021).
9. Watanabe, T., Watanabe, S. & Kawaoka, Y. Cellular networks involved in the influenza virus life cycle. *Cell Host Microbe* **7**, 427–439. https://doi.org/10.1016/j.chom.2010.05.008 (2010).
10. Lee, S. *et al.* The SARS-CoV-2 RNA interactome. *Mol. Cell* **81**, 2838-2850.e6 (2021).
11. Stukalov, A. *et al.* Multilevel proteomics reveals host perturbations by SARS-CoV-2 and SARS-CoV. *Nature* **594**, 246–252 (2021).
12. Tang, D. *et al.* VISDB: A manually curated database of viral integration sites in the human genome. *Nucleic Acids Res.* **48**, D633–D641 (2020).
13. Mihara, T. *et al.* Linking virus genomes with host taxonomy. *Viruses* **8**, 66 (2016).
14. Zhang, Y., Zmasek, C., Sun, G., Larsen, C. N. & Scheuermann, R. H. Hepatitis C virus database and bioinformatics analysis tools in the virus pathogen resource (ViPR). In *Methods in Molecular Biology, vol. 1911* 47–69 (Humana Press Inc., 2019).
15. Palmeira, L., Penel, S., Lotteau, V., Rabourdin-Combe, C. & Gautier, C. PhEVER: A database for the global exploration of virus-host evolutionary relationships. *Nucleic Acids Res.* **39**, 569 (2011).
16. Ammari, M. G., Gresham, C. R., McCarthy, F. M. & Nanduri, B. HPIDB 2.0: A curated database for host-pathogen interactions. *Database Oxf.* **2016**, 103 (2016).
17. Wang, Y. *et al.* ViMIC: A database of human disease-related virus mutations, integration sites and cis-effects. *Nucleic Acids Res.* **50**, D918–D927 (2022).
18. Hulo, C. *et al.* ViralZone: A knowledge resource to understand virus diversity. *Nucleic Acids Res.* **39**, 576 (2011).
19. Tang, Z. *et al.* MVIP: Multi-omics portal of viral infection. *Nucleic Acids Res.* **50**, D817–D827 (2022).
20. Katzenelenbogen, Y. *et al.* Coupled scRNA-Seq and intracellular protein activity reveal an immunosuppressive role of TREM2 in cancer. *Cell* **182**, 872-885.e19 (2020).
21. Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* **18**, 35–45. https://doi.org/10.1038/nri.2017.76 (2018).
22. Li, Q. *et al.* Developmental heterogeneity of microglia and brain myeloid cells revealed by deep single-cell RNA sequencing. *Neuron* **101**, 207-223.e10 (2019).
23. Wang, M. *et al.* Single-cell RNA sequencing analysis reveals sequential cell fate transition during human spermatogenesis. *Cell Stem Cell* **23**, 599-614.e4 (2018).

24. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
25. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
26. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
27. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
28. Clarke, Z. A. *et al.* Tutorial: Guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nat. Protocols* **16**, 2749–2764. https://doi.org/10.1038/s41596-021-00534-0 (2021).
29. Kharchenko, P. V. The triumphs and limitations of computational methods for scRNA-seq. *Nat. Methods* **18**, 723–732. https://doi.org/10.1038/s41592-021-01171-x (2021).
30. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
31. Hou, R., Denisenko, E. & Forrest, A. R. R. ScMatch: A single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics* **35**, 4688–4695 (2019).
32. Hu, H. *et al.* AnimalTFDB 3.0: A comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res.* **47**, D33–D38 (2019).
33. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41**, 452 (2013).
34. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
35. Zhao, Y. *et al.* TPM, FPKM, or normalized counts? A comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models repository. *J. Transl. Med.* **19**, 21 (2021).
36. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 12 (2014).
37. Dennis, G. *et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery*. *Genome Biology, vol. 4* http://dot.ped.med.umich.edu:2000/ (2003).
38. Kanehisa, M. & Goto, S. *KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, vol. 28* http://www.genome.ad.jp/kegg/ (2000).
39. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951. https://doi.org/10.1002/pro.3715 (2019).
40. Finak, G. *et al.* MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 1 (2015).
41. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, 25 (2009).
42. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
43. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
44. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, 9 (2008).
45. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
46. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. DeepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, 187 (2014).
47. Yu, G., Wang, L. G. & He, Q. Y. ChIP seeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
48. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26. https://doi.org/10.1038/nbt.1754 (2011).
49. Urban, M. *et al.* PHI-base: The pathogen-host interactions database. *Nucleic Acids Res.* **48**, D613–D620 (2020).
50. Scott-Browne, J. P. *et al.* Dynamic changes in chromatin accessibility occur in CD8+ T cells responding to viral infection. *Immunity* **45**, 1327–1340 (2016).
51. Macchietto, M. G., Langlois, R. A. & Shen, S. S. Virus-induced transposable element expression up-regulation in human and mouse host cells. *Life Sci. Alliance* **3**, 2 (2020).
52. Guo, X., Zhao, Y. & You, F. Identification and characterization of endogenous retroviruses upon SARS-CoV-2 infection. *Front. Immunol.* **15**, 145 (2024).
53. Cañadas, I. *et al.* Tumor innate immunity primed by specific interferon-stimulated endogenous retroviruses. *Nat. Med.* **24**, 1143–1150 (2018).
54. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **1979**(351), 1083–1087 (2016).

Acknowledgements

We thank the support from the Peking University High-performance Computing Platform for providing the computing clusters to facilitate the data analysis procedure in this research.

Author contributions

Xuefei Guo conducted and revised this research, analyzed most of the data, and built these websites. Yang Zhao and Fuping You downloaded and analyzed part of the data. Xuefei Guo and Yang Zhao wrote the main manuscript text. All authors reviewed the manuscript.

Funding

Peking University Medicine Fund (PKU2020LCXQ009 to YFP).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-65629-6.

Correspondence and requests for materials should be addressed to X.G.

Reprints and permissions information is available at www.nature.com/reprints.