NCBI News, September 2016

October 11th NCBI Minute: BLAST+ 2.5.0 with Support for HTTPS, accession.version Identifiers and Much More

Friday, September 30, 2016

On October 11th, the NCBI Minute will be a discussion of changes made to the BLAST standalone distribution due to the switch to HTTPS and the transition of the sequence databases to accession.version as the primary identifier.

Date and time: Tuesday, October 11, 2016 12:00 PM – 12:30 PM EDT

Registration URL: https://attendee.gotowebinar.com/register/6522718969008808193

The new BLAST+ 2.5.0 release supports both, and provides support for composition-based statistics with RPSTBLASTN, and has a new taxonomic organism report. You will learn how these changes improve your BLAST searches and the analysis of results.

After registering for the webinar, 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.

BLAST+ 2.5.0 released with support for HTTPS, accession.version and more

Friday, September 30, 2016

The new version of the BLAST+ executables offers support for HTTPS, accession.version as the primary sequence identifier, support for composition-based statistics with RPSTBLASTN, and a new taxonomic organism report. See more details about these updates on BLAST News. A full list of new features, improvements and bug fixes is available in the release notes.

October 5th webinar: NCBI at ASHG 2016

Friday, September 30, 2016

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

Next Wednesday, October 5th, NCBI staff will give a brief overview of our activities at this year's ASHG meeting related to ClinVar, dbGaP, GRCh38 and other topics, and how these will benefit ASHG attendees.

Date and time: Wednesday, October 5, 2016 1:00 PM - 2:00 PM EDT

Registration URL: http://bit.ly/2dKrCZj

The ASHG annual meeting will happen October 18-22 in Vancouver, British Columbia, Canada.

After registering for the webinar, 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.

Sequence Viewer 3.16 is now available

Thursday, September 29, 2016

Sequence Viewer 3.16 brings several new features, improvements and bug fixes to the graphical viewer, including improved compliance with HTTPS protocol requirements, a new track type, and improved labeling to avoid label duplication. For a full list of changes, see the release notes.

Sequence Viewer is a graphical view of sequences and color-coded annotations on regions of sequences stored in the Nucleotide and Protein databases.

GenBank release 215.0 is now available via FTP

Wednesday, September 28, 2016

GenBank release 215.0 (08/19/2016) has 196,120,831 traditional records containing 217,971,437,647 base pairs of sequence data. In addition, there are 359,796,497 WGS records containing 1,637,224,970,324 base pairs of sequence data, as well as 113,179,607 TSA records containing 103,399,742,586 base pairs of sequence data.

During the 66 days between the close dates for GenBank releases 214.0 and 215.0, the traditional portion of GenBank grew by 4,770,529,828 base pairs and by 1,657,259 sequence records. During the same period, 75,882 records were updated at an average of 26,260 traditional records added and/or updated per day.

Between releases 214.0 and 215.0, the WGS component of GenBank grew by 81,049,025,676 base pairs and by 9,518,416 sequence records. The TSA component of GenBank grew by 8,985,783,667 base pairs and by 8,502,546 sequence records.

The total number of sequence data files increased by 57 with this release. The divisions are as follows:

- BCT: 18 new files, now a total of 267
- CON: 7 new files, now a total of 351
- ENV: 1 new file, now a total of 93
- EST: 1 new file, now a total of 481
- GSS: 1 new file, now a total of 301
- HTG: 2 new files, now a total of 153
- INV: 3 new files, now a total of 144
- PAT: 11 new files, now a total of 263
- PLN: 8 new files, now a total of 134
- PRI: 2 new files, now a total of 55
- SYN: 1 new file, now a total of 9
- TSA: 1 new file, now a total of 229
- VRL: 1 new file, now a total of 48

For downloading purposes, please keep in mind that the uncompressed GenBank release 215.0 flatfiles require approximately 790 GB (sequence files only); the ASN.1 data require approximately 650 GB.

More information about GenBank release 215.0 is available in the release notes.

Genomes-announce listserv reactivated

Wednesday, September 28, 2016

The Genomes-announce email list has been reactivated. To sign up for the Genomes-announce Listserv, go to https://www.ncbi.nlm.nih.gov/mailman/listinfo/genomes-announce.

We will use the Genomes-announce Listserv primarily to announce changes to the NCBI Genomes FTP site and other genome data download channels. Announcements will be made about new data types, changes to data formats or data organization.

Read more about the Genomes-announce Listserv in this announcement.

NCBI to hold Developers' Forum September 28th

Friday, September 23, 2016

On Wednesday, September 28th, at 3 PM EDT, NCBI will hold a developers' forum for those who use large amounts of NCBI data. The forum will help us provide you with better access to NCBI data.

To join the forum, complete this short survey. An invitation will be extended to 21 people.

October 4-6: Stream the University of Michigan NCBI workshops

Thursday, September 22, 2016

On October 4th, 5th, and 6th, the University of Michigan's Taubman Health Sciences Library will host a series of NCBI workshops that can also be streamed remotely. The workshops are: *Navigating NCBI's Molecular Data Using the Integrated Entrez System and BLAST*, *A Practical Guide to NCBI BLAST* and *EDirect: Command Line Access to NCBI's Biomolecular Databases*. Please see the Taubman Health Sciences Library Remote Site Registration Page for details.

Introducing Magic-BLAST

Thursday, September 22, 2016

Magic-BLAST is a new tool for mapping large sets of next-generation RNA or DNA sequencing runs against a whole genome or transcriptome. Magic-BLAST executables for LINUX, MacOSX, and Windows as well as the source files are available on the FTP site.

Each alignment optimizes a composite score, taking into account simultaneously the two reads of a pair, and in case of RNA-Seq, locating the candidate introns and adding up the score of all exons. Sequencing reads can be provided as NCBI SRA accessions, FASTA or SRA files.

Magic-BLAST implements ideas developed in the NCBI Magic pipeline using the NCBI BLAST libraries. Magic-BLAST is under active development, and we expect the next few releases to occur on a monthly basis. Read more about Magic BLAST on the FTP site.

Scheduled: Next Round of HTTPS Tests

Wednesday, September 21, 2016

We have scheduled another round of HTTPS tests, following up from the initial tests performed on September 15. More information can be found on an NCBI Insights Blog post.

The schedule for these tests is as follows (all times are EDT):

Thursday, Sept 22
8:00 AM – 12:00 PM : redirect web pages from HTTP to HTTPS, same as the previous September 15 test
8:00 AM – 9:00 AM : redirect CGI's and API calls to HTTPS where possible, reject where not possible

Monday, Sept 26
8:00 AM – 10:00 AM : redirect web pages from HTTP to HTTPS with HSTS activated

using a 1-hour expiration
10:00 AM – 12:00 PM : redirect web pages from HTTP to HTTPS without HSTS

Tuesday, Sept 27
8:00 AM : Start continually redirecting web pages from HTTP to HTTPS

NCBI's Bryant and Bolton receive 2016 Herman Skolnik Award for PubChem database

Monday, September 19, 2016

On August 23, Drs. Stephen Bryant and Evan Bolton received the American Chemical Society (ACS) 2016 Herman Skolnik Award for their work in developing, maintaining, and expanding the National Center for Biotechnology Information's PubChem database of chemical substances and their biological activities. The award was presented at the ACS 252nd National Meeting & Exposition in Philadelphia.

The Herman Skolnik award is named after its first recipient, the founder of the Journal of Chemical Information and Computer Sciences, and "recognizes outstanding contributions to and achievements in the theory and practice of chemical information science and related disciplines," according to ACS.

In its announcement of the award, ACS said: "Under Bryant and Bolton's leadership, the PubChem team has created a world-class resource for chemical and biological information. PubChem is the first major public database to connect cheminformatics to bioinformatics and thereby provide a unique information resource for pharmaceutical research."

Introduced in 2004, PubChem currently includes more than 220 million chemical substance records for 90 million unique compounds. The database also contains biological screening results from more than 1.2 million bioassays for over 3.5 million tested substances. The information in PubChem is the result of collaborations with more than 250 academic and commercial organizations that have contributed their data. PubChem is integrated with many other NCBI other databases, with links to related information, such as compounds with similar structures, protein sequences, and relevant journal articles. This extensive network of links provides users with vast opportunities for exploration and for making discoveries. Each day, tens of thousands of researchers from university labs and pharmaceutical and biotech companies access PubChem.

September 21st webinar: Update on NCBI's Transition to HTTPS

Monday, September 19, 2016

Next Wednesday, September 21st, NCBI staff will discuss plans regarding the move to HTTPS-only services. This past week, we conducted the first of a series of HTTPS tests; in this webinar, we will talk about this and future tests that will help all of us prepare for this

Figure 1. Drs. Bryant and Bolton receive the American Chemical Society 2016 Herman Skolnik Award.

change. We will also briefly discuss circumstances surrounding proxy services and software dependent upon NCBI software, such as the SRA and C++ toolkits.

Date and time: Wednesday, September 21, 2016 12:00 PM EDT

Registration URL: https://attendee.gotowebinar.com/register/2680765856528138244

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.

New video on YouTube: Tree Viewer - Display Large Trees

Thursday, September 15, 2016

The newest video on the NCBI YouTube channel, *Tree Viewer: Display Large Trees*, demonstrates a new functionality in Tree Viewer - the ability to display much larger trees.

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

Genomes FTP site update (version 1.3) adds new data formats and more

Wednesday, September 14, 2016

NCBI has released a comprehensive update of all current genome assemblies in the Genomes FTP site, affecting data reported in the /genomes/genbank/, /genomes/refseq/ and /genomes/all/ FTP directories. This update adds nucleotide FASTA sequences of CDS and RNA features computed from the genome sequence, expands the scope of the /genbank/ data to include metagenomes, and more. The FTP content of nearly all "latest" GenBank and RefSeq assemblies was updated to reflect these changes between 5/11/2016 and 6/24/2016.

Genomes FTP version 1.3 includes the following changes:

- New files for genome assemblies with annotation:
  - Files named as *_cds_from_genomic.fna.gz provide nucleotide FASTA sequences corresponding to all CDS features annotated on the assembly, based on the genome sequence
  - Files named as *_rna_from_genomic.fna.gz provide nucleotide FASTA sequences corresponding to all RNA features annotated on the assembly, based on the genome sequence
- Metagenomes
  - Metagenomes have been added to the Assembly database and a new genome group directory (metagenomes) is now available: ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/metagenomes/
- Annotation hashes:
  - Reporting hash values and last changed dates for different aspects of the annotation data are useful to monitor for when annotation has changed in a way that is significant for a particular use case and warrants downloading the updated records. These data can be used to monitor for annotation changes compared to previously downloaded files based on comparison of the hash values, or for annotation changes since a particular point in time
  - A file named annotation_hashes.txt in each genome assembly directory provides hash values and last changed dates for different aspects of the annotation and descriptor data for that assembly
  - A file named annotation_hashes.txt in each organism group and species directory provides hash values and last changed dates for all assemblies from the organism group or species
- RepeatMasker files:

    - Data in the *_rm.out files is now generated using a newer version of RepeatMasker and repeat libraries (change from RepeatMasker version 3.3.0 to 4.0.6 and from RM database version 20120418 to version 20150807)
- GFF3 format:
    - genomic.gff.gz files for RefSeq eukaryotic genomes annotated with NCBI's Eukaryotic Genome Annotation Pipeline now incorporate 1-2 bp gaps or "micro-introns" to compensate for frameshifting indels in mRNA and CDS features where the indel is thought to represent a genome sequencing error and the gene is likely to produce a functional product
    - Some features are now represented with better or more appropriate Sequence Ontology (SO) terms in the GFF3 files, including the following:

#INSDC term	old SO term	new SO term
misc_feature	region	sequence_feature
variation	sequence_variant	sequence_alteration
conflict	region	sequence_conflict
mobile_element	region	mobile_genetic_element
rep_origin	region	origin_of_replication
telomere	region	telomere
centromere	region	centromere
regulatory/enhancer	region	enhancer
regulatory/promoter	region	promoter
GC_signal	GC_rich_region	GC_rich_promoter_region
N_region	region	N_region
S_region	region	S_region
V_region	region	V_region
unsure	region	sequence_uncertainty
virion	region	viral_sequence

- Feature table report:
    - Small improvements in the feature_table.txt.gz file, including fixing the occurrence of "?" strand
- GBFF format:
    - Small improvements in formatting of the GenBank flatfiles
- Assembly summary files:
    - A column reporting "Excluded from RefSeq" reasons has been added
- Assembly reports:
    - More metadata fields were added to the headers of the assembly, statistics and regions reports
    - Headers no longer include the the latest/suppressed/replaced status of the assembly

- Gzip compression:
  - A subtle change in gzip compression that has no effect on file contents but does subtly alter file sizes and md5checksums

Additional information about the genomes FTP site came 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.

RefSeq release 78 is now available

Monday, September 12, 2016

RefSeq release 78 is accessible online, via FTP and through NCBI's programming utilities. This full release incorporates genomic, transcript, and protein data available as of September 6, 2016 and contains 107,045,797 records, including 70,427,238 proteins, 16,172,490 RNAs, and sequences from 62,739 organisms. The release is provided in several directories as a complete dataset and also as divided by logical groupings.

More information about release 78 can be found in the release notes. For more information about the RefSeq project, please see the RefSeq homepage.

New on NCBI Insights: The Future of Existing GI Numbers at NCBI

Monday, September 12, 2016

The latest blog post on NCBI Insights discusses what will happen to existing GI numbers in records now that NCBI is phasing out Sequence GIs.

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

Leiden Open Variation Database to be retired September 30, 2016

Thursday, September 08, 2016

NCBI is retiring the Leiden Open Variation Database (LOVD) on September 30, 2016. LOVD has been used to capture information about novel human variants. We encourage past submitters of human genetic variations to LOVD to transfer their information to the ClinVar database.

If you would like to add new human variation data, please review our instructions on submitting to ClinVar. You may find our submission wizard eases the process.

While the LOVD site will be retired on September 30, 2016, an FTP archive will continue to store LOVD data for download after this date.

Figure 1. The PubMed Journals homepage.

New on NCBI Insights: Find, Browse and Follow Biomedical Literature with PubMed Journals

Wednesday, September 07, 2016

The latest blog post on NCBI Insights presents the new PubMed Journals, the latest experiment from PubMed Labs. PubMed Journals allows you to easily find and browse journals of interest, browse new articles, and more. Learn more about PubMed Journals, try it out, and leave us feedback on the blog post.

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

September 7th webinar: The E-Utilities in an Age without GI Numbers

Thursday, September 01, 2016

Next Wednesday, September 7th, NCBI will present a webinar that briefly describes NCBI's future plans for the E-utilities API in a time where GI numbers are no longer used as the primary identifiers for sequence records. You will learn how to convert GI numbers

to accession.version identifiers and how to quickly determine the most recent version of an accession. You'll also learn about a new E-utility parameter (to be released this fall) that allows these tools to work only with accession.version identifiers.

Date and time: Wednesday, September 7, 2016 12:00 PM EDT

Registration URL: https://attendee.gotowebinar.com/register/4529251610671340033

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.

October 24-26: Hackathon at Cold Spring Harbor Laboratory

Thursday, September 01, 2016

From October 24th to 26th, Cold Spring Harbor Laboratory (CSHL), with assistance from NCBI, will host a biomedical data science hackathon immediately before the Biological Data Science Conference at CSHL. The hackathon will primarily focus on writing functional software for advanced bioinformatics analysis of next generation sequencing data and metadata, but also may include analysis of other types of data, such as images or other molecular measurements.

This event is for students, postdocs and investigators or other already engaged in the creation of pipelines for genomic analyses from next generation sequencing data, imaging data or metadata.* The event is open to anyone selected for the hackathon and willing to travel to CSHL.**

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

** Attendees of the CSHL Biological Data Science Conference will be given preference, if space is limited. Also, there will be a nominal fee for attendees, partly to cover refreshments during the events.

Organization

Working groups of 5-6 individuals will be formed into four to five teams. These teams will build novel pipelines, tools, and visualizations to analyze large datasets within a cloud infrastructure. The potential subjects for this hackathon include:

- structural variant identification & analysis,
- genome assembly,
- single cell transcriptome & epigenome analysis,
- ultrafast genomic mapping,
- data encryption,
- deep learning technologies,

- and image analysis.

Please see the application for specific team projects.

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

Most of the datasets will come from the public repositories, primarily those 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 may be submitted to an appropriate journal such as the F1000Research hackathons channel. Continued development of the technology after the hackathon is also encouraged.

Application

To apply, complete this form (approximately 10 minutes to complete). Applications are due **September 16, 2016 by 4 PM ET**. Prior participants and applicants are especially encouraged to reapply.

The first round of accepted applicants will be notified on September 19th by 5 PM ET, and have until September 22nd 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.

Notes

Participants will need to bring their own laptop to this program. 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.

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 is currently available, although attendees will be notified if that changes.

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.