NCBI News, August 2015

NCBI annotates 250th eukaryote with Eukaryotic Genome Annotation Pipeline

Friday, August 28, 2015

This month, the NCBI Eukaryotic Genome Annotation Pipeline has annotated its 250th organism! Mammals dominate the list of annotated organisms with a total of 95, but NCBI has increased coverage of invertebrates - we've annotated 22 new insects since the beginning of 2015. See the full list of annotated organisms and request that your favorite organism(s) be annotated!

We make an effort to re-annotate genomes every two years so the latest annotation incorporates recently submitted RNA-Seq and Transcript Shotgun Assemblies as evidence and benefits from the latest software developments. Data produced by the Eukaryotic Genome Annotation Pipeline is available in the Reference Sequences (RefSeq) collection, BLAST non-redundant and organism-specific databases, and the Gene database; it is also downloadable from the NCBI FTP site.

September 2nd NCBI Minute: "Introducing SmartBLAST, a Rapid Protein Identification Tool"

Thursday, August 20, 2015

On September 2nd, NCBI staff will introduce SmartBLAST, a faster alternative to ordinary protein-protein BLAST searches for protein query sequence identification.

SmartBLAST reports the top three results from a separate database of high quality protein sequences, as well as the top two hits from nr. SmartBLAST also produces a full multiple alignment of the query sequence and results with mapped conserved domains. SmartBLAST is part of the new PubMed Labs initiative from NCBI.

Date and Time: September 2nd, 2015 12:15 PM EDT

Registration URL: https://attendee.gotowebinar.com/register/7468878402343662081

After the live presentation, the webinar will be uploaded to the NCBI YouTube channel. The webinar and any materials will also be archived on the Webinars and Courses page, where you can also find information about future webinars.

GenBank release 209.0 is now available via FTP

Wednesday, August 19, 2015

GenBank release 209.0 (8/14/2015) has 187,066,846 non-WGS, non-CON records containing 199,823,644,287 base pairs of sequence data. In addition, there are 302,955,543 WGS records containing 1,163,275,601,001 base pairs of sequence data, as well as 87,827,013 TSA records containing 69,360,654,413 base pairs of sequence data.

During the 57 days between the close dates for GenBank releases 208.0 and 209.0, the traditional (i.e., non-WGS/non-CON) portion of GenBank grew by 5,902,601,341 base pairs and by 2,047,494 sequence records. During that same period, 288,641 records were updated. An average of 40,985 traditional records were added and/or updated per day.

Between releases 208.0 and 209.0, the WGS component of GenBank grew by 124,338,390,780 base pairs and by 44,253,405 sequence records; the TSA component of GenBank grew by 8,663,181,843 base pairs and by 10,852,412 sequence records.

The total number of sequence data files increased by 59 with this release. The divisions are as follows:

- BCT: 9 new files, now a total of 196
- CON: 4 new files, now a total of 323
- ENV: 4 new files, now a total of 85
- GSS: 2 new files, now a total of 299
- INV: 1 new file, now a total of 129
- MAM: 19 new files, now a total of 28
- PAT: 6 new files, now a total of 229
- PHG: 1 new file, now a total of 3
- PLN: 2 new files, now a total of 114
- VRL: 1 new file, now a total of 37
- VRT: 10 new files, now a total of 56

For downloading purposes, please keep in mind that the uncompressed GenBank flatfiles are approximately 735GB (sequence files only); the ASN.1 data are approximately 600GB.

More information about GenBank release 209.0 is available in the release notes.

Tree Viewer 1.6 now available

Wednesday, August 19, 2015

NCBI Tree Viewer version 1.6 includes several new features, improvements and bug fixes, including the added ability to download data in Newick and Nexus formats. To see the full list of updates, see the Tree Viewer release notes.

NCBI Tree Viewer is a tool for viewing your own phylogenetic tree data.

New NCBI video: "NCBI's 1000 Genomes Browser: Introduction"

Wednesday, August 12, 2015

The newest video on the NCBI YouTube channel is an introduction to the 1000 Genomes Browser, which allows you to view variation and genotype data, and support sequence reads from the 1000 Genomes Project.

Subsequent videos will cover other functions, such as uploading data. For video updates, subscribe to our YouTube channel.

August 26th webinar: "Troubleshooting GenBank Submissions: Determining and Annotating Coding Regions (CDS) for Eukaryotic Genes"

Wednesday, August 12, 2015

In two weeks, NCBI staff will show you how to use BLAST to determine the locations of the coding sequences in your genomic submissions. You will also learn how to describe these coding regions with the GenBank submission tools BankIt and Sequin, annotate multiple transcript splice variants, and address problems with splice sites, internal stop codons in protein translations, and sequencing gaps that affect coding region annotation.

Date and Time: August 26, 2015 1PM EDT

Registration URL: https://attendee.gotowebinar.com/register/3143702023795693569

After the live presentation, the webinar will be uploaded to the NCBI YouTube channel. The webinar and any materials will also be archived on the Webinars and Courses page, where you can also find information about future webinars.

New NCBI Insights blog post: "SciENcv Updated to Support New NIH Biosketch Format"

Monday, August 10, 2015

The latest blog post on NCBI Insights will show you how to use SciENcv to convert your existing NIH Biosketch from the old format to the new format required for grant applications submitted with due dates after May 24, 2015.

Genomes FTP site update (version 1.2) expands taxonomic scope and more

Wednesday, August 05, 2015

NCBI has released a comprehensive update of all current genome assemblies in the Genomes FTP site, affecting data reported in the /genomes/all/, /genomes/genbank/, and /genomes/refseq/ FTP directories. This update expands the taxonomic scope of the /refseq/ data and adds a new report file, a data conversion script, and more. The FTP content of all "latest" GenBank and RefSeq assemblies was updated to reflect these changes.

Genomes FTP version 1.2 includes the following changes:

- Genome group directories:
  - Assembly summary files have been added to genbank and refseq genome group directories (e.g., ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/fungi/assembly_summary.txt)
  - Viral RefSeq genomes have been added to the Assembly database and a new genome group directory, /viral/, is now available: ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/viral/
- New feature table report:
  - Files named as *_feature_table.txt.gz are tab-delimited files reporting annotated features, coordinates, and attributes (including names). Feature types reported include gene, CDS, RNA (all types), operon, immunoglobulin C/V/N/S regions and V/D/J segments.
- WGS master files:

  - This record type has been added to the FTP release and is provided in GenBank flatfile format using the file name convention *_wgsmaster.gbff.gz
- Conversion script:
  - add_utrs_to_gff.py: Python script to add explicit UTR exon features, as inferred from the gene, mRNA, exon and CDS features, to GFF3 formatted data. Script location: ftp://ftp.ncbi.nlm.nih.gov/genomes/TOOLS/add_utrs_to_gff/
- GBFF format:
  - Genomic records in the CON division now include both a CONTIG line and the sequence. For example, the GenBank flatfile format for GG698602.1, a GenBank scaffold in the *Dialister invisus* DSM 15470 Assembly GCA_000160055.1, shows both the CONTIG and ORIGIN/sequence data:

```
CONTIG     join(ACIM02000001.1:1..1894898,gap(unk100),ACIM02000002.1:1..962)
ORIGIN
       1 caaggcttgg agcgacataa aactaatagg tcgaggtctt aacttaggaa caccgagaca (ETC.)
```

- GFF3:
  - Additional information about NCBI's GFF3 files is now available at ftp://ftp.ncbi.nlm.nih.gov/genomes/README_GFF3.txt
  - GFF files now include information on the gene biotype
- Assembly summary files:
  - Assembly levels have been simplified to four types: contig, scaffold, chromosome, and complete genome
- Assembly reports:
  - The length of each sequence has been added to the report
  - UCSC style names (e.g., chr1) have been added to the report for those sequences that have been matched to assemblies on the UCSC genomes FTP site (e.g., /genomes/all/GCF_000001405.30_GRCh38.p4/GCF_000001405.30_GRCh38.p4_assembly_report.txt)

Please note that RefSeq prokaryotic genome annotation is currently being refreshed. Once that process is complete, we will update the data in the Genomes FTP site.

CCDS release 19 for mouse added to Gene

Tuesday, August 04, 2015

The Consensus Coding Sequence (CCDS) update that compares NCBI's *Mus musculus* annotation release 105 to Ensembl's release 81 is now reflected in Gene. This update adds 1,003 new CCDS IDs and adds 148 Genes into the mouse CCDS set. CCDS release 19 includes a total of 24,834 CCDS IDs that correspond to 20,215 GeneIDs.

For information about CCDS, please visit the CCDS homepage.