

# Recent improvements to the NCBI BLAST service

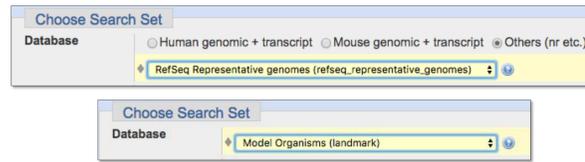
G Boratyn, C Camacho, A Fong, Y Merezhuk, T Rackers, Y Raytselis, J Ye, I Zaretskaya, P Cooper, W Matten, S McGinnis, T Tao, TL Madden  
Information Engineering Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health

## Abstract

Basic Local Alignment Search Tool (BLAST) is the most popular sequence similarity search tool. BLAST is a versatile tool that can search nucleotide or protein queries against either a protein or nucleotide database, translating the query or database as needed. The National Center for Biotechnology Information offers a number of ways to perform BLAST searches. First, the NCBI BLAST webpage is popular among users wishing to run a few searches. It offers a number of display options and links to other resources at the NCBI. Second, the stand-alone BLAST package allows power users to run searches on their own workstations or clusters. The standalone package also supports a large number of different types of reports. Finally, the NCBI offers a version of BLAST available at several cloud services. This offers access to a powerful computing environment without the expense of purchasing and maintaining hardware. We report on a number of BLAST improvements. We discuss recent changes to the NCBI BLAST webpage including an updated BLAST home page, a new fast protein search with SmartBLAST, and an updated taxonomy report. Additionally, the BLAST databases have been updated to better match user needs. We also discuss stand-alone BLAST changes including new report formats and improved multithreading utilization. This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

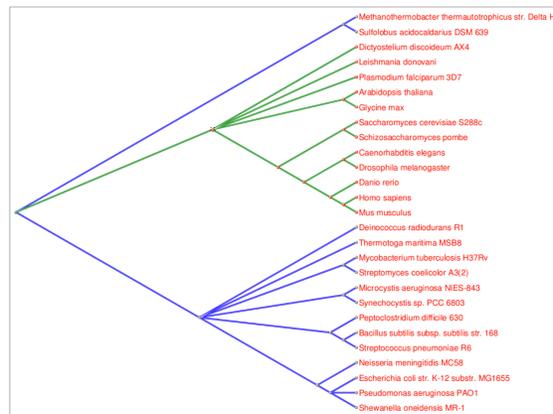
## Expanded Database Selection

Two new low-redundancy BLAST databases speed up BLAST searches and make results easier to interpret.



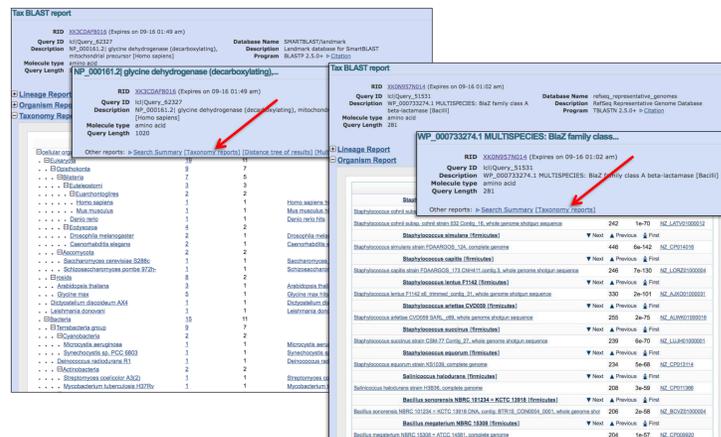
The nucleotide **RefSeq Representative genomes** database greatly reduces the redundancy in genome sequences as it contains only the high quality NCBI RefSeq Reference and Representative genome sequences. These are from broad taxonomic groups including eukaryotes, bacteria, archaea, viruses and viroids and are among the best quality genomes available at NCBI.

The protein **Model Organisms (landmark)** database contains proteomes from 27 genomes spanning a wide taxonomic range. These represent well-characterized organisms with well-annotated and accurate genomes. The Model Organisms database is useful for quick identification of unknown proteins and assessing the conservation across major groups of cellular organisms and is the database used in the initial search in SmartBLAST. The tree below shows the species with proteomes in the Model Organisms database.



## Updated BLAST Taxonomy Report

The updated Taxonomy report (Tax BLAST) available as a click from web results provides the taxonomic distribution of the BLAST matches. The images below show the Tax BLAST reports for searches against the Model Organisms and Representative Genomes databases.



## Expanded Output Formats

New specialized output formats are now available for the Web BLAST service and for standalone BLAST.

The **SAM** alignment format, popular for Next-Gen sequence analysis, is now available through the Downloads section on nucleotide BLAST results.



Many other useful formats are available for download from the Web service including the structure XML, XML2 and JSON formats.

## Formatting options for standalone BLAST

Standalone BLAST can generate the SAM format (nucleotide) and the structured output formats shown above (all BLAST programs) through options passed through the -outfmt flag. In addition the -outfmt flag can be used with the argument 6, 7 or 10 to produce tabular or comma delimited output with many options for specifying which fields should be shown. The examples below demonstrate some example of reports that can be generated either directly from the search programs or from the standalone blast\_formatter program. See the BLAST+ manual for complete details.

The first example shows the default output for the -outfmt 7. Only -outfmt 7 has the header information (lines starting with #). The second example uses -outfmt 6 and includes the sequence IDs and start/stop of the matches. The output also includes the scientific name for the database sequence. The BLAST database must be properly formatted for this task and an ancillary file (taxdb) must be downloaded from the NCBI FTP site.

```
blast_formatter -archive u00001.asn -outfmt 7 | head
# BLASTN 2.5.0+
# Query: U00001.1 Human homologue of S. pombe nuc2+ and A. nidulans bimA
# Database: nt
# Fields: query acc., subject acc., % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, ev
e, bit score
# 617 hits found
U00001.1 U00001 100.000 2592 0 0 1 2592 1 2592 0.0 4787
U00001.1 NM_001293089 99.884 2579 3 0 7 2585 89 2667 0.0 4747
U00001.1 NM_001256 99.768 2582 3 1 7 2585 89 2670 0.0 4732
U00001.1 S78234 99.768 2582 3 1 7 2585 49 2630 0.0 4732
U00001.1 XM_008961740 99.767 2579 6 0 7 2585 89 2667 0.0 4730
```

```
blast_formatter -archive u00001.asn -outfmt "6 qseqid saccvcr qstart qend sstart send sscinames" | head
U00001.1 U00001.1 1 2592 1 2592 Homo sapiens
U00001.1 NM_001293089.1 7 2585 89 2667 Homo sapiens
U00001.1 NM_001256.4 7 2585 89 2670 Homo sapiens
U00001.1 S78234.1 7 2585 49 2630 Homo sapiens
U00001.1 XM_008961740.1 7 2585 89 2667 Pan paniscus
U00001.1 XM_009431857.2 7 2585 89 2667 Pan troglodytes
U00001.1 XM_003809514.2 7 2585 89 2670 Pan paniscus
U00001.1 XM_511624.6 7 2585 89 2670 Pan troglodytes
U00001.1 XM_012497748.1 7 2585 89 2667 Nomascus leucogenys
U00001.1 XM_011525546.2 7 2585 76 2672 Homo sapiens
```

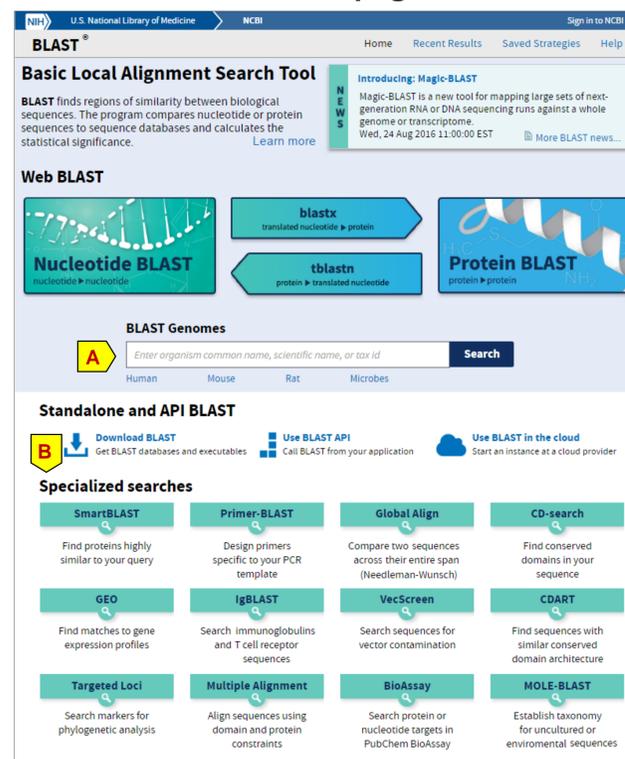
## Summary

The NCBI BLAST services have a number of new improvements that make them more flexible and powerful. Enhancements include new databases, report formats, and better handling of accession-only records. This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

## Links to Additional Information

BLAST Homepage: <https://blast.ncbi.nlm.nih.gov>  
BLAST Help Manual: <https://www.ncbi.nlm.nih.gov/books/NBK1762/>  
Standalone Manual: <https://www.ncbi.nlm.nih.gov/books/NBK279690/>  
Write to: [blast-help@ncbi.nlm.nih.gov](mailto:blast-help@ncbi.nlm.nih.gov)

## Enhanced BLAST Homepage



- A) The **BLAST Genomes** search box allows quick access to the best genomic data available for a target organism.
- B) Reorganized and clearly labeled **Specialized searches** make these BLAST related services and tools easier to identify and access.

## BLAST is Moving to HTTPS

As part of a US Government-wide change to provide a more secure experience NCBI BLAST is moving to HTTPS

### When?

The NCBI (and NLM) will be moving web sites and services, including web APIs, to HTTPS by September 30, 2016.

### What to do?

- The change will not affect interactive users of NCBI BLAST webpages.
- Users of the following services that access NCBI servers must change their scripts to use HTTPS
  - BioPerl, BioPython, BioJava
  - Users of the BLAST URL API
  - Commercial packages
- BLAST+ users who use the -remote option to send searches to NCBI must update to the BLAST+ 2.5.0 release (coming around September 26).

## Change to Accession-only Sequences

The GI number has been the main sequence identifier at the NCBI. Because of limitations of the GI, the NCBI is now moving to the accession.version (e.g., AARO4849.1) as the sole identifier. In the future, many sequences will only be assigned accessions. BLAST uses GIs for purposes such as limiting a search. Newer versions of BLAST (2.4.0 or newer) provide improved support for such tasks with accessions instead of GIs.

- Stand-alone search programs (e.g, blastn, blastp):
    - seqidlist option can use accessions rather than GIs.
  - Makeblastdb:
    - taxid\_map option will take a file of taxids and accessions.
  - Blastdbcmd:
    - FASTA output will drop the GI to have only accession.version
- Other improvements will appear in upcoming BLAST+ releases.