NCBI News, May 2017

NCBI to phase out support for non-human organism data in dbSNP and dbVar

Tuesday, May 09, 2017

Starting September 1, 2017, NCBI will not accept non-human variant data submissions to dbSNP and dbVar. Any non-human data that is already in the databases or that is submitted before September 1, 2017 will continue to be available via the dbSNP and dbVar FTP download sites.

NCBI will phase out support for non-human organisms in dbSNP and dbVar following this timeline:

- September 1, 2017 – dbSNP and dbVar stop accepting submissions of non-human variant data.
- November 1, 2017 – dbSNP and dbVar interactive websites and related NCBI services stop presenting non-human variant data. The data will, however, continue to be available for download on the dbSNP and dbVar FTP sites.

We would like to thank all the submitters and users who have supported dbSNP and dbVar throughout the years. If you want to submit non-human variation data now or after September 1, 2017, European Bioinformatics Institute (EBI) – one of our partners in the International Nucleotide Sequence Database Collaboration (INSDC) – is accepting these data in the European Variation Archive.

Eleven eukaryotic annotations added to RefSeq in April 2017

Monday, May 08, 2017

In April, the NCBI Eukaryotic Genome Annotation Pipeline released new annotations in RefSeq for the following eleven organisms:

- *Bombus terrestris* (buff-tailed bumblebee)
- *Ceratitis capitata* (Mediterranean fruit fly)
- *Athalia rosae* (coleseed sawfly)
- *Dendrobium catenatum* (a monocot)
- *Phalaenopsis equestris* (a monocot)
- *Orbicella faveolata* (stony coral)

NLM Citation: NCBI News [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 1991-2012. NCBI News, May 2017.

- *Pogona vitticeps* (central bearded dragon)
- *Oryzias latipes* (Japanese medaka)
- *Sesamum indicum* (sesame)
- *Jatropha curcas* (a eudicot)
- *Amborella trichopoda* (a flowering plant)

See more details on the Eukaryotic RefSeq Genome Annotation Status page.

NCBI to assist with NYGC Genomics Hackathon June 19-21

Monday, May 08, 2017

From June 19-21, 2017, the NCBI will assist in a bioinformatics hackathon at the New York Genome Center (NYGC). This hackathon will focus on advanced bioinformatics analysis of next generation sequencing (NGS) data, proteomics and metadata. To apply for this hackathon, complete this application (approximately 10 minutes to complete). Applications are due **Monday, May 22, 2017 by 5 PM ET**.

This event is for researchers, including students and postdocs, who are already engaged in the use of bioinformatics data or in the development of pipelines for bioinformatics analyses from high-throughput experiments. Some projects are available to other non-scientific developers, mathematicians or librarians.

The event is open to anyone selected for the hackathon and able to travel to the NYGC (see address below).

Potential subjects for this iteration are:

- Expanding and publicizing a Shiny app for visualizing protein correlation profiling data,
- building a pipeline for efficient partitioning of barcodes,
- creating a public JBrowse database for all Staphylococcus aureus genomes,
- simulating tumor genomes,
- associating somatic mutations with clinical outcomes,
- simplifying access to shared-data repositories from Python, and
- building a pipeline for searching for virus-associated protein domains in NGS datasets.

Please see the application for specific and evolving team projects.

Organization

There will be 5-7 teams comprised of 5-6 individuals. These teams will build pipelines and tools to analyze large datasets within a cloud infrastructure.

After a brief organizational session, teams will spend three days analyzing a challenging set of scientific problems related to a group of datasets. Participants will analyze and combine datasets in order to work on these problems.

Datasets

Datasets will come from public repositories or will be supplied by the project lead. During the hackathon, participants will have an opportunity to include other datasets and tools for analysis. Please note, if you use your own data during the hackathon, we ask that you submit it to a public database within **six months** of the end of the event.

Products

All pipelines and other scripts, software and programs generated in this course will be added to a public GitHub repository designed for that purpose. Manuscripts describing the design and usage of the software tools constructed by each team may be submitted to an appropriate journal such as the F1000Research hackathons channel.

Application

To apply, complete this form (approximately 10 minutes to complete). Applications are due Monday, May 22, 2017 by 5 PM ET.

Participants will be selected based on the experience and motivation they provide on the form. Prior participants and applicants are especially encouraged to apply.

The first round of accepted applicants will be notified on May 24th by 5 pm ET, and have until May 25th at 5 pm ET to confirm their participation. If you confirm, please make sure it is highly likely you can attend, as confirming and not attending bars other data scientists from attending this event. Please include a monitored email address, in case there are follow-up questions.

Note:

1. Participants will need to bring their own laptop to this program.
2. A working knowledge of scripting (e.g., Shell, Python, R) is necessary to be successful in this event. Employment of higher level scripting or programming languages may also be useful.
3. Applicants must be willing to commit to all three days of the event.
4. No financial support for travel, lodging or meals is available for this event.
5. The hackathon may extend into the evening hours on Monday and/or Tuesday. Please make any necessary arrangements to accommodate this possibility.

Please contact ben.busby@nih.gov with any questions.

Venue: New York Genome Center, 101 6th Ave, New York, NY 10013

GenBank release 219.0 is available via FTP

Thursday, May 04, 2017

GenBank release 219.0 (4/14/2017) has 200,877,884 traditional records containing 231,824,951,552 base pairs of sequence data. In addition, there are 451,840,147 WGS records containing 2,035,032,639,807 base pairs of sequence data, 165,068,542 TSA records containing 149,038,907,599 base pairs of sequence data, as well as 1,438,349 TLS records containing 636,923,295 base pairs of sequence data.

During the 60 days between the close dates for GenBank releases 218.0 and 219.0, the traditional portion of GenBank grew by 3,105,513,914 base pairs and by 1,536,507 sequence records. During that same period, 173,862 records were updated (an average of 28,506 added and/or updated per day).

Between releases 218.0 and 219.0, the WGS component of GenBank grew by 142,066,331,172 base pairs and by 42,349,750 sequence records. The TSA component of GenBank grew by 15,521,695,495 base pairs and by 13,637,057 sequence records. The TLS component of GenBank did not change.

The total number of sequence data files increased by 42 with this release. The divisions are as follows:

- BCT: 20 new files, now a total of 350
- CON: 3 new files, now a total of 359
- ENV: 2 new files, now a total of 97
- EST: 2 new files, now a total of 483
- INV: 1 new file, now a total of 153
- PAT: 7 new files, now a total of 290
- PHG: 1 new file, now a total of 4
- PLN: 2 new files, now a total of 145
- PRI: 1 new file, now a total of 56
- SYN: 1 new file, now a total of 10
- TSA: 1 new file, now a total of 230
- VRL: 1 new file, now a total of 48

For downloading purposes, please keep in mind that the uncompressed GenBank Release 219.0 flatfile require roughly 818 GB (sequence files only). The ASN.1 data require approximately 685 GB.

More information about GenBank release 219.0 is available in the release notes, as well as in the README files in the genbank (ftp.ncbi.nih.gov) and ASN.1 (ncbi-asn1) directories.

May 10th NCBI Minute: How to Locate and Use Human Genomes and Annotations from NCBI

Monday, May 01, 2017

In two weeks, NCBI staff will show you how to quickly find and download human genome annotations from both the web and the command line for incorporation into your workflows. We will also show you how to convert the accessions in these files to those used in other bioinformatics databases, as well as how to visualize these annotations on our Genome Data Viewer.

Date and time: Wednesday, May 10, 2017 12:00 PM - 12:30 PM EDT

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 from the Webinars and Courses page; you can also learn about future webinars on this page.