

# Public Databases in Health and Life Sciences

*“the potential to translate big data into big discovery”*



**Academic year 2023-2024**

# What is a database?

- A collection of related data elements
  - *Tables*
    - *columns (fields)*
    - *rows (records)*
  - *Documents*
    - *Key -> Value*
- Records retrieved using a query language
- Database technology is well established

## Relational Databases

### Rows (records)

- actual data
- whereas *fields* describe what data is stored, the *rows* of a table are where the actual data is stored

### Columns (fields)

- attributes of tables, e.g. for *citation* table, *title, journal, volume, author*

# How is information organized in databases?

## Accession numbers and Identifiers

An **Identifier** is essentially a name of a database, table, or table column.

- As the creator of the database, you are free to identify these objects as you please.
- The identifier can change (based on the curator)

Each record (row in the table) has a **unique identifier**, alone or combined with another column is unique for that table.

**The primary key (accession number or accession code).**

- The primary key should not change.
- Data is indexed according to this primary key
- The unique identifier serves to identify the data stored in this record across all the tables in the database (relational database).
- Usually, a string of letters and digits that uniquely identifies an entry in its database.
  - The accession number for TPIS\_CHICK in Uniprot/Swissprot is P00940

# How is information organized in databases?

## Accession numbers and Identifiers

Some DBs have both Identifiers and accession as unique for each entry in the DB.

In these cases, the main difference is that Identifiers are Human readable and accession are just "random" codes.

Example Pfam\_ID: MlaC; Pfam\_AC: PF05494

In some cases Identifiers are mutable (but remain unique) while accessions are not in UniprotKB accession P0AAP1 Use to have ADRA\_ECOLI as Id (Name) but now it is DGCC\_ECOLI

The fact that they are called "immutable" does not mean that can not be considered obsolete and removed from the database.

Sometimes accessions can have a version number, which means that something has changed, but whatever they represent remains

Examples:

In Pfam the current accession with version for the MlaC family is PF05494.15

In ncbi refseq nucleotide, the current accession for *Neisseria meningitidis* MC58 complete genome is NC\_003112.2

# What is a flat file database?

- Sequential collection of entries, stored in a set of text files
- Flat-File databases can be represented as holding all of their data in one table only (two-dimensional table)
- Files written in plain text, standard defined format. Examples:
  - Each line is a record. Fields are separated by delimiters: tabs, commas...
  - Each file is a record. Fields expressed as key->value (eg: json db)
- Searching issues!

Accession	Source	Gene	Mol Type
AF068625.2	Mus musculus	dnmt3a	mRNA
HD654844.1	Homo sapiens	hba1	mRNA
AD836734.3	Escherichia coli	recA	DNA
BD823723.5	Homo sapiens	hpo3	DNA
TF7823562.1	VIH	p17	cDNA
AS9832656.3	Homo sapiens	hbb	DNA
AF6723523.1	Danio rerio	ccf2	mRNA

# What is a relational database?

- A relational database contains multiple tables and defines the relationships between them.
- Virtually all use SQL (Structured Query Language) as a language for querying and maintaining

invoice_id	customer	product	price	quantity	total
1	Elmer	buckshot	\$2,00	2	\$4,00
2	Wiley	Acme snow machine	\$5,00	1	\$5,00
3	Elmer	shotgun	\$25,00	1	\$25,00
4	Bugs	carrots	\$0,50	20	\$10,00

customer_table		
name	address	notes
Elmer	Looney Tunes Dr.	likes hunting and opera
Wiley	Southwest desert	big mail order customer
Bugs	Rabbit Hole	likes to cross dress

product_table		
product	price	notes
carrots	\$ 0.50	
shotgun	\$ 25.00	oddly flexible
buckshot	\$ 2.00	
Acme snow machine	\$ 5.00	high defect rate

database scheme

# A common way of storing biological data in a structured manner is to use a relational database

GeneBank Flat File Format

GeneBank Flat File Format

**GeneBank Flat File Format**

LOCUS AF068625 200 bp mRNA linear ROD 06-DEC-1999

DEFINITION Mus musculus DNA cytosine-5 methyltransferase 3A (Dnmt3a) mRNA, complete cds.

ACCESSION AF068625 REGION: 1..200

VERSION AF068625.2 GI:6449467

KEYWORDS .

SOURCE Mus musculus (house mouse)

ORGANISM Mus musculus Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea; Muridae; Murinae; Mus.

REFERENCE1 (bases 1 to 200) , AUTHORS, TITLE, JOURNAL, etc.

REFERENCE2 (bases 1 to 200) , AUTHORS, TITLE, JOURNAL, etc.

REMARK Sequence update by submitter

COMMENT On Nov 18, 1999 this sequence version replaced gi:3327977.

FEATURES Location/Qualifiers

source 1..200 /organism="Mus musculus" /mol\_type="mRNA" /db\_xref="taxon:10090" /chromosome="12" /map="4.0 cM"

gene 1..>200 /gene="Dnmt3a"

ORIGIN 1 gaattccggc ctgctgccgg gccgccgac ccgccgggcc acacggcaga gccgcctgaa 61  
 gccacgcgt gaggctgcac tttccgagg gctgacatc agggctatg ttaagtctt 121 agctctgct  
 tacaaagacc acggcaatc ctctctgaa gccctcgag cccacagcg 181 ccctcgagc cccagcctgc//

tab1

Accession	Source	Gene	Mol Type
AF068625.2	Mus musculus	dnmt3a	mRNA
HD654844.1	Homo sapiens	hba1	mRNA
AD836734.3	Escherichia coli	recA	DNA
BD823723.5	Homo sapiens	hpo3	DNA
TF7823562.1	VIH	p17	cDNA
AS9832656.3	Homo sapiens	hbb	DNA
AF6723523.1	Danio rerio	ccf2	mRNA

tab2

Species	TaxID	Synonym
Homo sapiens	9606	Human
Mus musculus	10090	Mouse
Danio rerio	7955	Zebra fish
Escherichia coli	562	E. coli

## Essential aspects of primary and secondary databases.

	Primary database	Secondary database
<b>Synonyms</b>	Archival database	Curated database; knowledgebase
<b>Source of data</b>	Direct submission of experimentally-derived data from researchers (database staff organize but don't add additional information)	Results of analysis, literature research and interpretation, often of data in primary databases
	Once given a database accession number, the data in primary databases are never changed: they form part of the scientific record.	Continuously updated <u>Biocuration</u>

EMBL-EBI Train online  
Bioinformatics for the terrified

<https://www.ebi.ac.uk/training/online/courses/bioinformatics-terrified/what-makes-a-good-bioinformatics-database/primary-and-secondary-databases/>



## Definition and aims of biocuration

Biocuration involves the interpretation and integration of information relevant to biology into a database or resource that enables integration of the scientific literature as well as large data sets.

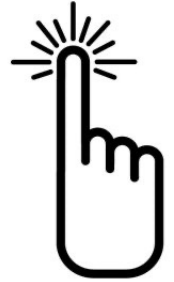
Primary goals of biocuration.

- Accurate and comprehensive representation of biological knowledge
- Easy access to this data for working scientists and a basis for computational analysis

# How to access the data in public databases ?

## Human Web interface (web based, small scale)

- Free text search
- Common mode of search are keywords with modifiers or identifiers
- Cross-references link the information of different databases
- You do not see the underlying database structure
- Output defined by host/provider



“click your way”

## Web services and Programmatic data access

- Application Programmers Interface (API)
- To approach database programmatically



Programming Utilities Web Service

## Download the data: File Transfer Protocol (FTP), rsync, http

- Flat files (script based, bulk data download)

## Fielded searching using any of the indexed fields (advanced searches)

Entering the phrase with a [field descriptions]:

robotic surgery [title]

Miller MJ [author]

“protein domain” [TI]

human [Organism]

insulin [Protein Name]

Combining fielded searching with booleans

enzymes [TI] NOT Gonzales P [AU]

human [Organism] AND insulin [Protein Name]

# Search for Field Descriptions are different in each Database

## NCBI

NC\_0000\*[Accession]  
Human[Organism]  
horse[taxonomy]  
neoplasms[MeSHTerms]  
prolactin[Protein Name]  
APOE[gene]  
srcdb\_refseq[Properties]  
2010/06[Publication Date]  
110:500[Sequence Length]  
gene\_symbol[sym]  
1.1.1.53[ecno]  
gbdiv\_est[PROP]  
: : :  
etc

## UNIPROT

accession:p62988  
organism:human  
taxonomy:40674  
keyword:neoplasms  
name:"prion protein"  
gene:HSPC233  
database:(type:pfam)  
created:[20121001 TO \*]  
length:[100 TO 500]  
go:0015629  
ec:3.2.1.23  
reviewed:yes  
: : :  
etc

[http://www.ncbi.nlm.nih.gov/entrez/query/static/help/Summary\\_Matrices.html#Search\\_Fields\\_and\\_Qualifiers](http://www.ncbi.nlm.nih.gov/entrez/query/static/help/Summary_Matrices.html#Search_Fields_and_Qualifiers)

<https://www.ncbi.nlm.nih.gov/books/NBK49540/>

<https://www.uniprot.org/help/query-fields>

# The NCBI is a comprehensive website for biologists (database of databases (of databases) )

<http://www.ncbi.nlm.nih.gov/gquery/>

- The National Center for Biotechnology Information (NCBI)
- Created in 1988 as a part of the National Library of Medicine at NIH
- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information
- Over 30 databases (primary, secondary, specialized, meta-databases, etc.)



# The NCBI home page

<http://www.ncbi.nlm.nih.gov/>

NCBI Resources ▾ How To ▾ masterbioinf My NCBI

NCBI National Center for Biotechnology Information

All Databases ▾  Search

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

## Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News](#) | [Blog](#)

### Submit

Deposit data or manuscripts into NCBI databases



### Download

Transfer NCBI data to your computer



### Learn

Find help documents, attend a class or watch a tutorial



### Develop

Use NCBI APIs and code libraries to build applications



### Analyze

Identify an NCBI tool for your data analysis task



### Research

Explore NCBI research and collaborative projects



## Popular Resources

- [PubMed](#)
- [Bookshelf](#)
- [PubMed Central](#)
- [PubMed Health](#)
- [BLAST](#)
- [Nucleotide](#)
- [Genome](#)
- [SNP](#)
- [Gene](#)
- [Protein](#)
- [PubChem](#)

## NCBI Announcements

- [Genome Workbench 2.10 now available](#)
- [Genome Workbench 2.10 includes reworked BLAST tool and new functionalities in Tree View. For t](#)
- [Sequence Viewer 3.11 now available](#)
- [Sequence Viewer 3.11, now available, contains a number of new features, improvements and bug fixes. incl](#)

# NCBI hosts over 30 databases



Resources

How To



National Center for  
Biotechnology Information

**NCBI Home**

**Resource List (A-Z)**

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

All Databases

- MeSH
- NCBI Web Site
- NLM Catalog
- Nucleotide**
- OMIM
- PMC
- PopSet
- Probe
- Protein
- Protein Clusters
- PubChem BioAssay
- PubChem Compound
- PubChem Substance
- PubMed
- PubMed Health
- SNP
- Sparcle
- SRA
- Structure
- Taxonomy

## to NCBI

Center for Biotechnology Information advances science and  
providing access to biomedical and genomic information.

[NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News](#) | [Blog](#)

### Submit

Submit data or  
sequences into  
NCBI databases

### Download

Transfer NCBI data  
to your computer



### Learn

Find help  
documents, attend  
a class or watch a  
tutorial



# How to access the NCBI data ?

## ***Entrez*: An Integrated Database Search and Retrieval System**

### **Human Web interface (web based, small scale)**

- Free text search
- List of identifiers (Batch *Entrez*)

### **Web service (Programmatic data access)**

- [Entrez Utilities Web Service \(NCBI\): The E-utilities](#)
  - [Entrez Direct: E-utilities on the Unix Command Line](#)

## **Searching sequence databases using a sequence query**

- BLAST

## **File Transfer Protocol (FTP)**

- Flat files (script based, bulk data download)



# ***Entrez: An Integrated Database Search and Retrieval System***

- Access all NCBI resources (Database Integration)
- *Entrez Databases*
  - All Molecular Database entries are organized by organism (Taxonomy Database).
  - Each record is assigned a UID “unique integer identifier” for internal tracking
  - Each record is indexed by data fields: [author], [title], [organism], and many others
  - Each record is given a Document Summary (DocSum).
  - Each record is manually or computationally assigned links to biologically related UIDs in and across databases.



## Batch Entrez

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Books  
Database Nucleotide File: Choose File no file selected Retrieve

### Batch Entrez

Given a file of Entrez accession numbers or other identifiers, Batch Entrez downloads the corresponding records.

#### Instructions

1. Start with a local file containing a list of accession numbers or identifiers
2. Select the database corresponding to the type of accession numbers or identifiers in your input file
3. Use the **Browse** or **Choose File...** button to select the input file
4. Press the **Retrieve** button to see a list of document summaries
5. Select a format in which to display the data for viewing, and/or saving
6. Select 'Send to file' to save the file.

#### Tips

- To download entire genome records, check the NCBI FTP site, instead of using Batch Entrez.
- Some lists of record identifiers can be tens of thousands of lines long, so Batch Entrez may not retrieve all records from one list. Split the list of identifiers into smaller files using a file splitting software or a file split command at the command prompt in UNIX or LINUX systems.
- When loading large numbers of genome records, put several thousand record identifiers per file, one per line, left-adjusted.
- Please note that Batch Entrez will check for duplicate identifiers when reporting results from a list that you have imported.
- When retrieving a list of Nucleotide accessions, you must select the specific component database from which the accessions or GIs were saved. For Nucleotide, choose either the CoreNucleotide, the EST or the GSS selection from the database menu. If you have a mixed list of nucleotide accessions or UIDs, you will need to run the Batch Entrez search three times. Select the database from the pull-down menu, CoreNucleotide, EST, and GSS separately.
- In all cases, be certain to select the database that corresponds to the identifiers you are uploading. For example, if you have saved a list of protein accession numbers, be sure to select the Protein database.

# Access Entrez through programs or scripts

Entrez Utilities Web Service (NCBI) **The E-utilities** <http://www.ncbi.nlm.nih.gov/books/NBK25500/>

• Entrez Programming Utilities are tools that provide access to Entrez data outside of the regular web query interface. You can access them by constructing the right URL and retrieving results. This is the base:

<https://eutils.ncbi.nlm.nih.gov/entrez/eutils/>

And then you add the desired tool with the appropriate parameters.

Some tools:

- **ESearch: Searches and retrieves primary IDs and retains results in the user's environment.**
- **EFetch: Retrieves records from one or more primary IDs or from the user's environment.**
- **Also: EGQuery, EInfo, ELink, ESpell, Esummary**

Example: Get the PubMed IDs (PMIDs) for articles about breast cancer published in Science in 2008:

[https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&term=science\[journal\] AND breast cancer AND 2008\[pdat\]&retmode=json](https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&term=science[journal] AND breast cancer AND 2008[pdat]&retmode=json)

• There is a set of programs implementing these functions. They allow to access the same services from the UNIX-like Command Line interface : Entrez Direct. (<https://www.ncbi.nlm.nih.gov/books/NBK179288/>)

Example: Download all protein sequences in fasta format from *Stenotrophomonas maltophilia* with the word “lactamase” in their title:

```
~]$ esearch -db protein -query "txid40324[porgn] AND lactamase[TI]"|efetch -format fasta
```

# NCBI:Molecular Sequence Databases

## Sequence Databases (Primary)

Nucleotide (GenBank)

PopSet

SRA, GSS

Protein

## Marker Databases

Single Nucleotide Polymorphisms (SNP's, dbSNP)

Sequence Tagged Sites (STS's, dbSTS)

Expressed Sequence Tags (EST's, dbEST)

NCBI Resources  How To

Sign

Nucleotide

Nucleotide

Advanced



## Nucleotide

The Nucleotide database is a collection of sequences from several sources including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

<https://www.ncbi.nlm.nih.gov/nucleotide/>

# NCBI: Derivative Databases

Nucleotide derived

Example: RefSeq, GENE

Protein-derived

Example: CDD

Structure-derived

Example: Structure

Human curated, compilation and correction of data

Example: RefSeq

Computationally Derived

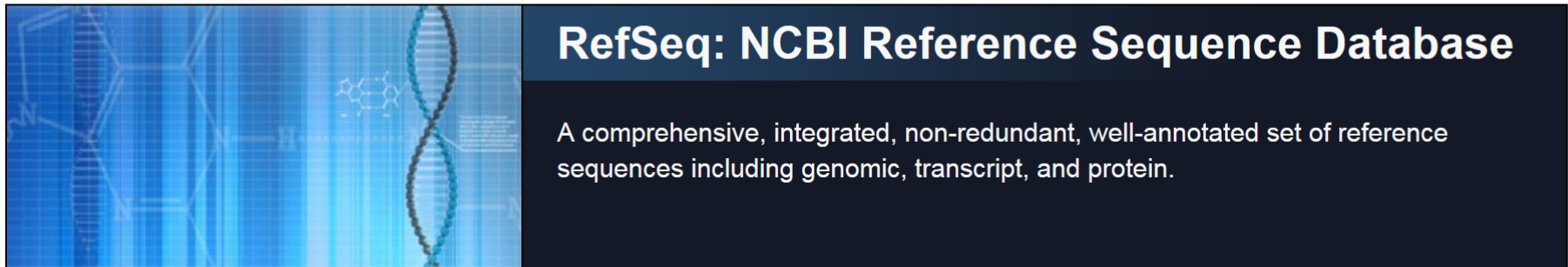
Example: UniGene

Combinations

Example: NCBI Genome Assembly

NCBI Resources How To Sign in to NCBI

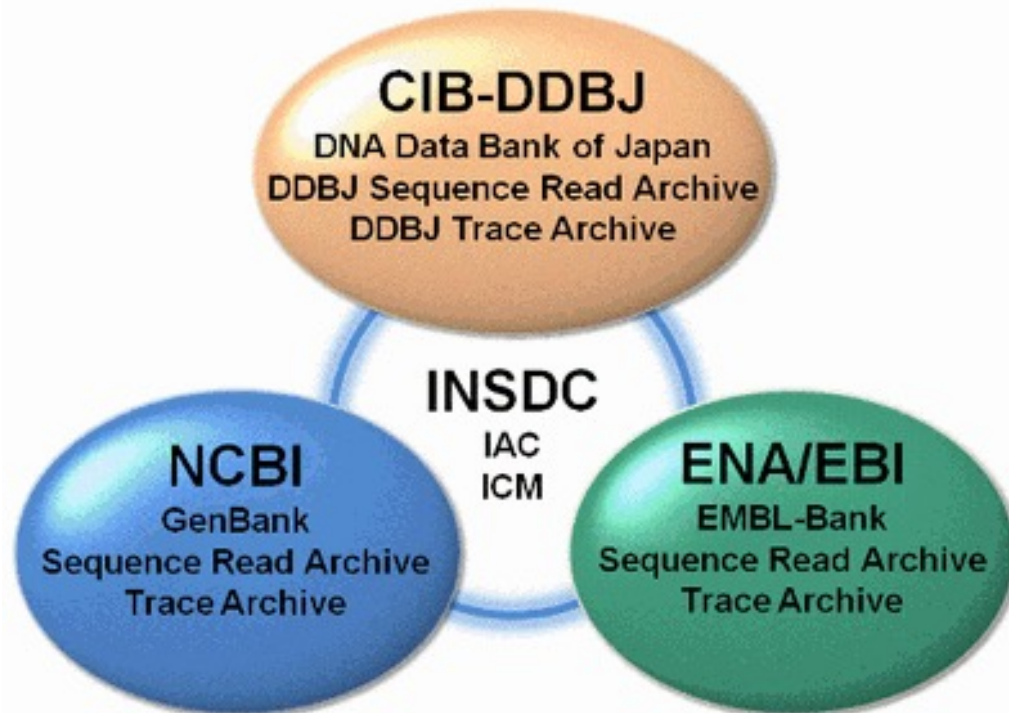
RefSeq RefSeq Search



**RefSeq: NCBI Reference Sequence Database**

A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.

## Primary (archival) nucleotide sequence databases



- The three databases are synchronized on a daily basis
- The accession numbers are consistent.
- There are no legal restriction in the usage of these databases. However, there are some patented sequences in the database

## Sequence submission to nucleotide databases

- Direct submissions from the authors:
  - Free submissions.
  - Authors can annotate the sequences.
  - Only minor staff supervision and quality assurance checks.
- Submissions through the Internet:
  - Web forms / Web services.
  - Email.
- Sequences shared/exchanged between the 3 centers on a daily basis:
  - The sequence content of the banks is identical.

# What is GenBank ?

- GenBank is the NIH (NCBI) genetic sequence database
- Nucleotide only sequence database (Beware: GenPept)
  - Example: <https://www.ncbi.nlm.nih.gov/nuccore/AM743169.1>
- Archival in nature
  - Historical
  - Reflective of submitter point of view (subjective)
  - Redundant
- GenBank Data
  - Direct submissions (traditional records)
  - Batch submissions (EST, GSS, STS)
  - ftp accounts (genome data)
- Three collaborating databases (all data from INSDC)
  - GenBank
  - DNA Database of Japan (DDBJ)
  - European Molecular Biology Laboratory (EMBL) Database



# The GenBank at the National Center for Biotechnology Information (NCBI)

<http://www.ncbi.nlm.nih.gov/nucleotide>

The image shows a screenshot of the NCBI website. At the top, there is a navigation bar with the NCBI logo and the text "National Center for Biotechnology Information". Below this, there is a main navigation menu with several categories: "NCBI Home", "Resource List (A-Z)", "All Resources", "Chemicals & Bioassays", "Data & Software", "DNA & RNA", "Domains & Structures", "Genes & Expression", "Genetics & Medicine", and "Genomes & Maps". A dropdown menu is open under "All Resources", showing a list of databases: "All Databases", "MeSH", "NCBI Web Site", "NLM Catalog", "Nucleotide" (highlighted in blue), "OMIM", "PMC", "PopSet", "Probe", "Protein", "Protein Clusters", "PubChem BioAssay", "PubChem Compound", "PubChem Substance", "PubMed", "PubMed Health", "SNP", "Sparcle", "SRA", and "Structure".

to NCBI

Center for Biotechnology Information advances science and providing access to biomedical and genomic information.

[NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News](#) | [Blog](#)

**Download**

Transfer NCBI data to your computer

**Learn**

Find help documents, attend a class or watch a tutorial

# The source databases for NCBI nucleotide and protein sequences

## Nucleotide (NCBI)

GenBank

EMBL

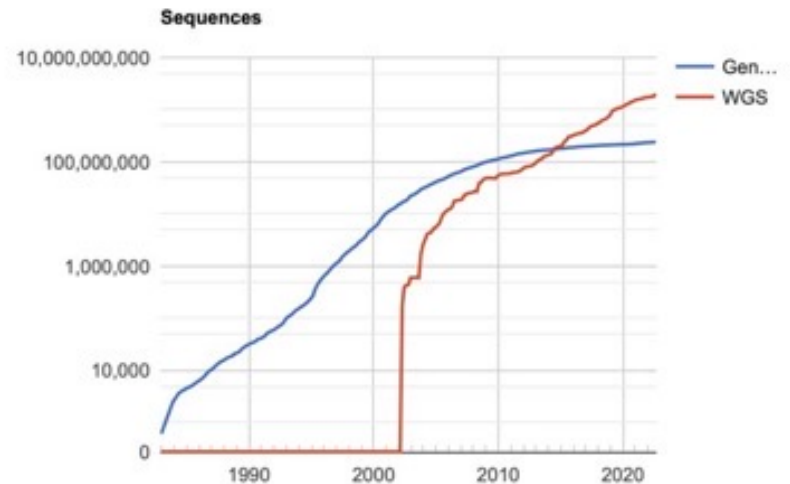
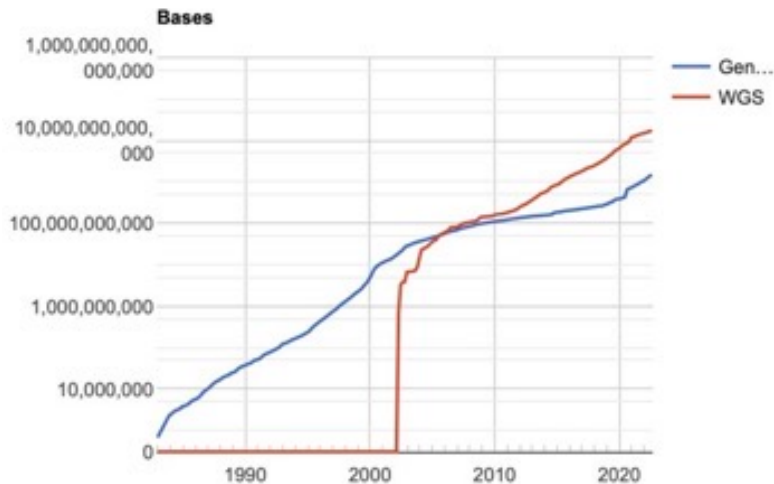
DDBJ

genbank[Filter]

EMBL[Filter]

ddbj[Filter]

# Growth of GenBank (1982-2022)



See the current GenBank release notes for up-to-date information.

<http://ftp.ncbi.nih.gov/genbank/gbrel.txt>

And GenBank statistics page:

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

# GenBank vs Nucleotide

**Nucleotide**: a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

**GenBank**: An archival database of primary nucleotide sequences that were directly sequenced by the submitter.

**RefSeq**: A curated, non-redundant database that includes genomic DNA, transcript (RNA), and protein products, for major organisms. The sequence data are derived from GenBank primary data, and the annotation is computational, from published literature, or from domain experts. All RefSeq ids have a **prefix**

**TPA**: A database designed to capture experimental or inferential results that support submitter-provided annotation for sequence data that the submitter did not directly determine but derived from GenBank primary data.

**PDB**: Repeat with me: "PDB is not a protein database" (**2UVG**, **4E1U**)

# What information is requested from authors when submitting a sequence?

	<u>Search Field Descriptions</u>
• The sequence	
• Automatic assignment of a unique <b><u>accession number</u></b>	[accession]
• Source	
• The organism	[organism]
• Type of molecule	[properties]
• References	
• Authors	[author]
• Publication	[journal]
• Coding Sequence (CDS) features	
• Gene	[gene name]
• Coordinates (automatic translation - protein sequence)	
• Genetic elements	[feature key]
• Protein	[protein name]
• Comments	[text word]
• etc., etc.,.....	

The diagram includes two callout boxes. The first, labeled 'Primary key', points to the 'accession number' field. The second, labeled 'Identifiers', points to a bracketed group of search field descriptions including '[gene name]', '[feature key]', '[protein name]', and '[text word]'.

# Advanced Searches

## Search Field Descriptions and Tags at NCBI

<b>[Accession]</b>	<b>[ACCN]</b>	The accession number assigned by NCBI.
<b>[All Fields]</b>	<b>[ALL]</b>	All terms from all search fields in the database.
<b>[Author]</b>	<b>[AU]</b>	All authors from all references in the records.
<b>[EC/RN Number]</b>	<b>[ECNO]</b>	Enzyme Commission (EC) number for an enzyme activity.
<b>[Feature Key]</b>	<b>[FKEY]</b>	Biological features listed in the Feature Table of the sequence records.
<b>[Gene Name]</b>	<b>[GENE]</b>	Gene names annotated on database records.
<b>[Issue]</b>	<b>[ISS]</b>	The issue number of the journals cited on sequence records
<b>[Journal]</b>	<b>[JOUR]</b>	The name of the journals cited on sequence records.
<b>[Keyword]</b>	<b>[KYWD]</b>	Keywords applied by submitter
<b>[Modification Date]</b>	<b>[MDAT]</b>	The date of most recent modification of a sequence record.
<b>[Organism]</b>	<b>[ORGN]</b>	The scientific and common names for the complete taxonomy of organisms
<b>[Properties]</b>	<b>[PROP]</b>	Molecular type, source database, and other properties of the sequence
<b>[Protein Name]</b>	<b>[PROT]</b>	The names of protein products as annotated on sequence records.
<b>[Publication Date]</b>	<b>[PDAT]</b>	The date that records were made public in Entrez.
<b>[Sequence Length]</b>	<b>[SLEN]</b>	The total length of the sequence
<b>[Text Word]</b>	<b>[WORD]</b>	Text on a sequence record that is not indexed in other fields.
<b>[Title]</b>	<b>[TI]</b>	Words and phrases found in the title of the sequence record.
<b>ETC.....</b>	<b>ETC.....</b>	ETC.....

<https://www.ncbi.nlm.nih.gov/books/NBK49540/>

# GenBank Flat Files (.gbff)

```
LOCUS       HUMCYPB                               linear   PRI 27-APR-1993
DEFINITION  Human cytochrome P-450
ACCESSION   M17398 J03472
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1 (bases 1 to 1868)
AUTHORS     Okino,S.T., Quattrocchi,L.C., Pendurthi,U.R., McBride,O.W. and
            Tukey,R.H.
TITLE       Characterization of multiple human cytochrome P-450 1 cDNAs. The
            chromosomal localization of the gene and evidence for alternate RNA
            splicing
JOURNAL     J. Biol. Chem. 262 (33), 16072-16079 (1987)
FEATURES             Location/Qualifiers
     source          1..1868
                    /organism="Homo sapiens"
                    /mol_type="mRNA"
                    /db_xref="taxon:9606"
     CDS             2..1183
                    /note="cytochrome P-450 1"
                    /codon_start=1
                    /protein_id="AAA35740.1"
                    /db_xref="GI:181328"
                    /translation="MEPFVVVLVLCLEFMLLFSLWRQSCRRRLKPPGPTPLPIIGNMLQ
IDVKDICKSFTNFSKVYGPVFTVYFGMNP I VV F H G Y E A V K E A L I D N G E E F S G R G N S P I
SQRITKGLGIIS
ASPCDPTFILG
FPLLIDCFPGTH
NQKSEFNIENLVGTVADLVAGTETTTSTTLRYGLLLLLLKHPEVTAKVQEEIDHVIGRH
RSPCMQDRSHMPYTDVAVVHEIQRYSDLVPTGVPHAVTTDTKFRNYLIPKSF DNKIMLA
A"
ORIGIN        337 bp
              1 aatggaacct tttgtggtcc tgggtcgtg tctctctttt atgcttctct tttcactctg
              61 gagacagagc tgtaggagaa ggaagctccc tctctgcccc actcctcttc ctattattgg
              121 aaatatgcta cagatagatg ttaaggacat ctgcaaatct ttcaccaatt tctcaaaagt
              181 ctatggtcct gtgttcaccg tgtattttgg catgaatccc atagtgggtg ttcatggata
              241 tgaggcagtg aaggaagccc tgattgataa tggagaggag tttctcggaa gaggcaattc
              301 cccaatatct caaagaatta ctaaaggact tggaatcatt tccagcaatg gaaagagatg
              361 gaaggagatc cggcgtttct ccctcacaaa cttgcggaat tttgggatgg ggaagaggag
              421 cattgaggac cgtgttcaag aggaagctca ctgccttggt gaggagtga gaaaaaccaa
              481 ggcttcacc cgtgatccca ctttcatcct ggcctgtgct cctcgaatg tgatctgctc
              541 ogttgttttc cagaaacgat ttgattataa agatcagaat tttctcacc tgat.....
```

Identifiers

Identification codes

Source and References

Feature Table

DNA sequence

type of molecule

accession

organism

division

journal

CDS coordinates

protein name

protein sequence

# NCBI: Understanding the mess (Databases)

1)

Go to:

<https://www.ncbi.nlm.nih.gov/guide/all/>

click on All Databases

find Nucleotide Database

find GenBank Database

find Reference Sequence (RefSeq)

find Protein Database

find GenPept (Where is it?)

2)

Count the previously reported Databases (58)

Go to: <https://www.ncbi.nlm.nih.gov>

open the Dropdown menu for the databases and count them (36)

Go to: <https://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi>

and count them (40)



# NCBI: Understanding the mess (Data and Databases)

Go to:

<https://www.ncbi.nlm.nih.gov/nuccore>

Search this:

Neisseria meningitidis MC58 complete[TI]

Open in new tab:

- Neisseria meningitidis MC58, complete genome
- Neisseria meningitidis MC58, complete sequence

Focus on AE002098.2

LOCUS, ACCESSION and VERSION

DBLINK

KEYWORDS (to compare with NC\_003112.2)

SOURCE

REFERENCE

COMENTS

FEATURES

## GenBank vs RefSeq

Compare AE002098.2 with NC\_003112.2

GenBank vs RefSeq

Search (Ctrl+F) NMB1736

Open [https://www.ncbi.nlm.nih.gov/nuccore/NG\\_011877](https://www.ncbi.nlm.nih.gov/nuccore/NG_011877)

Read COMENTS

# BLAST databases:

Nucl. Database	Content
<b>nt (default)</b>	All GenBank + EMBL + DDBJ + PDB sequences, excluding sequences from PAT, EST, STS, GSS, WGS, TSA and phase 0, 1 or 2 HTGS sequences. Non-redundant, records with identical sequences collapsed into a single entry.
rRNA/ITS data-bases	A collection of four databases: a 16S Microbial rRNA sequences from <a href="#">NCBI's Targeted Loci Projects</a> , an 18S and a 26S RNA rRNA data-bases for fungi, plus an ITS database for fungi.
refseq_rna	Curated (NM_, NR_) plus predicted (XM_, XR_) sequences from NCBI Reference Sequence Project.
refseq_representative_genomes	NCBI RefSeq Reference and Representative genomes across broad taxonomy groups including eukaryotes, bacteria, archaea, viruses and viroids. These genomes are among the best quality genomes available with minimum redundancy - one genome per species for eukaryotes and diverse isolates for the same species for others.
Refseq genome	This database contains NCBI RefSeq genomes across all taxonomy groups. It contains only the top-level sequences, i.e. the longest sequences representing any given part of the genomes, to reduce redundancy.
wgs	Assemblies of Whole Genome Shotgun sequences.
est	Database of GenBank + EMBL + DDBJ sequences from EST division
Human G+T	The genomic sequences plus curated and predicted RNAs from the current build of the human genome.
Mouse G+T	The genomic sequences plus curated and predicted RNAs from the current build of the mouse genome.
est	Database of GenBank + EMBL + DDBJ sequences from EST division
TSA	Transcriptome Shotgun Assemblies, assembled from RNA-seq SRA data
SRA	Nextgen sequences from NCBI's Sequence Read Archive (SRA), limit to specific subset required.
HTGS	Unfinished High Throughput Genomic Sequences; Sequences: phases 0, 1 and 2.
pat	Nucleotides from the Patent division of GenBank.
pdb	Nucleotide sequences from the 3-dimensional structure records from Protein Data Bank.
refseq_genomic	All genomic sequences from NCBI Reference Sequence Project, highly redundant.
Prot. Database	Content
<b>nr (default)</b>	Non-redundant GenBank CDS translations + RefSeq + PDB + SwissProt + PIR + PRF, excluding those in PAT, TSA, and env_nr.
refseq_protein	Protein sequences from NCBI Reference Sequence project.
Landmark	The landmark database includes <a href="#">proteomes from representative genomes</a> spanning a wide taxonomic range
swissprot	Last major release of the UniProtKB/SWISS-PROT protein sequence database (no incremental updates).
pat	Proteins from the Patent division of GenBank.
pdb	Protein sequences from the 3-dimensional structure records from the Protein Data Bank.
env_nr	Protein sequences translated from the CDS annotation of metagenomic nucleotide sequences.
tsa_nr	Protein sequences translated from CDSs annotated on transcriptome shotgun assemblies.

[https://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo\\_BLASTGuide.pdf](https://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf)

