Published online 17 October 2011 Nucleic Acids Research, 2012, Vol. 40, Database issue D593–D598
doi:10.1093/nar/gkr859

ViPR: an open bioinformatics database and analysis resource for virology research

Brett E. Pickett[1], Eva L. Sadat[1], Yun Zhang[1], Jyothi M. Noronha[1], R. Burke Squires[1], Victoria Hunt[1], Mengya Liu[2], Sanjeev Kumar[3], Sam Zaremba[3], Zhiping Gu[3], Liwei Zhou[3], Christopher N. Larson[4], Jonathan Dietrich[3], Edward B. Klem[3] and Richard H. Scheuermann[1,5,*]

[1]Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX, 75390, [2]Department of Statistical Science, Southern Methodist University, Dallas, TX, 75275, [3]Northrop Grumman Health IT Systems, Rockville, MD, 20850, [4]Vecna Technologies, Greenbelt, MD, 20770 and [5]Division of Biomedical Informatics, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA

Received August 15, 2011; Revised September 21, 2011; Accepted September 23, 2011

ABSTRACT

The Virus Pathogen Database and Analysis Resource (ViPR, www.ViPRbrc.org) is an integrated repository of data and analysis tools for multiple virus families, supported by the National Institute of Allergy and Infectious Diseases (NIAID) Bioinformatics Resource Centers (BRC) program. ViPR contains information for human pathogenic viruses belonging to the *Arenaviridae*, *Bunyaviridae*, *Caliciviridae*, *Coronaviridae*, *Flaviviridae*, *Filoviridae*, *Hepeviridae*, *Herpesviridae*, *Paramyxoviridae*, *Picornaviridae*, *Poxviridae*, *Reoviridae*, *Rhabdoviridae* and *Togaviridae* families, with plans to support additional virus families in the future. ViPR captures various types of information, including sequence records, gene and protein annotations, 3D protein structures, immune epitope locations, clinical and surveillance metadata and novel data derived from comparative genomics analysis. Analytical and visualization tools for metadata-driven statistical sequence analysis, multiple sequence alignment, phylogenetic tree construction, BLAST comparison and sequence variation determination are also provided. Data filtering and analysis workflows can be combined and the results saved in personal 'Workbenches' for future use. ViPR tools and data are available without charge as a service to the virology research community to help facilitate the development of diagnostics, prophylactics and therapeutics for priority pathogens and other viruses.

INTRODUCTION

Viral disease outbreaks tend to occur every several years in the human population (1–4). During such outbreaks, identification of the causative agent and comparative genomic analysis can be critical to limiting the spread of the virus and identifying suitable treatment options. While the first lines of defense against such outbreaks are clinical reporting and automated surveillance, 'wet lab' experimentation of viral isolates is also important. Bioinformatics tools and database resources that provide information about the genomic structures and phenotypic characteristics of known viruses can make this process more efficient by supporting data mining for the development of hypotheses worthy of in-depth laboratory experimentation on the emerging strain.

Sequence data deposited in archives such as GenBank are extremely valuable in the computational analysis of emerging viruses. These sequence records can be enhanced through integration with additional knowledge about the viral strains including: gene and protein annotations, immune epitope locations, 3D protein structures, clinical metadata, etc. Database resources and modern computing make storing and retrieving such a wealth of data tractable.

With foresight into the importance of providing access to such information, the National Institute of Allergy and Infectious Diseases (NIAID) within the US National Institutes of Health (NIH) implemented the Bioinformatics Resource Centers (BRC) for Infectious Diseases program (5,6) to develop open, integrated online resources for data about human pathogens. Five such centers currently exist, each supporting a different class of human pathogens or insect vectors (www.pathogenportal.org). The Virus Pathogen Database and

[*] To whom correspondence should be addressed. Tel: +1 214 648 4115; Fax: +1 214 648 4070; Email: richard.scheuermann@utsouthwestern.edu

Analysis Resource (ViPR; www.ViPRbrc.org) serves as the publicly accessible repository for viruses categorized as either A–C priority pathogens or viruses that adversely affect public health (http://www.niaid.nih.gov/topics/biodefenserelated/biodefense/research/pages/cata.aspx). ViPR is unique among other virus-centered databases in that it contains a wealth of integrated information for a large number of virus families that are pathogenic specifically to humans. This is in contrast with the NCBI viral genome project, which provides sequence data for all viruses (7) or the Human Immunodeficiency Virus sequence database (http://www.hiv.lanl.gov) and the Influenza Research Database (IRD, www.fludb.org) (8), which focus on a particular taxon or virus pathogen.

The goal of the BRC program is to provide the necessary data, bioinformatics tools and workflows to enhance ongoing basic and applied research. By integrating the data with a variety of computational analysis tools free of charge to the virology research community, complex analyses that take advantage of cross-referenced data within the ViPR database become possible.

DESCRIPTION

Summary of ViPR data

As of June 2011, ViPR holds sequence data and related information for over 50 000 virus strains from 912 species belonging to 70 genera and 14 families: *Arenaviridae*, *Bunyaviridae*, *Caliciviridae*, *Coronaviridae*, *Filoviridae*, *Flaviviridae*, *Hepeviridae*, *Herpesviridae*, *Paramyxoviridae*, *Picornaviridae*, *Poxviridae*, *Reoviridae*, *Rhabdoviridae* and *Togaviridae*. ViPR integrates data and other information from three different types of sources: (i) data transferred from public archives, (ii) data directly submitted by researchers and (iii) data derived through computational methods.

Data from public archives. The ViPR database integrates various types of data acquired from multiple publicly accessible database resources (Supplementary Table S1). Specifically, ViPR includes >64 000 genomic segment sequences from GenBank (9), >220 000 protein sequences from UniProt (10), >1400 experimentally determined T-cell and B-cell epitopes from the Immune Epitope Database (IEDB; www.iedb.org) (11), >2900 structures from the Protein Data Bank (PDB; www.pdb.org) (12) and >59 000 Gene Ontology annotations (GO, www.geneontology.org) (13) (Supplementary Table S2). All these data are updated with each bimonthly release and are directly searchable using intuitive web-based user interfaces.

Data from direct submission. Additional projects, including those that involve Dengue virus and SARS coronavirus strains being sequenced by the NIAID-sponsored Genomic Sequencing Centers for Infectious Diseases program, submit metadata associated with the sequence data directly to ViPR including human clinical information about disease symptoms, diagnostic test results, disease severity, etc. associated with infection. Searching for strains based on direct submission metadata for those viruses annotated with such data returns strain records and genome sequences matching the query criteria.

Novel derived and predicted data. These public and direct submission data are augmented with novel data derived from numerous comparative genomics and other bioinformatics analyses performed by the ViPR team. Examples of derived data at the strain level include computationally improved annotations and manually curated information, which are displayed with a Genome Map image and a Protein Information table on the Strain Details page. Selecting a gene or protein from the image or table loads the Gene/Protein Details page (Supplementary Figure S1), which contains derived data consisting of predicted immune epitopes determined using the NetCTL algorithm (14), protein domains and motifs predicted using InterProScan (15), molecular weight, isoelectric point, closest BLAST hits, homologous PDB structures, etc. These data are combined with the strain name, sequence data, virus taxonomy, host and country of isolation, collection date, as well as other information on the Gene/Protein Details pages.

The ViPR annotation process extends the information contained in the representative RefSeq strain for each virus species. For example, multiple sequence alignment is used to map homologous regions across related virus genomes in order to transfer sequence region annotations, including mature peptide cleavage sites on polyproteins, to the genomes lacking annotations. Virus Orthologous Clusters (VOC) annotations group sets of proteins having similar functions within large DNA virus families as determined using the OrthoMCL algorithm (16).

We have recently developed a novel Sequence Feature Variant Type (SFVT) component in ViPR that catalogs the precise location of characterized regions within virus proteins. This component is based on similar work performed for the human HLA proteins (17), and has been customized for the virology community. Information used to define the various ‘Sequence Features’ (SFs) was obtained from UniProt, GenBank, IEDB and the scientific literature. SFs are categorized as structural (e.g. alpha helices), functional (e.g. active sites), immune epitopes and sequence alteration positions, with current support for Hepatitis C (subtype 1a), Dengue (serotypes 1–4) and Pox (Vaccinia) viruses. Initial SF definitions for each characterized region have been inspected and validated by domain experts to ensure accuracy.

All sequence records in ViPR belonging to the designated taxon are searched to identify all unique amino acid sequence variations existing for a defined SF. Strains that repeat the same sequence variation pattern are assigned to the same ‘Variant Type’ (VT). All computed results are then stored in the database and can be accessed within ViPR. The Sequence Feature Details page displays additional information about the SF including the protein and strain from which it was originally identified, the observed VTs, hyperlinks to the homologous 3D protein structures, and the ability to search for a VT based on sequence. For simplicity, links to the relevant defined

SFs and VTs are found on the respective Protein Details pages. By subdividing strains based on the unique sequence variations in defined protein regions, SFVT analysis can rapidly identify genotypic polymorphisms that correlate with a specific phenotype at a finer level of granularity than was previously possible.

Search and analysis capabilities. To access ViPR data as well as analytical and visualization tools, the user begins by selecting a virus family on the home page (Supplementary Figure S2). ViPR has been designed in this way to manage the unique genomic structural and data type characteristics associated with different virus families. Queries for virus strains can be constructed to include genus, species, geographical and/or temporal points of isolation, virus host, clinical data (where available), etc. Alternatively, simple keyword searches and more advanced searches can be performed to retrieve the desired information. Virus genomes that match the query are displayed in a summary table on the Genome Search Results page (Supplementary Figure S3). Selecting any of the strains in the list will bring up the corresponding Strain Details page as described above.

The data in ViPR can be analyzed using a suite of comparative genomics analysis and visualization tools including phylogenetic tree reconstruction with the FastME, RAxML or PhyML algorithms (18–20), tree manipulation with Archaeopteryx (21) and evolutionary model optimization with modelTest (22); multiple sequence alignments calculated with MUSCLE (23) and visualized with JalView (24,25), 3D protein structure and sequence feature visualization with Jmol (26), a metadata-driven comparative genomics automated workflow, BLAST (27), sequence variation analysis using sequence logos (28), a protein sequence pattern matching tool and the automated Genome Annotation Transfer Utility (29). Several of these tools will be described in more detail within the context of the comparative analysis use case described below.

Scientific use case illustrating comparative analysis

ViPR development has been guided by various scientific use cases to define requirements for relevant data collection and storage, database queries, and analysis methods to improve both the integrated informatics and the supported analytical workflows. To showcase how the various types of data and associated metadata within ViPR can be used to explore sequence variation within a virus species, we will use Dengue virus (DENV) data in a comparative genomics use case.

Dengue virus is an arthropod-borne virus native to tropical regions of the world and endemic in areas where it colocalizes with the preferred *Aedes aegypti* mosquito vector. Because of these restrictions, it can be assumed that DENV infections reported in clinics located in non-tropical regions of the world are likely due to recent travel by the patient to an endemic area. These imported cases can thus establish viral lineage in new regions as a result of human travel. The CDC has recently demonstrated that the travel history of US residents having been clinically diagnosed with DENV between 1999 and 2000 validates such an explanation (30). As an example use case, we will extend the CDC study by performing an in-depth comparative genomics analysis of all DENV serotype 1–4 isolates taken between the years of 1999 and 2000, involving the following bioinformatics workflow: (i) identify sequence records using the ViPR Genome Search interface; (ii) save the matching sequence records as a working set in a personal Workbench; (iii) reconstruct a phylogenetic tree; (iv) visualize the multiple sequence alignment; (v) perform a metadata-driven statistical analysis of sequence variation; (vi) determine where these differentiating residues are located in relation to validated Sequence Features; and (vii) examine the 3D structure associated with a homologous protein from a related DENV-2 strain. Although this illustration is focused on Dengue virus, it should be noted that similar tasks can be performed for other virus species to address other biological questions by combining the wealth of relevant data with the suite of bioinformatics tools integrated into ViPR (Supplementary Figure S4).

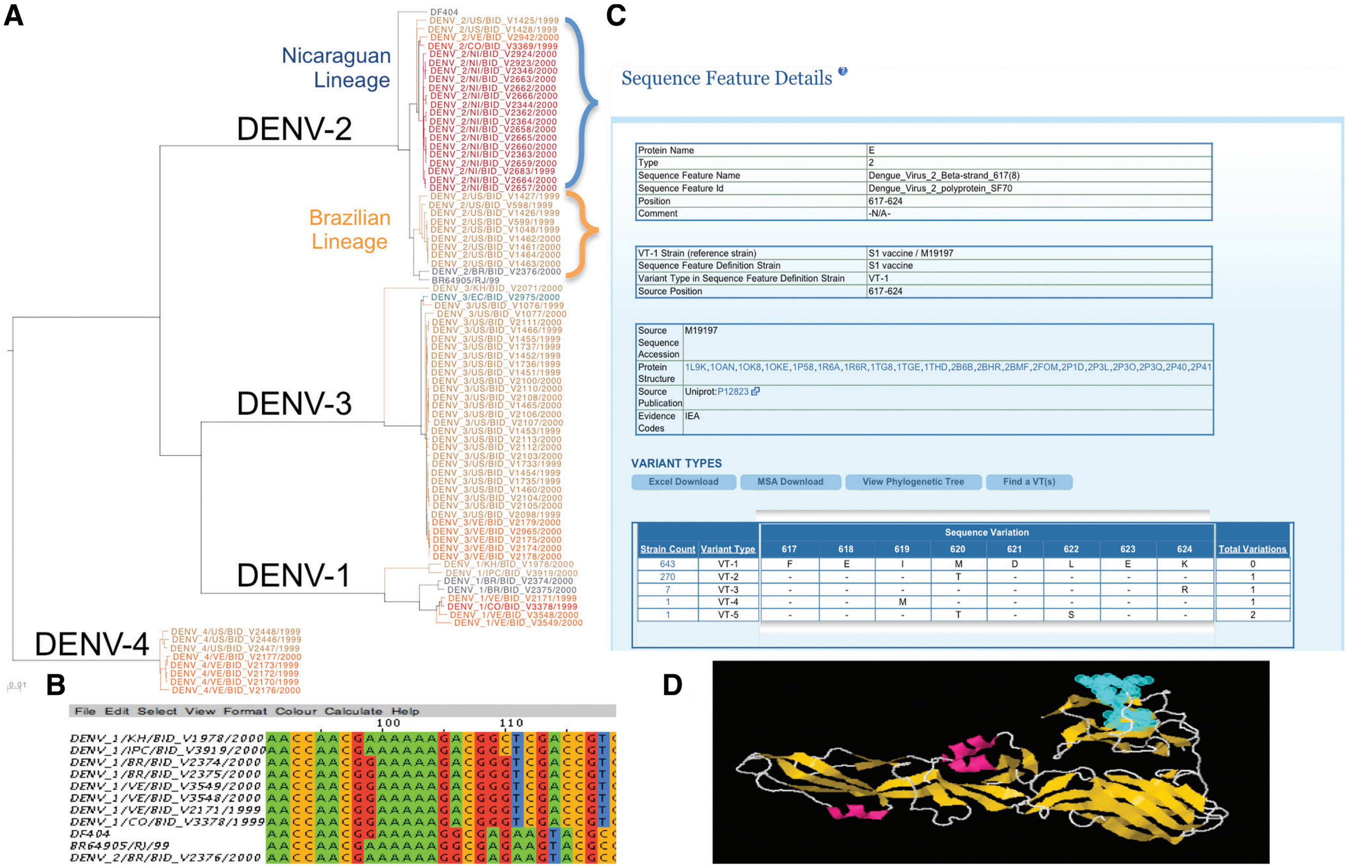
Search for relevant sequence records. To begin, a query is constructed for all DENV 1–4 records, isolated from humans, between the years 1999 and 2000. In June 2011, this specific query returned 82 whole genome records (8 DENV-1, 33 DENV-2, 33 DENV-3 and 8 DENV-4) from eight countries (Brazil, Cambodia, Colombia, Ecuador, Nicaragua, United States, Venezuela and Vietnam). The content on the Genome Search Results page can be sorted by clicking on any of the column headings. Selected records from this page can then be transferred to any of the relevant analysis and visualization tools using the ‘Run Analysis’ pull-down tab.

Save to the Workbench. The Workbench feature in ViPR is relatively novel among virus database resources and is used to store search and analysis results in designated personal workspaces on the ViPR server. Distinct personal workspaces can be established for each virus family by providing a valid email address and a password for login purposes. The Workbench provides an interface to construct ‘working sets’, consisting of the results from one or more searches, and allows the re-use of sequence data from various strains in multiple analyses (Supplementary Figure S5). Additionally, users can upload their own custom sequence data and other files to the Workbench for analysis using the various tools provided by the system, and can share items within their Workbench with collaborators, virtually. For the current use case, all the genome sequence records from the query result page described above are saved as a custom working set for subsequent analysis.

Phylogenetic tree calculation and visualization. Phylogenetic trees can be calculated on the ViPR server using sequence records selected from a query result, an existing working set or uploaded to the system through a web interface. Phylogenetic tree reconstructions can then be saved in the Workbench or downloaded in either Newick or phyloXML format (31). For the current use case, a FastME phylogenetic tree is constructed with the desired sequences by choosing the 'Quick Tree' option on the 'Generate Phylogenetic Tree' page. PhyML and RAxML are also provided as alternative phylogenetic tree reconstruction algorithms.

Once complete, the Archaeopteryx phylogenetic tree viewer can be used to quickly visualize, explore and manipulate the resulting tree using typical actions such as re-rooting, select subtrees, branch swapping, etc. This application has recently been customized to take advantage of the extensive sequence metadata within the ViPR database by allowing user-driven coloring of tree 'leaves' according to year, country of isolation or host from which the strain was originally isolated. The customized graphic can then be exported for enhanced interpretation and inclusion in publications. For the current use case, the phylogeny was calculated for all sequences matching our initial query and colored by the country of isolation. This tree reconstruction shows the major branching structures that separate sequences primarily according to serotype (Figure 1A). However, within the serotype 2 branch, two separate US lineages can be identified—one corresponding to a Brazilian lineage and one corresponding to a Nicaraguan lineage—suggesting that two different serotype 2 introduction events have occurred in the United States.

Multiple sequence alignment calculation and visualization. Sequence data from multiple sources, including search results, working sets and uploaded sequences in FASTA format, can be used as input to run a custom MUSCLE alignment on the ViPR server. After completion, ViPR assists users in viewing, exploring and modifying the label or sequence information within an alignment. Upon visual inspection with JalView, the alignment of nucleotide sequences from the Dengue virus use case shows that these genomes are well conserved overall with the serotype-specific relations observed in the phylogenetic analysis recapitulated in the sequence alignments (Figure 1B).



```
DENV_1/KH/BID_V1978/2000  AACCAACGAAAAAAGACGGCTCGACCGT
DENV_1/IPC/BID_V3919/2000 AACCAACGAAAAAAGACGGCTCGACCGT
DENV_1/BR/BID_V2374/2000  AACCAACGGAAAAAGACGGGTCGACCGT
DENV_1/BR/BID_V2375/2000  AACCAACGGAAAAAGACGGGTCGACCGT
DENV_1/VE/BID_V3549/2000  AACCAACGGAAAAAGACGGGTCGACCGT
DENV_1/VE/BID_V3548/2000  AACCAACGGAAAAAGACGGGTCGACCGT
DENV_1/VE/BID_V2171/1999  AACCAACGGAAAAAGACGGGTCGACCGT
DENV_1/CO/BID_V3378/1999  AACCAACGGAAAAAGACGGGTCGACCGT
DF404                     AACCAACGGAAAAAGGCGAGAAGTACGCC
BR64905/RJ/99             AACCAACGAAAAAAGGCGAGAAGTACGCC
DENV_2/BR/BID_V2376/2000  AACCAACGAAAAAAGGCGAGAAGTACGCC
```

Sequence Feature Details

Protein Name	E
Type	2
Sequence Feature Name	Dengue_Virus_2_Beta-strand_617(8)
Sequence Feature Id	Dengue_Virus_2_polyprotein_SF70
Position	617-624
Comment	-N/A-

VT-1 Strain (reference strain)	S1 vaccine / M19197
Sequence Feature Definition Strain	S1 vaccine
Variant Type in Sequence Feature Definition Strain	VT-1
Source Position	617-624

Source Sequence Accession	M19197
Protein Structure	1L9K,1OAN,1OK8,1OKE,1P58,1R6A,1R6R,1TG8,1TGE,1THD,2B6B,2BHR,2BMF,2FOM,2P1D,2P3L,2P3O,2P3Q,2P40,2P41
Source Publication	Uniprot:P12823
Evidence Codes	IEA

VARIANT TYPES

Excel Download | MSA Download | View Phylogenetic Tree | Find a VT(s)

		Sequence Variation								
Strain Count	Variant Type	617	618	619	620	621	622	623	624	Total Variations
643	VT-1	F	E	I	M	D	L	E	K	0
270	VT-2	-	-	-	T	-	-	-	-	1
7	VT-3	-	-	-	-	-	-	-	R	1
1	VT-4	-	-	M	-	-	-	-	-	1
1	VT-5	-	-	-	T	-	S	-	-	2

Figure 1. Online Bioinformatics Tools Provided in ViPR. Various bioinformatics analyses can be performed, visualized and stored on the ViPR server. (**A**) A phylogenetic tree, in Archaeopteryx, colored by country of isolation with dark gray, light brown, tan, orange, red, dark red, blue and light gray representing viruses isolated in Vietnam, United States, Cambodia, Venezuela, Colombia, Nicaragua, Ecuador and Brazil, respectively. Strains belonging to the four serotypes are indicated. The blue text and brace indicate DENV-2 strains belonging to a Nicaraguan lineage, whereas the orange text and brace delineate DENV-2 strains from a Brazilian lineage. (**B**) A portion of a multiple sequence alignment, visualized with JalView, constructed using MUSCLE from the nucleotide sequences for the DENV use case. (**C**) The Sequence Feature Details page showing the SF metadata and the VTs for a β-strand structure in the DENV-2 E protein, named Dengue_Virus_2_Beta-straind_617(8), spanning positions 617–624 of the polyprotein. (**D**) A DENV E protein structure (PDB: 1OK8) in the ViPR implementation of the Jmol structure viewer with α-helices in magenta, β-strands in yellow and the SF from (C) highlighted in cyan.

Metadata-driven comparative analysis tool for sequences. The metadata-driven Comparative Analysis Tool for Sequences (meta-CATS) is an automated workflow, developed by the ViPR team, to assist researchers in taking advantage of the breadth of sequence data and the accompanying metadata. Metadata is the information associated with the sequence record, including time and place of specimen isolation, host species, clinical symptoms, etc. By using statistics to simultaneously analyze the sequence and metadata, genotype–phenotype associations can be inferred. This tool allows users to quickly and easily select multiple sequences, align those sequences, divide them into multiple groups based on any one (or more) metadata type(s), perform automated statistical analyses on the sequences and view the results. The output from this tool has been validated using a previously published sequence set divided into two groups based on phylogenetic tree topology (32).

For the current use case, the DENV-2 polyprotein sequences matching the original search criteria are divided into two groups with 9 and 21 strains, respectively, according to the topology of the phylogenetic tree mentioned above. The meta-CATS analysis identified 37 homologous positions that significantly differ between the two defined DENV-2 lineages.

Sequence feature variant types. The SFVT component of ViPR can be used to identify the characterized structural and functional regions in the DENV polyprotein that contain the substitutions found to differentiate the metadata groups. For the current use case, polyprotein position 620 was one of 37 residues that were identified as significantly differing between the two DENV-2 lineages. This position is located within a structural SF named Dengue_Virus_2_Beta-strand_617(8), indicating that this region contains a β-strand protein secondary structure that begins at residue 617 of the polyprotein and continues for 8 residues (Figure 1C), with the Brazilian introduction corresponding to VT-1 and the Nicaraguan introduction to VT-2.

3D structure visualization. ViPR includes the Jmol protein structure viewer application to permit the rapid exploration and visualization of virus-related structures from the PDB. We have enhanced this tool by including the ability to highlight ligands and active sites on the displayed 3D protein structure. Options to customize the general appearance of the protein structure are provided. Individual residues within the protein structure(s) for each PDB file are mapped to homologous positions from UniProt records to make comparison between different structures and the associated amino acid sequence as simple as possible. In the future, highlighting of immune epitopes and other sequence features will be supported.

The ViPR implementation of the Jmol 3D structure viewer was used to explore how the domains within the DENV-2 E protein relate to the Dengue_Virus_2_Beta-strand_617(8) SF identified in the current use case by highlighting the region using the customization option (Figure 1D).

Conclusions from scientific use case. The workflow that was followed to explore the scientific use case confirms previous findings that DENV-2 in the Western Hemisphere exists as multiple endemic lineages that cocirculate among human and vector host populations (33,34). The meta-CATS analysis identified 37 amino acid variations that significantly distinguish these DENV-2 lineages, including one located within a known β-strand of the E protein. Additional investigation will be required to determine whether sequence differences in any of these individual positions or their combinations affect the secondary structure of the involved protein or play a role in dictating different phenotypic characteristics of these two virus lineages. The scientific use case addressed here underscores the ability of ViPR to assist in generating biologically relevant hypotheses that can then be tested experimentally.

CONCLUSIONS

The Virus Pathogen Database and Analysis Resource (ViPR, www.ViPRbrc.org) is an integrated repository of data and analysis tools for multiple virus families, supported by the National Institute of Allergy and Infectious Diseases (NIAID) Bioinformatics Resource Centers (BRC) program. The uniqueness of ViPR lies in (i) integrating data from many sources; (ii) encouraging the analysis of the extensive data contained within the system; (iii) combining the available tools to quickly perform complex analytical workflows; (iv) facilitating rapid hypothesis generation using bioinformatics methods for subsequent experimental testing; and (v) allowing data storage and sharing with collaborators in personal workbenches. By taking advantage of the powerful suite of resources provided within the ViPR BRC, virology researchers can streamline and expedite experimental discovery, for the ultimate goal of developing improved diagnostics, prophylactics and/or therapeutics for pathogenic viruses. The availability of such a resource can not only decrease the time required for scientific discovery at the 'bench', but also aid in translating those findings to the development of viable diagnostics, prophylactics and/or therapeutics.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online. Supplementary Table 1, Supplementary Figures 1–5.

ACKNOWLEDGEMENTS

First and foremost, we would like to acknowledge the contribution of all providers of data to public database resources like ViPR. Without their efforts in both generating the data and making it available through public data repositories, the potential for this kind of comprehensive integrative data analysis would not be possible. We also express our gratitude to the external experts who validated the SF definitions. We thank Andrei Gabrielian and Valentina Di Francesco for helpful insight through their oversight of the BRC program.

FUNDING

The development of the ViPR Bioinformatics Resource Center has been funded by N01-AI2008038 (to R.H.S.) from the US National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services. Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. CDC. (1999) From the Centers for Disease Control and Prevention. Outbreak of West Nile-like viral encephalitis–New York, 1999. *MMWR Morb. Mortal Wkly Rep.*, **48**, 845–849.
2. Ha,D.Q., Tien,N.T., Huong,V.T., Loan,H.T. and Thang,C.M. (2000) Dengue epidemic in southern Vietnam, 1998. *Emerg. Infect. Dis.*, **6**, 422–425.
3. CDC. (2003) From the Centers for Disease Control and Prevention. Severe acute respiratory syndrome–Taiwan, 2003. *JAMA*, **289**, 2930–2932.
4. Trifonov,V., Khiabanian,H. and Rabadan,R. (2009) Geographic dependence, surveillance, and origins of the 2009 influenza A (H1N1) virus. *N. Engl. J. Med.*, **361**, 115–119.
5. 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.
6. Lefkowitz,E.J., Upton,C., Changayil,S.S., Buck,C., Traktman,P. and Buller,R.M. (2005) Poxvirus Bioinformatics Resource Center: a comprehensive Poxviridae informational and analytical resource. *Nucleic Acids Res.*, **33**, D311–D316.
7. Bao,Y., Federhen,S., Leipe,D., Pham,V., Resenchuk,S., Rozanov,M., Tatusov,R. and Tatusova,T. (2004) National center for biotechnology information viral genomes project. *J. Virol.*, **78**, 7291–7298.
8. Squires,B., Macken,C., Garcia-Sastre,A., Godbole,S., Noronha,J., Hunt,V., Chang,R., Larsen,C.N., Klem,E., Biersack,K. *et al.* (2008) BioHealthBase: informatics support in the elucidation of influenza virus host pathogen interactions and virulence. *Nucleic Acids Res.*, **36**, D497–D503.
9. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Federhen,S. *et al.* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38–D51.
10. Apweiler,R., Martin,M.J., O'Donovan,C., Magrane,M., Alam-Faruque,Y., Antunes,R., Barrell,D., Bely,B., Bingley,M., Binns,D. *et al.* (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
11. Vita,R., Zarebski,L., Greenbaum,J.A., Emami,H., Hoof,I., Salimi,N., Damle,R., Sette,A. and Peters,B. (2010) The immune epitope database 2.0. *Nucleic Acids Res.*, **38**, D854–D862.
12. Berman,H., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**, 980.
13. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
14. Larsen,M.V., Lundegaard,C., Lamberth,K., Buus,S., Lund,O. and Nielsen,M. (2007) Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics*, **8**, 424.
15. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
16. Li,L., Stoeckert,C.J. Jr and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
17. Thomson,G., Marthandan,N., Hollenbach,J.A., Mack,S.J., Erlich,H.A., Single,R.M., Waller,M.J., Marsh,S.G., Guidry,P.A., Karp,D.R. *et al.* (2010) Sequence feature variant type (sfvt) analysis of the hla genetic association in juvenile idiopathic arthritis. *Pac. Symp. Biocomput.*, 359–370.
18. Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.
19. Desper,R. and Gascuel,O. (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.*, **9**, 687–705.
20. Stamatakis,A., Ludwig,T. and Meier,H. (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, **21**, 456–463.
21. Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.
22. Posada,D. and Crandall,K.A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.
23. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
24. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
25. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
26. Hanson,R. (2010) Jmol - a paradigm shift in crystallographic visualization. *J. Appl. Crystallogr.*, **43**, 1250–1260.
27. 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.
28. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
29. Tcherepanov,V., Ehlers,A. and Upton,C. (2006) Genome Annotation Transfer Utility (GATU): rapid annotation of viral genomes using a closely related reference genome. *BMC Genomics*, **7**, 150.
30. CDC. (2002) From the Centers for Disease Control and Prevention. Imported dengue–United States, 1999 and 2000. *MMWR Morb. Mortal Wkly Rep.*, **51**, 281–283.
31. Han,M.V. and Zmasek,C.M. (2009) phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, **10**, 356.
32. Pickett,B.E., Striker,R. and Lefkowitz,E.J. (2010) Evidence for separation of HCV subtype 1a into two distinct clades. *J. Viral Hepat*, **18**, 608–618.
33. Oliveira,M.F., Galvao Araujo,J.M., Ferreira,O.C. Jr, Ferreira,D.F., Lima,D.B., Santos,F.B., Schatzmayr,H.G., Tanuri,A. and Ribeiro Nogueira,R.M. (2010) Two lineages of dengue virus type 2, Brazil. *Emerg. Infect. Dis.*, **16**, 576–578.
34. Bennett,S.N., Holmes,E.C., Chirivella,M., Rodriguez,D.M., Beltran,M., Vorndam,V., Gubler,D.J. and McMillan,W.O. (2006) Molecular evolution of dengue 2 virus in Puerto Rico: positive selection in the viral envelope accompanies clade reintroduction. *J. Gen. Virol.*, **87**, 885–893.