Briefings in Bioinformatics, 22(2), 2021, 832–844

doi: 10.1093/bib/bbaa425
Advance Access Publication Date: 30 January 2021
Problem Solving Protocol

HVIDB: a comprehensive database for human–virus protein–protein interactions

Xiaodi Yang, Xianyi Lian, Chen Fu, Stefan Wuchty, Shiping Yang and Ziding Zhang

Corresponding authors: Shiping Yang, State Key Laboratory of Plant Physiology and Biochemistry, College of Biological Sciences, China Agricultural University, Beijing 100193, China. Tel.: +86 10 62733780, E-mail: shi_ping_yang@163.com; Ziding Zhang, State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China. Tel.: +86 10 62734376, E-mail: zidingzhang@cau.edu.cn

Abstract

While leading to millions of people's deaths every year the treatment of viral infectious diseases remains a huge public health challenge.Therefore, an in-depth understanding of human–virus protein–protein interactions (PPIs) as the molecular interface between a virus and its host cell is of paramount importance to obtain new insights into the pathogenesis of viral infections and development of antiviral therapeutic treatments. However, current human–virus PPI database resources are incomplete, lack annotation and usually do not provide the opportunity to computationally predict human–virus PPIs. Here, we present the Human–Virus Interaction DataBase (HVIDB, http://zzdlab.com/hvidb/) that provides comprehensively annotated human–virus PPI data as well as seamlessly integrates online PPI prediction tools. Currently, HVIDB highlights 48 643 experimentally verified human–virus PPIs covering 35 virus families, 6633 virally targeted host complexes, 3572 host dependency/restriction factors as well as 911 experimentally verified/predicted 3D complex structures of human–virus PPIs. Furthermore, our database resource provides tissue-specific expression profiles of 6790 human genes that are targeted by viruses and 129 Gene Expression Omnibus series of differentially expressed genes post-viral infections. Based on these multifaceted and annotated data, our database allows the users to easily obtain reliable information about PPIs of various human viruses and conduct an in-depth analysis of their inherent biological significance. In particular, HVIDB also integrates well-performing machine learning models to predict interactions between the human host and viral proteins that are based on (i) sequence embedding techniques, (ii) interolog mapping and (iii) domain–domain interaction inference. We anticipate that HVIDB will serve as a one-stop knowledge base to further guide hypothesis-driven experimental efforts to investigate human–virus relationships.

Key words: human–virus interaction; database; protein–protein interaction; annotation; prediction

Xiaodi Yang is a PhD student at the State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University. Her current research interests include protein bioinformatics and machine learning.
Xianyi Lian is a PhD student at the State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University. Her current research interests include protein bioinformatics and machine learning.
Chen Fu is a PhD student at the State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University. Her current research interests include protein bioinformatics and machine learning.
Stefan Wuchty is an Associate Professor at the Department of Computer Science and Biology, and a member of the Institute of Data Science and Sylvester Comprehensive Cancer Center at the University of Miami. His research interests revolve around systems and network biology.
Shiping Yang is a post-doctoral fellow at the State Key Laboratory of Plant Physiology and Biochemistry, College of Biological Sciences, China Agricultural University. His current research interests include protein bioinformatics and plant genomics.
Ziding Zhang is a Professor at the State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University. His research interests are protein bioinformatics and systems biology.
Submitted: 23 September 2020; **Received (in revised form):** 12 November 2020

Introduction

Viral infectious diseases remain a major threat to public health around the world. One important milestone toward understanding the pathogenesis of viral infections and developing therapeutic strategies is to unravel protein–protein interactions (PPIs) between the human host and various viral proteins since human–virus PPIs directly reach into viral infection pathways and host immune responses [1]. As infections with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) have triggered a global pandemic, treating COVID-19 remains a challenge as a consequence of limited knowledge about the molecular details of the ways SARS-CoV-2 infects host cells [2–5]. Experimental and prediction techniques have been rapidly applied to decipher the interactome between humans and SARS-CoV-2 [2–6], which has become an important entry point to explore the pathogenesis of SARS-CoV-2, identify potential drug targets and develop effective antiviral drugs [2, 4, 7, 8].

In past decades, low-throughput [e.g. co-immunoprecipitation (Co-IP)] and high-throughput [e.g. mass spectrometry (MS) and yeast two-hybrid (Y2H)] techniques [9–11] allowed the determination of human–virus PPIs on an unprecedented scale, providing an abundance of PPI data that have been stored in a series of state-of-the-art host-pathogen protein interaction databases. For example, VirHostNet [12] integrates intra- and inter-species (i.e. human–human, virus–virus and human–virus) PPIs and provides table-based and graph-based network visualization. VirusMentha [13] curates non-redundant host-virus PPI data and provides weekly automated data updates. PHISTO [14] incorporates host-pathogen interaction related information and allows users to access the functional and network topological properties of pathogen targeted human proteins. Following the strategy used in the STRING database [15], Viruses.STRING [16] quantitatively measures the reliability of interactions between viral and human proteins. In addition, some database resources are only designed for an individual virus species such as the NCBI HIV-1 Human Interaction Database [17], DenHunt [18] and HCVpro [19]. Considering the tremendous value of human–virus PPI data, the development of more advanced data resources to facilitate the research community is still desirable.

Current technical advances have accelerated the accumulation of experimental human–virus PPI data, providing an unprecedented opportunity for the development of reliable computational methods to predict human–virus PPIs. Computationally cost-effective PPI prediction methods can complement experimental efforts and allow us to capture the global landscape of human–virus interactomes more rapidly. Methods such as interolog mapping [20–23], domain–domain/motif interaction (DDI/DMI) inference [24–26], structural homology-based method [27, 28] and machine learning (ML) approaches [29–35] have been widely employed to predict potential interactions on a proteome-wide scale. Undoubtedly, providing prediction methods online in some human–virus PPI data resources can effectively help the users to computationally predict potential PPIs between a virus and the human host in the absence of abundant experimental evidence.

Here, we introduce a comprehensive human–virus PPI database, HVIDB (http://zzdlab.com/hvidb/), that (i) combines multiple human–virus PPI data resources and (ii) provides a computational platform to predict interactions between human and viral proteins. Designed as a powerful one-stop resource for human–virus interactions, our database components provide experimentally verified PPIs, 3D complex structures of protein interactions, virally targeted human complex information and manually collected host factor data. As for auxiliary data that revolve around human–virus interactions, our database further integrates and annotates interactions with differential expression information of human genes post-viral infections, human tissue-specific gene expression profiles and functional enrichment analysis of virus-targeted human proteins. Furthermore, our prediction platform integrates three state-of-the-art prediction methods including interolog mapping, domain–domain interaction (DDI) inference and a novel ML approach based on our previous work [36] that adopts a sequence embedding-based random forest method (i.e. doc2vec + RF).

Materials and methods

Collection of experimentally verified human–virus PPI data

We collected experimentally verified PPIs from five public databases (i.e. HPIDB [37], PHISTO [14], VirHostNet [12], VirusMentha [13] and PDB [38]) and recently published literature [6], allowing us to obtain 48 643 human–virus PPIs after removing self, genetic and redundant interactions, including 303 human-SARS-CoV-2 PPIs [6]. In particular, we mapped protein IDs from different databases to UniProt IDs, Entrez Gene IDs, gene names or protein names as query options in HVIDB.

Three-dimensional complex structures of human–virus PPIs

Three-dimensional complex structures in the HVIDB database were collected and processed from human–virus experimental complex structures in the PDB database [38] or predicted by using homology modeling of protein complexes (HMPC) [39, 40] (Supplementary Figure S1). Briefly, HMPC mainly captures (i) homologous template selection, (ii) monomer modeling and (iii) complex modeling. First, we selected the best homologous template for each protein by BLAST searching the PDB database. In particular, we set the thresholds to 30% sequence identity and 40% alignment coverage [39, 40] and considered the template candidate with the highest sequence identity as the best template. Note, that HMPC requires two monomer templates of each protein involved in a PPI that belong to different chains of the same protein complex in PDB. Second, we constructed a monomer model for each protein based on the selected template through Modeller (version 9.19) [41]. Finally, we combined the two monomer models into the final complex structure and calculated protein interaction sites based on these experimental/predicted complex structures, that are available to users for download. Further methodological details of complex structure construction and interaction site calculation are available in our previous publication [40].

Virally targeted human protein complexes

Collecting human protein complexes from CORUM (http://mips.helmholtz-muenchen.de/corum/) [42] and hu.MAP (http://hu.proteincomplexes.org/) [43], we obtained 2923 and 4588 human protein complexes, respectively. While complexes in CORUM are experimentally determined, hu.MAP also contains many complexes reconstructed through ML methods, providing better coverage. Pointing to virally targeted human protein complexes, we mapped human–virus PPIs to human protein complexes and

found 6633 virally targeted human protein complexes, revolving around 37 094 human–virus PPIs.

Differential expression analysis

To identify differentially expressed genes (DEGs) post-viral infections, we first manually screened human gene expression series of microarray/RNA-Seq experiments in human tissues and cell lines that were infected with viruses from the Gene Expression Omnibus (GEO) [44] and the Sequence Read Archive [45].

Each GEO series contains multiple expression samples capturing control and different infection conditions (e.g. tissues, cells, virus species and infection time points). In the same expression series, we further manually binned samples into different control-infection groups, that refer to specific infection conditions, and calculated corresponding DEGs. Following the strategy used in [46], we also manually curated different control-infection groups, capturing GEO series/platform, PubMed ID, tissues/cells, viral species and viral families. To ensure the quality of expression data, we only retained published GEO expression samples of both microarray and RNA-Seq experiments. As for microarray experiments, we only retained single-channel and discarded dual-channel microarray data. These manual filtering/annotation steps and DEG calculations allowed us to obtain DEGs in 411 control-infection groups from 95 microarray GEO series and 121 control-infection groups from 34 RNA-Seq GEO series, which include 4453 public human expression samples covering the infections of 20 viral families.

Microarray data

After normalization and $\log_2$ transformation of microarray expression values, DEGs were determined through the R package 'limma' [47], defined as genes with $|\log_2 FC| \geq 1.5$ and false discovery rate (FDR)-adjusted P-value ≤ 0.05. Note that all gene probes were converted to Entrez Gene IDs directly or by using the Ensembl BioMart tool [48]. If multiple probe sets were mapped to the same gene (i.e. a gene has been determined multiple times), we averaged corresponding gene expression values.

RNA-Seq data

To control the quality of reads, we first removed adapter sequences and low-quality ends with Trimmomatic [49] and aligned trimmed reads of each sample to the human h38 reference genome as of GENCODE [50] through HISAT2 [51]. Subsequently, we used StringTie [52] to assemble the transcriptome of each sample and estimated the expression levels of all genes. Furthermore, we utilized StringTie's script 'prepDE.py' to determine raw gene counts and determined DEGs through the R package 'DESeq2' [53], where we considered genes as differentially expressed when $|\log_2 FC| \geq 1.5$ and FDR-adjusted P-value ≤ 0.05. Note that all genes were represented by their Ensembl IDs.

Tissue-specific expression for virally targeted human proteins

While tissue-specific transcript-level expression values of 37 tissues were extracted from the Human Protein Atlas database [54], the prevalence of a gene expression in a (given group of) tissue(s) was categorized through 'tissue enriched,' 'group enriched,' 'tissue enhanced,' 'low tissue specificity' and 'not detected.' Finally, such human protein-coding gene expression data were mapped to human–virus PPIs providing tissue-specific expression information for each virally targeted human protein.

Enrichment analysis of virally targeted human proteins

Gene Ontology enrichment analysis

Gene Ontology (GO) annotation data of human proteins were downloaded from http://current.geneontology.org/ [55]. Using all human proteins mapped to three GO terms categories [i.e. cellular component, biological process (BP) and molecular function] as reference sets, enriched GO terms of viral targets were determined by hypergeometric tests, where corresponding P-values were Bonferroni corrected.

Kyoto Encyclopedia of Genes and Genomes pathway enrichment analysis. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway data were downloaded from https://www.genome.jp/kegg/ [56]. Using all human proteins in KEGG pathways as a reference set, enriched KEGG pathways of viral targets were identified by hypergeometric tests where corresponding P-values were Bonferroni corrected.

For each virally targeted human protein set, the above two types of enrichment analyses were preprocessed and corresponding results were deposited in HVIDB.

Host factors

To further understand how the human host responds to viral infections, we manually collected 3572 host factors of the human immunodeficiency virus (HIV), human papillomavirus (HPV), dengue virus (DENV), zika virus (ZIKV), ebola virus (EBOV) and influenza A virus subtype H1N1 from the literature. Host factors include host dependency factors (HDFs), that help viruses to infect human host cells, and host restriction factors (HRFs) that restrict viruses to perform their functions (e.g. hinder and limit viruses to invade human host cells or tissues). To investigate whether curated host factors are viral targets, we mapped 3572 host factors (2768 HDFs and 804 HRFs) to human–virus PPIs and obtained 6056 human–virus PPIs that involved host factors. In particular, host factors are shown not only at the viral target level but also at the virally targeted human complex level. Moreover, the complete host factor list can be obtained through the Download page of HVIDB.

Computational tools to predict human–virus PPIs

To determine whether a given human–virus protein pair interacts, we utilized a predictive framework integrating three individual prediction methods to detect potential interactions (Supplementary Figure S2). Sample selection and prediction methods are elaborated below.

Sample selection

To establish a gold-standard of high-quality PPIs, we excluded PPIs from large-scale MS experiments with only one experimental observation. Moreover, redundant PPIs and interactions between proteins with less than 30 amino acids, more than 5000 amino acids or non-standard amino acids were removed, allowing us to obtain 31 383 human–virus PPIs. Utilizing 'Dissimilarity-Based Negative Sampling' [29, 36], we compiled a negative PPI set that was 10 times larger than the positive sample set. Briefly, this negative sampling strategy stipulates

that a protein pair C–B should not be selected as a negative sample if viral proteins A and B have similar sequences, and A interacts with human protein C (i.e. protein pair C–A is a positive sample). Sets of 31 383 positive and 313 830 negative samples, thus compiled, were further used for model training and assessment.

Interolog mapping method

The core idea of interolog mapping is that a potential interaction between protein A and B occurs if their respective homologs A′ and B′ interact, defined as the interolog template of A–B. Briefly, we first collected known PPIs including all intra- and inter-species interactions from five popular protein interaction databases including IntAct [57], BioGrid [58], MINT [59], DIP [60] and HPIDB [37] to obtain an interolog template library. Subsequently, we employed the scoring scheme of HIPPIE [61] to assess the template quality. We assigned a quality score $S_{template}$ to each interolog template according to the experimental detection techniques used as well as the number of references reporting the PPI and the number of involved species. To identify interologs of a given human–virus protein pair, BLAST is employed to search for homologs by aligning all the sequences in the PPI template library. Specifically, we considered homologs if their sequence identity $\geq$0.3 and the alignment coverage of query protein $\geq$0.4. Similar to our previous work [62], the final interaction probability of a protein pair ($S_{Interolog}$) combines all the quality scores of the identified interolog templates through Bayes integration, $S_{Interolog} = 1 - \prod_{i=1}^{n}(1 - S_{template})$, where n is the number of involved interolog templates of the query PPI.

DDI inference method

The DDI inference method predicts the interaction probability of a query protein pair based on the detected interacting domain pairs. Briefly, we scanned each interacting protein for the presence of Pfam protein domains using HMMER [63] (E-value $\leq 10^{-5}$). Subsequently, we obtained co-occurrence domain pairs from known protein interactions to form a comprehensive DDI library. Domains of query human–virus protein pairs were also retrieved by searching the Pfam database [64] with the same threshold. Similar to the interolog mapping method, each domain pair in the DDI library was assigned a confidence score S_{EM} through the expectation maximization (EM) algorithm [65]. Finally, the interaction probability score of a protein pair (S_{DDI}) was determined by integrating the confident scores of DDIs involved in the query protein pairs through Bayes method, $S_{DDI} = 1 - \prod_{i=1}^{n}(1 - S_{EM})$, where n is the number of involved DDIs for the query PPI.

ML method

We employed our implemented doc2vec + RF approach to predict interactions between human and viral proteins. Specifically, we applied the doc2vec style learning approach that allowed us to effectively capture contextual information of interacting protein sequences through a 32-dimensional feature vector (details of the doc2vec + RF method are available in our previous study [36]). Based on such an interaction representation, we utilized the random forest algorithm to determine an interaction probability score (S_{ML}) ranging from 0 to 1.

To maximize prediction performance, we combined these individual scores (i.e. $S_{Interolog}$, S_{DDI} and S_{ML}) in a vector and trained a logistic regression model, reflecting the overall interaction probability for each human–virus protein pair.

Database construction

HVIDB is based on CentOS 7.4, Apache 2.4.6, MySQL 5.5.60 and PHP 5.4.16. The user interface charts and tables were generated based on several Javascript-based libraries, such as DataTable.js and echarts.js. A Javascript graph library Cytoscape.js [66] was used to display PPI networks. NGL [67], a WebGL-based 3D viewer, was utilized to display 3D complex structures of PPIs.

Results and discussion

Overall description of HVIDB

As major components, the HVIDB online resource includes a comprehensive data module and an online prediction platform for human–virus protein interactions. Currently, HVIDB provides 48 643 experimentally verified human–virus PPIs, 6633 virally targeted human host complexes, 3572 host factors, 6790 human tissue-specific gene expression profiles of viral targets, DEGs post-viral infections from 129 GEO series as well as 911 (474 experimentally verified and 437 predicted) 3D complex structures of human–virus PPIs and their corresponding 3D interaction sites. The main architecture of HVIDB (Figure 1) contains the following key features: (i) as for network information, HVIDB provides an associated human–virus PPI subnetwork for each query human/viral protein. Considering that viruses tend to target human complexes, human complex information is integrated into the identified subnetwork for an improved mechanistic understanding of viral infections. (ii) As for structural information, HVIDB provides experimentally verified/predicted 3D complex structure visualization of human–virus PPIs and corresponding interaction sites. (iii) HVIDB provides differential expression information of viral targets providing an in-depth understanding of the ways human host genes respond to viral infections. Tissue-specific gene expression of viral targets is further provided to indicate tissue specificity of the corresponding human–virus PPIs post-viral infections. (iv) As for functional annotations, enrichments of GO terms and KEGG pathways are provided for virally targeted human proteins when querying a viral protein. Moreover, our manually curated host dependency/restriction factors are mapped onto human–virus PPIs and virally targeted human complexes, which can be utilized as a complement to understand the functional roles of virally targeted human proteins. (v) HVIDB provides a comprehensive PPI prediction platform to rapidly predict potential interactions between query human–virus protein pairs in the absence of experimental information.

Searching interface for PPI associated resources and usability

HVIDB provides multiple searching/browsing modules, allowing users to easily access our multifaceted data (Figure 2). Users can center a search around human/viral proteins to obtain corresponding human–virus PPI networks (Figure 2A). Specifically, entries can be searched through various gene/protein IDs, symbols or keywords (e.g. UniProt ID, gene name and protein name). HVIDB also allows users to explore PPI data and associated auxiliary information by browsing through corresponding lists, covering human–virus PPIs, 3D complex structural information, virally targeted human complexes, differential expression information and host factors (Figure 2B). Furthermore, HVIDB cross-links PPI information with targets of other single viruses or viral

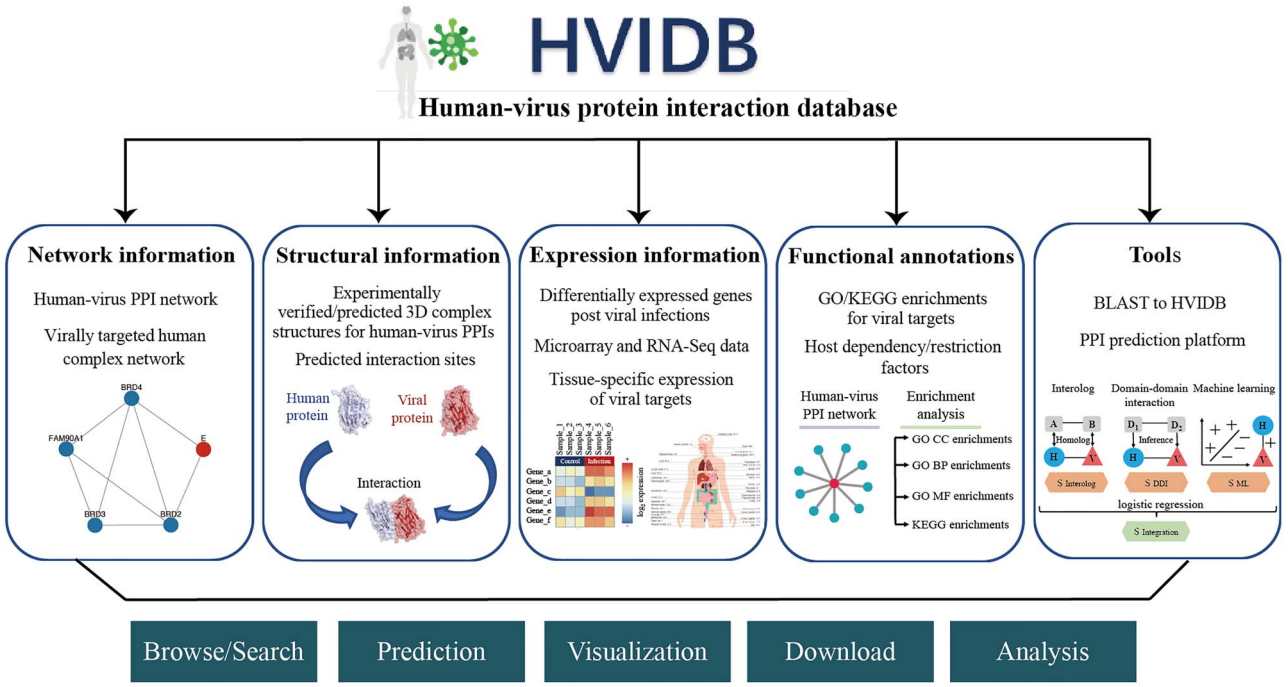

Figure 1. Basic functions of HVIDB. HVIDB provides information about **(i)** networks between viral and human proteins, **(ii)** structural characteristics of interactions, **(iii)** expression levels, **(iv)** functional annotations of targeted human proteins and **(v)** tools to predict PPIs. Specifically, HVIDB collects experimentally verified human–virus PPIs from expert-curated source databases and literature, as well as virally targeted human complexes. Structural information includes experimental/predicted 3D complex structures and interaction sites of human–virus PPIs. Expression information captures large-scale DEG data post-viral infections and tissue-specific gene expression information of viral targets while functional annotations mainly provide GO/KEGG enrichments of viral targets. HVIDB also provides manually curated host factor data as a complement to explore the functional roles of viral targets. High-performing human–virus PPI prediction methods are integrated in HVIDB, allowing the users to find potential interactions between query viral and human host proteins.

families (Figure 2B). In addition to protein and PPI based queries, HVIDB provides two individual modules for accessing DEGs post-viral infections and virally targeted human complexes. In particular, users can access DEG data through an individual interface (Figure 2C) by inputting/selecting a GEO series of cell lines/tissues with different time points post infections by viral families or data type (i.e. microarray data or RNA-Seq data). In a different way to search HVIDB, users can input human/viral protein or targeted complex ID/names to obtain relevant complex information (Figure 2D). Finally, HVIDB provides the opportunity to predict the presence of an interaction between a user-provided pair of human and viral proteins in FASTA format (Figure 2E).

In a concrete example that revolves around a single protein of interest, we use HVIDB to find PPIs that involve the influenza protein 'NS' using the UniProt ID 'P03495' as search term (Figure 3). HVIDB presents corresponding search results on separate pages, including visualization of the corresponding PPI network, enriched GO terms/KEGG pathways and interaction information table (Figure 3A). In particular, HVIDB provides detailed functional and structural information about the interacting protein (Figure 3B) as well as PPIs capturing interacting protein/PPI basic information, PPI comprehensive score, 3D complex structure/interaction sites, virally targeted human complexes, DEGs post-viral infections and tissue-specific expression information (Figure 3C–H). In addition, users can download related datasets from the Download Page in HVIDB for local use.

Performance of online PPI prediction platform

While we employed three individual methods to predict human–virus PPIs, we improved the prediction performance by combining results of the individual prediction methods through a logistic regression model. To comprehensively assess the method's performance, we randomly sampled 80% of all human–virus PPIs as a training dataset and considered the remaining 20% as an independent test set. Assessing the performance of the three individual methods and the logistic regression integration method through the area under the precision-recall curves (AUPRC), we observed that the integrative model (AUPRC = 0.877) slightly outperformed the best individual method (i.e. doc2vec + RF method, AUPRC = 0.866) (Figure 4A). At a recall control of 60%, the corresponding precision values of these four methods (i.e. interolog, DDI, ML and logistic regression) were 47.46%, 40.04%, 93.15% and 96.69%, respectively. However, query viral proteins may not occur in the training set in real applications. To provide a more rigorous performance assessment, we sampled a training set with 80% of all PPIs, assuring that viral proteins in the training data did not occur in the testing data set. As expected, the performance of each individual method in this strict testing framework decreased considerably (Figure 4B). However, the AUPRC of the integrative model still outperformed the best individual method by a larger margin (Figure 4B), further indicating the relevance of integrating individual prediction approaches. At a recall control of 60%, the corresponding precision values of these four methods (i.e. interolog, DDI, ML and logistic regression) were 43.75%, 35.18%, 65.38% and 79.55%, respectively. As an assessment of their reliability, HVIDB provides the corresponding individual prediction scores (i.e. $S_{Interolog}$, S_{DDI} and S_{ML}) and the logistic regression integrative score for each curated ' human–virus PPI.

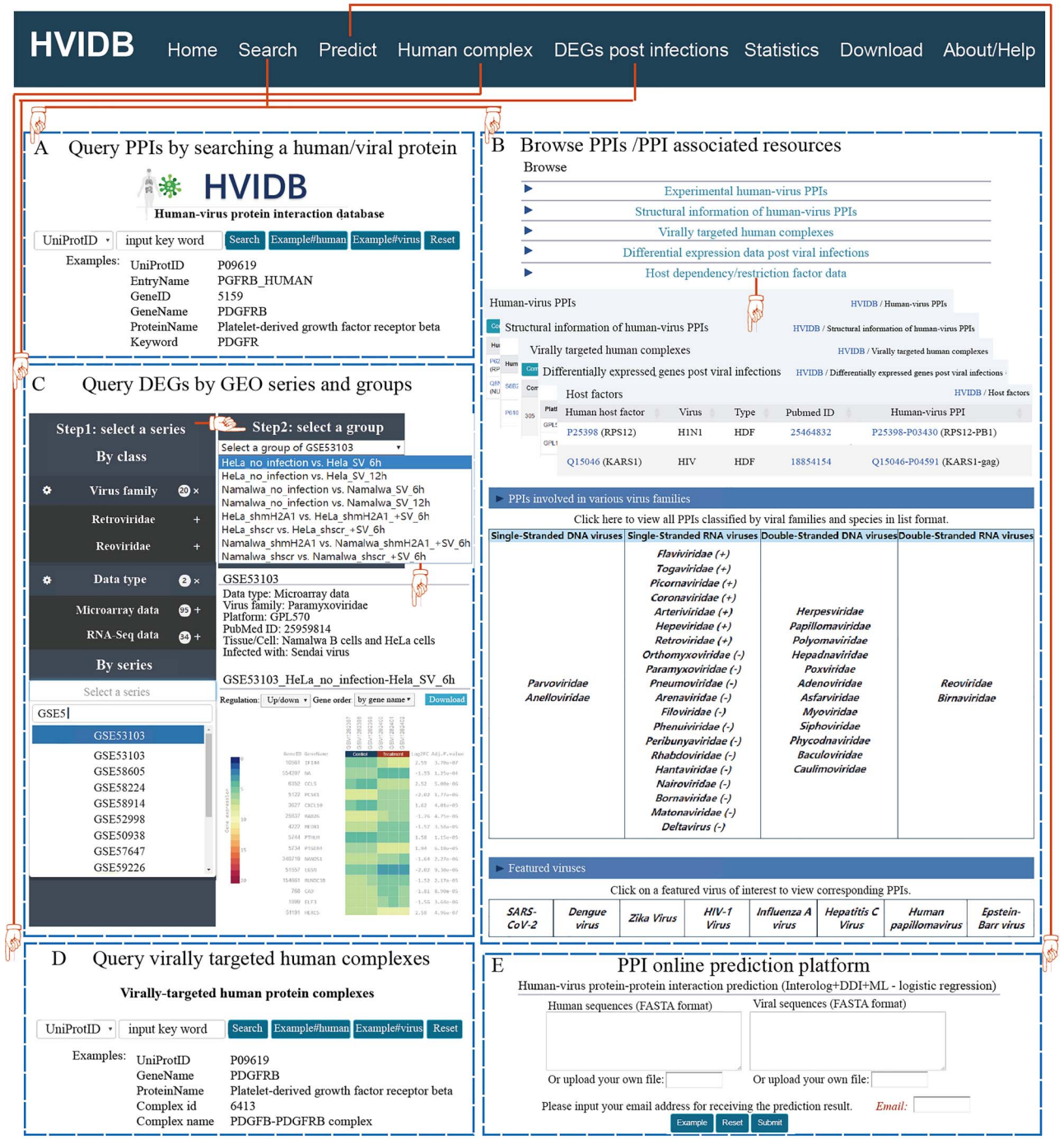

Figure 2. Main search and browse interfaces of HVIDB. (**A**) The main interface provides access to PPIs and corresponding metadata through querying a human host or viral protein. (**B**) Obtained human-virus PPIs are augmented with gene annotations, structural information of corresponding PPIs, virally targeted human complexes, host factors and DEGs in different tissues and cell lines post infections. Furthermore, HVIDB also allows users to find human genes that are targeted by other viruses (families). (**C**) DEGs post-viral infections can be queried through selecting/inputting a GEO series and a control-infection group. (**D**) Furthermore, HVIDB allows the users to search for virally targeted human complexes through a query with a human/viral protein or complex ID/name. (**E**) Providing a prediction platform, HVIDB allows the prediction of a potential interaction between a human and a viral protein query.

Human-ZIKV PPI analysis as an application case of HVIDB

To allow a more comprehensive understanding of the functionality of HVIDB, we present the results of a case study of human-ZIKV PPIs using HVIDB (Figure 5A). ZIKV is a mosquito-borne flavivirus transmitted through mosquitoes, resulting in serious complications, such as birth defects in babies and spontaneous loss of fetuses. Even worse, currently, there are no effective drugs available for clinical treatment. Although experimentally determined human-ZIKV PPIs appear in HVIDB, the putative number of the interactome is unknown. Therefore, we leveraged the prediction platform of HVIDB to find PPIs between 10 known

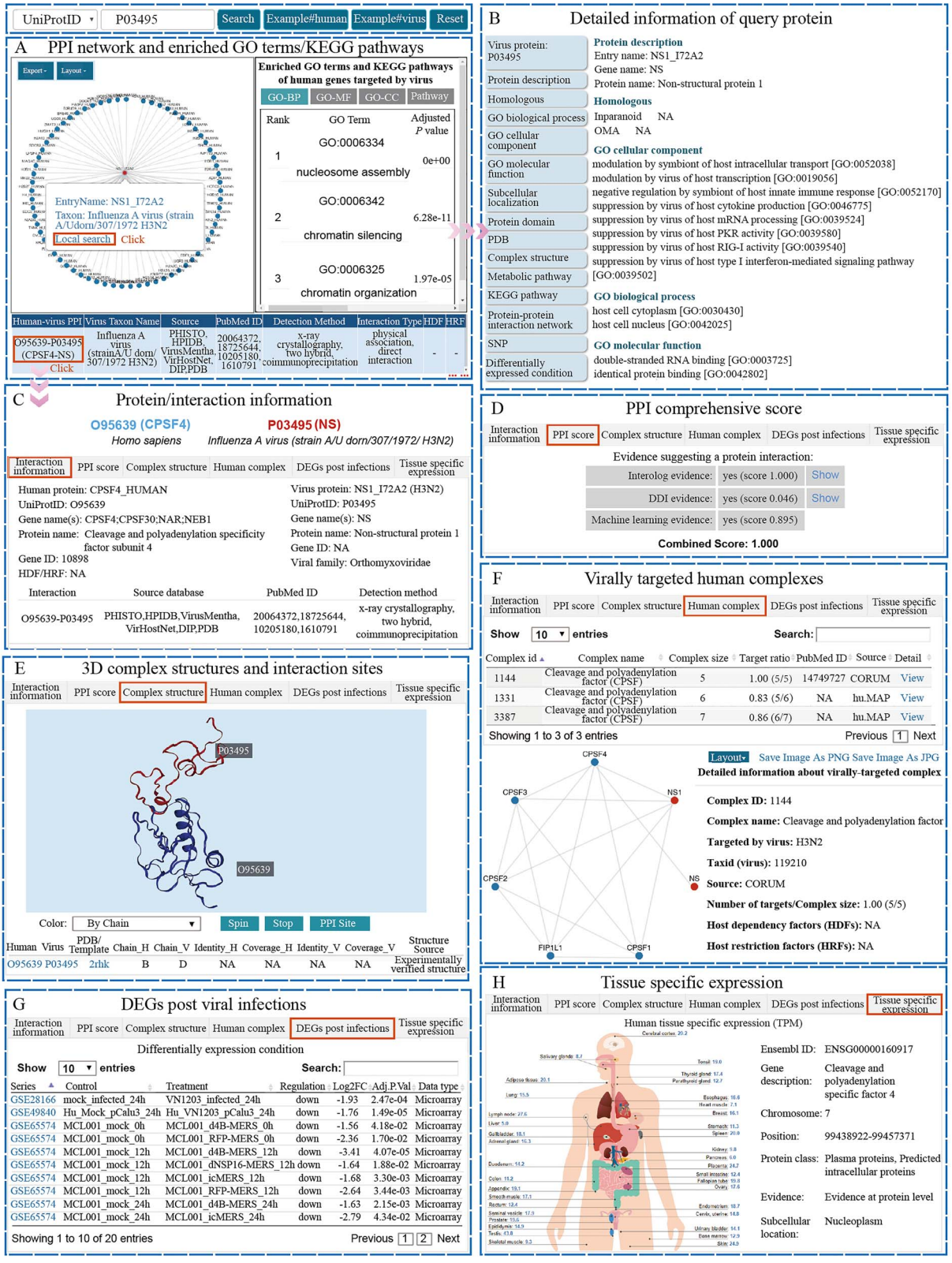

Figure 3. Search results with example query 'Influenza A virus NS protein.' In (**A**), we show the network representation and a scroll-down list of the corresponding interactions with targeted human proteins that are further analyzed with enriched GO terms and presence in KEGG pathways. (**B**) indicates detailed protein-specific functional and structural annotations of the query viral protein NS. As for an example of an interaction that involves NS, we consider CPSF4-NS. Specifically, HVIDB provides detailed multimodal information, including (**C**) basic meta-information, (**D**) reliability, (**E**) structural characteristics of the underlying PPI, as well as (**F**) protein complexes that CPSF4 is involved in, (**G**) differential expression levels and (**H**) tissue specificity of CPSF4.

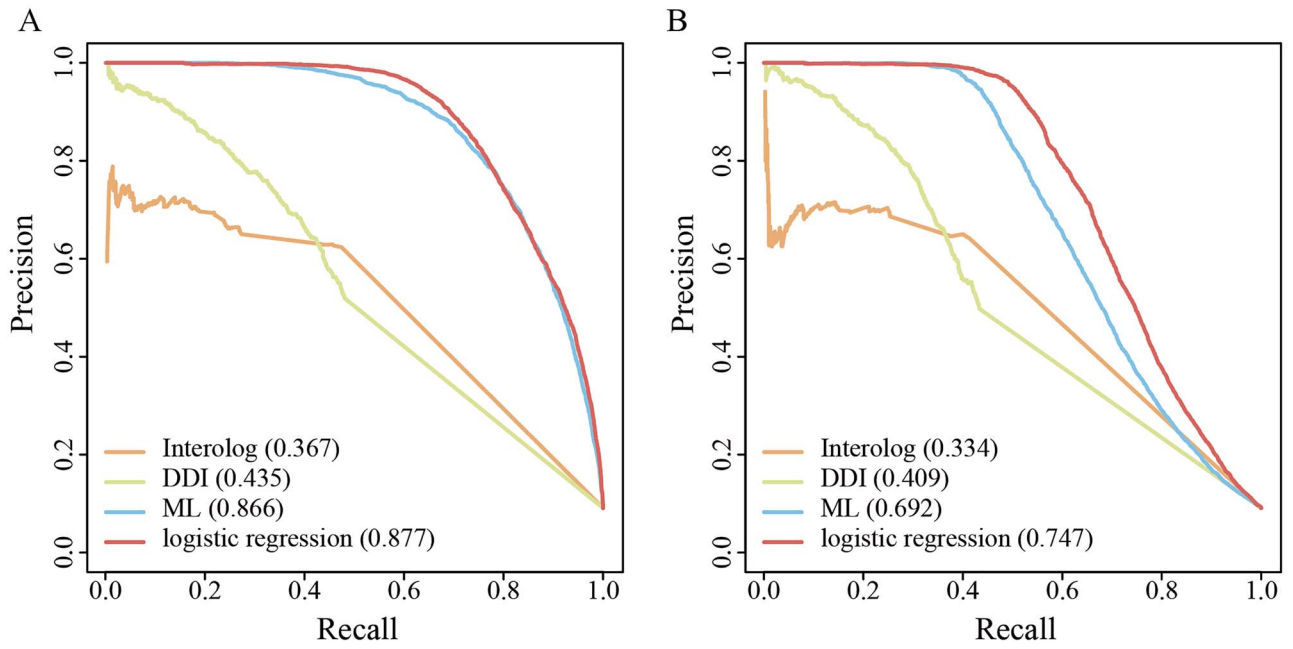

Figure 4. Performance of PPI prediction methods in HVIDB. **(A)** Randomly sampling 80% of all human-virus interactions as training data we assessed the performance of interolog mapping, DDI, and our ML methods as well as their integration through a logistic regression model by using the remaining 20% as testing data. AUPRC clearly shows that our doc2vec + RF ML approach outperformed other methods. **(B)** Refining our training dataset by only considering viral proteins that do not appear in the test data, we corroborate our initial observation. However, we also observe that the integration of all prediction methods through logistic regression offered a competitive prediction edge.

ZIKV proteins and 20 262 reviewed human proteins as annotated by SwissProt. Considering all sequence pairs of human-ZIKV proteins as input, we predicted the potential interactions through three individual methods and our integrative model. In particular, we selected the top 2000 predicted human-ZIKV interactions as high-confidence interactions as they pertain to a false positive rate <0.01 for follow-up analysis according to the logistic regression integration score. Notably, we found that a substantial fraction of 477 non-redundant, experimentally known human-ZIKV PPIs in HVIDB were confirmed by our predictions (Figure 5B). We further employed the rich data resources in HVIDB to analyze the combined set of 2102 human-ZIKV PPIs (Figure 5C; Supplementary Table S2). The enriched GO terms for virally targeted human proteins are shown in Supplementary Table S3, as a function of interactions with any given ZIKV protein. In general, ZIKV targeted human proteins are enriched for functions associated with apoptosis, cell cycle and immune response, which is consistent with known biological processes associated with ZIKV infections [68–70]. To investigate the human host responses to viral infections, we extracted RNA-Seq data related to ZIKV infections in HVIDB (GEO accession number: GSE93385), where the expression samples were annotated from primary human fetal brain-derived neural stem cells (cell line G010, K054 and K048), and mapped the corresponding DEGs to the 2102 human-ZIKV PPIs to obtain 32 up- and 16 down-regulated human genes post infections with ZIKV (Supplementary Table S4).

Furthermore, HVIDB allows the functional analysis of the targets of ZIKV non-structural protein NS5, showing that functions such as 'mRNA splicing, via spliceosome,' 'mRNA export from nucleus' and 'regulation of mRNA stability' are significantly enriched (Figure 5D; Supplementary Table S3), which is in line with nuclear localization of NS5. Previous observations indicate that NS5 in another flavivirus DENV inhibits the splicing of host mRNA [69], suggesting that ZIKV NS5 possibly also inhibits the splicing of host mRNA. Moreover, NS5 targets the interferon-stimulated gene factor 3 transcription complex (ISGF3) through an experimentally verified interaction with STAT2, which is a human host factor. The experimental verified viral target STAT2 is also identified by our prediction method (Figure 5E). Notably, we also predicted an interaction between NS5 and STAT1 in ISGF3. This innate immune-related protein has been identified as a drug target in TTD (http://db.idrblab.net/ttd) (Figure 5E) and can serve as a potential drug target for anti-ZIKV therapeutics. By further considering transcriptional regulation data, we hypothesize that NS5 may target the complex and results in the up-regulation of all subunits to inhibit the recruitment of interferon-stimulated gene factors (Figure 5E).

Comparison of HVIDB to existing human–virus PPI databases

In recent years, the rapid generation of large-scale human–virus PPI led to the accumulation of multifaceted data that revolves around human–virus interactomes. While a plethora of human–virus PPI databases have been released, we developed HVIDB as a source to provide comprehensively and thoroughly annotated human–virus PPI data. To compare HVIDB with other contemporary databases, we provide brief descriptions of some representative databases in Table 1. While impressive and practical in their own right, these contemporary database resources mainly represent storage platforms for human–virus PPIs, lacking detailed PPI-specific multimodal annotations that are conducive to further investigations of the pathogenesis of viral infections and specific dynamic mechanism of host immune responses. In addition, some databases only focus on a special

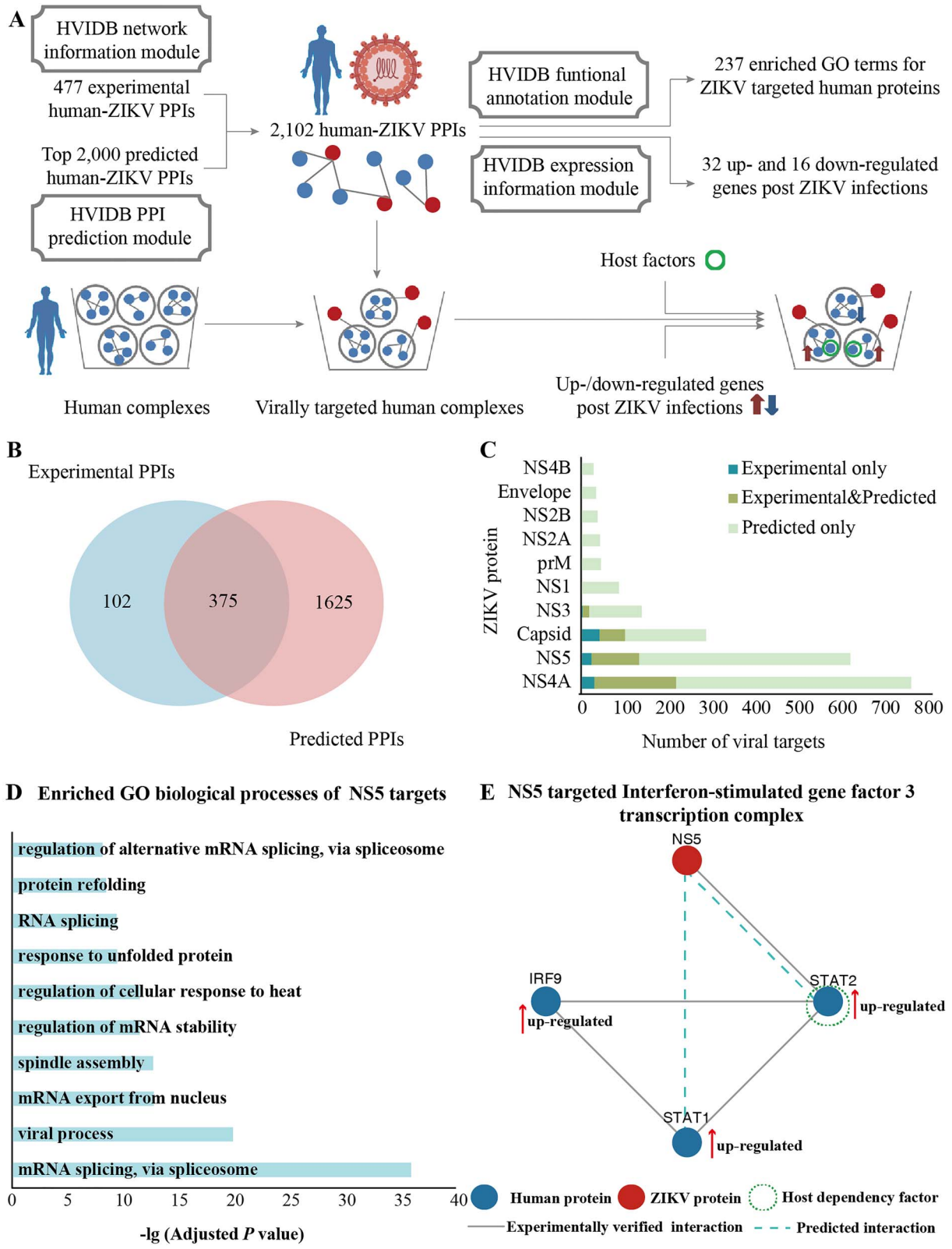

Figure 5. Case study of human-ZIKV PPIs. Focusing on the Zika virus, we show ways HVIDB can be employed to analyze the network of interactions between Zika and human host proteins. (**A**) HVIDB allowed us to find 2102 experimentally confirmed or predicted PPIs. Utilizing the functional and expression information in HVIDB we found 237 enriched GO terms and 48 differentially expressed human genes that were targeted by ZIKV. Furthermore, HVIDB allowed us to find viral targets in human protein complexes and annotate such genes as host factors. (**B**) The Venn diagram indicates that the prediction methods in HVIDB confirmed a substantial fraction of experimentally known PPIs. In (**C**), we show the number of experimentally verified and predicted human-ZIKV PPIs for each ZIKV protein. (**D**) As HVIDB provides the functional characterization of targets, we show the top 10 enriched GO–BP terms of NS5 targeted human proteins. (**E**) Determining a subnetwork of viral protein NS5 and its targeted interferon-stimulated gene factor 3 transcription complex, we used HVIDB to fully annotate the underlying PPIs and targeted proteins.

Table 1. Existing databases and resources related to human–virus PPIs

Type	Name	Host	Pathogen	Description	URL
Species-specific database	HCVpro	Only human	Only HCV	Numerous cross-reference links and extended reviews about HCV proteins	http://cbrc.kaust.edu.sa/hcvpro/
	NCBI HIV-1 Human Interaction Database	Only human	Only HIV-1	Sequence information and brief description of the interactions	http://www.ncbi.nlm.nih.gov/genome/viruses/retroviruses/hiv-1/interactions/
	DenHunt	Only human	Only DENV	Differentially expressed human genes post DENV infections.	http://proline.biochem.iisc.ernet.in/DenHunt/
Generic database	VirHostNet	Animal, human, plant	Only viruses	Interface website provides visualized protein interaction networks and associated experimental evidence	http://virhostnet.prabi.fr/
	VirusMentha	All hosts	Only viruses	Visualized interaction networks and weekly updated data. Mainly designed for PPI data curation	http://virusmentha.uniroma2.it/
	HPIDB	Animal, human, plant	Bacteria, fungi, viruses	Relatively comprehensive interaction information and BLAST tool to obtain homologous PPIs. Mainly designed for PPI data curation	https://hpidb.igbb.msstate.edu/index.html
	PHISTO	Only human	Bacteria, fungi, protozoa, viruses	GO/KEGG enrichments for queries of virally targeted human proteins. Mainly designed for PPI data curation/search	http://www.phisto.org/
	Viruses.STRING	All hosts	Only viruses	Scoring scheme refers to STRING database. Focuses on some specific viral strain	http://viruses.string-db.org/

virus species, and thus their applications are somehow limited. Comparatively, our proposed HVIDB focuses on all available protein interactions between human and any virus species and provides a large amount of multifaceted auxiliary information that is associated with human–virus PPIs. Based on these multifaceted annotation data, users can easily obtain reliable information on human–virus PPIs as well as conduct further in-depth functional analysis. Most strikingly, contemporary databases generally lack online PPI prediction tools, hampering users to access the complete, putative host-virus interactomes. To the best of our knowledge, HPIDB [37] provides a simple tool to predict potential interactions which only allows users to find homologous host-pathogen PPIs through basic sequence alignments. In contrast to the majority of contemporary databases, HVIDB also integrates an ML model using sequence embedding technique and two state-of-the-art PPI inference methods (i.e. interolog mapping and DDI inference method) that allow the reliable prediction of putative interactions between host and viral proteins. By considering the above advantages of HVIDB, we are confident that HVIDB can become a competitive data resource for human–virus interactomes.

Conclusions

HVIDB is a freely accessible resource providing comprehensive, downloadable human–virus PPI data to maximize its application. In contrast to existing human–virus PPI databases, HVIDB seamlessly integrates multifaceted data resources for human–virus PPIs, allowing users to explore the corresponding biological applications of human–virus interactomes. Furthermore, HVIDB provides an integrative human–virus PPI prediction platform, enabling users to accurately predict new interactions between human and viral proteins of interest. Our case study indicated the ease HVIDB allowed us to rapidly provide a more comprehensive landscape of the human-ZIKV interactome. Regarding future development, we will upgrade HVIDB regularly through integrating more data resources, developing better prediction tools and designing simple and more user-friendly interactive analysis interfaces. For instance, we will add known drug targets or drug information of viral targets to HVIDB, potentially accelerating drug target discovery and drug repositioning to combat deadly virally induced diseases. Furthermore, our current deep learning model to predict host-virus PPIs will serve as a starting point to develop a more powerful human–virus PPI prediction platform. Taken together, HVIDB can serve as a one-stop knowledge base to further guide hypothesis-driven experimental efforts to understand human–virus relationships and to develop antiviral treatments to tackle the continuous challenge of viral infections.

Key points

- HVIDB is a comprehensive web-based human–virus PPI data resource, providing rich human–virus PPI annotation data.
- HVIDB allows the users to explore corresponding biological implications of human–virus interactomes, while seamlessly integrating multifaceted data resources associated with human–virus PPIs.
- HVIDB provides a highly accurate and reliable integrative human–virus PPI prediction platform to predict new interactions between viral and human proteins.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Funding

This work was supported by the National Key Research and Development Program of China (2017YFC1200205).

Conflict of interest

The authors declare no conflict of interest.

References

1. Dyer MD, Murali TM, Sobral BW. The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathog* 2008;**4**:e32.
2. Tai W, He L, Zhang X, *et al*. Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. *Cell Mol Immunol* 2020;**17**:613–20.
3. Walls AC, Park Y, Tortorici MA, *et al*. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* 2020;**181**:281–92.
4. Wang C, Wang S, Li D, *et al*. Human intestinal defensin 5 inhibits SARS-CoV-2 invasion by cloaking ACE2. *Gastroenterology* 2020;**159**:1145–7.
5. Wrapp D, Wang N, Corbett KS, *et al*. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 2020;**367**:1260–3.
6. Gordon DE, Jang GM, Bouhaddou M, *et al*. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 2020;**583**:459–68.
7. Monteil V, Kwon H, Prado P, *et al*. Inhibition of SARS-CoV-2 infections in engineered human tissues using clinical-grade soluble human ACE2. *Cell* 2020;**181**:905–13.
8. Wang Q, Zhang Y, Wu L, *et al*. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell* 2020;**181**:894–904.
9. Watanabe T, Kawakami E, Shoemaker JE, *et al*. Influenza virus–host interactome screen as a platform for antiviral drug development. *Cell Host Microbe* 2014;**16**:795–805.
10. Do Kwon Y, Pancera M, Acharya P, *et al*. Crystal structure, conformational fixation and entry-related interactions of mature ligand-free HIV-1 Env. *Matur Struct Mol Biol* 2015;**22**:522–31.
11. Lee S, Salwinski L, Zhang C, *et al*. An integrated approach to elucidate the intra-viral and viral-cellular protein interaction networks of a gamma-herpesvirus. *PLoS Pathog* 2011;**7**:e1002297.
12. Guirimand T, Delmotte S, Navratil V. VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. *Nucleic Acids Res* 2015;**43**:D583–7.
13. Calderone A, Licata L, Cesareni G. VirusMentha: a new resource for virus-host protein interactions. *Nucleic Acids Res* 2015;**43**:D588–92.
14. Durmuş Tekir S, Çakir T, Ardiç E, *et al*. PHISTO: pathogen-host interaction search tool. *Bioinformatics* 2013;**29**:1357–8.
15. Szklarczyk D, Morris JH, Cook H, *et al*. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* 2017;**45**:D362–8.

16. Cook HV, Doncheva NT, Szklarczyk D, *et al*. Viruses.STRING: a virus-host protein-protein interaction database. *Viruses* 2018;**10**:519.
17. Ako-Adjei D, Fu W, Wallin C, *et al*. HIV-1, human interaction database: current status and new features. *Nucleic Acids Res* 2015;**43**:D566–70.
18. Karyala P, Metri R, Bathula C, *et al*. DenHunt—a comprehensive database of the intricate network of dengue-human interactions. *PLoS Negl Trop Dis* 2016;**10**:e0004965.
19. Kwofie SK, Schaefer U, Sundararajan VS, *et al*. HCVpro: hepatitis C virus protein interaction database. *Infect Genet Evol* 2011;**11**:1971–7.
20. Yu H, Luscombe NM, Lu HX, *et al*. Annotation transfer between genomes: protein-protein interrologs and protein-DNA regulogs. *Genome Res* 2004;**14**:1107–18.
21. He F, Zhang Y, Chen H, *et al*. The prediction of protein-protein interaction networks in rice blast fungus. *BMC Genomics* 2008;**9**:519.
22. Garcia-Garcia J, Schleker S, Klein-Seetharaman J, *et al*. BIPS: BIANA interolog prediction server. A tool for protein-protein interaction inference. *Nucleic Acids Res* 2012;**40**:147–51.
23. Ma S, Song Q, Tao H, *et al*. Prediction of protein-protein interactions between fungus (*Magnaporthe grisea*) and rice (*Oryza sativa L.*). *Brief Bioinform* 2019;**20**:448–56.
24. Dyer MD, Murali TM, Sobral BW. Computational prediction of host-pathogen protein-protein interactions. *Bioinformatics* 2007;**23**:i159–66.
25. Zhang A, He L, Wang Y. Prediction of GCRV virus-host protein interactome based on structural motif-domain interactions. *BMC Bioinformatics* 2017;**18**:145.
26. Ghadie MA, Lambourne L, Vidal M, *et al*. Domain-based prediction of the human isoform interactome provides insights into the functional impact of alternative splicing. *PLoS Comput Biol* 2017;**13**:e1005717.
27. Zhang QC, Petrey D, Deng L, *et al*. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* 2012;**490**:556–60.
28. Lasso G, Mayer SV, Winkelmann ER, *et al*. A structure-informed atlas of human-virus interactions. *Cell* 2019;**178**:1526–41.
29. Eid FE, Elhefnawi M, Heath LS. DeNovo: virus-host sequence-based protein-protein interaction prediction. *Bioinformatics* 2016;**32**:1144–50.
30. Emamjomeh A, Goliaei B, Zahiri J, *et al*. Predicting protein-protein interactions between human and hepatitis C virus via an ensemble learning method. *Mol Biosyst* 2014;**10**:3147–54.
31. Cui G, Fang C, Han K. Prediction of protein-protein interactions between viruses and human by an SVM model. *BMC Bioinformatics* 2012;**13**:S5.
32. Qi Y, Tastan O, Carbonell JG, *et al*. Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. *Bioinformatics* 2010;**26**:i645–52.
33. Yang S, Li H, He H, *et al*. Critical assessment and performance improvement of plant-pathogen protein-protein interaction prediction methods. *Brief Bioinform* 2019;**20**:274–87.
34. Lian X, Yang S, Li H, *et al*. Machine-learning-based predictor of human-bacteria protein-protein interactions by incorporating comprehensive host-network properties. *J Proteome Res* 2019;**18**:2195–205.
35. Mohamed TP, Carbonell JG, Ganapathiraju MK. Active learning for human protein-protein interaction prediction. *BMC Bioinformatics* 2010;**11**:S57.
36. Yang X, Yang S, Li Q, *et al*. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. *Comput Struct Biotechnol J* 2020;**18**:153–61.
37. Ammari MG, Gresham CR, McCarthy FM, *et al*. HPIDB 2.0: a curated database for host-pathogen interactions. *Database* 2016;**2016**:baw103.
38. Rose PW, Prlić A, Altunkaya A, *et al*. The RCSB protein data bank: integrative view of protein gene and 3D structural information. *Nucleic Acids Res* 2017;**45**:D271–81.
39. Li H, Yang S, Wang C, *et al*. AraPPISite: a database of fine-grained protein–protein interaction site annotations for *Arabidopsis thaliana*. *Plant Mol Biol* 2016;**92**:105–16.
40. Yang X, Yang S, Qi H, *et al*. PlaPPISite : a comprehensive resource for plant protein-protein interaction sites. *BMC Plant Biol* 2020;**20**:61.
41. Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 1993;**234**:779–815.
42. Giurgiu M, Reinhard J, Brauner B, *et al*. CORUM: the comprehensive resource of mammalian protein complexes — 2019. *Nucleic Acids Res* 2019;**47**:D559–63.
43. Drew K, Lee C, Huizar RL, *et al*. Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Mol Syst Biol* 2017;**13**:932.
44. Barrett T, Wilhite SE, Ledoux P, *et al*. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 2013;**41**:D991–5.
45. Leinonen R, Sugawara H, Shumway M. The sequence read archive. *Nucleic Acids Res* 2011;**39**:D19–21.
46. Qi H, Jiang Z, Zhang K, *et al*. PlaD: a transcriptomics database for plant defense responses to pathogens, providing new insights into plant immune system. *Genomics Proteomics Bioinformatics* 2018;**16**:283–93.
47. Ritchie ME, Phipson B, Wu D, *et al*. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47.
48. Kinsella RJ, Kähäri A, Haider S, *et al*. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* 2011, 2011;bar030.
49. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–20.
50. Frankish A, Diekhans M, Ferreira AM, *et al*. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* 2019;**47**:D766–73.
51. Kim D, Langmead B, Salzberg1 SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015;**12**:357–60.
52. Pertea M, Pertea GM, Antonescu CM, *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015;**33**:290–5.
53. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.
54. Uhlén M, Fagerberg L, Hallström BM, *et al*. Tissue-based map of the human proteome. *Science* 2015;**347**:1260419.
55. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* 2019;**47**:D330–8.
56. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000;**28**:27–30.
57. Orchard S, Ammari M, Aranda B, *et al*. The MIntAct project—IntAct as a common curation platform for 11

molecular interaction databases. *Nucleic Acids Res* 2014;**42**: D358–63.
58. Chatr-aryamontri A, Oughtred R, Boucher L, *et al*. The BioGRID interaction database: 2017 update. *Nucleic Acids Res* 2017;**45**:D369–79.
59. Licata L, Briganti L, Peluso D, *et al*. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 2012;**40**: D857–61.
60. Salwinski L, Miller CS, Smith AJ, *et al*. The database of Interacting Proteins: 2004 update. *Nucleic Acids Res* 2004;**32**:D449–51.
61. Schaefer MH, Fontaine J-F, Vinayagam A, *et al*. HIPPIE: integrating protein interaction networks with experiment based quality scores. *PLoS One* 2012;**7**:e31826.
62. Lian X, Yang X, Shao J, *et al*. Prediction and analysis of human-herpes simplex virus type 1 protein-protein interactions by integrating multiple methods. 2020. *Quant Biol* 2020;**8**:312–24.
63. Potter SC, Luciani A, Eddy SR, *et al*. HMMER web server: 2018 update. *Nucleic Acids Res* 2018;**46**:W200–4.
64. Finn RD, Bateman A, Clements J, *et al*. Pfam: the protein families database. *Nucleic Acids Res* 2014;**42**:D222–30.
65. Liu X, Huang Y, Liang J, *et al*. Computational prediction of protein interactions related to the invasion of erythrocytes by malarial parasites. *BMC Bioinformatics* 2014;**15**: 393.
66. Franz M, Lopes CT, Huck G, *et al*. Cytoscape.Js: a graph theory library for visualisation and analysis. *Bioinformatics* 2016;**32**:309–11.
67. Rose AS, Hildebrand PW. NGL viewer: a web application for molecular visualization. *Nucleic Acids Res* 2015;**43**:W576–9.
68. Li C, Xu D, Ye Q, *et al*. Zika virus disrupts neural progenitor development and leads to microcephaly in mice. *Cell Stem Cell* 2016;**19**:120–6.
69. Shah PS, Link N, Jang GM, *et al*. Comparative flavivirus-host protein interaction mapping reveals mechanisms of dengue and Zika virus pathogenesis. *Cell* 2018;**175**:1931–45.
70. Lima NS, Rolland M, Modjarrad K, *et al*. T cell immunity and Zika virus vaccine development. *Trends Immunol* 2017;**38**:594–605.