



## NCBI News, September 2014

### NCBI Sequence Viewer version 3.4 available

*Tuesday, September 30, 2014*

NCBI Sequence Viewer has recently been updated and now has improved visualization of graphs, sequence track and other text, as well as a reworked configuration dialog. A full list of new features, improvements and fixes is included in the [release notes](#).

Sequence Viewer provides a graphical view of sequences and color-coded annotations on regions of sequences stored in the Nucleotide and Protein databases.

### HIV-1, human interaction database updated

*Monday, September 29, 2014*

The HIV-1, human interaction database has been updated and is now on an improved [page](#). The improved interface includes help documentation and supports structured queries against [Gene](#), as well as browsing, filtering and downloading the protein and replication interaction data sets. The most recent data release (June 2014) includes 12,785 HIV-1, human protein-protein interactions for 3,142 human genes and 1,316 replication interactions for 1,250 human genes.

NCBI Resources How To

## Retroviruses

HIV-1 Human Interaction Database Browse About Help Publications Releases

### HIV-1 Human Interaction Database

- [About the database](#)
- [Help](#)
- [Publications](#)
- [Releases](#)

### Browse and Download Data

Protein and Replication Interactions

### Search NCBI Gene Records With HIV-1 Interaction Data

Searching human genes with all selected criteria. For HIV-1 genes, [click here](#).

Search criteria

Interaction type  Protein  Replication  Both

Gene Ontology (GO)

Protein domain name

Properties  Has phenotype  Has gene expression data  Has Homologene Cluster  Has >1 RefSeq transcript  Has biological pathways

Entrez Gene keywords

[Clear form](#)

**Figure 1.** The HIV-1 interactions database homepage.

The HIV-1, human interactions project collates published reports of two types of interactions: HIV-1, human protein interactions, and human gene knock-downs that affect virus replication which are reported as “replication interactions”.

# Virus Variation Resource pages for Ebolavirus, MERS coronavirus give quick and easy access to related sequences and other data

Friday, September 19, 2014

NCBI has created resource pages for [Ebolavirus](#) and [MERS coronavirus](#), giving users an easy way to find all sequences related to these pathogens. These pages aggregate links to virus data at NCBI and also provide important links out to other information at the CDC, WHO, and HealthMap.

**Ebolavirus Resource**

Ebolavirus human hemorrhagic fever/disease. Please see "Ebolavirus links" to the right for more information.

**Ebolavirus Resource Components**

The [Ebolavirus database](#) can be used to search and retrieve MERS genome and protein sequences based on standardized biological criteria.

The [Zaire ebolavirus reference genome](#) graphical display supports several interactive functions as described in the "How to use" section below.

**Ebolavirus database**  
 Search ebolavirus sequences  
 Help

**Other NCBI Ebolavirus Resources**  
 Ebolavirus publications  
 Ebolavirus genome browser  
 Ebolavirus taxonomy

**Ebolavirus links**  
 Health Map  
 CDC  
 WHO

**How to use the graphical viewer**

The [NCBI Graphical Sequence Viewer](#) displays sequence annotations using colored bars: **red bars** = gene features; **green bars** = coding regions; **black bars** = other sequence features such as mature peptides and conserved domains.

To learn more information about a given annotation feature, hover your mouse above it. A small pop-up window will appear that includes information about the feature, download options, and links to other NCBI databases.

Left click the tool icon to access a number of useful functions including genome sequence download, BLAST search and primer search.

Please, find detailed information about the use of this graphical display [here](#).

**Figure 1.** Ebolavirus Resource page. The main column gives a brief description of the resource and displays the NCBI Graphical Sequence Viewer. In the right column, from top to bottom, are links to: the [Ebolavirus database](#), other NCBI Ebolavirus resources, and links out to HealthMap, CDC, and WHO. The [MERS coronavirus resource page](#) has the same layout and links.

The Virus Variation resource pages all include a description of the resource components: the database, the reference genome graphical display and links to other resources, both external and within NCBI.

Dedicated Virus Variation databases for [Ebola virus](#) and [MERS coronavirus](#) have been developed. These databases allow searching for nucleotide and protein sequences by a variety of criteria including host, sequence patterns, region or country of isolation, and collection or release dates. The databases allow you to:

- Quickly find the sequences you need, through an intuitive search interface for all viral sequences using standardized protein/gene names and metadata
- Select the latest sequences based on date criteria or sorting of results
- Download sequences in many formats or find links to sequences in NCBI databases.

Visit the [Virus Variation homepage](#) to see resource pages for other viruses.

## Simplified FASTA headers included on new NCBI Genomes FTP site

*Wednesday, September 17, 2014*

Last month, a major revision of the [NCBI Genomes FTP site](#) was [announced](#). In response to user feedback, a new format for FASTA headers of genome, protein and transcript records has been implemented. This new format is limited to records in the `/all/`, `/refseq/`, and `/genbank/` directories on the new Genomes FTP site and does not affect the Nucleotide database web FASTA displays.

Now, instead of "`>gi|xx|dbsrc|accession.version|description`", the new format is simply "`>accession.version description`".

For example, the header on the FASTA record for Homo sapiens chromosome 1 was previously:

```
>gi|568336023|gb|CM000663.2| Homo sapiens chromosome 1, GRCh38 reference primary assembly.
```

On the new Genomes FTP site, the header is now:

```
>CM000663.2 Homo sapiens chromosome 1, GRCh38 reference primary assembly.
```

NCBI has traditionally used a compound FASTA sequence identifier string in which multiple IDs were separated by “|” characters. This format provides more information, but requires that the individual sequence identifiers be parsed out of the compound string. The simpler sequence identifier string is identical to that used in the GFF annotation files on the genomes FTP site. Providing sequence and annotation files with matching sequence identifiers supports their use in commonly used RNA-Seq analysis packages and in other analysis pipelines that rely on simple string comparison to match sequence identifiers.

More information about the revised Genomes FTP site, including the new FASTA header format, is available on the [Genomes Download FAQ page](#).

## RefSeq release 67 available on FTP

*Thursday, September 11, 2014*

The full [RefSeq release 67](#) is now available on the FTP site with over 61 million records describing 45,166,402 proteins, 8,163,775 RNAs, and sequences from 41,913 different NCBI TaxIDs.

More details about the RefSeq release 67 are included in the [release statistics](#) and [release notes](#). In addition, reports indicating the [accessions included](#) in the release and the [files installed](#) are available.

NCBI Resources How To

**Virus Variation** Ebolavirus

Contact us Help

Virus Variation home Virus resources

**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

Add query Show results

**Select sequence type:**

Protein  Nucleotide

**Search for keyword:**

Keyword Search in sequence pattern

**Define search set:**

| Species               | Host    | Region/Country | Genome Region                | Collection date   | Release date   |
|-----------------------|---------|----------------|------------------------------|-------------------|----------------|
| any                   | any     | any            | any                          | From: [ ] [ ] [ ] | [ ] [ ] [ ]    |
| Bundibugyo ebolavirus | Unknown | regions        | Nucleoprotein                | To: [ ] [ ] [ ]   | [ ] [ ] [ ]    |
| Tai Forest ebolavirus | Human   | Africa         | Polymerase complex protein   | [ ] [ ] [ ]       | [ ] [ ] [ ]    |
| Sudan ebolavirus      | Pig     | Asia           | Matrix protein               | [ ] [ ] [ ]       | [ ] [ ] [ ]    |
| Reston ebolavirus     |         | Europe         | Second secreted glycoprotein | Year Month Day    | Year Month Day |

Full-length genomes only

Add query Show results Clear form

**Figure 2.** Ebolavirus database. Users can get sequences by accession or browse by searching for a keyword, host, or region/country, among other options.

## Identical Protein Report Display option added to Protein database

*Tuesday, September 09, 2014*

A new display option has been added to the [Protein database](#) - the "Identical Protein Report". When viewing an individual record, this display allows you to access a list of all other identical proteins including those submitted as translations to GenBank, as well as RefSeq, UniProtKB/Swiss-Prot, PIR, PDB, and patented protein records.



[Display Settings:](#)  Identical Protein Report [Send to:](#)

## 60S ribosomal protein L23a [*Trypanosoma brucei*]

GenBank: AAX79509.1  
[GenPept](#) [FASTA](#) [Graphics](#)

RefSeq Selected Product: [XP\\_846140.1](#), 164 amino acids  
 Name: 60S ribosomal protein L23a [*Trypanosoma brucei brucei* strain 927/4 GUTat10.1]

| Source     | CDS Region in Nucleotide                     | Protein                     | Organism   | Superkingdom |
|------------|--|-----------------------------|--|--------------|
| RefSeq     | <a href="#">NC_007280.1:1362946-1363440+</a> | <a href="#">XP_846140.1</a> | <a href="#">Trypanosoma brucei brucei strain 927/4 GUTat10.1</a> | Eukaryota    |
| RefSeq     | <a href="#">XM_841047.1:1-495+</a>           | <a href="#">XP_846140.1</a> | <a href="#">Trypanosoma brucei brucei strain 927/4 GUTat10.1</a> | Eukaryota    |
| Swiss-Prot | N/A  | <a href="#">P41165.1</a>    | <a href="#">Trypanosoma brucei brucei</a>                        | Eukaryota    |
| PDB        | N/A  | <a href="#">3ZF7_X</a>      | <a href="#">Trypanosoma brucei brucei strain 927/4 GUTat10.1</a> | Eukaryota    |
| INSDC      | <a href="#">L21172.1:43-539+</a>             | <a href="#">AAC37186.1</a>  | <a href="#">Trypanosoma brucei</a>                               | Eukaryota    |
| INSDC      | <a href="#">AC105378.10:65131-65625+</a>     | <a href="#">AAX79509.1</a>  | <a href="#">Trypanosoma brucei</a>                               | Eukaryota    |
| INSDC      | <a href="#">CP000070.1:1362946-1363440+</a>  | <a href="#">AAZ12581.1</a>  | <a href="#">Trypanosoma brucei brucei strain 927/4 GUTat10.1</a> | Eukaryota    |
| INSDC      | <a href="#">FN554970.1:1431278-1431772+</a>  | <a href="#">CBH12694.1</a>  | <a href="#">Trypanosoma brucei gambiense DAL972</a>              | Eukaryota    |
| Patent     | N/A  | <a href="#">AAQ39714.1</a>  | Unknown  | Unknown      |
| Patent     | N/A  | <a href="#">ACC09504.1</a>  | Unknown  | Unknown      |

**Figure 1.** The “Identical Protein Report” display setting in the Protein database, showing identical proteins for 60S ribosomal protein L23 [*Trypanosoma brucei*].

As shown in Figure 1, the page title reflects the protein record from which you started. Beneath that, there is information on the suggested RefSeq preferred protein accession, protein length, and protein name. Identical proteins are presented in a tabular format that includes information on the database source (e.g., RefSeq, INSDC, etc.), the corresponding nucleotide CDS accession and location, the organism name, and the superkingdom. The displayed table can be downloaded for further use, and is also available through [Eutils](#).

The Identical Protein Report display setting provides important functions, such as:

- A mapping table between protein accessions and the nucleotide record(s) on which they are annotated, when relevant;
- For the RefSeq autonomous non-redundant protein dataset, a mapping table to the organisms to which the protein is relevant;
- And identification of highly conserved proteins when the identical protein sequence is found annotated on divergent species.