Chapter 1

Update on Genomic Databases and Resources at the National Center for Biotechnology Information

Tatiana Tatusova

Abstract

The National Center for Biotechnology Information (NCBI), as a primary public repository of genomic sequence data, collects and maintains enormous amounts of heterogeneous data. Data for genomes, genes, gene expressions, gene variation, gene families, proteins, and protein domains are integrated with the analytical, search, and retrieval resources through the NCBI website, text-based search and retrieval system, provides a fast and easy way to navigate across diverse biological databases.

Comparative genome analysis tools lead to further understanding of evolution processes quickening the pace of discovery. Recent technological innovations have ignited an explosion in genome sequencing that has fundamentally changed our understanding of the biology of living organisms. This huge increase in DNA sequence data presents new challenges for the information management system and the visualization tools. New strategies have been designed to bring an order to this genome sequence shockwave and improve the usability of associated data.

Key words Bioinformatics, Genome, Genome assembly, Database, Data management system, Sequence analysis

1 Introduction

Genome science together with many other research fields of life sciences had entered the Era of Large-scale Data Acquisition in the early 1990s. The Era was led by the fast accumulation of human genomic sequences and followed by similar data from other large model organisms. Microbial genomics has also been pursued into both metagenomics [1, 2] and pan-genomics [3]. Recent advances in biotechnology and bioinformatics led to a flood of genomic (and metagenomic) data and a tremendous growth in the number of associated databases. As of June 2015, NCBI Genome collection contains more than 35,000 genome sequence assemblies from almost 13,000 different organisms (species) representing all major taxonomic groups in Eukaryotes (Fig. 1), Prokaryotes (Fig. 2), and Viruses. Prokaryotic genomes are the most abundant and rapidly growth portion of assembled genomes data collection in public archives.

Oliviero Carugo and Frank Eisenhaber (eds.), *Data Mining Techniques for the Life Sciences*, Methods in Molecular Biology, vol. 1415, DOI 10.1007/978-1-4939-3572-7_1,

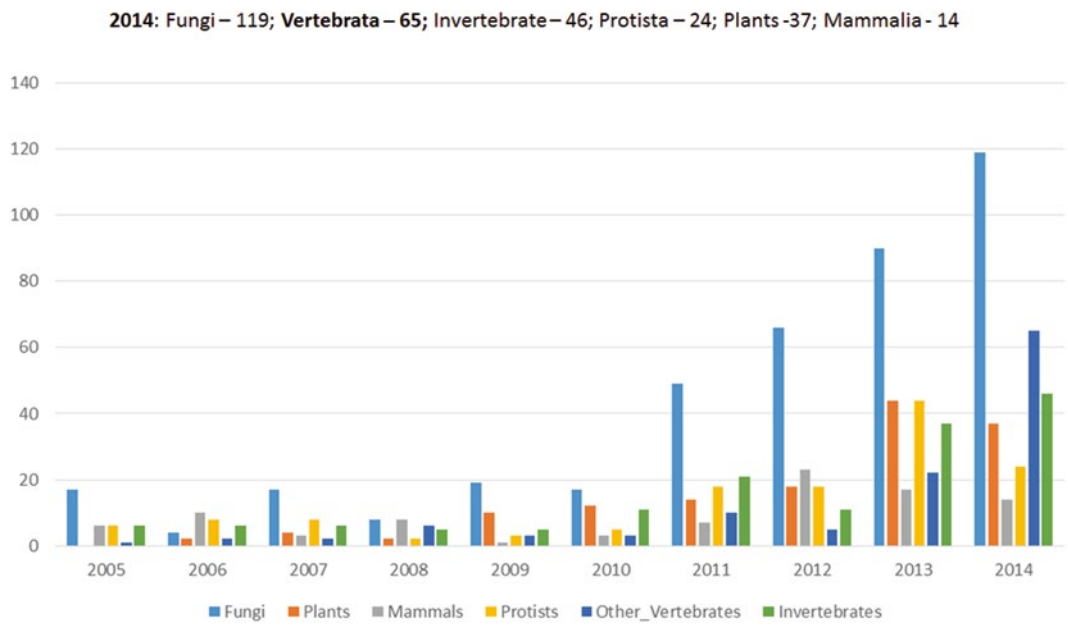


Fig. 1 Assembled eukaryotic genomes in public archives released by year by major taxonomic groups

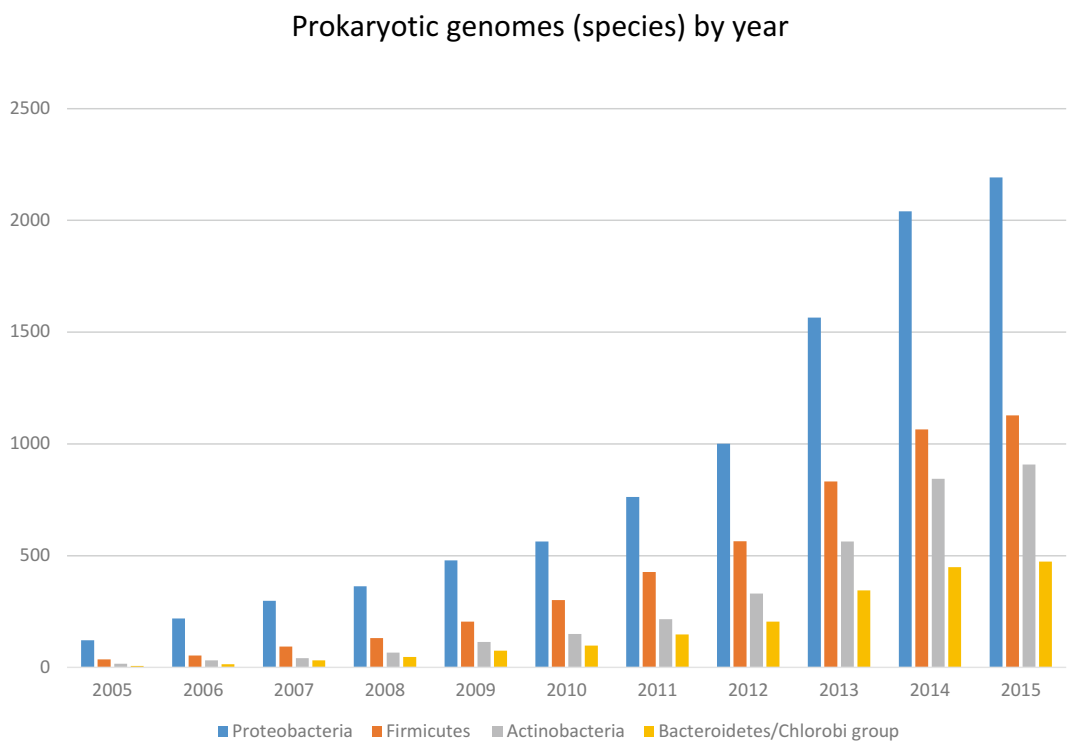


Fig. 2 Assembled prokaryotic genomes in public archives released by year by major phyla

NCBI, major public sequence archive, accepts primary submission from the large sequencing centers, small laboratories, and individual researches. Raw sequence data and alignments of the read data produced by Next Generation Sequencing (NGS) are stored in SRA (Sequence Read Archive) database. Massively parallel sequencing technologies (Illumina, 454, PacBio) [4] have opened an extensive new vista of research possibilities—personal genomics, human microbiome studies, analysis of bacterial and viral disease outbreaks, generating thousands of terabytes of short read data. More recently, new technologies of third and fourth generation sequencing [5] such as single cell molecule [6], nanopore-based [7] have been applied to whole-transcriptome analysis that opened a possibility for profiling rare or heterogeneous populations of cells. New generation sequencing platforms offer both high-throughput and long sequence reads. The new Pacific Bioscience RS (PacBio) third-generation sequencing platform offers high throughput of 50,000–70,000 reads per reaction and a read length over 3 kb. Oxford Nanopore released the MinION® device, a small and low-cost single-molecule nanopore sequencer, which offers the possibility of sequencing DNA fragments up to 60 kB. These advanced technologies may solve assembly problems for large and complex genomes [8] and allow to obtain a highly contiguous (one single contig) and accurate assemblies for prokaryotic genomes [9, 10]. NCBI Sequence Read Archive accepts data submission in many different formats originated from various platforms adding additional formats as they become available (*see* Submission section 2.1 below).

Assembled nucleotide sequence data and annotation with descriptive metadata including genome and transcriptome assemblies are submitted to the three public archive databases of the International Nucleotide Sequence Database Collaboration (INSDC, www.insdc.org)—European Nucleotide Archive (ENA) [11], GenBank [12], and the DNA Data Bank of Japan (DDBJ) [13]. Two new datatypes (GenBank divisions) have been recently introduced to accommodate the data from new sequencing technologies: (1) Whole Genome Shotgun (WGS) archives genome assemblies of incomplete genomes or chromosomes that are generally being sequenced by a whole genome shotgun strategy; (2) Transcriptome Shotgun Assembly (TSA) archives computationally assembled transcript sequences from primary read data. *See* Fig. 3 for the growth of sequence data in public archives.

As the volume and complexity of data sets archived at NCBI grow rapidly, so does the need to gather and organize the associated metadata. Although metadata has been collected for some archival databases, previously, there was no centralized approach at NCBI for collecting this information and using it across databases. The BioProject database [14] was recently established to facilitate organization and classification of project data submitted to NCBI, EBI and DDBJ databases. It captures descriptive information

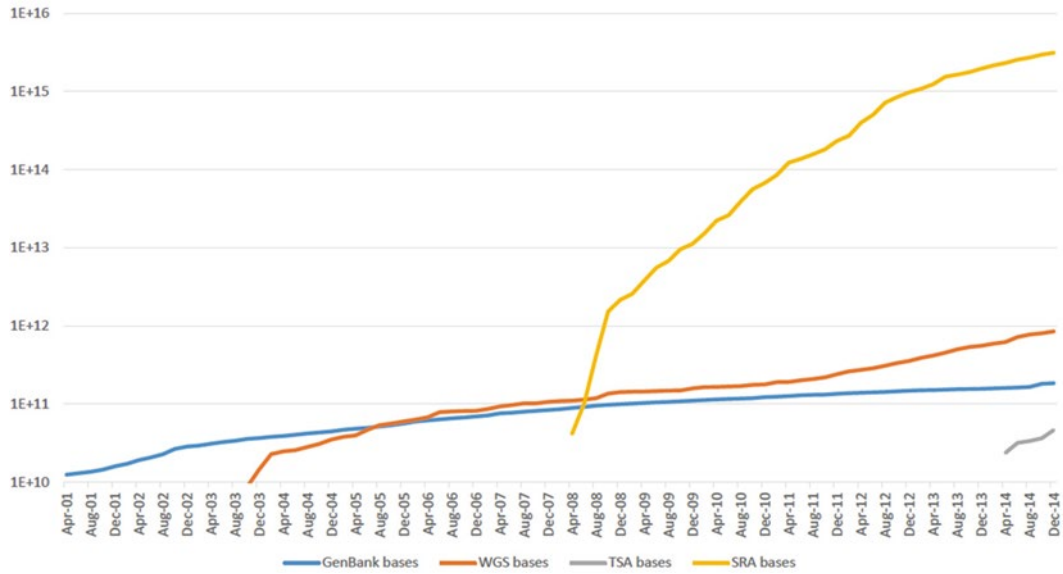


Fig. 3 Growth of NCBI sequence archives: GenBank, WGS, TSA, and SRA

about research projects that result in high volume submissions to archival databases, ties together related data across multiple archives and serves as a central portal by which to inform users of data availability. Concomitantly, the BioSample database [14] is being developed to capture descriptive information about the biological samples investigated in projects. BioProject and BioSample records link to corresponding data stored in archival repositories.

Additional information on biomedical data is stored in an increasing number of various databases. Navigating through the large number of genomic and other related "omic" resources and linking it to the metagenome (epidemiological, geographical) data becomes a great challenge to the average researcher. Understanding the basics of data management systems developed for the maintenance, search, and retrieval of the large volume of biological data will provide necessary assistance in navigating through the information space.

This chapter is an update of the previous report on NCBI genome sequence data management system [15]. The updated version provides a description of new and/or completely redesigned genomic resources that became available since the first 2008 edition. NCBI, as a primary public repository of biomolecular information, collects and maintains enormous amounts of heterogeneous data. The databases vary in size, data types, design, and implementation. They cover most of the genomic biology data types including sequence data (genomic, transcript, protein sequences); metadata describing the objectives and goals of the project and environmental, clinical, and epidemiological data that

is associated with the sample collections (BioSample); and related bibliographical data.

All these databases are integrated in a single data management system and use a common engine for search and retrieval. This provides researchers with a common interface and simplifies navigation through the large information space.

This chapter focuses on the primary genome sequence data and some related resources, but many other NCBI databases such as GEO, Epigenomics, dbSNP dbVar, and dbGaP, although related, are not in scope for the current review.

There are many different ways of accessing genomic data at NCBI. Depending on the focus and the goal of the research project or the level of interest, the user would select a particular route for accessing the genomic databases and resources. These are: (1) text searches, (2) direct genome browsing, (3) searches by sequence similarity, and (4) pre-computed results of analysis. All of these search types enable navigation through pre-computed links to other NCBI resources. Recently redesigned genome FTP directories provide easy access to the individual genome assemblies as well as to large datasets arranged in organism groups.

2 Primary Data Submission and Storage

The National Center for Biotechnology Information was established on November 4, 1988 as a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH) in order to develop computerized processing methods for biomedical research. As a national resource for molecular biology information, NCBI's mission is to develop automated systems for storing and analyzing knowledge about molecular biology, biochemistry, and genetics; facilitating the use of such databases and software by the research and medical community; coordinating efforts to gather biotechnology information both nationally and internationally; and performing research into advanced methods of computer-based information processing for analyzing the structure and function of biologically important molecules.

The fundamental sequence data resources at NCBI consist of both primary databases and derived or curated databases. Primary sequence databases such as SRA [16], GenBank [12] and metadata repositories such as BioProject and BioSample [17, 18] archive the original submissions that come from large sequencing centers or individual experimentalists. The database staff organizes the data but do not add additional information. Curated databases such as Reference Sequence Collection [14, 15] provide a curated/expert view by compilation and correction of the data. For more detailed information on all NCBI and database resources see also most recent NCBI databases review [19].

This section provides an overview of genome submission processing, from the management of data submission to the generation of publicly available data products.

2.1 Primary Raw Sequence Data: Sequence Read Archive (SRA)

Most of the data generated in genome sequencing projects is produced by whole genome shotgun sequencing, resulting in random short fragments - raw sequence reads.

For many years the raw sequence reads remained out of the public domain because the scientific community has focused its attention primarily on the end product: the fully assembled final genome sequence. As the analysis of genomic data progressed, the scientific community became more concerned with the quality of the genome assemblies and thought they'd need a place to store the primary sequence read data. Also, having all the read data in a single repository could also provide an option to combine reads from multiple sequencing centers and/or try different assembly algorithms on same public set of reads. Trace Archive has successfully served as a repository for the data produced by capillary-based sequencing technologies for many years. New parallel sequencing technologies (e.g., 454, Solexa, Illumina, ABI Solid,) have started to produce massive amounts of short sequence reads (20–100 kb). More recently, Pacific Biosystems (PacBio) and Oxford sequencing technologies have started producing much longer reads (10–15 kB on average) with really massive throughput.

In addition to raw sequence data SRA can store alignment information from high-throughput sequencing assembly projects. The alignments provide the important information on mapping the reads to the consensus or reference assembly as well as the duplicated and not-mapped reads. The importance of storing the alignment of raw reads to the consensus sequence in public archives was emphasized from the very beginning af large-scale genome sequencing projects [21]. Trace archive has an option to capture and display assembly alignments. More recently, the research community in collaboration with major sequence archives have developed the standard formats for the assembly data.

SAM, which stands for Sequence Alignment/Map format, is a TAB-delimited text format consisting of a header (optional) and an alignment. Typically, each alignment represents the linear alignment of a segment. BAM is a binary version of SAM format.

NCBI SRA submission portal is accepting assembly data files in SAM/BAM formats. For more details see online specification of SAM/BAM format at https://samtools.github.io/hts-specs/SAMv1.pdf

For more information on SRA submission protocol and data structure see SRA documentation at http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=doc and NCBI SRA Handbook at http://www.ncbi.nlm.nih.gov/books/NBK242623/

NGS parallel sequencing approach results in high level of redundancy in the sequence runs. By aligning the reads to the reference identical bases can be identified and collapsed, only the mismatching bases are stored. The original sequence may be restored by applying a function to the reference and the stored only differences. Continues growth of primary raw data requires further development of data compression and reducing the redundancy. At some point the need to store every read generated by sequencing machine may become unnecessary. The major objective of storing all primary data was the concern about data reproducibility. With the cost of sequencing dropping down so fast the cost of re-sequencing may become lower than the cost of storage of terabytes of data.

2.2 Primary Sequence Data—Genome and Transcriptome Assemblies Rapidly

Sequences assembled from raw machine reads are traditionally submitted to GenBank/EMBL/DDBJ consortium. Two new data types were recently created to accommodate assembled data from NGS projects.

TSA is an archive of computationally assembled sequences from primary data such as ESTs, traces and Next Generation Sequencing Technologies. The overlapping sequence reads from a complete transcriptome are assembled into transcripts by computational methods instead of by traditional cloning and sequencing of cloned cDNAs. The primary sequence data used in the assemblies must have been experimentally determined by the same submitter. TSA sequence records differ from EST and GenBank records because there are no physical counterparts to the assemblies. For more details see http://www.ncbi.nlm.nih.gov/genbank/tsa

Whole Genome Shotgun (WGS) projects are genome assemblies of incomplete genomes or incomplete chromosomes of prokaryotes or eukaryotes that are generally being sequenced by a whole genome shotgun strategy. A whole genome assembly may be a large hierarchical sequence structure: http://www.ncbi.nlm.nih.gov/genbank/wgs

Shotgun technology generates high volume of reads that represent random fragments of the original genome or transcriptome sequence. A computational process of the reconstructing of the original sequence by merging the fragments back together is called assembly. The resulting sequence (gapless contig) or a collection of sequences represents assembly as on object. In large genome sequencing project a set of contigs can be linked (by employing linking information) together forming scaffolds. Scaffolds can be mapped to the chromosome coordinates if physical or genetic mapping information is available.

Assembly instructions can be formally described in AGP format. A tab delimited file describes the assembly of a larger sequence object from smaller objects. The large object can be a contig, a scaffold (supercontig), or a chromosome. Each line (row) of the

AGP file describes a different piece of the object. These files are provided by primary submitters of Whole Genome Sequence (WGS) data. For details *see*: http://www.ncbi.nlm.nih.gov/projects/genome/assembly/agp/AGP_Specification.shtml.

2.3 Primary Metadata: BioProject, BioSample

As the diversity, complexity, inter-relatedness and rate of generation of the genome sequence data continue to grow, it is becoming increasingly important to capture scholarly metadata and allow the identification of various elements of a research project, such as grant proposals, journal articles, and data repository information. With the recent advances in biotechnology researches gain access to new types of molecular data. The genome studies have expanded from just genome sequencing to capturing structural genome variations, genetic and phenotypic data, epigenome, trasncriptome, exome sequencing and more.

The BioProject database [14] replaces NCBI's Genome Project database and reflects an expansion of project scope, a redesigned database structure and a redesigned website. The BioProject database organizes metadata for research projects for which a large volume of data is anticipated and provides a central portal to access the data once it is deposited into an archival database. A BioProject encompasses biological data related to a single initiative, originating from a single organization or from a consortium of coordinating organizations.

Project materials (sample) information is captured and given a persistent identifiers in BioSample database [14]. Given the huge diversity of sample types handled by NCBI's archival databases, and the fact that appropriate sample descriptions are often dependent on the context of the study, the definition of what a BioSample represents is deliberately flexible. Typical examples of a BioSample include a cell line, a primary tissue biopsy, an individual organism or an environmental isolate.

Together, these databases offer improved ways for users to query, locate, integrate, and interpret the masses of data held in NCBI's archival repositories.

2.4 Submission Portal

Submission portal is a single entry point that allows submitters to register a project (or a multiple projects) and deposit data to different NCBI databases. All primary data including metadata on the biological material, raw sequence reads, assembled genome, transcriptome, and functional genomic assays can be submitted using the same interface (https://submit.ncbi.nlm.nih.gov/).

The data submission process at NCBI include multiple steps aiming to ensure the data quality and integrity. Quality Control is implemented as a set of automatic validation checks followed by manual review by NCBI staff. Figure 4 shows the major steps of primary submission process.

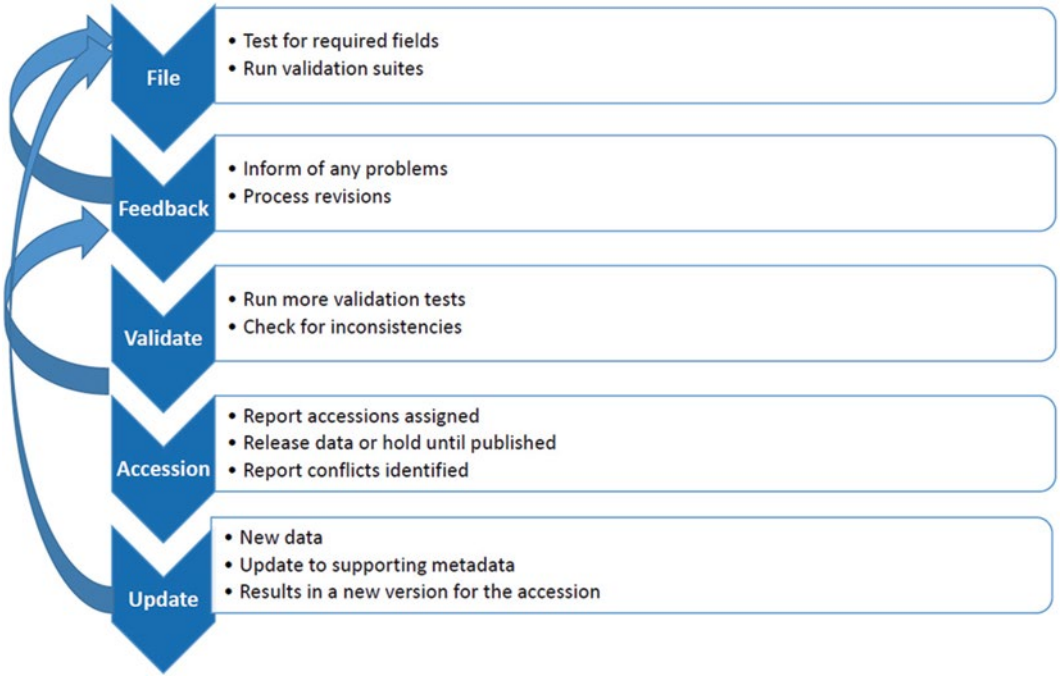


Fig. 4 Primary submission data processing

3 Text Search and Retrieval system

3.1 Basic Organizing principles

Entrez is the text-based search and retrieval system used at NCBI for all of the major databases and it provides an organizing principle for biomedical information.

Entrez integrates data from a large number of sources, formats, and databases into a uniform information model and retrieval system. The actual databases from which records are retrieved and on which the Entrez indexes are based have different designs, based on the type of data, and reside on different machines. These will be referred to as the "source databases". A common theme in the implementation of Entrez is that some functions are unique to each source database, whereas others are common to all Entrez databases.

An Entrez "node" is a collection of data that is grouped and indexed together. Some of the common routines and formats for every Entrez node include the term lists and posting files (i.e., the retrieval engine) used for Boolean Query, the links within and between nodes, and the summary format used for listing search results in which each record is called a DocSum. Generally, an Entrez query is a Boolean expression that is evaluated by the common Entrez engine and yields a list of unique ID numbers (UIDs), which identify records in an Entrez node. Given one or more UIDs, Entrez can retrieve the DocSum(s) very quickly.

3.1.1 Query Examples

Each Entrez database ("node") can be searched independently by selecting the database from the main Entrez Web page (http://www.ncbi.nlm.nih.gov/sites/gquery).

Typing a query into a text box provided at the top of the Web page and clicking the "Go" button will return a list of DocSum records that match the query in each Entrez category. These include, for example, nucleotides, proteins, genomes, publications (PubMed), taxonomy, and many other databases. The numbers of results returned in each category are provided on a single summary page and provide the user with an easily visible view of the results in each of ~35 databases. The results are presented differently in each database but within the same framework which includes the common elements such as search bar, display options, page formatting, and links.

In processing a query, Entrez parses the query string into a series of tokens separated by spaces and Boolean operators (AND, NOT, OR). An independent search is performed for each term, and the results are then combined according to the Boolean operators.

Query uses the following syntax:

term [field] OPERATOR term [field]

where "term" refers to the search terms, "field" to the Search Field defined by specific Entrez database, and "OPERATOR" to the Boolean Operators.

More sophisticated searches can be performed by constructing complex search strategies using Boolean operators, for example, in Genome database a query

(Bacteria[organism] OR Archaea[organism]) AND complete[Status]

will return all genome records (species) from bacteria and Archaea domain for which complete genome sequence assemblies are available.

The main goals of the information system are reliable data storage and maintenance, and efficient access to the information. The retrieval is considered reliable if the same information that was deposited can be successfully retrieved. The Entrez system goes beyond that by providing the links between the nodes and pre-computing links within the nodes. The links made within or between Entrez nodes from one or more UIDs (Unique IDentifier) is also a function across all Entrez source databases. Linking mechanisms are described in detail in the previous version [15]. On public facing Web pages links to other databases are presented to the user in **Find related data** section where the name of the related database can be selected from pulldown menu (Fig. 5a).

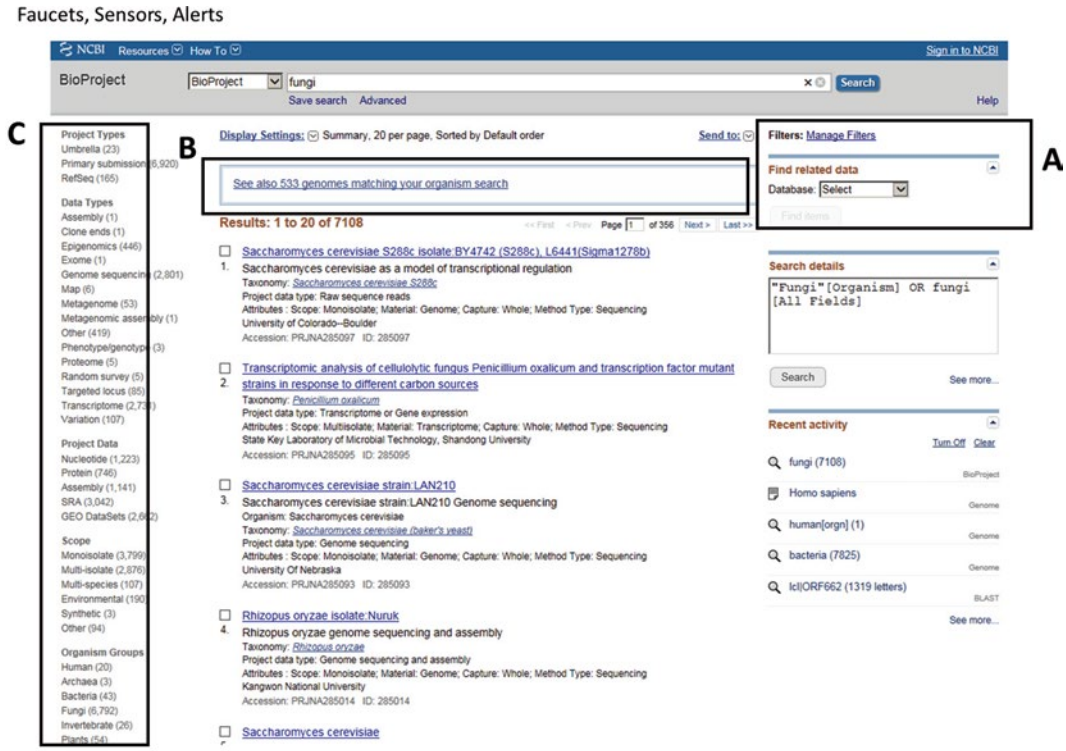


Fig. 5 New features improving the presentation of search results: (**a**) pre-computed links to related data in other resources; (**b**) sensor, a provisional navigation path based on the analysis of the search query; (**c**) Faucets (filters)

3.1.2 Towards Discovery: Sensors and Adds, Faucets (Filters) and Alerts

More recently, several new features have been developed aiming to help the researches with understanding the results of the search and provide a provisional navigation path that is based on the analysis of the search query (Fig. 5b). Various filters can be applied to the search results to limit the result set to subset of a particular interest. In the previous version these filters can be applied using complex Boolean Query or using a custom-designed **Limits** page. In the recently redesigned version all filters applicable to the results set are shown on the result page and are implemented as faucets providing upfront options to focus on the subset of interest (Fig. 5c).

Alert option provides a subscription to My NCBI which allows to retain user information and database preferences to provide customized services for many NCBI databases.

My NCBI subscription provides various useful features that allow to save searches and create automatic e-mail alerts when new results become available: save display format preferences and filter options, store recent activity searches and results for 6 months, and many more. More information about Entrez system can be found from NCBI online Help Manual at http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez.chapter.EntrezHelp.

3.2 Tools for Advanced Users

The Entrez Programming Utilities (E-Utils) are a set of eight server-side programs that provide a stable interface to the Entrez query and database system. The E-Utils use a fixed URL syntax that translates a standard set of input parameters into the values necessary for various NCBI software components to search for and retrieve data, and represent a structured interface to the Entrez system databases.

To access these data, a piece of software first posts an eUtils URL to NCBI, then retrieves the results of this posting, after which it processes the data as required. The software can thus use any computer language that can send a URL to the eUtils server and interpret the XML response, such as Perl, Python, Java, and C++. Combining e-Utils components to form customized data pipelines within these applications is a powerful approach to data manipulation. More information and training on this process is available through a course on NCBI Powerscripting: http://www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/course.html.

4 Genomic Databases; Public Reports

The genome sequencing era that started about 20 years ago has brought into being a range of genome resources. Genomic studies of model organisms give insights into understanding of the biology of humans enabling better prevention and treatment of human diseases. Comparative genome analysis leads to further understanding of fundamental concepts of evolutionary biology and genetics. Species-specific genomic databases comprise a lot of invaluable information on genome biology, phenotype, and genetics. However, primary genomic sequences for all the species are archived in public repositories that provide reliable, free, and stable access to sequence information. In addition NCBI provides several genomic biology tools and online resources, including group-specific and organism-specific pages that contain links to many relevant websites and databases.

4.1 Sequence Read Archive (SRA)

The access to the SRA data is provided through SRA Web browser and specialized SRA BLAST search application. NCBI has developed a set of tools that allow the users to download sequencing files directly from SRA database. For the detailed description of SRA Toolkit visit documentation page http://www.ncbi.nlm.nih.gov/Traces/sra/?view=toolkit_doc.

SRA data are organized in SRA studies, experiments and runs. SRA studies are registered in BioPorject database (*see* Subheading 2.3); the metadata include the aims and objectives of the project, title and brief description, and optional funding sources and publications. The description of biological material (sample) used in the experiment(s) within the study is captured and

maintained in BioSample database (*see* Subheading 2.4). It includes description of the sample (collection date and location, age, gender, cell line, etc.) as well as information on sequencing methods and instrumental models used in the experiments.

Multiple experiments can be performed with a same sample but using multiple samples in a same experiment is not allowed in the SRA data model. Data sets within an experiment are organized in runs usually associated with the sequencing libraries.

Run browser (http://www.ncbi.nlm.nih.gov/Traces/sra/?view=run_browser) allows the user to search data for a single run with the run accession number. SRA Run selector

Allow the user to search with accession(s) of the studies, samples, or experiments. The search, indexing, and Web presentation (Fig. 6) are implemented with the Solr database technology (http://lucene.apache.org/solr/).

Special version of BLAST is using megablast [21]—nucleotide blast version optimized for highly similar sequences (*see* Subheading 5 below)

4.2 NCBI Taxonomy

The NCBI Taxonomy Database [22] serves as the standard nomenclature and classification for the International Sequence Database (INSDC). Taxonomy was first indexed in Entrez in 1993—at the time there were just over 5000 species with formal scientific names represented in GenBank. As of June 2015 sequences from over 300 000 species are represented in INSDC. However, with common estimates of the species on the planet around two million the subset with sequence in GenBank represents only 15 % of the total.

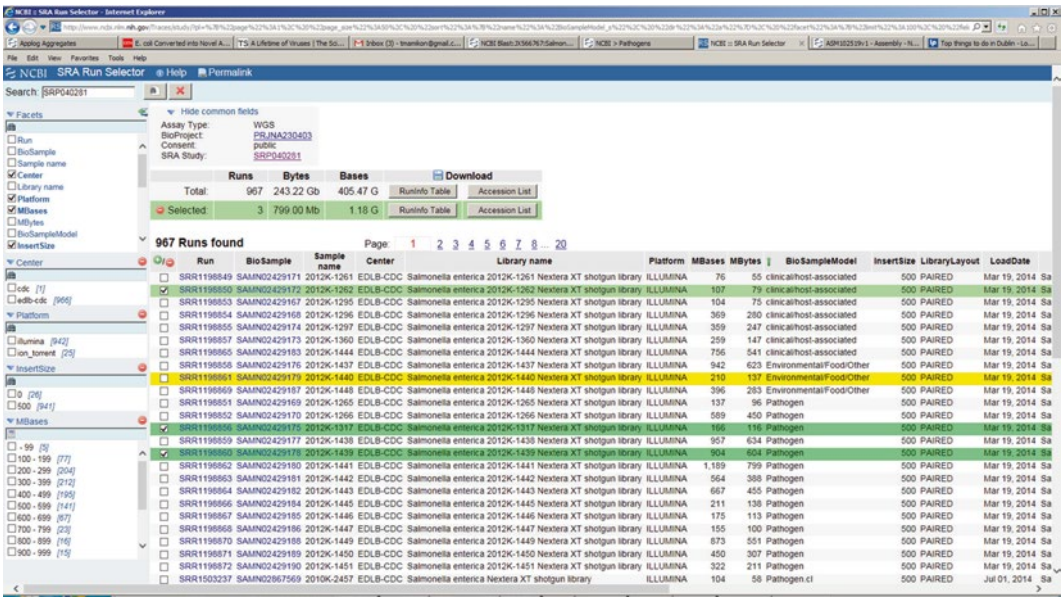


Fig. 6 SRA Run browser

Several initiatives (e.g., Barcode of Life) are explicitly focused on extending sequence coverage to all species of life.

Sequence entries in GenBank are identified with varying degrees of certainty. Some are taken from specimens (or cultures) that can be independently identified by a specialist—some of these come with species-level identifications (formal binomial names), the others get informal names of several sorts. Species with a formal name in the appropriate code of nomenclature are indexed in Taxonomy Entrez with the specified [property]. Taxonomy identifier often serves as the primary key that links together different data types related by organism. More recently, NCBI has started a project to curate sequence from type material [22]. Type material is the taxonomic device that ties formal names to the physical specimens that serve as exemplars for the species. For the prokaryotes these are strains submitted to the culture collections; for the eukaryotes they are specimens submitted to museums or herbaria. The NCBI Taxonomy Database (http://www.ncbi.nlm.nih.gov/taxonomy) now includes annotation of type material that is used to flag sequences from type in GenBank, Genomes, and BLAST (see below).

4.3 GenBank

GenBank is the NIH genetic sequence database, an archival collection of all publicly available DNA sequences [12]. Many journals require submission of sequence information to a database prior to publication to ensure an accession number will be available to appear in the paper. GenBank archives assembled nucleotide sequence data and annotations with descriptive metadata including genome and transcriptome assemblies. Due to the increasing volume of short genome survey sequences (GSS) and expressed sequence tags (EST) generated by high throughput sequencing studies the data in Nucleotide have been split into three search sets: GSS, EST and the rest of nucleotide sequences (nuccore).

These sequences are accessible via Web interface by text Query using Entrez. Searching any of the three databases will provide links to results in the other using sensor mechanism described above (*see* Subheading 2). Unless you know that you are trying to find a specific set of EST or GSS sequences, searching the Nucleotide database (http://www.ncbi.nlm.nih.gov/nuccore/) with general text Query will produce the most relevant results. You can always follow links to results in EST and GSS from the Nucleotide database results.

Quarterly GenBank releases are also downloadable via FTP (*see* Subheading 6).

As of June, 15 2015 GenBank release 208.0 (ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt) contains almost 194 billion bases in over 185 million sequence entries (compare to 80 million in 2008 at the time of previous addition). The data come from the large sequencing centers as well as from small experimental labs.

4.4 Whole Genome Shotgun (WGS)

Genome assembly especially for large eukaryotic genomes with highly repetitive sequence remains one of the major challenges in genome bioinformatics [8–10].

While the cost of sequencing drops down dramatically, the genome assembly still takes considerable amount of time and effort. In some research projects (comparative analysis, population and variation studies) the researches might work with a high quality reference assembly and bunch of lower quality variant genomes for the same species. Thousands of draft genomes are assembled up to the contig level only, sometimes with very low assembly quality (low N50/L50, large number of contigs). These genomes typically remain unannotated. These fragmented genomes with no annotation might not be very useful compared to complete genome with full gene/protein complement. However, the contig sequences can be used for comparative analysis. These draft contig-level assemblies are treated differently than traditional sequence records. The contigs are not loaded to the main sequence repository, general identifiers (GI number) are not assigned, contigs sequences are not indexed and therefore are not searchable in Entrez Nucleotide except for the master record (Fig. 7). These projects can be browsed by organism in a custom made viewer (http://www.ncbi.nlm.nih.gov/Traces/wgs/).

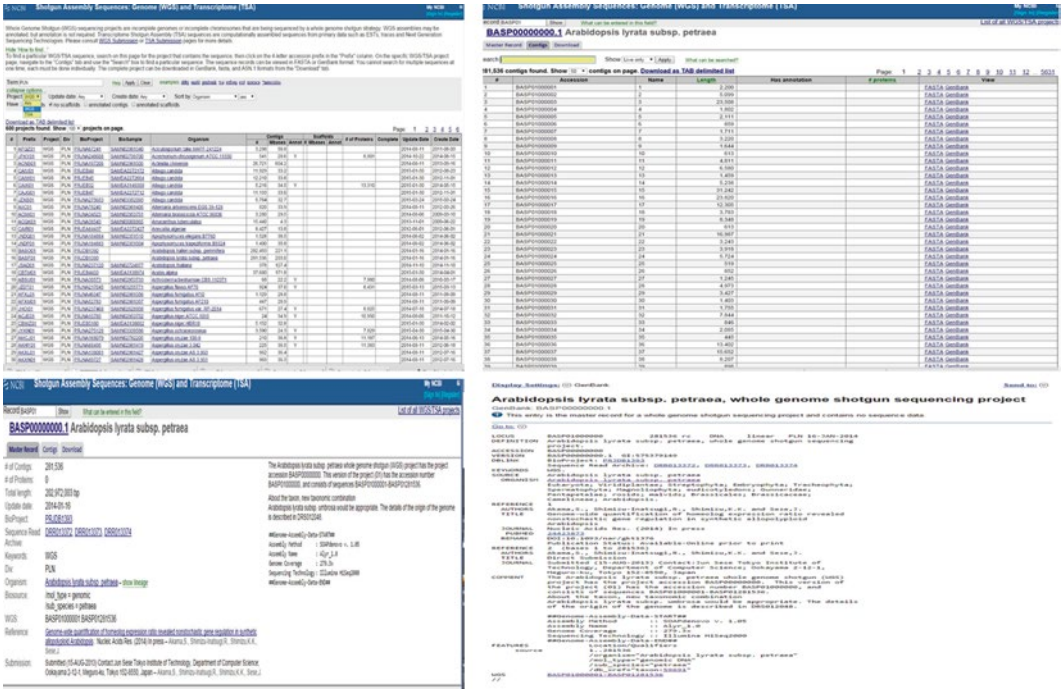


Fig. 7 WGS and TSA customer reports. (**a**) organism browser; (**b**) contig report; (**c**) customized WGS project overview; (**d**) traditional GenBank flat file view of WGS project master record

4.5 Genome Collection Database (Assembly)

The Assembly database (http://www.ncbi.nlm.nih.gov/assembly/) has information about the structure of assembled genomes as represented in an AGP file or as a collection of completely sequenced chromosomes. The database tracks changes to assemblies that are updated by submitting groups over time with a versioned Assembly accession number. The Web resource provides meta-data about assemblies such as assembly names (and alternate names), simple statistical reports of the assembly (type and number of contigs, scaffolds; N50s), and a history view of updates. It also tracks the relationship between an assembly submitted to the International Nucleotide Sequence Database Consortium (INSDC) and the assembly represented in the NCBI Reference Sequence (RefSeq) project. More information can be found at (http://www.ncbi.nlm.nih.gov/assembly/help/#find) 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 genus). Assemblies with partial genome representation can be found in Entrez Assembly database by using the following query:

Archaea[orgn] OR Bacteria[orgn] AND "partial genome representation"[Properties]

Some genome assemblies come from mixed cultures, hybrid organisms and chimeras; these "anomalous" assemblies do not represent an organism. These assemblies are valid results of the experimental studies and are legitimate genome records in GenBank; however, they should be filtered out in genome analysis and comparative genome studies These assemblies can be found in Entrez Assembly database by using the following query:

Archaea[orgn] OR Bacteria[orgn] AND "anomalous" [Properties]

Modern high-throughput sequencing technologies vary in the size of raw sequence reads and the patterns of sequencing errors. Despite many computational advances to genome assembly, complete and accurate assembly from raw sequence read data remains a major challenge. There are two major approaches that have been used: de novo assembly from raw sequence reads and reference guided assembly if the closest reference genome is available. The quality of genome assembly can be assessed using a number of different quality metrics. For many years N50 and L50 contig and scaffold lengths have been major measure of assembly quality. N50 defines the length of contig (or scaffold) for which the set of all contigs of that length or longer contains at least 50 % of the total size of the assembly (sum of the lengths of all contigs). L50 is the number of sequences evaluated at the point when the sum length exceeds 50 % of the assembly size. More recently, a number of different metrics have been suggested [22, 23]. Some of the standard

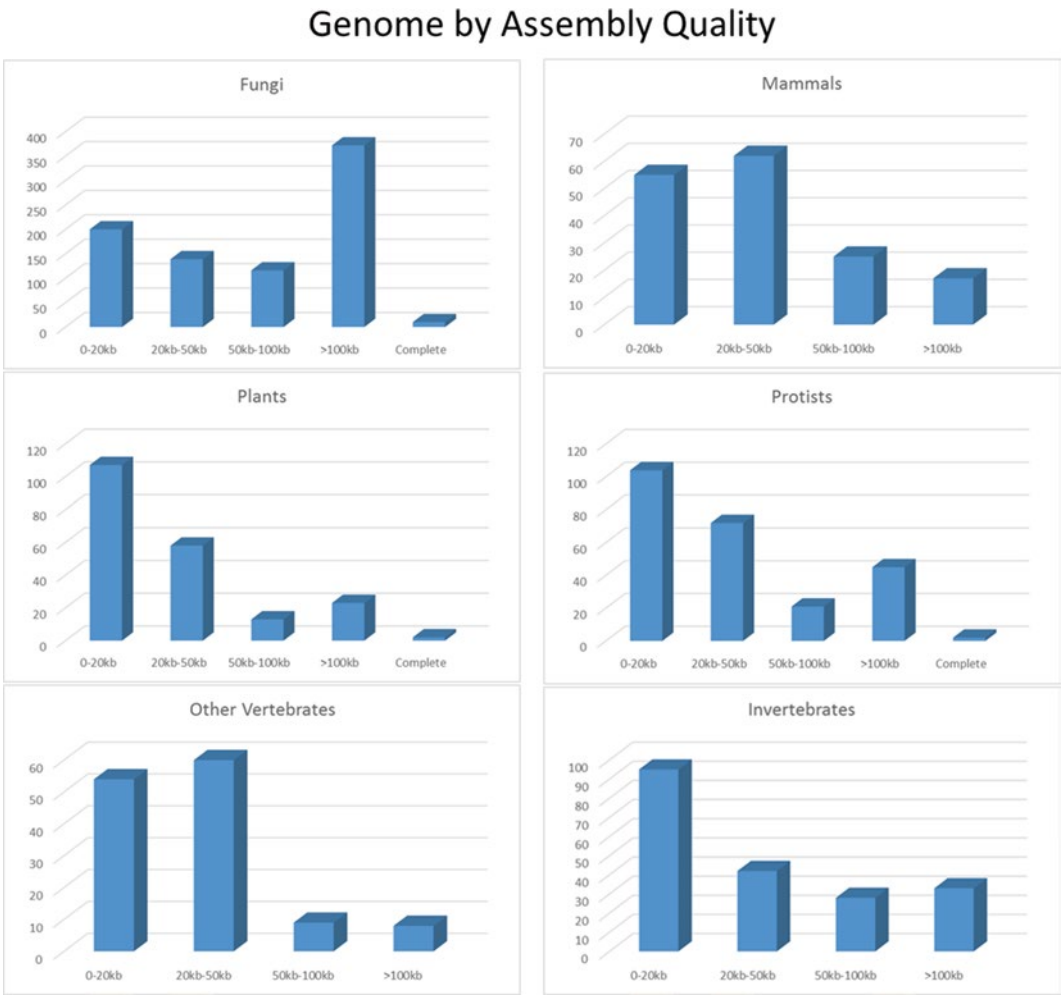


Fig. 8 Genome by assembly quality

global statistics measures and reference based statistics are used for quality assessment. Figure 8 illustrates the differences in N50 for eukaryotic genome assemblies.

The user can access to assembly data by using Entrez text searches from the home Web page: (http://www.ncbi.nlm.nih.gov/assembly/), or by browsing and filtering assemblies by organism, and download the data from the FTP site (*see* Subheading 6).

4.6 BioProject

The BioProject resource [14] became public in May 2011, replacing the older NCBI Genome Project database, which had been created to organize the genome sequences in GenBank [12] and RefSeq [16]. The BioProject database was created to meet the need for an organizational database for research efforts beyond just genome sequencing, such as transcriptome and gene expression, proteomics, and variation studies. However, because a BioProject is defined by its multiple attributes, there is flexibility for additional

types of projects in the future, beyond those that were included in 2011. The new BioProject database allows more flexible grouping of projects and can collect more data elements for each project, e.g., grant information and project relevance.

BioProjects describe large-scale research efforts, ranging from genome and transcriptome sequencing efforts to epigenomic analyses, genome-wide association studies (GWAS), and variation analyses. Data are submitted to NCBI or other INSDC-associated databases citing the BioProject accession, thus providing navigation between the BioProject and its datasets. Consequently, the BioProject is a robust way to access data across multiple resources and multiple submission time points, e.g., when there are different types of data that had been submitted to multiple databases, or sequential submissions deposited over the course of a research project. Web access to all publicly registered bioporjects (http://www.ncbi.nlm.nih.gov/bioproject/) has typical text search and browse by project data types options.

4.7 Genome

Entrez Genome, the integrated database of genomic information at the NCBI, organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations. The genome paradigm has transitioned from a single reference genome of an organisms to multiple genome representing the whole population. To reflect the change in the genome paradigm the Genome database has been completely redesign in 2013. In the past an entry in Genomes database used to represent a complete sequence of a single replicon such as a chromosome, organelle, or plasmid.

New Genome records pull together genome data at various levels of completion, ranging from recently registered projects with SRA/trace data to genomes represented by scaffolds/contigs or fully assembled chromosomes with annotation. Genome information is grouped by organism so that each record in the Entrez Genome database represents a taxonomic node at species level for the most part. In addition, group-specific pages provide links to relevant external websites and databases and to aggregated data and tools.

As of June 2015 Entrez Genomes houses a collection of almost 13,000 entries (organism level) for almost 55,000 assemblies. Table 1 shows the number of genome records (species) and assemblies in major taxonomic groups.

Table 1
Data statistics in major organism groups and data categories

Eukaryota	Prokaryota	Viruses	Organelles	Plasmids	Total	
1424	6726	4658	6268	1024	12,808 (unique)	Genome (species)
2291	35,211	4714	6821	5954	53,991	Assembly

Table 2
Differences between Entrez databases presenting genome and metagenome data

Database	Definition	Central portal	Grouping
Genome	Total genetic content contained within an organism	A single portal to genome sequence and annotation	Defined by organism
BioProject	A set of related data tied together with unique identifier	A higher order organization of the data deposited into several archival databases, it provides a central point to inform customers of data availability in these databases	Defined by submitter, by funding source, by named collaboration
BioSample	Biological material under investigation in a project. The attributes describe the role the sample holds in the project	A single portal to sample description and attributes	Defined by the context of experimental study
Taxonomy	The conception, naming, and classification of organism groups	Organism groups organized in a hierarchical classification	Rank-based biological classification: Kingdom, Phylum, Class, Order, Family, Genus, Species
Assembly	A data structure that maps the sequence data to a putative reconstruction of the target	Assembly structure, assembly version history	Primary data defined by submitter or Refseq data defined by NCBI staff
Nucleotide	A collection of genomic and transcript sequences	A single portal to all DNA and RNA sequences	Primary data defined by submitter or Refseq data defined by NCBI staff

The BioProject, Genome, and Assembly databases are interconnected and can be used to access and view genomes in different ways. Every prokaryotic and eukaryotic genome submission has BioProject, BioSample, Assembly, and GenBank accession numbers, so users can start in any of those resources and get to the others. The BioProject and BioSample databases allow users to find related datasets, e.g., multiple bacterial strains from a single isolation location, or the transcriptome and genome from a particular sample. The Assembly accession is assigned to the entire genome and is used to unambiguously identify the set of sequences in a particular version of a genome assembly from a particular submitter. Finally, the Genome database displays all of the genome assemblies in INSDC and RefSeq, organized by organism. A brief description of genome-related resources is summarized in Table 2.

Genome information is accessible via Entrez text-based search Query or by browsing sortable tables organized by organism and BioProject accession. Links to Genome records may be found in several other Entrez databases including Taxonomy, BioProject, Assembly, PubMed, Nucleotide, Gene and Protein. Accessing actual sequence data, for example, all the nucleotide sequence data from a particular WGS genome is easily found via the Genome database from the organism overview page or browser table in two ways. First, by using the link to the assembly database and following the link to the Nucleotide database located under the related information heading. Second, from the assembly database the link to the WGS browser provide access to a table with a list of contigs and statistics as well as GNU zipped archive files for download.

4.7.1 Genome Browser

The browser (http://www.ncbi.nlm.nih.gov/genome/browse/) divided in four tables each with genome attributes statistics. Table summaries include (1) an overview by organism and then lists of genome sequencing assemblies for (2) Eukaryotes, (3) Prokaryotes, (4) Viruses, (5) Organelles, and (6) Plasmids. Table displays can be filtered by lineage information and/or genome status and the results downloaded. Various filters can be applied to create a data set of interest. The Genome top level records can be filtered by the organisms groups at phylum and/or family level. Assemblies can be filtered by completeness and the highest level off assembly (complete, chromosome, scaffold, contig). Selected records can be downloaded in tab-delimited text files. The whole report can be downloaded from this FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/).

4.7.2 Entrez Text-Based Searches

The organism search is expected to be the most frequent method to look for the data in Entrez Genome. One of the features of the Entrez system is "organism sensor" allowing to recognize a organism name in a query. Advanced searching in Entrez Genome allows for refined Query by specifying "organism" (or short version "orgn") field in square brackets (e.g., yeast[orgn]). When a search term, for example "human" is recognized as an organism name, the original query is transformed to the organism-specific one "Homo sapiens [Organism] OR human[All Fields]". Take note, a nonspecific search term such as "human" can result in the listing of several genome records which contain the word "human" as part of the text. However, the Entrez Genome record for human will be at the top of the list since the query has been transformed (http://www.ncbi.nlm.nih.gov/genome/?term=human). Only the search term "human [orgn]" or latin binomial "*Homo sapiens*[orgn]" will provide the specific Genome page for *Homo sapiens* (http://www.ncbi.nlm.nih.gov/genome/?term=human+[orgn]). A list of all fields indexed for a more refined search is available in the Genome Advanced

Search Builder (http://www.ncbi.nlm.nih.gov/genome/advanced). The Limits page provides an easy way to limit a search by certain fields without having to use complex Boolean operations. Genome searches can be limited by organism groups or cell location of genetic content (chromosome, organelle, or plasmid).

4.7.3 Organelle and Plasmid

An organelle is a specialized structure that is enclosed within its own membrane inside a eukaryotic cell. The mitochondria and chloroplasts are maintained throughout the cell cycle, and replicate independently of the mitosis or meiosis of the cell. Mitochondrial and chloroplast DNA sequences are often used for phylogenetic analysis and population genetics, as well as cultivar identification and forensic studies. Due to the relatively small size, conserved gene order, and content of animal mitochondria, whole genome sequencing and comparisons across many species have been possible for many decades. NCBI maintains a special collection of reference organelle genomes. However, the organelle genome alone does not represent the full genetic content of an organism. The Entrez Genome organism search does not include organelle and plasmid data in the result listing but provides a short summary at the top of the search page linked to Genome records with organelle or plasmid data only. The search results are automatically weighted by scientific relevance, high quality genomes, model organisms will be shown at the top of the list. For example, the search for "Fungi" will result in 650 species-level records; Saccharomyces cerevisiae, the most studied model organism will be shown at the top of the list. The individual Genome report include Organism Overview, Genome Assembly and Annotation Report, and Protein Table. Organism Overview typically contains a short description of the organism, a picture if available, lineage as defined by NCBI Taxonomy, related publications, and summary statistics for the genome sequence data. For many species hundreds and thousands of genome assemblies are being sequenced and the number continues to grow. In Genome database a reference genome assembly is selected to serve as a single representative of a particular organism. A representative genome or genomes are chosen either by the community or calculated by comparative sequence analysis (see more details in [17]). Genomes of the highest quality sequence and annotation, often the most important isolates, historically used by the research community for clinical studies, experimental validation are marked as "Reference" genomes. One of the best known examples of the "Reference" genome is the genome of the non-pathogenic strain K-12 of Escherichia coli first obtained from a patient in 1922. The genome has been sequenced in 1997 [24] and was extensively curated by the research community ever since [25]. The genome information panel that provide a quick and easy access to the sequence data for the representative (or

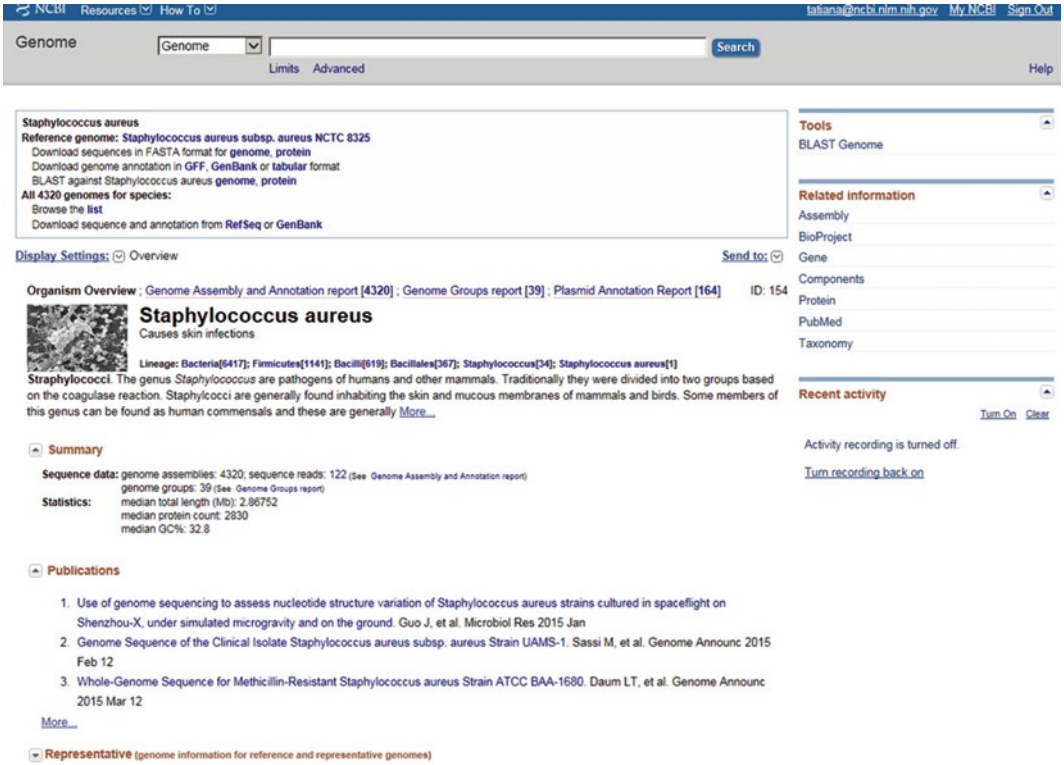


Fig. 9 Genome reports: organism overview

reference) genome is provided at the very top of the Organism Overview page (Fig. 9). The list of all representative and reference prokaryotic genomes is available at http://www.ncbi.nlm.nih.gov/genome/browse/reference/.

4.7.4 Graphical View

Since the genome structure and biology of eukaryotes, prokaryotes and viruses are very different, the genomic data display between taxonomic groups varies to some extent. Virus genomes are small enough to display the whole annotated genome in graphic form and also have links to a virus specific genome resource. Hundreds of prokaryotic genomes are available for particular species thus making the display of the relationship of prokaryote genomes from a specific bacterial species relevant. Genome relationships are displayed in the form of a dendrogram based on genomic BLAST scores (Fig. 10). In addition, the prokaryotic genome can be displayed in a graphic form (like a virus genome) when an individual strain is selected from the dendrogram or table. The human Genome record on the other hand has a detailed ideogram of the 24 chromosomes with links to MapViewer. The ideogram display

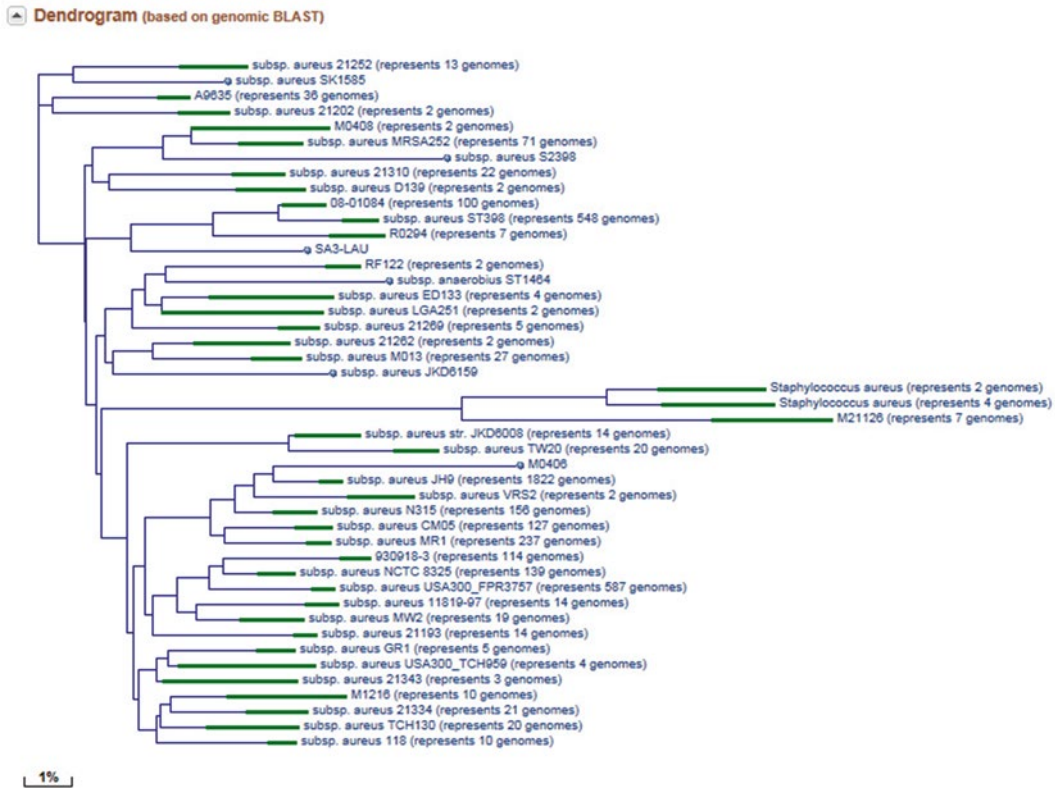


Fig. 10 Genome reports: BLAST based dendrogram

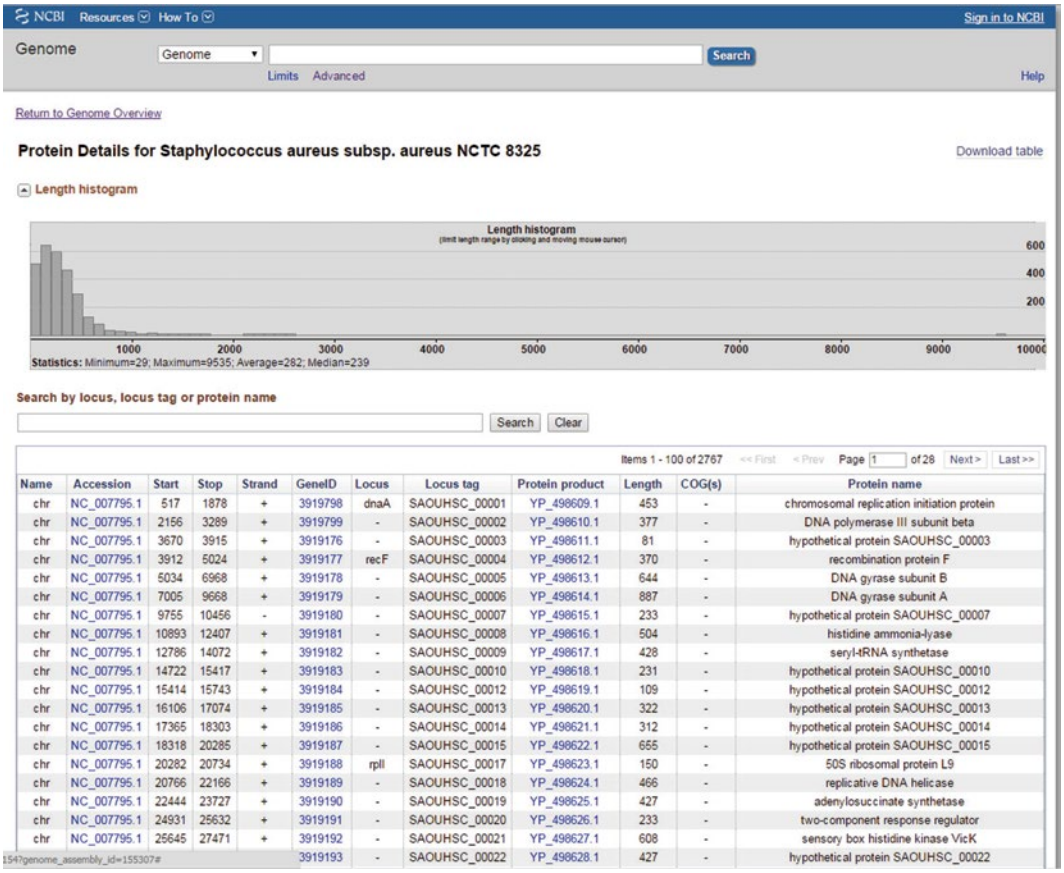
for a eukaryote organism originates from a single representative genome and may not represent the full variation in karyotypes observed in particular organisms (e.g., Fungi).

4.7.5 Assembly and Annotation Report

This section provides full details of the assembly and annotation (different feature types) for each assembly represented in a single Genome record (usually species). Microbial genomes represented by thousands of isolates are organized in clades and tight genomes groups calculated by sequence similarity as described in [17].

4.7.6 Protein Details Report

The Protein Details page provides a length histogram with descriptive statics (minimum, maximum, average and median) of all the relevant proteins listed in the table which expand over several pages (Fig. 11). Details about each protein are given in each row which includes protein name, accession, locus tag, location coordinates, strand info, length, a related structure link and links to other NCBI resources such as Entrez Gene, Protein, and Protein Clusters.



Name	Accession	Start	Stop	Strand	GeneID	Locus	Locus tag	Protein product	Length	COG(s)	Protein name
chr	NC_007795.1	517	1878	+	3919798	dnaA	SAOUHSC_00001	YP_498609.1	453	-	chromosomal replication initiation protein
chr	NC_007795.1	2156	3289	+	3919799	-	SAOUHSC_00002	YP_498610.1	377	-	DNA polymerase III subunit beta
chr	NC_007795.1	3670	3915	+	3919176	-	SAOUHSC_00003	YP_498611.1	81	-	hypothetical protein SAOUHSC_00003
chr	NC_007795.1	3912	5024	+	3919177	recF	SAOUHSC_00004	YP_498612.1	370	-	recombination protein F
chr	NC_007795.1	5034	6968	+	3919178	-	SAOUHSC_00005	YP_498613.1	644	-	DNA gyrase subunit B
chr	NC_007795.1	7005	9668	+	3919179	-	SAOUHSC_00006	YP_498614.1	887	-	DNA gyrase subunit A
chr	NC_007795.1	9755	10456	-	3919180	-	SAOUHSC_00007	YP_498615.1	233	-	hypothetical protein SAOUHSC_00007
chr	NC_007795.1	10893	12407	+	3919181	-	SAOUHSC_00008	YP_498616.1	504	-	histidine ammonia-lyase
chr	NC_007795.1	12786	14072	+	3919182	-	SAOUHSC_00009	YP_498617.1	428	-	seryl-tRNA synthetase
chr	NC_007795.1	14722	15417	+	3919183	-	SAOUHSC_00010	YP_498618.1	231	-	hypothetical protein SAOUHSC_00010
chr	NC_007795.1	15414	15743	+	3919184	-	SAOUHSC_00012	YP_498619.1	109	-	hypothetical protein SAOUHSC_00012
chr	NC_007795.1	16106	17074	+	3919185	-	SAOUHSC_00013	YP_498620.1	322	-	hypothetical protein SAOUHSC_00013
chr	NC_007795.1	17365	18303	+	3919186	-	SAOUHSC_00014	YP_498621.1	312	-	hypothetical protein SAOUHSC_00014
chr	NC_007795.1	18318	20285	+	3919187	-	SAOUHSC_00015	YP_498622.1	655	-	hypothetical protein SAOUHSC_00015
chr	NC_007795.1	20282	20734	+	3919188	rplI	SAOUHSC_00017	YP_498623.1	150	-	50S ribosomal protein L9
chr	NC_007795.1	20766	22166	+	3919189	-	SAOUHSC_00018	YP_498624.1	466	-	replicative DNA helicase
chr	NC_007795.1	22444	23727	+	3919190	-	SAOUHSC_00019	YP_498625.1	427	-	adenylosuccinate synthetase
chr	NC_007795.1	24931	25632	+	3919191	-	SAOUHSC_00020	YP_498626.1	233	-	two-component response regulator
chr	NC_007795.1	25645	27471	+	3919192	-	SAOUHSC_00021	YP_498627.1	608	-	sensory box histidine kinase VicK
					3919193	-	SAOUHSC_00022	YP_498628.1	427	-	hypothetical protein SAOUHSC_00022

Fig. 11 Genome reports: The Protein Details page provides a length histogram with descriptive statics and a table of protein information and links to related NCBI resources

5 Searching Data by Sequence Similarity (BLAST)

The Basic Local Alignment Search Tool (BLAST) [20] finds regions of local similarity between sequences. By finding similarities between sequences, scientists can infer the function of newly sequenced genes, predict new members of gene families, and explore evolutionary relationships. New features include searching against SRA experiments, easy access to genomic BLAST databases by using auto-complete organism query option, Redesigned BLAST pages include new limit options; and a Tree View option that presents a graphical dendrogram display of the BLAST results.

5.1 Exploring NGS Experiments with SRA-BLAST

SRA-BLAST offers two different ways of finding data sets to search. The BLAST service itself provides an autocomplete feature under "Choose Search Set" that finds matches to experiment, study, and run accessions as well as text from experiment descriptions.

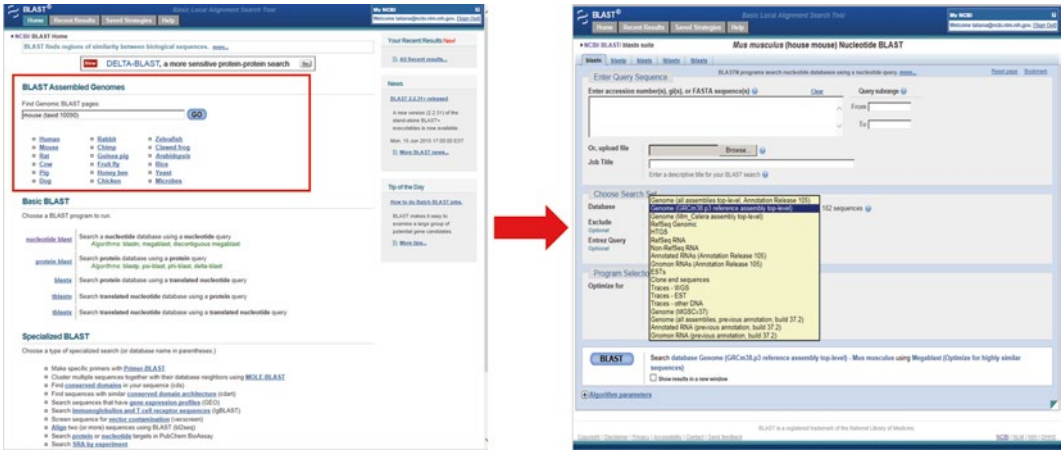


Fig. 12 New feature on BLAST home page provide an easy way to organism-specific blast with assembled genomes

You can now also use the Entrez SRA system to identify experiments of interest and load these as BLAST databases in SRA BLAST through the "Send to" menu from the SRA search results

5.2 BLAST with Assembled Eukaryotic Genomes

In addition to the BLAST home page, the BLAST search tool can also be found on the Organism Overview page of the Genome database. Accessed from the Organism Overview page this search tool has BLAST databases limited to genome data of that specific organism. For each organism, if the data exist, the following default list of organism specific BLAST databases are available to search against: HTG sequences, ESTs, clone end sequences, RefSeq genomic, RefSeq RNA, RefSeq protein, non-RefSeq RNA, and non-RefSeq protein. These BLAST databases are defined by Entrez Query. In addition, some organisms have custom databases available. Specifying the BLAST database to genomes of a specific taxon is not only limited to the search tool found on the Organism Overview page in the Genome database. BLAST databases can be limited to any taxonomic level at the BLAST home page. For example Fig. 12 shows how to start searching against mouse genome from BLAST home page and select the search set from all datasets available for mouse on the specialized Mouse BLAST page.

5.3 Microbial Genomic BLAST: Reference and Representatives

Microbial Genomic BLAST provides access to complete and Whole Genome Sequence (WGS) draft assemblies, and plasmids. Sequenced microbial genomes represent a large collection of strains with different levels of quality and sampling density. Largely because of interest in human pathogens and advances in sequencing technologies, there are rapidly growing sets of very closely related genomes representing variations within the species. Many bacterial species are represented

in the database in thousands of variant genomes. If the users are interested in multi-species comparative analysis they would need a single genomes which is designated to represent a species. Refseq group at NCBI has introduced new categories of "reference" and "representative" genomes defined as following.

Reference Genome—manually selected "gold standard" complete genomes with high quality annotation and the highest level of experimental support for structural and functional annotation. They include community curated genomes if the annotation quality meets "reference genome" requirements that are manually reviewed by NCBI staff (http://www.ncbi.nlm.nih.gov/genome/browse/reference/).
Representative Genome—representative genome for an organism (species); for some diverse species can be more than one. Corresponds to Sequence Ontology—[SO:0001505] [10] (www.ncbi.nlm.nih.gov/genome/browse/representative/).

The users interested in the organism diversity in BLAST results have an option to select a search database of reference and representative genomes only.

6 FTP Resources for Genome Data

NCBI has redesigned the genomes FTP site to expand the content and facilitate data access through an organized predictable directory hierarchy with consistent file names and formats. The updated site provides greater support for downloading assembled genome sequences and/or corresponding annotation data. The new FTP site structure provides a single entry point to access content representing either GenBank or Refseq data. More detailed information can be found at (http://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/).

Refseq dataset is organized by major taxonomic groups. It provides curated sequence records for genomes, transcripts, and proteins.

Download the curated RefSeq full release or daily updates (ftp://ftp.ncbi.nih.gov/refseq/).

7 Conclusion

The tremendous increase in genomic data in the last 20 years has greatly expanded our understanding of biology. Genome sequencing projects now span from draft assemblies, complete genomes, large-scale comparative genomic projects, and the new field of metagenomics where genetic material is recovered directly from environmental samples and the entire complement of DNA from a given ecological niche is sequenced. Although these provide an ever greater resource for studying biology, there is still a long way to go from the initial submission of sequence data to the

understanding of biological processes. By integrating different types of biological and bibliographical data, NCBI is building a discovery system that enables the researcher to discover more than would be possible from just the original data. By making links between different databases and computing associations within the same database, Entrez is designed to infer relationships between different data that may suggest future experiments or assist in interpretation of the available information. In addition, NCBI is developing the tools that provide users with extra layers of information leading to further discoveries.

Genomics is a very rapidly evolving field. The advance in sequencing technologies has lead to new data types which require different approaches to data management and presentation. NCBI continues to add new databases and develop new tools to address the issue of ever increasing amounts of information.

Acknowledgements

The authors would like to thank, in alphabetic order, Boris Fedorov and Sergei Resenchuk for their expertise and diligence in the design and maintenance of the databases highlighted in this publication and Stacy Ciufo for the helpful discussion and comments. These projects represent the efforts of many NCBI staff members along with the collective contributions of many dedicated scientists worldwide.

References

1. Matsen FA (2015) Phylogenetics and the human microbiome. Syst Biol 64(1):e26–e41, Review
2. Hedlund BP, Dodsworth JA, Murugapiran SK, Rinke C, Woyke T (2014) Impact of single-cell genomics and metagenomics on the emerging view of extremophile "microbial dark matter". Extremophiles 18(5):865–875, Review
3. Vernikos G, Medini D, Riley DR, Tettelin H (2015) Ten years of pan-genome analyses. Curr Opin Microbiol 23:148–154, Review
4. Henson J, Tischler G, Ning Z (2012) Next-generation sequencing and large genome assemblies. Pharmacogenomics 13(8):901–915
5. Wang Y, Navin NE (2015) Advances and applications of single-cell sequencing technologies. Mol Cell 58(4):598–609
6. Feng Y, Zhang Y, Ying C, Wang D, Du C (2015) Nanopore-based fourth-generation DNA sequencing technology. Genomics Proteomics Bioinformatics 13(1):4–16
7. Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, Mburu FM, Mantalas GL, Sim S, Clarke MF, Quake SR (2014) Quantitative assessment of single-cell RNA-sequencing methods. Nat Methods 11(1):41–46
8. Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM (2015) Assembling large genomes with single-molecule sequencing and locality sensitive hashing. Nat Biotechnol 33(6):623–630
9. Madoui MA, Engelen S, Cruaud C, Belser C, Bertrand L, Alberti A, Lemainque A, Wincker P, Aury JM (2015) Genome assembly using Nanopore-guided long and error-free DNA reads. BMC Genomics 16(1):327
10. Koren S, Phillippy AM (2015) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol 23:110–120
11. Silvester N, Alako B, Amid C, Cerdeño-Tárraga A et al (2015) Content discovery and retrieval services at the European Nucleotide

Archive. Nucleic Acids Res 43(Database issue): D23–D29

12. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2015) GenBank. Nucleic Acids Res 43(Database issue):D30–D35
13. Kodama Y, Mashima J, Kosuge T, Katayama T, Fujisawa T, Kaminuma E, Ogasawara O, Okubo K, Takagi T, Nakamura Y (2015) The DDBJ Japanese Genotype-phenotype Archive for genetic and phenotypic human data. Nucleic Acids Res 43(Database issue):D18–D22
14. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, Kimelman M, Pruitt KD, Resenchuk S, Tatusova T, Yaschenko E, Ostell J (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Res 40(Database issue):D57–D63
15. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database Collaboration (2012) The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res 40(Database issue): D54–D56
16. Pruitt KD, Brown GR, Hiatt SM et al (2014) RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 42(Database issue):D756–D763
17. Tatusova T, Ciufo S, Federhen S, Fedorov B, McVeigh R, O'Neill K, Tolstoy I, Zaslavsky L (2015) Update on RefSeq microbial genome resources. Nucleic Acids Res 43(Database issue):D599–D605
18. NCBI Resource Coordinators (2015) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 43 (Database issue):D6–D17
19. Salzberg SL, Church D, DiCuccio M, Yaschenko E, Ostell J (2004) The genome Assembly Archive: a new public resource. PLoS Biol 2(9), E285
20. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402, Review
21. Federhen S (2012) The NCBI Taxonomy database. Nucleic Acids Res 40:D13–D25
22. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29(8): 1072–1075
23. Rahman A, Pachter L (2013) CGAL: computing genome assembly likelihoods. Genome Biol 14(1):R8
24. Blattner FR, Plunkett G 3rd, Bloch CA et al (1997) The complete genome sequence of Escherichia coli K-12. Science 277(5331): 1453–1462
25. Riley M, Abe T, Arnaud MB, Berlyn MK et al (2006) Escherichia coli K-12: a cooperatively developed annotation snapshot--2005. Nucleic Acids Res 34(1):1–9