U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: NCBI News [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 1991-2012. NCBI News, May 2015.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/

NCBI News, May 2015

New NCBI YouTube video: "Genome Workbench: Import BAMs and Export Alignments"

Friday, May 29, 2015

This video on the NCBI YouTube channel shows you how to import BAM files, create a BAM file index, and export selected alignments using NCBI's Genome Workbench.

Subscribe to the NCBI YouTube channel to be notified of our new videos, which range from quick tips to full webinar presentations.

June 3rd webinar: "Troubleshooting GenBank Submissions: Coding Region Annotation"

Thursday, May 28, 2015

Next Wednesday, June 3rd, NCBI staff will show you how to troubleshoot internal stop codon errors encountered during coding region (CDS) annotation. The source of this problem can be in (1) improper frame/strand, or genetic code designation or (2) poor sequence quality. You will learn how to analyze your sequences and uncover problems with BLAST prior to submitting them to GenBank.

To sign up for this webinar, click here. Like all of our webinars, this will be posted on the NCBI YouTube account after the live presentation; you can subscribe to our YouTube channel to be notified of all our new videos.

To see upcoming webinars, as well as related materials and recordings from past webinars, please see the NCBI Webinars page.

NCBI to hold three-day genomics hackathon in August

Wednesday, May 27, 2015

From August 3-5, NCBI will host its second genomics hackathon focusing on advancing bioinformatics analysis of next generation sequencing data. This event is for students, postdocs and investigators already engaged in the use of pipelines for genomic analyses from next generation sequencing data.* Working groups of 5-6 individuals will be formed for twelve teams, in three sections. These groups will build pipelines to analyze large datasets within a cloud infrastructure. The sections for this iteration are:"RNA-Seq Normalization for Every Biologist", "Translational Genomics", and "Democratization of Genomics". Please see the application for specific team projects.

* Specific projects are available to other developers or mathematicians.

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 on or near the NIH main campus in Bethesda, Maryland.

Datasets

Datasets will come from the public repositories housed at 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 of the hackathon and describing participant processes, products and scientific outcomes will be submitted to an appropriate journal. A pre-print of the manuscript from the January NCBI/ADDS hackathon is available from bioRxiv.

Application

To apply, complete this form (approximately 10-15 minutes to complete). Applications are due **June 6th by 3pm Eastern time**. Participants will be selected from a pool of applicants; prior students and applicants will be given priority in the event of a tie. Please note: applicants are judged based on the motivation and experience outlined in the form itself. Accepted applicants will be notified on June 18th, by 2pm Eastern time, and have until June 22 at 5pm Eastern time to confirm their participation. 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 can be provided 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.

New NCBI YouTube Video: NCBI's Tree Viewer

Wednesday, May 27, 2015

This short video, "NCBI's Tree Viewer" on the NCBI YouTube channel is an introduction to Tree Viewer, a tool for viewing your own phylogenetic tree data. Tree Viewer is customizable and can be embedded in a wide variety of web pages.

Subscribe to the NCBI YouTube channel to be notified of our new videos, which range from quick tips to full webinar presentations.

New NCBI Insights blog post: "NCBI's First Hackathon: Advanced Bioinformatic Analysis of Next-Gen Sequencing Data"

Friday, May 22, 2015

In the latest blog post on NCBI Insights, we discuss the genomics hackathon NCBI hosted earlier this year, in conjunction with the NIH Office of Data Science. The goal was to have experienced genomics professionals create efficient pipelines for people who are new to this field.

Visit NCBI Insights, the official NCBI blog, for posts on what's new at NCBI, quick tips for using our tools and databases, and science feature stories.

May 26th webinar: "The NCBI Minute: Prokaryotic Genome Annotation Update"

Thursday, May 21, 2015

The next NCBI Minute on Tuesday, May 26th will cover recent improvements to the way we annotate and manage RefSeq bacterial and archaeal genomes at NCBI. We'll introduce you to the new annotation paradigm, provide tips on adapting your workflow, and point out how to find help and more information.

To sign up for this brief webinar, navigate to: https://attendee.gotowebinar.com/register/1585449954415535618.

The NCBI Minute is a series of short webinars that give a brief introduction to a specific topic or NCBI tool. To see upcoming webinars, as well as summaries, recordings on YouTube, and related materials from past webinars, please see the NCBI Webinars page.

New NCBI Insights blog post: "NCBI RefSeq's Antimicrobial Peptide Indexed Field: Facilitating Novel Antibiotic Discovery"

Thursday, May 21, 2015

The latest blog post on NCBI Insights introduces the RefSeq "*Protein has antimicrobial activity [prop]*" indexed field, which retrieves curated sequence annotations showing naturally occuring antimicrobial peptides (AMPs), making it easier for researchers to identify alternatives to traditional antibiotics.

Visit NCBI Insights, the official NCBI blog, for posts on what's new at NCBI, quick tips for using our tools and databases, and science feature stories.

Export data into Genome Workbench with Tree Viewer version 1.4

Tuesday, May 19, 2015

NCBI Tree Viewer 1.4 implements several new features, improvements and bug fixes, including an updated Download function, which now allows you to export data into Genome Workbench; you can also upload custom user-defined data in ASN.1 and Newick formats. To see the full list of updates, see the Tree Viewer release notes.

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

June 9th hands-on workshops at NLM will show users how to search NCBI's molecular databases

Monday, May 18, 2015

On June 9th, 2015, NCBI will present two workshops on searching NCBI molecular databases: "Accessing Genomes, Assemblies and Annotation Products" and "Human Variation and Medical Genetics Resources". Workshops will be held at the Lister Hill Center (Building 38A) Auditorium on the NIH campus.

NOTE: Participants must provide their own WiFi-ready laptop with a standard Web browser installed.

Accessing Genomes, Assemblies and Annotation Products

9am-12pm

You will learn how NCBI processes genome-level data and produces annotation through the prokaryotic and eukaryotic genome annotation pipelines. You will find, browse, and download genome-level data for your organism of interest and for environmental and organismal metagenomes using the Genome, BioProject and Assembly resources. In addition to assembled and annotated data, you will retrieve and download draft whole genome shotgun and next-generation sequencing data from the Nucleotide and Sequence Read Archive (SRA) databases. You will access results of precomputed analyses of genomes, as well as perform your own analyses of assembled and unassembled genomic data using NCBI's genome BLAST and SRA-BLAST services.

Accessing NCBI Human Variation and Medical Genetics Resources

1pm-4pm

You will learn to use and access resources associated with human sequence variations and phenotypes associated with specific human genes and phenotypes. The workshop will emphasize the Gene, MedGen and ClinVar resources to search by gene, phenotype and variant respectively. You will learn how to map variation from dbSNP and dbVar onto genes, transcripts, proteins and genomic regions and how to find genetic tests in GTR. You will also gain experience using additional tools and viewers including PheGenI, a browser for genotype associations, as well as the new Variation Viewer and the 1000 Genomes Browser. All of these provide useful ways to search for, map, and browse variants.

Register for one or both workshops at https://www.surveymonkey.com/s/W2ZPW6D.

If you have any questions, contact **courses@ncbi.nlm.nih.gov**. To see more educational offerings from NCBI, please visit the Learn page on our website.

Genome Workbench 2.9.0 now available

Thursday, May 14, 2015

As of May 5th, Genome Workbench 2.9.0 is available. New features include custom selections and search support for Tree Viewer, as well as improvements to Graphical Sequence View. For the full list of fixes, improvements and features, see the Genome Workbench release notes.

New NCBI Insights blog post - Accessing the Hidden Kingdom: Fungal ITS Reference Sequences"

Monday, May 11, 2015

NCBI staff, in collaboration with outside mycology experts, are curating a set of fungal sequences from internal transcribed spacer (ITS) regions of nuclear rRNA genes. These ITS sequences are especially useful for identifying and classifying fungal species by morphology, a difficult process when using traditional methods.

Read more about this fungal RefSeq Targeted Loci BioProject on the NCBI blog, NCBI Insights. To receive notice of new blog posts, you can sign up to the RSS feeds by clicking on the RSS links in the column on the right; you can also click the Follow tab that appears on the bottom of the screen when you visit NCBI Insights.

RefSeq release 70 is now available with re-annotated bacterial genomes for uniformity across genomes and species

Thursday, May 07, 2015

The full RefSeq release 70 is now available online, on the FTP site, and through NCBI's programming utilities, with 74,720,563 records describing 50,351,119 proteins, 11,310,700 RNAs, and sequences from 54,118 different organisms.

This release reflects a large update of complete bacterial RefSeq genomes, proteins and genes. In order to make genome annotation comparable across genomes and species, NCBI has re-annotated all RefSeq prokaryotic genomes using NCBI's genome annotation pipeline. Previously, it was possible that the same gene, in the same species, with an identical sequence for the gene's genomic region might be annotated with a different protein, simply because it was annotated using different methods. Now, the same gene in the same species with the same sequence will be annotated with exactly the same protein in RefSeq. If you'd like to learn more about the re-annotation project and what NCBI is doing to help you transition to using this new data, please see the RefSeq Re-annotation Project page.

In addition, each annotated CDS used to be tracked with a distinct RefSeq protein accession number. However, due to identical protein sequences being found on multiple re-annotated RefSeq genomes and extensive bacterial genome sequencing, the RefSeq prokaryotic protein dataset rapidly became very redundant. Rather than flood the protein database with thousands of completely identical proteins, NCBI has adopted the use of non-redundant WP proteins for RefSeq prokaryotic genomes annotated with NCBI pipelines, which we first announced in June 2013.

Now, if the identical protein sequence appears on more than one RefSeq genome, NCBI simply reuses the existing WP accession number instead of creating a new accession for each new occurrence and genome. As a result, over 7 million proteins were removed, significantly reducing protein redundancy for the prokaryotic dataset. Removed accessions are reported in release70.removed-records.gz and a supplemental data mapping file is available in the release-catalog directory (release70.bacterial-reannotation-report.txt.gz).

Here are some measures for four species that illustrate the significant reduction in protein record redundancy resulting from the use of non-redundant RefSeq proteins (WP_accessions).

Counts of annotated proteins:

Species	Genomes	Total Proteins	Total Unique WPs	Total Singleton WPs
Staphylococcus aureus	4,194	11,764,898	222,588	138,284
Escherichia coli	2,685	13,637,370	1,033,617	649,100
Mycobacterium tuberculosis	1,790	7,245,836	139,800	101,255
Salmonella enterica	918	4,099,013	294,106	194,982

Percent reduction in protein accessions:

Species	Genomes	Percent Reduction (WPs)	Percent Singleton WPs
Staphylococcus aureus	4,194	98%	62%
Escherichia coli	2,685	94%	63%

Table continued from previous page.

Mycobacterium tuberculosis	1,790	98%	72%
Salmonella enterica	918	93%	66%

Singletons per Genome:

Species	Average Protein Count	Singleton WPs per Genome	Percent Singleton per Genome
Staphylococcus aureus	2,814	33	1.17%
Escherichia coli	5,088	241	4.74%
Mycobacterium tuberculosis	4,046	56	1.38%
Salmonella enterica	4,485	212	4.72%

Definitions:

- "Total Proteins" counts the number of times non-redundant proteins accessions are annotated on the set of genomes for the species.
- "Total Unique WPs" counts the distinct number of non-redundant proteins used across all genomes. This is the truly non-redundant set of proteins for the species.
- "Total Singleton WPs" counts the number of non-redundant proteins used only once in the set of genomes for the species.
- "Percent Reduction" measures the compression in protein identifier space gained by using non-redundant protein accessions (WP_ prefix).
- "Percent Singleton WPs" measures the percent of all non-redundant proteins for that species that are used only once in that species.

If you'd like to learn more about non-redundant proteins and see an example of this new RefSeq protein record, please see the RefSeq non-redundant proteins page.

This is a first step toward managing data in a world where genomes are sequenced for assays, rather than to discover novel proteins. We appreciate that this is a new and major change for RefSeq prokaryotic genomes, but it is a necessary change to make as the number of disease-outbreak and other isolate sequencing continues to rapidly increase.

Protein records

In all bacterial genomes, except reference genomes and a small number which have yet to be re-annotated, protein accessions NP/YP have been replaced with non-redundant protein accession numbers (WP_).

- Over 7 million bacterial YP_ and NP_ RefSeq proteins were suppressed as complete bacterial genomes were re-annotated to conform to the new data model.
- Nearly 1 million non-redundant protein records were updated in March and April 2015 to improve protein names. These updates affected CDS "/product=" annotation details for all (>31,000) of the RefSeq bacterial genomes and included typographical corrections, name format standardization, and improved functional information.
- We have initiated a long-term project to validate and improve protein names for non-redundant protein records. In March and April, we validated names for approximately 2 million records using multiple support lines from Swiss-Prot, HMM analysis, domain architecture analysis, and NCBI staff curation.

Nucleotide records

- Over 6,400 new or re-annotated RefSeq bacterial genomes were released.

- All new complete or draft RefSeq prokaryote genomes now use the accession format rule NZ_<original_INDSC_accession>. Complete genomes that were already accessioned using the 'NC_' prefix will continue to use that accession number. Thus, the accession prefix is no longer an indicator of a complete bacterial genome. Information about genome completeness is provided in the record DEFINITION line, the Assembly resource, and FTP reports provided by Assembly and Genome resources.

Impact to NCBI Gene

Together with this re-annotation effort, the scope of bacterial genomes included in Gene has been changed to include only genomes designated as a "reference genome," or "representative genome" where there is a cluster of related assemblies to indicate that the chosen representative assembly will be stable. Individual gene features on each assembly are identified with a locus_tag that can be used as a unique identifier for the gene in publications, even if the assembly is out of scope for Gene.

Using this data:

a) Strain-specific protein datasets for individual RefSeq genomes can be obtained online, by FTP, and through NCBI's programming utilities. For more detailed instructions, please see the Prokaryotic RefSeq FAQ.

b) A graphical display of an annotated gene or protein can be accessed from the Nucleotide resource. Starting from a RefSeq genome record of interest, such as NC_002695.1, follow the link to 'Graphics', and search for the locus_tag or protein name of interest.

c) Conversely, if starting from an individual non-redundant protein record, information about the annotated genomic location and genome taxonomy is available by following the link to the Identical Protein report. When a non-redundant protein record has been annotated on multiple RefSeq genomes, this report page lists the set of genomes that contain that identical protein, the genomic coordinates of the annotated CDS, and the specific organism information of the annotated genomic record. Thus, this report page can be used to identify the taxonomic range in which that identical protein has been found. The protein report can be downloaded in tabular format by using the 'Send to' link, and can be accessed using NCBI's programming utilities.

Future plans

NCBI's future plans include:

- Organism classification and quality assurance: Work continues to identify misclassified genomes and contaminated genomes. Depending on the specific details of identified issues, additional RefSeq bacterial genomes may be suppressed or updated.
- Re-annotation of complete genomes: A small number of bacterial genomes have not yet been re-annotated at this time and will be in the near future. We also plan to re-annotate the archaeal RefSeq genomes in 2015.
- Protein names: We are working on providing improved names for the non-redundant (WP_ accessioned) bacterial protein dataset. We are leveraging multiple sources of information, including curated UniProtKB/Swiss-Prot records, HMMs, Domain and domain architecture, publications and manual curation.
- Partial proteins: We are re-examining the prokaryotic genome annotation pipeline logic with regards to providing a non-redundant protein record for partial coding sequences.

Documentation

NCBI has created documentation to explain these changes in detail:

- RefSeq Re-annotation Project: An explanation of what the re-annotation project is, why and how it was done, and how we will facilitate your transition to the new annotation data.
- RefSeq non-redundant proteins: A description of this new protein record type with examples.
- Prokaryotic RefSeq Genomes: The prokaryotic RefSeq genomes policy, as well as definitions for reference genomes and representative genomes.
- Prokaryotic annotation pipeline: An explanation of the prokaryotic genome annotation process at NCBI.
- Prokaryotic RefSeq FAQ
- Supplemental data mapping file: An FTP file in the release-catalog directory (release70.bacterial-reannotation-report.txt.gz) has been prepared for re-annotated complete genomes that have recently transitioned to using the new non-redundant proteins. This file reports the old protein accession and GI, the annotated CDS coordinates, the old locus_tag and NCBI GeneID values and maps that to the current non-redundant protein accession and GI, the new locus_tag and NCBI GeneID (if available), the current CDS annotation coordinates, and indicates if the original protein identically matches or is similar to the replacement non-redundant protein.
- Supplemental report of suppressed assemblies: An FTP file in the release-catalog directory (release70.addedQA-SuppressedAssemblies.txt) reports details for a subset of bacterial genomes that were suppressed in March 2015 following an expansion of QA metrics and curatorial review. This report illustrates some of the reasons for suppression.

If you have more questions or specific questions that are not addressed in the documentation, you can write to the Help Desk at info@ncbi.nlm.nih.gov or use the feedback form on the RefSeq page.

May 13th webinar: "Introducing dbVar, the NCBI Database of Large-Scale Genetic Variation

Thursday, May 07, 2015

Next Wednesday, May 13th, NCBI staff will introduce dbVar, NCBI's database of genomic structural variation. In addition to describing the database's scope and features, we will also show you how to find, display and interpret dbVar records of interest.

To sign up for this webinar, click here. Like all of our webinars, this will be posted on the NCBI YouTube account after the live presentation; you can subscribe to our YouTube channel to be notified of all our new videos.

To see a list of upcoming webinars, as well as YouTube recordings and related materials from past webinars, please see the NCBI Webinars page.