NCBI News, February 2016

NCBI to assist Brandeis University in hosting Boston genomics hackathon in April

Monday, February 29, 2016

From April 25 to 27, NCBI will assist Brandeis University in hosting a genomics hackathon focusing on advanced bioinformatics analysis of next-generation sequencing data and metadata. This event is for students, postdocs, investigators and other researchers already engaged in the use of pipelines for genomic analyses from next-generation sequencing data or metadata.* Researchers and/or data scientists from the Boston area are especially encouraged to apply, but the event is open to anyone selected for the hackathon and able to travel to Brandeis.

Working groups of 5-6 individuals will be formed into five or six teams. These teams will build pipelines and tools to analyze large datasets within a cloud infrastructure. The potential subjects for this iteration are:

1. Network Analysis of Variants
2. Structural Variation*
3. RNA-Seq
4. Streaming Data and Metadata*
5. Neuroscience/Immunity
6. Command-line user-interface design*

Please see the application for specific team projects.

**Some projects are available to other non-scientific developers, mathematicians or librarians.*

Organization

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 the public repositories housed at the NCBI. During the course, participants will have an opportunity to include other datasets and tools for analysis.

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

Please note, if you use your own data during the course, 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. A manuscript outlining the design and usage of the software tools constructed by each team will be submitted to an appropriate journal.

Application

To apply, complete this form (approximately 10 minutes to complete). Applications are due **March 22nd by 5 PM ET**. Participants will be selected from a pool of applicants based on the experience and motivation they provide on the form. Prior participants and applicants are especially encouraged to reapply.

Accepted applicants will be notified on March 24th by 2 PM ET; applicants have until **March 27th at 9 AM 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: Participants will need to bring their own laptop to this program. A working knowledge of scripting (e.g., Shell, Python) is necessary to be successful in this event. Employment of higher level scripting or programming languages may also be useful. Applicants must be willing to commit to all three days of the event. *No financial support for travel, lodging or meals is available for this event.* Also note that the course 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.

If you are interested in having NCBI facilitate a regional hackathon hosted at your institution, please fill out this form.

GenBank release 212.0 available via FTP

Friday, February 26, 2016

GenBank release 212.0 (2/13/2016) is now available online, on the FTP site and through NCBI's programming utilities. Release 212.0 has 190,250,235 non-WGS, non-CON records containing 207,018,196,067 base pairs of sequence data. In addition, there are 333,012,760 WGS records containing 1,399,865,495,608 base pairs of sequence data, as well as 92,132,318 TSA records containing 81,932,555,094 base pairs of sequence data.

During the 61 days between the close dates for GenBank releases 211.0 and 212.0, the traditional (i.e., non-WGS, non-CON) portion of GenBank grew by 3,079,084,996 base pairs and by 1,017,310 sequence records. During that same period, 395,404 records were updated. An average of 23,159 'traditional' records were added and/or updated per day.

Between releases 211.0 and 212.0, the WGS component of GenBank grew by 101,999,877,243 base pairs and by 15,890,603 sequence records; the TSA component of GenBank grew by 4,349,215,918 base pairs and by 4,643,779 sequence records. The total number of sequence data files increased by 29 with this release. The divisions are as follows:

- BCT: 10 new files, now a total of 224
- CON: 3 new files, now a total of 330
- ENV: 1 new files, now a total of 89
- EST: 2 new files, now a total of 480
- INV: 1 new files, now a total of 136
- PAT: 4 new files, now a total of 246
- PLN: 3 new file, now a total of 125
- PRI: 4 new file, now a total of 54
- ROD: 1 less file, now a total of 31
- VRL: 1 new file, now a total of 40
- VRT: 1 new file, now a total of 61

For downloading purposes, please keep in mind that the uncompressed GenBank flat files require approximately 756 GB (sequence files only); the ASN.1 data require approximately 619 GB. More information about GenBank release 212.0 is available in the release notes and in the README files in the genbank and ASN.1 directories.

March 2nd webinar: NCBI Resources for Cancer Researchers

Wednesday, February 24, 2016

Next Wednesday, NCBI staff will discuss the facets of NCBI resources relevant to cancer in a live webinar. The databases and tools included in this overview are: BLAST, GenBank, DNA-Seq, RNA-Seq, Epigenomics and metagenomics datasets, as well as tools and APIs at NCBI that can be used to extract relevant subsets of data for cancer research.

Date and time: March 2, 2016 1:00-2:00 PM EST

Registration link: https://attendee.gotowebinar.com/register/3717666889216708353

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 find information about future webinars on this page.

Zika virus resource page provides access to nucleotide, protein sequences from latest outbreak

Tuesday, February 23, 2016

The new Zika virus resource page makes it easy to find and analyze relevant sequence data. The page includes links to the following Zika virus data at NCBI: nucleotide and protein sequences, the reference genome with updated mature peptide annotation, and publications.

In addition, a Zika database will be added to the NCBI Virus Variation resource. This database will use specialized pipelines to annotate genes, proteins and mature peptides, and standardize sample metadata. With this specialized database, you'll be able to:

- Quickly find the sequences you need, through an intuitive search interface for all viral sequences using standardized protein/gene names and metadata,
- Select the latest sequences based on date criteria or sorting of results,
- Download sequences in many formats or find links to sequences in NCBI databases, and
- Analyze sequences using multiple sequence alignments and phylogenetic trees.

Stay tuned to NCBI News or our social media channels - particularly Facebook, Twitter and LinkedIn - for updates on the specialized Zika virus database.

dbSNP Build 146 for salmon, barrel medic, cottonwood and mouse now available

Tuesday, February 23, 2016

dbSNP Build 146 data for salmon, barrel medic, cottonwood and mouse are available now on the web and FTP.

New salmon (*Salmo salar*) information:

- FTP

- Entrez
- Assembly: ICSASG_v2 (GCF_000233375.1)
- New SS: 1,342,320
- New RS: 1,029,869

New barrel medic (*Medicago truncatula*) information:

- FTP
- Entrez
- Assembly: MedtrA17_3.5 (GCF_000219495.1)
- New SS: 0
- New RS: 0

New cottonwood (*Populus trichocarpa*) information:

- FTP
- Entrez
- Assembly: Poptr1_1 (GCF_000002775.1)
- New SS: 17,902,170
- New RS: 9,505,665

New mouse (*Mus musculus) information:*

- FTP
- Entrez
- Assembly: GRCm38.p3 (GCF_000001635.23)
- New SS: 107,682
- New RS: 14,352

dbSNP, NCBI's Short Genetic Variations database, catalogs short variations in nucleotide sequences from a wide range of organisms.

Rotavirus resource uses standardized metadata and annotations, suite of tools to make it easier to search, download and analyze sequences

Friday, February 19, 2016

The new NCBI Rotavirus resource, part of the Virus Variation family of NCBI resources, provides users with a unique, metadata-driven search interface that leverages advanced data management pipelines.

Sequence annotations and descriptive metadata are mapped to a standardized vocabulary within this resource, making it much easier to find and analyze sequences of interest. Searches can also be restricted to sequences from complete genome sets or to sequence sets containing specific combinations of segments and segment genotypes.

Finally, a suite of tools allows users to build alignments and phylogenetic trees from selected sequences, and users can also download sequences with customized titles/deflines based on standardized metadata.

New video on the NCBI YouTube channel: Eukaryotic Genome Data Curation at NCBI

Friday, February 19, 2016

A recording of the January 5th webinar is now on the NCBI YouTube channel. In addition, the webinar question and answer session has been summarized in a document available on FTP.

In this webinar, three RefSeq biocurators discuss aspects of eukaryotic organism data curation. Topics covered include sequence analysis, functional annotation, data validation and community collaboration. To make it easier to locate specific sections or topics within the video, we added a table of contents at the 9 second mark.

The video is also included in the NCBI Webinars playlist, which you can save to your own Playlists collection on YouTube for quick access - just click on the plus sign marked Save.

Subscribe to the NCBI YouTube channel to receive alerts about our videos, which range from quick tips to full presentations.

For information about upcoming webinars, stay tuned to NCBI News and the Courses and Webinars page.

NCBI Insights blog post: Professors: "NCBI can help you streamline your teaching and research efforts"

Wednesday, February 17, 2016

The latest blog post on NCBI Insights points out the many tasks an NCBI account can help with, from storing and automating searches to creating bibliographic collections and more.

While this post is written to highlight how professors can gain from an NCBI account, these tips apply to everyone who signs up for an NCBI account. We encourage users to bookmark this blog post and refer to it whenever needed.

NCBI Insights is the official NCBI blog, where we share quick tips and what's new at NCBI.

New video on the NCBI YouTube channel: Viral resources at NCBI

Thursday, February 11, 2016

In the newest video on the NCBI YouTube channel, *Viral resources at NCBI*, Dr. Rodney Brister, head of the viral resources group at NCBI, gives a detailed tour of the many tools and databases publicly available to anyone studying viruses.

Subscribe to the NCBI YouTube channel to receive alerts about new videos ranging from quick tips to full presentations.

NCBI to assist Louisiana State University in South and Southeast regional genomics hackathon

Monday, February 08, 2016

From March 21st to 23rd, NCBI will assist Louisiana State University (LSU) in hosting a regional genomics hackathon in Shreveport, LA. This event is for students, postdocs, investigators and other researchers already engaged in the use of pipelines for genomic analyses from next-generation sequencing data or metadata.* Researchers and/or data scientists from the South and Southeast United States are especially encouraged to apply, but the event is open to anyone selected for the hackathon and able to travel to Shreveport.

* Some projects are available to other non-scientific developers, mathematicians or librarians.

Working groups of 5-6 individuals will be formed into five or six teams. These teams will build pipelines and tools to analyze large datasets within a cloud infrastructure. The potential subjects for this iteration are:

1. Network Analysis of Variants
2. Structural Variation
3. RNA-Seq
4. Streaming Data and Metadata
5. Neuroscience/Immunity
6. Command-line user-interface design

Please see the application for specific team projects.

Organization

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. This course will take place at the Louisiana State Health Sciences Center - Shreveport in Shreveport, Louisiana and is hosted by the two LSU institutions in town: LSU Health Sciences Center and LSU Shreveport.

Datasets

Datasets will come from the public repositories housed at the NCBI. During the course, participants will have an opportunity to include other datasets and tools for analysis. Please note, if you use your own data during the course, 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. A manuscript outlining the design and usage of the software tools constructed by each team will be submitted to an appropriate journal.

Application

To apply, complete this form, which takes approximately 10 minutes to complete. Applications are due **February 2/19/16 by 5 pm ET**. Participants will be selected from a pool of applicants based on the experience and motivation they provide on the form. Prior students and applicants are encouraged to reapply. Accepted applicants will be notified on February 22 by 2 pm ET, and have until February 26 at 9 am 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: Participants will need to bring their own laptop to this program. A working knowledge of scripting (e.g., Shell, Python) is necessary to be successful in this event. Employment of higher level scripting or programming languages may also be useful. Applicants must be willing to commit to all three days of the event. No financial support for travel, lodging or meals is available for this event. Also note that the course 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. Finally, if you are interested in having NCBI facilitate a regional hackathon hosted at your institution, please fill out this form.

Variation Viewer 1.5 adds facet toggling, updated backend data

Thursday, February 04, 2016

Variation Viewer 1.5 provides several new features, improvements and bug fixes, including facet toggling in Variant Table, updated backend data to dbSNP human build 146, dbVar (December 2015) and ClinVar (January 2016), and more. A full list of changes to Variation Viewer is available in the release notes.

Variation Viewer is a tool for navigating variant data in dbSNP, dbVar and ClinVar in a genomic context.

February 17th webinar: "Five ways to submit next-gen sequencing data to NCBI's Sequence Read Archive (SRA)"

Wednesday, February 03, 2016

In two weeks, NCBI will present a webinar on SRA submissions, discussing five different ways to submit next-generation sequencing data to SRA, including the new SRA submission portal (beta) which allows you to submit data via FTP and the Aspera command line, Illumina's BaseSpace, MOTHUR, and the iPlant Collaborative.

Date and time: February 17, 2016 1:00-2:00 PM EST

Registration link: https://attendee.gotowebinar.com/register/6510651823186558978

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 find information about future webinars on this page.