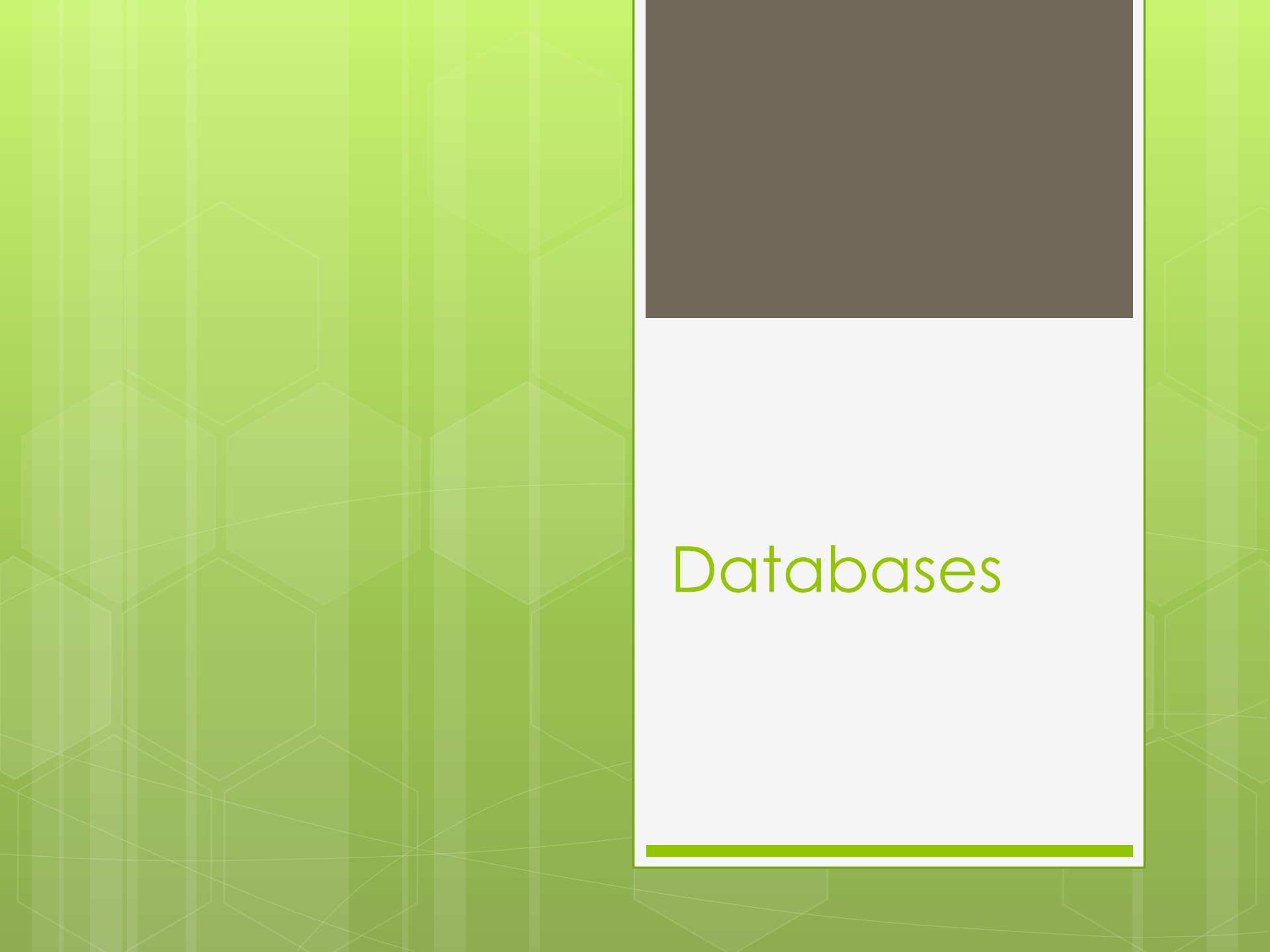




Databases

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 strung 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

ACCACTTTTCACAATCTGCTAGCAAAGGTTATGCAGCGCGTGAACATGATCATGGCAGAATCACCAGGCGCT
CATCACCATCTGCCTTTTATAGGATATCTACTCAGTGCTGAATGTACAGTTTTTCTTGATCATGAAAACGCC
AACAAAATTCTGAATCGGCCAAAGAGGTATAATTTCAGGTAAATTGGAAGAGTTTGTTC AAGGGAACCTTG
AGAGAGAATGTATGGAAGAAAAGTGTAGTTTTTGAAGAAGCACGAGAAGTTTTTGA AAAACACTGAAAGAAC
AACTGAATTTTGG AAGCAGTATGTTGATGGAGATCAGTGTGAGTCCAATCCATGTTTAAATGGCGGCAGT
TGCAAGGATGACATTAATTCCTATGAATGTTGGTGTCCCTTTGGATTTGAAGGAAAGA AACTGTGAATTAG
ATGTAACATGTAACATTAAGAATGGCAGATGCGAGCAGTTTTTGTAAAAATAGTGCTGATAACAAGGTGGT
TTGCTCCTGTACTGAGGGATATCGACTTGCAGAAAACCGAAGTCCCTGTGAACCAGCAGTGCCATTTCCA
TGTGGAAGAGTTTTCTGTTTCACAAACTTCTAAGCTCACCCGTGCTGAGACTGTTTTTCTGATGTGGACT
ATGTAAATTCTACTGAAGCTGAAACCATTTTTGGATAACATCACTCAAAGCACCC AATCATTTAATGACTT
CACTCGGGTTGTTGGTGGGAGAAGATGCCAAACCAGGTCAATTCCCTTGGCAGGTTGTTTTGAATGGTAAA
GTTGATGCATTCTGTGGAGGCTCTATCGTTAATGAAAAATGGATTGTA AACTGCTGCCCACTGTGTTGAAA
CTGGTGT TAAAAATTACAGTTGTCTGCAGGTGAACATAATATTGAGGAGACAGAACATACAGAGCAAAAAGCG
AAATGTGATTTCGAATTATTCCTCACCAACA ACTACAATGCAGCTATTAATAAGTACAACCATGACATTGCC
CTTCTGGA AACTGGACGAACCCTTAGTGCTAAACAGCTACGTTACACCTATTTGCATTGCTGACAAGGAAT
ACACGAACATCTTCCTCAAATTTGGATCTGGCTATGTAAGTGGCTGGGGGAAGAGTCTTCCACAAAGGGAG
ATCAGCTTTAGTTCTTCAGTACCTTAGAGTTCCACTTGTTGACCGAGCCACATGTCTTCGATCTACAAAG
TTCACCATCTATAACAACATGTTCTGTGCTGGCTTCCATGAAGGAGGTTAGAGATTCAATGTCAAGGAGATA
GTGGGGGACCCCATGTTACTGAAGTGG AAGGGACCAGTTTCTTAACTGGAATTATTAGTCTGGGGTGAAGA
GTGTGCAATGAAAGGCCAAATATGGAATATATACCAAGGTATCCCGGTATGTCAACTGGATTAAAGGAAAAA
ACAAAGCTCACTTAATGAAAGATGGATTTCCAAGGTTAATTCATTGGAATTGAAAATTAAACAGGGCCTCT
CACTAACTAATCACTTTCCCATCTTTTGT TAGATTTGAATATATACATTCTATGATCATTGCTTTTTTCTC
TTTACAGGGGAGAAATTT CATATTTTACCTGAGCAAATTGATTAGAAAATGGAACC ACTAGAGGAATATAA
TGTGTTAGGAAATTACAGTCATTTCTAAGGGGCCAGCCCTTGACAAAATTGTGAAGTTAAATTCCTCCACT
CTGTCCATCAGATACTATGGTCTTCCACTATGGCAACTA ACTCACTCAATTTTTCCCTCCTTAGCAGCATT
CCATCTTCCC GATTTCTTTGCTTCTCCAACCAAAACATCAATGTTTATTAGTTCTGTATACAGTACAGG
ATCTTTGGTCTACTCTATCACAAGGCCAGTACCACACTCATGAAGAAAGAACACAGGAGTAGCTGAGAGG
CTAA AACTCATCAAAAACACTACTCCTTTTCTCTACCCCTATTCCTCAATCTTTTACCTTTTCCAATCC
CAATCCCCCAAATCAGTTTTTCTCTTTCTTACTCCCTCTCTCCCTTTTACCCCTCCATGGTCGT TAAAGGAG
AGATGGGGGAGCATCATCTGTTATACTTCTGTACACAGTTATACATGTCTATCAAACCCAGACTTGGCTTC
CGTAGTGGAGACTTGCTTTTTCAGAACATAGGGATGAAGTGAAGGTGCCTGAAAAGTTTGGGGGAAAAGTTT
CTTTCAGAGAGTTAAGTTATTTTATATATATATAATATATATAATAAATAATAATAATAATAATAATA
TAGTGTGTGTGTATGCGTGTGTGTAGACACACACGCATACACACATATAATGGAAGCAATAAGCCATTCT
AAGAGCTTGTATGGTTATGGAGGTCTGACTAGGCATGATTTACGGAAGGCAAGATTGGCATATCATTGTA
ACTAAAAAAGCTGACATTGACCCAGACATATTG TACTCTTTCTAAAAATAATAATAATAATAATGCTAACAGA
AAGAAGAGAACC GTTCGTTTGCAATCTACAGCTAGTAGAGACTTTGAGGGAAGAATTCAACAGTGTGTCTT
CAGCAGTGTTCAGAGCCAAGCAAGAAGTTGAAGTTGCCTAGACAGAGGACATAAGTATCATGTCTCCTT
TAACTAGCATACCCG AAGTGGAGAAGGGTGCAGCAGGCTCAAAGGCATAAGTCAATTCCAATCAGCCAAC
TAAGTTGTCCTTTTCTG GTTTTCGTGTTCCACCATGGAACATTTTGATTATAGTTAATCCTTCTATCTTGAA
TCTTCTAGAGAGTTGCTGACCAACTGACGTATGTTTCCCTTTGTGAATTAATAAACTGGTGTCTG GTTC
AT

Figure 1.3 Practical Bioinformatics (© Garland Science 2013)

An example of DNA sequence

ACCACTTTTCAACAATCTGCTAGCAAAGGTTATGTCAGCGCGTGAACATGATCATGGCAGAATCACCAGGCCT
CATCACCATCTGCCTTTTAGGATATCTACTCACTGCTGAATGTACAGTTTTTCTTGATCATGAAAACGCC
AACAAAATTTCTGAATCGGCCAAAGAGGTATAATTCAAGGTAAATTGGAAGAGTTTGTTC AAGGGAACCTTG
AGAGAGAATGTATGGAAGAAAAGTGTAGTTTTGAAGAAGCAGAGAAGTTTTTGA AAAACACTGAAAGAAC
AACTGAATTTTGG AAGCAGTATGTTGATGGAGATCAGTGTGAGTCCAATCCATGTTTAAATGGCGGCAGT
TGCAAGGATGACATTAATTCCCTATGAATGTTGGTGTCCCTTTGGATTTGAAGGAAAGA AACTGTGAATTAG
ATGTAACATGTAACATTAAGAATGGCAGATGCGAGCAGTTTTTGTAAAAATAGTGCTGATAACAAGGTGGT
TTGCTCCTGTACTGAGGGATATCGACTTGCAGAAAACCAGAAAGTCCTGTGAACCAGCAGTGCCATTTCCA
TGTGGAAGAGTTTCTGTTTCACAAACTTCTAAGCTCACCCGTGCTTGAGACTGTTTTTCTTGATGTGGACT
ATGTAAATTCTACTGAAGCTGAAACCATTTTTGGATAACATCACTCAAAGCACCCCAATCATTTAATGACTT
CACTCGGGTTGTTGGTGGAGAAGATGCCAAACCAGGTCAATTCCTTTGGCAGGTTGTTTTTGAATGGTAAA
GTTGATGCATTCTGTGGAGGCTCTATCGTTAATGAAAAATGGATTGTAAGTGTGCTGCCCACTGTGTTGAAA
CTGGTGT TAAAAATTACAGTTGTCTGCAGGTGAACATAATATTGAGGAGACAGAACATACAGAGCAAAAGCG
AAATGTGATTTCGAATTATTCCCTCACCACAAC TACAATGCAGCTATTAATAAGTACAACCATGACATTGCC
CTTCTGGAAC TGGACGAACCCCTTAGTGCTAAACAGCTACGTTACACCTATTTG CATTGCTGACAAGGAAT
ACACGAACATCTTCCTCAAATTTGGATCTGGCTATGTAAGTGGCTGGGGGAAGAGTCTTCCACAAAGGGAG
ATCAGCTTTAGTTTCTTCAGTACCTTAGAGTTCCACTTGTGTGACCGAGCCACATGTCTTCGATCTACAAAG
TTCAACCATCTATAACAACATGTTCTGTGCTGGCTTCCATGAAGGAGGTAGAGATT CATGTCAAGGAGATA
GTGGGGGACCCCACTTACTGAAGTGG AAGGGACCAGTTTCTTAACTGGAATTATTAGCTGGGGTGAAGA
GTGTGCAATGAAAGGCAAAATATGGAATATATACCAAGGTATCCCGGTATGTCAACTGGATT AAGGAAAAA
ACAAAGCTCACTTAAAGATGGATTTC CAAGGTTAATTCAATTGGAATTGAAAATTAACAGGGCCTCT
CACTAACTAATCTTTCCCATCTTTTGT TAGATTGGAATATATACATTCTATGATCATTGCTTTTTTCTC
TTTACAGGGGAGAAATTT CATATTTTACCTGAGCAAATGATTAGAAAAATGGAACCACTAGAGGAATATA
TGTGTTAGGAAATTACAGTCATTTCTAAGGGCCCAGCCCTTGACAAAATTGTGAAGTTAAATTCTCCACT
CTGTCCATCAGATACTATGGTTCTCCACTATGGCAACTAACTCACTCAATTTTCCCTCCTTAGCAGCATT
CCATCTTCCCGATCTTCTTTGCTTCTCCAACCAAAACATCAATGTTTATTAGTTCTGTATACAGTACAGG
ATCTTTGGTCTACTCTATCACAAGGCCAGTACCACACTCATGAAGAAAGAACACAGGAGTAGCTGAGAGG
CTAAAACTACATCAAAAACACTACTCCTTTTCTCTACCTATTTCTCAATCTTTTACCTTTTCCAAATCC
CAATCCCCAAATCAGTTTTTCTCTTTCTTACTCCCTCTCTCCCTTTTACCCTCCATGGTCGTTAAAGGAG
AGATGGGGGAGCATCATTCTGTTATACTTCTGTACACAGTTATACATGTCTATCAAACCCAGACTTGCTTC
CGTAGTGGAGACTTGCTTTTTCAGAACATAGGGATGAAGTAAGGTGCCTGAAAAGTTTGGGGGAAAAGTTT
CTTTCAGAGAGTTAAGTTATTTTATATATATATAATATATATAATAATATAACAATATAAATATA
TAGTGTGTGTATGCGTGTGTGTATAGACACACGCATACACACATATAATGGAAGCAATAAGCCATTCT
AAGAGCTTGTATGGTTAGGTTCTGACTAGGCATGATTTACGGAAGGCAAGATTGGCATATCATTGTA
ACTAAAAAAGCTGACATTGACCCAGACATATTGTACTCTTTCTAAAAAATAATAATAATAATGCTAACAGA
AAGAAGAGAACCGTTTCGTTTGCAATCTACAGCTAGTAGAGACTTTGAGGAAGAATTCAACAGTGTGTCTT
CAGCAGTGTTT CAGAGCCAAGCAAGAAGTTGAAGTTGCCTAGACGAGGACATAAGTATCATGTCTCCCT
TAAGTAGCATACCCCGAAGTGGAGAAGGTTGCAGCAGGCTCAAAGGCATAAGTCAATTCAGCCCAAT
TAAGTTGTCCTTTTCTGTTTTCTGTTTCACCATGGAACATTTTGATTATAGTTAATCCTTCTATCTTGAA
TCTTCTAGAGAGTTGCTGACCAACTGACGTATGTTTCCCTTTGTGAATTAATAAACTGGTGTCTGTTTC
AT

Figure 1.4 Practical Bioinformatics (© Garland Science 2013)

An example of DNA sequence

ACCACTTTTCACAATCTGCTAGCAAAGGTT
ATG CAG CGC GTG AAC ATG ATC ATG GCA GAA TCA CCA GGC CTC ATC ACC ATC TGC CTT TTA GGA TAT CTA CTC AGT
GCT GAA TGT ACA GTT TTT CTT GAT CAT GAA AAC GCC AAC AAA ATT CTG AAT CGG CCA AAG AGG TAT AAT TCA GGT
AAA TTG GAA GAG TTT GTT CAA GGG AAC CTT GAG AGA GAA TGT ATG GAA GAA AAG TGT AGT TTT GAA GAA GCA CGA
GAA GTT TTT GAA AAC ACT GAA AGA ACA ACT GAA TTT TGG AAG CAG TAT GTT GAT GGA GAT CAG TGT GAG TCC AAT
CCA TGT TTA AAT GGC GGC AGT TGC AAG GAT GAC ATT AAT TCC TAT GAA TGT TGG TGT CCC TTT GGA TTT GAA GGA
AAG AAC TGT GAA TTA GAT GTA ACA TGT AAC ATT AAG AAT GGC AGA TGC GAG CAG TTT TGT AAA AAT AGT GCT GAT
AAC AAG GTG GTT TGC TCC TGT ACT GAG GGA TAT CGA CTT GCA GAA AAC CAG AAG TCC TGT GAA CCA GCA GTG CCA
TTT CCA TGT GGA AGA GTT TCT GTT TCA CAA ACT TCT AAG CTC ACC CGT GCT GAG ACT GTT TTT CCT GAT GTG GAC
TAT GTA AAT TCT ACT GAA GCT GAA ACC ATT TTG GAT AAC ATC ACT CAA AGC ACC CAA TCA TTT AAT GAC TTC ACT
CGG GTT GTT GGT GGA GAA GAT GCC AAA CCA GGT CAA TTC CCT TGG CAG GTT GTT TTG AAT GGT AAA GTT GAT GCA
TTC TGT GGA GGC TCT ATC GTT AAT GAA AAA TGG ATT GTA ACT GCT GCC CAC TGT GTT GAA ACT GGT GTT AAA ATT
ACA GTT GTC GCA GGT GAA CAT AAT ATT GAG GAG ACA GAA CAT ACA GAG CAA AAG CGA AAT GTG ATT CGA ATT ATT
CCT CAC CAC AAC TAC AAT GCA GCT ATT AAT AAG TAC AAC CAT GAC ATT GCC CTT CTG GAA CTG GAC GAA CCC TTA
GTG CTA AAC AGC TAC GTT ACA CCT ATT TGC ATT GCT GAC AAG GAA TAC ACG AAC ATC TTC CTC AAA TTT GGA TCT
GGC TAT GTA AGT GGC TGG GGA AGA GTC TTC CAC AAA GGG AGA TCA GCT TTA GTT CTT CAG TAC CTT AGA GTT CCA
CTT GTT GAC CGA GCC ACA TGT CTT CGA TCT ACA AAG TTC ACC ATC TAT AAC AAC ATG TTC TGT GCT GGC TTC CAT
GAA GGA GGT AGA GAT TCA TGT CAA GGA GAT AGT GGC GGA CCC CAT GTT ACT GAA GTG GAA GGG ACC AGT TTC TTA
ACT GGA ATT ATT AGC TGG GGT GAA GAG TGT GCA ATG AAA GGC AAA TAT GGA ATA TAT ACC AAG GTA TCC CGG TAT
GCT AAC TGG ATT AAG GAA AAA ACA AAG CTC ACT TAA
TGAAAGATGGATTTCGAAGGTTAATTCATTGGAATTGAAAATTAGGAGGCGCTCTCACTAATACTACTTTCCCATCTTTTGTAGATTGAAATATATACA
TTCTATGATCATTGCTTTTCTCTTTACAGGGGAGAATTTTCATATTTTACCTGAGCAAATTGATTAGAAAATGGAACCACTAGAGGAATAAATGTGTTAGG
AAATTACAGTCATTTCTAAGGGCCCCAGCCCTTGACAAAATTGTGAAGTTAAATTCTCCACTCTGTCCATCAGATACTATGGTTCTCCACTAGGCACTAAC
TCACTCAATTTTCCCTCCTTAGCAGCATTCCATCTTCCCGATCTTCTTTGCTTCTCCAACCAAAACATCAATGTTTATTAGTTCTGTATACAGTACAGGATC
TTTGGTCTACTCTATCACAAGGCCAGTACCACACTCATGAAGAAAGAACACAGGAGTAGCTGAGAGGCTAAAACACTCATCAAAAACACTACTCCTTTTCTCT
ACCCTATTTCCTCAATCTTTTACCTTTTCCAAATCCCAATCCCCAAATCAGTTTTTCTCTTTCTTACTCCCTCTCTCCCTTTTACCCTCCATGGTCGTTAAAG
GAGAGATGGGGAGCATCATTCTGTTATACTTCTGTACACAGTTATACATGTCTATCAAACCCAGACTTGCTTCCGTAGTGGAGACTTGCTTTTTCAGAACATA
GGGATGAAGTAAGGTGCCTGAAAAGTTTGGGGGAAAAGTTTCTTTTCAGAGAGTTAAGTTATTTTATATATATAATATATATATAAAATATATAATATACAAT
ATAAATATATAGTGTGTGTGTATGCGTGTGTGTAGACACACACGCATACACACATATAATGGAAGCAATAAGCCATTCTAAGAGCTTGTATGGTTATGGAGG
TCTGACTAGGCATGATTTACGAAGGCAAGATTGGCATATCATTGTAATAAAAAAGCTGACATTGACCCAGACATATTGTACTCTTTCTAAAAATAATAAT
AATAATGCTAACAGAAAGAAGAGAACCGTTTCGTTTGCAATCTACAGCTAGTAGAGACTTTGAGGAAGAATTCAACAGTGTGTCTTCAGCAGTGTTCAGAGCC
AAGCAAGAAGTTGAAGTTGCCTAGACCAGAGGACATAAGTATCATGTCTCTTTTAACTAGCATAACCCGAAGTGGAGAAGGGTGCAGCAGGCTCAAAGGCAT
AAGTCATTCCAATCAGCCAACCTAAGTTGTCTTTTCTGGTTCCTGTTTCACCATGGAACATTTTGATTATAGTTAATCCTTCTATCTTGAATCTTCTAGAGA
GTTGCTGACCAACTGACGTATGTTTCCCTTTGTGAATTAATAAATGGTGTCTGTTTCAT

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

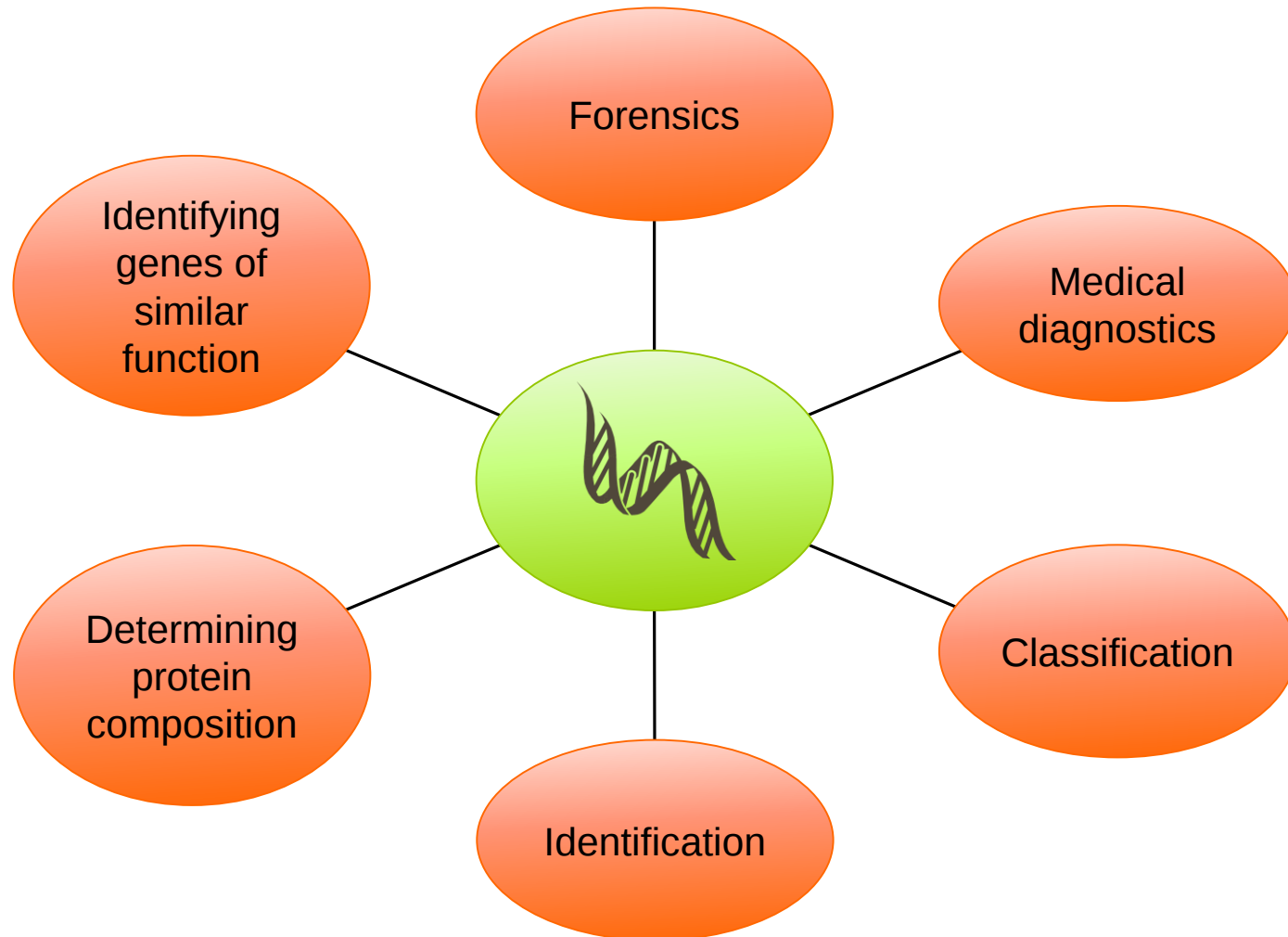
Uses of sequence analysis

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

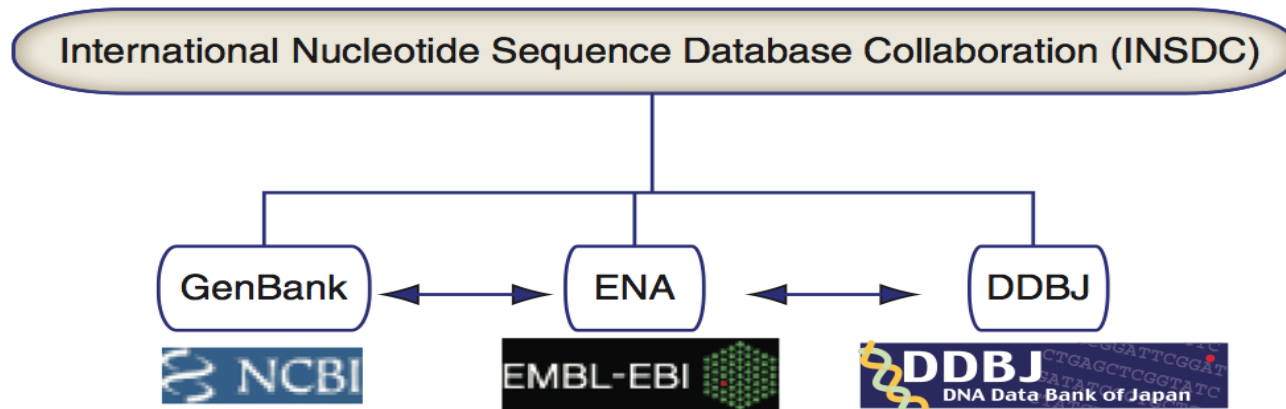
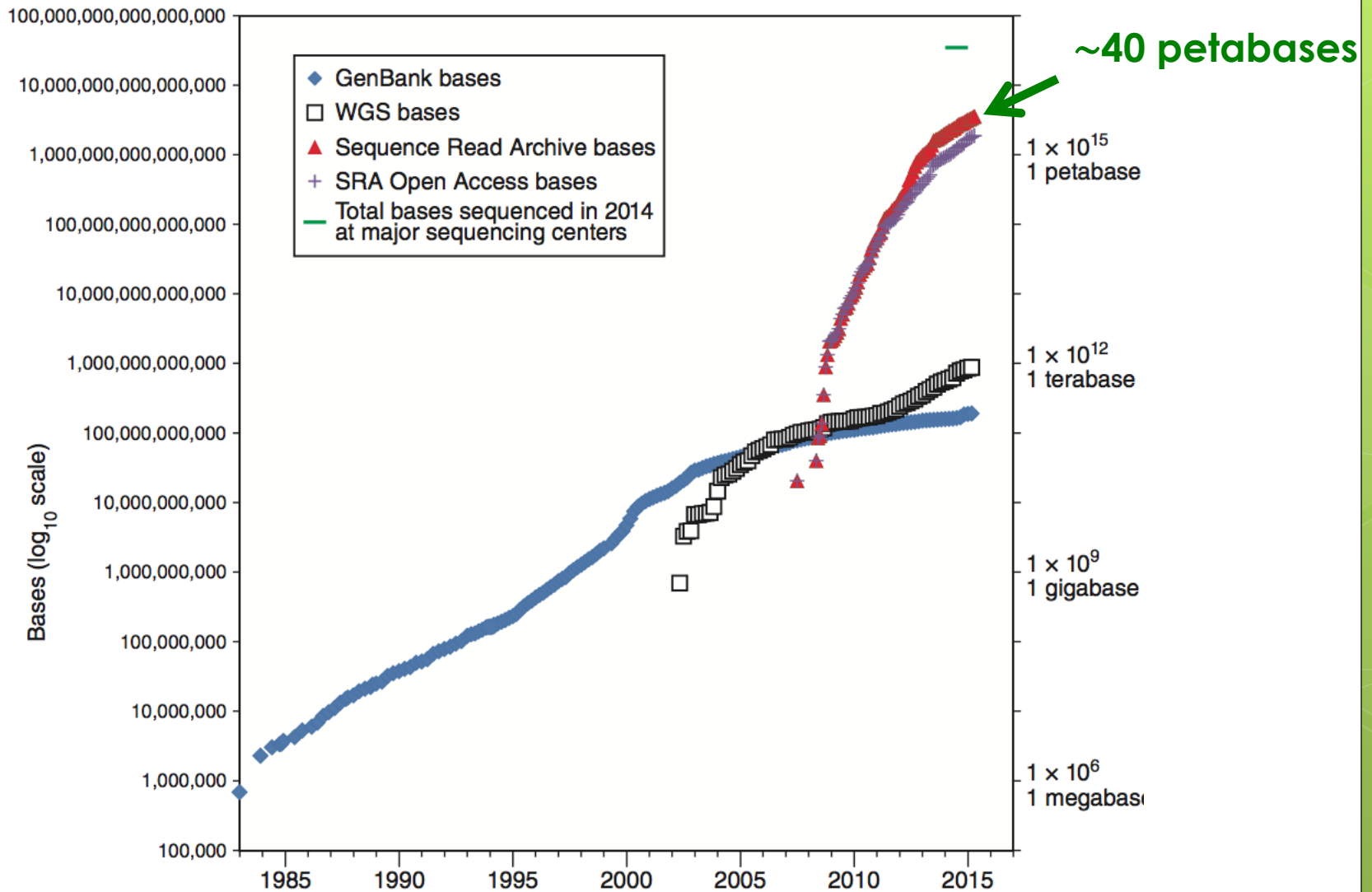


FIGURE 2.1 The nucleotide collections of GenBank at NCBI, EMBL-Bank at the European Bioinformatics Institute, and DDBJ at the DNA Data Bank of Japan are all coordinated by the International Nucleotide Sequence Database Collaboration (INSDC).

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

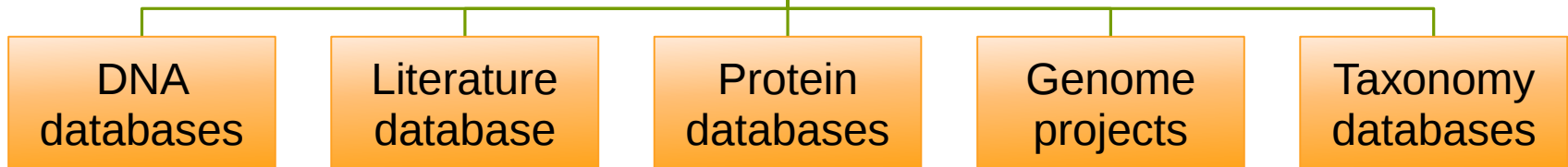
- **Entrez:** integrates scientific literature, DNA, and protein sequence databases, 3D protein structure data, population study datasets, and assemblies of complete genomes into a tightly coupled system
- **PubMed:** search service from the National Library of Medicine (NLM) that provides access to 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

Simple 'one box' search.



Entrez Main Screen

 NCBI

 *Entrez, The Life Sciences Search Engine*

HOME | SEARCH | SITE MAP

PubMed | All Databases | Human Genome | GenBank | Map Viewer | BLAST

Search across databases [Help](#)

Welcome to the Entrez cross-database search page

 PubMed: biomedical literature citations and abstracts 	 Books: online books 
 PubMed Central: free, full text journal articles 	 OMIM: online Mendelian Inheritance in Man 
 Site Search: NCBI web and FTP sites 	 OMIA: online Mendelian Inheritance in Animals 

 CoreNucleotide: Core subset of nucleotide sequence records 	 dbGaP: genotype and phenotype 
 EST: Expressed Sequence Tag records 	 UniGene: gene-oriented clusters of transcript sequences 
 GSS: Genome Survey Sequence records 	 CDD: conserved protein domain database 
 Protein: sequence database 	 3D Domains: domains from Entrez Structure 
 Genome: whole genome sequences 	 UniSTS: markers and mapping data 
 Structure: three-dimensional macromolecular structures 	 PopSet: population study data sets 
 Taxonomy: organisms in GenBank 	 GEO Profiles: expression and molecular abundance profiles 
 SNP: single nucleotide polymorphism 	 GEO DataSets: experimental sets of GEO data 
 Gene: gene-centered information 	 Cancer Chromosomes: cytogenetic databases 
 HomoloGene: eukaryotic homology groups 	 PubChem BioAssay: bioactivity screens of chemical substances 
 GENSAT: gene expression atlas of mouse central nervous system 	 PubChem Compound: unique small molecule chemical structures 
 Probe: sequence-specific reagents 	 PubChem Substance: deposited chemical substance records 
 Genome Project: genome project information 	 Protein Clusters: a collection of related protein sequences 

 Journals: detailed information about the journals indexed in PubMed and other Entrez databases 	 MeSH: detailed information about NLM's controlled vocabulary 
 NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections 	

⚠ 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](#)

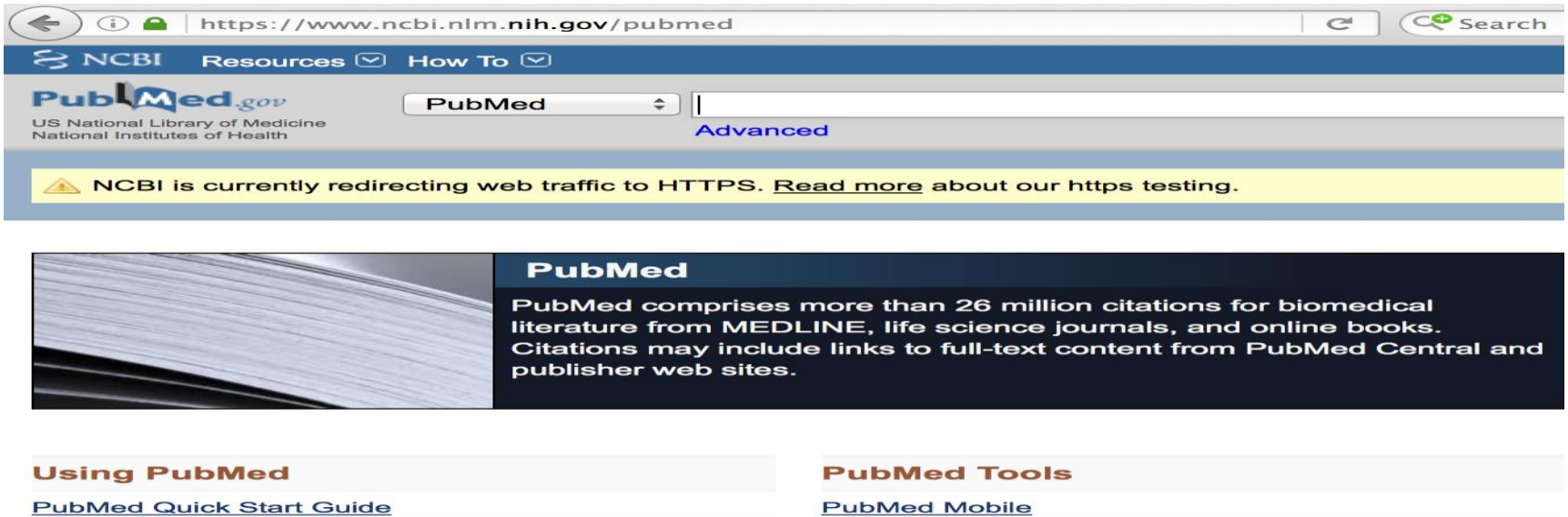
[PubChem](#)

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.



The screenshot shows the NCBI PubMed website. The browser address bar displays <https://www.ncbi.nlm.nih.gov/pubmed>. The NCBI logo and navigation links (Resources, How To) are visible. The PubMed logo and text "US National Library of Medicine National Institutes of Health" are present. A search bar with the text "PubMed" and a dropdown menu is shown, along with a link to "Advanced" search. A yellow banner message states: "NCBI is currently redirecting web traffic to HTTPS. [Read more](#) about our https testing." Below this, a section titled "PubMed" features an image of a book and text describing the database's scope: "PubMed comprises more than 26 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites." At the bottom, there are two sections: "Using PubMed" with a link to the "PubMed Quick Start Guide", and "PubMed Tools" with a link to "PubMed Mobile".

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

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

Advanced

NCBI is currently redirecting web traffic to HTTPS. [Read more](#) about our https testing.

PubMed

PubMed comprises more than 26 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

Using PubMed

[PubMed Quick Start Guide](#)

PubMed Tools

[PubMed Mobile](#)

Format: Abstract ▾

Send to

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 I¹, Alshahrani M², Almutairi T³.

+ 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

 **NCBI** [Resources](#)  [How To](#) 

 **PubMed**  [Advanced](#) [Search](#) [Help](#)



PubMed

PubMed comprises more than 25 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

PubMed Commons



Featured comment - Jan 22
A Messori highlights literature on frequently documented drug-drug interactions in patients with cancer. [1.usa.gov/1k5AsDd](https://pubmed.ncbi.nlm.nih.gov/1k5AsDd)

Using PubMed

- [PubMed Quick Start Guide](#)
- [Full Text Articles](#)
- [PubMed FAQs](#)
- [PubMed Tutorials](#)
- [New and Noteworthy](#) 

PubMed Tools

- [PubMed Mobile](#)
- [Single Citation Matcher](#)
- [Batch Citation Matcher](#)
- [Clinical Queries](#)
- [Topic-Specific Queries](#)

More Resources

- [MeSH Database](#)
- [Journals in NCBI Databases](#)
- [Clinical Trials](#)
- [E-Utilities \(API\)](#)
- [LinkOut](#)

You are here: [NCBI](#) > [Literature](#) > [PubMed](#)

[Write to the Help Desk](#)

GETTING STARTED

- [NCBI Education](#)
- [NCBI Help Manual](#)
- [NCBI Handbook](#)
- [Training & Tutorials](#)
- [Submit Data](#)

RESOURCES

- [Chemicals & Bioassays](#)
- [Data & Software](#)
- [DNA & RNA](#)
- [Domains & Structures](#)
- [Genes & Expression](#)
- [Genetics & Medicine](#)
- [Genomes & Maps](#)
- [Homology](#)
- [Literature](#)
- [Proteins](#)
- [Sequence Analysis](#)
- [Taxonomy](#)
- [Variation](#)

POPULAR

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

FEATURED

- [Genetic Testing Registry](#)
- [PubMed Health](#)
- [GenBank](#)
- [Reference Sequences](#)
- [Gene Expression Omnibus](#)
- [Map Viewer](#)
- [Human Genome](#)
- [Mouse Genome](#)
- [Influenza Virus](#)
- [Primer-BLAST](#)
- [Sequence Read Archive](#)

NCBI INFORMATION

- [About NCBI](#)
- [Research at NCBI](#)
- [NCBI News](#)
- [NCBI FTP Site](#)
- [NCBI on Facebook](#)
- [NCBI on Twitter](#)
- [NCBI on YouTube](#)

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA
[Policies and Guidelines](#) | [Contact](#)



Results of Pubmed

Article types

Clinical Trial

Review

Customize ...

Text availability

Abstract

Free full text

Full text

PubMed

Commons

Reader comments

Trending articles

Publication dates

5 years

10 years

Custom range...

Species

Humans

Other Animals

[Clear all](#)

[Show additional filters](#)

Summary ▾ 20 per page ▾ Sort by Most Recent ▾

Send to: ▾ Filters: [Manage Filters](#)

Search results

Items: 1 to 20 of 56

<< First < Prev Page 1 of 3 Next > Last >>

- ☐ [The bacterial cell cycle regulator GcrA is a \$\sigma 70\$ cofactor that drives gene expression from a subset of methylated promoters.](#)

Haakonsen DL, Yuan AH, Laub MT.
Genes Dev. 2015 Nov 1;29(21):2272-86. doi: 10.1101/gad.270660.115.
PMID: 26545812
[Similar articles](#)

- ☐ [32 Electrostatic properties of T7-like phages promoters for host bacterial and native viral RNA polymerases.](#)

Osyrov AA, Kamzolova SG.
J Biomol Struct Dyn. 2015;33 Suppl 1:19-20. doi: 10.1080/07391102.2015.1032581. No abstract available.
PMID: 26103243
[Similar articles](#)

- ☐ [Base flipping in open complex formation at bacterial promoters.](#)

3. Karpen ME, deHaseth PL.
Biomolecules. 2015 Apr 28;5(2):668-78. doi: 10.3390/biom5020668. Review.
PMID: 25927327 **Free PMC Article**

New feature

Try the new Display Settings option -
[Sort by Relevance](#)

Titles with your search terms

Bacterial genetics by flow cytometry: rapid isolation of *Salmonella typhi* [Mol Microbiol. 1996]

Regulation at complex **bacterial promoters**: how bacteria use different [Curr Opin Microbiol. 2004]

Identification of an UP element consensus sequence for **ba** [Proc Natl Acad Sci U S A. 1998]

[See more...](#)

Find related data

Database:

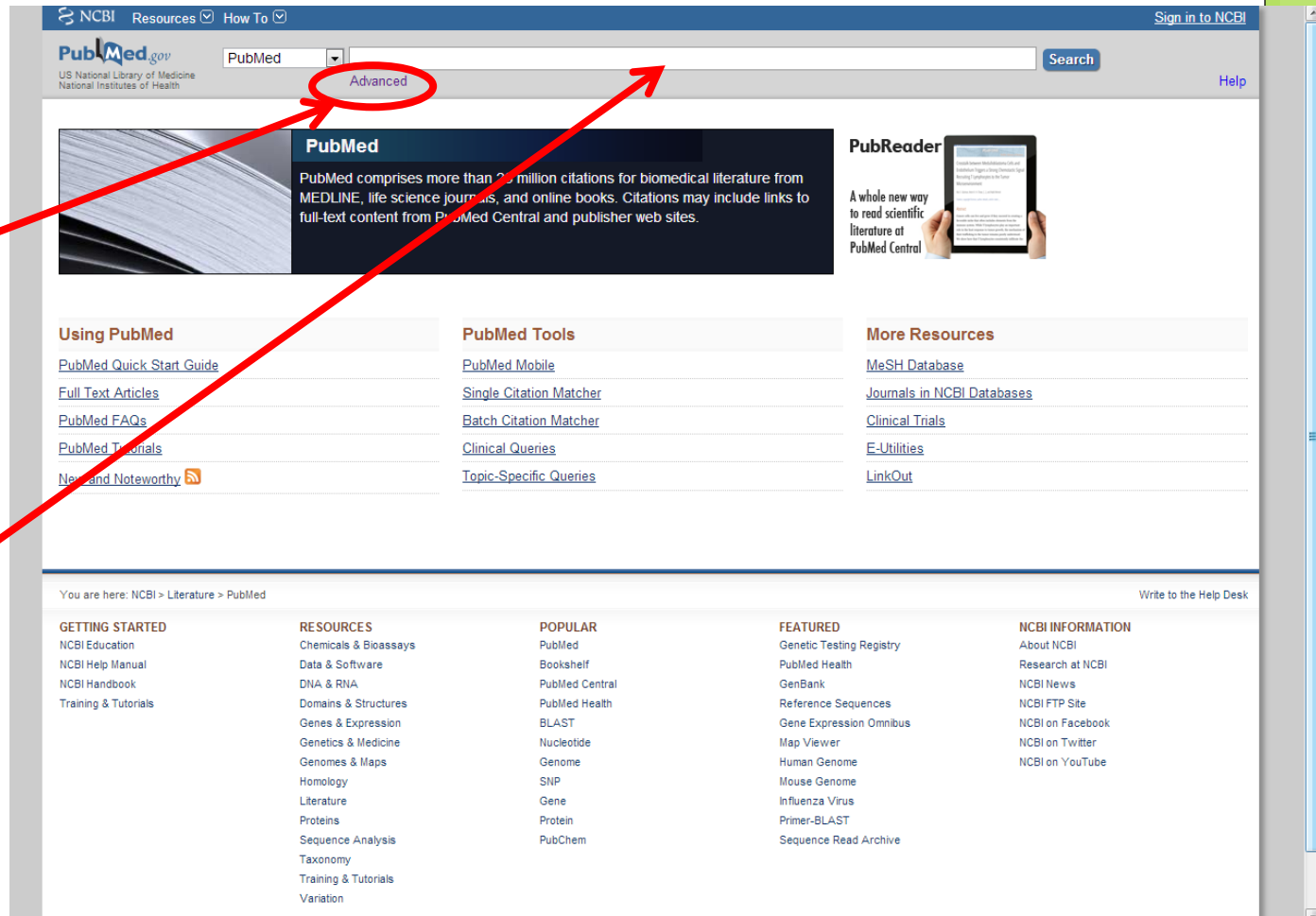
[Find items](#)

Using Pubmed

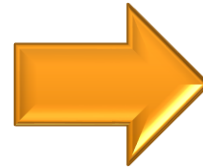
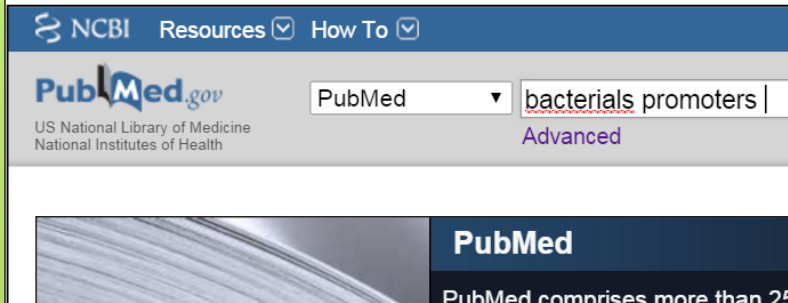
Go to <http://www.ncbi.nlm.nih.gov/pubmed>

Click here
for more
options

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



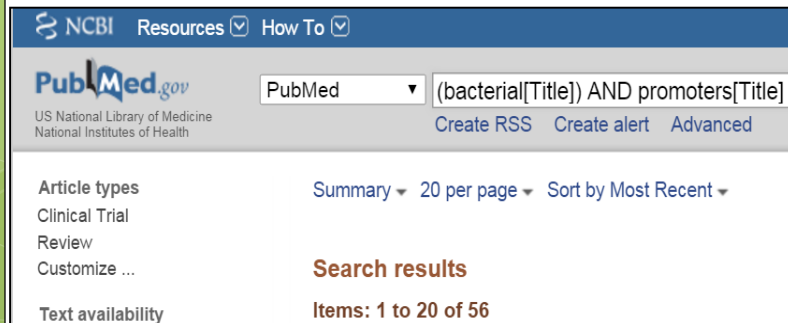
Using pubmed



Text MUST be correct



8108 articles are there



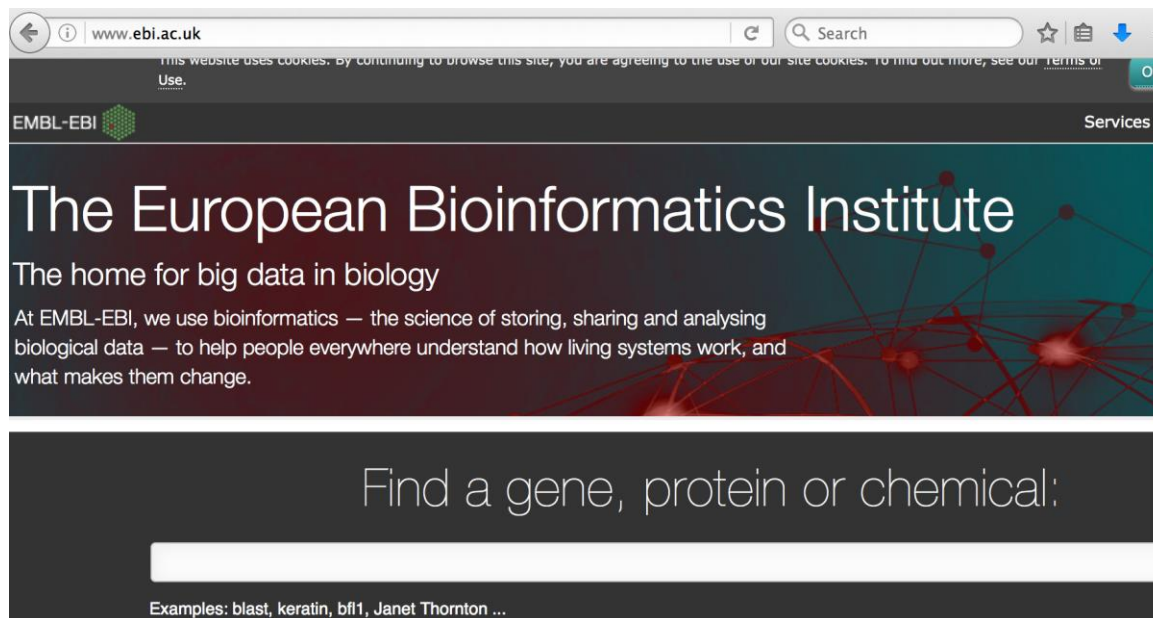
56 articles after using advanced option

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

https://www.ncbi.nlm.nih.gov/gene/?term=beta+globin

NCBI Resources How To Sign in to NCBI

Gene beta globin Search

Create RSS Create alert Advanced

NCBI is currently redirecting web traffic to HTTPS. [Read more](#) about our https testing.

Gene sources Genomic Categories Alternatively spliced Unannotated genes Non-coding Protein-coding Pseudogene Sequence Content CCDS Ensembl RefSeq RefSeqGene Status Current

Tabular 20 per page Sort by Relevance Send to: Hide sidebar

Search results

Items: 1 to 20 of 130

See also 8 discontinued or replaced items.

Name/Gene ID	Description	Location	Aliases	MIM
☐ HBB ID: 3043	hemoglobin subunit beta [<i>Homo sapiens</i> (human)]	Chromosome 11, NC_000011.10 (5225466..5227071, complement)	CD113t-C, beta-globin	141900
☐ hbg1 ID: 394453	hemoglobin subunit gamma 1 [<i>Xenopus tropicalis</i>]	Chromosome 9, NC_030685.1 (36978881..36980393, complement)	beta-globin, hbb1, hbga, hbgr, hsggl1	

clear

Filters: [Manage Filters](#)

Results by taxon

Top Organisms [\[Tree\]](#)

- Homo sapiens* (40)
- Mus musculus* (28)
- Danio rerio* (6)
- Salmo salar* (6)
- Rattus norvegicus* (5)
- All other taxa (45)

[More...](#)

Find related data

Database: [Select](#)

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 HGNC
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

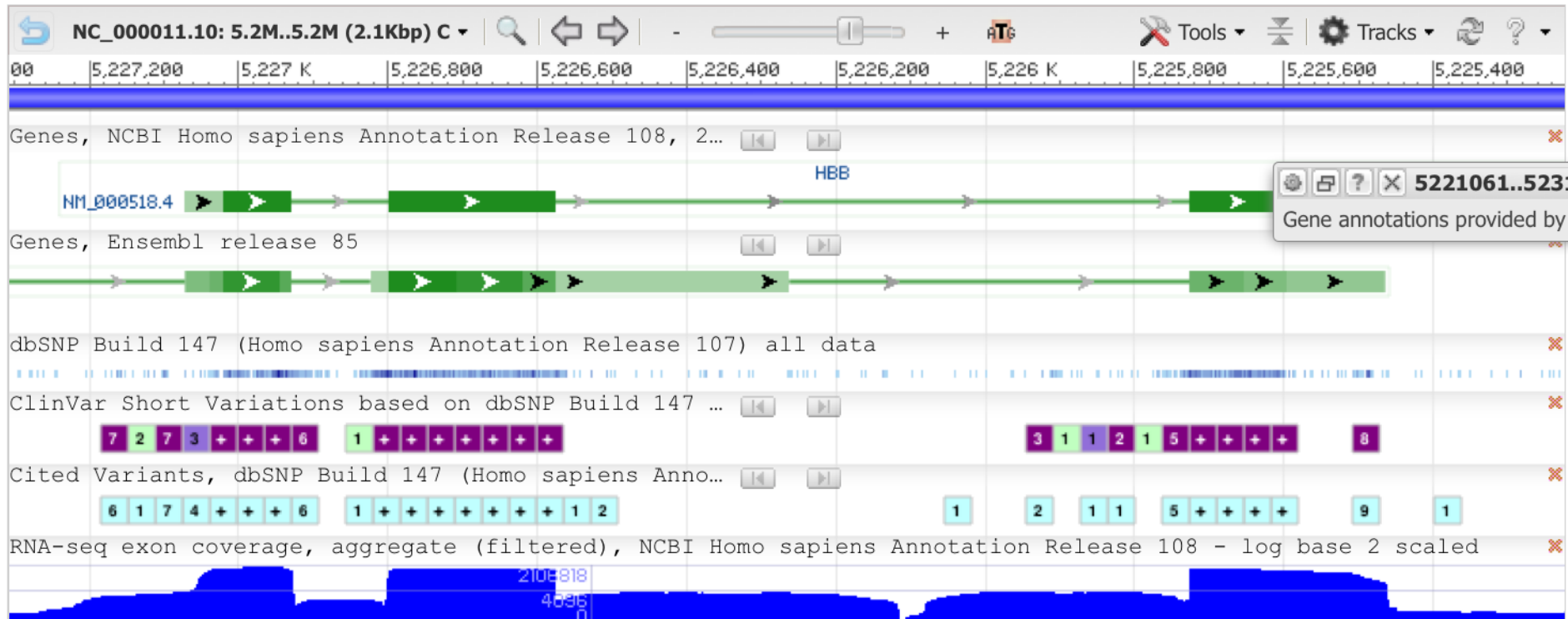
Access to Information via Gene Resource at NCBI

Genomic regions, transcripts, and products

Go to [reference sequence details](#)

Genomic Sequence: NC_000011.10 Chromosome 11 Reference GRCh38.p7 Primary Assembly

Go to nucleotide: [Graphics](#) [FASTA](#) [GenBank](#)



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](#)

>gi|568815587:c5227071-5225466 Homo sapiens chromosome 11, GRCh38.p7 Primary Assembly

ACATTTGCTTCTGACACAACCTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGA
GGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGC
AGGTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAACTGGGCATGTGGAGACAGAGAAG
ACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGG
TGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGG
CAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGAC
AACCTCAAGGGCACCTTTGCCCACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACT
TCAGGGTGAGTCTATGGGACGCTTGATGTTTTCTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAG
GAAGGGGATAAGTAACAGGGTACAGTTTAGAATGGGAAACAGACGAATGATTGCATCAGTGTGGAAGTCT
CAGGATCGTTTTAGTTTCTTTTATTTGCTGTTTCATAACAATTGTTTTCTTTTGTTTAATTCTTGCTTTCT
TTTTTTTTCTTCTCCGAATTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAAAGGAAATA

description

sequence

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

FASTA files may contain one or more sequences

→ >crab_anap1 ALPHA CRYSTALLIN B CHAIN (ALPHA(B)-CRYSTALLIN).
MDITIHNP LIRRP LFSWL APSRIFDQIFGEHLQESELLPASPSLSPFLMR
SPIFRMP SWLETGLSEMRLEKDKFSVNLDVKHFSPEELKVKVLGDMVEIH
GKHEERQDEHGFIAREFN RKYRIPADVDPLTITSSLSLDGVLTVSAPRKQ
SDVPERSIPITREEKPAIAGAQRK

→ >crab_bovin ALPHA CRYSTALLIN B CHAIN (ALPHA(B)-CRYSTALLIN).
MDIAIHHPWIRRPFFPFHSPSRLFDQFFGEHLLESDFPASTSLSPFYLR
PPSFLRAPSWIDTGLSEMRLEKDRFSVNLDVKHFSPEELKVKVLGDVIEV
HGKHEERQDEHGFI SREFHRKYRIPADVDPLAITSSLSSDGVLTVNGPRK
QASGPERTIPITREEKPAVTAAPKK

→ >crab_chick ALPHA CRYSTALLIN B CHAIN (ALPHA(B)-CRYSTALLIN).
MDITIHNP LVRRLFSWLTPSRIFDQIFGEHLQESELLPTSPSLSPFLMR
SPFFRMP SWLETGLSEMRLEKDKFSVNLDVKHFSPEELKVKVLGDMIEIH
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/usermanual.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

YOUR MITOCHONDRIAL HVR I SEQUENCE

16126C, 16147T, 16183C, 16189C, 16294T, 16296T, 16297C,
16304C, 16519C

```
ATTCTAATTTAAACTATTCTCTGTTCTTTCATGGGGAAGCAGATTTGGGTACCA
CCCAAGTATTGACTCACCCATCAACAACCGCTATGTATTTTCGTACATTACTGCC
AGCCACCATGAATATTGCACGGTACCATAAATACTTGATTCACCTGTAGTACATAA
AAACCCAATCCACATCAAACCCCCCCCCATGCTTACAAGCAAGTACAGCAAT
CAACCCTCAACTATCACACATCAACTGCAACTCCAAAGCCACCCCTCACCCAC
TAGGATACCAACAAACCTACCCATTCTCTAAACAGCACATAGTACATAAAGCCATTT
ACCGTACATAGCACATTACAGTCAAATCCCTTCTCGTCCCCATGGATGACCCC
CCTCAGATAGGGGTCCCTTGACCACCATCCTCCGTGAAATCAATATCCCGCAC
AAGAGTGCTACTCTCCTCGCTCCGGGGCCATAACACTTGGGGGTAGCTAAAGT
GAACTGTATCCGACATCTGGTTCCTACTTCAGGGCCATAAAGCCTAAATAGCCC
ACACGTTCCCCTTAATAAGACATCACGATG
```

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

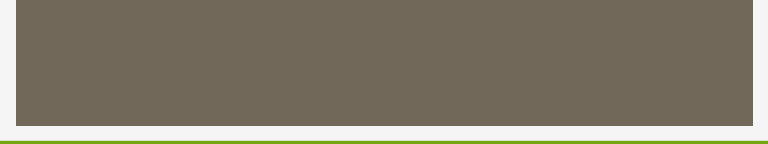
GENOME BROWSERS

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.

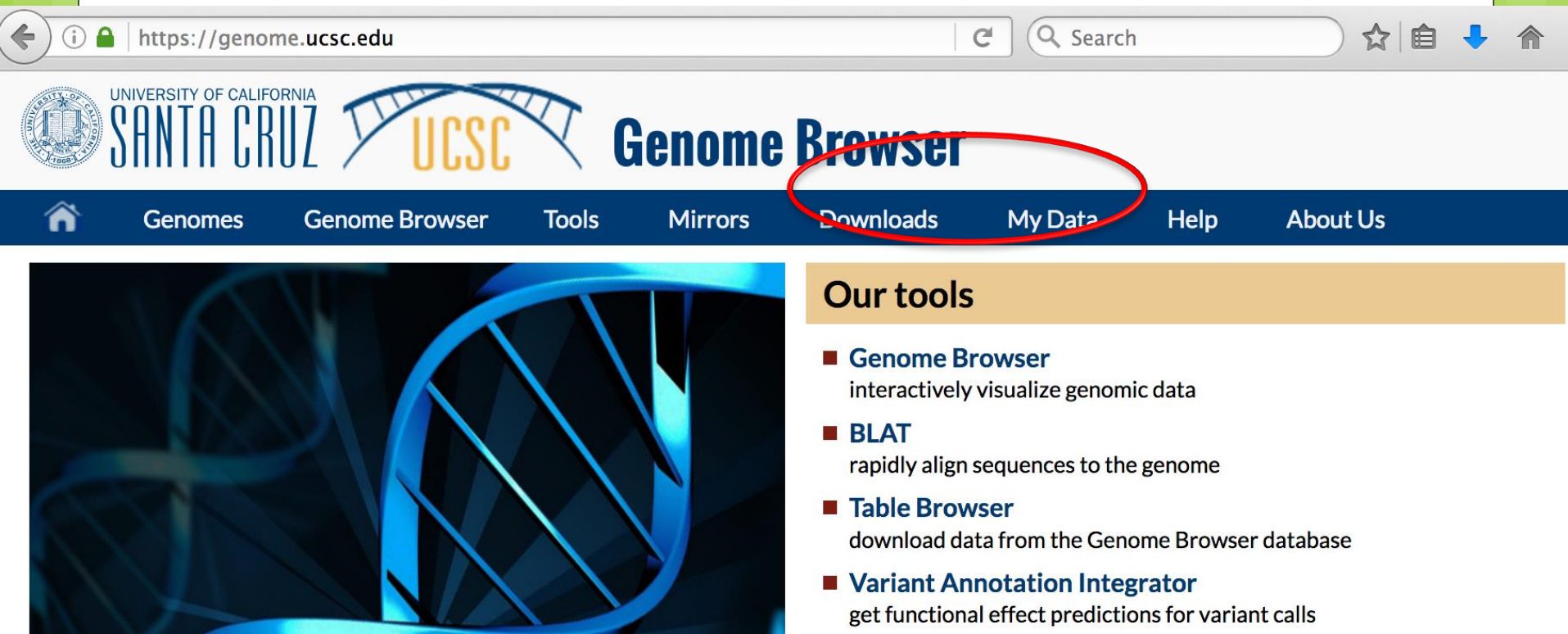
The University of California, Santa Cruz (UCSC) Genome Browser

- 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



The screenshot shows the UCSC Genome Browser website. The browser's address bar displays <https://genome.ucsc.edu>. The website header includes the University of California Santa Cruz logo, the UCSC logo, and the text "Genome Browser". A navigation bar below the header contains links: Home, Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Help, and About Us. The "Downloads" and "My Data" links are circled in red. On the left side of the page is a large image of a DNA double helix. On the right side, under the heading "Our tools", there is a list of 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



UNIVERSITY OF CALIFORNIA
SANTA CRUZ



Genome Browser Gateway

[Genomes](#)[Genome Browser](#)[Tools](#)[Mirrors](#)[Downloads](#)[My Data](#)[Help](#)[About Us](#)

Browse/Select Species

POPULAR SPECIES



Human



Mouse



Rat



Fruitfly



Worm



Yeast

REPRESENTED SPECIES

Find Position

Human Assembly

Position/Search Term

Current position: chr11:5,225,464-5,227,071

GO

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

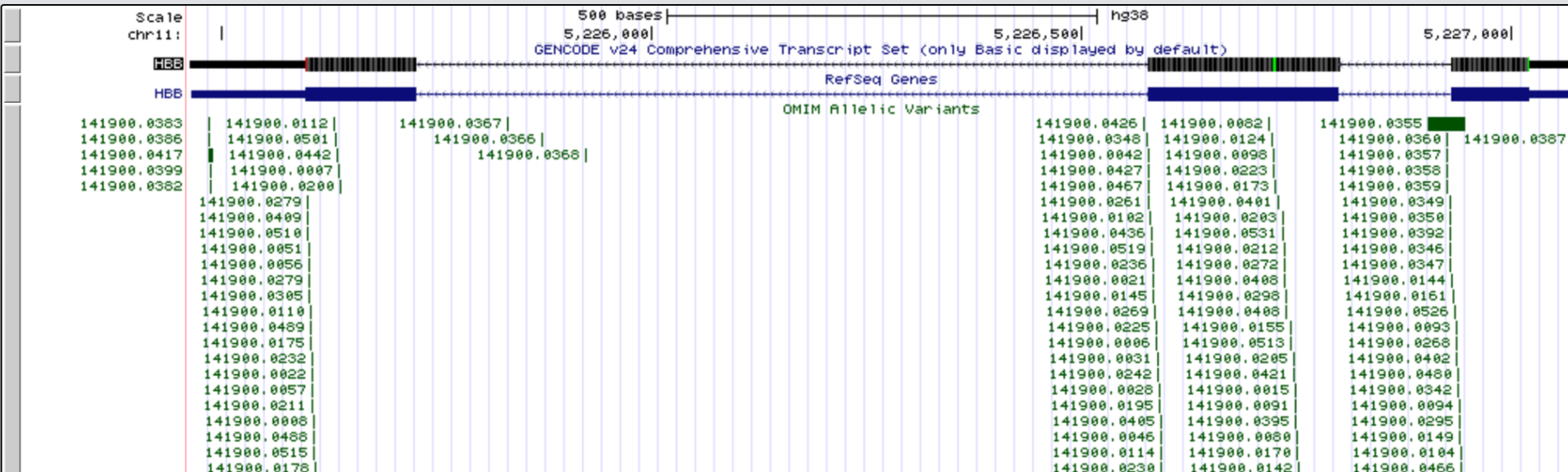
[Genomes](#)[Genome Browser](#)[Tools](#)[Mirrors](#)[Downloads](#)[My Data](#)[View](#)[Help](#)[About Us](#)

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr11:5,225,464-5,227,071 1,608 bp.

go



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.



You have been redirected to your nearest mirror. [Click here to go back to www.ensembl.org](#)

[BLAST/BLAT](#)[BioMart](#)[Tools](#)[Downloads](#)[Help & Documentation](#)[More ▾](#)

Search: for

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](#)

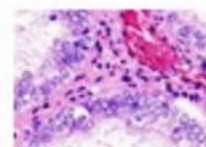
Still using Human GRCh37?



Variant Effect Predictor



Gene expression in different tissues



Find SNPs and other variants for my gene

```
GTRTATACATT  
CCTRAAAGTCT  
CTTCTAAATT  
GRAACATTTTC
```

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

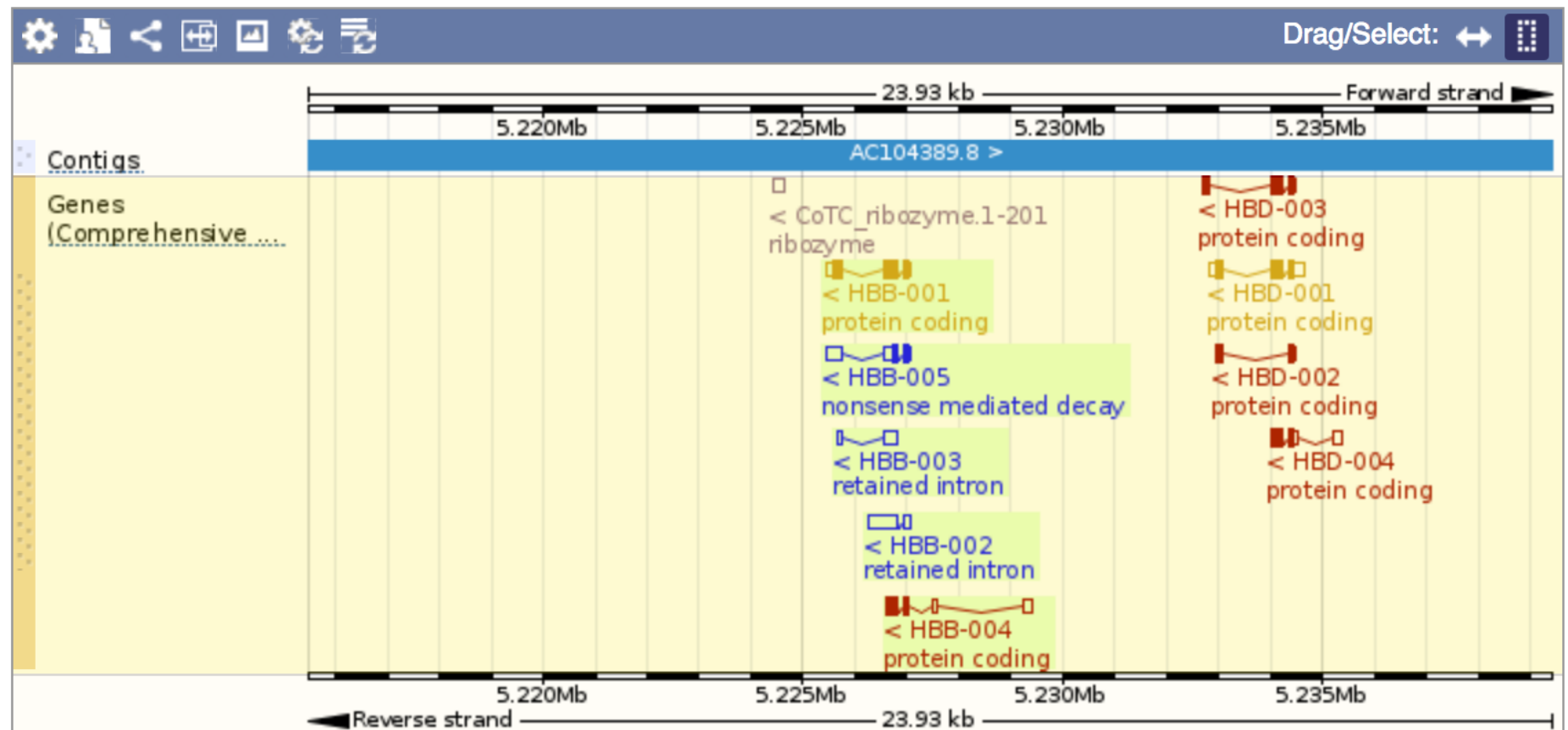
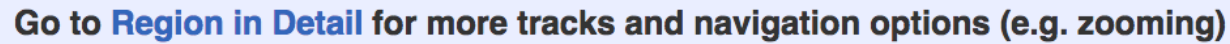
[HBB](#) (HGNC Symbol)

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](#)



The Map Viewer provides a wide variety of genome mapping and seq

Search

Search:

for:

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
<i>Callithrix jacchus</i>	white-tufted-ear marmoset
<i>Chlorocebus sabaeus</i>	green monkey
<i>Gorilla gorilla</i>	western gorilla
<i>Homo sapiens</i>	human
<i>Macaca fascicularis</i>	crab-eating macaque
<i>Macaca mulatta</i>	rhesus macaque
<i>Nomascus leucogenys</i>	northern white-cheeked gibbon
<i>Otolemur garnettii</i>	small-eared galago
<i>Pan paniscus</i>	pygmy chimpanzee
<i>Pan troglodytes</i>	chimpanzee

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

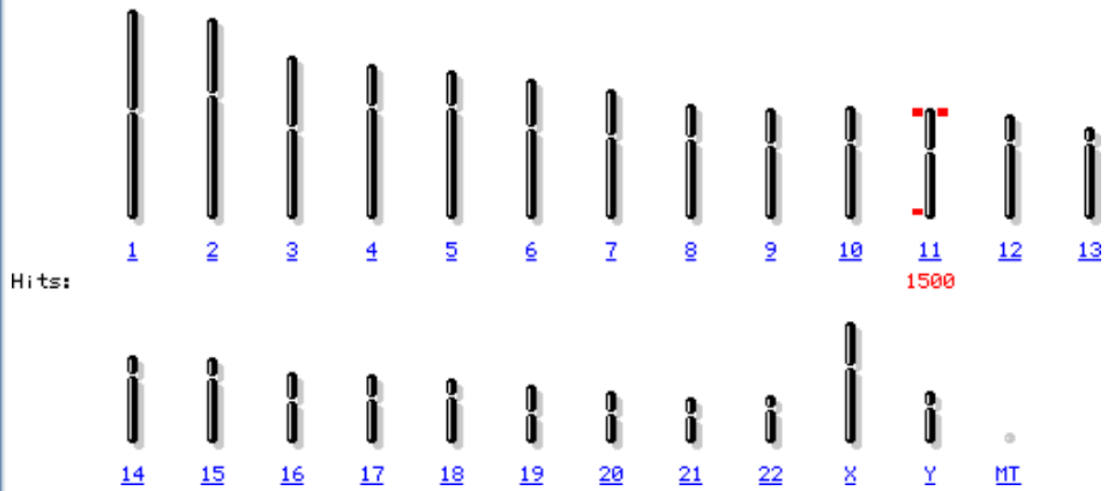
NCBI Resources

- Assembly
- CCDS
- Gene
- Genome
- RefSeq

Organism Data in GenBank

EST

Homo sapiens (human) genome view
[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	Browser extensible data	Any size; often millions of rows
BED		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	Wiggle	Any size
PSL		Any size
WIG		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



UNIVERSITY OF CALIFORNIA
SANTA CRUZ



Genome Browser



Genomes

Genome Browser

Tools

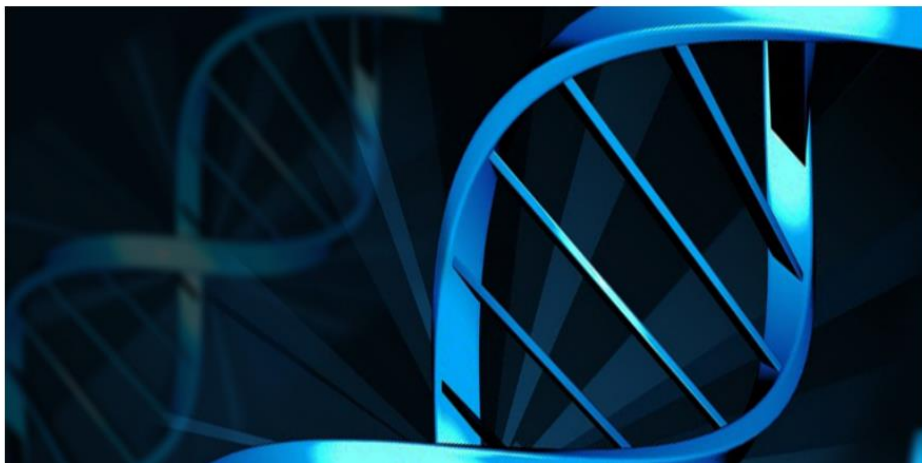
Mirrors

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

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

All Databases ▾

- Conserved Domains
- dbGaP
- dbVar
- Epigenomics
- EST
- Gene
- Genome
- GEO DataSets
- GEO Profiles
- GSS
- GTR
- HomoloGene
- MedGen
- MeSH
- NCBI Web Site
- NLM Catalog
- Nucleotide**
- OMIM
- PMC
- PopSet
- Proteins

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

NCBI

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

Home | Mission | Organization | Research | NCBI News

Learn how to accomplish specific tasks at NCBI

Submit data to GenBank or other NCBI databases

Popular Resources

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

NCBI Twitter feed

Keep up-to-date on data updates, resource announcements, and other information about what is going on at the NCBI.

GO

1 2 3 4 5 6 7 8

NCBI Announcements

NCBI's next webinar is The Statistics of Local Pairwise Sequence Alignment, Parts 1 and 2

Jan 13, 2011

On Thursday, January 22nd, Stephen

Database

Query

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide mRNA Search

Save search Advanced

Send to file

Show additional filters

Species
Animals
Plants
Fungi
Protists
Bacteria
Archaea
Viruses
More ...

Molecule types
genomic DNA/RNA
mRNA
rRNA
More ...

Source databases
GenBank
RefSeq
More ...

Genetic compartments
Chloroplast
Mitochondrion
Plasmid
Plastid

Sequence length
Custom range...

Release date
Custom range...

Display Settings: Summary, 20 per page, Default order

Send to: Filters: Manage Filters

Found 117333794 nucleotide sequences. Nucleotide (41519086) EST (75808994) GSS (5714)

Results: 1 to 20 of 41519086

<< First < Prev Page 1 of 2075955 Next > Last >>

1. [Arabidopsis thaliana chromosome 4 sequence](#)
18,585,056 bp linear DNA
Accession: CP002687.1 GI: 332656411
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

2. [Arabidopsis thaliana chromosome 1 sequence](#)
30,427,000 bp linear DNA
Accession: CP002684.1 GI: 332648804
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

3. [Arabidopsis thaliana chromosome 5 sequence](#)
26,975,502 bp linear DNA
Accession: CP002688.1 GI: 332002898
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

4. [Arabidopsis thaliana chromosome 2 sequence](#)
23,459,830 bp linear DNA
Accession: CP002686.1 GI: 332640072
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

5. [Arabidopsis thaliana chromosome 3 sequence](#)
19,698,289 bp linear DNA
Accession: CP002685.1 GI: 330250293
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

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

Entry

Display GenBank or Fasta

Change the number per page & others

Filters to minimizes number of results

NCBI abbreviations for nucleotide

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

NCBI abbreviations for protein

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

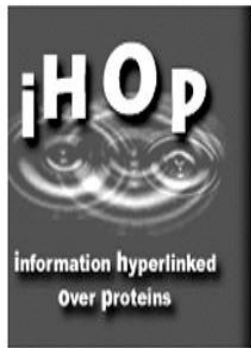
Predicted & Known Proteins

- **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
 Life cycles of successful genes <i>Trends in Genetics</i>				↓
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 !\[\]\(42ac8b6ce98c050538664bdd9f7933cf_img.jpg\) ...](#)

[Most recent information for WWC1 !\[\]\(62302595baed42c1fa7f7c9fb1f95ae9_img.jpg\) ...](#)

[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**

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]  

Conclusion

- Many databases and resources are available, some as websites and some (such as R packages or NCBI E-Utilities) via programming 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.