NCBI News, August 2016

Genomes FTP site data organization to change on September 20, 2016

Tuesday, August 30, 2016

NCBI is moving the contents of the "all" and "ASSEMBLY_REPORTS/All" directories on the Genomes FTP site. Currently, listing the contents of these two directories is impractical because they contain many thousands of directories or files.

Reorganization of ftp://ftp.ncbi.nlm.nih.gov/genomes/all

The genome assembly directories currently directly under "all" will be moved into a new 4-level structure under genomes/all.

Two new directories under "all" will be named for the accession prefix (GCA or GCF). These directories will contain another three levels of directories named for digits 1-3, 4-6 & 7-9 of the assembly accession, creating paths like *genomes/all/GCA/xxx/xxx/xxx/* and *genomes/all/GCF/xxx/xxx/xxx/*. For example:

- The data currently in genomes/all/GCA_000001405.23_GRCh38.p8 will be moved to genomes/all/GCA/000/001/405/GCA_000001405.23_GRCh38.p8.
- The data currently in genomes/all/GCF_001696305.1_UCN72.1 will be moved to genomes/all/GCF/001/696/305/GCF_001696305.1_UCN72.1.

Schedule of changes

On September 20, 2016:

- New directories genomes/all/GCA, genomes/all/GCF and the three levels of directories named for groups of digits in the assembly accession will be added.
- Individual genome assembly data directories directly under genomes/all will be moved into the new directory structure under genomes/all/GCA & GCF.
- Assembly data directories directly under genomes/all will be replaced by symbolic links to the corresponding directory in the new structure.
- The old and new data organizations will be maintained in parallel for 6 weeks.

On December 1, 2016:

NLM Citation: NCBI News [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 1991-2012. NCBI News, August 2016.

- The old paths to individual genome assembly data directories directly under genomes/all will be removed.
- All access to genome assembly data under genomes/all/ will need to use the genomes/all/GCA/xxx/xxx/xxx/ & genomes/all/GCF/xxx/xxx/xxx/ paths.

Impact

Users who access genome assembly data by any of the following methods will not be affected by this change:

- Following a link to "Download the GenBank assembly" or "Download the RefSeq assembly" from an Assembly details page
- Navigating the genomes/genbank or genomes/refseq paths of the genomes FTP site
- Using the ftp_path provided in the assembly_summary.txt files provided on the genomes FTP site

Users who mirror all data under genomes/all will get two copies of the data for each genome assembly during the transition period, unless they modify their scripts to only take data from genomes/all/GCA & GCF.

Other changes:

- Scripts that retrieve data using hard-coded paths to individual genome assembly directories directly under genomes/all will fail after the transition period
- Links from non-NCBI web pages to individual genome assembly directories directly under genomes/all will fail after the transition period
- Published paths to individual genome assembly directories directly under genomes/all will fail after the transition period

Removal of ftp://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/All

First, the assembly reports currently under genomes/ASSEMBLY_REPORTS/All will be moved into the assembly data directories in the new directory hierarchy under genomes/all/GCA & genomes/all/GCF described above, replacing the symbolic links to the assembly report files that currently exist in this location. The assembly report files in the assembly data directories will retain the name previously provided by the symbolic link.

- {assembly_accession.version}.assembly.txt will appear as {assembly_accession.version}_{assembly_name}_assembly_report.txt
- {assembly_accession.version}.stats.txt will appear as {assembly_accession.version}_{assembly_name}_assembly_stats.txt
- {assembly_accession.version}.regions.txt will appear as {assembly_accession.version}_{assembly_name}_assembly_regions.txt

Then, the genomes/ASSEMBLY_REPORTS/All directory will be removed.

Schedule

On September 20, 2016:

- The assembly reports currently under genomes/ASSEMBLY_REPORTS/All will be moved into the assembly data directories, replacing the symbolic links currently in the data directories.
- The assembly reports under genomes/ASSEMBLY_REPORTS/All will be replaced by symbolic links to the corresponding report in the assembly data directory.
- The old and new data organizations for assembly reports will be maintained in parallel for 6 weeks.

On December 1, 2016:

- The old paths to assembly reports under genomes/ASSEMBLY_REPORTS/All will be removed.
- The genomes/ASSEMBLY_REPORTS/All directory will be removed.
- All access to assembly reports will need to use the genomes/all/GCA/, genomes/all/GCF, genomes/genbank or genomes/refseq paths to the individual assembly data directories.

Impact

Users who access assembly reports by any of the following methods will not be affected by this change:

- Following a link to "Download the full sequence report" from an Assembly details page
- From an assembly data directory under the genomes/genbank or genomes/refseq path on the genomes FTP site

Attempts to access assembly reports using the genomes/ASSEMBLY_REPORTS/All path will fail after the transition period.

Additional information about the genomes FTP site can be found in the genomes FTP README file and in the genomes FTP FAQ.

Subscribe to the genomes-announce mail list to be informed of changes to the NCBI genomes FTP site.

dbVar July 2016 data release includes new 1000 Genomes Phase III structural variants

Monday, August 29, 2016

The dbVar July 2016 data release includes 1,455,032 new Variant regions, 13,961,956 Variant calls and 6 new studies. See a list of the studies, including descriptions and links to the data in the release notes.

Follow the dbVar RSS feed for monthly releases.

August 31st NCBI Minute: Downloading Genome Data from the NCBI FTP Site

Thursday, August 25, 2016

In the next NCBI Minute, we will teach you how to use the Web and the command line to quickly access and download genomic sequence and annotation files for a species, metagenome or taxonomic group of interest.

Date and time: Wednesday, August 31, 2016 12:00 PM EDT

Registration URL: https://attendee.gotowebinar.com/register/8835228315982188801

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the NCBI YouTube channel. Any related materials will be accessible on the Webinars and Courses page; you can also learn about future webinars on this page.

VAST+ update provides refined alignments

Tuesday, August 23, 2016

The new version of VAST+ provides a refined structure-based alignment of similar macromolecular complexes and displays the 3D superpositions in the recently launched iCn3D.

See the MMDB news page for more detail about how VAST+ now works.

September 12th class at NLM: EDirect - Command Line Access to NCBI's Biomolecular Databases

Monday, August 22, 2016

On September 12, 2016, NCBI staff will discuss EDirect in a class at the National Library of Medicine. During the optional first hour of this workshop (9-10 AM), you will get a basic introduction to the Unix/Linux command line interface. The main workshop (10 AM - Noon) will cover how to use EDirect to set up simple pipelines to retrieve and process data from PubMed, Gene and the Nucleotide and Protein sequence databases. We will provide access to EDirect installed in a Linux environment on a cloud service.

Date and time: Monday, September 12, 2016 9:00 AM EDT

Registration link: https://www.surveymonkey.com/r/5NCWLK6

NOTE: This is an in-person class at the National Library of Medicine on the NIH campus in Bethesda, MD, USA. **The course is limited to 22 participants. Participants must bring their own laptop.**

The EDirect suite of programs allows easy command line access for searching and retrieving literature (PubMed) and accessing NCBI's biomolecular (Gene, Nucleotide, sequence databases, etc.) records. Its advantages include direct command line access to NCBI's databases without writing Perl or Python scripts, construction of custom pipelines for processing data, built-in batch access, and the ability to generate highly flexible custom output reports.

HIV-1 datasets in Gene updated

Thursday, August 11, 2016

NCBI has updated the HIV-1 interaction datasets available in Gene with data provided by the Southern Research Institute.

The protein interactions dataset now has:

- 7,762 interacitons;
- 15,665 interaction descriptions;
- 3,729 proteins encoded by 3,649 human genes;
- and 6,690 publications.

The replications interactions dataset now has:

- 1,325 interactions;
- 1,439 interaction decriptions;
- 1,325 proteins encouded by 1,325 human genes;
- and 125 publications.

Data are also available at the RefSeq HIV-1 website and the GeneRIF FTP site.