Databases

Objective of this lecture

The **goal** of this lecture is to **introduce the databases** that **store** these data and strategies to **extract** information from them.

What is a Gene and Genome

- A **genome** is the complete set of DNA.
- A **gene** is a section of DNA that makes up the building plans for physical traits.

  Genes vary a lot in size:

  - Humans: average 3000 bp.
  - Largest 2.4 million bp.
  - A **base pair** (**bp**) is a unit consisting of two nucleobases bound to each other by hydrogen bonds.
- Genes are separated by sequences with unknown function.

- Genomes of 1000s of organisms have been completely sequenced (**300,000 species!**)
- Publicly available databanks now contain quadrillions (>10^{15}) of nucleotides of DNA sequence data, soon to be quintillions (>10^{18} bases).

What is a DNA Sequence?

- The DNA double helix is made up of a series of chemical bases stung along a sugar backbone.
- There are 4 bases usually represented by the letters A, T, C and G.
- The linear sequence in which these bases occur determines all the instructions for building an organism.

Features of DNA Sequence analysis

- **Detecting open reading frames (ORF):**
  - Initial codon: ATG
  - Stop codon: TGA, TAA, TAG
- **Features may be used as indicators of potential protein coding regions in DNA:**
  - Sufficient ORF length
  - Recognition of flanking Kozak consensus sequence ((gcc) gccRccAUGG)
  - Patterns of codon usage
  - A general preference for G/C over A/T in third base (wobble) position of a codon
  - Ribosome binding sites
- **Alignment with homologous protein sequences**

An example of DNA sequence

ACCACTTTCACAATCTGCTAGCAAAGGTTATGCAGCGCGTGAACATGATCATGGCAGAATCACCAGGCCT
CATCACCATCTGCCTTTTAGGATATCTACTCAGTGCTGAATGTACAGTTTTTCTTGATCATGAAAACGCC
AACAAAATTCTGAATCGGCCAAAGAGGTATAATTCAGGTAAATTGGAAGAGTTTGTTCAAGGGAACCTTG
AGAGAGAATGTATGGAAGAAAAGTGTAGTTTTGAAGAAGCACGAGAAGTTTTTGAAAACACTGAAAGAAC
AACTGAATTTTGGAAGCAGTATGTTGATGGAGATCAGTGTGAGTCCAATCCATGTTTAAATGGCGGCAGT
TGCAAGGATGACATTAATTCCTATGAATGTTGGTGTCCCTTTGGATTTGAAGGAAAGAACTGTGAATTAG
ATGTAACATGTAACATTAAGAATGGCAGATGCGAGCAGTTTTGTAAAAATAGTGCTGATAACAAGGTGGT
TTGCTCCTGTACTGAGGGATATCGACTTGCAGAAAACCAGAAGTCCTGTGAACCAGCAGTGCCATTTCCA
TGTGGAAGAGTTTCTGTTTCACAAACTTCTAAGCTCACCCGTGCTGAGACTGTTTTTCCTGATGTGGACT
ATGTAAATTCTACTGAAGCTGAAACCATTTTGGATAACATCACTCAAAGCACCCAATCATTTAATGACTT
CACTCGGGTTGTTGGTGGAGAAGATGCCAAACCAGGTCAATTCCCTTGGCAGGTTGTTTTGAATGGTAAA
GTTGATGCATTCTGTGGAGGCTCTATCGTTAATGAAAAATGGATTGTAACTGCTGCCCACTGTGTTGAAA
CTGGTGTTAAAATTACAGTTGTCGCAGGTGAACATAATATTGAGGAGACAGAACATACAGAGCAAAAGCG
AAATGTGATTCGAATTATTCCTCACCACAACTACAATGCAGCTATTAATAAGTACAACCATGACATTGCC
CTTCTGGAACTGGACGAACCCTTAGTGCTAAACAGCTACGTTACACCTATTTGCATTGCTGACAAGGAAT
ACACGAACATCTTCCTCAAATTTGGATCTGGCTATGTAAGTGGCTGGGGAAGAGTCTTCCACAAAGGGAG
ATCAGCTTTAGTTCTTCAGTACCTTAGAGTTCCACTTGTTGACCGAGCCACATGTCTTCGATCTACAAAG
TTCACCATCTATAACAACATGTTCTGTGCTGGCTTCCATGAAGGAGGTAGAGATTCATGTCAAGGAGATA
GTGGGGGACCCCATGTTACTGAAGTGGAAGGGACCAGTTTCTTAACTGGAATTATTAGCTGGGGTGAAGA
GTGTGCAATGAAAGGCAAATATGGAATATATACCAAGGTATCCCGGTATGTCAACTGGATTAAGGAAAAA
ACAAAGCTCACTTAATGAAAGATGGATTTCCAAGGTTAATTCATTGGAATTGAAAATTAACAGGGCCTCT
CACTAACTAATCACTTTCCCATCTTTTGTTAGATTTGAATATATACATTCTATGATCATTGCTTTTTCTC
TTTACAGGGGAGAATTTCATATTTTACCTGAGCAAATTGATTAGAAAATGGAACCACTAGAGGAATATAA
TGTGTTAGGAAATTACAGTCATTTCTAAGGGCCCAGCCCTTGACAAAATTGTGAAGTTAAATTCTCCACT
CTGTCCATCAGATACTATGGTTCTCCACTATGGCAACTAACTCACTCAATTTTCCCTCCTTAGCAGCATT
CCATCTTCCCGATCTTCTTTGCTTCTCCAACCAAAACATCAATGTTTATTAGTTCTGTATACAGTACAGG
ATCTTTGGTCTACTCTATCACAAGGCCAGTACCACACTCATGAAGAAAGAACACAGGAGTAGCTGAGAGG
CTAAAACTCATCAAAAACACTACTCCTTTTCCTCTACCCTATTCCTCAATCTTTTACCTTTTCCAAATCC
CAATCCCCAAATCAGTTTTTCTCTTTCTTACTCCCTCTCTCCCTTTTACCCTCCATGGTCGTTAAAGGAG
AGATGGGGAGCATCATTCTGTTATACTTCTGTACACAGTTATACATGTCTATCAAACCCAGACTTGCTTC
CGTAGTGGAGACTTGCTTTTCAGAACATAGGGATGAAGTAAGGTGCCTGAAAAGTTTGGGGGAAAAGTTT
CTTTCAGAGAGTTAAGTTATTTTATATATATAATATATATATAAAATATATAATATACAATATAAATATA
TAGTGTGTGTGTATGCGTGTGTGTAGACACACACGCATACACACATATAATGGAAGCAATAAGCCATTCT
AAGAGCTTGTATGGTTATGGAGGTCTGACTAGGCATGATTTCACGAAGGCAAGATTGGCATATCATTGTA
ACTAAAAAAGCTGACATTGACCCAGACATATTGTACTCTTTCTAAAAATAATAATAATAATGCTAACAGA
AAGAAGAGAACCGTTCGTTTGCAATCTACAGCTAGTAGAGACTTTGAGGAAGAATTCAACAGTGTGTCTT
CAGCAGTGTTCAGAGCCAAGCAAGAAGTTGAAGTTGCCTAGACCAGAGGACATAAGTATCATGTCTCCTT
TAACTAGCATACCCCGAAGTGGAGAAGGGTGCAGCAGGCTCAAAGGCATAAGTCATTCCAATCAGCCAAC
TAAGTTGTCCTTTTCTGGTTTCGTGTTCACCATGGAACATTTTGATTATAGTTAATCCTTCTATCTTGAA
TCTTCTAGAGAGTTGCTGACCAACTGACGTATGTTTCCCTTTGTGAATTAATAAACTGGTGTTCTGGTTC
AT

Figure 1.3 Practical Bioinformatics (© Garland Science 2013)

An example of DNA sequence

Figure 1.4 Practical Bioinformatics (© Garland Science 2013)

An example of DNA sequence

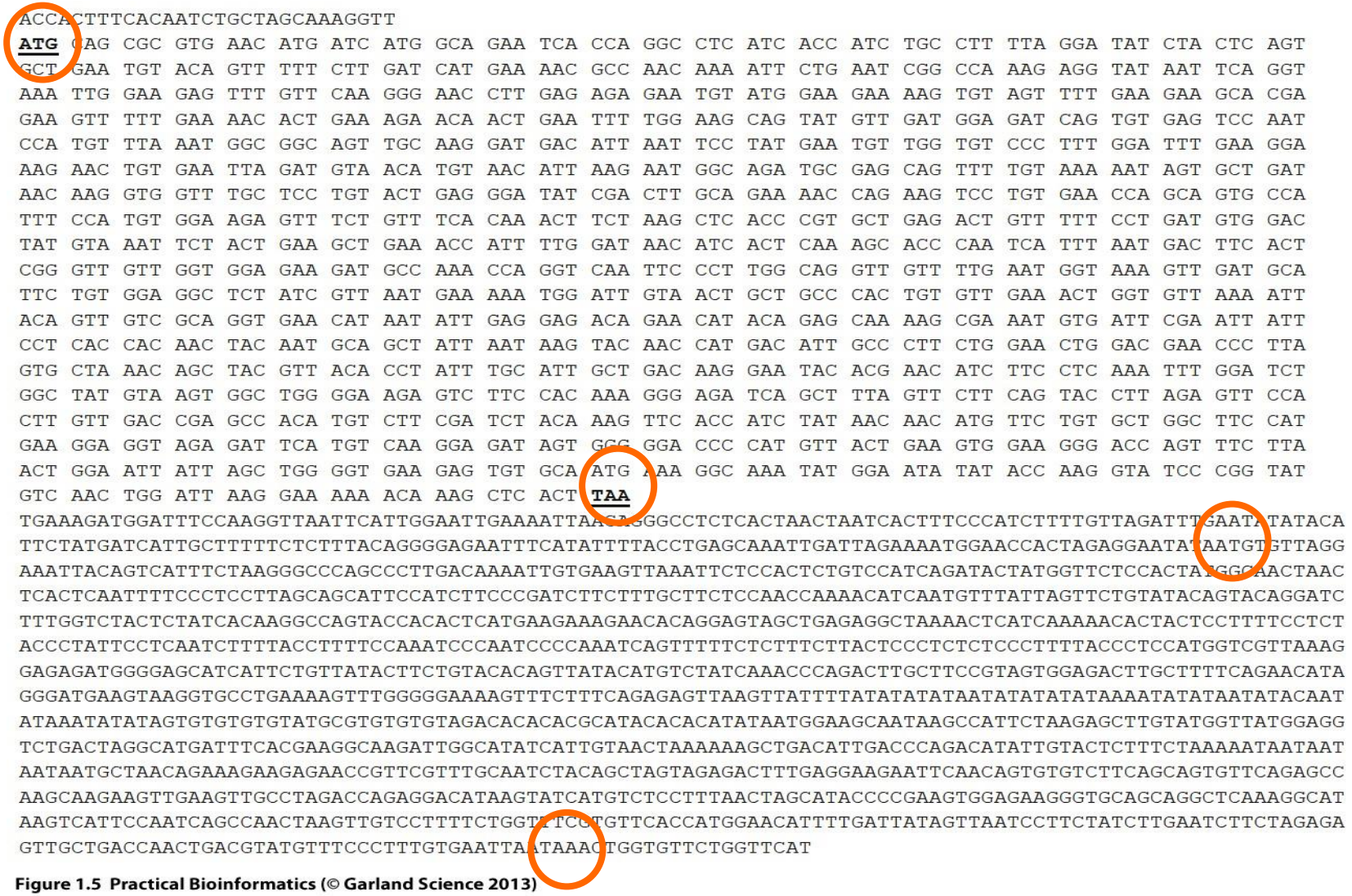


Figure 1.5 Practical Bioinformatics (© Garland Science 2013)

What is a Protein Sequence?

- Proteins are complex molecules which control most aspects of cell biology.
- Constructed of small subunits called amino acids.
- There are 20 main types of amino acids that form proteins.
- Assembled by 'reading' (or *translating*) the DNA sequence.
- Every set of 3 bases (e.g. ATG) corresponds to an amino acid.
- So a protein is built up one amino acid at a time according to the DNA blueprint.

Sequence analysis

- **Sequence analysis** is the process of subjecting a DNA, RNA or peptide sequence to any of a wide range of analytical methods to understand its features, function, structure, or evolution.

Applications

Sequence analysis can be used in following fields of molecular biology:

- The comparison of sequences in order to find similarity, often to infer if they are related (homologous)
- Identification of intrinsic features of the sequence such as active sites, post translational modification sites, gene-structures, reading frames, distributions of introns and exons and regulatory elements.
- Identification of sequence differences and variations such as point mutations and single nucleotide polymorphism (SNP) in order to get the genetic marker.
- Revealing the evolution and genetic diversity of sequences and organisms
- Identification of molecular structure from sequence alone

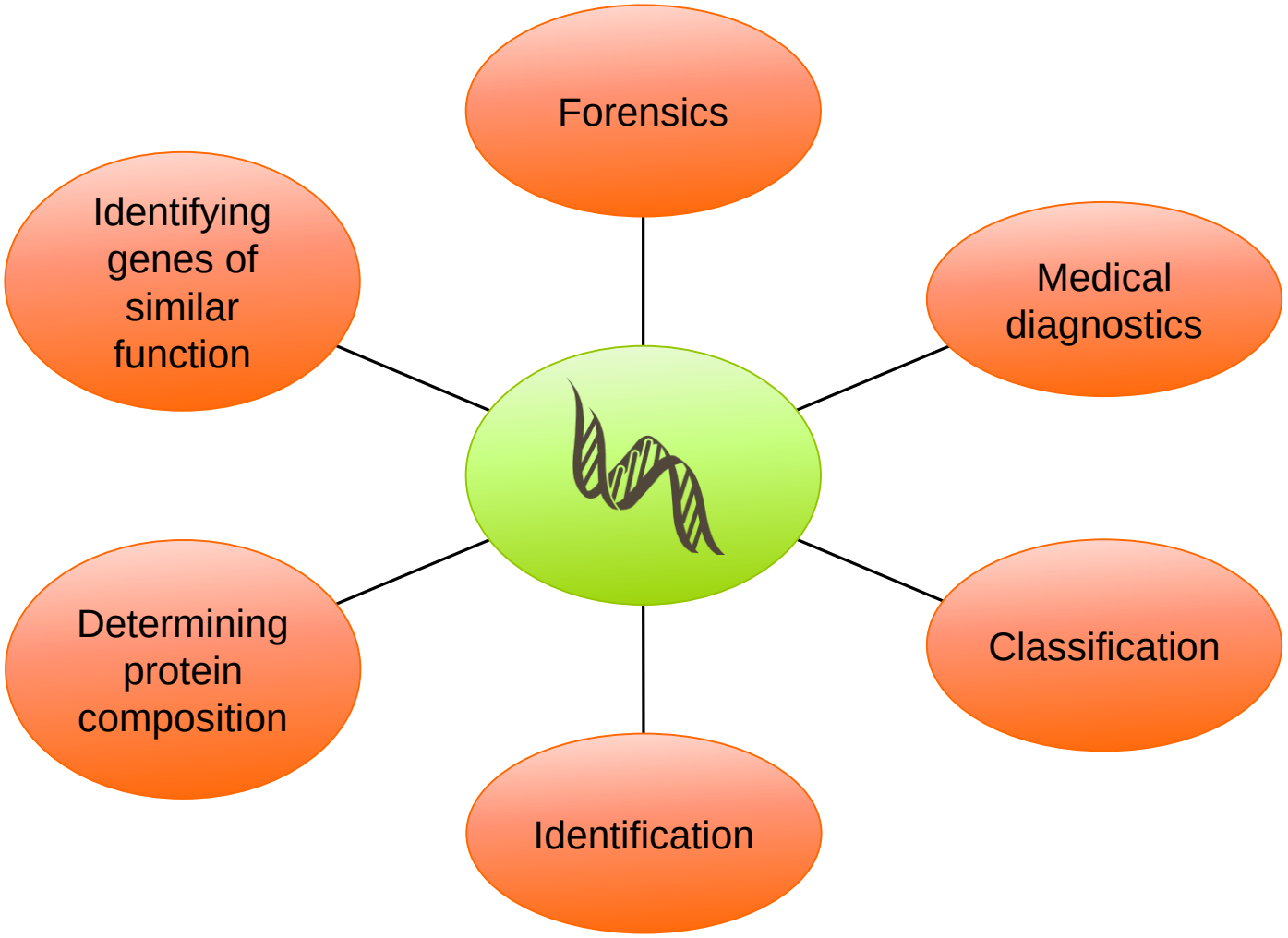
Uses of sequence analysis

1. Proteomics (determining protein structure).
2. Sequence Alignment.
3. Genomics.
4. Evolutionary Biology (Phylogenetics).
5. Systems biology.

Looking at DNA sequences

- Analysis of DNA or protein sequences is a frequent requirement of research.
  - Locating genes within a sequence.
  - Comparing two sequences for similarity.
  - Searching for similar genes (orthologues) in other organisms.

Using a DNA Sequence In



Centralized Databases Store DNA Sequences

- We begin with **three main sites** that have been responsible for storing nucleotide sequence data from 1982 to the present

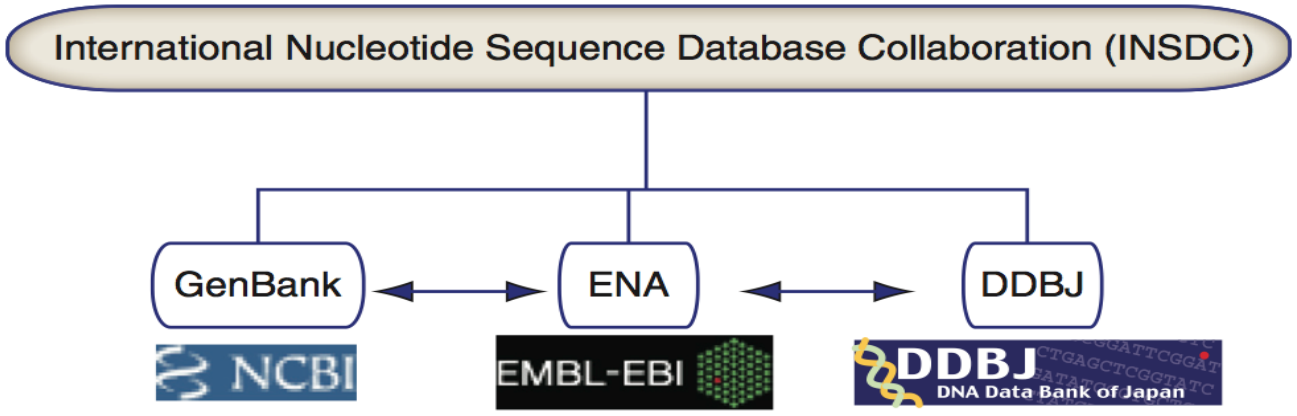
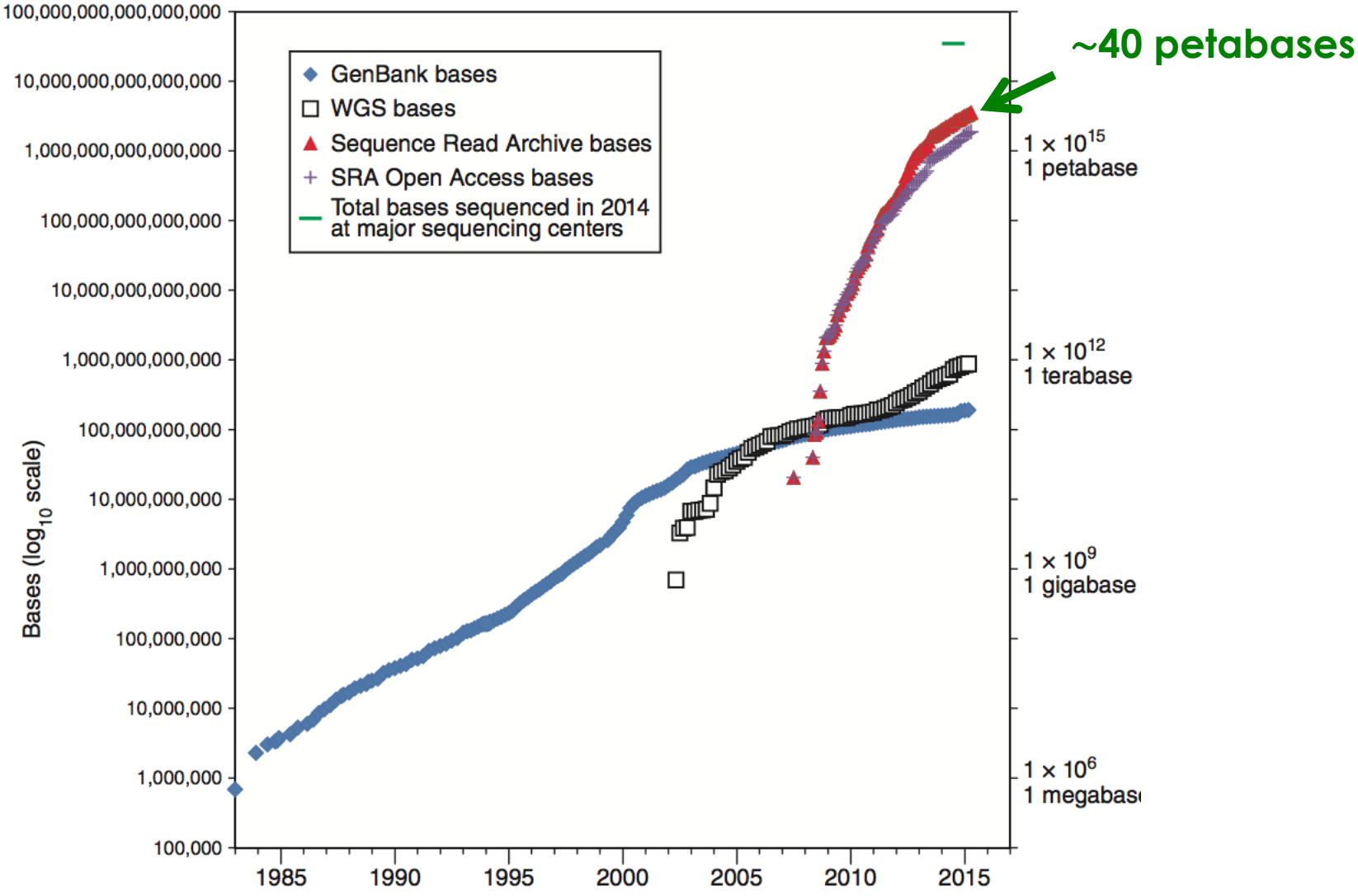


FIGURE 2.1 The nucleotide collections of GenBank at NCBI, EMBL-Bank at the European Bioinformatics Institute, and DDBJ at the DNA Data Bank of Japan are all coordinated by the International Nucleotide Sequence Database Collaboration (INSDC).

Growth of DNA in Repositories



What is database?

- It is a collection of data from different sources.
- Each stored data is called entry.
- Used to retrieve data (or entrez).
- Searchable with keyword (called query).

Why do we need databases?

- To arrange these large number of sequences, i.e., to make a library and catalogue of genes, RNA and proteins
- To remove redundant sequences. So, we have only unique sequences per gene.
- To annotate (name) the sequences. So, we know what the protein does or where the gene starts and ends.

Scales of DNA Base Pairs

TABLE 2.1 Scales of DNA base pairs.

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	

GenBank

- A database consisting of DNA and protein sequences.
- Contains bibliographic and biological annotation.
- Data are available free of charge from NCBI.
- Over 310,000 different species!

GenBank

- Over **1000** new species added **per month**
- Over the past 30 years the number of bases in GenBank has doubled approximately **every 18 months**
- GenBank received submissions since 1982, including sequences from thousands of individual submitters.

Organisms in GenBank

TABLE 2.3 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

Source: GenBank, NCBI, http://www.ncbi.nlm.nih.gov/Taxonomy/txstat.cgi.

NCBI and EBI

Two of **the main centralized bioinformatics hubs**:

- The National Center for Biotechnology Information (NCBI)
- The European Bioinformatics Institute (EBI).

In many cases those sites begin with similar raw data and then provide distinct ways of organizing, analyzing, and displaying data across a broad range of bioinformatics applications.

NCBI

- Creates public databases.
- Conducts research in computational biology
- Develops software tools for analyzing genome data
- Disseminates biomedical information

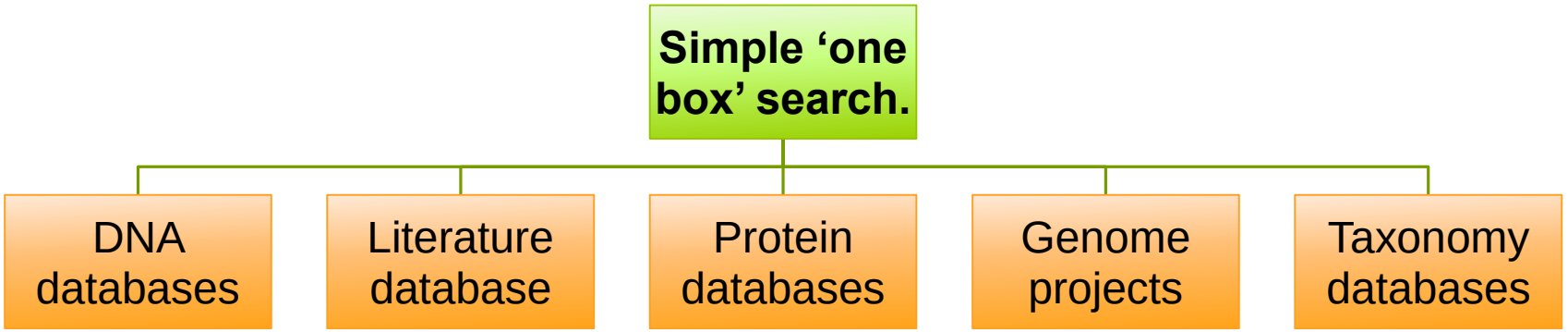
NCBI Resources

- **Entrez**: integrates scientific literature, DNA, and protein sequence databases, 3D protein structure data, population study datasets, and assemblies of complete genomes into a tightly coupled system
- **PubMed**: search service from the National Library of Medicine (NLM) that provides access to over 24 million citations
- **BLAST**: (Basic Local Alignment Search Tool) is NCBI's sequence similarity search
- **OMIM**: Online Mendelian Inheritance in Man (OMIM) is a catalog of human genes and genetic disorders
- **Books**: NCBI offers about 200 books online
- **other**

NCBI Entrez

Entrez:

- Accessed from the home page of NCBI
- Provides links to results from **40** different NCBI **databases**
  - scientific literature, DNA, and protein sequence databases, 3D protein structures, population study datasets, and assemblies of complete genomes



Entrez Main Screen

https://www.ncbi.nlm.nih.gov

Search

NCBI Resources How To Sign in to NCBI

NCBI
National Center for Biotechnology Information

All Databases Search

NCBI is currently redirecting web traffic to HTTPS. Read more about our https testing.

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

Welcome to NCBI

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

About the NCBI | Mission | Organization | NCBI News | Blog

Submit

Deposit data or manuscripts into NCBI databases

Download

Transfer NCBI data to your computer

Learn

Find help documents, attend a class or watch a tutorial

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI PubMed

PubMed: provides access to over 24 million citations.

PubMed citations come from

1) MEDLINE indexed journals (largest subset of PubMed)
2) Journals/manuscripts deposited in PMC, and
3) NCBI Bookshelf.

J Infect Public Health. 2016 Sep 15. pii: S1876-0341(16)30146-0. doi: 10.1016/j.jiph.2016.09.007. [Epub ahead of print]

Building predictive models for MERS-CoV infections using data mining techniques.

Al-Turaiki I[1], Alshahrani M[2], Almutairi T[3].

Author information

Abstract

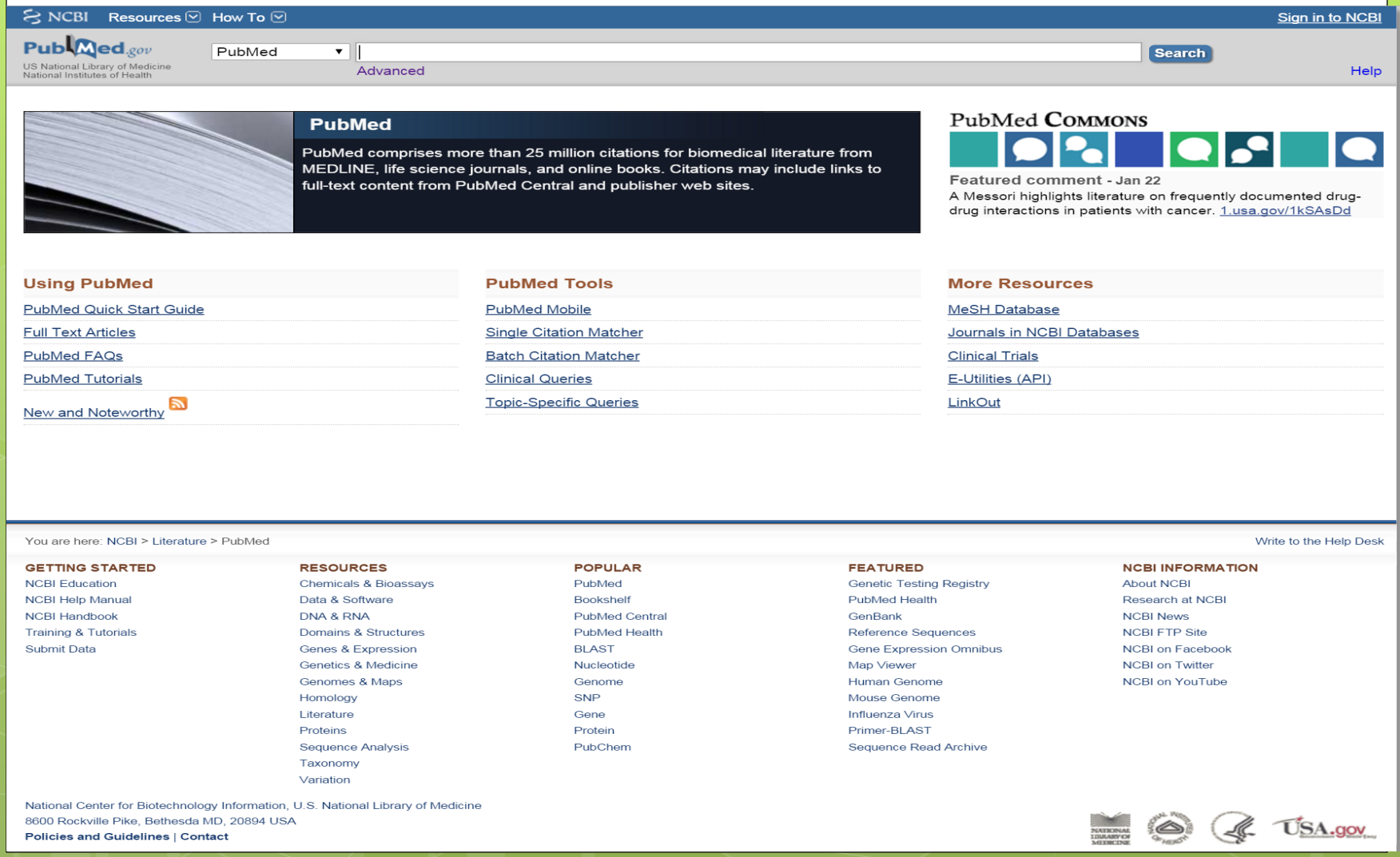
BACKGROUND: Recently, the outbreak of MERS-CoV infections caused worldwide attention to Saudi Arabia. The novel virus belongs to the coronaviruses family, which is responsible for causing mild to moderate colds. The control and command center of Saudi Ministry of Health issues a daily report on MERS-CoV infection cases. The infection with MERS-CoV can lead to fatal complications, however little information is known about this novel virus. In this paper, we apply two data mining techniques in order to better understand the stability and the possibility of recovery from MERS-CoV infections.

METHOD: The Naive Bayes classifier and J48 decision tree algorithm were used to build our models. The dataset used consists of 1082 records of cases reported between 2013 and 2015. In order to build our prediction models, we split the dataset into two groups. The first group combined recovery and death records. A new attribute was created to indicate the record type, such that the dataset can be used to predict the recovery from MERS-CoV. The second group contained the new case records to be used to predict the stability of the infection based on the current status attribute.

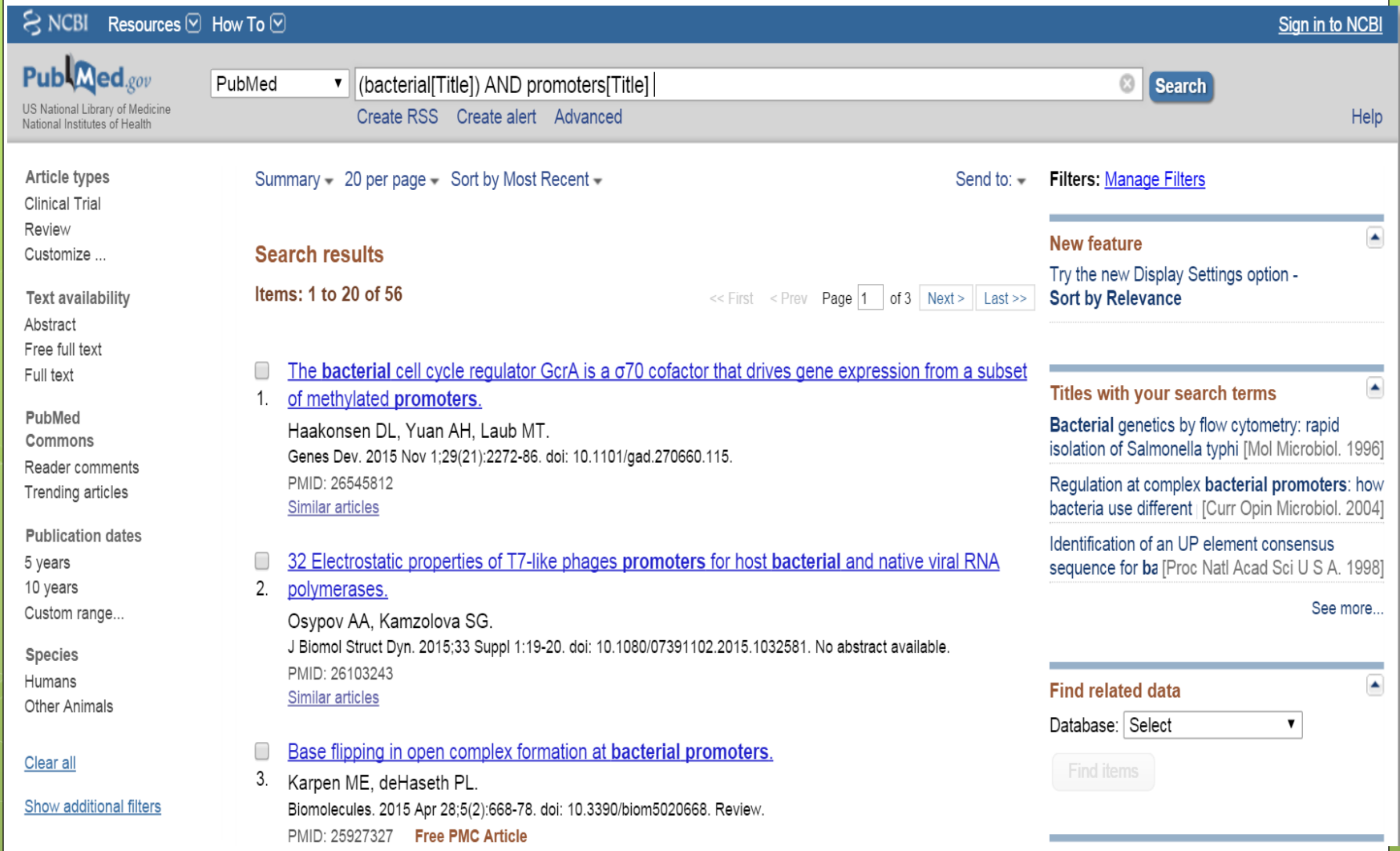
What is the difference between pubmed and medline?

- Pubmed is an interface used to search Medline, as well as additional biomedical content. Medline is an interface for searching only Medline content.
- Pubmed is more user-friendly and allows you to search through more content than Ovid Medline. However, Medline allows you to perform a more focused search. You will get slightly different results by searching in each database.
- In Pubmed, in addition to Medline articles, you will have access to PubMedCentral papers, which are full text articles deposited to promote open access, and articles that are "in press" that is, prior to being indexed with MeSH terms, and articles submitted by publishers, "ahead of print." This is why if you search for the same term in Medline and in Pubmed, you may obtain as many as ten thousand more articles in Pubmed.

Pubmed database



Results of Pubmed

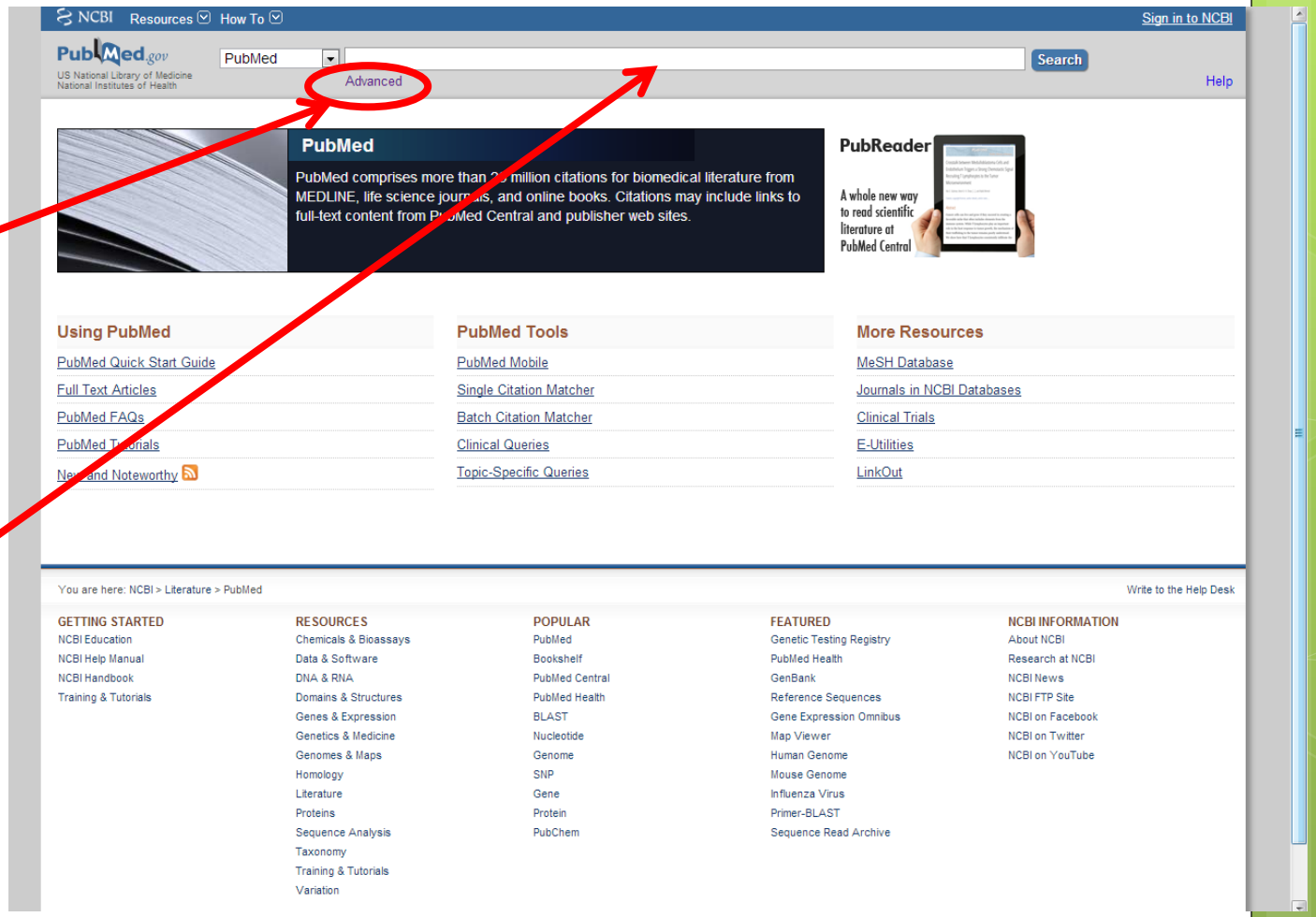


Using Pubmed

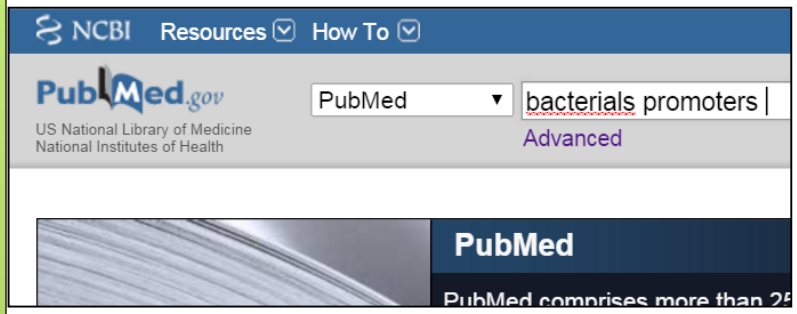
Go to http://www.ncbi.nlm.nih.gov/pubmed

Click here for more options

Type the name of a topic, author or even an institute



Using pubmed



Text MUST be correct

8108 articles are there

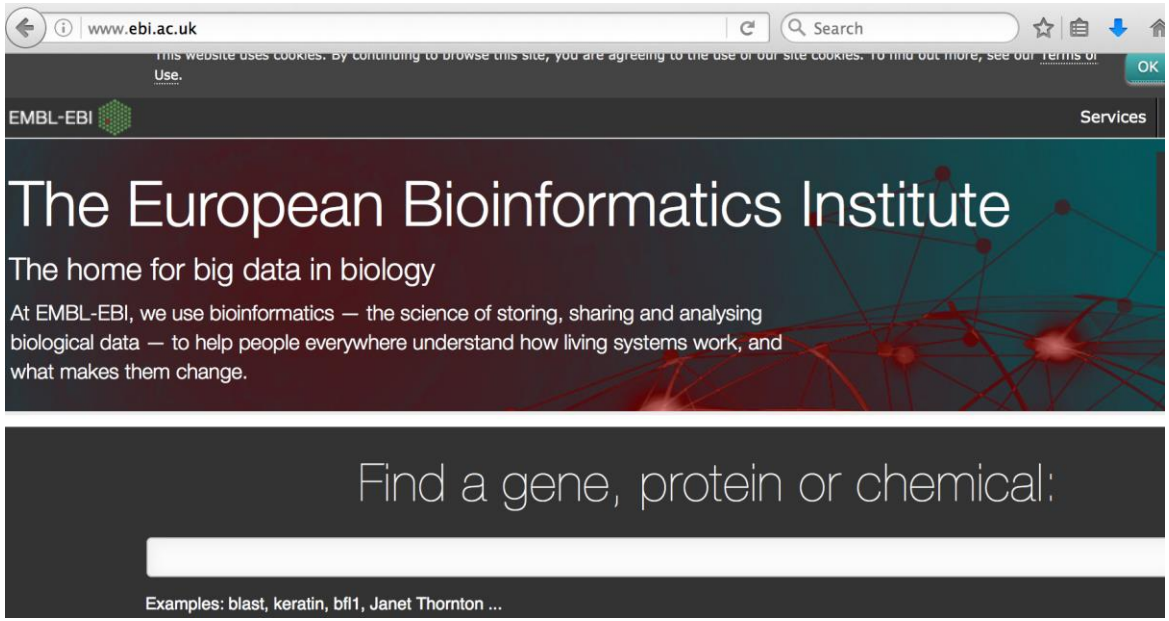
56 articles after using advanced option

NCBI

- **BLAST:** sequence similarity search
- **OMIM:** a catalog of human genes and genetic disorders.
- **Books**: NCBI offers about 200 books online.

EBI

- The EBI website is comparable to NCBI in its scope and mission, and it represents a complementary, independent resource.



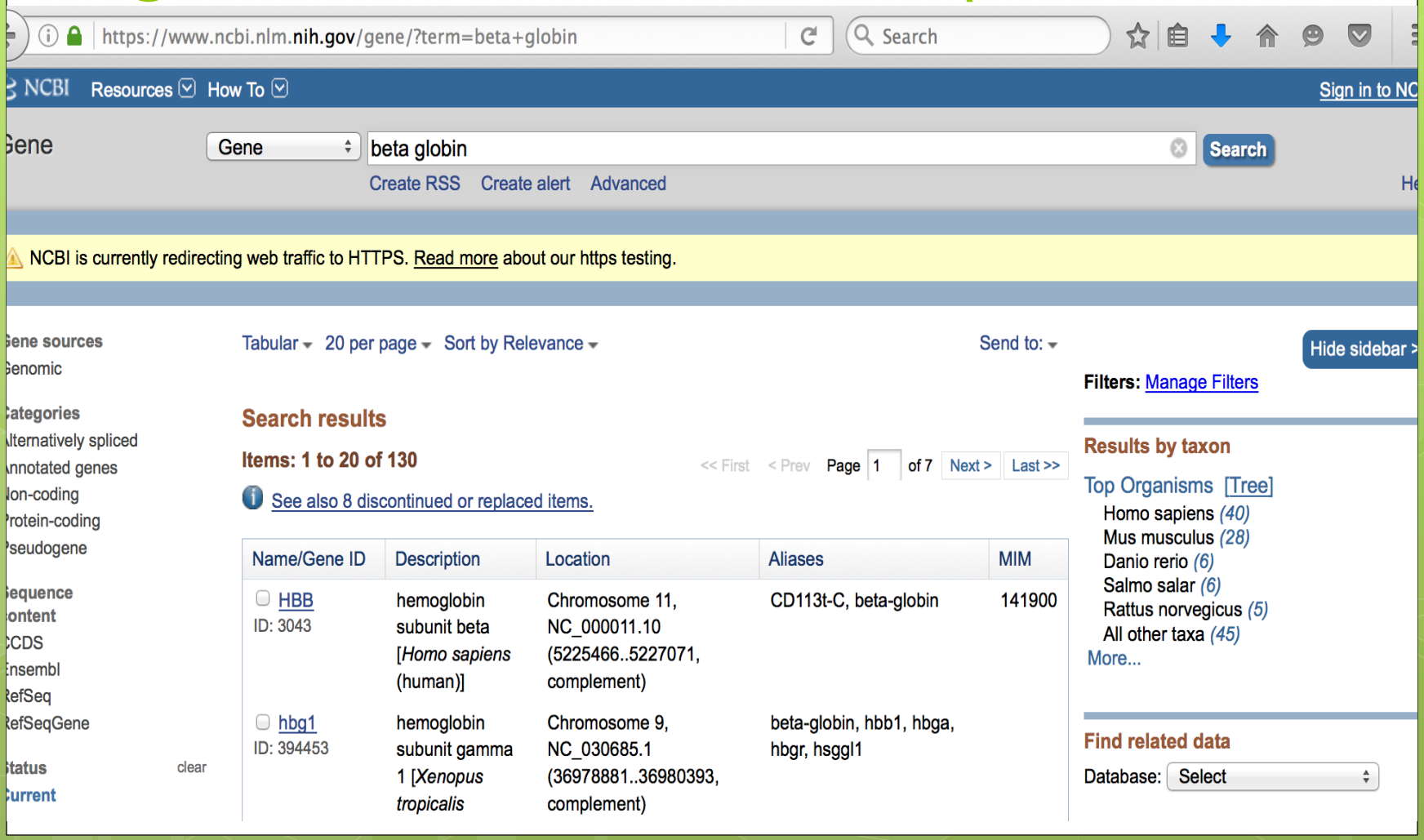
EBI

- EBI features **six core molecular databases**:
  - EMBL-Bank is the repository of DNA and RNA sequences that is complementary to GenBank and DDBJ
  - Swiss-Prot and TrEMBL are two protein databases
  - MSD is a protein structure database
  - Ensembl is one of the main genome browsers
  - ArrayExpress is for gene expression data

Access to Information via Gene Resource at NCBI

- To illustrate the use of NCBI Gene we search for human **beta globin**.

Result of a search for "beta globin" in NCBI Gene (via an



NCBI Gene entry for human beta globin

HBB hemoglobin subunit beta [*Homo sapiens* (human)]

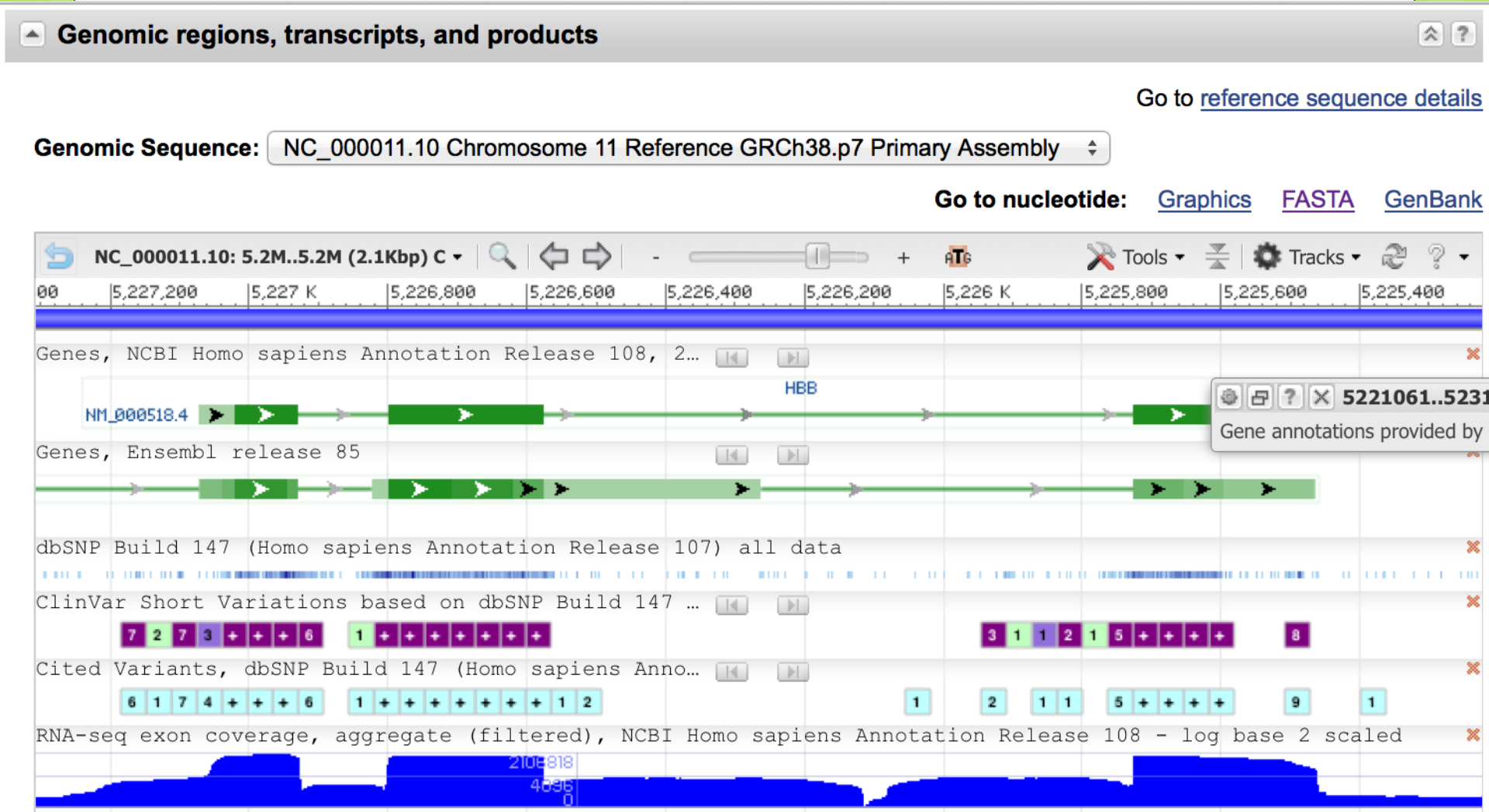
Gene ID: 3043, updated on 22-Sep-2016

Information is provided on the gene structure and chromosomal location, as well as a summary of the protein's function.

Summary

Official Symbol	HBB provided by
Official Full Name	hemoglobin subunit beta provided by HGNC
Primary source	HGNC:HGNC:4827
See related	Ensembl:ENSG00000244734 HPRD:00786; MIM:141900; Vega:OTTHUMG00000066678
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	CD113t-C; beta-globin
Summary	The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin, Hb A. The normal adult hemoglobin tetramer consists of two alpha chains and two beta chains. Mutant beta globin causes sickle cell anemia. Absence of beta chain causes beta-zero-thalassemia. Reduced amounts of detectable beta globin causes beta-plus-thalassemia. The order of the genes in the beta-globin cluster is 5'-epsilon -- gamma-G -- gamma-A -- delta -- beta--3'. [provided by RefSeq, Jul 2008]
Orthologs	all

Access to Information via Gene Resource at NCBI



Sequence format in databases

- A sequence format defines the permitted layout and content of text in a file.
- The **FASTA format** is a very widely used format. It consists of a header line starting with a ”>” character followed by a code identifying the sequence and, very often, some text describing the sequence. The header line is followed by one or more lines containing the sequence itself.
- FASTA files may contain one or more sequences

FASTA Format

Homo sapiens chromosome 11, GRCh38.p7 Primary Assembly

NCBI Reference Sequence: NC_000011.10

GenBank Graphics

description

```
>gi|568815587:c5227071-5225466 Homo sapiens chromosome 11, GRCh38.p7 Primary
Assembly
ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGA
GGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGC
AGGTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCATGTGGAGACAGAGAAG
ACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGG
TGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGG
CAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGAC
AACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACT
TCAGGGTGAGTCTATGGGACGCTTGATGTTTTCTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAG
GAAGGGGATAAGTAACAGGGTACAGTTTAGAATGGGAAACAGACGAATGATTGCATCAGTGTGGAAGTCT
CAGGATCGTTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTTGTTTAATTCTTGCTTTCT
TTTTTTTTCTTCTCCGCAATTTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAAAGGAAATA
```

sequence

FASTA is both an alignment program and a commonly used sequence format

FASTA files may contain one or more sequences

```
>crab_anapl ALPHA CRYSTALLIN B CHAIN (ALPHA(B)-CRYSTALLIN).
MDITIHNPLIRRPLFSWLAPSRIFDQIFGEHLQESELLPASPSLSPFLMR
SPIFRMPSWLETGLSEMRLEKDKFSVNLDVKHFSPEELKVKVLGDMVEIH
GKHEERQDEHGFIAREFNRKYRIPADVDPLTITSSLSLDGVLTVSAPRKQ
SDVPERSIPITREEKPAIAGAQRK
>crab_bovin ALPHA CRYSTALLIN B CHAIN (ALPHA(B)-CRYSTALLIN).
MDIAIHHPWIRRPFFPFHSPSRLFDQFFGEHLLESDLFPASTSLSPFYLR
PPSFLRAPSWIDTGLSEMRLEKDRFSVNLDVKHFSPEELKVKVLGDVIEV
HGKHEERQDEHGFISREFHRKYRIPADVDPLAITSSLSSDGVLTVNGPRK
QASGPERTIPITREEKPAVTAAPKK
>crab_chick ALPHA CRYSTALLIN B CHAIN (ALPHA(B)-CRYSTALLIN).
MDITIHNPLVRRPLFSWLTPSRIFDQIFGEHLQESELLPTSPSLSPFLMR
SPFFRMPSWLETGLSEMRLEKDKFSVNLDVKHFSPEELKVKVLGDMIEIH
GKHEERQDEHGFIAREFSRKYRIPADVDPLTITSSLSLDGVLTVSAPRKQ
SDVPERSIPITREEKPAIAGSQRK
```

Some identifiers used in fasta description

Database Name	Abbreviations
GenBank	gb
EMBL Data Library	emb
DNA Database of Japan	dbj
SWISS-PROT	sp
Protein Data Bank	pdb
NCBI Reference Sequence	ref

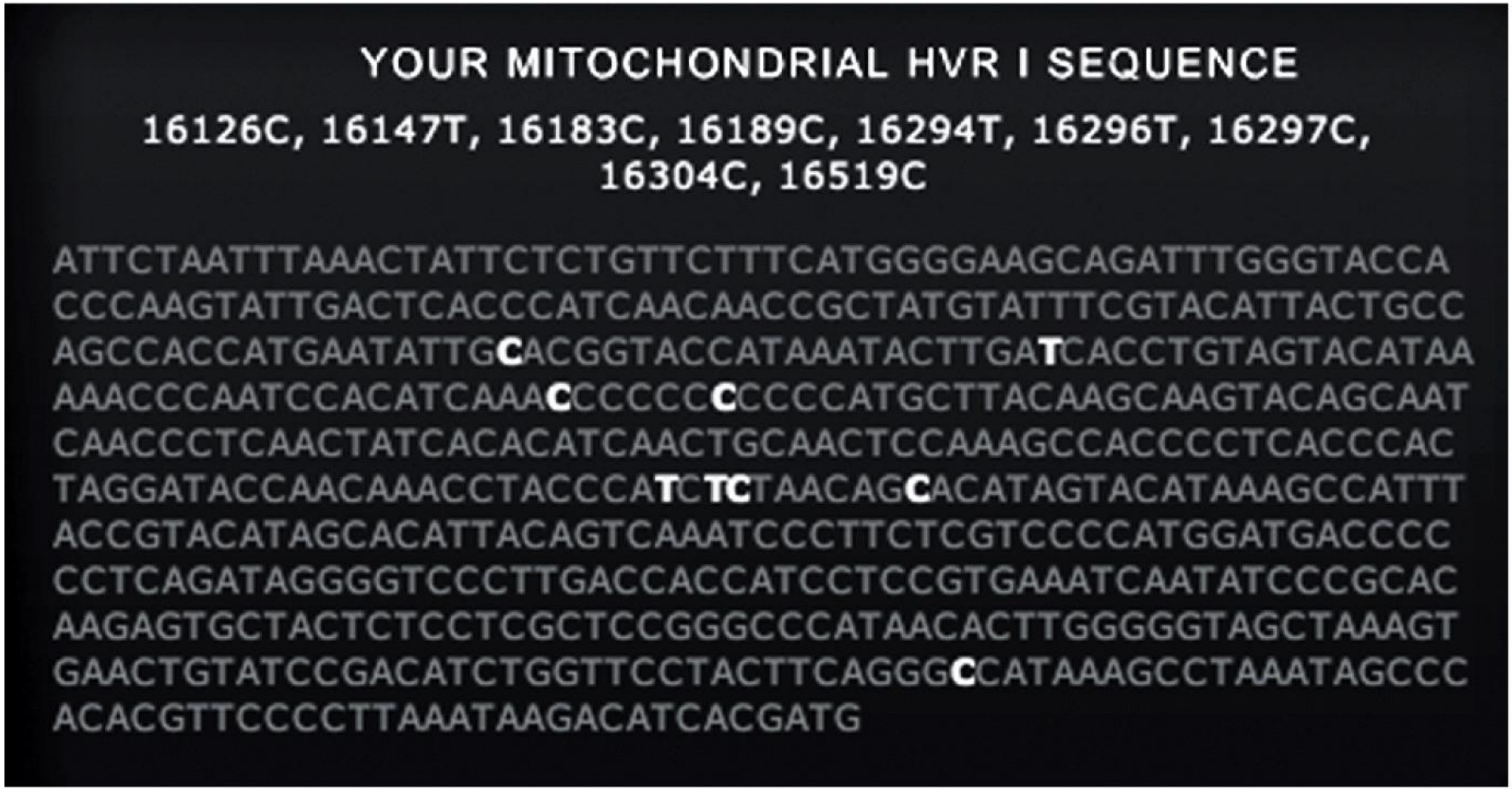
Other formats

- Beyond FASTA, the most widespread sequence formats are those used by the major sequence databases:
- ***EMBL***http://www.ebi.ac.uk/embl/Documentation/User_manual/format.html
- ***GenBank***http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html
- ***SwissProt***http://ca.expasy.org/sprot/userman.html#whatis

HVR sequence

A **hypervariable region (HVR)** is a location within nuclear DNA or the D-loop of mitochondrial DNA in which base pairs of nucleotides repeat (in the case of nuclear DNA) or have substitutions (in the case of mitochondrial DNA). Changes or repeats in the hypervariable region are highly polymorphic.

Example of HVR sequence



Box 1.1 Practical Bioinformatics (© Garland Science 2013)

COMMAND-LINE ACCESS TO DATA AT NCBI

- The websites of NCBI, EBI, Ensembl, and other bioinformatics sites offer convenient access to resources through a web browser
- *An alternative is to use command-line tools.*

ACCESS TO INFORMATION GENOME BROWSERS

Genome browsers are **databases** with a **graphical** interface that presents a representation of **sequence information** and other data as a **function** of **position** across the chromosomes.

- Three principal genome browsers (Ensembl, UCSC, and NCBI)

ACCESS TO INFORMATION GENOME BROWSERS

Genome build refers to an **assembly** in which DNA sequence is **collected and arranged** to reflect the sequence along each chromosome. For a given organism's genome, a build is released only occasionally (typically every few years)

ACCESS TO INFORMATION GENOME BROWSERS

Annotation is the **assignment** of **information** such as the start and stop position of **genes**, exons, repetitive DNA elements, or other features.

- Best to use the most recent available build.
- Earlier builds have richer annotation

Annotations (Features)

- It is a type of metadata (sideway database)
- It is the complete information required to be known about that particular sequence (could be DNA, RNA or protein).

In case of nucleotides, annotation is the process of:

- Identifying the locations of genes, coding regions and other specific locations that are of importance in a DNA sequence or genome.

- Associating relevant information with those locations (e.g. determining what the identified genes do).

In case of proteins, annotation is the process of:

- Describing regions or sites of interest in the protein sequence, such as post-translational modifications, binding sites, enzyme active sites, local secondary structure or other characteristics.

The University of California, Santa Cruz (UCSC) Genome Browser

- Provides graphical views of chromosomal locations at various levels of resolution (from several base pairs up to hundreds of millions of base pairs spanning an entire chromosome).
- Each chromosomal view is accompanied by horizontally oriented annotation tracks.

The University of California, Santa Cruz (UCSC) Genome Browser

- There are hundreds of available user-selected tracks in categories such as mapping and sequencing, phenotype and disease associations, genes, expression, comparative genomics, and genomic variation.
- These annotation tracks offer the Genome Browser tremendous depth and flexibility.

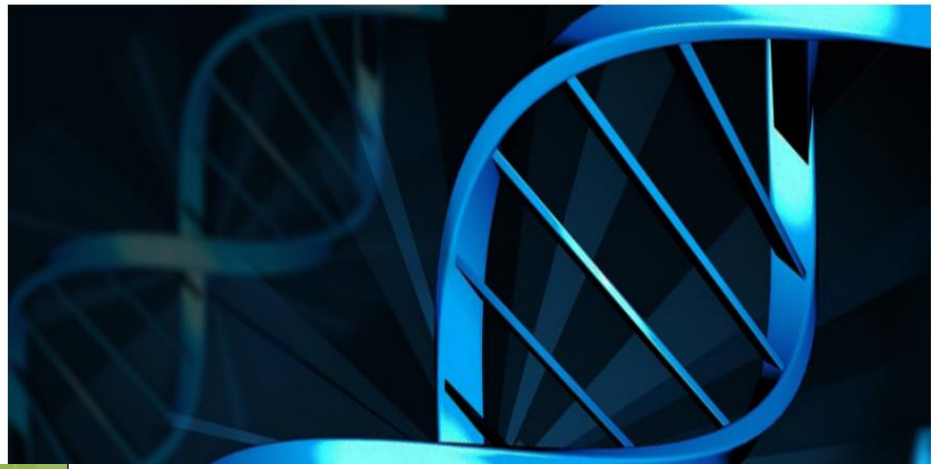
The University of California, Santa Cruz (UCSC) Genome Browser

https://genome.ucsc.edu

Search

UNIVERSITY OF CALIFORNIA SANTA CRUZ UCSC Genome Browser

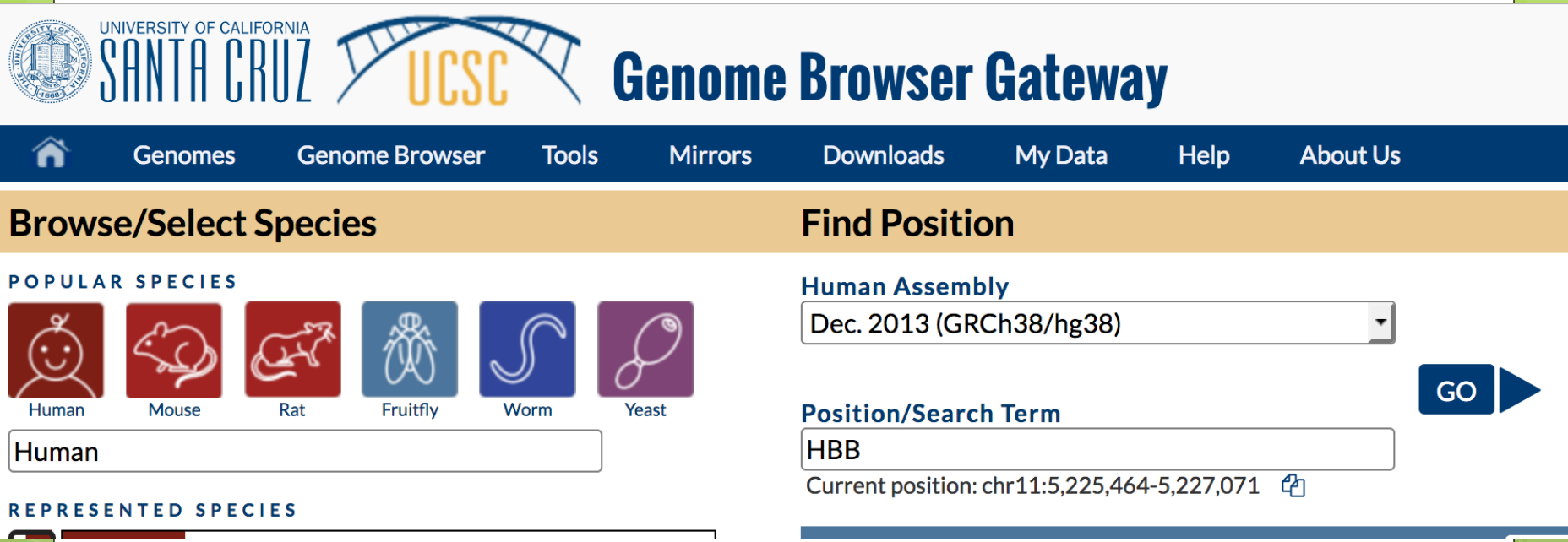
Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us



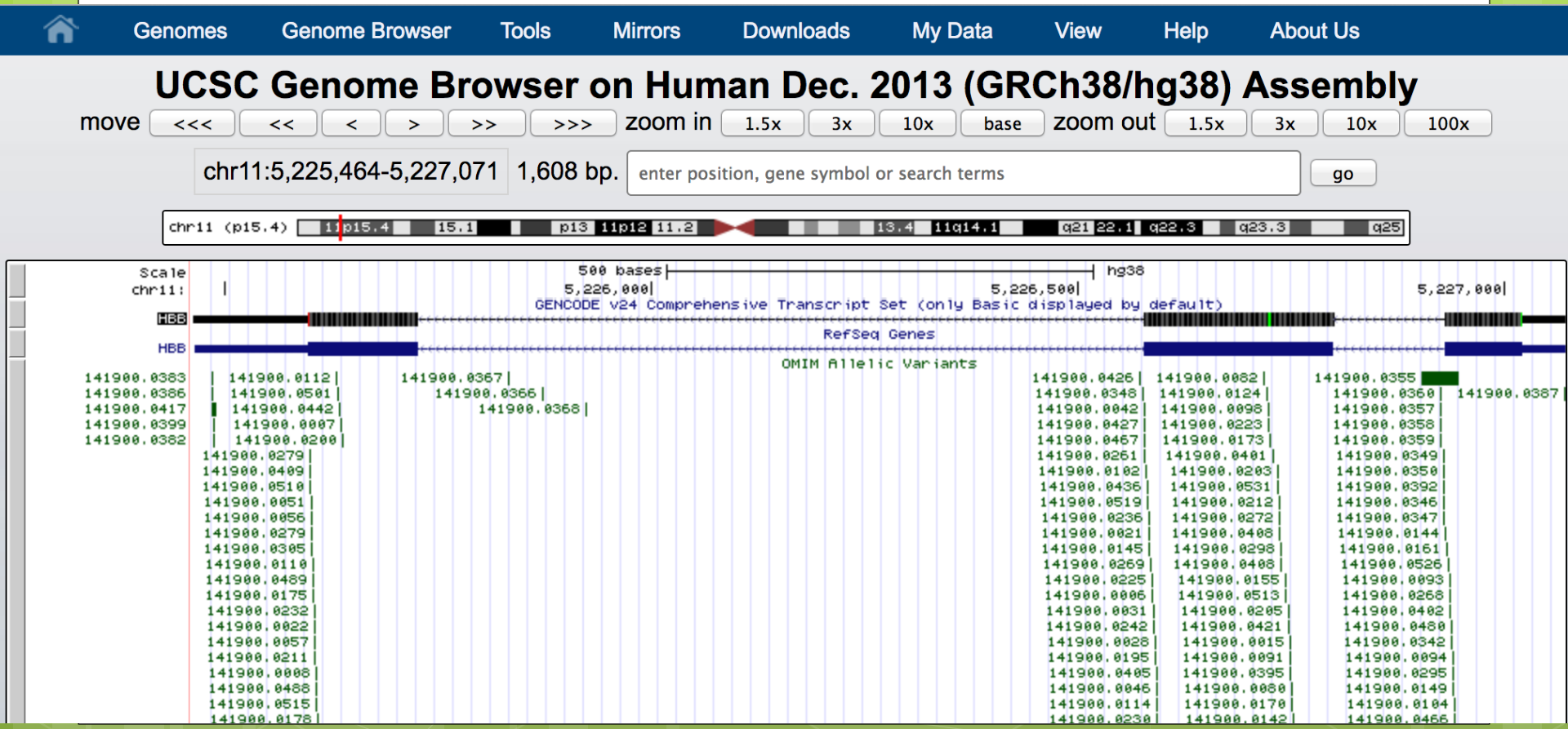
Our tools

- **Genome Browser**
  interactively visualize genomic data
- **BLAT**
  rapidly align sequences to the genome
- **Table Browser**
  download data from the Genome Browser database
- **Variant Annotation Integrator**
  get functional effect predictions for variant calls

The University of California, Santa Cruz (UCSC) Genome Browser



The University of California, Santa Cruz (UCSC) Genome Browser



The Ensembl Genome Browser

- To many users, it is comparable in scope and importance to the UCSC Genome Browser, and it is often useful for new users to visit both sites.

asia.ensembl.org/index.html

Search

You have been redirected to your nearest mirror. **Click here to go back to www.en**

e!Ensembl ASIA

BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | More ▾

Search: All species for Go

e.g. **BRCA2** or **rat 5:62797383-63627669** or **rs699** or **coronary heart disease**

Browse a Genome

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Popular genomes

Human

Human

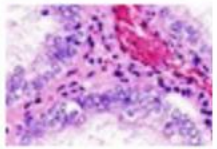
Still using Human GRCh37?

Go to e!GRCh37

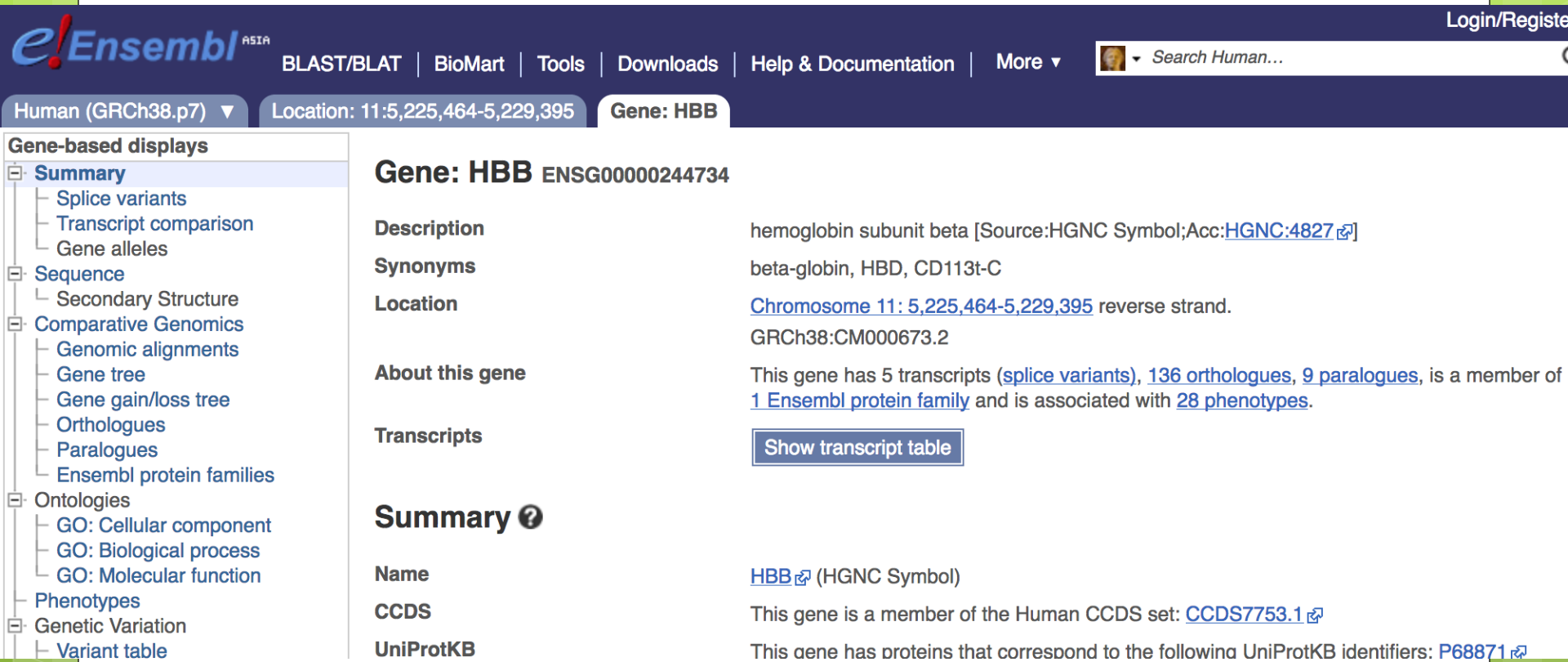
Variant Effect Predictor

VeP

Gene expression in different tissues

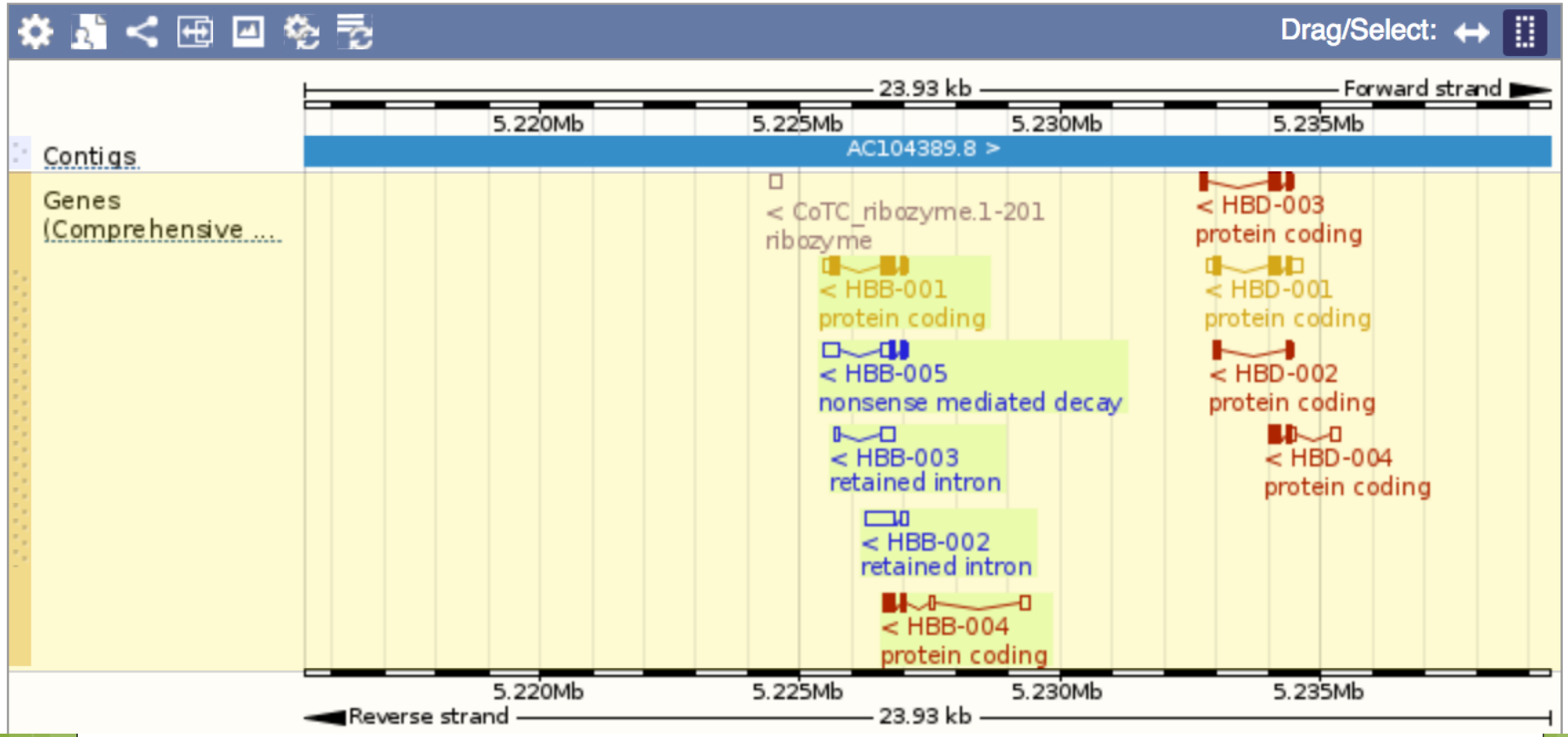


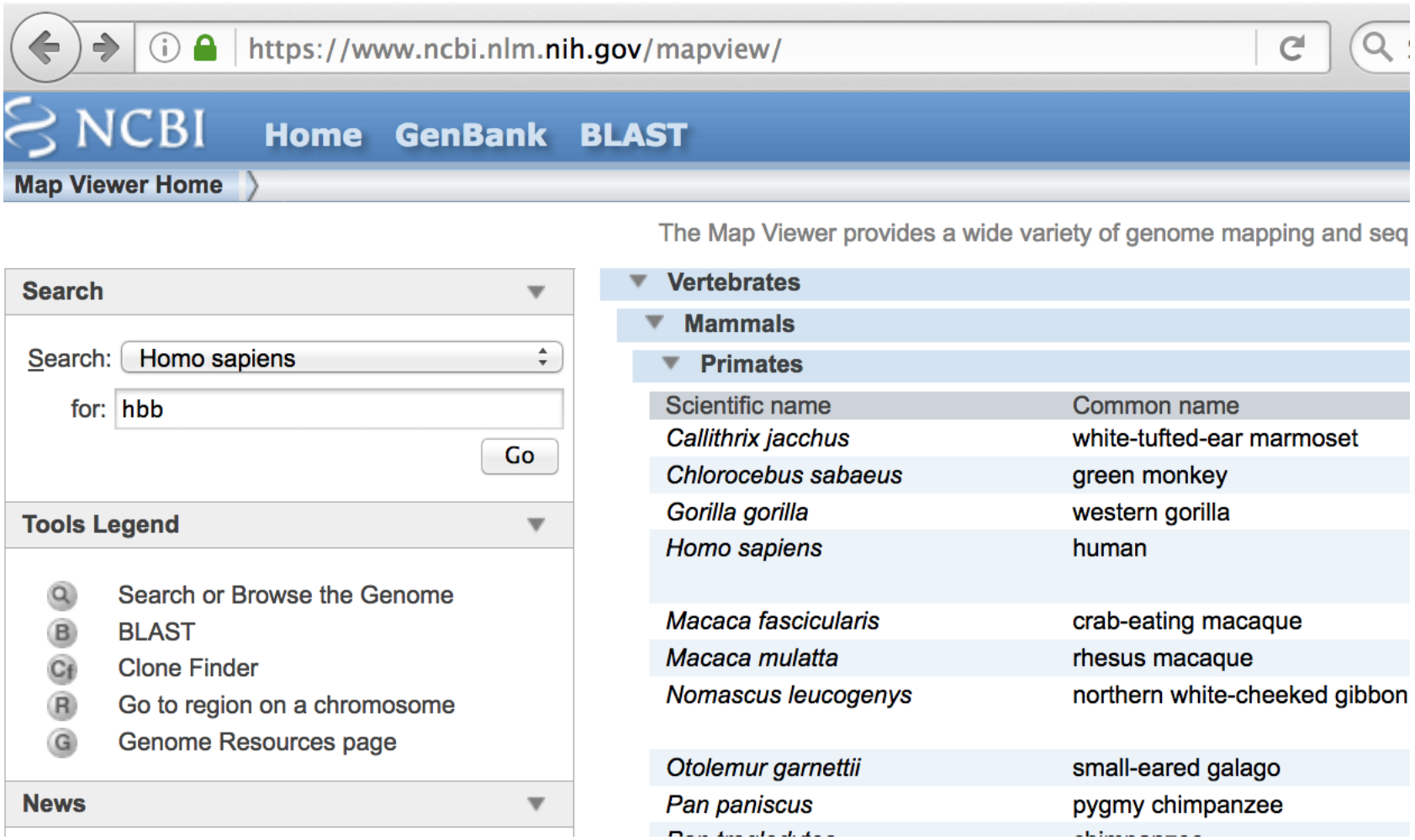
Find SNPs and other variants for my gene


Ensembl ASIA
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | More
Search Human...
Login/Registe
Human (GRCh38.p7)
Location: 11:5,225,464-5,229,395
Gene: HBB
Gene-based displays
Summary
Splice variants
Transcript comparison
Gene alleles
Sequence
Secondary Structure
Comparative Genomics
Genomic alignments
Gene tree
Gene gain/loss tree
Orthologues
Paralogues
Ensembl protein families
Ontologies
GO: Cellular component
GO: Biological process
GO: Molecular function
Phenotypes
Genetic Variation
Variant table
Gene: HBB ENSG00000244734
Description
hemoglobin subunit beta [Source:HGNC Symbol;Acc:HGNC:4827]
Synonyms
beta-globin, HBD, CD113t-C
Location
Chromosome 11: 5,225,464-5,229,395 reverse strand.
GRCh38:CM000673.2
About this gene
This gene has 5 transcripts (splice variants), 136 orthologues, 9 paralogues, is a member of 1 Ensembl protein family and is associated with 28 phenotypes.
Transcripts
Show transcript table
Summary
Name
HBB (HGNC Symbol)
CCDS
This gene is a member of the Human CCDS set: CCDS7753.1
UniProtKB
This gene has proteins that correspond to the following UniProtKB identifiers: P68871

Havana gene: OTTHUMG00000000078

Go to Region in Detail for more tracks and navigation options (e.g. zooming)




https://www.ncbi.nlm.nih.gov/mapview/
NCBI
Home
GenBank
BLAST
Map Viewer Home
The Map Viewer provides a wide variety of genome mapping and seq
Search
Search: Homo sapiens
for: hbb
Go
Tools Legend
Search or Browse the Genome
BLAST
Clone Finder
Go to region on a chromosome
Genome Resources page
News
Vertebrates
Mammals
Primates
Scientific name
Common name
Callithrix jacchus
white-tufted-ear marmoset
Chlorocebus sabaeus
green monkey
Gorilla gorilla
western gorilla
Homo sapiens
human
Macaca fascicularis
crab-eating macaque
Macaca mulatta
rhesus macaque
Nomascus leucogenys
northern white-cheeked gibbon
Otolemur garnettii
small-eared galago
Pan paniscus
pygmy chimpanzee

NCBI

NCBI Map Viewer

PubMed | Nucleotide | Protein | Genome | Gene | Structure | PopSet | Ta

Search for hbb on chromosome(s) assembly All Find

Map Viewer

Map Viewer Home
Map Viewer Help
Human Maps Help
FTP

NCBI Resources

Assembly
CCDS
Gene
Genome
RefSeq

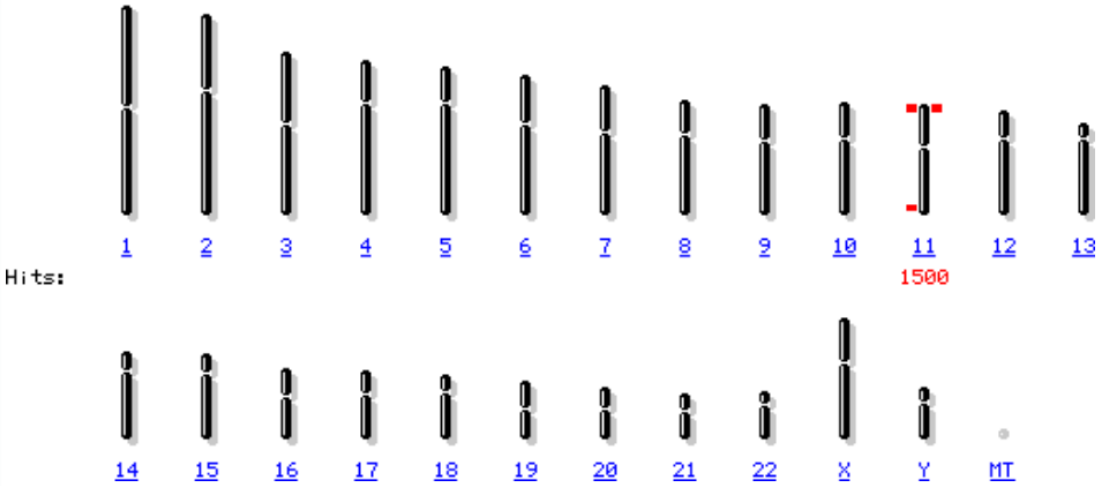
Organism Data in GenBank

EST

Homo sapiens (human) genome view

BL

Annotation Release 107 statistics Switch to previous build



Search results for query "hbb": 1500 hits shown (out of 5149 found)

Hits shown: 1 - 100

Chr	Assembly	Match	Map element	Type
11	reference	all matches		

LARGE-SCALE QUERIES OF REGIONS AND FEATURES

- In many cases we are interested in a single gene.
- In many other cases we want to know about large collections of genes, proteins, or indeed any other element.
- Example:
  - What is the complete set of human globin genes?
  - To which chromosomes are they assigned?
  - How many exons are on chromosome 11, and how many repeat elements occur in each exon?

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

LARGE-SCALE QUERIES OF REGIONS AND FEATURES

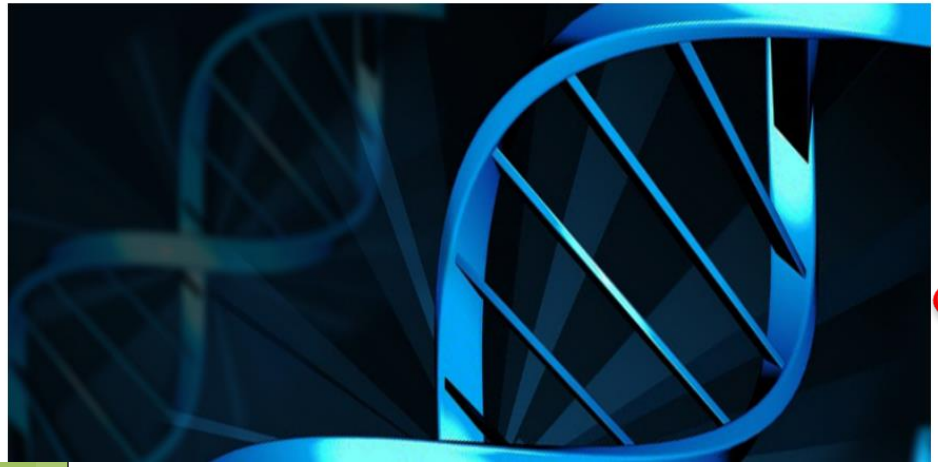
- Collect information one gene at a time?
  - tedious, inefficient, and error-prone
- There are many bioinformatics tools that allow us to collect genome-wide information.
  - UCSC Table Browser

https://genome.ucsc.edu

Search

UNIVERSITY OF CALIFORNIA SANTA CRUZ UCSC **Genome Browser**

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us



Our tools

- **Genome Browser**
  interactively visualize genomic data
- **BLAT**
  rapidly align sequences to the genome
- **Table Browser**
  download data from the Genome Browser database
- **Variant Annotation Integrator**
  get functional effect predictions for variant calls

Conclusion

- Bioinformatics is an emerging field whose defining feature is the accumulation of biological information in databases.
- The three major traditional DNA databases – GenBank, EMBL-Bank, and DDBJ – are adding several million new sequences each year as well as billions of nucleotides.
- Next-generation sequencing technology is producing vastly greater amounts of DNA.

Nucleotide Database

- **NCBI's sequence databases accept genome data from sequencing projects from around the world and serve as the cornerstone of bioinformatics research.**

GenBank:

- An annotated collection of all publicly available nucleotide and amino acid sequences.

EST database:

- A collection of expressed sequence tags, or short, single-pass sequence reads from mRNA (cDNA).

GSS database:

- A database of genome survey sequences, or short, single-pass genomic sequences.

HomoloGene:

- A gene homology tool that compares nucleotide sequences between pairs of organisms in order to identify putative orthologs.

Nucleotide Database

HTG database:

- A collection of high-throughput genome sequences from large-scale genome sequencing centers, including unfinished and finished sequences.

SNPs database:

- A central repository for both single-base nucleotide substitutions and short deletion and insertion polymorphisms.

RefSeq:

- A database of non-redundant reference sequences standards, including genomic DNA contigs, mRNAs, and proteins for known genes. Multiple collaborations, both within NCBI and with external groups, supports data-gathering efforts.

Nucleotide Database

STS database:

- A database of sequence tagged sites, or short sequences that are operationally unique in the genome.

UniSTS:

- A unified, non-redundant view of sequence tagged sites (STSs).

UniGene:

- A collection of ESTs and full-length mRNA sequences organized into clusters, each representing a unique known or putative human gene annotated with mapping and expression information and cross-references to other sources.

  UniGene computationally identifies transcripts from the same locus; analyzes expression by tissue, age, and health status; and reports related proteins (protEST) and clone resources.

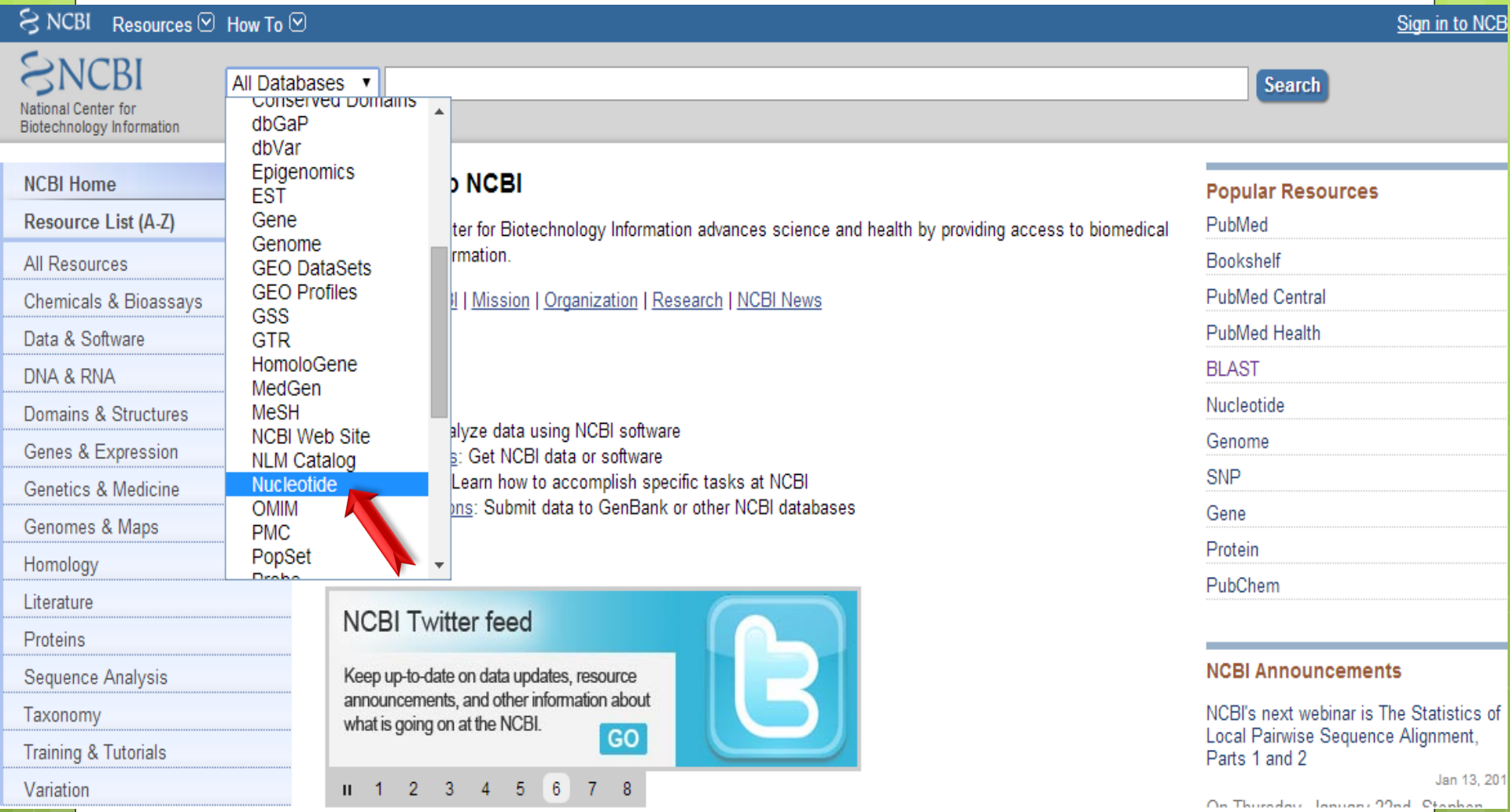
Single Nucleotide Polymorphism (SNP) database

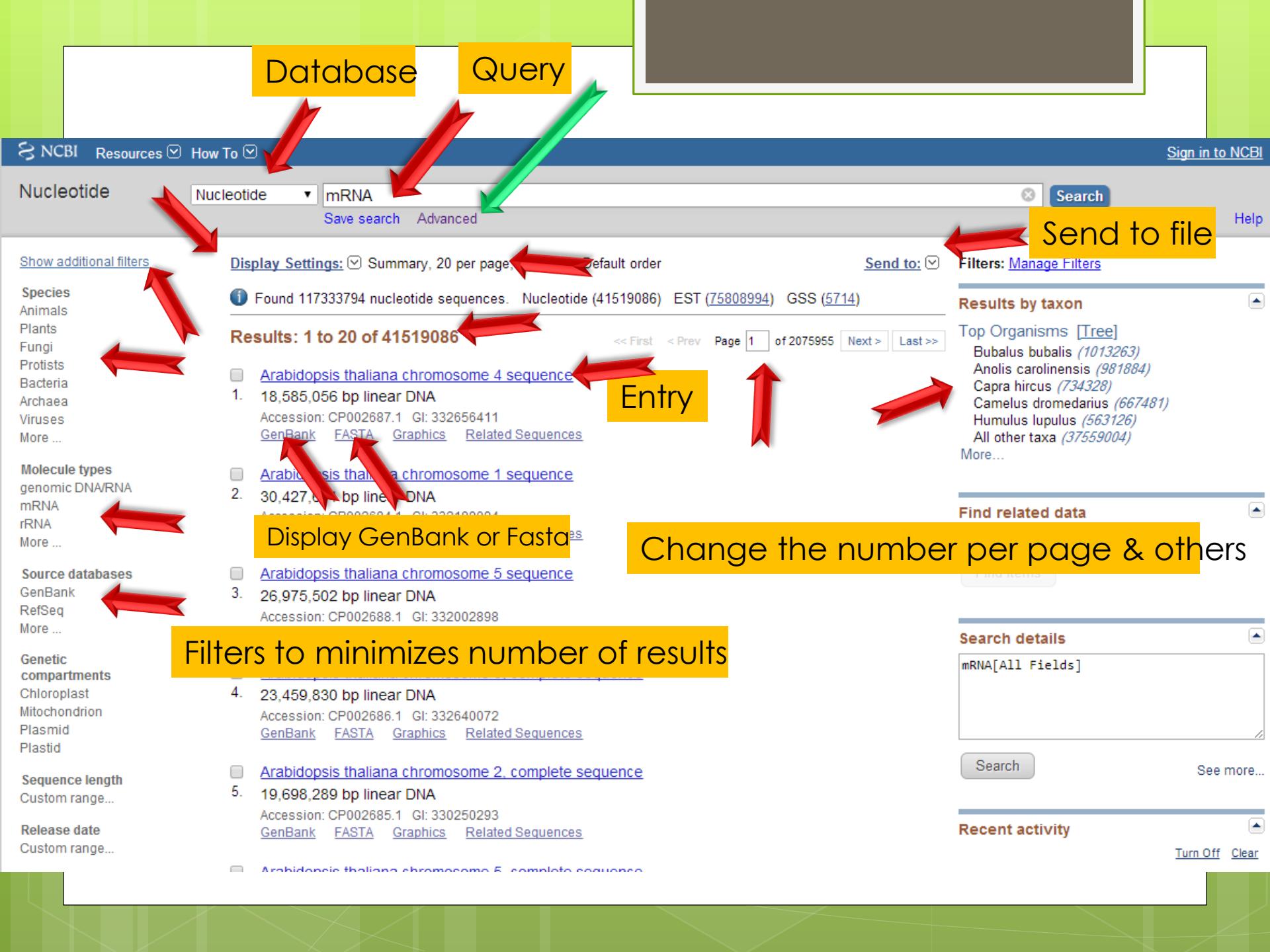
- The **SNP Database** (also known as dbSNP) is an **archive** for genetic variation within and across different species developed and hosted by NCBI in collaboration with the National Human Genome Research Institute (NHGRI).
  - *Polymorphism* in biology occurs when two or more clearly different phenotypes exist in the same population of a species: related to biodiversity, genetic variation and adaptation
- The dbSNP accepts apparently neutral polymorphisms, polymorphisms corresponding to known phenotypes, and regions of no variation.
- It was created in September 1998 to supplement GenBank (NCBI's nucleic acid and protein sequences)

Single Nucleotide Polymorphism (SNP) database

- Its goal is to act as a **single database that contains all identified genetic variation**, which can be used to investigate a wide variety of genetically based natural phenomenon. Specifically, access to the molecular variation cataloged within dbSNP aids basic research such as **physical mapping, population genetics, investigations into evolutionary relationships, as well as being able to quickly and easily quantify the amount of variation at a given site of interest**.
- Applied research, genetic engineering, drug discovery, etc.

Nucleotide Database




Database
Query
Send to file
Entry
Display GenBank or Fasta
Change the number per page & others
Filters to minimizes number of results
NCBI Resources How To
Sign in to NCBI
Nucleotide
Nucleotide
mRNA
Search
Save search
Advanced
Help
Show additional filters
Display Settings: Summary, 20 per page, Default order
Send to:
Filters: Manage Filters
Species
Animals
Plants
Fungi
Protists
Bacteria
Archaea
Viruses
More ...
Found 117333794 nucleotide sequences. Nucleotide (41519086) EST (75808994) GSS (5714)
Results: 1 to 20 of 41519086
<< First < Prev Page 1 of 2075955 Next > Last >>
Arabidopsis thaliana chromosome 4 sequence
1. 18,585,056 bp linear DNA
Accession: CP002687.1 GI: 332656411
GenBank FASTA Graphics Related Sequences
Arabidopsis thaliana chromosome 1 sequence
2. 30,427,
bp line
DNA
Molecule types
genomic DNA/RNA
mRNA
rRNA
More ...
Source databases
GenBank
RefSeq
More ...
Arabidopsis thaliana chromosome 5 sequence
3. 26,975,502 bp linear DNA
Accession: CP002688.1 GI: 332002898
Genetic compartments
Chloroplast
Mitochondrion
Plasmid
Plastid
4. 23,459,830 bp linear DNA
Accession: CP002686.1 GI: 332640072
GenBank FASTA Graphics Related Sequences
Arabidopsis thaliana chromosome 2, complete sequence
5. 19,698,289 bp linear DNA
Accession: CP002685.1 GI: 330250293
GenBank FASTA Graphics Related Sequences
Sequence length
Custom range...
Release date
Custom range...
Results by taxon
Top Organisms [Tree]
Bubalus bubalis (1013263)
Anolis carolinensis (981884)
Capra hircus (734328)
Camelus dromedarius (667481)
Humulus lupulus (563126)
All other taxa (37559004)
More...
Find related data
Search details
mRNA[All Fields]
Search
See more...
Recent activity
Turn Off Clear

NCBI abbreviations for nucleotide

Category	Description
NC_########	DNA
NM_########	mRNA
XM_########	Predicted mRNA

NCBI abbreviations for protein

Category	Description
NP_########	Protein
XP_########	Predicted Protein

Predicted & Known Proteins

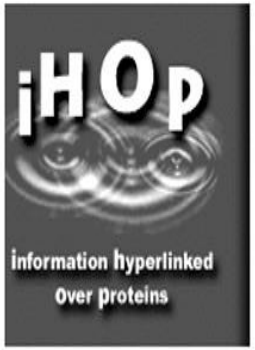
- **Predicted** proteins are predicted by gene identification tools (software) and they share sequence similarity to well-characterized proteins.

- **Known** proteins are experimentally proven to exist and are linked to a known gene.

Information Hyperlinked Over Proteins (iHOP)

- iHOP provides fast, accurate, comprehensive, and up-to-date summary information on more than 80, 000 biological molecules by automatically extracting key sentences from millions of PubMed documents.
- **It** is an online text-mining service that provides a gene-guided network to access PubMed abstracts.
- iHOP (Information Hyperlinked over Proteins) allows researchers to explore a network of gene and protein interactions **based on published scientific literature**. For each gene search, iHOP reports sentences from abstracts associating it with other genes, links out to full abstracts, and reports **experimental evidence** for the interactions, if available. You can also select sentences to create and visualize your own gene model

http://www.ihop-net.org/UniPub/iHOP/



Symbol	Name	Synonym/ DB-reference	Organism	Results
[redacted]		Life cycles of successful genes *Trends in Genetics*		
WWC1	WW and C2 domain containing 1	**KIBRA**	Homo sapiens	
Wwc1	WW, C2 and coiled-coil domain containing 1	**KIBRA**	Mus musculus	

Figure 2.7a Practical Bioinformatics (© Garland Science 2013)

Symbol	Name	Synonyms	Organism
WWC1	WW and C2 domain containing 1	FLJ10865, FLJ23369, HBeAg-binding protein 3, HBEBP3, HBEBP36, KIAA0869, KIBRA, Kidney and brain protein, Protein KIBRA, WW domain-containing protein 1	Homo sapiens

WikiGenes	edit this page new
UniProt	Q8IX03, Q7Z4G8, Q6MZX4
IntAct	Q8IX03
PDB Structure	2Z0U
OMIM	610533
NCBI Gene	23286
NCBI RefSeq	NP_001155133, NP_001155134
NCBI RefSeq	NM_001161662, NM_001161661
NCBI UniGene	23286
NCBI Accession	DR001014, AK296323

more than 2,800 organisms, 110,000 genes, 23.4 million sentences.
...always up to date – every day.

Homologues of WWC1 ...

Interaction information for WWC1 ...

Most recent information for WWC1 ...

Enhanced PubMed/Google query ...

WARNING: Please keep in mind that gene detection is done automatically and can exhibit a certain error. Read more about synonym ambiguity and the iHOP confidence value ☆☆☆.

Find in this Page

Sentences in this view contain definitions for WWC1 - Definitions are available whenever you see this symbol - Read more.

For a summary overview of the information in this page click here. new

Show all

Order by relevance

We also found that KIBRA☆-DLC1☆ interaction is mandatory for the recruitment and transactivation **functions** of ER☆ or DLC1☆ to the target chromatin. [2006]

Finally we found that KIBRA☆ **interacts** with histone H3 via its glutamic acid [?]-rich region and that such interaction might play a mechanistic role in conferring an optimal ER☆ transactivation function as well as the proliferation of ligand-stimulated breast cancer cells. [2006]

KIAA0513☆ **interacts** with KIBRA☆, HAX-1☆, and INTS4☆, which also interact with proteins involved in neuroplasticity, apoptosis, and cytoskeletal regulation. [2006]

Figure 2.7b Practical Bioinformatics (© Garland Science 2013)

Conclusion

- Many databases and resources are available, some as websites and some (such as R packages or NCBI E-Utilities) via programing languages.
- There is no single correct way to find information; many approaches are possible.
- Resources are closely interrelated, providing links between the databases.
- Bioinformatics databases are evolving extremely rapidly.
- Each January, the first issue of the journal *Nucleic Acids Research* includes nearly 100 brief articles on databases.