# NCBI News, January 2014

## Human CCDS release 15 now available on web and FTP

*Monday, January 27, 2014*

The Consensus Coding Sequence (CCDS) update for Homo sapiens annotation release 105 is now available on the CCDS website and FTP site. This release adds 349 new CCDS IDs to the human CCDS dataset and is based on comparative analysis of NCBI Homo sapiens annotation release 105 and Ensembl release 74. The human CCDS dataset now includes 29,045 proteins that correspond to 18,683 genes.

This is the final CCDS update for human that is based on the human reference assembly GRCh37. The next CCDS update for human will be based on the updated assembly GRCh38 and is tentatively expected to be released in July 2014.

## RefSeq release 63 now available

*Tuesday, January 21, 2014*

The full RefSeq release 63 is now available with nearly 50 million records describing 37,371,278 proteins, 5,760,653 RNAs, and sequences from 33,485 different organisms.

Some important updates include the following:

**Directory name change**: The RefSeq release directory “microbial” will be removed. Two new directories, “archaea” and “bacteria” will be added. This change will appear in release 65 in May 2014.

**WGS process flow change**: WGS accessions will no longer be processed on a per-project (WGS prefix) basis. Instead, these accessions will be processed and packaged the same as non-WGS accessions. This will significantly reduce the number of files in the /complete/ and (new) /archaea/ and /bacteria/ directories. Therefore, there will no longer be a series of files named like “microbialNZ_*”. Instead, all WGS scaffolds will be found in concatenated files just like all other accession data. We will continue to provide a separate file for the WGS master records. This change will appear in release 65 in May 2014.

**Human Genome GRCh38 Annotation plans**: The Genome Reference Consortium released an updated assembly for the human reference genome (GRCh38) in late December 2013. NCBI annotation of the RefSeq copy of this assembly is currently in

NLM Citation: NCBI News [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 1991-2012. NCBI News, January 2014.

progress. We anticipate releasing annotation in early to mid-February and including it in RefSeq release 64 in March 2014.

More details about RefSeq release 63 is included in the release statistics and release notes. In addition, reports indicating the accessions included in the release and the files installed are available.

## Taxonomy database now shows type material, sequences from type specimens and strains now labeled in Entrez

*Tuesday, January 21, 2014*

The naming, classification and identification of organisms traditionally relies on the concept of type material, which defines the representative examples ("name-bearing") of a species. For larger organisms, the type material is often a preserved specimen in a museum drawer, but the type concept also extends to type bacterial strains as cultures deposited in a culture collection. Of course, modern taxonomy also relies on molecular sequence information to define species. In many cases, sequence information is available for type specimens and strains. Accordingly, the NCBI has started to curate type material from the Taxonomy database, and are using this data to label sequences from type specimens or strains in the sequence databases. The figure below shows type material as it appears in the NCBI taxonomy entry and a sequence record for the recently described African monkey species, *Cercopithecus lomamiensis*.

Sequence from type material is particularly important because the species identification is virtually certain to be correct. The Entrez query "sequence from type"[filter] can be used to retrieve these sequence entries and can be used in combination with other queries as in the following examples.

**By Organism**

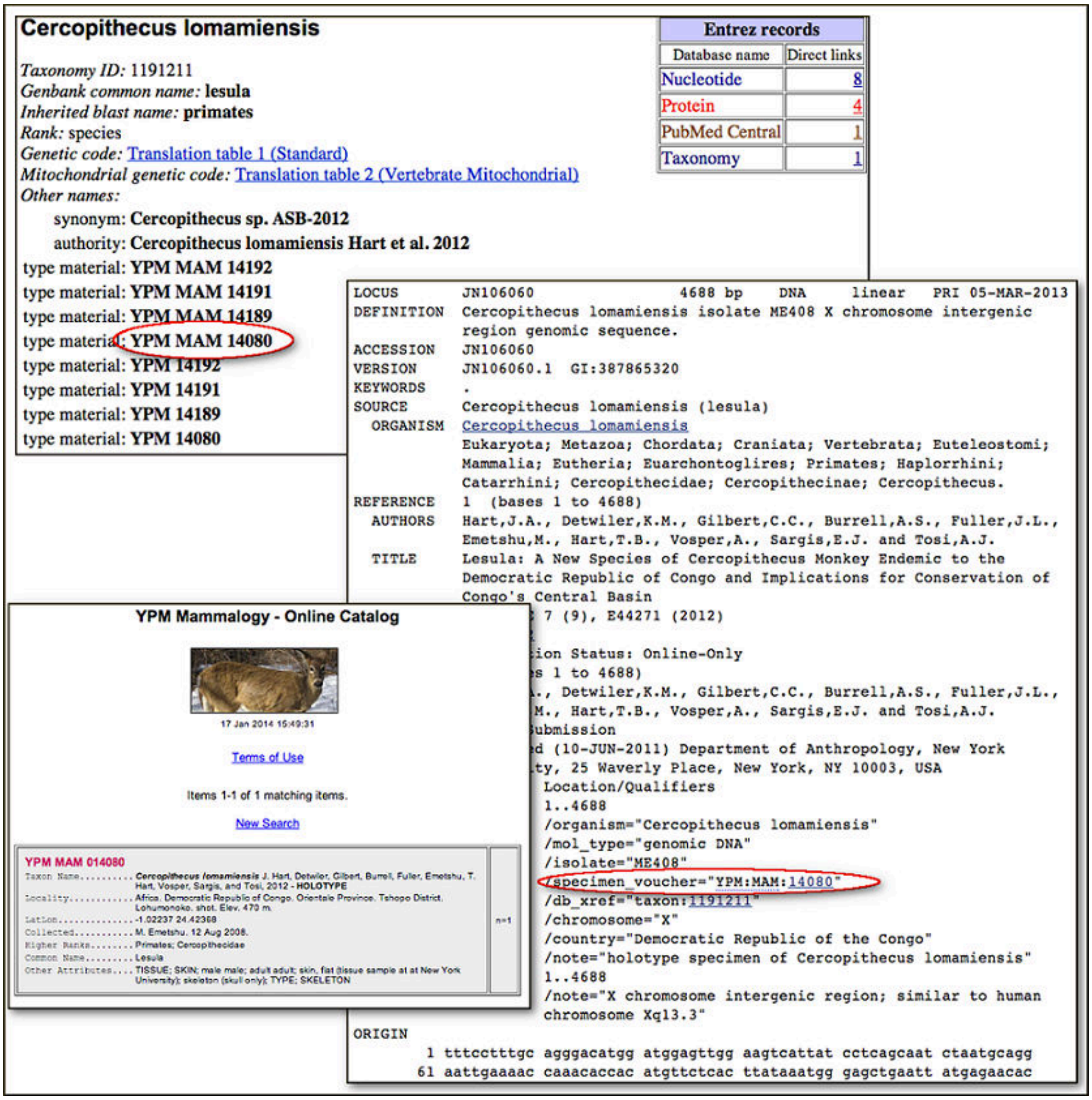
- sequence from type [filter] AND bacteria [Organism]
- sequence from type [filter] AND animals [Organism]

**By Collection**

- "sequence from type"[filter] AND collection cbs (type strains at the CBS culture collection)
- "sequence from type"[filter] AND collection mcz (type specimen at the Museum of Comparative Zoology)

**By Author**

- "sequence from type"[filter] AND hedges sb
- "sequence from type"[filter] AND Baldwin c



**Cercopithecus lomamiensis**

*Taxonomy ID:* 1191211
*Genbank common name:* **lesula**
*Inherited blast name:* **primates**
*Rank:* species
*Genetic code:* Translation table 1 (Standard)
*Mitochondrial genetic code:* Translation table 2 (Vertebrate Mitochondrial)
*Other names:*
synonym: **Cercopithecus sp. ASB-2012**
authority: **Cercopithecus lomamiensis Hart et al. 2012**
type material: **YPM MAM 14192**
type material: **YPM MAM 14191**
type material: **YPM MAM 14189**
type material: **YPM MAM 14080**
type material: **YPM 14192**
type material: **YPM 14191**
type material: **YPM 14189**
type material: **YPM 14080**

| Entrez records | |
|---|---|
| Database name | Direct links |
| Nucleotide | 8 |
| Protein | 4 |
| PubMed Central | 1 |
| Taxonomy | 1 |

```
LOCUS       JN106060                4688 bp    DNA     linear   PRI 05-MAR-2013
DEFINITION  Cercopithecus lomamiensis isolate ME408 X chromosome intergenic
            region genomic sequence.
ACCESSION   JN106060
VERSION     JN106060.1  GI:387865320
KEYWORDS    .
SOURCE      Cercopithecus lomamiensis (lesula)
  ORGANISM  Cercopithecus lomamiensis
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Cercopithecidae; Cercopithecinae; Cercopithecus.
REFERENCE   1  (bases 1 to 4688)
  AUTHORS   Hart,J.A., Detwiler,K.M., Gilbert,C.C., Burrell,A.S., Fuller,J.L.,
            Emetshu,M., Hart,T.B., Vosper,A., Sargis,E.J. and Tosi,A.J.
  TITLE     Lesula: A New Species of Cercopithecus Monkey Endemic to the
            Democratic Republic of Congo and Implications for Conservation of
            Congo's Central Basin
                     7 (9), E44271 (2012)

                     ion Status: Online-Only
                     s 1 to 4688)
                     ., Detwiler,K.M., Gilbert,C.C., Burrell,A.S., Fuller,J.L.,
                     M., Hart,T.B., Vosper,A., Sargis,E.J. and Tosi,A.J.
                     ubmission
                     d (10-JUN-2011) Department of Anthropology, New York
                     ty, 25 Waverly Place, New York, NY 10003, USA
                     Location/Qualifiers
                     1..4688
                     /organism="Cercopithecus lomamiensis"
                     /mol_type="genomic DNA"
                     /isolate="ME408"
                     /specimen_voucher="YPM:MAM:14080"
                     /db_xref="taxon:1191211"
                     /chromosome="X"
                     /country="Democratic Republic of the Congo"
                     /note="holotype specimen of Cercopithecus lomamiensis"
                     1..4688
                     /note="X chromosome intergenic region; similar to human
                     chromosome Xq13.3"
ORIGIN
        1 tttcctttgc agggacatgg atggagttgg aagtcattat cctcagcaat ctaatgcagg
       61 aattgaaaac caaacaccac atgttctcac ttataaatgg gagctgaatt atgagaacac
```

**YPM Mammalogy - Online Catalog**

17 Jan 2014 15:49:31

Terms of Use

Items 1-1 of 1 matching items.

New Search

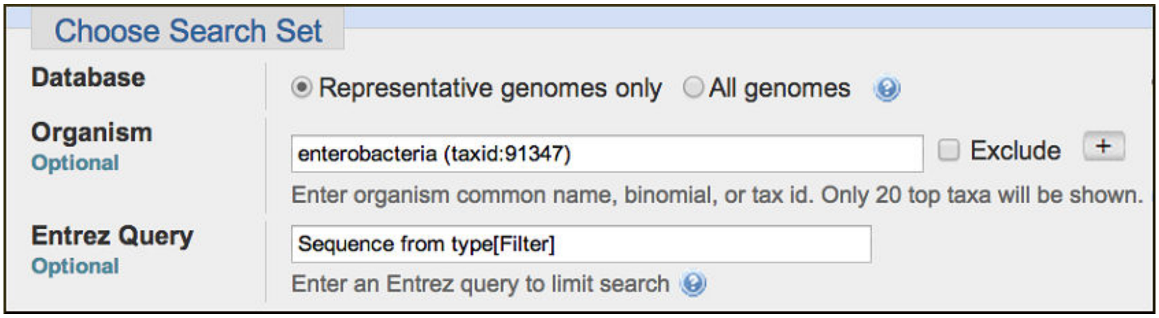
YPM MAM 014080

```
Taxon Name.........Cercopithecus lomamiensis J. Hart, Detwiler, Gilbert, Burrell, Fuller, Emetshu, T.
                   Hart, Vosper, Sargis, and Tosi, 2012 - HOLOTYPE
Locality...........Africa. Democratic Republic of Congo. Orientale Province. Tshopo District.
                   Lohumonoko. shot. Elev. 470 m.
LatLon.............-1.02237 24.42368                                              n=1
Collected..........M. Emetshu. 12 Aug 2008.
Higher Ranks.......Primates; Cercopithecidae
Common Name........Lesula
Other Attributes...TISSUE; SKIN; male male; adult adult; skin, flat (tissue sample at at New York
                   University); skeleton (skull only); TYPE; SKELETON
```

**Figure 1**. Type material information as it appears in the NCBI Taxonomy database (upper left) and nucleotide database (lower right) for *Cercopithecus lomamiensis*. Both records refer to the type material housed in the Yale Peabody Museum Mammalogy collection (lower left, YPM MAM 140180).

As shown in the figure below, “sequence from type”[filter] is also useful as an Entrez query to limit BLAST searches to reliably identified sequences, particularly when working with prokaryotes.

Stay tuned for more developments and added features coming to the Taxonomy database.



Choose Search Set

Database: ● Representative genomes only ○ All genomes

Organism (Optional): enterobacteria (taxid:91347) □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query (Optional): Sequence from type[Filter]

Enter an Entrez query to limit search

**Figure 2**. The "Choose Search Set" section of the Microbial Genomes BLAST page. The optional Organism limit, enterobacteria,  and the Entrez query 'Sequence from type[Filter]' restricts the search to sequences from enterobacteria type strains.

## New NCBI Insights blog: NCBI Remap tool helps you transition to newest human reference genome assembly, GRCh38

*Thursday, January 16, 2014*

NCBI's Genome Remapping Service (NCBI Remap) allows you to map annotation data from one genomic assembly to another for a selected set of organisms. This may be particularly helpful in updating your annotations for the human reference genome assembly, which has recently been updated to the new version, GRCh38. The newest blog post on NCBI Insights describes how Remap works and how it allows you to analyze data in the context of the newest genome assembly. Finally, the blog post provides links to Remap-related documentation including an overview, FAQs, and a YouTube tutorial.

## Sequence Viewer PDF rendering available - YouTube video tutorial

*Tuesday, January 14, 2014*

NCBI has created a YouTube video tutorial showing you how to generate a PDF rendering of your Sequence Viewer display. The short clip takes you through the downloading process, and shows you what you can do with your file after creating a PDF.

## Genome Workbench Update 2.7.12

*Tuesday, January 14, 2014*

Genome Workbench 2.7.12 has been released. The update includes several new features like a faster Tree Renderer, redesigned data import and an improved GFF format reader. The release notes include more information on features, fixes and improvements.

# VAST+ released: Find similar 3D structures for macromolecular complexes

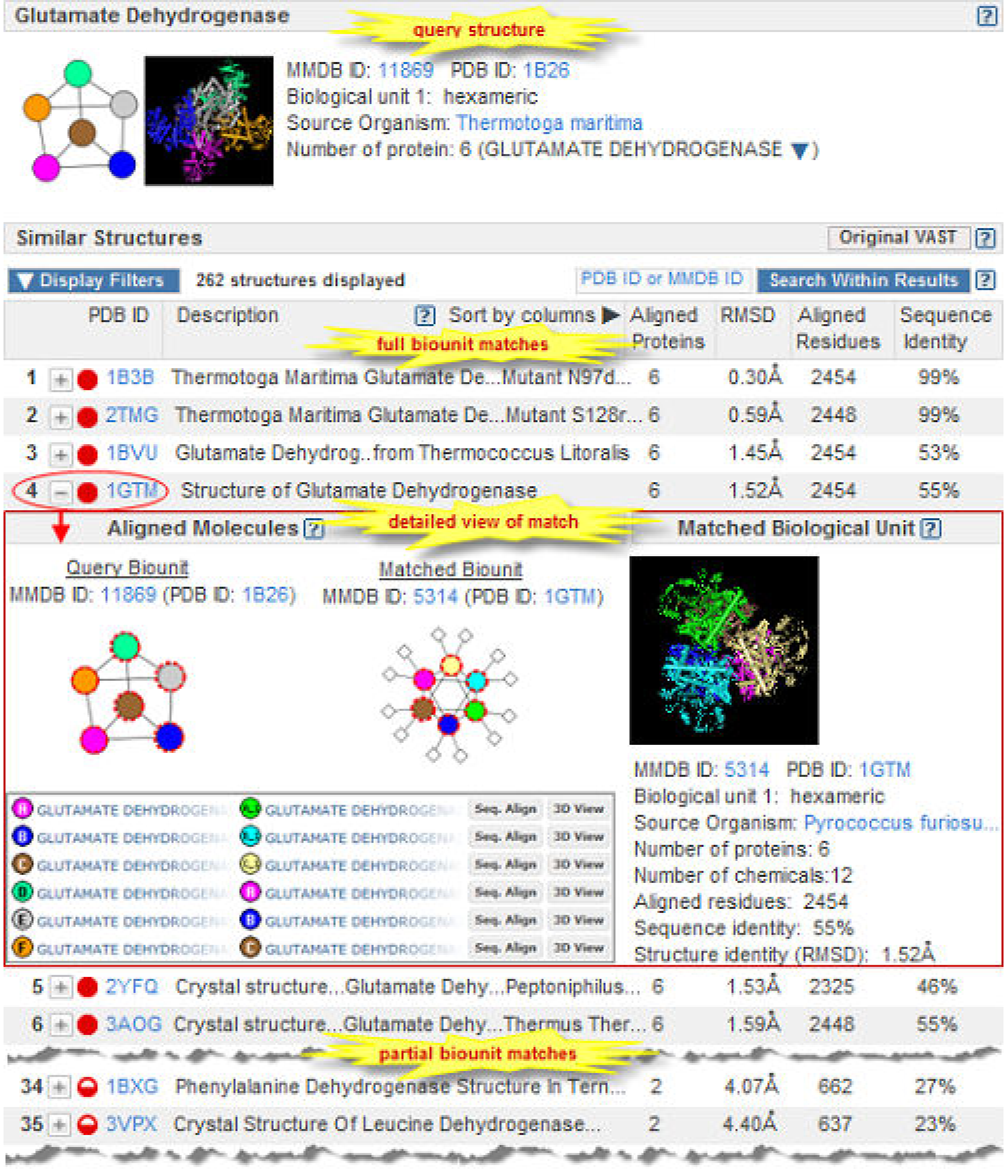
*Thursday, January 09, 2014*

VAST+ is a new tool designed to identify macromolecules that have similar 3-dimensional structures with an emphasis on finding similar macromolecular complexes. The similarities are calculated using purely geometric criteria without regard to sequence similarity, and therefore can identify distant homologs.

This new tool is built upon the original Vector Alignment Search Tool (VAST) and expands the capabilities of that program by taking into account the biological unit ("biounit") of each structure, not just individual protein chains or their substructures.

A recent publication provides detailed information about the VAST+ algorithm. In addition, the extensive VAST+ help document includes a comparison of original VAST and VAST+, as well as examples of how can this tool can be used to learn more about proteins.

*Please note that in order to view the 3D superpositions of similar biological units, you must install the most recent version of the NCBI molecular viewing software, Cn3D 4.3.1.*



Glutamate Dehydrogenase [?]

MMDB ID: 11869 PDB ID: 1B26
Biological unit 1: hexameric
Source Organism: Thermotoga maritima
Number of protein: 6 (GLUTAMATE DEHYDROGENASE ▼)

Similar Structures — Original VAST [?]

▼ Display Filters — 262 structures displayed — PDB ID or MMDB ID — Search Within Results [?]

| | PDB ID | Description [?] Sort by columns ▶ | Aligned Proteins | RMSD | Aligned Residues | Sequence Identity |
|---|---|---|---|---|---|---|
| 1 ⊞ ● | 1B3B | Thermotoga Maritima Glutamate De...Mutant N97d... | 6 | 0.30Å | 2454 | 99% |
| 2 ⊞ ● | 2TMG | Thermotoga Maritima Glutamate De...Mutant S128r... | 6 | 0.59Å | 2448 | 99% |
| 3 ⊞ ● | 1BVU | Glutamate Dehydrog..from Thermococcus Litoralis | 6 | 1.45Å | 2454 | 53% |
| 4 ⊟ ● | 1GTM | Structure of Glutamate Dehydrogenase | 6 | 1.52Å | 2454 | 55% |

Aligned Molecules [?] — Matched Biological Unit [?]

Query Biounit
MMDB ID: 11869 (PDB ID: 1B26)

Matched Biounit
MMDB ID: 5314 (PDB ID: 1GTM)

| | | | |
|---|---|---|---|
| Ⓐ GLUTAMATE DEHYDROGEN | [illegible] GLUTAMATE DEHYDROGEN | Seq. Align | 3D View |
| Ⓑ GLUTAMATE DEHYDROGEN | [illegible] GLUTAMATE DEHYDROGEN | Seq. Align | 3D View |
| Ⓒ GLUTAMATE DEHYDROGEN | [illegible] GLUTAMATE DEHYDROGEN | Seq. Align | 3D View |
| Ⓓ GLUTAMATE DEHYDROGEN | Ⓐ GLUTAMATE DEHYDROGEN | Seq. Align | 3D View |
| Ⓔ GLUTAMATE DEHYDROGEN | Ⓑ GLUTAMATE DEHYDROGEN | Seq. Align | 3D View |
| Ⓕ GLUTAMATE DEHYDROGEN | Ⓒ GLUTAMATE DEHYDROGEN | Seq. Align | 3D View |

MMDB ID: 5314 PDB ID: 1GTM
Biological unit 1: hexameric
Source Organism: Pyrococcus furiosu...
Number of proteins: 6
Number of chemicals:12
Aligned residues: 2454
Sequence identity: 55%
Structure identity (RMSD): 1.52Å

| | PDB ID | Description | Aligned Proteins | RMSD | Aligned Residues | Sequence Identity |
|---|---|---|---|---|---|---|
| 5 ⊞ ● | 2YFQ | Crystal structure...Glutamate Dehy...Peptoniphilus... | 6 | 1.53Å | 2325 | 46% |
| 6 ⊞ ● | 3AOG | Crystal structure...Glutamate Dehy...Thermus Ther... | 6 | 1.59Å | 2448 | 55% |
| 34 ⊞ ◒ | 1BXG | Phenylalanine Dehydrogenase Structure In Tern... | 2 | 4.07Å | 662 | 27% |
| 35 ⊞ ◒ | 3VPX | Crystal Structure Of Leucine Dehydrogenase... | 2 | 4.40Å | 637 | 23% |

Figure. VAST+ search results for *Thermotoga maritima* Glutamate Dehydrogenase (PDB ID 1B26) with detailed view of a match to 1GTM.

## For more information:

- VAST+ home page
- Publication in Nucleic Acids Research
- Vast+ examples
- Cn3D home page

## NCBI Insights blog: A Librarian's Guide to NCBI - an intensive training course for medical librarians to be offered April 2014

*Wednesday, January 08, 2014*

The NCBI in partnership with the National Library of Medicine Training Center (NTC) will offer the Librarian's Guide to NCBI course on the NIH campus in April 2014. This will be the second presentation of the course; it was previously offered in the spring of 2013 (NCBI Insights April 11 and May 6, 2013). After the course, we will post lecture slides and hands-on practical exercises on the education area of the NCBI FTP site and video tutorials of the course lectures will be available on the NCBI YouTube channel. Materials from the 2013 course are available, as well as lecture videos for the expression module. More information, including prerequisites, is available in the newest NCBI Insights blog post.

## Mouse genome annotation release 104 available

*Wednesday, January 08, 2014*

The mouse (*Mus musculus*) genome annotation has recently been updated to annotation release 104 and is now available in the Nucleotide, Protein sequence and Gene databases, is searchable using BLAST, and can be downloaded from the FTP site.

Mouse annotation release 104, based on the sequence assemblies GRCm38.p2 (GCF_000001635.22) and Mm_Celera (GCF_000002165.2), identifies a total of 35,389 genes, as well as 100,581 transcripts on GRCm38.p2. RNA-Seq data from 31 distinct BioSample accessions were aligned to assist in gene prediction. More statistics are available in the annotation report.

See what other annotation runs are in progress on the Eukaryotic Genome Annotation Pipeline status page.

## BLAST+ 2.2.29 now available

*Tuesday, January 07, 2014*

Stand-alone BLAST version 2.2.29+ is now available for download from the FTP site. BLAST 2.2.29+ provides a number of important improvements and bug fixes. Some improvements include improved blastn batch query performance, source releases build optimized multi-thread binaries by default, and improved multithreading by better dividing the BLAST database among threads. The BLAST Release notes lists more upgrades and fixes.