LVA 320.004

Genome data analysis

Computer lab session 1

Theresa Schwarz, MSc

theresa.schwarz@jku.at
Institute of Biophysics, JKU

Overview

- ☐ **Grading/Requirements**
- ☐ Introduction to databases
- ☐ Examples → browsing through NCBI
- ☐ How to write a Report

Grading

Final Exam	50%
Small Assay	10%
Reports	**35%**
Attendance / Participation in class	5%

Grading scale:

1: 100-90%, 2: 89-80%, 3: 79-70%, 4: 69-60%, 5: 59-0%

Resources

- PDF files of the lecture available in MOODLE
- digital recordings of the lectures available in MOODLE
- Access to MOODLE: https://moodle.jku.at/jku2015/index.php

Overview

- ☐ Grading/Requirements
- ☐ **Introduction to databases**
- ☐ Examples → browsing through NCBI
- ☐ How to write a Report

Genome data analysis

?

Genome data analysis

- genetic material of an organism

- chromosomes
  - centromeres
  - telomeres

- genes = coding regions
  - DNA – RNA – protein
- non-coding regions = "junk DNA" ?
  - DNA – ncRNA – x
    miRNA
  - repetitive DNA → satellite DNA
  - transposons

- nucleotide sequence
  - the human genome:
    3×10^9 bp
    3 billion bp = 3 Milliarden bp

```
GGATGCGTCGGGTGCGGGTTCCTTCCGAGTTCCCTGG
GTCGGGTACGGGTGCCTTCCGAGTTCCCTGGAACGGG
TTCGGGTACGGGTTCCTTCCGAGTTCCCTGGAACGGG
GGATGCTTCGGGTACGGGTTCCTTCCGAGTTCCCTGG
TCAGTGCGTCTTTCAGTAATGAGAATCCGCCGAACAC
ACGGGACGCCATAGAGGGTGAGAGCCCCGTCTGGTAG
GGATGCTTCGGGTGCGGGTACCTACTGAGTTCCCTGG
TTTGTAGAGGATGCTTCGAGTTATGGTTCCTTCCGAG
ATCTGGCTCCTTCGGGGTCCGAGTTGTAATTTGTAGA
TGAAGCGGCAACAGCTCAAATTTGAAATCTGGCTCCT
GATGCCTCGGGTACGTGTGCCTACTGAGTTCCCTGGA
GGATGCTTCGGGTACGGGTGCCTACTGAGTTCCCTGG
GACGCCATAAAGGGTGAGAGCCCCCGTCTGGTAGGACA
GGATGCTTCGGGTGCGGGTTCCTTCCGAGTTCCCTGG
AGGATGCCTCGGGTGCGGGTTCCTTCCGAGTTCCCTG
GATGCTTCGAGTTATGGTTCCTTCCGAGTTCCCTGGA
```

Theresa Schwarz

DNA

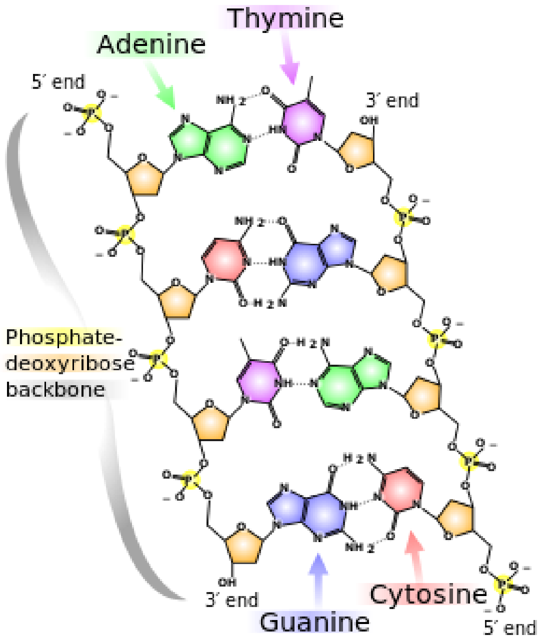
- DNA = **D**eoxyribo**n**ucleic **a**cid

- **Structure**:
  - Phosphate Backbone
  - deoxyribose (sugar)
  - Bases:
    - A (adenine)
    - G (guanine)
    - T (thymine)
    - C (cytosine)

A (adenine), G (guanine): purines

T (thymine), C (cytosine): pyrimidines

Phosphate Backbone, deoxyribose (sugar), Bases: **nucleotide**



- A $=$ T
- G $\equiv$ C

pairing via hydrogen bonds

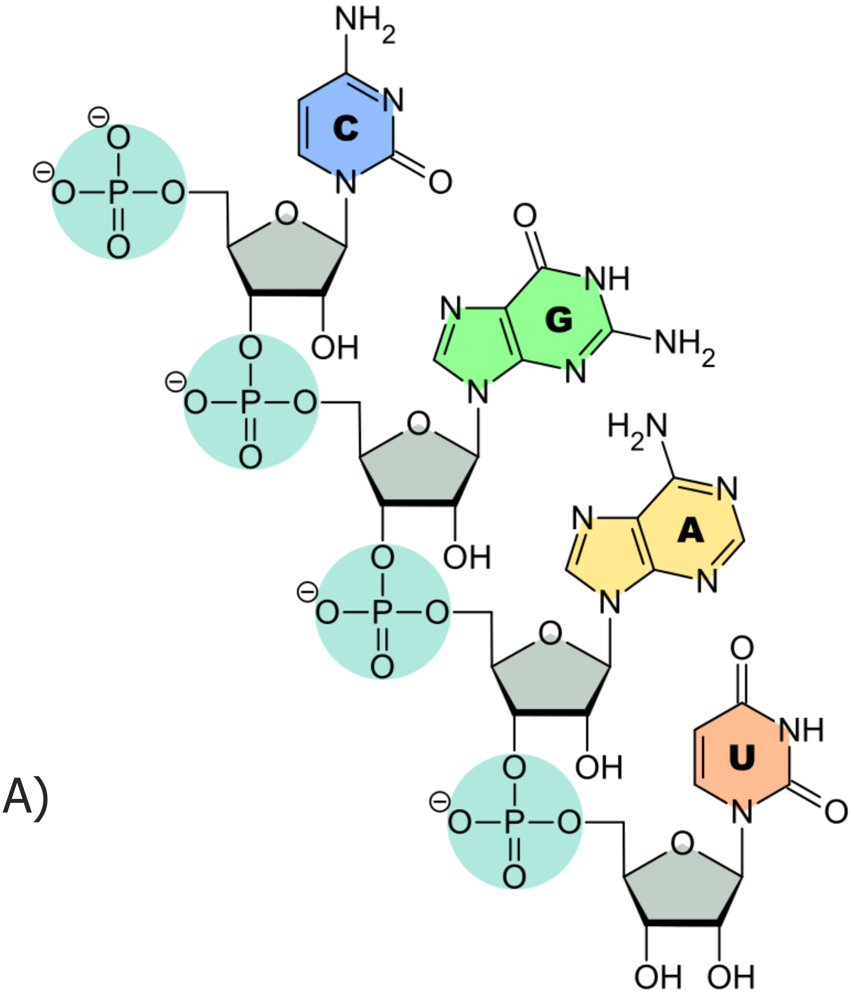
- Coding regions: **genes**
- Non-coding regions

RNA

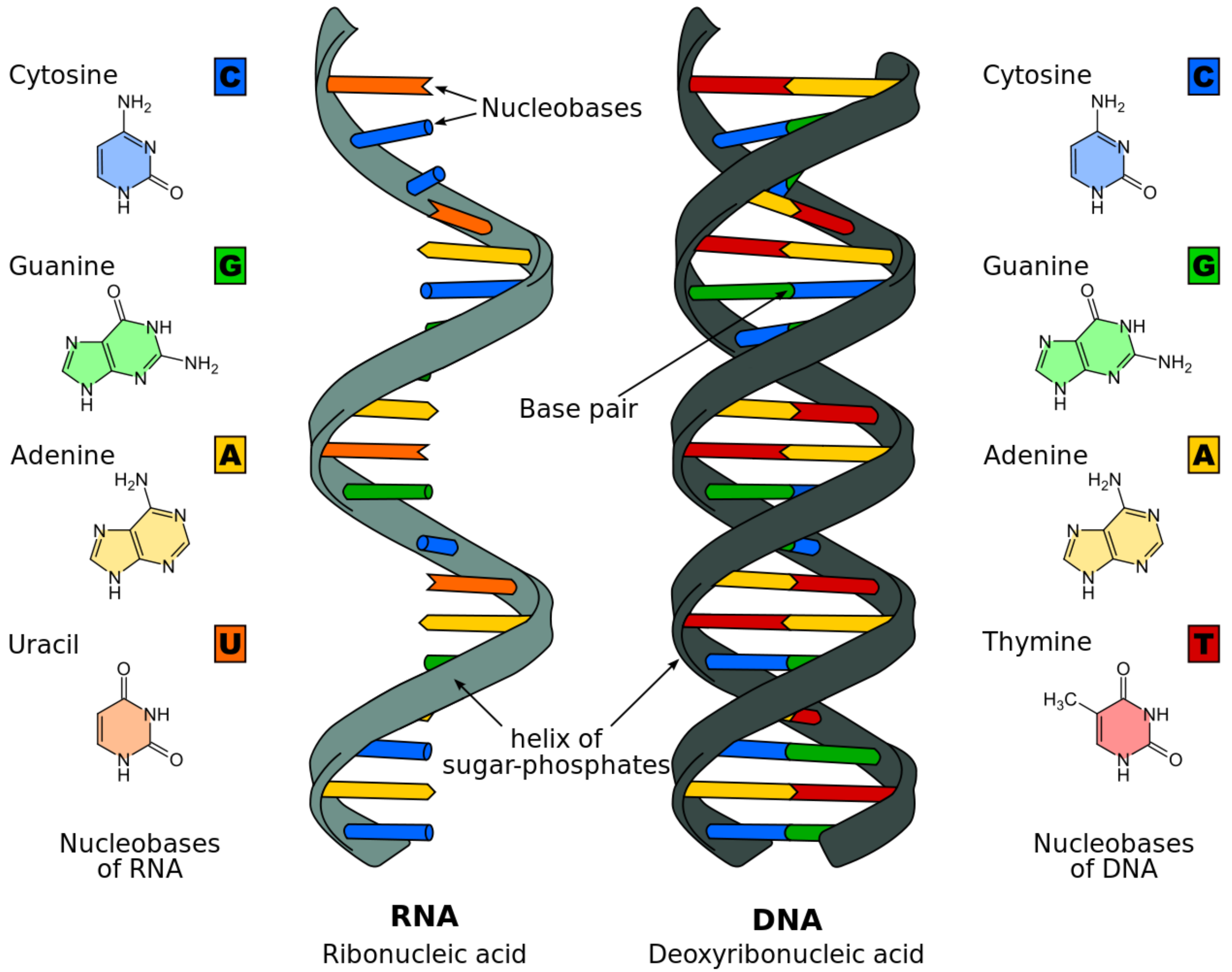
- RNA= **R**ibo**n**ucleic **a**cid

- **Structure**:
  - Phosphate Backbone
  - ribose (sugar)
  - Bases: A (adenine)
    G (guanine)
    U (uracil)
    C (cytosine)

- **Types**:
  - coding: mRNA (messenger RNA)
  - non-coding: rRNA (ribosomal RNA)
    tRNA (transfer RNA)
    miRNA (micro RNA)
    snRNA (small nuclear RNA) …



DNA vs. RNA



Nucleobases of RNA

Nucleobases of DNA

RNA
Ribonucleic acid

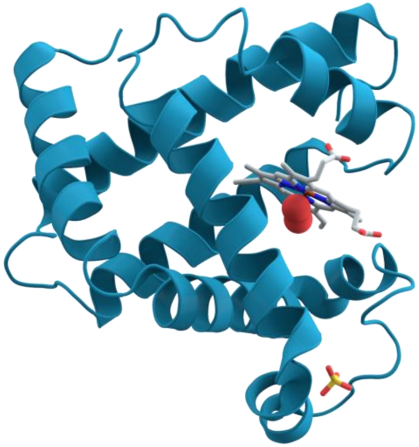
DNA
Deoxyribonucleic acid

Proteins

- DNA is **transcribed** to mRNA which is **translated** to amino acid sequence = PROTEIN

- **Nucleotide sequence**
  atgccaagtacgcccctattgacgtcaatgacggtaaatggcccgcctgg

- **Codon/Base-triplets (mRNA)** coding for one amino acid



Myoglobin

- **Amino acid sequence**
  - 3-letter code
    Met-Pro-Ser-Thr-Pro-Pro-Ile-Asp-Val-Asn-Asp-Gly-Lys-Trp-Pro
  - single letter code
    MPSTPPIDVNDGKWPAWHYAQYMTLWDFPTWQYIYVLVIAIT

Amino acid codes

Name	3-Letter	1-Letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I

Name	3-Letter	1-Letter
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Genome data analysis

- genetic material of an organism

- nucleotide sequence
- different species
- genes

- chromosomes
  - centromeres
  - telomeres

- genes = coding regions
- non-coding regions = "junk DNA" ?
  - DNA – RNA – protein
  - DNA – ncRNA – x
    miRNA
  - repetitive DNA → satellite DNA
  - transposons

- nucleotide sequence
  - the human genome:
    3x10^9 bp
    3 billion bp = 3 Milliarden bp

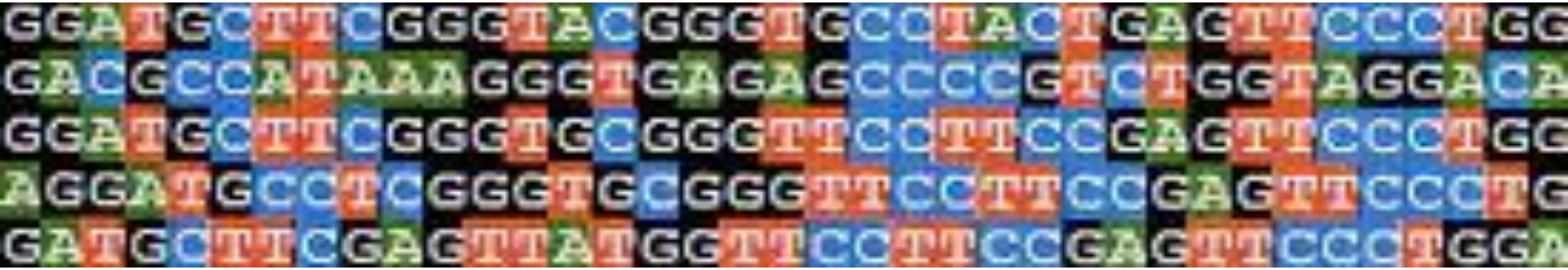
Example: DNA sequencing

GGATGCGTCGGGTGCGGGTTCCTTCCGAGTTCCCTGG
GTCGGGTACGGGTGCCTTCCGAGTTCCCTGGAACGGG
TTCGGGTACGGGTTCCTTCCGAGTTCCCTGGAACGGG
GGATGCTTCGGGTACGGGTTCCTTCCGAGTTCCCTGG

You will get DNA fragments of 1.000 base pairs

\---

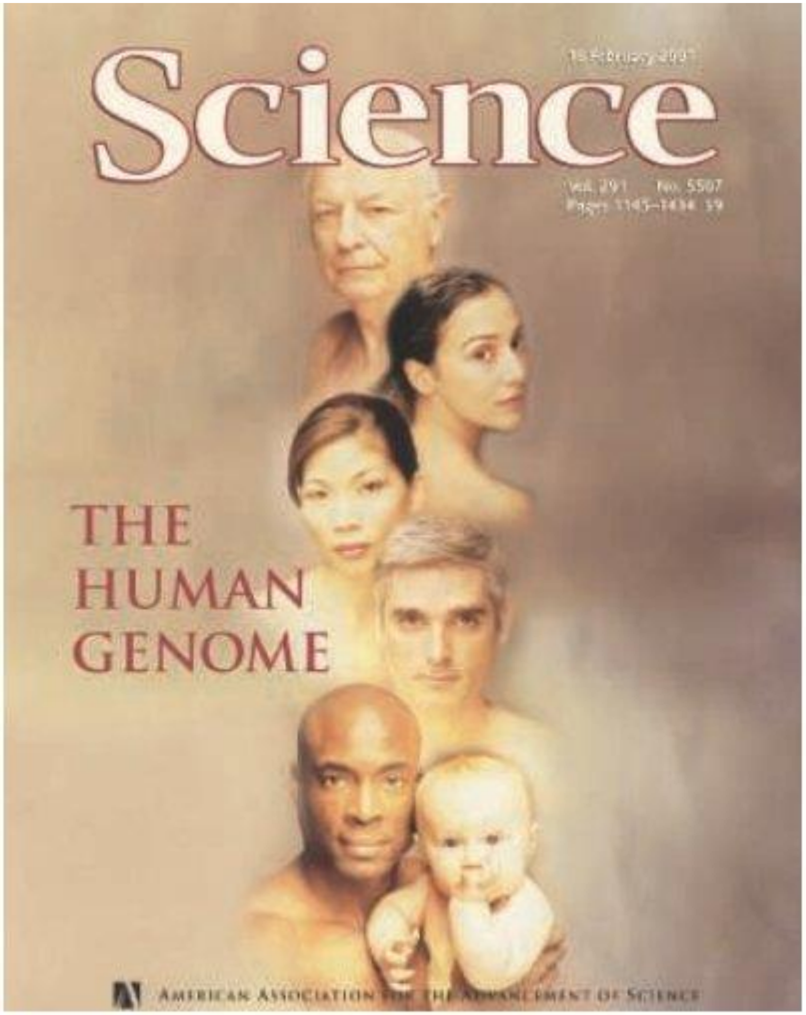
The human genome has 3.000.000.000 base pairs

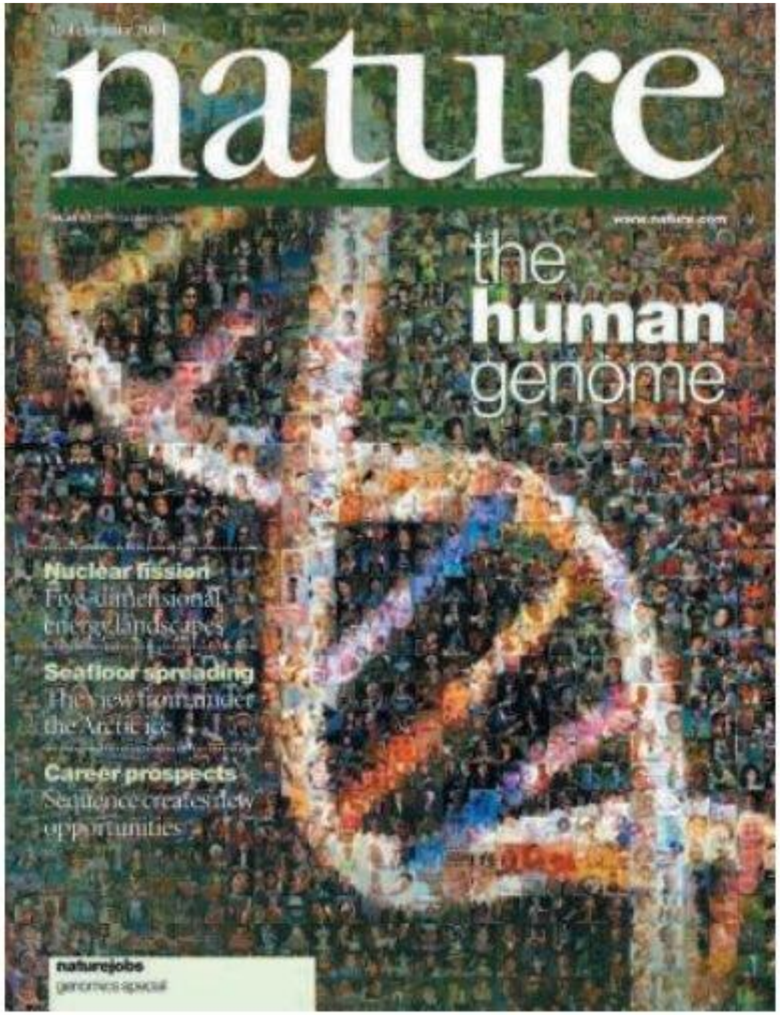


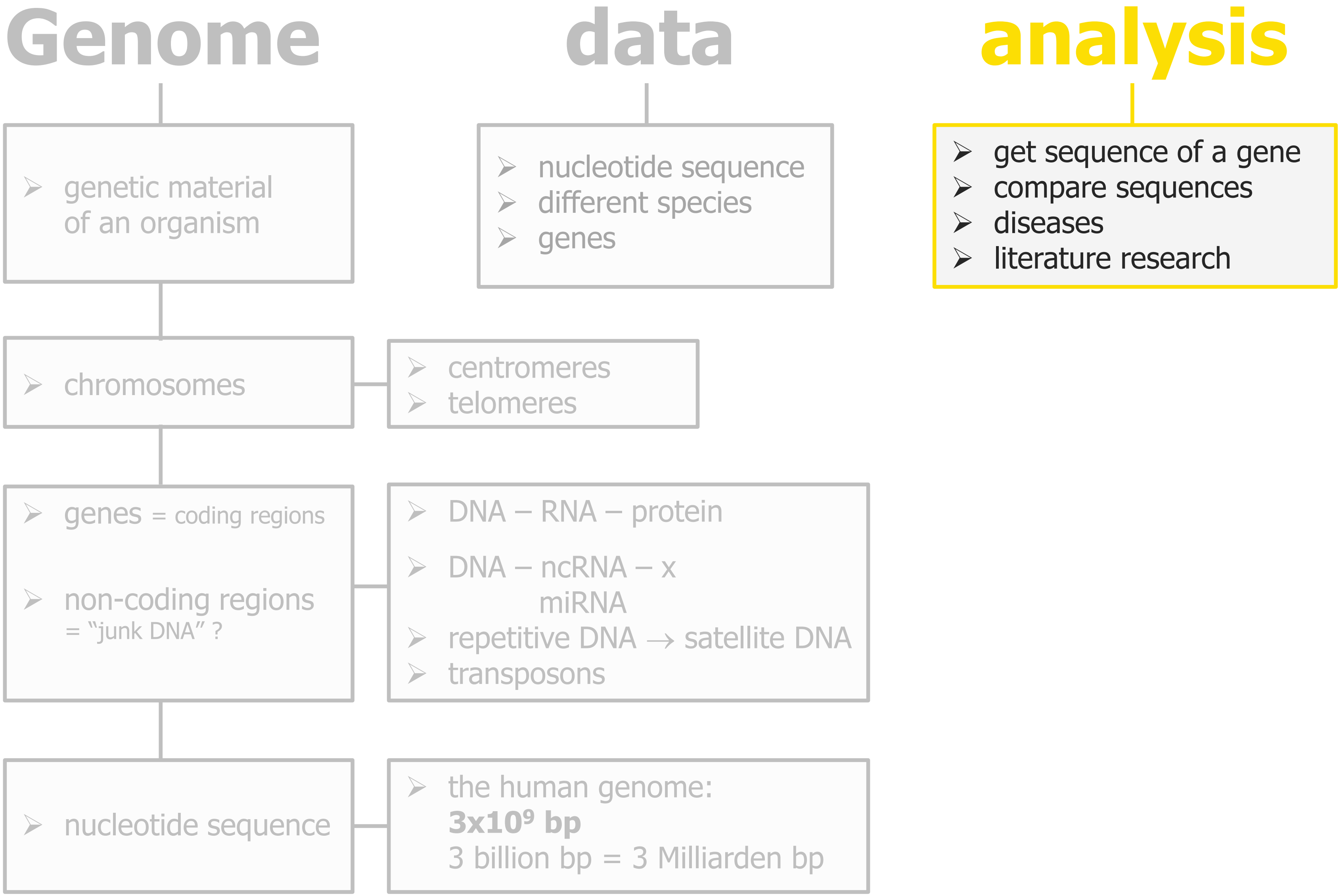
GGATGCTTCGGGTACGGGTGCCTACTGAGTTCCCTGG
GACGCCATAAAGGGTGAGAGCCCCCGTCTGGTAGGACA
GGATGCTTCGGGTGCGGGTTCCTTCCGAGTTCCCTGG
AGGATGCCTCGGGTGCGGGTTCCTTCCGAGTTCCCTG
GATGCTTCGAGTTATGGTTCCTTCCGAGTTCCCTGGA

Sequence of the human genome

- first published in February 2001







Genome data analysis

Genome

- genetic material of an organism
- chromosomes
  - centromeres
  - telomeres
- genes = coding regions
- non-coding regions = "junk DNA" ?
  - DNA – RNA – protein
  - DNA – ncRNA – x
    miRNA
  - repetitive DNA → satellite DNA
  - transposons
- nucleotide sequence
  - the human genome:
    3x10^9 bp
    3 billion bp = 3 Milliarden bp

data

- nucleotide sequence
- different species
- genes

analysis

- get sequence of a gene
- compare sequences
- diseases
- literature research

Genome data analysis

Genome

- genetic material of an organism

- chromosomes
  - centromeres
  - telomeres

- genes = coding regions
  - DNA – RNA – protein
- non-coding regions = "junk DNA" ?
  - DNA – ncRNA – x
    miRNA
  - repetitive DNA → satellite DNA
  - transposons

- nucleotide sequence
  - the human genome:
    $3x10^9$ bp
    3 billion bp = 3 Milliarden bp

data

- nucleotide sequence
- different species
- genes

analysis

- get sequence of a gene
- compare sequences
- diseases
- literature research

Online databases

- **What are genomic databases and why are they useful?**

- Collection of genomic information, online available
  (DNA profiles, protein specifications, …)

- get information of gene / genomic region / protein / etc. of interest you need for your research

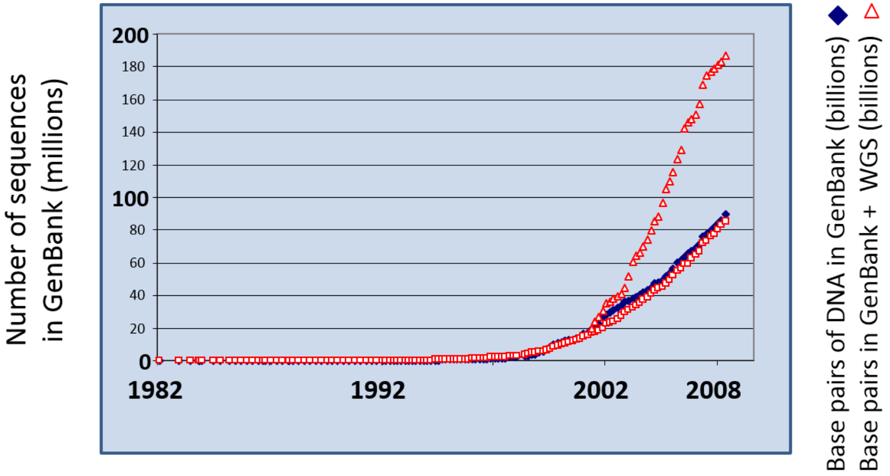
- analyze obtained data and compare with available data

- …

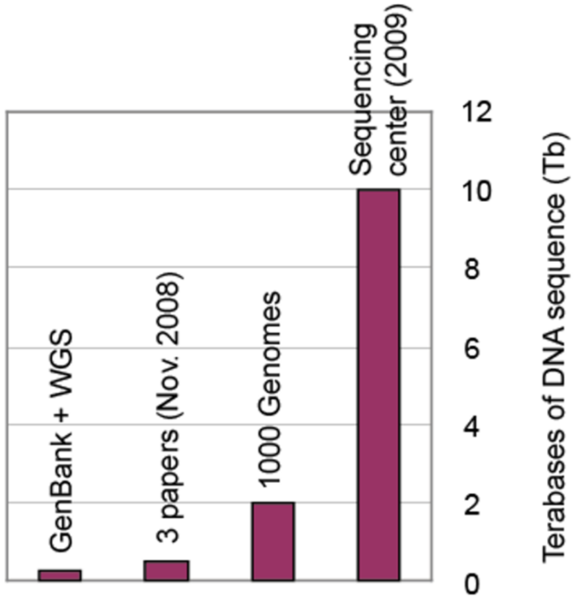
Necessity of DNA databases

- Due to advances in DNA sequencing technologies the number of sequenced genomes rose drastically.

Growth of GenBank + Whole Genome Shotgun (1982-November 2008): we reached 0.2 Tb



Arrival of next-generation sequencing: In two years we have gone from 0.2 Tb to 71 Tb (71,000 Gbases) (November 2010)



There are three major public DNA databases

http://www.ensembl.org

http://www.ddbj.nig.ac.jp/

http://www.ncbi.nlm.nih.gov/genbank

EMBL ⟷ **GenBank** ⟷ **DDBJ**

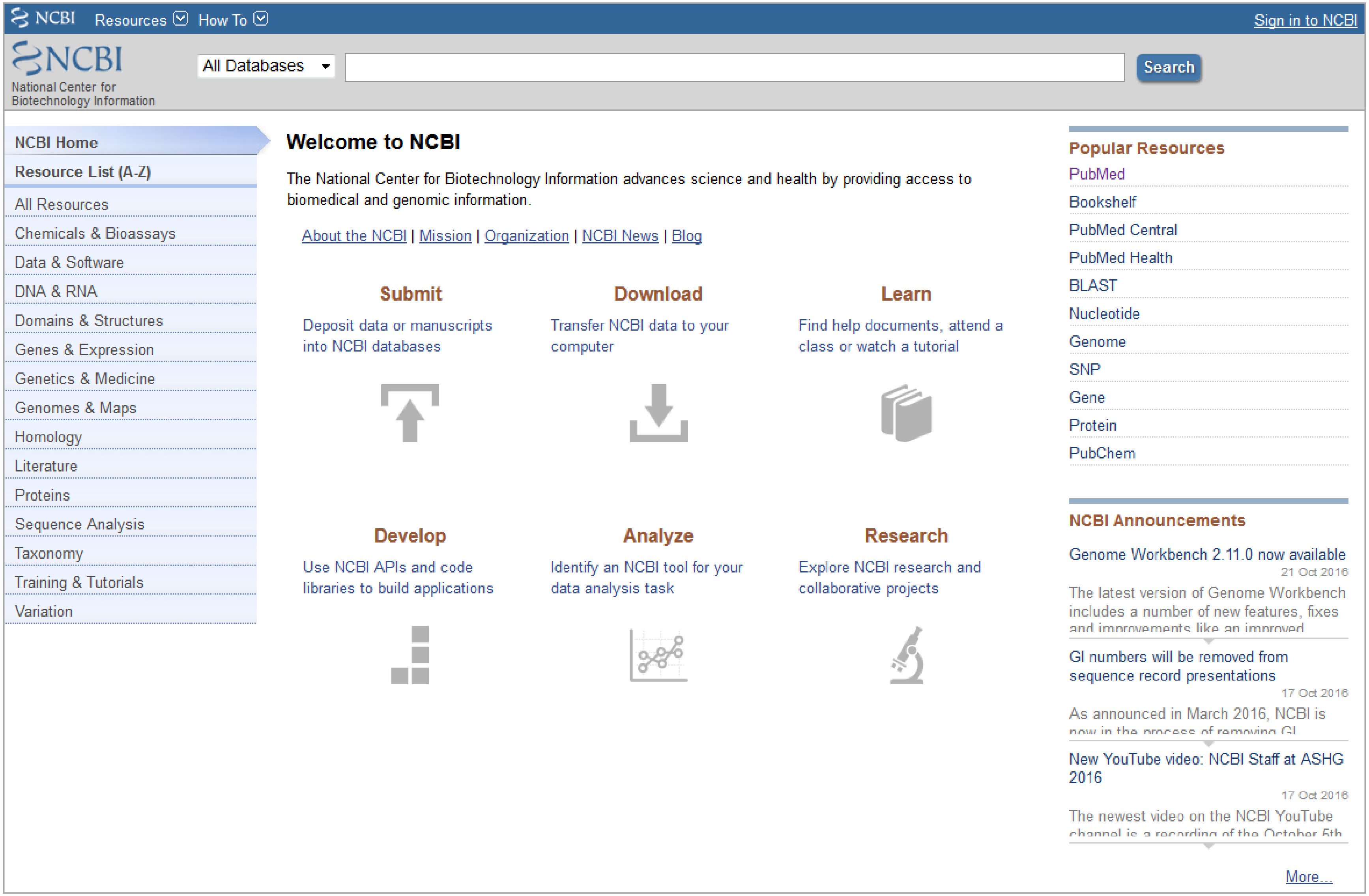
Housed at EBI
European Bioinformatics Institute

Housed at NCBI
National Center for Biotechnology Information

Housed in Japan
DNA Database of Japan

NCBI homepage

- http://www.ncbi.nlm.nih.gov/



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

Develop

Use NCBI APIs and code libraries to build applications

Analyze

Identify an NCBI tool for your data analysis task

Research

Explore NCBI research and collaborative projects

Popular Resources

NCBI Announcements

Genome Workbench 2.11.0 now available
21 Oct 2016
The latest version of Genome Workbench includes a number of new features, fixes and improvements like an improved

GI numbers will be removed from sequence record presentations
17 Oct 2016
As announced in March 2016, NCBI is now in the process of removing GI

New YouTube video: NCBI Staff at ASHG 2016
17 Oct 2016
The newest video on the NCBI YouTube channel is a recording of the October 5th

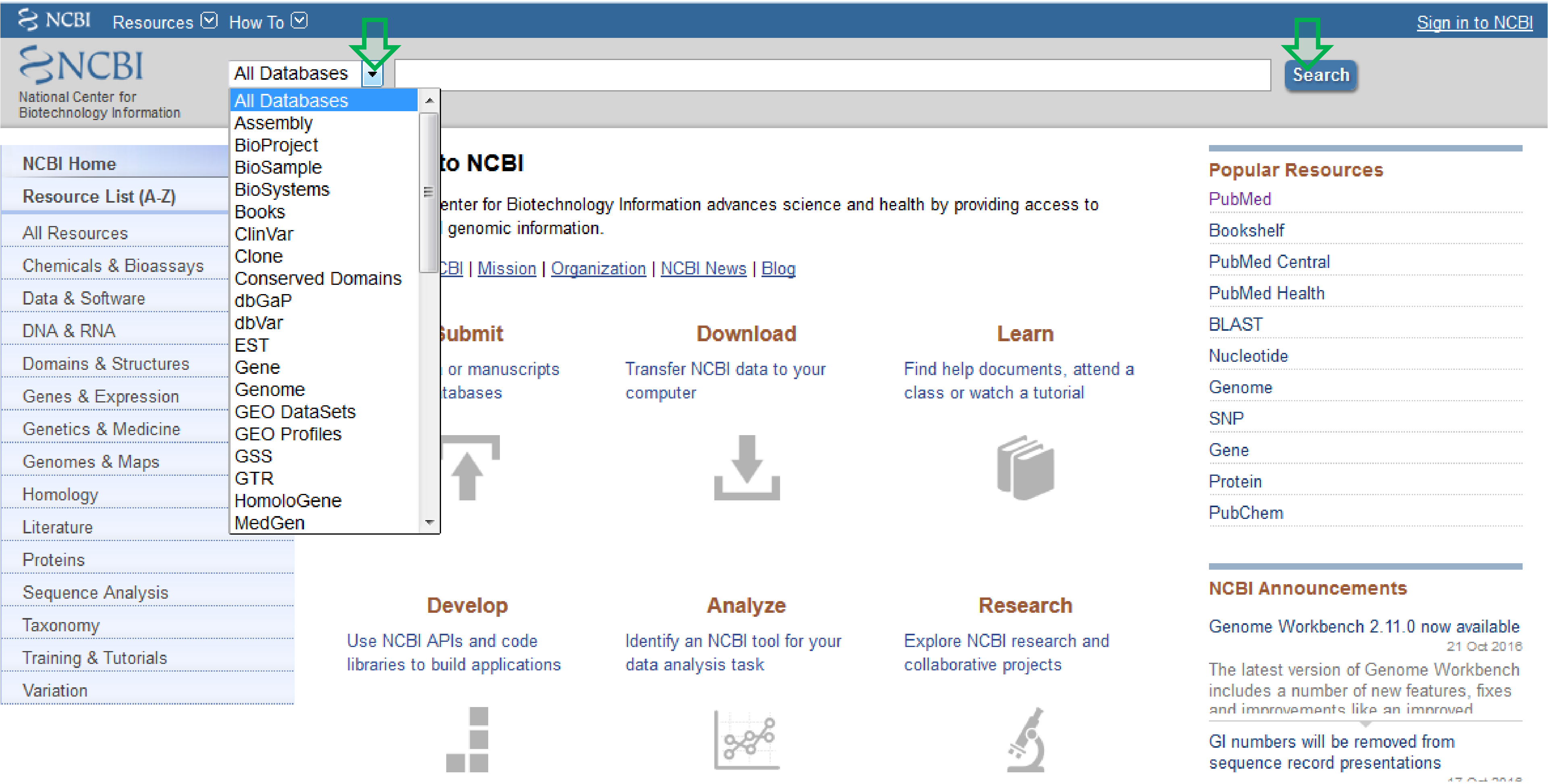
More...

NCBI databases

Use the dropdown menu

NCBI is a collection of many databases

And click "Search" to get details



to NCBI

enter for Biotechnology Information advances science and health by providing access to genomic information.

ubmit
or manuscripts tabases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

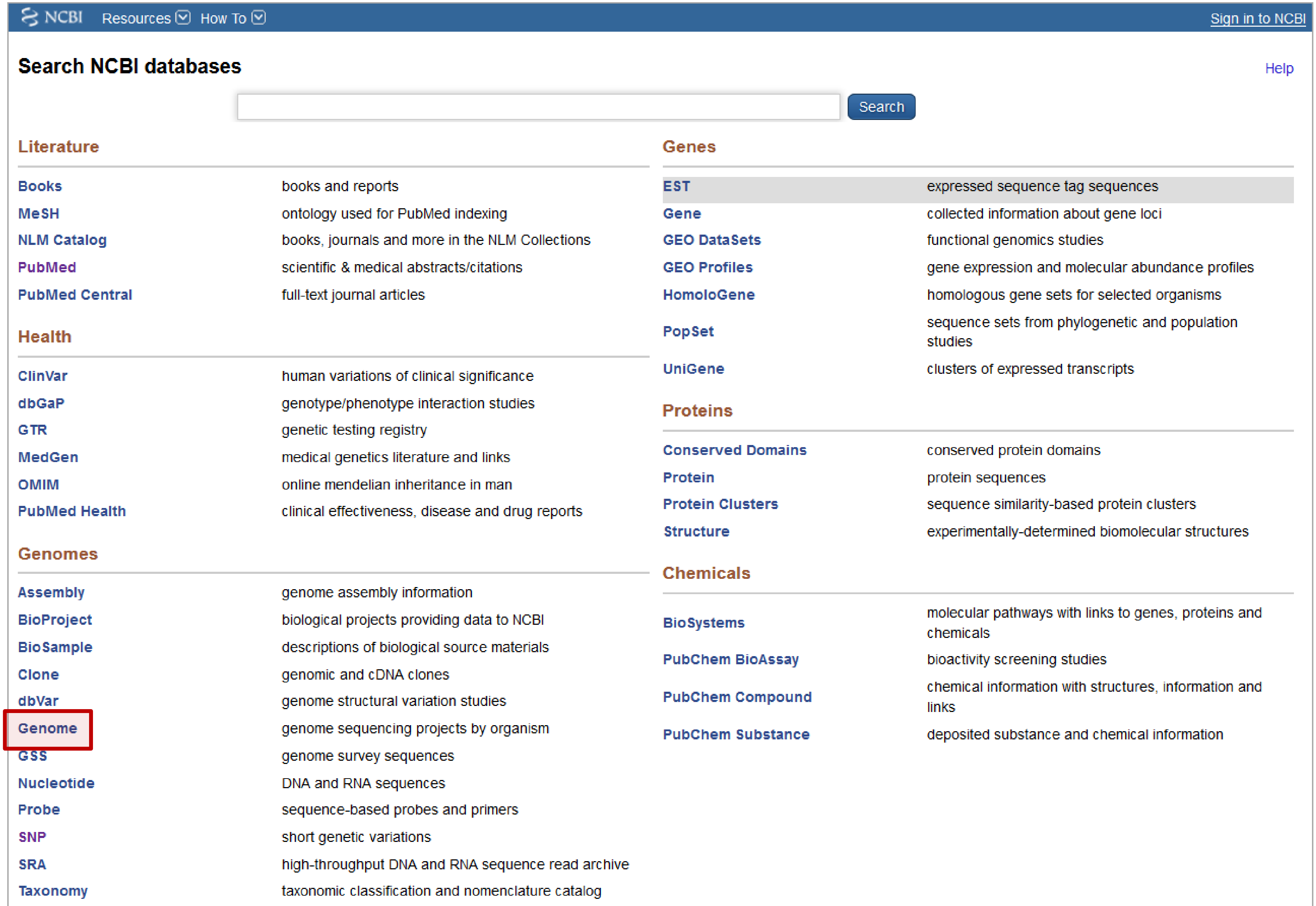
NCBI Announcements

Genome Workbench 2.11.0 now available
21 Oct 2016

The latest version of Genome Workbench includes a number of new features, fixes and improvements like an improved

GI numbers will be removed from sequence record presentations

NCBI databases



Search NCBI databases

Literature

Books	books and reports
MeSH	ontology used for PubMed indexing
NLM Catalog	books, journals and more in the NLM Collections
PubMed	scientific & medical abstracts/citations
PubMed Central	full-text journal articles

Health

ClinVar	human variations of clinical significance
dbGaP	genotype/phenotype interaction studies
GTR	genetic testing registry
MedGen	medical genetics literature and links
OMIM	online mendelian inheritance in man
PubMed Health	clinical effectiveness, disease and drug reports

Genomes

Assembly	genome assembly information
BioProject	biological projects providing data to NCBI
BioSample	descriptions of biological source materials
Clone	genomic and cDNA clones
dbVar	genome structural variation studies
Genome	genome sequencing projects by organism
GSS	genome survey sequences
Nucleotide	DNA and RNA sequences
Probe	sequence-based probes and primers
SNP	short genetic variations
SRA	high-throughput DNA and RNA sequence read archive
Taxonomy	taxonomic classification and nomenclature catalog

Genes

EST	expressed sequence tag sequences
Gene	collected information about gene loci
GEO DataSets	functional genomics studies
GEO Profiles	gene expression and molecular abundance profiles
HomoloGene	homologous gene sets for selected organisms
PopSet	sequence sets from phylogenetic and population studies
UniGene	clusters of expressed transcripts

Proteins

Conserved Domains	conserved protein domains
Protein	protein sequences
Protein Clusters	sequence similarity-based protein clusters
Structure	experimentally-determined biomolecular structures

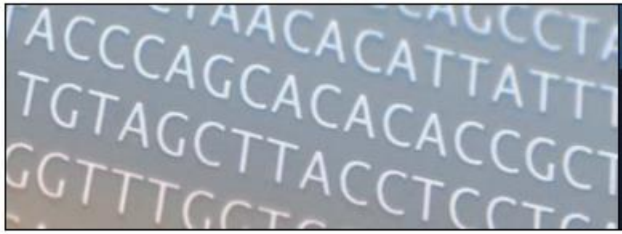
Chemicals

BioSystems	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	bioactivity screening studies
PubChem Compound	chemical information with structures, information and links
PubChem Substance	deposited substance and chemical information

NCBI databases

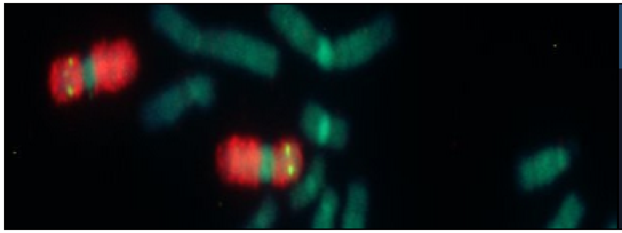
Genome

This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations.



Nucleotide

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

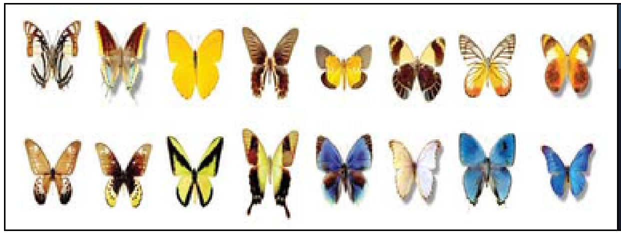


Gene

Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

Protein

The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.



Taxonomy

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.

How does NCBI get the data?

- Scientist will publish/submit their data → e.g. new sequences
- Curators will verify the data and link them to the different databases

---

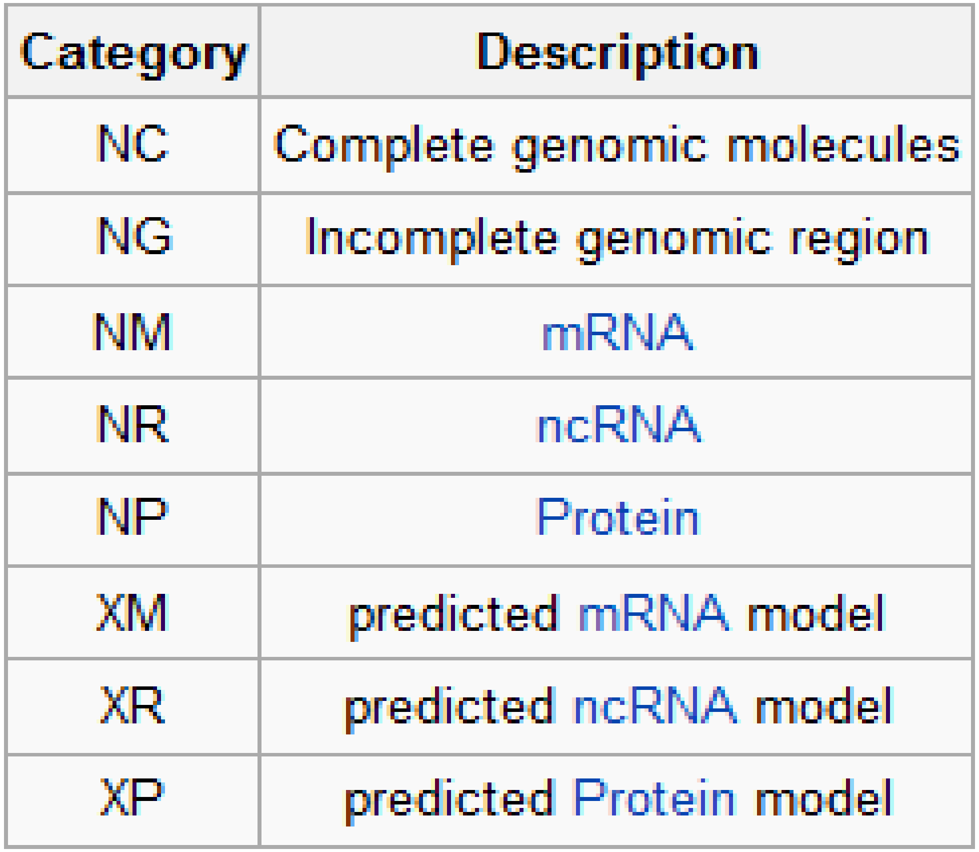
- The most reliable sequences are called Reference Sequences (RefSeq)
- RefSeqs are the most stable references for studies since they are
  - **non-redundant**
  - **well-annotated (introns, exons)**
  - **curated, reviewed**

Identification numbers

GeneIDs	Gene ID: 11830
gi (gene info identifier) numbers	gi | 147905620 |
mRNA accession numbers	NM_020227
mRNA version numbers	NM_020227.2
Uni gene accession numbers	UniGene Mm.45580
EST identification numbers	BY704936.1
GeneBank protein numbers	AAC02945
PDB (protein data bank) accession numbers	1KT7

Accession numbers & prefixes

- **NC**_000005.10
- **NM**_020227.2
- **NP**_064612.2



Category	Description
NC	Complete genomic molecules
NG	Incomplete genomic region
NM	mRNA
NR	ncRNA
NP	Protein
XM	predicted mRNA model
XR	predicted ncRNA model
XP	predicted Protein model

FASTA sequence

- Text-based format representing nucleotide or amino acid sequences
- Standard format in the field of bioinformatics
- Databases and programs recognize a nucleotide or amino acid sequence when it is written in FASTA format

- FASTA sequence characteristics:

  > „greater than" sign marks the start of a FASTA sequence
  followed by a description of the sequence in the same line
  followed by a line break (↵ ENTER)

  in the second line comes the nucleotide or amino acid sequence in the single letter code

FASTA sequence

Examples:

- >nucleotide_sequence_XY
  CTCACTAAAAAAGATAAGACAACAAACAAGCATCCAAAAGAATGCTCATGGCACTTGTCATAG
  GGAAAAGCAAATTAATGCAACAATGGTATACCACTGAGCATCTACTAATAGACAACAAACAAGC
  ATCCAATTAATGCAACAAT

- >amino_acid_sequence_XY
  MSPEKSQEESPEEDTERTERKPMVKDAFKDISIYFTKEEWAEMGDWEKTRYRNVKRNYNALITI
  GLRATRPAFMCHRRQAIKLQVDDTEDSDEEWTPRQQVKPPWMALRVEQRHRRQAIKLQVDD
  TEDSDEEKDISIYFTKEEWAEMGD

Overview

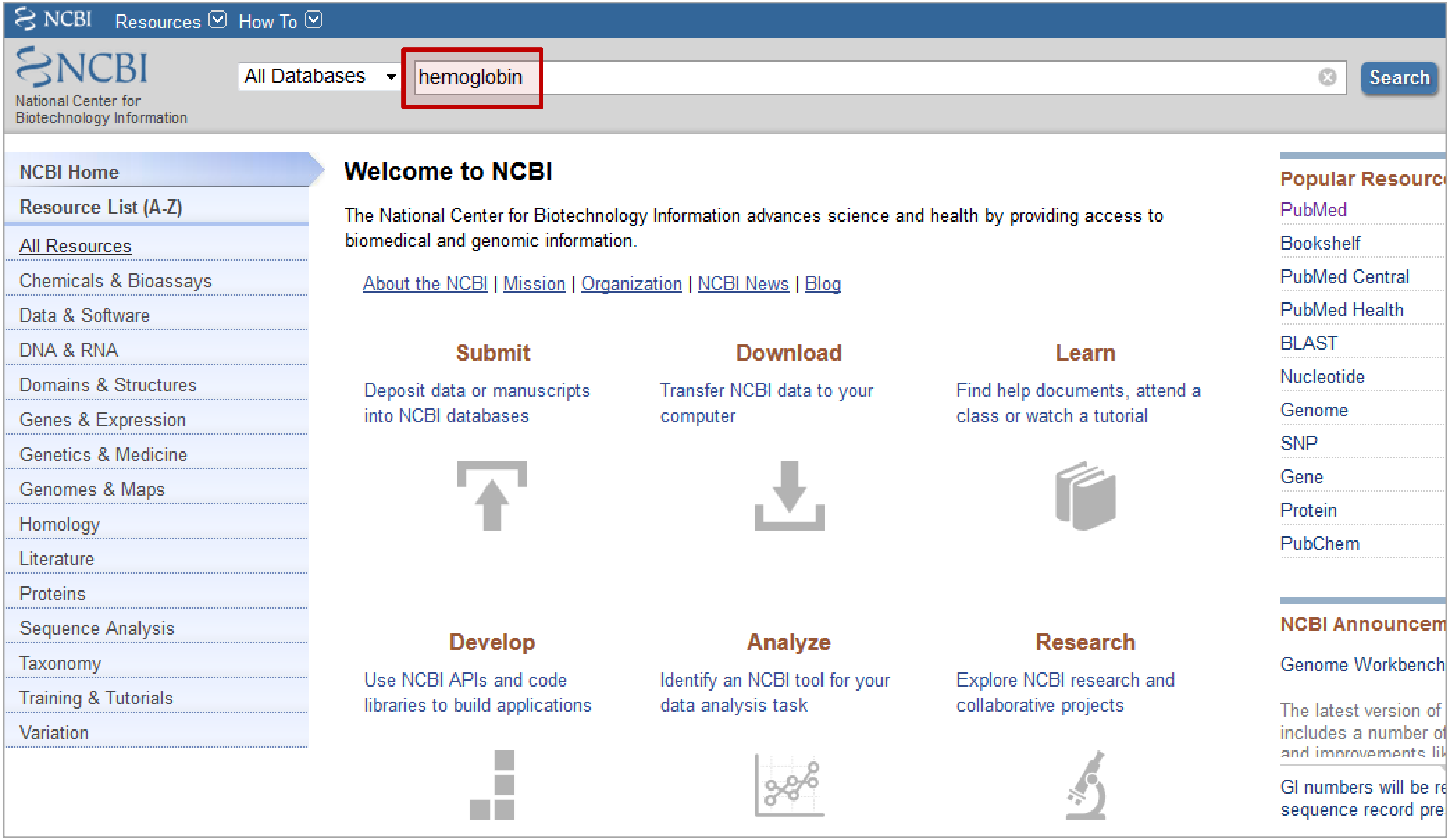
- ☐ Grading/Requirements
- ☐ Introduction to databases
- ☐ **Examples → browsing through NCBI**
- ☐ How to write a Report

NCBI databases

Today we will get familiar with the following NCBI databases:

- Gene Database
- CCDS Database
- UniGene Database
- Taxonomy Database
- Nucleotide Database
- Protein Database
- OMIM Database

NCBI – search through databases



- **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

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

Develop

Use NCBI APIs and code libraries to build applications

Analyze

Identify an NCBI tool for your data analysis task

Research

Explore NCBI research and collaborative projects

Popular Resourc

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

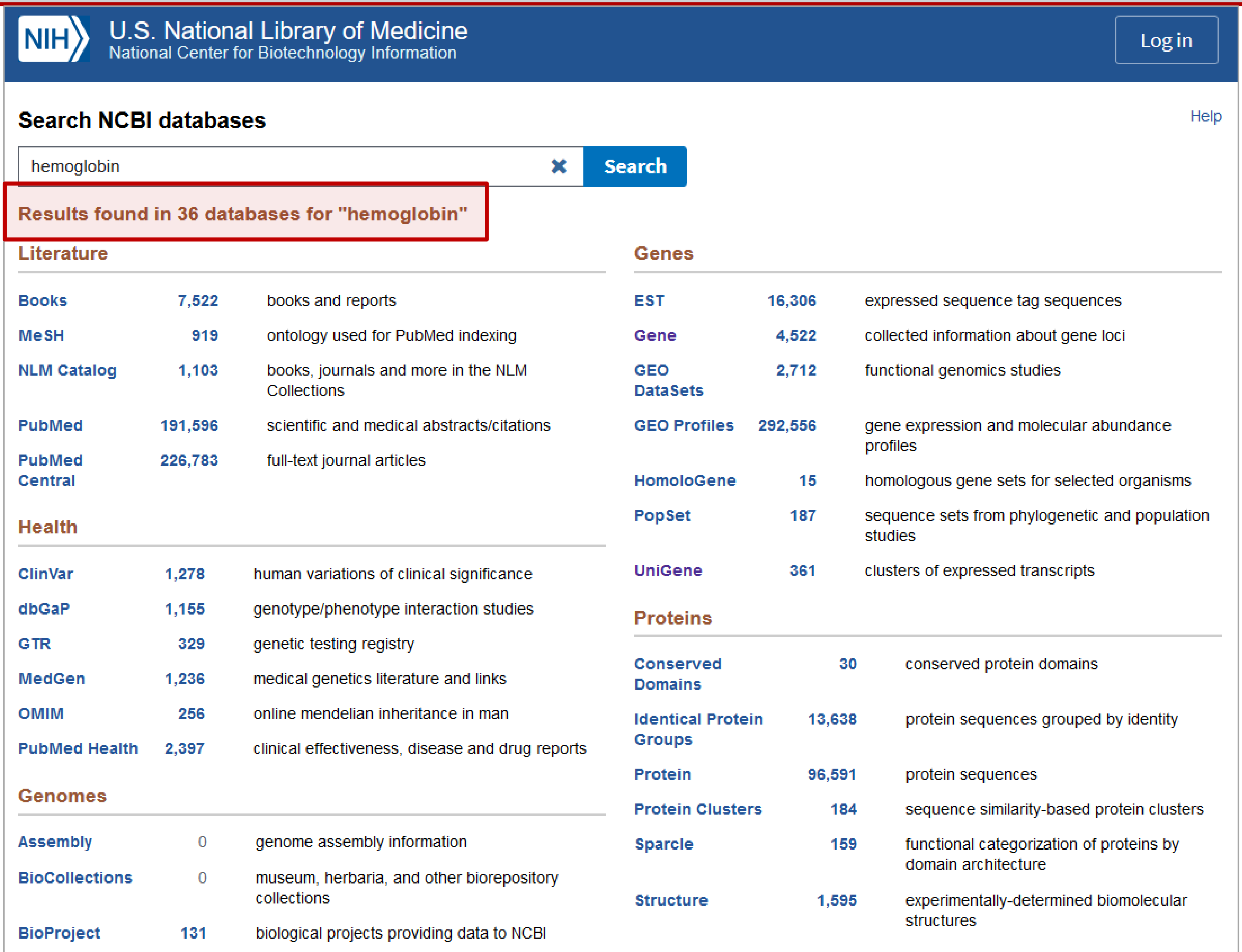
NCBI Announcem

Genome Workbench

The latest version of includes a number of and improvements lik

GI numbers will be re sequence record pre

NCBI – search through databases



Search NCBI databases

Results found in 36 databases for "hemoglobin"

Literature

Books	7,522	books and reports
MeSH	919	ontology used for PubMed indexing
NLM Catalog	1,103	books, journals and more in the NLM Collections
PubMed	191,596	scientific and medical abstracts/citations
PubMed Central	226,783	full-text journal articles

Health

ClinVar	1,278	human variations of clinical significance
dbGaP	1,155	genotype/phenotype interaction studies
GTR	329	genetic testing registry
MedGen	1,236	medical genetics literature and links
OMIM	256	online mendelian inheritance in man
PubMed Health	2,397	clinical effectiveness, disease and drug reports

Genomes

Assembly	0	genome assembly information
BioCollections	0	museum, herbaria, and other biorepository collections
BioProject	131	biological projects providing data to NCBI

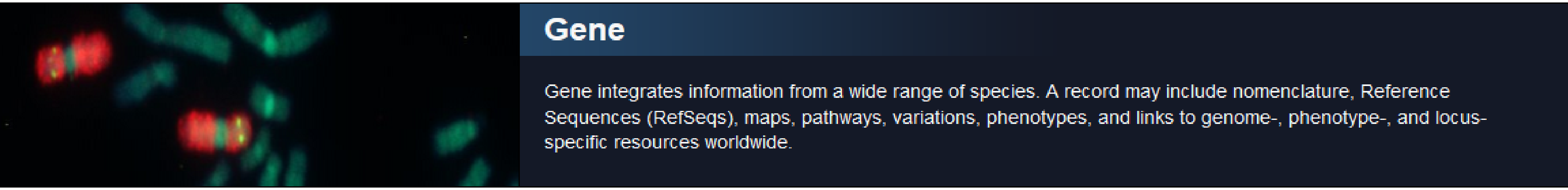
Genes

EST	16,306	expressed sequence tag sequences
Gene	4,522	collected information about gene loci
GEO DataSets	2,712	functional genomics studies
GEO Profiles	292,556	gene expression and molecular abundance profiles
HomoloGene	15	homologous gene sets for selected organisms
PopSet	187	sequence sets from phylogenetic and population studies
UniGene	361	clusters of expressed transcripts

Proteins

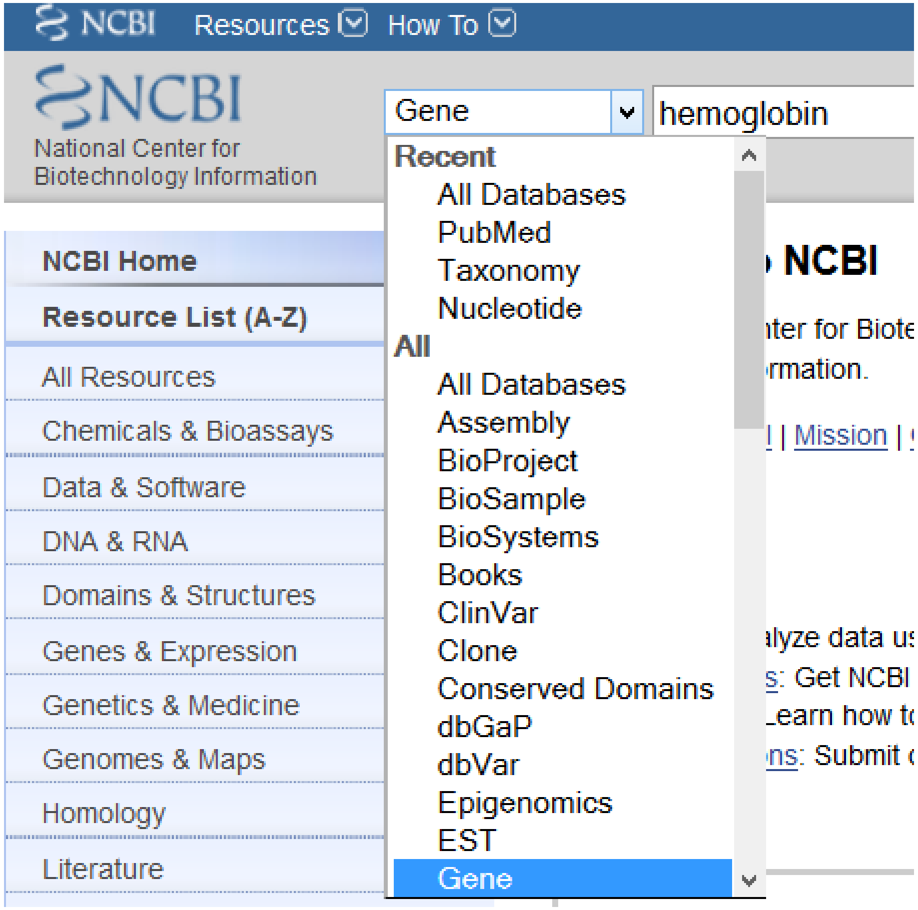
Conserved Domains	30	conserved protein domains
Identical Protein Groups	13,638	protein sequences grouped by identity
Protein	96,591	protein sequences
Protein Clusters	184	sequence similarity-based protein clusters
Sparcle	159	functional categorization of proteins by domain architecture
Structure	1,595	experimentally-determined biomolecular structures

Gene Database



Gene

Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

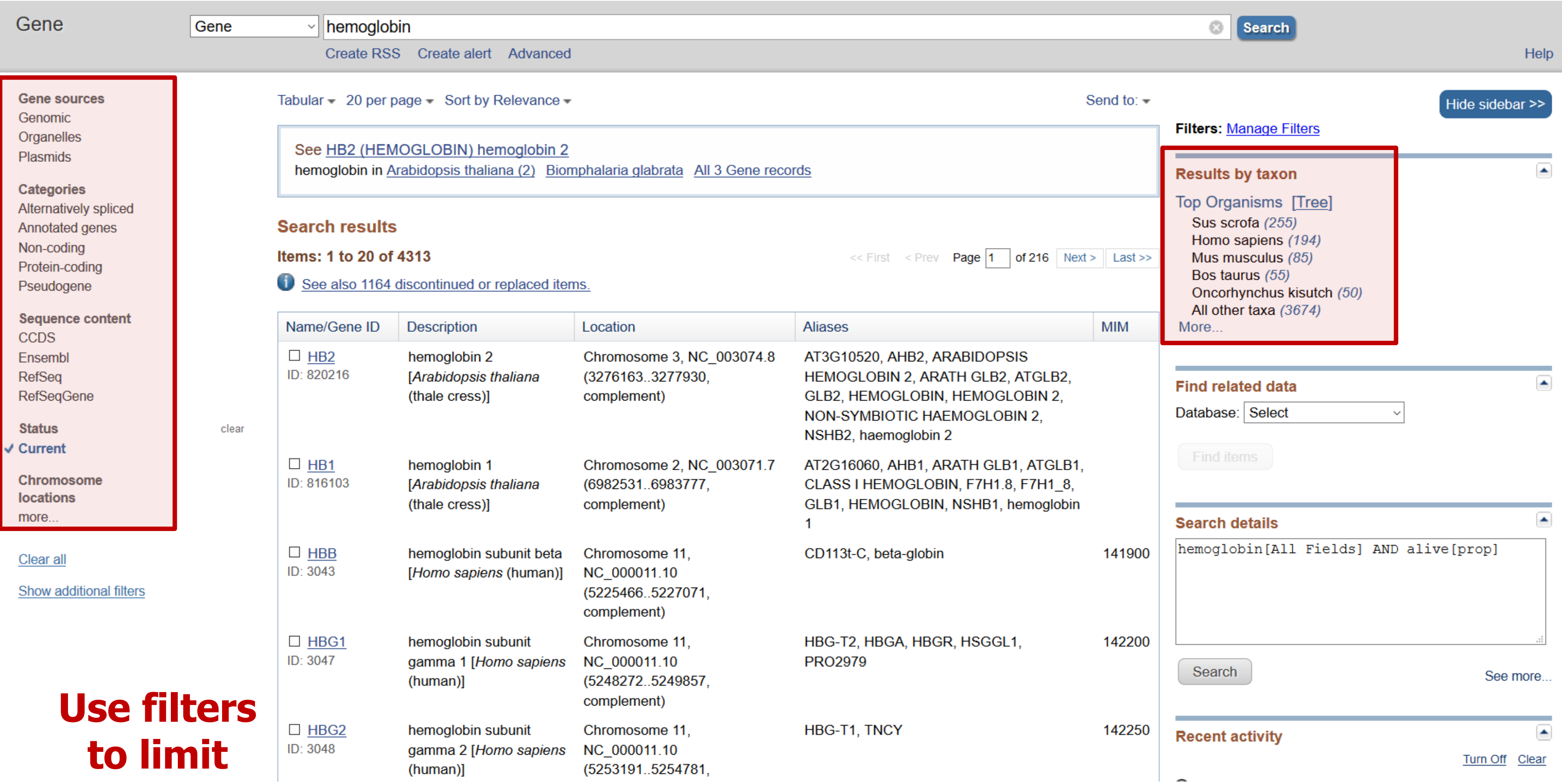


Gene

hemoglobin

choose the
Gene Database
and search for
hemoglobin

Gene Database



hemoglobin

Gene sources
Genomic
Organelles
Plasmids

Categories
Alternatively spliced
Annotated genes
Non-coding
Protein-coding
Pseudogene

Sequence content
CCDS
Ensembl
RefSeq
RefSeqGene

Status
✓ Current

Chromosome locations
more...

See HB2 (HEMOGLOBIN) hemoglobin 2
hemoglobin in Arabidopsis thaliana (2) Biomphalaria glabrata All 3 Gene records

Search results

Items: 1 to 20 of 4313

Page 1 of 216

See also 1164 discontinued or replaced items.

Name/Gene ID	Description	Location	Aliases	MIM
☐ HB2 ID: 820216	hemoglobin 2 [*Arabidopsis thaliana* (thale cress)]	Chromosome 3, NC_003074.8 (3276163..3277930, complement)	AT3G10520, AHB2, ARABIDOPSIS HEMOGLOBIN 2, ARATH GLB2, ATGLB2, GLB2, HEMOGLOBIN, HEMOGLOBIN 2, NON-SYMBIOTIC HAEMOGLOBIN 2, NSHB2, haemoglobin 2	
☐ HB1 ID: 816103	hemoglobin 1 [*Arabidopsis thaliana* (thale cress)]	Chromosome 2, NC_003071.7 (6982531..6983777, complement)	AT2G16060, AHB1, ARATH GLB1, ATGLB1, CLASS I HEMOGLOBIN, F7H1.8, F7H1_8, GLB1, HEMOGLOBIN, NSHB1, hemoglobin 1	
☐ HBB ID: 3043	hemoglobin subunit beta [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (5225466..5227071, complement)	CD113t-C, beta-globin	141900
☐ HBG1 ID: 3047	hemoglobin subunit gamma 1 [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (5248272..5249857, complement)	HBG-T2, HBGA, HBGR, HSGGL1, PRO2979	142200
☐ HBG2 ID: 3048	hemoglobin subunit gamma 2 [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (5253191..5254781,	HBG-T1, TNCY	142250

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
Sus scrofa *(255)*
Homo sapiens *(194)*
Mus musculus *(85)*
Bos taurus *(55)*
Oncorhynchus kisutch *(50)*
All other taxa *(3674)*
More...

Find related data

Database: Select

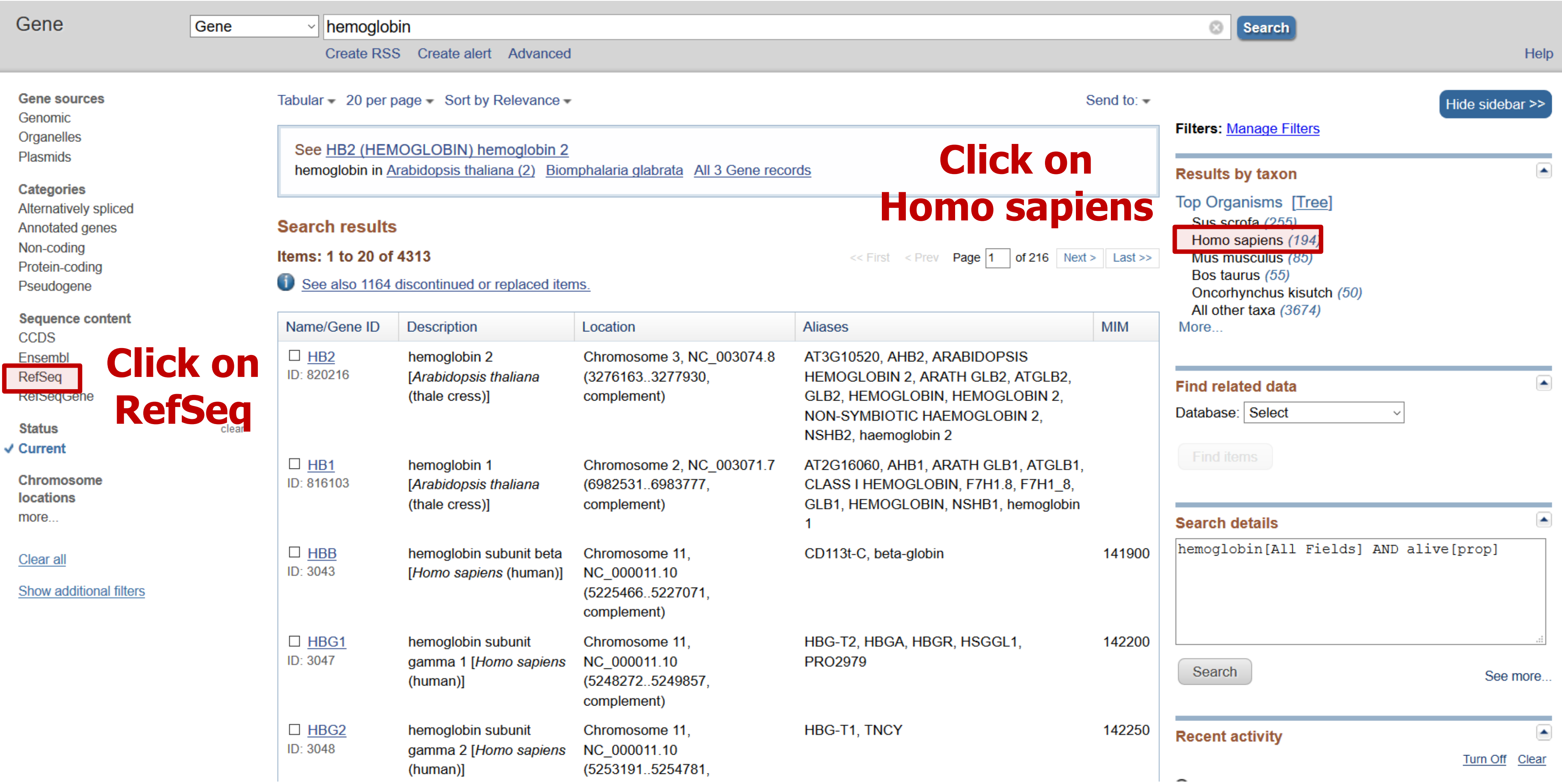
Search details

```
hemoglobin[All Fields] AND alive[prop]
```

Recent activity

Use filters to limit your search

Gene Database



Search: hemoglobin

See HB2 (HEMOGLOBIN) hemoglobin 2
hemoglobin in Arabidopsis thaliana (2) Biomphalaria glabrata All 3 Gene records

Search results

Items: 1 to 20 of 4313

See also 1164 discontinued or replaced items.

Name/Gene ID	Description	Location	Aliases	MIM
☐ HB2 ID: 820216	hemoglobin 2 [*Arabidopsis thaliana* (thale cress)]	Chromosome 3, NC_003074.8 (3276163..3277930, complement)	AT3G10520, AHB2, ARABIDOPSIS HEMOGLOBIN 2, ARATH GLB2, ATGLB2, GLB2, HEMOGLOBIN, HEMOGLOBIN 2, NON-SYMBIOTIC HAEMOGLOBIN 2, NSHB2, haemoglobin 2	
☐ HB1 ID: 816103	hemoglobin 1 [*Arabidopsis thaliana* (thale cress)]	Chromosome 2, NC_003071.7 (6982531..6983777, complement)	AT2G16060, AHB1, ARATH GLB1, ATGLB1, CLASS I HEMOGLOBIN, F7H1.8, F7H1_8, GLB1, HEMOGLOBIN, NSHB1, hemoglobin 1	
☐ HBB ID: 3043	hemoglobin subunit beta [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (5225466..5227071, complement)	CD113t-C, beta-globin	141900
☐ HBG1 ID: 3047	hemoglobin subunit gamma 1 [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (5248272..5249857, complement)	HBG-T2, HBGA, HBGR, HSGGL1, PRO2979	142200
☐ HBG2 ID: 3048	hemoglobin subunit gamma 2 [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (5253191..5254781,	HBG-T1, TNCY	142250

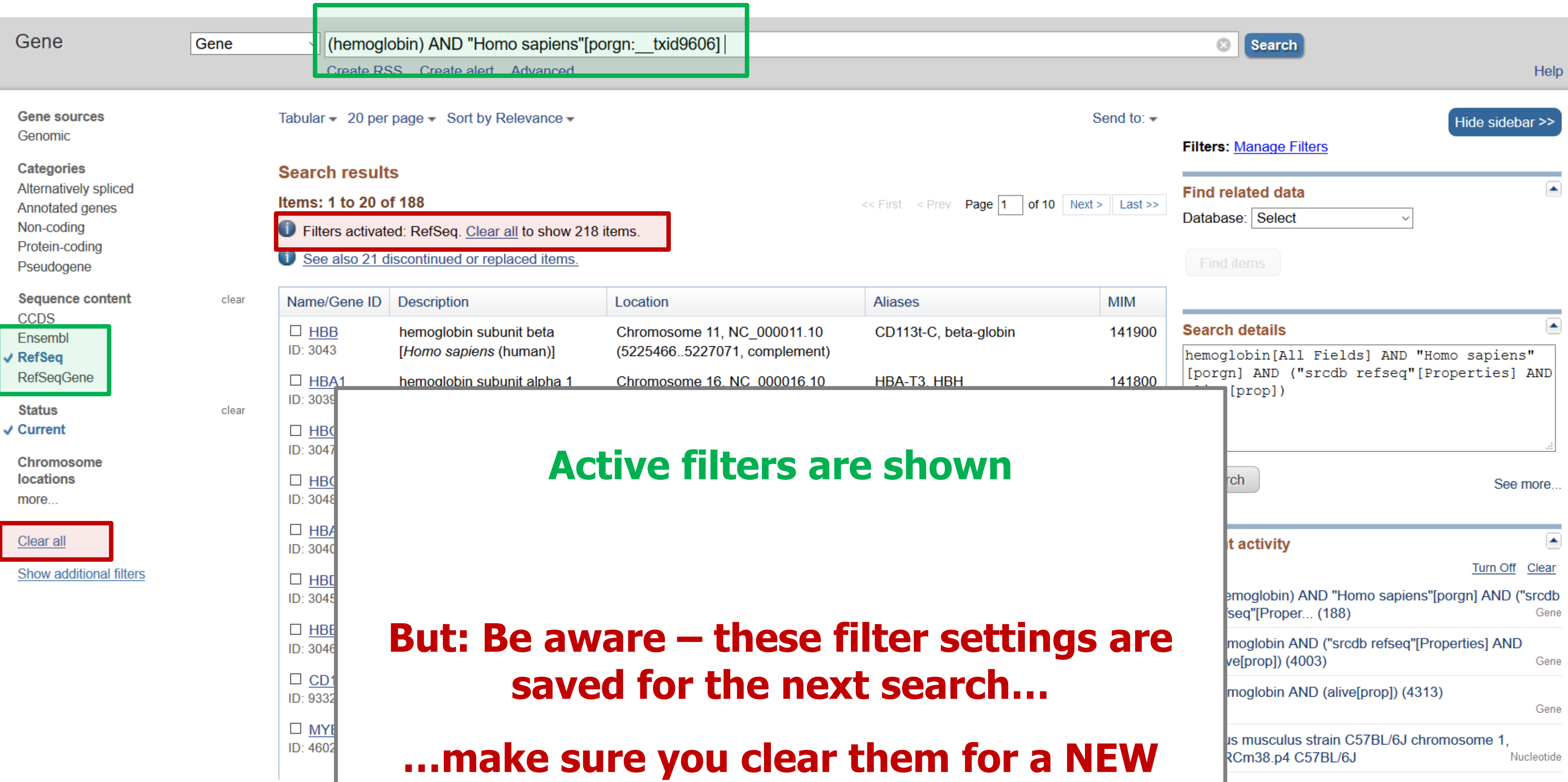
```
hemoglobin[All Fields] AND alive[prop]
```

Click on Homo sapiens

Click on RefSeq

Theresa Schwarz

Gene Database



Gene | Gene | (hemoglobin) AND "Homo sapiens"[porgn:__txid9606] | Search

Create RSS Create alert Advanced Help

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

Gene sources
Genomic

Categories
Alternatively spliced
Annotated genes
Non-coding
Protein-coding
Pseudogene

Sequence content clear
CCDS
Ensembl
✓ **RefSeq**
RefSeqGene

Status clear
✓ **Current**

Chromosome locations
more...

Clear all

Show additional filters

Search results

Items: 1 to 20 of 188

<< First < Prev Page 1 of 10 Next > Last >>

Filters activated: RefSeq. Clear all to show 218 items.

See also 21 discontinued or replaced items.

Name/Gene ID	Description	Location	Aliases	MIM
☐ HBB ID: 3043	hemoglobin subunit beta [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (5225466..5227071, complement)	CD113t-C, beta-globin	141900
☐ HBA1 ID: 3039	hemoglobin subunit alpha 1	Chromosome 16, NC_000016.10	HBA-T3, HBH	141800
☐ HBC ID: 3047				
☐ HBC ID: 3048				
☐ HBA ID: 3040				
☐ HBD ID: 3045				
☐ HBE ID: 3046				
☐ CD1 ID: 9332				
☐ MYE ID: 4602				

Filters: Manage Filters

Find related data
Database: Select
Find items

Search details

```
hemoglobin[All Fields] AND "Homo sapiens"
[porgn] AND ("srcdb refseq"[Properties] AND
        [prop])
```

rch See more...

t activity
Turn Off Clear

emoglobin) AND "Homo sapiens"[porgn] AND ("srcdb fseq"[Proper... (188) Gene

moglobin AND ("srcdb refseq"[Properties] AND ve[prop]) (4003) Gene

moglobin AND (alive[prop]) (4313) Gene

us musculus strain C57BL/6J chromosome 1, RCm38.p4 C57BL/6J Nucleotide

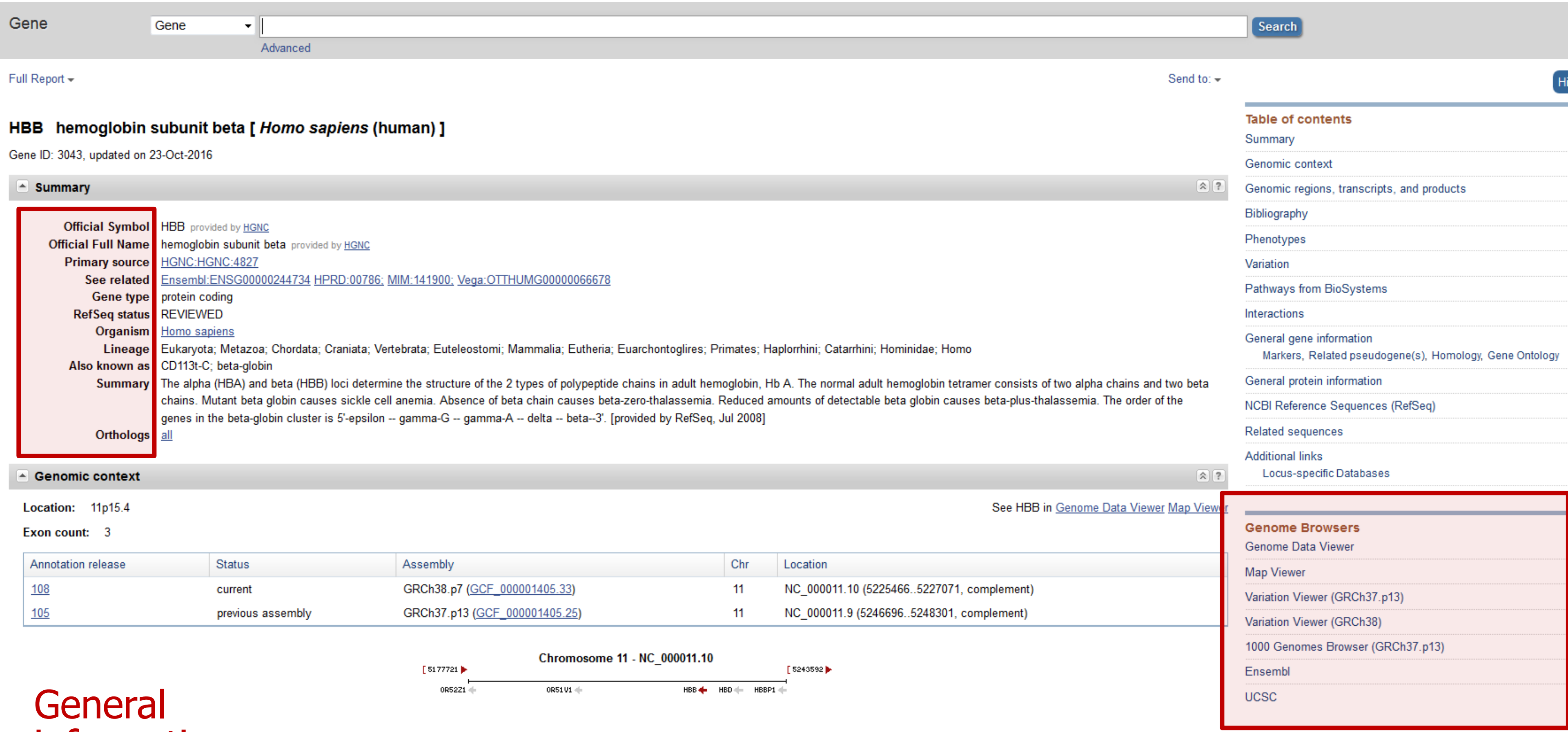
Active filters are shown

But: Be aware – these filter settings are saved for the next search…

…make sure you clear them for a NEW search

→ Click on first entry **HBB**

Gene Database



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

Gene ID: 3043, updated on 23-Oct-2016

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

Genomic context

Location: 11p15.4

Exon count: 3

See HBB in Genome Data Viewer Map Viewer

Annotation release	Status	Assembly	Chr	Location
108	current	GRCh38.p7 (GCF_000001405.33)	11	NC_000011.10 (5225466..5227071, complement)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	11	NC_000011.9 (5246696..5248301, complement)

Chromosome 11 - NC_000011.10

Table of contents

Summary

Genomic context

Genomic regions, transcripts, and products

Bibliography

Phenotypes

Variation

Pathways from BioSystems

Interactions

General gene information

Markers, Related pseudogene(s), Homology, Gene Ontology

General protein information

NCBI Reference Sequences (RefSeq)

Related sequences

Additional links

Locus-specific Databases

Genome Browsers

Genome Data Viewer

Map Viewer

Variation Viewer (GRCh37.p13)

Variation Viewer (GRCh38)

1000 Genomes Browser (GRCh37.p13)

Ensembl

UCSC

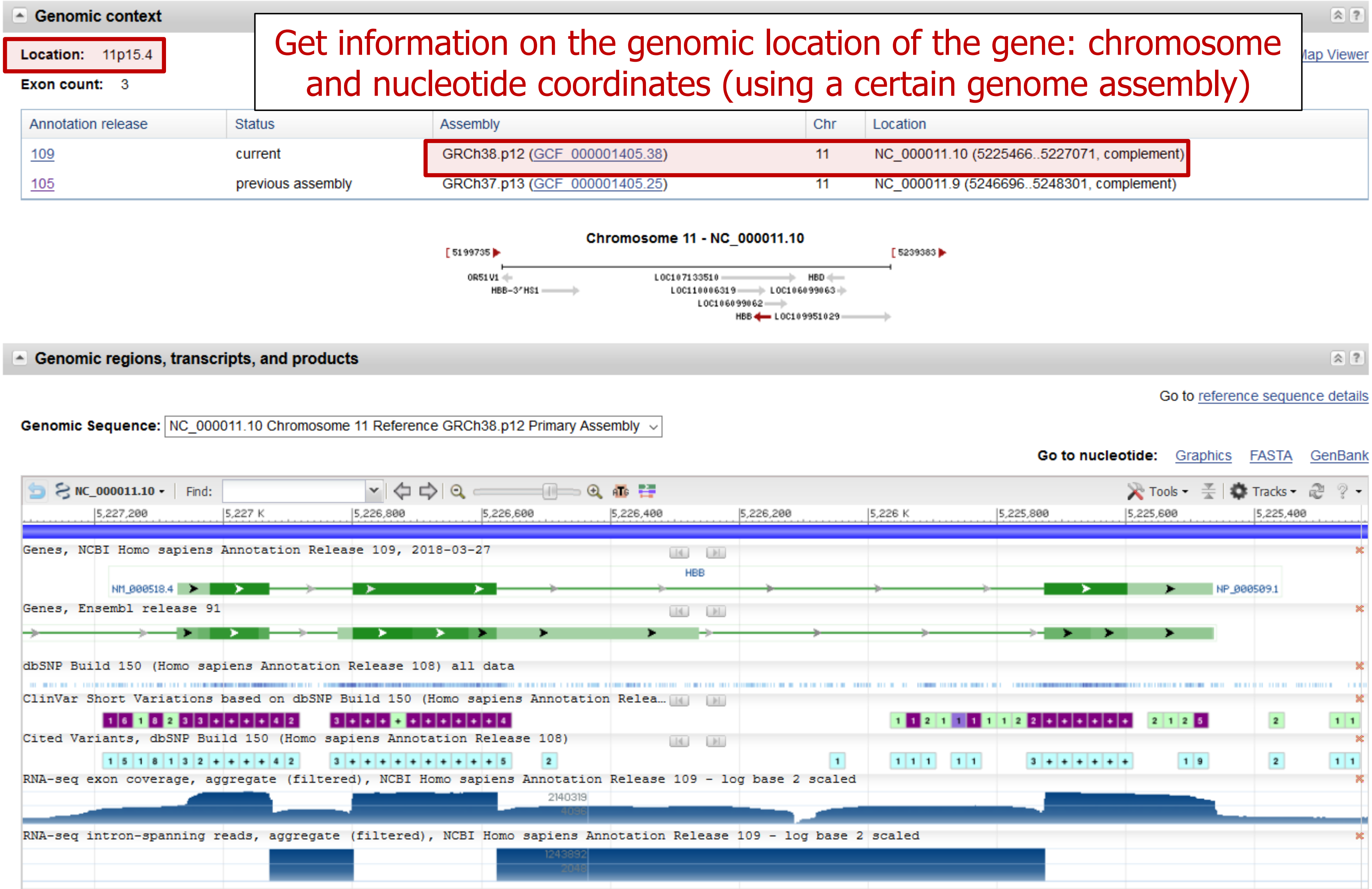
General information about this gene

At the right-hand side you will find links to other HBB database entries

Theresa Schwarz

Gene Database

→ Scroll down to **Genomic context**



Get information on the genomic location of the gene: chromosome and nucleotide coordinates (using a certain genome assembly)

Genomic context

Location: 11p15.4

Exon count: 3

Annotation release	Status	Assembly	Chr	Location
109	current	GRCh38.p12 (GCF_000001405.38)	11	NC_000011.10 (5225466..5227071, complement)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	11	NC_000011.9 (5246696..5248301, complement)

Chromosome 11 - NC_000011.10

Genomic regions, transcripts, and products

Go to reference sequence details

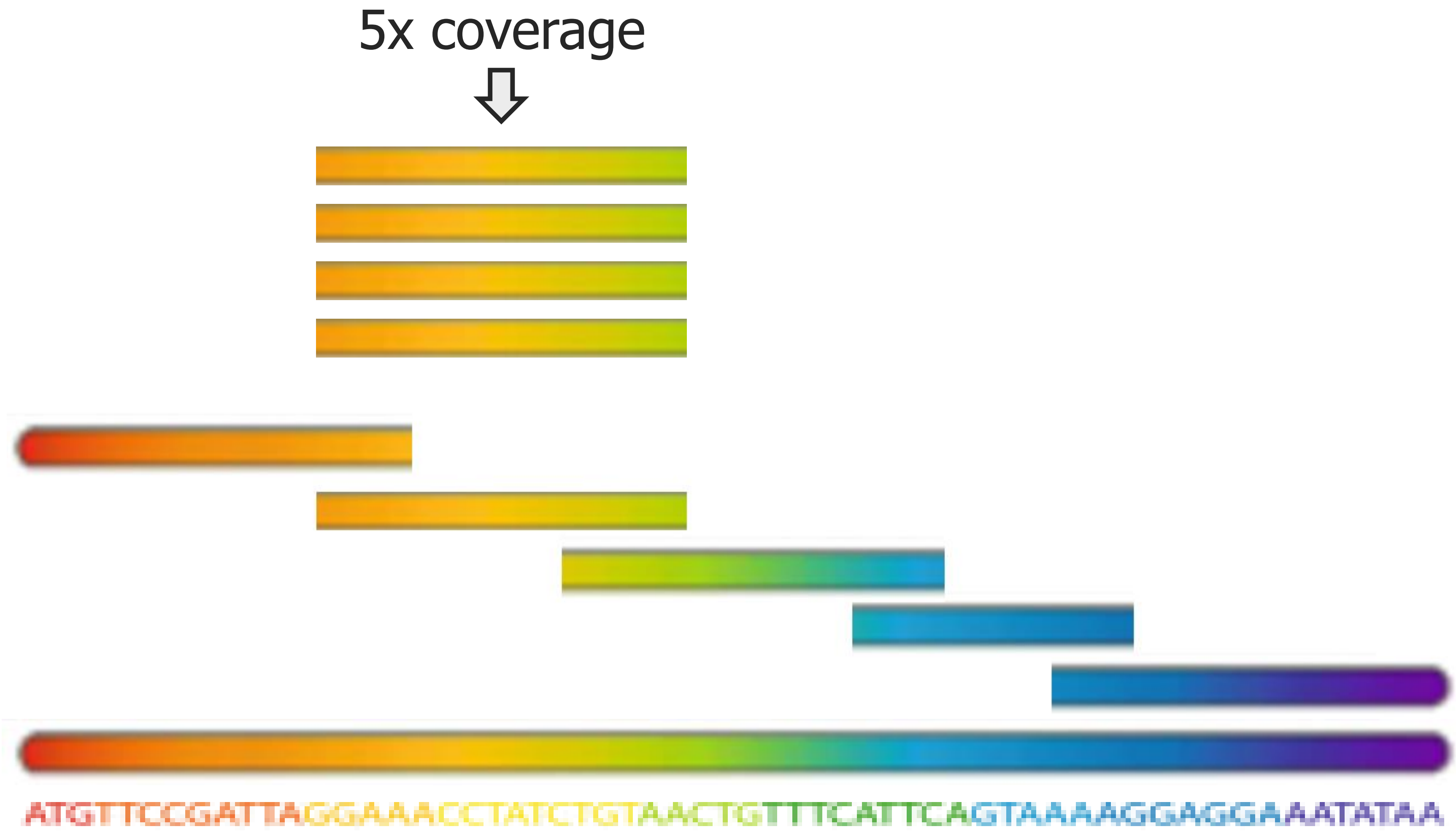
Genomic Sequence: NC_000011.10 Chromosome 11 Reference GRCh38.p12 Primary Assembly

Go to nucleotide: Graphics FASTA GenBank

What is a genome assembly?

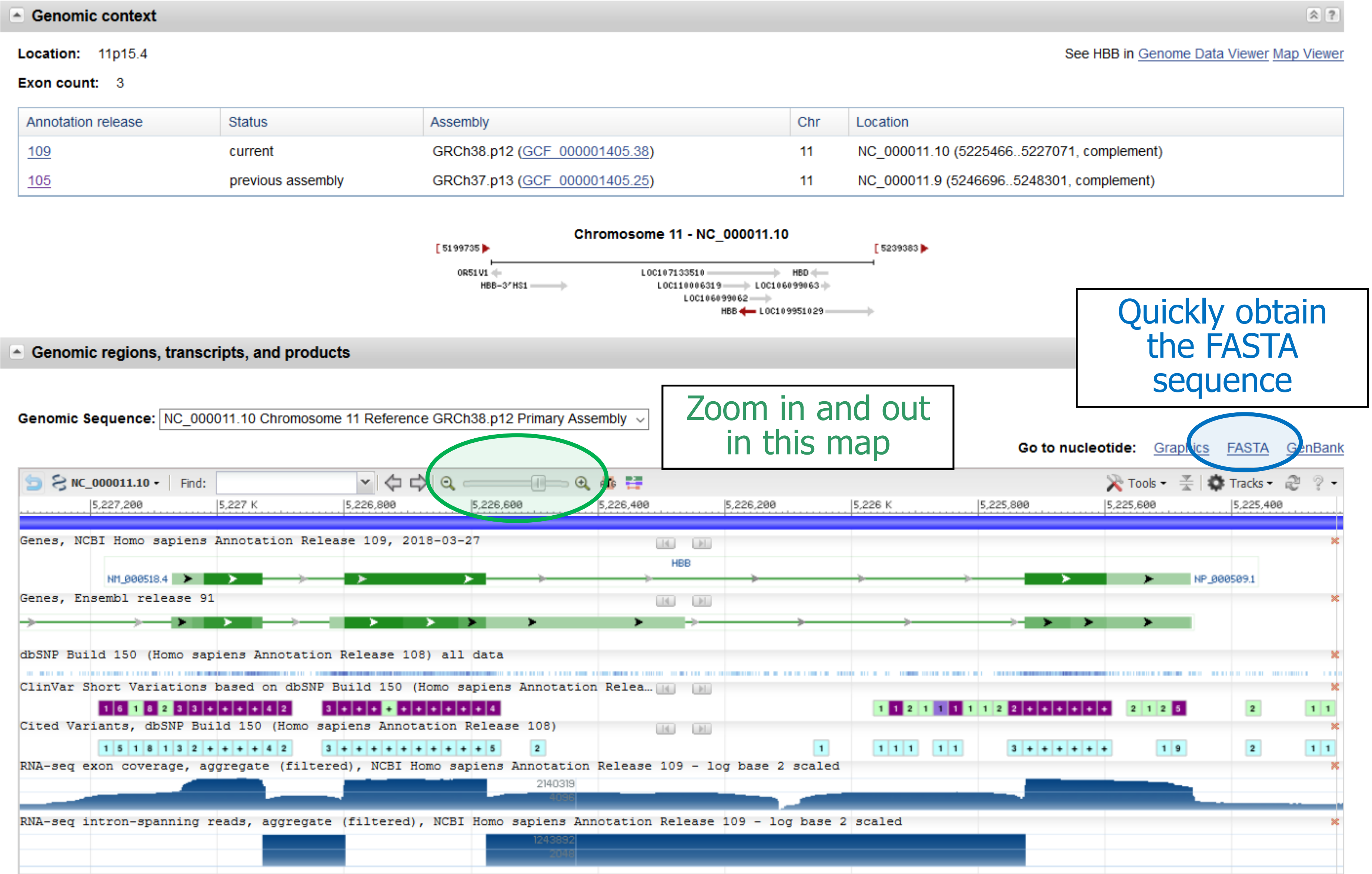
- Assembly of short DNA sequences to create a representation of the chromosomes of a certain species
- Current version for the human genome: GRCh38.p12
  = Genome Reference Consortium Human Build 38patch release 12
- Nucleotide positions can change between different genome assemblies

What is a genome assembly?



Gene Database

→ Click on **FASTA**



Genomic context

Location: 11p15.4

See HBB in Genome Data Viewer Map Viewer

Exon count: 3

Annotation release	Status	Assembly	Chr	Location
109	current	GRCh38.p12 (GCF_000001405.38)	11	NC_000011.10 (5225466..5227071, complement)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	11	NC_000011.9 (5246696..5248301, complement)

Genomic regions, transcripts, and products

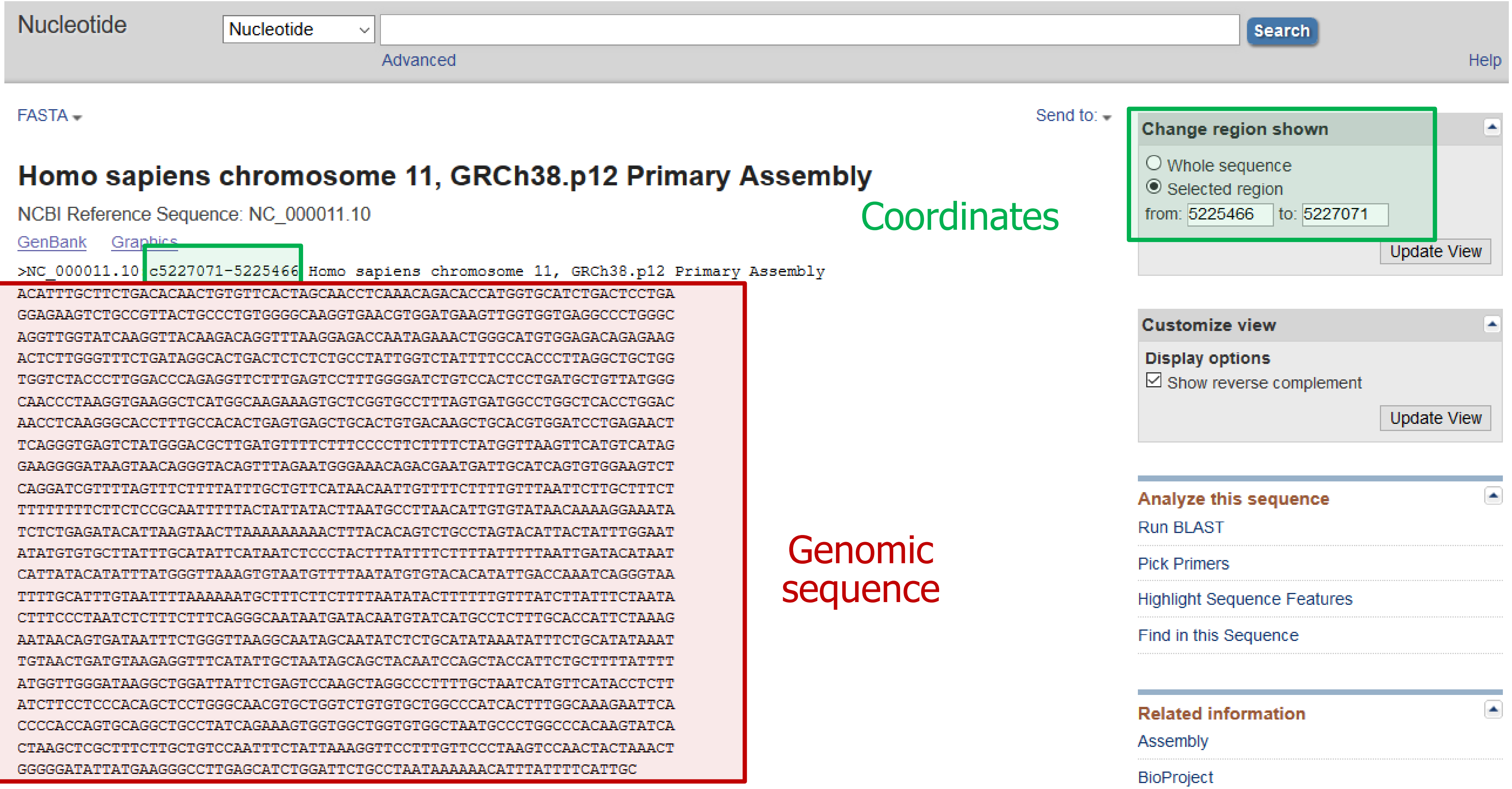
Genomic Sequence: NC_000011.10 Chromosome 11 Reference GRCh38.p12 Primary Assembly

Go to nucleotide: Graphics FASTA GenBank

Zoom in and out in this map

Quickly obtain the FASTA sequence

Gene Database



Homo sapiens chromosome 11, GRCh38.p12 Primary Assembly

NCBI Reference Sequence: NC_000011.10

```
>NC_000011.10 c5227071-5225466 Homo sapiens chromosome 11, GRCh38.p12 Primary Assembly
ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGA
GGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGC
AGGTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCATGTGGAGACAGAGAAG
ACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGG
TGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGG
CAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGAC
AACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACT
TCAGGGTGAGTCTATGGGACGCTTGATGTTTTCTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAG
GAAGGGGATAAGTAACAGGGTACAGTTTAGAATGGGAAACAGACGAATGATTGCATCAGTGTGGAAGTCT
CAGGATCGTTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTTGTTTAATTCTTGCTTTCT
TTTTTTTTCTTCTCCGCAATTTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAAAGGAAATA
TCTCTGAGATACATTAAGTAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACTATTTGGAAT
ATATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTTATTTTCTTTTATTTTTAATTGATACATAAT
CATTATACATATTTATGGGTTAAAGTGTAATGTTTTAATATGTGTACACATATTGACCAAATCAGGGTAA
TTTTGCATTTGTAATTTTAAAAAATGCTTTCTTCTTTTAATATACTTTTTTGTTTATCTTATTTCTAATA
CTTTCCCTAATCTCTTTCTTTCAGGGCAATAATGATACAATGTATCATGCCTCTTTGCACCATTCTAAAG
AATAACAGTGATAATTTCTGGGTTAAGGCAATAGCAATATCTCTGCATATAAATATTTCTGCATATAAAT
TGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGCTACAATCCAGCTACCATTCTGCTTTTATTTT
ATGGTTGGGATAAGGCTGGATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCATACCTCTT
ATCTTCCTCCCACAGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCA
CCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCA
CTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACT
GGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGC
```

Coordinates

Change region shown

○ Whole sequence
● Selected region
from: 5225466 to: 5227071

Customize view

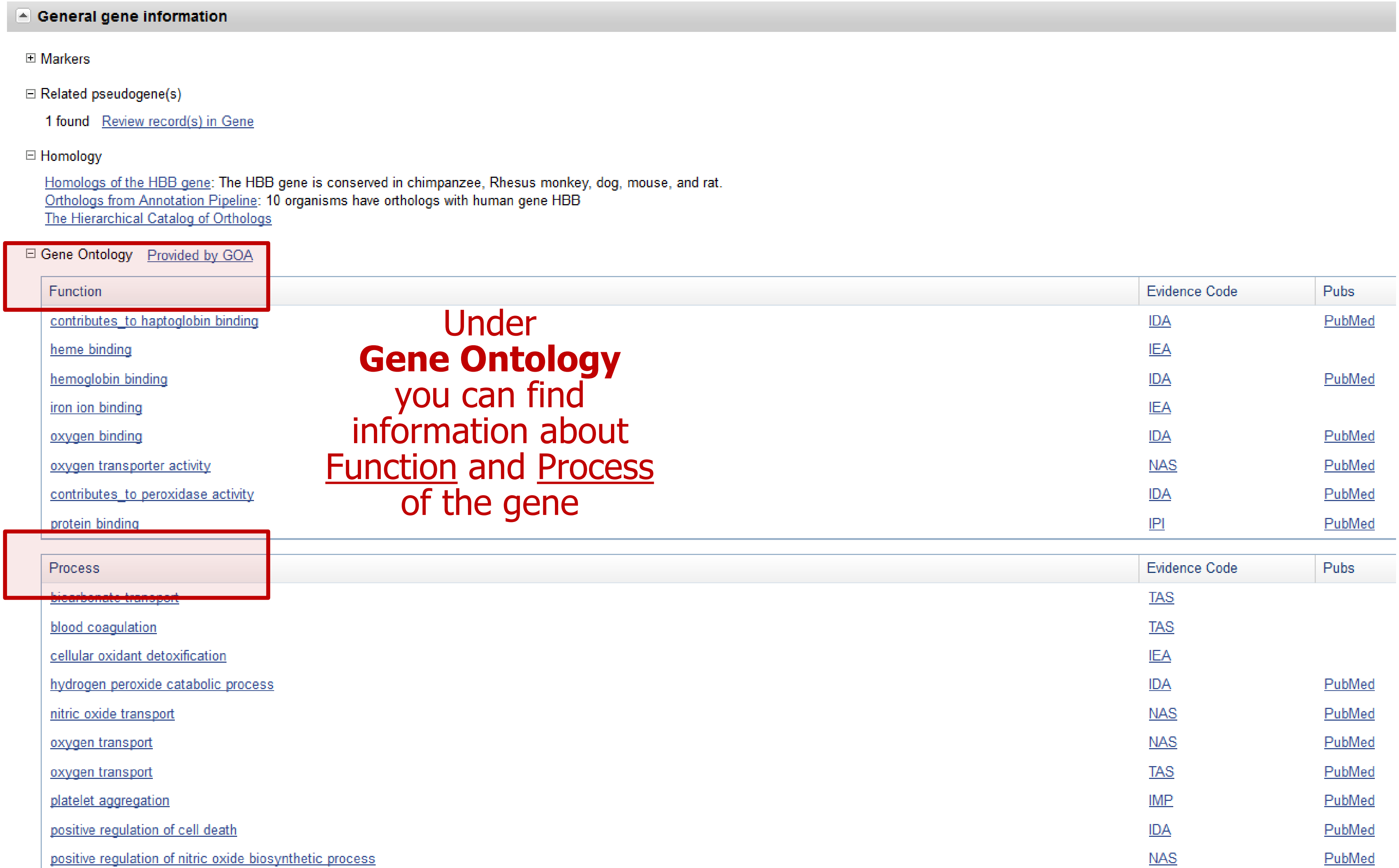
Display options

☑ Show reverse complement

Genomic sequence

Gene Database

→ Go back and scroll down to **General gene information**



General gene information

⊞ Markers

⊟ Related pseudogene(s)

1 found Review record(s) in Gene

⊟ Homology

Homologs of the HBB gene: The HBB gene is conserved in chimpanzee, Rhesus monkey, dog, mouse, and rat.
Orthologs from Annotation Pipeline: 10 organisms have orthologs with human gene HBB
The Hierarchical Catalog of Orthologs

⊟ Gene Ontology Provided by GOA

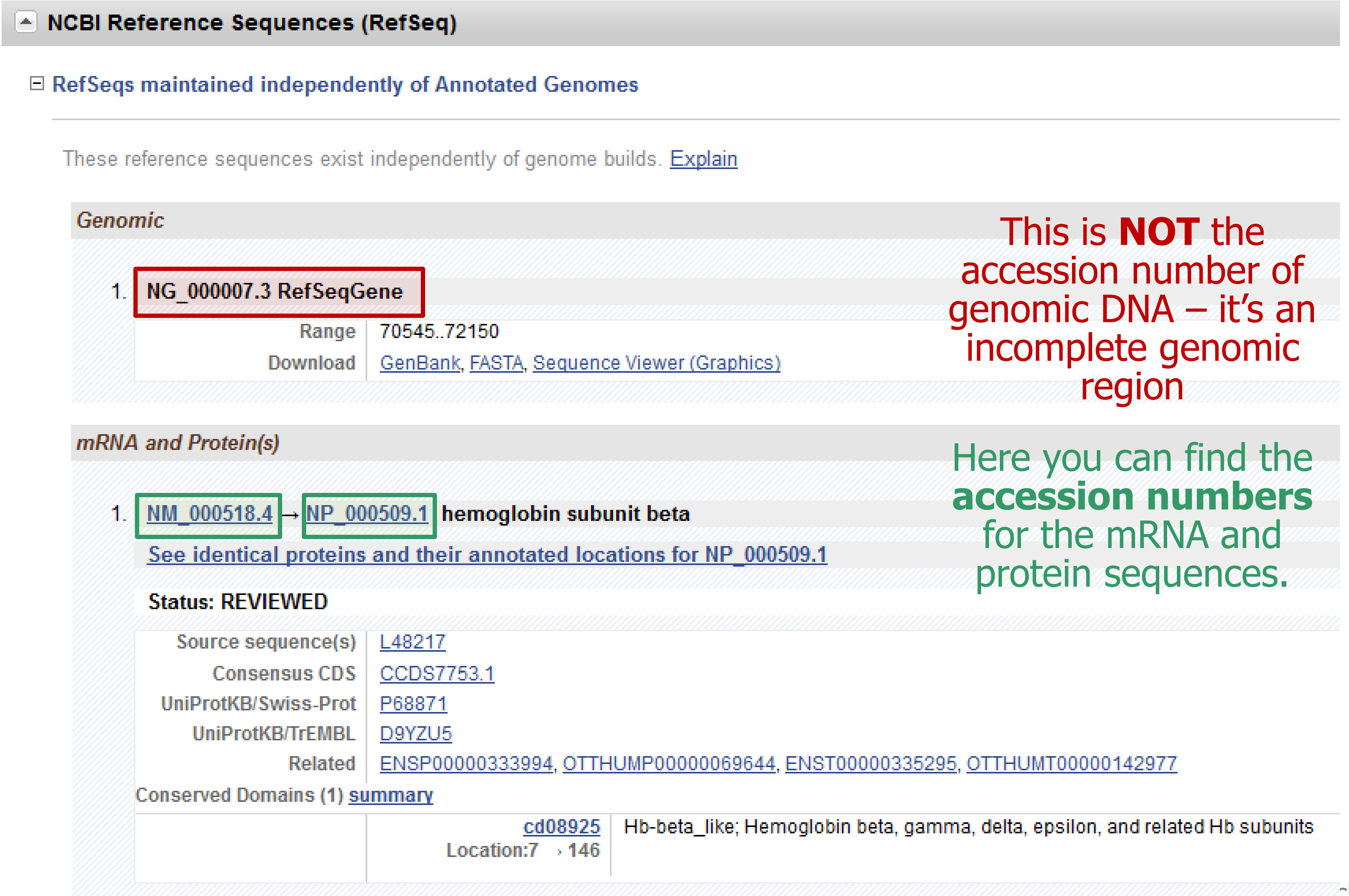
Function	Evidence Code	Pubs
contributes_to haptoglobin binding	IDA	PubMed
heme binding	IEA	
hemoglobin binding	IDA	PubMed
iron ion binding	IEA	
oxygen binding	IDA	PubMed
oxygen transporter activity	NAS	PubMed
contributes_to peroxidase activity	IDA	PubMed
protein binding	IPI	PubMed

Process	Evidence Code	Pubs
bicarbonate transport	TAS	
blood coagulation	TAS	
cellular oxidant detoxification	IEA	
hydrogen peroxide catabolic process	IDA	PubMed
nitric oxide transport	NAS	PubMed
oxygen transport	NAS	PubMed
oxygen transport	TAS	PubMed
platelet aggregation	IMP	PubMed
positive regulation of cell death	IDA	PubMed
positive regulation of nitric oxide biosynthetic process	NAS	PubMed

Under
Gene Ontology
you can find
information about
Function and Process
of the gene

Gene Database

→ Scroll down to **NCBI Reference Sequences (RefSeq)**



NCBI Reference Sequences (RefSeq)

RefSeqs maintained independently of Annotated Genomes

These reference sequences exist independently of genome builds. Explain

Genomic

1. NG_000007.3 RefSeqGene

Range	70545..72150
Download	GenBank, FASTA, Sequence Viewer (Graphics)

This is **NOT** the accession number of genomic DNA – it's an incomplete genomic region

mRNA and Protein(s)

1. NM_000518.4 → NP_000509.1 **hemoglobin subunit beta**

See identical proteins and their annotated locations for NP_000509.1

Status: REVIEWED

Source sequence(s)	L48217
Consensus CDS	CCDS7753.1
UniProtKB/Swiss-Prot	P68871
UniProtKB/TrEMBL	D9YZU5
Related	ENSP00000333994, OTTHUMP00000069644, ENST00000335295, OTTHUMT00000142977

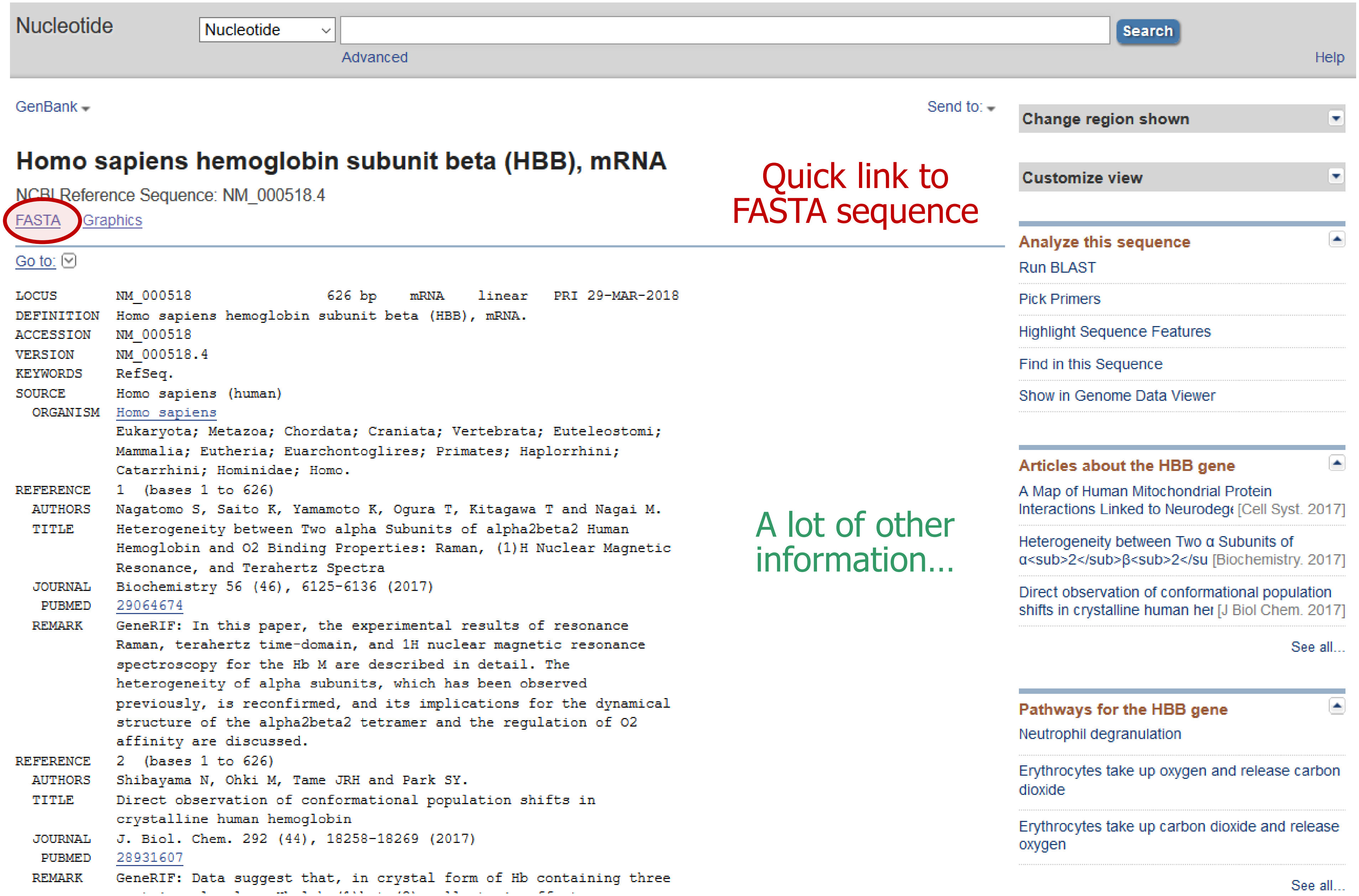
Conserved Domains (1) summary

cd08925 Location:7 › 146	Hb-beta_like; Hemoglobin beta, gamma, delta, epsilon, and related Hb subunits

Here you can find the **accession numbers** for the mRNA and protein sequences.

→ Click on the **mRNA** accession number

Gene Database



Homo sapiens hemoglobin subunit beta (HBB), mRNA

NCBI Reference Sequence: NM_000518.4

```
LOCUS       NM_000518                626 bp    mRNA    linear   PRI 29-MAR-2018
DEFINITION  Homo sapiens hemoglobin subunit beta (HBB), mRNA.
ACCESSION   NM_000518
VERSION     NM_000518.4
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 626)
  AUTHORS   Nagatomo S, Saito K, Yamamoto K, Ogura T, Kitagawa T and Nagai M.
  TITLE     Heterogeneity between Two alpha Subunits of alpha2beta2 Human
            Hemoglobin and O2 Binding Properties: Raman, (1)H Nuclear Magnetic
            Resonance, and Terahertz Spectra
  JOURNAL   Biochemistry 56 (46), 6125-6136 (2017)
   PUBMED   29064674
  REMARK    GeneRIF: In this paper, the experimental results of resonance
            Raman, terahertz time-domain, and 1H nuclear magnetic resonance
            spectroscopy for the Hb M are described in detail. The
            heterogeneity of alpha subunits, which has been observed
            previously, is reconfirmed, and its implications for the dynamical
            structure of the alpha2beta2 tetramer and the regulation of O2
            affinity are discussed.
REFERENCE   2  (bases 1 to 626)
  AUTHORS   Shibayama N, Ohki M, Tame JRH and Park SY.
  TITLE     Direct observation of conformational population shifts in
            crystalline human hemoglobin
  JOURNAL   J. Biol. Chem. 292 (44), 18258-18269 (2017)
   PUBMED   28931607
  REMARK    GeneRIF: Data suggest that, in crystal form of Hb containing three
```

Analyze this sequence

- Run BLAST
- Pick Primers
- Highlight Sequence Features
- Find in this Sequence
- Show in Genome Data Viewer

Articles about the HBB gene

- A Map of Human Mitochondrial Protein Interactions Linked to Neurodege [Cell Syst. 2017]
- Heterogeneity between Two α Subunits of α₂β₂</su [Biochemistry. 2017]
- Direct observation of conformational population shifts in crystalline human he [J Biol Chem. 2017]

See all...

Pathways for the HBB gene

- Neutrophil degranulation
- Erythrocytes take up oxygen and release carbon dioxide
- Erythrocytes take up carbon dioxide and release oxygen

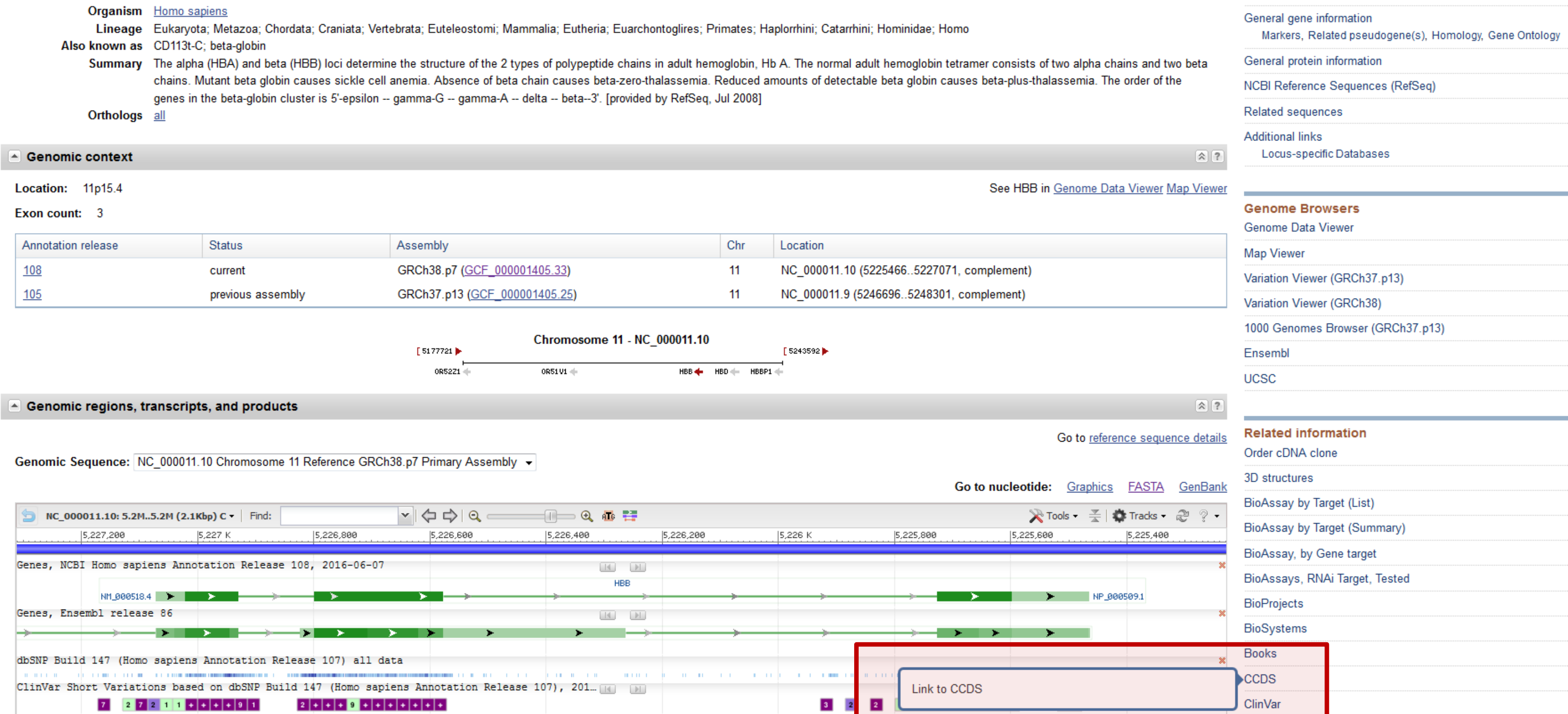
See all...

Quick link to FASTA sequence

A lot of other information...

Consensus coding sequence (CCDS)

Go back and use the link at the right-hand side to get to the CCDS Database!



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

Genomic context

Location: 11p15.4

Exon count: 3

See HBB in Genome Data Viewer Map Viewer

Annotation release	Status	Assembly	Chr	Location
108	current	GRCh38.p7 (GCF_000001405.33)	11	NC_000011.10 (5225466..5227071, complement)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	11	NC_000011.9 (5246696..5248301, complement)

Chromosome 11 - NC_000011.10

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

General gene information
Markers, Related pseudogene(s), Homology, Gene Ontology
General protein information
NCBI Reference Sequences (RefSeq)
Related sequences
Additional links
Locus-specific Databases

Genome Browsers
Genome Data Viewer
Map Viewer
Variation Viewer (GRCh37.p13)
Variation Viewer (GRCh38)
1000 Genomes Browser (GRCh37.p13)
Ensembl
UCSC

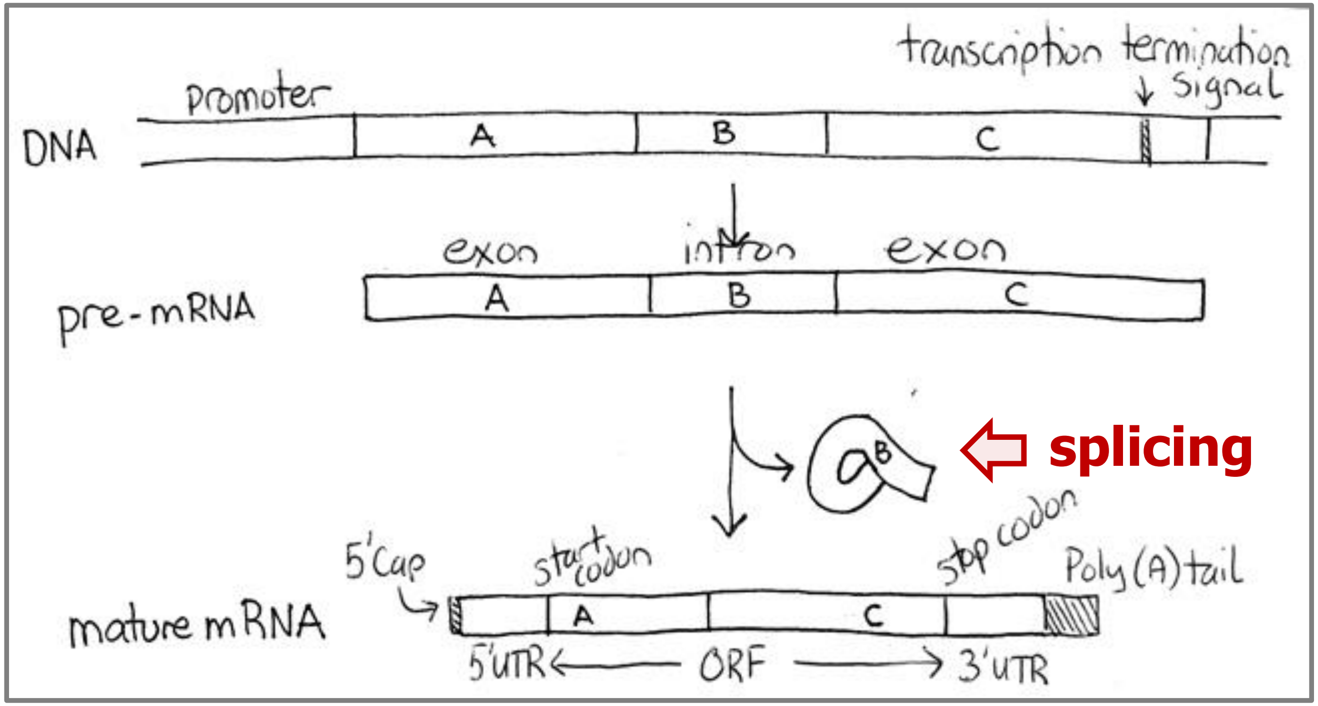
Related information
Order cDNA clone
3D structures
BioAssay by Target (List)
BioAssay by Target (Summary)
BioAssay, by Gene target
BioAssays, RNAi Target, Tested
BioProjects
BioSystems
Books
CCDS
ClinVar

Link to CCDS

Consensus coding sequence (CCDS)

- Allows you to analyze the **coding region** of genes = ORFs
- ORF = open-reading-frame

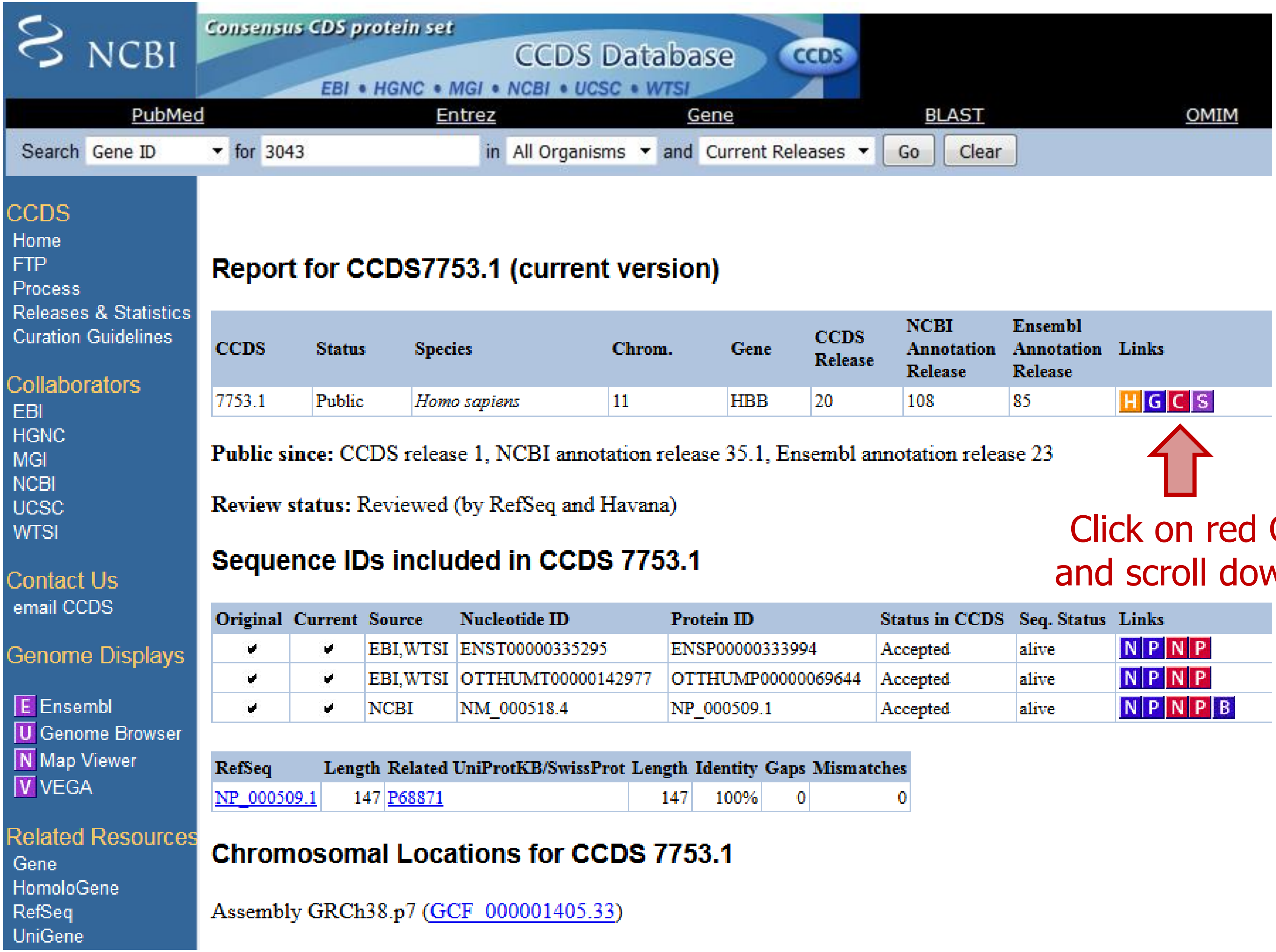
genomic DNA vs. mRNA sequence



⇦ **splicing**

- An **mRNA sequence** contains the
  - 5′ UTR (untranslated region)
  - ORF (open reading frame) = **coding sequence** → CCDS database
  - 3` UTR (untranslated region)
  - Poly-A tail

Consensus coding sequence (CCDS)



Report for CCDS7753.1 (current version)

CCDS	Status	Species	Chrom.	Gene	CCDS Release	NCBI Annotation Release	Ensembl Annotation Release	Links
7753.1	Public	*Homo sapiens*	11	HBB	20	108	85	H G C S

Public since: CCDS release 1, NCBI annotation release 35.1, Ensembl annotation release 23

Review status: Reviewed (by RefSeq and Havana)

Sequence IDs included in CCDS 7753.1

Original	Current	Source	Nucleotide ID	Protein ID	Status in CCDS	Seq. Status	Links
✓	✓	EBI,WTSI	ENST00000335295	ENSP00000333994	Accepted	alive	N P N P
✓	✓	EBI,WTSI	OTTHUMT00000142977	OTTHUMP00000069644	Accepted	alive	N P N P
✓	✓	NCBI	NM_000518.4	NP_000509.1	Accepted	alive	N P N P B

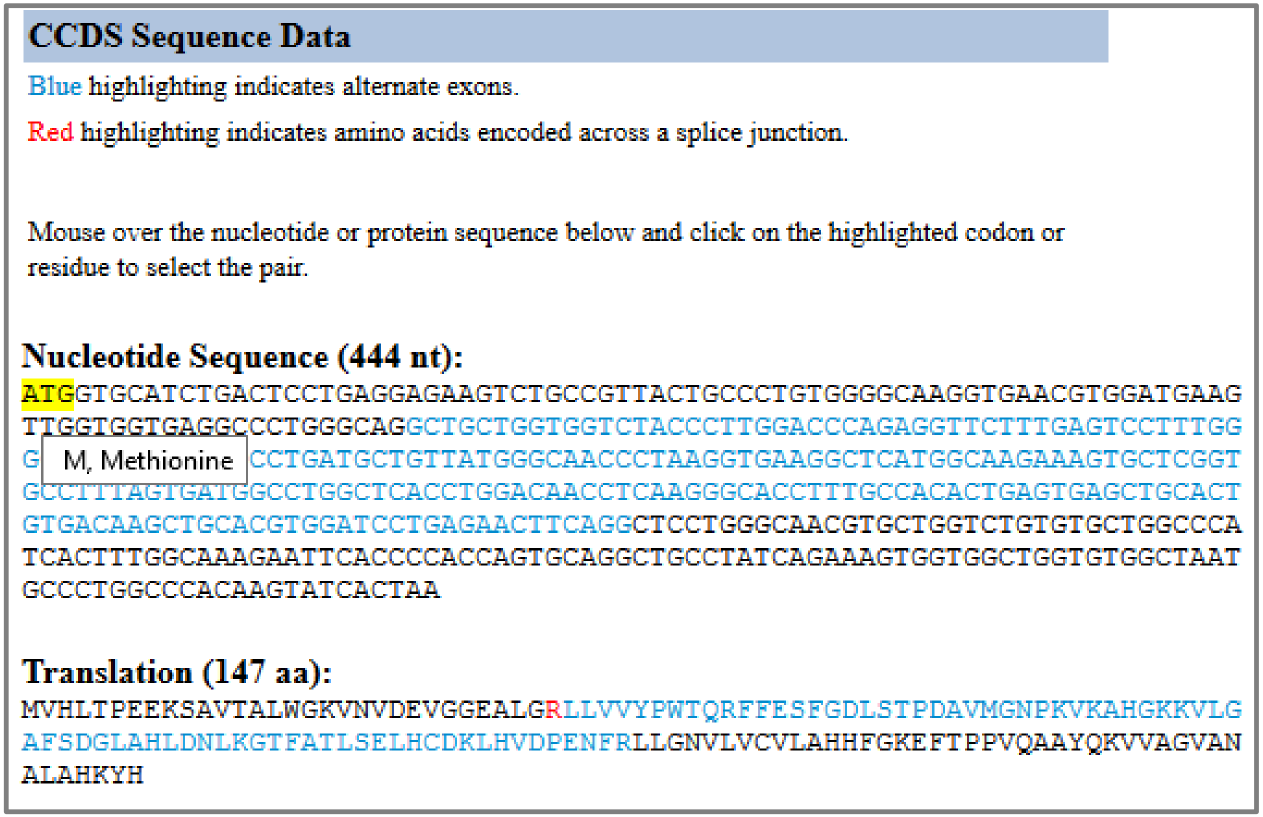
RefSeq	Length	Related UniProtKB/SwissProt	Length	Identity	Gaps	Mismatches
NP_000509.1	147	P68871	147	100%	0	0

Chromosomal Locations for CCDS 7753.1

Assembly GRCh38.p7 (GCF_000001405.33)

Click on red C and scroll down

Consensus coding sequence (CCDS)



CCDS Sequence Data

Blue highlighting indicates alternate exons.

Red highlighting indicates amino acids encoded across a splice junction.

Mouse over the nucleotide or protein sequence below and click on the highlighted codon or residue to select the pair.

Nucleotide Sequence (444 nt):

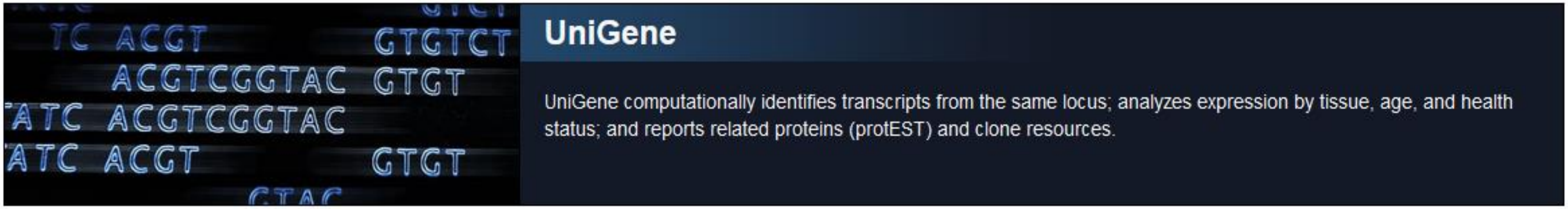
```
ATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG
TTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGG
G M, Methionine CCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCA
TCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
GCCCTGGCCCACAAGTATCACTAA
```

Translation (147 aa):

```
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

- Allows you to analyze the coding region of genes = ORFs
- Shows base triplets in nucleotide sequence and respective amino acid sequence
- Alternating exons are shown in black or blue

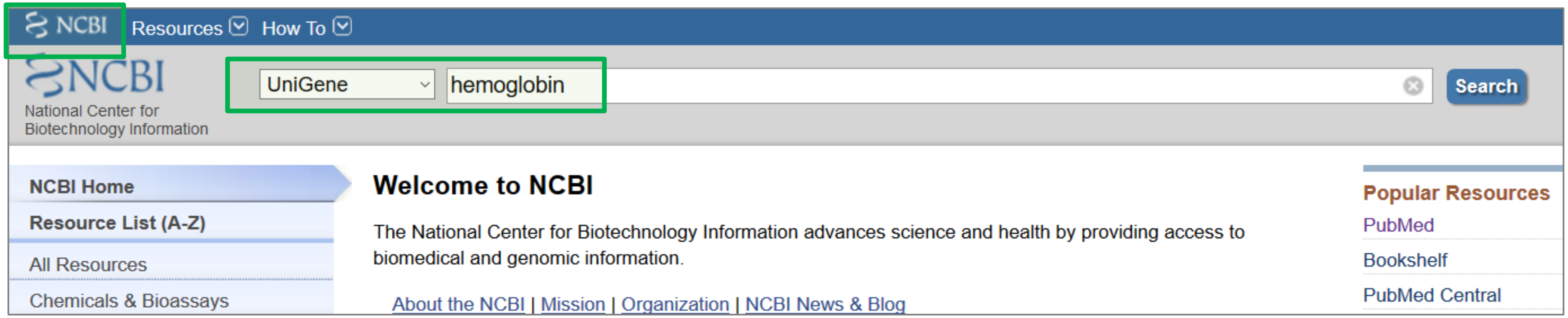
UniGene Database (expression of genes)



UniGene

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

- UniGene is a database with information on
  - **where in the body**
  - **when in development**
  - **how abundantly** a gene is expressed.



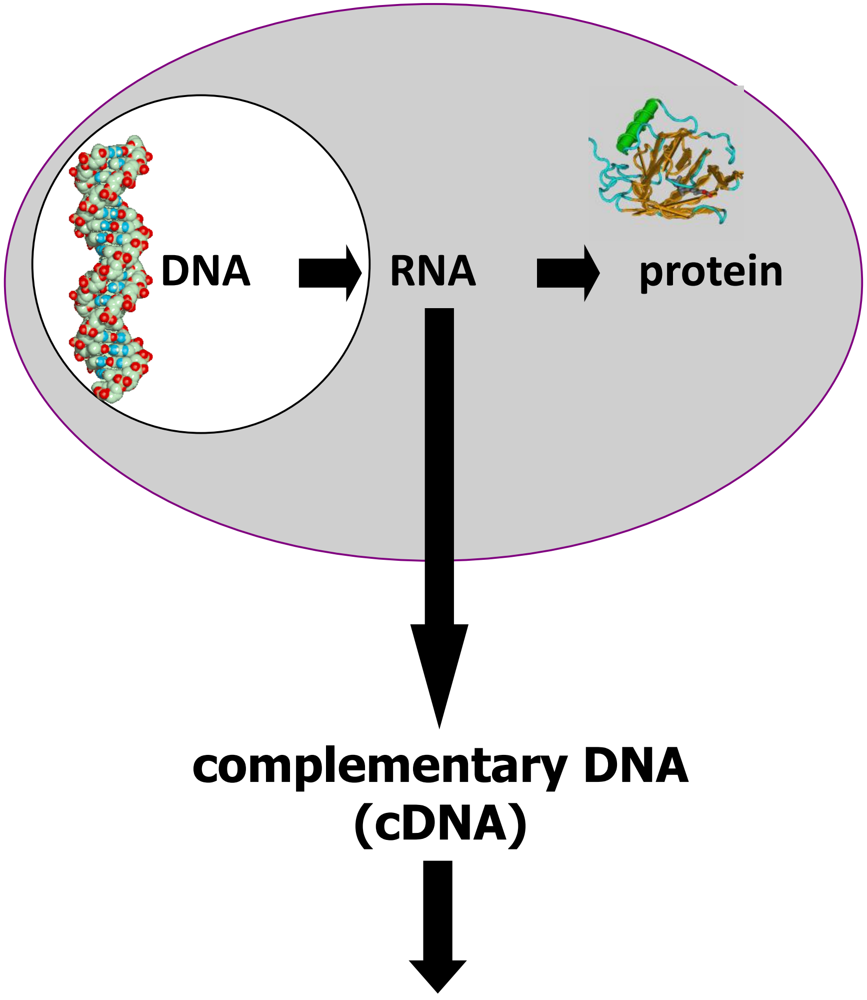
UniGene | hemoglobin

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

UniGene and ESTs

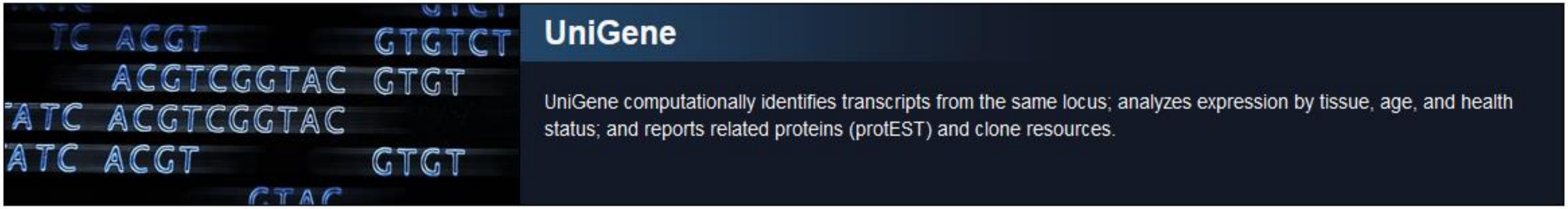


complementary DNA (cDNA)

ESTs

- ESTs = Expressed sequence tags
- ESTs are short DNA sequences, produced by reverse transcription of mRNA
- Represent small parts of genes that are expressed in certain tissues

UniGene Database (expression of genes)



UniGene

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

- Search for **hemoglobin** in the **UniGene** database and select ***Hemoglobin, beta***

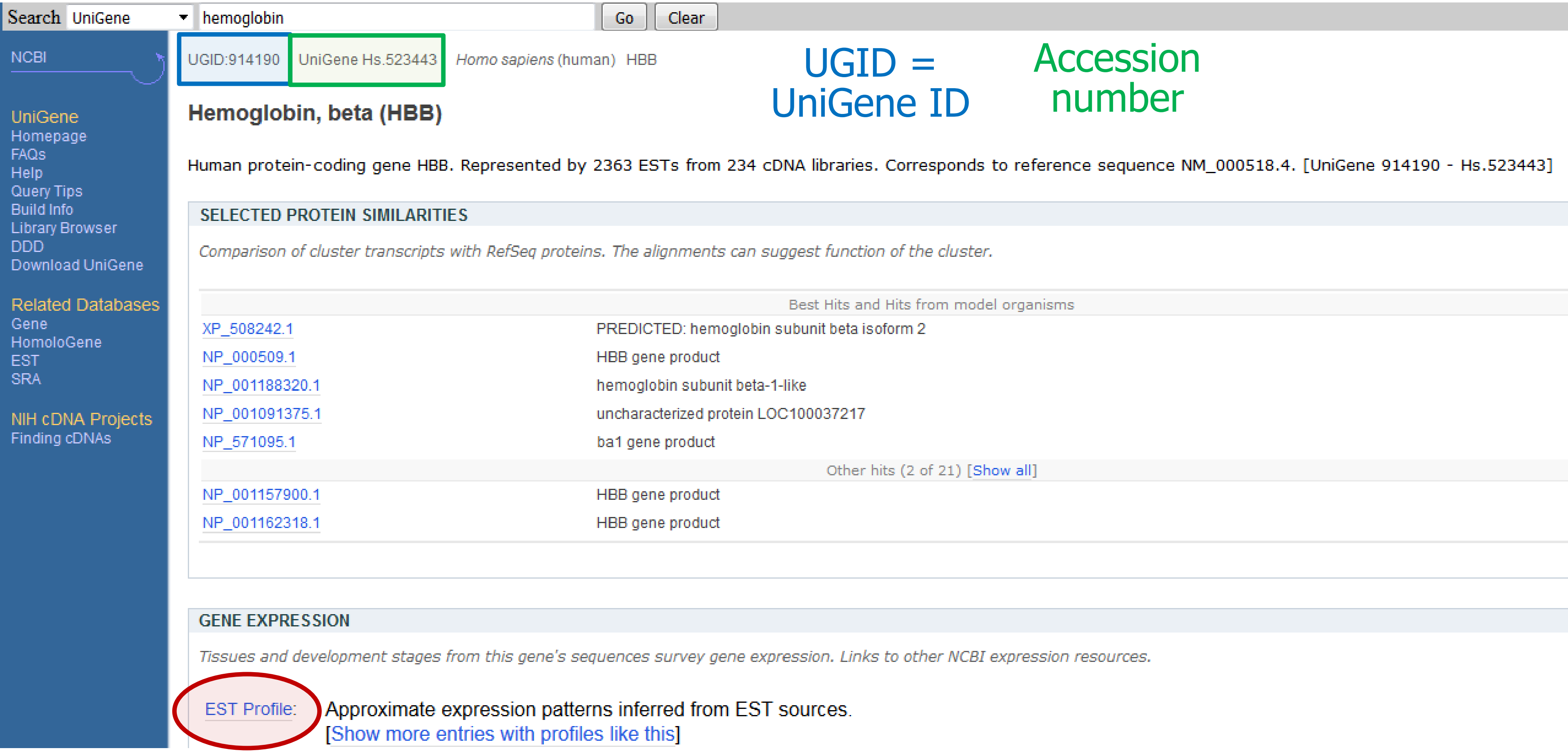
hemoglobin

Search results

Items: 1 to 20 of 361

1. ☐ Hemoglobin, beta
   HBB, Homo sapiens
   Hs.523443: 2383 sequences.
   Order cDNA clone
2. ☐ Hemoglobin, alpha 2
   HBA2, Homo sapiens
   Hs.654744: 4125 sequences.
   Order cDNA clone
3. ☐ Hemoglobin, alpha 1
   HBA1, Homo sapiens
   Hs.449630: 1311 sequences.
   Order cDNA clone

Expressed sequence tags (ESTs)



UGID:914190 UniGene Hs.523443 *Homo sapiens* (human) HBB

UGID = UniGene ID

Accession number

Hemoglobin, beta (HBB)

Human protein-coding gene HBB. Represented by 2363 ESTs from 234 cDNA libraries. Corresponds to reference sequence NM_000518.4. [UniGene 914190 - Hs.523443]

SELECTED PROTEIN SIMILARITIES

Comparison of cluster transcripts with RefSeq proteins. The alignments can suggest function of the cluster.

Best Hits and Hits from model organisms	
XP_508242.1	PREDICTED: hemoglobin subunit beta isoform 2
NP_000509.1	HBB gene product
NP_001188320.1	hemoglobin subunit beta-1-like
NP_001091375.1	uncharacterized protein LOC100037217
NP_571095.1	ba1 gene product
Other hits (2 of 21) [Show all]	
NP_001157900.1	HBB gene product
NP_001162318.1	HBB gene product

GENE EXPRESSION

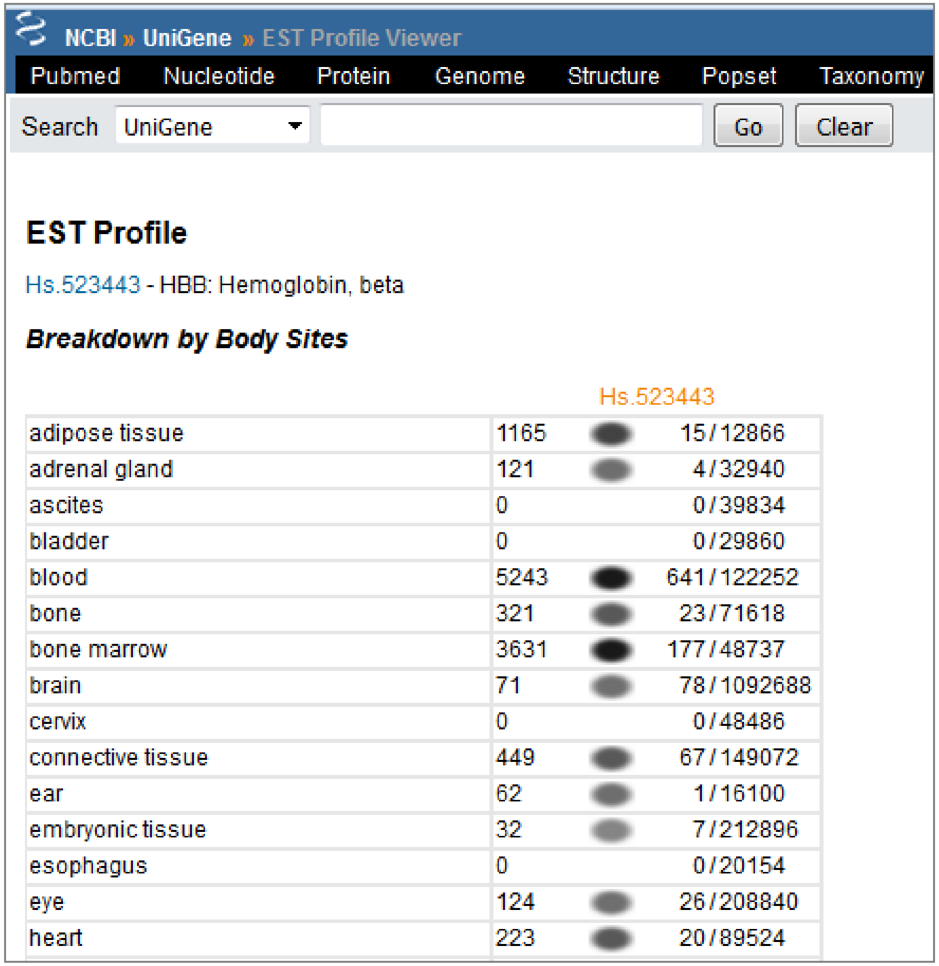
Tissues and development stages from this gene's sequences survey gene expression. Links to other NCBI expression resources.

EST Profile: Approximate expression patterns inferred from EST sources.
[Show more entries with profiles like this]

EST profile

→ Click on **EST profile**

ESTs profile

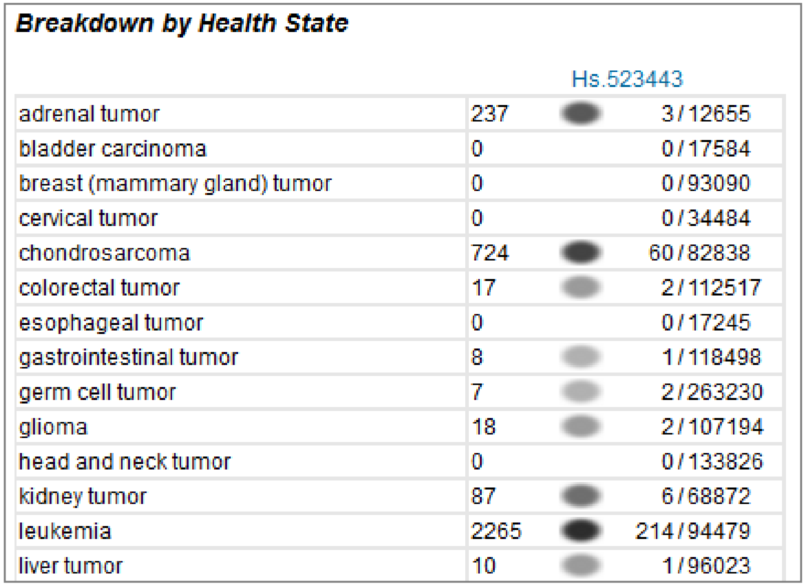


EST Profile

Hs.523443 - HBB: Hemoglobin, beta

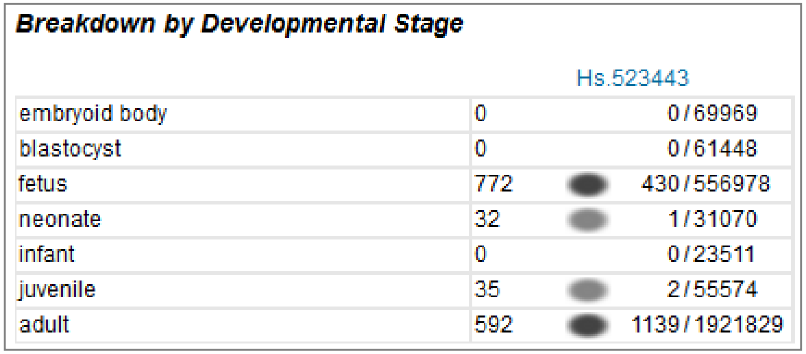
Breakdown by Body Sites

	Hs.523443		
adipose tissue	1165	●	15 / 12866
adrenal gland	121	●	4 / 32940
ascites	0		0 / 39834
bladder	0		0 / 29860
blood	5243	●	641 / 122252
bone	321	●	23 / 71618
bone marrow	3631	●	177 / 48737
brain	71	●	78 / 1092688
cervix	0		0 / 48486
connective tissue	449	●	67 / 149072
ear	62	●	1 / 16100
embryonic tissue	32	●	7 / 212896
esophagus	0		0 / 20154
eye	124	●	26 / 208840
heart	223	●	20 / 89524



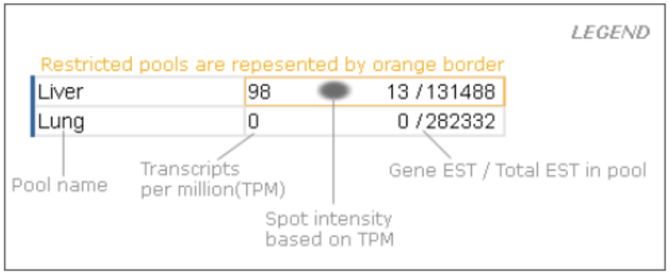
Breakdown by Health State

	Hs.523443		
adrenal tumor	237	●	3 / 12655
bladder carcinoma	0		0 / 17584
breast (mammary gland) tumor	0		0 / 93090
cervical tumor	0		0 / 34484
chondrosarcoma	724	●	60 / 82838
colorectal tumor	17	●	2 / 112517
esophageal tumor	0		0 / 17245
gastrointestinal tumor	8	●	1 / 118498
germ cell tumor	7	●	2 / 263230
glioma	18	●	2 / 107194
head and neck tumor	0		0 / 133826
kidney tumor	87	●	6 / 68872
leukemia	2265	●	214 / 94479
liver tumor	10	●	1 / 96023



Breakdown by Developmental Stage

	Hs.523443		
embryoid body	0		0 / 69969
blastocyst	0		0 / 61448
fetus	772	●	430 / 556978
neonate	32	●	1 / 31070
infant	0		0 / 23511
juvenile	35	●	2 / 55574
adult	592	●	1139 / 1921829



LEGEND

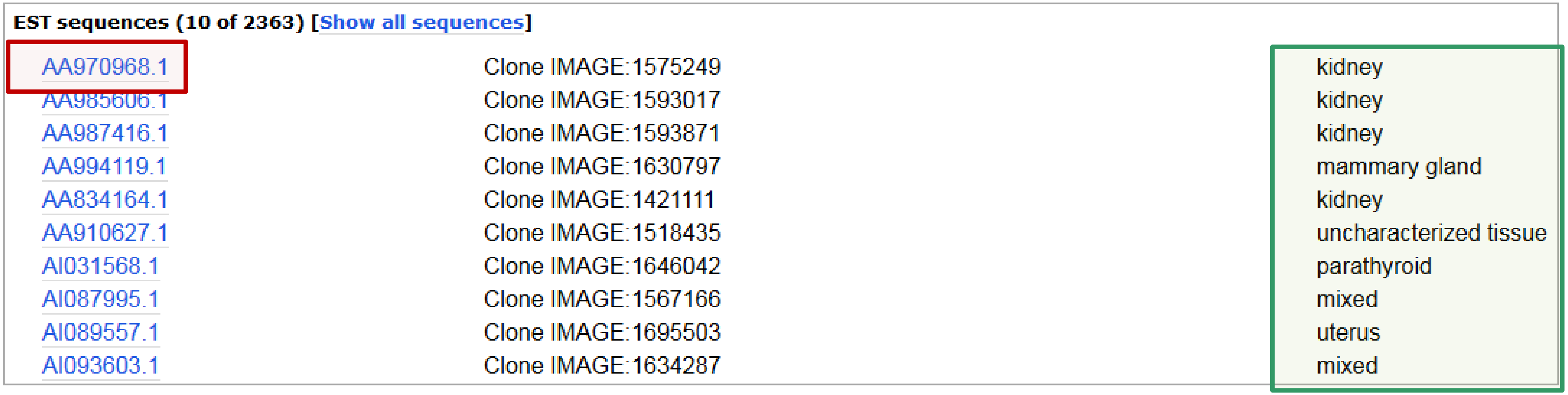
Restricted pools are repesented by orange border

Liver	98	●	13 / 131488
Lung	0		0 / 282332

Theresa Schwarz

Expressed sequence tags (ESTs)

→ Go back and scroll down to **EST sequences**



EST sequences (10 of 2363) [Show all sequences]

AA970968.1	Clone IMAGE:1575249	kidney
AA985606.1	Clone IMAGE:1593017	kidney
AA987416.1	Clone IMAGE:1593871	kidney
AA994119.1	Clone IMAGE:1630797	mammary gland
AA834164.1	Clone IMAGE:1421111	kidney
AA910627.1	Clone IMAGE:1518435	uncharacterized tissue
AI031568.1	Clone IMAGE:1646042	parathyroid
AI087995.1	Clone IMAGE:1567166	mixed
AI089557.1	Clone IMAGE:1695503	uterus
AI093603.1	Clone IMAGE:1634287	mixed

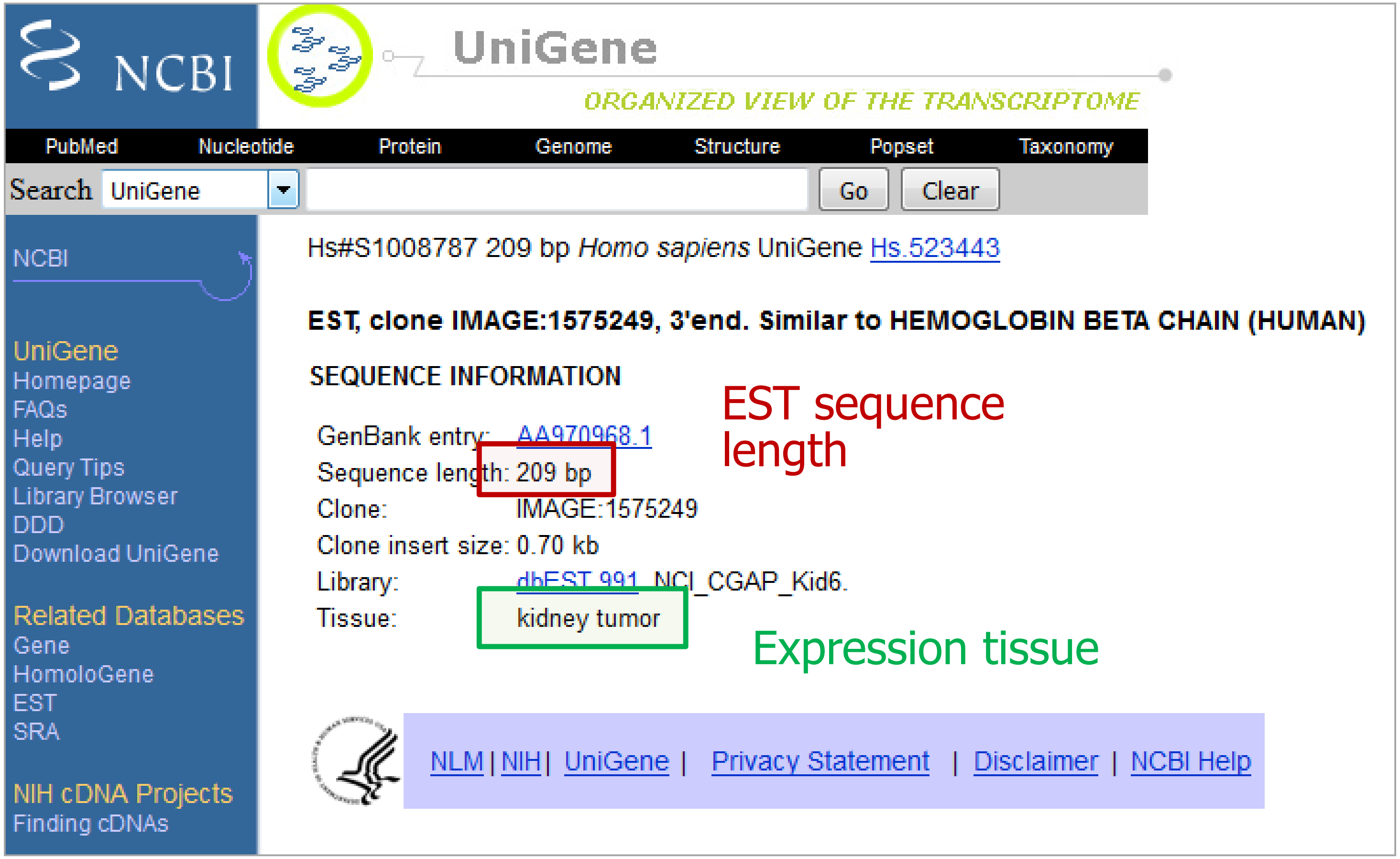
EST sequence identification numbers

Where is the EST expressed

→ Click on one **EST sequence**

Th

Expressed sequence tags (ESTs)



Hs#S1008787 209 bp *Homo sapiens* UniGene Hs.523443

EST, clone IMAGE:1575249, 3'end. Similar to HEMOGLOBIN BETA CHAIN (HUMAN)

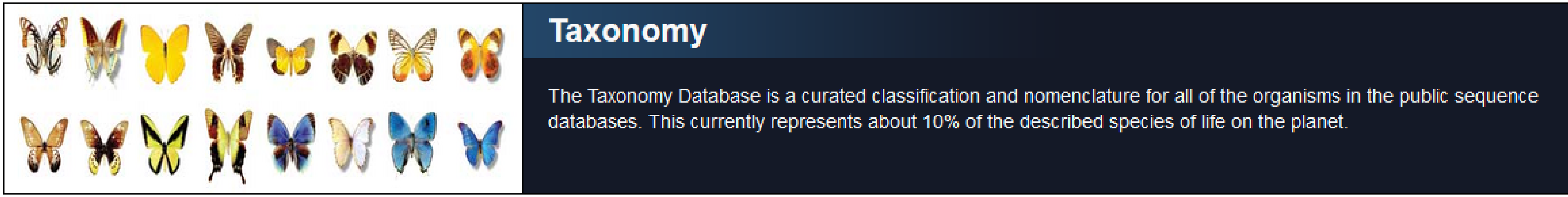
SEQUENCE INFORMATION

GenBank entry: AA970968.1
Sequence length: 209 bp
Clone: IMAGE:1575249
Clone insert size: 0.70 kb
Library: dbEST 991 NCI_CGAP_Kid6.
Tissue: kidney tumor

EST sequence length

Expression tissue

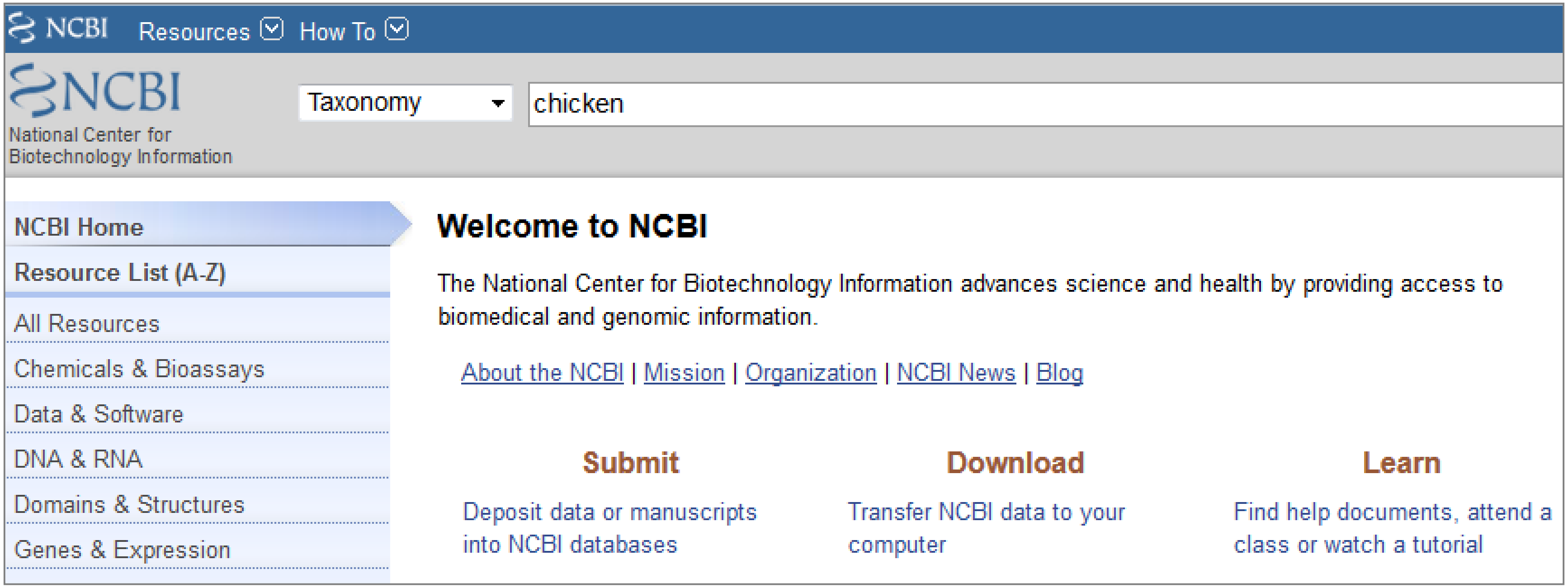
Taxonomy Database



Taxonomy

The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet.

- Search for **chicken** in the **taxonomy database**



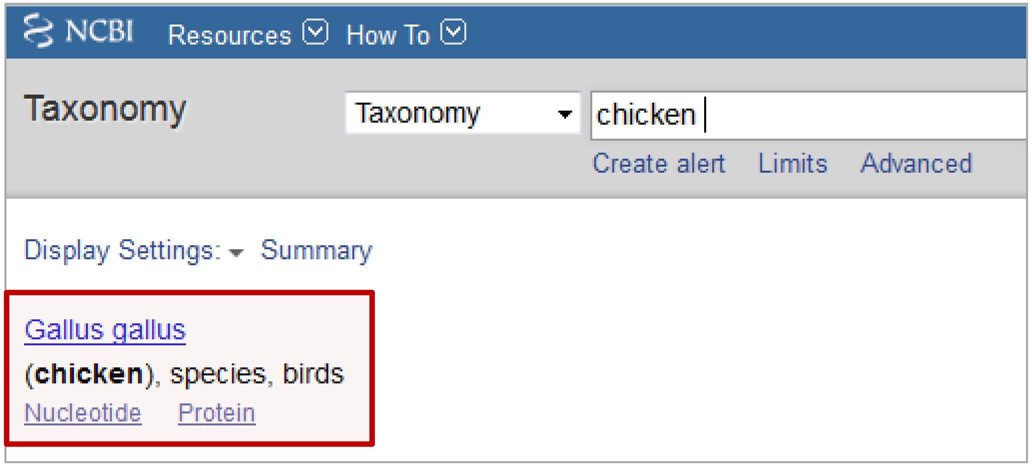
Welcome to NCBI

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	Download	Learn
Deposit data or manuscripts into NCBI databases	Transfer NCBI data to your computer	Find help documents, attend a class or watch a tutorial

Taxonomy Database



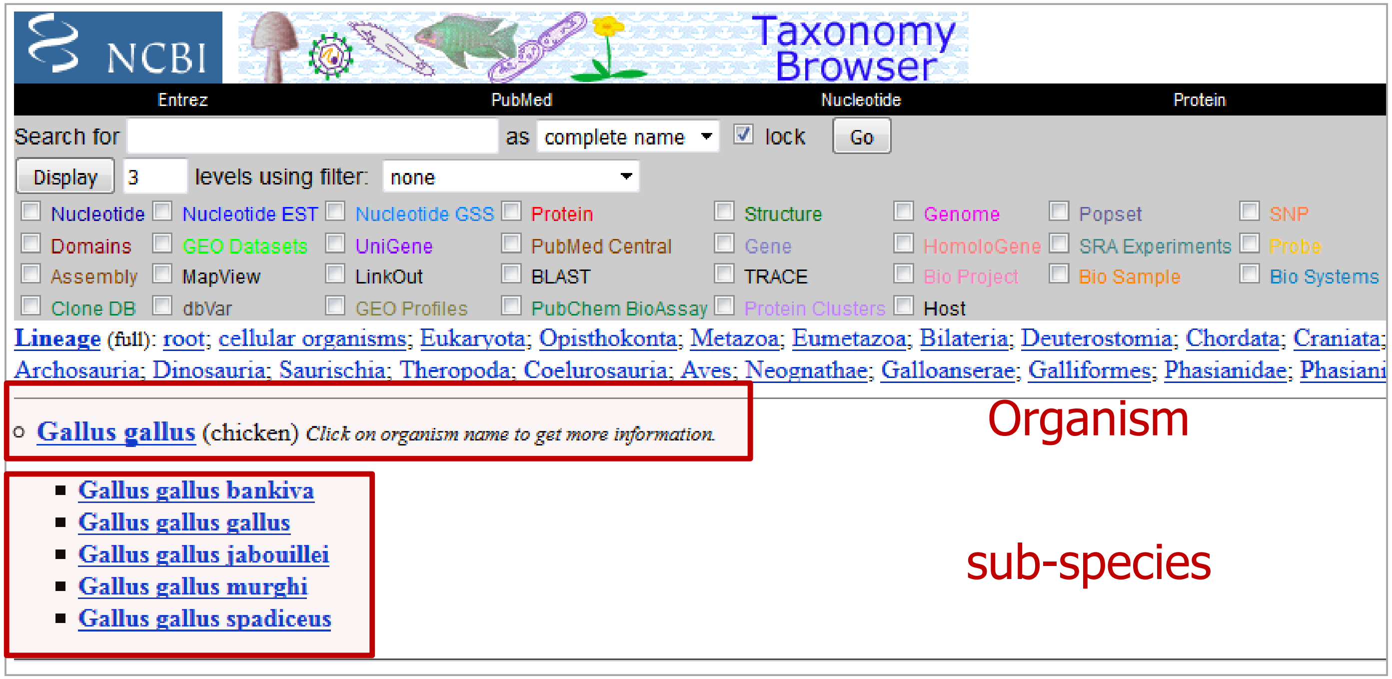
Taxonomy

Taxonomy ▾ chicken

Gallus gallus

(**chicken**), species, birds

Nucleotide Protein



Search for as complete name ▾ ☑ lock

3 levels using filter: none ▾

☐ Nucleotide ☐ Nucleotide EST ☐ Nucleotide GSS ☐ Protein ☐ Structure ☐ Genome ☐ Popset ☐ SNP
☐ Domains ☐ GEO Datasets ☐ UniGene ☐ PubMed Central ☐ Gene ☐ HomoloGene ☐ SRA Experiments ☐ Probe
☐ Assembly ☐ MapView ☐ LinkOut ☐ BLAST ☐ TRACE ☐ Bio Project ☐ Bio Sample ☐ Bio Systems
☐ Clone DB ☐ dbVar ☐ GEO Profiles ☐ PubChem BioAssay ☐ Protein Clusters ☐ Host

Lineage (full): root; cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Archosauria; Dinosauria; Saurischia; Theropoda; Coelurosauria; Aves; Neognathae; Galloanserae; Galliformes; Phasianidae; Phasiani

- **Gallus gallus** (chicken) *Click on organism name to get more information.*
  - **Gallus gallus bankiva**
  - **Gallus gallus gallus**
  - **Gallus gallus jabouillei**
  - **Gallus gallus murghi**
  - **Gallus gallus spadiceus**

Organism

sub-species

Theresa Schwarz

→ Click on **Gallus gallus**

Taxonomy Database

Search for as complete name ☑ lock

3 levels using filter: none

Gallus gallus

Taxonomy ID: 9031

Genbank common name: **chicken**

Inherited blast name: **birds**

Rank: species

Genetic code: Translation table 1 (Standard)

Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)

Other names:

common name: **chickens**

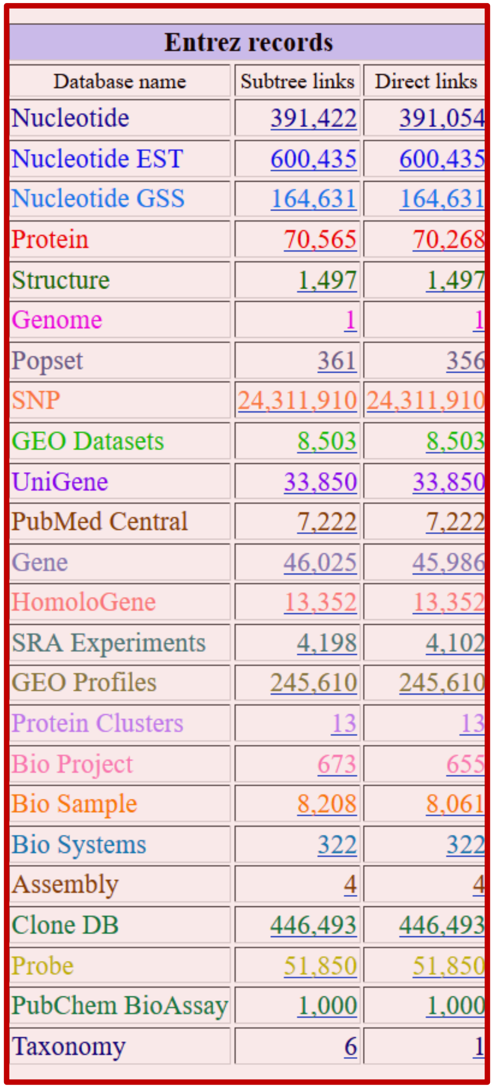
common name: **bantam**

includes: **red junglefowl**

includes: **dwarf Leghorn chickens**

misnomer: **Gallus gallus domesticus**

misnomer: **Gallus domesticus**



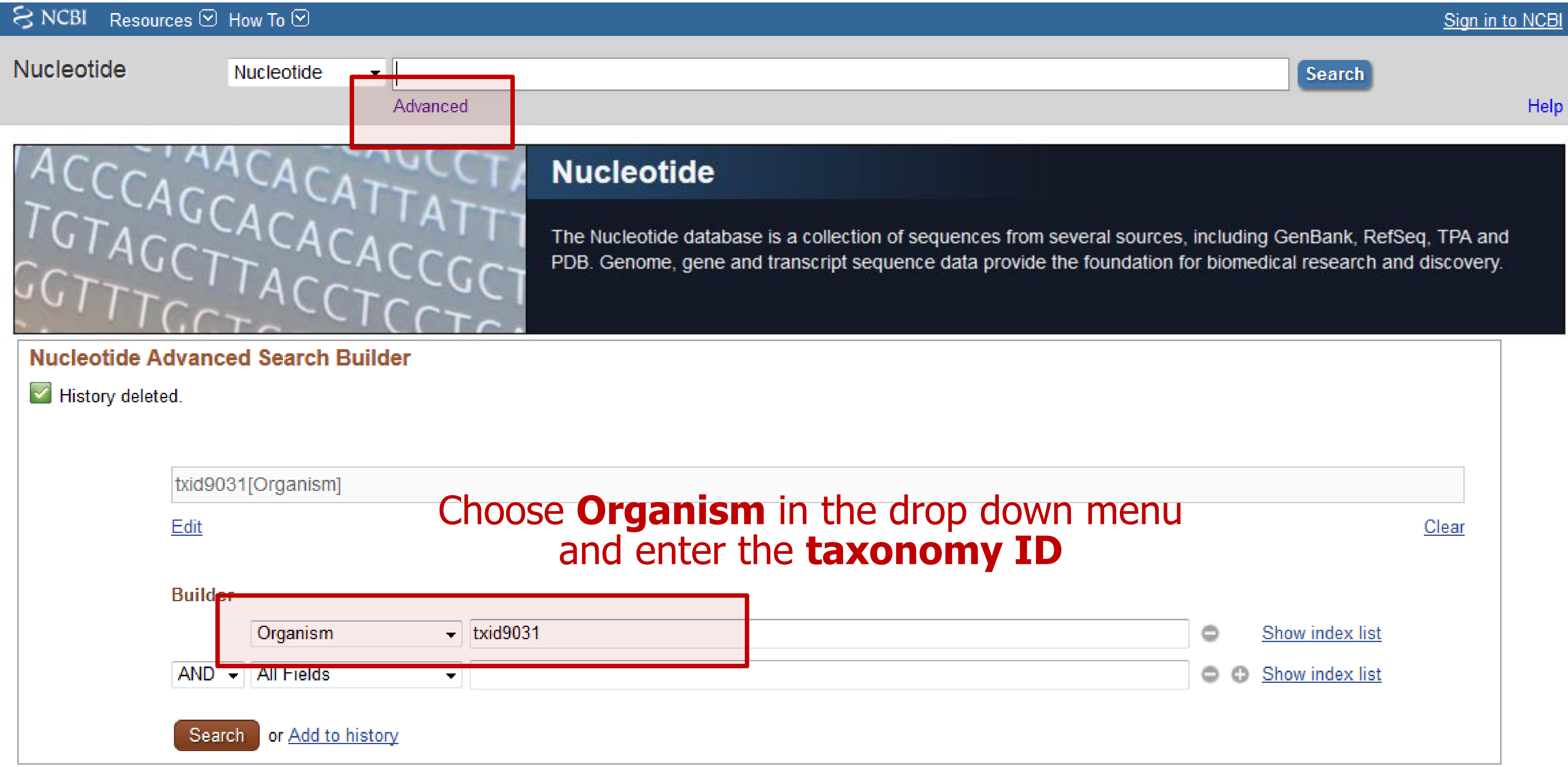
Entrez records

Database name	Subtree links	Direct links
Nucleotide	391,422	391,054
Nucleotide EST	600,435	600,435
Nucleotide GSS	164,631	164,631
Protein	70,565	70,268
Structure	1,497	1,497
Genome	1	1
Popset	361	356
SNP	24,311,910	24,311,910
GEO Datasets	8,503	8,503
UniGene	33,850	33,850
PubMed Central	7,222	7,222
Gene	46,025	45,986
HomoloGene	13,352	13,352
SRA Experiments	4,198	4,102
GEO Profiles	245,610	245,610
Protein Clusters	13	13
Bio Project	673	655
Bio Sample	8,208	8,061
Bio Systems	322	322
Assembly	4	4
Clone DB	446,493	446,493
Probe	51,850	51,850
PubChem BioAssay	1,000	1,000
Taxonomy	6	1

right-hand side:
direct links to other
databases

Nucleotide Database

- Search for the organism *gallus gallus* in the **Nucleotide** database using **advanced search**



Nucleotide

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

Nucleotide Advanced Search Builder

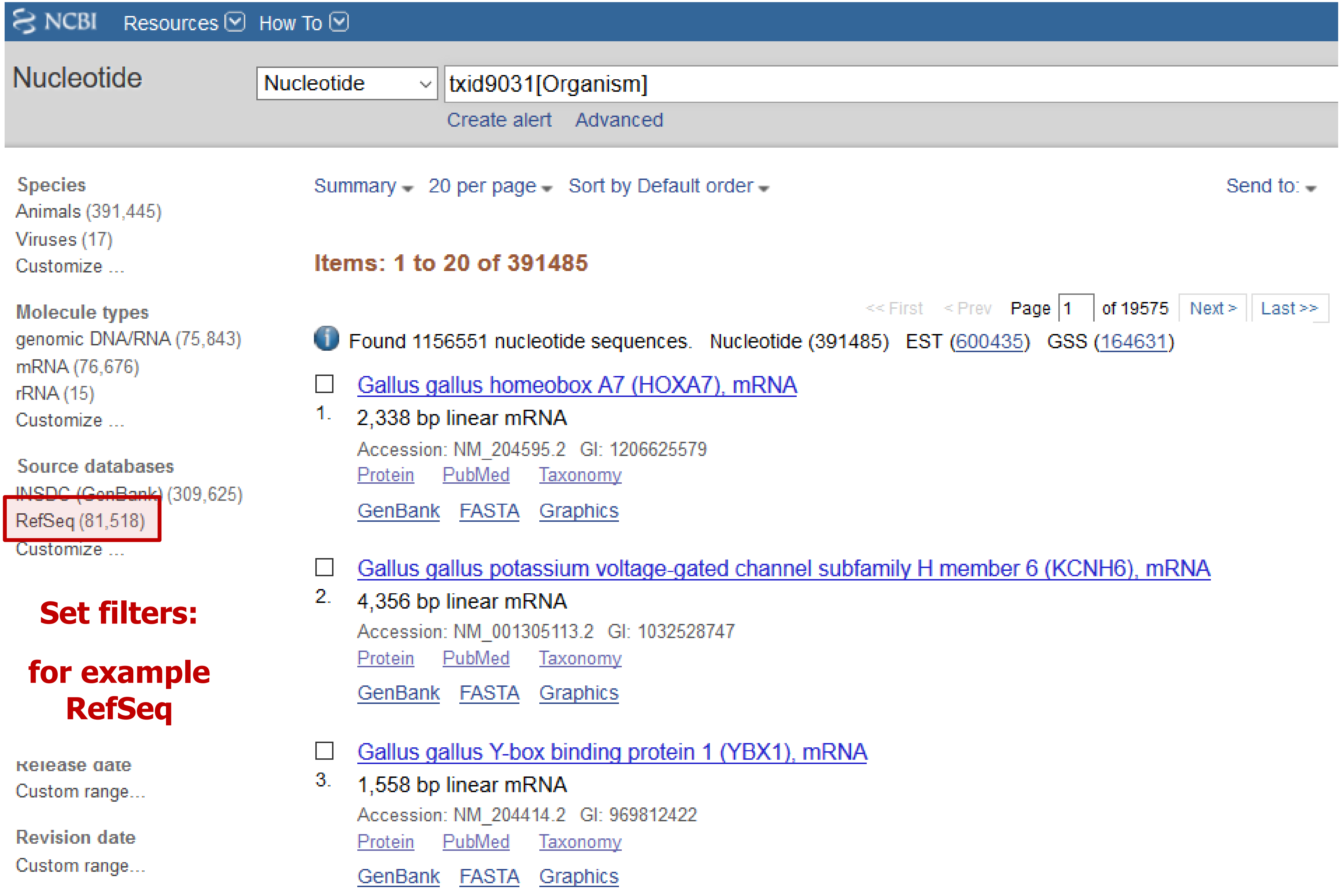
☑ History deleted.

txid9031[Organism]

Organism: txid9031

Choose **Organism** in the drop down menu and enter the **taxonomy ID**

Nucleotide Database



txid9031[Organism]

Species
Animals (391,445)
Viruses (17)
Customize ...

Molecule types
genomic DNA/RNA (75,843)
mRNA (76,676)
rRNA (15)
Customize ...

Source databases
INSDC (GenBank) (309,625)
RefSeq (81,518)
Customize ...

Set filters:

for example RefSeq

Release date
Custom range...

Revision date
Custom range...

Items: 1 to 20 of 391485

Page 1 of 19575

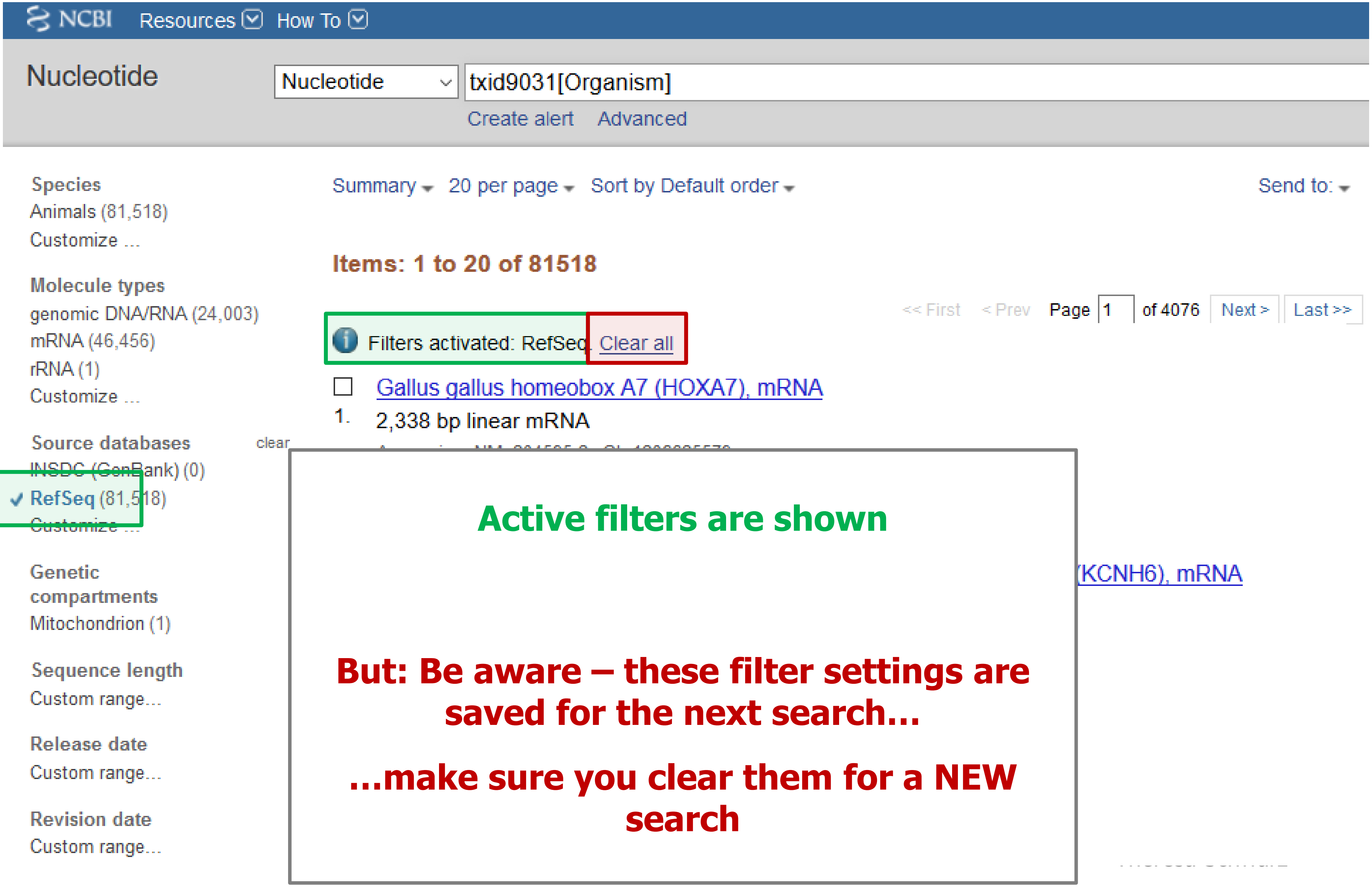
Found 1156551 nucleotide sequences. Nucleotide (391485) EST (600435) GSS (164631)

☐ Gallus gallus homeobox A7 (HOXA7), mRNA
1. 2,338 bp linear mRNA
Accession: NM_204595.2 GI: 1206625579
Protein PubMed Taxonomy
GenBank FASTA Graphics

☐ Gallus gallus potassium voltage-gated channel subfamily H member 6 (KCNH6), mRNA
2. 4,356 bp linear mRNA
Accession: NM_001305113.2 GI: 1032528747
Protein PubMed Taxonomy
GenBank FASTA Graphics

☐ Gallus gallus Y-box binding protein 1 (YBX1), mRNA
3. 1,558 bp linear mRNA
Accession: NM_204414.2 GI: 969812422
Protein PubMed Taxonomy
GenBank FASTA Graphics

Nucleotide Database



txid9031[Organism]

Species
Animals (81,518)
Customize ...

Molecule types
genomic DNA/RNA (24,003)
mRNA (46,456)
rRNA (1)
Customize ...

Source databases
INSDC (GenBank) (0)
✓ RefSeq (81,518)
Customize ...

Genetic compartments
Mitochondrion (1)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

Items: 1 to 20 of 81518

Filters activated: RefSeq. Clear all

☐ 1. Gallus gallus homeobox A7 (HOXA7), mRNA
2,338 bp linear mRNA

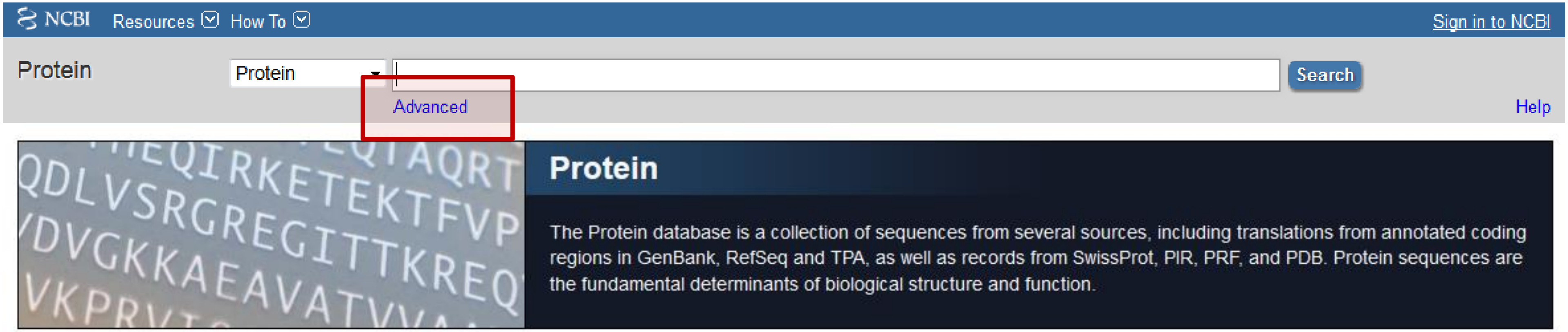
(KCNH6), mRNA

Active filters are shown

But: Be aware – these filter settings are saved for the next search…

…make sure you clear them for a NEW search

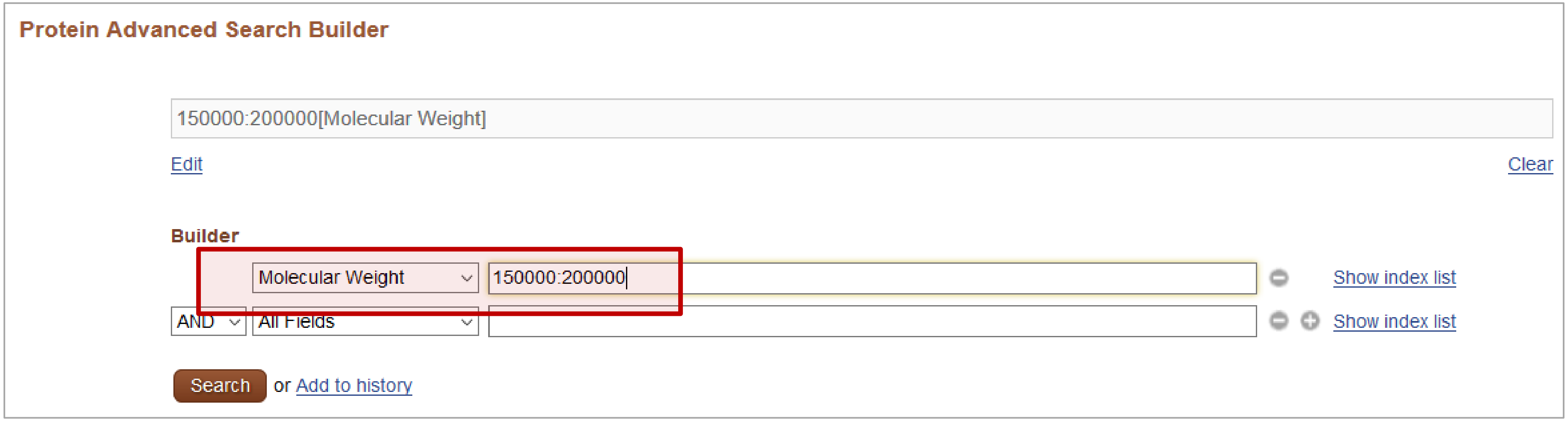
Protein Database



Protein

The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.

- How many proteins in **humans** are between **150 and 200kDa**?

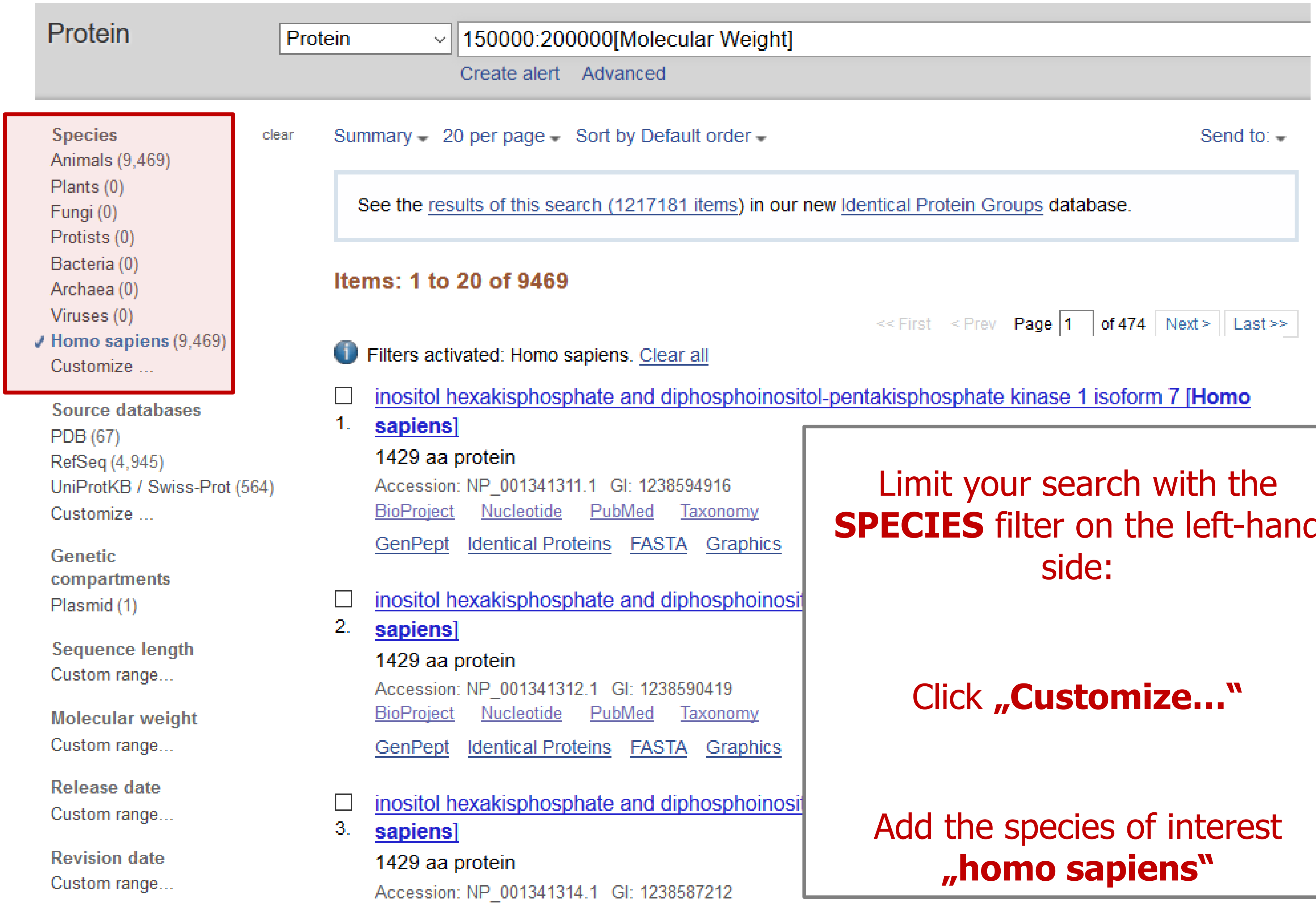


Protein Advanced Search Builder

150000:200000[Molecular Weight]

Molecular Weight: 150000:200000

Protein Database



Protein

Protein ▾ 150000:200000[Molecular Weight]

Create alert Advanced

Species
Animals (9,469)
Plants (0)
Fungi (0)
Protists (0)
Bacteria (0)
Archaea (0)
Viruses (0)
✓ Homo sapiens (9,469)
Customize ...

clear

Source databases
PDB (67)
RefSeq (4,945)
UniProtKB / Swiss-Prot (564)
Customize ...

Genetic compartments
Plasmid (1)

Sequence length
Custom range...

Molecular weight
Custom range...

Release date
Custom range...

Revision date
Custom range...

Summary ▾ 20 per page ▾ Sort by Default order ▾ Send to: ▾

See the results of this search (1217181 items) in our new Identical Protein Groups database.

Items: 1 to 20 of 9469

<< First < Prev Page 1 of 474 Next > Last >>

Filters activated: Homo sapiens. Clear all

☐ 1. inositol hexakisphosphate and diphosphoinositol-pentakisphosphate kinase 1 isoform 7 [**Homo sapiens**]
1429 aa protein
Accession: NP_001341311.1 GI: 1238594916
BioProject Nucleotide PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

☐ 2. inositol hexakisphosphate and diphosphoinosit
sapiens]
1429 aa protein
Accession: NP_001341312.1 GI: 1238590419
BioProject Nucleotide PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

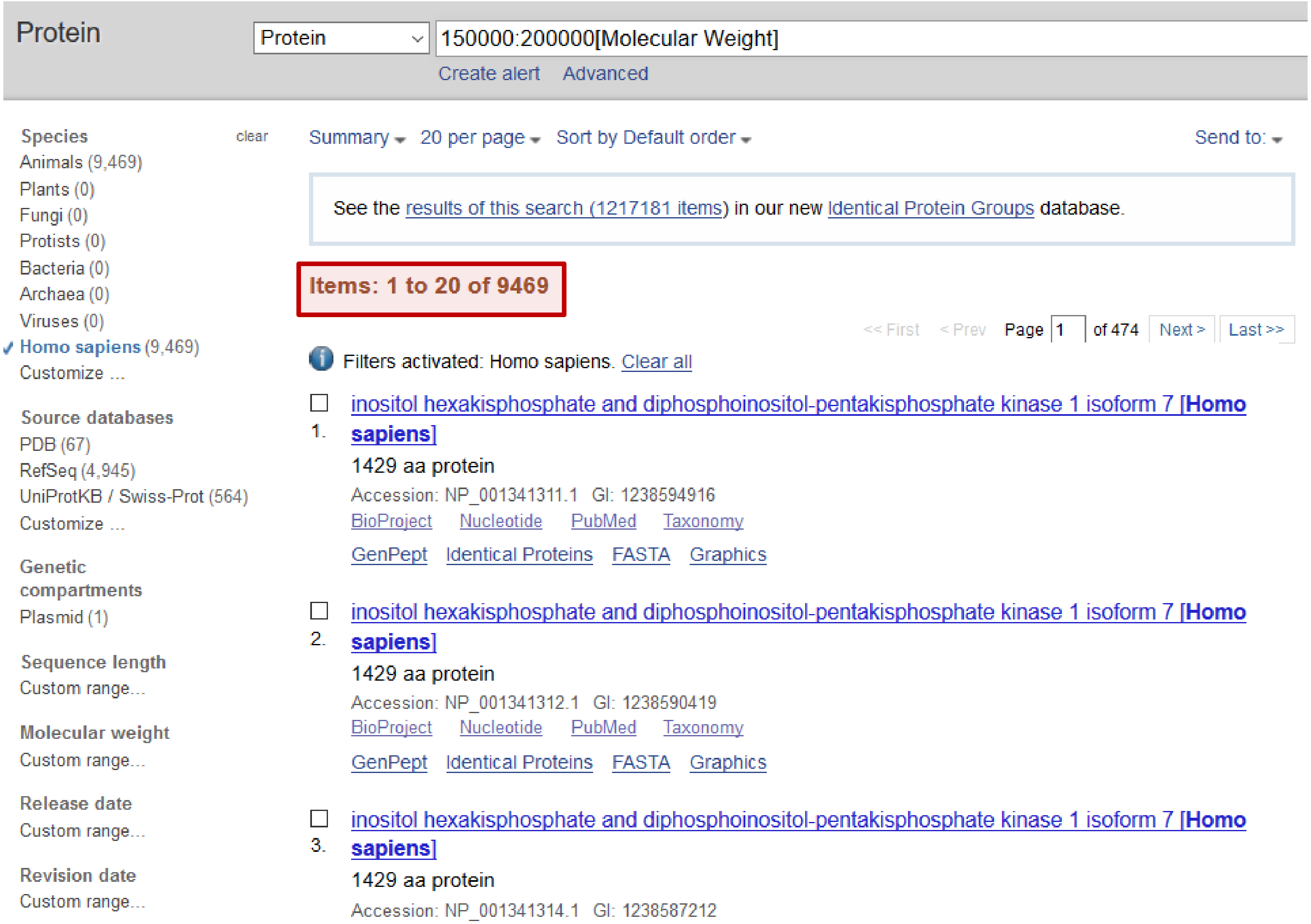
☐ 3. inositol hexakisphosphate and diphosphoinosit
sapiens]
1429 aa protein
Accession: NP_001341314.1 GI: 1238587212

Limit your search with the **SPECIES** filter on the left-hand side:

Click **„Customize…“**

Add the species of interest **„homo sapiens“**

Protein Database



Protein

Protein | 150000:200000[Molecular Weight]

Create alert Advanced

Species clear
- Animals (9,469)
- Plants (0)
- Fungi (0)
- Protists (0)
- Bacteria (0)
- Archaea (0)
- Viruses (0)
- ✓ Homo sapiens (9,469)
- Customize ...

Source databases
- PDB (67)
- RefSeq (4,945)
- UniProtKB / Swiss-Prot (564)
- Customize ...

Genetic compartments
- Plasmid (1)

Sequence length
- Custom range...

Molecular weight
- Custom range...

Release date
- Custom range...

Revision date
- Custom range...

Summary 20 per page Sort by Default order Send to:

See the results of this search (1217181 items) in our new Identical Protein Groups database.

Items: 1 to 20 of 9469

<< First < Prev Page 1 of 474 Next > Last >>

Filters activated: Homo sapiens. Clear all

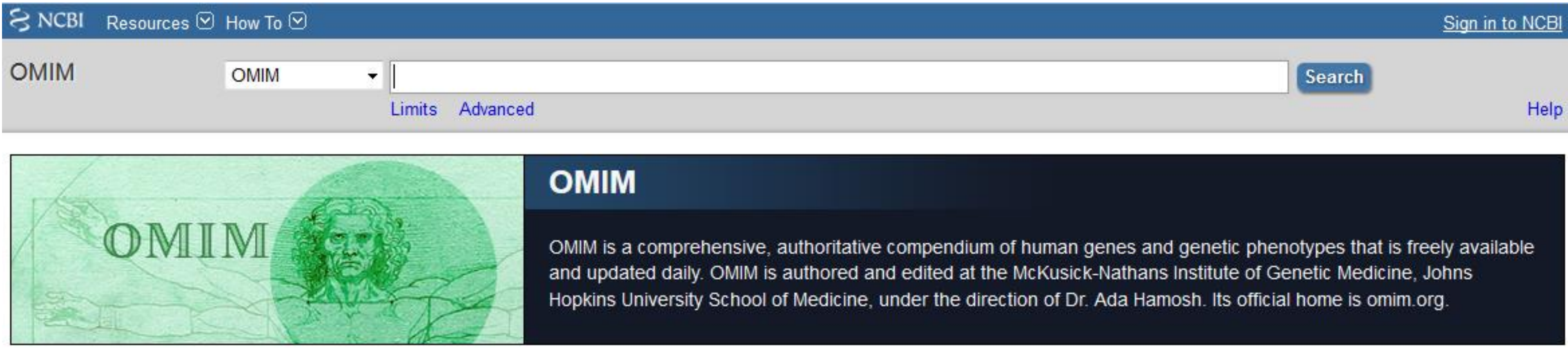
☐ 1. inositol hexakisphosphate and diphosphoinositol-pentakisphosphate kinase 1 isoform 7 [**Homo sapiens**]
1429 aa protein
Accession: NP_001341311.1 GI: 1238594916
BioProject Nucleotide PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

☐ 2. inositol hexakisphosphate and diphosphoinositol-pentakisphosphate kinase 1 isoform 7 [**Homo sapiens**]
1429 aa protein
Accession: NP_001341312.1 GI: 1238590419
BioProject Nucleotide PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

☐ 3. inositol hexakisphosphate and diphosphoinositol-pentakisphosphate kinase 1 isoform 7 [**Homo sapiens**]
1429 aa protein
Accession: NP_001341314.1 GI: 1238587212

9469 proteins in the human organism have a molecular weigth between 150-200kDa.

OMIM database



OMIM

OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. OMIM is authored and edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, under the direction of Dr. Ada Hamosh. Its official home is omim.org.

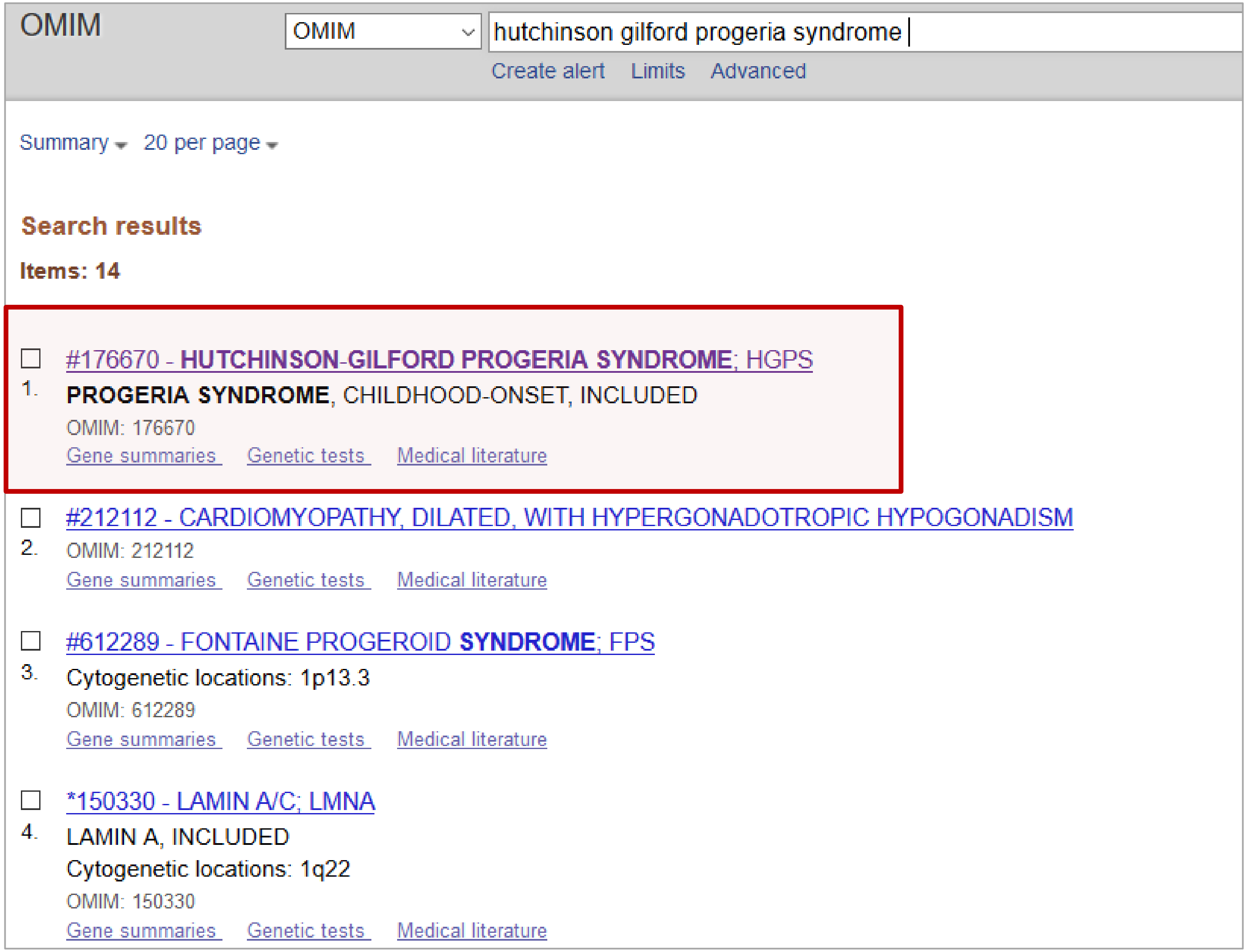
OMIM is a great source to look up information about **genetic diseases**

OMIM = Online Mendelian Inheritance in Man

created by Dr. Victor McKusick; led by Dr. Ada Hamosh at JHMI

- Search for the **hutchinson gilford progeria syndrome**

OMIM database



hutchinson gilford progeria syndrome

Search results

Items: 14

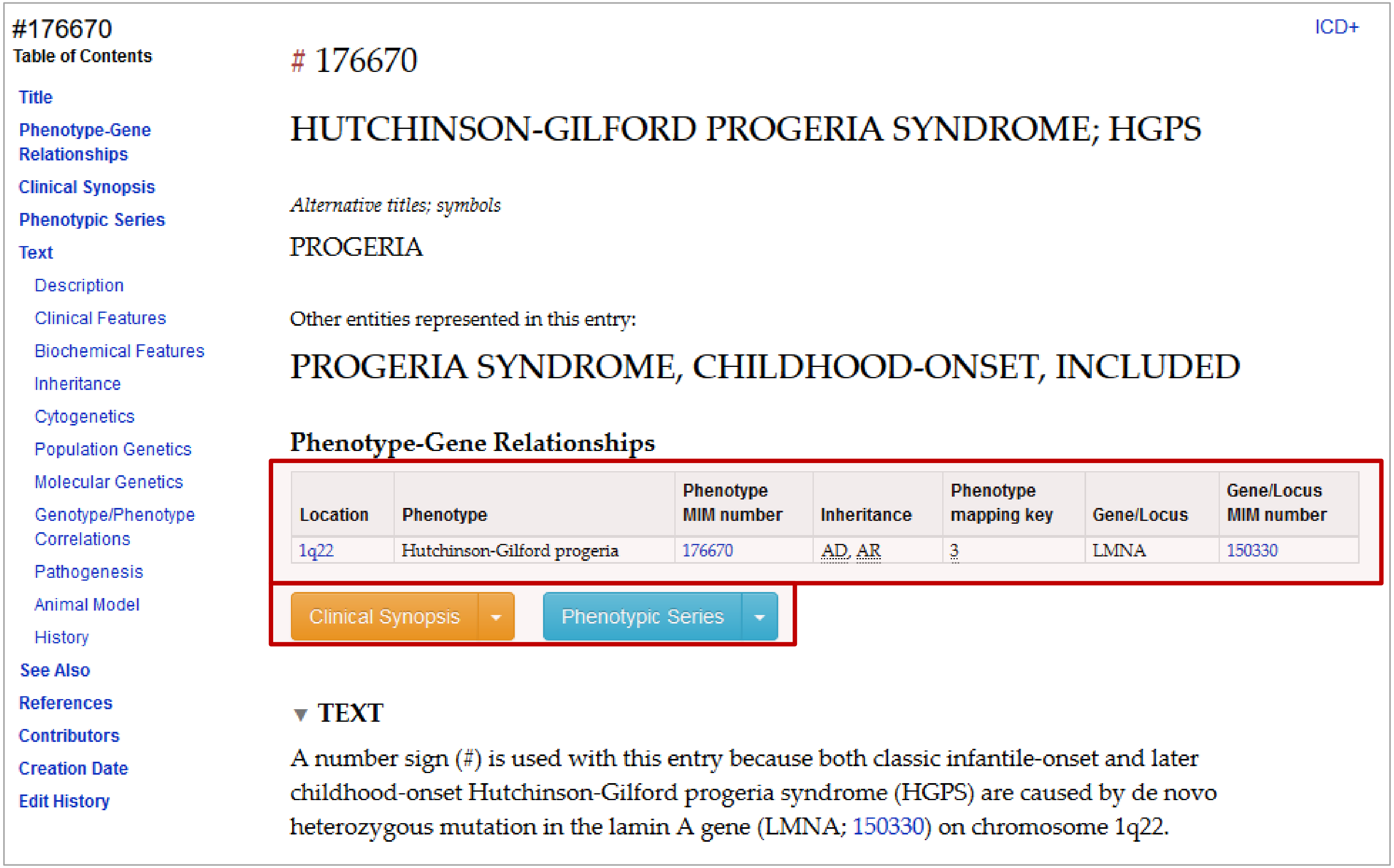
☐ 1. #176670 - **HUTCHINSON-GILFORD PROGERIA SYNDROME**; HGPS
PROGERIA SYNDROME, CHILDHOOD-ONSET, INCLUDED
OMIM: 176670
Gene summaries Genetic tests Medical literature

☐ 2. #212112 - CARDIOMYOPATHY, DILATED, WITH HYPERGONADOTROPIC HYPOGONADISM
OMIM: 212112
Gene summaries Genetic tests Medical literature

☐ 3. #612289 - FONTAINE PROGEROID **SYNDROME**; FPS
Cytogenetic locations: 1p13.3
OMIM: 612289
Gene summaries Genetic tests Medical literature

☐ 4. *150330 - LAMIN A/C; LMNA
LAMIN A, INCLUDED
Cytogenetic locations: 1q22
OMIM: 150330
Gene summaries Genetic tests Medical literature

OMIM database



#176670
Table of Contents

- Title
- Phenotype-Gene Relationships
- Clinical Synopsis
- Phenotypic Series
- Text
  - Description
  - Clinical Features
  - Biochemical Features
  - Inheritance
  - Cytogenetics
  - Population Genetics
  - Molecular Genetics
  - Genotype/Phenotype Correlations
  - Pathogenesis
  - Animal Model
  - History
- See Also
- References
- Contributors
- Creation Date
- Edit History

176670

HUTCHINSON-GILFORD PROGERIA SYNDROME; HGPS

Alternative titles; symbols

PROGERIA

Other entities represented in this entry:

PROGERIA SYNDROME, CHILDHOOD-ONSET, INCLUDED

Phenotype-Gene Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key	Gene/Locus	Gene/Locus MIM number
1q22	Hutchinson-Gilford progeria	176670	AD, AR	3	LMNA	150330

▼ TEXT

A number sign (#) is used with this entry because both classic infantile-onset and later childhood-onset Hutchinson-Gilford progeria syndrome (HGPS) are caused by de novo heterozygous mutation in the lamin A gene (LMNA; 150330) on chromosome 1q22.

**Here you can find a lot of information about the disease.
Scroll down to see more.**

Overview

- ☐ Grading/Requirements
- ☐ Introduction to databases
- ☐ Examples → browsing through NCBI
- ☐ **How to write a Report**

Reports

- After each lecture you get exercises you have to address in a written format = **Reports**

- Prepare reports in **groups of 2**

- Upload the report to MOODLE until next **Monday 08:00 a.m.** (as a PDF file)

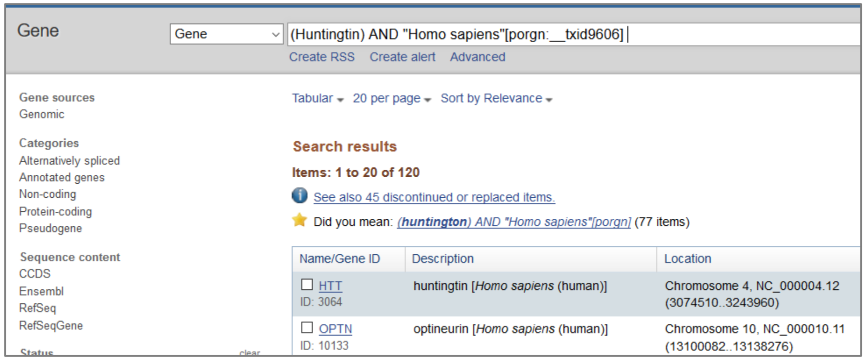
- Proper formatting:
  - Name the file "**Report#1_surname1_surename2**"
  - Insert **report number** and your **name** to the header

1. Explain in key words how you get your results
2. Show your results and critical steps in a representative screenshot
3. Write down your result/answer to the question

Example

Search for human Huntingtin in the Gene database and name the different RefSeq accession numbers for genomic DNA, mRNA and protein.

Go to NCBI -> select Gene database and search for huntingtin -> set filter to homo sapiens -> select first entry HTT -> scroll down to Genomic context and to NCBI RefSeq to find the accession numbers



Gene | Gene | (Huntingtin) AND "Homo sapiens"[porgn:__txid9606]

Create RSS Create alert Advanced

Gene sources
Genomic

Categories
Alternatively spliced
Annotated genes
Non-coding
Protein-coding
Pseudogene

Sequence content
CCDS
Ensembl
RefSeq
RefSeqGene

Status clear

Tabular 20 per page Sort by Relevance

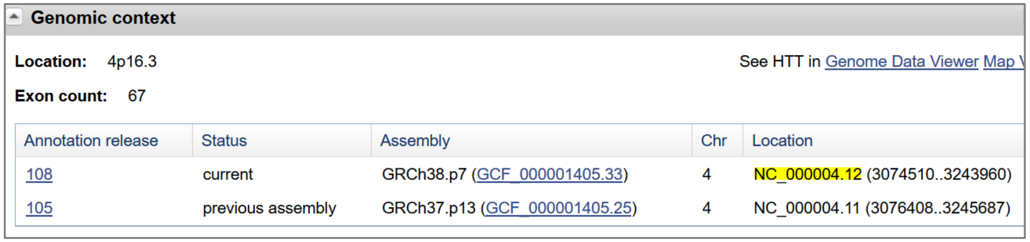
Search results

Items: 1 to 20 of 120

See also 45 discontinued or replaced items.

Did you mean: *(huntington) AND "Homo sapiens"[porgn]* (77 items)

Name/Gene ID	Description	Location
☐ HTT ID: 3064	huntingtin [*Homo sapiens* (human)]	Chromosome 4, NC_000004.12 (3074510..3243960)
☐ OPTN ID: 10133	optineurin [*Homo sapiens* (human)]	Chromosome 10, NC_000010.11 (13100082..13138276)



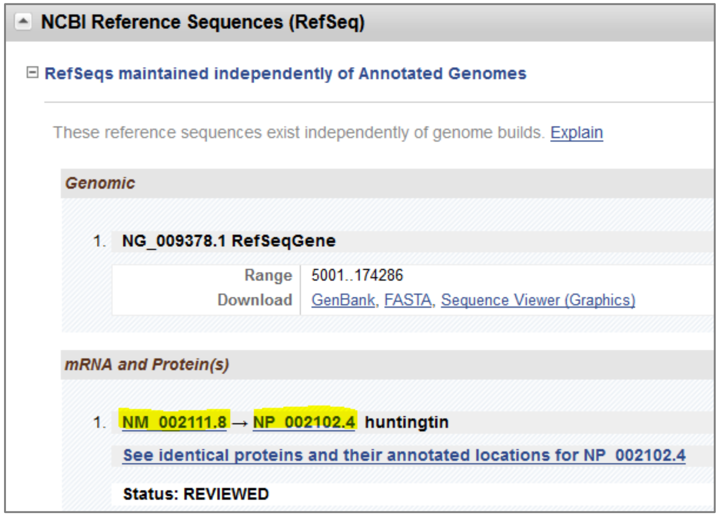
Genomic context

Location: 4p16.3

Exon count: 67

See HTT in Genome Data Viewer Map V

Annotation release	Status	Assembly	Chr	Location
108	current	GRCh38.p7 (GCF_000001405.33)	4	NC_000004.12 (3074510..3243960)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	4	NC_000004.11 (3076408..3245687)



NCBI Reference Sequences (RefSeq)

RefSeqs maintained independently of Annotated Genomes

These reference sequences exist independently of genome builds. Explain

Genomic

1. **NG_009378.1 RefSeqGene**

Range 5001..174286
Download GenBank, FASTA, Sequence Viewer (Graphics)

mRNA and Protein(s)

1. **NM_002111.8 → NP_002102.4 huntingtin**

See identical proteins and their annotated locations for NP_002102.4

Status: REVIEWED

The accession numbers are the following:

Genomic DNA	NC_000004.12
mRNA	NM_002111.8
Protein	NP_002102.4

- QUESTIONS?
- Please, download **Report #1** from MOODLE
- Upload the Report until next Monday 8:00 a.m.

GOOD LUCK!