



UNIVERSIDADE FEDERAL DA BAHIA
Instituto Multidisciplinar em Saúde
Campus Anísio Teixeira



Bancos de dados mais usados em Bioinformática

NCBI



O NCBI é dividido em vários bancos de dados específicos. Cada banco de dados armazena informações e apresenta links com outros bancos do próprio NCBI e bancos externos.



Entrez (GQuery) é o sistema de busca do NCBI que faz pesquisa em mais de 30 banco de dados. Esse sistema faz busca nas seguintes divisões do NCBI:

Literature
Health
Genome
Genes
Proteins
Chemicals



UNIVERSIDADE FEDERAL DA BAHIA
Instituto Multidisciplinar em Saúde
Campus Anísio Teixeira



NCBI Resources How To Sign In to NCBI

Search NCBI databases Help

Search

Literature	Genes
Books books and reports	EST expressed sequence tag sequences
MeSH ontology used for PubMed indexing	Gene collected information about gene loci
NLM Catalog books, journals and more in the NLM Collections	GEO DataSets functional genomics studies
PubMed scientific & medical abstracts/citations	GEO Profiles gene expression and molecular abundance profiles
PubMed Central full-text journal articles	HomoloGene homologous gene sets for selected organisms
Health	PopSet sequence sets from phylogenetic and population studies
ClinVar human variations of clinical significance	UniGene clusters of expressed transcripts
dbGaP genotype/phenotype interaction studies	Proteins
GTR genetic testing registry	Conserved Domains conserved protein domains
MedGen medical genetics literature and links	Protein protein sequences
OMIM online mendelian inheritance in man	Protein Clusters sequence similarity-based protein clusters
PubMed Health clinical effectiveness, disease and drug reports	Structure experimentally-determined biomolecular structures
Genomes	Chemicals
Assembly genome assembly information	BioSystems molecular pathways with links to genes, proteins and chemicals
BioProject biological projects providing data to NCBI	PubChem BioAssay bioactivity screening studies
BioSample descriptions of biological source materials	PubChem Compound chemical information with structures, information and links
Clone genomic and cDNA clones	PubChem Substance deposited substance and chemical information
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	

A busca no Entrez é realizada em todos os banco de dados ligados ao NCBI. O Entrez ainda aceita pesquisa específicas para cada banco ou pesquisas mais refinadas usando operadores.

Exemplo:

(CASP1) **AND** Homo Sapiens

O operador **AND** diz ao Entrez para retornar somente onde encontrar o termo CASP1 e Homo sapiens como organismo

(CASP1) **AND** Homo Sapiens[**Organism**]

O operador **AND** diz ao Entrez para retornar somente onde encontrar o termo CASP1 e Homo sapiens como organismo.

Em cada banco é possível montar uma busca refinada usando ferramenta de busca avançada. Vamos usar o banco gene para fazer essa busca avançada.



UNIVERSIDADE FEDERAL DA BAHIA
Instituto Multidisciplinar em Saúde
Campus Anísio Teixeira



Gene/NCBI

NCBI Resources How To Sign in to NCBI

Gene Gene Search Help

Advanced

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.

<https://www.ncbi.nlm.nih.gov/gene>

Figura. Página de entrada do Gene. Clicar em "Advanced"

Builder

Selecionar

"Gene name" e digitar CASP1

Usar o Operador **AND**

"Organism" e digitar Homo sapiens

Vejam que a busca já está sendo montada acima do "**Builder**"

Ao terminar a busca clicar em "**Search**".



UNIVERSIDADE FEDERAL DA BAHIA
Instituto Multidisciplinar em Saúde
Campus Anísio Teixeira



NCBI Resources How To Sign in to NCBI

Gene Home Help

Gene Advanced Search Builder

Showing Current items.

(CASP1[Gene Name]) AND Homo sapiens[Organism]

Edit Clear

Builder

Gene Name	CASP1	Show index list
AND Organism	Homo sapiens	Show index list
AND All Fields		Show index list

Search or Add to history

Agora experimentem montar diferentes buscas, usando outros operadores (AND/OR/NOT), para outros genes e outras características disponíveis para busca "Advanced".

Vamos buscar o Gene TLR9 (Toll Like Receptor 9) no banco Gene/NCBI para ver quais informações esse banco de dados retorna.

Official Symbol TLR9 provided by [HGNC](#)

Official Full Name toll like receptor 9 provided by [HGNC](#)

Primary source [HGNC:HGNC:15633](#)

See related [Ensembl:ENSG00000239732](#) [MIM:605474](#); [Vega:OTTHUMG00000158106](#)

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 CD289

Summary The protein encoded by this gene is a member of the Toll-like receptor (TLR) family which plays a fundamental role in pathogen recognition and activation of innate immunity. TLRs are highly conserved from *Drosophila* to humans and share structural and functional similarities. They recognize



UNIVERSIDADE FEDERAL DA BAHIA
Instituto Multidisciplinar em Saúde
Campus Anísio Teixeira



pathogen-associated molecular patterns (PAMPs) that are expressed on infectious agents, and mediate the production of cytokines necessary for the development of effective immunity. The various TLRs exhibit different patterns of expression. This gene is preferentially expressed in immune cell rich tissues, such as spleen, lymph node, bone marrow and peripheral blood leukocytes. Studies in mice and human indicate that this receptor mediates cellular response to unmethylated CpG dinucleotides in bacterial DNA to mount an innate immune response. [provided by RefSeq, Jul 2008]

Orthologs [mouse](#) [all](#)

Pathways from BioSystems

KEGG

REACTOME

WikiPathways

Homology

Homologs of the TLR9 gene: The TLR9 gene is conserved in chimpanzee, Rhesus monkey, dog, cow, mouse, rat, zebrafish, and frog.

Orthologs from Annotation Pipeline: 72 organisms have orthologs with human gene TLR9

Map Viewer (Mouse, Rat)

The Hierarchical Catalog of Orthologs

Gene Ontology

Function

Process

Component



UNIVERSIDADE FEDERAL DA BAHIA
Instituto Multidisciplinar em Saúde
Campus Anísio Teixeira



Nucleotide/NCBI

The screenshot shows the NCBI Nucleotide search page. At the top, there are navigation links for 'NCBI', 'Resources', and 'How To', along with a 'Sign in to NCBI' link. Below this is a search bar with 'Nucleotide' in the dropdown menu and a 'Search' button. A 'Help' link is also visible. The main content area features a dark background with the word 'Nucleotide' in white. To the left, there is a blurred image of DNA sequence data. To the right, a text box explains that the Nucleotide database is a collection of sequences from various sources like GenBank, RefSeq, TPA, and PDB, used for biomedical research.

<https://www.ncbi.nlm.nih.gov/nucleotide>

O banco Nucleotide armazena sequências e informações associadas a cada uma das sequências. A pesquisa nesse banco pode retornar várias sequências depositadas por diferentes pesquisadores.

Podemos selecionar as sequências desejadas e baixar para analisar mutações, diversidade, presença de regiões conservadas.

A busca pelo termo TLR9 no banco nucleotide retorna 3258 entradas (Busca realizada 26/01/17).

A descrição "**partial cds**" significa que a sequência do gene não está completa, o pesquisador que depositou não obteve a sequência completa e submeteu somente uma parte do gene. A descrição "**complete cds**" significa que foi depositado a sequência completa do gene. CDS significa a região codificante do gene.

Vamos observar a entrada 33 dessa busca.



Homo sapiens toll like receptor 9 (TLR9), mRNA

NCBI Reference Sequence: NM_017442.3

[FASTA](#) [Graphics](#)

Go to:

LOCUS NM_017442 3922 bp mRNA linear PRI 06-OCT-2016
DEFINITION Homo sapiens toll like receptor 9 (TLR9), mRNA.
ACCESSION NM_017442
VERSION NM_017442.3
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 3922)
AUTHORS Newman ZR, Young JM, Ingolia NT and Barton GM.
TITLE Differences in codon bias and GC content contribute to the balanced
expression of TLR7 and TLR9
JOURNAL Proc. Natl. Acad. Sci. U.S.A. 113 (10), E1362-E1371 (2016)
PUBMED [26903634](#)
REMARK GeneRIF: Data show that differences in codon bias limit Toll-like
receptor 7 (TLR7) expression relative to Toll-like receptor 9
(TLR9).

Figura. Informações iniciais sobre a entrada número 33 após a busca usando o termo TLR9.



UNIVERSIDADE FEDERAL DA BAHIA
Instituto Multidisciplinar em Saúde
Campus Anísio Teixeira



```
4091-3922          BC032713.1          2139-3384
FEATURES
source             Location/Qualifiers
                  1..3922
                  /organism="Homo sapiens"
                  /mol_type="mRNA"
                  /db_xref="taxon:9606"
                  /chromosome="3"
                  /map="3p21.2"
gene              1..3922
                  /gene="TLR9"
                  /gene_synonym="CD289"
                  /note="toll like receptor 9"
                  /db_xref="GeneID:54106"
                  /db_xref="HGNC:HGNC:15633"
                  /db_xref="HPRD:05685"
                  /db_xref="MIM:605474"
exon              1..637
                  /gene="TLR9"
                  /gene_synonym="CD289"
                  /inference="alignment:Splign:1.39.8"
misc feature      563..565
                  /gene="TLR9"
                  /gene_synonym="CD289"
                  /note="upstream in-frame stop codon"
CDS               635..3733
                  /gene="TLR9"
                  /gene_synonym="CD289"
                  /codon_start=1
                  /product="toll-like receptor 9 precursor"
                  /protein_id="NP_059138.1"
                  /db_xref="CCDS:CCDS2848.1"
                  /db_xref="GeneID:54106"
                  /db_xref="HGNC:HGNC:15633"
                  /db_xref="HPRD:05685"
                  /db_xref="MIM:605474"
                  /translation="MGFCRSALHPLSLLVQAIMLAMTLALGTLPAFLPCELQPHGLVN
                  CNWLFLKSVPHFSMAAPRGNVTSLSLSSNRIHHLHDSDFAHLP SLRHLN LKWNCPPVG
```

Figura. Informações sobre as características já descritas para essa sequência.

Observem que mesmo dentro do banco Nucleotide nós temos acesso a sequência de proteína codificada por esse gene. Dentro da "Feature"/"CDS"/"translation".

O formato GenBank guarda toda a informação sobre as proteínas depositadas no banco Nucleotide do NCBI.



Protein/NCBI

<https://www.ncbi.nlm.nih.gov/protein>

O banco Protein armazena sequências de proteínas e sequências que foram traduzidas automaticamente de sequências nucleotídicas codificadoras. O banco também traz informações associadas a cada uma das sequências. A pesquisa nesse banco pode retornar várias sequências depositadas por diferentes pesquisadores.

Podemos selecionar as sequências desejadas e baixar para analisar alterações na sequência de proteína, diversidade, presença de regiões conservadas.

A busca pelo termo TLR9 retorna 8 entradas (Busca realizada 26/01/17).

O termo "partial" significa que a sequência da proteína não está completa.

Vamos observar as informações da entrada 20 da busca realizada com o termo TLR9.

GenPept ▾

TLR9 [Homo sapiens]

GenBank: AAZ95520.1

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

Go to:

```
LOCUS          AAZ95520          1032 aa          linear          PRI 25-NOV-2009
DEFINITION     TLR9 [Homo sapiens].
ACCESSION     AAZ95520
VERSION       AAZ95520.1
DBSOURCE      accession DQ019999.1
KEYWORDS      .
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 1032)
AUTHORS       Georgel,P., Macquin,C. and Bahram,S.
TITLE         The heterogeneous allelic repertoire of human toll-like receptor
              (TLR) genes
JOURNAL       PLoS ONE 4 (11), E7803 (2009)
PUBMED        19924287
REMARK        Publication Status: Online-Only
```

Figura. Informações sobre a entrada 20 da busca TLR9.



UNIVERSIDADE FEDERAL DA BAHIA
Instituto Multidisciplinar em Saúde
Campus Anísio Teixeira



FEATURES

```
source          1..1032
                /organism="Homo sapiens"
                /db_xref="taxon:9606"
                /chromosome="3"
                /map="3p21.3"

Protein         1..1032
                /product="TLR9"

Region         43..64
                /region_name="leucine-rich repeat"
                /note="leucine-rich repeat [structural motif]"
                /db_xref="CDD:275380"

Region         64..135
                /region_name="LRR_8"
                /note="Leucine rich repeat; pfam13855"
                /db_xref="CDD:290566"

Region         65..88
                /region_name="leucine-rich repeat"
                /note="leucine-rich repeat [structural motif]"
                /db_xref="CDD:275380"

Region         89..124
                /region_name="leucine-rich repeat"
                /note="leucine-rich repeat [structural motif]"
                /db_xref="CDD:275380"
```

Figura. Features da entrada 20 da busca por TLR9.

No final da entrada encontramos a sequência da proteína.

O formato GenPept guarda toda a informação sobre as proteínas depositadas no banco **Proteins** do NCBI.



UNIVERSIDADE FEDERAL DA BAHIA
Instituto Multidisciplinar em Saúde
Campus Anísio Teixeira



Taxonomy/NCBI

<https://www.ncbi.nlm.nih.gov/taxonomy>

Banco de dados com informações sobre a classificação taxonômica das espécies. Esse banco de dados apresenta somente informações taxonômicas de espécies que têm informações moleculares depositadas no NCBI.

Informações sobre a espécie *Homo sapiens* no banco de dados Taxonomy.

Homo sapiens

Taxonomy ID: 9606

Genbank common name: **human**

Inherited blast name: **primates**

Rank: species

Genetic code: [Translation table 1 \(Standard\)](#)

Mitochondrial genetic code: [Translation table 2 \(Vertebrate Mitochondrial\)](#)

Other names:

synonym: **humans**

common
name: **man**

authority: **Homo sapiens Linnaeus,
1758**

[Lineage\(full \)](#)

[cellular organisms](#); [Eukaryota](#); [Opisthokonta](#); [Metazoa](#); [Eumetazoa](#); [Bilateria](#); [Deuterostomia](#); [Chordata](#); [Craniata](#); [Vertebrata](#); [Gnathostomata](#); [Teleostomi](#); [Euteleostomi](#); [Sarcopterygii](#); [Dipnotetrapodomorpha](#); [Tetrapoda](#); [Amniota](#); [Mammalia](#); [Theria](#); [Eutheria](#); [Boreoeutheria](#); [Euarchontoglires](#); [Primates](#); [Haplorrhini](#); [Simiiformes](#); [Catarrhini](#); [Hominoidea](#); [Hominidae](#); [Homininae](#); [Homo](#)

A classificação apresenta quantitativo de dados em outros banco de dados com links para essas informações na tabela "Entrez records".



UNIVERSIDADE FEDERAL DA BAHIA
Instituto Multidisciplinar em Saúde
Campus Anísio Teixeira



Entrez records		
Database name	Subtree links	Direct links
Nucleotide	14,112,440	14,112,396
Nucleotide EST	8,705,106	8,705,106
Nucleotide GSS	1,762,817	1,761,491
Protein	1,069,030	1,068,738
Structure	33,029	33,029
Genome	1	1
Popset	23,505	23,504
SNP	165,297,523	165,297,523
Domains	24	24
GEO Datasets	1,244,816	1,244,816
UniGene	130,056	130,056
PubMed Central	553,832	553,829
Gene	219,667	219,594
HomoloGene	18,713	18,713
SRA Experiments	742,541	742,319
Probe	27,382,435	27,382,435
Assembly	90	90
Bio Project	32,438	32,427
Bio Sample	2,516,813	2,516,682
Bio Systems	3,070	3,070
Clone DB	17,567,413	17,567,413



UNIVERSIDADE FEDERAL DA BAHIA
Instituto Multidisciplinar em Saúde
Campus Anísio Teixeira



dbVar	<u>5,064,734</u>	<u>5,064,734</u>
GEO Profiles	<u>61,958,910</u>	<u>61,958,910</u>
PubChem BioAssay	<u>311,510</u>	<u>311,502</u>
Protein Clusters	<u>13</u>	<u>13</u>
Taxonomy	<u>3</u>	<u>1</u>

Ainda podemos encontrar informações sobre projetos genomas e outros links para bancos taxonômicos nessa página.

Veja as informações que você pode encontrar para outras espécies.

Mus musculus

Taxonomy ID: 10090

Genbank common name: house mouse

Inherited blast name: rodents

Arabidopsis thaliana

Taxonomy ID: 3702

Genbank common name: thale cress



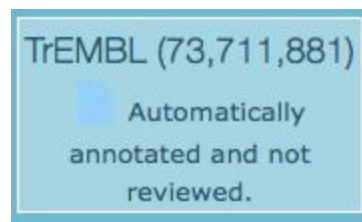
<http://www.uniprot.org>



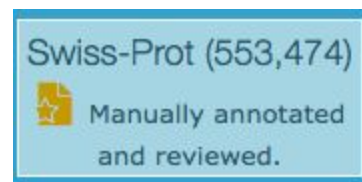
UNIVERSIDADE FEDERAL DA BAHIA
Instituto Multidisciplinar em Saúde
Campus Anísio Teixeira



O uniprot fornece informações e sequências sobre proteínas. As sequências e informações funcionais das proteínas podem vir de projetos genomas (sequências codificantes) ou podem ser depositadas por curadores do UniProt. Isso cria duas divisões dentro do Uniprot, o "**UniProtKB/TrEMBL**" e o "**UniProtKB/Swiss-Prot**". As entradas do Uniprot apresentam número de acesso alfanumérico, exemplo Q9NR96.



O TrEMBL são proteínas que foram adicionadas no Uniprot de forma automática e não passou por um processo de revisão dos curadores (Status **Unreviewed**). Essas informações devem ser usadas com mais cuidado.



O Swiss-Prot tem um crescimento lento porque todas as sequências que estão nessa subdivisão passaram pela anotação manual e revisão de um curador (Status **Reviewed**). Essas informações são confiáveis. As entradas trazem informações sobre resultados experimentais, características e conclusões científicas.

Informações

Protein

Gene

Organism

Status (**Reviewed** ou **Unreviewed**)

- Annotation score:

- "**Experimental evidence at protein level**" ou "**Protein predicted**"

Function

GO - Molecular function

GO - Biological process

Subcellular location



UNIVERSIDADE FEDERAL DA BAHIA
Instituto Multidisciplinar em Saúde
Campus Anísio Teixeira



Pesquisem sobre TLR9 no uniprot e vejam as informações que vocês podem encontrar.
Entrem no primeiro resultado (TLR(humano - Acesso Q9NR96).



UNIVERSIDADE FEDERAL DA BAHIA
Instituto Multidisciplinar em Saúde
Campus Anísio Teixeira



Protein Data Bank

www.rcsb.org

Banco de dados que armazena informações sobre estrutura tridimensional das proteínas. Nessa banco são armazenadas estruturas das proteínas que foram avaliadas por cristalografia Raio-X, Espectroscopia por ressonância magnética nuclear (NMR). Atualmente existem 126278 estruturas depositadas no PDB.