NCBI News, February 2015

March 5th webinar: "NCBI and the NIH Public Access Policy: PubMed Central submissions, My NCBI, My Bibliography and SciENcv"

Wednesday, February 25, 2015

Next Thursday, March 5th, NCBI will host a webinar outlining how to use My NCBI to report public access policy compliance for NIH grant holders. Topics will include the NIH Public Access Policy, NIHMS and PubMed Central submissions, creating My NCBI accounts, use of My Bibliography to report compliance to eRA Commons and using SciENcv to create BioSketches.

To register for this webinar, go here.

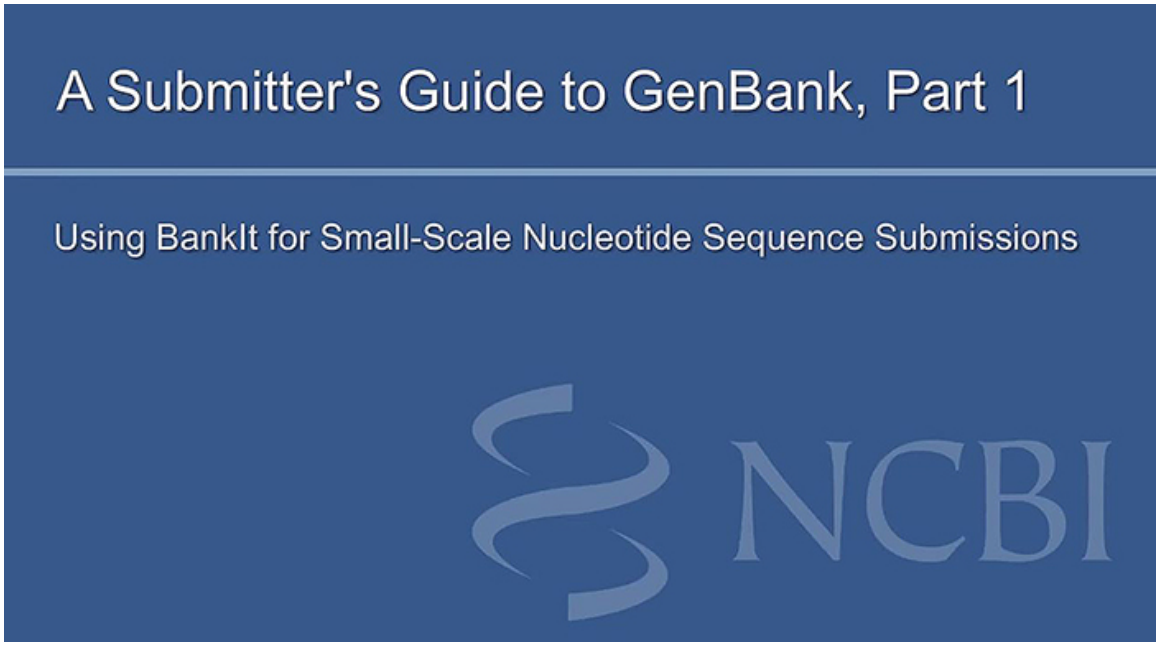
Please see the NCBI Webinars page for a list of upcoming webinars as well as recordings (on YouTube) and related materials from past webinars.

"A Submitter's Guide to GenBank" webinar parts 1 and 2 on YouTube

Friday, February 20, 2015

If you missed the recent "A Submitter's Guide to GenBank" webinar series, Parts 1 and 2 have been uploaded to our official YouTube account. In addition, we have prepared a PDF of the question and answer sessions conducted after each presentation, available via FTP.

NLM Citation: NCBI News [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 1991-2012. NCBI News, February 2015.



These webinars outline the process of using BankIt. a web-based submission tool at NCBI, to submit sequence data to the GenBank database. The first part is a demonstration on using BankIt forms to complete a submission of a single or a few nucleotide sequences, while the second part shows you how to use BankIt file inputs to complete a submission of nucleotide sequences that require multiple features for each sequence.

The NCBI YouTube channel provides presentations and tutorials about our biomolecular and biomedical literature databases and tools. See the Webinars playlist to watch presentations you may have missed or to rewatch your favorite video. A list of past and upcoming webinars, as well as related materials, is available on the NCBI Webinars page.

NCBI Insights blog: How to delegate authority to others to edit/ create your profile and Collections

Thursday, February 19, 2015

The latest blog post on the NCBI Insights blog shows you how to send a delegate invitation that will allow a colleague to view and edit your My Bibliography collection as well as view, edit and create profiles in your SciENcv.

NCBI Insights offers posts that cover new developments and events at NCBI, as well as tips on using our resources and tools. We encourage you to join the conversation at NCBI Insights.

NCBI webinar on February 25: The Next Generation of Access to Sequencing Data: Using NCBI's SRA Toolkit to Access Data from dbGaP and SRA

Wednesday, February 18, 2015

Next Wednesday, February 25th, NCBI staff will present a webinar on the SRA Toolkit, a system for accessing the approximately 3.4 Petabases of next-generation genomic and expressed sequence data housed in the NCBI Sequence Read Archive (SRA).

As data sets grow larger, mining information and performing comparisons directly from structured databases becomes increasingly necessary. The SRA Toolkit is not only capable of dumping data out as fastq or sam files, but also provides direct analysis and comparison from specific genomics regions across hundreds or thousands of samples.

In the webinar, we will show examples of configuration and use of the Toolkit for both public SRA and controlled access data associated with studies in the Database of Genotypes and Phenotypes (dbGaP).

To register for this webinar, please go here.

NCBI Genomes FTP site update adds analysis sets and other data

Wednesday, February 18, 2015

Several improvements have recently been implemented in three sections of the NCBI Genomes FTP site: GenBank and RefSeq (both browsable), and the "all" genomes FTP directory (not browsable; however, it can be used for scripted downloads based on assembly directory name).

A range of new content is available for download on the FTP site:

- Analysis sets for human GRCh38 and mouse GRCm38.p3 are in the GenBank assembly directories. These sets contain FASTA and GFF files with modified sequence identifiers and index files, which make these data convenient for analysis with next generation sequencing tools. Please refer to the provided documentation (human | mouse) for a complete description.
- A text file, assembly_summary.txt, has been added to each species directory. This file is a species-specific subset of the comprehensive assembly summary files provided in the "Assembly Reports" folder. The file content includes information on release dates, submitter and assembly names, assembly accession version, assembly status, RefSeq category, full FTP path (see below) and associated meta-data, including BioProject and BioSample identifiers. Example: *Saccharomyces cerevisiae*

- The full FTP path has been added as the last column in assembly_summary_refseq.txt and assembly_summary_genbank.txt files provided in the "Assembly Reports" folder.
- A small number of assemblies that have a large number of contigs were omitted from the first release of the new FTP site. These assemblies are now available and include *Triticum aestivum* (see Assembly GCA_000334095.1) and *Locusta migratoria* (see Assembly GCA_000516895.1).

In addition, files for all "latest" assemblies were regenerated to make the following changes:

- Removal of erroneously reported CDD features in RNA flat files
- Inclusion of missing strand information for some features on the forward strand and added plus signs ("+") in column 7 in updated GFF3 files
- Correct representation of multi-interval non-trans-spliced tRNA features on GFF3 files. Each multi-interval non-trans-spliced tRNA feature is now represented by a single feature (line) of type tRNA and multiple nested features of type exon (one for each interval).

NCBI staff continues to work on fully replacing the original /genbank/genomes/ and /genomes/ FTP content. As previously announced, we plan to remove content from the older FTP directories by the end of March. We will not remove content from the historical areas until it is available in the new areas. Note that some content may be available in a different file name or format or sub-directory in the newer FTP directories.

Please refer to the FTP README.txt files and the NCBI Genomes FTP FAQs to learn more.

GenBank release 206.0 is now available via FTP

Tuesday, February 17, 2015

Release 206.0 (2/13/2015) has 181,336,445 non-WGS, non-CON records containing 187,893,826,750 base pairs of sequence data. In addition, there are 205,465,046 WGS records containing 873,281,414,087 base pairs of sequence data, as well as 66,706,014 TSA records containing 49,765,340,047 base pairs of sequence data.

During the 63 days between the close dates for GenBank releases 205.0 and 206.0, the non-WGS, non-CON portion of GenBank grew by 2,955,763,136 base pairs and by 2,040,676 sequence records. During that same period, 164,936 records were updated; an average of 35,010 non-WGS, non-CON records were added and/or updated per day. Between releases 205.0 and 206.0, the WGS component of GenBank grew by 24,303,492,065 base pairs and by 5,163,496 sequence records. The TSA component of GenBank also grew; 3,708,919,144 base pairs and 4,070,397 sequence records were added.

The total number of sequence data files increased by 46 with this release. The divisions are as follows:

- BCT: 10 new files, now a total of 169
- CON: 11 new files, now a total of 303
- ENV: 2 new files, now a total of 80
- GSS: 4 new files, now a total of 293
- INV: 1 new file, now a total of 133
- PAT: 3 new files, now a total of 217
- PLN: 1 new file, now a total of 96
- TSA: 13 new files, now a total of 172
- VRL: 1 new file, now a total of 45

For downloading purposes, please keep in mind that the uncompressed GenBank flatfiles are approximately 700 GB (sequence files only), and the ASN.1 data require approximately 572 GB.

More information about GenBank release 206.0, including important changes in this release and upcoming changes, is available in the release notes.

Mouse, cow and zebrafish added to dbSNP build 142

Thursday, February 12, 2015

Three organisms are now available in dbSNP build 142: mouse, cow and zebrafish. This data is indexed in Entrez and is available by FTP.

New mouse (mus musculus) information on FTP and Entrez:

- Assembly: GRCm38.p2 (GCF_000001635.22)
- New RS: 9323191
- Total RS: 80429085

New cow (bos taurus) information on FTP and Entrez:

- Assembly: Bos_taurus_UMD_3.1 (GCF_000003055.4)
- New RS: 11509794
- Total RS: 85027819

New zebrafish (danio rerio) information on FTP and Entrez:

- Assembly: Zv9 (GCF_000002035.4)
- New RS: 16326757
- Total RS: 17765748

1000 Genomes Browser updated to include Phase 3 May 2013 call set

Tuesday, February 10, 2015

1000 Genomes Browser version 3.4 is now available. This update includes variant and genotype calls from the Phase 3 May 2013 call set. For a full list of browser features, see the Release Notes. A detailed browser user guide is also available.

The browser will continue to provide access to data from the Phase 1 March 2012 call set.