Nucleic Acids Research Advance Access published October 23, 2008

Nucleic Acids Research, 2008, 1–5
doi:10.1093/nar/gkn729

The RNA Virus Database

Robert Belshaw[1,*], Tulio de Oliveira[1,2], Sidney Markowitz[3] and Andrew Rambaut[4]

[1]Department of Zoology, University of Oxford, Oxford OX1 3PS, UK, [2]South African National Bioinformatics Institute, University of the Western Cape, Cape Town 7535, South Africa, [3]Bioinformatics Institute, University of Auckland, Auckland 1142, New Zealand and [4]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, UK

Received August 12, 2008; Revised September 30, 2008; Accepted October 1, 2008

ABSTRACT

The RNA Virus Database is a database and web application describing the genome organization and providing analytical tools for the 938 known species of RNA virus. It can identify submitted nucleotide sequences, can place them into multiple whole-genome alignments (in species where more than one isolate has been fully sequenced) and contains translated genome sequences for all species. It has been created for two main purposes: to facilitate the comparative analysis of RNA viruses and to become a hub for other, more specialised virus Web sites. It is available at the following four mirrored sites.

- **http://virus.zoo.ox.ac.uk/rnavirusdb**
- **http://hivweb.sanbi.ac.za/rnavirusdb**
- **http://bioinf.cs.auckland.ac.nz/rnavirusdb**
- **http://tree.bio.ed.ac.uk/rnavirusdb**

INTRODUCTION

Viruses are divided into two similar-sized groups depending on whether the virus particle contains DNA or RNA, and, as causes of human fatality, RNA viruses are by far the more important (1). New viral diseases continue to appear as a result of several changes in human activity: travel, population growth, interaction with wild habitats etc. Well-known novel, or emergent, RNA diseases include severe acute respiratory syndrome (SARS) (2), human immunodeficiency virus 1 (HIV-1) (3), and may come to include Avian influenza H5N1 virus (4). These emergent diseases are an important factor behind the increase in the number of genome sequences that NCBI treats as representing new species (Figure 1). In 2005, more than 200 new virus species were submitted to GenBank (more recent dates are less reliable because there is typically a delay between submission and public availability). As more emergent viruses appear, it is important to have a site that allows their genomes to be compared to those of known viruses. The origin of most major infectious diseases is unknown because of our ignorance of the diversity of pathogens in wild animals. This restricts our ability to both predict risks and develop treatments (5).

Despite some advances (6,7), the evolutionary history of RNA viruses is in general poorly known, especially the deep phylogenetic relationships between virus families (8,9). We believe that one of the reasons for this is a lack of easily available translated genes and genomes for all species, and the lack of aligned genome sequences representing different isolates of the same species.

In addition to the need to facilitate greater comparative analysis of RNA viruses is the need to link together the existing virus Web sites and their underlying databases. There are many Web sites that provide genomic data, tools for genetic analysis and/or biological information for some viruses (see 'Links' on our site home page). The RNA Virus Database is intended to complement these other sites by providing basic genomic information and tools for all RNA viruses and linking the user to more specialist sites, where they exist, e.g. for HIV-1 and hepatitis C virus (HCV), we provide links to sites such as the Los Alamos Laboratory on the main page for each of these viruses (find by typing HIV-1 or HCV into the search window in the top toolbar). For such viruses, we do not duplicate the work of other groups by attempting to display the available diversity of genomes. We intend that the RNA Virus Database should develop further as a hub for other sites and we therefore encourage other workers to contact us with details of their sites that they wish linked to ours. Also, we encourage workers to 'adopt a virus' and improve and/or expand the information that we provide for individual species. This can be done by emailing us or getting involved directly in developing the database, which is an Open Source project available at our GoogleCode site (http://code.google.com/p/rnavirusdb).

Some of the data and tools on the RNA Virus Database can be found elsewhere, but not all of them can, e.g. NCBI's Genome site provides genomic overviews of virus species and pairwise alignments of other isolates to the reference sequence, but it does not provide multiple alignments or complete translated genomes as we do. Similarly, its general Entrez site provides pair-wise alignments of the query sequence and similar sequences in the

*To whom correspondence should be addressed. Tel: +44 1865 281997; Fax: +44 1865 310447; Email: robert.belshaw@zoo.ox.ac.uk

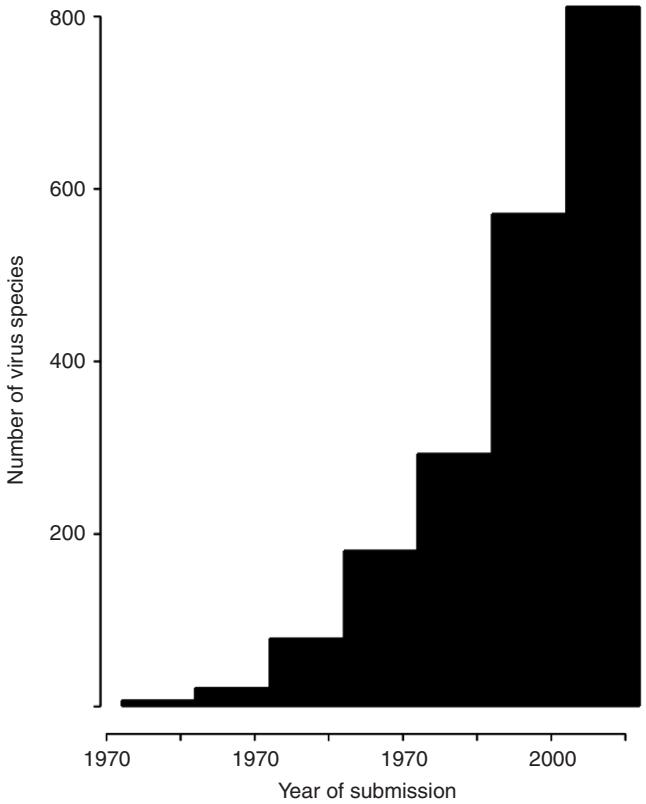

Figure 1. Submission of new virus species to GenBank between 1970 and 2006. Dates are the earliest given in the accession (either of submission or publication). Submissions after 2006 are excluded because accessions are made public typically only following publication and thus the frequency of submissions is more recent time periods is underestimated.

database, plus a phylogenetic tree calculated from those distances; however, no multiple alignment is built. We also corrected the (few) errors in the GenBank entries, and our database records features such as RNA editing (10) that make genome translation problematic.

We have, therefore, created the RNA Virus Database as a user-friendly site devoted to RNA viruses, providing essential genomic data and tools (discussed in more detail below) and links to the other virus Web sites. The three main features are as follows.

- Provide multiple whole-genome alignments, gene and whole-genome translations for all RNA virus species
- Identification and taxonomic searching facility
- Guidance to other web resources.

WHOLE-GENOME ALIGNMENTS

We provide multiple alignments of whole genomes (as nucleotides) for all species where GenBank contains multiple representatives. Our database, thus, currently has multiple alignments for approximately half the species (available from the main page of any virus species under 'Download alignment'). These alignments were made by downloading from GenBank all complete (or near-complete) genomes using the BioPerl GenBank modules (11). Accidental mismatches were excluded by performing a preliminary alignment using BlastAlign (12), which is designed to cope with non- or poorly homologous sequences and reports such matches. The sequences that showed clear homology to the NCBI reference sequences (we used a cutoff of a maximum of 40% of positions being represented by gaps in the BlastAlign alignment), up to a maximum of 50, were then aligned using ClustalW (default parameter values) (13). For species for which we have at least three sequences, a neighbor-joining tree was then constructed using PAUP (with HKY-adjusted genetic distances) (14), and this tree is displayed both as a pdf [via the TreeGraph program (15)] and using FigTree, which is a new Java-based tree-drawing application created by one of us (Rambaut,A., unpublished data). FigTree will also display strain and isolate information as well as accession numbers.

IDENTIFY RELATIONSHIPS OF SUBMITTED SEQUENCES

Our site allows virus nucleotide or amino acid sequences submitted by the user to be identified or, if the query is a new species, its closest relative to be found. In addition, the genomic location of any matched region of the library sequences is shown. For this we use NCBI's suite of BLAST programs (16) (go to the 'BLAST' link on the toolbar of the home page). Once the most closely related reference species has been located, the query sequence can then be placed into a whole-genome multiple alignment for that species (where such an alignment is present) in order to show the query's phylogenetic relationships to the genomes in our database (go to 'Align your sequence' on the virus species main page). An example of this process is illustrated in Figure 2. Two procedures are available here for building the new multiple alignment. (i) A BLAST of the query to the reference species sequence provides coordinates from the resulting pair-wise alignment. These coordinates are then used to select homologous regions from the reference multiple alignment, and a new multiple alignment is then built using ClustalW along with a phylogenetic tree using PAUP as described above. (ii) BlastAlign (described above) is used to generate a new multiple alignment using the query sequence plus the sequences from the reference multiple alignment.

Another database table that is available as part of the download, 'isolates', includes the biological data of the isolates used in the whole-genome multiple alignments [except for HIV-1, HCV and hepatitis B virus (HBV), where we used manually built multiple alignments; for these three species, accession numbers for the isolates are given in the Supplementary Data as AccessionHIVHCVHBV.xls].

TRANSLATED GENES AND GENOMES

We provide amino acid sequences for all virus genes (or, more strictly, for all Open Reading Frames) plus complete translated genomes for each virus species (go to 'Proteins' on the toolbar of any virus species main page).

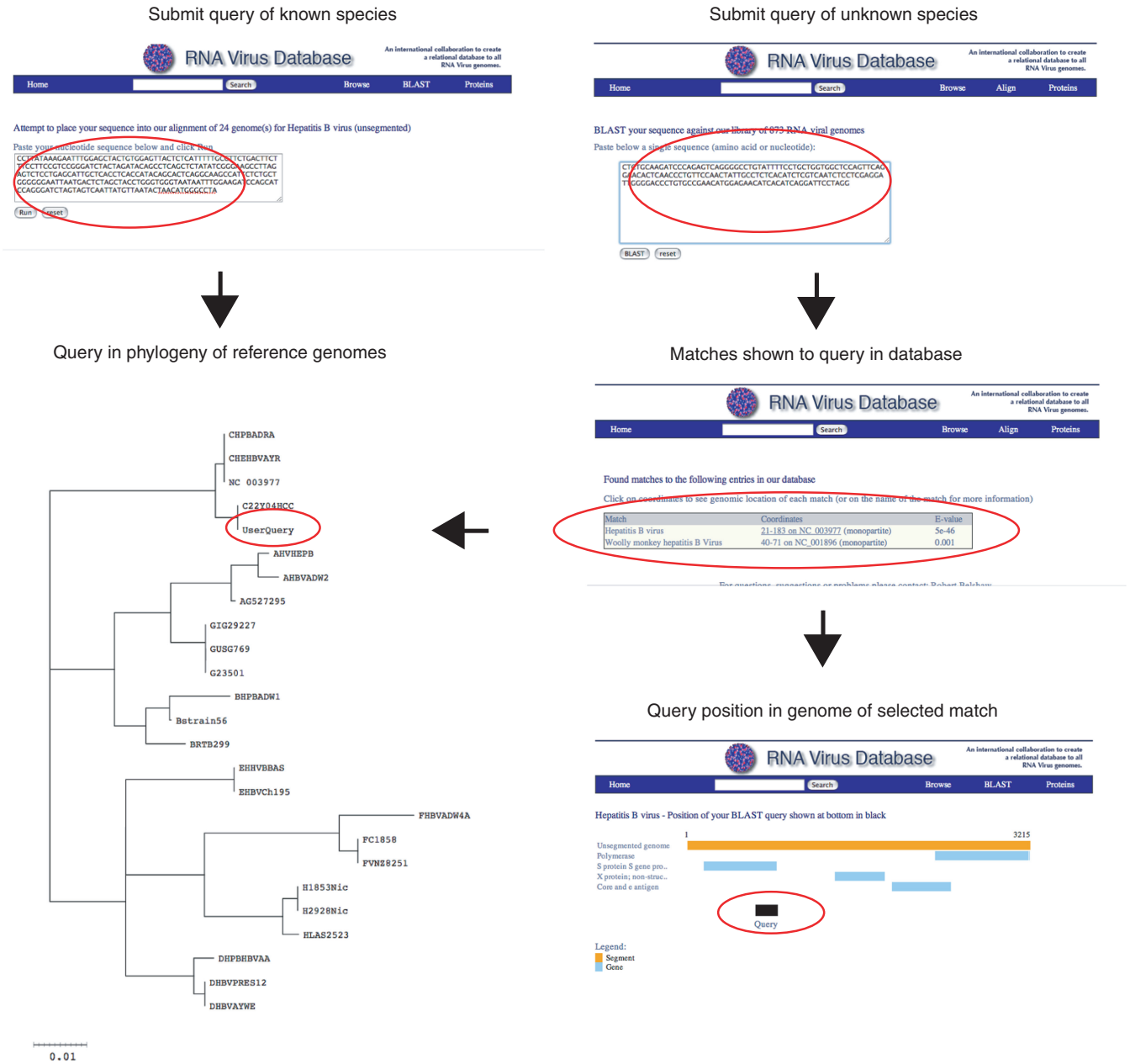

Figure 2. Screenshots illustrating use of the RNA Virus Database to investigate a submitted virus nucleotide sequence.

The translated genomes are intended to facilitate phylogenetic analysis of more distantly related viruses (9). One feature that makes annotation of RNA viruses difficult is that most species have some gene overlap (17), i.e. where the same nucleotides code for two different genes by being read in two different frames. We, therefore, allow the user to select from three possible options for dealing with this feature: (i) have overlapped regions excised from the translated genome, (ii) have one only of any overlapped amino acid sequences represented or (iii) have all the amino acids sequences present, with overlapped sequences placed sequentially (and thus the nucleotides represented twice).

Using a key word search of the GenBank entries, and a standard reference work (18) where this did not reveal a match, we have placed most of the genes into functional groups (go to 'Proteins' on the toolbar of the home page).

DATABASE STRUCTURE AND DATA SOURCES

The RNA Virus Database is a PHP web application on top of a mySQL database. The underlying database has eight tables linked as shown in Figure 3. PHP is available at http://www.php.net, but should come pre-installed on UNIX machines. MySQL is available free from http://www.mysql.com. All data have been taken from the NCBI's Genome and Nucleotide sites at the following two URLs.

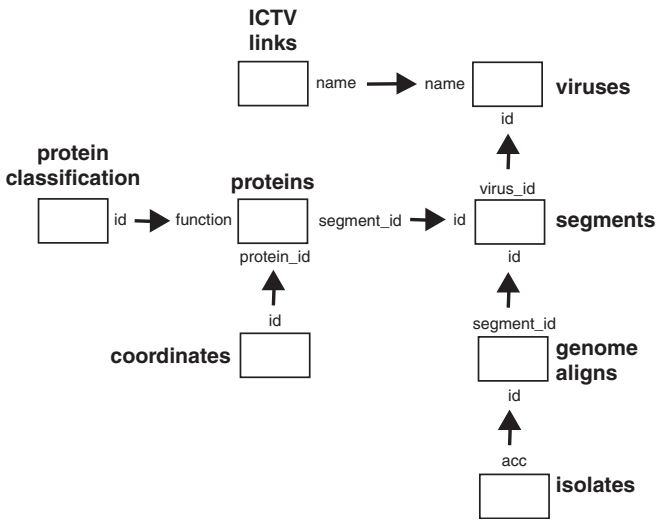

Figure 3. Relationships of tables in the underlying mySQL database. Table names are in large bold font and interlinking column names are in small regular font.

http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/10239.html.
http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide

Species names, nucleotide sequences and accession numbers were downloaded directly from GenBank using BioPerl modules, while further details of the virus—gene coordinates, taxonomic affinities etc.—were subsequently extracted from the flatfile of all GenBank entries that can be downloaded from the NCBI Genome site (19). Our approach was to treat all entries in the NCBI Virus Genome sites as species and to follow their taxonomic classification, although we only give virus type (e.g. single-stranded positive-sense, retrotranscribing), family and genus at our site. We, therefore, follow NCBI's inclusion of hepadnaviruses, which include HBV, and caulimoviruses among the RNA viruses despite their mature virion containing DNA rather than RNA (their possession of reverse transcriptase clearly places them biologically and evolutionarily among the reverse-transcribing group of RNA viruses).

AVAILABILITY, FUTURE EXTENSIONS AND UPDATES

The URLs of our mirrors are given in Abstract. The PHP scripts can be accessed using subversion (http://subversion.tigris.org/) from our GoogleCode site at http://code.google.com/p/rnavirusdb. The mySQL database (a gzipped 16 Mb dump) may also be downloaded from the same site for installation on the user's computer if required. A README with installation instructions is present among the PHP scripts. If required, the database can subsequently be updated by other users following instructions and Perl scripts given in the Supplementary Material (perl_scripts.tar). This updating involves a series of short intervening manual steps (we find that complete automation of such processes is inefficient).

We intend to update the database on at least a 6-monthly basis in order to include newly discovered viruses, and are currently working to incorporate biological and epidemiological data. We also intend to release soon a DNA virus database using the same format. As discussed in Introduction, we are keen to collaborate with other groups over further developments of our database.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are also very grateful to Oli Pybus and Alexei Drummond for their help.

FUNDING

Wellcome Trust (to R.B.); EU Marie Curie Fellowship scheme (to T.d.O.); the Royal Society (to A.R.).

Conflict of interest statement. None declared.

REFERENCES

1. Belshaw,R., Gardner,A., Rambaut,A. and Pybus,O.G. (2008) Pacing a small cage: mutation and RNA viruses. *Trends Ecol. Evol.*, **23**, 188–193.
2. Sleigh,A. (2005) Twenty-first century plague: the story of SARS. *Nature*, **435**, 886–887.
3. Rambaut,A., Posada,D., Crandall,K.A. and Holmes,E.C. (2004) The causes and consequences of HIV evolution. *Nat. Rev. Genet.*, **5**, 52–61.
4. Kuiken,T., Holmes,E.C., McCauley,J., Rimmelzwaan,G.F., Williams,C.S. and Grenfell,B.T. (2006) Host species barriers to influenza virus infections. *Science*, **312**, 394–397.
5. Wolfe,N.D., Dunavan,C.P. and Diamond,J. (2007) Origins of major human infectious diseases. *Nature*, **447**, 279–283.
6. Gorbalenya,A.E. (2001) Big nidovirus genome—when count and order of domains matter. *Adv. Exp. Med. Biol.*, **494**, 1–17.
7. Le Gall,O., Christian,P., Fauquet,C.M., King,A.M.Q., Knowles,N.J., Nakashima,N., Stanway,G. and Gorbalenya,A.E. (2008) Picornavirales, a proposed order of positive-sense single-stranded RNA viruses with a pseudo-T=3 virion architecture. *Arch. Virol.*, **153**, 715–727.
8. Koonin,E.V. and Dolja,V.V. (1993) Evolution and taxonomy of positive-strand RNA vruses—implications of comparative-analysis of amino acid sequences. *Crit. Rev. Biochem. Mol. Biol.*, **28**, 375–430.
9. Zanotto,P.M.D., Gibbs,M.J., Gould,E.A. and Holmes,E.C. (1996) A reevaluation of the higher taxonomy of viruses based on RNA polymerases. *J. Virol.*, **70**, 6083–6096.
10. Hausmann,S., Garcin,D., Delenda,C. and Kolakofsky,D. (1999) The versatility of paramyxovirus RNA polymerase stuttering. *J. Virol.*, **73**, 5568–5576.
11. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G.R., Korf,I., Lapp,H. *et al.* (2002) The bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
12. Belshaw,R. and Katzourakis,A. (2005) BlastAlign: a program that uses blast to align problematic nucleotide sequences. *Bioinformatics*, **21**, 122–123.
13. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) Clustal-W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

14. Swofford, D. (1998) *PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods) Version 4*. Sinauer, Sunderland, MA.
15. Mueller,J. and Mueller,K. (2004) TREEGRAPH: automated drawing of complex tree figures using an extensible tree description format. *Mol. Ecol. Notes*, **4**, 786–788.
16. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol.*, **215**, 403–410.
17. Belshaw,R., Pybus,O.G. and Rambaut,A. (2007) The evolution of genome compression and genomic novelty in RNA viruses. *Genome Res.*, **17**, 1496–1504.
18. van Regenmortel,M.H.V., Fauquet,C.M., Bishop,D.H.L., Carstens,E.B., Estes,M.K., Lemon,S.M., Maniloff,J., Mayo,M.A., McGeoch,D.J., Pringle,C.R. *et al.* (2000) *Virus Taxonomy: Classification and Nomenclature of Viruses*. Academic Press, San Diego, CA.
19. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.