

# Public Databases in Health and Life Sciences

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



# Public Databases in Health and Life Sciences



## Topic 2. Biological raw data in public sequence databanks.

- Repositories for primary nucleotide sequences.
- Nucleic acid submission and accession: The GenBank and the ENA.
- Tools for online databases search and retrieval (part II).
- Models for sequence-related information
- Nucleic acids sequence formats.

## Practical session #2

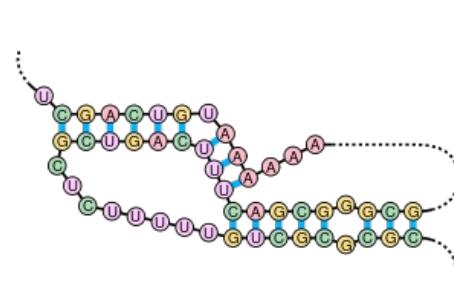
# Primary structures vs. Primary databases

Nucleotides

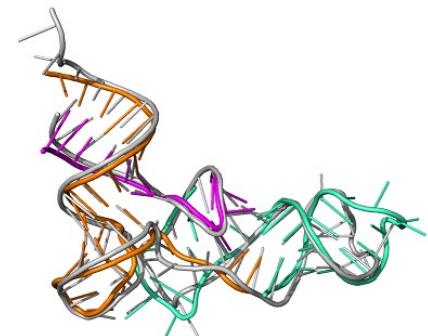
Primary structure

```
>XN_003444  
AACCACCTTAGA  
CAGATAGACAG  
ATAGAGACAAA  
AGCTT
```

Secondary structure

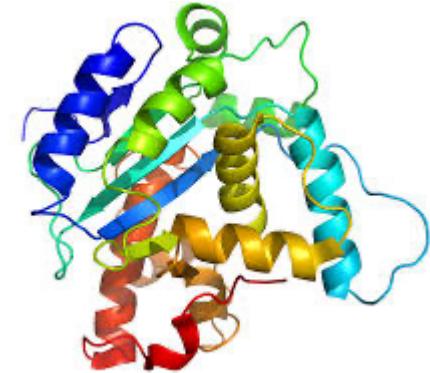
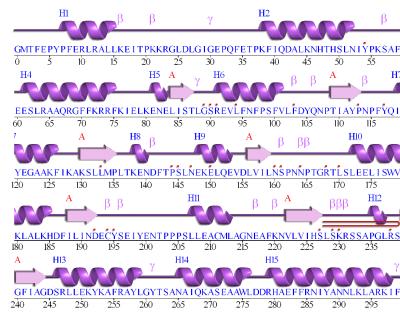


Tertiary structure



Proteins

```
>XP_003444  
MATGWYALAAN  
CNAAAFKLAGEC  
NVTQALAANFMI  
LIANLAAY
```



# Primary (archival) nucleotide sequence databases

The International Nucleotide Sequence Database (INSDC) (<http://www.insdc.org/>) consists of the following databases.

DDBJ (DNA Data Bank of Japan)

EMBL/ENA (European Bioinformatics Institute)

GenBank (National Center for Biotechnology Information)

The screenshot shows the INSDC homepage with a blue header featuring the INSDC logo and the text "International Nucleotide Sequence Database Collaboration". Below the header is a navigation bar with links for "ABOUT INSDC", "POLICY", "ADVISORS", and "DOCUMENTS". On the left side, there are logos for NCBI, DDBJ, and ENA. The main content area contains two sections: "International Nucleotide Sequence Database Collaboration" which discusses the collaboration between DDBJ, ENA, and GenBank, and "How to submit data" which provides instructions for data submission to the databases.

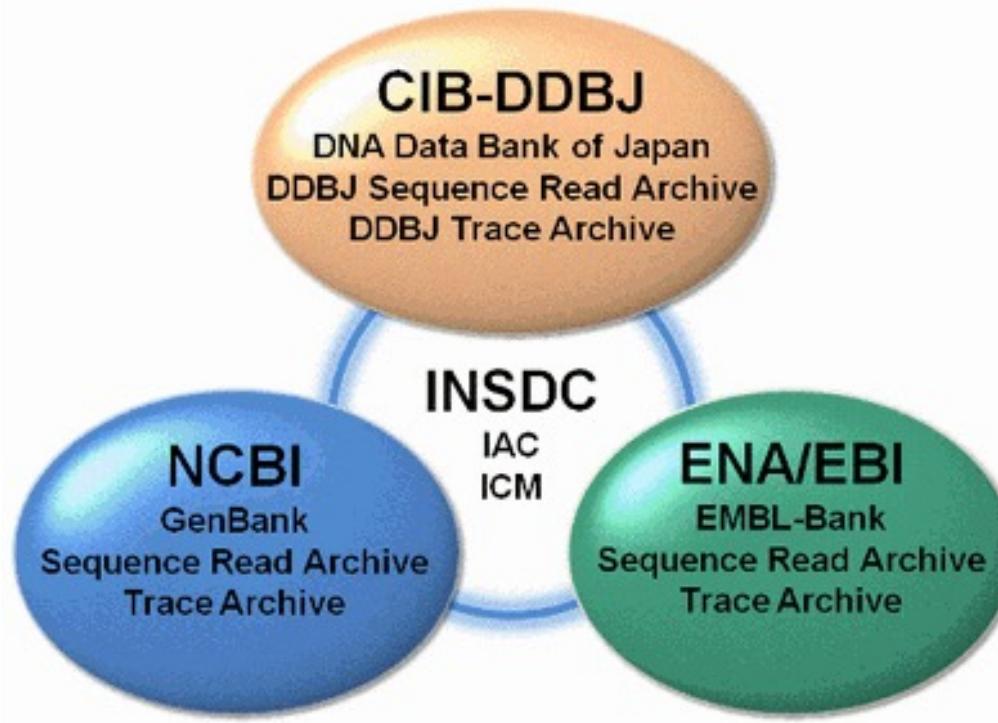
**International Nucleotide Sequence Database Collaboration**

- The International Nucleotide Sequence Databases (INSD) have been developed and maintained collaboratively between [DDBJ](#), [ENA](#), and [GenBank](#) for over 18 years.
- The INSDC advisory board, the [International Advisory Committee](#), is made up of members of each of the databases' advisory bodies. At their most recent meeting, members of this committee unanimously endorsed and reaffirmed the existing data-sharing policy of the three databases that make up the INSDC, which is stated below.
- Individuals submitting data to the international sequence databases should be aware of [INSDC policy](#).

**How to submit data**

- For full details of how to submit data to the databases, please select a collaborating partner.
- [DDBJ](#), [ENA](#), [GenBank](#)
- The INSDC Feature Table Definition Document is available [here](#).

# Primary (archival) nucleotide sequence databases



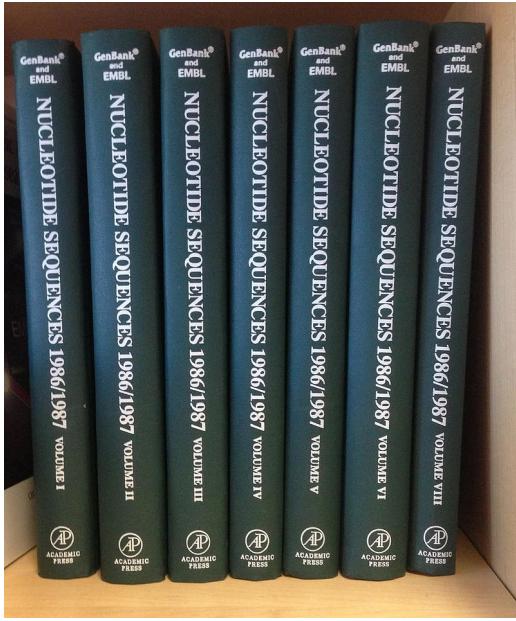
- The three databases are synchronized on a daily basis
- The accession numbers are consistent.
- There are no legal restriction in the usage of these databases. However, there are some patented sequences in the database

## **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/nuccore>

NCBI Resources How To

NCBI 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

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

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

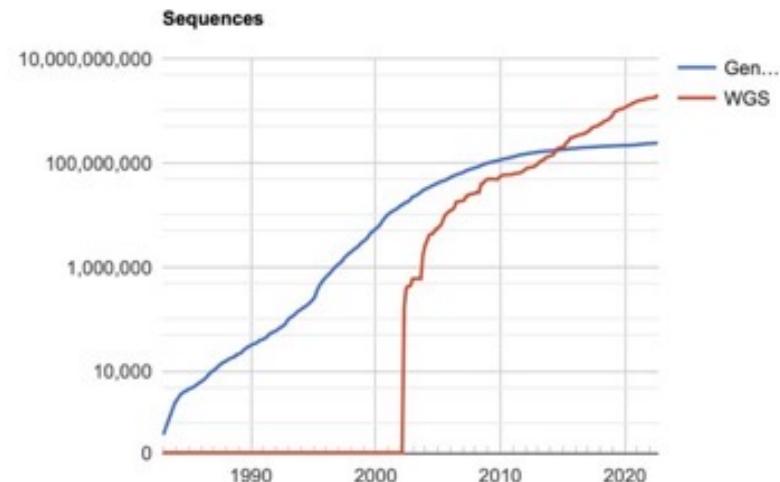
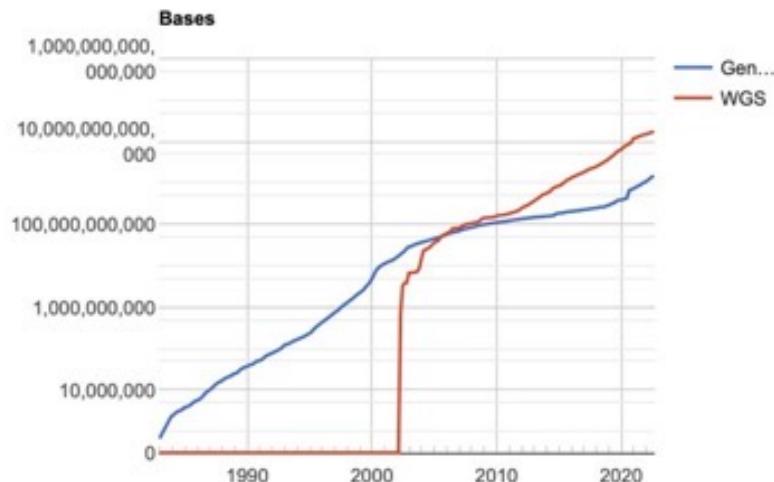
DDBJ

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/>

# There are several ways to search and retrieve data from GenBank.

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

- Free text search



- Advanced searches (search tags and regular expressions)
- List of identifiers (Batch *Entrez*)

## Search sequence databases using a sequence query

- BLAST (Basic Local Alignment Search Tool)

## Web service (Programmatic data access)

- *Entrez* Utilities Web Service (NCBI): The E-utilities

## File Transfer Protocol (FTP)

- Flat files (bulk data download)

<ftp://ftp.ncbi.nlm.nih.gov/genbank>

<http://ftp.ncbi.nlm.nih.gov/genbank>

# **GenBank organization: taxonomical divisions**

PRI - primate sequences

ROD - rodent sequences

MAM - other mammalian sequences

VRT - other vertebrate sequences

INV - invertebrate sequences

PLN - plant, fungal, and algal sequences

BCT - bacterial sequences

VRL - viral sequences

PHG - bacteriophage sequences

SYN - synthetic sequences

UNA - unannotated sequences

ENV - environmental sampling sequences

**gbdiv\_pri[PROP]**

## **Others**

PAT - patent sequences

# **GenBank organization: high-throughput or functional divisions and other sub-databases**

EST – Expressed Sequence Tag (1<sup>st</sup> pass single read cDNA)

gbdv\_htg[PROP]

GSS – Genome Survey Sequences (1<sup>st</sup> pass single read gDNA)

HTG – High Throughput Genomic (incomplete sequences of genomic clones)

STS – Sequence Tagged Site (PCR-based mapping)

HTC - High-throughput cDNA

TSA - Transcriptome shotgun data (transcriptome shotgun assembly projects)

WGS - Whole genome shotgun data (whole genome shotgun projects)

"wgs"[Properties]

<ftp://ftp.ncbi.nih.gov/genbank/wgs/>

## **Others**

CON - Contains no sequence data, but rather instructions for the assembly of reads.

# 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.

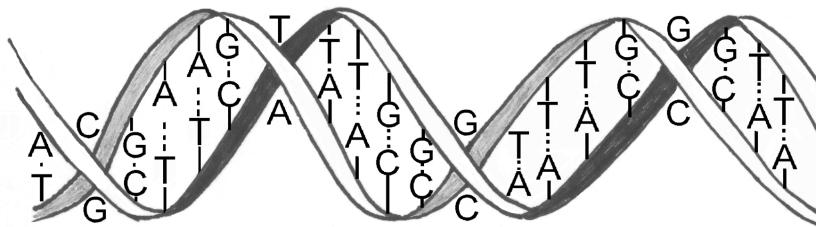
**RefSeq:** A curated, non-redundant database that includes genomic DNA, transcript (RNA), and protein products, for major organisms. The sequence data are derived from GenBank primary data, and the annotation is computational, from published literature, or from domain experts. All RefSeq ids have a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

# The DNA sequence as a line of text

agcttctttggcaaacggtt



Orientation to read →

5' agcttctttggcaaacggtt 3'

Direct

3' tcgaagaaaccgtttgccaa 5'

Complement

← Orientation to read

5' aaccgtttgccaaagaagct 3'

Reverse complement

It's all relative

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

- The sequence
  - Automatic assignment of a unique **accession number** [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.,.....

Identifiers

Primary key

# Advanced Searches

## Search Field Descriptions and Tags at NCBI

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

# Nucleotide Advanced Search Builder

History deleted.

(Cytochrome-b[Gene Name]) AND human[Primary Organism]

[Edit](#)

## Builder

Gene Name:  [Show index list](#)

AND: Primary Organism:  [Show index list](#)

AND: [Accession](#) [All Fields](#)

AND: **Assembly** [Show index list](#)

[Search](#)

## History

There is

You are h

## GETTING

NCBI Ed

NCBI He

NCBI Ha

Training &

Submit D

Sequence Length

Strain

Substance Name

Text Word

Title

Volume

## Nucleotide Database

### RECENTS

Accessions & Bioassays

Software

Protein

Publication Date

SeqID String

Expression

Chemistry & Medicine

Geographic & Maps

### POPULAR

PubMed

Bookshelf

PubMed Central

BLAST

Nucleotide

Genome

SNP

Gene

Protein

PubChem

### FEATURED

Genetic Testing Registry

GenBank

Reference Sequences

Gene Expression Omnibus

Genome Data Viewer

Human Genome

Mouse Genome

Influenza Virus

Primer-BLAST

Sequence Read Archive

<https://www.ncbi.nlm.nih.gov/books/NBK25500/>



## Entrez Programming Utilities Help [Internet].

[Show details](#)

[Contents](#)

[Search this book](#)

## The E-utilities In-Depth: Parameters, Syntax and More

Eric Sayers, PhD.

[Author Information](#)

Created: May 29, 2009; Last Update: October 24, 2018.

Estimated reading time: 18 minutes

### Introduction

Go to: [\(dropdown\)](#)

This chapter serves as a reference for all supported parameters for the E-utilities, along with accepted values and usage guidelines. This information is provided for each E-utility in sections below, and parameters and/or values specific to particular databases are discussed within each section. Most E-utilities have a set of parameters that are required for any call, in addition to several additional optional parameters that extend the tool's functionality. These two sets of parameters are discussed separately in each section.

### General Usage Guidelines

Go to: [\(dropdown\)](#)

Please see [Chapter 2](#) for a detailed discussion of E-utility usage policy. The following two parameters should be included in all E-utility requests.

#### tool

Name of application making the E-utility call. Value must be a string with no internal spaces.

#### email

E-mail address of the E-utility user. Value must be a string with no internal spaces, and should be a valid e-mail address.

#### api\_key – enforced in December 2018

In December 2018, NCBI will begin enforcing the practice of using an API key for sites that post more than 3 requests per second. Please see [Chapter 2](#) for a full discussion of this new policy.

### E-utilities DTDs

Go to: [\(dropdown\)](#)

With the exception of EFetch, the E-utilities each generate a single XML output format that conforms to a DTD specific for that utility. Links to these DTDs are provided in the XML headers of the E-utility returns.

ESummary version 2.0 produces unique XML DocSums for each Entrez database, and as such each Entrez database has a unique DTD for version 2.0 DocSums. Links to these DTDs are provided in the version 2.0 XML.

EFetch produces output in a variety of formats, some of which are XML. Most of these XML formats also conform to DTDs or schema specific to the relevant Entrez database. Please follow the appropriate link below for the PubMed DTD:

- [PubMed DTD June 2018 – current PubMed DTD](#)
- [PubMed DTD January 2019 – forthcoming DTD](#)

GenBank as any other database is a body of information stored in two dimensions (rows and columns) = Tables

## Fields (identifiers) in columns

Entries (sequences) in rows

Accession (ID)	Sequence	Mol type	Organism	Authors	Genes	Feat

How all this information is stored?

How is it shown to users?

### Human mitochondrial cytochrome b gene, partial cds

GenBank: M28016 1

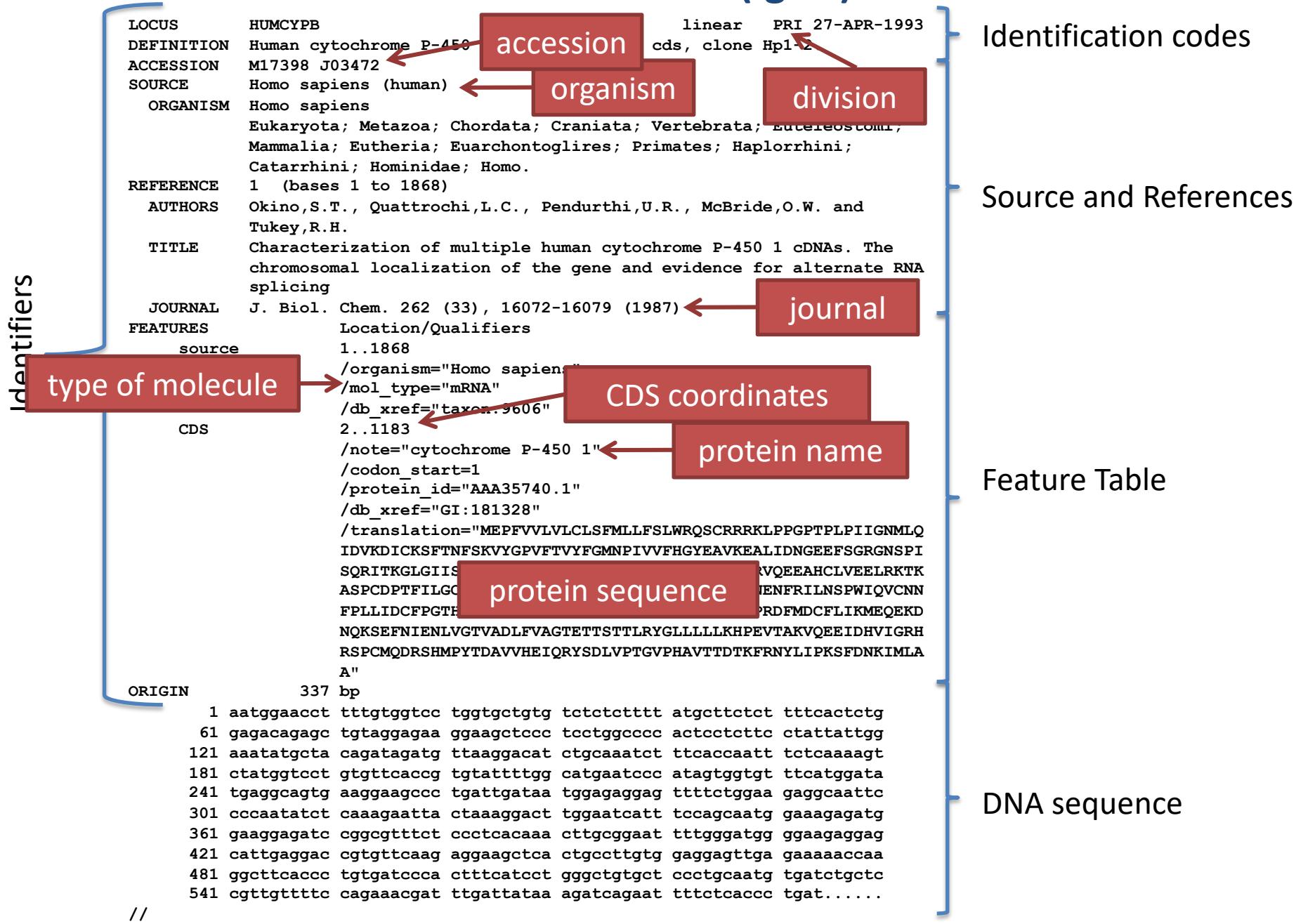
GenBank FASTA Graphics

```

Seg-entry ::= set {
    level 1,
    class nuc-prot,
    descr {
        source {
            genome mitochondrial,
            org {
                taxname "Homo sapiens",
                common "human",
                db {
                    {
                        db "taxon",
                        tag id 9606
                    }
                },
                orgname {
                    name binomial {
                        genus "Homo",
                        species "sapiens"
                    },
                    lineage "Eukaryota; Metazoa; Chordata; Craniata; Vertebrates; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini; Hominidae; Homo",
                    gcode 1,
                    mgcode 2,
                    div "PRI"
                }
            },
            subtype {
                {
                    subtype cell-line,
                    name "Maja"
                },
                {
                    subtype cell-type,
                    name "lymphoblastoid"
                },
                {
                    subtype tissue-lib,
                    name "Provided by Drs John Trowsdale and Janet Lee"
                }
            }
        },
        pub {
            pub {
                gen {
                    serial-number 1
                },
                muid 86064879,
                article {
                    title {
                        name "Serendipitous cloning of a mitochondrial cDNA polymorphism."
                    },
                    authors {
                        names std {
                            {
                                name name {
                                    last "Spurr",
                                    initials "N.K."
                                }
                            },
                            {
                                name name {
                                    last "Bodmer",
                                    initials "W.F."
                                }
                            }
                        }
                    }
                },
                from journal {
                    title {
                        iso-jta "Mol. Biol. Med.",
                        ml-jta "Mol Biol Med",
                        issn "0735-1313",
                        name "Molecular biology & medicine."
                    }
                }
            }
        }
    }
}

```

# GenBank Flat Files (.gbff)



## FASTA file

>M17398.1 Human cytochrome P-450 1 mRNA, complete cds, clone Hp1-2  
AATGGAACCTTTGTGGTCCTGGTGCCTGCTCTCTTTATGCTTCTCTTCACTCTGGAGACAGAGC  
TGTAGGAGAAGGAAGCTCCCTCCTGGCCCCACTCCTCTTCTATTATTGAAATATGCTACAGATAGATG  
TTAAGGACATCTGCAAATCTTACCAATTCTCAAAAGTCTATGGCCTGTGTTACCGGTATTTGG  
CATGAATCCCATACTGGTGTTCATGGATATGAGGCAGTGAAAGGAAGCCCTGATTGATAATGGAGAGGAG  
TTTCTGGAAGAGGCAATTCCCCAATATCTCAAAGAATTACTAAAGGACTTGAATCATTCCAGCAATG  
GAAAGAGATGGAAGGGAGATCCGGCTTCTCCCTCACAAACTTGCAGAATTTGGATGGGAAGAGGAG  
CATTGAGGACCGTGTCAAGAGGAAGCTCACTGCCCTGTGGAGGAGTTGAGAAAAACCAAGGCTTCACCC  
TGTGATCCCACCTTCATCCTGGCTGTGCTCCCTGCAATGTGATCTGCTCCGTTGTTCCAGAAACGAT  
TTGATTATAAAGATCAGAATTTCACCCCTGATGAAAAGATTCAATGAAAACCTCAGGATTCTGAACTC  
CCCATGGATCCAGGTCTGCAATAATTCCCTCTACTCATTGATTGTTCCCAGGAACTCACAACAAAGTG  
CTTAAAAATGTTGCTCTTACACGAAGTTACATTAGGGAGAAAGTAAAAGAACACCAAGCATCAGGATG  
TTAACAACTCTCGGGACTTTATGGATTGCTCCTGATCAAAATGGAGCAGGAAAAGGACAACCAAAAGTC  
AGAATTCAATATTGAAAACCTGGTGGACTGTAGCTGATCTATTGTTGCTGGAACAGAGACAACAAGC  
ACCACTCTGAGATATGGACTCCTGCTCCTGTAAGCACCAGAGGTACAGCTAAAGTCCAGGAAGAGA  
TTGATCATGTAATTGGCAGACACAGGAGCCCTGCATGCAGGATAGGAGCCACATGCCTTACACTGATGC  
TGTAGTGCACGAGATCCAGAGATACAGTGACCTGTCCCCACCGGTGTGCCCATGCAGTGACCACTGAT  
ACTAAGTTCAGAAACTACCTCATCCCCAAGAGCTTGTATAACAAGATAATGCTGGCTGCATAAAACTAGG  
GCACAACCATAATGGCATTACTGACTTCCGTGCTACATGATGACAAAGAATTCTCATCCAAATATCTT  
TGACCCCTGGCCACTTCTAGATAAGAATGGCAACCTTAAGAAAAGTGACTIONTCTCATGCCTTCTCAGCA  
GGAAAACGAATTGTCAGGAGAAGGACTTGCCTGATGGAGCTATTTTATTCTAACCAATTTCAGCA  
AGAACTTAAACCTGAAATCTGATGATTAAAGAACCTCAATAACTACTGCAGTTACCAAGGGATTGT  
TTCTCTGCCACCCATACCAAGATCTGCTCATCCCTGTGAAGAATGCTAGCCCATCTGGCTGCTGAT  
CTGCTATCACCTGCAACTCTTTTATCAAGGACATTCCACTATTATGCTTCTGACCTCTCATCA  
AATCTCCCATTCACTCAATATCCCATAAGCATCCAAACTCCATTAGGAGAGTTGTCAGGTCACTGCA  
CAAATATATCTGCAATTATTCAACTCTGTAACACTTGTATTAATTGCTGCATATGCTAATACCTTCTA  
ATGCTGACTTTTAATATGTTACTGTAAAACACAGAAAAGTGATTAATGAATGATAATTAGTCAT  
TTCTTTGTGAATGTGCTAAATAAAAGTGTATTAATTGCTGGTTCA

## The FASTA format

- FASTA is the name of a popular sequence alignment-and-database-scanning program created by W.R. Pearson and D.J. Lipman in 1988 [PNAS, 85:2444-2448].
- The sequences used by FASTA have to obey the following format:

> (mandatory) definition line (My\_identifiers)

ARCGTCRGCKINTANDRGCKINTAND  
CKINTANDARCGTCRGCKINTANDRG  
CKINTAND

- The line starting with > (the definition line) should contain a short definition or identifiers. The lines that follow it contain the DNA or protein sequence (in one-letter code). Usually, this line has a specific format decided by each database
- Sequences are expected to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes. No numbers allowed! (blank characters are ignored)
- Because FASTA is easy to parse, It is now the default input format for much sequence analysis software.

## FASTA file (ENA)

>ENA|M17398|M17398.1 Human cytochrome P-450 1 mRNA, complete cds, clone Hp1-2.

```
AATGGAACCTTGTGGCCTGGTGCTGTCTCTCTTTATGCTTCTCTTTCACTCTG  
GAGACAGAGCTGTAGGAGAAGGAAGCTCCCTCTGGCCCCACTCCTCTTCTATTATTGG  
AAATATGCTACAGATAGATGTTAAGGACATCTGCAAATCTTCACCAATTCTCAAAAGT  
CTATGGCCTGTGTTCACCGTGATTTGGCATGAATCCCATAGGGTGTTCATGGATA  
TGAGGCAGTGAAGGAAGCCCTGATTGATAATGGAGAGGAGTTCTGGAAGAGGCAATT  
CCCAATATCTCAAAGAATTACTAAAGGACTTGAATCATTCCAGCAATGGAAAGAGATG  
GAAGGGAGATCCGGCGTTCTCCCTCACAAACTTGCAGAATTTGGGATGGGGAGAGGAG  
CATTGAGGACCCTGTTCAAGAGGAAGCTCACTGCCCTGTGGAGGAGTTGAGAAAAACAA  
GGCTTCACCCCTGTGATCCCACCTTCATCCTGGCTGTGCTCCCTGCAATGTGATCTGCTC  
CGTTGTTTCCAGAACGATTGATTATAAGATCAGAATTTCACCCGATGAAAAG  
ATTCAATGAAAACCTCAGGATTCTGAACCTCCCCATGGATCCAGGTCTGCAATAATTCCC  
TCTACTCATTGATTGTTCCAGGAACTCACAACAAAGTGTAAATGTTGCTTTAC  
ACGAAGTTACATTAGGGAGAAAGTAAAAGAACACCAAGCATTGGATGTTAACAAATCC  
TCGGGACTTATGGATTGCTCCTGATCAAATGGAGCAGGAAAGGACAACCAAAAGTC  
AGAATTCAATATTGAAAACCTGGTTGGCACTGTAGCTGATCTATTGTTGCTGGAACAGA  
GACAACAAGCACCACTCTGAGATATGGACTCCTGCTCTGCTGAAGCACCAGAGGTAC  
AGCTAAAGTCCAGGAAGAGATTGATCATGTAATTGGCAGACACAGGAGCCCTGCATGCA  
GGATAGGAGCCACATGCCTTACACTGATGCTGTAGTGCACGAGATCCAGAGATACTG  
CCTTGTCCCCACCGGTGTGCCCCATGCAGTGACCACTGATACTAAGTTAGAAACTACCT  
CATCCCCAAGAGCTTGATAACAAGATAATGCTGGCTGCATAAAACTAGGGCACAAACCAT  
AATGGCATTACTGACTTCCGTGCTACATGATGACAAAGAATTCTTAATCCAAATATCTT  
TGACCCCTGCCACTTCTAGATAAGAATGGCAACTTTAAGAAAAGTGAACCTTCA  
TTTCTCAGCAGGAAAACGAATTGTCAGGAGAAGGACTGCCCCATGGAGCTATT  
ATTTCTAACCAACAAATTACAGAACTTTAACCTGAAATCTGTTGATGATTAAAGAACCT  
CAAACTACTGCAGTTACCAAAGGGATTGTTCTCTGCCACCCCTCATACCAGATCTGCTT  
CATCCCTGTGAGAAATGCTAGCCCATCTGGCTGCTGATCTGCTATCACCTGCAACTCT  
TTTTTATCAAGGACATTCCACTATTATGCTTCTCTGACCTCTCATCAAATCTCCCA  
TTCACTCAATATCCATAAGCATCCAAACTCCATTAAAGGAGAGTTGTTCAAGGTCACTG  
CAAATATATCTGCAATTATTCATACTCTGTAACACTTGTATTAAATTGCTGCATATGCTAA  
TACTTTCTAACGACTTTAAATGTTACTGTAACACAGAAAAGTGAATTAA  
TGAATGATAATTAGTCCATTCTTTGTGAATGTGCTAAATAAAAGTGTATTAAATTG  
CTGGTTCA
```

## Used for both nucleotides and proteins

> My nucleotide sequence

```
AATGGAACCTTTGTGGCCTGGTGCTGTCTCTCTTTATGCTTCTCTTTCACTCTGGAGACAGA  
GCTGTAGGAGAAGGAAGCTCCCTCCTGGCCCCACTCCTCTTCATTATTGGAAATATGCTACAGATA  
GATGTTAAGGACATCTGCAAATCTTCACCAATTCTCAAAGTCTATGGCCTGTGTTCACCGTGT  
TTTGGCATGAATCCCAGTAGTGGTGTTCATGGATATGAGGCAGTGAAGGAAGCCCTGATTGATAATG  
GAGAGGAGTTTCTGGAAGAGGCAATTCCCCAATATCTCAAAGAATTACTAAAGGACTTGGAAATCATT  
TCCAGCAATGGAAAGAGATGGAAGGAGATCCGGCGTTCTCCCTCACAAACTTGCAGAATTTGGGAT  
GGGAAGAGGAGCATTGAGGACCGTGTCAAGAGGAAGCTCACTGCCTGTGGAGGAGTTGAGAAAAAA  
CCAAGGCTTCACCCTGTGATCCCACTTCATCCTGGGCTGTGCTCCCTGCAATGTGATCTGCTCCGTT  
GTTTCCAGAACGATTGATTATAAGATCAGAATTTCACCCCTGATGAAAAGATTCAATGAAAAA  
CTTCAGGATTCTGAACTCCCCATGGATCCAGGTCTGCAATAATTCCCTCTACTCATTGATTGTTCC  
CAGGAACTACAACAAAGTCTAAAAATGTTGCTCTACACGAAGTTACATTAGGGAGAAAGTAAAAA  
GAACACCAAGCATCACTGGATGTTAACAAATCCTCGGGACTTATGGATTGCTTCTGATCAAAATGGA  
GCAGGAAAAGGACAACCAAAAGTC
```

> My protein sequence

```
LCLYTHIGRNIYYGSYLYSETWNTGIMLLITMATAFMGYVLPWGQMSFWGATVITNLFSAI PYIGTN  
LVEWIWGGFSVDKATLNRRFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIK  
DFLGLLILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALF  
LSIVILGLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTWIGSQPV EYPYTIIGQMASILYFSIIL  
AFLPIAGXIENY
```

## Sequence formats

---

ASN.1

DNAStrider

EMBL

Fitch

GCG

GenBank/GB

IG/Stanford

MSF

NBRF

Olsen

PAUP/NEXUS

Pearson/Fasta

Phylip

PIR/CODATA

Plain/Raw

Pretty

Zuker

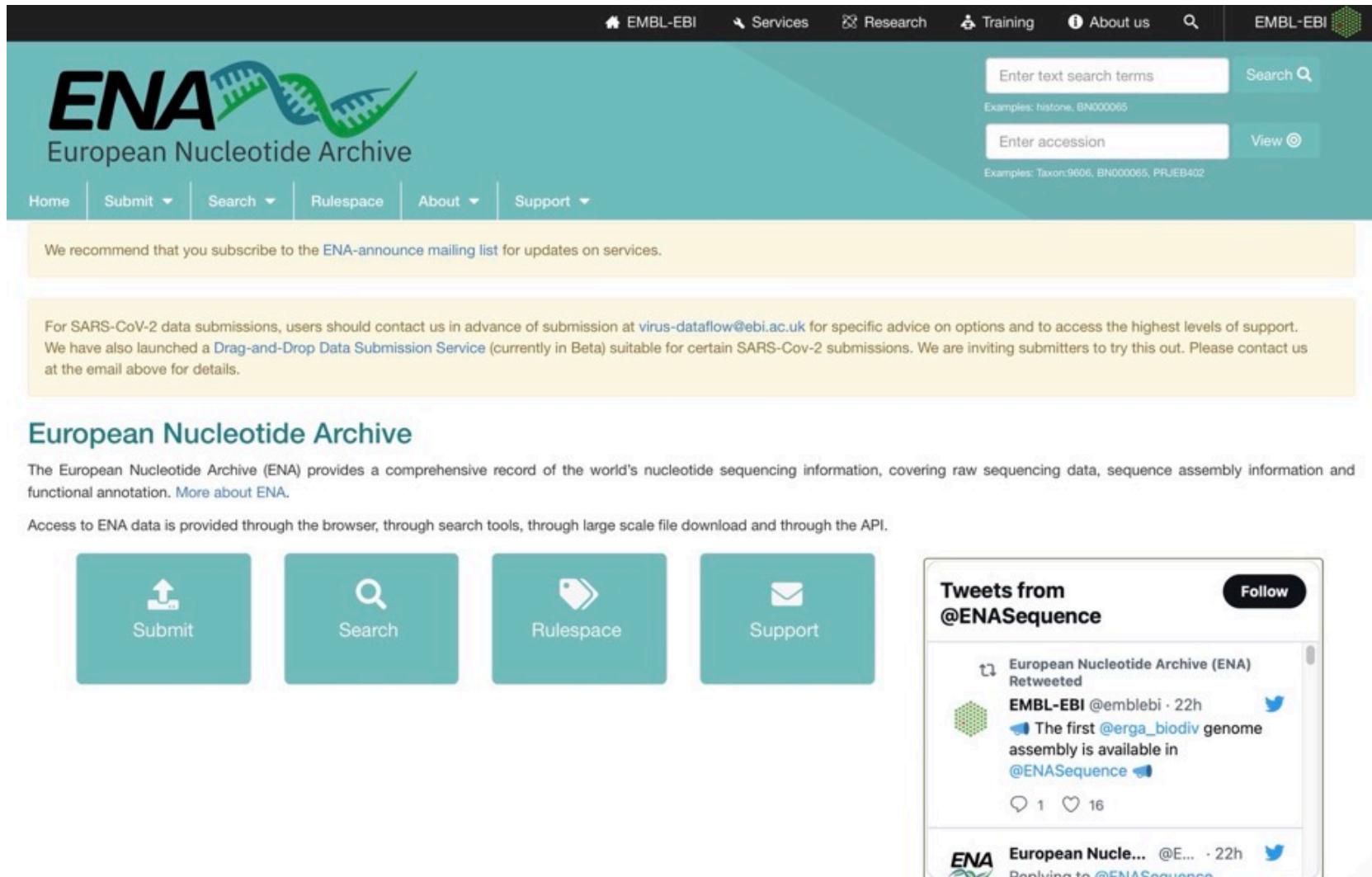
**NOTE:**

- FASTA is a popular sequence format

# The EMBL-EBI nucleotide repository

## ENA

<http://www.ebi.ac.uk/ena/>



The screenshot shows the ENA homepage with a dark header bar containing links for EMBL-EBI, Services, Research, Training, About us, and a search icon. The main navigation bar below has links for Home, Submit, Search, Rulespace, About, and Support. A large green banner features the ENA logo with a DNA helix and the text "European Nucleotide Archive". To the right of the banner are two search boxes: one for text search terms and another for accession numbers, both with examples provided. Below the banner, a yellow box contains a recommendation to subscribe to the ENA-announce mailing list. A larger yellow box further down provides information about SARS-CoV-2 data submissions, mentioning the email address virus-dataflow@ebi.ac.uk and a Drag-and-Drop Data Submission Service. The main content area is titled "European Nucleotide Archive" and describes the archive's purpose: providing a comprehensive record of nucleotide sequencing information. It also mentions the API and various access methods. Below this, four teal buttons offer "Submit", "Search", "Rulespace", and "Support". To the right, a Twitter sidebar displays tweets from the account @ENASequencing, including one from EMBL-EBI announcing a genome assembly.

Home | Submit | Search | Rulespace | About | Support

We recommend that you subscribe to the ENA-announce mailing list for updates on services.

For SARS-CoV-2 data submissions, users should contact us in advance of submission at [virus-dataflow@ebi.ac.uk](mailto:virus-dataflow@ebi.ac.uk) for specific advice on options and to access the highest levels of support. We have also launched a [Drag-and-Drop Data Submission Service](#) (currently in Beta) suitable for certain SARS-CoV-2 submissions. We are inviting submitters to try this out. Please contact us at the email above for details.

## European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. [More about ENA](#).

Access to ENA data is provided through the browser, through search tools, through large scale file download and through the API.

Submit | Search | Rulespace | Support

**Tweets from @ENASequencing**

European Nucleotide Archive (ENA) Retweeted  
EMBL-EBI @emblebi · 22h  
The first @erga\_biodiv genome assembly is available in @ENASequencing

ENAS European Nucle... · 22h

# There are several ways to search and retrieve data from GenBank.

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

- Free text search



- Advanced searches (search tags and regular expressions)
- List of identifiers (Batch *Entrez*)

## Search sequence databases using a sequence query

- BLAST (Basic Local Alignment Search Tool)

## Web service (Programmatic data access)

- *Entrez* Utilities Web Service (NCBI): The E-utilities

## File Transfer Protocol (FTP)

- Flat files (bulk data download)

<ftp://ftp.ncbi.nlm.nih.gov/genbank>

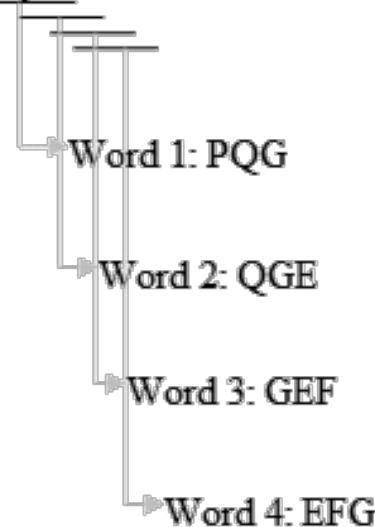
<http://ftp.ncbi.nlm.nih.gov/genbank>

# How does BLAST work?

Generate a list of k-words

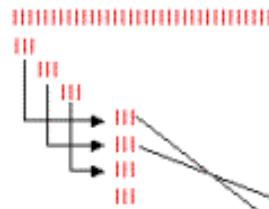
Not all of them, “low-complexity” regions  
(pe:PPCDPPPPPDKKKKDDGPP) are not used to generate k-words

Query sequence: PQGEFG



## BLAST Heuristic Mechanism

query sequence length L

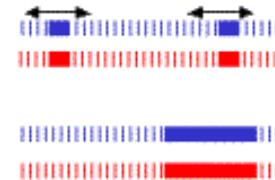


compile list of **high scoring** words (w) from query  
usually w = 3 for proteins  
there are  $L - w + 1$  words per length of sequence

- Search for these "words"

- Extend the alignment

- Stop extension when the accumulated score falls below a threshold (*dropoff value*)



compare word list with sequences in database and identify exact matches

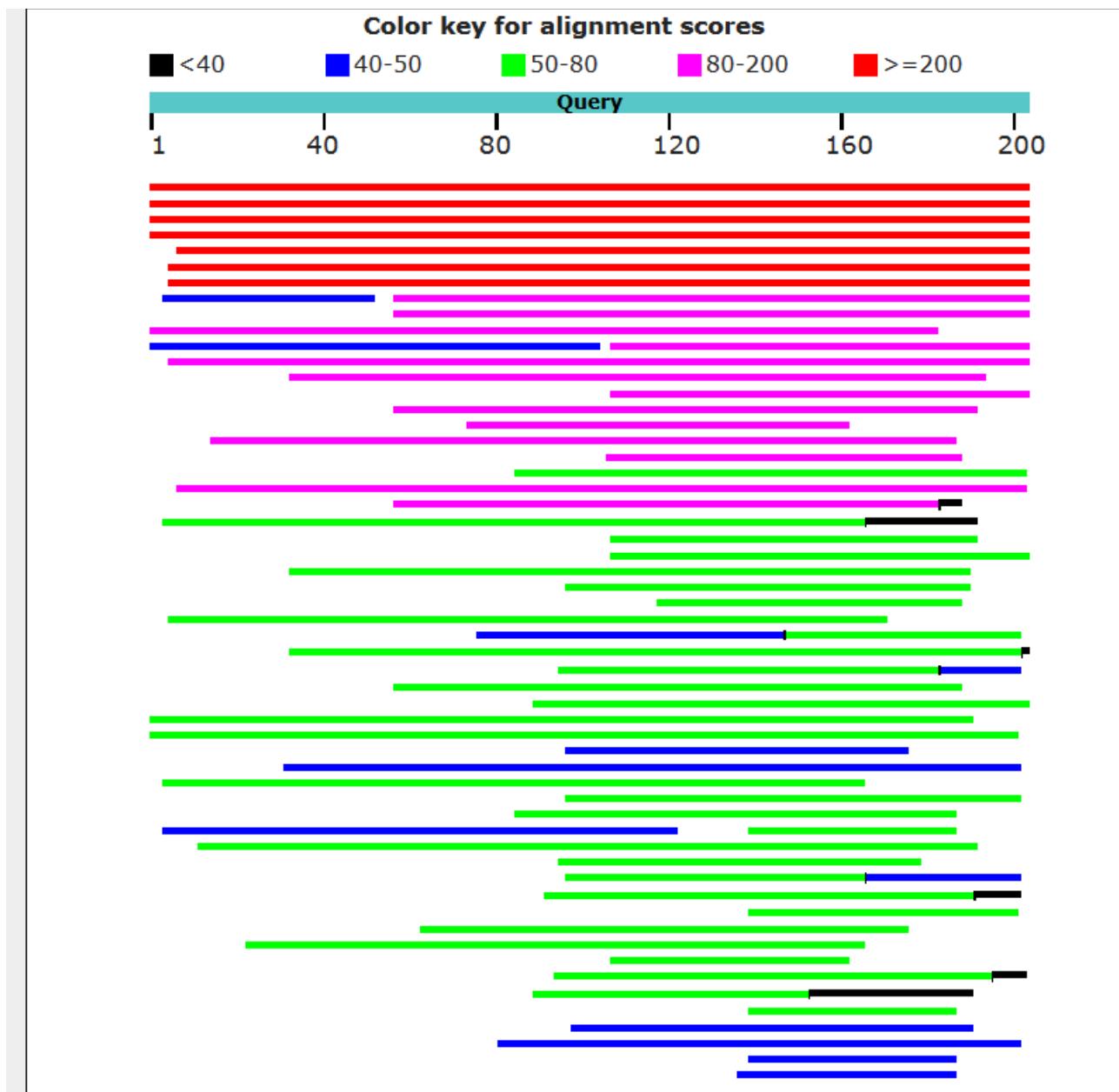
extend matches in both directions to find alignments scoring greater than a threshold

high scoring segment pair - HSP

# BLOSUM 62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																			C	
S	-1	4																		S	
T	-1	1	5																	T	
P	-3	-1	-1	7																P	
A	0	1	0	-1	4															A	
G	-3	0	-2	-2	0	6														G	
N	-3	1	0	-2	-2	0	6													N	
D	-3	0	-1	-1	-2	-1	1	6												D	
E	-4	0	-1	-1	-1	-2	0	2	5											E	
Q	-3	0	-1	-1	-1	-2	0	0	2	5										Q	
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									H	
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								R	
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							K	
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						M	
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					I	
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				L	
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			V	
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		F	
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	Y	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

# The BLAST output: The NCBI graphical format



# The BLAST output

- Produces local alignments: only a portion of each sequence could be aligned.
- Calculates similarity for biological sequences.
  - Score, % identity and % similarity
- Uses statistical theory to determine if a match might have occurred by chance. (Expected value ( $E$ ) )

PREDICTED: putative uncharacterized protein FLJ44672 [Gorilla gorilla gorilla]

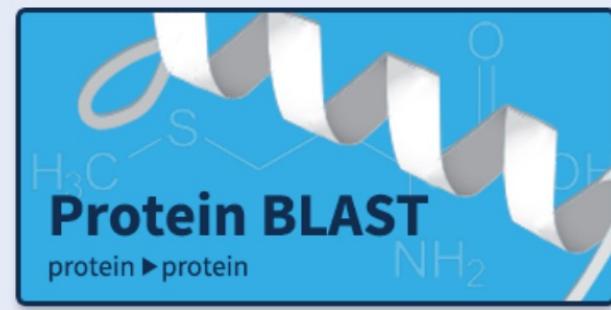
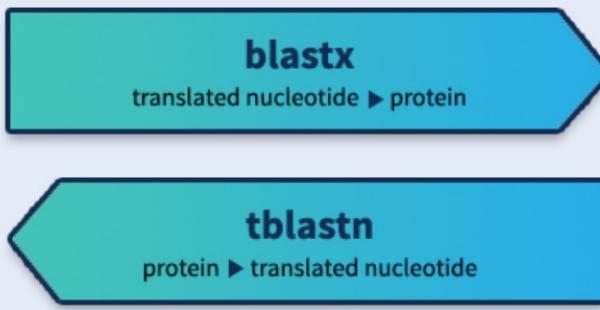
Sequence ID: [XP\\_004045538.1](#) Length: 213 Number of Matches: 1

Range 1: 1 to 213 <a href="#">GenPept</a> <a href="#">Graphics</a>					<a href="#">▼ Next Match</a>	<a href="#">▲ Previous Match</a>
Score	Expect	Method	Identities	Positives	Gaps	
281 bits(720)	7e-94	Compositional matrix adjust.	180/213(85%)	183/213(85%)	10/213(4%)	
Query 1	MPLLASPGPAPACWRPVEAQPLPQQWALHAHLLPRRGGLLPGSRLGAAPAGPAPASR-PS				59	
Sbjct 1	MPLLASPGPAPACWRP+EAQPLPQQWALHAHL PRRGGLLPGSR+GAAP GPAPASR PS					
Query 60	MPLLASPGPAPACWRPLEAQPLPQQWALHAHLSRRGGLLPGSRVGAAPTGPAPASRRPS				60	
Sbjct 61	GGQAHASGRPLPAWRLLLLCMGSRSCTSSGRPLQTQL-----FLPAASVGPDRCRQVG				110	
Query 111	GGQAHASGRPLPAWRLLLLCMGSR CTSS RPLQ Q FLPAAS GPDCRQVG					
Sbjct 121	GGQAHASGRPLPAASAGPDCRQVGLSRDSSCLPSASVGPSRPQVGLPRPSSGLSAASAKVP				120	
Query 171	LSRDSSCLPAASVGPDCHQVGLSMMDSSCLPVTSVGPSRPQVGLPRPSSGLSAASAKVP				170	
Sbjct 181	LSRDSSCLPAASAGPDCRQVGLSRDSSCLPSASVGPSRPQVGLPRPSSGLSAASSSAKVP				180	
Query 171	RVSLSPRSSCLPVASFGPAQFMPPGGLPRPHF 203					
Sbjct 181	RV LSR SSCLPVASF PAQ +PPGGL RP F					
Sbjct 181	RVRLSRSSSCLPVASFSPAQLIPPGGLSRPRF 213					

# BLAST Basic programs

Web BLAST

<https://blast.ncbi.nlm.nih.gov/Blast.cgi>



**nucleotide blast**    Search a nucleotide database using a nucleotide query

**protein blast**    Search protein database using a protein query

**blastx**    Search protein database using a translated nucleotide query

**tblastn**    Search translated nucleotide database using a protein query

**tblastx**    Search translated nucleotide database using a translated nucleotide query

# The BLAST family of programs

QUERY  
SEQUENCE

DATABASE

Nucleic Acid

*blastn*



conceptual  
protein  
translations

*tblastx*

*blastx*

Nucleic Acids



conceptual  
protein  
translations

Peptide/Protein

*tblastn*

*blastp*

Proteins/Peptides



# BLAST databases:

Nucl. Database	Content
nt (default)	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 databases	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 databases 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
nr (default)	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)

# NCBI Datasets

[https://www.ncbi.nlm.nih.gov/datasets/docs/v2/getting\\_started/](https://www.ncbi.nlm.nih.gov/datasets/docs/v2/getting_started/)

National Library of Medicine  
National Center for Biotechnology Information

Search NCBI ... Log in

NCBI Datasets Taxonomy Genome Gene Command-line tools Documentation

Eukarya / Metazoa / Chordata / Mammalia / Primates / Hominidae / Homo

### Homo sapiens

Human (*Homo sapiens*) is a species of primate in the family *Hominidae* (great apes).

Browse taxonomy

Current scientific name: *Homo sapiens*  
Common name: human  
Taxonomic rank: species  
NCBI Taxonomy ID: 9606

For more details see [NCBI Taxonomy](#)  
View the legacy [Genome page](#)

#### Genome

Browse all 1,090 genomes

Reference genome: GRCh38.p14  
Genome Reference Consortium (2022). RefSeq GCF\_000001405.40

Download

Genome size: 3.1 Gb  
Contig N50: 57.9 Mb  
Genes: 59,444

View full annotation report  
View in [Genome Data Viewer \(GDV\)](#)

Includes updated and unannotated genes

#### Database links

	Nucleotide	Protein
All nucleotide sequences	28,487,720	Protein sequences 1,923,501
Genomic sequences	15,849,089	Conserved domains 83
mRNA sequences	9,925,357	3D structures 63,642

#### GEO Datasets

	Datasets	Sequence Read Archive (SRA)
All SRA experiments	1,772	4,526,538
Series	93,426	DNA 2,780,051

Current gene set pie chart: 55.1% Other, 25.8% Non-coding, 18.1% Protein-coding, 0.8% Pseudogenes, 0.2% Small RNAs

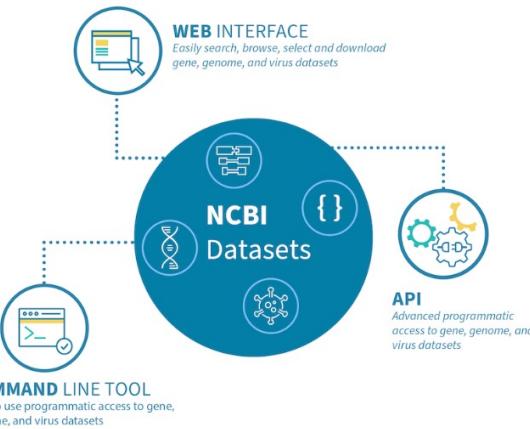
View all genes

Variation resources: ClinVar, dbSNP

External links: Encyclopedia of Life, GBIF, iNaturalist, Wikipedia

NCBI Datasets is a new resource that lets you easily gather data from across NCBI databases. You have the choice of getting the data through three interfaces:

- NCBI Datasets website
- Command-line tools
- API : Accessible with UNIX tools (such as wget and curl).



National Library of Medicine  
National Center for Biotechnology Information

Search NCBI ... Log in

NCBI Datasets Taxonomy Genome Gene Command-line tools Documentation

#### Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

Selected taxa: Homo sapiens (human) Enter one or more taxonomic names

Filters

Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotat...	Size	Action
GRCh38.p14 	GCA_000001405.29	GCF_000001405.40	Homo sapiens (human)	<a href="#">NCBI RefSeq</a>	3,01		
T2T-CHM13v2.0	GCA_009914755.4	GCF_009914755.1	Homo sapiens (human)	<a href="#">NCBI RefSeq</a>	3,1		
Ash1_v2.2	GCA_011064465.2		Homo sapiens (human)	NA24385 (isol...)	Submitter	3,1	
hg01243.v3.0	GCA_018873775.2		Homo sapiens (human)	HG01243 (isol...)	Submitter	3,1	
Han1	GCA_024586135.1		Homo sapiens (human)	HG00621 (isol...	Submitter	3,1	
HG002.pat.cur.20211005	GCA_021950905.1		Homo sapiens (human)	NA24385 (isol...	Submitter	2,91	