doi: 10.1093/nar/gkw1065

Virus Variation Resource – improved response to emergent viral outbreaks

Eneida L. Hatcher, Sergey A. Zhdanov, Yiming Bao, Olga Blinkova, Eric P. Nawrocki, Yuri Ostapchuck, Alejandro A. Schäffer and J. Rodney Brister[*]

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

Received September 15, 2016; Revised October 20, 2016; Editorial Decision October 21, 2016; Accepted: October 28, 2016

ABSTRACT

The Virus Variation Resource is a value-added viral sequence data resource hosted by the National Center for Biotechnology Information. The resource is located at http://www.ncbi.nlm.nih.gov/genome/viruses/variation/ and includes modules for seven viral groups: influenza virus, *Dengue virus*, *West Nile virus*, *Ebolavirus*, MERS coronavirus, *Rotavirus A* and *Zika virus*. Each module is supported by pipelines that scan newly released GenBank records, annotate genes and proteins and parse sample descriptors and then map them to controlled vocabulary. These processes in turn support a purpose-built search interface where users can select sequences based on standardized gene, protein and metadata terms. Once sequences are selected, a suite of tools for downloading data, multi-sequence alignment and tree building supports a variety of user directed activities. This manuscript describes a series of features and functionalities recently added to the Virus Variation Resource.

INTRODUCTION

Genome sequences have the potential to define evolutionary relationships, elucidate disease determinants and inform public health policy decisions. The public databases that comprise the International Nucleotide Sequence Database Consortium (INSDC) are an invaluable resource to a variety of genome-related sequence analysis projects (1). This collaboration between the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute and the DNA Databank of Japan supports free and unrestricted access to stored sequence data that are maintained as part of the scientific record. As nucleotide sequencing efforts extend into the future, the archival INSDC databases will support comparisons between samples collected over generations and provide infrastructure to study the evolution and impact of viruses in real time. Despite this potential, there are fundamental issues with archival databases that can only be resolved through resources that provide enhanced data such as the NCBI Virus Variation Resource (http://www.ncbi.nlm.nih.gov/genome/viruses/variation/), which is described in this manuscript.

GenBank records (2) and other INSDC sequence records are archival by design, and changes to them can be made only by one of the original submitters. Hence, it is likely that the gene and protein annotations and information about the source of the sequence will remain unchanged after a sequence is deposited in an INSDC database. This is problematic because even if communities develop sequence annotation standards, the pace of biochemical and genetic research effectively guarantees that annotations become outdated as new genetic features are characterized and naming conventions change. For example, while it has been known for some time that flavivirus genomes encode a polyprotein that is cleaved into mature peptides, sometimes with two rounds of cleavage (3–6), recently, several flavivirus proteins have been identified that are translated (at least partially) from alternative reading frames (7). These alternative reading frame proteins and mature peptides, especially the products of the second round of cleavage, are not annotated in the vast majority of current GenBank records for flavivirus genomes.

The limitations of an archival database can be illustrated by considering a common way in which it might be used – to obtain all of the nucleotide sequences that encode a particular gene of interest. Take, for example, the RNA-dependent RNA polymerase (RdRp) of the Ebolavirus. One would need to know that this gene is also sometimes called L-protein or L-polymerase and search the database with all three names to find all relevant protein sequences. In addition, not all genes or proteins are annotated in all database entries, so one would still likely miss some potential sequences. Alternatively, a nucleotide BLAST search could be performed using the RdRp coding region from the Zaire ebolavirus Reference Sequence (RefSeq accession number NC_002549.1). However, when matching sequences are obtained, there would still be no indication of potential prob-

[*] 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 on behalf of Nucleic Acids Research 2016.

lems with the sequences, such as frameshifts, which may affect the biological function of the resulting protein. Even when an annotation pipeline is available to validate retrieved sequences, several additional steps would be needed to associate metadata, such as country of isolation or host, to the sequences.

Issues regarding the long term usability of sequence data were addressed in the NCBI Influenza Virus Resource (8). This resource leveraged machine processing of GenBank records, human curation and a unique search and retrieval interface to build a value-added user experience where researchers could search for sequences using defined, standardized terms (Table 1). An annotation pipeline was added later to standardize gene and protein annotation and nomenclature across all sequences. This feature supports not only standardized annotation of sequences when submitted, but also provides a mechanism to update previously submitted sequences as new genes and proteins are described. In many ways, the NCBI Influenza Virus Resource paved a path for a variety of other resources that share the common goal of making viral sequence data more accessible (9–12). These include the NCBI Virus Variation Resource where the Influenza Virus Resource data model was extended to include dengue and West Nile viruses (13,14). While the initial release of this resource provided a range of functionalities, the necessity of in-house annotation pipelines and internally developed tools imposed long development cycles making it difficult to quickly provide new modules in response to emerging outbreaks and associated nucleotide sequencing efforts.

Here, we document a series of updates and improvements designed to make viral sequences more easily accessible and usable through the Virus Variation Resource, a value-added database, as well as tools that make it simple to analyze genomic relationships. The resource now includes expanded data processing pipelines and analysis tools, and supports selection and retrieval of nucleotide and protein sequences from four new viral groups: Ebolaviruses, MERS coronavirus, rotavirus, and Zika virus (Table 2). The latest package of updates includes a variety of features designed to improve data usability and ease data retrieval. New processes have been added to parse source descriptor terms from GenBank records and map these to controlled vocabulary, and the resource now supports retrieval of sequences based on standardized isolation source and host terms in addition to standardized gene and protein names. A new set of filters has also been developed to identify laboratory isolates, vaccine strains or environmental samples so that they can be included or excluded from searches. A variety of updates have been made to the search interface and results table to better leverage these features, and a new set of multi-sequence alignment and tree building tools has been implemented to allow robust analysis of retrieved sequences.

The Virus Variation model

The NCBI Virus Variation Resource provides users with a convenient way in which to search, download, and analyze viral nucleotide and protein sequences. The resource includes data processing pipelines that retrieve sequences from GenBank, provide standardized gene and protein annotation, and map sequence source descriptors (i.e. metadata) to uniform vocabularies. This data processing enables users to select sequences based on standardized gene, protein and metadata terms using a purposely-designed interface. Once selected, sequences can then be downloaded with the standardized metadata in a variety of formats or analyzed using web-based alignment and tree building tools. There are currently seven discrete Virus Variation modules—*Dengue virus*, *Ebolavirus*, influenza virus, MERS coronavirus, *Rotavirus A*, *West Nile virus,* and *Zika virus*—and these include a total of nearly 550 000 nucleotide sequences (see Table 2). Example usages of the resources for dengue virus, Ebolavirus, and rotavirus are Klema *et al.* (15), Bell *et al.* (16), Agbemabiese *et al.* (17), respectively.

Rapid deployment model

Current development efforts have focused on expanding the Virus Variation model to include more viruses, enhancing the functionality of the resource and providing rapid support to emergent sequencing efforts. This last point has been particularly relevant over the past several years as emerging viral outbreaks of Ebola and Zika viruses and others have quickly led to large sequencing efforts. There was a clear need to support these sequencing efforts with bioinformatics resources, but timelines prevented traditional development paths where new virus modules and features were added over the course of months. The first rapid deployment of a Virus Variation module was during the western African Ebola virus outbreak that began in December of 2013. The outbreak was declared a Public Health Emergency of International Concern by the World Health Organization on August 8, 2014 (http://www.who.int/mediacentre/news/statements/2014/ebola-20140808/en/). By September, a Virus Variation Resource specific to Ebolaviruses was available to help access the sequences that had begun to pour into the INSDC databases. Similarly, a Virus Variation Resource module was developed in September 2014 in response to the outbreak of Middle East respiratory syndrome-related coronavirus (MERS-CoV). Most recently, this rapid response model was repeated for the Zika virus module, which was put in place in March 2016. This need-based deployment strategy is likely a model for future efforts, and much of our current development is geared toward harmonizing processes and interfaces among individual data and software modules so as to provide more support for more virus species within the resource and to respond more efficiently to emergent large-scale sequencing efforts.

Sequence annotation

Accurate gene and protein annotation is necessary both to identify sequences of interest and to analyze them. The Virus Variation Resource employs annotation pipelines that support consistent gene and protein naming. Initial processing for each annotation pipeline is the same: Newly released GenBank records are retrieved hourly based on their listed taxonomy. Retrieved sequences are compared to nucleotide references for that virus group using BLASTN, and the best match is determined (8,13,18). This step confirms species

Table 1. Summary of data enhancements in the Virus Variation Resource

INSDC/GenBank	Virus Variation Resource
Inconsistent and/or out of date gene/protein names present in INSDC sequence records	Gene and protein sequences are validated and given consistent, up to date names
Annotation is often incomplete in INSDC sequence records, especially for mature peptides	All proteins and mature peptides annotated and possible sequence errors reported
Non-standardized source descriptor (metadata) vocabulary and formatting within INSDC sequence records	Source descriptors are parsed from several fields within INSDC sequence records and mapped to standardized terms with correct spelling
Source metadata potentially missing from INSDC sequence records	Source metadata can be added manually from literature
Drug resistance and/or high virulence sequence polymorphisms may not be annotated in INSDC Influenza virus sequence records	Documented drug resistance and high virulence sequence variations are detected and can be retrieved
Sequence searches based on metadata terms and gene/protein names can be difficult	Complex searches can be performed through a convenient user interface
Once sequences are retrieved, users must perform some data analysis locally or on a third party site	Selected sequences can be aligned or visualized as a tree within the resource
Download formats for sequences and metadata limited for some uses	Sequences can be downloaded in a variety of formats with customized metadata fields

Table 2. Publically available sequence content of Virus Variation Resource (as of September 1, 2016)

Virus module	Species/Types included	Nucleotide seq.	Complete genomes	Protein seq.
Dengue virus	*Dengue virus* types 1, 2, 3 and 4	18 495	4140	17 635
Ebolavirus	*Zaire ebolavirus, Bundibugyo ebolavirus, Sudan ebolavirus, Reston ebolavirus; Tai Forest ebolavirus*	1849	1318	14 407
Influenza virus	*Influenza A virus, Influenza B virus, Influenza C virus*	471 603	33 717	624 541
MERS coronavirus	*Middle East respiratory syndrome-related coronavirus*	730	320	3269
Rotavirus	*Rotavirus A*	49 186	1169	49 607
West Nile virus	*West Nile virus* genotypes 1 and 2; *Kunjin virus*	4184	1675	3678
Zika virus	*Zika virus*	386	111	345

taxonomy, identifies segment assignment if applicable and provides information about the lineage, genotype, type or subtype. The references used are listed in Table 3, and sequences that fail to match a reference within established metrics are pushed to a curation interface where they can be reviewed manually.

Once a sequence has been matched to a reference, one of three pipelines is employed to determine the span of gene and protein features and to assign standardized names to these features. The first pipeline uses a reference protein guided approach based on the Prosplign tool as described previously (8,13,18). Here, protein reference sequences are aligned with potential translations of the query sequence. The highest scoring translation alignment to any protein reference is then chosen and parsed to determine that it meets specific criteria – the presence of a start codon, exact matches to mature peptide cleavage sites or premature stop sites. Post transcriptional and translational exceptions can be accounted for by this tool by adjusting parameters and allowing multiple transitions from different open reading frames to be assembled into a single alignment. One advantage of this approach is that new viruses can be incorporated by adding new reference protein sequences and adjusting the criteria used for validating a particular translation. Such was the case for Zika virus annotation where the existing dengue virus pipeline was updated with new Zika virus reference sequences (see Table 3).

A second approach to gene and protein annotation was implemented in the Ebola virus and MERS coronavirus rapid deployment modules. Here, there was a need to quickly develop a pipeline that could validate the annotation on GenBank records and assign consistent gene and protein names so that these could be accurately used as search criteria. To accomplish this, a BLAST-based pipeline was developed that compares genes and proteins as annotated on GenBank records to reference proteins derived from the best reference nucleotide match. If a protein matches the reference sequence with >70% identity as measured by BLASTP then the presence of this protein is stored. Genes are validated in the same manner using BLASTN and reference nucleotide sequences. Sequences with genes and proteins that cannot be validated are pushed to the curation interface where they can be manually examined. Ultimately these approaches support both search and analysis functionality but are not capable of generating standardized annotation across all sequences belonging to a particular virus.

Our experience has emphasized the importance of accurate annotation pipelines that can be applied to new viruses rapidly in response to emergent needs. Though our current pipelines are effective, they are also very specific to particular viruses and application to new viruses requires much work developing reference sequences, defining processing parameters and manually reviewing annotation results. With that in mind we are now implementing a new, third approach to annotation that can be adapted rapidly when needed and is scalable to multiple virus groups. This new approach is built around two important considerations.

Table 3. Reference sequences employed by Virus Variation

Virus module	Reference sequences
Dengue virus	NC_001477, NC_001474, NC_001475, NC_002640
Ebolavirus	NC_014372, NC_014373, NC_004161, NC_006432, NC_002549
Influenza virus	References are created by Virus Variation staff as needed, and a comprehensive list is maintained here: ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/ANNOTATION/
MERS coronavirus	NC_019843
Rotavirus	References are selected and maintained by the Rotavirus Classification Working Group (27,28) and updates can be found here: https://rega.kuleuven.be/cev/viralmetagenomics/virus-classification/newgenotypes
West Nile virus	NC_009942, NC_001563
Zika virus	NC_012532

First, it uses annotations contained within the so-call Reference Sequence records (19) that are created by our group to represent important taxonomic and sequence space groups. The nucleotide and protein sequences within these records can be invaluable for the unambiguous assignment of sequences to defined groups and can also serve as repositories of reference sequence feature annotation maintained by in-house curation efforts often in collaboration with other scientists (20–24). Second, this approach includes a comprehensive list of error flags that provide extensive information about sequences and can provide warnings about potential problems. This error coding not only allows staff to quickly sort through thousands of annotations during the development of new pipelines, but also provides potential criteria for the selection or filtering of sequences to resource users.

This new approach was used to annotate polyprotein and mature peptide genomic intervals in West Nile virus (WNV), and this annotation will be available soon through the Virus Variation Resource. These annotations were calculated as follows: First, GenBank West Nile sequences were classified as one of the two common lineages of WNV (lineage 1 or lineage 2) using a combination of BLASTN (25) against the two RefSeq sequences and expert knowledge. The principal characteristic that distinguishes lineage 1 from lineage 2 is that the additional protein WARF4 occurs only in lineage 1 WNV genomes and is believed to occur in most of them (7). There is some evidence that a small proportion of WNV genomes do not fit neatly into lineage 1 or lineage 2 (7), but these were classified as lineage 2 in our annotations. Second, the annotation pipeline built a covariance model (CM) for each of 16 mature peptides present in the NC_009942 RefSeq annotation and for the 15 mature peptides in the NC_001563 RefSeq. The CMs are built using the cmbuild program of the Infernal homology search software package (26). Infernal is typically used for modeling the sequence and secondary structure of RNAs, and because the sequences we are modeling lack structure (i.e. basepairs between positions), the CMs we created are effectively identical to sequence-only profile hidden Markov models. In the current version of our pipeline, each model was derived from the single RefSeq nucleotide sequence encoding each mature peptide. Third, the CMs built from the RefSeq to which that genome was assigned were used to predict each mature peptide coding sequence using Infernal's cmscan program.

The annotation software then runs a variety of validation checks and produces error codes that assist in curation of sequences. For example, the pipeline checks for the existence of any in-frame stop codons within the predicted regions. If one or more is found, the prediction boundaries are modified to terminate at the 5′-most stop found. Coding sequence (CDS) coordinates are determined implicitly based on the predicted mature peptide coordinates. Lineage 1 (NC_009942) has three CDSs and lineage 2 (NC_001563) has two CDSs. For each CDS, the predictions for the corresponding mature peptides that make up each CDS are tested for consistency by ensuring that mature peptide coding sequences that are adjacent (separated by 0 nucleotides) in the RefSeq are also adjacent in the predictions. The start position of the first mature peptide and end position of the final mature peptide that comprise each CDS are then used as the start and stop position for that CDS. CDS annotations are not made if the mature peptide consistency check fails. In addition to checking for early stop codons and the adjacency of mature peptide coding sequences, the annotation pipeline identifies other unusual or unexpected features in each sequence and reports those as 'error codes'. There are 17 possible error codes, which provide an easy way for users to gauge the quality of each sequence and its annotations, and should facilitate the selection of subsets of the sequence data that meet specific user-defined quality standards. A more detailed description of the new annotation pipeline and error flags will be included in full detail eventually in a separate manuscript, as well as in the help documents available at the Virus Variation Resource.

Source metadata processing

Another important aspect of sequence analysis is to place a given sequence within biological, temporal and geospatial contexts. Such associations can provide profound health policy and scientific insights, but unfortunately, descriptors that provide information about the source of nucleotide sequences are notoriously inconsistent. To resolve this issue, the Virus Variation database loading pipeline parses GenBank records, identifies important metadata terms, such as sample isolation host, date, country and source, and maps these to a standardized vocabulary using a hierarchical approach. For example, isolation host terms are first identified in the host field and failing that, then isolate or strain fields, then isolation source, note and finally organism name.

This vocabulary mapping strategy follows the INSDC practice of separating isolation host from source. In this convention host refers to an organism—and hence has an organism's name that can be mapped to the NCBI taxonomy tree—and isolation source refers to a physical, environmental or local geographic location (1). For human pathogens isolation source often refers to a host tissue or bodily fluid, and the Virus Variation vocabulary mapping strategy attempts to combine similar clinical terms into biologically relevant groups. For example, the parsed terms 'serum,' 'plasma' and 'lymphocytes' are all mapped to the standardized vocabulary term 'blood'. To support more efficient data retrieval, host terms are mapped in a hierarchy, and once a species term such as '*Accipiter cooperii*' is identified, it is mapped to both the group name 'Bird' and the common name 'Accipiter.'

Other metadata terms such as those for disease associations and clinical/laboratory manipulations are more difficult to parse. To this end, laboratory isolates, vaccine strains and environmental samples are identified by searching for key terms, such as 'tissue culture' or 'sewage,' from all fields. Disease terms for dengue virus are also found using a similar strategy. In all cases these strategies require extensive examination of sequence records and documentation of specific terms that can be accurately mapped to controlled vocabulary gleaned from established ontologies such as the Environmental Ontology (https://bioportal.bioontology.org/ontologies/ENVO) and the Infectious Disease Ontology (https://bioportal.bioontology.org/ontologies/IDO). This process is supported by a curation interface that lists records where parsing fails to identify expected terms, leading to good old-fashioned manual curation and the identification of new terms, common misspellings, regional spelling differences and the manual incorporation of metadata from relevant literature into the Virus Variation database. In total, these vocabulary remapping strategies can have a profound impact on data usability as large numbers of parsed terms can be mapped to controlled vocabularies (Table 4).

Search interface

The Virus Variation annotation and metadata mapping pipelines create standardized terms that can then be leveraged by the resource search interface. A link to this interface can be found on the home page of each virus module, which also includes links to help documents, other NCBI resources, and relevant external resources (for an example, please see http://www.ncbi.nlm.nih.gov/genome/viruses/variation/dengue/). To access the search interface from the module home page, select the link to 'Search nucleotide and protein sequences.' Here, users can select between protein and nucleotide searches (see Figure 1). When searching protein sequences, selecting 'Full-length sequences only' filter, limits retrieved sequences to those with a complete coding region as determined to the relevant reference. The same filter limits nucleotide searches to full-length genomes, where the completeness of a given genome is operationally determined by comparing the genes/proteins present on a given sequence to those on the relevant, full-length reference genome. Currently, non-coding, terminal regions are not included in this determination.

During both protein and nucleotide searches, users can define explicitly the genomic regions present on retrieved sequences using drop-down menus that support multiple selections. Additionally, sequences can be filtered using standardized source metadata terms for host, region/country and isolation source using similar pull down menus. The host and country menus are arranged so that aggregate terms are listed in the top portion of the menu and more discrete terms below. In addition to these common filters, there are module-specific filters for species, types, and disease for Ebolaviruses and dengue virus respectively. The influenza virus module also provides some module-specific search options. For example, a user can select 'Full length only' to include sequences with complete coding regions or 'Full length plus' to include sequences with complete coding regions, but no start and/or stop codon. Several other specific filters are also available on the influenza module search interface, such as H and N subtypes, minimum or maximum sequence lengths, and inclusion or exclusion of pandemic H1N1 viruses.

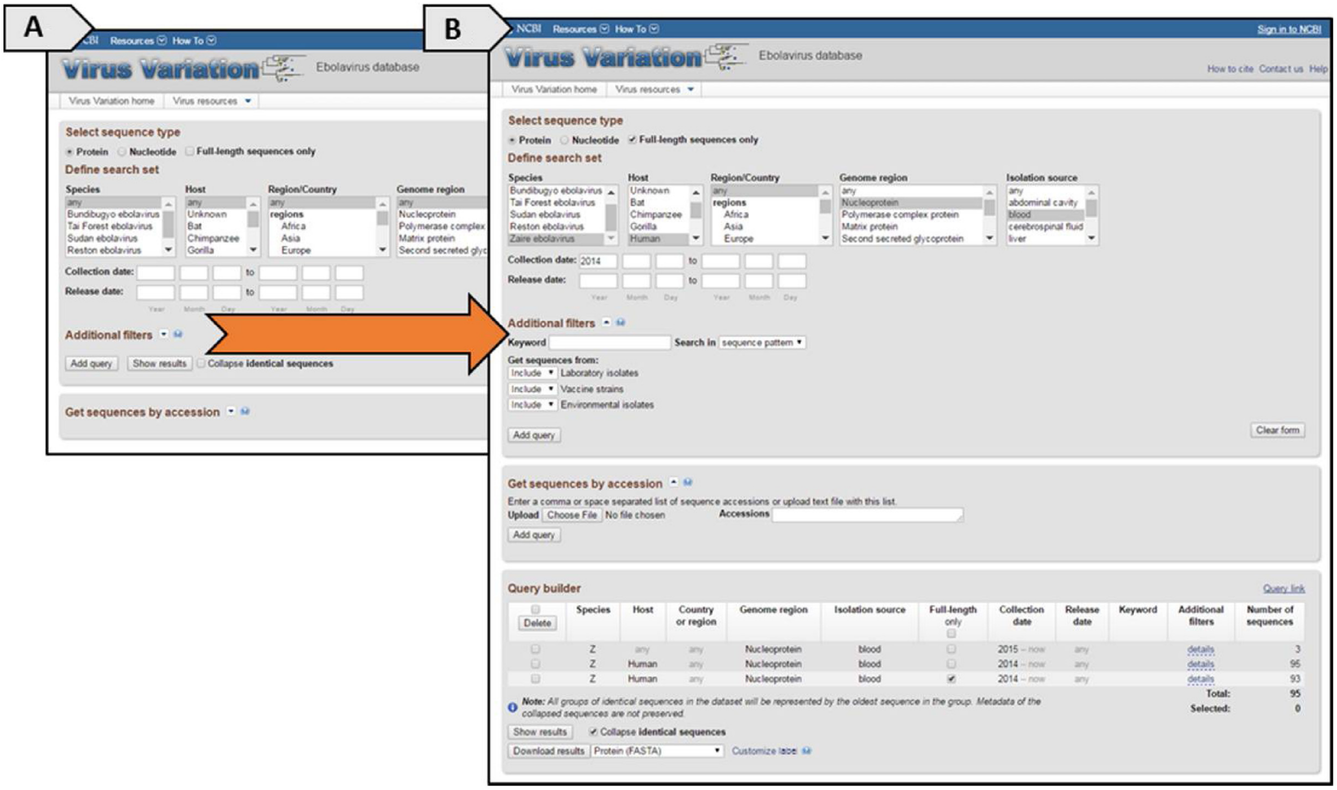
A second set of functions and filters is included within the 'Additional filters' menu. Here users can search for keywords in the GenBank record deflines or strings within sequences. There are also filters to include or exclude laboratory isolates, vaccine strains, and environmental isolates. One can also select specific rotavirus segment types based on assignment by the Rotavirus Classification Working Group (27,28), or by selecting specific sequences by GenBank accession. Once the parameters for a specific search are selected, a user can choose to add the query to the query builder and define another search, or they can go directly to the results. Several searches can be run and added to the query builder where the combination of filters and number of retrieved sequences is displayed for each search. The number of unique sequences can be displayed using the 'collapse identical sequences' checkbox. Individual searches can then be selected and/or combined and sent to the results page for further refinement and analysis.

Results page

The results page supports selection of sequences from the search set for analysis or download. Search parameters are displayed at the top of the results page, and a table displays retrieved sequences and associated metadata. The individual columns within the table can be selected to display specific sets of metadata and hyperlinked GenBank and BioSample accessions (29). BioSample records store an extended set of sample descriptors and are linked to Sequence Read Archive (SRA) (30) records, allowing users to easily find sequence read data associated with retrieved GenBank sequences when available. One new feature is the ability to collapse identical retrieved sequences for all viruses as described in the preceding section. When identical sequences are collapsed on the query page, they will be represented by a single sequence on the results page with the number of collapsed sequences shown in the 'Identical sequences' column (see Figure 2). Clicking the arrows in the 'Identical sequences' column displays the individual sequences and makes them selectable. Users can now customize sequence titles including the FASTA defline of downloaded sequences and tree labels using the 'Customize label' tool. The defline can be modified to include various types of data such as the sequence accession number, calculated genomic

Table 4. Number of GenBank sequences where non-standard metadata terms were mapped to standardized vocabulary

Virus module	Total sequences processed	Isolation country	Isolation host	Isolation source
Dengue virus	18 909	1321	6361	7402
Ebolavirus	1849	598	56	588
Influenza virus	472 050	267 955	380 384	n.a.
MERS coronavirus	730	5	95	327
Rotavirus	49 186	15 823	17 166	19 009
West Nile virus	4184	2143	1253	1329
Zika virus	386	86	127	148



Select sequence type: ● Protein ○ Nucleotide ☑ Full-length sequences only

Define search set: Species: Zaire ebolavirus; Host: Human; Region/Country: any; Genome region: Nucleoprotein; Isolation source: blood

Collection date: 2014 to

Release date: to

Additional filters: Keyword; Search in: sequence pattern

Get sequences from: Include Laboratory isolates; Include Vaccine strains; Include Environmental isolates

Get sequences by accession: Enter a comma or space separated list of sequence accessions or upload text file with this list. Upload: Choose File No file chosen; Accessions

Query builder

□	Species	Host	Country or region	Genome region	Isolation source	Full-length only □	Collection date	Release date	Keyword	Additional filters	Number of sequences
□	Z	any	any	Nucleoprotein	blood	□	2015 – now	any		details	3
□	Z	Human	any	Nucleoprotein	blood	□	2014 – now	any		details	95
□	Z	Human	any	Nucleoprotein	blood	☑	2014 – now	any		details	93
										Total:	95
										Selected:	0

Note: All groups of identical sequences in the dataset will be represented by the oldest sequence in the group. Metadata of the collapsed sequences are not preserved.

☑ Collapse identical sequences

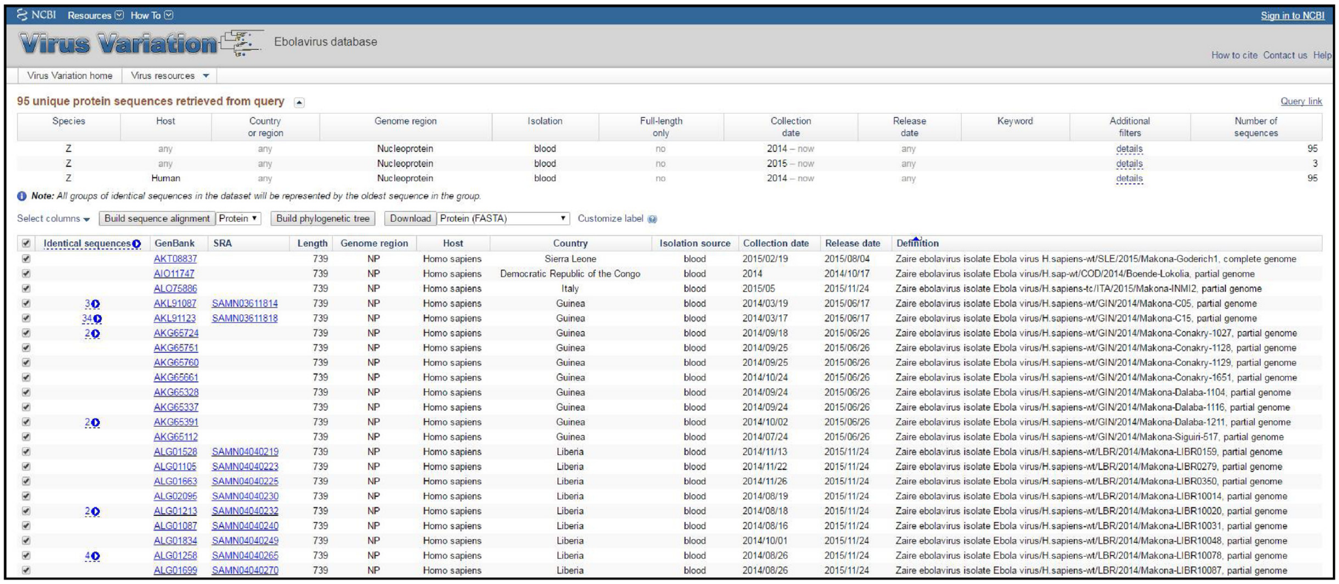
Download results: Protein (FASTA)

Figure 1. Virus Variation Resource search interface page. (**A**) The Ebolavirus module search interface prior to selection of filters and hidden elements. (**B**) The Ebolavirus module search interface with all elements opened and several example searches displayed in the query builder. The search page is divided into three elements. The first element supports selection of protein or nucleotide sequences based on standardized metadata terms generated by processing pipelines described in the text. Menus support filtering of sequences based on gene or protein names, host, isolation country and isolation source, and collection and release dates ranges can be set with text boxes. Additional filters are accessible with a drop-down arrow revealing options for environmental or laboratory isolates, vaccine strains, keyword or sequence string searches, and optional menus tailored to specific viruses. The second element supports searches based on GenBank accessions – either using the text box or by uploading a text file of accessions. The third element includes the query builder where the number of sequences retrieved from individual searches can be viewed by clicking one of the 'Add query' buttons. When multiple searches are added to the Query Builder, the total number of unique sequence records is also summed. A checkbox is provided that allows identical sequences to be collapsed and represented by the oldest sequence on the results table. Clicking the 'Show results' button opens a separate browser tab and displays all of the sequences meeting the criteria in each of the checked queries in the results interface.

region, host, isolation source, collection date or country, as well as field-separators such as pipes or slashes. User-selected titles will also be displayed in multi-sequence alignments and trees as described in the following section.

Analysis tools

Users can build multiple sequence alignments or trees from selected sequences, and these in turn can be downloaded in various formats. The influenza module uses previously described tools for these functions (8,13,31), but a new set of tools has been developed for other viruses. Multiple sequence alignments are constructed using an optimized version of MUSCLE, and rooted trees are generated using the Unweighted Pair Group Method with six base nucleotide or amino acid k-mers (32) (see Figure 3). The multiple sequence alignment display includes a navigation map above the alignment, a variation histogram and a consensus sequence. Characters are colored to indicate variable positions. The alignment can be downloaded in FASTA, Clustal, Phylip, NEXUS, or ASN.1 formats. The tree display supports a variety of layouts including rectangular and slanted cladograms, radial trees and circular trees, the image can be downloaded as a PDF, and the tree file can be down-



95 unique protein sequences retrieved from query

Species	Host	Country or region	Genome region	Isolation	Full-length only	Collection date	Release date	Keyword	Additional filters	Number of sequences
Z	any	any	Nucleoprotein	blood	no	2014 – now	any		details	95
Z	any	any	Nucleoprotein	blood	no	2015 – now	any		details	3
Z	Human	any	Nucleoprotein	blood	no	2014 – now	any		details	95

Note: *All groups of identical sequences in the dataset will be represented by the oldest sequence in the group.*

☑	Identical sequences	GenBank	SRA	Length	Genome region	Host	Country	Isolation source	Collection date	Release date	Definition
☑		AKT08837		739	NP	Homo sapiens	Sierra Leone	blood	2015/02/19	2015/08/04	Zaire ebolavirus isolate Ebola virus H.sapiens-wt/SLE/2015/Makona-Goderich1, complete genome
☑		AIO11747		739	NP	Homo sapiens	Democratic Republic of the Congo	blood	2014	2014/10/17	Zaire ebolavirus isolate Ebola virus/H.sap-wt/COD/2014/Boende-Lokolia, partial genome
☑		ALO75886		739	NP	Homo sapiens	Italy	blood	2015/05	2015/11/24	Zaire ebolavirus isolate Ebola virus/H.sapiens-tc/ITA/2015/Makona-INMI2, partial genome
☑	3	AKL91087	SAMN03611814	739	NP	Homo sapiens	Guinea	blood	2014/03/19	2015/06/17	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/GIN/2014/Makona-C05, partial genome
☑	34	AKL91123	SAMN03611818	739	NP	Homo sapiens	Guinea	blood	2014/03/17	2015/06/17	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/GIN/2014/Makona-C15, partial genome
☑	2	AKG65724		739	NP	Homo sapiens	Guinea	blood	2014/09/18	2015/06/26	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/GIN/2014/Makona-Conakry-1027, partial genome
☑		AKG65751		739	NP	Homo sapiens	Guinea	blood	2014/09/25	2015/06/26	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/GIN/2014/Makona-Conakry-1128, partial genome
☑		AKG65760		739	NP	Homo sapiens	Guinea	blood	2014/09/25	2015/06/26	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/GIN/2014/Makona-Conakry-1129, partial genome
☑		AKG65661		739	NP	Homo sapiens	Guinea	blood	2014/10/24	2015/06/26	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/GIN/2014/Makona-Conakry-1651, partial genome
☑		AKG65328		739	NP	Homo sapiens	Guinea	blood	2014/09/24	2015/06/26	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/GIN/2014/Makona-Dalaba-1104, partial genome
☑		AKG65337		739	NP	Homo sapiens	Guinea	blood	2014/09/24	2015/06/26	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/GIN/2014/Makona-Dalaba-1116, partial genome
☑	2	AKG65391		739	NP	Homo sapiens	Guinea	blood	2014/10/02	2015/06/26	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/GIN/2014/Makona-Dalaba-1211, partial genome
☑		AKG65112		739	NP	Homo sapiens	Guinea	blood	2014/07/24	2015/06/26	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/GIN/2014/Makona-Siguiri-517, partial genome
☑		ALG01528	SAMN04040219	739	NP	Homo sapiens	Liberia	blood	2014/11/13	2015/11/24	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/LBR/2014/Makona-LIBR0159, partial genome
☑		ALG01105	SAMN04040223	739	NP	Homo sapiens	Liberia	blood	2014/11/22	2015/11/24	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/LBR/2014/Makona-LIBR0279, partial genome
☑		ALG01663	SAMN04040225	739	NP	Homo sapiens	Liberia	blood	2014/11/26	2015/11/24	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/LBR/2014/Makona-LIBR0350, partial genome
☑		ALG02095	SAMN04040230	739	NP	Homo sapiens	Liberia	blood	2014/08/19	2015/11/24	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/LBR/2014/Makona-LIBR10014, partial genome
☑	2	ALG01213	SAMN04040232	739	NP	Homo sapiens	Liberia	blood	2014/08/18	2015/11/24	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/LBR/2014/Makona-LIBR10020, partial genome
☑		ALG01087	SAMN04040240	739	NP	Homo sapiens	Liberia	blood	2014/08/16	2015/11/24	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/LBR/2014/Makona-LIBR10031, partial genome
☑		ALG01834	SAMN04040249	739	NP	Homo sapiens	Liberia	blood	2014/10/01	2015/11/24	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/LBR/2014/Makona-LIBR10048, partial genome
☑	4	ALG01258	SAMN04040265	739	NP	Homo sapiens	Liberia	blood	2014/08/26	2015/11/24	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/LBR/2014/Makona-LIBR10078, partial genome
☑		ALG01699	SAMN04040270	739	NP	Homo sapiens	Liberia	blood	2014/08/26	2015/11/24	Zaire ebolavirus isolate Ebola virus/H.sapiens-wt/LBR/2014/Makona-LIBR10087, partial genome

Figure 2. Virus Variation Resource results interface page. The results interface search criteria at the top of the page and a table of retrieved sequences below. There is a row of functions directly above the table of retrieved sequences that supports a number of actions. For example, users can select the visible columns in the results table using the 'Select columns' link, or quickly display multiple sequence alignments of selected sequences using the 'Build sequence alignment' button. There is also an option to customize sequence labels before downloading them or building trees. Individual GenBank or BioSample records listed in the table can be reviewed by clicking the hyperlinked accessions. If identical sequences were collapsed, they can be expanded to view individual accessions by clicking the blue arrow in the 'Identical sequences' column.

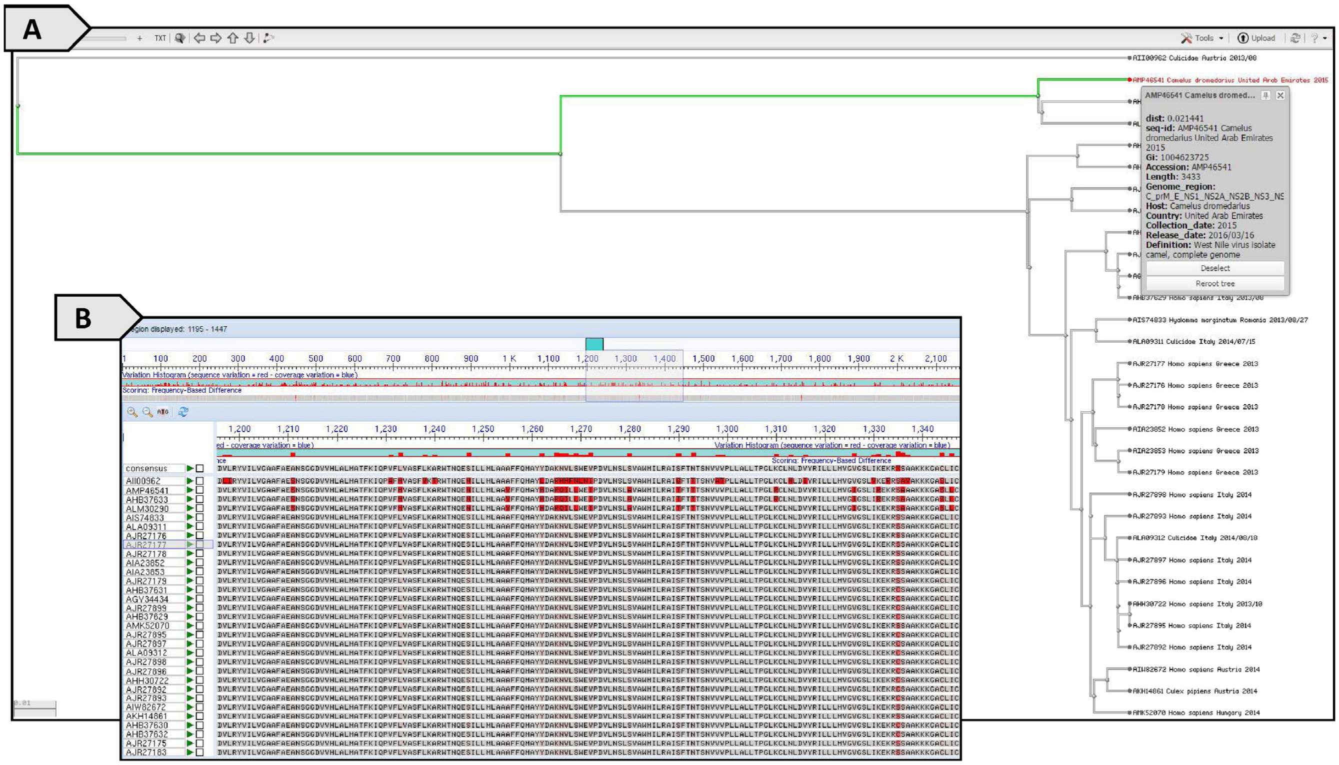
loaded in ASN text or binary, Newick, or NEXUS formats. These options are accessible through the 'Tools' menu in the viewer. The data labels on multi-sequence alignments and trees can be customized from the results table before the tree is calculated using the 'Customize label' options, making it easier to identify the distribution of sample/sequence characteristics. When certain download formats are selected, customized labels will be included in the downloaded files (FASTA and ASN.1 for the multiple alignments, and all files for the trees). A URL is also provided to make sharing a tree easy.

FUTURE DIRECTIONS

The Virus Variation Resource described here provides a number of features that improve the usability of archival sequence data. The resource now includes more than 20% of the GenBank sequences that are assigned viral taxonomy. Further improvement will be dependent on which viruses are added in the future and on updates to the various pipelines, interfaces and tools so that they can further support user needs. Our plan is to increase the pace at which new virus species are added to the Virus Variation Resource, and we are currently developing layers of data processing – the least transformative of which could be applied across all viral sequences but still provide basic information about a sequence. The search interface and data displays will be revised so that they better support user-required comparative genomic functions across a much larger number of viral species from the same query page. We also intend to support searches based on author names and more detailed sample information, such as clinical symptoms or laboratory handling. Though we will begin parsing the potentially rich metadata data sets from BioSample records, the success of this effort will ultimately rest on improved community awareness and more consistent submission of metadata to public databases.

Given the unbridled growth and clear potential of nucleotide sequencing efforts, one must assume the current Virus Variation Resource is just scratching the surface of future bioinformatic needs. The current resource model is suited to viruses that have experimentally validated annotation, and similar modules are in development for additional viral species. However, the vast majority of viruses do not have strong experimental evidence for protein coding regions, making it difficult to build a Virus Variation module including an annotation pipeline. In these cases annotation will need to be inferred from related, experimentally studied viruses, requiring new approaches and better ways of standardizing gene and protein information across multiple groups of viruses. Our current annotation pipeline development is directed toward these goals, and we intend to extend public access to these pipelines beyond our current influenza virus module. We also intend to reveal resource-derived annotation as tracks on multiple sequence alignments, making annotated sequences available for download and improving access to our data sets. This will also enable users to limit downloads and multiple sequence alignments to selected mature peptides for polyprotein sequences, and trees to be built from selected genomic regions.

Finally, there are a variety of enhancements to our tools under development. We are developing improved tree visualizations that support better search and markup functions, similar to those currently used in the influenza virus module. Some limitations of the tree function will be addressed at a later time by giving the user the option of viewing the quick tree which is currently offered, or a more


A
B
TXT
Tools
Upload
AII00962 Culicidae Austria 2013/08
AMP46541 Camelus dromedarius United Arab Emirates 2015
AHB37629 Homo sapiens Italy 2013/08
AIS74833 Hyalomma marginatum Romania 2013/08/27
ALA09311 Culicidae Italy 2014/07/15
AJR27177 Homo sapiens Greece 2013
AJR27176 Homo sapiens Greece 2013
AJR27178 Homo sapiens Greece 2013
AIA23852 Homo sapiens Greece 2013
AIA23853 Homo sapiens Greece 2013
AJR27179 Homo sapiens Greece 2013
AJR27898 Homo sapiens Italy 2014
AJR27893 Homo sapiens Italy 2014
ALA09312 Culicidae Italy 2014/08/18
AJR27897 Homo sapiens Italy 2014
AJR27896 Homo sapiens Italy 2014
AHH30722 Homo sapiens Italy 2013/10
AJR27895 Homo sapiens Italy 2014
AJR27892 Homo sapiens Italy 2014
AIW82672 Homo sapiens Austria 2014
AKH14861 Culex pipiens Austria 2014
AMK52070 Homo sapiens Hungary 2014
Deselect
Reroot tree
egion displayed: 1195 - 1447
Variation Histogram (sequence variation = red - coverage variation = blue)
Scoring: Frequency-Based Difference
consensus
AII00962
AMP46541
AHB37633
ALM30290
AIS74833
ALA09311
AJR27176
AJR27177
AJR27178
AIA23852
AIA23853
AJR27179
AHB37631
AGY34434
AJR27899
AHB37629
AMK52070
AJR27895
AJR27897
ALA09312
AJR27898
AJR27896
AHH30722
AJR27892
AJR27893
AIW82672
AKH14861
AHB37630
AHB37632
AJR27175
AJR27183

AMP46541 Camelus dromed...

dist: 0.021441
seq-id: AMP46541 Camelus dromedarius United Arab Emirates 2015
Gi: 1004623725
Accession: AMP46541
Length: 3433
Genome_region: C_prM_E_NS1_NS2A_NS2B_NS3_NS
Host: Camelus dromedarius
Country: United Arab Emirates
Collection_date: 2015
Release_date: 2016/03/16
Definition: West Nile virus isolate camel, complete genome

Figure 3. Virus Variation Resource tree and multi-sequence alignment displays. (**A**) A sample tree is shown depicting the use of standardized metadata terms as sequence labels. The tree was built from 31 West Nile virus complete polyprotein sequences collected since 2013. Sequence labels are based on GenBank accessions, host, country of isolation and isolation date. Left clicking a node highlights the lineage, and hovering over a node with the cursor displays a menu that includes descriptors for that particular sample, including GenBank accession and available standardized metadata terms for host, country, isolation source, etc. The menu also includes a function to reroot the tree around that sequence. (**B**) A multi-sequence alignment is shown for the same 31 West Nile polyprotein sequences. Individual GenBank accessions are listed to the left next to sequences. Left clicking the accession displays a menu that includes the standardized metadata label chosen in the results interface, a link to the sequence in GenBank, a function to use that sequence as an anchor for the alignment. Differences between residues in a given sequence and the consensus are highlighted in red. A histogram above the alignment shows coverage in blue and the frequency of changes in red.

sophisticated combination of MUSCLE-multiple sequence alignment and phylogenetic tree. We are also interested in supporting BLAST-based searches within our data sets to support more precise sequence associations. Ultimately, the presumed very large sequencing datasets of the future will ultimately require better ways to evaluate data retrieved from searches which, in turn, will require better integration of search functions with data visualizations such as trees.

Members of the scientific community are encouraged to contact the NCBI Help Desk (ncbi-help@ncbi.nlm.nih.gov) to make suggestions to improve the Virus Variation Resource, or to assist with establishing annotation or metadata standards.

FUNDING

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. Cochrane,G., Karsch-Mizrachi,I., Takagi,T. and International Nucleotide Sequence Database Collaboration. (2016) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, **44**, D48–D50.
2. Benson,D.A., Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2015) GenBank. *Nucleic Acids Res.*, **43**, D30–D35.
3. Paul,D. and Bartenschlager,R. (2015) Flaviviridae replication organelles: Oh, what a tangled web we weave. *Annu. Rev. Virol.*, **2**, 289–310.
4. Lingala,S. and Ghany,M.G. (2015) Natural history of Hepatitis C. *Gastroenterol. Clin. North Am.*, **44**, 717–734.
5. McVey,D.S., Wilson,W.C. and Gay,C.G. (2015) West Nile virus. *Rev. Sci. Tech.*, **34**, 431–439.
6. Bavia,L., Mosimann,A.L., Aoki,M.N. and Duarte Dos Santos,C.N. (2016) A glance at subgenomic flavivirus RNAs and microRNAs in flavivirus infections. *Virol. J.*, **13**, 84–103.
7. Faggioni,G., Pomponi,A., De Santis,R., Masuelli,L., Ciammaruconi,A., Monaco,F., Di Gennaro,A., Marzocchella,L., Sambri,V., Lelli,R. *et al.* (2012) West Nile alternative open reading frame (N-NS4B/WARF4) is produced in infected West Nile Virus (WNV) cells and induces humoral response in WNV infected individuals. *Virol. J.*, **9**, 283–296.
8. 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.
9. Foley,B., Leitner,T., Apetrei,C., Hahn,B., Mizrachi,I., Mullins,J., Rambaut,A., Wolinsky,S. and Korber,B. (2013) *HIV Sequence Compendium 2013*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, New Mexico, LA-UR 13-26007.
10. Greene,J.M., Collins,F., Lefkowitz,E.J., Roos,D., Scheuermann,R.H., Sobral,B., Stevens,R., White,O. and Di Francesco,V. (2007) National

Institute of Allergy and Infectious Diseases bioinformatics resource centers: new assets for pathogen informatics. *Infect. Immun.*, **75**, 3212–3219.

11. Pickett,B.E., Sadat,E.L., Zhang,Y., Noronha,J.M., Squires,R.B., Hunt,V., Liu,M., Kumar,S., Zaremba,S., Gu,Z. *et al.* (2012) ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.*, **40**, D593–D598.
12. Van Doorslaer,K., Tan,Q., Xirasagar,S., Bandaru,S., Gopalan,V., Mohamoud,Y., Huyen,Y. and McBride,A.A. (2013) The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Res.*, **41**, D571–D578.
13. 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**, 65–71.
14. Brister,J.R., Bao,Y., Zhdanov,S.A., Ostapchuck,Y., Chetvernin,V., Kiryutin,B., Zaslavsky,L., Kimelman,M. and Tatusova,T.A. (2014) Virus Variation Resource–recent updates and future directions. *Nucleic Acids Res.*, **42**, D660–D665.
15. Klema,V.J., Ye,M., Hindupur,A., Teramoto,T., Gottipati,K., Padmanabhan,R. and Choi,K.H. (2016) Dengue virus nonstructural protein 5 (NS5) assembles into a dimer with a unique methyltransferase and polymerase interface. *PLoS Pathog.*, **12**, e1005451.
16. Bell,A., Lewandowski,K., Myers,R., Wooldridge,D., Aarons,E., Simpson,A., Vipond,R., Jacobs,M., Gharbia,S. and Zambon,M. (2015) Genome sequence analysis of Ebola virus in clinical samples from three British healthcare workers, August 2014 to March 2015. *Euro Surveill.*, **20**, 6–10.
17. Agbemabiese,C.A., Nakagomi,T., Doan,Y.H., Do,L.P., Damanka,S., Armah,G.E. and Nakagomi,O. (2016) Genomic constellation and evolution of Ghanaian G2P[4] rotavirus strains from a global perspective. *Infect. Genet. Evol.*, **45**, 122–131.
18. Bao,Y., Bolotov,P., Dernovoy,D., Kiryutin,B. and Tatusova,T. (2007) FLAN: a web server for influenza virus genome annotation. *Nucleic Acids Res.*, **35**, W280–W284.
19. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
20. Matthijnssens,J., Ciarlet,M., McDonald,S.M., Attoui,H., Banyai,K., Brister,J.R., Buesa,J., Esona,M.D., Estes,M.K., Gentsch,J.R. *et al.* (2011) Uniformity of rotavirus strain nomenclature proposed by the Rotavirus Classification Working Group (RCWG). *Arch. Virol.*, **156**, 1397–1413.
21. Brister,J.R., Bao,Y., Kuiken,C., Lefkowitz,E.J., Le Mercier,P., Leplae,R., Madupu,R., Scheuermann,R.H., Schobel,S., Seto,D. *et al.* (2010) Towards viral genome annotation standards, report from the 2010 NCBI annotation workshop. *Viruses*, **2**, 2258–2268.
22. Brister,J.R., Le Mercier,P. and Hu,J.C. (2012) Microbial virus genome annotation-mustering the troops to fight the sequence onslaught. *Virology*, **434**, 175–180.
23. Kuhn,J.H., Andersen,K.G., Bao,Y., Bavari,S., Becker,S., Bennett,R.S., Bergman,N.H., Blinkova,O., Bradfute,S., Brister,J.R. *et al.* (2014) Filovirus RefSeq entries: evaluation and selection of filovirus type variants, type sequences, and names. *Viruses*, **6**, 3663–3682.
24. Brister,J.R., Ako-Adjei,D., Bao,Y. and Blinkova,O. (2015) NCBI viral genomes resource. *Nucleic Acids Res.*, **43**, D571–D577.
25. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
26. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
27. Matthijnssens,J., Ciarlet,M., Rahman,M., Attoui,H., Banyai,K., Estes,M.K., Gentsch,J.R., Iturriza-Gomara,M., Kirkwood,C.D., Martella,V. *et al.* (2008) Recommendations for the classification of group A rotaviruses using all 11 genomic RNA segments. *Arch. Virol.*, **153**, 1621–1629.
28. Maes,P., Matthijnssens,J., Rahman,M. and Van Ranst,M. (2009) RotaC: a web-based tool for the complete genome classification of group A rotaviruses. *BMC Microbiol.*, **9**, 238–241.
29. Barrett,T., Clark,K., Gevorgyan,R., Gorelenkov,V., Gribov,E., Karsch-Mizrachi,I., Kimelman,M., Pruitt,K.D., Resenchuk,S., Tatusova,T. *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.*, **40**, D57–D63.
30. Kodama,Y., Shumway,M., Leinonen,R. and International Nucleotide Sequence DatabaseCollaboration. (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
31. 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**, 237–243.
32. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.