BMC Bioinformatics

DATABASE

Open Access

HIHISIV: a database of gene expression in HIV and SIV host immune response

Raquel L. Costa[1]*, Luiz Gadelha[2]*, Mirela D'arc[3], Marcelo Ribeiro-Alves[4], David L. Robertson[5], Jean-Marc Schwartz[6], Marcelo A. Soares[7,8] and Fábio Porto[1]

*Correspondence: quelopes@lncc.br; luiz.gadelha@dkfz-heidelberg.de

[1] DEXL Lab, National Laboratory for Scientific Computing, Petrópolis, Brazil
[2] German Human Genome-Phenome Archive (GHGA, W620), German Cancer Research Center (DKFZ), Heidelberg, Germany
[3] Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
[4] Instituto Nacional de Infectologia Evandro Chagas, Oswaldo Cruz Foundation (Fiocruz), Rio de Janeiro, Brazil
[5] MRC-University of Glasgow Centre for Virus Research, University of Glasgow, Glasgow, UK
[6] Division of Evolution, Infection and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester M13 9PT, UK
[7] Departamento de Genética, Instituto de Biologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro 21941-902, Brazil
[8] Programa de Oncovirologia, Divisão de Pesquisa Translacional, Instituto Nacional do Câncer, Rio de Janeiro 20230-130, Brazil

Abstract

In the battle of the host against lentiviral pathogenesis, the immune response is crucial. However, several questions remain unanswered about the interaction with different viruses and their influence on disease progression. The simian immunodeficiency virus (SIV) infecting nonhuman primates (NHP) is widely used as a model for the study of the human immunodeficiency virus (HIV) both because they are evolutionarily linked and because they share physiological and anatomical similarities that are largely explored to understand the disease progression. The HIHISIV database was developed to support researchers to integrate and evaluate the large number of transcriptional data associated with the presence/absence of the pathogen (SIV or HIV) and the host response (NHP and human). The datasets are composed of microarray and RNA-Seq gene expression data that were selected, curated, analyzed, enriched, and stored in a relational database. Six query templates comprise the main data analysis functions and the resulting information can be downloaded. The HIHISIV database, available at https://hihisiv.github.io, provides accurate resources for browsing and visualizing results and for more robust analyses of pre-existing data in transcriptome repositories.

Keywords: HIV-1, SIV, Immune response, High-throughput gene expression data

Background

The evolutionary history of human immunodeficiency virus type 1 (HIV-1) is closely linked with that of simian immunodeficiency virus (SIV), with four cross-species introductions described from chimpanzees (HIV-1 M and N) and gorillas (HIV-1 O and P) into humans. For years, the general thought was that all African monkey species do not develop immunodeficiency syndrome when infected with SIVs, namely natural hosts, similar to the human cases of non-progressive and controller patients [1–6]. However, little is known about the in vivo pathogenicity of SIV because it is very difficult to monitor infected primates in the wild. Only captive animal data were initially analyzed and it was assumed that all naturally infected monkeys develop a non-pathogenic phenotype, such as sooty mangabeys and African green monkeys that are widely studied in captivity. In captivity, macaque species, endemic to Asia, can be experimentally infected with virus strains from sooty mangabeys and progress to an AIDS-like phase. The clinical hallmarks of disease in these (non-naturally) infected macaques is similar to the human immunodeficiency disease caused by HIV-1 [7, 8]. Importantly, epidemiological surveys in free-ranging chimpanzees showed that SIVcpz has a substantial negative impact on health, reproduction and lifespan of infected animals indicating natural SIV infections are not necessarily non-pathogenic.

One of the major challenges in the study of these viruses is understanding the complex interplay with the host's immune system, which tends to involve the depletion of CD4 + T lymphocytes. In the last few decades, advances in high-throughput technologies has enabled the analysis of large numbers of transcripts simultaneously and inferring subsets of these transcripts that are associated with the same biological condition [9]. Although the amount of data is constantly increasing in transcriptome repositories, such as GEO (Gene Expression Omnibus), ArrayExpress, and SRA (Sequence Read Archive), the data are not processed, standardized, or integrated.

While several HIV databases offer an integrated approach, providing a comprehensive toolkit for exploring host genome dynamics, understanding latency mechanisms, and decoding the functional implications of differentially expressed genes during HIV infection [10–12], there remains a notable gap in databases that also organize and present SIV gene expression data.

To address this gap, we introduce HIHISIV, a database dedicated to the host immune response in the expression profiles of both SIV and HIV infections. Developed to support researchers in identifying molecular signatures, co-expressed genes, experimental design, and essential host factors. Moreover, it serves as a valuable framework for researchers in generating hypotheses and advancing our understanding of these infections. The database is composed of microarray and RNA-Seq gene expression data of viral infection by SIV and HIV hosts retrieved from the GEO repository. The datasets were manually curated and reanalyzed using standardized methods and the metadata and results were annotated and enriched with ontology terms. The HIHISIV database is currently in version 2.0 and includes 63 transcriptome experiments stored in a relational database. Furthermore, HIHISIV has a user-friendly web interface where co-expression networks can be visualized as graph and tabular data views and the dataset can be downloaded for further analysis.

Construction and content

Datasets and data analysis

Dataset selection

The initial datasets consist of microarray (1-color mRNA) and RNA-Seq (mRNA) gene expression data related to viral infections with SIV or HIV-1, retrieved from the GEO from the National Center for Biotechnology Information (NCBI). We compiled the initial database by selecting searches for 'SIV' or 'HIV-1' projects. Each dataset was manually examined to identify and separate the groups for contrast (reference *versus* test). Additionally, we established exclusion criteria for GEO projects or samples, and the following criteria were adopted for removing projects or samples:

(i) Lack of sample information preventing the differentiation of groups for comparison (e.g., host type); (ii) Projects or samples related to antiretroviral therapy (ART), vaccination, monoclonal antibodies, or cell cultures; (iii) Insufficient number of samples (less than 3 per group).

It's important to note that each of these criteria serves as an exclusion factor, thus, any datasets meeting any of these conditions were excluded.

As a proof of concept to create the pipeline and compose the conceptual model, we selected one short-read RNA-Seq project (GEO id: GSE119234) and applied the same exclusion criteria.

During the curation process, we encountered various terminologies that refer to the same entity. To ensure standardization across different datasets and establish consistency, we have adopted the following nomenclature:

- natural host: refers to natural host primates such as African green monkeys (*Chlorocebus aethiops*) and sooty mangabeys (*Cercocebus atys*);
- non-natural host: refers to non-natural host primates such as rhesus macaque (*Macaca mulatta*);
- non-progressor: refers to a human that has not progressed to immunodeficiency.
- uninfected (EFO_0001460): uninfected class is a disposition in which the bearer is not known to be affected by a disease within the context of a study.
- acute infection (IDO:0000627): an infectious disorder that is the physical basis for an unfolding acute infectious disease course.
- chronic infection (IDO:0000628): an infection that persists for an extended period of time.

In each project, the data were compared pairwise, and each comparison was named as an *experiment*. In some cases, these projects resulted in more than one experiment. For example, the project GSE7157 worked with non-natural species during three different phases of the infection (uninfected, acute and chronic infections). This project was divided into three different pairwise experiments (uninfected *versus* acute: GSE7157_d1; uninfected *versus* chronic: GSE7157_d2, and acute *versus* chronic: GSE7157_d3). The complete list of experiments is shown in the *Experiments list* navigation bar entry on the database web page and in the Additional file 1.

Metadata and ontologies

To enrich the experimental information with ontology terms, we employed the GEOquery library, for retrieving metadata (including phenodata) from GEO datasets. The extracted metadata consisted of semi-structured information, including essential details in the experiments such as title, summary, and overall design. We used SpaCy model, a natural language processing (NLP) tool, to identify automatically the key elements from these metadata [13]. After that, the spurious or irrelevant terms were carefully filtered and the remaining terms were mapped to their corresponding biological ontologies using rols library [14] to access and query the Ontology Lookup Service (OLS) improving the interoperability and integration of the HIHISIV database with other sources.

Microarray and RNA-Seq data analysis

Due to the heterogeneity of experiments conducted by different research teams (assay platforms, virus strains, tissues etc.), the analyses were conducted separately through the pipeline delineated in a dataflow shown in Fig. 1. The list of experiments with the corresponding normalization matrix and phenodata is available in the HIHISIV database.

All microarray data (1-color mRNA) were downloaded from the Affymetrix platform [15]. The datasets were normalized using the MAS5 normalization method implemented in the *Affy* R package (version 1.78.2) [16] resulting in a normalized matrix with the probes. For each experiment, a pair was compared. The first condition was considered as the reference and the second condition as the test, for instance, uninfected (reference) *versus* acute infection (test).

The detection of the differentially expressed genes (DEGs) was based on normalized datasets by the fitting of the gene-wise linear model (for each probe) followed by moderated t-tests implemented in the *limma* package in R (version 3.56.2) [17]. We calculated the up and down-regulated genes between the sample pairs through the adjusted p-value (default for queries: false-discovery rate (FDR) ≤ 0.05) and the log fold-change measures (default for queries: $|\text{log-FC}| > 1$). We kept the complete results of the analysis in the database and probes mapped to more than one entrez_gene_id were inserted as separate tuples in the database. Orthologous genes in *Macaca mulatta* and *Homo sapiens* were mapped using the *bioRmat* R library [18].

To establish a standardized pipeline, an RNA-Seq (mRNA) dataset was collected from the SRA repository (SRS from NextSeq 550 *Homo sapiens*) [19]. Firstly, the data were transformed to Fastq using the SRA toolkit (version 2.10.5) and the quality of reads was checked using FastQC. RSEM *rsem-calculate-expression* (version 1.3) was used to align the transcriptome dataset in the genome reference (UCSC hg19) and after the mapping, the DEGs were evaluated using the library *limma* (limma-voom). The annotation was conducted by the R libraries *TxDBHsapiens*, *HomoSapiens,* and *GenomicFeatures* (version 1.52.1) [20]. The gene identifiers ('ucsc_id') were kept.

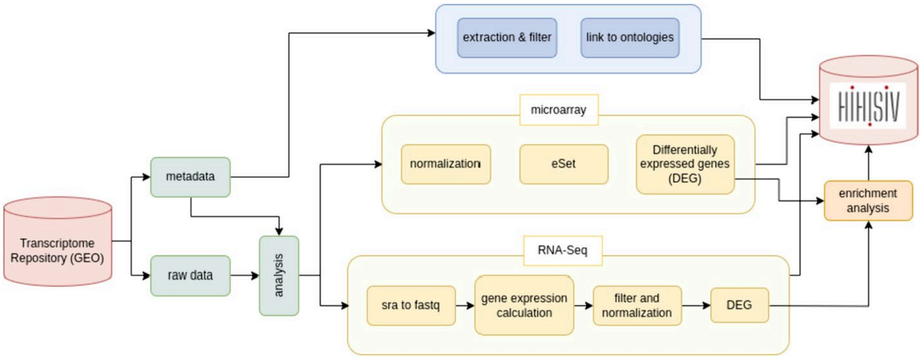


Fig. 1 Dataflow used to generate data for HIHISIV showing metadata and raw data input, their RNA-Seq, microarray, and enrichment analyses, the metadata extraction of relevant entities, and insertion of all results in the database

Enrichment analysis, ingestion into database, and query design

We conducted enrichment analysis using the hypergeometric test implemented in the GOStats package [21] to identify relevant Gene Ontology (GO) terms associated with biological processes [22]. The gene subset was defined by the set of differentially expressed (DE) genes obtained from pairwise comparisons between a test and a reference condition (threshold criteria: adj_*p* value ≤ 0.05 and |log-FC|> 1). To mitigate potential bias, as demonstrated by Timmons, Szkop and Gallagher [23], we applied a more restrictive cut-off (*p*-value < 0.001) to reduce detection bias.

The results of the analysis were ingested into the database (the technical description can be found in Additional file 2) and the single gene expression and co-expression information, presented visually as networks, are extracted by querying the database tables in SQL.

System architecture, implementation, and access

The HIHISIV database was designed and implemented as a multiple-component framework, as represented in Fig. 2. The database was implemented and instantiated in a PostgreSQL (version 15.3) server following the same conceptual model described previously (Additional file 2). The web application provides a user-friendly interface implemented in the *Streamlit* (version 1.25.0) Python library with predefined parameterizable queries to be executed in the database. It enables the analysis of gene expression levels in HIV-1 and SIV hosts. The results are displayed in tables or networks, and all the data can be downloaded by the users. Other Python libraries used in the implementation include *Psycopg2* (version 2.9.5), for connecting to the PostgreSQL database and executing queries; *Pandas* (version 1.5.3) for manipulating data frames; and *NetworkX* (version 3.1) for building networks.

The static web page contains a description of how the database was built and a link to the source data that were used. The web page was written using the Jekyll framework. Finally, the Nginx proxy was used to route the requests from the Internet to the web app. For each of these components, a docker container was built and docker-compose was used to connect and deploy them. The system is hosted at the Data Extreme Laboratory (DEXL) of the National Laboratory for Scientific Computing (LNCC) in Brazil.

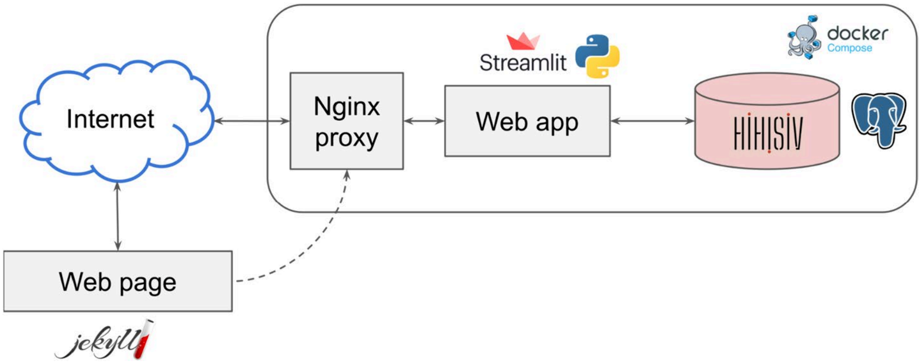


Fig. 2 The HIHISIV framework system architecture

Support for the FAIR principles

The FAIR principles [24] are considered a gold standard for research data management. In this work, we follow best practices for improving the support for them in the HIHISIV database. For improving findability, basic metadata using schema.org [25] was added to the headers of the HIHISIV web page, allowing for indexing by search engines and dataset repositories. HIHISIV data and metadata are openly accessible through its URL (https://hihisiv.github.io) using the HTTP protocol. When applicable, terms used both in the database schema and in its web interface follow Gene Ontology [22] and other ontologies such as EFO (The Experimental Factor Ontology), UBERON (Uber-anatomy ontology), PATO (the Phenotype And Trait Ontology), MONDO (Mondo Disease Ontology) and NCIT (The National Cancer Institute Thesaurus) improving interoperability. For better reusability, the database contains references to the original data used in the analyses. The source code for the analysis workflows, the database schema and initialization scripts, and the web application are available on GitHub (https://github.com/quelopes/HIHISIV). This article describes v2.0 of the database, which is archived on Zenodo under 10.5281/zenodo.7093185.

Utility and discussion

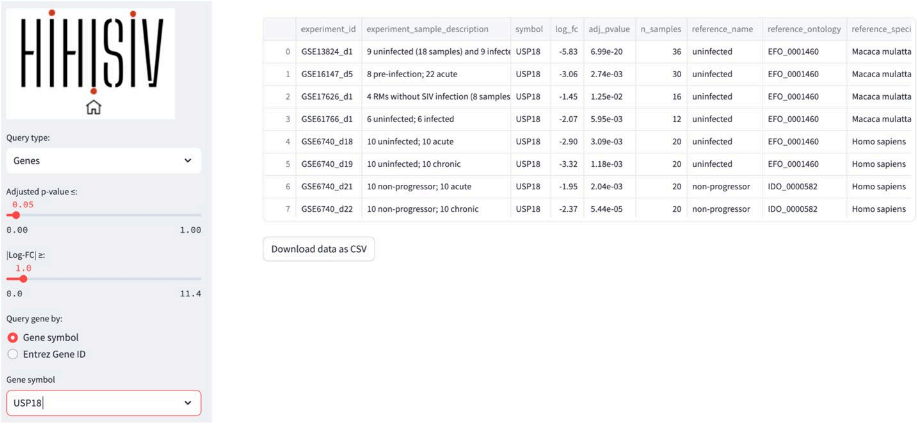
Querying in the user interface

The current version of the database (2.0) comprises a total of 62 experiments sourced from the 14 GEO transcriptome repository. The database includes a total of 1,816 differentially expressed genes (DEGs), considering gene symbols. Additionally, transcript IDs were retained in the results and are displayed in the query interface, totaling 15,026 unique transcript IDs. The criteria for this count were set as adjusted p-value ≤ 0.05 and $|\text{log-FC}| > 1$. However, it is noteworthy that these thresholds were primarily applied to highlight significant genes found in the analysis. Nonetheless, all results were retained in the database to provide flexibility for adjustments by researchers. Concerning Gene Ontology (GO) Biological Process terms, we selected a total of 1249 unique terms when applying the thresholds outlined in the 'Enrichment analysis, ingestion into database, and query design' section.

HIHISIV can be queried through a web interface in which the user can configure and adjust the parameters of a query. There are six query interfaces with major analytic functions, which are '*Gene*', '*Transcripts*', '*Biological Process (GO)*', '*Ontology terms*', '*Single gene co-expression network*', and '*Gene set co-expression network*'. The resulting information can be exported in a comma-separated values (CSV) file for further data manipulation. Public access and documentation are freely available through the database web page. Next, we provide a set of illustrative examples in a biological context to demonstrate how researchers can utilize HIHISIV to explore relevant datasets, identify key genes associated with immune responses, analyze co-expression networks, and elucidate relevant biological processes related to SIV and HIV-1 infections.

Gene

In this mode, the user can select the conditions under which a gene of interest shows differential expression. As an example, Fig. 3 presents the results for the gene USP18,



Query type:
Genes

Adjusted p-value ≤:
0.05
0.00 1.00

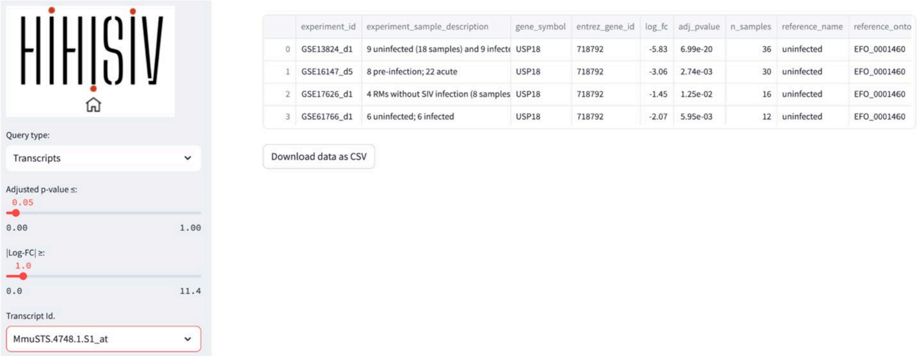
|Log-FC| ≥:
1.0
0.0 11.4

Query gene by:
● Gene symbol
○ Entrez Gene ID

Gene symbol
USP18

	experiment_id	experiment_sample_description	symbol	log_fc	adj_pvalue	n_samples	reference_name	reference_ontology	reference_speci
0	GSE13824_d1	9 uninfected (18 samples) and 9 infecte	USP18	-5.83	6.99e-20	36	uninfected	EFO_0001460	Macaca mulatta
1	GSE16147_d5	8 pre-infection; 22 acute	USP18	-3.06	2.74e-03	30	uninfected	EFO_0001460	Macaca mulatta
2	GSE17626_d1	4 RMs without SIV infection (8 samples	USP18	-1.45	1.25e-02	16	uninfected	EFO_0001460	Macaca mulatta
3	GSE61766_d1	6 uninfected; 6 infected	USP18	-2.07	5.95e-03	12	uninfected	EFO_0001460	Macaca mulatta
4	GSE6740_d18	10 uninfected; 10 acute	USP18	-2.90	3.09e-03	20	uninfected	EFO_0001460	Homo sapiens
5	GSE6740_d19	10 uninfected; 10 chronic	USP18	-3.32	1.18e-03	20	uninfected	EFO_0001460	Homo sapiens
6	GSE6740_d21	10 non-progressor; 10 acute	USP18	-1.95	2.04e-03	20	non-progressor	IDO_0000582	Homo sapiens
7	GSE6740_d22	10 non-progressor; 10 chronic	USP18	-2.37	5.44e-05	20	non-progressor	IDO_0000582	Homo sapiens

Fig. 3 'Gene' query example for the gene *USP18,* showing the conditions under which the gene was selected (adj_*p* value ≤ 0.05 and |log-FC|) > 1)



Query type:
Transcripts

Adjusted p-value ≤:
0.05
0.00 1.00

|Log-FC| ≥:
1.0
0.0 11.4

Transcript Id.
MmuSTS.4748.1.S1_at

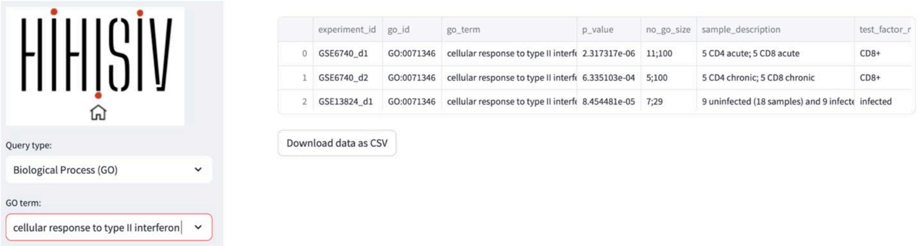
	experiment_id	experiment_sample_description	gene_symbol	entrez_gene_id	log_fc	adj_pvalue	n_samples	reference_name	reference_onto
0	GSE13824_d1	9 uninfected (18 samples) and 9 infecte	USP18	718792	-5.83	6.99e-20	36	uninfected	EFO_0001460
1	GSE16147_d5	8 pre-infection; 22 acute	USP18	718792	-3.06	2.74e-03	30	uninfected	EFO_0001460
2	GSE17626_d1	4 RMs without SIV infection (8 samples	USP18	718792	-1.45	1.25e-02	16	uninfected	EFO_0001460
3	GSE61766_d1	6 uninfected; 6 infected	USP18	718792	-2.07	5.95e-03	12	uninfected	EFO_0001460

Fig. 4 'Transcripts' query example shows the result for the probe id 'Mmu.6048.1.S1_s_at' (adj_*p* value ≤ 0.05 and |log-FC|> 1)

(ubiquitin specific peptidase 18) that belongs to the ubiquitin-specific proteases (UBP). The *USP18* gene and its protein product, USP18, are known to be involved in the innate immune response to viral infections, particularly in response to type I interferons, which are cytokines produced by the body during viral infections [12]. The applied thresholds for differential expression analysis were adjusted p-value (adj_*p* value) ≤ 0.05 and |log-FC| > 1. The results show this gene was DE in seven experiments, for example in the comparison rhesus uninfected *versus* acute infection (GSE16147_d5, GSE17626_d1, GSE61766_d1), humans uninfected *versus* acute infection (GSE6740_d18) and human non-progressor *versus* acute infection (GSE6740_d21).

Transcript

This mode is similar to the previous one, but instead of querying by gene symbol or entrez_gene_id, the user may be interested in a specific probe or transcript id. In the example shown in Fig. 4, the probe id '*Mmu.STS.4748.a.s1_at*' is selected based on the criteria of adj_*p* value ≤ 0.05 and |log-FC| > 1. As this probe id is specific to microarray platforms in *M. mulatta* (GPL3535—Affymetrix Rhesus Macaque Genome Array), the results show only experiments that used this platform.



Query type:

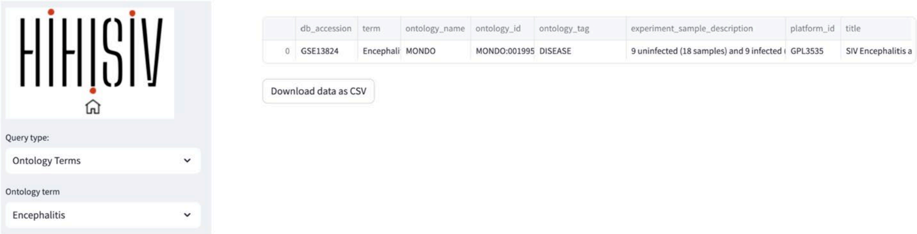
Biological Process (GO)

GO term:

cellular response to type II interferon

	experiment_id	go_id	go_term	p_value	no_go_size	sample_description	test_factor_r
0	GSE6740_d1	GO:0071346	cellular response to type II interfe	2.317317e-06	11;100	5 CD4 acute; 5 CD8 acute	CD8+
1	GSE6740_d2	GO:0071346	cellular response to type II interfe	6.335103e-04	5;100	5 CD4 chronic; 5 CD8 chronic	CD8+
2	GSE13824_d1	GO:0071346	cellular response to type II interfe	8.454481e-05	7;29	9 uninfected (18 samples) and 9 infecte	infected

Fig. 5 An example of a 'Gene Ontology' query shows the result for the biological process domain for the GO term 'cellular response to type II interferon'



Query type:

Ontology Terms

Ontology term

Encephalitis

	db_accession	term	ontology_name	ontology_id	ontology_tag	experiment_sample_description	platform_id	title
0	GSE13824	Encephali	MONDO	MONDO:001995	DISEASE	9 uninfected (18 samples) and 9 infected (	GPL3535	SIV Encephalitis a

Fig. 6 'Ontology term' mode shows the project (GSE13824) associated with the term 'Encephalitis'

Gene ontology (biological process)

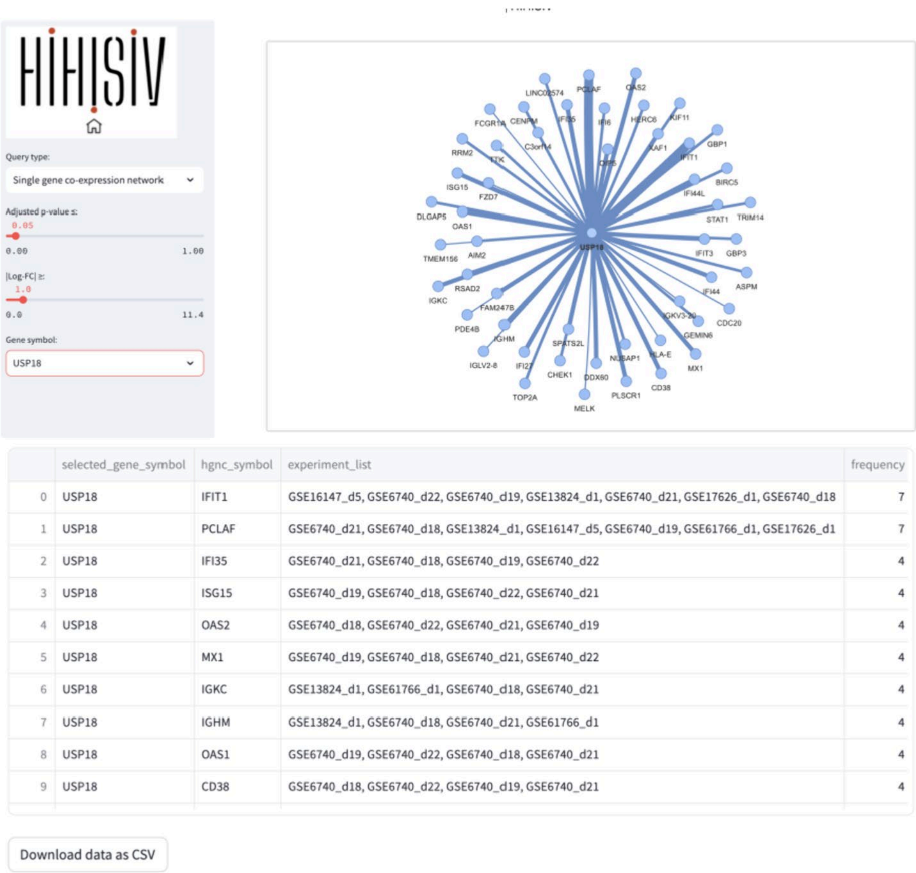
This mode presents the results of the enrichment for GO terms related to BP for each comparison derived from the DEG analysis. By utilizing this mode, researchers can identify experiments that exhibit enrichment in specific BP terms of interest. For instance, in Fig. 5, the BP term '*cellular response to type II interferon*' (GO:0071346) was found enriched in the experiments GSE6740_d1 (CD4 + *versus* CD8 + in acute infection), GSE6740_d2 (CD4 + *versus* CD8 + in chronic infection) and GSE13824_d1 (uninfected *versus* infected). Additionally, this mode provides information about the genes that were enriched within this specific ontology. To ensure statistical robustness, only the results with a significant p-value < 0.001 are displayed in the database.

Ontology terms

By integrating ontology terms with the experimental information, our approach provided a comprehensive understanding of the underlying biological contexts and enriched the dataset's metadata with valuable semantic annotations. In the '*Ontology term*' mode the example in Fig. 6 shows the experiment associated with the term '*Encephalitis*' (project GSE13824).

Single gene co-expression network

This query mode '*Single gene co-expression network*' represents the genes that were co-expressed with a selected target gene. In the example shown in Fig. 7, we are using the same gene as in the 'Gene' mode, *USP18* (adj_*p* value ≤ 0.05 and |log-FC|> 1). The result displays a set of genes that are co-expressed with *USP18* in different experiments (e.g., *IFIT1, PCLAF, IFI35, ISG15* and *OAS2*). The thickness of the connection lines between genes depends on the number of co-expressed experiments that the target gene has with



Query type: Single gene co-expression network

Adjusted p-value ≤: 0.05 (0.00 – 1.00)

|Log-FC| ≥: 1.0 (0.0 – 11.4)

Gene symbol: USP18

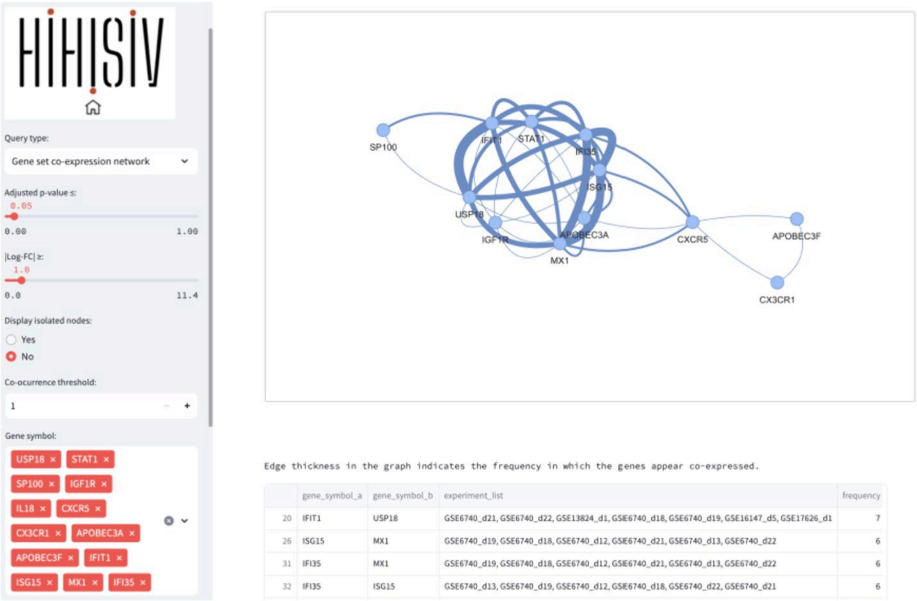
	selected_gene_symbol	hgnc_symbol	experiment_list	frequency
0	USP18	IFIT1	GSE16147_d5, GSE6740_d22, GSE6740_d19, GSE13824_d1, GSE6740_d21, GSE17626_d1, GSE6740_d18	7
1	USP18	PCLAF	GSE6740_d21, GSE6740_d18, GSE13824_d1, GSE16147_d5, GSE6740_d19, GSE61766_d1, GSE17626_d1	7
2	USP18	IFI35	GSE6740_d21, GSE6740_d18, GSE6740_d19, GSE6740_d22	4
3	USP18	ISG15	GSE6740_d19, GSE6740_d18, GSE6740_d22, GSE6740_d21	4
4	USP18	OAS2	GSE6740_d18, GSE6740_d22, GSE6740_d21, GSE6740_d19	4
5	USP18	MX1	GSE6740_d19, GSE6740_d18, GSE6740_d21, GSE6740_d22	4
6	USP18	IGKC	GSE13824_d1, GSE61766_d1, GSE6740_d18, GSE6740_d21	4
7	USP18	IGHM	GSE13824_d1, GSE6740_d18, GSE6740_d21, GSE61766_d1	4
8	USP18	OAS1	GSE6740_d19, GSE6740_d22, GSE6740_d18, GSE6740_d21	4
9	USP18	CD38	GSE6740_d18, GSE6740_d22, GSE6740_d19, GSE6740_d21	4

Fig. 7 Example of 'Single gene co-expressed network' query for the gene *USP18*. The connections in this gene represent the genes co-expressed according to the parameters (adj_*p* value ≤ 0.05 and |log-FC|> 1) and the thickness of the connection varies according to the number of co-expressed experiments that the target gene has with the resulting gene

the resulting gene. For instance, *IFIT1* (interferon induced protein with tetratricopeptide repeats 1) is a gene that encodes a protein that may inhibit viral replication and translational initiation (provided by RefSeq, Aug 2012). This gene is co-expressed with the *USP18* gene in seven experiments, including GSE16147_d5, GSE13824_d1, and GSE6740_d22. The query result can be visualized as a graph representing the gene co-expression network as well as a tabular format.

Gene set co-expression network

In this query mode, instead of single gene co-expression, a set of genes can be entered and the result displays a visual network representation. The associated table facilitates a comprehensive understanding of the co-expression patterns and experimental associations among the genes of interest. For instance, in the example shown in Fig. 8, we used the gene mentioned in the previous examples, *USP18*, along with other genes such as *STAT1*, *SP100, IFI35, APOBEC3A, MX1,* and *CXCR5* (adj_*p* value ≤ 0.05 and |log-FC|> 1). The query result reveals an interconnected network of genes associated with the immune response, particularly in defense against pathogens such as viruses. Additionally, the result is presented as a table indicating the experiments associated



Query type: Gene set co-expression network

Adjusted p-value ≤: 0.05 (0.00 – 1.00)

|Log-FC| ≥: 1.0 (0.0 – 11.4)

Display isolated nodes:
○ Yes
● No

Co-ocurrence threshold: 1

Gene symbol: USP18 × STAT1 × SP100 × IGF1R × IL18 × CXCR5 × CX3CR1 × APOBEC3A × APOBEC3F × IFIT1 × ISG15 × MX1 × IFI35 ×

```
Edge thickness in the graph indicates the frequency in which the genes appear co-expressed.
```

	gene_symbol_a	gene_symbol_b	experiment_list	frequency
20	IFIT1	USP18	GSE6740_d21, GSE6740_d22, GSE13824_d1, GSE6740_d18, GSE6740_d19, GSE16147_d5, GSE17626_d1	7
26	ISG15	MX1	GSE6740_d19, GSE6740_d18, GSE6740_d12, GSE6740_d21, GSE6740_d13, GSE6740_d22	6
31	IFI35	MX1	GSE6740_d19, GSE6740_d18, GSE6740_d12, GSE6740_d21, GSE6740_d13, GSE6740_d22	6
32	IFI35	ISG15	GSE6740_d13, GSE6740_d19, GSE6740_d12, GSE6740_d18, GSE6740_d22, GSE6740_d21	6

Fig. 8 Result of the query module 'Gene set co-expressed network' (adj_*p* value ≤ 0.05 and |log-FC|> 1) shows an interconnected network of genes linked to the immune response, particularly in defense against pathogens such as viruses. The table result provides information on the experiments associated with each pair of connections

with each pair of connections. The user can also choose if disconnected nodes, i.e., genes that are not co-expressed with other genes, are also displayed in the network representation.

Conclusions

We present the HIHISIV database that provides a comprehensive integrated view of immune host response in SIV and HIV hosts. The data made available through HIHISIV, following best practices for supporting the FAIR principles, is based on aggregation of metadata and a workflow for analyzing microarray and RNA-Seq datasets and annotations. The workflow identifies differentially expressed genes in the different studies analyzed and adds other types of interactions and relevant roles that these genes have. The HIHISIV database contains a web page with an easy-to-use interface for biologists to search and browse for genes and experimentally testable new hypotheses of molecular mechanisms related to the infection process in HIV/SIV and host types. The database also has additional information about viruses, documentation, experiments list, and external sources. Our objective is to continue the development of the HIHISIV around the querying, metadata, analysis functionality, and addition of new datasets from main transcriptome repositories.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05740-7.

Additional file 1 includes a table of the datasets/experiments in the current HIHISIV database (v2.0).

Additional file 2 comprises the database design and description details.

Acknowledgements
The authors would like to thank the Brazilian National Laboratory for Scientific Computing for providing computational resources used in the analyses.

Author contributions
RLC: methodology, conceptualization, software, writing—original draft, writing—review and editing; LG: methodology, conceptualization, software, writing—original draft, writing—review and editing; MDA: methodology, conceptualization, software, writing—original draft, writing—review and editing; MRA: methodology, conceptualization, writing—original draft, writing—review and editing, supervision; DLR: methodology, conceptualization, writing—original draft, writing—review and editing, supervision; JMS: methodology, conceptualization, writing—original draft, writing—review and editing, supervision; MAS: methodology, conceptualization, writing—original draft, writing—review and editing; FP: methodology, conceptualization, writing—original draft, writing—review and editing, supervision.

Funding
Open Access funding enabled and organized by Projekt DEAL. This work has been supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of GHGA – The German Human Genome-Phenome Archive (www.ghga.de, Grant Number 441914366 (NFDI 1/1)). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials
The HIHISIV v2.0 database is archived on Zenodo under doi: 10.5281/zenodo.7093185. It is available online for querying and download of query results at https://hihisiv.github.io.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors have no competing interests as defined by BMC, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Received: 25 November 2023 Accepted: 11 March 2024
Published online: 22 March 2024

References

1. Cichutek K, Norley S. Lack of immune suppression in SIV-infected natural hosts. AIDS. 1993;7:S25-36.
2. Rey-Cuillé MA, Berthier JL, Bomsel-Demontoy MC, Chaduc Y, Montagnier L, Hovanessian AG, et al. Simian immunodeficiency virus replicates to high levels in sooty mangabeys without inducing disease. J Virol. 1998;72(5):3872–86.
3. Chakrabarti LA, Lewin SR, Zhang L, Gettie A, Luckay A, Martin LN, et al. Normal T-cell turnover in sooty mangabeys harboring active simian immunodeficiency virus infection. J Virol. 2000;74(3):1209–23.
4. Hahn BH, Shaw GM, De KM, Cock, Sharp PM. AIDS as a zoonosis: scientific and public health implications. Science. 2000;287(5453):607–14.
5. Broussard SR, Staprans SI, White R, Whitehead EM, Feinberg MB, Allan JS. Simian immunodeficiency virus replicates to high levels in naturally infected African Green Monkeys without inducing immunologic or neurologic disease. J Virol. 2001;75(5):2262–75.
6. Silvestri G, Sodora DL, Koup RA, Paiardini M, O'Neil SP, McClure HM, et al. Nonpathogenic SIV infection of sooty mangabeys is characterized by limited bystander immunopathology despite chronic high-level viremia. Immunity. 2003;18(3):441–52.
7. Brenchley JM, Silvestri G, Douek DC. Nonprogressive and progressive primate immunodeficiency lentivirus infections. Immunity. 2010;32(6):737–42.
8. Brenchley JM, Paiardini M. Immunodeficiency lentiviral infections in natural and non-natural hosts. Blood. 2011;118(4):847–54.
9. Ota M, Fujio K. Multi-omics approach to precision medicine for immune-mediated diseases. Inflamm Regen. 2021;41(1):23.
10. Mohammadi P, Desfarges S, Bartha I, Joos B, Zangger N, Muñoz M, et al. 24 Hours in the life of HIV-1 in a T cell line. PLoS Pathog. 2013;9(1):e1003161. https://doi.org/10.1371/journal.ppat.1003161.
11. Bartha I, McLaren PJ, Ciuffi A, et al. GuavaH: a compendium of host genomic data in HIV biology and disease. Retrovirology. 2014;11:6. https://doi.org/10.1186/1742-4690-11-6.
12. Li C, Ramarathinam S, Revote J, et al. HIVed, a knowledgebase for differentially expressed human genes and proteins during HIV infection, replication and latency. Sci Rep. 2017;7:45509. https://doi.org/10.1038/srep45509.
13. Honnibal, M. spaCy (Version v3.0.6). Retrieved from https://spacy.io/
14. Gatto L. *Rols: an R interface to the ontology lookup service*. R package version 2.28.0 2023; http://lgatto.github.com/rols/.

15. Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG. The affymetrix GeneChip platform: an overview. Methods Enzymol. 2006;410:3–28. https://doi.org/10.1016/S0076-6879(06)10001-4.
16. Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy–analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307–15.
17. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47–e47.
18. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4(8):1184–91.
19. Austin JW, Buckner CM, Kardava L, Wang W, Zhang X, Melson VA, Swanson RG, Martins AJ, Zhou JQ, Hoehn KB, Fisk JN, Dimopoulos Y, Chassiakos A, O'Dell S, Smelkinson MG, Seamon CA, Kwan RW, Sneller MC, Pittaluga S, Doria-Rose NA, McDermott A, Li Y, Chun TW, Kleinstein SH, Tsang JS, Petrovas C, Moir S. Overexpression of T-bet in HIV infection is associated with accumulation of B cells outside germinal centers and poor affinity maturation. 2019. Sci Transl Med.;11(520):eaax0904.
20. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan M, Carey V. Software for computing and annotating genomic ranges. PLoS Computational Biology, 9; 2013.
21. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007;23(2):257–8. https://doi.org/10.1093/bioinformatics/btl567.
22. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9.
23. Timmons JA, Szkop KJ, Gallagher IJ. Multiple sources of bias confound functional enrichment analysis of global -omics data. Genome Biol. 2015;16:186. https://doi.org/10.1186/s13059-015-0761-7.
24. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
25. Guha RV, Brickley D, Macbeth S. Schema.org: evolution of structured data on the web. Commun ACM. 2016;59(2):44–51.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.