NCBI News, December 2013

New human genome assembly (GRCh38) released!

Tuesday, December 24, 2013

On December 24th, the Genome Reference Consortium (GRC) submitted a new assembly for the human genome (GRCh38) to GenBank. These data are now available in the Assembly database with accession GCA_000001405.15 and are also available on the FTP site. Please note the GRC provides these assemblies as unannotated sequences.

Now that the GRC sequences are available in GenBank, our Reference Sequence (RefSeq) Genome Annotation Group has downloaded these sequences and has begun processing them using our eukaryotic annotation pipeline. The resulting human chromosome sequences will continue to have the RefSeq accessions NC_000001-NC_000024, but their versions will increment as the update to the GRCh38 assembly includes a sequence change for all chromosomes. The process of annotating the human genome generally takes about 2 weeks. When this is complete, we will incorporate these sequences into various analysis and display tools, such as human genome BLAST, NCBI Remapping Service, and various genome viewers. Thus, at the end of this process each chromosome will be represented by both an unannotated sequence in GenBank (the original GRC data) and an annotated sequence in the RefSeq collection.

Please check back frequently for updates on the NCBI News and our social media sites (NCBI Twitter Channel, NCBI Facebook Page, NCBI Announce RSS Feed, NCBI Announce Email ListServ) as this process unfolds.

In addition, we have a series of posts on the NCBI Insights Blog site on topics such as how NCBI processes genome annotations, a tip to remap annotations from older assemblies to GRCh38, and highlighting some loci that have changed significantly in the new assembly.

Annotation reports now generated for recently annotated organisms

Monday, December 23, 2013

The NCBI Eukaryotic Genome Annotation Pipeline now publishes a report to accompany each new annotation. This report provides statistics on the annotation products, such as the number of genes, the number and length of coding and non-coding transcripts, and

NLM Citation: NCBI News [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 1991-2012. NCBI News, December 2013.

the number of transcripts per gene. It also presents statistics on the protein and transcript alignments that were used by the gene prediction process.

See the annotation reports generated for rat and potato:
http://www.ncbi.nlm.nih.gov/genome/annotation_euk/Rattus_norvegicus/104
http://www.ncbi.nlm.nih.gov/genome/annotation_euk/Solanum_tuberosum/100

Annotation reports for other recently annotated organisms can be accessed from the Eukaryotic Genome Annotation Pipeline status page.

Learn more about how the eukaryotic genome annotation pipeline works:http://www.ncbi.nlm.nih.gov/news/12-17-2013-new-handbook-chapters-genome-annotation-pipelines/

Meet PubMed Commons: The new comments forum in PubMed

Thursday, December 19, 2013

If you are one of the millions of people who visit PubMed today, be on the look-out for something different. On each abstract page, there's now a section called PubMed Commons. It's a forum for scientific discussion on publications open to any authors in the world's largest biomedical literature database.

Read more at PubMed Commons Blog.

Rat genome annotation release 104

Wednesday, December 18, 2013

The rat (*Rattus norvegicus*) 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.

Rat annotation release 104, based on the sequence assemblies Rnor_5.0 (GCF_000001895.4, reference) and Rn_Celera (GCF_000002265.2), identifies a total of 31,451 genes. In addition, 64,745 transcripts were identified on Rnor_5.0. A new annotation pipeline step in this update is the alignment of RNA-Seq data from 85 distinct BioSample accessions to assist in gene prediction.

More statistics are available in the Rat Annotation Release 104 Report.

See what other annotation runs are in progress on the Eukaryotic genome annotation pipeline status page.

New NCBI Handbook chapters: Eukaryotic and prokaryotic genome annotation pipelines

Tuesday, December 17, 2013

In order to increase the utility of genomic information, we provide gene annotation and other features on Reference Sequence (RefSeq) genome records. Genome annotation is a multi-step process that includes prediction of protein-coding genes, as well as other functional genome units such as structural RNAs, tRNAs, small RNAs, pseudogenes, control regions, direct and inverted repeats, insertion sequences, transposons, and other mobile elements.

Depending upon the genome, the identification of key genomic features and their locations on RefSeq genome records are provided by outside sources (the submitter's annotation copied from the GenBank genomic sequence records or curated annotation provided by a model organism database, like FlyBase or WormBase), or are generated by annotation pipelines developed at NCBI specifically for eukaryotic or for prokaryotic genomes.

An overview of each pipeline is available in our web documentation In addition to web documentation of our eukaryotic genome annotation pipeline and prokaryotic genome annotation process.

Our newest NCBI Handbook Chapters on the eukaryotic and prokaryotic annotation pipelines describe the processes in greater detail, including information on algorithms, history, annotation standards and special considerations like multiple annotation assemblies:

- Eukaryotic Genome Annotation Handbook chapter
- Prokaryotic Genome Annotation Pipeline Handbook chapter

We also provide eukaryotic genome annotation policies and the status of genomes in the current pipeline, as well as information about prokaryotic genome annotation standards.

Sequence Viewer has been updated

Tuesday, December 17, 2013

NCBI Sequence Viewer provides a graphical view of sequences and color-coded annotations on regions of sequence stored in the Nucleotide and Protein databases. Sequence Viewer has recently been updated and now has better loading and management of uploaded custom tracks, improved naming of downloaded files including sequence ranges and file extensions, and easier embedding in external Web sites.

A full list of new features, improvements and fixes is available at: http://www.ncbi.nlm.nih.gov/tools/sviewer/release-notes/

GenBank release 199 now available

Monday, December 16, 2013

GenBank Release 199 is now available through NCBI's Entrez and BLAST services.

Release 199.0 (12/10/2013) has 169,331,407 non-WGS, non-CON records containing 156,230,531,562 base pairs of sequence data. In addition, there are 133,818,570 WGS records containing 556,764,321,498 base pairs of sequence data.

During the 54 days between the close dates for GenBank Releases 198.0 and 199.0, the non-WGS/non-CON portion of GenBank grew by 1,054,036,863 base pairs and by 996,011 sequence records. During the same period, 494,249 records were updated; an average of 27,597 non-WGS/non-CON records per day were added and/or updated. Between releases 198.0 and 199.0, the WGS component of GenBank grew by 20,922,153,757 base pairs and by 3,615,365 sequence records.

The total number of sequence data files increased by 18 with this release. The divisions are as follows:

- BCT: 2 new files, now a total of 114
- CON: 5 new files, now a total of 231
- ENV: 2 new files, now a total of 67
- EST: 1 new file, now a total of 475
- GSS: 1 new file, now a total of 279
- PAT: 2 new files, now a total of 199
- PLN: 1 new file, now a total of 65
- TSA: 2 new files, now a total of 147
- VRL: 2 new files, now a total of 29

For downloading purposes, please keep in mind that the GenBank flatfiles are approximately 618 GB (sequence files). ASN.1 data are approximately 508 GB.

Change to Accession Format: As mentioned in the GenBank Release 198.0 news story, CON-division WGS scaffolds will have new format accession numbers. For an example of the new accession format, please see Section 1.3.2 of the GenBank release notes. We do not currently plan to update existing records with the new accession format.

NCBI Video: Submitting manuscripts on NIHMS

Thursday, December 05, 2013

NCBI's latest YouTube video takes you through the manuscript submission process on the NIH Manuscript Submission System (NIHMS), step-by-step. NIHMS enables publishers, authors, and principal investigators to submit manuscripts for processing and archiving in PubMed Central.

PMCID - PMID - Manuscript ID - DOI Converter Upgraded

Tuesday, December 03, 2013

We have upgraded the PMCID - PMID - Manuscript ID - DOI Converter. The updated ID Converter API allows you to convert IDs for publications referenced in PubMed and PMC.

The ID Converter tool allows you to convert IDs for publications referenced in PubMed and PMC. You can also cross-reference Open Access NIH Manuscript Submission IDs (NIHMS) and Digital Object Identifiers (DOIs) often used by publishers. For example, these identifiers refer to the same publication:

PMCID: PMC3702208

PMID: 24288678

NIHMS: NIHMS518180

DOI: 10.1007/s00213-013-3057-1

This tool uses an underlying web service, which is also publicly available for those needing programmatic access to this data. For more information, see the ID Converter API documentation.