

# Bioinformatics

Chap\_02

Introduction to Biological Databases

## Introduction

The development of databases to handle the vast amount of molecular biological data is thus a fundamental task of bioinformatics.

## WHAT IS A DATABASE?

- ✓ A **database** is a **computerized archive** used to **store** and **organize** data in such a way that information can be **retrieved easily** via a variety of search criteria.
- The **chief objective** of the **development** of a **database** is to **organize data** in a **set of structured records** to **enable easy retrieval of information**.
- Each **record** is called an **entry**.
- To **retrieve a particular record from the database**, a user can **specify** a particular **piece of information**, called **value**, to be found in a particular field and expect the computer to retrieve the **whole data** record. This process is called **making a query**.
- **Another objective** of the development of a database is **knowledge discovery**.

# BIOLOGICAL DATABASES

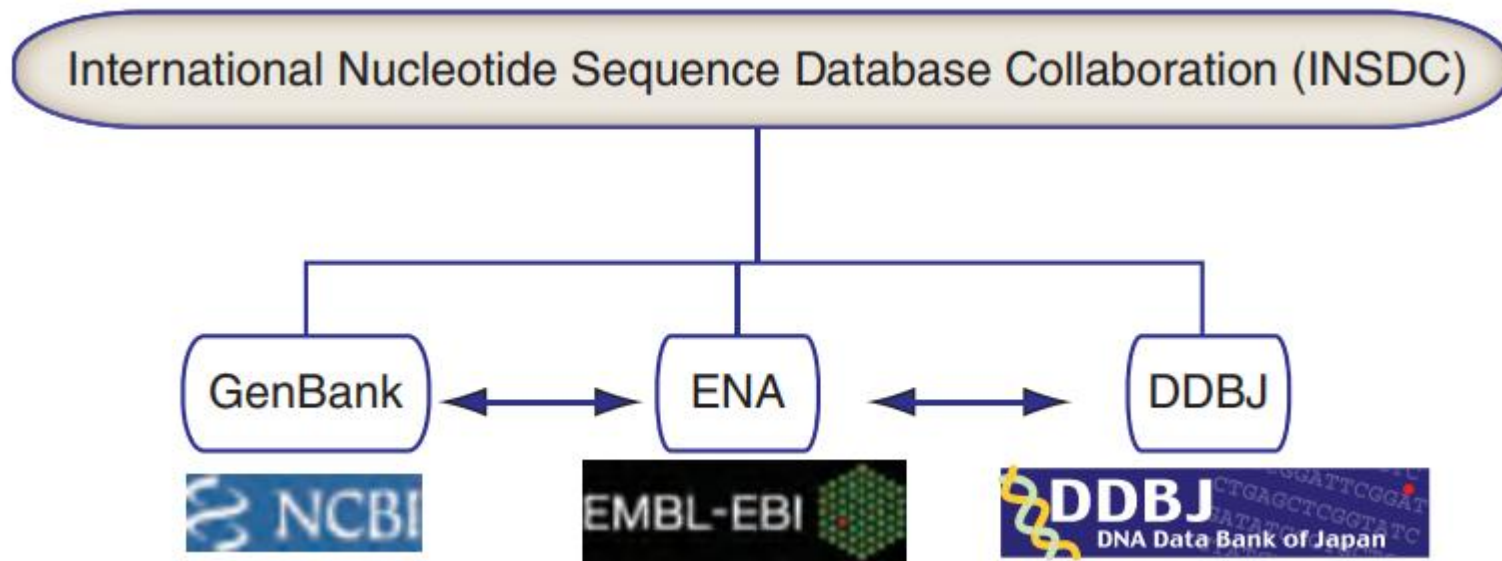
Despite the obvious drawbacks of using flat files in database management, **many biological databases still use this format.**

**Based on** their **contents**, **biological databases** can be roughly divided into three categories:

- **Primary databases**; they contain **original biological data** (e.g. **GenBank** and **Protein Data Bank (PDB)**).
- **Secondary databases**; they contain **computationally processed** or **manually curated information**, based on original information from primary databases (e.g. **SWISS-Prot** and **Protein Information Resources (PIR)**).
- **Specialized databases** are those that cater to a **particular research interest** (e.g. **Flybase**, **HIV sequence database**, and **Ribosomal Database Project**).

## Centralized Databases Store DNA Sequences

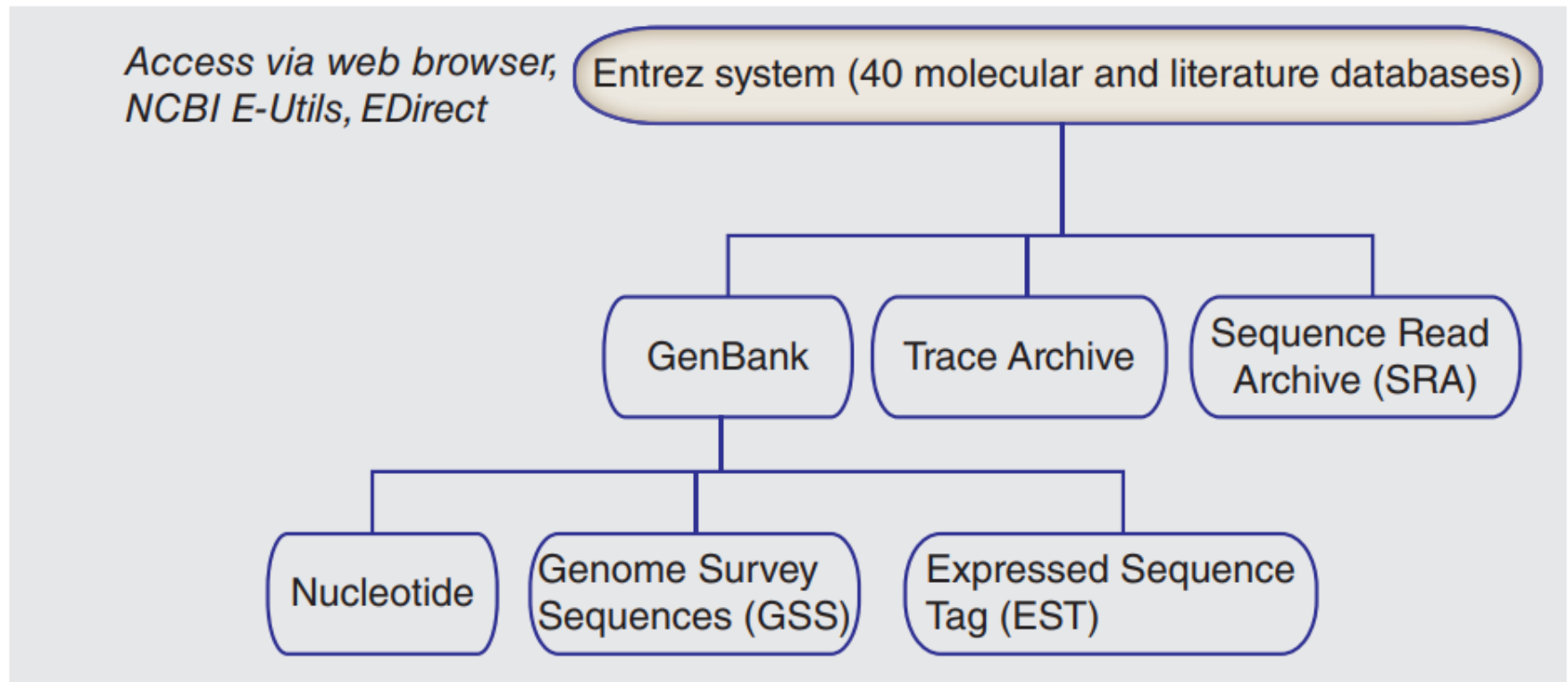
How much DNA sequence is stored in public databases? Where are the data stored?



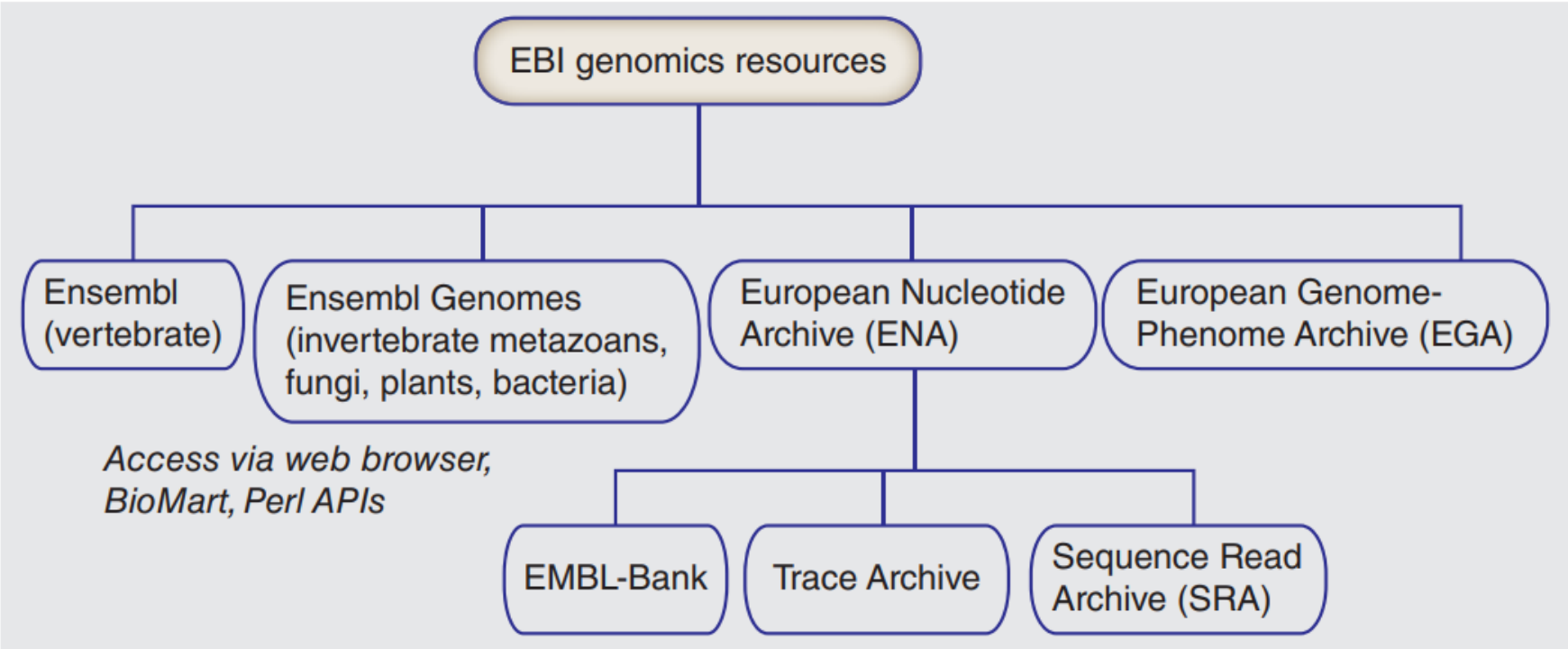
- **GenBank** at the National Center for Biotechnology Information (**NCBI**) of the National Institutes of Health (**NIH**) in **Bethesda** (NCBI Resource Coordinators)
- European Molecular Biology Laboratory (**EMBL**)-**Bank** Nucleotide Sequence Database (**EMBL-Bank**), part of the European Nucleotide Archive (**ENA**) at the European Bioinformatics Institute (**EBI**) in **Hinxton, England**
- DNA Database of Japan (**DDBJ**) at the **National Institute of Genetics** in **Mishima**

NCBI, EBI, and DDBJ offer many dozens of other resources for the study of sequence data

## National Center for Biotechnology Information

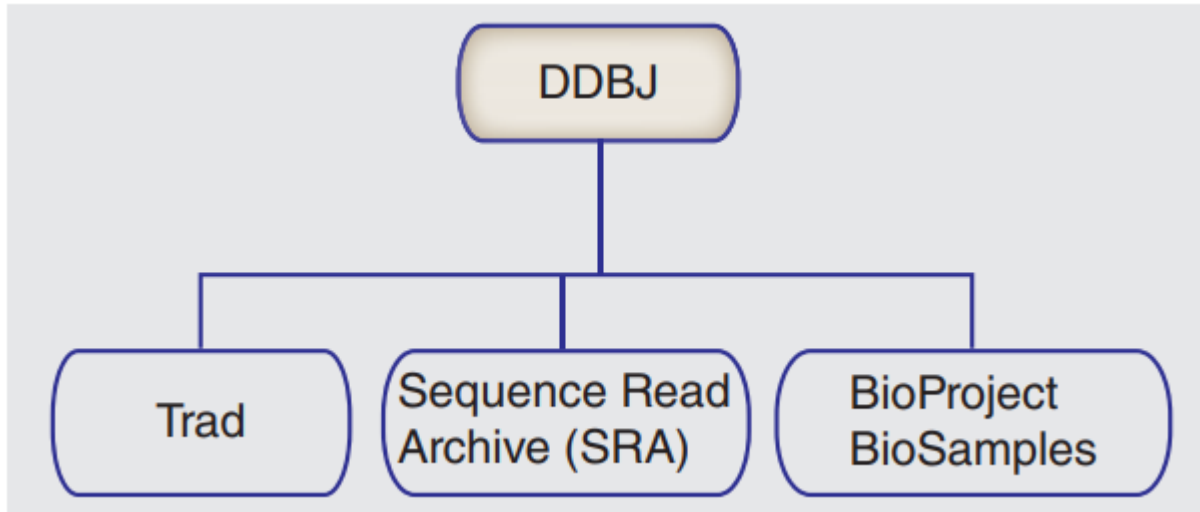


- The **Trace Archive** on **NCBI** (National Center for Biotechnology Information) is a **repository** that stores **raw data** from **DNA sequencing experiments**, particularly from **older sequencing technologies** like **Sanger sequencing** and **capillary electrophoresis**.
- The **Sequence Read Archive (SRA)** stores **next-generation sequence data (NGS)**.



Within **ENA**, **EMBL-Bank** includes the **same raw sequence data as GenBank** at NCBI. Similar data are also housed in the Trace Archive and SRA.

## DNA Database of Japan



Traditional (**Trad**) division shares the **same raw sequence data** with **GenBank** and **EMBL-Bank** daily

## Sequence submission to databanks:

**Data Submission Process** – How researchers submit sequence records.

Researchers can submit records **directly** to **NCBI**, **EBI**, and **DDBJ**

**Quality Control Measures** – **Guidelines** and **reconciliation projects** (e.g., RefSeq).

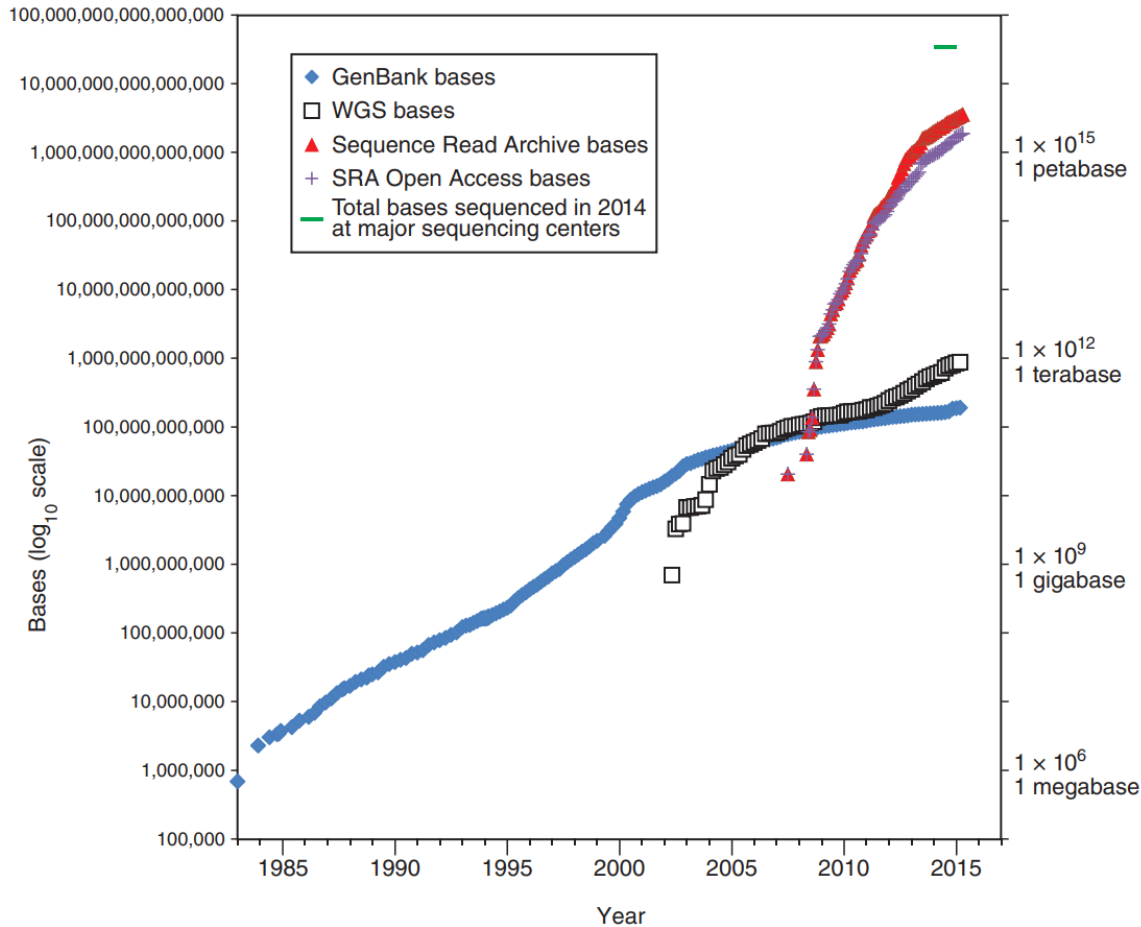
- **Guidelines** ensure **quality control** during submission.
- Projects like RefSeq **reconcile** differences between submitted entries.

**GenBank and Automation** – Use of the **tbl2asn** tool for **sequence record creation**.

NCBI provides the command-line tool **tbl2asn** to automate GenBank sequence record creation.

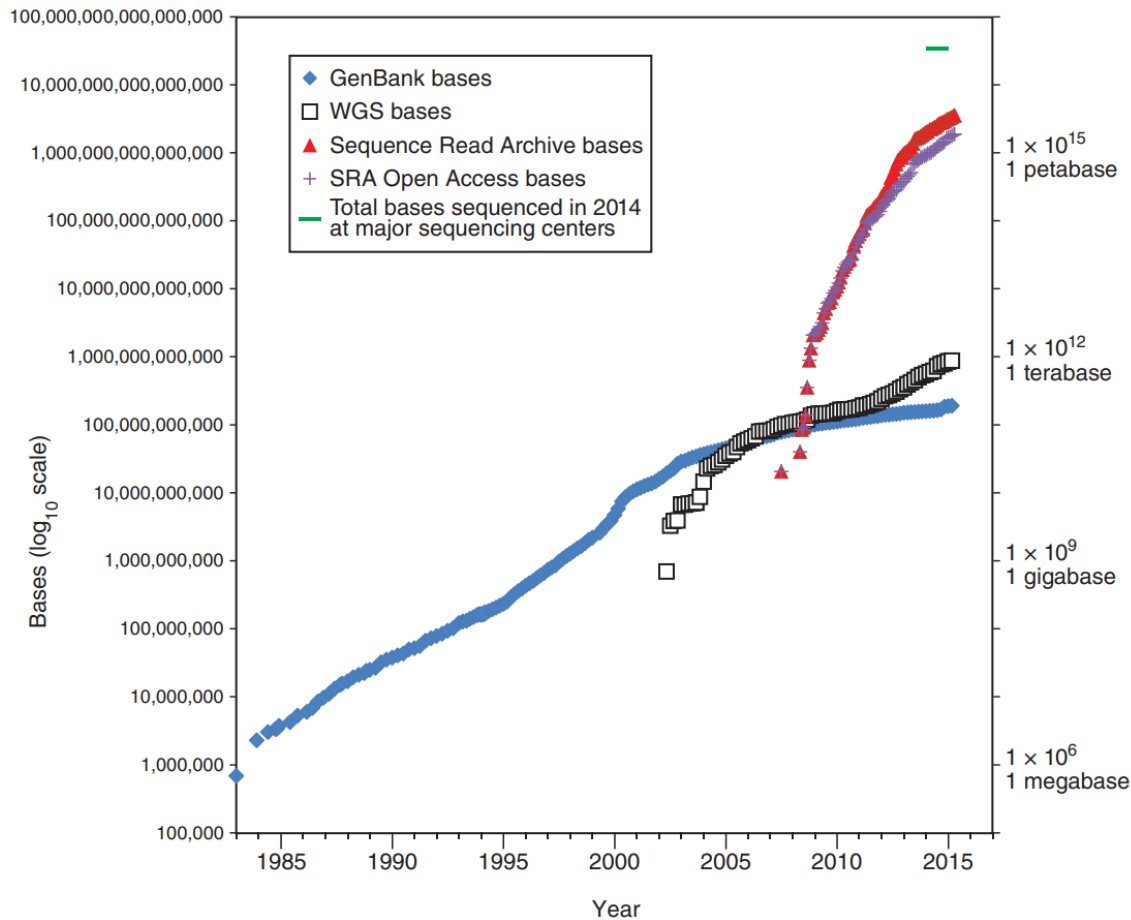


## Growth of DNA sequence in repositories



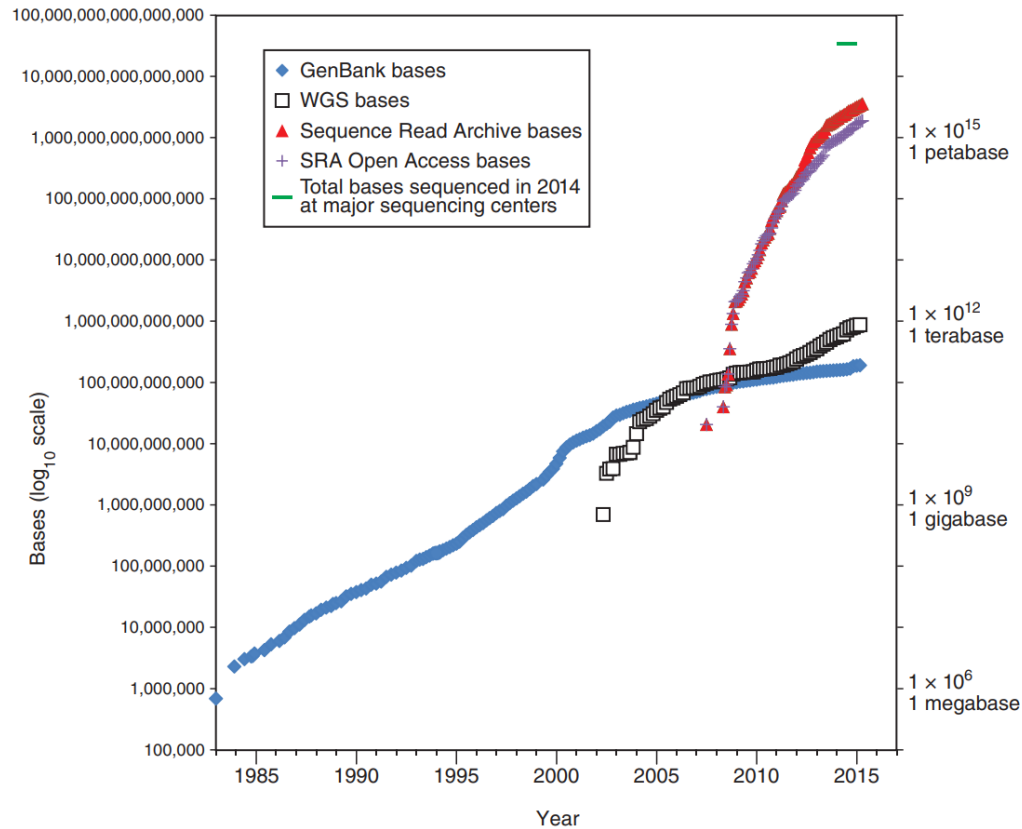
Growth of DNA Repositories: Overview of **GenBank's growth** since 1982, **doubling** every **18 months** (mirrored by EMBL-Bank and DDBJ).

## Growth of DNA sequence in repositories



**Whole-Genome Shotgun (WGS) Data:** Explanation of **WGS** sequences, introduced in **2002**, and their **exclusion** from **main** GenBank/EMBL-Bank/DDBJ releases, with **base pairs surpassing** GenBank's holdings.

# Growth of DNA sequence in repositories



**Sequence Read Archive (SRA):** Introduction to SRA, containing vastly more data (**3000x more bases**) than GenBank and WGS combined, driven by **short-read next-generation sequencing** (50–400 base pairs).

- **Data Accessibility:** Distinction between **open-access SRA data** (e.g., sequences from various organisms) and **restricted human-derived data**, requiring ethical approval for access

To make sense of such large numbers of bases of DNA we can look at several specific examples (see Table)

Base pairs	Unit	Abbreviation	Example
1	1 base pair	1 bp	
1000	1 kilobase pair	1 kb	Size of a typical coding region of a gene
1,000,000	1 megabase pair	1 Mb	Size of a typical bacterial genome
$10^9$	1 gigabase pair	1 Gb	The human genome is 3 billion base pairs
$10^{12}$	1 terabase pair	1 Tb	
$10^{15}$	1 petabase pair	1 Pb	

# Contents of DNA, RNA, and Protein Databases

## Overview of DNA, RNA, and Protein Databases:

Introduction to **GenBank** (and its equivalence to DDBJ and EMBL-Bank) as a **repository** of **public DNA** and **protein** sequences, **excluding next-generation data**, with **bibliographic** and **biological annotations**, **freely accessible** via NCBI.

## Diversity of Organisms in GenBank:

Representation of over **310,000** species, with **1000+** new species added **monthly**, spanning **bacteria**, **archaea**, **eukaryotes**, and **viruses**, as detailed in the following Table.

### Taxa represented in GenBank

Ranks	Higher taxa	Genus	Species	Lower taxa	Total
Archaea	143	140	525	0	808
Bacteria	1,370	2,611	13,331	819	18,131
Eukaryota	20,443	67,606	297,207	22,608	407,864
Fungi	1,550	4,620	29,450	1,128	36,748
Metazoa	14,670	45,517	145,044	11,428	216,659
Viridiplantae	2,622	14,680	113,529	9,789	140,620
Viruses	618	442	2,349	0	3,409
All taxa	22,603	70,806	313,443	23,427	430,279

**Focus on Model Organisms:** Emphasis on commonly studied mammals (**human, mouse, cow**), vertebrates (**chicken, frog**), and plants (**corn, rice, bread wheat, wine grape**) in biological research.

Overview of the 10 most sequenced organisms in GenBank (excluding chloroplast/mitochondrial sequences), featuring key model organisms.

**Table. Ten most sequenced organisms in GenBank.**

Entries	Bases	Species	Common name
20,614,460	17,575,474,103	<i>Homo sapiens</i>	Human
9,724,856	9,993,232,725	<i>Mus musculus</i>	Mouse
2,193,460	6,525,559,108	<i>Rattus norvegicus</i>	Rat
2,203,159	5,391,699,711	<i>Bos taurus</i>	Cow
3,967,977	5,079,812,801	<i>Zea mays</i>	Maize
3,296,476	4,894,315,374	<i>Sus scrofa</i>	Pig
1,727,319	3,128,000,237	<i>Danio rerio</i>	Zebrafish
1,796,154	1,925,428,081	<i>Triticum aestivum</i>	Bread wheat
744,380	1,764,995,265	<i>Solanum lycopersicum</i>	Tomato
1,332,169	1,617,554,059	<i>Hordeum vulgare subsp. vulgare</i>	Barley

To help organize the available information, each **sequence name** in a **GenBank** record is followed by its [data file division and primary accession number](#).

1. **PRI**: primate sequences
2. **ROD**: rodent sequences
3. **MAM**: other mammalian sequences
4. **VRT**: other vertebrate sequences
5. **INV**: invertebrate sequences
6. **PLN**: plant, fungal, and algal sequences
7. **BCT**: bacterial sequences
8. **VRL**: viral sequences
9. **PHG**: bacteriophage sequences
10. **SYN**: synthetic sequences
11. **UNA**: unannotated sequences
12. **EST**: expressed sequence tags
13. **PAT**: patent sequences
14. **STS**: sequence-tagged sites
15. **GSS**: genome survey sequences
16. **HTG**: high-throughput genomic sequences
17. **HTC**: high-throughput cDNA sequences
18. **ENV**: environmental sampling sequences
19. **CON**: constricted sequences
20. **TSA**: transcriptome shotgun assembly sequences

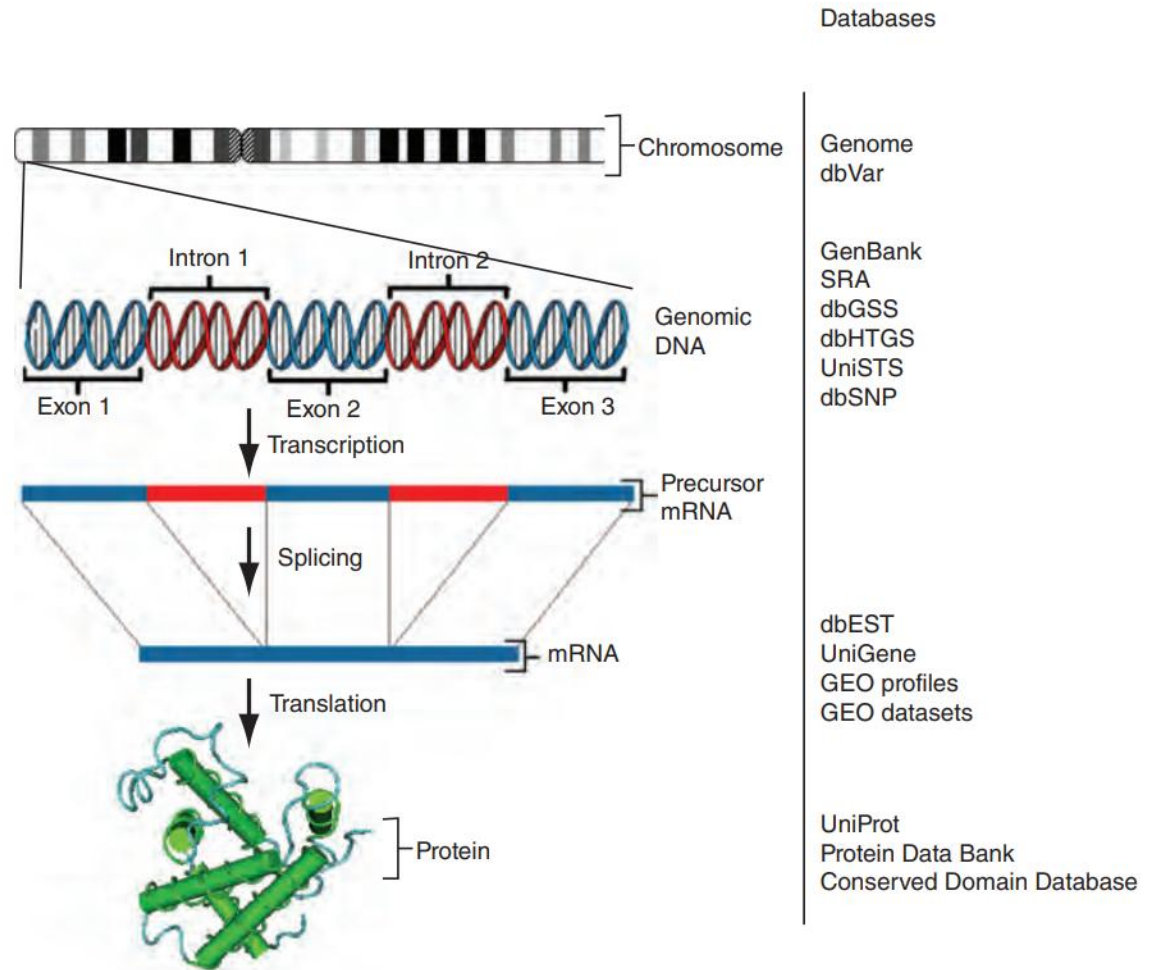
## Types of Data in GenBank/EMBL-Bank/DDBJ

There are **enormous molecular sequences** in the **DDBJ**, **EMBL-Bank**, and **GenBank** databases.

**Q:** We want to find out the sequence of human beta-globin.

**Sequence Data Types:** Distinction between **DNA**, **RNA-based**, and **protein** sequences **stored** in **separate databases**, each represented in multiple forms (e.g., DNA as a gene, RNA as mRNA, protein).

**RNA to cDNA Conversion:** Explanation of RNA's **instability** leading to its conversion to **complementary DNA** (cDNA) for **storage** in **databases**.



**Naming Nuances:** Clarification that "**hemoglobin gene**" may be a **misnomer**, as **globin genes** encode **globin proteins** which combine with heme to form hemoglobin; suggestion that "**globin, beta**" might be **more accurate**.



## Genomic DNA Databases

- **Gene Definition:** A gene (e.g., **human HBB** on chromosome 11) is a **DNA sequence** with **regulatory regions**, **exons**, and **introns**; typically **10–100 kb** in size.
- **Location:** Genes are **part** of **chromosomes** and can be **cloned** in large DNA fragments (e.g., **cosmids**, **BACs**, **YACs**).
- **BACs & YACs:** Bacterial Artificial Chromosomes (up to 200 kb) and Yeast Artificial Chromosomes are **vectors** for **sequencing large genomic portions**.

**Basic kinds of DNA sequence data in GenBank, EMBL-Bank, and DDBJ:**

### ***DNA-Level Data: Sequence-Tagged Sites (STSs)***

- **Definition:** STSs are **short** (~500 bp) **genomic landmarks** with **sequence** and **mapping data** (**Probe database**, NCBI).
- **Coverage:** Obtained from **hundreds** of **organisms** (e.g., primates, rodents).
- **Utility:** Useful for **mapping studies** due to polymorphic short sequence repeats.

**Note:** The **NCBI Probe database** was **retired** in April **2020** and has **no direct replacement**.

Users seeking similar information may now utilize other databases or services provided by NCBI, such as **Gene** or **GenBank**.

## ***DNA-Level Data: Genome Survey Sequences (GSSs)***

- **NCBI Search Structure:** Nucleotide database splits into **GSS**, **ESTs**, and **CoreNucleotide**.
- **GSS Overview:** **Genomic-origin sequences** (unlike **ESTs** from cDNA/mRNA).
- **Data Types:**
  - Random single-pass reads
  - Cosmid/BAC/YAC end sequences
  - Exon-trapped genomic sequences
  - *Alu* PCR sequences

**Note:** 38 million GSS entries are from over **1000 organisms** (February 2015). The **top four organisms** account for about **one-third of all entries** (the mouse *Mus musculus*, a marine metagenome collection, the maize *Zea mays*, and human).

## ***DNA-Level Data: High-Throughput Genomic Sequence (HTGS)***

- **Purpose:** Provides **rapid access** to "unfinished" genomic sequences.
- **Collaboration:** Coordinated by **DDBJ**, **EMBL**, and **GenBank**.
- **Source:** Generated by **high-throughput sequencing centers**.

## RNA-Level Data Overview

- Transition from DNA: **Shift** from **DNA sequence data** (GenBank, EMBL-Bank, DDBJ) to **RNA-level data**.
- **Focus:** RNA data derived from **expressed genes** and stored as **cDNA**.

## cDNA Databases

- **Gene Expression:** **Protein-coding genes**, **pseudogenes**, and **noncoding genes** transcribed into RNA, **varying** by **tissue** (e.g., liver) and **developmental stage**.
- **cDNA Process:** RNA **purified** and **converted** to **stable cDNA**; **beta globin** represented as **expressed sequence tags (ESTs)** in databases.

## Expressed Sequence Tags (ESTs)

- **dbEST Overview:** Division of GenBank with **single-pass cDNA sequences** (300–800 bp) from various organisms (e.g., **human brain, rat liver**).
- **Characteristics:** Randomly selected, one-strand sequenced, **prone to errors**; early **EST efforts identified hundreds of novel genes**.
- **Categories & Stats:** Divided into **human** (~8.7M ESTs), **mouse**, and **others** (Table); ~400 ESTs per **human protein-coding gene** (20,300 genes).

### Top ten organisms for which ESTs have been sequenced.

Organism	Common name	Number of ESTs
<i>Homo sapiens</i>	Human	8,704,790
<i>Mus musculus + domesticus</i>	Mouse	4,853,570
<i>Zea mays</i>	Maize	2,019,137
<i>Sus scrofa</i>	Pig	1,669,337
<i>Bos taurus</i>	Cattle	1,559,495
<i>Arabidopsis thaliana</i>	Thale Cress	1,529,700
<i>Danio rerio</i>	Zebrafish	1,488,275
<i>Glycine max</i>	Soybean	1,461,722
<i>Triticum aestivum</i>	Wheat	1,286,372
<i>Xenopus (Silurana) tropicalis</i>	Western clawed frog	1,271,480

## UniGene Project

- **Goal:** Cluster ESTs into **nonredundant, gene-specific sets (one UniGene cluster per gene)**.
- **Variability:** Clusters **range from 1 EST (rarely expressed genes) to tens of thousands (highly expressed, e.g., ~2400 ESTs for beta-globin)**.
- **Scope:** Represents **142 organisms** across **19 phyla** (Table).

Phylum	Number of species	Example
Chordata	42	<i>Equus caballus</i> (horse)
Echinodermata	2	<i>Strongylocentrotus purpuratus</i> (purple sea urchin)
Arthropoda	19	<i>Apis mellifera</i> (honey bee)
Mollusca	2	<i>Aplysia californica</i> (California sea hare)
Annelida	2	<i>Alvinella pompejana</i>
Nematoda	2	<i>Caenorhabditis elegans</i> (nematode)
Platyhelminthes	3	<i>Schistosoma mansoni</i>
Porifera	1	<i>Amphimedon queenslandica</i>
Cnidaria	3	<i>Nematostella vectensis</i> (starlet sea anemone)
Ascomycota	5	<i>Neurospora crassa</i>
Basidiomycota	1	<i>Filobasidiella neoformans</i>
Codonosigidae	1	<i>Monosiga ovata</i>
Streptophyta	50	<i>Zea mays</i> (maize)
Chlorophyta	2	<i>Chlamydomonas reinhardtii</i>
Apicomplexa	1	<i>Toxoplasma gondii</i>
Bacillariophyta	1	<i>Phaeodactylum tricornutum</i>
Oomycetes	2	<i>Phytophthora infestans</i> (potato late blight agent)
Dictyosteliida	1	<i>Dictyostelium discoideum</i> (slime mold)
Ciliophora	2	<i>Paramecium tetraurelia</i>

**Table. 19 Phyla and 142 organisms represented in UniGene.**

## UniGene Challenges

- **Discrepancy:** ~20,300 human genes **vs.** 130,000 UniGene clusters.
- **Reasons:**
  - **Low-Level Transcription:** 64,000 clusters with 1 EST, 100,000 with 1–4 ESTs (possible rare events, per ENCODE project).
  - **Cloning Artifacts:** Some cDNA may not reflect true transcripts; alternative splicing may create false clusters.
  - **Cluster Overlap:** Multiple clusters may belong to one gene, expected to consolidate as genome sequencing improves.

## Access to Protein Databases

- **Purpose:** Focus on **retrieving protein sequences**.
- **NCBI Protein Database:** Includes **translated GenBank coding regions + external sources** (UniProt, PIR, SWISS-PROT, PRF, PDB).
- **EBI Access:** Offers **protein data** via similar **major databases**.

## UniProt Overview

- **Definition:** Universal Protein Resource (**UniProt**) – **comprehensive, centralized protein catalog** (est. 2002).
- **Collaboration:** Combines **three** key databases:
  - **Swiss-Prot:** Expert-curated, best-annotated protein data.
  - **TrEMBL:** Automated annotations for proteins not in Swiss-Prot, driven by genome sequencing.
  - **PIR:** Expert-curated Protein Sequence Database.

## UniProt Structure

- **Three Layers:**
  - **UniProtKB:** Central database with:
    - UniProtKB/Swiss-Prot (manual annotations).
    - UniProtKB/TrEMBL (computational annotations).
  - **UniRef:** Nonredundant clusters (**50%**, **90%**, or **100%** identity) based on UniProtKB.
  - **UniParc:** Stable, nonredundant archive from diverse sources (e.g., RefSeq, Ensembl, patents).

## Accessing UniProt

- **Access Points:** UniProt website, EBI, or ExPASy.
- **Example:** Search for beta-globin yields dozens of results.
- **Limitation:** RefSeq accessions **are not displayed**, making prototype sequence identification unclear.



## Central Bioinformatics Resources

- **Focus:** Overview of **NCBI** and **EBI** as **key bioinformatics hubs**.
- **Context:** Follows discussion on DNA, RNA, and protein data in centralized databases.
- **Relation:** DNA repositories (NCBI, EBI, DDBJ) outlined in the previous Figure

## Introduction to NCBI

- **Mission:** **Creates** databases, conducts **computational biology** research, **develops tools**, and shares biomedical info.
- **Key Resources:**
  - **PubMed:** Access to 24M+ citations in MEDLINE and linked journals.
  - **Entrez:** Integrates literature, DNA/protein sequences, 3D structures, and genomes (PubMed included).

## NCBI Tools and Databases (Part 1)

- **BLAST:** Sequence similarity search tool for nucleotide/protein databases
- **OMIM:** Catalog of human genes and genetic disorders with PubMed and sequence links
- **Books:** ~200 searchable online books linked to PubMed (see recommended reading).

## NCBI Tools and Databases (Part 2)

- **Taxonomy:** Browser for archaea, bacteria, eukaryotes, and viruses with genetic codes and molecular data ([LINK](#)).
- **Structure:** Molecular Modelling Database (**MMDB**) with 3D structures from **PDB**, tools like **Cn3D**, **PDBeast**, and **VAST** for **visualization** and **comparison**.

The **entry** for **Homo sapiens** at the **NCBI Taxonomy Browser** displays **information** about the **genus** and **species** as well as a **variety** of **links** to Entrez records.

A list of **proteins**, **genes**, **DNA sequences**, **structures**, or other data types **restricted** to this **organism** can be obtained by following these links. This can be a **useful strategy** to find a **protein** or **gene** from a particular organism (e.g., a species or subspecies of interest), **excluding** data from all other species.

The screenshot shows the NCBI Taxonomy Browser interface for the organism *Homo sapiens*. The search bar at the top contains "Homo sapiens" and the search criteria are set to "as complete name". The "Display" level is set to "0" and the filter is "none".

**Homo sapiens**

*Taxonomy ID:* 9606  
*Genbank common name:* **human**  
*Inherited blast name:* **primates**  
*Rank:* species  
*Genetic code:* [Translation table 1 \(Standard\)](#)  
*Mitochondrial genetic code:* [Translation table 2 \(Vertebrate Mitochondrial\)](#)  
*Other names:*  
common name: **man**  
authority: **Homo sapiens Linnaeus, 1758**

[Lineage\( full \)](#)  
[cellular organisms](#); [Eukaryota](#); [Opisthokonta](#); [Metazoa](#); [Eumetazoa](#); [Bilateria](#); [Deuterostomia](#); [Chordata](#); [Craniata](#); [Vertebrata](#); [Gnathostomata](#); [Teleostomi](#); [Euteleostomi](#); [Sarcopterygii](#); [Tetrapoda](#); [Amniota](#); [Mammalia](#); [Theria](#); [Eutheria](#); [Euarchontoglires](#); [Primates](#); [Haplorrhini](#); [Simiiformes](#); [Catarrhini](#); [Hominoidea](#); [Hominidae](#); [Homininae](#); [Homo](#)

Entrez records		
Database name	Subtree links	Direct links
Nucleotide	<a href="#">10,217,570</a>	<a href="#">10,217,541</a>
Nucleotide EST	<a href="#">8,704,803</a>	<a href="#">8,704,803</a>
Nucleotide GSS	<a href="#">1,729,196</a>	<a href="#">1,727,870</a>
Protein	<a href="#">696,378</a>	<a href="#">696,243</a>
Structure	<a href="#">20,041</a>	<a href="#">20,041</a>
Genome	<a href="#">1</a>	<a href="#">1</a>
Popset	<a href="#">22,687</a>	<a href="#">22,687</a>
SNP	<a href="#">63,228,028</a>	<a href="#">63,228,028</a>
Domains	<a href="#">12</a>	<a href="#">12</a>
GEO Datasets	<a href="#">475,213</a>	<a href="#">475,213</a>
UniGene	<a href="#">130,045</a>	<a href="#">130,045</a>
UniSTS	<a href="#">328,844</a>	<a href="#">328,844</a>
PubMed Central	<a href="#">11,154</a>	<a href="#">11,148</a>
Gene	<a href="#">43,470</a>	<a href="#">43,433</a>
HomoloGene	<a href="#">18,473</a>	<a href="#">18,473</a>
SRA Experiments	<a href="#">53,471</a>	<a href="#">53,469</a>
Probe	<a href="#">24,258,933</a>	<a href="#">24,258,933</a>
Assembly	<a href="#">25</a>	<a href="#">25</a>
Bio Project	<a href="#">13,443</a>	<a href="#">13,442</a>
Bio Sample	<a href="#">812,246</a>	<a href="#">812,243</a>
Bio Systems	<a href="#">2,518</a>	<a href="#">2,518</a>
dbVar	<a href="#">2,517,546</a>	<a href="#">2,517,546</a>
Epigenomics	<a href="#">4,186</a>	<a href="#">4,186</a>
GEO Profiles	<a href="#">27,034,750</a>	<a href="#">27,034,750</a>
Protein Clusters	<a href="#">13</a>	<a href="#">13</a>
Taxonomy	<a href="#">3</a>	<a href="#">1</a>

## The European Bioinformatics Institute (EBI)

- **Overview:** [EBI](#) as a complementary, independent bioinformatics hub to NCBI.
- **Mission:** Similar scope to NCBI, offering unique data organization, analysis, and display.

## EBI Core Databases

- **Six Key Databases:**
  - **EMBL-Bank:** DNA/RNA sequence repository (complements GenBank, DDBJ).
  - **Swiss-Prot:** Expert-annotated protein database.
  - **TrEMBL:** Automated protein annotations.
  - **MSD:** Protein structure database.
  - **Ensembl:** Major genome browser.
  - **ArrayExpress:** Gene expression repository (with NCBI's GEO)

## NCBI vs. EBI Integration

- **Approach:** Both sites use similar raw data but differ in data handling and visualization.
- **Utility:** Valuable to explore both for comprehensive insights (e.g., gene structure/function).
- **Integration:** Increasing linkage between NCBI and EBI for seamless data access.

## Ensembl Project

- **Origin:** Launched in 1999 for human genome annotation.
- **Scope:** Now covers 70+ vertebrate species.
- **Expansion:** Related projects include hundreds of species (insects to bacteria).

Ensembl is a joint project of the EBI and WTSI ([http:// www.ensembl.org](http://www.ensembl.org)).

### Related Ensembl projects include:

Metazoans (<http://metazoa.ensembl.org/>)

Plants (<http://plants.ensembl.org/>)

Fungi (<http://fungi.ensembl.org/>)

Protists (<http://protists.ensembl.org/>)

Bacteria (<http://bacteria.ensembl.org/>)

## Access to Information: Accession Numbers to Label and Identify Sequences

### Access to Information: Accession Numbers

- **Purpose:** Essential for identifying and extracting gene/protein data from databases.
- **Definition:** Strings (4–12 characters) tagging DNA, protein, or other molecular records.
- **Formats:** Vary by database (see the BOX, next page); indicate data type (nucleotide/protein).
- **Challenges:** Thousands of accession numbers per molecule (e.g., beta globin ESTs); quality varies (full vs. partial, errors, SNPs, splice variants).

### Sequence Identifiers

- **GenInfo (GI) Numbers:** Unique, consecutive numbers per sequence (e.g., NM\_000518.4 → GI:28302128).
- **Versioning:** Accession suffixes (e.g., .4) indicate updates; older versions have different GIs.

## BOX 2.2 TYPES OF ACCESSION NUMBERS

Type of Record	Sample Accession Format
GenBank/EMBL/DDBJ nucleotide sequence records	One letter followed by five digits (e.g., X02775); two letters followed by six digits (e.g., AF025334).
GenPept sequence records (which contain the amino acid translations from GenBank/EMBL/DDBJ records that have a coding region feature annotated on them)	Three letters and five digits (e.g., AAA12345).
Protein sequence records from SwissProt and PIR	Usually one letter and five digits (e.g., P12345). SwissProt numbers may also be a mixture of numbers and letters.
Protein sequence records from the Protein Research Foundation	A series of digits (often six or seven) followed by a letter (e.g., 1901178A).
RefSeq nucleotide sequence records	Two letters, an underscore bar, and six or more digits (e.g., mRNA records (NM_*): NM_006744; genomic DNA contigs (NT_*): NT_008769).
RefSeq protein sequence records	Two letters (NP), an underscore bar, and six or more digits (e.g., NP_006735).
Protein structure records	PDB accessions generally contain one digit followed by three letters (e.g., 1TUP). They may contain other mixtures of numbers and letters (or numbers only). MMDB ID numbers generally contain four digits (e.g., 3973.)

Many accession numbers include a suffix (e.g., .1 in NP\_006735.1), indicating a version number.

## The Reference Sequence (RefSeq) Project

- **Goal:** Provide the best representative sequence for each nonmutated gene transcript/protein.
- **Nonredundancy:** Reduces GenBank's redundancy (e.g., one RefSeq vs. hundreds of GenBank entries per gene).
- **Splice Variants:** Multiple RefSeq entries for genes with variants (e.g., myoglobin: NM\_005368.2, NM\_203377.1, NM\_203378.1).
- **Curation:** Managed by NCBI, with status levels (predicted, provisional, reviewed).
- **Formats:** Recognizable (e.g., NP\_000509 for protein, NM\_006744 for mRNA; **Table 2.7, Table 2.8**).



**Table 2.7** Formats of accession numbers for RefSeq entries. There are currently 22 different RefSeq accession formats. The methods include expert manual curation, automated curation, or a combination. Abbreviations: BAC, bacterial artificial chromosome; WGS, whole-genome shotgun

Molecule	Accession format	Genome
Complete genome	NC_123456	Complete genomic molecules, including genomes, chromosomes, organelles, and plasmids
Genomic DNA	NW_123456 or NW_123456789	Intermediate genomic assemblies
Genomic DNA	NZ_ABCD12345678	Collection of whole-genome shotgun sequence data
Genomic DNA	NT_123456	Intermediate genomic assemblies (BAC and/or WGS sequence data)
mRNA	NM_123456 or NM_123456789	Transcript products; mature mRNA protein-coding transcripts
Protein	NP_123456 or NM_123456789	Protein products (primarily full-length)
RNA	NR_123456	Noncoding transcripts (e.g., structural RNAs, transcribed pseudogenes)

**Table 2.8** RefSeq accession numbers corresponding to human beta globin. Adapted from <http://www.ncbi.nlm.nih.gov/refseq/about/>

Category	Accession	Size	Description
DNA	NC_000011.9	135,006,516 bp	Genomic contig
DNA	NM_000518.4	626 bp	DNA corresponding to mRNA
DNA	NG_000007.3	81,706 bp	Genomic reference
protein	NP_000509.1	147 amino acids	Protein

## RefSeq Limitations & Solutions

- **Issue:** Version number changes (e.g., NM\_000518.3 to .4) cause ambiguity in variant reporting.
- **Locus Reference Genomic (LRG):** Stable reference sequences without version numbers, independent of genome updates.
- **RefSeqGene:** Expanded RefSeq to address versioning issues.

## Consensus Coding Sequence (CCDS) Project

- **Purpose:** Define a core set of high-quality protein-coding sequences.
- **Collaboration:** EBI, NCBI, Wellcome Trust Sanger Institute, UCSC.
- **Scope:** Limited to human/mouse genomes; more curated than RefSeq.
- **Strength:** “Gold standard” annotations with extensive expert manual review.

## Vertebrate Genome Annotation (VEGA) Project

- **Goal:** High-quality, manual annotation of human, mouse, and selected vertebrate genomes.
- **Features for HBB Search:**
  - **Transcript View:** cDNA, coding sequences, protein domains.
  - **Gene View:** Orthologs, alternative alleles.

## Access to Information via Gene Resource at NCBI

### Access to Information via NCBI Gene

- **Purpose:** Navigate DNA/protein sequences using interconnected databases.
- **NCBI Gene:** Major portal (formerly Entrez Gene/LocusLink) for curated genetic loci data.
- **Features:** Official nomenclature, aliases, sequence accessions, phenotypes, EC/OMIM numbers, UniGene clusters, HomoloGene, map locations, and external links.

### Searching NCBI Gene (Example: Beta Globin)

- **Search Process:** Use NCBI Gene to search for human beta globin ([LINK](#)).
  - Restrict searches by organism using “limits” tab.
  - Access related databases via “Links” button.
- **Results ([LINK](#)):**
  - Table of contents for beta globin entry.
  - Links to NCBI/external databases (e.g., Ensembl, UCSC).
  - Official symbol (HBB), name, gene structure, function, RefSeq/GenBank accession numbers.

## NCBI Protein Records

- **Default Display:** Standard NCBI Protein record for beta globin ([LINK](#)), ([LINK2](#)).
- **Formats:** Switch to FASTA format for sequences ([LINK](#)).
- **Coding Sequence (CDS):** Access nucleotides encoding protein (start: ATG, stop: TAG/TAA/TGA) for alignment/phylogeny.

## NCBI Gene vs. Nucleotide/Protein Resources

- **Direct Search:** Use NCBI Nucleotide/Protein for specific sequences, with filters (organism, RefSeq).
- **Advantages of NCBI Gene:**
  - Official gene names and chromosomal locations.
  - Comprehensive RefSeq DNA/protein variant list.

## Comparison of NCBI Gene and UniGene

- **UniGene Overview:** One cluster per gene (e.g., HBB → Hs.523443) with GenBank ESTs, mapping, homologies, and expression data.
- **Common Features:** Both link to OMIM, homologs, mapping, RefSeq.
- **Differences:**
  - **UniGene:** Detailed expression data (cDNA library regions).
  - **UniGene:** Lists ESTs for detailed study.
  - **Gene:** More stable, less collapsible than UniGene.
  - **Gene:** Fewer, but more curated entries.

## NCBI Gene and HomoloGene

- **HomoloGene Overview:** Groups annotated proteins from sequenced eukaryotic genomes using BLASTP.
- **Search Example:** “Hemoglobin” yields matches for myoglobin, alpha/beta globin.
- **Beta Globin Results:**
  - Lists proteins with RefSeq numbers (human, chimpanzee, dog, mouse, chicken).
  - Summarizes pairwise alignment scores.
  - Downloads sequences (DNA, mRNA, protein) and displays protein alignments.

## Command-Line Access to NCBI Data

- **Overview:** Alternative to web browsers; use command-line tools like Entrez Direct (EDirect) for NCBI's Entrez databases.
- **Applications:** BLAST, sequence alignment, phylogeny, DNA/RNA analysis, genome comparisons, and annotation.

## Benefits of Command-Line Software

- **Operating Systems:** Windows, Mac OS, Unix/Linux; Linux preferred for bioinformatics.
- **Linux Advantages:**
  - Free, customizable, flexible.
  - Handles large datasets (unlike Excel's row limits and auto-reformatting).
- **Access:** Use Linux locally or via Secure Shell (SSH) clients like PuTTY on Windows.

## Introduction to EDirect

- Definition: Suite of Perl scripts for Unix (Linux, Mac OS, Cygwin on Windows).
- Installation: Simple, creates `edirect` folder in the home directory ([Box 2.4](#)).
- Functions:
  - Navigation: `esearch`, `elink`, `efilter`.
  - Retrieval: `esummary`, `efetch`.
  - Extraction: `xtract` (XML fields).
  - Others: `epost` (upload IDs/accessions).



## Direct Example 1 – PubMed Search

- **Task:** Search PubMed for “Pevsner J AND GNAQ”, fetch summaries, display or save to file (example1.out).

- **Command:**

```
$ esearch -db pubmed -query "pevsner j AND gnaq" | efetch -format docsum > example1.txt
```

- **Utilities:** Use `less` for paginated view, `head` for first lines, `man/info` for help.

## EDirect Example 2 – Query Counts

- **Task:** Search PubMed for “Pevsner J” and display result counts without fetching.

- **Command:**

```
$ esearch -db pubmed -query "pevsner j" | less
```

- **Results:** Shows counts (e.g., Pevsner J: 99 articles, hemoglobin: ~155,000, bioinformatics: ~131,000).

### EDirect Example 3 – Author Rankings

- **Task:** Find top authors in bioinformatics software using MeSH and title/abstract.

- **Command:**

```
$ esearch -db pubmed -query "bioinformatics [MAJR] AND software [TIAB]" |  
efetch -format xml | xtract ... | sort-uniq-count-rank.
```

- **Output:** Lists top authors (e.g., Aebersold R: 29 articles, Deutsch EW: 22).

### EDirect Example 4 – Protein Search

- **Task:** Search Protein database for “hemoglobin”, fetch FASTA format, view first 6 lines.

- **Command:**

```
$ esearch -db protein -query "hemoglobin" | efetch -format fasta | head -6
```

- **Flexibility:** Can search any Entrez database.

## EDirect Example 5 – Related Data

- **Task:** Find PubMed articles on “hemoglobin”, link to related articles, then to proteins.

- **Command:**

```
$ esearch -db pubmed -query "hemoglobin" | elink -related | elink -target  
protein
```

## EDirect Example 6 – Chromosome Data

- **Task:** List genes on human chromosome 16 with start/stop positions, save to file.

- **Command:**

```
$ esearch -db gene -query "16[chr] AND human[orgn] ..." | esummary | xtract  
... > example6.out
```

- **View:** Use `head -5` to see first five lines of output.

## EDirect Example 7 – Taxonomy Data

- **Task:** Find taxonomic family and BLAST division for model organisms (e.g., *E. coli*, *H. sapiens*).
- **Steps:**
  - Create `organisms.txt` listing organisms.
  - Write `taxonomy.sh` script to query taxonomy database.
  - Make script executable: `$ chmod ugo+rx taxonomy.sh`
  - Run: `$ cat organisms.txt | ./taxonomy.sh`
- **Output:** Lists organism, family, division (e.g., *E. coli* → Enterobacteriaceae, enterobacteria).

## Introduction to Genome Browsers

- **Definition:** Databases with graphical interfaces displaying sequence and annotation data along chromosomes
- **Scope:** Covers viral, bacterial, archaeal, and eukaryotic genomes (Chapters 16–20)
- **Importance:** Essential for organizing genomic information
- **Focus:** Three main browsers – [Ensembl](#), [UCSC](#), NCBI

## Understanding Genome Builds

- **Definition:** Assembly of DNA sequences to reflect chromosome arrangement, with annotations (e.g., gene start/stop, repeats).
- **Frequency:** Released every few years; newer builds may have less annotation than older ones.
- **Management:** Genome Reference Consortium (GRC) oversees human, mouse, zebrafish builds (e.g., GRCh38/hg38, 2013).
- **Challenges:**
  - Coordinates (e.g., HBB on chr11: different in GRCh37 vs. GRCh36).
  - Gaps in repetitive regions (e.g., telomeres, centromeres).
  - Representation of structural variants, polymorphisms.
  - Error rates (e.g., 1 in 100,000 bases → 30,000 errors in 3B bases).

## Genome Build Complexities

- **Example Issue:** Major histocompatibility complex (MHC) lacks a single consensus due to diversity.
- **Solutions:**
  - Primary/alternate loci (e.g., HLA-DRB3 on alternate).
  - Patches (e.g., GRCh37.p10) correct errors, add alternate loci, minimize coordinate changes.

## UCSC Genome Browser

- **Overview:** Supports 36+ vertebrate/invertebrate genomes; widely used for human/mouse.
- **Features:**
  - Graphical views at multiple resolutions (base pairs to chromosomes).
  - Hundreds of annotation tracks (e.g., genes, expression, variation).
- **Example Use ([LINK](#)):**
  - Search “hbb” → RefSeq entry for beta globin on chr11.
  - Analyze gene, mRNA, protein, regulatory elements.
- **Resources:** BLAT (Chapter 5), Table Browser, variation analysis.

(a) Specifying the genome, assembly, and gene (or region or feature)

group	genome	assembly	position	search term	
Mammal	Human	Feb. 2009 (GRCh37/hg19)	chr21:33,031,597-33,041,570	hbb	<input type="button" value="submit"/>
					HBB (Homo sapiens hemoglobin, beta (HBB), mRNA.) HBBP1 (Homo sapiens hemoglobin, beta pseudogene 1 (HBBP1), mRNA.)
<a href="#">Click here to reset</a> the browser user interface settings to default					
<input type="button" value="track search"/>		<input type="button" value="add custom tracks"/>		<input type="button" value="track hubs"/> <input type="button" value="configure tracks and display"/>	

(b) Selecting a gene

## UCSC Genes

[HBB \(uc001mae.1\) at chr11:5246696-5248301](#) - Homo sapiens hemoglobin, beta (HBB), mRNA.  
[HBD \(uc001maf.1\) at chr11:5254059-5255858](#) - Homo sapiens hemoglobin, delta (HBD), mRNA.  
[RBM17 \(uc010qav.2\) at chr10:6131309-6159422](#) - Homo sapiens RNA binding motif protein 17 (RBM17), transcript variant 2, mRNA.  
[RBM17 \(uc001ijb.3\) at chr10:6130949-6159422](#) - Homo sapiens RNA binding motif protein 17 (RBM17), transcript variant 1, mRNA.  
[HBA1 \(uc002cfx.1\) at chr16:226679-227520](#) - Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA.  
[HBA2 \(uc002cfv.4\) at chr16:222846-223709](#) - Homo sapiens hemoglobin, alpha 2 (HBA2), mRNA.  
[HBBP1 \(uc001mag.3\) at chr11:5263185-5264822](#) - Homo sapiens hemoglobin, beta pseudogene 1 (HBBP1), non-coding RNA.  
[TMEM158 \(uc011baf.2\) at chr3:45265956-45267814](#) - Homo sapiens transmembrane protein 158 (gene/pseudogene) (TMEM158), mRNA.

## RefSeq Genes

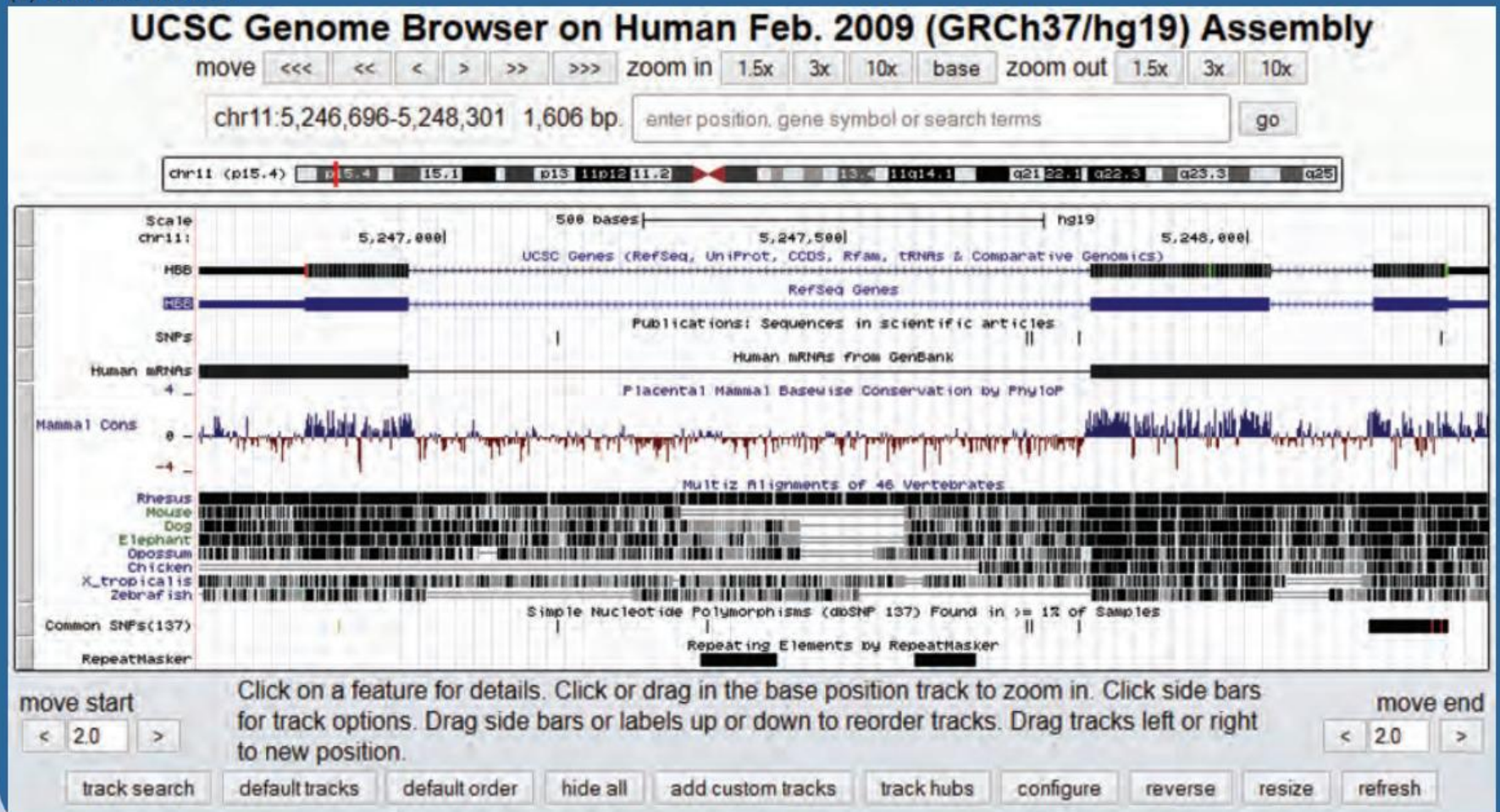
[HBB at chr11:5246696-5248301](#) - (NM\_000518) hemoglobin subunit beta  
[HBBP1 at chr11:5263185-5264822](#) - (NR\_001589)

## Using the UCSC Genome Browser.

- Select from dozens of organisms (mostly vertebrates) and assemblies, then enter a query such as “beta globin” (shown here) or an accession number or chromosomal position.
- By clicking submit, a list of known genes as well as RefSeq genes is displayed.



(c) Genome browser



(c) Following the link to the RefSeq gene for beta globin, a browser window is opened showing 1606 base pairs on human chromosome 11. A series of horizontal tracks are displayed including a list of RefSeq genes and Ensembl gene predictions; exons are displayed as thick bars, and arrows indicate the direction of transcription (from right to left, toward the telomere or end of the short arm of chromosome 11).

## Ensembl Genome Browser

- **Overview:** Comprehensive for eukaryotic genomes; comparable to UCSC.
- **Goals:** Auto-analyze, annotate, and present genomic data (Chapter 15).

### Example Use:

- Search “hbb” → Links to beta globin protein/gene, DNA sequence, external databases.
- **Features:** Stable identifiers (Table).

**Table** Ensembl stable identifiers. For **human entries**, the prefix is **ENS**, while other common species prefixes include **ENSBTA** (cow **Bos taurus**), **ENSMUS** (mouse **Mus musculus**), **ENSRNO** (rat **Rattus norvegicus**), and **FB** (fruit fly **Drosophila melanogaster**).

Feature prefix	Definition	Human beta globin example
E	exon	ENSE00001829867
FM	protein family	ENSFM002500000000136
G	gene	ENSG00000244734
GT	gene tree	ENSGT00650000093060
P	protein	ENSP00000333994
R	regulatory feature	ENSR00000557622
T	transcript	ENST00000335295

*Source:* Ensembl Release 76; Flicek *et al.* (2014). Reproduced with permission from Ensembl.

## NCBI Map Viewer

- **Overview:** Chromosomal maps (physical/genetic) for metazoans, fungi, plants (Chapter 20).
- **Query Types:** Text-based (e.g., “beta globin”) or sequence-based (BLAST, Chapter 4).
- **Detail Levels:**
  - Organism home page.
  - Genome view (chromosome ideograms).
  - Map view (variable resolution).
  - Sequence view (gene annotations, sequences).
- **Access:** Via NCBI Gene HBB entry ([LINK](#)), with tools/configure options.

## Accessing Sequence Data – Challenges

- **Purpose:** Explore practical challenges in accessing data for specific genes/proteins.
- **Examples:**
  - Human histones: Complex due to biological diversity.
  - HIV-1 pol protein: Complex due to extensive variants.

## Histones – Overview

- **Background:** Small nuclear proteins (12–20 kDa) interacting with DNA.
- **Types:** Five major subtypes (H2A, H2B, H3, H4 core; H1 linker) + variants.
- **Challenge:** 470,000 histone entries in NCBI Protein (April 2015).

## Histones – Narrowing the Search

- **Step 1:** Species Filter: Use NCBI Protein/Taxonomy Browser with Homo sapiens ID (9606) → >8000 human histones, >2000 with RefSeq.
- **Step 2:** Refine Query: Exclude deacetylases/acetyltransferases → >1700 RefSeq entries.
  - **Query:**  

```
txid9606[Organism:exp] AND histone[All Fields] NOT deacetylase NOT acetyltransferase
```

## Histones – Search Strategies

- **Options:**
  - **NCBI Gene:** Summaries via RefSeq (similar to globin example, Fig. 2.9).
  - **Random Selection:** Risk of unrepresentative histone.
  - **Specialized Databases:** Histone Sequence Database → 113 human histone genes, including 56 on chromosome 6p.
  - **Protein Family Databases:** Pfam, InterPro for family descriptions (Chapters 6, 12).

## HIV-1 pol – Overview

- **Background:** Reverse transcriptase gene (pol) in HIV-1, an RNA-dependent DNA polymerase.
- **Challenge:** >500,000 nucleotide entries for “hiv-1” in NCBI, >3000 with RefSeq.

## HIV-1 pol – Narrowing the Search

- **Step 1:** General Search: “hiv-1” in NCBI → Too many hits due to re-sequencing and cross-species references.
- **Step 2:** Species Filter: Restrict to HIV → One RefSeq entry (NC\_001802.1, 9181 bases, 9 genes including gag-pol).
- **Value of RefSeq:** Provides a common reference sequence amidst thousands of variants.

## HIV-1 pol – Alternative Strategies

- **Options:**

- From Entrez results, access genome/assembly/taxonomy pages → Single NCBI Genome record.
- Genome annotation report → Table of 9 genes/proteins, including:
  - gag-pol precursor (NP\_057849.4, 1435 aa).
  - Mature pol protein (NP\_789740.1, 995 aa).

## HIV-1 pol – Database Limitations

- **Inappropriate Databases:**

- UniGene: Excludes viral records.
- OMIM: Limited to human entries (e.g., HIV susceptibility genes).

- **Indirect Links:** UniGene/OMIM may link to related eukaryotic reverse transcriptases.

## Large-Scale Data Queries – Overview

- **Focus:** Shift from single gene (e.g., HBB) to large-scale queries of genes, proteins, or genomic elements.
- **Example Questions:**
  - What are all human globin genes?
  - Which chromosomes host them?
  - How many exons/repeats on chromosome 11?
- **Challenge:** Manual collection is inefficient; need tools for genome-wide data.

## Tools for Large-Scale Queries

- **Resources:**
  - **Ensembl:** BioMart for querying multiple databases.
  - **UCSC:** Genome Browser and Table Browser for tabular data.
- **Access:** Both integrated via Galaxy for enhanced analysis.
- **Comparison:** Complementary tools with different formats but overlapping data.



## The BioMart Project

- **Purpose:** Easy access to vast data across databases.
- **Principles:**
  - Data Agnostic Modeling: Imports diverse datasets, uses relational schema to link queries (e.g., gene names) to annotations.
  - Data Federation: Integrates distributed databases into a single virtual database.
- **Coverage:** Includes RefSeq, Ensembl, HGNC, LRG, UniProt, CCDS, etc.
- **Usage:** Explored via Computer Labs 2.4–2.6 and R package biomaRt (Chapter 8).

## UCSC Table Browser

- **Purpose:** Provides tabular data matching UCSC Genome Browser visualizations.
- **Example Use** (Fig. 2.13):
  - Set genome to human (GRCh37/hg19), track to RefSeq genes.
  - Define region (e.g., whole genome, ENCODE, custom).
  - Choose output format (e.g., BED) and send to Galaxy/Great.
- **Features:** Download/view tables, query, summarize output size.

## Custom Tracks – BED Files

- **Purpose:**
  - Select specific data (e.g., microRNAs near exons).
  - Upload experimental data (e.g., microarray results).
- **BED Format** (Fig. 2.13c):
  - Required Fields: Chromosome, start, end positions.
  - Optional Fields: Name, score, strand, thickStart/thickEnd, itemRgb, blockCount/blockSizes/blockStarts.
- **Applications:** Used in next-generation sequencing (Chapter 9) with BEDTools for overlap analysis.
- **Formats:** Many supported by Ensembl/UCSC (Table 2.10).

**Table 2.10** File formats for custom tracks used at Ensembl and/or UCSC. Two definitions of GTF (from Ensembl and UCSC) are given

File Format	Definition	Typical file size
BAM	Browser extensible data	Any size; often millions of rows
BED		Any size; often dozens to thousands or millions of rows
BedGraph		Any size
bigBed		Any size
GFF/GTF	General feature format, General transfer format Gene transfer format	Any size
MAF	Wiggle	Any size
PSL		Any size
WIG		Any size
BAM		Very large
BigWig	Binary alignment/map	Very large
VCF		Very large
	Variant call format	Very large

## Custom Tracks – Caveats

- **Considerations:**
  - **Chromosome naming:** “11” vs. “chr11” (Fig. 2.13c).
  - **Counting schemes:** Zero-based vs. one-based (Table 2.11, Box 2.5).
    - **Example:** HBB start in UCSC Genome Browser (1-based) = 5,246,696; Table Browser (0-based) = 5,246,695.
- **Importance:** One-nucleotide differences critical for variant analysis.

## Galaxy – Web-Based Analysis Platform

- **Purpose:** Integrates BioMart, UCSC Table Browser, and other tools for high-throughput research.
- **Interface:** Three panels – tools (left), display (center), history (right).
- **Advantages:**
  - Large collection of tools for importing/analyzing high-throughput data.
  - Web-based access to command-line software.
  - Reproducible research via documented, sharable workflows.

## Galaxy – Example Use

- **Process:**
  - Select “Get Data” → UCSC Table Browser → Query “hbb” → Set format to protein sequence → Send to Galaxy.
  - View sequence in history panel, analyze further with hundreds of tools.
- **Applications** (Explored in Book):
  - Pairwise alignment (Chapter 3).
  - Genomic DNA alignments (Chapter 6).
  - Microsatellite extraction (Chapter 8).
  - Next-generation sequencing (Chapter 9): FASTQ import, FASTQC, BAM/VCF analysis.
  - RNA-seq analysis (Chapter 11): Tools like Bowtie, BWA.

## Access to Biomedical Literature – Overview

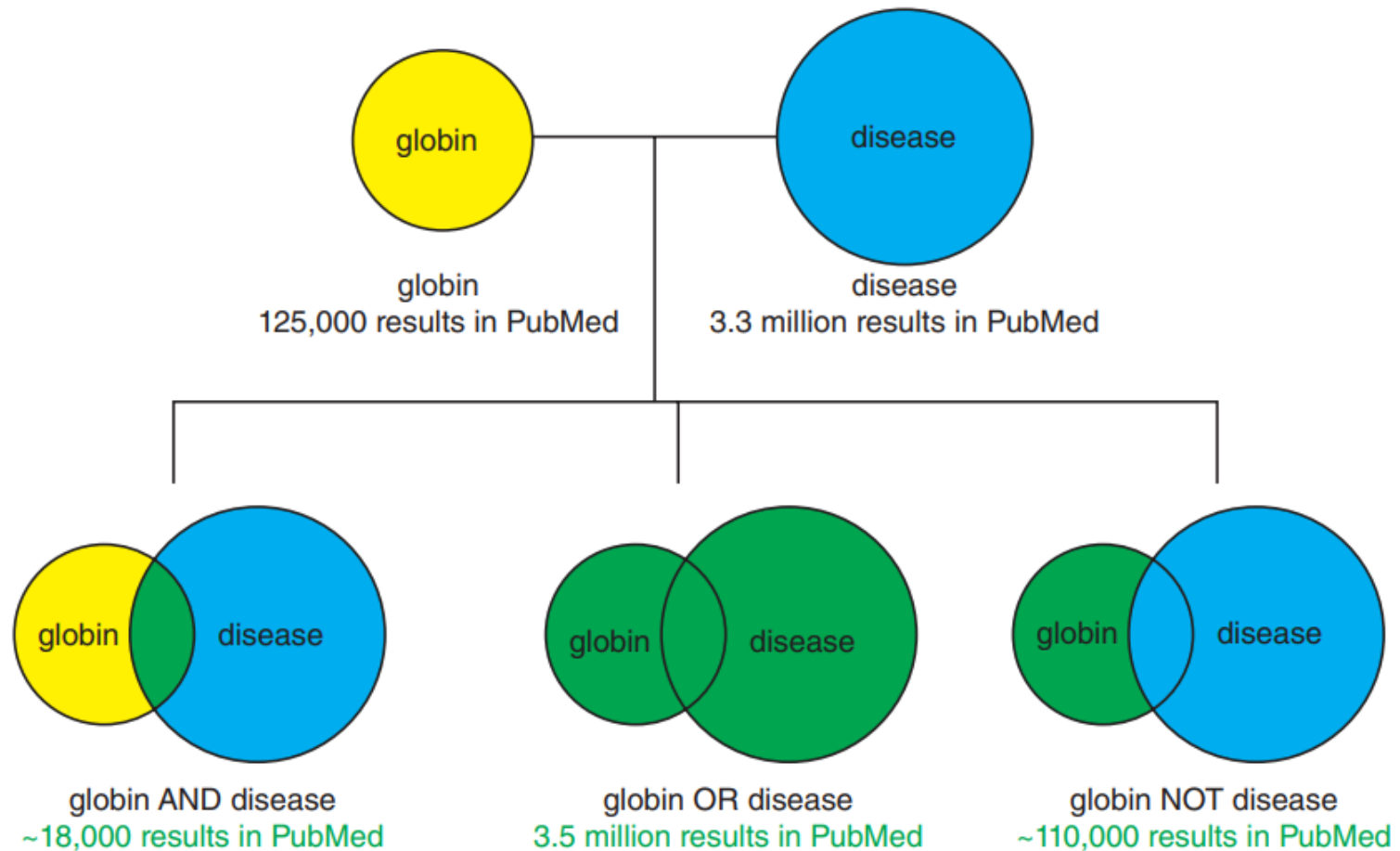
- **National Library of Medicine (NLM):** World's largest medical library.
- **MEDLINE:** Bibliographic database (since 1971) with >24M references from >5600 biomedical journals.
- **PubMed:** Free access to MEDLINE, developed by NCBI, with additional features:
  - Links to full-text articles.
  - Integration with NCBI molecular databases (DNA/protein sequences, genome maps, 3D structures).

## Example of PubMed Search

- **Search Query:** “beta globin” → ~6700 entries.
- **Refinement Tools:**
  - Filters: Restrict to free articles via PubMed Central (sidebar).
  - Boolean Operators: Use AND, OR, NOT to refine searches (Box 2.6).

## Box 2.6 Venn Diagrams of Boolean Operators AND, OR, and NOT

- The **AND** command restricts the search to entries that are **both present** in a **query**.
- The **OR** command allows **either one** or **both** of the **terms** to be present.
- The **NOT** command **excludes query results**.



## Medical Subject Headings (MeSH) Browser

- **Purpose:** Controlled vocabulary thesaurus with >30,000 descriptors to focus/expand searches.
- **Access:** From PubMed or NCBI homepage → Select MeSH → Search “beta globin”.
- **Features:**
  - Suggests related topics (e.g., “beta-Globins”).
  - Structures searches for specific information.

## Strategies for Effective Searches

- **Goals** (Lewitter, 1998; Fielding & Powell, 2002):
  - Balance sensitivity (find relevant articles) and specificity (exclude irrelevant ones).
- **Techniques:**
  - Combine text keywords with MeSH terms for poorly indexed subjects.
  - Use truncations (e.g., “therap\*” → therapy, therapist, therapeutic) to capture variations.



# Primary Databases

Public sequence databases for raw nucleic acid sequence data:

- **GenBank**
  - **European Molecular Biology Laboratory (EMBL) database**
  - **DNA Data Bank of Japan (DDBJ)**
- ✓ All of them are **freely available** on the Internet.
- ✓ Most of the data in the databases are **contributed directly by authors** with a **minimal level of annotation**.
- ✓ **Sequence submission** to either GenBank, EMBL, or DDBJ is a **precondition** for **publication** in most scientific journals **to ensure** the **fundamental molecular data to be made freely available**.
- ✓ These three public databases **closely collaborate** and **exchange new data daily**. They together constitute the **International Nucleotide Sequence Database Collaboration**.

## Public 3-D structure database for macromolecules:

- For the **three-dimensional structures** of **biological macromolecules**, there is only **one centralized database**, the **PDB**.
- This database archives **atomic coordinates of macromolecules** (both proteins and nucleic acids) determined by **x-ray crystallography** and **NMR**.
- It uses a **flat file format** to represent protein name, authors, experimental details, secondary structure, cofactors, and atomic coordinates.

## Secondary Databases

- **Sequence annotation information** in the **primary database** is often **minimal**.
- So, we need **secondary databases, computationally processed sequence information derived from the primary databases**.

Note:

The **amount of computational processing** work **varies greatly** among the **secondary databases**; some are simple archives of **translated sequence data from identified open reading frames in DNA**, whereas others provide **additional annotation and information related to higher levels of information regarding structure and functions**.

## Secondary Databases

### SWISS-PROT

- The **sequence data** are mainly derived from **TrEMBL**, a database of translated nucleic acid sequences stored in the EMBL database.
- The **protein annotation** includes **function**, **domain structure**, **catalytic sites**, **cofactor binding**, **posttranslational modification**, **metabolic pathway information**, **disease association**, and **similarity with other sequences**.
- **Much** of this **information** is obtained from **scientific literature** and **entered** by **database curators**.
- The features such as **very low redundancy** and **high level of integration with other primary and secondary databases** make **SWISS-PROT** **very popular among biologists**.

Note:

A recent effort to **combine SWISS-PROT, TrEMBL, and PIR** led to the creation of the **UniProt database**, which has **larger coverage** than any one of the three databases while at the same time maintaining the original SWISS-PROT feature of **low redundancy, cross-references, and a high quality of annotation**.

The **UniProt Knowledgebase (UniProtKB)** is the **central hub** for the **collection** of **functional information on proteins**, with **accurate, consistent** and **rich annotation**.

### **Mandatory information**

The **amino acid sequence, protein name or description, taxonomic data and citation information**

### **Additional information**

The **widely accepted biological ontologies, classifications and cross-references**, and **clear indications of the quality of annotation** in the form of evidence attribution of experimental and computational data.

### **The UniProt Knowledgebase consists of two sections:**

**UniProtKB/Swiss-Prot** (reviewed, manually annotated): a section containing manually-annotated records with **information extracted from literature and curator-evaluated computational analysis**

**UniProtKB/TrEMBL** (unreviewed, automatically annotated): a section with computationally analyzed records that **await full manual annotation**.

## Where do the protein sequences come from?

More than **95 %** of the **protein sequences** provided by **UniProtKB** are **derived** from the **translation of the coding sequences (CDS)** which have been **submitted** to the **public nucleic acid databases**, the **EMBL-Bank/GenBank/DDBJ databases (INSDC)**. **All these sequences, as well as the related data submitted by the authors, are automatically integrated into UniProtKB/TrEMBL.**

## What are the differences between UniProtKB/Swiss-Prot and UniProtKB/TrEMBL?

**UniProtKB/TrEMBL** (unreviewed) **contains protein sequences** associated with **computationally generated annotation** and **large-scale functional characterization**.

**UniProtKB/Swiss-Prot** (reviewed) is a **high quality manually annotated** and **non-redundant protein sequence database**, which brings together **experimental results, computed features** and **scientific conclusions**.

## What is manual annotation?

Manual annotation consists of a critical review of experimentally proven or computer-predicted data about each protein, including the protein sequences. Data are continuously updated by an expert team of biologists.

## Other secondary databases:

- **Pfam**

- **Blocks**

They contain **aligned protein sequence information** as well as derived **motifs** and **patterns**, which can be used **for classification of protein families** and **inference of protein functions**.

- **DALI** is a **protein secondary structure database** that is **vital for protein structure classification** and **threading analysis** to identify **distant evolutionary relationships among proteins**.



## Specialized Databases

Specialized databases normally serve a **specific research community** or **focus on a particular organism**. The content of these databases may be sequences or other types of information.

Many **genome databases** that are taxonomic specific fall within this category.  
e.g. **Flybase**, **WormBase**, **AceDB**, and **TAIR**

There are specialized databases that **contain original data derived from functional analysis**. For example, **GenBank EST database** and **Microarray Gene Expression Database** at the **European Bioinformatics Institute (EBI)** are some of the gene expression databases available.

# Interconnection between Biological Databases

## In the biological community

- Need for **connection** between the **secondary** and **specialized databases** and **primary databases**
- Need to **get information** from both **primary** and **secondary databases** to **complete** a task
- ✓ The **main barrier** to linking different biological databases: **format incompatibility** between current biological databases

## Solution: using

- Common **Object Request Broker Architecture (COBRA)**, allowing **database programs** at **different locations** to **communicate** in a network through an “**interface broker**” **without** having to ***understand*** each other's **database structure**
- **eXtensible Markup Language (XML)**, helping in **bridging** databases

# PITFALLS OF BIOLOGICAL DATABASES

One of the **problems associated** with **biological databases** is **overreliance** on **sequence information** and **related annotations**, without understanding the reliability of the information.

## Why we must not rely on biological databases?

### ➤ Errors in nucleotide sequences

Sources of errors:

- **Sequencing errors**; these errors cause **frame shifts** that make **whole gene identification difficult** or **protein translation impossible**
- **Contamination with sequences from cloning vectors**

Exceptional care should be taken when dealing with more dated sequences (before the 1990s). Sequence quality has been greatly improved since.

# PITFALLS OF BIOLOGICAL DATABASES

## ➤ Redundancy

### Reasons of redundancy:

- **Repeated submission** of **identical** or **overlapping sequences** by the same or different authors
- **Revision of annotations**
- **Dumping** of expressed sequence tags (EST) data
- **Poor database management**

This makes some **primary databases** excessively **large** and **unwieldy** for information retrieval.

### Solutions:

- Creation a **non-redundant database**, called **RefSeq**, by the **National Center for Biotechnology Information (NCBI)**
- For **protein sequences**, **minimal redundancy** in **SWISS-PROT database** compared to most other databases

# PITFALLS OF BIOLOGICAL DATABASES

## ➤ Erroneous annotations

### Reasons:

- **Different names** for the **same gene**, resulting in multiple entries and confusion about the data
- **The same name** for **unrelated genes**

### Solutions:

- **Reannotation** of **genes** and **proteins** using a set of **common, controlled vocabulary** to describe a gene or protein; A **prominent example** of such systems is **Gene Ontology**
- **Genuine disagreement** between **researchers** in the field
- **Imprudent assignment** of **protein functions** by **sequence submitters**
- **Errors caused** by **omissions** or **mistakes** in **typing**

**Errors** in **annotation** can be particularly **damaging** because the large majority of **new sequences** are **assigned functions** based on **similarity** with **sequences** in the databases that are **already annotated**.

# INFORMATION RETRIEVAL FROM BIOLOGICAL DATABASES

Two types of **retrieval systems** for **biological data**

- **Entrez**
- **Sequence Retrieval Systems (SRS)**

To perform complex queries in a database often requires the use of **Boolean operators**.

- ✓ **AND**: means that the **search result** must **contain both words**
- ✓ **OR**: means to search for **results** containing **either word or both**
- ✓ **NOT**: **excludes** results containing **either one of the words**
- ✓ **( )**: if **multiple** words and relationships are involved, **items contained within parentheses are executed first**.
- ✓ **Quotes** can be used to **specify a phrase**.

Most **search engines** of **public biological databases** use some form of this Boolean logic.

# Entrez (Global Query Cross-Database Search System)

The **NCBI developed** and **maintains Entrez**, a biological database retrieval system.

Entrez is a **federated search engine**, or **web portal**.

As a gateway, Entrez allows **text-based searches** for a **wide variety of data**, including

- **annotated genetic sequence information**
- **structural information**
- **citations and abstracts**
- **full papers**
- **taxonomic data**
- **and so on**

**Key feature of Entrez:** the **ability** to integrate information

This is highly convenient: users **do not have to visit multiple databases** located in disparate places. For example, in a **nucleotide sequence page**, one may find **cross-referencing links** to the **translated protein sequence**, **genome mapping data**, or to the **related PubMed literature information**, and to **protein structures** if available.

# Entrez (Global Query Cross-Database Search System)

Effective use of Entrez requires an **understanding** of the **main features** of the **search engine**.

**database**.

- **Preview/Index**; connect different searches with the **Boolean operators** and uses a **string of logically connected keywords** to perform a new search. The search can also be limited to a particular search field (e.g., gene name or accession number).
- **History**; provide **record** of the **previous searches** so that the user can **review**, **revise**, or **combine** the results of earlier searches.
- **Clipboard**; **store search results** for later viewing for a limited time. To store information in the Clipboard, the “**Send to Clipboard**” function should be used.



# Entrez (Global Query Cross-Database Search System)

## PubMed

One of the **databases** accessible from Entrez is a **biomedical literature database** known as **PubMed**, which contains abstracts and in some cases the full text articles from nearly **4,000** journals.

An **important feature** of PubMed is the **retrieval of information** based on **medical subject headings (MeSH) terms**.

The **MeSH system** consists of a **collection** of more than **20,000 controlled and standardized vocabulary terms** used for indexing articles. In other words, it is a **thesaurus** that helps **convert search keywords** into **standardized terms** to describe a concept.

By doing so, it **allows “smart” searches** in which a **group of accepted synonyms** are **employed** so that the **user** not only **gets exact matches**, but also **related matches** on the same topic that otherwise might have been missed.

Another way to broaden the retrieval is by using the **“Related Articles” option**.

By using this feature, **articles** on the same topic that were **missed in the original search** can be **retrieved**.

# OMIM

Online Mendelian Inheritance in Man (**OMIM**) is a non-sequence-based database of human disease genes and human genetic disorders.

Each **entry** in **OMIM** contains **summary information** about **a particular disease** as well as **genes related to the disease**. The text contains **numerous hyperlinks** to **literature citations**, **primary sequence records**, as well as **chromosome loci** of the disease genes. The database can serve as an excellent starting point to study genes related to a disease.

## Taxonomy Database

A **taxonomy database** that **contains** the **names** and **taxonomic positions** of over **100,000** **organisms** with **at least one nucleotide** or **protein sequence** represented in the **GenBank** database.

The taxonomy database has a **hierarchical classification** scheme. The **root level** is **Archaea**, **Eubacteria**, and **Eukaryota**. The database **allows** the **taxonomic tree** for a **particular organism** to be displayed. The **tree** is based on **molecular phylogenetic data**, namely, the **small ribosomal RNA data**.