Nucleotide database

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.

In this Lab, NCBI's Nucleotide database is used to provide DNA and RNA sequences of any gene.

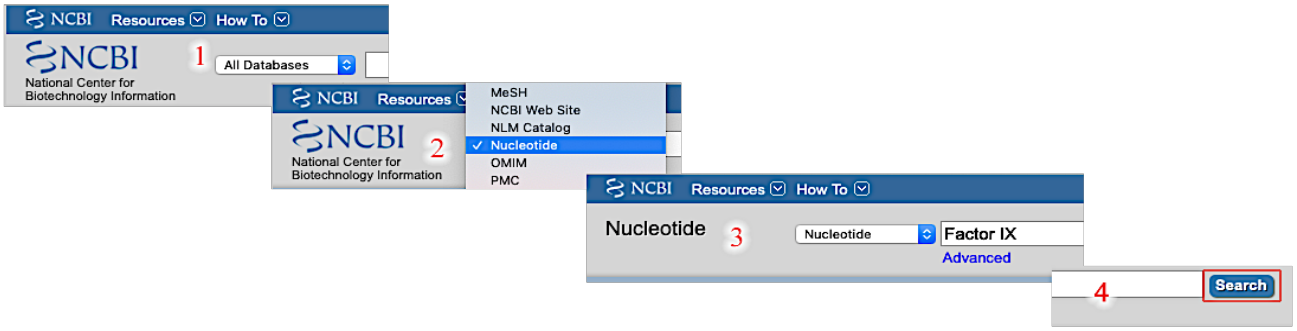
To exhibit how to use and search this database, a search on Factor IX gene was contucted for further guidance and explanation:

♦ How to search for a gene sequence?

e.g.1 *Retrieve the mRNA transcript sequence for the human* ***Factor IX*** *gene.*

1- Since a nucleotide sequence is required, your search should be on **nucleotide database.**

2- Search with gene name or symbol.

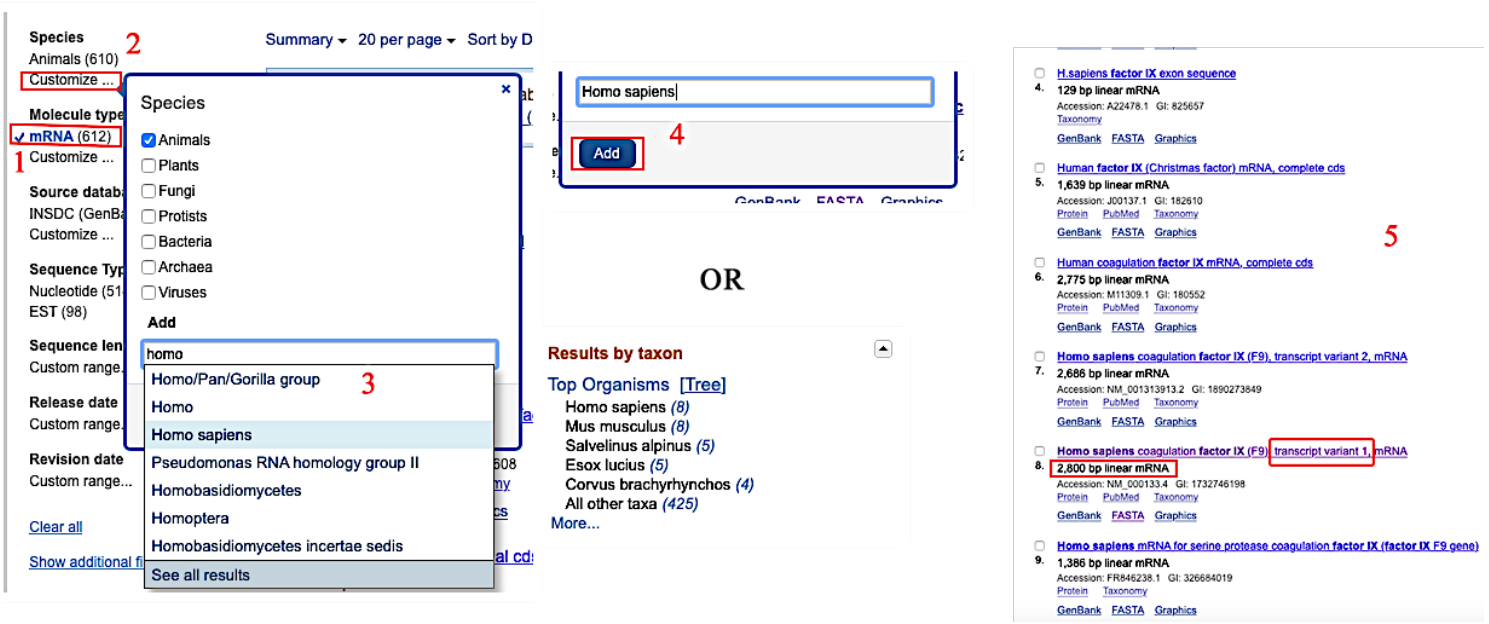


Tips to retrieve specific subsets of records.

1- Choose the **organism either by** left-hand side filter or Top Organisms list in the right-hand column of search results.

2- Choose the **molecule type** in the left-hand side filter.

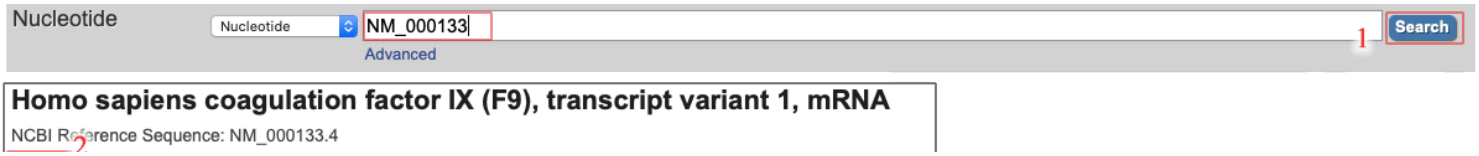
3- **Look for the longest sequence, this show the full sequence of the gene.**



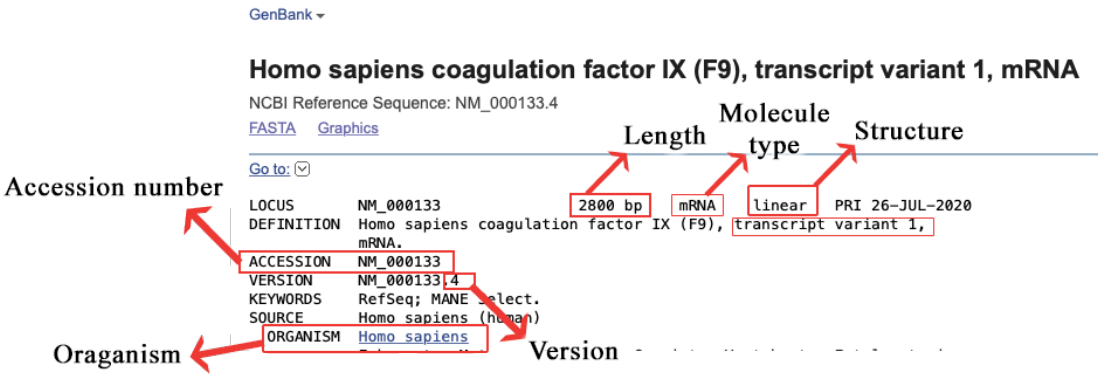

```
Transcript Variant: This variant (1) represents the longer
transcript and encodes the longer isoform (1).
```

OR → To get a certain sequence, a search by accession number is conducted:

An accession number in bioinformatics is a unique identifier given to a DNA or protein sequence record to allow for tracking of different versions of that sequence record and the associated sequence over time in a single data repository.



- **The search results after choosing your mRNA transcript is explained as follows:**



- **How to locate and highlight the biological function and the features of a nucleotide sequence.**
  - **Gene function is presented in "summary":**

```
Summary: This gene encodes vitamin K-dependent coagulation factor
IX that circulates in the blood as an inactive zymogen. This factor
is converted to an active form by factor XIa, which excises the
activation peptide and thus generates a heavy chain and a light
chain held together by one or more disulfide bonds. The role of
this activated factor IX in the blood coagulation cascade is to
activate factor X to its active form through interactions with Ca+2
ions, membrane phospholipids, and factor VIII. Alterations of this
gene, including point mutations, insertions and deletions, cause
factor IX deficiency, which is a recessive X-linked disorder, also
called hemophilia B or Christmas disease. Alternative splicing
results in multiple transcript variants encoding different isoforms
that may undergo similar proteolytic processing. [provided by
RefSeq, Sep 2015].
```

  - **These features include:** Origin of the gene (organism), molecule type, locus, gene length, Exon, **CDS** (Coding Sequences), and the **amino acid sequence.**

```
FEATURES             Location/Qualifiers
     source          1..2800
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="X"
                     /map="Xq27.1"
     gene            1..2800
                     /gene="F9"
                     /gene_synonym="F9 p22; FIX; HEMB; P19; PTC; THPH8"
                     /note="coagulation factor IX"
                     /db_xref="GeneID:2158"
                     /db_xref="HGNC:HGNC:3551"
                     /db_xref="MIM:300746"
     exon            1..114
                     /gene="F9"
1                    /gene_synonym="F9 p22; FIX; HEMB; P19; PTC; THPH8"
                     /inference="alignment:Splign:2.1.0"
     CDS             27..1412
                     /gene="F9"
                     /gene_synonym="F9 p22; FIX; HEMB; P19; PTC; THPH8"
                     /EC_number="3.4.21.22"
                     /note="isoform 1 preproprotein is encoded by transcript
                     variant 1; Christmas factor; plasma thromboplastic
                     component; factor 9; factor IX F9; plasma thromboplastin
                     component"
                     /codon_start=1
                     /product="coagulation factor IX isoform 1 preproprotein"
```

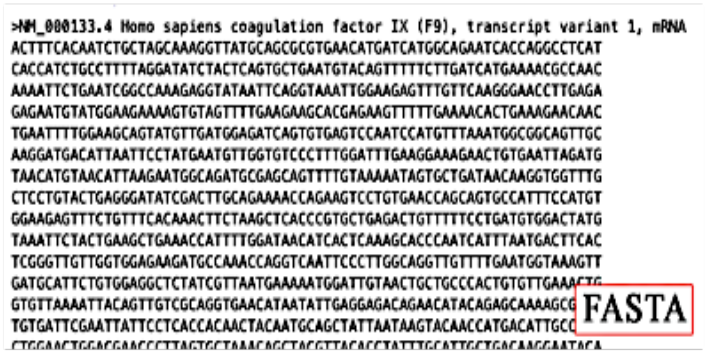
```
CDS                  27..1412
                     /gene="F9"
                     /gene_synonym="F9 p22; FIX; HEMB; P19; PTC; THPH8"
                     /EC_number="3.4.21.22"
                     /note="isoform 1 preproprotein is encoded by transcript
                     variant 1; Christmas factor; plasma thromboplastic
                     component; factor 9; factor IX F9; plasma thromboplastin
                     component"
                     /codon_start=1
                     /product="coagulation factor IX isoform 1 preproprotein"
                     /protein_id="NP_000124.1"
                     /db_xref="CCDS:CCDS14666.1"
                     /db_xref="GeneID:2158"
                     /db_xref="HGNC:HGNC:3551"
2                    /db_xref="MIM:300746"
                     /translation="MQRVNMIMAESPGLITICLLGYLLSAECTVFLDHENANKILNRP
                     KRYNSGKLEEFVQGNLERECMEEKCSFEEAREVFENTERTTEFWKQYVDGDQCESNPC
                     LNGGSCKDDINSYECWCPFGFEGKNCELDVTCNIKNGRCEQFCKNSADNKVVCSCTEG
                     YRLAENQKSCEPAVPFPCGRVSVSQTSKLTRAETVFPDVDYVNSTEAETILDNITQST
                     QSFNDFTRVVGGEDAKPGQFPWQVVLNGKVDAFCGGSIVNEKWIVTAAHCVETGVKIT
                     VVAGEHNIEETEHTEQKRNVIRIIPHHNYNAAINKYNHDIALLELDEPLVLNSYVTPI
                     CIADKEYTNIFLKFGSGYVSGWGRVFHKGRSALVLQYLRVPLVDRATCLRSTKFTIYN
                     NMFCAGFHEGGRDSCQGDSGGPHVTEVEGTSFLTGIISWGEECAMKGKYGIYTKVSRY
                     VNWIKEKTKLT"
```

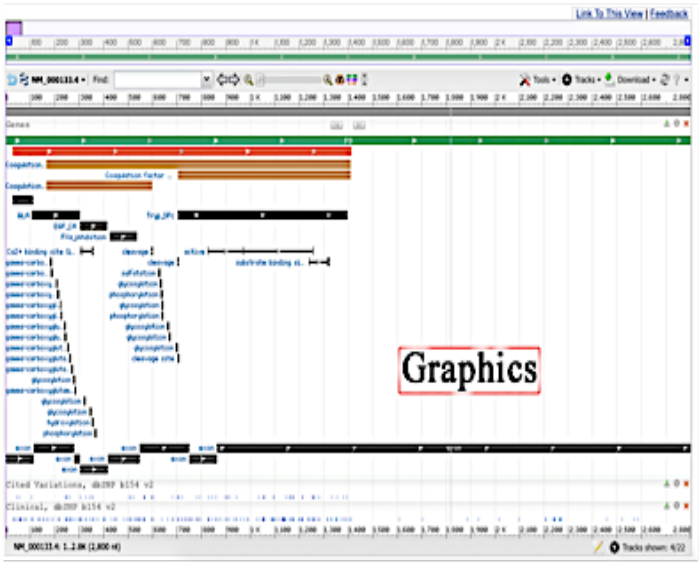
Amino acid sequence

♦ How to retrieve FASTA format and graphical view of the sequence.

- ✓ **FASTA format** is pronounced "**fast A**", and stands for "**FAST-All**", because it works with any alphabet. This is a text-based **format** for representing either nucleotide **sequences** or peptide **sequences**, in which base pairs or amino acids are represented using single-letter codes. A **sequence** in **FASTA format** begins with a single-line description, followed by lines of **sequence** data.

- ✓ **Both FASTA format and graphical view of the sequence** can be retrieved from the hyperlinks under the gene's name





Note:

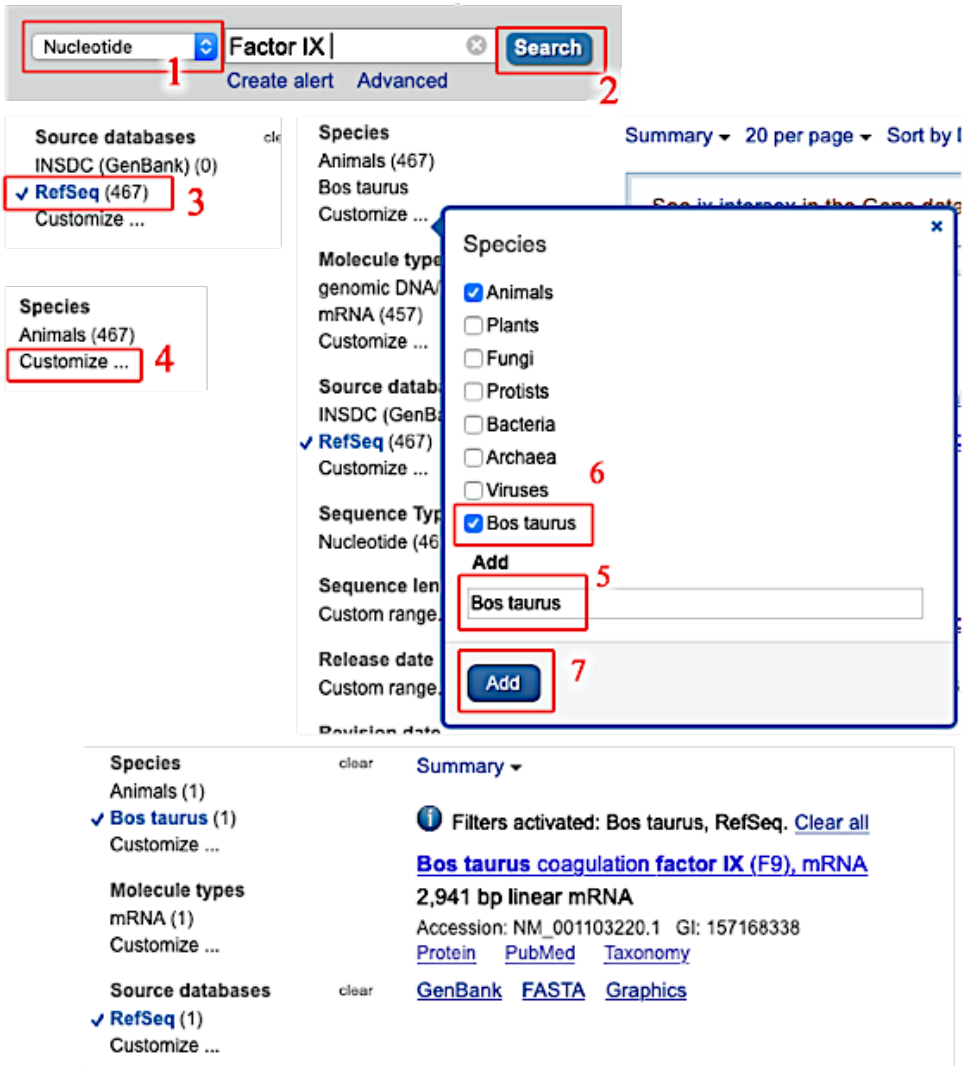
The graphical view is presented as lines with different colors:

- ✓ *Green → total gene*
- ✓ *Red → coding region*
- ✓ *Golden brown →mature peptide (after Post translational modification)*

♦ How to use the RefSeq database and search for a particular organism

RefSeq database is a comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.

e.g.2 *Search for Factor IX (Factor 9) mRNA in Bos taurus using RefSeq database.*



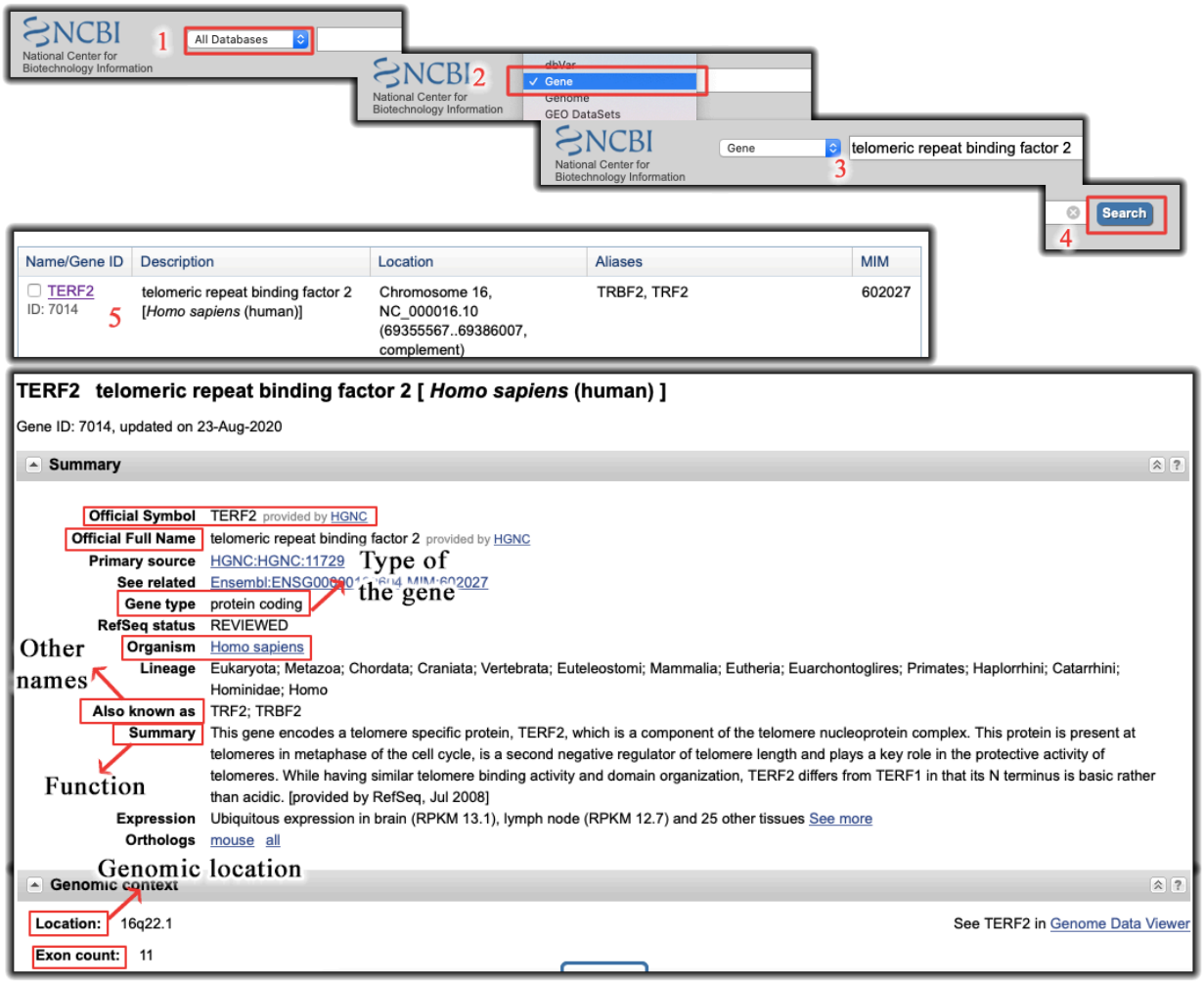
Gene database

Gene database 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.

♦ How to search for a gene in Gene database?

e.g 3 *Conduct a search on human* ***telomeric repeat binding factor 2 gene to provide gene-related information such as*** *official gene symbol, type, function, other names, genomic location of the gene*

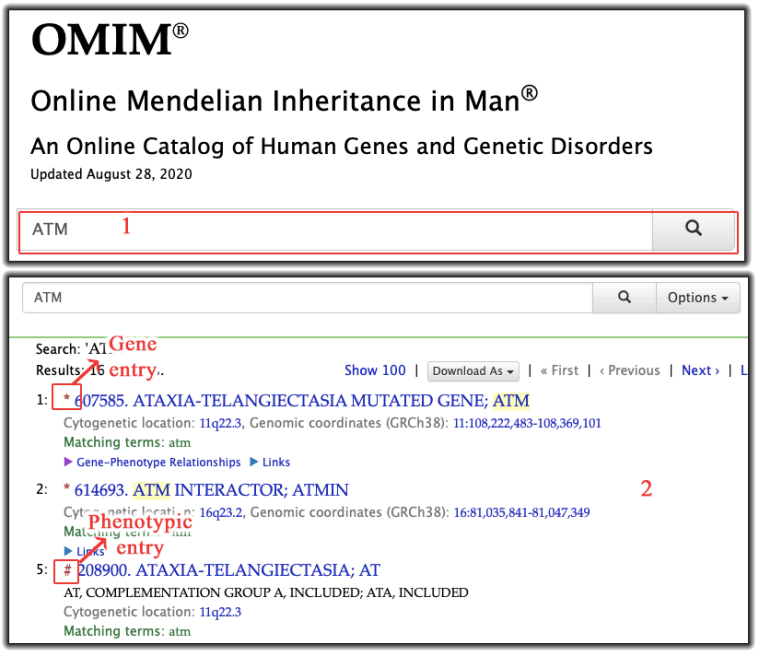
Since gene- related information are required → NCBI's gene database is used



Online Mendelian Inheritance in Man (OMIM) database

OMIM is an online comprehensive, authoritative ***catalog of human genes and genetic disorders*** that is freely available and updated daily. Its official home is omim.org.

- **How to search for a specific gene in OMIM?**



MIM Number Prefix:

- ☐ * gene with known sequence
- ☐ + gene with known sequence and phenotype
- ☐ # phenotype description, molecular basis known
- ☐ % mendelian phenotype or locus, molecular basis unknown

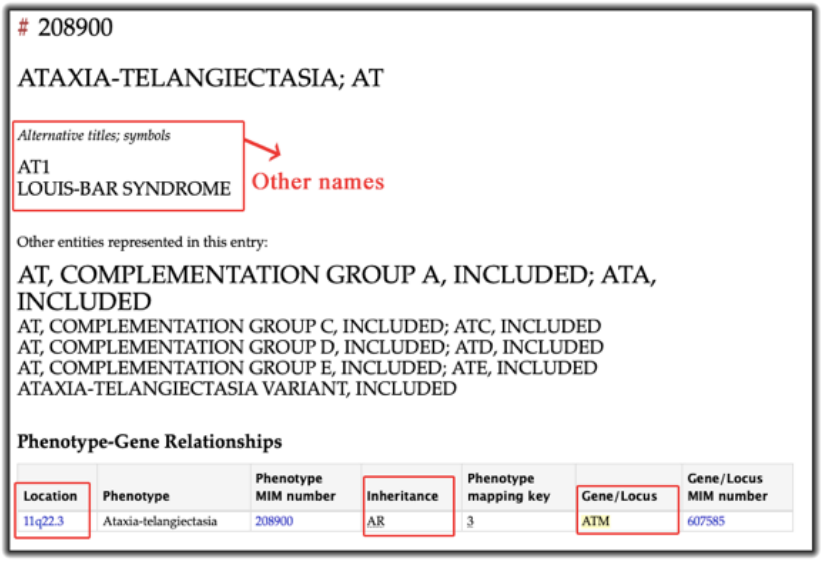
Note:

Prefixes are helpful during your search, so look for the following if:

- ✓ * → information on the gene
- ✓ # → diseases or phenotypes related to the gene

e.g 4 *Can you find any* ***disease*** *associated with* ***ATM?*** *Give a brief description of the disease*

Search for the gene → choose (#); since what you're interested in finding is a disease



208900

ATAXIA-TELANGIECTASIA; AT

Alternative titles; symbols

AT1
LOUIS-BAR SYNDROME

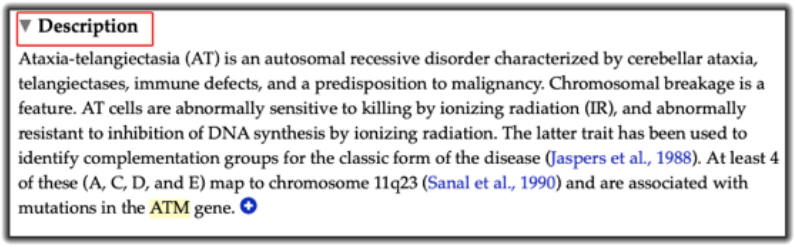
Other names

Other entities represented in this entry:

AT, COMPLEMENTATION GROUP A, INCLUDED; ATA, INCLUDED
AT, COMPLEMENTATION GROUP C, INCLUDED; ATC, INCLUDED
AT, COMPLEMENTATION GROUP D, INCLUDED; ATD, INCLUDED
AT, COMPLEMENTATION GROUP E, INCLUDED; ATE, INCLUDED
ATAXIA-TELANGIECTASIA VARIANT, INCLUDED

Phenotype-Gene Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key	Gene/Locus	Gene/Locus MIM number
11q22.3	Ataxia-telangiectasia	208900	AR	3	ATM	607585

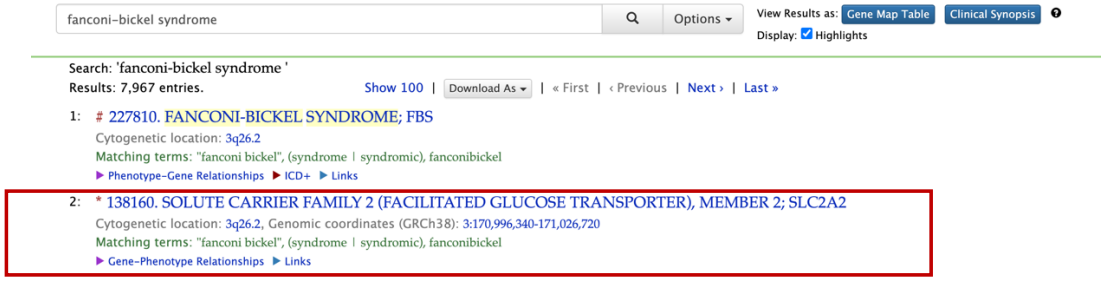


Description

Ataxia-telangiectasia (AT) is an autosomal recessive disorder characterized by cerebellar ataxia, telangiectases, immune defects, and a predisposition to malignancy. Chromosomal breakage is a feature. AT cells are abnormally sensitive to killing by ionizing radiation (IR), and abnormally resistant to inhibition of DNA synthesis by ionizing radiation. The latter trait has been used to identify complementation groups for the classic form of the disease (Jaspers et al., 1988). At least 4 of these (A, C, D, and E) map to chromosome 11q23 (Sanal et al., 1990) and are associated with mutations in the ATM gene.

e.g 5 *find the* ***gene*** *that associated with* ***fanconi-bickel syndrome***. What is the **function** of the gene?

Search for the disease→ choose (*); since what you're interested in finding is a gene

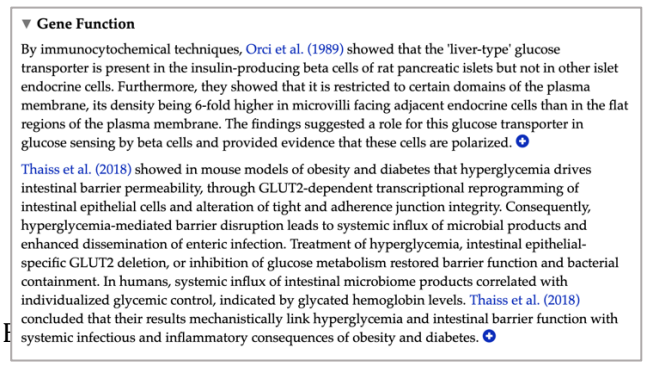


fanconi-bickel syndrome — Options — View Results as: Gene Map Table | Clinical Synopsis — Display: Highlights

Search: 'fanconi-bickel syndrome '
Results: 7,967 entries. Show 100 | Download As | « First | ‹ Previous | Next › | Last »

1: # 227810. FANCONI-BICKEL SYNDROME; FBS
Cytogenetic location: 3q26.2
Matching terms: "fanconi bickel", (syndrome | syndromic), fanconibickel
Phenotype-Gene Relationships | ICD+ | Links

2: * 138160. SOLUTE CARRIER FAMILY 2 (FACILITATED GLUCOSE TRANSPORTER), MEMBER 2; SLC2A2
Cytogenetic location: 3q26.2, Genomic coordinates (GRCh38): 3:170,996,340-171,026,720
Matching terms: "fanconi bickel", (syndrome | syndromic), fanconibickel
Gene-Phenotype Relationships | Links



Gene Function

By immunocytochemical techniques, Orci et al. (1989) showed that the 'liver-type' glucose transporter is present in the insulin-producing beta cells of rat pancreatic islets but not in other islet endocrine cells. Furthermore, they showed that it is restricted to certain domains of the plasma membrane, its density being 6-fold higher in microvilli facing adjacent endocrine cells than in the flat regions of the plasma membrane. The findings suggested a role for this glucose transporter in glucose sensing by beta cells and provided evidence that these cells are polarized.

Thaiss et al. (2018) showed in mouse models of obesity and diabetes that hyperglycemia drives intestinal barrier permeability, through GLUT2-dependent transcriptional reprogramming of intestinal epithelial cells and alteration of tight and adherence junction integrity. Consequently, hyperglycemia-mediated barrier disruption leads to systemic influx of microbial products and enhanced dissemination of enteric infection. Treatment of hyperglycemia, intestinal epithelial-specific GLUT2 deletion, or inhibition of glucose metabolism restored barrier function and bacterial containment. In humans, systemic influx of intestinal microbiome products correlated with individualized glycemic control, indicated by glycated hemoglobin levels. Thaiss et al. (2018) concluded that their results mechanistically link hyperglycemia and intestinal barrier function with systemic infectious and inflammatory consequences of obesity and diabetes.

♦ How to minimize your search results in OMIM.

e.g. 6 *can you find any records associated with ataxia telangiectasia? Make sure your strategy does not result in a number greater than 300 hits.*