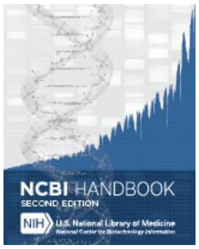
NLM Citation: Tatusova T, Ciufo S, Fedorov B, et al. About Prokaryotic Genome Processing and Tools. 2014 Jan 23. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013-.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



About Prokaryotic Genome Processing and Tools

Tatiana Tatusova, PhD,[1] Stacy Ciufo, PhD,[1] Boris Fedorov, PhD,[1] Kathleen O'Neill, PhD,[1] Igor Tolstoy,[1] and Leonid Zaslavsky, PhD[1]

Created: January 23, 2014.

Scope

RefSeq Prokaryotic Genome Project

As of October 2013, the prokaryotic genome dataset contains more than 15,000 genomes from almost 4,500 species representing a wide range of organisms. They include many important human pathogens, but also organisms that are of interest for non-medical reasons, biodiversity, epidemiology, and ecology. There are obligate intracellular parasites, symbionts, free-living microbes, hyperthermophiles and psychrophiles, and aquatic and terrestrial microbes, all of which have provided a rich insight into evolution and microbial biology and ecology. There is almost a 20-fold range of genomes sizes, spanning from ultra-small 45 kb archaeal genome of *Candidatus Parvarchaeum acidiphilum* recently obtained from mine drainage metagenome project (1) to the largest (14,7 Mb) strain of *Sorangium cellulosum*, an alkaline-adaptive epothilone producer (2).

The NCBI Reference Sequence (RefSeq) prokaryotic genome collection represents assembled genomes with different levels of quality and sampling density. Largely because of interest in human pathogens and advances in sequencing technologies (3), there are rapidly growing sets of very closely related genomes representing variations within the species. Some bacteria are often indistinguishable by means of current typing techniques. Whole-genome sequencing may provide improved resolution to define transmission pathways and characterize outbreaks. In order to support genome pathogen detection projects, RefSeq is changing the scope of the prokaryotic genome project to include all genomes submitted to public archives. Next generation technologies are changing the conventional use of microbial genome sequencing. Not so long ago genome sequencing projects were focused on a single bacterium, isolated and cultured from a single initial sample. Most of the sequencing technologies require DNA library preparation (DNA extraction and purification) followed by amplification and random shotgun sequencing. More recently new approaches have been developed that skip some of these steps. Metagenome sequencing shifted the focus from a single bacterium to multi-isolate and multi-species bacterial populations found in a single environmental sample. Individual organisms are not isolated and cultured but can be assembled computationally. RefSeq is taking a conservative approach of representing the genomic sequence of a single organism. Metagenomic assembly usually represents not a single organism but rather a composition of bacterial population. Metagenomic assemblies are not taken into RefSeq, however, that policy may change as the technologies and methods evolve. Single-cell sequencing technology (4) is another new technology that is being used to expand the catalog of uncultivated microorganisms. Genome assemblies that are generated from single-cell sequencing are taken into RefSeq when they meet basic validation criteria (see Quality Control).

Author Affiliation: 1 NCBI.

RefSeq Prokaryotic Re-annotation Project

Historically, RefSeq prokaryotic genomes relied on annotation submitted to one of the archival sequence databases maintained by the International Nucleotide Sequence Database (INSD) Collaboration . RefSeq curation focused primarily on the correction of protein names using protein clusters (first COG (5), later PRK (6)). Some attempts to correct the start sites were made but were not comprehensive and were based on manual review that didn't scale when the number of genomes grew to many thousands. The problem of missing genes has not been addressed at all. The result was inconsistent annotation even in closely related genomes with a good reference genome such as *Escherichia coli*. (7). To address these problems, NCBI developed its own prokaryotic genome annotation and analysis pipeline (PGAAP) that has been successfully used for many genomes submitted to GenBank in the last 5 years. This pipeline produces more consistent and high quality automatic annotation that in many cases surpasses the original author-provided annotation.

More recently, we have re-designed the PGAAP pipeline using a more structured framework that enables faster processing of batches of bacterial genomes and integrates additional automated quality checks. All RefSeq genomes, newly or previously submitted, will be re-annotated using the updated pipeline to further improve consistency and quality in the RefSeq prokaryotic genomes dataset. This process will include both existing RefSeq genomes and new submissions of complete and draft (WGS) bacterial genomes that meet basic quality threshholds.

RefSeq Targeted Loci Project

The small subunit ribosomal RNAs (16S in prokaryotes and 18S in eukaryotes) are useful phylogenetic markers that have been used extensively for evolutionary analyses. The large subunit ribosomal RNAs (23S and 5S in prokaryotes and 28S in eukaryotes) have also been used for evolutionary analyses although to a lesser extent than the 16S or 18S. The 16S bacterial/archaeal ribosomal RNA project is the result of an international collaboration with the Ribosomal Database Project (8)), GreenGenes, and Silva ribosomal RNA databases that curate and maintain sequence datasets for these markers. The fungal 18S and 28S ribosomal RNA projects are the result of an international collaboration with the Fungal Tree of Life project.

History

RefSeq Prokaryotic Genome Project History

The scale of genome sequencing and the production of data have reached astounding proportions since the completion of the first microbial genome of *Haemophilus influenzae* Rd KW20 (9) was released in 1995 and both the number of genomes and the number of unique genera for which a completely sequenced genome is available are increasing rapidly. We can see a shift in paradigm around 2010 when genome sequencing shifted from a single sample of a bacterial organism to hundreds of samples of bacterial populations.

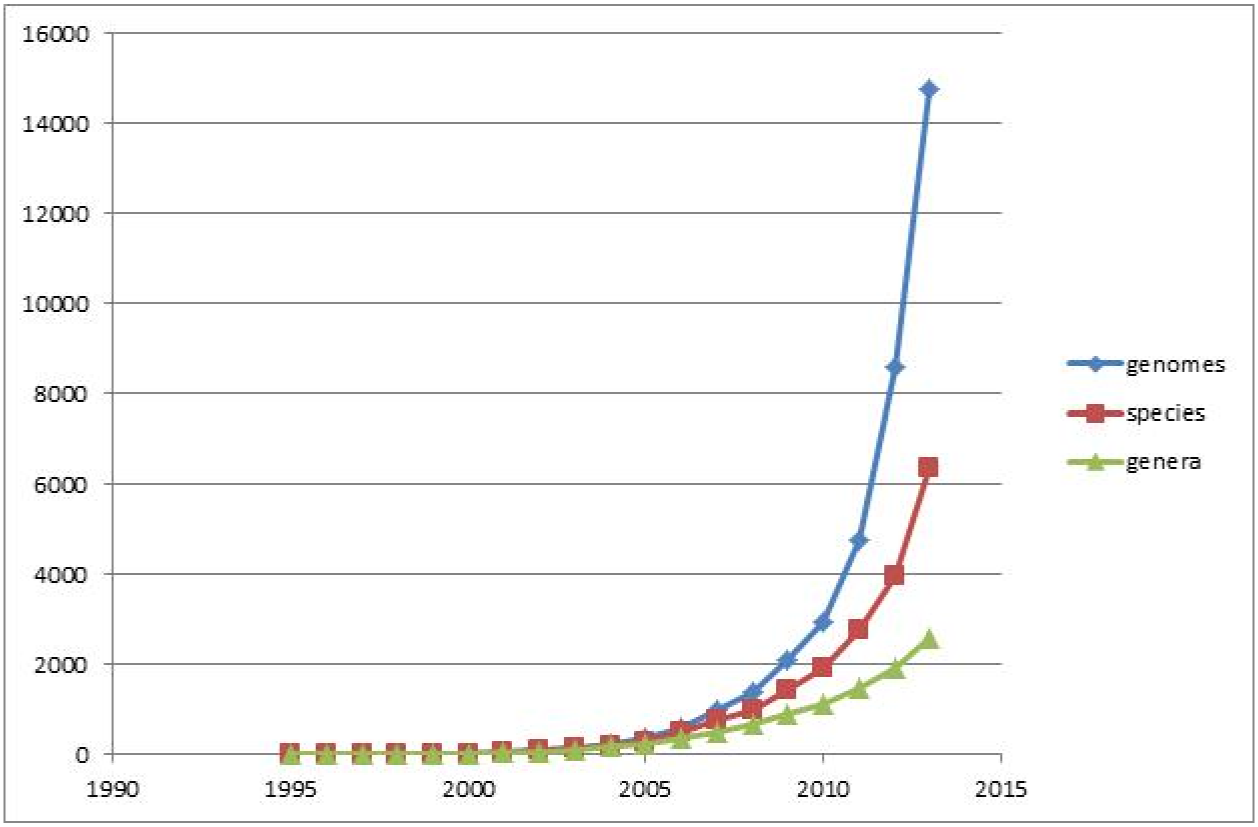


Figure 1. Growth of genomes, species, and genera: rapid growth of the number of isolates with relatively slow growth of new genera. Note that the data does not include assemblies from environmental studies where the number of novel species is growing much faster.

RefSeq Targeted Loci Project History

The Targeted Loci Project initiated in 2009 as a database of molecular markers used for phylogenetic analyses and identification of bacteria, archaea, and fungi. The initial project consisted of 16S ribosomal RNA from bacterial and archaeal type strains and has expanded to include 23S and 5S ribosomal RNA from bacterial and archaeal genomes as well as 18S and 28S ribosomal RNA from fungi.

Data Model

Genome and Assembly

RefSeq prokaryotic genomes are organized in several new categories based on curated attributes and assembly and annotation quality measures.

Reference genomes—manually selected "gold standard" complete genomes with high quality annotation and the highest level of experimental support for structural and functional annotation. Available at: http://www.ncbi.nlm.nih.gov/genome/browse/reference/

Representative genomes—representative genome for an organism (species); for some diverse species there can be more than one. For example, pathogenic and non-pathogenic *E. coli* will each be assigned a reference. Available at: www.ncbi.nlm.nih.gov/genome/browse/representative/

Variant genomes—all other genomes from individual samples representing genome variations within the species. Corresponds to Sequence Ontology- [SO:0001506].

How to define a genome?

BioProject ID can no longer define a genome for many multi-isolate and multi-species projects.

Taxonomy ID (taxid) can no longer define a genome since a unique taxid will not be assigned for individual strains and isolates. The collection of DNA sequences of an individual sample (isolate) will be represented by unique BioSample ID and if raw sequence reads are assembled and submitted to GenBank they will get a unique Assembly accession. The Assembly accession is specific for a particular genome submission and provides a unique ID for the set of sequence accessions representing the genome. Therefore, sequence data associated with a BioSample ID could be assembled with two different algorithms which may be submitted resulting in two sets of GenBank accessions, each with its own Assembly accession.

For example, BioProject PRJNA203445 is a multi-species project with multiple strains and isolates of different food pathogens. Each isolate has its own BioSample ID and each assembled genome has its own Assembly accession. An isolate of *Listeria monocytogenes* strain R2-502 was registered as BioSample SAMN02203126, and its genome is comprised of GenBank accessions CP006595-CP006596, which are in the Assembly database as accession GCA_000438585.

Autonomous Proteins

In order to manage the flood of identical proteins arising from annotation of variant genomes and decrease existing redundancy from bacterial genomes, NCBI is introducing a new protein data type in the RefSeq collection signified by a “WP” accession prefix. WP accessions provide non-redundant identifiers for protein sequences. This new data type is provided through NCBI’s genome annotation pipeline but is managed independently of the genome sequence data to ensure the dataset remains non-redundant.

We are doing this for two major reasons: 1) WP protein records represent a non-redundant protein collection that provides information about the protein sequence and name with linked information to genomic context and taxonomic sample; 2) use of WP accessions allows us to avoid creating millions of redundant protein records in the RefSeq collection. See more details at ftp://ftp.ncbi.nlm.nih.gov/refseq/release/announcements/WP-proteins-06.10.2013.pdf

Targeted Loci

16S ribosomal RNA project: Archaea and Bacteria

Initially, the 16S ribosomal RNA project compared curated, near-full-length16S sequences that corresponded to bacterial and archaeal type strains and from all contributing databases. RefSeq records corresponding to the original INSD submission were created from sequences and taxonomic assignments that were in agreement in all databases. Curation provided additional information, such as culture collection information or type strain designations and corrections to the sequence or taxonomy as compared to the original INSD submission. The 16S ribosomal RNA project has been expanded to include full length 16S ribosomal sequences from complete and incomplete genomes to provide representatives at the species level for the entire taxonomic range of bacteria and archaea.

23S ribosomal RNA project: Archaea and Bacteria

The 23S ribosomal RNA project includes full length 23S sequences from complete and incomplete genomes to provide representatives at the species level for the entire taxonomic range of bacteria and archaea.

5S ribosomal RNA project: Archaea and Bacteria

The 5S ribosomal RNA project includes full length 5S sequences from complete and incomplete genomes to provide representatives at the species level for the entire taxonomic range of bacteria and archaea.

18S and 28S ribosomal RNA projects

18S and 28S markers that correspond to type specimens and near full length sequences from all contributing databases were compared. RefSeq records corresponding to the original INSD submission were created from sequences and taxonomic assignments that were in agreement in all databases. The RefSeqs may contain corrections to the sequence or taxonomy as compared to the original INSD submission, and may have additional information added that is not found in the original.

Dataflow

The source of the genomic sequence in the RefSeq collection is a primary sequence record in the INSD public archives. Genomic sequences (nucleotide) in prokaryotic RefSeqs are identical copies of the underlying primary INSD records.

Quality Control

Genome representation: only assemblies with full representation of the genome of the organism are taken into RefSeq. Many genomes assemblies coming from single cell sequencing technology give only partial representation of DNA in a cell, ranging from 10% to 90%. Genome representation can be validated by comparative analysis if other genomes are available in closely related groups (species or genera). For novel phyla or kingdoms, some indirect criteria are applied (presence of universally conserved genes and total genome size)

Genomes and Genome Groups (Clades)

Historically, prokaryotic organisms were organized by classical taxonomic ranking system (species, genus, family, order, and phylum). Unlike eukaryotes, prokaryotes do not have clear definition of a species. Delineation of prokaryotic species was originally based on phenotypic information, pathogenicity, and environmental observations. More recently, several complementary approaches have been developed including molecular techniques such as DNA-DNA hybridization, phylogenetic markers (16S or universally conserved genes), and whole genome comparison (ANI – average nucleotide identity). Sequence-based methods using single-copy universally conserved genes are used for for delineation of prokaryotic species. We have implemented a similar approach to define bacterial clades based on comparison of universally conserved ribosomal proteins (markers).

Table 1. 23 universally conserved markers, shown here by Cluster ID.

1	30S ribosomal protein S12
2	30S ribosomal protein S7
3	30S ribosomal protein S2
4	50S ribosomal protein L11
5	50S ribosomal protein L1
6	50S ribosomal protein L3
7	50S ribosomal protein L22
8	30S ribosomal protein S3
9	50S ribosomal protein L14
10	50S ribosomal protein L5
11	30S ribosomal protein S8
12	50S ribosomal protein L6
13	30S ribosomal protein S5
14	30S ribosomal protein S13

Table 1. continued from previous page.

15	30S ribosomal protein S11
16	50S ribosomal protein L13
17	30S ribosomal protein S9
18	30S ribosomal protein S15
19	30S ribosomal protein S17
20	50S ribosomal protein L16
21	50S ribosomal protein L15
22	50S ribosomal protein L18
23	30S ribosomal protein S4

The pipeline for calculating genome clades consists of three major components. The first step is collecting the input data from NCBI main sequence repositories. The genomic data are dynamic: hundreds of new genomes and assembly updates are submitted to NCBI each day. We create a snapshot of all live genome assemblies and their nucleotide sequence components (chromosomes, scaffolds, and contigs) and store them in an internal database with a date stamp. The genome dataset is organized into large groups, phyla and super-phyla as defined by NCBI Taxonomy, see ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/CLADES/Phyla.txt.

Assemblies are then filtered by quality and passed to the processing script. Ribosomal protein markers are predicted in every genome to overcome problems with the submitted genome annotations (missing and/or incorrect annotations) and to normalize the predicted markers data set. Marker predictions are performed by aligning reference protein markers against full genome assemblies. Assemblies with at least 17 markers are passed to the next step. Genome distance is calculated as an average of pairwise protein distances of markers shared in a pair of genomes. Finally, agglomerative hierarchical clustering trees are built within phylum-level groups. Clades at the species level are calculated using a species-aware algorithm. Sub-clade trees are provided by cutting out the trees at the distance of 0.25.

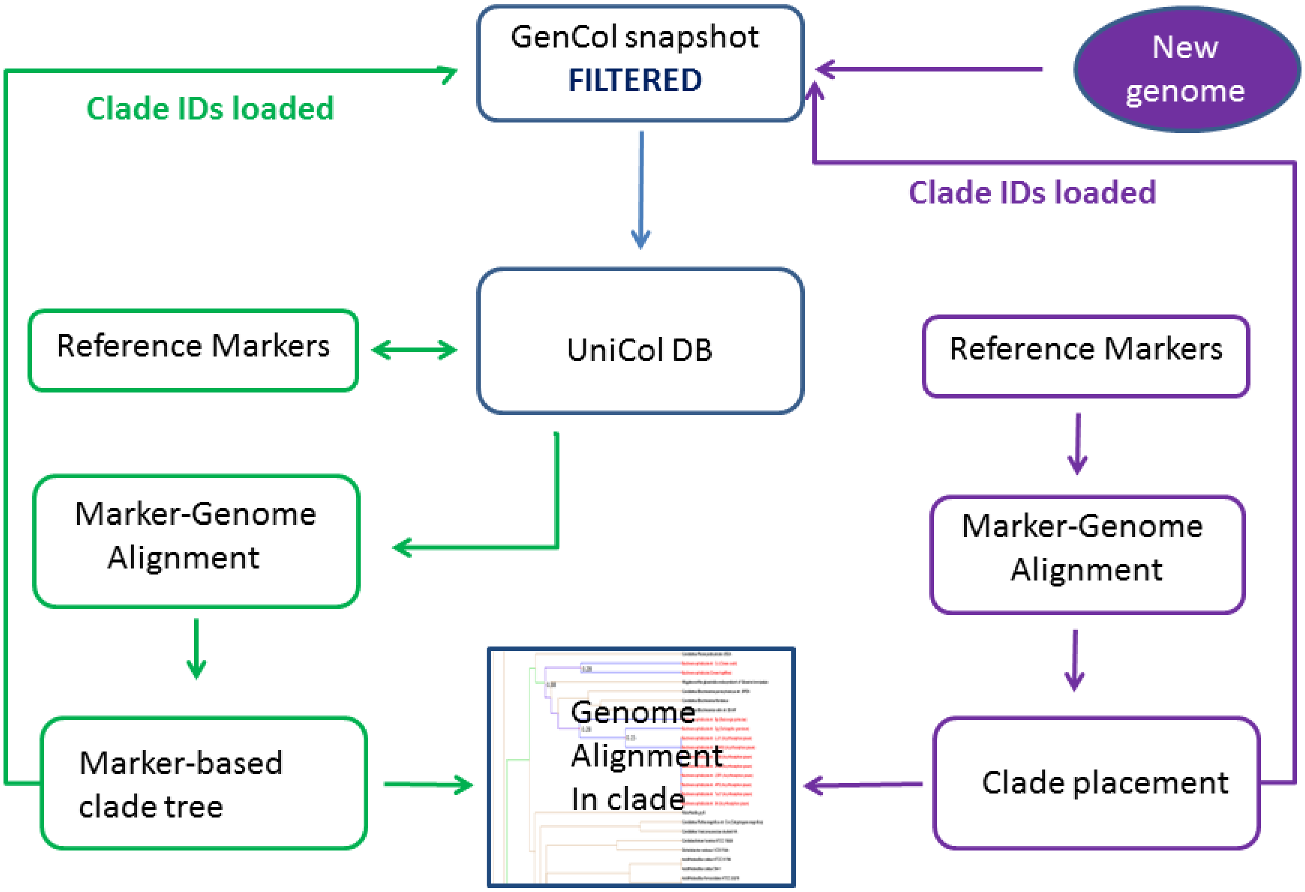


Figure 2. Calculating genome groups (clades) using universally conserved clusters. Snapshots are made every 6 months. The "FILTERED" step removes: partial assemblies, chimera, hybrid, mixed-cultured, metagenome assemblies.

Re-Annotation

The goal of the RefSeq re-annotation project is to improve the quality, normalize annotation, and reduce redundancy by creating reference sets of genomes, genes, and proteins. Information about the NCBI Prokaryotic Genome Annotation Pipeline is available here.

Improved consistency in RefSeq annotation across prokaryote genomes will provide a common ground for experimental and computational analysis. However, automatic pipelines cannot replace the manual curation and experimental validation of unusual features and biological artifacts such as ribosomal slippage, pseudogenes, mobile elements, Insertion Elements (IS), and rare non-standard start codons. Connecting the experimental studies of individual genes or gene families to genome annotation—despite some attempts to make it automatic—continues to be a laborious manual process. We do not want to overwrite manual curation with automatic prediction so continue to emphasize an integrated approach of computation supplemented by curation. We also encourage the research community to submit experimental data and manually curated data to RefSeq. There are two ways of making a contribution to the prokaryotic RefSeq collection:

1. **Organism/genome experts**—submit regular updates of your community-curated or experimentally validated genome annotation to GenBank and RefSeq.

For example: *Escherichia coli* K-12, *Mycobacterium tuberculosis*, *Bacillus subtilis*, *Pseudomonas aeruginosa* PAO1, and *Salmonella enterica* LT2. These genomes are all in the the RefSeq "Reference genome" dataset

2. **Gene or gene family experts/experimental data providers**—partial annotation updates to RefSeq records. For example: proteomics, ColleCT, REBASE. The update can be implemented as an automatic pipeline, or experimental validation studies published in journals indexed in PubMed can submit an annotated citation (a GeneRif) through NCBI's Gene resource.

There are ongoing efforts to establish relationships with the research community to provide accurate and up-to-date annotation for specific organisms or metabolic pathways. The "gold standard" Reference genome annotation is a result of the comparison of community annotation and NCBI automatic annotation reviewed by RefSeq curators.

Access

New genomes processed for RefSeq, are made public in NCBI resources and added to FTP directories daily.

Genome Groups

Reference Genomes: http://www.ncbi.nlm.nih.gov/genome/browse/reference/

Representative Genomes: www.ncbi.nlm.nih.gov/genome/browse/representative/

Complete list of prokaryotic genomes is available in Entrez Genome browser: http://www.ncbi.nlm.nih.gov/genome/browse/

Text version of the table can be downloaded from the FTP site: ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/

Species-level clades : ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/CLADES/

Reference set of universally conserved markers: ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/MARKERS

Sequence Data

Complete genomes : ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria

Draft genome assemblies: ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria_DRAFT

Plasmids: ftp://ftp.ncbi.nlm.nih.gov/genomes/Plasmids—this directory contains a complete list of all plasmids that are submitted as part of whole genome or individual complete plasmids that are sequenced and submitted separate from the chromosomes in plasmid targeted studies.

The genomes FTP area supports users who are interested in downloading data for one or a specific subset of organisms and/or in downloading the data that corresponds to an annotated genome. Users who are interested in comprehensive downloads can do so via the existing RefSeq release: ftp://ftp.ncbi.nlm.nih.gov/refseq/release/

Genomes are automatically linked to many other databases and resources in Entrez. These include Bioproject, Biosample, Assembly, PubMed, Taxonomy, and many others. See the Genome chapter for details.

Related Tools and Resources

Entrez links allow navigation through different databases. Metadata and sequence data are stored separately but are easily linked.

Resources

Taxonomy

The Taxonomy database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet. The NCBI taxonomy group began assigning strain-level taxids for prokaryotes with complete genome sequences as a convenience for those at INSDC institutes and their users when that sequencing was a major achievement. That practice was extended to prokaryotes with draft genome sequences and to some eukaryotic microbial organisms, e.g., yeasts. With high-throughput sequencing, manual curation of strain-level taxids is no longer possible. Therefore, the practice will be discontinued in January 2014. However, the thousands of existing strain-level taxids will remain, and we will continue to add informal strain-specific names for genomes from specimens that have not been identified to the species level, e.g., "*Rhizobium sp.* CCGE 510" and "*Salpingoeca sp.* ATCC 50818". The strain information will continue to be collected and displayed. Submitters of genome sequences will be required to register a BioSample ID for each organism that they are sequencing. Available at: http://www.ncbi.nlm.nih.gov/taxonomy/

BioSample

The BioSample database contains descriptions of biological source materials used in experimental assays. The BioSample record includes strain information and other metadata, such as culture collection and isolation information, as appropriate. The BioSample accession will be included as a "DBLINK" on GenBank records, and the GenBank records themselves will continue to display the strain in the source information. Available at: http://www.ncbi.nlm.nih.gov/biosample/

BioProject

A BioProject record is a collection of biological data related to a single initiative, originating from a single organization or from a consortium. A BioProject record provides users a single place to find links to the diverse data types generated for that project. Available at: http://www.ncbi.nlm.nih.gov/bioproject/

Assembly

Each genome assembly is loaded to the Assembly database and assigned an Assembly accession. The Assembly accession is specific for a particular genome submission. Genome assemblies are hierarchical. The shortest assembly components are contigs. Contigs are assembled into longer scaffolds, and scaffolds are assembled into chromosomes if there is sufficient mapping information. The Assembly resource provides the information on genome assembly structure and statistics. Available at: http://www.ncbi.nlm.nih.gov/assembly/

Protein Clusters

This collection of related protein sequences (clusters) consists of proteins derived from the annotations of whole genomes, organelles and plasmids. It currently limited to Archaea, Bacteria, Plants, Fungi, Protozoans, and Viruses. Available at: http://www.ncbi.nlm.nih.gov/proteinclusters/

Microbial Genome Resources:

The Microbial Genome resource page provides a central hub for many of NCBI's tools and resources. These include the gene prediction tools GeneMark (10) and Glimmer (11), and a statement of availability for the NCBI Genome Annotation Pipeline, now available as part of the Genbank submission process.

An expandable menu lists all of NCBI's prokaryotic genome tools and resources, guides for genome submission, information about NCBI's Annotation Workshop, and dynamically updated statistics on the the growth of microbial genome data.

The page also contains an expandable taxonomic tree with all bacterial and archaeal genomes with submitted genome data, and includes both complete and draft genomes. Available at: http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html

Tools

Microbial Genomes BLAST has new database options including "Representative genomes," now the default database, and "All genomes." Representative genomes provide a smaller, less redundant set of records for a given bacterial species. These representatives are selected by the research community and NCBI computational processes and are especially helpful for microbial species that are highly represented by genomes for numerous strains in NCBI databases, such as *Escherichia coli*. The "All genomes" option offers the choice of Complete genomes, Draft genomes, or Complete plasmids. You can search these sets individually or in any combination. The microbial BLAST report also has a new "Genome" link to the species page in Entrez Genome in the alignments section of the BLAST report.

Concise BLAST includes a representative protein from each cluster. This allows a more comprehensive taxonomic BLAST search while eliminating much of the noise from similiar species: http://www.ncbi.nlm.nih.gov/genomes/static/conciseblasthelp.html

The Submission Check Tool is available for users to check the validity of their genome data prior to submission to Genbank. Checks include gene overlaps, RNA overlaps, partial overlaps, frameshifts, truncated proteins, missing RNAs, and RNA strand mismatches.

gMap is a graphical representation of pre-computed genomic comparisons of closely related strains. Syntenic blocks are detected through analysis of BLAST hits between every pair of the input sequences. Hits are split or combined to keep the number and lengths of syntenic blocks in accordance with the length of selected genomic intervals, as well as to ensure consistency of the blocks across multiple sequences. The results are displayed in a simple graphic that shows color-coded and numbered segments indicating similarity between two or more genomes. This tool can be used to visually detect chromosomal similarities, rearrangements, and the above-mentioned genomic islands, as well as smaller insertions or deletions. Care must be taken when interpreting results from incomplete genomes as hit coverage may be affected by the number of contigs and the gaps between them. Other tools that are useful for examining pairwise genomic comparisons are GenePlot and HitPlot which are linked on this page.

ProtMap is a graphical gene neighborhood tool that displays clickable, linked genes upstream and downstream of the target. This resource is useful in identifying paralogs.

GenePlot combines protein-sequence similarity searches with sequence location, unlike gMap, which is solely based on nucleotide sequence similarity. This tool can be used to detect syntenic regions or chromosomal rearrangements in closely related species, as well as contiguous regions in distantly related organisms. Small insertions or deletions and major genomic islands in closely related species are also identifiable using this tool and a table of the best hits between both organisms is available. Geneplot provides a more detailed view of the pairwise comparison of two genomes.

TaxPlot compares two reference proteomes to a query proteome and thus provides a three-way comparison of proteome similarities. A single protein can be searched for and highlighted, and entire COG functional categories can be examined, allowing a three-way comparison of a single category of proteins. This tool can be used to detect potentially horizontally transferred genes by using a distantly related organism in comparison to two closely related strains.

tRNAscan-SE (12) is a program for detection of tRNA genes in genomic sequence.

References

1. Fujishima K, Sugahara J, Miller CS, Baker B, Di Giulio M, Tomita M, Banfield JF, Kanai A. A novel three-unit tRNA splicing endonuclease found in ultra-small Archaea possesses broad substrate specificity. Nucleic Acids Res. 2011 Dec;39(22):9695–704. PubMed PMID: 21880595.
2. Han K, Li ZF, Peng R, Zhu LP, Zhou T, Wang LG, Li SG, Zhang XB, Hu W, Wu ZH, Qin N, Li YZ. Extraordinary expansion of a Sorangium cellulosum genome from an alkaline milieu. Sci Rep. 2013;3:2101. PubMed PMID: 23812535.
3. Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol. 2012 Sep;10(9):599–606. PubMed PMID: 22864262.
4. Blainey PC. The future is now: single-cell genomics of bacteria and Archaea. FEMS Microbiol Rev. 2013 May;37(3):407–27. PubMed PMID: 23298390.
5. Koonin EV. The Clusters of Orthologous Groups (COGS) Database: Phylogenetic Classification of Proteins from Complete Genomes. The NCBI Handbook (Internet). 2002.
6. O'Neill K, Klimke W, Tatusova T. Protein Clusters: A collection of proteins grouped by sequence similarity and function. NCBI Help Manual. 2007.
7. Poptsova MS, Gogarten JP. Using comparative genome analysis to identify problems in annotated microbial genomes. Microbiol. 2010;156:190917. PubMed PMID: 20430813.
8. Maidak BL, Olsen GJ, Larsen N, Overbeek R, MacCaughey MJ, Woese CR. The Ribosomal Database Project (RDP). Nucleic Acids Res. 1995;24:82–85. PubMed PMID: 8594608.
9. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995 Jul 28;269(5223):496–512. PubMed PMID: 7542800.
10. Borodovsky M, McIninch J. GeneMark: parallel gene recognition for both DNA strands. Computers & Chemistry. 1993;17(2):123–133.
11. Salzberg S, Delcher A, Kasif S, White O. Microbial gene identification using interpolated Markov models. Nucleic Acids Research. 1998;26(2):544–548. PubMed PMID: 9421513.
12. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. (1997) Mar 1;25(5):955-64.