CMPT 830 - Bioinformatics and Computational Biology

Chapter 4: Sequence Databases

Ian McQuillan and Tony Kusalik
kusalik@cs.usask.ca

University of Saskatchewan

October 9, 2019

Sequence Databases

Three major publicly accessible sequence databases

1. **Genbank** at the National Center for Biotechnology Information (NCBI) at the National Institute of Health (NIH).
2. European Molecular Biology Laboratory **EMBL** Nucleotide Sequence Database at the European Bioinformatics Institute (EBI).
3. The DNA Database of Japan **DDBJ** at the national Institute of Genetics.

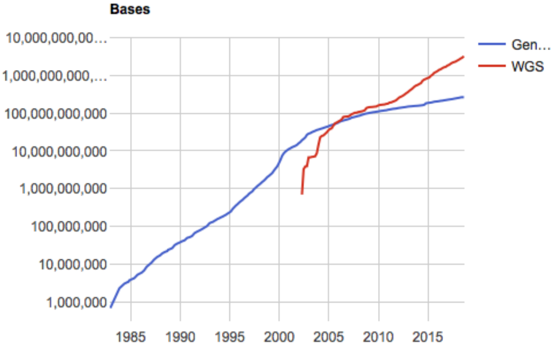
There are hundreds of other public sequence databases that are more specialized. Some specialize in a species, or a type of data.

Sequence Databases

- The three databases exchange sequences every day, so in terms of data, it does not matter which you use.
- However, each website has their own benefits, features and built-in programs.
- We'll mostly use NCBI.

GenBank

- GenBank currently holds over 260 billion nucleotides from 208 million sequences.
- Growth has been roughly exponential.



picture from http://www.ncbi.nlm.nih.gov/genbank/statistics

NCBI and Entrez

- NCBI: National Center for Biotechnology is located at `http://www.ncbi.nlm.nih.gov`
- The sequence records in GenBank are accessible by NCBI's retrieval system, **Entrez**.
- Entrez: `http://www.ncbi.nlm.nih.gov/gquery/`
- Entrez is a database retrieval system that covers 40 biological databases.
- NCBI integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and biomedical journal literature via PubMed.

Searching

- GenBank contains a number of distinct databases. For example, two are “Core Nucleotide” (the main collection) and dbEST (Expressed Sequence Tags).
- It is possible to search with Entrez from the main Entrez page, and it will tell you the number of ‘hits’ in the various databases at NCBI.
- Or, it is possible to go into an individual database and search just that database.

How to Search?

- It is possible to search Entrez with a key word or a phrase contained in the record (this allows to search for information that has been annotated on the sequence).
- If you know the sequence itself, use BLAST to find the record.
- BLAST results are NCBI records.
- Or, you can search Entrez with the ‘key’ associated with each record.

Accession Number

- Each NCBI database record has a unique identifier (the ‘key’) called an *accession number* (shared with DDBJ and EBI/EMBL).
- Records can be updated, giving a new *version number* by increasing a number after a decimal of the accession number.
- For example, `NM_000518` is an accession number, and `NM_000518.1` and `NM_000518.2` are version numbers.

Accession Number

- Search NCBI for `NM_000518` (in the nucleotide database).
- Search for version 3?
- Searching for accession (without version) will take you to newest version.
- But it's possible to find old versions.

Sequence Databases

Typically, sequence databases come in two flavours:

1. Primary Databases - experimental results, not the consensus of a population.

2. Secondary Databases - curated reviews, perhaps multiple sequences collapsed into fewer records.

Some Databases at NCBI

There are quite a number of different databases stored (separately) at NCBI. We won't use them all, but here are some:

- Nucleotide Database – includes all nucleotide sequences from GenBank, RefSeq (to be explained later) and others.

- Protein Database - includes protein information from a variety of databases including RefSeq, GenPept, Swiss-Prot, PIR and others (also to be explained later).

- Entrez Gene – a searchable database of genes, focusing on genomes that have been completely sequenced. Includes lots of information regarding entries such as nomenclature, location on the chromosome, gene products, phenotypes, and links to external databases. More on this one later.

More Databases at NCBI

Lastly:

- Trace Archive – older repository of raw sequencing data, mainly derived from Sanger sequencing.

- SRA – The Sequence Read Archive stores raw sequence data from "next-generation" sequencing technologies. It is shared between NCBI, EBI and DDBJ.

These have a **LOT** of data.

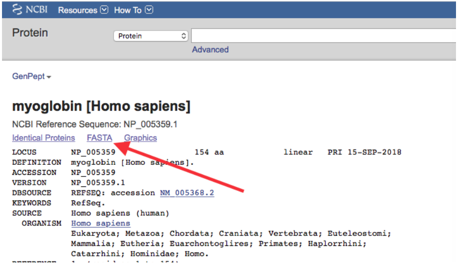
Viewing Records

When viewing records, it is possible to switch between:

- GenBank/GenPept flat file format,
- FASTA format.

When a record is clicked, the flat file is automatically loaded. It is default.

Switch to FASTA



myoglobin [Homo sapiens]

NCBI Reference Sequence: NP_005359.1

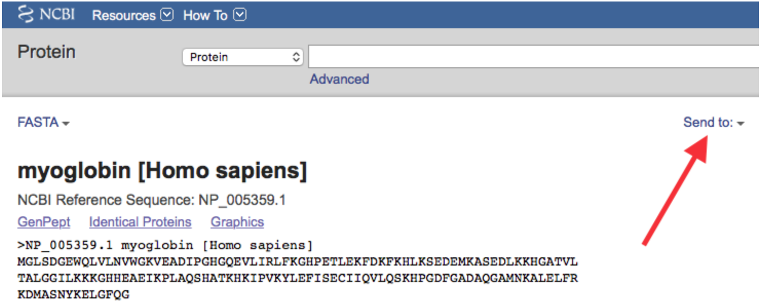
```
LOCUS       NP_005359                154 aa            linear   PRI 15-SEP-2018
DEFINITION  myoglobin [Homo sapiens].
ACCESSION   NP_005359
VERSION     NP_005359.1
DBSOURCE    REFSEQ: accession NM_005368.2
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
```

- Start by searching the Protein database for accession number `NP_005359`.
- This is a display of the flat file.
- From the flat file, it is possible to switch to FASTA view.

FASTA file format

- FASTA is a file format used to store sequence information.
- The first line is the *definition line*. It starts with '$>$'.
- Typically, after the '$>$', we have some sort of identifer, including, if in GenBank, the accession number. Often, we also get a description in this line.

FASTA example



myoglobin [Homo sapiens]

NCBI Reference Sequence: NP_005359.1

GenPept Identical Proteins Graphics

```
>NP_005359.1 myoglobin [Homo sapiens]
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVL
TALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFR
KDMASNYKELGFQG
```

- This is what displays after clicking FASTA.
- It is possible to copy it to the clipboard (if you want to paste somewhere else).
- Or, by clicking ‘send to’, there is an option to save it as a file on your computer.

FASTA file format

- All other lines contain sequence, usually 60 characters per line (except perhaps the last line which might have fewer characters).

- More information on FASTA file format here: `http://www.ncbi.nlm.nih.gov/BLAST/fasta.shtml`

- Why is this format useful?

FASTA file format

- A lot of times we want to use the **sequence** and do something with it.
- Programs that do these tasks need to scan the **sequence**, and they almost all take FASTA files as input.
- Or you can use them with bioinformatics apps running on your computer.
- It can be viewed on any computer using a plain text editor.

Flat Files

- Each sequence database (NCBI/EMBL/DDBJ) stores files in their own internal format.

- Each also has its own “flat file” (i.e. plain text) format. (although it is possible to convert from one to another)

- We will very briefly go over the GenBank flat file format.

Flat Files

GenBank flat files consist of three main parts:

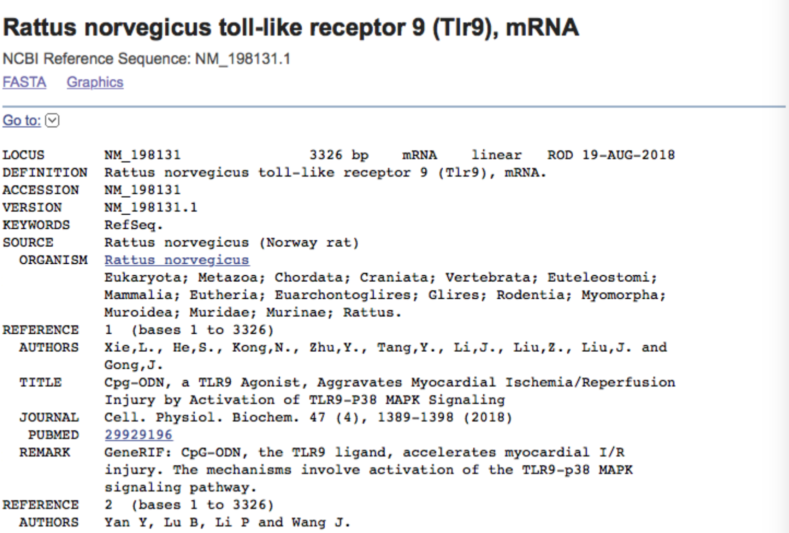
1. the header (information applying to the record),
2. the features (annotations on the record),
3. the nucleotide sequence.

The last line ends with //.

A sample flat file and information on what each entry means is at `http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html`.

GenBank Header

Here is a header.

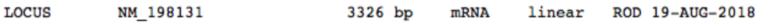


Rattus norvegicus toll-like receptor 9 (Tlr9), mRNA

NCBI Reference Sequence: NM_198131.1

```
LOCUS       NM_198131               3326 bp    mRNA    linear   ROD 19-AUG-2018
DEFINITION  Rattus norvegicus toll-like receptor 9 (Tlr9), mRNA.
ACCESSION   NM_198131
VERSION     NM_198131.1
KEYWORDS    RefSeq.
SOURCE      Rattus norvegicus (Norway rat)
  ORGANISM  Rattus norvegicus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha;
            Muroidea; Muridae; Murinae; Rattus.
REFERENCE   1  (bases 1 to 3326)
  AUTHORS   Xie,L., He,S., Kong,N., Zhu,Y., Tang,Y., Li,J., Liu,Z., Liu,J. and
            Gong,J.
  TITLE     Cpg-ODN, a TLR9 Agonist, Aggravates Myocardial Ischemia/Reperfusion
            Injury by Activation of TLR9-P38 MAPK Signaling
  JOURNAL   Cell. Physiol. Biochem. 47 (4), 1389-1398 (2018)
   PUBMED   29929196
  REMARK    GeneRIF: CpG-ODN, the TLR9 ligand, accelerates myocardial I/R
            injury. The mechanisms involve activation of the TLR9-p38 MAPK
            signaling pathway.
REFERENCE   2  (bases 1 to 3326)
  AUTHORS   Yan Y, Lu B, Li P and Wang J.
```

GenBank Flat File Header



```
LOCUS       NM_198131                 3326 bp    mRNA    linear   ROD 19-AUG-2018
```

- The first line is the ‘LOCUS’ line.
- The first entry (in this case, `NM_198131`) is the accession number.
- Next, we have the length of the sequence.
- Then, the molecule type.
- Then, there is a “division code” (some examples: PRI for primates, ROD for rodents, MAM for other mammals, PLN for plants, BCT for bacterial sequences).
- Then, the latest release date of the record.

Header

Also, in the header

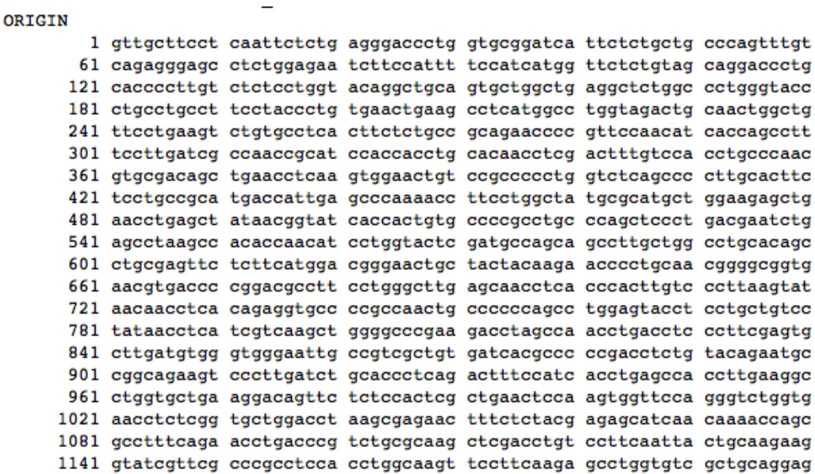
- the organism, with taxonomy,
- literature references associated with that record,
- comments on the function.

GenBank flat file

```
     CDS             97..3195
```

- The feature table, is the direct representation of the biological information in the record.
- One of the most important features is the coding sequence (CDS above).
- The coding sequence represents the parts of the gene that codes for the corresponding protein.
- It'll have the exons in it, but not the introns.
- The numbers are positions of the sequence with dots representing ranges, and the commas representing the non-coding parts.

Flat Files



```
ORIGIN
        1 gttgcttcct caattctctg agggaccctg gtgcggatca ttctctgctg cccagtttgt
       61 cagagggagc ctctggagaa tcttccattt tccatcatgg ttctctgtag caggaccctg
      121 caccccttgt ctctcctggt acaggctgca gtgctggctg aggctctggc cctgggtacc
      181 ctgcctgcct tcctaccctg tgaactgaag cctcatggcc tggtagactg caactggctg
      241 ttcctgaagt ctgtgcctca cttctctgcc gcagaacccc gttccaacat caccagcctt
      301 tccttgatcg ccaaccgcat ccaccacctg cacaacctcg actttgtcca cctgcccaac
      361 gtgcgacagc tgaacctcaa gtggaactgt ccgccccctg gtctcagccc cttgcacttc
      421 tcctgccgca tgaccattga gcccaaaacc ttcctggcta tgcgcatgct ggaagagctg
      481 aacctgagct ataacggtat caccactgtg ccccgcctgc ccagctccct gacgaatctg
      541 agcctaagcc acaccaacat cctggtactc gatgccagca gccttgctgg cctgcacagc
      601 ctgcgagttc tcttcatgga cgggaactgc tactacaaga accccctgcaa cggggcggtg
      661 aacgtgaccc cggacgcctt cctgggcttg agcaacctca cccacttgtc ccttaagtat
      721 aacaacctca cagaggtgcc ccgccaactg ccccccagcc tggagtacct cctgctgtcc
      781 tataacctca tcgtcaagct ggggcccgaa gacctagcca acctgacctc ccttcgagtg
      841 cttgatgtgg gtgggaattg ccgtcgctgt gatcacgccc ccgacctctg tacagaatgc
      901 cggcagaagt cccttgatct gcaccctcag actttccatc acctgagcca ccttgaaggc
      961 ctggtgctga aggacagttc tctccactcg ctgaactcca agtggttcca gggtctggtg
     1021 aacctctcgg tgctggacct aagcgagaac tttctctacg agagcatcaa caaaaccagc
     1081 gcctttcaga acctgacccg tctgcgcaag ctcgacctgt ccttcaatta ctgcaagaag
     1141 gtatcgttcg cccgcctcca cctggcaagt tccttcaaga gcctggtgtc gctgcaggag
```

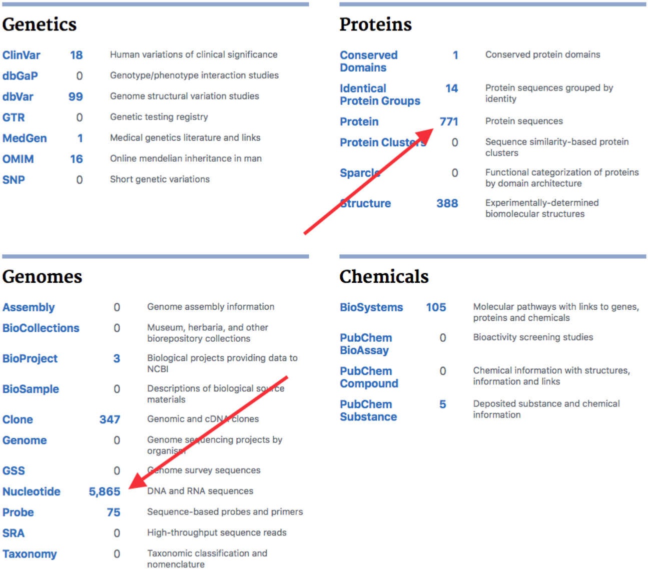
- Then at the bottom (after the header and features), is the sequence itself.

Data Redundancy in GenBank

- Many sequences in GenBank (especially in the Nucleotide and Protein database) (and DDBJ/EMBL) are represented many times. It is **extremely redundant**.

- There are also partial sequences, incomplete records and erroneous records.

Redundancy



Genetics

ClinVar	18	Human variations of clinical significance
dbGaP	0	Genotype/phenotype interaction studies
dbVar	99	Genome structural variation studies
GTR	0	Genetic testing registry
MedGen	1	Medical genetics literature and links
OMIM	16	Online mendelian inheritance in man
SNP	0	Short genetic variations

Proteins

Conserved Domains	1	Conserved protein domains
Identical Protein Groups	14	Protein sequences grouped by identity
Protein	771	Protein sequences
Protein Clusters	0	Sequence similarity-based protein clusters
Sparcle	0	Functional categorization of proteins by domain architecture
Structure	388	Experimentally-determined biomolecular structures

Genomes

Assembly	0	Genome assembly information
BioCollections	0	Museum, herbaria, and other biorepository collections
BioProject	3	Biological projects providing data to NCBI
BioSample	0	Descriptions of biological source materials
Clone	347	Genomic and cDNA clones
Genome	0	Genome sequencing projects by organism
GSS	0	Genome survey sequences
Nucleotide	5,865	DNA and RNA sequences
Probe	75	Sequence-based probes and primers
SRA	0	High-throughput sequence reads
Taxonomy	0	Taxonomic classification and nomenclature

Chemicals

BioSystems	105	Molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	0	Bioactivity screening studies
PubChem Compound	0	Chemical information with structures, information and links
PubChem Substance	5	Deposited substance and chemical information

Approaches to Redundancy

- There are various approaches to reducing redundancy.

- Curate a database so that each gene or sequence in each organism only gets one entry.

- Use computer algorithms to “collapse” records if the sequences are exactly the same, or are “similar”.

RefSeq

- NCBI developed the *RefSeq* collection which is a curated secondary database.
- http://www.ncbi.nlm.nih.gov/refseq/
- The entries are curated by the staff at NCBI.
- The goal is to provide the best representative sequence for each normal transcript and protein.

RefSeq

- Instead of potentially having 100 different accession numbers for a gene in an organism, in RefSeq, the goal is to have 1 accession number.

- (There can be more than 1 accession number even in RefSeq if the same gene occurs at multiple loci of the genome for example.)

- This is often what one actually wants to have.

- Thus, it has nonredundant sequences, including DNA, transcripts and protein products for **select** organisms.

- This also includes all of the human genome assemblies.

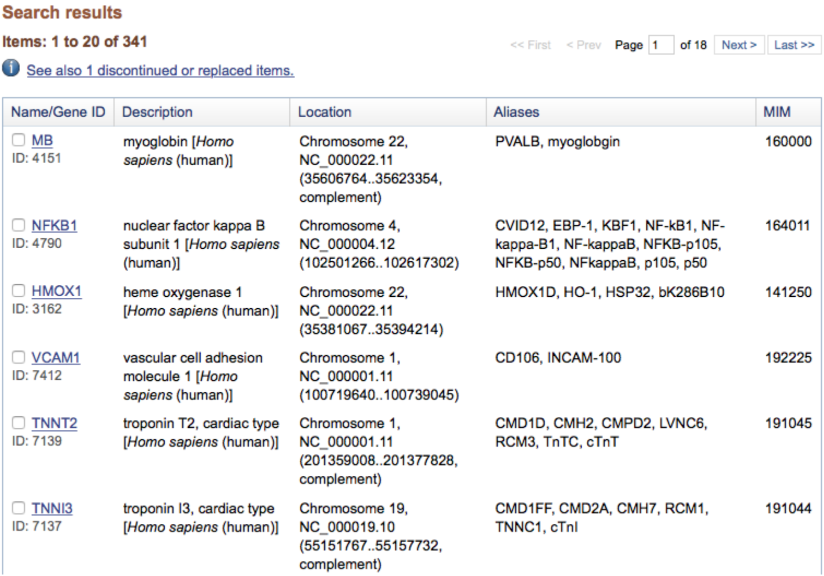
RefSeq

- It is possible to search the Nucleotide or Protein databases, and then restrict to RefSeq afterwards.
- Search the Nucleotide database for human `TLR9`, then restrict to RefSeq.
- RefSeq is a great resource.

Entrez Gene

- *Entrez Gene* is often a good starting point for finding information.
- Can be found at `http://www.ncbi.nlm.nih.gov/gene`.
- It is a curated databased containing information regarding various genetic loci.
- From there, information can be found on nomenclature, phenotypes, and map locations.

Entrez Gene



Search results

Items: 1 to 20 of 341

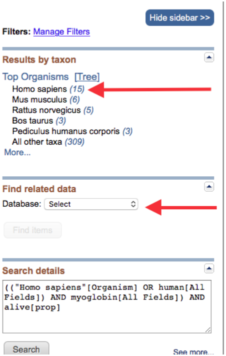
See also 1 discontinued or replaced items.

Name/Gene ID	Description	Location	Aliases	MIM
□ MB ID: 4151	myoglobin [*Homo sapiens* (human)]	Chromosome 22, NC_000022.11 (35606764..35623354, complement)	PVALB, myoglobgin	160000
□ NFKB1 ID: 4790	nuclear factor kappa B subunit 1 [*Homo sapiens* (human)]	Chromosome 4, NC_000004.12 (102501266..102617302)	CVID12, EBP-1, KBF1, NF-kB1, NF-kappa-B1, NF-kappaB, NFKB-p105, NFKB-p50, NFkappaB, p105, p50	164011
□ HMOX1 ID: 3162	heme oxygenase 1 [*Homo sapiens* (human)]	Chromosome 22, NC_000022.11 (35381067..35394214)	HMOX1D, HO-1, HSP32, bK286B10	141250
□ VCAM1 ID: 7412	vascular cell adhesion molecule 1 [*Homo sapiens* (human)]	Chromosome 1, NC_000001.11 (100719640..100739045)	CD106, INCAM-100	192225
□ TNNT2 ID: 7139	troponin T2, cardiac type [*Homo sapiens* (human)]	Chromosome 1, NC_000001.11 (201359008..201377828, complement)	CMD1D, CMH2, CMPD2, LVNC6, RCM3, TnTC, cTnT	191045
□ TNNI3 ID: 7137	troponin I3, cardiac type [*Homo sapiens* (human)]	Chromosome 19, NC_000019.10 (55151767..55157732, complement)	CMD1FF, CMD2A, CMH7, RCM1, TNNC1, cTnI	191044

Entrez Gene

- It's possible to restrict to specific organisms (top right).

- It's possible to find information on myoglobin in other databases.



Filters: Manage Filters

Results by taxon

Top Organisms [Tree]

Homo sapiens (15)
Mus musculus (6)
Rattus norvegicus (5)
Bos taurus (3)
Pediculus humanus corporis (3)
All other taxa (309)
More...

Find related data

Database: Select

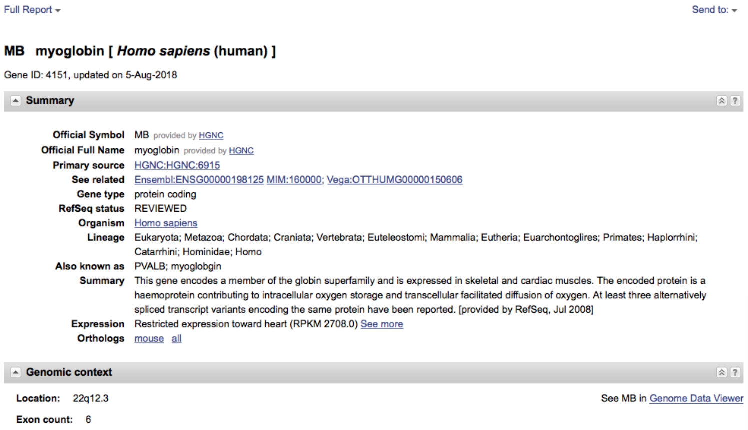
Search details

```
(("Homo sapiens"[Organism] OR human[All
Fields]) AND myoglobin[All Fields]) AND
alive[prop]
```

Entrez Gene

- If you follow the first hit, you get Entrez Gene entry for human myoglobin.
- The entry is here: `http://www.ncbi.nlm.nih.gov/gene/4151`

Entrez Gene for Human Myoglobin



MB myoglobin [*Homo sapiens* (human)]

Gene ID: 4151, updated on 5-Aug-2018

Summary

Official Symbol	MB provided by HGNC
Official Full Name	myoglobin provided by HGNC
Primary source	HGNC:HGNC:6915
See related	Ensembl:ENSG00000198125 MIM:160000; Vega:OTTHUMG00000150606
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	PVALB; myoglobgin
Summary	This gene encodes a member of the globin superfamily and is expressed in skeletal and cardiac muscles. The encoded protein is a haemoprotein contributing to intracellular oxygen storage and transcellular facilitated diffusion of oxygen. At least three alternatively spliced transcript variants encoding the same protein have been reported. [provided by RefSeq, Jul 2008]
Expression	Restricted expression toward heart (RPKM 2708.0) See more
Orthologs	mouse all

Genomic context

Location: 22q12.3

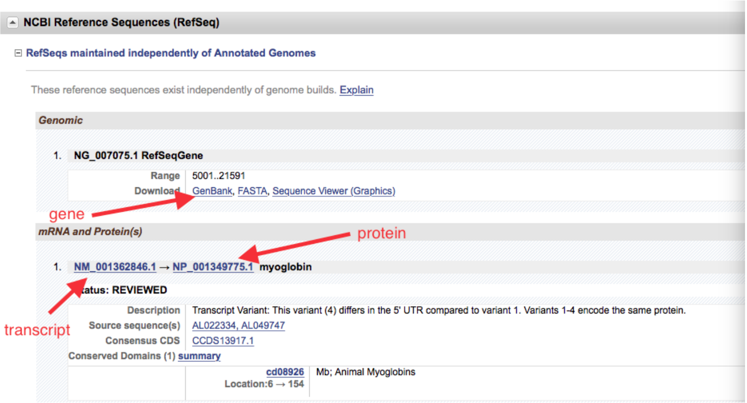
See MB in Genome Data Viewer

Exon count: 6

The summary section indicates the “Official Symbol” of that gene (MB), the gene type, chromosomal location, and a summary of its function.

Entrez Gene for Human Myoglobin

Scroll down, and see the RefSeq accession numbers for the DNA sequence encoding the full gene, the longest myoglobin transcript, and the protein entry.



NCBI Reference Sequences (RefSeq)

RefSeqs maintained independently of Annotated Genomes

These reference sequences exist independently of genome builds. Explain

Genomic

1. **NG_007075.1 RefSeqGene**

Range	5001..21591
Download	GenBank, FASTA, Sequence Viewer (Graphics)

gene

mRNA and Protein(s)

protein

1. NM_001362846.1 → NP_001349775.1 **myoglobin**

Status: REVIEWED

Description	Transcript Variant: This variant (4) differs in the 5' UTR compared to variant 1. Variants 1-4 encode the same protein.
Source sequence(s)	AL022334, AL049747
Consensus CDS	CCDS13917.1

transcript

Conserved Domains (1) summary

cd08926 Location:6 → 154	Mb; Animal Myoglobins

Downloading

- It is possible to download entire databases from NCBI.
- For example, one can download entire sequenced genomes, the complete proteome for an organism, or all of RefSeq.
- Info here: https://www.ncbi.nlm.nih.gov/guide/genomes-maps/.
- Download whole genomes here: ftp://ftp.ncbi.nih.gov/genomes/.

Expression

- The genome of every cell is (almost) the same.
- But, genes gets transcribed into RNA in differing quantities.
- For each gene, the amount of RNAs depends on the organism, the tissue type, condition, treatments, and temporal stages.

Expression

- It is common to assess differences in expression of each gene between different conditions or tissues.

- Diseased vs. non-diseased, treated vs. non-treated, tissue A vs. tissue B.

- RNA-seq is a technique to sequence RNAs in a biological sample (the sequence plus quantity).

- These sequence datasets are available on the *Gene Expression Omnibus* (GEO) `https://www.ncbi.nlm.nih.gov/geo/`.

Protein Databases

- Protein data searched by NCBI comes from several sources.
- NCBI contains translated coding regions from nucleotide records in GenBank.
- NCBI also searches protein sequences from some external databases, such as the Protein Information Resource (PIR), Swiss-Prot, TrEMBL, Protein Research Foundation (PRF), the Protein Data Bank (PDB).

Protein Sequence Databases

- The first main protein database was Swiss-Prot.
- Swiss-Prot was supplemented with TrEMBL (translation of EMBL nucleotide sequences).
- If a coding sequence is not indicated on a nucleic acid record, it will not lead to the creation of a record in the protein databases.
- If a coding region in a DNA sequence database contains incorrect information, this will be passed on to other nucleotide and protein records.

Protein Sequence Databases

Two databases, Swiss-Prot and TrEMBL, combined together under one umbrella, UniProt.

Their website can be found at

http://www.uniprot.org/

Protein Sequence Databases

There are three major parts of UniProt:

1. The UniProt Archive (or UniParc):
   - New sequences appear daily.
   - Redundancy is eliminated by collapsing sequences that are identical (regardless if they are in separate species).
   - Also contains sequence data from International Protein Index, RefSeq, FlyBase and WormBase.
2. The UniProt Knowledgebase:
   - Curated with much care.
   - Redundancy is eliminated by collapsing all protein products derived from a certain gene in a certain species into a single record.
   - Complete proteomes available.

Protein Sequence Databases

3. The UniProt nonredundant reference database (or UniRef): contains nonredundant information from UniParc and UniProt Knowledgebase.
    - Eliminates redundancy, even across multiple species by collapsing into a single record.
    - UniRef100 collapses into one record if they share 11 consecutive residues that are the same.
    - UniRef90 merges further (roughly), if sequences are 90% similarity.
    - UniRef50 merges (roughly) if sequences are 50% similar.