NCBI News, February 2017

NCBI Insights | PubMed Citations: A New, Faster Process for Correcting Errors

Tuesday, February 28, 2017

The latest blog post on NCBI Insights introduces users to the PubMed Data Management System (PMDM), which allows publishers to correct PubMed citation data directly. Authors should contact journal publishers to correct PubMed citation mistakes.

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

March 1st NCBI Minute: Setting up new data alerts with MyNCBI

Thursday, February 23, 2017

Next Wednesday, March 1, 2017, NCBI will present a short webinar that will show you have to use MyNCBI alerts to be notified when new citations of interest appear in traditional sequence databases, as well as SRA and GEO.

Date and time: Wednesday, March 1, 2017 12:00 PM – 12:15 PM EST

Register

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.

Bottlenose dolphin annotation release 101

Wednesday, February 22, 2017

Annotation Release 101 for the bottlenose dolphin (Tursiops truncatus) is out in RefSeq! This annotation was based on the NIST Tur_tru v1 assembly, which has a four-fold increase in contiguity from the assembly used in the previous annotation. Over four billion RNA-Seq reads from skin and blood tissue were used for gene prediction. As a result of these improvements, the percent of partially-represented protein-coding genes went down from 24% to 4%. Over 2500 genes that were fragmented in the previous assembly were merged into complete genes. A total of 24,026 genes were annotated, and 17,096 of them were protein-coding. A full report on the annotation can be found here.

Figure 1. *Tursiops truncatus*, **the bottlenose dolphin.**

This improved genomic resource for the dolphin will allow NIST to develop standardized research methods, produce reference data and tools, and perform bioanalytical measurements on dolphins and other marine organisms. Dolphins are important sentinel organisms for the health status of the marine environment and their study expands knowledge on cognition, communication, acoustics, conservation, and hydrodynamics.

Annotation Release 101 is available for download and formatted for BLAST searches.

New video on YouTube: Embed the NCBI Sequence Viewer into Your Pages

Tuesday, February 21, 2017

The newest video on the NCBI YouTube channel introduces the Sequence Viewer embedding API. A few quick examples illustrate how easy it is to embed Sequence Viewer into your own pages.

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

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

NLM Webinar series: "Insider's Guide to Accessing NLM Data: EDirect for PubMed"

Friday, February 17, 2017

Beginning February 21, 2017, the National Library of Medicine (NLM) will present the three-part webinar series "Insider's Guide to Accessing NLM Data: EDirect for PubMed."

This series of workshops will introduce new users to the basics of using EDirect to access exactly the PubMed data you need, in the format you need. Over the course of three 90-minute sessions, students will learn how to use EDirect commands in a Unix environment to access PubMed, design custom output formats, create basic data pipelines to get data quickly and efficiently, and develop simple strategies for solving real-world PubMed data-gathering challenges. No prior Unix knowledge is required; novice users are welcome!

This series of classes involves hands-on demonstrations and exercises, and we encourage students to follow along. Before registering for these classes, we strongly recommend that you:

- Watch the first Insider's Guide class "Welcome to E-utilities for PubMed", or be familiar with the basic concepts of APIs and E-utilities.
- Be familiar with structured XML data (basic syntax, elements, attributes, etc.)
- Have access to a Unix command-line environment on your computer (for more information, see our Installing EDirect page).
- Install the EDirect software (for more information, see our Installing EDirect page).

Due to the nature of this class, registration will be limited to 50 students per offering.

Registration is currently open for the February/March 2017 series:

- Part 1: Getting PubMed Data: Tuesday, February 21, 1-2:30 PM ET
- Part 2: Extracting Data from XML: Tuesday, February 28, 1-2:30 PM ET
- Part 3: Building Practical Solutions: Tuesday, March 7, 1-2:30 PM ET

Students are expected to attend Part 1, Part 2, and Part 3 in a single series.

To register, and for more information, visit https://goo.gl/DVOh6M.

Tree Viewer version 1.12 implements new API to markup trees

Tuesday, February 14, 2017

Tree Viewer version 1.12 has several improvements, updates and bug fixes, including a new API, PDF rendering, and Tree Viewer macro language. The Tree Viewer release notes list all updates.

NCBI Tree Viewer is a tool for viewing your own phylogenetic tree data.

Interim annotation updates for the human GRCh37p.13 and GRCh38.p10 assemblies

Tuesday, February 14, 2017

Updates to the annotation of the human GRCh37.p13 and GRCh38.p10 assemblies are now available for download by anonymous FTP. These annotation updates contain features projected from the current known RefSeq transcripts and curated genomic sequences (with accession prefixes NM_ or NR_, and NG_ respectively) placed on either the GRCh37.p13 or GRCh38.p10 assembly.

The GRCh37.p13 annotation is being provided to help support members of the clinical community who are still dependent on the old GRCh37 (hg19) assembly. However, users should be cautious about using these annotation results, especially in regions that were extensively revised in GRCh38. See the corresponding README file for more details including details on genes that are no longer annotated in the update.

The two annotations started with the same set of RefSeq transcripts, and differences in which RefSeqs are annotated reflect improvements in the GRCh38 assembly, as well as some genes and functionally distinct alleles that were relocated between the chromosomes of the primary assembly and alt loci scaffolds.

Annotation is available in GFF3 format, as well as alignments of current RefSeq transcripts to the genome in both GFF3 and BAM formats. The annotations are also available in NCBI's genome browsers such as Variation Viewer and 1000 Genomes Browser, including in either the "Genes" recommended track set or from the track selection dialog (search for "interim").

Please send questions, comments, and suggestions concerning these updates to refseq-admin@ncbi.nlm.nih.gov or use the Feedback link from Entrez Gene reports.

February 22nd webinar: Introducing the Multiple Sequence Alignment Viewer

Monday, February 13, 2017

Next Wednesday, February 22, 2017, NCBI will present a webinar on the Multiple Sequence Viewer (MSAV). In this webinar, you will learn how to display alignment data from many sources, including NCBI BLAST results, as well as precomputed multiple alignments of your own data. You will also see how to embed the viewer in your own web pages or share a link to a particular alignment display.

Date and time: Wednesday, February 22, 2017 12:00 PM – 12:30 PM EST

Registration URL: https://attendee.gotowebinar.com/register/7663489773270563843

The MSAV is a versatile web application that helps you visualize and interpret multiple sequence alignments for both nucleotide and protein sequences. You can use the viewer to explore sequence conservation, investigate variation or troubleshoot assembly or sequencing errors.

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.

SmartBLAST updated to provide more information, database matches

Monday, February 13, 2017

The SmartBLAST service has recently been updated to emphasize matches to the landmark database, which comprises the proteomes from 26 well-curated genomic assemblies. The display also now presents more information about conserved domains and details about the query.

SmartBLAST quickly finds the closest relatives to a protein query and evaluates the phylogenetic relationship among the query and matched sequences. You can start a SmartBLAST search from the SmartBLAST page or the BLAST home page. Read more about SmartBLAST on NCBI Insights.

Sequence Viewer 3.19 is now available

Monday, February 13, 2017

Sequence Viewer 3.19 has several new features, improvements and bug fixes, including a new aggregate track type, improved display of projected features and cleaned alignments, and a new manual for using embedded API. For a full list of changes, see the Sequence Viewer 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.

New NCBI Insights post: New Web Services for Comparing and Grouping Sequence Variants

Thursday, February 09, 2017

The latest post on the NCBI Insights blog introduces new web services for comparing and grouping variants. Geneticists, dataflow engineers, and anyone who needs to compare genetic variants can use these services.

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

NCBI to host genomics hackathon March 20-22

Thursday, February 02, 2017

From March 20th to 22nd, the NCBI will host a genomics hackathon on the NIH campus. To apply for this hackathon, complete this application (approximately 10 minutes to complete). Applications are due **February 22nd by 1 PM ET.**

This hackathon will primarily focus on advanced bioinformatics analysis of next generation sequencing data and metadata. This event is for students, postdocs and investigators or other researchers already engaged in the use of genomics data or pipelines for genomic analyses from next generation sequencing data. However, there are some projects available to other non-scientific developers, mathematicians or librarians. The event is open to anyone selected for the hackathon who is able to travel to NIH.

Organization

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

The potential subjects for this iteration are:

- GA4GH - NCBI API Integration,
- Global Screening Arrays,
- Graph Genome Information Extraction,
- Single Cell Methylation Data,
- Generation of an automated gff3 parser,
- integration of Immport metadata with specific genomic datasets,
- And several others.

Please see the application for specific and evolving 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

Datasets will come from 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.

Application

To apply, complete this application (approximately 10 minutes to complete). Applications are due **February 22nd by 1 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.

The first round of accepted applicants will be notified on February 24th by 5 pm ET, and have until February 27th at 1 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.

Notes

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.