Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Database and Analytical Resources for Viral Research Community

Sujal Phadke and Saichetana Macherla, J. Craig Venter Institute, La Jolla, CA, United States
Richard H Scheuermann, J. Craig Venter Institute, La Jolla, CA, United States; University of California, San Diego, CA, United States; La Jolla Institute for Immunology, La Jolla, CA, United States; and Global Virus Network, Baltimore, MD, United States

© 2021 Elsevier Ltd. All rights reserved.

This is an update of K. McLeod, C. Upon, Virus Databases, In Reference Module in Biomedical Sciences, Elsevier Inc., 2017, doi:10.1016/B978-0-12-801238-3.95728-3.

Significance of Viral Databases

Viral disease outbreaks are an ongoing threat to public health. Every few years, viral pathogens, including various influenza strains, SARS and MERS coronaviruses, Ebola, and most recently Zika virus, have caused considerable personal and economic loss (see "Relevant Websites section"). Identification of causative agents, clinical reporting of infections, and epidemiological surveillance are all critical during these outbreaks. Such efforts to identify the causes of these infectious outbreaks lead to a wealth of information about the viruses and their host. Databases allow storage and analysis of this information to fuel further wet-lab experimentation on the causative agents and comparative analyses of their genomes. Such research efforts are important for predicting, preventing and limiting future outbreaks.

Comprehensive databases such as Viral Pathogen Resource (ViPR, see "Relevant Websites section") and Influenza Research Database (IRD, see "Relevant Websites section") have been instrumental in providing a one-stop-shop for data and analytical tools for basic and applied research in virology. The significance of such databases is evident in multiple use cases that demonstrate the utility of the resources. For example, data and bioinformatics tools from ViPR and IRD have facilitated research into detection and diagnostics of viral pathogens, prediction of viral hosts and environmental reservoirs, viral evolution, development of vaccines, discovery of genomic determinants of virulence, and anti-viral drug development. Importantly, these resources allow investigators at all levels of training and expertise to easily perform their desired analyses and to contribute critical information about infectious disease outbreaks.

Overview of Viral Databases and Analytical Tools

Viruses infect all kingdoms of life. In this article, we focus on database resources for viruses that infect humans and other animals. We call attention to databases such as Plant Viruses Online (see "Relevant Websites section") and the Prokaryotic Virus Ortholog Groups (pVOGs) (see "Relevant Websites section") for readers interested in viruses that infect other host organisms, which are out of scope of this article. The landscape of databases and analytical tools available for human virology research is guided by research and development goals for priority pathogens. The available resources can be categorized as databases that store specific data types and bioinformatics webtools that offer specific analytical capabilities. These two essential functions have also been combined and integrated in comprehensive resources such as ViPR and IRD.

Types of Databases

Several types of databases are available for virology research that can be distinguished based on the type of data they contain or the pathogen area of focus (**Table 1**). For instance, many popular databases focus on storing information about specific biomolecules, such as gene and protein sequences, immune epitopes, or protein structures. These databases can be further distinguished as sequence archives such as GenBank (see "Relevant Websites section"), and UniProt (see "Relevant Websites section"), where data is deposited by the primary investigators and curated DBs such as RefSeq (see "Relevant Websites section") that integrate additional knowledge (e.g., annotations) with sequence records to provide an enhanced knowledgebase. Biomolecule information other than sequences is also stored in other databases, including the Protein Data Bank (PDB; see "Relevant Websites section"), which stores 3D structural data, the Immune Epitope Database (IEDB; see "Relevant Websites section"), which catalogs experimental data on B cell and T cell epitopes studied in humans and other animals and the Virus Particle Explorer (VIPERdb; see "Relevant Websites section"), which stores the structures of viruses with icosahedral virions.

Virology databases have also been designed to focus on particular taxa of viral pathogens. For example, recognizing hepatitis B virus as a major public health problem worldwide, the Hepatitis B Virus Database (HBVDb; see "Relevant Websites section") has been designed to facilitate research on the genetic variability of HBV and its resistance to treatment. HBVDb allows the analysis of annotated sequences for genotyping and drug resistance profiling. Similarly, a collection of databases for research on the Human Immunodeficiency Virus (HIV) are available (see "Relevant Websites section") that contain comprehensive data on genome and protein sequences and immunological epitopes. Because influenza virus poses perhaps the most persistent major global public health threat, several databases are dedicated to research on influenza. For instance, the Global Initiative on Sharing All Influenza Data (GISAID; see "Relevant Websites section") is an access-controlled resource of influenza sequence information and related epidemiological data. FluNet (See Relevant Websites section) is a global web-based influenza surveillance data collection, maintained at the World Health Organization (WHO) and available for tracking the movement of flu viruses globally. The Influenza Virus Resource (see "Relevant Websites section") supports the search and analysis of

Table 1 List of databases and webtools

Category	*Name*	*Types of data/services*	*Weblink*
Databases	GenBank	Gene and genome sequences	https://www.ncbi.nlm.nih.gov/genbank/
	UniProt	Protein sequences	https://www.uniprot.org
	RefSeq	Curated genome and protein sequences	https://www.ncbi.nlm.nih.gov/refseq/
	Protein Data Bank (PDB)	3D protein structures	www.pdb.org
	Immune Epitope Database (IEDB)	Experimental data on B cell and T cell epitopes	www.iedb.org
	Virus Particle Explorer (VIPERdb)	Structures of viruses with icosahedral virions	http://viperdb.scripps.edu/
	Viral Pathogen Resource (ViPR)	Comprehensive collection of multiple data types on high priority human pathogenic and related viruses and an integrated suite of analytical and visualization capabilities	https://www.viprbrc.org/
	Influenza Research Database (IRD)	Comprehensive collection of influenza virus-related data and an integrated suite of analytical and visualization capabilities	https://www.fludb.org
	Global Initiative on Sharing All Influenza Data (GISAID)	Access-controlled resource of influenza sequence information and related epidemiological data	https://www.gisaid.org
	FluNet	Global web-based influenza surveillance data collection	https://www.who.int/influenza/gisrs_laboratory/flunet/en/
	Influenza Virus Resource	Influenza genomic and protein sequences	https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database
	Hepatitis B Virus Database (HBVDb)	Nucleotide and protein sequence information for analysis of drug resistance profiling	https://hbvdb.ibcp.fr/HBVdb/
	Human Immunodeficiency Virus (HIV)	Genome and protein sequences and immunological epitopes	https://www.hiv.lanl.gov/content/index
	ViralZone	highly curated and extensive virus knowledgebase	https://viralzone.expasy.org
	Prokaryotic Virus Ortholog Groups (pVOGs)	Bacteriophage protein orthology information	http://dmk-brain.ecn.uiowa.edu/VOG/
	Plant Viruses Online	Plant virus gene and protein sequences and structures	http://sdb.im.ac.cn/vide/refs.htm
Webtools	IDSeq	Real-time pathogen detection from metagenomes	https://idseq.net
	Virome	Viral detection from environmental metagenomes	http://virome.dbi.udel.edu
	VirusDetect	viral detection from small RNA datasets using both *de novo* and reference-based assemblies	http://virusdetect.feilab.net/cgi-bin/virusdetect/index.cgi
	Viral Genome ORF Reader (VIGOR)	Homology-driven viral gene prediction and genome annotation	https://github.com/JCVenterInstitute/VIGOR4
	STRING-Viruses	Assessment of viral-host protein-protein interactions using visualization tools such as Cytoscape	http://viruses.string-db.org
	NextStrain	Rapid, real-time tracking and prediction of spatio-temporal spread of infection during infectious outbreaks	https://nextstrain.org

influenza genomic and protein sequences at National Center for Biotechnology Information (NCBI). The Influenza Research Database (IRD; see "Relevant Websites section") provides the most comprehensive collection of influenza virus-related data and an integrated suite of analytical and visualization capabilities for research on influenza virus.

In contrast to the aforementioned resources that are focused on a particular data type or virus, the ViPR resource (see "Relevant Websites section") is unique in that it provides cross-referenced data of multiple types on all high priority human pathogenic viruses that pose a threat to public health, except HIV. Each virus family has a dedicated portal within ViPR that offers intuitive, customized search interfaces and analytical options tailored for each of the virus families. ViralZone (see "Relevant Websites section") provides access to a highly curated and extensive knowledgebase about a wide range of viruses.

Types of Bioinformatics Webtools

Research in virology is heavily dependent on data mining using sophisticated bioinformatics tools. With the foresight into the importance of such capabilities, several dedicated webtools are available for the users to conduct various types of analyses on viral genomes. For instance, tools such as IDSeq (see "Relevant Websites section"), Virome (see "Relevant Websites section") and VirusDetect (see "Relevant Websites section") allow detection of viruses from deep-sequencing of metagenomics samples. IDSeq is designed with the aim of real-time pathogen detection from metagenomes. Virome focuses on environmental metagenomes, whereas VirusDetect specifically uses small RNA datasets to detect viruses using both de novo and reference-based assemblies.

Once a novel virus isolate or variant is detected, tools such as the Viral Genome ORF Reader (VIGOR; see "Relevant Websites section") enable genome annotation. VIGOR is a homology-driven viral gene prediction program that yields predicted proteins and mature peptides for newly sequenced isolates and variants of human virus. The software uses a set of highly curated databases enabling VIGOR to annotate a given viral genome. Currently VIGOR supports gene prediction and annotation of about 25 different virus taxonomic groups.

Tools are also available to study the viral pathogen in the context of its host environment. For instance, STRING-Viruses (see "Relevant Websites section") is a webtool available as part of the STRING database that allows assessment of protein-protein interactions using visualization tools such as Cytoscape. This webtool is particularly important for studying how viral proteins interact with host proteins during various stages of infection. Likewise, NextStrain (see "Relevant Websites section") is a webtool that enables rapid, real-time tracking of evolving pathogen populations during infectious outbreaks. NextStrain is an open source system that tracks mutation marker data on pathogen phylogenies to make inferences about epidemiologically-relevant parameters such as spatio-temporal spread of the infection within a host population.

Virus Pathogen Database and Analysis Resource (ViPR) and Influenza Research Database (IRD)

For the remainder of this article, we focus on describing two related database and analytical resources available for research on human viral pathogens – ViPR and IRD – as examples for how these types of resources are developed and used. For a more comprehensive list of other available virus database and analysis resources, we encourage the reader to explore additional information about resources listed at ViralZone (see "Relevant Websites section").

The National Institute of Allergy and Infectious Diseases (NIAID) at the U.S. National Institutes of Health (NIH) implemented the Bioinformatics Resource Centers (BRCs) for Infectious Diseases program to support research on priority pathogens of humans. As a result, the BRC focused on viral pathogens has developed the ViPR and IRD resources as publicly-accessible online repositories for viruses that adversely affect public health with the aim of integrating research and surveillance data. ViPR (see "Relevant Websites section") is unique amongst viral-centered databases in offering a wealth of information on a large number of viral families. In contrast, IRD (see "Relevant Websites section") is a parallel resource that is focused exclusively on Influenza virus. The objective of both resources is to provide virus data and analytical capabilities to advance the understanding of virus transmission, pathogenesis, and host range, and to support the development of diagnostics and therapeutic interventions.

Sources of Data

The ViPR and IRD databases integrate data from three sources (**Table 2**):

Data Aggregated From Public Data Archives

The ViPR and IRD databases capture various data types from multiple publicly-accessible data archives. ViPR and IRD integrate genomic sequence information from GenBank (see "Relevant Websites section"), protein sequences from UniProt (see "Relevant Websites section"), protein structures from the Protein Data Bank (PDB; see "Relevant Websites section"), experimentally determined T-cell and B-cell epitopes from the Immune Epitope Database (IEDB; see "Relevant Websites section"), and Gene Ontology annotations from the GO database (GO, see "Relevant Websites section"). All data types are regularly updated and are searchable using their original accession numbers within intuitive web-based user interfaces.

Direct Submission of Novel Data

In some cases, active research projects supported by the U.S. National Institutes of Health and other interested parties submit data and related metadata directly to ViPR and IRD. For instance, NIAID-funded Systems Biology Consortium for Infectious Diseases research programs submit a variety of different transcriptomic, proteomic, and metabolomic datasets that investigate in vivo and in vitro host responses to viral infections. The Genomic Sequencing Centers for Infectious Diseases (GCID) program submit detailed structured metadata, including clinical information such as disease symptoms, severity, and diagnostic test outcomes, that are linked with sequence records of the corresponding virus isolate obtained from GenBank. IRD serves as the repository for the influenza human and animal surveillance data collected by the Centers of Excellence for Influenza Research and Surveillance (CEIRS) program.

Derived and Predictive Data

The IRD and ViPR development team generates and integrates unique derived data from bioinformatics analysis pipelines performed in-house, tailored specifically for a given taxonomic groups. Derived data include improved and consistent metadata annotations including strain name, clade and genotype information, virus taxonomy, host and country of isolation, and collection date. For instance, the ViPR annotation process extends information available in the representative RefSeq strain for each species. The process uses multiple sequence alignment to map homologous regions across related viral genomes to map mature peptide cleavage sites on

Table 2 Sources of IRD and ViPR data

	Data source/algorithm	*Data type*	*Component*
Imported public data	NCBI - GenBank	Genome sequences/annotations	ViPR and IRD
	NCBI - RefSeq	Genome sequences/annotations	ViPR and IRD
	Immune Epitope Database (IEDB)	Curated epitopes	ViPR and IRD
	UniProt	Protein annotations	ViPR and IRD
	RCSB Protein Data Bank (PDB)	Protein 3D structures	ViPR and IRD
	Catalytic site atlas	Active sites	ViPR and IRD
	PATRIC & VBRC Bioinformatics Resource Centers	Orthologs	ViPR
	AVIBase	Bird taxonomy	IRD
Data submitted directly	NIAID Genome Sequencing Centers	Clinical metadata	ViPR and IRD
	NIAID Systems Biology program	Host factor data	ViPR and IRD
	ViPR-funded driving biological projects	Host factor data	ViPR
	NIAID Centers of Excellence for Influenza Research and Surveillance (CEIRS)	Surveillance records,	IRD
	NIAID Centers of Excellence for Influenza Research and Surveillance (CEIRS)	Serology test records	IRD
ViPR/IRD generated data	NCBI BlastP	Sequence similarities	ViPR and IRD
	InterProScan	Domains/motifs	ViPR and IRD
	NetCTL	Predicted CTL epitopes	ViPR and IRD
	ViPR pipeline	Mature peptides	ViPR
	ViPR custom algorithm	Isoelectric point and molecular weight	ViPR
	ViPR curation	Sequence feature variant types	ViPR
	IRD pipeline	SNP/consensus sequence	IRD
	IRD custom algorithm	Isoelectric point and molecular weight	IRD
	IRD curation	Sequence feature variant types	IRD
	IRD curation	PCR primers & probes	IRD
	IRD algorithm	PA-X protein annotation	IRD
	IRD tool	H5N1 clade classification	IRD
	IRD curation	Flu season assignment	IRD
	IRD tool	2009 pH1N1 sequence classification	IRD

polyproteins. Likewise, a custom annotation pipeline is used in IRD to predict open reading frames and sequences for variants of influenza proteins including PA-X, PA-N155, PA-N182, M42, NS3 and PB1-40. The predicted variant proteins can be retrieved from the Nucleotide and Protein Sequence Search pages. Various tree-based clade classification tools are also available and have been used to predict clades and genotypes of pathogenic strains of several viruses including Zika, rotaA, and Hepatitis C virus in ViPR and H1N1, H5N1 and swine H1 strains in IRD. Furthermore, Sequence Features (SFs) are derived using information integrated from UniProt, GenBank, IEDB and the scientific literature followed by inspection and validation by domain experts. SFs are protein regions with important structural, functional, immune epitopes, or sequence alteration characteristics. Once the SF protein regions are defined, the extent of sequence variation observed in each region is determined as a series of Variant Types (VTs). Lastly, the Host factor component of IRD/ViPR contains a variety of derived data that gives insights about the systems-level infection dynamics. For instance, host factor biosets are group of genes/proteins/metabolites that are significantly differentially expressed/abundant at different times post infection. Data models derived using Weighted Gene Coexpression Network Analysis (WGCNA) are available to aid identification of co-expressed genes that may be functionally related, tightly co-regulated or members of similar pathway. The set of co-expressed genes can also be visualized as Cytoscape networks where nodes represent genes and edges represents the strength of co-expression.

Data Summary

Table 3 ViPR and IRD offer frequent updates on all data types. Genome sequence data are updated daily (IRD) or weekly (ViPR) while all other data types are updated with each bimonthly release. As of September 23, 2019, ViPR provides data on 667,249 virus strains from nearly 6126 viral species belonging to 20 families including *Arenaviridae, Caliciviridae, Coronaviridae, Fimoviridae, Filoviridae, Flaviviridae, Hantaviridae, Hepeviridae, Herpesviridae, Nairoviridae, Paramyxoviridae, Peribunyaviridae, Phasmaviridae, Phenuviridae, Picornaviridae, Poxviridae, Reoviridae, Rhabdoviridae and Togaviridae.* It contains sequences from nearly 883,170 genomes, out of which upwards of 110,742 are complete genome sequences. Sequence data on >2,100,000 proteins are also available and contains various attributes including annotations, mature peptide data, experimentally determined epitopes, etc. ViPR contains a total of 16,945 3D protein structures from PDB and 61,816 experimentally-determine immune epitopes. Table 3 displays a breakdown of available data; details may be found at the link (see "Relevant Websites section").

Table 3 Data summary. Numbers of various data types available in ViPR and IRD as of September 23, 2019 are shown. All data types are regularly updated

Data category	*Attribute*	*ViPR*	*IRD*
Genome information	Species	6,126	
	Genomes/segments	883,170	751,002
	Complete genome segments	110,742	419,482
	Proteins	2,143,646	1,184,929
	Mature peptides	243,538	
	Strains	667,249	161,216
	Strains with predicted genotypes	140,817	
	Strains with predicted segments	28,570	
	Genomes with clinical metadata (NIAID GSCID, manual curation)	3,931	
	Orthology group	9,385	
	Functional annotation	1,242,728	
	PubMed references	350,036	180,692
	Sequence feature	1,659	5,629
Number of proteins with specified annotations	Proteins	2,143,646	1,184,929
	Epitopes from IEDB	61,816	3,885
	PDB Files	16,945	1,748
	Pfam domains	1,699,884	1,104,118
	Other domains/motifs	1,295,044	760,672
	GO IDs biological process	155,156	599,355
	GO IDs molecular function	177,460	575,346
	GO IDs localization	262,063	663,293
	GO IDs		792,355
	EC numbers	33,636	
	Proteins with predicted epitopes*	1,769,976	1,089,613
	BlastP Alignments*	1,914,679	
Number of strains with specified annotation	Strains with predicted pH1N1 classification		69,668
	Strains with predicted H5 clade classification		9,042
Number of segments with specified annotation	Total segment sequences		751,002
	Polymorphism data*		437,685
	PubMed references		180,692
Number of Surveillance Samples	Hosts sampled		1,055,511
	Samples		1,233,703
	Samples tested for flu		1,211,204
	Flu-positive samples		80,040
	Samples with sequence data		9,686
	Samples with serology data		339
	Samples with structured metadata		1,233,703

The IRD holds 751,002 total influenza genome segment sequences, 1748 PDB structures and 1,184,929 proteins with predicted epitopes. Also, the IRD is unique in providing host factor datasets generated from experimental infections of host organism and cell lines with various viral strains. These cover a range of pathogens in the *Orthomyxoviridae* and *Coronaviridae* families. Currently, 66 datasets from four types of "omics" experiments (transcriptomics, proteomics, lipidomics, and metabolomics) are provided. Out of 66 datasets 34 have been analyzed for the WGCNA data models and 25 have the Cytoscape network visualization implemented. IRD is also unique in offering human and animal surveillance data collected by Centers of Excellence for Influenza Research and Surveillance (CEIRS). Additionally, in collaboration with the Global Animal Disease Information System the of Food and Agricultural Organization of United Nations, IRD has established links between sequence records in IRD and disease outbreak event records in EMPRES-I (see "Relevant Websites section").

KEY DATA FEATURES of ViPR and IRD
- Comprehensive and up-to-date
- Consistent annotations
- Well-curated Influenza variant proteins
- Mature peptides for *Flaviviridae*, *Coronaviridae,* and *Picornaviridae*
- Well-curated metadata about geographic locations, host species, date of isolation, etc.
- Unique data types including host factor, sequence features and animal and human surveillance information

Data Curation

ViPR and IRD offer highly-curated data that has been vetted using computational and manual curation strategies. For instance, an in-house curation and annotation pipeline provides curated sequences from which sequence anomalies have been detected for potential removal during downstream analysis. Along with the sequence data, the ViPR team has manually-curated the scientific literature to provide improved and consistent annotations of metadata including the geographic location, year, and host for many clinically-relevant taxonomic groups. The highly curated strain level data are displayed with a Genome Map image and a Protein Information table from which detailed structural and functional information for a given gene/protein can be obtained. ViPR utilizes RefSeq strains to extend the manually-curated annotations to strains belonging to the same taxon. Furthermore, RefSeq sequences are used to construct virus ortholog groups and their associated annotations, which enable identification of proteins with similar function within a given virus taxon. ViPR and IRD also offer curated data on T-cell and B-cell immune epitopes and their predicted positioning on protein structures from the IEDB. Data curation in ViPR and IRD continues to grow and expand beyond sequence and strain level information. For example, both databases offer curated antiviral drug data from the DrugBank (see "Relevant Websites section"), including the descriptive drug information, 3D structures for target complexes, interaction sites as sequence features and antiviral resistance mutations to aid in assessing the risk of anti-viral drug resistance development.

Data Retrieval

Search Interface

ViPR offers customized search interfaces to allow for the retrieval of selected genomic, structural and other data records using specific metadata for different virus families. Users initiate the search by selecting a virus family on the home page. A user can narrow the search data specific to a virus strain by querying the database using genus or species, geographical location and date of isolation, virus host, and clinical or experimental data. The user also has an option to cast a wider net using keywords with further narrowing using advanced search options. Once a strain or set of strains is selected, detailed genomic and protein sequence information and associated annotation can be accessed. These data can then be directly analyzed using any of the appropriate tools available from within ViPR.

Because IRD is dedicated to influenza viruses, the search interface design is guided by the availability of influenza strain-specific data. Users can query the database using the branching logic inherent in the database. For instance, users can search for complete or partial genome and segment sequences, and proteins by directly entering the name of the strain(s) of interest. Users can also choose amongst the several metadata fields such as host, geographic location and the date of pathogen isolation. Once a particular taxon, strain or metadata category is selected as a search criterion, additional search criteria appear dynamically to allow the users to perform more focused searches. Moreover, users can also use the advanced search options to refine the search results with the more fine-grained search criteria. For instance, users can choose to view data on strains isolated in specific months of a given year(s) or limit search to specific host attributes such as gender and age and choose to limit their search to specific specimen type, laboratory strains and organism detection method. Lastly, users can customize their viewing options to specific display fields through advanced search menu. An example of the various search options is shown in **Fig. 1**.

Application Programming Interfaces (API)

ViPR and IRD provide users an option to retrieve certain data types using command line utilities via Application Programming Interfaces (APIs). Specifically, the sequence search API allows users to retrieve sequences and associated metadata using GenBank and protein accession IDs. The retrieved sequences can be obtained in either FASTA or JASON formats with user-defined associated metadata. The surveillance API allows retrieval of surveillance records and metadata from host surveillance samples. IRD allows users to submit sequences for large phylogenetic analysis jobs through an API to the high-performance computing environment provided by the NSF-sponsored Cyber-Infrastructure for Phylogenetic Research (CIPRES) Gateway. Tree calculations are made using the high-performance computing environment and the resulting phylogenetic tree is returned for visualization using the Archaeopteryx tool in IRD.

Analysis and Visualization Capabilities

ViPR and IRD host a comprehensive suite of bioinformatics tools for data analysis and visualization, closely integrated with the supported data. These include popular webtools in bioinformatics constructed by the ViPR team or contributed by users/collaborators. Examples of the types of analysis that can be performed and the webtools that are available are described below. For a complete list of analytical tools, the reader is directed to the ViPR (see "Relevant Websites section") and IRD (see "Relevant Websites section") homepages.

Sequence Annotation

The sequence annotation pipelines allow users to upload and annotate genomic sequences to predict segment type, CDS location, and genotype information, and to identify possible sequencing artifacts.

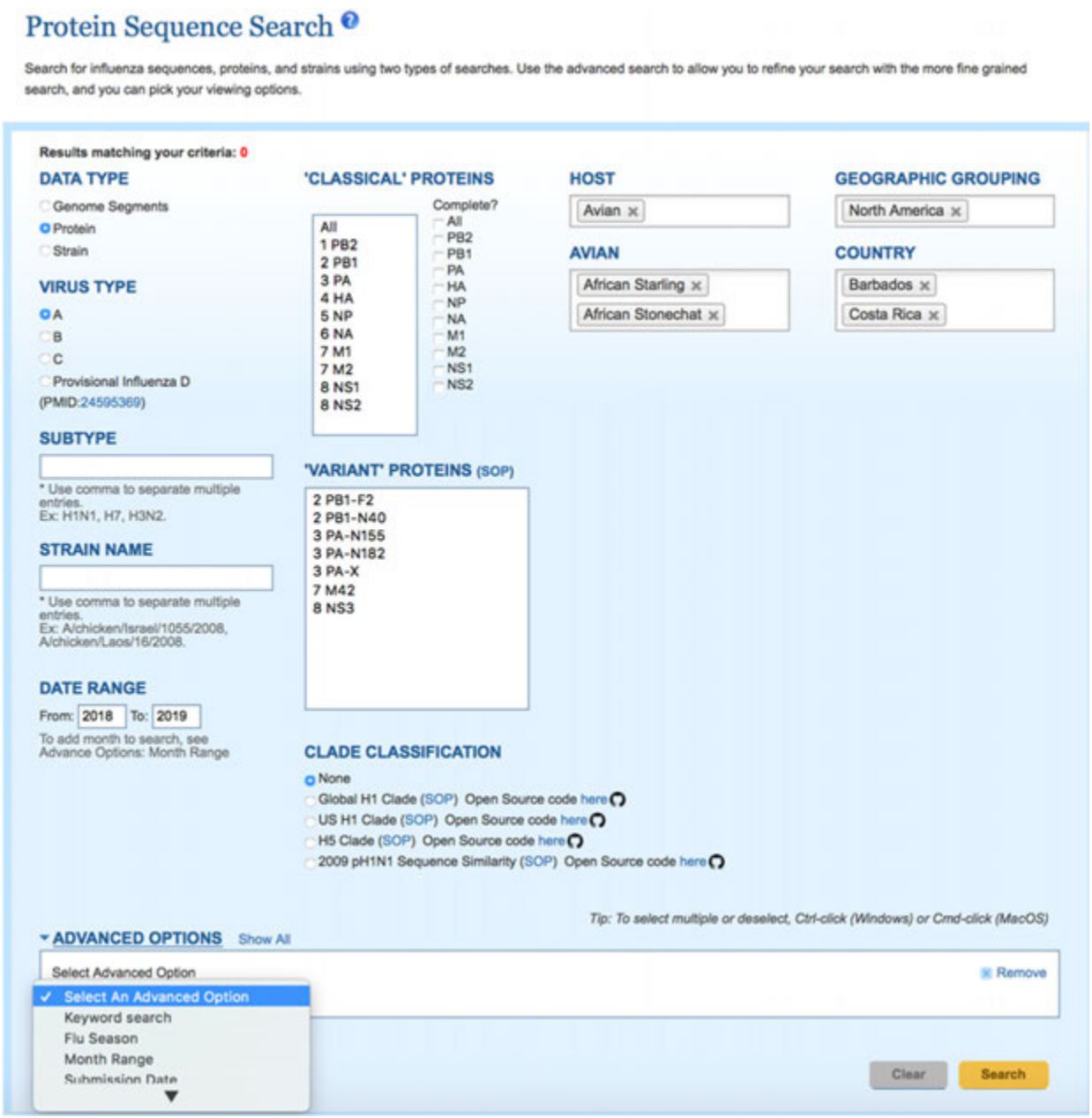

Fig. 1 Protein search interface in IRD. The search page supports queries based on "classical" as well as "variant" proteins and associated metadata. A search query can be made more specific by choosing various query features. For example, users can search for specific strain(s) by entering the strain name and subtypes in the appropriate search fields. Additionally, choosing a type of host, such as avian, brings a drop down menu from which the user can choose one or more species to make the search criterion more specific. Users may also choose to limit their search results by geographic region(s) by choosing one or more countries in the dropdown menu. Search results may be limited to a specific date range by putting in the years or by choosing a month range through advanced options. Multiple other search criteria such as keyword search, submission date, host gender and age etc. are available through advanced options to make the search results more specific.

Sequence Search and Alignments

Users can use popular tools such as BLAST and MUSCLE within ViPR and IRD. Sequences can be selected from a search result or a working set in their personal workbenches. Users can also perform manual exploration and curation of sequence alignments including relabeling the sequences and adding sequence features. After an alignment is completed, users have an option to download the input sequences and output files in a variety of formats or pass the alignment to another tool including SNP analysis or meta-CATS.

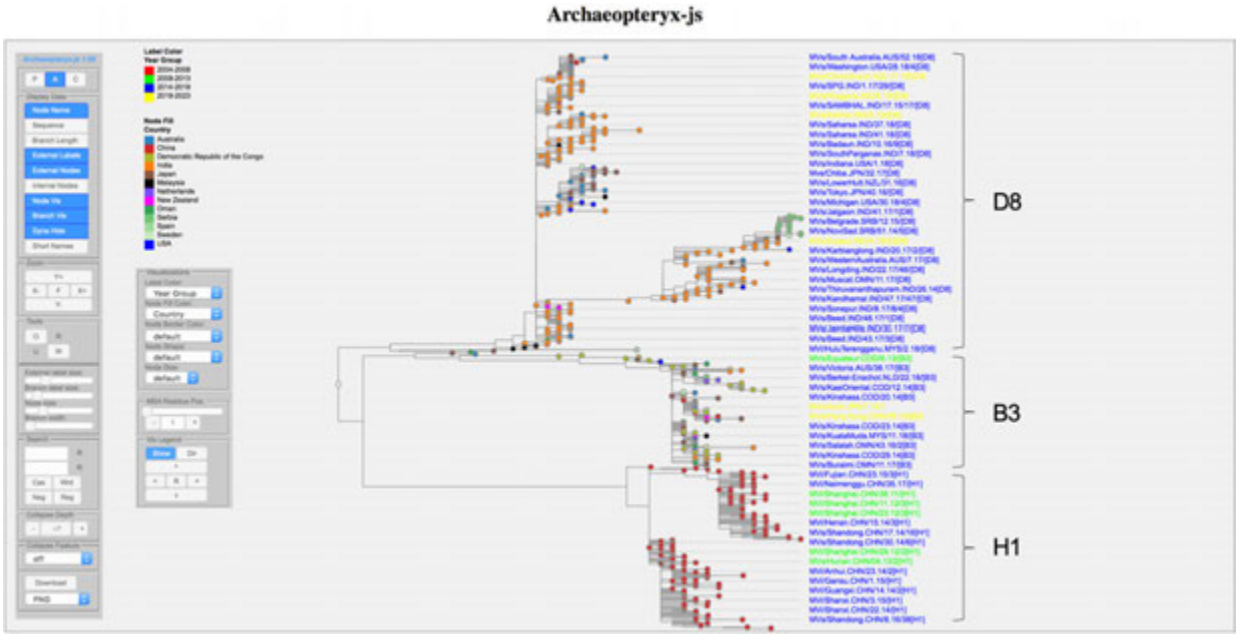

Fig. 2 Example of a phylogenetic tree constructed in ViPR. The search interface was used to retrieve unique sequences belonging to the 450 bp region that codes for the C-terminal 150 amino acids of the N nucleocapsid protein of the human measles *Morbillivirus*. A total of 554 unique sequences were obtained. Multiple sequence alignments were performed with MUSCLE and phylogenetic relationships inferred using RAxML for visualization with Archaeopteryx.js. The legend shows options for users to customize the tree visuals and highlight desired metadata in Archaeopteryx. For instance, the phylogram display with aligned labels has been chosen from the top left panel for improved readability. Likewise, the "Dyna Hide" option on the left panel has been selected to only show representative sequence names. Names of the nodes have been color coded to indicate the year of isolation. The node color indicates the country of origin. The available sequences separate into 3 clades belonging to subtypes D8, B3 and H1. The subtypes are represented in the parentheses following the names of the sequences. While the subtypes D8 and B3 are more globally circulating, causing infections across different countries, the subtype H1 is predominantly found in China. Moreover, in a given country, strains belonging to the same subtype have repeatedly caused measles infections over several years indicating pathogen persistence in the population, likely due to imperfect and inadequate vaccination practices.

Phylogenetic Tree Reconstruction

Users can infer phylogenetic relationships using RAxML and FastME algorithms, and visualize the results using a customized visualization tool developed in house called Archaeopteryx, which allows users to color-code and annotate various tree branches and nodes using available metadata, such as geographical location, host, isolation year, and amino acid residues at user-selected positions (e.g., Fig. 2).

Metadata-Driven Comparative Genomics

The meta-CATS tool provides a statistical analysis of sequences to identify genome and protein positions that show significantly-different residue distributions between groups of sequences using the Chi-squared statistic. Sequences can be segregated into groups manually or automatically based on selected metadata values, such as year of isolation, geographical location, host species, etc. Thus, a user can put the analysis of sequences in the context of infection and infer association of variations in a genomic region with a particular infection characteristic.

Analysis and Visualization of 3D Protein Structures

Users can search for protein structures using multiple types of queries, including PDB IDs, gene symbol, Entrez ID, UniProt accession, and gene product names. Furthermore, search can be restricted to include only proteins with experimentally determined epitopes, experimentally determined active sites and proteins with sequence features. Additionally, users can use advanced options to query the database using theoretical structures. Once a particular structure from the search results is selected, users can customize the general appearance of protein structures. For example, users can highlight ligands, active sites, epitopes and sequence features on the 3D structures. Individual residues within the protein structures are mapped to homologous positions from UniProt records, which allows comparison between protein structures. Annotating a 3D structure with important residues and regions of interest can yield testable hypotheses about the functional relevance of the protein. Lastly, users can download the highlighted protein structure as a publication quality image file or a structure movie.

Genome Annotation Using VIGOR

Users can use the VIGOR software tool along with its collection of highly-curated reference databases for different viruses to predict viral protein open reading frames and sequences, and to identify typical viral transcriptional and translational exceptions including RNA editing, stop codon read-throughs and ribosomal slippage.

Virus Genotype/Clade Classification

ViPR and IRD offer two types of user/community contributed tools for virus classification. A clade classification tool infers clades for a query sequence from its position within a reference phylogenetic tree. Currently, clade classification is available for Zika virus, and Hepatitis C Virus (HCV) in ViPR and swine H1 and H5N1 influenza viruses in IRD. The H5N1 classification tool uses phylogenetic analysis to classify HA sequences according to the WHO H5 classification scheme. The H5N1 classifier has been verified to have $>98\%$ accuracy for sequences of at least 300 nucleotides of HA1. On the other hand, the H1N1 classification tool in IRD is a robust application of BLAST to recognize sequences closely related to pandemic sequences. BLAST-based classification is also available for classifying rotaA virus sequences in ViPR.

HA Subtype Numbering Conversion

IRD has implemented an HA subtype numbering conversion tool that allows users to convert HA sequence coordinates among any selected subtypes based on protein structure alignment rather than sequence-based alignment. Using this tool, the user can convert the coordinates of an HA protein sequence to the corresponding coordinates in other subtypes, to compare substitutions associated with phenotypic changes and to identify cross-reactive immune epitopes. The tool can also be integrated with sequence variation analysis and meta-CATS.

KEY ANALYSIS FEATURES
- Seamless integration of data and analysis/visualization tools
- Analysis of user data in combination with database data

TOP ANALYSIS TOOLS BASED ON USAGE
- Multiple sequence alignment
- BLAST
- HA numbering
- SNP analysis
- Phylogenetic tree reconstruction
- H5N1 classifier
- H1 classifier

Workbench

Users can establish personal workspaces under the "workbench" feature within the IRD and ViPR. This tool provides an interface that allows users to save previous search or analysis results, which enables users to re-use their work without re-running the analysis. It also allows users to combine multiple analyses. Users can upload and save their own private data and metadata to their

Fig. 3 Links to multifaceted user support available in ViPR.

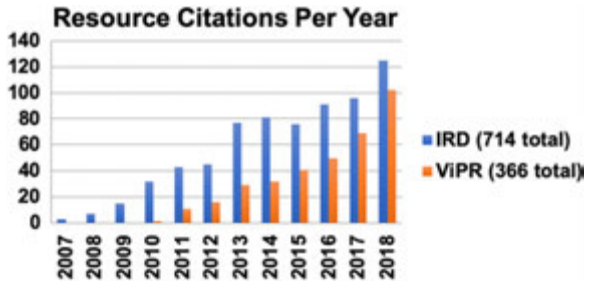

Fig. 4 The number of peer-reviewed citations for the ViPR and IRD resources.

workspace to be analyzed using the analytical and visualization tools provided by IRD and ViPR. The saved data and analysis results can be shared with collaborators through their workbench accounts.

User Support

The IRD and ViPR databases are open access resources and can be used and shared without restrictions. The databases offer multifaceted user support (**Fig. 3**). Users can report a problem or ask a question using the forms provided online. The development and management teams of both IRD and ViPR are responsive to questions from the helpdesk and to suggestions for enhancements. Users can join a newsletter mailing list to get information about updates of the resources.

Both IRD and ViPR provide extensive tutorials, training modules and manuals. For additional support, the development and management teams engage in outreach sessions that include webinars, tutorials, and training workshops at various geographical locations. For an expert user, the analytical tools developed by the ViPR/IRD team are also available on GitHub, which avails the user with an option of using the tools outside of the IRD and ViPR resources on their preferred platform.

Usage Statistics

The ViPR and IRD databases continue to provide critical resources in several research studies as evident by the increasing number of citations in the scientific literature (**Fig. 4**). Together, the two databases have been cited in 1080 publications as of May 10, 2019. The number of new sessions initiated per week in 2018 (Google Analytics) tallies at 1488 at ViPR and 1482 at IRD. Importantly, these sessions have been documented from 181 countries for ViPR and 174 countries for IRD.

Summary and Conclusions

Virology research is dependent on timely availability of reliable data on viral pathogens, their hosts and the infection/outbreak dynamics. ViPR and IRD offer comprehensive, highly curated data on human viral pathogens along with an intuitive search interface and seamless integration of the data with analytical and visualization tools. The resources are available freely without restrictions. The availability of such resources streamlines and expedites experimental discovery advancing the ultimate goal of developing improved diagnostics and therapeutics for priority pathogenic viruses.

Further Reading

Adhikari, U.K., *et al.*, 2018. Immunoinformatics approach for epitope-based peptide vaccine design and active site prediction against polyprotein of emerging Oropouche Virus. Journal of Immunology Research. doi:10.1155/2018/6718083.
Afelt, A., *et al.*, 2018. Bats, Bat-Borne Viruses, and Environmental Changes. IntechOpen.
Andreani, J., *et al.*, 2019. Atypical cowpox virus infection in smallpox-vaccinated patient, France. Emerging Infectious Diseases 25 (2), 212–219.
Babayan, S.A., *et al.*, 2018. Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. Science 362 (6414), 577–580. doi:10.1126/science.aap9072.
Brown, D.M., *et al.*, 2018. Contemporary circulating enterovirus D68 strains show differential viral entry and replication in human neuronal cells. mBio 9 (5), doi:10.1128/mBio.01954-18.
Carter, K., et al. Anti-Chikv Antibodies and Uses Thereof, US Patent. US20180127487A1.
Claes, *et al.*, 2014. The EMPRES-i genetic module: A novel tool linking epidemiological outbreak information and genetic characteristics of influenza viruses. Database. bau008.
Dutta, S.K., *et al.*, 2018. Chikungunya virus: Genomic microevolution in Eastern India and its in-silico epitope prediction. 3 Biotech 8, 318.
Greene, J.M., *et al.*, 2007. National institute of allergy and infectious diseases bioinformatics resource centers: New assets for pathogen informatics. Infection and Immunity 75, 3212–3219.
Langfelder, P., Hovrath, S., 2008. WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics 9, 559–572.
Lee, A.J., *et al.*, 2017. Identification of diagnostic peptide regions that distinguish zika virus from related mosquito-borne flaviviruses. PLoS One 12 (5), e0178199. doi:10.1371/journal.pone.0178199.
Lee, *et al.*, 2015. Diversifying selection analysis predicts antigenic evolution of 2009 pandemic H1N1 influenza A virus in humans. Journal of Virology 89, 5427–5440. doi:10.1128/JVI.03636-14.

Mette, *et al*., 2005. An integrative approach to CTL epitope prediction: A combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. European Journal of Immunology 35 (8), 2295–2303. doi:10.1002/eji.200425811.
Miller, M.A. *et al*., 2010. Creating the CIPRES science gateway for inference of large phylogenetic trees. In: Proceedings of the Gateway Computing Environments Workshop (GCE), pp. 1–8. New Orleans, LA.
Mock, F., *et al*., 2019. Viral host prediction with deep learning. bioRxiv. doi:10.1101/575571.
Peng, M., *et al*., 2018. Luteolin escape mutants of dengue virus map to prM and NS2B and reveal viral plasticity during maturation. Antiviral Research 154, 87–96.
Pickett, B.E., *et al*., 2011. ViPR: An open bioinformatics database and analysis resource for virology research. Nucleic Acids Research 40 (Database issue), D593–D598. (PMID: 22006842).
Pickett, B.E., *et al*., 2012. Virus pathogen database and analysis resource (ViPR): A comprehensive bioinformatics database and analysis resource for the coronavirus research community. Viruses. 4 (11), 3209–3226.
Pickett, B.E., *et al*., 2013. Metadata-driven comparative analysis tool for sequences (meta-CATS): An automated process for identifying significant sequence variations that correlate with virus attributes. Virology 447 (1–2), 45–51. doi:10.1016/j.virol.2013.08.021.
Pinsky, B.A., *et al*., 2019. Methods and Reagents for Detection of Chikungunya Virus and Zika Virus, United States Patent Application. 20190024195.
Zhang, Y., *et al*., 2017. Influenza research database: An integrated bioinformatics resource for influenza research. Nucleic Acids Research 45 (Database issue), D466–D474. doi:10.1093/nar/gkw857.
Zou, C., *et al*., 2019 4. Virulence difference of five type I dengue viruses and the intrinsic molecular mechanism. PLOS: Neglected Tropical Diseases 13 (3), e0007202.

Relevant Websites

https://emergency.cdc.gov/recentincidents/index.asp
Centers for Disease Control and Prevention.
http://www.drugbank.ca
DrugBank.
http://empres-i.fao.org/empres-i
EMPRES-i - FAO.
https://www.who.int/influenza/gisrs_laboratory/flunet/en/
FluNet - WHO.
https://www.ncbi.nlm.nih.gov/genbank/
GenBank Overview - NCBI - NIH.
www.geneontology.org
Gene Ontology Resource.
https://www.gisaid.org
GISAID - Global Initiative on Sharing All Influenza Data.
https://idseq.net
IDseq.
www.iedb.org
IEDB.org: Free epitope database and prediction resource.
https://www.fludb.org
Influenza Research Database.
www.fludb.org
Influenza Research Database.
https://github.com/JCVenterInstitute/VIGOR4
JCVenterInstitute/VIGOR4: VIGOR4 - GitHub.
https://www.hiv.lanl.gov/content/index
Los Alamos National Laboratory.
https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database
National Center for Biotechnology Information.
https://nextstrain.org
Nextstrain.org.
http://sdb.im.ac.cn/vide/refs.htm
Plant Viruses Online: Descriptions and Lists from the VIDE Database.
http://dmk-brain.ecn.uiowa.edu/VOG/
pVOGs - Prokaryotic Virus Orthologous Groups.
https://www.ncbi.nlm.nih.gov/refseq/
RefSeq: NCBI Reference Sequence Database - NIH.
http://viruses.string-db.org
STRING Viruses.
https://hbvdb.ibcp.fr/HBVdb/
The Hepatitis B Virus database.
https://www.uniprot.org
UniProt.Org.
http://viperdb.scripps.edu/
VIPERdb - The Scripps Research Institute.
https://www.viprbrc.org/brc/dataSummary.spg?decorator=vipr
ViPR.
https://viralzone.expasy.org
ViralZone root - ExPASy.
https://viralzone.expasy.org/677
Virology links ~ ViralZone page.
http://virome.dbi.udel.edu
VIROME.

http://virusdetect.feilab.net/cgi-bin/virusdetect/index.cgi
VirusDetect.
https://www.viprbrc.org/
Virus Pathogen Database and Analysis Resource.
www.viprbrc.org
Virus Pathogen Database and Analysis Resource.
www.pdb.org
wwPDB: Worldwide Protein Data Bank.