NCBI News, August 2014

Milestone: NCBI annotates 150th eukaryotic genome

Thursday, August 28, 2014

NCBI has now completed the genome annotation for 150 different organisms. The 150th organism is the Upper Galilee mountains blind mole rat (*Nannospalax galili*), a rodent of particular interest because of its resistance to cancer.

NCBI began annotating eukaryotic genomes in 2000, and now has complete genome annotations for 150 different organisms, including:

- 74 mammals,
- 39 other vertebrates,
- 21 invertebrates,
- and 16 plants.

Among these organisms, 40 were annotated for the first time in 2014. Data produced by the Eukaryotic Genome Annotation Pipeline is available in the Reference Sequences (RefSeq) collection, BLAST non-redundant and organism-specific databases, Gene database, and on the NCBI FTP site.

View genomes currently in progress and browse the list of all eukaryotes ever annotated by NCBI using the Eukaryotic Genome Annotation Pipeline. Need a public genome annotated? Make a request!

The new NCBI Genomes FTP site is here!

Tuesday, August 26, 2014

NCBI has released a major revision of the genomes FTP site. The new FTP site structure provides a single entry point to access sequence and annotation content of both GenBank and RefSeq genomes data. The FTP site can be accessed directly for FTP, or from links provided in NCBI's Assembly database.

The initial release of the redesigned genomes FTP site adds three new directories, namely '*genbank*', '*refseq*', and '*all*' to the existing ftp area – **ftp://ftp.ncbi.nlm.nih.gov/genomes/**. It includes >28,000 GenBank and >17,000 RefSeq assemblies ranging from archaea to human and provides a consistent core set of files for the sequence and annotation products. Additional file formats will be added in future updates.

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

The revised FTP site offers several advantages including:

- comprehensive provision of GenBank and RefSeq genomes data available in NCBI's Assembly database
- provision of a consistent core set of files including:
  - FASTA format for genomic sequences, accessioned transcript products, and accessioned protein products
  - GenBank/GenPept format for genomic, transcript, and protein records
  - GFF (version 3) format for annotated genomic records
  - Md5checksums for all files provided per assembly
- **consistent use of accession.version as the primary sequence ID for both GFF and FASTA files**; this facilitates the use of these data in some public domain RNAseq read mapping tools.

To give those with automated tools time to update, we plan to maintain the older content and structure of the preexisting /genomes/ FTP site in parallel with the new structure until March 1, 2015. The older content will be archived or deleted after that date. Please contact info@ncbi.nlm.nih.gov if you have concerns or questions about these changes.

More information on the initial release and documentation of file formats is available in the following FTP README files:

- genomes
- GenBank & RefSeq
- assembly structure.

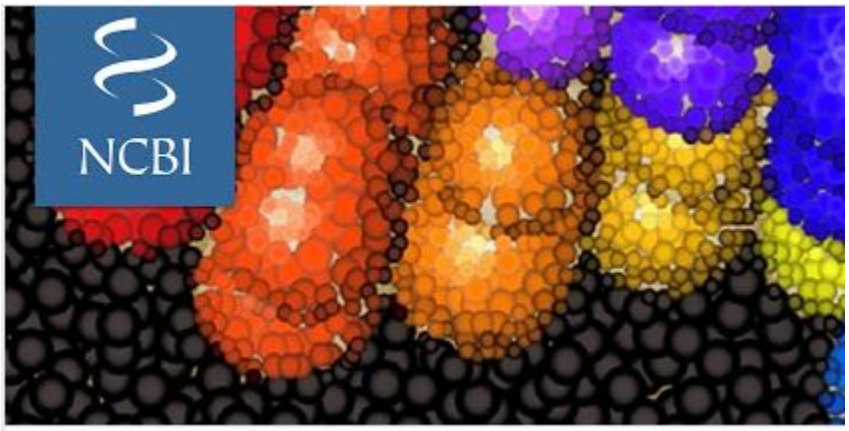
There is also a Genomes FTP FAQ available at www.ncbi.nlm.nih.gov/genome/doc/ftpfaq.

New NCBI YouTube video: Downloading FASTA sequences in Sequence Viewer

Friday, August 22, 2014

The newest video on NCBI's YouTube channel is a quick tutorial on downloading FASTA sequences for certain gene features using NCBI's Sequence Viewer.

This video is one of several on the Sequence Viewer playlist. Subscribe to the NCBI YouTube channel for notifications on all new videos and to see videos on My NCBI, Variation Viewer, E-Utilities and many more of the programs and services NCBI provides.



Rat annotation release 105 now on Gene, FTP, sequence and BLAST databases

Friday, August 22, 2014

Rat (*Rattus norvegicus*) assemblies Rnor_6.0 (GCF_000001895.5, reference) and Rn_Celera (GCF_000002265.2) are annotated in release 105.

This annotation was produced by the Eukaryotic Genome Annotation Pipeline and is available in the sequence and BLAST databases, in Gene, and on the FTP site.

RNA-Seq data from 340 distinct BioSample accessions were aligned to help gene prediction. A total of 29,998 genes and 61,506 transcripts were identified on Rnor_6.0. More statistics are available in the annotation report.

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

New NCBI Insights blog: How to comply with NIH Public Access Policy

Thursday, August 21, 2014

The newest blog post on NCBI Insights, "Advice for NIH Grantees: How to Comply with NIH Public Access Policy" guides grantees through the compliance process, from determining whether the Public Access policy applies to their publication to tracking compliance status with My Bibliography.

"BLAST in the Cloud!" is the newest video on the NCBI Webinars YouTube playlist

Wednesday, August 20, 2014

The video recording of July's NCBI webinar, "BLAST in the Cloud", is live on the NCBI Webinars YouTube playlist, complete with closed captioning.

The NCBI Webinars playlist contains video recordings of past webinars, and will be updated regularly. Each of the videos on the playlist is accessible from the NCBI Webinars page as well.

GenBank release 203.0 is now available via FTP

Wednesday, August 20, 2014

Release 203.0 (8/16/2014) has 174,108,750 non-WGS, non-CON records containing 165,722,980,375 base pairs of sequence data. In addition, there are 189,080,419 WGS records containing 774,052,098,731 base pairs of sequence data.

During the 65 days between the close dates for GenBank Releases 202.0 and 203.0, the non-WGS/non-CON portion of GenBank grew by 3,900,134,732 base pairs and by 755,674 sequence records. During that same period, 403,182 records were updated; an average of 17,828 non-WGS/non-CON records were added and/or updated per day. Between releases 202.0 and 203.0, the WGS component of GenBank grew by 54,470,139,988 base pairs and by 13,301,355 sequence records.

The total number of sequence data files increased by 41 with this release. The divisions are as follows:

- BCT: 6 new files, now a total of 142
- CON: 11 new files, now a total of 278
- ENV: 1 new file, now a total of 74
- EST: 1 new file, now a total of 476
- INV: 1 new file, now a total of 40
- PAT: 1 new file, now a total of 210
- PLN: 16 new files, now a total of 86
- PRI: 1 new file, now a total of 48

- SYN: 1 new file, now a total of 8
- VRL: 1 new file, now a total of 32
- VRT: 1 new file, now a total of 33

For downloading purposes, please keep in mind that the uncompressed GenBank flatfiles are approximately 652 GB (sequence files only). The ASN.1 data require approximately 544 GB.

More information about GenBank Release 203.0 and coming changes are available in the release notes.

Genome Workbench 2.8.0 released

Wednesday, August 20, 2014

Genome Workbench 2.8.0 is available, as of August 18th. New features include exporting alignments to tab delimited format, and new flexible broadcasting between bio trees. For the full list of fixes, improvements and features, see the Genome Workbench release notes.

"NCBI's OSIRIS: Quality Assurance for DNA Forensic Profiling" webinar on September 17th

Tuesday, August 19, 2014

On September 17th, NCBI will host a webinar that will cover OSIRIS, an open-source forensics analysis program. **To sign up for this webinar, please go here.**

Identification of people, animals and tissues by forensic, identification and stem cell transplant engraftment laboratories is typically done by analyzing PCR-amplified short tandem repeats (STRs). OSIRIS minimizes analysis time and increases accuracy by identifying artifacts. In addition, OSIRIS can process thousands of samples per day, and the program can also be used by medical and biological research laboratories to identify and validate tissue and cell lines.

The FBI has accepted OSIRIS as a validated expert system, and it is used by the U.S. Army Criminal Investigation Laboratory. OSIRIS can be downloaded here.

To see past and upcoming webinars, please visit the NCBI Webinars page.

UniVec build 8.0 now available for VecScreen searches and FTP

Friday, August 01, 2014

UniVec, NCBI's non-redundant database of vector sequences, has been updated to build 8.0, which enables searches run using NCBI's VecScreen tool to detect more of the foreign sequences introduced during the cloning or sequencing process. UniVec build 8.0 is also available via FTP.

This build added 2 complete vector sequences and 257 adapter and primer sequences, including a large number of oligonucleotides used in next-generation sequencing protocols, bringing the total number of sequences represented in the UniVec database to 2,282.

UniVec is a non-redundant database of sequences commonly attached to cDNA or genomic DNA during the cloning process. UniVec primarily consists of the unique segments from a large number of vectors but also includes many linker, adapter and primer sequences. Redundant sub-sequences have been eliminated from the database to make searches more efficient and to simplify interpretation of the results. For more details, see the UniVec page.