

# Chapter 4

## Managing Sequence Data

Christopher O’Sullivan, Benjamin Busby, and Ilene Karsch Mizrahi

### Abstract

Nucleotide and protein sequences are the foundation for all bioinformatics tools and resources. Researchers can analyze these sequences to discover genes or predict the function of their products. The INSDC (International Nucleotide Sequence Database—DDBJ/ENA/GenBank + SRA) is an international, centralized primary sequence resource that is freely available on the Internet. This database contains all publicly available nucleotide and derived protein sequences. This chapter discusses the structure and history of the nucleotide sequence database resources built at NCBI, provides information on how to submit sequences to the databases, and explains how to access the sequence data.

**Key words** Sequence database, GenBank, SRA, INSDC, RefSeq, Next generation sequencing

---

### 1 Structure and History of Sequence Databases at NCBI

The National Center for Biotechnology Information (NCBI) is responsible for building and maintaining the sequence databases Sequence Read Archive (SRA), GenBank, and RefSeq. GenBank and SRA are primary archival resources that collect data from researchers as part of the publication process. RefSeq is a secondary database built from data submitted to the primary archives but with added curation. GenBank and SRA are part of the International Nucleotide Sequence Database Collaboration (INSDC) a centralized public sequence resource that encompasses raw sequence reads, assembled sequences and derived annotations.

#### 1.1 INSDC

The INSDC [1] is a partnership between GenBank/SRA at NCBI in the USA [2] <https://www.ncbi.nlm.nih.gov/>, the European Nucleotide Archive (ENA) at EMBL-EBI in Europe [3] <http://www.ebi.ac.uk/ena>, and DNA Databank of Japan (DDBJ) in Japan [4] <http://www.ddbj.nig.ac.jp/>. For more than 25 years, the three partners have maintained this active, successful collaboration for building the nucleotide sequence databases. As new sequencing technologies have emerged, additional archives have been

incorporated to store raw and aligned sequence read data. Representatives from the three databases meet annually to discuss technical and biological issues affecting the databases. Ensuring that sequence data from scientists worldwide is freely available to all is the primary mission of this group. As part of the publication process, scientists are required to deposit sequence data in a public repository; the INSDC encourages publishers of scientific journals to enforce this policy to ensure that sequence data associated with a paper is freely available from an international resource for research and discovery. For example, in 2011, analysis of genomic sequence data in INSDC led to the identification of the enterohemorrhagic *Escherichia coli* that caused numerous deaths in Germany [5].

The INSDC website, <http://www.insdc.org>, contains links to the member sites, data release policy, the Feature Table Document (which outlines features and syntax for sequence records), lists of controlled vocabularies and procedures that are important for the collaborators, data submitters and database users.

Though each of the three INSDC partners has its own set of tools for submission and retrieval, data is exchanged regularly so that all content is made available at each of the member sites.

The following table (Table 1) from the INSDC website contains links for retrieval of data at the three partner sites (*see Note 1*).

In this chapter, we discuss NCBI’s contribution to INSDC, including information about submission processing and data usage. Similar tools and processes can be found at the other partner sites.

## 1.2 SRA/GEO

The NCBI SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Illumina, Applied Biosystems SOLiD, Complete Genomics, Pacific Biosciences SMRT, Nanopore, Ion Torrent, and Roche 454. The

**Table 1**  
**Links for data retrieval from DDBJ, EMBL-EBI, and NCBI**

Data type	DDBJ	EMBL-EBI	NCBI
Next generation reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive	European Nucleotide Archive (ENA)	Trace Archive
Annotated sequences	DDBJ	European Nucleotide Archive (ENA)	GenBank
Samples	BioSample	European Nucleotide Archive (ENA)	BioSample
Studies	BioProject	European Nucleotide Archive (ENA)	BioProject

data stored in the NCBI SRA are suitable for the reanalysis of data that supports publications as well as data that supports assembly, annotation, variation, and expression data submissions to other NCBI archives. As of October 2016, SRA contains more than 9 PetaBases ( $9 \times 10^{15}$  bases) from nearly 49,000 different source organisms and metagenomes. Approximately half of the total is controlled-access human clinical sequence data supporting dbGaP studies. SRA growth is explosive, doubling approximately every 12 months. Statistics are updated daily and available from the SRA home page (<https://trace.ncbi.nlm.nih.gov/Traces/sra/>). SRA stores descriptive metadata and sequence data separately using distinct accession series: the SRA Experiment and Run. The SRA Experiment record is used for search and display. It contains details describing sequence library preparation, molecular and bioinformatics workflows and sequencing instruments. The SRA Run contains sequence, quality, alignment, and statistics from a specific library preparation for a single biological sample. Every SRA Experiment references a BioSample and a BioProject.

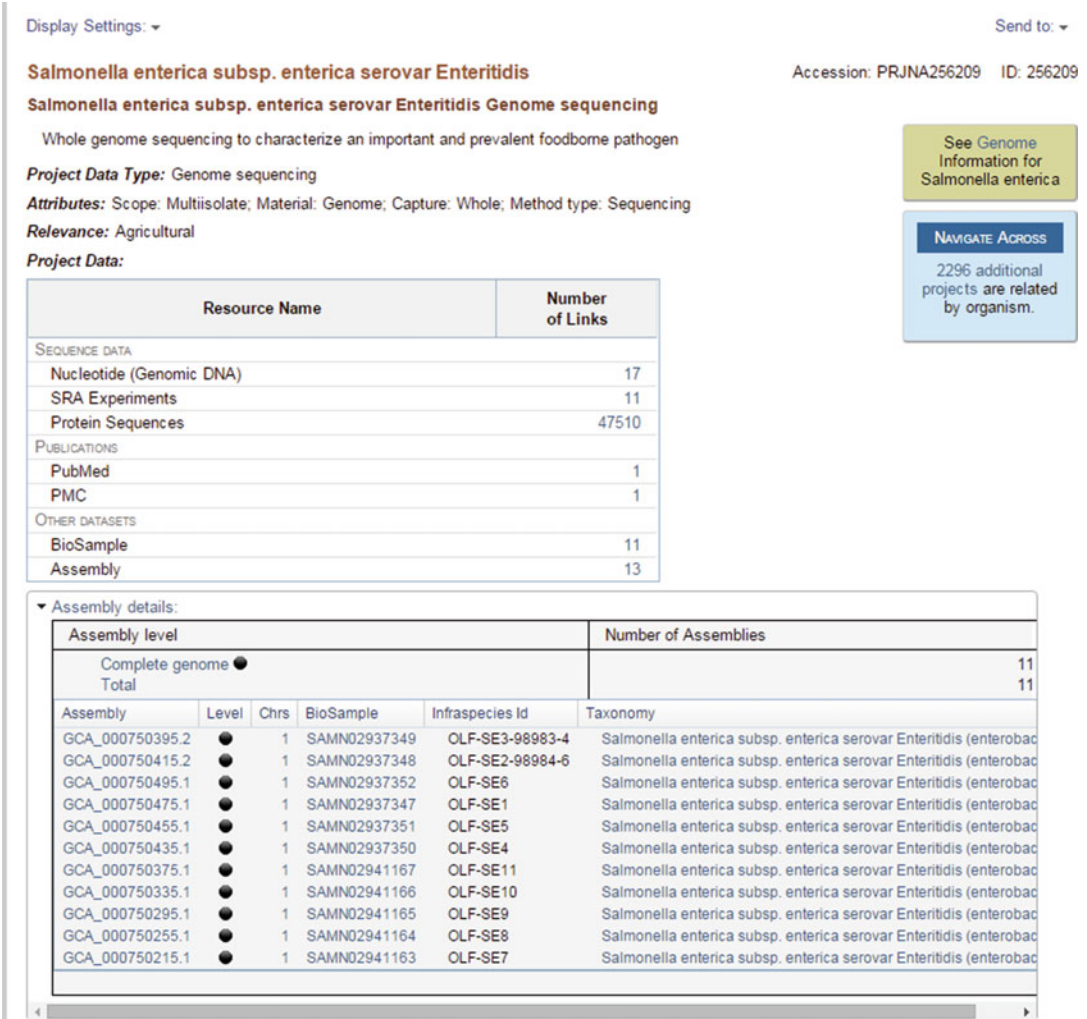
GEO [6], The Gene Expression Omnibus, is a public functional genomics data repository supporting MIAME-compliant data submissions. An integral part of NCBI's primary data archives, GEO stores profiles of gene expression, coding and noncoding RNA, Chromatin Immunoprecipitation, genome methylation, genome variation, and SNP arrays derived from microarrays and next-generation sequencing. Raw sequence data submitted in support of GEO profiles is stored in SRA. The profiles and underlying sequence data are linked via BioSample and BioProject references.

### 1.3 **BioProject**

A BioProject is a collection of biological data related to a single initiative, such as a grant, manuscript, consortia project or other collection. A BioProject record provides users a single place to find links to the diverse data types generated for that project (Fig. 1). For instance, a multiisolate genome sequencing project for a food-borne pathogen *Salmonella enterica* subsp. *enterica* serovar Enteritidis contains links to the SRA reads, the genome assemblies, the BioSamples, and the publications.

### 1.4 **BioSample**

The BioSample database contains descriptive information about biological source materials from which data stored in INSDC sequence archives are derived. BioSample records describe cell lines, whole organisms, and environmental isolates using structured and consistent attribute names and values. The information is important to provide context to derived data to facilitate analysis and discovery. It also acts to aggregate and integrate disparate data sets submitted to any number of resources.



**Fig. 1** BioProject report from Entrez. This report provides information about an initiative for sequencing *Salmonella enterica*, a common foodborne pathogen. The report includes links to the GenBank assemblies, the SRA reads, the BioSamples, and the publications. There are also links to other related projects

### 1.5 GenBank

GenBank [2] is a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation. GenBank is maintained and distributed by the NCBI, a division of the National Library of Medicine (NLM) at the US National Institutes of Health (NIH) in Bethesda, MD. NCBI builds GenBank from several sources including the submission of sequence data from individual researchers and from the bulk submission of whole genome shotgun (WGS), transcriptome shotgun assembly (TSA) and other high-throughput data from sequencing centers. GenBank doubles in size every 18 months. In October 2016, Release 216 contained over 220 billion bases in 197 million

sequences. WGS and TSA projects contribute another terabase of sequence data. GenBank contains sequences from over 373,000 named species. Full releases of GenBank are made every 2 months beginning in the middle of February each year. Between full releases, daily updates are provided on the NCBI FTP site. Detailed statistics for the current release can be found in the GenBank release notes <ftp://ftp.ncbi.nlm.nih.gov/genbank/README.genbank>.

## 1.6 Genomes

The first complete microbial genome was submitted to GenBank in 1995 [7]. Since then, more than 151,000 cellular genomes have been released into the public archive. These genomes are derived from organisms from all branches of the tree of life. Microbial genomes, those from bacteria, archaea and fungi, are relatively small in size compared with their eukaryotic counterparts, ranging from hundreds of thousands to millions of bases. Nonetheless, these genomes contain thousands of genes, coding regions, and structural RNAs. Many of the genes in microbial genomes have been identified only by similarity to other genomes and their gene products are often classified as hypothetical proteins. Each gene within a genome is assigned a `locus_tag`: a unique identifier for a particular gene in a particular genome. Since the function of many of the genes is unknown, `locus_tags` have become surrogate gene identifiers. Submitters of prokaryote genomes are encouraged to use the NCBI Prokaryotic Genome Annotation Pipeline [8] to annotate genome submissions. This service provides standardized annotation for genomes, which can easily be compared across genomes. This pipeline is also used to annotate prokaryotic genomes for RefSeq.

The first genome sequences were built by sequencing overlapping clones and then assembling the genome by overlap. Many bacterial genomes and the first human genome were sequenced in this manner. In 2001, a new approach for sequencing complete genomes was introduced: Whole Genome Shotgun (WGS) data are generated by breaking the genome into random fragments for sequencing and then computationally assembling them to form contigs which can be assembled into larger structures called scaffolds. All of the contig sequences from a single genome assembly, along with the instructions for building scaffolds, are submitted together as a single WGS project. WGS has become such a dominant sequencing technique that more nucleotides of WGS sequence have been added to INSDC than from all of the other divisions since the inception of the database.


For each project, a master record is created that contains information that is common among all the records of the sequencing projects, such as the biological source, submitter information, and publication information. Each master record includes links to the range of accession numbers for the individual contigs in the assembly and links to the range of accessions for the scaffolds from the

project. The NCBI Assembly resource (<https://www.ncbi.nlm.nih.gov/assembly/>) stores statistics about each genome assembly. Each genome assembly is also linked to a BioProject and BioSample.

In response to the anticipated rapid growth in the submission of highly redundant prokaryotic genome sequences from clinical samples and other large sequencing projects, a new model for prokaryotic protein sequences has been employed at NCBI in which a single nonredundant protein record is used for all identical protein sequences annotated on different genome records. The accession numbers for these nonredundant proteins begins with WP. In Entrez Protein, there is a link to an Identical Protein report (Fig. 2) that displays the database source, chromosomal localization, protein accession, protein name, organism, and other taxonomic information for this protein. At the time of this writing only RefSeq bacterial genomes are being annotated with the nonredundant proteins but in the future, these will be used for GenBank annotation, as well. However, the identical protein report encompasses both GenBank and RefSeq genomes.

**1.7 Metagenomes and Environmental Sample Sequencing**

The microbial biodiversity of an environment can be discerned by studying the genome and transcriptome sequences of the microorganisms living in that environment. This field of study is known as metagenomics. Sequencing of 16S ribosomal RNA is a popular way to detect bacterial and archaeal species present in a community.

 This record is a non-redundant protein sequence. Please [read more here](#).

**MULTISPECIES: transcriptional regulator [Afipia]**

NCBI Reference Sequence: WP\_006020932.1  
[GenPept](#) [FASTA](#) [Graphics](#)

RefSeq Selected Product: WP\_006020932.1, 121 amino acids  
 Name: MULTISPECIES: transcriptional regulator [Afipia]

Source	CDS Region in Nucleotide	Protein	Name	Organism	Strain	Superkingdom
RefSeq	<a href="#">NZ_AVBK01000016.1:30820-31185</a> (-)	<a href="#">WP_006020932.1</a>	MULTISPECIES: transcriptional regulator [Afipia]	<a href="#">Afipia sp.</a> <a href="#">NBIMC_P1-C1</a>	NBIMC_P1-C1	<a href="#">Bacteria</a>
RefSeq	<a href="#">NZ_AVBL01000027.1:52526-52891</a> (-)	<a href="#">WP_006020932.1</a>	MULTISPECIES: transcriptional regulator [Afipia]	<a href="#">Afipia sp.</a> <a href="#">NBIMC_P1-C2</a>	NBIMC_P1-C2	<a href="#">Bacteria</a>
RefSeq	<a href="#">NZ_AVBM01000025.1:30822-31187</a> (-)	<a href="#">WP_006020932.1</a>	MULTISPECIES: transcriptional regulator [Afipia]	<a href="#">Afipia sp.</a> <a href="#">NBIMC_P1-C3</a>	NBIMC_P1-C3	<a href="#">Bacteria</a>
RefSeq	<a href="#">NZ_KB375282.1:2382517-2382882</a> (-)	<a href="#">WP_006020932.1</a>	MULTISPECIES: transcriptional regulator [Afipia]	<a href="#">Afipia broomeae</a> <a href="#">ATCC 49717</a>	ATCC 49717	<a href="#">Bacteria</a>
INSDC	<a href="#">AGWX01000003.1:378060-378425</a> (-)	<a href="#">EKS38374.1</a>	polar-differentiation response regulator divK [Afipia broomeae ATCC 49717]	<a href="#">Afipia broomeae</a> <a href="#">ATCC 49717</a>	ATCC 49717	<a href="#">Bacteria</a>
INSDC	<a href="#">KB375282.1:2382517-2382882</a> (-)	<a href="#">EKS38374.1</a>	polar-differentiation response regulator divK [Afipia broomeae ATCC 49717]	<a href="#">Afipia broomeae</a> <a href="#">ATCC 49717</a>	ATCC 49717	<a href="#">Bacteria</a>

**Fig. 2** The identical protein report that can be accessed from Entrez Protein (<https://www.ncbi.nlm.nih.gov/protein>) record. It is a tabular display of all of the protein sequences in GenBank with the identical sequence. It includes links to coding regions, the genome records, the protein product name as it is cited in the genome record and the source organisms that have this protein

Alternatively, one can sequence the whole metagenome or whole transcriptome of the environmental sample and assemble genomes and transcripts from the sample to understand the diversity in the environment. Clustered and assembled 16S rRNA, in addition to assembled metagenome and metatranscriptome sequences, can be submitted to GenBank. It is preferable that the raw reads are submitted to SRA as well. All of the datasets from a single environment, such as mouse gut or seawater, should cite a single BioProject so that the entirety of the project will be detectable to users simultaneously from the NCBI BioProject Database.

RNA sequence analysis or transcriptome analysis is an important mechanism for studying gene expression of a particular organism or tissue type. The transcriptome provides insight into the particular genes that are expressed in specific tissues, disease states or under varied environmental conditions. The transcriptome can also be used to map genes and understand their structure in the genome. Computationally assembled transcript sequences should be deposited into TSA and the underlying sequence reads in SRA so that researchers can use this data to make important scientific discoveries. The structure of TSA records is similar to WGS with a similar accessioning scheme, a master record and sequence overlap contigs.

## 1.8 The GenBank Sequence Record

A GenBank sequence record is most familiarly viewed as a flat file where the data and associated metadata is structured for human readability. The format specifications are as follows.

### 1.8.1 Definition, Accession and Organism

The top of a GenBank flat file is shown in Fig. 3.

The first token of the LOCUS field is the locus name. At present, this locus name is the same as the accession number, but in the past, more descriptive names were used. For instance, HUMHBB is the locus name for the human beta-globin gene in the record with accession number U01317. This practice was abandoned in the mid-1990s when the number of sequences deposited had increased to a point where it was too cumbersome to generate these manually. Following the locus name is the sequence length, the molecule type, the topology (linear or circular), the division

```

LOCUS      KM527068                963 bp   DNA     linear   BCT 13-DEC-2014
DEFINITION Salinispora tropica strain CNS197 Cas1 protein I-E (cas1) gene,
           complete cds.
ACCESSION  KM527068
VERSION   KM527068.1
KEYWORDS   .
SOURCE     Salinispora tropica
  ORGANISM Salinispora tropica
           Bacteria; Actinobacteria; Micromonosporales; Micromonosporaceae;
           Salinispora.

```

**Fig. 3** The top of a GenBank flat file

code and the date of last modification. Records are assigned to divisions based on the source taxonomy or the sequencing strategy that was used (*see* **Note 2** for list of GenBank divisions). The DEFINITION line gives a brief description of the sequence including information about the source organism, gene(s) and molecule information. The ACCESSION is the database-assigned identifier, which has one of the following formats: two letters and six digits (e.g., KM123456) or one letter and five digits for INSDC records; four letters and eight digits for WGS and TSA records (e.g., ABCD01012345); and two letters, an underscore, and six to eight digits for RefSeq records (e.g., NM\_123456). The VERSION line contains the accession number and sequence version.

WGS and TSA projects are assigned a four-letter project ID which serves as the accession prefix for that project. This is followed by a two-digit assembly version, and a six to eight-digit contig id. For example, the *Neurospora crassa* genome is stored in project accession AABX, the first version of the genome assembly is AABX01000000 and AABX01000111 is the 111th contig.

Historically, the KEYWORD field in the GenBank record was used as a summary of the information present in the record. It was a free text field and may have contained gene name, protein name, tissue localization, etc. Information placed on this line was more appropriately placed elsewhere in the sequence record. GenBank strongly discourages the use of keywords to describe attributes of the sequence. Instead, GenBank uses a controlled vocabulary on the KEYWORD line to describe different submission types or divisions. The list of INSDC sanctioned keywords is available at <http://www.insdc.org/documents/methodological-keywords>.

The SOURCE and ORGANISM lines contain the taxonomic name and the taxonomic lineage, respectively, for that organism.

### 1.8.2 Reference Section

The next section of the GenBank flat file contains the bibliographic and submitter information (Fig. 4):

The REFERENCE section contains published and unpublished references. Many published references include a link to a PubMed ID number that allows users to view the abstract of the cited paper in PubMed. The last REFERENCE cited in a record reports the names of submitters of the sequence data and the location where the work was done.

The COMMENT field may have submitter provided comments about the sequence or a table that contains structured metadata for the record. This example has sequencing and assembly methodology but there are other structured comments with isolation source or other phenotypic information. In addition, if the sequence has been updated, then the COMMENT will have a link to the previous version.



```

REFERENCE 1 (bases 1 to 963)
AUTHORS   Wietz,M., Millan-Aguinaga,N. and Jensen,P.R.
TITLE     CRISPR-Cas systems in the marine actinomycete Salinispora: linkages
          with phage defense, microdiversity and biogeography
JOURNAL   BMC Genomics 15 (1), 936 (2014)
PUBMED    25344663
REMARK    Publication Status: Online-Only
REFERENCE 2 (bases 1 to 963)
AUTHORS   Wietz,M., Millan-Aguinaga,N. and Jensen,P.R.
TITLE     Direct Submission
JOURNAL   Submitted (08-SEP-2014) Scripps Institution of Oceanography,
          University of California San Diego, 9500 Gilman Drive, La Jolla, CA
          92093-0204, USA
COMMENT   ##Assembly-Data-START##
          Assembly Method      :: Velvet v. 1.1.04; ALLPATHS-LG v. R41043
          Sequencing Technology :: Illumina
          ##Assembly-Data-END##

```

**Fig. 4** Reference and comment sections of a GenBank flat file

### 1.8.3 Features and Sequence

The FEATURES section contains a source feature, which has additional information about the source of the sequence and the organism from which the DNA was isolated. There are approximately 50 different standard qualifiers that can be used to describe the source. Some examples are /strain, /chromosome, and /host. Following the source feature are annotations that describe the sequence, such as gene, CDS (coding region), mRNA, rRNA, variation, and others. Like the source feature, other features can be further described with feature-specific qualifiers. For example, a mandatory qualifier for the CDS feature is a /translation which contains the protein sequence. Following the Feature section is the nucleotide sequence itself. The specification for the Feature Table can be found at <http://www.insdc.org/documents/feature-table> (see **Note 3**). An example is included as Fig. 5.

### 1.9 Updates and Maintenance of the Database

An INSDC record can be updated by the submitter any time new information is acquired. Updates can include: adding new sequence, correcting existing sequence, adding a publication, or adding new annotation. Information about acceptable update formats can be found at <https://www.ncbi.nlm.nih.gov/genbank/update>. The new, updated record replaces the older one in the database and in the retrieval and analysis tools. However, since GenBank is archival, a copy of the older record is maintained in the database. The sequence in the GenBank record is versioned. For example, KM527068.1 is the first version of the sequence in GenBank record KM527068. When the sequence is modified or updated, the accession version gets incremented. So the accession in the sample record will become KM527068.2 after a sequence update. The base accession number does not change as this is a stable identifier for this record. A COMMENT is added to the updated GenBank flat file that indicates when the sequence is

```

FEATURES             Location/Qualifiers
     source           1..963
                     /organism="Salinispora tropica"
                     /mol_type="genomic DNA"
                     /strain="CNS197"
                     /isolation_source="marine sediment"
                     /db_xref="taxon:168695"
     gene            1..963
                     /gene="cas1"
     CDS             1..963
                     /gene="cas1"
                     /note="CRISPR/Cas system-associated protein Cas1"
                     /codon_start=1
                     /transl_table=11
                     /product="Cas1 protein I-E"
                     /protein_id="AIZ06592.1"
                     /translation="MSTSAQRRLAAPT LAMLPRVADSL SFLYADIVRIVQDDTGVLAQ
VDTTKGTERVYLPTAALSCLLLGPGT SITHHALSTLARHGTTVVCVSGVVR CYAGIT
PTSLTTNWLEKQARCWADDNTRLQVAVRMYEHRFGEAVPEGTTLAQLRGM EGQRMKVL
YRLLAQKYRTGKFRNYPDSKWDTQDPVNLALSAASACLYGVVHAVV LALGCSPALGF
VHSGTQHAFVYDIADLYKAKVTVPLAFAMSTSAQPERDVRRLKCDDFRLLKLMPTIVT
DIQRLLDPDSTPKRQRPVAEVTALWDPEM GAMPSPGVNYSSDPWD"

ORIGIN
   1 atgagcacca ggcgccagcg gcgactcgcc gcaccgaccc tggccatgct gccccgcgtg
  61 gcggaactcg tcagcttcct ctacgccgac atcgttcgga tcgtccaaga cgacaccgga
 121 gtcctcgccc aggtcgacac aaccaagggg accgaacgcg tctacctacc caccgccgcc
 181 ctgagttgcc ttctcctcgg acccggcacc tcgatcacc accacgccct gtccaccctc
 241 gcccgccacg gcaccaccgt cgtctgcgtc ggctccgggtg tcgtccgctg ttacgccggc
 301 atcaccccca cctcctgac caccaactgg ctgaaaaagc aggcccgctg ctggcgccgac
 361 gacaacaccc gcctacaggt agcagtacgg atgtatgagc atcgcttcgg cgaagccgtg
 421 cccgaaggca ccacgctggc ccagcttcgt ggcatggaag gccagcgcac gaaagtgtc
 481 taccgcctgc tggccagaaa atatcgaacc ggcaaatcc gccgcaacta tgaccccgagc

```

**Fig. 5** Features section of a GenBank flat file

updated and provides a link to the older version of the sequence. Annotation changes do not increment the accession version. However, these changes can be detected using the Sequence Revision History, which can be accessed from the Display Setting menu while viewing a sequence record in Entrez (Fig. 6).

## 1.10 Pitfalls of an Archival Primary Sequence Database

### 1.10.1 Bad Annotation and Propagation

Because INSDC is an archival primary sequence database, submitters “own” their annotation and are primarily responsible for ensuring that it is correct. Database staff review records for accuracy and inform submitters of any problems. If entries with poor annotation appear in the database, they may be used by other scientists to validate their own data and possibly to prepare a submission to the database which can lead to the propagation of bad data to subsequent entries in the database. During processing, the GenBank annotation staff checks the sequences and annotation for biological validity. For instance, does the conceptual translation of a coding region match the amino acid sequence provided by the submitter? Does the sequence cluster with the taxonomically related sequences? Does the submitter’s description of the

Revision History ▾ Send ▾

Show difference between **I** and **II** as GenBank/GenPept ▾ Compare

[Sus scrofa mitochondrial NAD+isocitrate dehydrogenase 3 beta \(IDH3B\) gene, complete cds, alternatively spliced, nuclear gene for mitochondrial product](#)

7,457 bp linear DNA

Accession: DQ507858.2 GI: 154818163

Current status: **live**

I	II	Version	GI	Accession	Update Date	Action
<input checked="" type="radio"/>	<input type="radio"/>	2	154818163	DQ507858.2	Feb 27, 2012 08:45 AM	
<input type="radio"/>	<input checked="" type="radio"/>	2	154818163	DQ507858.2	Dec 24, 2009 11:16 AM	
<input type="radio"/>	<input type="radio"/>	2	154818163	DQ507858.2	Jul 24, 2008 10:58 AM	
<input type="radio"/>	<input type="radio"/>	2	154818163	DQ507858.2	Aug 8, 2007 08:43 AM	
<input type="radio"/>	<input type="radio"/>	1	98283611	DQ507858.1	May 22, 2006 12:06 AM	

Accession [DQ507858](#) was first seen at NCBI on May 22, 2006 12:06 AM

**Fig. 6** Sequence Revision History page allows users to retrieve older versions of a sequence record prior to it being updated. Sequence changes are indicated by incrementing the version number. One can view the modifications that were made during an update for two versions of a sequence record by choosing a version in columns I and II and then clicking the Show button

sequence agree with the results of a BLAST similarity search? If problems are detected, the submitter is contacted to correct their submission. Since 2013, if a submitter is unable or unwilling to correct their entry, the sequences are flagged as UNVERIFIED with a comment indicating that the “GenBank staff is unable to verify source organism and sequence and/or annotation provided by the submitter.” These unverified sequences are excluded from the BLAST databases. While this practice attempts to minimize problematic sequences in GenBank, there is still legacy data and submissions processed through automated pipelines that may not be flagged unverified even though they may have problems.

### 1.10.2 Sequence Contamination

The GenBank staff actively removes vector and linker contamination from sequence submissions when it is discovered. GenBank submissions are screened using a specialized BLAST database—UniVec (<https://www.ncbi.nlm.nih.gov/tools/vecscreen/>)—to detect vector contamination. Assembled genomes are also screened for contamination by sequence from other organisms, for instance, detection of stretches of human DNA in a bacterial assembly.

### 1.11 RefSeq

The Reference Sequence (RefSeq) database at NCBI (<https://www.ncbi.nlm.nih.gov/RefSeq/>) provides a comprehensive, integrated, nonredundant well-annotated set of reference sequences including genomic DNA, transcript (RNA), and protein sequences. RefSeq records are derived from INSDC records and can be a synthesis of information from a number of sources. The relationship between the primary data in GenBank and RefSeq is analogous to the relationship between a research paper and a review

article. Each sequence is annotated as accurately as possible with the correct organism name, the correct gene symbol for that organism, and reasonable names for proteins where possible. In some cases, RefSeq records are created in collaboration with authoritative groups who are willing to provide annotations or links to phenotypic or organism-specific resources. For others, the RefSeq staff assembles the best view of the organism that they can put together based on data from INSDC and other public sources. INSDC records are selected based on a number of criteria, validated, corrected, and sometimes re-annotated before inclusion in the RefSeq collection.

---

## 2 Submission of Sequence Data to NCBI Archives

Submission of sequence data to INSDC is required by most journals as a condition of publication. A unified portal for the submission of all sequence related data and metadata is being developed. The Submission Portal (<https://submit.ncbi.nlm.nih.gov>) offers both Web wizards to guide a user through the process of submitting and a programmatic interface for the more advanced user. At the time of publication, wizards are available for the submission of SRA, genomes, transcriptomes, ribosomal RNA sequences and their associated BioProject and BioSample. Submitters create BioSample and BioProject records for SRA and GenBank submissions at the beginning of the process. While additional wizards are developed for other sequence submission types, existing submission tools, like BankIt, and Sequin for GenBank are still supported.

### 2.1 SRA Submissions

SRA processes multiple TeraBytes of sequence data every day. Like GenBank, SRA submissions come from a variety of sources including small labs, core facilities, and genome sequencing centers. Low to mid volume submissions are initiated via NCBI sra submission portal (<https://submit.ncbi.nlm.nih.gov/subs/sra/>) where submitters enter descriptions of the samples and libraries that they intend to upload. Data files may be uploaded from the browser but are typically delivered separately using ftp or Aspera FASP protocol (<https://downloads.asperasoft.com/connect2/>). High volume automated submission pipelines use dedicated upload accounts to deliver data files and bulk metadata submission via programmatically generated xml files.

NCBI works to help labs that do significant amounts of sequencing, covered under the Genomic Data Submission policy, comply with that policy. Submission to SRA, GEO, or dbGaP (<https://www.ncbi.nlm.nih.gov/gap>) or GenBank qualify as acceptable submissions. Resource specific help desks assist individuals with technical issues regarding those submissions (*see Note 4*).

SRA accepts a variety of file formats (*see* **Note 4**). Following best formatting practices will help prevent delays or errors during data loading. Multiple sequencer runs from a library should be split into distinct SRA Runs or submitted as distinct Read Groups in bam format in order to retain batch information. It is best to submit fastq files with the original header formatting. Modification or replacement of the systematic identifiers generated by the instrument may lead to errors or delays in Submission processing. Bam files containing reads aligned to a reference genome are the preferred submission format. Please ensure that submitted bam files have robust header information, including Program (@PG) and Read group (@RG) fields. In addition, alignments to high quality (chromosome level) genomic reference assemblies are strongly recommended. Methods for pre-submission validation of data files can be found in SRA File Format Guide and should ensure successful loading to SRA. Additional detail and the most current information on SRA submissions can be found on SRA documentation and home pages (*see* **Note 4**).

## **2.2 GenBank Submissions**

GenBank processes thousands of new sequence submissions per month from scientists worldwide. GenBank submissions come from a variety of submitters, from small laboratories doing research on a particular gene or genes to genome centers doing high-throughput sequencing. The “small scale” submissions may be a single sequence or sets of related sequences, and with annotation. Submissions include mRNA sequences with coding regions, fragments of genomic DNA with a single gene or multiple genes, a viral or organelle complete genome or ribosomal RNA gene clusters. If part of the nucleotide sequence encodes a protein, a coding sequence (CDS) feature and resulting conceptual translation are annotated in the record. Each nucleotide and protein sequence is assigned an accession number that serves as a permanent identifier for the sequence records.

Submitters can specify that their sets of sequences that span the same gene or region of the genome are biologically related by classifying them as environmental sample, population, or phylogenetic sets. Each sequence within a set is assigned its own accession number and can be viewed independently in Entrez. However, each set is also included in Entrez PopSet (<https://www.ncbi.nlm.nih.gov/popset/>), allowing scientists to view the relationship among the set’s sequences through an alignment.

BankIt (<https://www.ncbi.nlm.nih.gov/WebSub/?tool=genbank>) has a series of wizards to guide a submitter through the submission process to ensure that the appropriate data, like sequence and annotations, metadata, sample isolation information and experimental details, are submitted. Users upload files containing sequences and source information and annotations in tabular format. These web-based tools have quality assurance and

validation checks that report the any problems back to the submitter for resolution.

NCBI maintains two submission tools that a user can download to their own computer to prepare their submission, Sequin (<https://www.ncbi.nlm.nih.gov/Sequin/>) and tbl2asn (<https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2>). Sequin is a stand-alone application that can be used for the annotation and analysis of nucleotide sequences. It also has a series of wizards to guide users through the submission process. tbl2asn is a command-line executable that combines input sequences and tables to generate files appropriate for submission to GenBank (*see Note 5*). The input files necessary for submission include: nucleotide and amino acid sequences in FASTA format (*see Note 6*), tables for source organism information, tables for source information and feature annotation (*see Note 7*). For submitting multiple, related sequences (e.g., those in a phylogenetic or population study), these tools accept the output of many popular multiple sequence-alignment packages, including FASTA + GAP, PHYLIP, and NEXUS. Submitters can use the alignments to propagate annotation from a single record to the other records in the set. Prior to submission to the database, the submitter is encouraged to validate their submission and correct any errors that may exist.

Direct submissions to GenBank are analyzed by the annotation staff. The first level of review occurs before accession numbers are assigned to ensure that the submissions meet the minimal criteria. Sequences should be of a minimal length, sequenced by, or on behalf of, the group submitting the sequence, and they must represent molecules which exist in nature (not a consensus sequence or a mix of genomic or mRNA sequence).

The GenBank annotation staff checks all submissions for:

- (a) Is the sequence of the appropriate length and derived from a single molecule type (not a mix of genomic DNA and mRNA)?
- (b) Biological validity: Does the conceptual translation of a coding region match the amino acid sequence provided by the submitter? Is the source organism name present in NCBI's taxonomy database? Does the sequence cluster with those of closely related organisms? Does the submitter's description of the sequence agree with the results of a BLAST similarity search against other sequences?
- (c) Is the sequence free of vector contamination?
- (d) Is the sequence or accession published? If so, can a PubMed ID be added to the record so that the sequence and publication can be linked in Entrez?
- (e) Formatting and spelling.

If there are problems with the sequence or annotation, the annotator works with the submitter by email to correct the problems.

Completed entries are sent to the submitter for a final review before their release into the public database. Submitters may request that their sequence remain confidential until a specified future release date. Records will be held until that date or when the accession number or the sequence is published, whichever is first.

As the volume of submissions is increasing, the GenBank staff is no longer able to manually review every submission that is received. Classes of sequence data, such as 16S ribosomal RNA from bacteria and archaea are processed and released automatically after undergoing a series of rigorous validation checks.

Microbial genomes are subjected to checks for completeness, the quality of the assembly and the quality of the annotation.

---

### 3 Finding Sequence Data in SRA and GenBank

#### 3.1 Direct Entrez SRA Query

Entrez (<https://www.ncbi.nlm.nih.gov/sra/>) is the primary NCBI Search and Retrieval system, and one of the primary points of access to SRA data. Since Entrez SRA indexing includes descriptive metadata contained in submitted SRA experiments, you can query Entrez using terms found in SRA experiment metadata. Metadata contained in an SRA experiment includes a description of sequencing libraries using controlled vocabularies for library concepts such as strategy, source material, and capture techniques as well as sequencing platform and instrument models (*see Note 8*).

#### 3.2 Simple Entrez Query

To perform a direct Entrez query using SRA metadata search terms, go to the top of the NCBI home page (<https://www.ncbi.nlm.nih.gov/>) and select “SRA” from the drop-down list of available databases and enter a query (for example, “salmonella”) in the search box. The SRA display page of results for your query will include records for each matching study, with links to read/run data, project descriptions, etc.

#### 3.3 Advanced Entrez Query

Go to the top of the NCBI home page (<https://www.ncbi.nlm.nih.gov/>) and select “SRA” from the drop-down list of available databases, then click the “Advanced” link located just below the search bar at the top of the page. The SRA Advanced Search Builder page (<https://www.ncbi.nlm.nih.gov/sra/advanced>) will appear and on this page you can construct a complex SRA query by selecting multiple search terms from a large number of fields and qualifiers such as accession number, author, organism, text word, publication date, and properties (paired-end, RNA, DNA, etc.). See the Advanced Search Builder video tutorial (<https://www.youtube.com/watch?v=...>).

[com/watch?v=dnCRQ1cobdc&feature=relmfu](http://www.ncbi.nlm.nih.gov/Traces/sra/watch?v=dnCRQ1cobdc&feature=relmfu)) for information about how to use existing values in fields and combine them to achieve a desired result.

### 3.4 SRA Home Page Query

You can also search SRA through the coordinated use of the “Browse” and “Search” tabs on the SRA home page (<https://www.ncbi.nlm.nih.gov/Traces/sra/>). The SRA Web interface allows the user to:

- (a) Access any data type stored in SRA independently of any other data type (e.g., accessing read and quality data without the intensity data).
- (b) Access reads and quality scores in parallel.
- (c) Access related data from other NCBI resources that are integrated with SRA.
- (d) Retrieve data based on ancillary information and/or sequence comparisons.
- (e) Retrieve alignments in “vertical slices” (showing underlying layered data) by reference sequence location.
- (f) Review the descriptions of studies and experiments (metadata) independently of experimental data.

### 3.5 The SRA Run Selector

The SRA Run Selector (<https://www.ncbi.nlm.nih.gov/Traces/study/>) can be used to view, filter and sort a metadata table from Entrez search results, or a list of accessions (e.g., SRA BioSample, BioProject, or dbGaP accession) pasted into the field at the top of the page. The Run Selector will dynamically generate a metadata table from library preparation and sample attributes.

The SRA Run Selector displays those attributes common to all selected Runs at the top of the page and displays variable values as columns. The columns are sortable and values can be used to filter the table content using the “Facets” box. Once you have a filtered set of data that you wish to work with, there are three options for saving the results:

1. Click “permalink” from the top of the page and copy the URL for later use or embedding.
2. Click the “RunInfo Table” button to download a tab-delimited text file of all or only selected metadata.
3. Click the “Accession List” button to download a list of Run accessions that can be used with the SRA toolkit to analyze and download the sequence and alignment data.

#### 3.5.1 Putting It All Together with Bioproject

Bioproject can aggregate several data types (e.g., a genome assembly and a transcriptome) by study (e.g., NIH grant). You can access BioProject records by browsing, querying, or downloading in Entrez, or by following a link from another NCBI database.



### 3.6 *BioProject Browsing*

To browse through BioProject content, go to the “By project attributes” (<https://www.ncbi.nlm.nih.gov/bioproject/browse/>) hyperlink from the BioProject home page (<https://www.ncbi.nlm.nih.gov/bioproject/>). You can browse by major organism groups, project type (umbrella projects vs. primary submissions), or project data type. The table includes links to the NCBI Taxonomy database (<https://www.ncbi.nlm.nih.gov/taxonomy/>), where additional information about the organism may be available, and to the BioProject record.

### 3.7 *BioProject Query*

You can perform a search in BioProject like you would in any other Entrez database, namely by searching for an organism name, text word, or BioProject accession (PRJNA31257), or by using the Advanced Search page to build a query restricted by multiple fields. Search results can be filtered by Project Type, Project Attributes, Organism, or Metagenome Groups, or by the presence or absence of associated data in one of the data archives.

Table 2 contains some representative searches:

### 3.8 *BioProject Linking*

You can also find BioProject records by following links from archival databases when the data cites a BioProject accession. You can find links to BioProject in several databases including SRA, Assembly, BioSample, dbVar, Gene, Genome, GEO, and Nucleotide (which includes GenBank and RefSeq nucleotide sequences). Large consortia also LinkOut (<https://www.ncbi.nlm.nih.gov/projects/linkout/>) from BioProject to their resources.

**Table 2**  
**Some example BioProject searches**

Find BioProjects by . . .	Search text example(s)
A species name	Escherichia coli [organism]
Project data type	“metagenome” [Project Data Type]
Project data type and Taxonomic Class	“transcriptome” [Project Data Type] AND Insecta [organism]
Publication	“19643200” [PMID]
Submitter organization, consortium, or center	JGI [Submitter Organization]
Sample scope and material used	“scope environment” [Properties] AND “material transcriptome” [Properties]
A BioProject database Identifier	PRJNA33823 or PRJNA33823 [bioproject] or 33823 [uid] or 33823 [bioproject]

### 3.9 *GenBank*

GenBank and Refseq nucleotide sequences records can be retrieved from Entrez Nucleotide (<https://www.ncbi.nlm.nih.gov/nucleotide/>). EST and GSS records are searched and retrieved from <https://www.ncbi.nlm.nih.gov/est/> and <https://www.ncbi.nlm.nih.gov/gss/> or by choosing the appropriate database from the pull-down menu at the top of most NCBI Web pages. Entrez Protein (<https://www.ncbi.nlm.nih.gov/protein/>) is a collection of protein sequences from a variety of sources, including translations from annotated coding regions in INSDC and RefSeq, UniProt and PDB. The Entrez retrieval system has a network of links that join entries from each of the databases. For example, a GenBank record in Nucleotide can have links to the Taxonomy, PubMed, PubMed Central, Protein, PopSet, BioProject, Gene, and Genome databases. Within a GenBank flat file hyperlinks to Taxonomy, BioProject, BioSample, and PubMed databases are displayed if the links are present. Links to external databases can be made by LinkOut or by cross-references (db\_xrefs) within the entry. By taking advantage of these links, users can make important scientific discoveries. These links are critical to discovering the relationship between a single piece of data and the information available in other databases.

In Entrez, sequence data can be viewed in a number of different formats. The default and most readable format is the GenBank flat file view. The graphical view, which eliminates most of the text and displays just sequence and biological features, is another display option. Other displays of the data, for instance XML, ASN.1, or FASTA formats, are intended to be more computer-readable.

### 3.10 *Genomes*

Genome assembly sequences can be accessed in Entrez from a number of entry points including BioProject, Nucleotide, Genome and Assembly. However, sometimes it is not straightforward to understand which sequences contribute to a complete genome assembly from some of these resources. The Assembly database (<https://www.ncbi.nlm.nih.gov/assembly>) provides users with detailed information and statistics for each genome assembly including links to download the sequences and annotation. One can search for organism, and even strain, then use the links to our FTP site (for either GenBank or RefSeq) provided in the upper right hand corner (Fig. 7). The FTP site contains nucleotide and protein sequences in FASTA format, in addition to annotation data in .gff and GenBank flatfile format.

---

## 4 Downloading the Sequence Data in SRA and GenBank

### 4.1 *The SRA Toolkit*

The SRA Toolkit is a collection of tools and libraries for using the SRA archive file format. SRA utilities have the ability to locate and download data on-demand from NCBI servers, removing the need for a separate download step, and most importantly, downloading

Assembly

Assembly  Search

Advanced Browse by organism Help

Display Settings: ☑ Full Report

**ASM19595v2**

Organism name: *Mycobacterium tuberculosis* H37Rv  
 Intraspecific name: Strain: H37Rv  
 BioSample: SAMEA3138326  
 Submitter: Sanger Institute  
 Date: 2013/02/01  
 Assembly level: Complete Genome  
 Genome representation: full  
 RefSeq category: reference genome  
 GenBank assembly accession: GCA\_000195955.2 (latest)  
 RefSeq assembly accession: GCF\_000195955.2 (latest)  
 RefSeq assembly and GenBank assembly identical: yes

IDs: 538048 [UID] 538028 [GenBank] 538048 [RefSeq]

History (Show revision history)

Global statistics

Total sequence length	4,411,532
Total assembly gap length	0
Total number of chromosomes and plasmids	1

Send to: ☺

See Genome Information for *Mycobacterium tuberculosis*

There are 2097 assemblies for this organism. See more

Access the data

- Download the full sequence report
- Download the statistics report
- GenBank FTP site
- RefSeq FTP site

Assembly Information

- Assembly Help
- Assembly Basics
- NCBI Assembly Data Model

Related Information

- BioProject
- BioSample
- Genome
- Nucleotide INSDC
- Nucleotide RefSeq
- PubMed
- Taxonomy

**Fig. 7** The NCBI Assembly resource can be searched for taxonomic and sequence information pertaining to whole genomes and scaffolds submitted to NCBI. Sequences can be downloaded from the GenBank and RefSeq FTP sites that are accessible from the NCBI Assembly pages

only required data. This feature can reduce the bandwidth, storage, and the time taken to perform tasks that use less than 100 % of the data contained in a run. Utilities are provided for delivering data in commonly used text formats such as fastq and sam. Additional information on using, configuring, and building the toolkit is maintained on the NCBI github repository (*see Note 9*).

## 4.2 Programmatic Interaction with the SRA Toolkit

We have developed a new, domain-specific API for accessing reads, alignments and pileups produced from Next Generation Sequencing called NGS (*see Note 9*). The API itself is independent from any particular back-end implementation, and supports use of multiple back-ends simultaneously. It also provides a library for building new back-end “engines.” The engine for accessing SRA data is contained within the sister repository *ncbi-vdb*.

The API is currently expressed in C++, Java, and Python languages. The design makes it possible to maintain a high degree of similarity between the code in one language and code in another—especially between C++ and Java.

## 4.3 BioProject Download

In addition to the Entrez Web interface and the BioProject browse page, you can download the entire BioProject database and the database .xsd schema from the FTP site: <ftp://ftp.ncbi.nlm.nih.gov/bioproject/>, or use Entrez Programming Utilities (E-utilities) to programmatically access public BioProject records.

#### 4.4 GenBank FTP

There is a bimonthly release of GenBank, which is available from the NCBI FTP site (<ftp://ftp.ncbi.nih.gov/genbank/>). Between releases, there is a daily dump of the sequences loaded into the database to the FTP site (<ftp://ftp.ncbi.nih.gov/genbank/daily-nc/>). Genome assemblies are retrievable by organism or by taxonomic group in a variety of formats on the FTP site (<ftp://ftp.ncbi.nih.gov/genomes/>).

The assembly database has FTP URLs in the upper right hand corner of each page (*see* Fig. 7).

---

## 5 Conclusion

Managing the ever increasing quantity of nucleotide and protein sequence data generated by an evolving array of platforms, methods and institutions requires the ability to accept a variety of source formats, extract the information and transform it to a common archival standard for display and computational access. To achieve this, NCBI collects and validates sequence data and provides easily accessible public websites and application interfaces. Sequence data records are enhanced with links to other NCBI resources such as taxonomy and PubMed. INSDC defines standard elements for the sequence data records to ensure that the information can be submitted to and retrieved from any of the collaborating archives, regardless of the tools used to collect or display the data. The integrity of the sequence data and annotation is confirmed by validation steps both at the submission and the processing stages. GenBank provides methods for updating, correcting, or adding additional information to existing sequence records, before and after they become publicly available. Finally, sequence data is not useful unless it is easily available to researchers in their labs and at their desks so NCBI provides these public users with multiple tools to search, discover, retrieve, and analyze sequence data and educational resources to help users understand it (*see* **Note 10**).

---

## 6 Notes

1. The table on the INSDC homepage contains links to the resources at the partner websites.
  - (a) DDBJ Sequence Read Archive: <http://trace.ddbj.nig.ac.jp/dra/>
  - (b) DDBJ Capillary Reads: <http://trace.ddbj.nig.ac.jp/dta/>
  - (c) DDBJ Annotated Sequences: <http://www.ddbj.nig.ac.jp/>
  - (d) DDBJ Samples: <http://trace.ddbj.nig.ac.jp/biosample/>
  - (e) DDBJ Studies: <http://trace.ddbj.nig.ac.jp/bioproject/>

- (f) ENA Sequence Read Archive, Annotated Sequences, Samples and Studies: <http://www.ebi.ac.uk/ena/submit/data-formats>
  - (g) NCBI Sequence Read Archive: <https://www.ncbi.nlm.nih.gov/sra/>
  - (h) NCBI Capillary Reads: <https://www.ncbi.nlm.nih.gov/Traces/>
  - (i) NCBI Annotated Sequences: <https://www.ncbi.nlm.nih.gov/genbank/>
  - (j) NCBI Samples: <https://www.ncbi.nlm.nih.gov/biosample/>
  - (k) NCBI Studies: <https://www.ncbi.nlm.nih.gov/bioproject/>
2. GenBank records are grouped into 19 divisions; either by taxonomic groupings or by a specific technological approach, such as WGS or TSA. Sequences in the technique-based divisions often have a specific keyword in the record from a controlled list <http://www.insdc.org/documents/methodological-keywords>. Entrez can be specifically queried for sequence records in these divisions. For example, if one wanted to retrieve actin sequences for all non-primate mammalian species, one could search for *Actin AND "gbdiv MAM" [prop]*
- The GenBank divisions are listed in Table 3.
3. The DDBJ/EMBL/GenBank Feature Table: Definition, which can be found at [http://insdc.org/feature\\_table.html](http://insdc.org/feature_table.html), lists all allowable features and qualifiers for a DDBJ/EMBL/GenBank record. This document gives information about the format and conventions, as well as examples, for the usage of features in the sequence record. Value formats for qualifiers are indicated in this document. Qualifiers may be
- (a) free text
  - (b) controlled vocabulary or enumerated values
  - (c) sequence
- Other syntax related to the flat file is described in this document. The document also contains reference lists for the following controlled vocabularies:
- (a) Nucleotide base codes (IUPAC)
  - (b) Modified base abbreviations
  - (c) Amino acid abbreviations
  - (d) Modified and unusual Amino Acids
  - (e) Genetic Code Tables
4. Help emails and documentation, as well as fact sheets, are available from the following links:

**Table 3**  
**Traditional taxonomic GenBank divisions**

<b>Code</b>	<b>Description</b>
BCT	Bacterial sequences
PRI	Primate sequences
MAM	Other mammalian sequences
VRT	Other vertebrate sequences
INV	Invertebrate sequences
PLN	Plant, fungal, and algal sequences
VRL	Viral sequences
PHG	Bacteriophage sequences
SYN	Synthetic and chimeric sequences
UNA	Unannotated sequences, including some WGS sequences obtained via environmental sampling methods
<i>Nontraditional GenBank divisions</i>	
PAT	Patent sequences
EST	EST division sequences, or expressed sequence tags, are short single pass reads of transcribed sequence
STS	STS division sequences include anonymous STSs based on genomic sequence as well as gene-based STSs derived from the 3' ends of genes. STS records usually include primer sequences, annotations and PCR reaction conditions
GSS	GSS records are predominantly single reads from bacterial artificial chromosomes ("BAC-ends") used in a variety of clone-based genome sequencing projects
ENV	The ENV division of GenBank, for non-WGS sequences obtained via environmental sampling methods in which the source organism is unknown
HTG	The HTG division of GenBank contains unfinished large-scale genomic records that are in transition to a finished state. These records are designated as Phase 0–3 depending on the quality of the data. Upon reaching Phase 3, the finished state, HTG records are moved into the appropriate taxonomic division of GenBank
HTC	The HTC division of GenBank accommodates high-throughput cDNA sequences. HTCs are of draft quality but may contain 5' UTRs and 3' UTRs, partial coding regions, and introns
CON	Large records that are assembled from smaller records, such as eukaryotic chromosomal sequences or WGS scaffolds, are represented in the GenBank "CON" division. CON records contain sets of assembly instructions to allow the transparent display and download of the full record using tools such as NCBI's Entrez
TSA	Transcriptome shotgun data are transcript sequences assembled from sequences deposited in the NCBI Trace Archive, the Sequence Read Archive (SRA), and the EST division of GenBank

SRA Submission Quick Start Guide: <https://www.ncbi.nlm.nih.gov/books/NBK47529/>

SRA File Format Guide: <https://www.ncbi.nlm.nih.gov/books/NBK242622/>

GEO Submission Guide: <https://www.ncbi.nlm.nih.gov/geo/info/submission.html>

GenBank Submissions Handbook: <https://www.ncbi.nlm.nih.gov/books/NBK51157/>

SRA homepage: <https://trace.ncbi.nlm.nih.gov/Traces/sra/>

dbGaP homepage: <https://www.ncbi.nlm.nih.gov/gap/>

GEO homepage: <http://www.ncbi.nlm.nih.gov/geo/>

GenBank homepage: <https://www.ncbi.nlm.nih.gov/genbank/>

Questions about data archives and submissions can be sent to: [submit-help@ncbi.nlm.nih.gov](mailto:submit-help@ncbi.nlm.nih.gov); [gb-admin@ncbi.nlm.nih.gov](mailto:gb-admin@ncbi.nlm.nih.gov)

General NCBI questions can be sent to the NCBI helpdesk: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov)

5. GenBank Submission tools are available on NCBI's FTP site (Sequin: <ftp://ftp.ncbi.nih.gov/sequin/>, tbl2asn: [ftp://ftp.ncbi.nih.gov/toolbox/ncbi\\_tools/converters/by\\_program/tbl2asn/](ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/converters/by_program/tbl2asn/)). Additional information about these programs can be found on the NCBI website (sequin: <https://www.ncbi.nlm.nih.gov/Sequin/> tbl2asn: <http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2>). Both of these submission utilities contain validation software that will check for problems associated with your submission. The validator can be accessed in Sequin from the Search->Validate menu item or in tbl2asn using `-v` in the command line.
6. When preparing a submission by Sequin or tbl2asn, information about the sequence can be incorporated into the Definition Line of the FASTA formatted sequence. FASTA format is simply the raw sequence preceded by a definition line. The definition line begins with a `>` sign and is followed immediately by the sequence identifier and a title. Information can be embedded into the title which Sequin and tbl2asn use to construct a submission. Specifically, you can enter organism and strain or clone information in the nucleotide definition line and gene and protein information in the protein definition line using name-value pairs surrounded by square brackets. Example:
 

```
>myID [organism=Drosophila melanogaster] [strain=Oregon R] [clone=abc1].
```

7. For both submission and updates, the submitter should prepare tabular files with source metadata and feature information in the following formats. These files can easily be incorporated into the GenBank records using any of the submission tools.

Source information (i.e., strain, cultivar, country, specimen\_voucher) is to be provided in a two-column tab-delimited table, for example:

Sequence id.	Strain	Country
AYxxxxxx	82	Spain
AYxxxxxy	ABC	France

Nucleotide sequence should be submitted in FASTA format:

```
>AYxxxxxx
```

```
cggtaataatggaccttggacccggcaagcggagagac
```

```
>AYxxxxxy
```

```
ggaccttggacccggcaagcggagagaccggtaataat
```

Feature annotation should be submitted as a tab-delimited five-column feature table.

Column 1: Start location of feature

Column 2: Stop location of feature

Column 3: Feature key

Column 4: Qualifier key

Example:

>Feature Sc_16				
1	7000	REFERENCE		
		PubMed		8849441
<1	1050	gene		
		gene		ATH1
<1	1009	CDS		
		product		acid trehalase
		product		Ath1p
		codon_start		2
<1	1050	mRNA		

In the future, GFF3 format will be supported as well.

8. Table 4.



**Table 4**  
**SRA experimental enumeration values and definitions**

<b>Strategy</b>	<b>Sequencing strategy used in the experiment</b>
WGA	Random sequencing of the whole genome following non-PCR amplification
WGS	Random sequencing of the whole genome
WXS	Random sequencing of exonic regions selected from the genome
RNA-Seq	Random sequencing of whole transcriptome
miRNA-Seq	Random sequencing of small miRNAs
WCS	Random sequencing of a whole chromosome or other replicon isolated from a genome
CLONE	Genomic clone based (hierarchical) sequencing
POOLCLONE	Shotgun of pooled clones (usually BACs and Fosmids)
AMPLICON	Sequencing of overlapping or distinct PCR or RT-PCR products
CLONEEND	Clone end (5', 3', or both) sequencing
FINISHING	Sequencing intended to finish (close) gaps in existing coverage
ChIP-Seq	Direct sequencing of chromatin immunoprecipitates
MNase-Seq	Direct sequencing following MNase digestion
DNase-Hypersensitivity	Sequencing of hypersensitive sites, or segments of open chromatin that are more readily cleaved by DNaseI
Bisulfite-Seq	Sequencing following treatment of DNA with bisulfite to convert cytosine residues to uracil depending on methylation status
Tn-Seq	Sequencing from transposon insertion sites
MRE-Seq	Methylation-sensitive restriction enzyme sequencing strategy
MeDIP-Seq	Methylated DNA immunoprecipitation sequencing strategy
MBD-Seq	Direct sequencing of methylated fractions sequencing strategy
OTHER	Library strategy not listed (please include additional info in the "design description")
<b>Source</b>	<b>Type of genetic source material sequenced</b>
GENOMIC	Genomic DNA (includes PCR products from genomic DNA)
TRANSCRIPTOMIC	Transcription products or non genomic DNA (EST, cDNA, RT-PCR, screened libraries)
METAGENOMIC	Mixed material from metagenome
METATRANSCRIPTOMIC	Transcription products from community targets
SYNTHETIC	Synthetic DNA
VIRAL RNA	Viral RNA

(continued)

**Table 4**  
**(continued)**

<b>Strategy</b>	<b>Sequencing strategy used in the experiment</b>
OTHER	Other, unspecified, or unknown library source material (please include additional info in the “design description”)
<b>Selection</b>	<b>Method of selection or enrichment used in the Experiment</b>
RANDOM	Random shearing or other “shotgun” method
PCR	Source material was selected by designed primers
RANDOM PCR	Source material was selected by randomly generated primers
RT-PCR	Source material was selected by reverse transcription PCR
HMPR	Hypo-methylated partial restriction digest
MDA	Multiple displacement amplification
MSLL	Methylation spanning linking library
cDNA	Complementary DNA
ChIP	Chromatin immunoprecipitation
MNase	Micrococcal nuclease (MNase) digestion
DNase	Deoxyribonuclease (MNase) digestion
Hybrid Selection	Selection by hybridization in array or solution
Reduced Representation	Reproducible genomic subsets, often generated by restriction fragment size selection, containing a manageable number of loci to facilitate re-sampling
Restriction Digest	DNA fractionation using restriction enzymes
5-methylcytidine antibody	Selection of methylated DNA fragments using an antibody raised against 5-methylcytosine or 5-methylcytidine (m5C)
MBD2 protein methyl-CpG binding domain	Enrichment by methyl-CpG binding domain
CAGE	Cap-analysis gene expression
RACE	Rapid amplification of cDNA ends
Padlock probes capture method	Circularized oligonucleotide probes
other	Other library enrichment, screening, or selection process (please include additional info in the “design description”)

## 9. SRA software tools documentation and downloads

NCBI distributes a variety of software tools from our github repository:

<https://github.com/ncbi>

Common SRA utilities are in the sra-tools repo:

<https://github.com/ncbi/sra-tools/wiki>

APIs for software development and examples are in the ngs repo:

<https://github.com/ncbi/ngs/wiki>

<https://github.com/ncbi/ngs-tools>

10. NCBI presents educational resources in several formats. First, we provide video tutorials as live webinars—both “full length” (30–60 min) and in “NCBI minute” (5–10 min format). We make these products approximately monthly and weekly. Announcements of upcoming webinars can be accessed at <https://www.ncbi.nlm.nih.gov/home/coursesandwebinars.shtml>. The recorded webinars are made public 2-3 weeks after the conclusion of the webinar and are available on our YouTube channel <https://www.youtube.com/user/NCBINLM>. Short, standalone videos are also available on our YouTube channel. Readers are encouraged to subscribe to get updated information about our video resources. Additionally, we have fact sheets about each resource that are available at <http://www.ncbi.nlm.nih.gov/home/documentation.shtml> and more extensive documentation about each resource available in the NCBI Handbook at <https://www.ncbi.nlm.nih.gov/books/NBK143764/>.

---

## Acknowledgement

This research was supported by the Intramural Research Program of the NIH, NLM, NCBI.

## References

1. Karsch-Mizrachi I, Nakamura Y, Cochrane G (2012) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res* 40 (Database issue):D33–D37
2. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2015) GenBank. *Nucleic Acids Res* 43(Database issue):D30–D35
3. Silvester N, Alako B, Amid C, Cerdeno-Tarraga A, Cleland I, Gibson R et al (2015) Content discovery and retrieval services at the European Nucleotide Archive. *Nucleic Acids Res* 43(Database issue):D23–D29
4. Kodama Y, Mashima J, Kosuge T, Katayama T, Fujisawa T, Kaminuma E et al (2015) The DDBJ Japanese Genotype-phenotype Archive for genetic and phenotypic human data. *Nucleic Acids Res* 43(Database issue):D18–D22
5. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A et al (2011) Prospective genomic characterization of the

- German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One* 6(7): e22751
6. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M et al (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 41(Database issue):D991–D995
  7. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269 (5223):496–512
  8. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Ciufu S, Li W (2013) Prokaryotic genome annotation pipeline. In: *The NCBI Handbook*, 2nd edn. [Internet], Bethesda, MD. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK174280/>