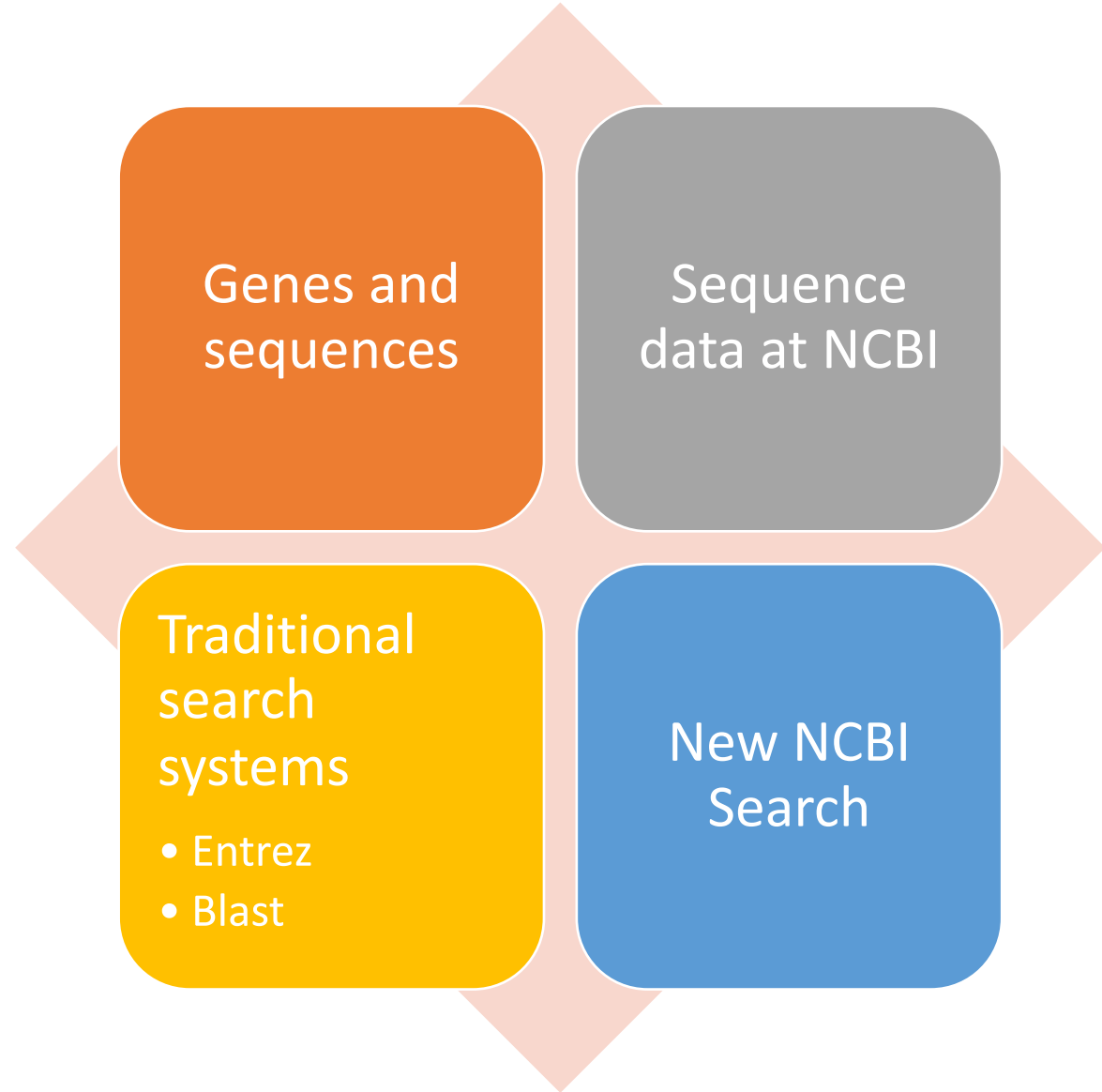




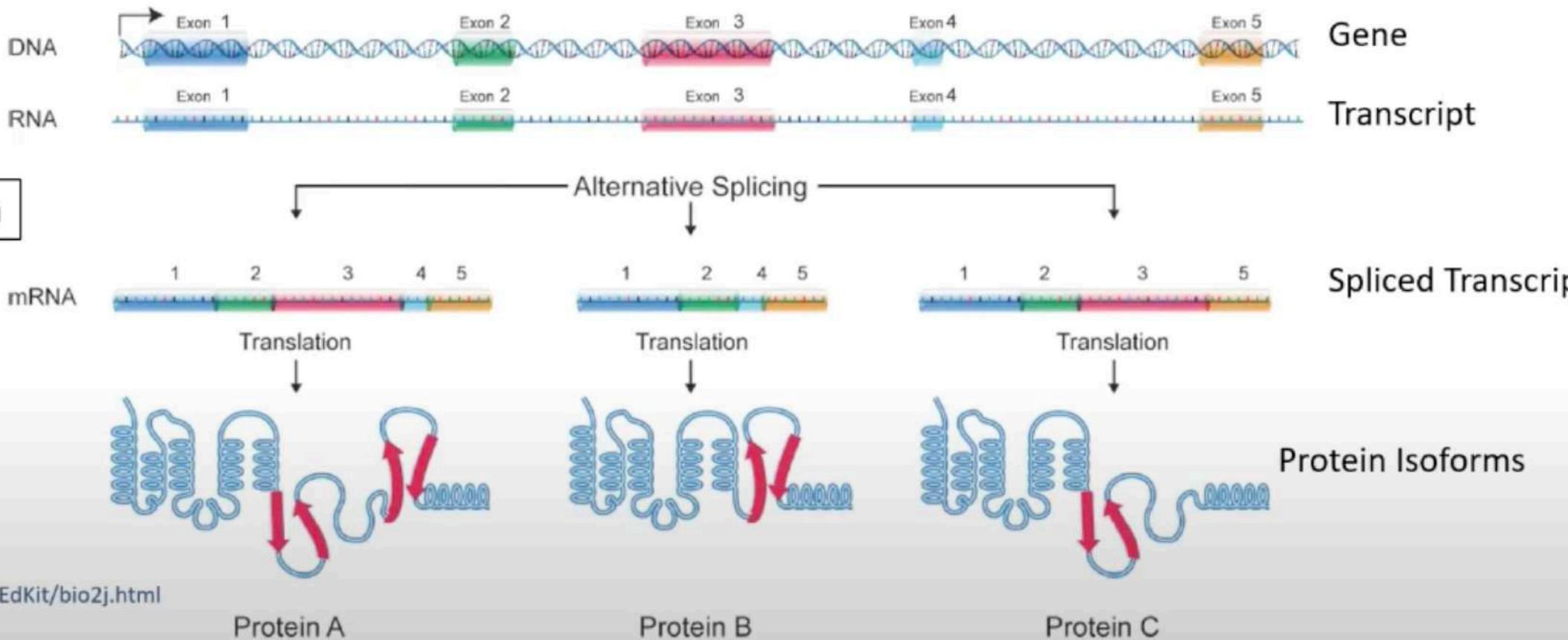
Introduction to NCBI

Assoc. Prof. Ilker BUYUK

Outline

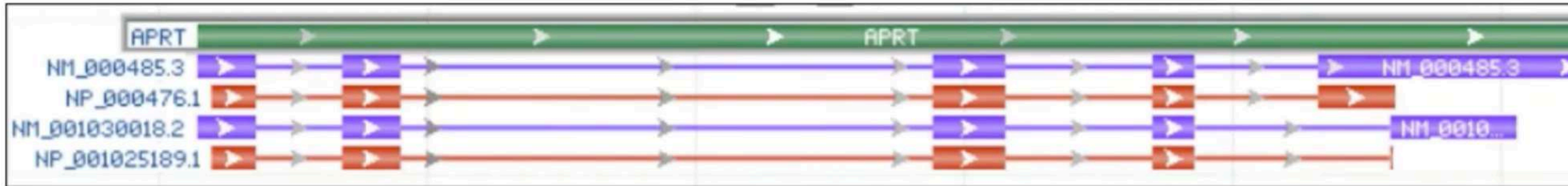


Transcription, splicing & translation



<http://www.genome.gov/EdKit/bio2j.html>

Gene and Transcripts at NCBI



Gene

chromosome 16: NC_000016.10 (88,809,339..88,811,928)

RefSeq Gene: NG_008013.1

Gene products

Two RefSeq Transcripts (splice forms) NM_000485.3, NM_001030018.2

Two RefSeq Proteins (isoforms) NP_000476.1, NP_001025189.1

Sources of sequence data at NCBI

- Submitted nucleotide sequences and corresponding proteins
 - International Sequence Database Colloration (INSDC)
 - Standard Sequences
 - **GenBank** -- US Sequence Database at NCBI
 - **European Nucleotide Archive** at EBI
 - **DNA Databank of Japan** at NIG
 - Next-Gen sequencing reads
 - Sequence Read Archive (SRA)
- High quality curated DNA and protein records
 - **NCBI Reference Sequences**
 - Swiss-Prot
 - Proteins from structures (PDB)

Example sequences for APRT

Human non-model RefSeqs (based on INSDC submissions)

Gene products

Transcripts

Proteins

NM_000485.3 → NP_000476.1

NM_001030018.2 → NP_001025189.1

Genomic

Primary Assembly

NC_000016.10 (88,809,339..88,811,928)

RefSeqGene

NG_008013.1 (5,007..7,596)

Rhesus Model RefSeqs (based on gene prediction and next-gen read data)

Transcripts

Proteins

XM_028841000.1 → XP_028696833.1

XM_001089867.4 → XP_001089867.2

XM_015126795.2 → XP_014982281.2

GenBank accessions

Transcripts

BC107151.2

BC106894.2

BC106895.2

Genomic

AY306126.1

AC092384.5

M16446.1

U04709.1

U09817.1

Other Protein

PDB

UniProtKB

1ORE_A P07741.2

4X44_A

6FCI_A

4X45_A

6FCH_A

6FD4_A



PDB sequence to structure links

adenine phosphoribosyltransferase isoform a [Homo sapiens]

NCBI Reference Sequence: NP_000476.1

[Identical Proteins](#) [FASTA](#) [Graphics](#)

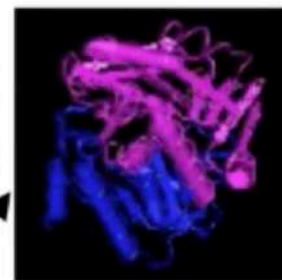
Go to:

LOCUS NP_000476 180 aa linear PRI 22-AUG-2019
DEFINITION adenine phosphoribosyltransferase isoform a [Homo sapiens].
ACCESSION NP_000476
VERSION NP_000476.1
DBSOURCE REFSEQ: accession [NM_000485.3](#)
KEYWORDS RefSeq; RefSeq Select.
SOURCE Homo sapiens (human)
ORGANISM [Homo sapiens](#)
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.

Customize view

Analyze this sequence

Protein 3D Structure



Crystal Stru
Human AP
type in com
PDB: 6HG
Source: H
sapiens
Method: X

Diffraction
Resolution: 1.55 Å

NCBI Structure

RefSeq

NP_000476

BLAST Similarity

6HGS_A

PDB protein chain

Web Sequence Databases

- Nucleotide www.ncbi.nlm.nih/nucore/
 - RNA, Genomic DNA sequences
 - 409 million (non-bulk) records
 - 348 million submitted (INSDC),
 - 61 million NCBI Reference Sequences
- Protein www.ncbi.nlm.nih/protein/
 - 774 million records
 - 619 million INSDC coding region translations
 - 153 million NCBI Reference Sequences
 - 2 million imported from outside
 - Includes 440 thousand from 3-D structures (PDB)

Does not including next-gen data (SRA)



Searching NCBI
data



NCBI Web Search Systems

The screenshot shows the NCBI homepage with a search bar at the top containing 'All Databases' and 'bacteria'. Below the search bar, there are navigation links for 'NCBI Home' and 'Resource List (A-Z)'. The main content area features a 'Welcome to NCBI' message and a 'Popular Resources' section with links to 'PubMed' and 'NCBI Home'. A sidebar on the left lists various biological domains and structures, including 'Genes & Expression', 'Genetics & Medicine', 'Genomes & Maps', 'Homology', 'Literature', 'Proteins', 'Sequence Analysis', 'Taxonomy', 'Training & Tutorials', and 'Variation'. The main content area is divided into three columns: 'Develop' (Use NCBI APIs and code libraries to build applications), 'Analyze' (Identify an NCBI tool for your data analysis task), and 'Research' (Explore NCBI research and collaborative projects). There is also a 'Nucleotide' section with links to 'Genome', 'SNP', 'Gene', 'Protein', and 'PubChem'. A 'NCBI News & Blog' section is visible at the bottom right.

Entrez text search system

40-plus integrated databases
Free text and database-specific fielded searches
The PubMed search engine

The screenshot shows the 'Standard Nucleotide BLAST' interface. At the top, there are tabs for 'blastn', 'blastp', 'blastx', 'tblastn', and 'tblastx'. The main section is titled 'Enter Query Sequence' and contains a text input field with the following sequence:

```
>gn|SRA|SRR5483149.1  
CCACAACTCCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATC  
CAGCCATGCCGCGTGTGTGAAGAAGGTCTTCGGATTGTAAAGCACTTAAAGTTGGGAGGA  
AGGGCAGTTACCTAATACGTAATTGTTTTGACGTACCAGACAATAAGCACCGGCTAAC  
TCTGTGCCAGCAGCCGCGTAATACAGAGGGTCAAGCGTTAATCGGAATTACTGGGCGT
```

 Below the input field, there are options to 'Or, upload file' and 'Job Title' (gn|SRA|SRR5483149.1). There is a checkbox for 'Align two or more sequences' and a 'Choose Search Set' dropdown menu. The 'Database' dropdown is open, showing options like 'Genomic plus Transcript', 'Human genomic plus transcript (Human G+T)', 'Mouse genomic plus transcript (Mouse G+T)', 'Other Databases', 'Nucleotide collection (nr/nt)', '16S ribosomal RNA sequences (Bacteria and Archaea)', 'Reference RNA sequences (refseq_rna)', and 'RefSeq Representative genomes (refseq_representative_genomes)'. The 'Organism' dropdown is also visible.

BLAST — Basic Local Alignment Search Tool

- Sequence similarity search tools
- Finds related nucleotide and protein sequences
 - Designed to find homologs
 - Used for other sequence analysis tasks

Search NCBI

all[sb]

X

Search

NCBI Databases

Results found in 37 databases for all[sb]

www.ncbi.nlm.nih.gov/search/

Literature

| | |
|----------------|------------|
| Bookshelf | 684,491 |
| MeSH | 277,799 |
| NLM Catalog | 1,590,993 |
| PubMed | 29,116,027 |
| PubMed Central | |

Genes

| | |
|--------------|-------------|
| Gene | 22,548,633 |
| GEO DataSets | 2,890,526 |
| GEO Profiles | 128,414,055 |
| HomoloGene | 141,268 |

Genetics

| | |
|---------|-------------|
| ClinVar | 472,230 |
| dbGaP | 1,217 |
| dbSNP | 672,043,185 |
| dbVar | 5,227,847 |

Proteins

| | |
|----------------------|-------------|
| Conserved Domain | |
| Identical Protein Gr | |
| Protein | 605,141,829 |
| Protein Clusters | 1,137,329 |
| Sparcle | 134,972 |
| Structure | 145,836 |

| | |
|------------|-------------|
| BioProject | 324,908 |
| BioSample | 9,515,814 |
| Genome | 40,781 |
| Nucleotide | 271,455,209 |
| Probe | 32,407,923 |
| SRA | 6,817,421 |
| Taxonomy | 2,150,126 |

| | |
|-------------------|-------------|
| PubChem Compound | 96,548,602 |
| PubChem Substance | 247,922,033 |

- 40 plus integrated literature and molecular databases
- Formal syntax for precise queries
- Links to related data and analyses

The Entrez System

- Where to begin??
- Typically people use only a few for direct searching
 - Start in central hubs
 - Gene
 - Assembly
 - Taxonomy
 - BioProject

BLAST results: finding related sequences

The image shows a BLAST search interface with two main panels. The left panel, titled "Sequences producing significant alignments", lists 11 sequences with checkboxes. The right panel provides a detailed view of the top match: **Ictalurus punctatus adenine phosphoribosyltransferase (aprt), mRNA** (Sequence ID: [NM_001200489.2](#), Length: 969, Number of Matches: 1). It displays sequence alignment statistics and a detailed alignment of the query sequence (103-336) with the subject sequence (325-480). The alignment shows a high degree of identity, with several amino acid differences highlighted in pink.

Sequences producing significant alignments

- [Bos taurus adenine phosphoribosyltransferase \(APRT\), mRNA](#)
- [Rattus norvegicus adenine phosphoribosyltransferase \(Aert\), mRNA](#)
- [Mus musculus adenine phosphoribosyltransferase \(Aert\), mRNA](#)
- [Gallus gallus adenine phosphoribosyltransferase \(APRT\), transcript variant 1, mRNA](#)
- [Gallus gallus adenine phosphoribosyltransferase \(APRT\), transcript variant 2, mRNA](#)
- [Taeniopygia guttata adenine phosphoribosyltransferase \(APRT\), mRNA](#)
- [PREDICTED: Sclerogages formosus adenine phosphoribosyltransferase \(aprt\), mRNA](#)
- [Ictalurus punctatus adenine phosphoribosyltransferase \(aprt\), mRNA](#)
- [Zea mays uncharacterized LOC100285146 \(LOC100285146\), mRNA](#)
- [Saccharomyces cerevisiae S288C adenine phosphoribosyltransferase APT2 \(APT2\), partial mRNA](#)
- [Zea mays uncharacterized LOC100282568 \(LOC100282568\), mRNA](#)

Ictalurus punctatus adenine phosphoribosyltransferase (aprt), mRNA
Sequence ID: [NM_001200489.2](#) Length: 969 Number of Matches: 1

Range 1: 325 to 480 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

| Score | Expect | Identities | Gaps | Strand |
|---------------|--------|--------------|-----------|-----------|
| 84.2 bits(92) | 3e-14 | 112/156(72%) | 0/156(0%) | Plus/Plus |

```
CDS:adenine phosphor 103  L E Y G K A E L E I Q K D A L E P G Q R
Query                 336  CTGGAGTACGGAAGGCTGAGCTGGAGATTCAGAAAGACGCCCTGGAGCCAGGACAGAGG
Sbjct                 325  CTCGAGTATGCCACGGCTGAGGTGGAGATTCAGGTGGACGCTGTGGACCCTGGACAGAAG
CDS:adenine phosphor 106  L E Y A T A E V E I Q V D A V D P G Q K
Query                 396  GTGGTCGTCGTGGATGATCTGCTGGCCACTGGTGAACCATGAACGCTGCCTGTGAGCTG
Sbjct                 385  GTTCTCGTCATCGATGACCTGCTGGCTACCGGAGGGACACTGTGTGCAGCGTGTGAGCTG
CDS:adenine phosphor 126  V L V I D D L L A T G G T L C A A C E L
Query                 456  CTGGGCCCGCCTGCAGGCTGAGGTCTGGAGTGCCTG 491
Sbjct                 445  ATGAAGAAGCAGAAGGCTCAGGTGCTGGGCTGCTTG 480
CDS:adenine phosphor 146  M K K Q K A Q V L G C L
```


A new search experience ...

GENOME ASSEMBLY Was this helpful?

Measles morbillivirus reference genome
Measles morbillivirus (Host: vertebrates, human)
ssRNA(-)
RefSeq assembly: GCF_000854845.1
RefSeq genomic segments (1) RefSeq Proteins (8) PubMed (98)

Reference Genomes

[BLAST](#) [Entrez Genome](#) [NCBI Virus](#) [Download](#)

Search results for: Horse

Genome assemblies

GENOME ASSEMBLY Was this helpful?

EquCab3.0
Equus caballus (horse)
University of Louisville (January 2018)
RefSeq assembly: GCF_002863925.1
PubMed (3)

[Genome Browser](#) [BLAST](#) [Download](#)

Search results for: Homo sapiens hemoglobin subunit beta

Genes & sequences

GENE Was this helpful?

HBB – hemoglobin subunit beta
Homo sapiens (human)
Processed peptides: LVV-hemorphin-7, Spinorphin
Also known as: CD113t-C, ECT6, beta-globin
GeneID: 3043
RefSeq transcripts (1) RefSeq proteins (1) RefSeqGene (2) PubMed (746)

[Orthologs](#) [Genome Browser](#) [BLAST](#) [Download](#)

Search results for: Homo sapiens mitochondrion

ORGANELLE Was this helpful?

Human mitochondrial reference genome
Homo sapiens
Included in the human reference assembly (GRCh38.p13)
RefSeq: NC_012920.1 Length: 16,569 bp circular
Gene (37) RefSeq protein (13) PopSet (598) PubMed (2)

[Genome Browser](#) [Primer-BLAST](#) [Download](#)

See all archival Homo sapiens complete mitochondrial genomes (48,239)

We know what you are looking for!

Search results for: BLAST

WEB RESOURCE

Web resources

BLAST - Basic Local Alignment Search Tool

A tool for comparing an amino acid or nucleotide sequence to an entire sequence library, identifying regions of high sequence similarity.

[blastn](#) [blastp](#) [blastx](#) [Primer-BLAST](#)

New NCBI Search makes finding genes and sequences easier

Search NCBI

APRT

Search results for: APRT

GENE Was this help

APRT – adenine phosphoribosyltransferase
Homo sapiens (human)
Also known as: AMP, APRTD
GeneID: 353

[RefSeq transcripts \(2\)](#) [RefSeq proteins \(2\)](#) [RefSeqGene \(1\)](#) [PubMed \(69\)](#)

Rapid access to genes and sequences

RefSeq Sequences

Showing 2 of 2 (by status, accession number)

| Transcript | nt | Protein | aa | Isoform | Status |
|--------------------------------|-----|--------------------------------|-----|---------|--|
| NM_000485.3 | 931 | NP_000476.1 | 180 | a | <input type="button" value="REFSEQ SELECT"/> |
| NM_001030018.2 | 667 | NP_001025189.1 | 134 | b | curated |

More options
Search the Nucleotide and Protein databases for more Homo sapiens APRT sequences

[RefSeq transcripts \(2\)](#) [RefSeq proteins \(2\)](#)
[Archival GenBank transcript sequences \(7\)](#) [Archival GenBank protein sequences \(9\)](#)

Orthologs / Similar genes partially replaces BLAST

New Search makes finding related genes easier

NCBI Orthologs [How was this calculated?](#)

0 items

Genes Literature

1 genes for: channel catfish (*Ictalurus punctatus*)

[Add to cart](#) [Protein alignment](#) [Download](#)

0 selected.

| Species | Gene | Architecture |
|--|--|--------------|
| <input type="checkbox"/> <i>Ictalurus punctatus</i> channel catfish | apt adenine phosphoribosyltransferase | |

| RefSeq transcripts (1) | RefSeq proteins (1) | Architecture |
|------------------------|---------------------|--------------|
| NM_001200489.2 | NP_001187418.1 | |

[Genome Browser](#) [InterPro](#)

Similar genes [How was this calculated?](#)

0 items

Genes

1 genes for: fruit fly (*Drosophila melanogaster*)

[Add to cart](#) [Protein alignment](#) [Download](#)

0 selected.

| Species | Gene | Ortholog Set | Architecture |
|--|--|--------------|--------------|
| <input type="checkbox"/> <i>Drosophila melanogaster</i> fruit fly | Apt Adenine phosphoribosyltransferase | | |

| RefSeq transcripts (3) | RefSeq proteins (3) | Architecture |
|------------------------|---------------------|--------------|
| NM_057289.3 | NP_476637.1 | |
| NM_001299979.1 | NP_001286908.1 | |
| NM_001299980.1 | NP_001286909.1 | |

[Genome Browser](#) [InterPro](#)

SEARCH THE TAXONOMY TREE

- Opisthokonta
 - animals
 - priapulids
 - arthropods
 - nematodes
 - segmented worms
 - molluscs
 - brachiopods
 - echinoderms
 - chordates
 - flatworms
 - cnidarians
 - placozoans
 - fungi

Orthologs of vertebrate genes
Similar genes for all eukaryotes

And comparing sequences easier

Cart

Protein alignment Download Remove all

5 selected.

| | Species | Gene | RNA | Protein | Remove |
|-------------------------------------|---|---|-----|---------|--------|
| <input checked="" type="checkbox"/> | <i>Homo sapiens</i> human | APRT adenine phosphoribosyltransferase | 2 | 2 | |
| <input checked="" type="checkbox"/> | <i>Drosophila melanogaster</i> fruit fly | Aprt Adenine phosphoribosyltransferase | 3 | 3 | |
| <input checked="" type="checkbox"/> | <i>Danio rerio</i> zebrafish | aprt adenine phosphoribosyltransferase | 1 | 1 | |
| <input checked="" type="checkbox"/> | <i>Caenorhabditis elegans</i> | T19B4.3 Adenine phosphoribosyltransferase | 1 | 1 | |
| <input checked="" type="checkbox"/> | <i>Schistosoma haematobium</i> | MS3_09305 Adenine phosphoribosyltransferase | 1 | 1 | |

Protein alignment

one sequence per gene (5)
 all sequences per gene (8)

Align

Live
demonstrations

Sequences for adenine ribosyltransferase

The known item search for APRT

- Finding related sequences
 - Orthologs
 - Similar Genes
 - Accessing COBALT
 - Downloading genes
- BLAST 2 Sequences
- Protein BLAST for bacterial adenine ribosyltransferase
- Structure of APRT protein in iCn3D



Thanks for listening