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.
Introduction

A database is a collection of information organized in such a way that it supports searching and manipulations such as data sorting and retrieval. Most current databases are relational, in that the various related data types are stored in the rows and columns of specific tables. This facilitates searching and retrieval of data. In virology, databases have evolved from lab notebooks to simple spreadsheets and presently, to a variety of large sequence databases (non-virus specific) and virus-specific databases. In recent years, new sequencing technologies using massively parallel techniques have lead to plummeting costs for these services, which in turn have created a huge growth in the business of collecting virus genome sequence data. Managing the growing mountains of sequence data presents challenges beyond simply cataloguing and storing data. For example, new tools for data retrieval, comparison and visualization are needed to help virologists mine these resources. Alongside the sequence databases there has also been significant growth in database development in many other areas of the biological sciences. Gene expression (micro-array and RNAseq), protein structure, proteomic and gene ontology databases are a few examples that also impinge on virology. Although these databases aim to capture and store knowledge for researchers to use, the provision of easy access and retrieval of data sets in manageable and usable formats is by no means a trivial task.

With the advent of the big-data era our need for comprehensive, well-designed virus databases is also on the rise. A database must collect and store data within a structure that is highly organized, accessible, and searchable by the user, while allowing for extraction and manipulation of the data. The computer programs that aid in the organization of this information are known as ‘Database Management Systems’ (DBMS). They provide the programmer with the code and tools required to design, edit and manage the database and its desired functionalities. However, regardless of the quality of the DBMS program used, the utility of the database is determined by the overall quality of the design. Quality design is oriented around the needs of the user and the unique characteristics of the virus, supporting both the structure of the input data and the types of queries that will be applied to filter the data.

Just as different kinds of data require different databases, the various families of viruses also need different database types. This primarily stems from the needs of researchers that arise from the significant differences that exist between the various virus families. Fortunately, general database requirements tend to follow virus genome size, and small < 30 kb < large can be used as a crude classification scheme. For example, human immunodeficiency virus (HIV), hepatitis C virus (HCV), influenza virus and dengue virus are all relatively small RNA viruses with genomes that encode <15 genes. Although these viruses are dramatically different from each other, each of the isolates from a particular species normally has the same set of genes; therefore, differences are generally limited to single nucleotide polymorphisms (SNPs) with occasional recombination events. In contrast, the large DNA viruses, such as those in the Herpesviridae, Poxviridae, and Iridoviridae families, possess a wide variety of non-essential genes; the comparative analysis of these genomes often revolves around the presence or absence of sets of genes. Thus, very different types of analyses are performed on the small and large viruses leading to important differences in database structure and processing requirements. Another important difference that tends to be inversely related to genome size is the number of genomes that have been sequenced. In the past, this has been due to a combination of cost, technological challenges and research needs, so it will be interesting to see how this changes as the new sequencing technologies become more commonplace. As of writing (Jan 2016) there are approximately 19000 influenza virus, 2400 HCV, 5800 HIV, 270 poxvirus and 580 herpesvirus genomes fully sequenced.

*Change History: April 2016. K McLeod and C Upton has made changes throughout the text.*
This review will focus on the basics of database organization and presenting a current overview of available virus databases; it is aimed at the virologist rather than database guru. Discussions will be illustrated with a limited number of virus databases, including our own - the Viral Orthologous Clusters (VOCs) database, which forms the core of Virology.ca that supports researchers working with a variety of large dsDNA viruses.

**Databases**

**Design**

Depending where you are in the world, GenBank, (maintained by the National Center for Biotechnology Information, NCBI, USA), the EMBL Data Library (maintained in the UK) or the DNA Databank of Japan may first come to mind when thinking about bioinformatics databases. However, these large repositories of genomic, and other, sequence information have a relatively simple flat-file structure. When submitting a viral genome sequence to one these databases, the sequence data must be annotated with various information including the positions of coding DNA sequences (CDS). Flat-file databases can only be queried for information that is explicitly stated within the file. For example, queries can be made by keywords in the protein name, but not derivative characteristics such as gene length. In contrast, a relational database stores additional information (calculated or predicted), which is derived from data within the flat-file and organized into tables with defined relationships. In this way, relational databases permit queries related to these derivative characteristics. For example, the calculation and storage of protein size permits queries for 200 aa < proteins < 300 aa in length. When building such a relational database, the designer must define the structure of all of the data records to be stored and their association to each other. The database schema illustrates these relationships and acts as a map to the organization of the data storage locations, that is specific containers (dataset tables) for storing each type of dataset. Fig. 1 illustrates a sample of the schema for the VOCs RDMS. Schema are not literal representations of data organization, but they are useful for helping the database designers and users understand the basic structure, and allowing them to collaborate on the specifics of database design and structure.

Various file formats aim to capture this viral sequence data and associated knowledge, including GenBank and XML formats. GenBank is a flat-file format, which offers the significant advantage of a file format that is human-readable. Similarly, these files can be worked with using standard text editors or word-processing programs. A downside is that minor changes to the format of GenBank files can cause software tools that read these files (parsers) to fail if they have not been programmed to handle these differences. In attempts to more strictly manage the formatting of information while giving the flexibility of using various data types, some tools use an extensible markup language (XML). XML formatting explicitly defines relationships among data; data fields are embedded within each other to define their relationships. Although XML files are excellent for representing relationships in a form that can be read by a computer, they are more difficult for users to understand.

A relational database management system (RDBMS) functions to assist the designer in the creation of these databases. They consist of the relational database and the software required to manage the database. Some of the most popular are Oracle, MySQL, Microsoft SQL Server and PostgreSQL. However, if a database is to be useful to the virology research community, then there must be at least one additional component in this system – an easy to use interface to enable a virologist to interact with the database, i.e. send queries to the database and retrieve information in a variety of formats. Furthermore, the database itself may form part of an analysis pipeline. In this case, the information that is retrieved from the database is sent directly to an analysis tool, which subsequently returns the processed information to the user. Rather than the database returning a file of the sequences to be formatted before submission to a program such as ClustalO or MUSCLE for alignment, it might be desirable to have the database send the sequences directly to the alignment tool. The tool can then run the desired function before returning the processed data within a visualization tool that simplifies editing of the final MSA. Although simplification of this process with default parameters is clearly valuable to the user, over-simplification can lead to restricted function, i.e. a lack of ability to alter analysis parameters. Therefore, it is valuable to provide a variety of export options allowing users to obtain formats that can be used directly with multiple analysis tools and permit full range of parameter access, as discussed in later sections. Design of the user interface with potential queries and flexibility of data retrieval in mind is key to the utility of the database.

Even with the use of an RDBMS, the design of relational databases is not trivial and is best left to those with expertise and experience. Generally, the databases are optimized for two main functionalities: 1) storing and updating structured data with high integrity and 2) providing tools for searching, retrieving, summarizing and possibly analyzing the data. The database schema depicts the various data storage tables, which are based on data types such as: genome/gene/protein sequence, taxonomy or gene properties (position in genome, nucleotide/amino acid composition); the VOCs database also places all genes within an ortholog table (Fig. 1). The relationships between these tables are explicitly defined and are not limited to one-to-one relationships. For example, genome-to-gene relationships show a one-to-many relationship, whereas gene-to-ortholog relationships have a many-to-one relationship.

An additional aspect of any discussion on database design is maintenance, which can be viewed from several aspects. One aspect of maintenance is the process of maintaining the software and hardware over the life of the database - often the most difficult part is justifying and obtaining funding in a research environment. Over time, bugs need fixing, hardware fails and systems are upgraded – all are part of an ongoing process associated with many bioinformatics projects. A second view of maintenance, which is more integral to database design, concerns the data: 1) how will data be imported into the database? 2) will data need to be edited
Again, these questions have a natural division along the lines of virus genome type. For example, some form of automated process is essential to import the thousands of genomes required for an influenza or HCV (small genomes, fixed gene complement) resource, but these small sequences are unlikely to need subsequent modification. In contrast, a resource such as Virology.ca, which deals with large poxviruses, needs to contend with far less genomes, but must handle changes to genome annotation both before and after import into the database due to large, less well characterized, dsDNA genomes. Therefore, while a software script to automatically collect new genome GenBank files and insert them into a database might be feasible for an influenza virus database, this process cannot be used for poxvirus genomes in Virology.ca. The complexity of the latter frequently leads to variations (sometimes errors) in annotation protocols. Furthermore, new discoveries surrounding gene function mean that old genomes must sometimes be re-annotated. Therefore, key to the design of database systems like Virology.ca is the inclusion of a manual, user-friendly database-editing tool that will allow virologists to enter and maintain the data.

Another frequent need for database editing in Virology.ca results from the assignment of genes to ortholog groups. This grouping of evolutionarily related genes simplifies comparative studies of viruses, helping researchers discern relationships and shared function. For example, use of the highlight orthologs option in the Viral Genome Organizer (VGO) visualization tool (Fig. 2) assists in the comparison and understanding of genome synteny. Although orthology is simply the prediction of a common ancestor between genes of different species, the extreme diversity among viruses in a single taxonomic family creates difficulties in the accurate assignment of genes to ortholog groups. This and the existence of paralogs (related genes within a single genome) lead to a need for the ability to easily edit these ortholog assignments.

As alluded to above, one of the greatest challenges in working with bioinformatics databases is the availability of long term support. Even well designed databases deteriorate over time, so repairs and the need to accommodate evolving data sets lead to a

---

![Database Schema](image-url)  
**Fig. 1** Partial database schema from VOCs database at Virology.ca, created with MySQL Workbench.
need for some level of constant maintenance so that such databases remain functional and relevant. Unfortunately, this work is
sometimes viewed as a non-discovery based science and funding is problematic. As a result, the fate of many unfunded databases is
a slow fade into obsolescence. Dead links and non-functional tools on main webpages characterize these neglected databases.

 Genome Annotation

Although many bioinformatics databases may be in some way relevant to viruses, the majority of virologists tend to think of, and
also use most frequently, genome sequence databases. As a result, it would be hard to find a virologist who hasn’t performed a
BLAST search of these databases. Similarly, any virologist who has partially or completely sequenced a viral genome is familiar with
the process of annotating the sequence for submission to GenBank (or local equivalent) before their paper can be published. Three
levels of annotated information exist: basic information within a GenBank file that can be directly transferred into databases; values
calculated or predicted from the basic data; and curated information, beyond that provided when first developing the GenBank file,
added manually by users. The GenBank sequence file usually contains basic virus identification information, affiliations of the
annotator, CDS locations within the genome, metadata, gene and product type/names, as well as mRNA splicing information, if
appropriate. Many virus databases simply import annotations from GenBank sequence files and do not provide any further
information. However, the sequence data itself can be used to calculate various values that might be required by the end-user such
as A+T content, protein pI, molecular mass of proteins and amino acid content. Additionally, the data may be used as input for
predictive tools, such as those that look for functional motifs that are associated with enzymatic function or protein localization.
Manual curation can also provide additional information. For example, there is generally no information describing the likely
reliability of the annotations or the source of the information (experimental data, software prediction or scientific literature) within
a GenBank file. Therefore, inclusion of this data requires manual curation by users. Unfortunately, it is rare that this kind of
supplementary information is incorporated into databases in a manner that can be searched in a meaningful way.

 Keeping record of the reliability of annotations is especially important when investigating the genes of unknown function,
frequently found within the varied large dsDNA viruses. Functional assignment of the associated proteins relies heavily on the
transfer of function from orthologous proteins. Orthologous proteins, which may only be 20–30% identical, often do not match
the same set of functional motifs, and vary in their ability to match distantly-related proteins that might give important clues as to
function. In such investigations, it is useful to work with a set of diverse orthologs rather than the gene from a favourite virus.

As noted above, accurate genome annotation is critical to providing a useful bioinformatics database. However, this must be
built upon a foundation of accurate, and preferably complete, genome sequencing. The complexity of the annotation problem
increases exponentially with genome size and novelty. Small viruses have only a few genes and usually do not vary in gene content
within a species, whereas poxviruses range in size from 134 to 360 kb and encode between 129 and 335 genes. Large viruses, such
as these, bring two additional complications to the annotation problem: 1) only a subset of the viral genes are essential, so some
orthologs may be functional in one species but fragmented in another and, 2) some of the more diverse viruses have > 100 genes
currently annotated as "hypothetical", and even the DNA polymerase protein may share only 32% aa identity with its closest
relative. These problems are even greater in the Megaviruses that have genomes greater than 1000 kb, illustrating the increase of
annotation issues with genome size. Without the use of either a closely related genome, or multiple more distant relatives with
similar gene sets, for reference genomes in annotation, it is very difficult to predict which ORFs are: true genes, gene fragments, or
random small ORFs. Similar problems are encountered when virologists try to annotate the extremely diverse genomes of bacteriophages. Although some groups share common physical virion structures, often only a few genes are found in common between related groups of phage. This is due to their non-linear evolutionary pathway; bacteriophages commonly undergo recombination events generating mosaic gene distributions among genomes. Thus, the complexity of annotation relates directly to viral genome size and diversity.

As a result, it is relatively straightforward to map annotations from one small genome to another closely related genome. Additionally, this process can often be fully automated as part of a pipeline to import newly sequenced genomes into a database. Indeed, some resources such as the Influenza Research Database (IRD) provide a freely available, interactive version of their influenza annotation pipeline. The IRD pipeline aligns input nucleotide sequences against a consensus sequence profile to “identify possible sequencing errors, determine the influenza type, segment number, and for segments 4 and 6 of type A, the subtype, and translate the nucleotide sequence”.

In stark contrast, the annotation and import process for large viruses is much more time consuming and requires a human annotator with experience with the virus’ biology. At Virology.ca, we created the Genome Annotation Transfer Utility (GATU) to help the annotators with this process. The function of GATU is to remove easy, but tedious, genome annotation steps from the workflow of the annotators allowing them to focus on the more difficult decisions. GATU uses BLAST to search the target genome for each of the genes present in the reference genome. If they are matched over sufficient length and with sufficient identity, the gene is automatically scored for the target genome. Both partial matches and potential novel genes are flagged in the GATU interface for the annotator to investigate further and make a final decision; tools are included in GATU to make this process easier.

In addition to GATU, a large number of genome annotation tools have been developed and several are listed in Table 1. The majority of these tools depend upon similarity searches against previously characterized genes, which can result in under-annotation of unique or highly divergent genes. For this reason, the use of a general list of genes may be more useful than using a specific reference virus. Some annotation tools, especially for prokaryotic and eukaryotic genomes, attempt gene prediction - searching for promoter-like sequences and other gene characteristics, however, these are less useful for the annotation of viral genes.

As described above, the semi-automated process used by GATU places the human annotator in direct contact with the raw data and leaves the annotator in control of the decision making process. This is invaluable in the annotation of large and complex viruses, such as poxviruses. Fig. 3, panel A shows the organization of a normal poxvirus gene, and panels B, C and D illustrate 3 classes of mutations that often complicate the annotation of poxvirus genes and require annotator input. In panel B, a mutation has introduced a premature STOP codon that leads to a C-terminal truncation (25%) of the protein product. Although this change might alter a protein’s function, most annotators would probably annotate this altered gene. In panel C, a mutation has introduced a premature STOP codon that leads to a C-terminal truncation (80%) of the protein product (labeled Ci). In this example the greatly shortened protein is very unlikely to be functional, however, because the second part of the gene is still present and represents 70% of the original gene this is sometimes annotated (Gii) even though it is very unlikely to be translated from the mRNA. In panel D, a mutation has not changed the gene or promoter, but instead extended the ORF to a position upstream of the promoter. Although the protein product is not changed, the larger ORF is sometimes erroneously annotated. These annotation errors illustrate the need for more complex automated evaluation of annotation results or the continued input of the human annotator.

Database Limitations and Errors

Although the various databases and associated sequence searching/analysis tools comprise a very valuable set of resources, it is also important to recognize their limitations. Researchers need to remain wary of the various types of information because regardless of the source, errors and out-of-date files can fall through the cracks. Even the primary repositories are not immune to such issues. The two primary sources of errors are the genome sequencing and genome annotation processes. If available, reference genomes can be very valuable and be used to predict/correct potential errors either automatically or manually. For example, a nucleotide substitution that introduces a STOP codon into the middle of the poliovirus single large ORF will certainly not be missed because it is not consistent with a functional genome. However, many of the genes present in the large DNA viruses such as poxviruses are non-essential; they can be, and often are, truncated in some viruses and complete in others. Thus, researchers must remain aware of potential errors, regardless of the source or genome type.

Annotation errors come in a variety of forms:

- **Errors due to sequencing errors** – often introduction or loss of STOP codons.
- **Errors of omission** - examples include genes only recently discovered, and therefore not annotated in older data sets.
- **Errors of ignorance** – a simple lack of knowledge: ortholog/paralog relationships not understood or viruses mis-named.
- **Errors of propagation** – the transfer of annotation errors from older data to new genomes.
- **Errors of opinion** – based upon varying annotation criteria. For example, decision to annotate ORFs >40 codons versus annotate ORFs >50 codons.
### Table 1  
Virus databases, tools, and resources

| Resource name   | Sponsor                        | URL                          | Description                                                                 |
|-----------------|--------------------------------|------------------------------|-------------------------------------------------------------------------------|
| **Sequence database** |                                |                              |                                                                                |
| ACLAME          | University of Brussels         | http://aclame.ulb.ac.be/     | Database of mobile genetic elements                                          |
| ATIVS           | NRI, Taiwan                    | http://influenza.nhri.org.tw/ATIVS/ | Influenza virus surveillance                                                   |
| AVpdb           | IMT, India                     | http://crdd.osdd.net/servers/avpdb/ | Database of antiviral peptides                                                |
| AVPpred         | IMT, India                     | http://crdd.osdd.net/servers/avpred/ | Prediction of antiviral peptides                                              |
| hNAberr         | Sanford Med Res. Inst.         | http://hnaberr.org/          | HIV Broadly Neutralizing Antibodies                                            |
| CAPHI           | NHRI                           | http://bioinfo-dbb.nhri.org.tw/capih/ | Protein interactions for HIV-1                                                |
| COGs            | NCBI                           | http://www.ncbi.nlm.nih.gov/COG/ | Phylogenetic classification of proteins                                       |
| Bat Assoc. Viruses | Pittsburgh Phage Institute     |                              | Database of Bat Associated Viruses                                            |
| euhCovdb        | CNRS and U of Lyon             |                              | European hepatitis C virus database                                           |
| eurest          | Karolinska Inst. Sweden        |                              | Clinical data to predict anti-HIV therapy                                     |
| ExPaSy          | SIB                            | http://www.expasy.org/       | Analysis tools for protein sequences                                          |
| FLAVIdb         | Dana-Farber Cancer Inst        |                              | Flavivirus database                                                           |
| HBVRegDB        | University of Otago, NZ        |                              | HBV Regulatory Sequence Database                                              |
| HCV DB Project  | NIAID                          |                              | HCV sequence and immunology                                                    |
| Hepatitis B Virus DB | IBCP                           |                              | HBV sequences and viral resistance                                            |
| HIV/EBola DB    | Nati Inst. Genetics, Japan     |                              | Database for hepatitis B, C, and E                                            |
| HIVdb           | IMT, India                     |                              | HFV sequence, immunological data                                              |
| HIV Drug Resist. | Stanford University           |                              | Database of HIV inhibiting peptides                                            |
| HIV Sequence    | Los Alamos National Database   |                              | HIV drug resistance database                                                   |
| Database        | Laboratories; NIAID           |                              | HIV sequence, resistance, immunology, and vaccine trials databases            |
| HIV-1, Human Protein Interactions | SRI, Birmingham, Alabama; NCBI |                              | Interactions of HIV-1 proteins with those of the host cell                    |
| HIVSirDB        | IMT, India                     |                              | A database of HIV inhibiting siRNAs                                           |
| HTLV-1 Molecular Epidemiology DB | Brazilian Ministry of Health |                              | Clinical, phylogenetic and epidemiological data from HTLV-1                  |
| Influenza Resource | NCBI                           |                              | Influenza sequence database                                                   |
| IRD             | NIAID                          |                              | Exponentially verified IRES structures                                        |
| IRESite         | Charles U, Czech Rep.          |                              | Analysis platform for influenza virus                                          |
| IVDV            | Beijing Genomics Inst          |                              | Influenza Sequence & Epitope Database                                          |
| KISED           | Korean NHI                     |                              | NCBI resources for the study of viruses                                       |
| NCBI Genomes    | NCBI                           |                              | Search flu sequences by type, location                                         |
| OpenFlu database | SIB                            |                              | Papillomavirus sequence and structures                                        |
| PaVE            | NIAID                          |                              | List of all available phage genomes                                           |
| Phage Genomes   | ENA                            |                              | Phylogenetic exploration of viruses                                           |
| PhEVER          | PRABI                          |                              | Database of gene regulation in phage                                           |
| phiSITE         | Inst. of Mol Biol             |                              | RNA Virus Database                                                             |
| RNA Virus DB    | BioAfrica                      |                              | SARS Coronavirus sequence database                                            |
| SARS Cov. Res.  | NCBI                           |                              | Viroids and viroid-like RNAs                                                   |
| Subviral RNA    | University of Ottawa           |                              | Pathogen sugar-binding database                                               |
| Sugar Bind DB   | SIB                            |                              | HBV database - sequence, molecular, clinical and epidemiological data          |
| Repository for HBV Strain Data | HPA, Department of Health (UK) |                              | Viral protein sequences and tools                                              |
| VaYzMoI0        | AFMB, France                   |                              | Analytical tools for large DNA viruses                                         |
| Virology.ca     | U of Victoria, Canada          |                              | Icosahedral virus capsid structures                                            |
| VIPERdb         | Scripps                        |                              | Predicted viral mRNA hairpins                                                 |
| Vir-Mir Database | IBS, Taiwan                    |                              | Protein sequences in RNA viruses                                               |
| VIRAD           | University of Cambridge        |                              | RNA Virus Database                                                             |
| ViralIDB - HUG  | Harbin Med U, China            |                              | Virus-host ncRNA interaction database                                          |
| ViRBase         | Oklahoma State U              |                              | Oligonucleotides for PCR                                                       |
| ViRolo          | IMT, India                     |                              | Database of viral siRNA and shRNA                                             |
| ViRSiRNAdb      | NCBI                           |                              | Virus Pathogen Resource                                                       |
| VIPR            | NIAID                          |                              | Database of viral sequences                                                    |
| Virus Variation | NCBI                           |                              | Human and viral proteins interactions                                          |
| VirusMint       | U of Rome Tor Vergata          |                              | HIV life cycle and host interaction                                            |
| Resources       |                                |                              |                                                                                |
| Bio Afrika      | South African MRC             |                              |                                                                                |
| Resource name                          | Sponsor                      | URL                               | Description                                                                 |
|---------------------------------------|------------------------------|-----------------------------------|-----------------------------------------------------------------------------|
| BioHealthBase (Influenza)             | Texas Southwestern Medical Center; NIAID | [http://www.biohealthbase.org/](http://www.biohealthbase.org/) | Bioinformatics resources for influenza                                      |
| Infectious Diseases BRCs              | NIAID, NIH                   | [http://www.niaid.nih.gov/labsandresources/resources/dmid/brc](http://www.niaid.nih.gov/labsandresources/resources/dmid/brc) | Genomics resources for viruses in category A, B, and C priority pathogens Bioinformatics resources for Dengue |
| Dengue Virus                          | Broad Institute              | [http://www.broadinstitute.org/annotation/viral/Dengue](http://www.broadinstitute.org/annotation/viral/Dengue) | Collects and locates human disease-related viral integration sites Ebola virus sequences. |
| Dr. VIS                               | Peking Union Medical College Hospital | [http://www.bioinfo.org/drvis/](http://www.bioinfo.org/drvis/) | Analysis of human viral pathogens Database for Bacteriophage               |
| Ebola Resource                        | NCBI                         | [http://www.ncbi.nlm.nih.gov/genome/viruses/variation/ebola](http://www.ncbi.nlm.nih.gov/genome/viruses/variation/ebola) | List of largest sequenced viruses Resources for HIV systems biology Caulimoviridae, Geminiviridae, and Potyviridae database |
| Epidemic                              | U. of Edinburgh, UK          | [http://epidemic.bio.ed.ac.uk/](http://epidemic.bio.ed.ac.uk/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| Félix d’Hérelle Ref Center for phage | Université Laval             | [http://www.phage.ulaval.ca/](http://www.phage.ulaval.ca/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| Giant Virus                           | Struc & Genom Info Lab       | [http://www.giantvirus.org/top.html](http://www.giantvirus.org/top.html) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| HIV Systems Biol. ILTAB               | U. of Pennsylvania           | [http://hivsystemsbiology.org/](http://hivsystemsbiology.org/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| Influenza Genome Sequencing Project   | Danforth Plant Sciences, St. Louis, Missouri | [http://www.danforthcenter.org/iltab/](http://www.danforthcenter.org/iltab/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| OpenHelix                             | [http://www.openhelix.com/](http://www.openhelix.com/) | Learn to use Bioinformatics Resources Repository of the virus fossil record Ebola sequence browser Resource for Viral hemorrhagic fevers |
| Paleovirology                         | University of Glasgow        | [http://bioinformatics.cvr.ac.uk/paleovirology/](http://bioinformatics.cvr.ac.uk/paleovirology/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| UCSC Ebola Portal                     | U. of California SC          | [http://genome.ucsc.edu/ebolaPortal/](http://genome.ucsc.edu/ebolaPortal/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| Viral Hemorrhagic Fever Consortium    | Tulane University            | [http://vhfc.org/home](http://vhfc.org/home) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| Viral Zone                            | SIB                          | [http://viralzone.expasy.org/](http://viralzone.expasy.org/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| VirMut                                | U. of Valencia, Spain        | [http://www.uv.es/rsanjuan/virmut](http://www.uv.es/rsanjuan/virmut) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| WHO Ebola stats                       | WHO                          | [http://apps.who.int/gho/data/node.rawdata/ebola-sitrep](http://apps.who.int/gho/data/node.rawdata/ebola-sitrep) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| Analitical tools                      |                              |                                   | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| Base-By-Base                          | U. of Victoria, Canada       | [http://athena.bioc.uvic.ca/virology-ca-tools/base-by-base/](http://athena.bioc.uvic.ca/virology-ca-tools/base-by-base/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| CBS Prediction Servers                | Technical University of Denmark | [http://www.cbs.dtu.dk/services/](http://www.cbs.dtu.dk/services/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| FluGenome                             | CDC                          | [http://www.flugenome.org/](http://www.flugenome.org/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| GATU                                  | U. of Victoria, Canada       | [http://atlas.bioc.uvic.ca/virology-ca-tools/gatuf](http://atlas.bioc.uvic.ca/virology-ca-tools/gatuf) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| GiRaF                                 | Genom Inst of Singapore     | [http://www.cs.cmu.edu/~cckingsf/software/gira/](http://www.cs.cmu.edu/~cckingsf/software/gira/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| HESAS                                 | Pusan National University, South Korea | [http://www.primate.or.kr/HESAS/](http://www.primate.or.kr/HESAS/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| HIV Therapy Simulator                 | University of Washington    | [https://sites.google.com/site/hivsimulator/](https://sites.google.com/site/hivsimulator/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| HIVCD                                 | Brigham Young U             | [http://sourceforge.net/projects/hivcd/](http://sourceforge.net/projects/hivcd/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| IVA                                   | Sanger Institute            | [http://sanger-pathogens.github.io/iva/](http://sanger-pathogens.github.io/iva/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| Jdotter                               | U. of Victoria, Canada       | [http://athena.bioc.uvic.ca/virology-ca-tools/jdotter/](http://athena.bioc.uvic.ca/virology-ca-tools/jdotter/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| jpHMM                                 | University of Gottingen      | [http://jphmm.gobics.de/](http://jphmm.gobics.de/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| LearnCoil-VMF                         | MIT                          | [http://groups.csail.mit.edu/cb/learncoil-vmf/cgi-bin/vmf.cgi](http://groups.csail.mit.edu/cb/learncoil-vmf/cgi-bin/vmf.cgi) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| META VIR                              | Blaise Pascal University     | [http://metavir-meb.univ-bpclermont.fr/](http://metavir-meb.univ-bpclermont.fr/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| NCBI Genotyping                       | NCBI                         | [http://www.ncbi.nlm.nih.gov/projects/genotyping/](http://www.ncbi.nlm.nih.gov/projects/genotyping/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| PASC                                  | NCBI                         | [http://www.ncbi.nlm.nih.gov/sutils/pasc/](http://www.ncbi.nlm.nih.gov/sutils/pasc/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| PhAnToMe                              | NCBI                         | [http://www.ncbi.nlm.nih.gov/sutils/phantom/](http://www.ncbi.nlm.nih.gov/sutils/phantom/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| PhyloType                              | CBI France                   | [http://web-mab.lirmm.fr/phylotype/](http://web-mab.lirmm.fr/phylotype/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| PrimerHunter                          | U. of Connecticut           | [http://dna.engr.uconn.edu/?page_id=85](http://dna.engr.uconn.edu/?page_id=85) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| PriSM                                 | Broad Institute             | [http://www.broadinstitute.org/annotation/viral/seq/specialprojects/primerDesign.cgi](http://www.broadinstitute.org/annotation/viral/seq/specialprojects/primerDesign.cgi) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| RDP4                                  | University of Cape Town      | [http://web.cbio.uct.ac.za/~darren/rdp](http://web.cbio.uct.ac.za/~darren/rdp) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| RetroTector                           | Uppsala U, Sweden           | [http://retrotector.neuro.uu.se/](http://retrotector.neuro.uu.se/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| RotaC                                 | U of Leuven, Belgium        | [http://rotac.regatools.be/](http://rotac.regatools.be/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |
| SeqMap                                | Indiana University          | [http://seqmap.compbio.iupui.edu/](http://seqmap.compbio.iupui.edu/) | General virus resource Spontaneous mutations rate for viruses Ebola sequence browser Resource for Viral hemorrhagic fevers |

(Continued)
Table 1  (Continued)

| Resource name | Sponsor | URL | Description |
|---------------|---------|-----|-------------|
| siVirus       | University of Tokyo | http://sivirus.rnai.jp/ | Antiviral siRNA design |
| VFS           | Hong Kong Bioinformatics Centre | http://sourceforge.net/projects/viralfusionseq/ | Discover viral integration events and fusion transcripts |
| VIPS          | CYC University, Taiwan | http://140.135.61.250/vips/ | Viral IRES prediction system |
| ViReMa        | Scripps | http://sourceforge.net/projects/virema/ | Recombination or fusion detection |
| VirHostNet    | PRABI | http://virhostnet.prabi.fr/ | Visualisation of virus-host interactions |
| ViroBLAST     | University of Washington | http://indra.mullins.microbiol.washington.edu/viroblast/viroblast.php | Comparison and contamination on viral research |
| viRome        | Edinburgh Genomics | http://www.ark-genomics.org/bioinformatics/virome | Visualization of NGS data |
| VIROME        | University of Delaware | http://virome.dbi.uel.edu/ | Resource for Metagenome Exploration |
| VirsiRNApred  | IMT, India | http://crdd.osdd.net/servers/virsirnapred/ | Viral siRNA efficacy prediction server |
| Virus Genotyping Tools | Africa Centre for Health and Population Studies | http://bioafrica.mrc.ac.za/rega-genotype/html/ | Virus genotyping tools |
| VirusFinder   | Vanderbilt University | http://bioinfo.mc.vanderbilt.edu/VirusFinder/ | Characterize intra-host viruses |
| VirusHunter   | WashU, St. Louis | http://pathology.wustl.edu/VirusHunter/ | Identify viral sequence origin |
| VirusSeq      | The University of Texas | http://odin.mdacc.tmc.edu/~xsu1/VirusSeq.html | Detect viral integration |
| Visitor       | Pasteur Institute | http://drosophila.org/gedlab/?page_id=254 | Viral siRNA sequences in Drosophila |
| Vita          | NCT University, Taiwan | http://vita.mbc.nctu.edu.tw/ | Predict viral miRNA targets in hosts |
| Lists of databases | | | | |
| NAR 2016 Database Issue | Journal: Nucleic Acids Research | http://nar.oxfordjournals.org/content/44/D1.toc | Issue describing molecular biology databases |
| NAR 2015 Web Server Issue | Journal: Nucleic Acids Research | http://nar.oxfordjournals.org/content/43/W1.toc | Issue describing molecular biology websites |
| General sites | | | | |
| DDBJ          | DNA Data Bank of Japan | http://www.ddbj.nig.ac.jp/ | Nucleotide and protein sequences |
| EMBL          | Euro Mol Biol Lab | http://www.ebi.ac.uk/ena | Nucleotide and protein sequences |
| GenBank       | NCBI | http://www.ncbi.nlm.nih.gov/genbank/ | Nucleotide and protein sequences |
| Assembly Archive | NCBI | http://www.ncbi.nlm.nih.gov/Traces/assembly/ | Repository of sequence assembly files |
| Trace Archive  | NCBI | http://www.ncbi.nlm.nih.gov/Traces/ | NCBI repository of sequence trace files |
| Viral Genomes  | NCBI | http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi | List of all virus complete genomes |
| PubMed        | NML, NIH | http://www.pubmed.gov/ | Biomedical literature database |
| UniProt       | EMBL | http://www.ebi.ac.uk/uniprot/ | Protein sequence repository |
| Taxonomy      | | | | |
| ICTVnet       | ICTV | http://www.ictvonline.org/ | ICTV official website |
| ICTvdb        | ICTV | http://ictvdb.bio-mirror.cn/index.htm | ICTV database of virus taxonomy |
| Taxonomy Browser Functional Motifs | NCBI | http://www.ncbi.nlm.nih.gov/Taxonomy/ | Taxonomy-based retrieval of sequences |
| interPro      | EBI | http://www.ebi.ac.uk/interpro/ | Protein families, domains, and motifs |
| CDD           | NCBI | http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml | Conserved Domain Database |
| PFAM          | WashU, St. Louis | http://pfam.wustl.edu/ | Proteins and Hidden Markov models |
| PRINTS        | U of Manchester, UK | http://www.bioinf.man.ac.uk/ddbbrowser/PRINTS/ | Conserved protein motifs |
| PROSITE       | SIB | http://ca.expasy.org/prosite/ | Database of proteins |
| VOCs          | U of Victoria, Canada | http://athena.bioc.uvic.ca/virology-ca-tools/vocs/ | Viral orthologous clusters database |
| Structural    | | | | |
| PDB           | Struct. Bioinformatics | http://www.rcsb.org/ | 3D structure database |
| PDBe          | EBI | http://www.ebi.ac.uk/pdbe/ | Macromolecular structure database |
| Dali          | U. of Helsinki, Finland | http://ekhidna.biocenter.helsinki.fi/dali/ | Database of 3D structure comparisons |
| The Big Picture Book of Viruses | Tulane University | http://www.virology.net/Big_Virology/ | Catalog of virus pictures on the Internet |
| Virus World Other | | | | |
| GO            | GO Consortium | http://www.geneontology.org/ | Images of virus capsids |
| NC80          | Stanford University | http://bioontology.org/ | Vocabularies to describe gene products |
| GEO           | NCBI | http://www.ncbi.nlm.nih.gov/geo/ | Technologies for biomedical ontologies |
| | | | | | |
Table 1  (Continued)

| Resource name | Sponsor | URL | Description |
|---------------|---------|-----|-------------|
| IEDB          | La Jolla Institute for Allergy and Immunology | http://www.iedb.org/ | Dissemination of immune epitope information |
| IntAct        | EBI     | http://www.ebi.ac.uk/intact/ | Analysis system for protein interactions |
| Medically oriented |        |     |             |
| AIDSinfo      | US Dept of Health | http://aidsinfo.nih.gov/ | Information on HIV/AIDS research |
| NCID          | US CDC  | http://www.cdc.gov/ncidod/diseases/ | US CDC disease index |
| CIPHI Fact Sheets | CIPHI  | http://www.ciphi.on.ca/resources | Information for public health and viruses |
| CSEI          | St Louis University | http://www.bioterrorism.slu.edu/ | Information on emerging infections |
| Diseases & Conditions | New York State | http://www.health.ny.gov/diseases/ | General public information |
| ENIVD         | European Commission | http://www.enivd.org/ | Diagnostics of ‘Imported’ Viral Diseases |
| EPR           | WHO     | http://www.who.int/csr/disease/en/ | Global alert and response system |
| European CDC  | The European Union | http://ecdc.europa.eu | Disease Prevention and Control |
| MedScape      | WebMD   | http://www.medscape.com/ | Information and educational tools |
| Manual of Infectious Diseases | Merck & Co., Inc. | http://www.merckmanuals.com/professional/infectious-diseases | Online reference of infectious diseases |
| MMWR          | CDC     | http://www.cdc.gov/mmwr/ | Morbidity and Mortality Weekly Report |
| VRC           | NIH     | http://www.vrc.nih.gov/ | Vaccine research center |

Fig. 3  Common annotation problems. Panel A: normal gene; Panel B: small C-term truncation; Panel C: large C-term truncation; Panel D: 5’ extension to ORF. Blue boxes show ORFs; Red box/P/arrow indicates promoter and transcript, which do not change; * indicates STOP codons; Colored boxes/letters indicate predicted polypeptides.
Another limitation of most databases is the absence of evidence for particular annotations. Examples of evidence types include: experimental, literature and sequence relationship to another virus/gene. However, even when evidence notes are present, there may be problems since experiments may eventually be discredited and/or require updates. Curating such a system, essentially with the annotations as a living document, would also be incredibly labor intensive.

The lack of an unambiguous and controlled naming standard that is carried across all viruses and databases results in variable descriptions and data that are difficult to query for specific characteristics. For example, software may not be programmed to recognise synonymous terms such as "ssDNA," "single strand DNA," "single-stranded DNA," and "single stranded DNA" as the same. Where possible, import systems should have locks on permitted words, based on a universal, controlled vocabulary, in order to reduce this ambiguity. For example, to only permit "ssDNA". Presently, the Gene Ontology Consortium (GO; www.geneontology.org) offers a controlled vocabulary that is unique and informative. Since 2006, GO has expanded and standardised viral terms and, where appropriate, has cross-referenced and aligned these with the ViralZone (viralzone.expasy.org) and UniProtKB-keywords (among other resources). In addition, although some terms will still be applicable to only particular viruses (eg, the phages, see viral head-tail joining; GO:0098005), GO viral terms are designed to be species-neutral where possible. However, complete implementation of a consistent and accurate vocabulary, such as that presented by GO, will only work if scientists choose to sustain it through their own participation. But the reward is an overarching, standardized vocabulary for all species and databases that simplifies comparisons and reduces misspelling and ambiguity.

Database Searching

Although the quality of raw data, annotations, and database structure is clearly important, a database is only as useful as its search tools. The database must be able to execute the questions posed by virologists, which for user convenience is often accomplished through the use of a Graphical User Interface (GUI). Although it is impossible to predict all of the queries that might be requested, the system should be flexible enough to provide a reasonably close search that may require minor post-search data processing. Therefore, it is important for database designs to accommodate the needs of the researcher and the characteristics of the virus type. For example, in a study examining the genomic variation of H1N1 influenza viruses obtained from humans between the years of 2000–2010, the researchers would have specific search parameters regarding influenza A subtype, host species, and year of isolation. If the search interface did not permit all of these search parameters, then the researcher would be left with an arduous task of manually sorting the results that, depending upon the volume of information and computer skills of the user, may be far too time consuming. Another aspect of the searching, which is virus-specific, is that for viruses with high sequencing volumes at similar times and locations, such as influenza, there may be many identical genomes in the database. Therefore, databases dealing with this type of virus need to have an exclude identical genomes filter. Examples illustrating typical database queries for small and large genomes are discussed below and displayed in Table 2.

Although it would be preferable to have database resources supported long term so that they can respond to users requests for new queries etc., it is not necessarily an efficient use of resources to build every feature requested by users into the database interface. Clearly a cost-benefit analysis must be performed on the requests for features to be included in the software so that money and effort can be targeted at the most-used database functions. However, the system must be able to provide users that have one-off analyses some basic filters to work with, while users must accept that it may off-load some data analysis to them. This review will discuss database-searching features of individual virus databases in following sections.

Database Output (Utilizing Information)

Sequence databases usually provide users with similar basic options for output. FASTA formatting of nucleotide and protein sequences is a standard because multiple sequences can be incorporated into one file, and they can be read by many bioinformatics programs. However, most annotations are stripped out of these files. In contrast, GenBank files contain gene annotations, but these may be problems since experiments may eventually be discredited and/or require updates. Curating such a system, essentially with the annotations as a living document, would also be incredibly labor intensive.

The lack of an unambiguous and controlled naming standard that is carried across all viruses and databases results in variable descriptions and data that are difficult to query for specific characteristics. For example, software may not be programmed to recognise synonymous terms such as “ssDNA,” “single strand DNA,” “single-stranded DNA,” and “single stranded DNA” as the same. Where possible, import systems should have locks on permitted words, based on a universal, controlled vocabulary, in order to reduce this ambiguity. For example, to only permit “ssDNA”. Presently, the Gene Ontology Consortium (GO; www.geneontology.org) offers a controlled vocabulary that is unique and informative. Since 2006, GO has expanded and standardised viral terms and, where appropriate, has cross-referenced and aligned these with the ViralZone (viralzone.expasy.org) and UniProtKB-keywords (among other resources). In addition, although some terms will still be applicable to only particular viruses (eg, the phages, see viral head-tail joining; GO:0098005), GO viral terms are designed to be species-neutral where possible. However, complete implementation of a consistent and accurate vocabulary, such as that presented by GO, will only work if scientists choose to sustain it through their own participation. But the reward is an overarching, standardized vocabulary for all species and databases that simplifies comparisons and reduces misspelling and ambiguity.

### Database Searching

Although the quality of raw data, annotations, and database structure is clearly important, a database is only as useful as its search tools. The database must be able to execute the questions posed by virologists, which for user convenience is often accomplished through the use of a Graphical User Interface (GUI). Although it is impossible to predict all of the queries that might be requested, the system should be flexible enough to provide a reasonably close search that may require minor post-search data processing. Therefore, it is important for database designs to accommodate the needs of the researcher and the characteristics of the virus type. For example, in a study examining the genomic variation of H1N1 influenza viruses obtained from humans between the years of 2000–2010, the researchers would have specific search parameters regarding influenza A subtype, host species, and year of isolation. If the search interface did not permit all of these search parameters, then the researcher would be left with an arduous task of manually sorting the results that, depending upon the volume of information and computer skills of the user, may be far too time consuming. Another aspect of the searching, which is virus-specific, is that for viruses with high sequencing volumes at similar times and locations, such as influenza, there may be many identical genomes in the database. Therefore, databases dealing with this type of virus need to have an exclude identical genomes filter. Examples illustrating typical database queries for small and large genomes are discussed below and displayed in Table 2.

Although it would be preferable to have database resources supported long term so that they can respond to users requests for new queries etc., it is not necessarily an efficient use of resources to build every feature requested by users into the database interface. Clearly a cost-benefit analysis must be performed on the requests for features to be included in the software so that money and effort can be targeted at the most-used database functions. However, the system must be able to provide users that have one-off analyses some basic filters to work with, while users must accept that it may off-load some data analysis to them. This review will discuss database-searching features of individual virus databases in following sections.

### Database Output (Utilizing Information)

Sequence databases usually provide users with similar basic options for output. FASTA formatting of nucleotide and protein sequences is a standard because multiple sequences can be incorporated into one file, and they can be read by many bioinformatics programs. However, most annotations are stripped out of these files. In contrast, GenBank files contain gene annotations, but these may be problems since experiments may eventually be discredited and/or require updates. Curating such a system, essentially with the annotations as a living document, would also be incredibly labor intensive.

| Queries for Influenza Viruses | Queries for Poxviruses |
|-------------------------------|------------------------|
| Find all H1 FluA viruses      | Find all gene families in CPXV |
| Find segment 3 from all FluA viruses | Find all gene families present in CPXV but absent from MPXV |
| Find all FluA viruses isolated in Belgium | Find all viruses containing TNF binding protein |
| Find segment 3 from FluA viruses isolated in Belgium between the years of 1990–2005; exclude incomplete sequences | Find orthologs of gene #28 from MPXV-Zaire |
| Find all FluA viruses isolated from swine | Find viruses without hemagglutinin |
| Find all proteins from segment 5; exclude redundant sequences | Find genes unique to Yokapoxivirus |
| Find swine FluA genomes isolated in USA 1990–2005; exclude incomplete sequences | Find all genes containing the motif ATGGGTTAGANNNNNNTATAGTGA |
| Find all FluA H1 proteins isolated 1990–2005, exclude H1N1 pandemic sequences | Find all proteins containing the motif TGKHYTLMA |
files can be tricky for software to read due to non-standard formatting errors. The most basic output format is comma-separated values (csv), a tabular output that can often be read by other software, or even spreadsheet programs such as Microsoft Excel. Yet, for many researchers data in these formats is very tedious to use. Therefore, databases are frequently paired with visualization and analysis tools, or permit the export of data in a format that can be accepted by external programs. For example, the Virology.ca database, VOCs, is linked to the Base-By-Base (BBB) visualization tool, which displays and allows editing of MSAs. Additionally, the VOCs database provides users with a series of standard export options to permit use in other applications: 1) write selected sequences to a FASTA file, 2) write selected sequences to GenBank files, and 3) send selected sequences to a window in FASTA format, allowing easy copy/paste functions.

**Virus-associated Databases**

Databases are commonly divided by data type; however, this is not as simple as it might seem. For example, a series of databases, each dedicated to a family of viruses, might all need to support many different types of molecular biological data. Alternatively, databases managing a particular data type (eg, sequence or virion structure) would need to deal with many taxonomically different viruses. Table 1 presents a summary of both of these database types.

Although this article aims to provide an updated overview of the biological databases relevant to viruses, as a print publication it is important to note that this resource will become quickly out-dated. New databases will be developed, others will gain features or improved algorithms for annotation and, sadly, a lack of funds will be the demise of someone’s favourite database.

When searching for new or updated biological databases, a good place to start is the Nucleic Acids Research Database issue (<http://www.oxfordjournals.org/our_journals/nar/database/c/>). Alternatively, *Google may be your friend*, but it will help if the name of the database is not easily confused with a variety of other Internet resources. The Virus Pathogen Resource (ViPR, pronounced viprbrc) is particularly tricky to find if you don’t know it’s a Bioinformatics Resource Center and therefore you should look for viprbrc. The search is further complicated by the existence of VIPERdb, a separate database of virus capsid structures and the VIPRE Antivirus software tool.

**Nucleotide Sequence Data**

Although this review focuses on virus databases, most of these are reliant on other generic databases as the source of their data. The three main repositories for nucleotide sequence data are GenBank from The National Center for Biotechnology Information (NCBI), DNA Data Bank of Japan (DDBJ), and The European Bioinformatics Institute, a part of the European Molecular Biology Laboratory (EMBL-EBI). Collectively, the databases form The International Nucleotide Sequence Database Collaborative (INSDC). All nucleotide sequence data submitted to the INSDC is shared among the databases. This provides up-to-date public access to nucleotide sequence data that can be accessed through any of the three interfaces. In an attempt to keep the databases up-to-date, most scientific journals require that genome sequences are submitted to one of the INSDC databases prior to publication. INSDC also hosts databases for raw sequencing data and alignment information used to create the final genome sequences, including the Trace Archive for capillary reads, and Sequence Read Archive for Next Generation Sequence reads. This information helps in the reproducibility of genome assemblies in cases requiring review, and allows the user to analyse whether unexpected results are characteristic of the virus or a systematic artifact. However, advances in sequencing technology raise the question of the value of saving raw sequencing data; re-sequencing of samples (given their availability) is fast, easy, cheap and increasingly accurate.

GenBank submissions are owned by the original submitters who retain exclusive editing permission. Therefore, if sequence errors are detected or new genes discovered within a genome by other research groups, it might be impossible to update the original submission. To deal with this problem, NCBI has created the Reference Sequences resource (RefSeq), which offers up-to-date reference genomes for taxonomically diverse organisms, including viruses. Each non-redundant RefSeq file is treated to ongoing curation by NCBI staff and/or collaborators (specified within the file) thereby keeping it current and correctly formatted with regards to sequence, annotation and citation data. New revisions are given distinct accession numbers. RefSeq files are not limited to nucleotide sequences, but also offer transcript and protein sequence data.

**Protein Sequence Data**

The Universal Protein Resource (UniProt) complements the function of the INSDC, providing a comprehensive protein sequence and annotation database consortium of three collaborators: EMBL-EBI, the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). The database is divided into several branches: UniProtKB, the protein knowledgebase with two subsections, TrEMBL (translated EMBL Nucleotide sequence data library) that stores automatically annotated proteins prior to review and Swiss-Prot, containing proteins that have been manually annotated and reviewed, and often have associated literature; UniParc functions as an archive, sorting new, revised and obsolete sequences with a non-redundant numbering scheme allowing outdated UniProt references from past literature to be traceable; and UniRef100, 90 and 50 branches that cluster proteins into groups of 100%, 90% and 50% aa identity, respectively. To ensure up-to-date public access, new protein sequences must be
submitted to UniProt prior to publication, the new protein sequences associated with a genome submitted to GenBank are automatically added.

**Structural**

Knowledge of a protein’s 3D structure can assist in the prediction of its functions and interactions, which are important aspects of understanding viral processes and drug and vaccine-design. As a result, 3D structures have been determined for many viral proteins, and in some cases, of the complete virion. The Protein Data Bank (PDB) collects all biochemical structures, but searches can be limited to the structures of viral proteins. In recent years, the number of viral structures in PDB has grown significantly. At the end of 2015, almost 7% of the total of 114,000 structures were viral in origin. The PDB also offers a variety of visualization and analysis tools to help the occasional user.

There are also several virus-specific structural databases that provide specialized virus-related information. These include: VIPERdb, which is maintained at the Scripps Research Institute and is a database for icosahedral virus capsid structures; The Big Picture Book of Viruses, a large catalog of virus pictures with associated information; Virus World, a summary of pictures and links available from PDB and VIPERdb organized by virus name; and Viral Protein Structure Resource (ViPs), a database that aims to provide a central source for all viral protein structures and provides a genome map feature allowing the user to determine which genes have structural information available. In addition, ViPs statistics summarize this structural coverage as compared between species to give a broad overview of the structural coverage of various viruses.

**Virus-specific Databases**

Some viruses have their own dedicated resources, often required by extremely high data volumes or a requirement for specific feature sets. Several of these will be discussed below; however, as an aside, there are also aspects of the standard databases that provide useful compartmentalization of data to make working with specific viruses easier. For example, search interfaces such as NCBI-BLAST and Advanced Nucleotide Search offer comprehensive tools for filtering search results, including limiting BLAST search results to a particular taxonomic group, for example Poxviridae. Additionally, GenBank offers a virus specific branch of the database that limits searches to within viral families and offers access to analytical tools and viral resources (www.ncbi.nlm.nih.gov/genome/viruses/).

The term “virus database” is widely applied, but as noted above it encompasses a variety of database formats with equally varied data organization and analysis tools dedicated to viruses: the virus-specific databases. Understanding these differences can help categorize the various virus-specific databases, and address which types of information can be drawn from each. The examples discussed below present the varying complexity of these databases.

At the most basic level, a website, can be used to present a collection of links to sequences in files or more traditional databases. Examples include Giantvirus.org, which simply provides a list of the viruses with the 100 largest genomes and their accession numbers. Although these resources usually deal with small data sets, they can help offer a more visually appealing and manageable access point for researchers. Although the same files can be easily stored on a local desktop computer, there is an important benefit to accessing from a website – the data should be the same/unaltered. Since most researchers tend to collect dozens of versions of sequences, often with various edits, accessing data from a website ensures the sequence is what it is supposed to be.

A relational database organizes and stores the data within the program, but is usually accessed remotely using a web-interface or specific program. Regardless of the interface, the database offers the ability to filter, collect and retrieve genomic data sets based upon their shared characteristics, thereby providing greater utility to the user. In addition to complete genomes, these databases often contain the gene and protein sequences with associated search utilities, allowing the user to query different data sets. For example, when NCBI BLAST searches the sequence databases, it is possible to filter the results by keywords and taxonomy categories. This type of filtering is very useful when searching for distantly related sequences within a specific virus family because confounding matches from other organisms are removed. Similarly, UniProt offers equivalent search utilities for filtering protein sequences, including narrowing database searches to particular branches of the database (UniRef, UniProtKB, UniParc, etc.). BLAST and keyword searches.

The design of virus-specific databases offers the ability to customize data sets and search features appropriate to the virus, adding calculated and curated information not present in the genome’s GenBank file. Examples of this genome-related data could include: G+C% content of genes and genomes, codon composition of genes, pl of proteins, predicted MW of proteins and amino acid composition of proteins. Some databases, such as VOCs, cluster orthologous genes within the database. This functions to simplify the retrieval of sets of related genes prior to performing comparative analyses. To illustrate this point, it is much easier to retrieve all DNA polymerase orthologs and select the entire set for alignment than to collect them individually.

The remainder of this section will compare and contrast three types of virus database resources: the Influenza Virus Resource at NCBI; the Virus Pathogen Resource Bioinformatics Resource Center (ViPRbrc) supported by the NIH, which supports a variety of viruses; and the author’s Virology.ca resource, which supports large dsDNA viruses.

The Influenza Virus Resource database at NCBI is part of a larger resource that also provides similar tools for dengue virus, West Nile virus, Middle East respiratory syndrome coronavirus, ebolavirus and rotavirus. The search interface supports the query of sequences (protein, CDS or nucleotide) by influenza type, genome segment, serotype, collection date, host and country of origin.
Additional filters can remove sequences that are not full length (of segment), that are not part of a full genome set, or that are identical to another virus. In addition, pandemic H1N1 sequences can be included or excluded, as can sequences from lab strains, vaccines or certain other projects. Once selected, sequences can be exported to a local computer or aligned with subsequent generation of a phylogenetic tree; the interface is completely web-based. The NCBI site also provides a web interface for automated annotation of influenza genome segments and the retrieval of complete data sets comprising thousands of sequences. Finally, the resource includes a utility to facilitate the submission of new sequences to the GenBank database.

ViPRbrc is funded by NIH to support research on viruses on the NIAID Category A–C Priority Pathogen lists, and those causing (re)emerging infectious diseases. It provides databases and tools for the analysis of ss(+)RNA viruses (Caliciviridae, Coronavirusidae, Flaviviridae, Hepaviridae, Picornaviridae and Togaviridae), ss(−)RNA viruses ( Arenaviridae, Bunyaviridae, Filoviridae, Paramyxoviridae and Rhabdoviridae), dsRNA viruses (Reoviridae) and dsDNA viruses (Herpesviridae and Poxviridae). In addition to the standard genome and gene/protein level information sets, the database also provides a variety of curated information and search tools for 3D protein structure, immune epitopes, protein domains and motifs, host factor experiments and sequence feature variants. Data is sourced from reliable external primary sources including GenBank, UniProt, Immune Epitope Database, and Protein Data Bank. Although the database selection, retrieval and analysis tools are still web-based, they are comprehensive and advanced. Of the analyses provide graphical output (e.g., genome maps), but have limited comparative tools for further analysis. A useful addition to the tools at ViPRbrc is the inclusion of a personalized, private Workbench for storage and sharing of search and analysis results on their server. This allows the user to store a variety of retrieved sequence sets, eliminating the need to repeat the tedious selection process, which in turn likely increases accuracy of the work. Offering an alternative to the Influenza Virus Resource at NCBI, ViPR also provides the Influenza Research Database (IRD). The IRD provides tools to identify the clade of H5N1 genomes and to annotate influenza genome segments. Data can also be exported from the database in a variety of formats, including GFF3, which is a tab-delimited format for describing genomic features.

The primary focus of ViPRbrc—viruses with small genomes (RNA)—is reflected in the analysis tools that are provided. Due to the high conservation of essential gene sets in small RNA viruses, genome analyses and comparisons rely heavily on the creation of MSAs and SNP analysis. To this aim, ViPRbrc tools allow BLAST searches of custom (virus-specific) databases and the creation and visualization of MSAs and phylogenetic trees, and one of the more valuable services provided is the tool for Analysis of Sequence Variation, that is SNPs (also amino acid variation) along sequences of an MSA.

In contrast to ViPRbrc, Virology.ca is dedicated to the support of large dsDNA viruses and as noted previously, this involves the provision of different tools to perform comparative genomics analyses. Although the genomes are 20–30 times the size of the RNA viruses, the current poxvirus database at Virology.ca contains less than 270 genomes, with the most studied virus species having 30–50 representatives. The poxviruses, which will be used to illustrate the functionality of Virology.ca, have between about 130 and 300 genes depending on the genome size. However, only about 40 are present in all poxvirus genomes, but this conserved set of core genes increases to about 80 if poxviruses that infect insects are excluded from the calculation. The remaining non-essential genes are often associated with host range and virulence functions. In the Orthopoxvirus genus, the biggest genomes tend to be found in viruses with the widest host range; it is thought that the restriction of viruses to a limited host range is associated with the loss of genes or gene function. Therefore, in large viruses, researchers are frequently interested in characterizing gene presence rather than comparing SNPs between genes (Table 2).

Virology.ca uses a MySQL database called Viral Orthologous Clusters (VOCs) to store the data imported from GenBank files, which is processed to create a relationship table that describes families of gene orthologs. These sets of gene orthologs are used extensively in the analysis of poxvirus genomes by the Virology.ca tools. In contrast to the other resources discussed, Virology.ca uses Java tools to access the database and provide analysis and visualisation of the data. The Java programs are able to provide more functionality for a variety of comparative analyses than a web-based interface (Table 3). For example, the VOCs GUI allows for simple retrieval of all genes of a poxvirus into an interactive table, which allows selection, display and sorting of

| Module | Description |
|--------|-------------|
| Viral Orthologous Clusters (VOCs) | An easy-to-use Java GUI used to access the VOCs genome database. Starting point for gene analysis pipelines |
| Viral Genome Organizer (VGO) | An easy-to-use Java genome browser for viruses that graphically displays and performs searches within a genome and talks to the VOCs database for display of annotations and genome comparisons |
| Base-By-Base (BBB) | A tool used to align/visualize/edit and search comments-genes-proteins-genomes, and visualize differences between the sequences. The program can also be used to view RNA-Seq data and analyze for recombination. BBB offers its own format of data storage, an XML file (.bbb), that allows for additional features such as the storage of user comments, primer annotations, and genome MSAs with gene annotations |
| JDotter | A program for generating dotplots; suitable for whole genomes, sub-genomes or protein sequences |
| Genome Annotation Transfer Utility (GATU) | A tool used to annotate genomes based on a closely related reference genome. Translations of predicted ORFs from the genome to be annotated are BLASTed against the proteins of the reference genome in order to find corresponding genes. GATU also suggests novel genes to the human annotator, who has last word on the annotation process |
| Sequence Searcher (SSeq): | An easy-to-use Java tool for searching protein and DNA sequences for user-specified sequence motifs. |
different data types (e.g., protein size, pI and amino acid content). These features are helpful when working with the large DNA viruses because of the uncertainty associated with the annotation of various genes – genes that are predicted to encode small proteins with a very high or very low pI and unusual amino acid composition are likely to represent annotation errors. Indeed, many of the features present in Base-By-Base and VGO, the tools that display the genome sequences and genome maps, respectively are devoted to helping solve annotation issues. For the large DNA viruses, accurate annotation is very important because a common investigation asks, *Why is virus X more virulent than virus Y?* The associated database query becomes *Which genes are in virus X, and absent from virus Y?*

As Table 1 indicates, there are a variety of specialized databases/tools devoted to various aspects of virology. So far, we have divided discussion issues by genome type; however, two additional databases should be mentioned. The first is the Los Alamos Human Immunodeficiency Virus (HIV; retrovirus; small RNA genome; [www.hiv.lanl.gov/]) sequence database, which has to manage several unique features, including virus drug resistance data, epidemiological-geographical data, recombination, and clinical data. The database has 3 main branches: 1) the sequence database, which serves a function much like the other resources, but includes clinical information, pre-made genome MSAs, a map-based geographical search tool, a subtyping tool and a series of phylogenetic tools; 2) a unique immunology database for detecting, searching and analyzing variability between CTL and Th epitopes, mapping antibody binding sites and escape mutants; 3) a unique Vaccine Database containing searchable information about previous HIV/SIV vaccine trials. The HIV drug resistance database is based at Stanford ([http://hivdb.stanford.edu](http://hivdb.stanford.edu)) and associates genotype with treatment, resistance and clinical data. A subtyping tool reads genome sequences and attempts to predict drug resistance.

The second database does not deal directly with biological data, but rather the taxonomy and classification of viruses. The International Committee on Taxonomy of Viruses ([ICTV; http://www.ictvonline.org/virustaxonomy.asp](http://www.ictvonline.org/virustaxonomy.asp)) is recognized as the official group for assigning viral taxonomy and nomenclature to the species level and higher. Divisions within species, such as clades, subgroups, strains and isolates are outside the scope of ICTV. This information, defining relationships among viruses, aids in the classification and understanding of newly discovered viruses. The classification scheme of the ICTV is assigned based upon viral morphology, structural characteristics and sequence-based comparisons, which have largely replaced serotyping. Assignments are supported by verifiable data, and the consensus of experts who are organized into a series of subcommittees and associated study groups for each viral species. The committees assign new isolates to existing taxonomic groups, or establish new taxonomic groups to support unique viruses. The current release, the 2014 ICTV taxonomy, comprises 7 orders, 104 families, 23 subfamilies, 14 genera, and 3186 species. However, there are 78 families of viruses that not assigned to an order, including one called *Unassigned*, which contains 14 genera.

### Conclusions and Thoughts About the Future

As genome sequencing prepares to enter its third revolution in as many decades, the mass of accumulated sequence data continues to grow exponentially. More than ever, there is an overwhelming need to organize this giant hairball of nucleotide strings. Sequences need to be organized to facilitate the process of similarity searching, for example with BLAST, for related sequences and also organized by functional or source (human, rodent, virus) relationships. However, it’s neither feasible nor affordable to create independent database resources for every organism, therefore resources like ViPRbrc that can function with a variety of viruses may become more commonplace.

With respect to genome sequencing, one of the greatest problems lies with the huge volume of raw data (sequencing reads) that is associated with any final genome sequence. The authors recently received 3 GB of compressed sequencing data, a mix of host and virus sequences, to assemble a 150 kb poxvirus genome; it is estimated that genomic data will soon become the world’s largest consumer of disk storage. Therefore, one question that is frequently asked is, *when will it be cheaper to recollect the data than store the data?* Although storage of raw sequencing data is a somewhat different problem than the organization of genomes to allow maximum use by researchers, it provides a valuable illustration as to where our capacity for genome sequencing is heading. Perhaps, in the not too distant future, it will not be unusual to have our own genome sequenced, as well as our various organ microbiomes and the genomes of any pathogens we are infected by. To deal with the benefits of the new sequencing technology’s ability to generate massive sequence coverage, which helps to reduce errors in the initial sequencing process and reveals the natural variation among the genomes of a virus population, we must also develop new annotation strategies such as how to annotate SNPs in virus genomes and gene fragments in the non-essential gene sets of large viruses. The goal of this type of annotation is to provide the virologist with more information in comparative studies. Often large viruses with different virulence phenotypes are compared through asking, *what genes are in virus A and absent in virus B?* The questioner is really asking about the presence of functional genes, but would likely be interested in knowing the mechanism of the deletion. For example, whether the gene was inactivated by a series of deletions or a single nucleotide change that could be the result of a sequencing error.

Not only is the number of genomes requiring annotation increasing exponentially, but the nature of the data is also changing with *metagenomics’ ability to determine* all of the nucleotide sequences from a crude environmental sample without the need to culture microbes in the laboratory or probe for a particular gene. Since many organisms fail to grow in the laboratory, we cannot grow viruses that are unique to them; thus, metagenomics provides a valuable window into the previously unknown diversity of viruses in the environment. Even with large-scale environmental sequencing under way, most metagenomic sequences appear to be unrelated to currently known sequences. These unique sequences represent a vast new pool of genes and functions for the research community.
However, metagenomics also comes with its own list of special problems and limitations. Most issues arise in assembly, gene prediction or the integration of metadata to sequence data. Genomic assembly can be difficult because the metagenomic short reads often provide less coverage than traditional sequencing. As well, misassembly of sequences due to repetitive DNA sequences is increased by the large number of different genomes in an environmental sample and the varying relative abundance of individual genomes. Although assembly against a reference genome is helpful, this approach can’t be used with environmental samples full of unknown microbial species. These issues of genome misassembly and lack of reference genomes also complicate gene prediction, which impedes accurate genome annotation. Furthermore, metadata is often left out of annotations, yet it is required for the data to be put into a useful, comparable, biological context. For example, in phylogeographical studies of epidemics and new outbreaks, a growing area of bioinformatic analysis, dates and location of collection are critical values. Thus, it is crucial that annotation of metadata be included as a mandatory and standardized component of genome submission.

With the accumulation of more and more diverse sequences from metagenomic sequencing of environmental samples, the organization of gene/protein sequences into Clusters of Orthologous groups (COGs: http://www.ncbi.nlm.nih.gov/COG/) and Viral COGs (VOGs) becomes especially useful for defining relationships. Even when matches between orthologs may be at the limit of detection, the greater the number of sequences in a cluster, the greater the chance of making new matches by virtue of transitive sequence comparisons, that is SeqA matches SeqB, but not SeqC; if SeqB also matches SeqC then a transitive relationship between SeqA and SeqC is established. NCBI offers protMap, an application that provides visualization of all genomes containing genes within a COG or VOG of interest, highlighting and centering the genomes to display the orthologs. In a 2016 publication, EMBL released eggNOG 4.5 (eggngodb.embl.de/), essentially COGs remade with improved algorithms, with support for more than 2000 organisms, and the addition of a branch containing 352 viral proteomes. Several features of this updated version may help increase the knowledge and use of ortholog groups by the virology community: the new web interface is simplified for less experienced users and offers new browsing and visualization capabilities; the new version has improved consistency and impact through incorporation and linking of Gene Ontology terms, KEGG pathways and UniProt and SMART/Pfam domains to ortholog group assignments; and the algorithms and pipelines for assignment are made freely available on the website, encouraging individuals to apply the techniques to their own datasets.

In recent years, the growth in both the volume and types of data that can be considered bioinformatic in nature has forced the scientific community to consider how much of a role that the manual forms of genome annotation, curation and maintenance can continue to play in assigning knowledge to new genomes. A compromise may be found in the maintenance of manually curated reference genomes, and development of programs to aid in increasing the accuracy of automated annotations of new sequences. Finally, the question remains whether funding will be made available for tailoring databases to the needs of virology researchers working on the wide array of bioinformatics challenges that these new data sets are generating.

Further Reading

An evolutionary view of the arms race between protein kinase R and large DNA viruses
The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection
Systems biology: A tool for charting the antiviral landscape
V-GAP: Viral genome assembly pipeline
ViralEpi v1.0: a high-throughput spectrum of viral epigenomic methylation profiles from diverse diseases
Whole-Genome Sequencing Analysis from the Chikungunya Virus Caribbean Outbreak Reveals Novel Evolutionary Genomic Elements
Automatic segmentation of Potyviridae family polyproteins
Computational approaches to predict bacteriophage-host relationships
CoVaMa: Co-Variation Mapper for disequilibrium analysis of mutant loci in viral populations using next-generation sequence data
Developing expertise in bioinformatics for biomedical research in Africa
EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data
Ebolavirus comparative genomics
eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences
EPIPOX: Immunoinformatic Characterization of the Shared T-Cell Epitome between Variola Virus and Related Pathogenic Orthopoxviruses
FluKB: A Knowledge-Based System for Influenza Vaccine Target Discovery and Analysis of the Immunological Properties of Influenza Viruses
HPMCMD: the database of human microbial communities from metagenomic datasets and microbial reference genomes.
Limits and patterns of cytomegalovirus genomic diversity in humans
Phage Phenomics: Physiological Approaches to Characterize Novel Viral Proteins
Plant Virus Metagenomics: Advances in Virus Discovery
The potential of whole genome NGS for infectious disease diagnosis
Time dependence of evolutionary metrics during the 2009 pandemic influenza virus outbreak
Unexpected features of the dark proteome
Unraveling the Web of Viroinformatics: Computational Tools and Databases in Virus Research 2015 25428870
Vaxvec: The first web-based recombinant vaccine vector database and its data analysis 2015 26403370
Virus evolution and transmission in an ever more connected world 2014 26702033
HIV-1, human interaction database: current status and new features 2014 25378338
Immunoinformatics: a brief review 2014 25048118
NCBI Viral Genomes Resource 2014 25428358
VIRBase: a resource for virus–host ncRNA-associated interactions 2014 25274736
VirHostNet 2.0: surfing on the web of virus/host molecular interactions data 2014 25392406
VirusMentha: a new resource for virus-host protein interactions 2014 25217587
AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses 2013 24285301
Virus Variation Resource—recent updates and future directions 2013 24304891
Bioinformatics for Analysis of Poxvirus Genomes 2012 22688771
HBVdb: a knowledge database for Hepatitis B Virus 2012 23125365
Influenza Research Database: an integrated bioinformatics resource for influenza research and surveillance 2012 22260278
The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata 2012 22135293
The LANL hemorrhagic fever virus database, a new platform for analyzing biothreat viruses 2012 22064861
The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis 2012 23093593
ViPR: an open bioinformatics database and analysis resource for virology research 2012 22006842
ViralZone: recent updates to the virus knowledge resource 2012 23193299
Bioinformatics Resource and Tools for Phage Display 2011 21245805
HCVpro: Hepatitis C virus protein interaction database 2011 21930248
Reactome pathway analysis to enrich biological discovery in proteomics data sets 2011 21751369
Immunoinformatics: an integrated scenario 2010 20722763
Viral Genomics and Bioinformatics 2010 3185590
phiSITE: database of gene regulation in bacteriophages 2009 19900969

Acknowledgements

The authors would like to thank Chad Smithson for help assembling the collection of virus databases. This work was supported by a Natural Sciences and Engineering Research Council of Canada Discovery Grant to CU.

Conflicts of Interest: The authors are the developers of Virology.ca.