



Viral Genome Resources at the NCBI

National Center for Biotechnology Information ■ National Library of Medicine ■ National Institutes of Health ■ Department of Health and Human Services

At the National Center for Biotechnology Information (NCBI), biological discoveries begin with Entrez, a text-based search and retrieval system which can be used to search NCBI's growing collection of linked databases. Currently, there are 33 Entrez databases, which can either be searched independently or simultaneously with **Entrez Global Query**
<http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi>

Using either of the following Entrez queries on the Global Query page, viruses [organism] or viroids [organism], it is possible to quickly obtain information about the distribution of organism-specific data in each of the Entrez databases, and to link to specific subsets of those data. Shown below are the number of viral- and viroid-specific records in Entrez on July 12, 2007:

Entrez Query: viruses[organism]				Entrez Query: viroids[organism]	
Database	# of Records	Database	# of Records	Database	# of Records
Protein	655,833	Popset	5,260	Nucleotide	1,428
Nucleotide	518,527	Structure	2,860	PubMed Central	116
Gene	55,775	Genome Sequences	2,835	Taxonomy	76
PubMed Central	46,264	Domains	917	Genome Sequences	36
Taxonomy	25,869	GEO Datasets	27	Popset	20
GEO Expressions	12,303	Genome Projects	2	Gene	3
3D Domains	12,190	UniSTS	1		

A brief summary of some of these databases and their content is provided here.

Entrez Genome

www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html

Entrez Genome is a database which contains genomic data from organisms ranging from viruses, bacteria and fungi to plants and mammals. The data in Entrez Genome represents a subset of the sequences in NCBI's Reference Sequence (RefSeq) database www.ncbi.nlm.nih.gov/RefSeq/), a collection of reference sequence standards for genomic, transcript, protein, and non-coding RNA sequences. The data in Entrez Genome is linked to other Entrez databases and resources and provide scientists with a solid foundation for genomic analysis.

Currently, there are 2825 Reference sequences for 1867 viral genomes and 36 Reference Sequences for viroids. In addition, there are 9300 other complete viral sequences and 222 complete viroid sequences linked to their corresponding reference sequences. The entire collection of virus and viroid reference genomes can be retrieved from Entrez Genome using the Entrez query

viruses [organism] OR viroids [organism]

Each reference genome record is identified by an accession number which begins with the prefix NC_ and is followed by a series of 6 digits. The records are maintained by NCBI, and always provide up to date information.

The viral reference genomes are also available for download via the RefSeq FTP site, <ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/>

Entrez Gene

www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

Entrez Gene is a curated database of descriptive information about genetic loci in more than 4,400 taxa. Currently, almost 1,900 different viral taxa are represented in Entrez Gene, and these taxa are shown in the table to the right.

The database can be searched using the name of an organism or taxa, accession number, gene name or symbol, or a PubMed identifier (PMID). Depending on the organism and the availability of genome data, Entrez Gene records may contain information about the genomic structure and function of the gene, Reference Sequences, GenBank records and links to other Entrez databases, including Genome, Nucleotide, Protein, Conserved Domains, Taxonomy, and PubMed/PubMed Central.

The scientific community may contribute to NCBI's gene annotation efforts by submitting information which links the publications in PubMed to gene function.

Taxon	# of Viral Taxa in Entrez Gene
ssRNA	650
dsDNA viruses, no RNA stage	597
ssDNA	344
dsRNA	102
Retro-transcribing	95
Satellites	87
unclassified phages	20
unclassified archaeal	5
unclassified viruses	5
Deltavirus	1
Eggplant latent viroid (Avsunviroidae)	1

To submit a Gene Reference into Function (GeneRIF) to NCBI, see <http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html>

Entrez Genome Project

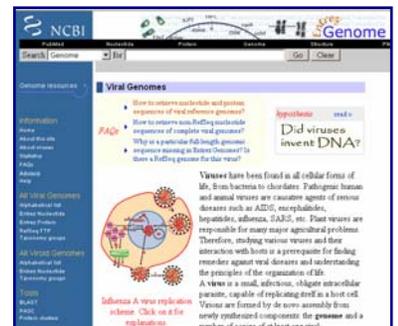
www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj

Entrez Genome Project provides organism-specific overviews of complete and in-progress large-scale sequencing, assembly, annotation, and mapping projects. The Viral Genome Project pages are currently under development. When available, these pages will provide information about the links to the project data, the institution doing the sequencing, publications, and other NCBI databases and tools, including BLAST.

Viral Genomes Home Page

<http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html>

The Viral Genomes home page is the central point of access to viral-specific data and data analysis tools. This page provides links to all available viral genomes, which are listed alphabetically and also organized by families or taxa. Educational information, including a short introduction to viruses and a scheme of Influenza A virus replication, are also provided.

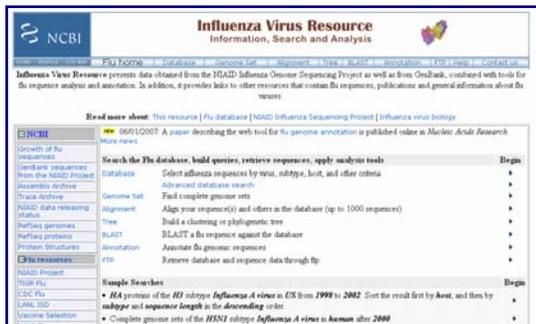


Influenza Virus Genome Resources and Tools

www.ncbi.nlm.nih.gov/genomes/FLU/

The viral genome sequences generated from the Influenza Genome Sequence Project are deposited in GenBank, and this data is the main component of the Influenza Virus Resource.

Here it is possible to construct queries, retrieve sequences, find complete genome sets, BLAST a flu sequence against the database, construct multiple sequence alignments, and build clustering or phylogenetic trees. In addition to the sequence data, this resource provides an Influenza Virus Sequence Annotation Tool, as well as links to other flu-related NCBI resources, including the Trace Archive, Assembly Archive, Reference Sequences, Viral Structures, and others. Links to recent publications on flu research and flu sequence updates in GenBank are also provided.



The Influenza Virus Sequence Annotation Tool

<http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/annotation.cgi>

The Influenza Virus Sequence Annotation Tool, or FLAN (for FLU ANnotation), is a web application for user-provided Influenza A and Influenza B virus sequences. It can predict protein sequences encoded by a flu sequence and produce a feature table that can be used for making a sequence submission to GenBank.

Trace Assembly Archive

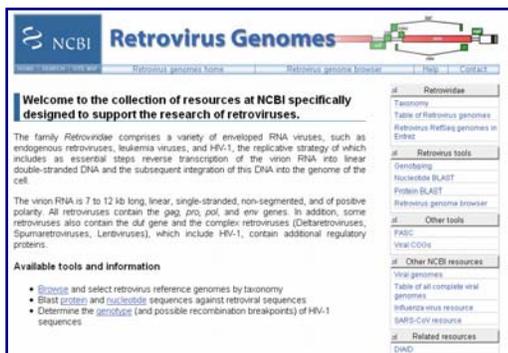
www.ncbi.nlm.nih.gov/Traces/assembly/assmbrowser.cgi

The NCBI Trace Assembly Archive links the raw sequence information found in the Trace Archive with assembly information found in publicly available sequence repositories, such as GenBank, EMBL and DDBJ. As of July 12, 2007, the Assembly Archive contained 1,745 assembled Influenza A viral genomes. Trace assemblies within the Assembly Archive may be examined in detail using the "Assembly Viewer", which enables viewing of multiple sequence alignments as well as the actual sequence chromatograms.

Retrovirus Genome Resources and Tools

<http://www.ncbi.nlm.nih.gov/retroviruses/>

The Retrovirus Genome Resource provides convenient access to retroviral RefSeq genomes, other retroviral genomes, publications linked to these genomes and links to other NCBI and selected non-NCBI tools and databases.



Viral Genotyping Tool

<http://www.ncbi.nlm.nih.gov/projects/genotyping/formpage.cgi>

This web-based tool can be used to identify the genotype of HIV, HTLV, Hepatitis B and C viruses and polioviruses. The Viral Genotyping Tool uses BLAST to compare the query sequence to a set of reference sequences of known viral genotypes. This approach is especially useful for the detection and analysis of recombinant sequences.

Pairwise Sequence Comparison (PASC) Tool

<http://www.ncbi.nlm.nih.gov/sutils/pasc/viridty.cgi?textpage=overview>

PASC is a web tool for analyzing the distribution of pair-wise identities between the completed genomes of several viral families, and serves as a good reference for taxonomic classification based on sequence similarities. The distribution of percent similarities of resulting from each pair-wise global alignment is graphically represented in the form of a histogram. The number of pairs at each percentage is plotted, and the distribution forms characteristic peaks for viruses at the isolate, species, genus, and subfamily levels.

Clusters of Related Viral Proteins or VOGS

<http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/vog.html>

Proteins from completed viral genomes are clustered by sequence similarity based on BLAST pair-wise alignments. When mapped back to viral genomes, this information supports evolutionary and functional studies, and can also be used to improve the annotation of the reference genome sequences. These groups of related viral proteins will soon also become available in Entrez Protein Clusters.

SARS Coronavirus Resources

<http://www.ncbi.nlm.nih.gov/genomes/SARS/SARS.html>

The SARS Coronavirus Resource page provides access to data and information relevant to the SARS Coronavirus. This page includes links to the most recent sequence data and publications, results of pre-computed sequence analysis, including genomic, protein, and 3D structures, and links to selected external SARS Coronavirus-related resources.

Educational Resources: Entrez Bookshelf

<http://www.ncbi.nlm.nih.gov/sites/entrez?db=Books>

The Bookshelf database is an on-line collection of biomedical books whose content can be searched from within Entrez. Currently there are more than 90 books on the Bookshelf including:

- *Retroviruses (1997); Cold Spring Harbor Laboratory Press*
- *Antiretroviral Resistance in Clinical Practice (2006); Mediscript Ltd.*
- *Vaccines: Chapter 6, Smallpox and Vaccinia (1999); W.B. Saunders & Co.*
- *Immunology (2001); Garland Science*
- *Molecular Biology of the Cell (2002); Garland Science*

Announce lists

http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html

<http://www.ncbi.nlm.nih.gov/feed/>

To be informed of any changes and updates to these and other resources at the NCBI, subscribe to one of NCBI's resource-specific email announce lists or RSS feeds.

For more information

To obtain more information about any of our databases, services, or programs, contact the NCBI Help Desk by email at info@ncbi.nlm.nih.gov, or by calling 301-496-2475, 8.30 am-5.30 pm (EST), Mon-Fri.