D576–D582 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 14 October 2010*
doi:10.1093/nar/gkq901

ViralZone: a knowledge resource to understand virus diversity

Chantal Hulo[1], Edouard de Castro[1], Patrick Masson[1], Lydie Bougueleret[1], Amos Bairoch[2,3], Ioannis Xenarios[1,4] and Philippe Le Mercier[1,*]

[1]Swiss-Prot group, Swiss Institute of Bioinformatics, Centre Médical Universitaire, CH-1211 Geneva 4, [2]CALIPHO group, Swiss Institute of Bioinformatics, Centre Médical Universitaire, CH-1211 Geneva 4, [3]Département de biologie structurale et bioinformatique, University of Geneva, CH-1211 Geneva 4 and [4]Vital-IT Group, Swiss Institute of Bioinformatics, Quartier Sorge – Bâtiment Génopode, 1015, Lausanne, Switzerland

Received August 13, 2010; Revised September 17, 2010; Accepted September 22, 2010

ABSTRACT

The molecular diversity of viruses complicates the interpretation of viral genomic and proteomic data. To make sense of viral gene functions, investigators must be familiar with the virus host range, replication cycle and virion structure. Our aim is to provide a comprehensive resource bridging together textbook knowledge with genomic and proteomic sequences. ViralZone web resource (www.expasy.org/viralzone/) provides fact sheets on all known virus families/genera with easy access to sequence data. A selection of reference strains (RefStrain) provides annotated standards to circumvent the exponential increase of virus sequences. Moreover ViralZone offers a complete set of detailed and accurate virion pictures.

INTRODUCTION

Viruses are presumably the most abundant biological entities on the planet, with the total number of virus particles exceeding by 10 times the total number of cells (25). Many viruses have a relatively small genome encoding for a few proteins: one of the smallest being the circovirus with a 1.7-kb genome coding only two proteins (11). Despite their apparent simplicity, viral biochemistry and replication mechanisms are more varied than those seen in the entire bacterial, plant and animal kingdoms (15, 19). Nearly every possible method for encoding information in nucleic acid is exploited by viruses, from single-stranded DNA to double-stranded RNA. Each of the 83 virus families has a different replication strategy which calls for unique proteins and unique enzymes (19). For example the replication cycles of Human herpesvirus 1 (HHV-1) and Ebolavirus (EBOV) have nothing in common (Figure 1). The dsDNA HHV-1 genome encodes 73 proteins and replicates in the host nucleus where new viral genomes are encapsidated before budding through the endoplasmic reticulum and then into vesicles that will release the virion out of the cell (6,26). The EBOV ssRNA genome encodes eight proteins, replicates in the host cell cytoplasm using its own RNA-dependent RNA polymerase complex and buds directly at the plasma membrane (2). These two disparate replication cycles only illustrate the tremendous variety of viral molecular biology. As a result, it is crucial to have a clear vision of a specific virus' biology in order to understand its genome and protein functions. Yet this information is hardly available outside academic books.

To help solve this problem, The Swiss-Prot virus annotation team has developed a website dedicated to viruses: ViralZone (www.expasy.org/viralzone). The concept of this website is to link specific knowledge for each virus family with viral protein and genomic sequences. All the available information is presented in a concise and accessible virus fact sheet. The fact sheets contain condensed information about genome, replication cycle, taxonomy and epidemiology as well as graphics describing virion organization, genome transcription and translation strategies. The whole site comprises 426 fact sheets covering the whole known virosphere: 83 families, 334 genera and nine additional pages dedicated to important species like Influenza H1N1 or HIV-1.

VIRUS TAXONOMY

Unlike Luca for cellular organism (14), there is no presumed common ancestor for viruses (12). Therefore

*To whom correspondence should be addressed. Tel: +41223795870; Fax: +41223795858; Email: philippe.lemercier@isb-sib.ch

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

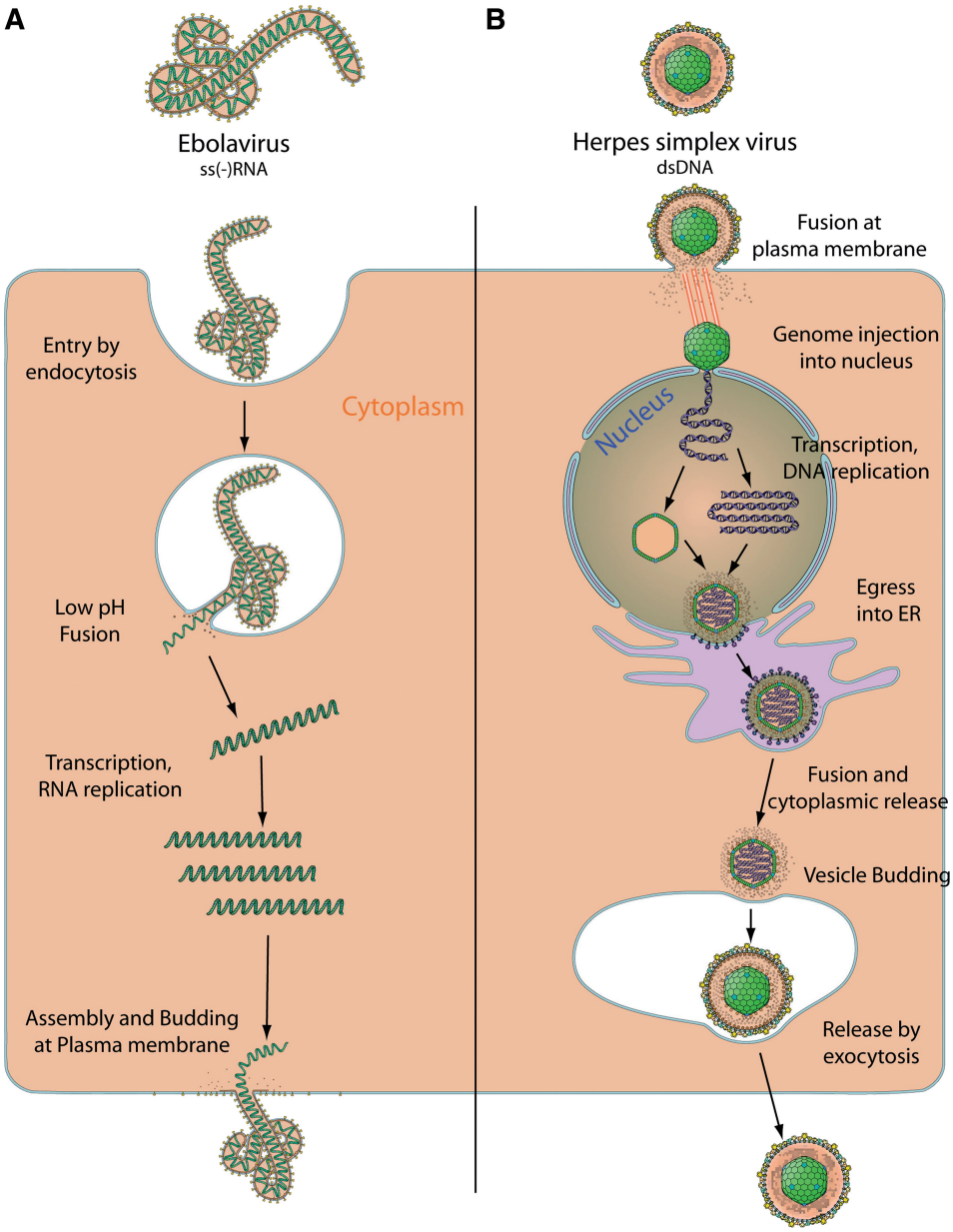

Figure 1. Diversity of viral replication: example of Ebolavirus versus Herpesvirus. (**A**) Ebolavirus is a negative single stranded RNA virus which replicates in the cytoplasm. It enters target cell by endocytosis, then penetrates into the cytoplasm by low pH fusion in the endosome. The viral RNA-dependent-RNA polymerase transcribes and replicates the viral genome in cell cytoplasm. Assembly and budding occur at the plasma membrane. (**B**) Herpes Simplex virus is a double stranded DNA virus which enters target cell by fusion at the plasma membrane, releasing the viral capsid in the cytoplasm. The capsid is transported to the nucleus where it injects the genomic DNA into the cell's nucleus. The viral genome circularises and conducts transcription and replication. New viral capsids are assembled in the nucleus, and bud into the endogenous reticulum (ER). These new virions fuse with ER membranes to release capsids into the cell cytoplasm. A second budding occurs at cell vesicle, and new virions are eventually released by exocytosis.

current virus classification comprises seven independent classes, according to the Baltimore system (4). This classification is based on the nature of the nucleic acids in the virion particle: dsDNA, ssDNA, dsRNA, ss(+)RNA, ss(−)RNA, ssRNA(RT) or ssDNA(RT).

Virus abundance on earth is higher than initially expected and recent studies have unveiled millions of

Table 1. The growing number of virus taxons

Year	Orders	Families	Genera	Reference
1995	2	54	184	ICTV 6th report (20)
2000	3	63	240	ICTV 7th report (9)
2005	3	71	282	ICTV 8th report (10)
2009	6	84	333	ICTV online, www.ictvonline.org

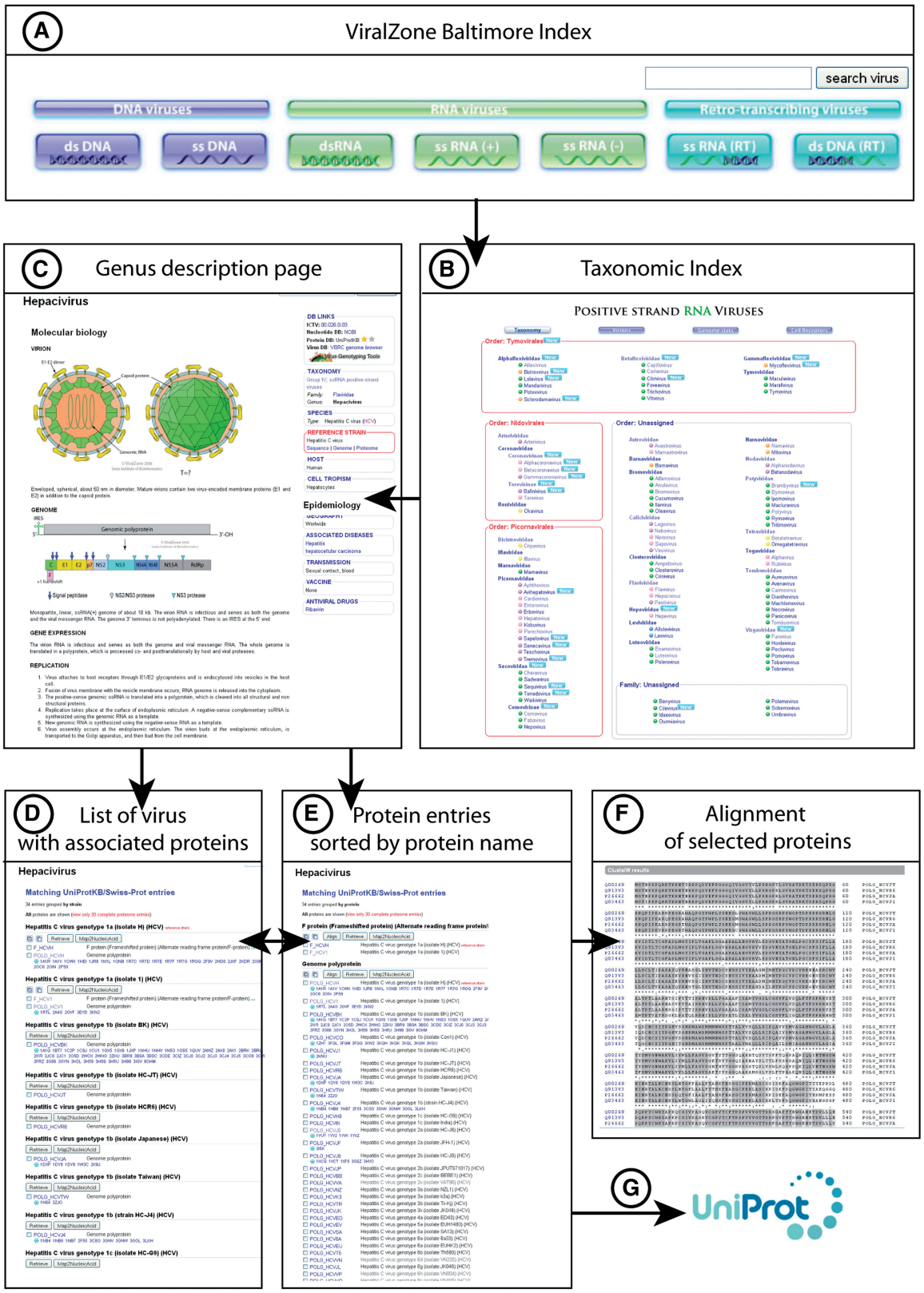

Figure 2. (**A**) ViralZone Baltimore Index. (**B**) Taxonomic index for ssRNA(+) viruses, classified by order, family then genus. Colour spots indicate the host infected by each virus genera: pink for human and other vertebrates, purple for non-human vertebrates, green for plants, yellow for invertebrates, orange for eukaryotic microorganism, and blue for prokaryotes; (**C**) Genus fact sheet. (**D**) List of viruses referenced in UniProtKB/Swiss-Prot along with correspondent protein entries displayed by default under the fact sheet. (**E**) The list of entries sorted by protein names. (**F**) Alignment obtained after selection of protein entries in (E). (**G**) Each Swiss-Prot protein entry gives a direct link to UniProt web site to access to the full details of protein annotation.

viruses per millilitre of seawater and billions per cubic centimeter in nearshore surface sediments (24,16); most of them are unidentified. As virus discovery accelerates, virus taxonomy has to be modified and completed each year (Table 1). In ViralZone, the starting point to access virus fact sheets are the seven Baltimore taxonomic pages (4) containing the whole list of known virus families and genera (Figure 2A). This list is reviewed each year as new viruses are constantly being described (8). The advantage of a website is that it can be incrementally updated while it can take years to publish new reference books. For example, the International Committee on Taxonomy of Viruses (ICTV) published important taxonomic changes on August 2009 on its website and the ViralZone taxonomy was updated accordingly only one month later.

From a public health point of view, providing comprehensive knowledge for all known virus genera turns out to be extremely useful when a new pathogen emerges out of a neglected virus family. A recent example is provided by the Xenotropic Moloney murine leukaemia virus-Related Virus (XMRV), which has recently raised the interest of the scientific community for its potential involvement in prostate cancer (22) and/or chronic fatigue syndrome (18). Since specific gammaretrovirus resources on the web were scarce, a direct consequence has been a dramatic increase in the number of hits to the corresponding ViralZone page (www.expasy.org/viralzone/all_by_species/67.html) that reached close to 2000 visitors in November–December 2009 (source: Google Analytics).

HOSTS

Virus host ranges can be quite narrow, e.g. the human hepatitis B virus which is strictly restricted to Human, or very large e.g. the rabies virus which seems to be able to successfully infect any mammal. Knowing the host tropism is essential to understanding the viral molecular biology. For example a dsDNA viral genome are transcribed differently in a bacteria or in a eukaryote. Moreover virus host range has a dramatic importance for public health, as illustrated by zoonosis like SARS, Ebola or Influenza that are caused by viruses able to mutate and cross hosts barriers, thus threatening the human population. For all these reasons the display of virus host tropism is highlighted in ViralZone. The hosts are indicated by a colour code for each virus genera on the taxonomy pages (Figure 2B). ViralZone display of hosts is restricted to the natural reservoirs. Vectors hosts, dead-end or laboratory hosts are not described here except if a human host/cell-line is involved.

Virus families can be browsed 'by host', allowing users to easily identify which viral families infect Humans, non-human vertebrates, plants, eukaryotic microorganisms, archaea or bacteria. A complete list of all major virus species able to infect humans is accessible through the ViralZone home page (www.expasy.org/viralzone/all_by_species/678.html).

GENUS AND FAMILY VIRUS FACT SHEETS

Virus fact sheets provide concise and specific information on molecular biology, taxonomy, hosts, and epidemiological data (Figure 2C). The first tab: 'General' describes molecular biology, virion and genome organization, followed by a step-by-step description of the viral cell infection cycle. The database section allows easy access to NCBI nucleotide and UniProtKB protein entries, as well as to specific virology databases such as VIPR (http://www.viprbrc.org) VIPERdb (7,23), Descriptions of Plant Viruses (1) and VBRC (www.vbrc.org, virology.ca). Host and cell tropism are generally indicated, but the latter might be absent since this kind of data can be unknown, or difficult to access in the literature. Cell receptor(s) for virus entry are also listed and links are provided to relevant publication(s). Finally epidemiological data briefly describe associated diseases and virus transmission as well as vaccines (if available) or antiviral drugs effective against this virus.

Under the 'Proteins by Strain' tab, strains and/or isolates are listed together with the proteins they encode. This list displays all related UniProtKB/Swiss-Prot entries (Figure 2D). These are manually curated entries with data extracted from publications. All the proteins annotated for a given virus strain or isolate are accessible at once for a given virus. Alternatively, the 'Proteins by Name' tab displays clusters of proteins having the same name and function (Figure 2E). This sorting is possible because the protein entries have been manually curated and have a coherent naming system. Calling the ClustalW alignment software (Figure 2F) directly from this page allows the user to align a set of these proteins to quickly generate a general alignment of any given viral protein family. Reference strain entries are clearly indicated, giving a landmark to users looking for the optimal data for a given viral protein family.

VIRION PICTURES

Virions are very diverse in shape and structure; they can be enveloped with one to several lipid bilayers or naked, and the genetic material can be protected by one, two or even three capsids showing helical or icosahedric symmetry and whose size is often related to genome length: from 17 nm (Porcine circovirus, 1.7 kb) to 400 nm in diameter (Mimivirus, 1200 kb). Virion pictures and diagrams can be found in Virology books or publications, but often with heterogeneous quality, colours and resolution. We created 160 original virion diagrams for ViralZone covering all known viral families and genera described to date. All the figures share the same concept and resolution, with defined colours for each part of the viral particle. Virions with icosahedric capsid symmetry are represented first as seen from a cross-section, then with 3D-like picture showing precise capsid architecture (Figure 3). Structural proteins are coloured in the same way in virion and genome pictures.

All these pictures are available to the scientific community, and freely accessible on the ViralZone web site.

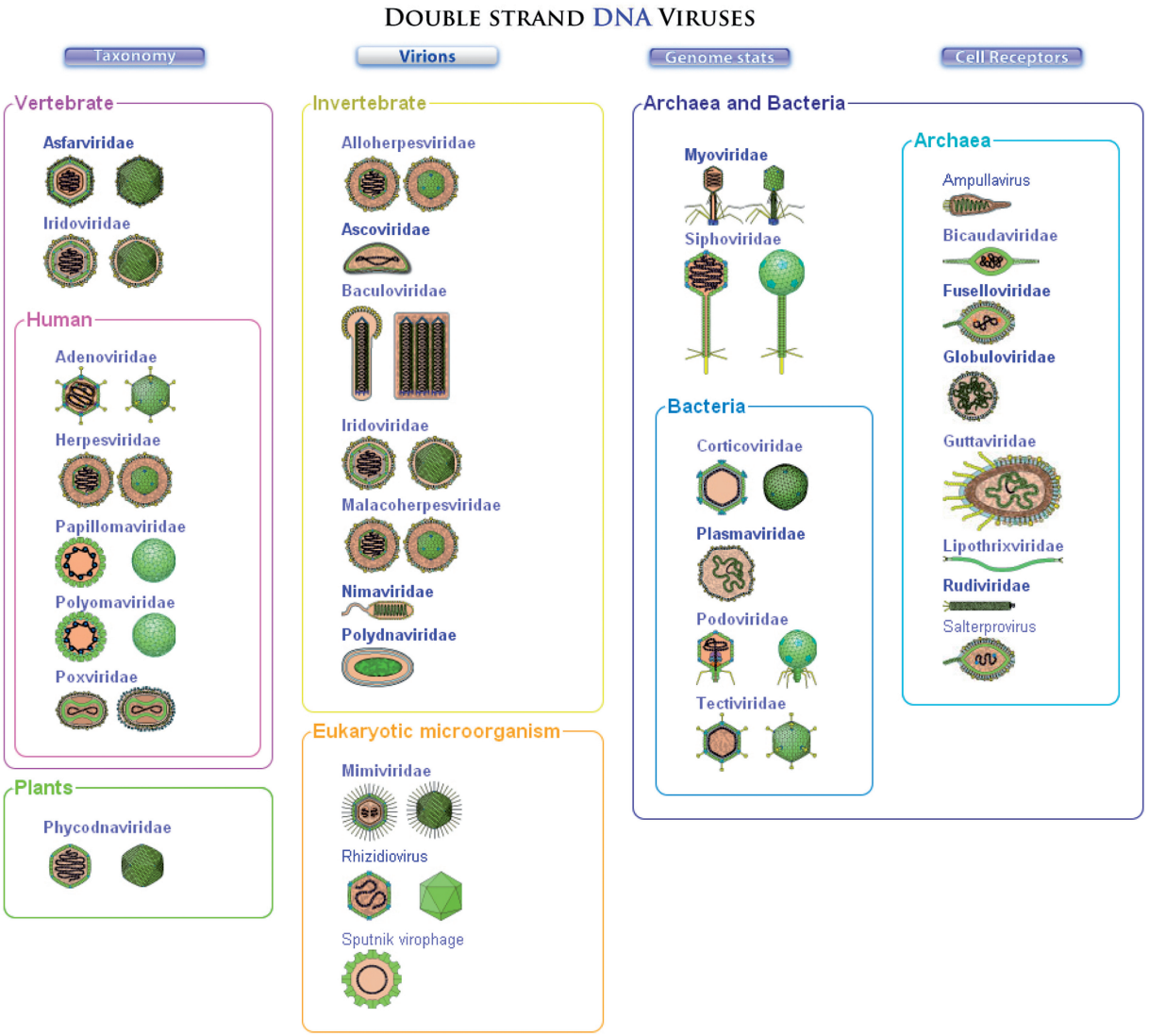

Figure 3. This page of ViralZone displays small virion picture for all dsDNA viruses (www.expasy.org/viralzone/all_by_species/748.html). Clicking on virus family or orphan genus name gives access to the virus description page, with full size picture of virion.

Table 2. Most represented species in UniProtKB (release 15.12)

Position	ORFs	Species	Prot/genome	Complete genome equivalents
1	313 532	Human immunodeficiency virus 1	9	34 329
2	113 396	Influenza A virus	11	10 309
3	95 799	Oryza sativa subsp. japonica (Rice)	40 577	2
4	77 225	Homo sapiens (Human)	20 500	4
...	...	...		...
6	74 067	Hepatitis C virus	2	37 034
...	...	...		...
17	34 367	Hepatitis B virus (HBV)	5	6873

REFERENCE STRAINS: COPING WITH THE EXPONENTIAL INCREASE OF VIRUS SEQUENCES

Virus genomes are relatively small, mostly <50 kb, and are therefore easy and relatively inexpensive to sequence.

This has resulted in an exponential increase in the number of new virus isolates deposited in the sequence databases. Of the 851 503 viral protein entries in UniProtKB, the species with the greatest number of open reading frames (ORF) deposited in UniProtKB is HIV-1 with 313 532 different ORFs, while the human proteome only accounts for 77 225 entries (Table 2) (UniProt release 15.12).

As manually annotating all UniProtKB viral proteins is not achievable, we selected about one reference strain (RefStrain) per genus to be fully curated. These RefStrains have been preferentially chosen in the genus type species, and belong to the NCBI Reference Sequences database (RefSeq) in which viral genomes have been manually reviewed (5). The 355 RefStrains selected account for 12 576 proteins, which are representative of the diversity of all virus genera and can be reasonably easily maintained in an annotated and updated form to reflect ongoing research. These RefStrains are now accessible through each ViralZone fact sheet which provides links to the corresponding RefSeq genome and UniProtKB virus proteome. RefStrains allow users to know which sequences to look at in order to have the best and most up-to-date information for any given virus, and can serve as templates to correctly annotate all similar viruses, an area of high interest to the bioinformatics community.

FUTURE DIRECTIONS

ViralZone is regularly updated with new information extracted from publications and scientific meetings abstracts. Users are also actively contributing by sending feedback, minor corrections and ideas to viralzone@isb-sib.ch. Future improvements will permit the further development of the viral molecular biology section, which will in turn be linked from fact sheets. Replication cycles are for the moment described in text format for all virus fact sheets but pictures would be more suited. An example of such picture is already accessible for the *Inoviridae* family (www.expasy.org/viralzone/all_by_species/675.html). The design of a virus specific controlled Gene Ontology (GO) that will both facilitate gene analysis and enhance data exchange between viral sequence databases is under way in collaboration with the GO consortium (13,27).

CONCLUSION

ViralZone is a freely accessible web resource that offers accurate and concise virus information for all known viruses. It displays high quality virion pictures available to all the scientific community. The site also functions as a hub for all scientists interested in virus knowledge, by bringing together virus metadata with genomic and protein sequence databases. Indeed the ViralZone web resource has already been cited as a source of data in several publications (3,17,21,28), dozens of thesis, many scientific web sites, and the virion figures are already widely used to support communication and teaching in Virology.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Swiss Federal Government through the Federal Office of Education; Science grant Swiss Institute of Bioinformatics (SIB) (www.isb-sib.ch). Funding for open access charge: Swiss Institute of Bioinformatics.

Conflict of interest statement. None declared.

REFERENCES

1. Adams,M.J. and Antoniw,J.F. (2006) DPVweb: a comprehensive database of plant and fungal virus genes and genomes. *Nucleic Acids Res.*, **34(Database issue)**, D382–D385.
2. Ascenzi,P., Bocedi,A., Heptonstall,J., Capobianchi,M.R., Di Caro,A., Mastrangelo,E., Bolognesi,M. and Ippolito,G. (2008) Ebolavirus and Marburgvirus: insight the Filoviridae family. *Mol. Aspects Med.*, **29**, 151–185.
3. Bahir,I., Fromer,M., Prat,Y. and Linial,M. (2009) Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Mol. Syst. Biol.*, **5**, 311.
4. Baltimore,D. (1971) Expression of animal virus genomes. *Bacteriol. Rev.*, **35**, 235–241.
5. 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.
6. Boehmer,P.E. and Nimonkar,A.V. (2003) Herpes virus replication. *IUBMB Life.*, **55**, 13–22.
7. Carrillo-Tripp,M., Shepherd,C.M., Borelli,I.A., Venkataraman,S., Lander,G., Natarajan,P., Johnson,J.E., Brooks,C.L. 3rd and Reddy,V.S. (2009) VIPERdb2: an enhanced and web API enabled relational database for structural virology. *Nucleic Acids Res.*, **37(Database issue)**, D436–D442.
8. Carstens,E.B. and Ball,L.A. (2009) Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses (2008). *Arch. Virol.*, **154**, 1181–1188.
9. Fauquet,C.M. and Mayo,M.A. (2001) The 7th ICTV report. *Arch. Virol.*, **146**, 189–194.
10. Fauquet,C.M., Mayo,M.A., Maniloff,J., Desselberger,U. and Ball,L.A. (2005) *Virus Taxonomy. Classification and Nomenclature of Viruses*, 8th ICTV Report, Academic Press, Elsevier.
11. Finsterbusch,T. and Mankertz,A. (2009) Porcine circoviruses–small but powerful. *Virus Res.*, **143**, 177–183.
12. Forterre,P. (2006) The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res.*, **117**, 5–16.
13. Gene Ontology Consortium. (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38(Database issue)**, D331–D335.
14. Glansdorff,N., Xu,Y. and Labedan,B. (2008) The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biol Direct.*, **3**, 29.
15. Koonin,E.V. (2009) On the origin of cells and viruses: primordial virus world scenario. *Ann. NY Acad. Sci.*, **1178**, 47–64.
16. Kristensen,D.M., Mushegian,A.R., Dolja,V.V. and Koonin,E.V. (2009) New dimensions of the virus world discovered through metagenomics. *Trends Microbiol.*, **18**, 11–19.
17. Liechti,R., Gleizes,A., Kuznetsov,D., Bougueleret,L., Le Mercier,P., Bairoch,A. and Xenarios,I. (2010) OpenFluDB, a database for human and animal influenza virus. *Database*, doi: 10.1093.

18. Lombardi,V.C., Ruscetti,F.W., Das Gupta,J., Pfost,M.A., Hagen,K.S., Peterson,D.L., Ruscetti,S.K., Bagni,R.K., Petrow-Sadowski,C., Gold,B. *et al*. (2009) Detection of an infectious retrovirus, XMRV, in blood cells of patients with chronic fatigue syndrome. *Science*, **326**, 585–589.
19. Macnaughton,T.B. and Lai,M.M. (2006) HDV RNA replication: ancient relic or primer? *Curr. Top Microbiol. Immunol.*, **307**, 25–45.
20. Murphy,F.A., Fauquet,C.M., Bishop,D.H.L., Ghabrial,S.A., Jarvis,A.W., Martelli,G.P., Mayo,M.A. and Summers,M.D. (1995) *Virus Taxonomy. Classification and Nomenclature of Viruses. Sixth Report of the International Committee on Taxonomy of Viruses*. Springer, Wien New York.
21. Saxena,S.K., Mishra,N., Saxena,R. and Saxena,S. (2009) Swine flu: influenza A/H1N1 2009: the unseen and unsaid. *Future Microbiol.*, **4**, 945–947.
22. Schlaberg,R., Choe,D.J., Brown,K.R., Thaker,H.M. and Singh,I.R. (2009) XMRV is present in malignant prostatic epithelium and is associated with prostate cancer, especially high-grade tumors. *Proc Natl Acad Sci USA*, **106**, 16351–16356.
23. Shepherd,C.M., Borelli,I.A., Lander,G., Natarajan,P., Siddavanahalli,V., Bajaj,C., Johnson,J.E., Brooks,C.L. 3rd and Reddy,V.S. (2006) VIPERdb: a relational database for structural virology. *Nucleic Acids Res.*, **34(Database issue)**, D386–D389.
24. Suttle,C.A. (2005) Viruses in the sea. *Nature*, **437**, 356–361.
25. Suttle,C.A. (2007) Marine viruses–major players in the global ecosystem. *Nat. Rev. Microbiol.*, **5**, 801–812.
26. Taylor,T.J., Brockman,M.A., McNamee,E.E. and Knipe,D.M. (2002) Herpes simplex virus. *Front Biosci.*, **7**, D752–D764.
27. The Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet*, **25**, 25–29.
28. Van Den Wollenberg,D.J., Van Den Hengel,S.K., Dautzenberg,I.J., Kranenburg,O. and Hoeben,R.C. (2009) Modification of mammalian reoviruses for use as oncolytic agents. *Expert Opin. Biol. Ther.*, **9**, 1509–1520.