Database Integrati

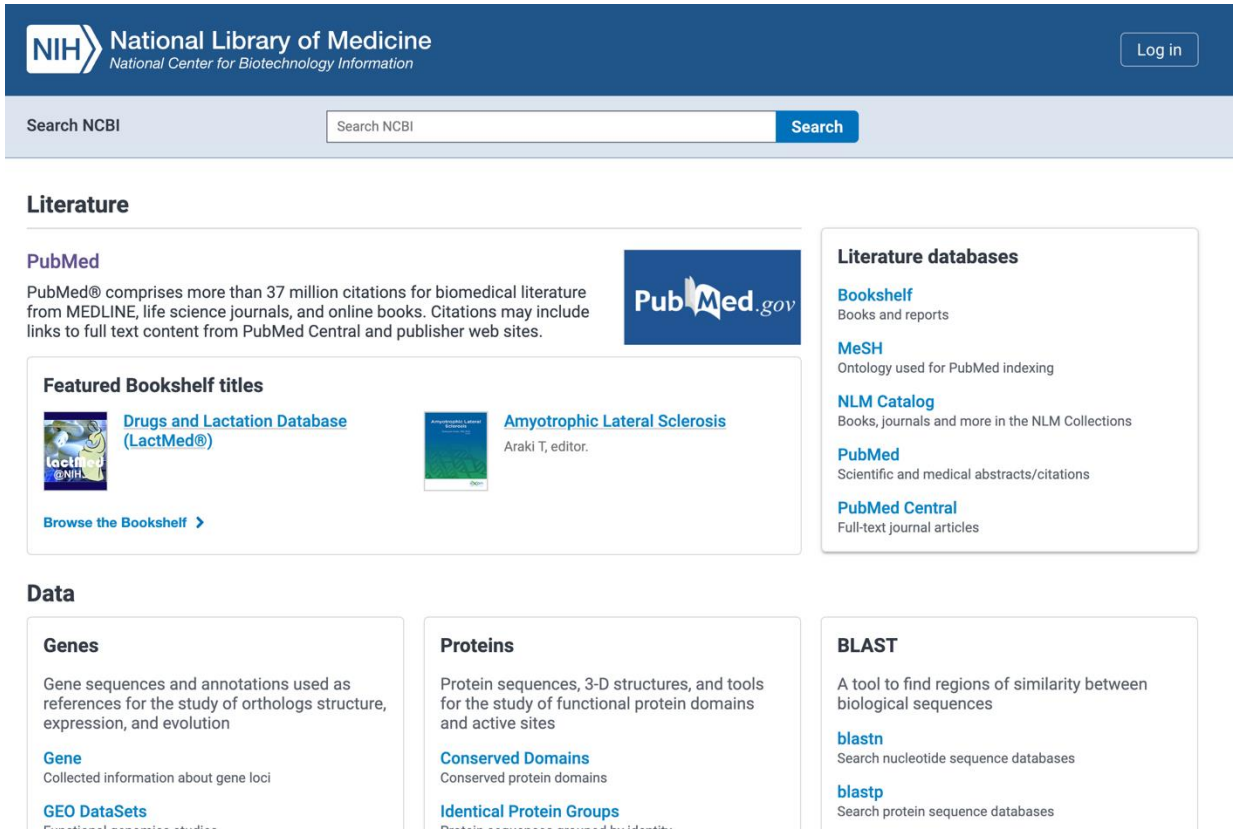
Entrez & Ensembl

Sistema Entrez

Entrez- Sistema integrato (USA) di banche dati biologiche

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

Questi sistemi consentono di accedere facilmente alle informazioni contenute nelle banche dati biologiche, permettendo di selezionare il database di interesse, imporre dei limiti, individuare dati correlati.



Welcome to the Entrez cross-database search page

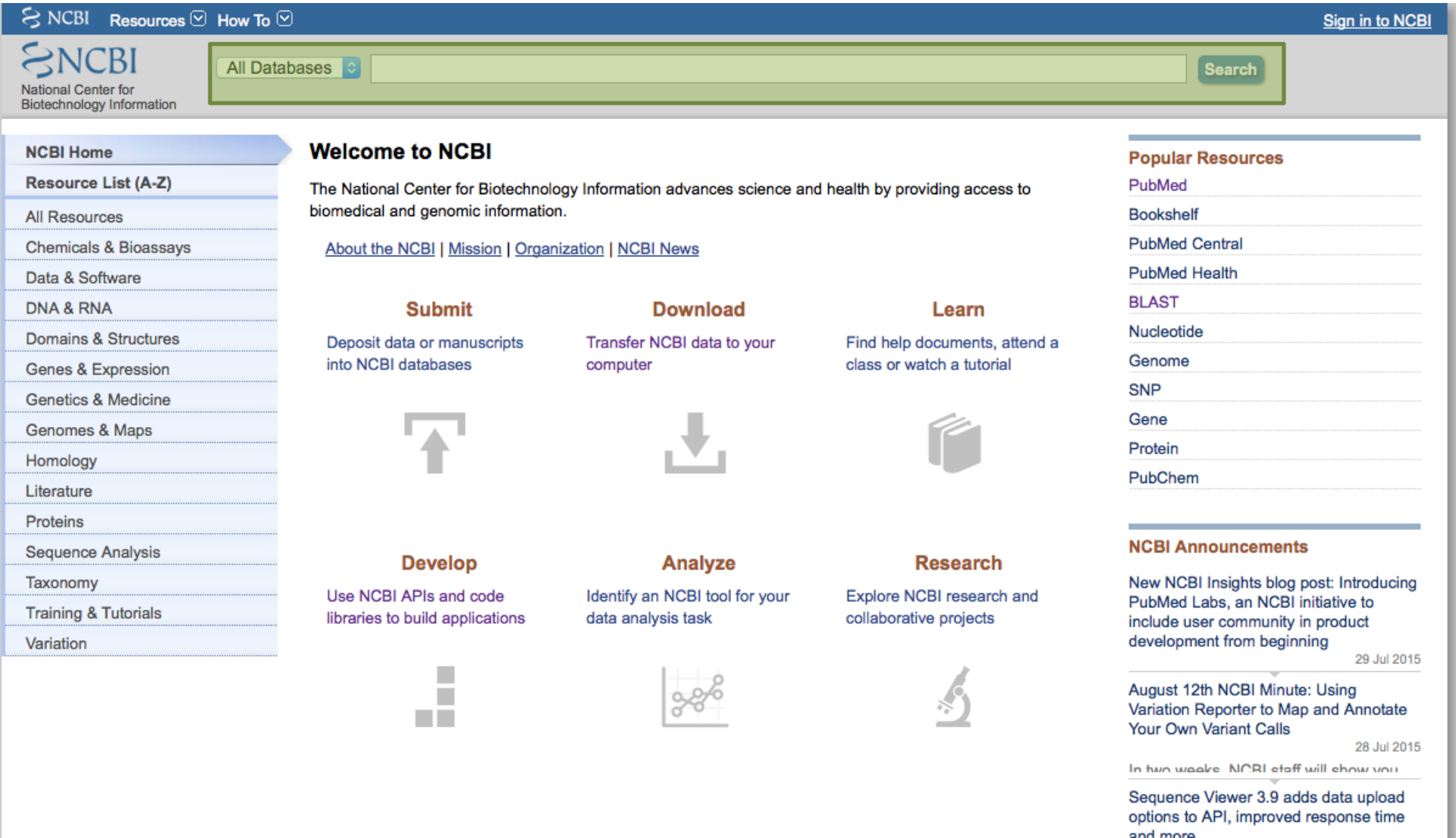
- **PubMed:** biomedical literature citations and abstracts
- **PubMed Central:** free, full text journal articles
- **Site Search:** NCBI web and FTP sites
- **Books:** online books
- **OMIM:** online Mendelian Inheritance in Man
- **OMIA:** online Mendelian Inheritance in Animals

- **Nucleotide:** Core subset of nucleotide sequence records
- **EST:** Expressed Sequence Tag records
- **GSS:** Genome Survey Sequence records
- **Protein:** sequence database
- **Genome:** whole genome sequences
- **Structure:** three-dimensional macromolecular structures
- **Taxonomy:** organisms in GenBank
- **SNP:** single nucleotide polymorphism
- **Gene:** gene-centered information
- **SRA:** Sequence Read Archive
- **BioSystems:** Pathways and systems of interacting molecules
- **HomoloGene:** eukaryotic homology groups
- **GENSAT:** gene expression atlas of mouse central nervous system
- **Probe:** sequence-specific reagents
- **Genome Project:** genome project information
- **dbGaP:** genotype and phenotype
- **UniGene:** gene-oriented clusters of transcript sequences
- **CDD:** conserved protein domain database
- **3D Domains:** domains from Entrez Structure
- **UniSTS:** markers and mapping data
- **PopSet:** population study data sets
- **GEO Profiles:** expression and molecular abundance profiles
- **GEO DataSets:** experimental sets of GEO data
- **Cancer Chromosomes:** cytogenetic databases
- **PubChem BioAssay:** bioactivity screens of chemical substances
- **PubChem Compound:** unique small molecule chemical structures
- **PubChem Substance:** deposited chemical substance records
- **Protein Clusters:** a collection of related protein sequences
- **Peptidome:** MS/MS proteomic experiments

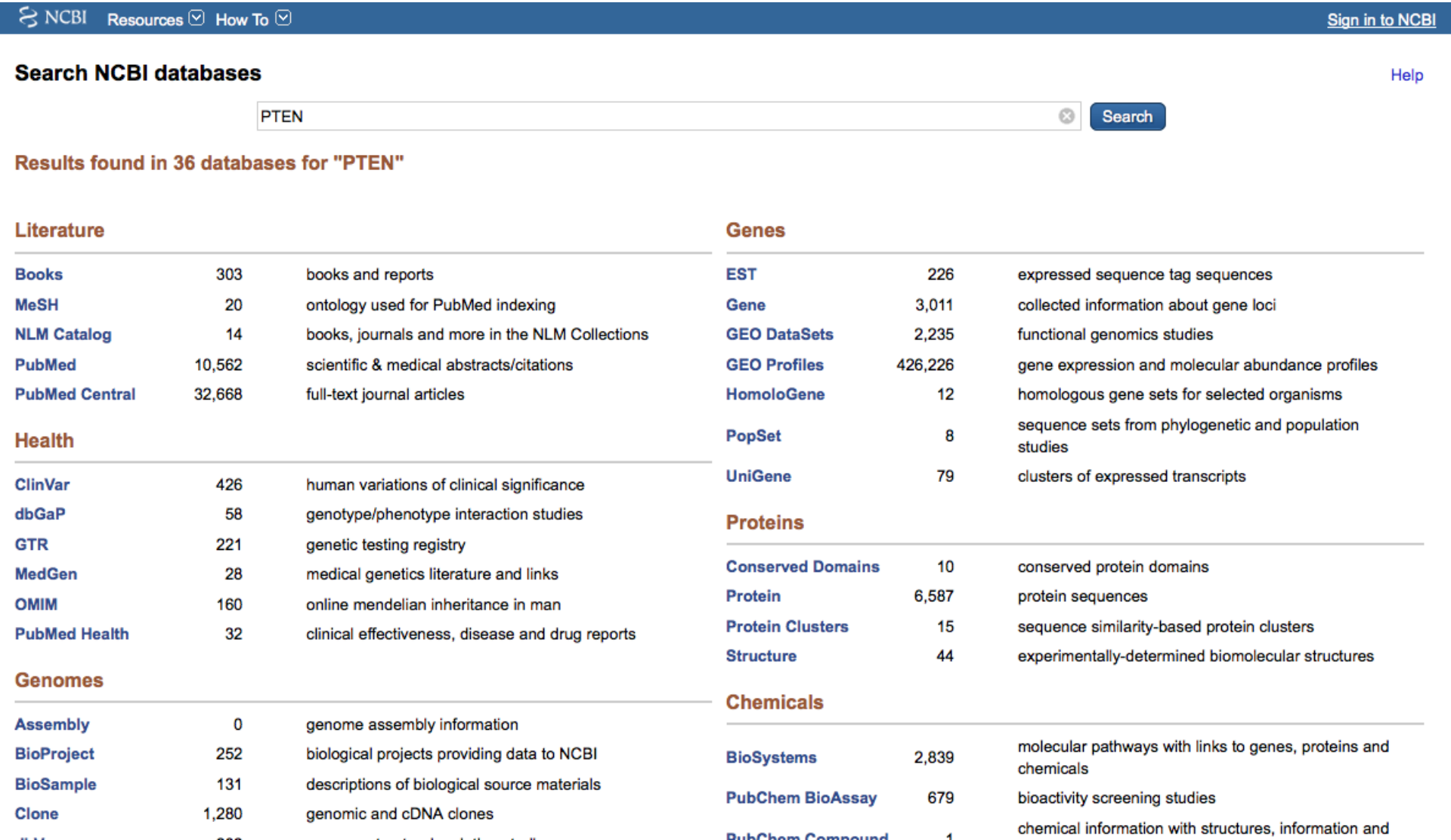
- **Journals:** detailed information about the journals indexed in PubMed and other Entrez databases
- **NLM Catalog:** catalog of books, journals, and audiovisuals in the NLM collections
- **MeSH:** detailed information about NLM's controlled vocabulary

National Center for Biotechnology Information

https://www.ncbi.nlm.nih.gov/



National Center for Biotechnology Information



NCBI Resources How To Sign in to NCBI

Search NCBI databases Help

PTEN Search

Results found in 36 databases for "PTEN"

Literature

Books	303	books and reports
MeSH	20	ontology used for PubMed indexing
NLM Catalog	14	books, journals and more in the NLM Collections
PubMed	10,562	scientific & medical abstracts/citations
PubMed Central	32,668	full-text journal articles

Health

ClinVar	426	human variations of clinical significance
dbGaP	58	genotype/phenotype interaction studies
GTR	221	genetic testing registry
MedGen	28	medical genetics literature and links
OMIM	160	online mendelian inheritance in man
PubMed Health	32	clinical effectiveness, disease and drug reports

Genomes

Assembly	0	genome assembly information
BioProject	252	biological projects providing data to NCBI
BioSample	131	descriptions of biological source materials
Clone	1,280	genomic and cDNA clones

Genes

EST	226	expressed sequence tag sequences
Gene	3,011	collected information about gene loci
GEO DataSets	2,235	functional genomics studies
GEO Profiles	426,226	gene expression and molecular abundance profiles
HomoloGene	12	homologous gene sets for selected organisms
PopSet	8	sequence sets from phylogenetic and population studies
UniGene	79	clusters of expressed transcripts

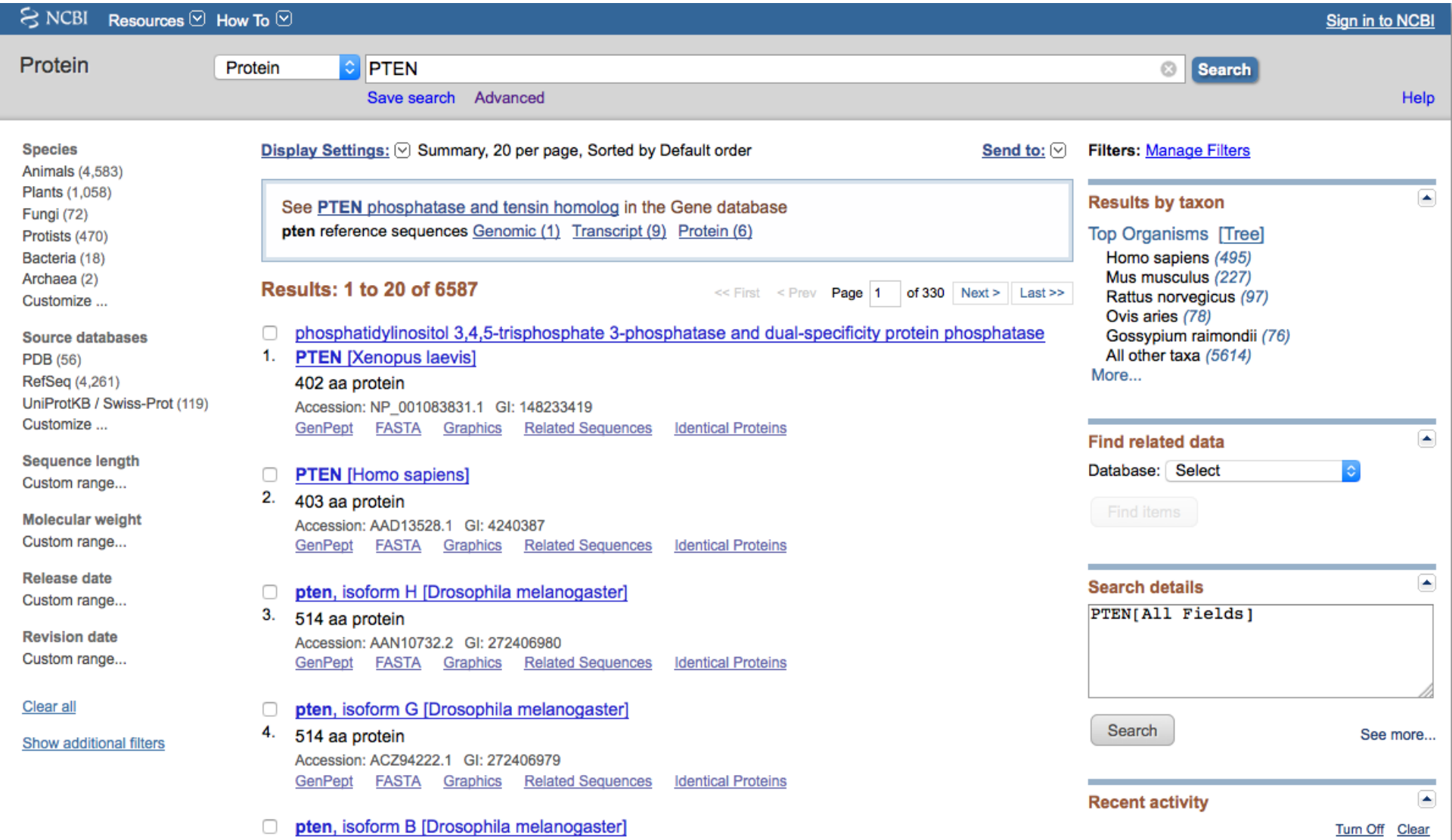
Proteins

Conserved Domains	10	conserved protein domains
Protein	6,587	protein sequences
Protein Clusters	15	sequence similarity-based protein clusters
Structure	44	experimentally-determined biomolecular structures

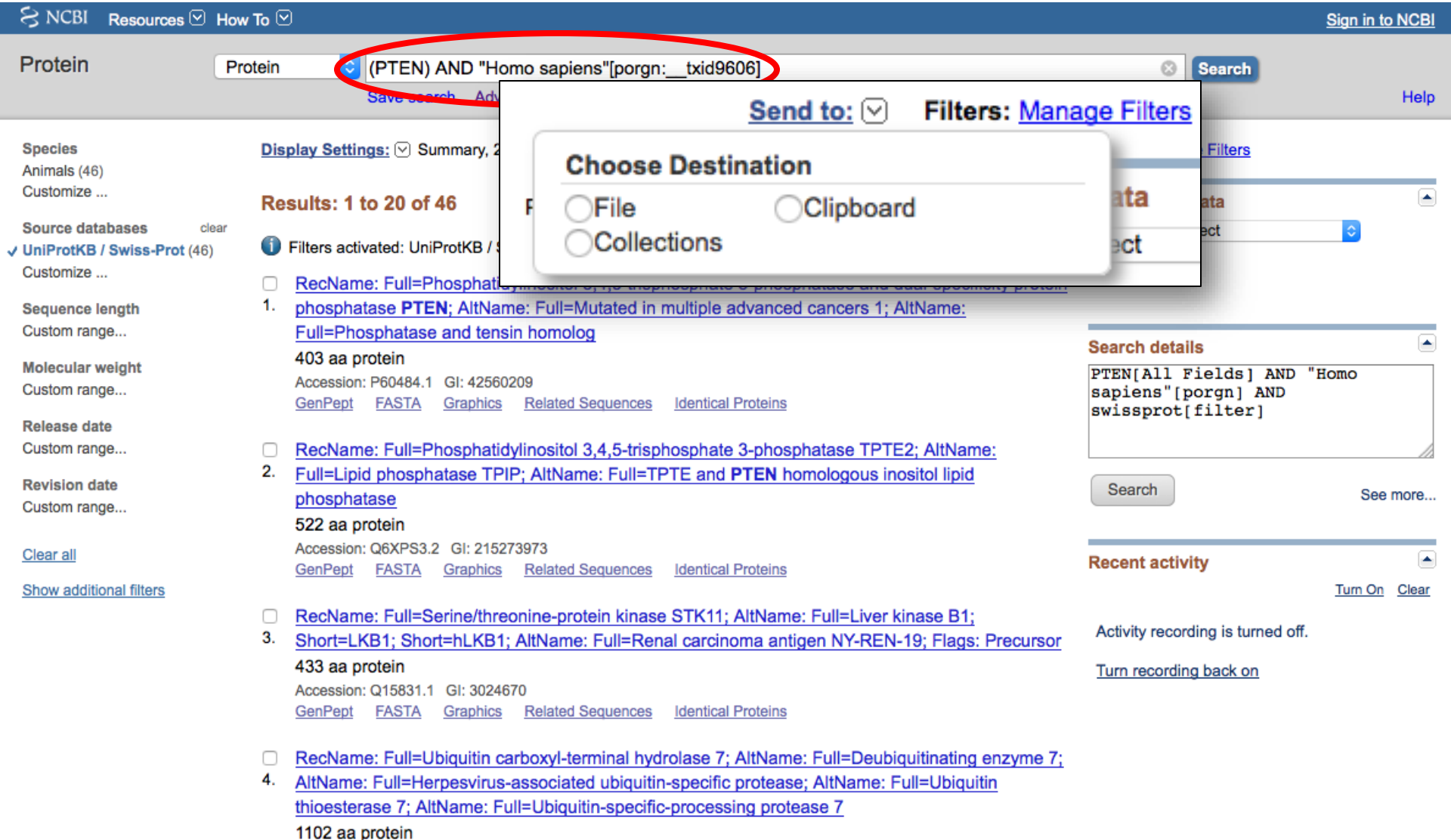
Chemicals

BioSystems	2,839	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	679	bioactivity screening studies
PubChem Compound	1	chemical information with structures, information and

National Center for Biotechnology Information

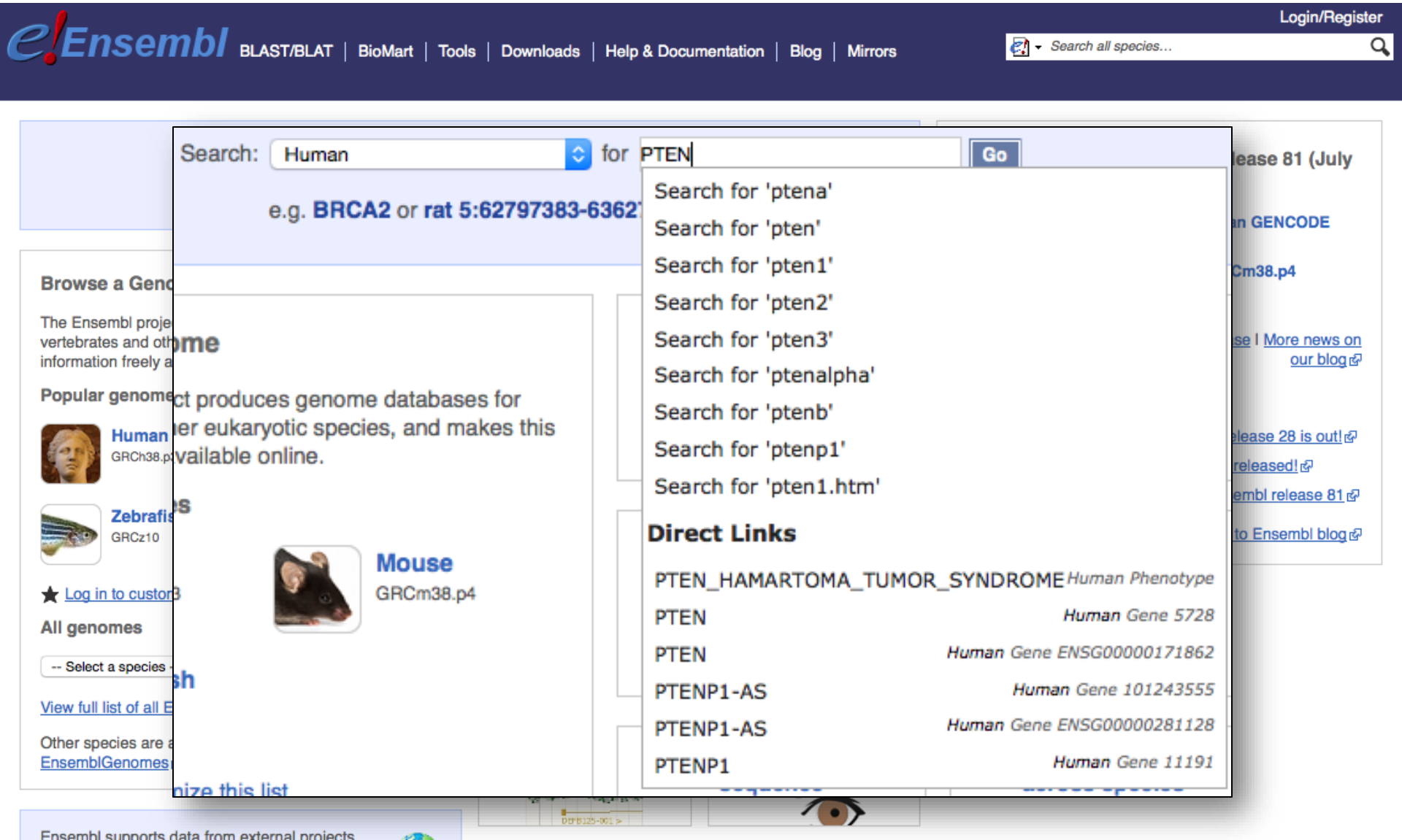


National Center for Biotechnology Information

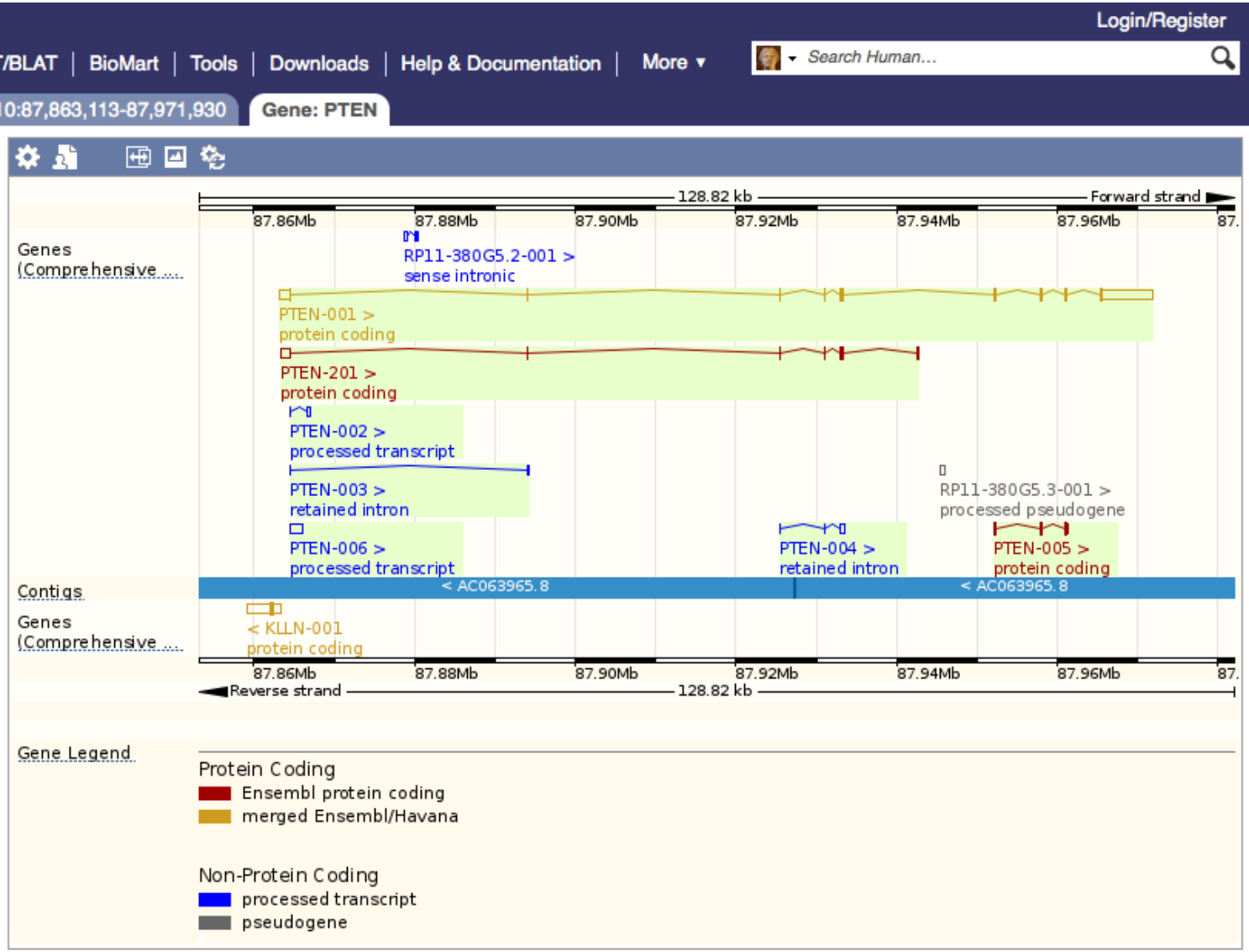


ENSEMBL

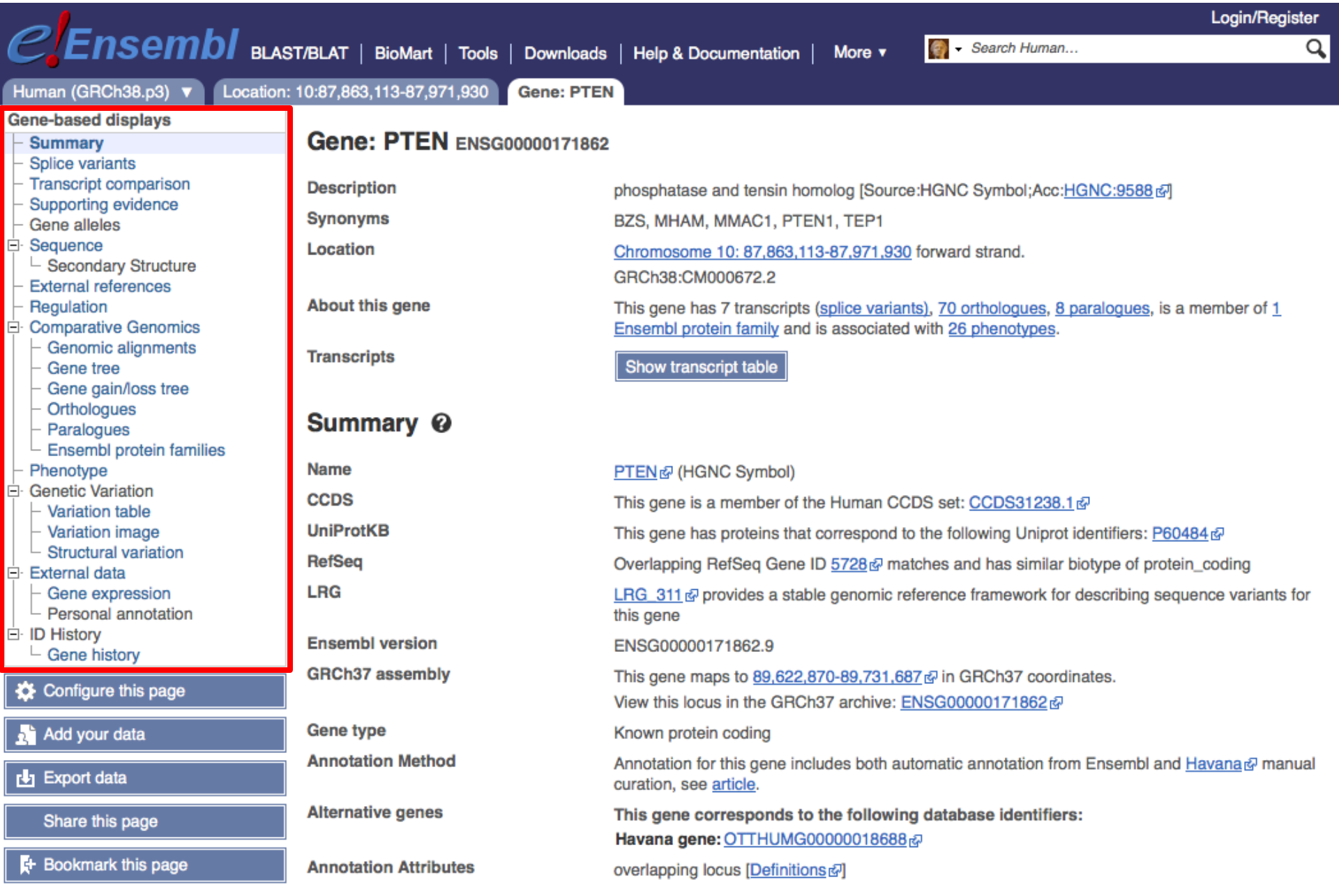
http://www.ensembl.org/

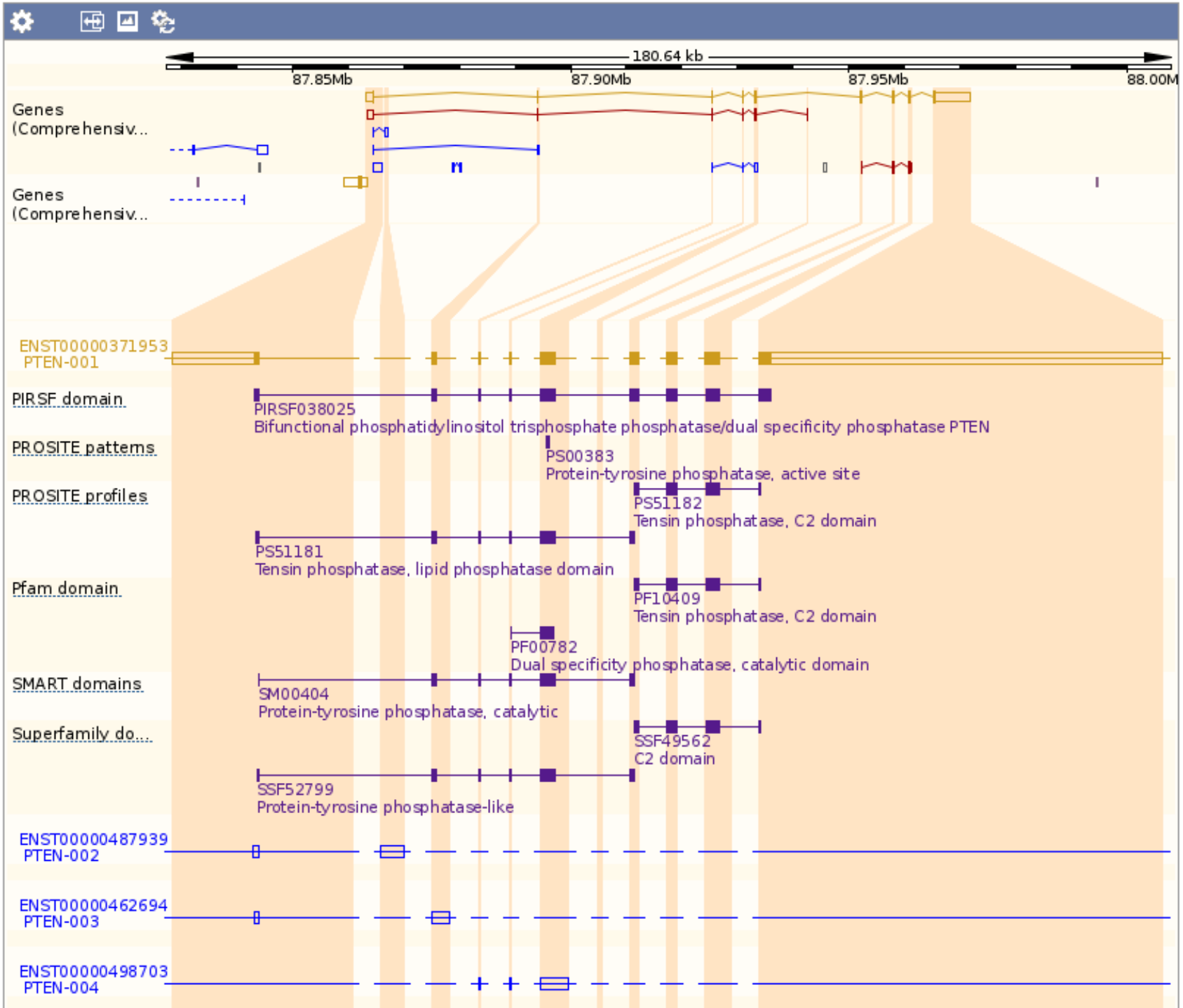


ENSEMBL


Ensembl
BLAST/BLAT
BioMart
Tools
Downloads
Help & Documentation
More
Login/Register
Search Human...
Human (GRCh38.p3)
Location: 10:87,863,113-87,971,930
Gene: PTEN
128.82 kb
Forward strand
87.86Mb
87.88Mb
87.90Mb
87.92Mb
87.94Mb
87.96Mb
Genes (Comprehensive ...
RP11-380G5.2-001 >
sense intronic
PTEN-001 >
protein coding
PTEN-201 >
protein coding
PTEN-002 >
processed transcript
PTEN-003 >
retained intron
RP11-380G5.3-001 >
processed pseudogene
PTEN-006 >
processed transcript
PTEN-004 >
retained intron
PTEN-005 >
protein coding
Contigs
< AC063965.8
< AC063965.8
Genes (Comprehensive ...
< KLLN-001
protein coding
Reverse strand
128.82 kb
Gene Legend
Protein Coding
Ensembl protein coding
merged Ensembl/Havana
Non-Protein Coding
processed transcript
pseudogene

ENSEMBL


Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | More
Login/Register
Search Human...
Human (GRCh38.p3)
Location: 10:87,863,113-87,971,930
Gene: PTEN
Gene-based displays
Summary
Splice variants
Transcript comparison
Supporting evidence
Gene alleles
Sequence
Secondary Structure
External references
Regulation
Comparative Genomics
Genomic alignments
Gene tree
Gene gain/loss tree
Orthologues
Paralogues
Ensembl protein families
Phenotype
Genetic Variation
Variation table
Variation image
Structural variation
External data
Gene expression
Personal annotation
ID History
Gene history
Configure this page
Add your data
Export data
Share this page
Bookmark this page
Gene: PTEN ENSG00000171862
Description
phosphatase and tensin homolog [Source:HGNC Symbol;Acc:HGNC:9588]
Synonyms
BZS, MHAM, MMAC1, PTEN1, TEP1
Location
Chromosome 10: 87,863,113-87,971,930 forward strand.
GRCh38:CM000672.2
About this gene
This gene has 7 transcripts (splice variants), 70 orthologues, 8 paralogues, is a member of 1 Ensembl protein family and is associated with 26 phenotypes.
Transcripts
Show transcript table
Summary
Name
PTEN (HGNC Symbol)
CCDS
This gene is a member of the Human CCDS set: CCDS31238.1
UniProtKB
This gene has proteins that correspond to the following Uniprot identifiers: P60484
RefSeq
Overlapping RefSeq Gene ID 5728 matches and has similar biotype of protein_coding
LRG
LRG_311 provides a stable genomic reference framework for describing sequence variants for this gene
Ensembl version
ENSG00000171862.9
GRCh37 assembly
This gene maps to 89,622,870-89,731,687 in GRCh37 coordinates.
View this locus in the GRCh37 archive: ENSG00000171862
Gene type
Known protein coding
Annotation Method
Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see article.
Alternative genes
This gene corresponds to the following database identifiers:
Havana gene: OTTHUMG00000018688
Annotation Attributes
overlapping locus [Definitions]


180.64 kb
87.85Mb
87.90Mb
87.95Mb
88.00M
Genes (Comprehensiv...
Genes (Comprehensiv...
ENST00000371953 PTEN-001
PIRSF domain
PIRSF038025
Bifunctional phosphatidylinositol trisphosphate phosphatase/dual specificity phosphatase PTEN
PROSITE patterns
PS00383
Protein-tyrosine phosphatase, active site
PROSITE profiles
PS51182
Tensin phosphatase, C2 domain
PS51181
Tensin phosphatase, lipid phosphatase domain
Pfam domain
PF10409
Tensin phosphatase, C2 domain
PF00782
Dual specificity phosphatase, catalytic domain
SMART domains
SM00404
Protein-tyrosine phosphatase, catalytic
Superfamily do...
SSF49562
C2 domain
SSF52799
Protein-tyrosine phosphatase-like
ENST00000487939 PTEN-002
ENST00000462694 PTEN-003
ENST00000498703 PTEN-004

Summary of variant consequences in ENSG00000171862

Switch to tree view

Show All entries | Filter

Number of variant consequences		Type	Description
3	Show	Transcript ablation	A feature ablation whereby the deleted region includes a transcript feature (SO:0001893)
147	Show	Splice donor variant	A splice variant that changes the 2 base region at the 5' end of an intron (SO:0001575)

Synonymous variant consequences

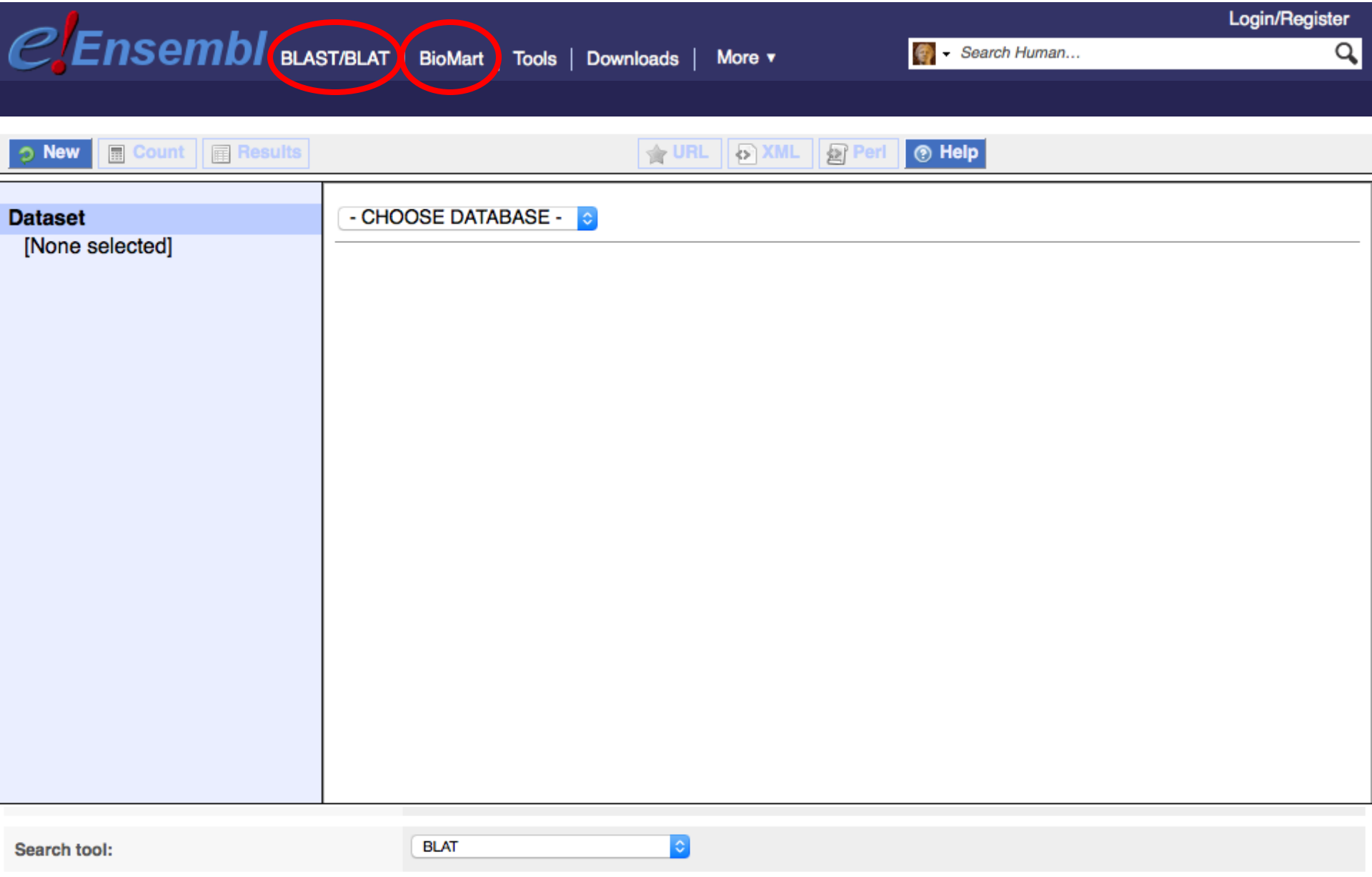
[back to top]

Show All entries | Show/hide columns (5 hidden) | Filter

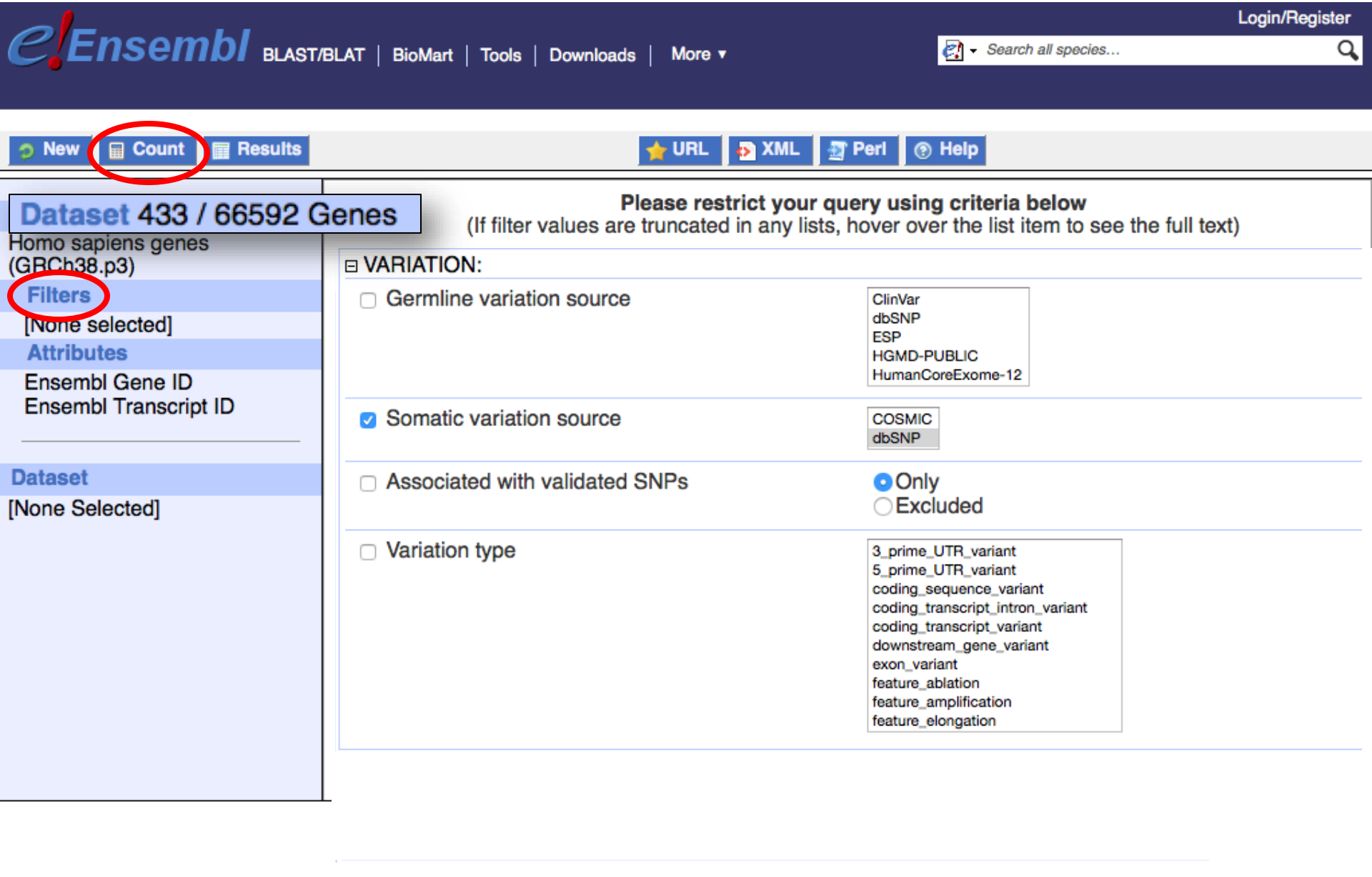
ID	Chr: bp	Alleles	Global MAF	Source	Type	AA	AA coord	Transcript
COSM1561221	10:87864478	C/T	-	COSMIC	Synonymous variant	A	3	ENST00000371953
COSM1349460	10:87864487	A/G	-	COSMIC	Synonymous variant	K	6	ENST00000371953
COSM1349463	10:87864499	C/T	-	COSMIC	Synonymous variant	S	10	ENST00000371953
COSM1349464	10:87864514	A/G	-	COSMIC	Synonymous variant	R	15	ENST00000371953
COSM346197	10:87864517	T/C	-	COSMIC	Synonymous variant	Y	16	ENST00000371953
COSM1349465	10:87864520	A/G	-	COSMIC	Synonymous variant	Q	17	ENST00000371953
COSM4016447	10:87864529	A/G	-	COSMIC	Synonymous variant	G	20	ENST00000371953
COSM1561218	10:87864532	C/T	-	COSMIC	Synonymous variant	F	21	ENST00000371953
COSM1349468	10:87864536	T/C	-	COSMIC	Synonymous variant	L	23	ENST00000371953
COSM1745950	10:87864547	C/A	-	COSMIC	Splice region variant, Synonymous variant	T	26	ENST00000371953
COSM1349471	10:87894026	T/C	-	COSMIC	Splice region variant, Synonymous variant	Y	27	ENST00000371953

1151	Show	Missense variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved (SO:0001583)
317	Show	Splice region variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron (SO:0001630)
0	-	Incomplete terminal codon variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed (SO:0001626)
254	Hide	Synonymous variant	A sequence variant where there is no resulting change to the encoded amino acid (SO:0001819)

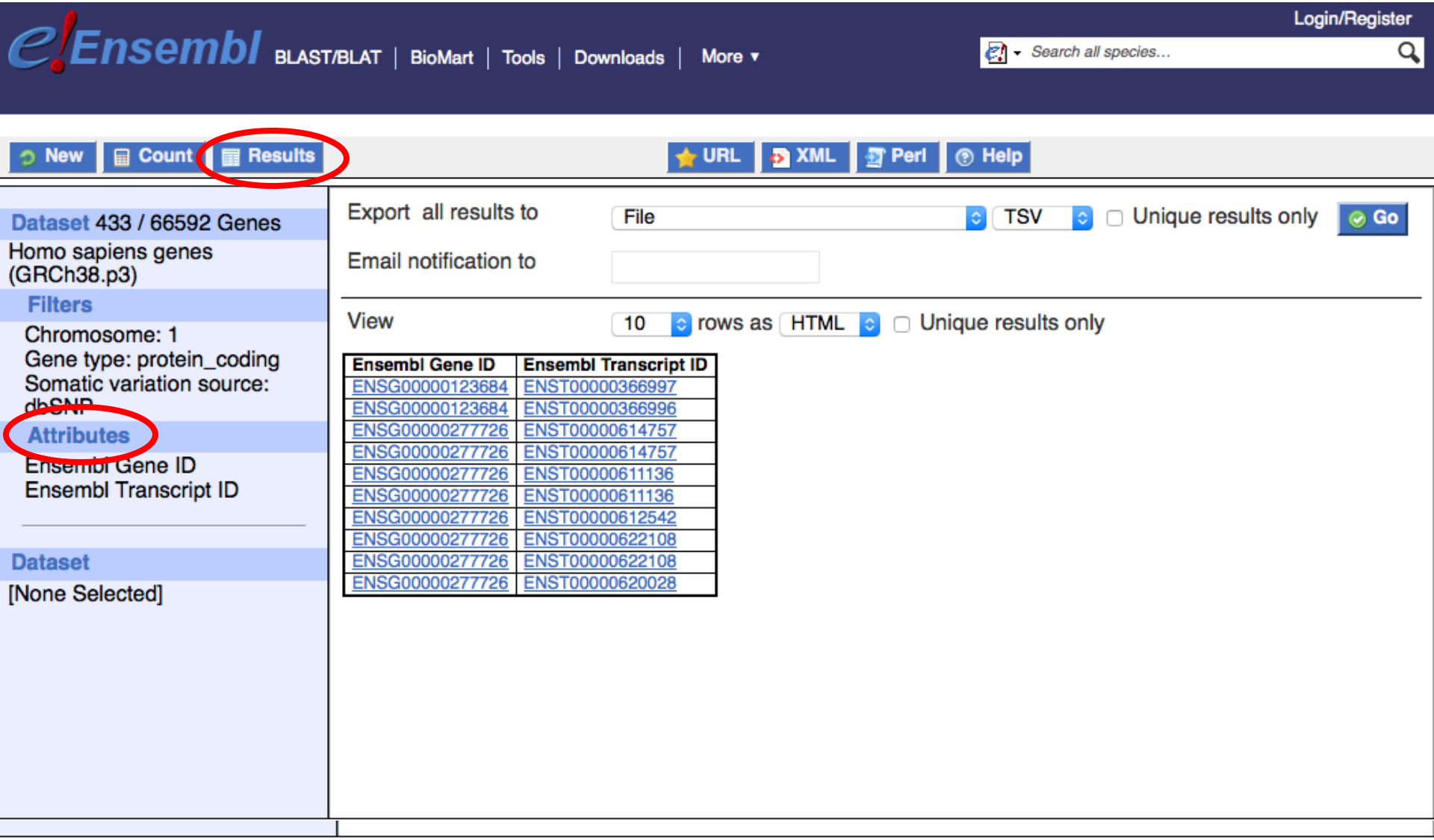
ENSEMBL


e!Ensembl
BLAST/BLAT
BioMart
Tools
Downloads
More
Login/Register
Search Human...
New
Count
Results
URL
XML
Perl
Help
Dataset
[None selected]
- CHOOSE DATABASE -
Search tool:
BLAT

ENSEMBL

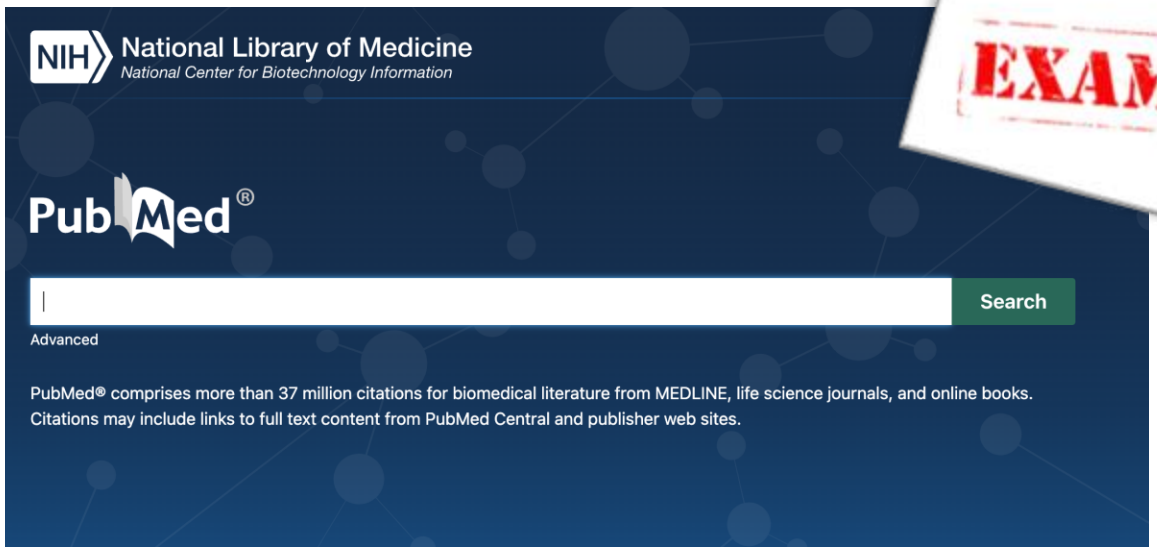

Ensembl
BLAST/BLAT
BioMart
Tools
Downloads
More
Login/Register
Search all species...
New
Count
Results
URL
XML
Perl
Help
Dataset 433 / 66592 Genes
Homo sapiens genes (GRCh38.p3)
Filters
[None selected]
Attributes
Ensembl Gene ID
Ensembl Transcript ID
Dataset
[None Selected]
Please restrict your query using criteria below
(If filter values are truncated in any lists, hover over the list item to see the full text)
VARIATION:
Germline variation source
ClinVar
dbSNP
ESP
HGMD-PUBLIC
HumanCoreExome-12
Somatic variation source
COSMIC
dbSNP
Associated with validated SNPs
Only
Excluded
Variation type
3_prime_UTR_variant
5_prime_UTR_variant
coding_sequence_variant
coding_transcript_intron_variant
coding_transcript_variant
downstream_gene_variant
exon_variant
feature_ablation
feature_amplification
feature_elongation

ENSEMBL



Ensembl Gene ID	Ensembl Transcript ID
ENSG00000123684	ENST00000366997
ENSG00000123684	ENST00000366996
ENSG00000277726	ENST00000614757
ENSG00000277726	ENST00000614757
ENSG00000277726	ENST00000611136
ENSG00000277726	ENST00000611136
ENSG00000277726	ENST00000612542
ENSG00000277726	ENST00000622108
ENSG00000277726	ENST00000622108
ENSG00000277726	ENST00000620028

SCAN ME



EXAMPLE

Torniamo al nostro lavoro sulla balena minke (scansionate il QR code)

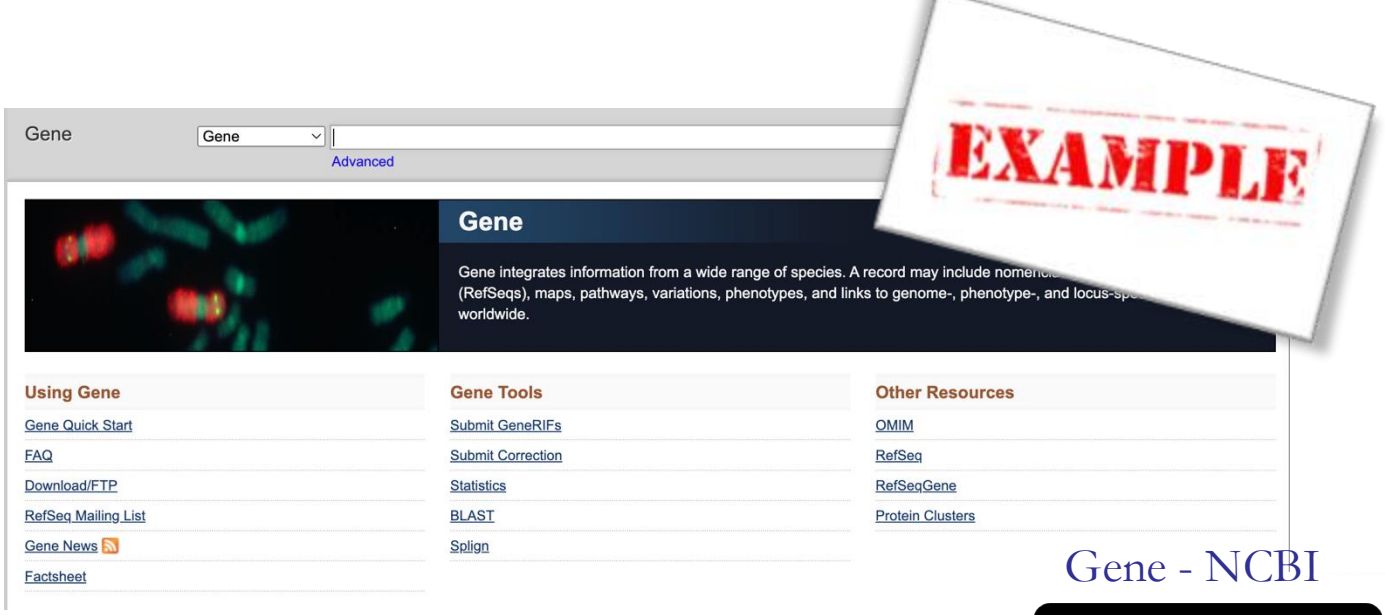
1. Andate a cercare la figura 2. Il testo sotto la figura 2 parla dell'espansione della famiglia genica delle Peroxiredoxins (PRDX)

Una **famiglia genica** è un gruppo di geni che condividono una sequenza simile e derivano da un antenato comune. I geni all'interno di una famiglia genica spesso codificano proteine con funzioni correlate o simili, e le loro somiglianze strutturali riflettono la loro origine evolutiva condivisa.

Le famiglie geniche si formano generalmente attraverso **processi di duplicazione genica**, in cui un gene viene copiato all'interno del genoma. Dopo la duplicazione, le copie possono evolvere in modo indipendente.

SCAN ME

lavoro balena



Gene - NCBI

Torniamo al nostro lavoro sulla balena minke (scansionate il QR code)

1. Andate a cercare la figura 2. Il testo sotto la figura 2 parla dell'espansione della famiglia genica associato alla resistenza allo stress e al redox signalling. Che famiglia è ?
2. Andate sul database gene e provate a inserire il nome della famiglia. Che cosa trovate ? Quanti membri della famiglia ci sono in Uomo ?
3. Andate su PRDX1, quanti trascritti ha questo gene ?
4. Provate ora a guardare la stessa cosa su Ensembl. Quanti trascritti ha PRDX1 in Ensembl ?
5. Cosa è successo ?

Ensembl

Ensembl vs Gene

Il database **GENE** e il database **Ensembl** possono avere un numero diverso di trascritti per lo stesso gene a causa di differenze che derivano principalmente dal modo in cui i dati sono raccolti, filtrati, aggiornati e annotati.

1.Differenti fonti di dati:

I due database possono utilizzare diverse fonti per identificare e annotare i trascritti. Ad esempio, Ensembl combina dati di RefSeq, GENCODE, RNA-Seq e può avere annotazioni più aggiornate o dettagliate rispetto a GENE.

2.Criteri di annotazione:

1. Ensembl adotta un approccio automatizzato per generare annotazioni di trascritti, mentre GENE ha un processo più curato manualmente, con criteri diversi per cosa considerare un trascritto valido.
2. Le regole di filtraggio per definire cosa è un "trascritto funzionale" o un "isoforma valida" possono variare, quindi un trascritto annotato in uno dei due database potrebbe non essere presente nell'altro.

3.Aggiornamenti e versioni:

Le versioni del database vengono aggiornate periodicamente. Se uno dei due database ha subito un aggiornamento recente, potrebbero esserci differenze tra i trascritti che rispecchiano le nuove scoperte scientifiche o i cambiamenti nell'annotazione.

4.Integrazione di RNA-Seq e dati sperimentali:

Ensembl spesso include trascritti derivati da dati di RNA-Seq che non sono presenti in GENE. Questo può portare a un numero maggiore di trascritti perché vengono considerate anche varianti rare o non convenzionali.

5.Trascritti predetti vs. confermati:

Alcuni trascritti presenti in Ensembl potrebbero essere predetti e non ancora confermati sperimentalmente, mentre GENE potrebbe includere solo trascritti con evidenze solide.

DATABASE PRIMARI

Sono i contenitori di tutte le sequenze prodotte nel mondo e rese disponibili alla comunità scientifica. Memorizzano la sequenza e poche altre informazioni generiche associate (laboratorio, data, specie, descrizione…)

EMBL datalibrary → Europa

GenBank → USA

DDBJ → Giappone

Si aggiornano quotidianamente

SWISSPROT sviluppata a Ginevra

TrEMBL (TRanslated EMBL)

PIR (Protein Information Resource).

DATABASE SECONDARI

Contengono il risultato di analisi eseguite sulle sequenze contenute nei database primari arricchendo il dato di informazioni utili.

Es. Da SWISSPROT primario di sequenze di amminoacidi, sono costruiti i database secondari PROSITE e PFAM

http://au.expasy.org/prosite/

(Database of protein families and domains)

http://www.sanger.ac.uk/Software/Pfam/

(Protein families database of alignments and HMMs)





Problema della ridondanza

① Molti database proteici e nucleotidici contengono sequenze appartenenti a famiglie geniche o versioni di geni omologhi di organismi differenti.
② Molti gruppi possono sottomettere la stessa sequenza.
③ Differenti approcci sperimentali possono produrre sequenze simili (ad esempio da genomico e da cDNA).
④ Splicing alternativi.

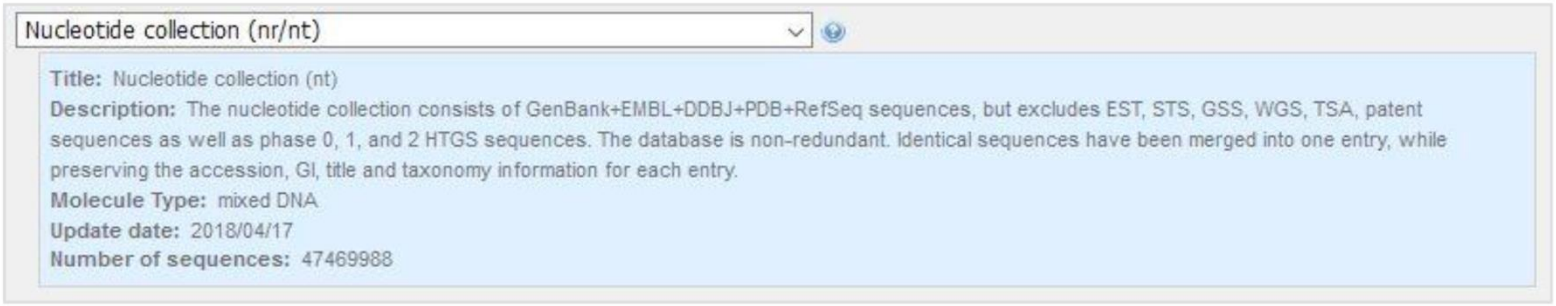
L`utilizzo di database ridondanti comporta almeno tre potenziali fonti di errore:

- ✓ Le analisi statistiche soffriranno di un bias
- ✓ Apparenti correlazioni in differenti posizioni specifiche delle sequenze possono essere un artefatto dovuto ad un bias nella composizione del campione.
- ✓ Se un dataset è usato per predizioni di certe caratteristiche, e le sequenze per calibrare la predizioni sono troppo correlate tra loro, l`apparente capacità predittiva può essere sovrastimata (riconosce il particolare, non il generale).

Problema della ridondanza

Per queste ragioni si sono resi necessari i **DATABASE COMPOSTI** che prendono i dati da più sorgenti, generalmente con formati diversi e **costituiscono un nuovo database specializzato, in genere "pulito" e non ridondante**.

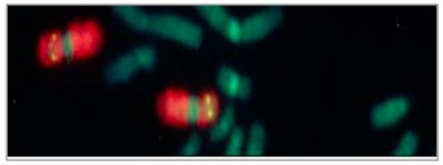
Esempio: **NRdb *(Non Redundant Database)***
Database di acidi nucleici o proteine in cui l'informazione è stata ottimizzata in modo non ripetitivo (non ridondante).



- In **RefSeq** sono rappresentate, in modo non ridondante, tutte le sequenze **genomiche**, di **mRNA** e di **proteine**.



- **Gene** è un sottoinsieme di RefSeq con rappresentate solo le sequenze **geniche**.



Database Proteici Primari

Uniprot è un **consorzio** che nasce dalla collaborazione tra European Bioinformatics Institute (EBI), SIB Swiss Institute of Bioinformatics e Protein Information Resource (PIR).

