Virus Variation Resource—recent updates and future directions

J. Rodney Brister*, Yiming Bao, Sergey A. Zhdanov, Yuri Ostapchuck, Vyacheslav Chetvernin, Boris Kiryutin, Leonid Zaslavsky, Michael Kimelman and Tatiana A. Tatusova

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received November 1, 2013; Revised November 12, 2013; Accepted November 13, 2013

ABSTRACT

Virus Variation (http://www.ncbi.nlm.nih.gov/genomes/VirusVariation/) is a comprehensive, web-based resource designed to support the retrieval and display of large virus sequence datasets. The resource includes a value added database, a specialized search interface and a suite of sequence data displays. Virus-specific sequence annotation and database loading pipelines produce consistent protein and gene annotation and capture sequence descriptors from sequence records then map these metadata to a controlled vocabulary. The database supports a metadata driven, web-based search interface where sequences can be selected using a variety of biological and clinical criteria. Retrieved sequences can then be downloaded in a variety of formats or analyzed using a suite of tools and displays. Over the past 2 years, the pre-existing influenza and Dengue virus resources have been combined into a single construct and West Nile virus added to the resultant resource. A number of improvements were incorporated into the sequence annotation and database loading pipelines, and the virus-specific search interfaces were updated to support more advanced functions. Several new features have also been added to the sequence download options, and a new multiple sequence alignment viewer has been incorporated into the resource tool set. Together these enhancements should support enhanced usability and the inclusion of new viruses in the future.

INTRODUCTION

‘So many sequences and yet, so little metadata’ might as well be the official slogan to the dawn of the sequencing age. Often sequence source descriptors such as host, isolation place and time, and other metadata are missing from International Nucleotide Sequence Database Collaboration (INSDC) (1) sequence records. Though metadata can sometimes be inferred from information found within the sequence record or found in the text of a research article, associating this derived metadata with the original sequence record is difficult in practice. Even when metadata are readily available, without universally accepted standards, varied but synonymous terms can hinder retrieval of relevant sequences from public database searches. Lack of data standardization extends beyond metadata and sequence annotations are often inconsistent, with the same protein annotated in different ways among different sequence records—a major impediment to sequence analysis.

Of course metadata and sequence annotation standards are but the tip of the iceberg. With so many sequences now available in public databases, under the best of circumstances, database queries often produce very large datasets, forcing the user to weed through pages of traditional text based displays. Indeed, one could argue that the explosion of sequence data now threatens to blow up traditional models of data storage, retrieval and display. This realization and the argument that such broad issues require equally broad solutions led to the development of the NCBI Virus Variation Resource (http://www.ncbi.nlm.nih.gov/genomes/VirusVariation/) (2). This comprehensive, value added web resource includes three elements—a specialized database, a unique search interface and a suite of tools and displays—all designed to support large sequence datasets.

VIRUS VARIATION 2.0

The current Virus Variation Resource is an outgrowth of the NCBI Influenza Virus Resource created in 2004 (3) in response to the National Institute of Allergy and

*To whom correspondence should be addressed. Tel: +1 301 594 6099; Fax: +1 301 402 9651; Email: jamesbr@ncbi.nlm.nih.gov

Published by Oxford University Press 2013. This work is written by US Government employees and is in the public domain in the US.
Infectious Diseases (NIAID) Influenza Genome Sequencing Project (4). The resource was initially designed to enhance the usability of very large influenza sequence datasets, and a number of features were introduced to facilitate sequence retrieval. Among these was the development of a metadata driven search interface (Figure 1). Sequence descriptors such as country of isolation, host and protein name are parsed from GenBank (5) records during database loading using advanced strategies. These machine processes are augmented with human curation allowing data found in publications and other sources to be associated with sequences in the database. The resultant metadata are mapped to controlled vocabulary lists, and consistent terms are stored in the database, providing a single term for synonymous and misspelled ones. These metadata terms are then displayed among several menus providing users with a straightforward but comprehensive search interface.

![Influenza virus database search interface](image)

**Figure 1.** Influenza virus database search interface. The search interface is shown with the ‘Additional filters’ selection open. A number of search criteria have been selected, and two separate searches added to the ‘Query builder’. Selected search results are indicated by the check box to the left of the ‘Query builder’ display and can be downloaded directly in a variety of formats or loaded into the Virus Variation search results interface by depressing the ‘Show results’ button.
through which users can retrieve nucleotide and protein sequences based on a number of biological and clinical criteria.

The number of sequences in the database has grown substantially as influenza continues to be a major human pathogen and as surveillance networks and virus sequencing efforts are maintained around the world (6,7). There are now more than 292,000 individual influenza nucleotide sequences in the database, including more than 17,100 complete genome sets. The value added influenza data model was first extended to a functionalities while leveraging shared backend support for influenza coding regions and other sequence features can be systematically annotated prior to INSDC database submission using the Flu Annotation Pipeline (FLAN) (15). This pipeline is publicly available from the Virus Variation web pages and first types (or genotypes) sequences by BLAST alignment to a set of virus-specific nucleotide references and then annotates protein coding regions using reference protein sequence sets specific to each virus subtype (15). Specifically, FLAN maintains a set of reference nucleotide sequences that are used to classify input influenza sequences by type (A, B or C), identify specific segments (1 through 8) and—when applicable—subtype influenza A hemagglutinin and neuraminidase segments (reference sequences available at ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/ANNOTATION/blastDB.fasta). Corresponding reference protein sets are then aligned to translated input sequences and protein coding regions predicted using the ‘Protein to nucleotide alignment tool’ (ProSplign) (15). The FLAN is continually being updated to support community needs. For example, the Influenza virus annotation tool now supports Influenza C sequences in addition to A and B, and can predict the recently discovered PA-X protein coding sequences.

The annotation pipelines for DENV and WNV are integrated into the database loading pipeline and are very similar to FLAN. The reference nucleotide records and corresponding protein sets used to annotate the two viruses are listed in Table 1. Currently, protein coding regions are extracted from INSDC records and mature peptides annotated by the pipeline and stored in the database. However, this dependency on submitted protein annotations has several shortcomings, not least of which is the inability to update protein annotations in response to evolving biological knowledge, and we are in the process of moving to a fully de novo annotation model. In the new model, all features will be annotated directly by
internal pipelines using an improved version of the NCBI ‘Protein to nucleotide alignment tool’ (ProSplign 2). Annotations will also be updated on a regular basis and consistent annotation maintained irrespective of sequence submission dates or changing annotation standards.

NEW FEATURES

A number of new features have been added to Virus Variation since the last published description of the resource. The resource web pages have been updated, including the database search interface (Figure 1). This interface now supports searches using multiple GenBank accessions as well as keyword searches for sequence patterns, strain names/definition lines and influenza drug resistance mutations. Search menus have been updated and support multiple selections, so several proteins, hosts or geographic locations can be added to a single set of search criteria. In the influenza query page there is now the option to select sequences from northern temperate, southern temperate and tropical regions in addition to the country and continent selections used throughout the resource. Searches can also be limited by both collection date and GenBank release date including year, month and day.

Several sets of virus specific filters have been added to the search interfaces to enhance usability. The ‘Full-length genomes only’ filter used in the DENV virus and WNV search interfaces limits retrieved mature protein sequences to those that are part of a complete polyprotein coding sequence (all mature proteins). On the influenza page the ‘Full-length only’ filter limits searches to protein or nucleotide sequences that include a complete coding region, from start codon to stop. A second, ‘Full-length plus’ filter restricts the search to both full-length protein or nucleotide sequences and nearly complete sequences missing only the start and/or stop codons. Complete, nearly complete and partial sequences are marked in search results. A set of ‘Additional filters’ have been added to the influenza query page, and users can now limit searches to those sequences that have a specified day and/or month in the collection date field. Users can also ‘Include’, ‘Exclude’ or ‘Only’ retrieve sequences from WHO recommended ‘Vaccine strains’, pandemic (H1N1) 2009 viruses, sequence sets with ‘Mixed subtypes’, ‘Lineage defining strains’ of well-defined lineages/clades. Currently, virus prototypes include those for the Victoria and Yamagata

![Figure 2](image-url)
lineages of influenza B viruses, and the H5N1 and H9N2 subtypes of influenza A viruses. The ‘Required segments’ filter limits retrieved sequences to those where all the selected segments of the same virus isolate exist in the database.

The Virus Variation search interface allows the user to build complicated datasets containing sequences retrieved using different criteria. To do this, the results from each individual database search are added to the ‘Query builder’ section at the bottom of the search interface (Figure 1), then one or more search sets selected for display on the Virus Variation search result page or direct download. The search result page displays sequences retrieved from search sets along with several sortable metadata columns and supports selection of individual sequences for download or further analysis (Figure 2). Identical sequences can be collapsed in the search results and represented by the oldest sequence in the group. Results can be downloaded as a table in XML, CSV or tab-delimited formats, or users can also download a GenBank accession list or FASTA file of selected sequences. The definition line of FASTA sequences can now be customized in the downloaded files, and users can replace original GenBank definition lines with a number of fields including host, country, date, serotype, patient age or gender, viral mutations and CDS location.

The resource sequence analysis tool set has been improved to enhance visualization of large datasets and facilitate discovery activities. A new multiple sequence alignment viewer (Figure 3) has been integrated into DENV virus and WNV resources and will soon be available for influenza virus. This tool is based on the NCBI Genome Workbench multiple sequence alignment viewer and includes a variation histogram above the alignment as well as a feature table that highlights mature protein boundaries and other important sequence features. There are a number of usability features integrated into the viewer such as selectable alignment scoring methods for individual nucleotides/amino acid residues, link outs to associated GenBank records and selectable alignment...
anchor sequence—either consensus or any sequence in the alignment. Alignments displayed in the viewer can also be downloaded in FASTA, Clustal, Phylip and Nexus formats for use locally or with other tools. The Virus Variation tree builder tool (16) has also been updated for all viruses, and GenBank accession numbers can be downloaded through the tree builder tool by selecting the branch of interest on the tree.

**FUTURE DIRECTIONS**

The long term plan is to increase the coverage of virus sequences in the Virus Variation Resource. The flexibility of the resource should support a number of diverse viral pathogens and provide consistently annotated sequence datasets with standardized isolate descriptors. This will require continued tweaking of metadata parsing strategies and development of new virus-specific sequence annotation modules. As these annotation modules are added to our core pipeline for use by the resource, they will be made publicly available. We will also explore approaches to increase user outreach and leverage community knowledge to improve data curation, reference sequence assignment and resource usability.

**FUNDING**

This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. Funding for open access charge: Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Conflict of interest statement. None declared.

**REFERENCES**

1. Nakamura, Y., Cochrane, G. and Karsch-Mizrachi, I. (2013) The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res., 41, D21–D24.

2. Resch, W., Zaslavsky, L., Kiryutin, B., Rozanov, M., Bao, Y. and Tatusova, T.A. (2009) Virus variation resources at the National Center for Biotechnology Information: dengue virus. BMC Microbiol., 9, 65.

3. Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J. and Lipman, D. (2008) The influenza virus resource at the National Center for Biotechnology Information. J. Virol., 82, 596–601.

4. Fauci, A.S. (2005) Race against time. Nature, 435, 423–424.

5. Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. (2013) GenBank. Nucleic Acids Res., 41, D36–D42.

6. Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J. and Lipman, D. (2008) The influenza virus resource at the National Center for Biotechnology Information. J. Virol., 82, 596–601.

7. Zaslavsky, L., Bao, Y. and Tatusova, T.A. (2008) Visualization of large influenza virus sequence datasets using adaptively aggregated trees with sampling-based subscale representation. BMC Bioinformatics, 9, 237.