Online Resources and Sequence Retrieval

Heiko Schmidt

Center for Integrative Bioinformatics Vienna (CIBIV)
http://www.cibiv.at

Nucleotide Databases

Three major public DNA databases

EMBL/ENA ⟷ **GenBank** ⟷ **DDBJ**

The underlying raw DNA sequences are identical

Nucleotide Databases

INSDC: The 3 major public DNA databases
(Intl. Nucleotide Sequence Database Collaboration)

EMBL/ENA ⟷ **GenBank** ⟷ **DDBJ**

European Nucleotide Archive housed at EBI European Bioinformatics Institute (Hinxton, UK)	Housed at NCBI National Center for Biotechnology Information (USA)	DNA DataBank of Japan housed at NIG National Institute of Genetics (Japan)

Nucleotide Databases

GenBank	An annotated collection of all publicly available nucleotide and amino acid sequences
Sequence Read Archive (SRA)	sequencing data from high-throughput sequencing platforms like Illumina, Roche 454, PacBio
Genome	sequence and map data from whole genomes of organisms (573 Archaea, 8573 Bacteria, 1936 Eukaryotes, 5482 Viruses), complete and in progress
HomoloGene	A gene homology database that compares nucleotide sequences between pairs of organisms to identify putative orthologs
Taxonomy	names and phylogenetic lineages of more than 360,000 organisms

Nucleotide Databases

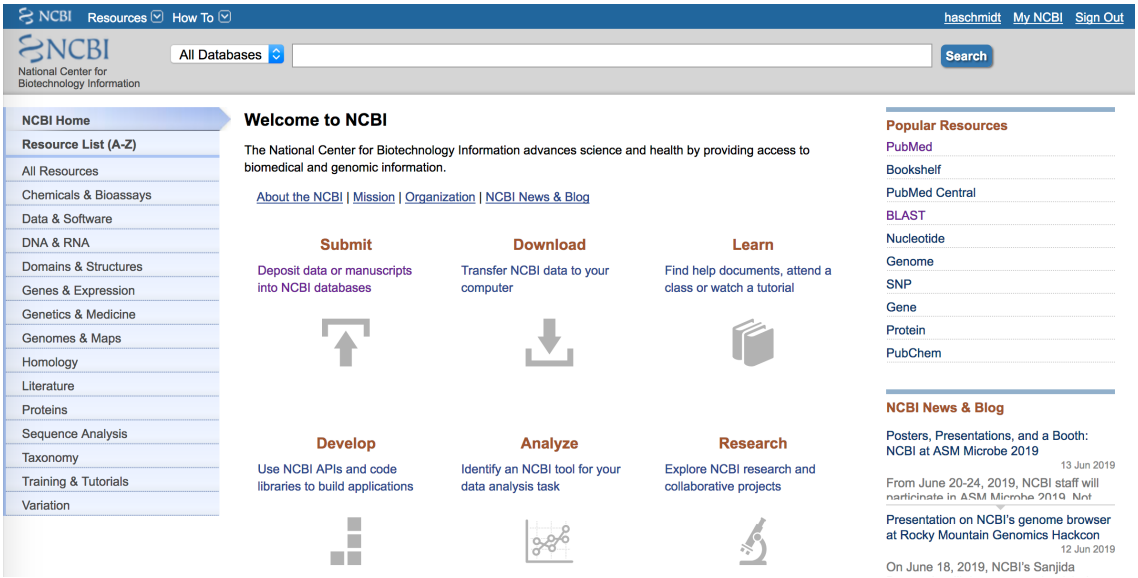
dbSNP	A central repository for both single-base nucleotide substitutions and short deletion and insertion polymorphisms
RefSeq	non-redundant reference sequences standards (including genomic DNA, mRNAs, and proteins for known genes) for genome annotation, gene identification, and comparative analyses
Gene	Gene 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.
ClinVar	archive of reports of clinically relevant human genetic variants and their relationships to phenotypes, with supporting evidence

NCBI

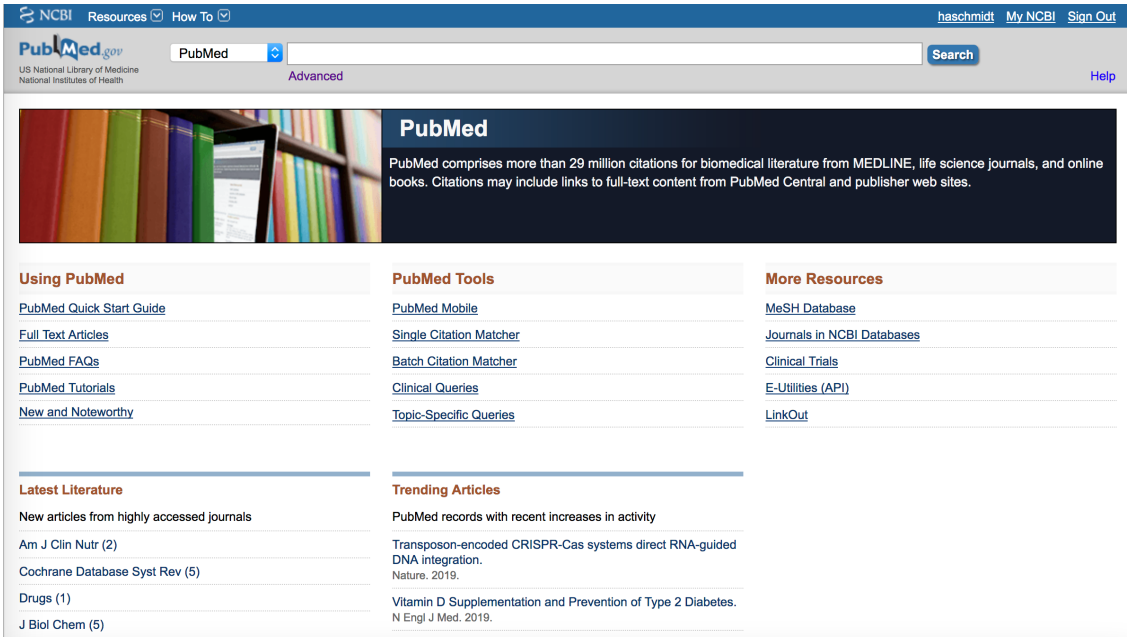
National Center for Biotechnology Information (NCBI)

www.ncbi.nlm.nih.gov

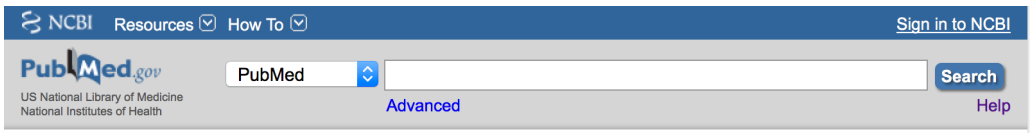
NCBI



NCBI PubMed



NCBI PubMed



PubMed is…

National Library of Medicine's search service
- in citations and abstracts for biomedical literature from MEDLINE, life science journals, and online books.
- covers a total of over 29 million entries (reference and abstracts)
- links to participating online journals

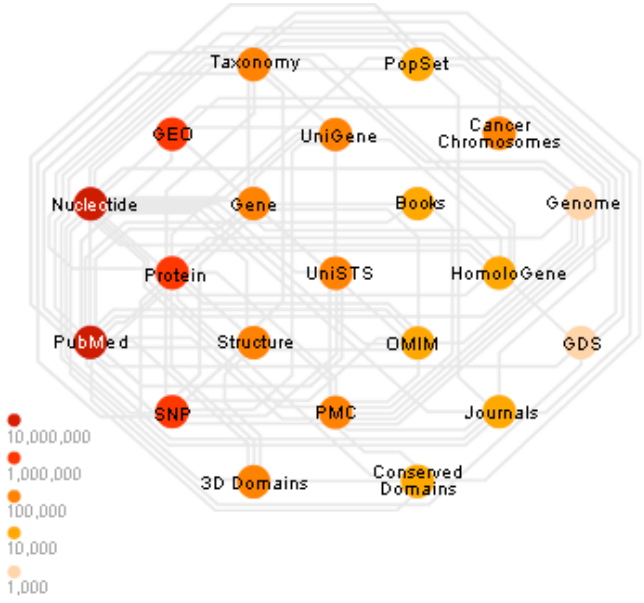
NCBI (Entrez) System

The NCBI/Entrez system integrates…

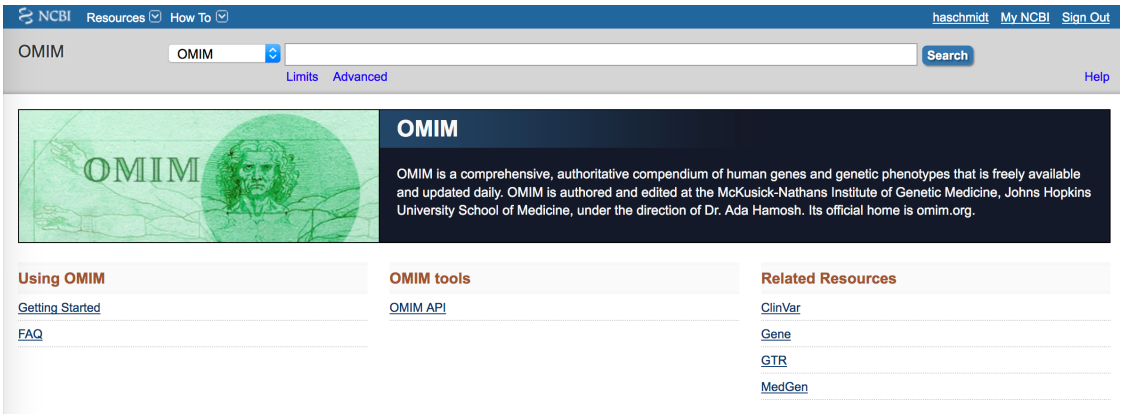
- the scientific literature
- DNA and protein sequence databases
- 3D protein structure data
- population study data sets
- assemblies of complete genomes
- many other databases and tools

NCBI database integration

Entrez is a search and retrieval system that integrates NCBI databases



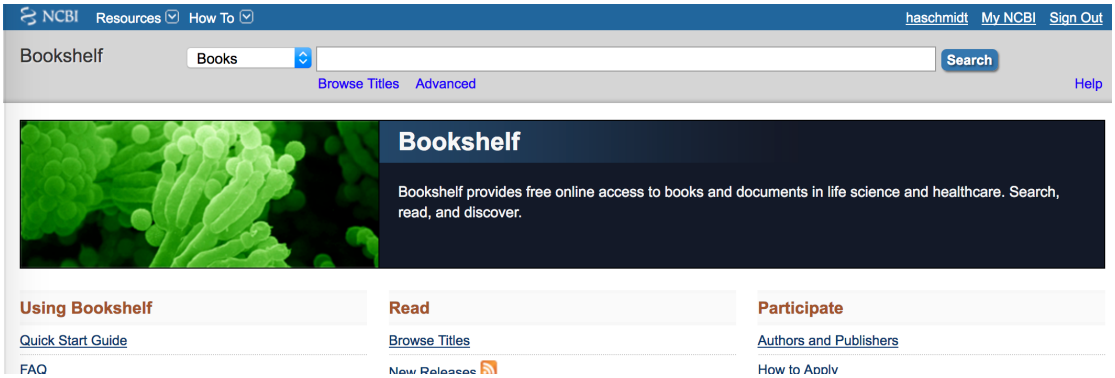
NCBI OMIM



OMIM is…

- Online Mendelian Inheritance in Man
- catalog of human genes and genetic disorders
- manually curated

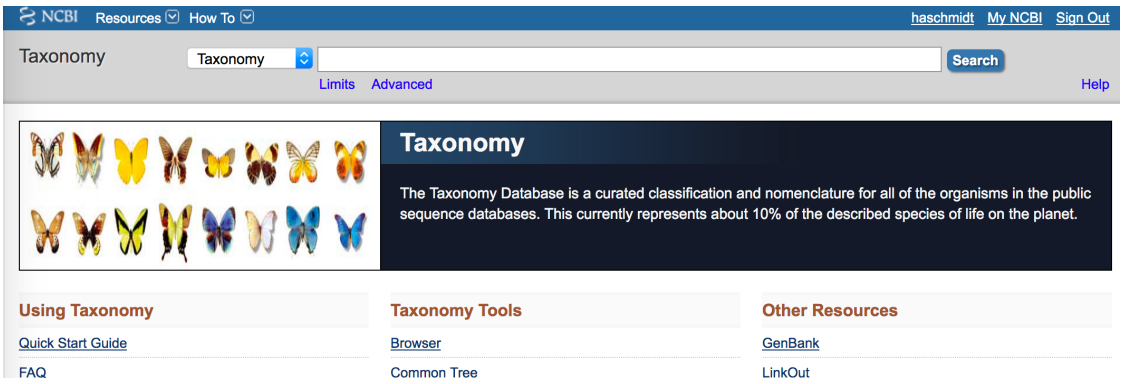
NCBI Bookshelf



Bookshelf is…

- searchable resource of on-line books

NCBI Taxonomy



Taxonomy Browser is…

- browser for the major divisions of living organisms (archaea, bacteria, eukaryota, viruses)
- taxonomy information such as genetic codes
- data sources for each taxon
- molecular data on extinct organisms

Accessing Information

Accession numbers are labels for sequences

NCBI includes databases (such as GenBank) that contain information on DNA, RNA, or protein sequences.

You may want to acquire information beginning with a query such as the name of a protein of interest, or the raw nucleotides comprising a DNA sequence of interest.

DNA sequences and other molecular data are tagged with accession numbers that are used to uniquely identify a sequence or other record relevant to molecular data.

Accessing Information

What is an accession number?
An accession number is label that used to identify a sequence. It is a string of letters and/or numbers that corresponds to a molecular sequence.

Examples (all for retinol-binding protein, RBP4) from different resouces:

- X02775 GenBank genomic DNA sequence
- NT_030059 Genomic contig
- Rs7079946 dbSNP (single nucleotide polymorphism)

- N91759.1 An expressed sequence tag (1 of 170)
- NM_006744 RefSeq DNA sequence (from a transcript)

- NP_007635 RefSeq protein
- AAC02945 GenBank protein
- Q28369 SwissProt protein
- 1KT7 Protein Data Bank structure record

NCBI Databases

Several ways to access DNA and protein sequences

[1] **NCBI Gene with RefSeq**
https://www.ncbi.nlm.nih.gov/
[2] European Bioinformatics Institute (EBI) with InterPro and Ensembl (separate from NCBI)
https://www.ebi.ac.uk/
https://www.ebi.ac.uk/interpro/
http://www.ensembl.org/

NCBI Databases

Several ways to access DNA and protein sequences

[3] ExPASy, SIB Bioinformatics Resource Portal (separate from NCBI, Switzerland)
https://www.expasy.org/
[4] Universal Protein Resource (UniProt) (separate from NCBI)
https://www.uniprot.org/

NCBI Databases

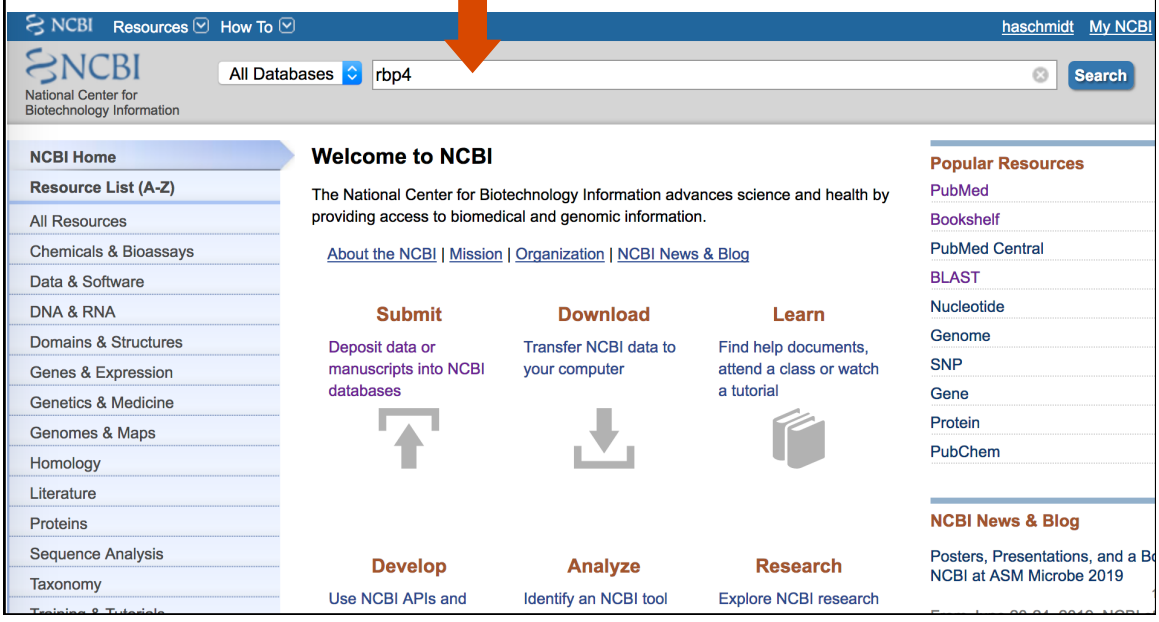
NCBI Gene with RefSeq

NCBI Gene is a great starting point: it collects key information on each gene/protein from major databases. It covers all major organisms.

RefSeq provides a curated, optimal accession number for each DNA (NM_006744) or protein (NP_007635) entry

NCBI Databases

At the NCBI home page enter "rbp4" and hit "Search"



NCBI Databases: (2019)

CIBIV
Center for Integrative Bioinformatics Vienna

Literature

Bookshelf	9
MeSH	5
NLM Catalog	2
PubMed	977
PubMed Central	2,034

Genes

Gene	380
GEO DataSets	2,073
GEO Profiles	9,450
HomoloGene	3
PopSet	2
UniGene	17

Genetics

ClinVar	29
dbGaP	0
dbSNP	6,488
dbVar	234
GTR	33
MedGen	4
OMIM	11

Proteins

Conserved Domains	0
Identical Protein Groups	14
Protein	565
Protein Clusters	0
Sparcle	1
Structure	44

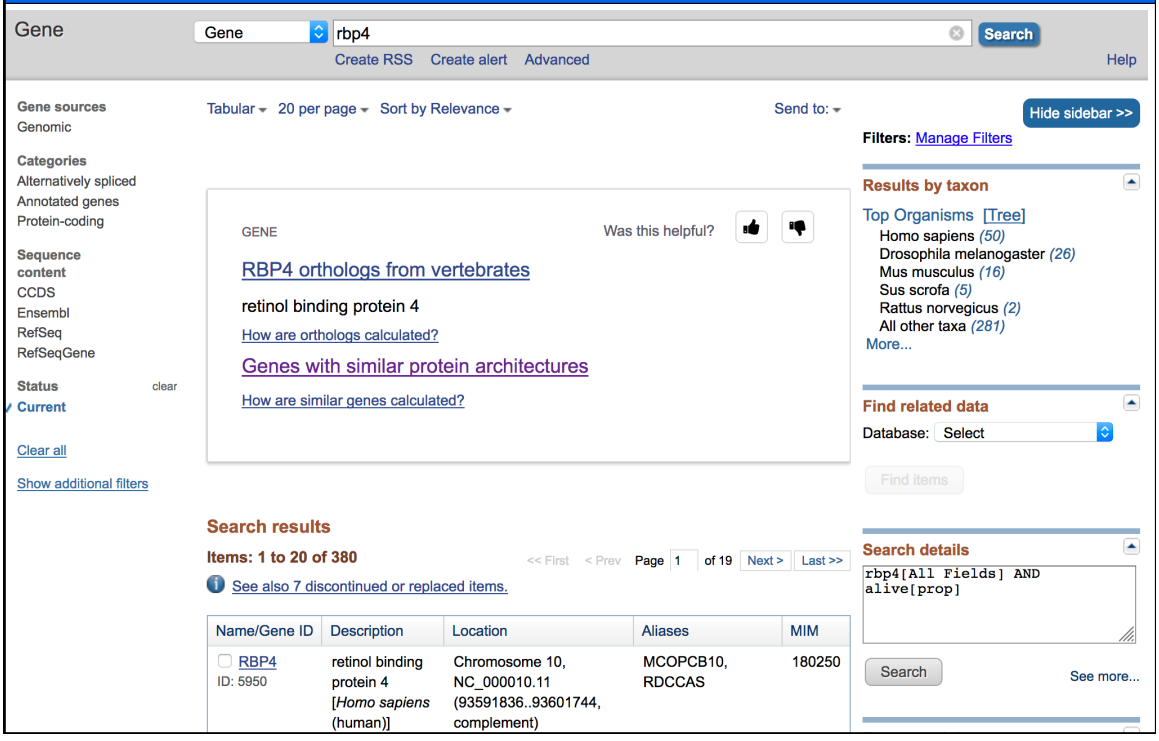
Genomes

Assembly	0
BioCollections	0
BioProject	2
BioSample	2,065
Genome	60
Nucleotide	1,068
Probe	153
SRA	3,756
Taxonomy	0

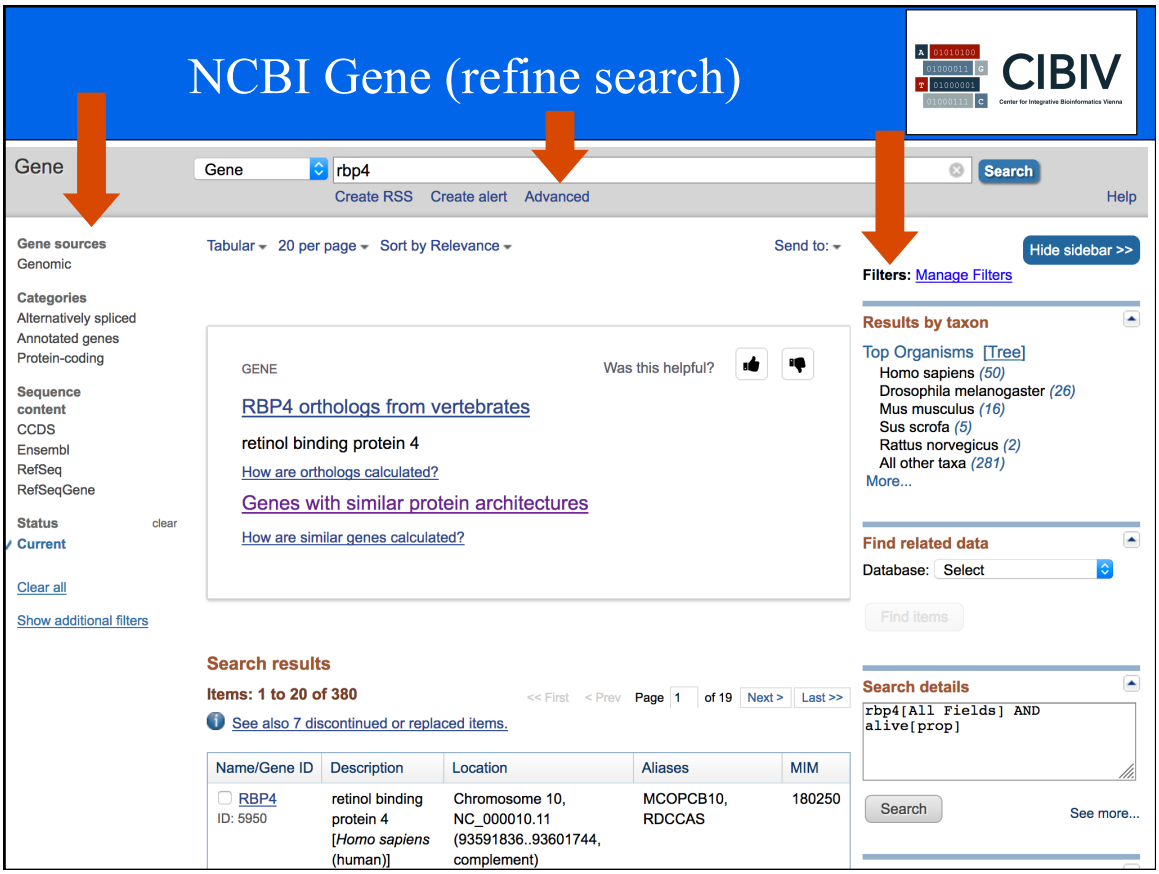
Chemicals

BioSystems	418
PubChem BioAssay	136
PubChem Compound	0
PubChem Substance	122

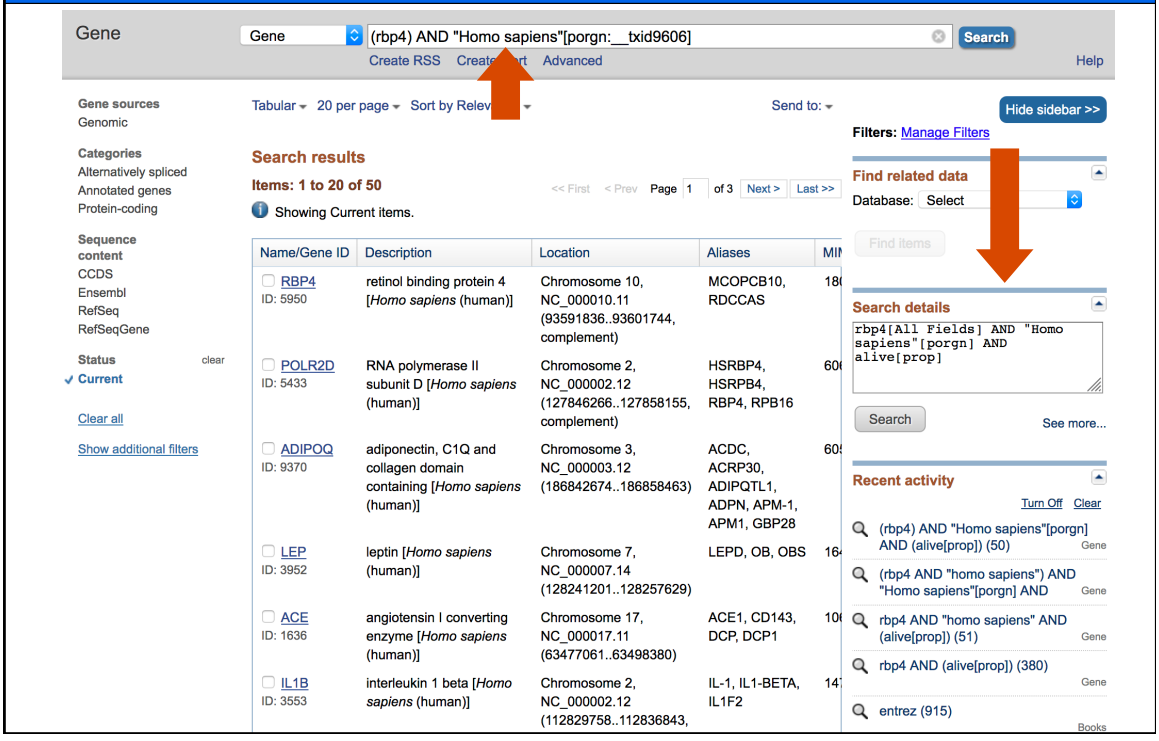
NCBI Gene



NCBI Gene (refine search)



Filters reduce the number of entries



Gene entry (summary)

RBP4 retinol binding protein 4 [*Homo sapiens* (human)]

Gene ID: 5950, updated on 28-May-2019

Summary

Official Symbol	RBP4 provided by HGNC
Official Full Name	retinol binding protein 4 provided by HGNC
Primary source	HGNC:HGNC:9922
See related	Ensembl:ENSG00000138207 MIM:180250
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	RDCCAS; MCOPCB10
Summary	This protein belongs to the lipocalin family and is the specific carrier for retinol (vitamin A alcohol) in the blood. It delivers retinol from the liver stores to the peripheral tissues. In plasma, the RBP-retinol complex interacts with transthyretin which prevents its loss by filtration through the kidney glomeruli. A deficiency of vitamin A blocks secretion of the binding protein posttranslationally and results in defective delivery and supply to the epidermal cells. [provided by RefSeq, Jul 2008]
Expression	Biased expression in liver (RPKM 2545.8) and fat (RPKM 387.7) See more
Orthologs	mouse all

Gene entry (genomic region)

Genomic regions, transcripts, and products

Go to reference sequence details

Genomic Sequence: NC_000010.11 Chromosome 10 Reference GRCh38.p12 Primary Assembly

Go to nucleotide: Graphics FASTA GenBank

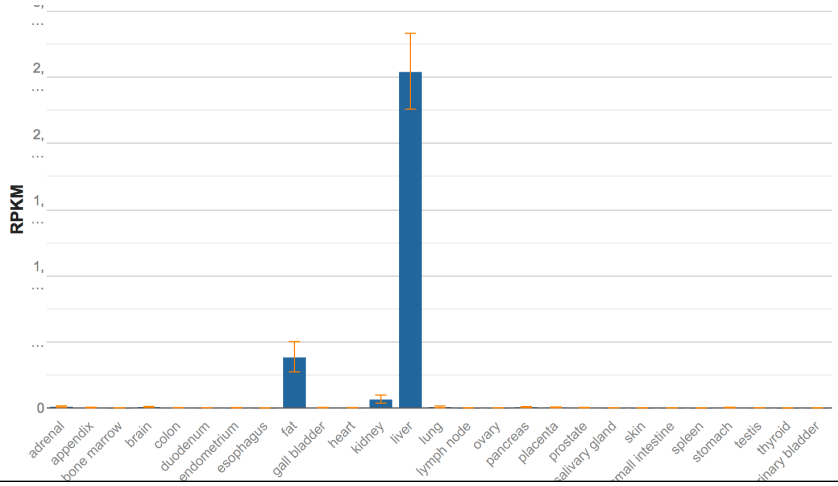
Gene entry (expression)

Expression

See details

HPA RNA-seq normal tissues

- Project title: HPA RNA-seq normal tissues
- Description: RNA-seq was performed of tissue samples from 95 human individuals representing 27 different tissues in order to determine tissue-specificity of all protein-coding genes
- BioProject: PRJEB4337
- Publication: PMID 24309898
- Analysis date: Wed Apr 4 07:08:55 2018



Gene entry (overview)

RBP4 retinol binding protein 4 [*Homo sapiens* (human)]

Gene ID: 5950, updated on 28-May-2019

- **Summary**
- **Genomic context**
- **Genomic regions, transcripts, and products**
- **Expression**
- **Bibliography**
- **Phenotypes**
- **Variation**
- **Pathways from BioSystems**
- **Interactions**
- **General gene information**
- **General protein information**
- **NCBI Reference Sequences (RefSeq)**
- **Related sequences**
- **Additional links**

Gene entry (protein via RefSeq)

mRNA and Protein(s)

1. NM_001323517.1 → NP_001310446.1 **retinol-binding protein 4 isoform a precursor**

Status: REVIEWED

Description	Transcript Variant: This variant (2) and variant 1 both encode isoform a.
Source sequence(s)	AL356214
Consensus CDS	CCDS31249.1
UniProtKB/Swiss-Prot	P02753
Related	ENSP00000360522.1, ENST00000371467.5

Conserved Domains (1) summary

pfam00061 Location:39 → 177	Lipocalin; Lipocalin / cytosolic fatty-acid binding protein family

2. NM_001323518.1 → NP_001310447.1 **retinol-binding protein 4 isoform b**

Status: REVIEWED

Source sequence(s)	AL356214, BC020633, BG565176, BI712834
Consensus CDS	CCDS81488.1
UniProtKB/Swiss-Prot	P02753
UniProtKB/TrEMBL	Q5VY30
Related	ENSP00000360524.2, ENST00000371469.2

Conserved Domains (1) summary

pfam00061 Location:37 → 175	Lipocalin; Lipocalin / cytosolic fatty-acid binding protein family

3. NM_006744.4 → NP_006735.2 **retinol-binding protein 4 isoform a precursor**

See identical proteins and their annotated locations for NP_006735.2

Status: REVIEWED

Description	Transcript Variant: This variant (1) and variant 2 both encode isoform a.
Source sequence(s)	AL356214, BC020633, BX495987, X00129
Consensus CDS	CCDS31249.1
UniProtKB/Swiss-Prot	P02753
Related	ENSP00000360519.3, ENST00000371464.8

Conserved Domains (1) summary

pfam00061 Location:39 → 177	Lipocalin; Lipocalin / cytosolic fatty-acid binding protein family

Protein entry (top)

Protein | Protein | Search | Advanced | Help

GenPept | Send to:

retinol-binding protein 4 isoform a precursor [Homo sapiens]

NCBI Reference Sequence: NP_001310446.1

Identical Proteins | FASTA | Graphics

Go to:

```
LOCUS       NP_001310446              201 aa            linear   PRI 28-MAY-2019
DEFINITION  retinol-binding protein 4 isoform a precursor [Homo sapiens].
ACCESSION   NP_001310446
VERSION     NP_001310446.1
DBSOURCE    REFSEQ: accession NM_001323517.1
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  (residues 1 to 201)
  AUTHORS   Aref-Eshghi E, Hurley O, Sun G, Simms A, Godwin M, Duke P, Araee M,
            Mahdavian M and Asghari S.
  TITLE     Genetic associations in community context: a mixed model approach
            identifies a functional variant in the RBP4 gene associated with
            HDL-C dyslipidemia
  JOURNAL   BMC Med. Genet. 19 (1), 205 (2018)
   PUBMED   30497399
  REMARK    GeneRIF: A mixed model approach identifies a functional variant in
            the RBP4 gene associated with HDL-C dyslipidemia the in
            Newfoundland and Labrador population.
            Publication Status: Online-Only
REFERENCE   2  (residues 1 to 201)
  AUTHORS   Wang H, Zhou P, Zou D, Liu Y, Lu X and Liu Z.
  TITLE     The role of retinol-binding protein 4 and its relationship with sex
            hormones in coronary artery disease
  JOURNAL   Biochem. Biophys. Res. Commun. 506 (1), 204-210 (2018)
   PUBMED   30342852
  REMARK    GeneRIF: RBP4 levels were significantly decreased and positively
```

Change region shown

Customize view

Analyze this sequence

Run BLAST

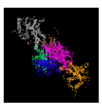
Identify Conserved Domains

Highlight Sequence Features

Find in this Sequence

Show in Genome Data Viewer

Protein 3D Structure



Complex of TTR and RBP4 and Oleic Acid
PDB: 2WQA
Source: Homo sapiens
Method: X-ray Diffraction
Resolution: 2.85 Å

See all 22 structures...

Articles about the RBP4 gene

Genetic associations in community context: a mixed r [BMC Med Genet. 2018]

The role of retinol-binding protein 4 and its rel [Biochem Biophys Res Commun. 2018]

Elevated RBP4 plasma levels were

Protein entry (bottom: features, sequence)

```
FEATURES             Location/Qualifiers
     source          1..201
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="10"
                     /map="10q23.33"
     Protein         1..201
                     /product="retinol-binding protein 4 isoform a precursor"
                     /note="retinol-binding protein 4, interstitial; RBP; PRBP;
                     plasma retinol-binding protein; retinol binding protein 4,
                     plasma"
                     /calculated_mol_wt=21072
     sig_peptide     1..18
                     /inference="COORDINATES: ab initio prediction:SignalP:4.0"
                     /calculated_mol_wt=1956
     mat_peptide     19..201
                     /product="retinol-binding protein 4 isoform a"
                     /calculated_mol_wt=21072
     Region          39..177
                     /region_name="Lipocalin"
                     /note="Lipocalin / cytosolic fatty-acid binding protein
                     family; pfam00061"
                     /db_xref="CDD:306552"
     Site            139
                     /site_type="methylation"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="Omega-N-methylarginine.
                     {ECO:0000250|UniProtKB:Q00724}; propagated from
                     UniProtKB/Swiss-Prot (P02753.3)"
     CDS             1..201
                     /gene="RBP4"
                     /gene_synonym="MCOPCB10; RDCCAS"
                     /coded_by="NM_001323517.1:171..776"
                     /note="isoform a precursor is encoded by transcript
                     variant 2"
                     /db_xref="CCDS:CCDS31249.1"
                     /db_xref="GeneID:5950"
                     /db_xref="HGNC:HGNC:9922"
                     /db_xref="MIM:180250"
ORIGIN
        1 mkwvwallll aalgsgraer dcrvssfrvk enfdkarfsg twyamakkdp eglflqdniv
       61 aefsvdetgq msatakgrvr llnnwdvcad mvgtftdted pakfkmkywg vasflqkgnd
      121 dhwivdtdyd tyavqyscrl lnldgtcads ysfvfsrdpn glppeaqkiv rqrqeelcla
      181 rqyrlivhng ycdgrsernl l
//
```

Protein entry (top)

Protein | Protein | Search | Advanced | Help

GenPept | Send to:

retinol-bind protein 4 isoform a precursor [Homo sapiens]

NCBI Reference Sequence: NP_001310446.1

Identical Proteins | FASTA | Graphics

Go to:

```
LOCUS       NP_001310446              201 aa            linear   PRI 28-MAY-2019
DEFINITION  retinol-binding protein 4 isoform a precursor [Homo sapiens].
ACCESSION   NP_001310446
VERSION     NP_001310446.1
DBSOURCE    REFSEQ: accession NM_001323517.1
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  (residues 1 to 201)
  AUTHORS   Aref-Eshghi E, Hurley O, Sun G, Simms A, Godwin M, Duke P, Araee M,
            Mahdavian M and Asghari S.
  TITLE     Genetic associations in community context: a mixed model approach
            identifies a functional variant in the RBP4 gene associated with
            HDL-C dyslipidemia
  JOURNAL   BMC Med. Genet. 19 (1), 205 (2018)
   PUBMED   30497399
  REMARK    GeneRIF: A mixed model approach identifies a functional variant in
            the RBP4 gene associated with HDL-C dyslipidemia the in
            Newfoundland and Labrador population.
            Publication Status: Online-Only
REFERENCE   2  (residues 1 to 201)
  AUTHORS   Wang H, Zhou P, Zou D, Liu Y, Lu X and Liu Z.
  TITLE     The role of retinol-binding protein 4 and its relationship with sex
            hormones in coronary artery disease
  JOURNAL   Biochem. Biophys. Res. Commun. 506 (1), 204-210 (2018)
   PUBMED   30342852
  REMARK    GeneRIF: RBP4 levels were significantly decreased and positively
```

Change region shown

Customize view

Analyze this sequence

Run BLAST

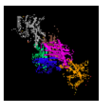
Identify Conserved Domains

Highlight Sequence Features

Find in this Sequence

Show in Genome Data Viewer

Protein 3D Structure



Complex of TTR and RBP4 and Oleic Acid

PDB: 2WQA

Source: Homo sapiens

Method: X-ray Diffraction

Resolution: 2.85 Å

See all 22 structures...

Articles about the RBP4 gene

Genetic associations in community context: a mixed r [BMC Med Genet. 2018]

The role of retinol-binding protein 4 and its rel [Biochem Biophys Res Commun. 2018]

Elevated RBP4 plasma levels were

Protein entry (sequence)

FASTA ▾ Send to: ▾

retinol-binding protein 4 isoform a precursor [Homo sapiens]

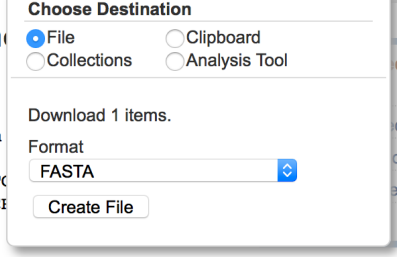
NCBI Reference Sequence: NP_001310446.1

GenPept Identical Proteins Graphics

```
>gi|1021087280|ref|NP_001310446.1| retinol-binding protein 4 isoform a precursor
[Homo sapiens]
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQ
MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRL
LNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
```

Protein entry (download, FASTA format)

FASTA ▾ Send to: ▾



retinol-binding protein 4 isoform a precursor [Hom

NCBI Reference Sequence: NP_001310446.1

GenPept Identical Proteins Graphics

```
>gi|1021087280|ref|NP_001310446.1| retinol-binding protein 4 isoform
[Homo sapiens]
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETG
MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCR
LNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
```

Protein 3D Structu

FASTA format

A sequence in FASTA format

- begins with ">" and a single-line description,
- followed by lines of sequence data.

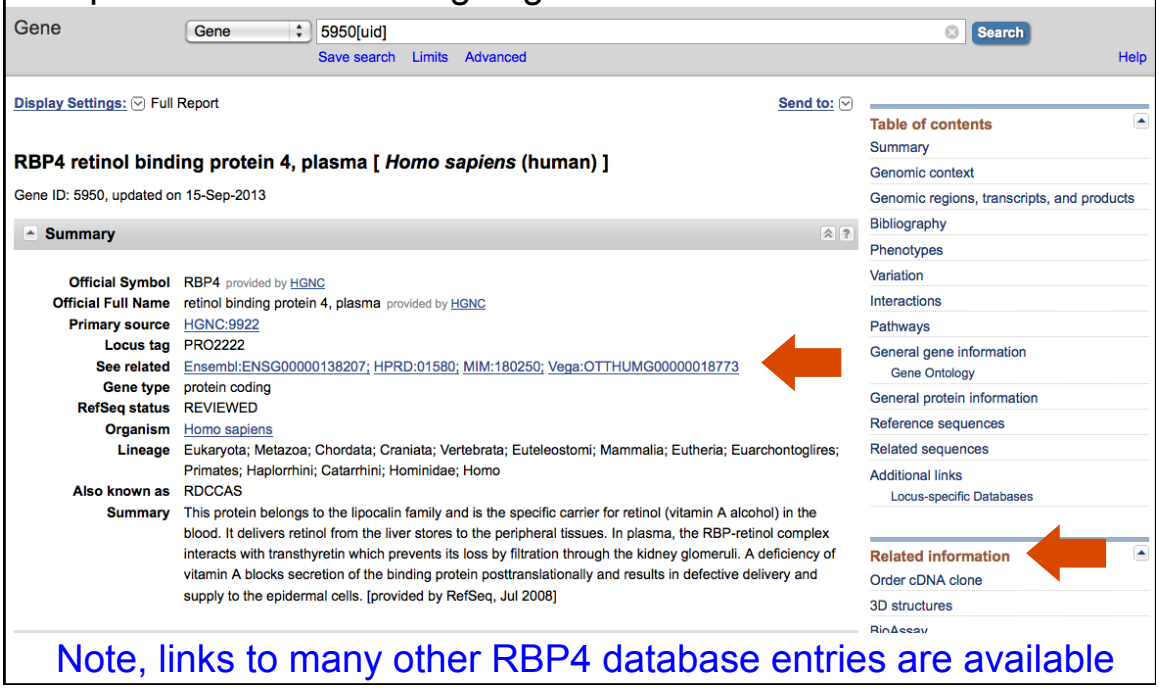
It is recommended that all lines of text be shorter than 80 characters in length.

An example:

```
>P02753
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDP
EGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTED
PAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADS
YSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNL
L
```

Gene entry (summary)

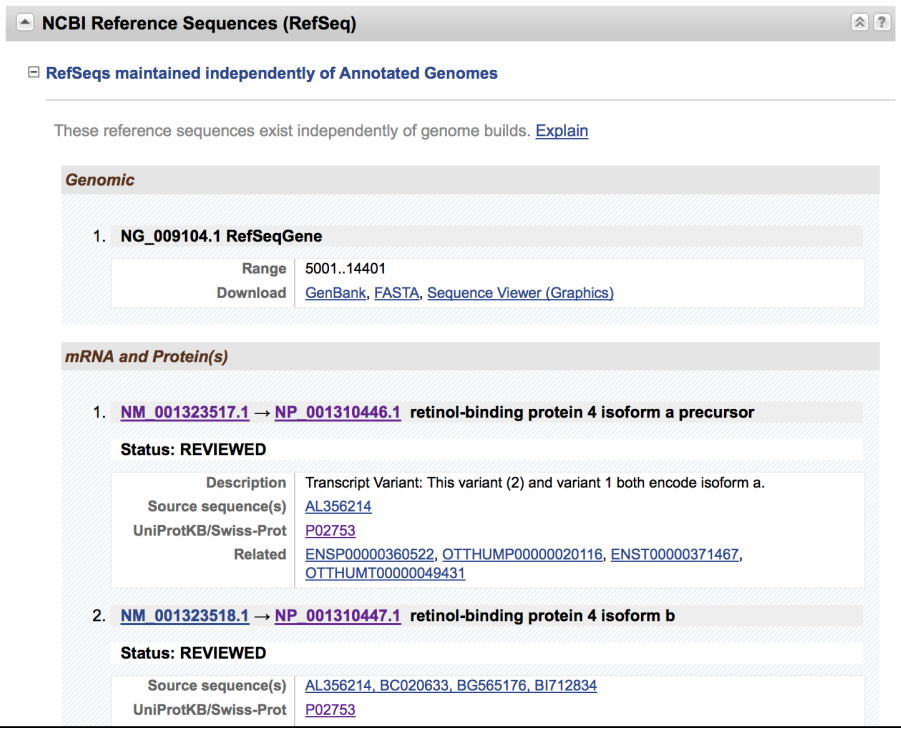
http://www.ncbi.nlm.nih.gov/gene/5950



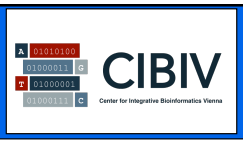
Note, links to many other RBP4 database entries are available

Gene entry (links from genomic region)

Gene entry (links to related sequences)



Gene entry (links in all categories)



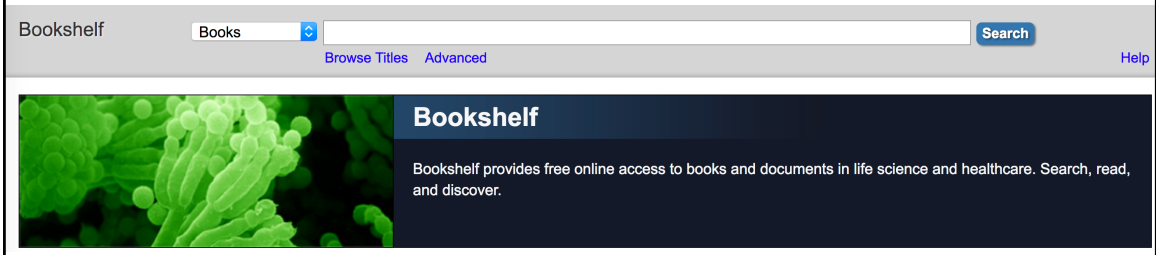
RBP4 retinol binding protein 4 [*Homo sapiens* (human)]

Gene ID: 5950, updated on 28-May-2019

- **Summary**
- **Genomic context**
- **Genomic regions, transcripts, and products**
- **Expression**
- **Bibliography**
- **Phenotypes**
- **Variation**
- **Pathways from BioSystems**
- **Interactions**
- **General gene information**
- **General protein information**
- **NCBI Reference Sequences (RefSeq)**
- **Related sequences**
- **Additional links**

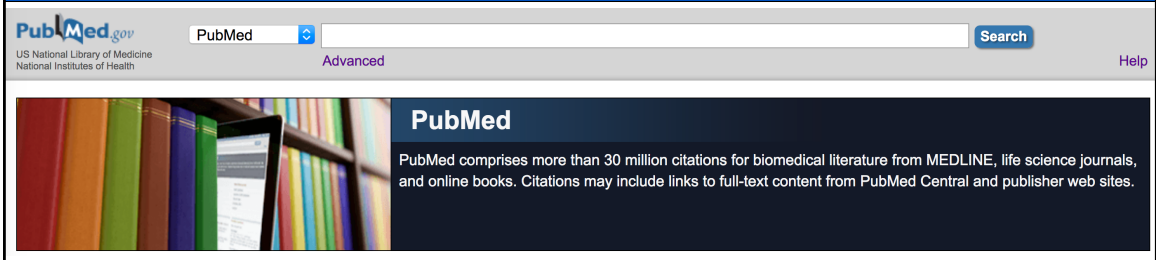
Access to Biomedical Literature

NCBI Online Books



NCBI Bookshelf provides free online access to books and documents in life science and healthcare.

PubMed to find literature information



PubMed is the NCBI gateway to MEDLINE.

- MEDLINE contains bibliographic citations
- and author abstracts from over 5,000 journals
- published in the United States and in 70 foreign
- countries.

- It has over 25 million records dating back to the 1960ies.

MeSH indexing in PubMed

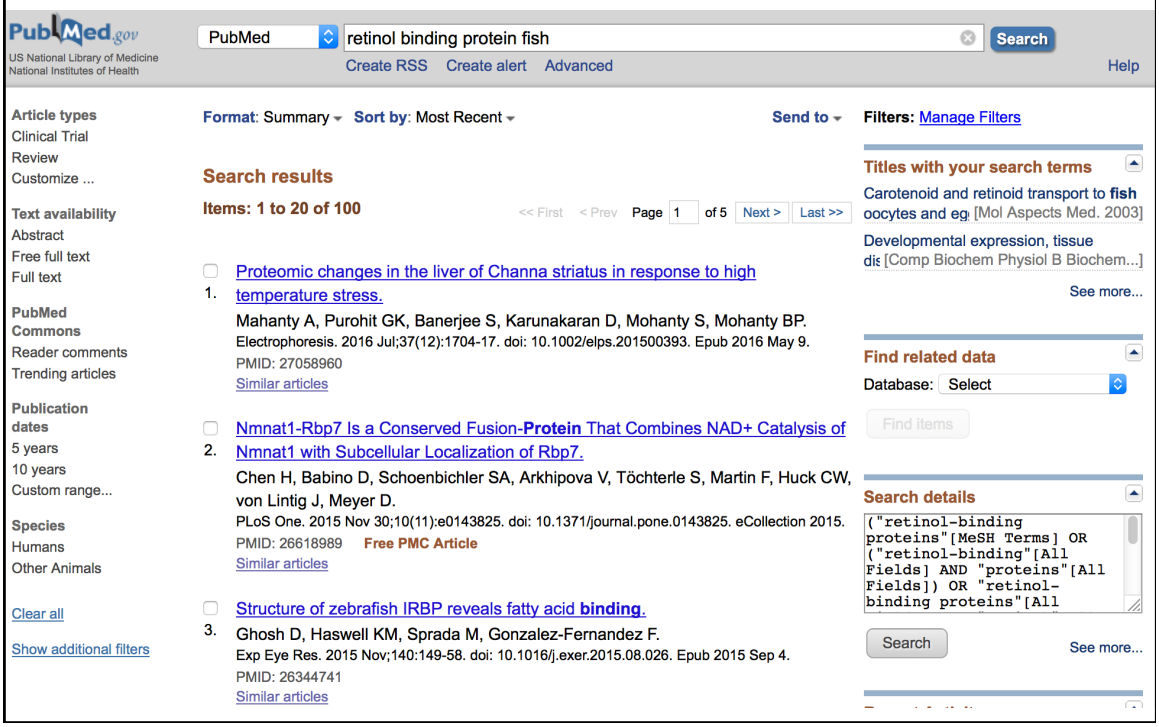
MeSH is the acronym for "Medical Subject Headings."

MeSH is the list of the vocabulary terms used for subject analysis of biomedical literature at NLM.

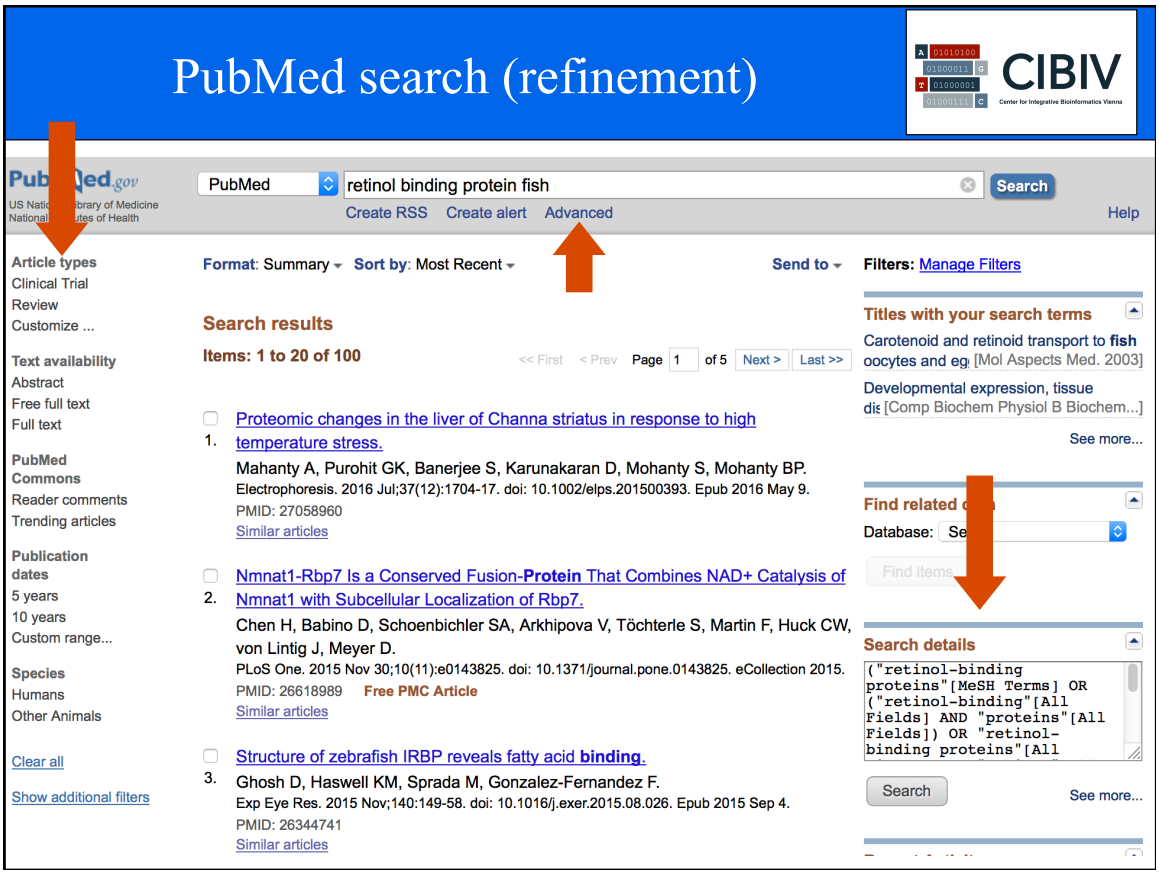
MeSH vocabulary is used for indexing journal articles for MEDLINE.

The MeSH controlled vocabulary imposes uniformity and consistency to the indexing of biomedical literature.

PubMed search



PubMed search (refinement)



PubMed entry

PubMed search strategies

Try the tutorials ("How To" and "help" at top)

Use boolean queries (capitalize AND, OR, NOT)

```
lipocalin AND disease
```

Utilize search fields (author [au], 1st author [1au], date of publication [dp], title [ti], abstract [ab], title+abstract [tiab], etc):

```
retinol binding protein[TIAB] AND chen[1AU]
```

Try using "Advanced"

Try "Similar articles" for similar articles or "LinkOut" to Entrez information and external resources

Tools for Sequence Analysis

NCBI Tools for Sequence Analysis

BLAST: The Basic Local Alignment Search Tool (BLAST) for comparing gene and protein sequences against others in public databases, comes in several types including PSI-BLAST, PHI-BLAST, and BLAST 2 sequences.

NCBI Tools for Sequence Analysis

Conserved Domain Database (CDD) A collection of sequence alignments and profiles representing protein domains conserved in molecular evolution. The CD Search Service can be used to search CDD.

NCBI Tools for Sequence Analysis

NCBI Gene Find information on sequence analyses for a particular gene and organism.

NCBI Protein Same, but protein-centered.

NCBI Genome Find information about genomes and genome projects.

NCBI Tools for Sequence Analysis

Gene Expression Omnibus (GEO) GEO provides several tools to assist with the visualization and exploration of curated GEO data.

ORF Finder A graphical analysis tool that finds all open reading frames of a selected minimum size in a user's sequence or in a sequence already in the database.

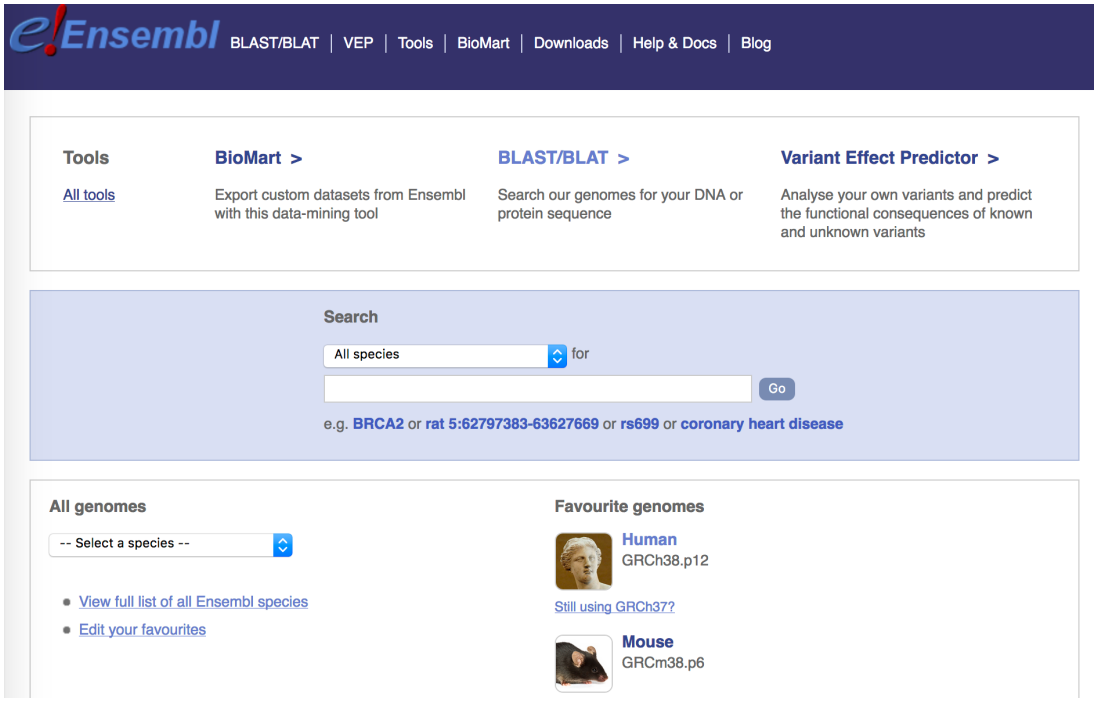
NCBI Tools for Sequence Analysis

Trace Archive Developed to store the raw sequence data underlying sequences generated by various genome projects.

Sequence Read Archive SRA store large amounts of short read data of next-generation sequencing runs generated by various genome projects.

VecScreen A tool for identifying segments of a nucleic acid sequence that may be of vector, linker, or adapter origin before using Tools for Sequence Analysis or submission.

Genomes via Ensembl



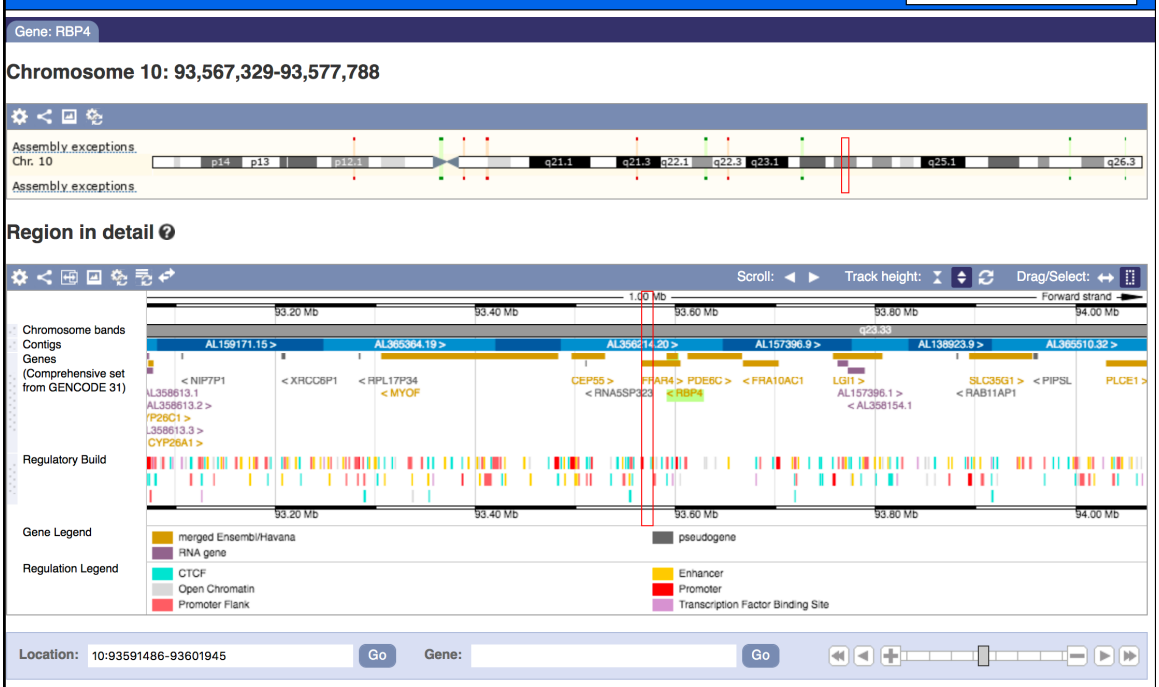
Genomes via Ensembl

- Ensembl is a genome annotation system, developed jointly by the EBI and the Wellcome Trust Sanger Institute.
- Originally used for the annotation, analysis and display of vertebrate genomes since 2000.
- Since 2009, complemented by the creation of five new sites, for bacteria, protists, fungi, plants and invertebrate metazoa.
- A single collection of interfaces for accessing and comparing genome-scale data from species of scientific interest from across the taxonomy.
- integrative power of Ensembl tools for comparative analysis, data mining and visualisation across genomes.

Tools at Ensembl/EnsemblGenomes

- **Searching Ensembl** – search for genes, keywords etc and have them displayed in the genomic context.
- **BLAST/BLAT** – sequence search against the genomes in Ensembl.
- **BioMart** – a data-mining tool to export custom datasets or conversion tables from Ensembl.
- **Variant Effect Predictor (VEP)** – determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions.
- **Downloads** – Ensembl provides means to download the whole genomic data, i.e. genomic/transcript/peptide sequences (FASTA format), feature annotation files (GFF+GTF format), etc.

Genome Viewing at Ensembl



Genomic context of the RBP4 gene in the human genome.

Genome Viewing – more details