Genome data analysis

Computer lab session 3

Theresa Schwarz MSc

theresa.schwarz@jku.at
Institute of Biophysics, JKU
Gruberstraße 40, 4020 Linz

Overview

- ❑ **Genome browsers**
  - ➢ **UCSC**
  - ➢ **ENSEMBL**

- ❑ **BLAST**
  - ➢ **Pairwise alignments**
  - ➢ **Database alignments**
  - ➢ **Primer-BLAST**

Overview

- ❑ **Genome browsers**
  - ➢ **UCSC**
  - ➢ **ENSEMBL**

- ❑ **BLAST**
  - ➢ **Pairwise alignments**
  - ➢ **Database alignments**
  - ➢ **Primer-BLAST**

Genome browsers

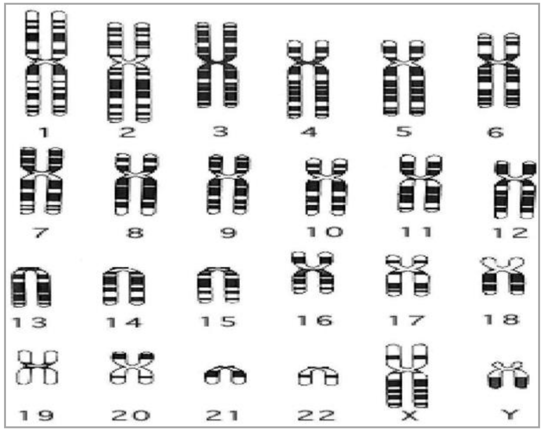
- **UCSC** Genome Browser (University of California Santa Cruz)
  https://genome.ucsc.edu/

- **ENSEMBL** (EMBO-Heidelberg/EBI-Cambridge)
  http://www.ensembl.org/

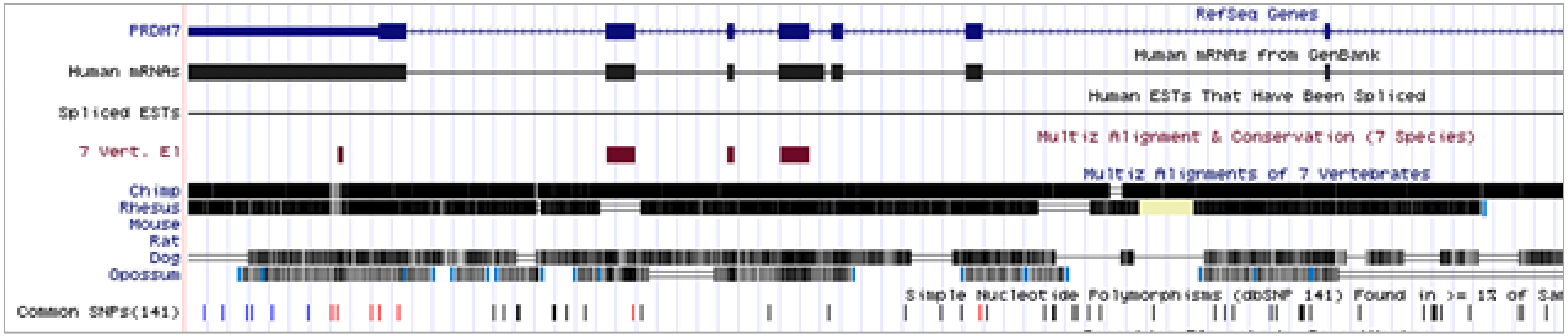
- **NCBI** (NIH, US) Genome Map Viewer
  https://www.ncbi.nlm.nih.gov/mapview/

Genome browsers

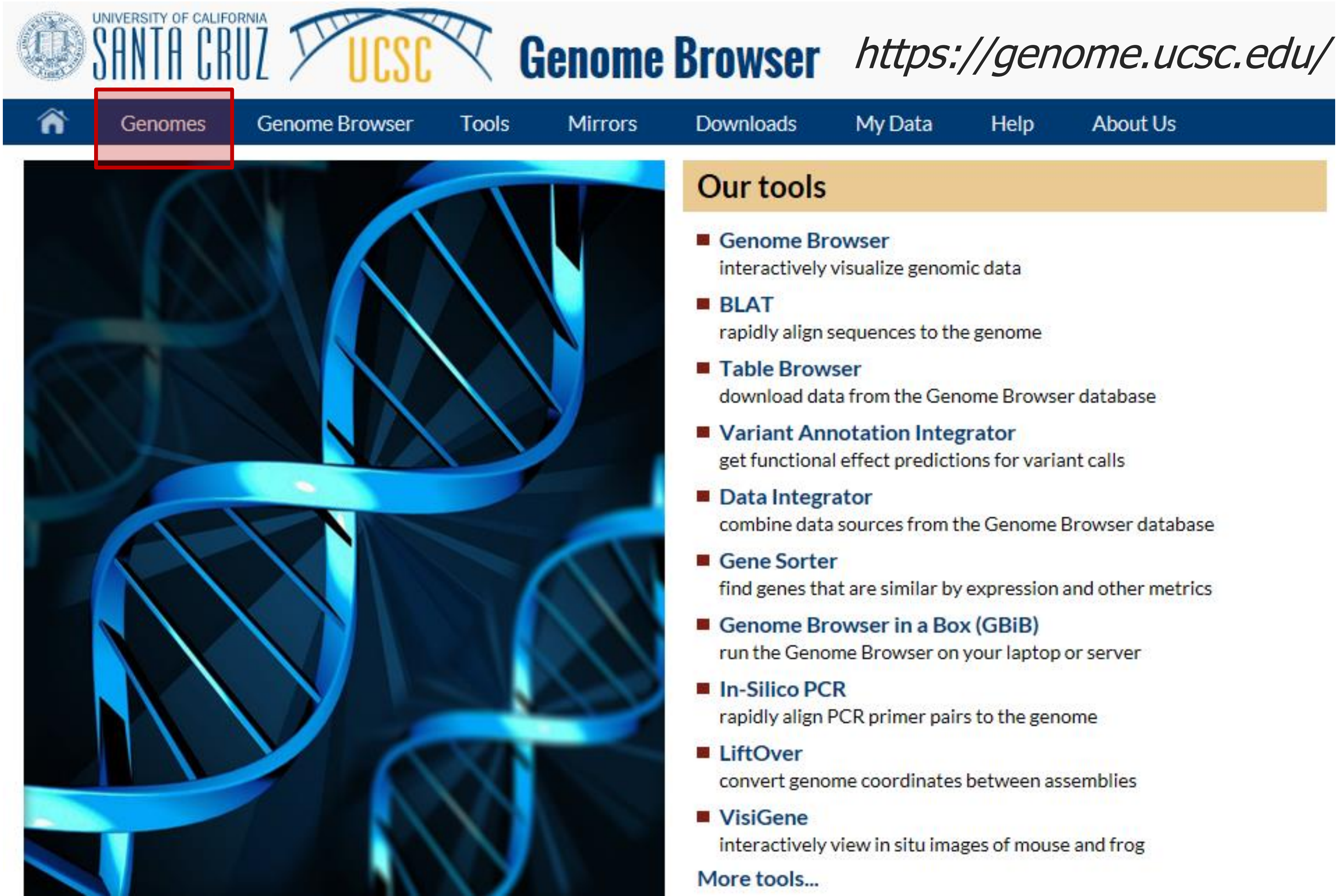
- Genomic DNA is organized in chromosomes.



- Genome browsers display ideograms (pictures) of chromosomes. Users can select '**annotation tracks**' that display many kinds of information.


RefSeq Genes
PRDM7
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Multiz Alignment & Conservation (7 Species)
7 Vert. El
Multiz Alignments of 7 Vertebrates
Chimp
Rhesus
Mouse
Rat
Dog
Opossum
Simple Nucleotide Polymorphisms (dbSNP 141) Found in >= 1% of Sa
Common SNPs(141)

UCSC Genome Browser



https://genome.ucsc.edu/

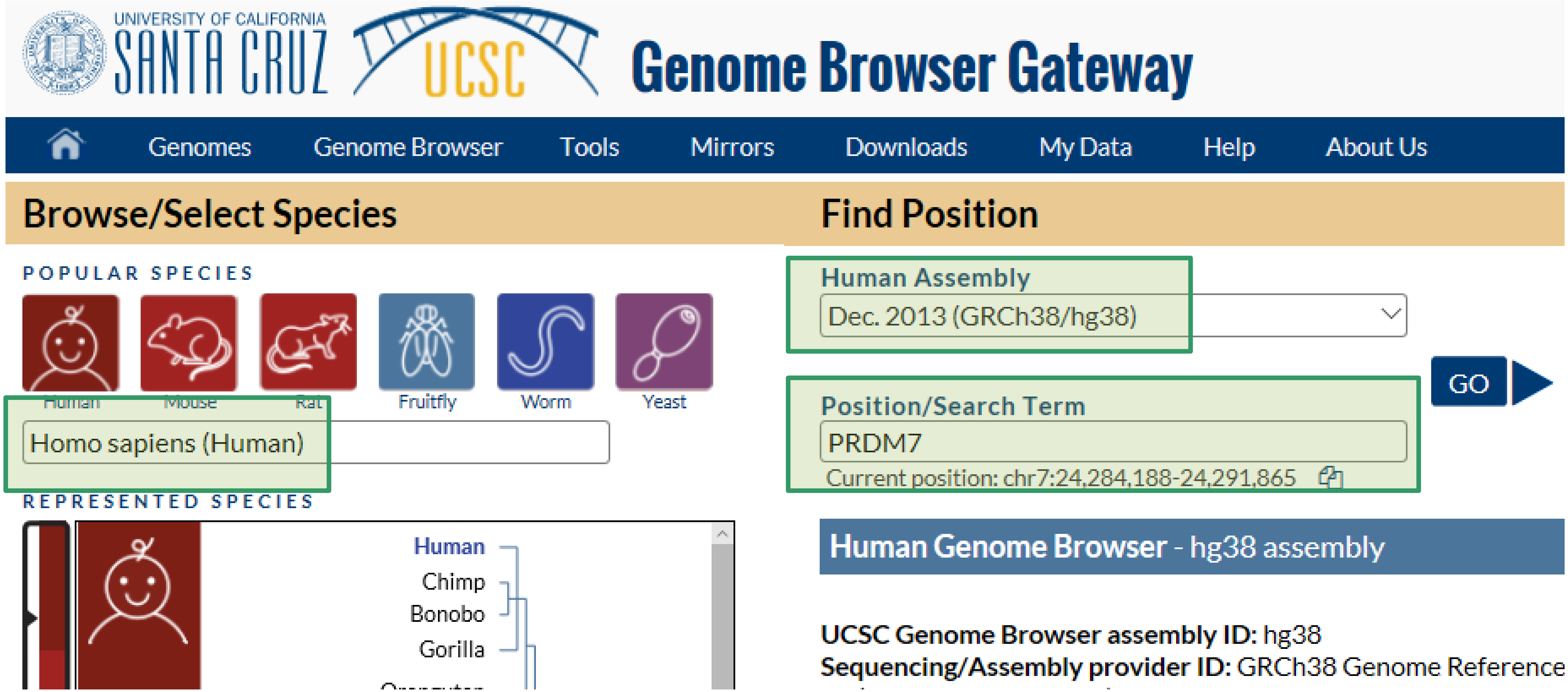
Our tools

- **Genome Browser**
  interactively visualize genomic data
- **BLAT**
  rapidly align sequences to the genome
- **Table Browser**
  download data from the Genome Browser database
- **Variant Annotation Integrator**
  get functional effect predictions for variant calls
- **Data Integrator**
  combine data sources from the Genome Browser database
- **Gene Sorter**
  find genes that are similar by expression and other metrics
- **Genome Browser in a Box (GBiB)**
  run the Genome Browser on your laptop or server
- **In-Silico PCR**
  rapidly align PCR primer pairs to the genome
- **LiftOver**
  convert genome coordinates between assemblies
- **VisiGene**
  interactively view in situ images of mouse and frog

More tools...

UCSC Genome Browser

Search for the **human** protein **PRDM7** using the **newest genome assembly**.



Homo sapiens (Human)

Human Assembly: Dec. 2013 (GRCh38/hg38)

Position/Search Term: PRDM7

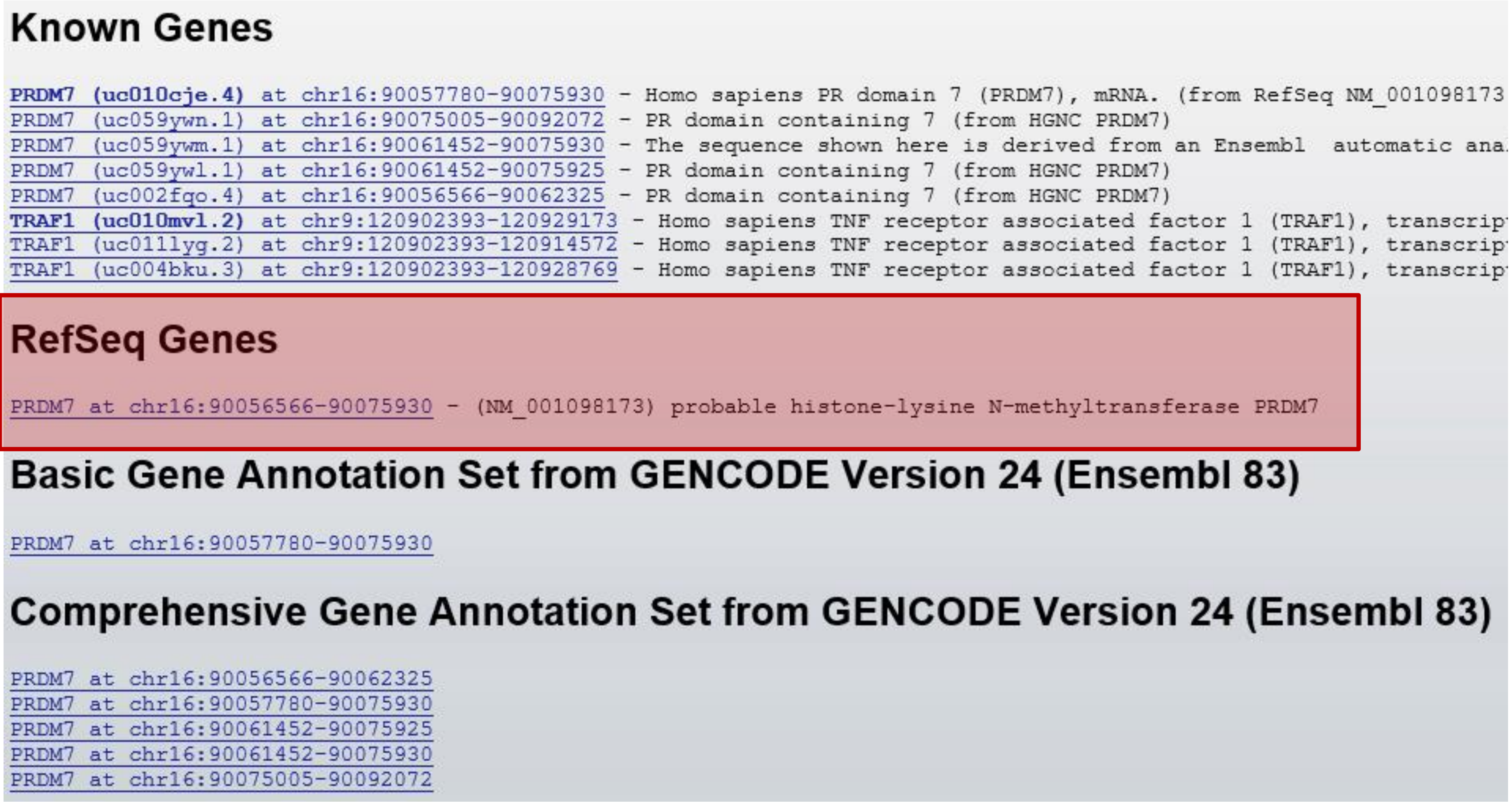
Current position: chr7:24,284,188-24,291,865

Human Genome Browser - hg38 assembly

UCSC Genome Browser assembly ID: hg38
Sequencing/Assembly provider ID: GRCh38 Genome Reference

Theresa Schwarz

UCSC Genome Browser



Known Genes

PRDM7 (uc010cje.4) at chr16:90057780-90075930 - Homo sapiens PR domain 7 (PRDM7), mRNA. (from RefSeq NM_001098173

PRDM7 (uc059ywn.1) at chr16:90075005-90092072 - PR domain containing 7 (from HGNC PRDM7)

PRDM7 (uc059ywm.1) at chr16:90061452-90075930 - The sequence shown here is derived from an Ensembl automatic ana

PRDM7 (uc059ywl.1) at chr16:90061452-90075925 - PR domain containing 7 (from HGNC PRDM7)

PRDM7 (uc002fqo.4) at chr16:90056566-90062325 - PR domain containing 7 (from HGNC PRDM7)

TRAF1 (uc010mvl.2) at chr9:120902393-120929173 - Homo sapiens TNF receptor associated factor 1 (TRAF1), transcrip

TRAF1 (uc011lyg.2) at chr9:120902393-120914572 - Homo sapiens TNF receptor associated factor 1 (TRAF1), transcrip

TRAF1 (uc004bku.3) at chr9:120902393-120928769 - Homo sapiens TNF receptor associated factor 1 (TRAF1), transcrip

RefSeq Genes

PRDM7 at chr16:90056566-90075930 - (NM_001098173) probable histone-lysine N-methyltransferase PRDM7

Basic Gene Annotation Set from GENCODE Version 24 (Ensembl 83)

PRDM7 at chr16:90057780-90075930

Comprehensive Gene Annotation Set from GENCODE Version 24 (Ensembl 83)

PRDM7 at chr16:90056566-90062325

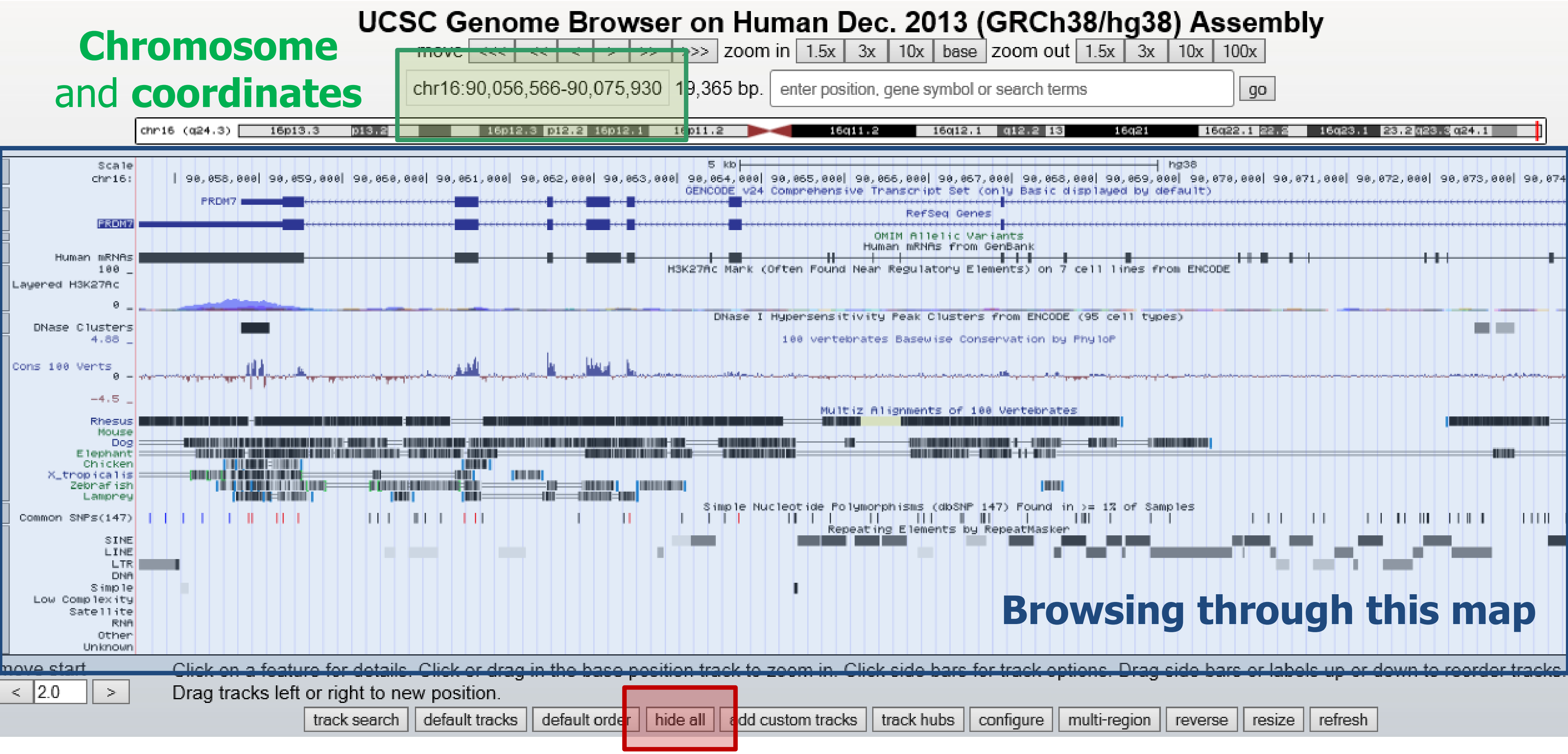
PRDM7 at chr16:90057780-90075930

PRDM7 at chr16:90061452-90075925

PRDM7 at chr16:90061452-90075930

PRDM7 at chr16:90075005-90092072

UCSC Genome Browser



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

chr16:90,056,566-90,075,930 19,365 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Chromosome and **coordinates**

Browsing through this map

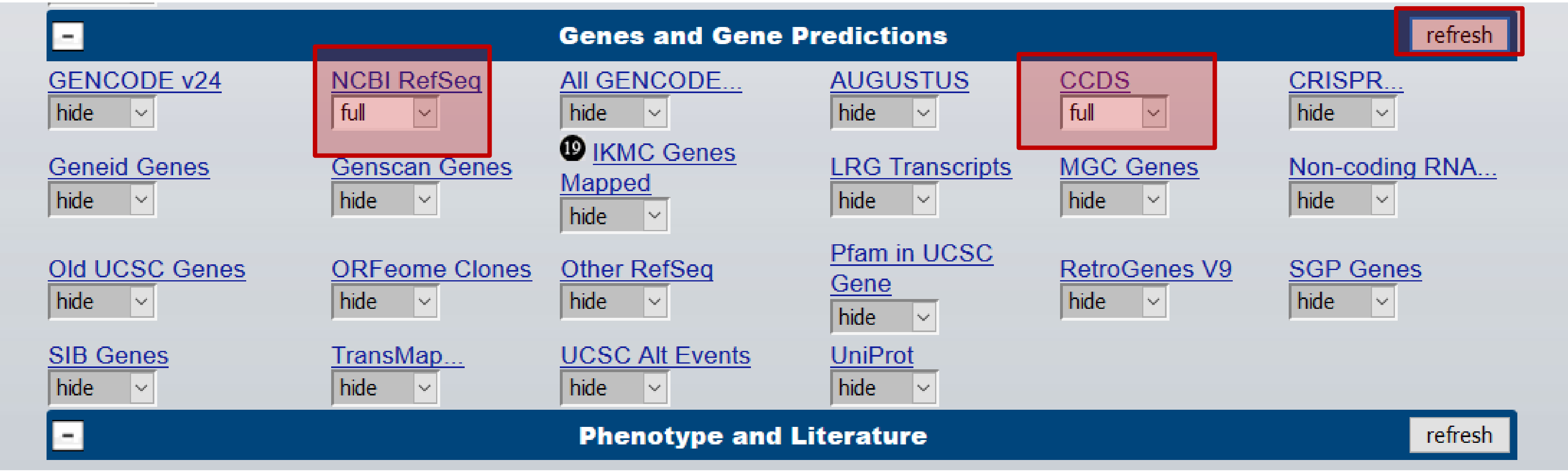
You can **hide all** tracks

→ **Scroll down to select your settings**

UCSC Genome Browser

You can choose all 'tracks' (settings) you want.

- Choose:
  - **NCBI RefSeq (full)**
  - **CCDS (full)**
- To update your settings select '**refresh**'

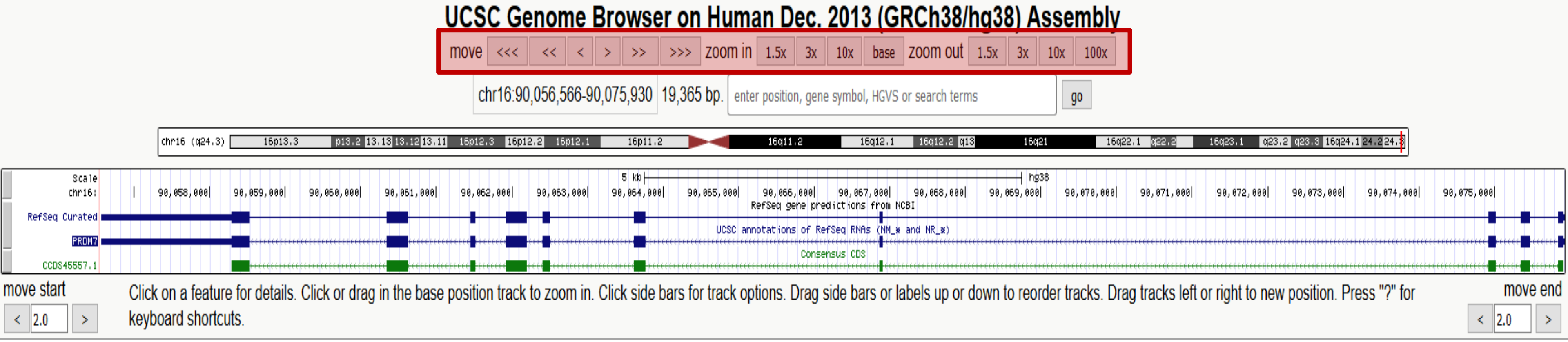


Genes and Gene Predictions — refresh

GENCODE v24: hide	NCBI RefSeq: full	All GENCODE...: hide	AUGUSTUS: hide	CCDS: full	CRISPR...: hide
Geneid Genes: hide	Genscan Genes: hide	19 IKMC Genes Mapped: hide	LRG Transcripts: hide	MGC Genes: hide	Non-coding RNA...: hide
Old UCSC Genes: hide	ORFeome Clones: hide	Other RefSeq: hide	Pfam in UCSC Gene: hide	RetroGenes V9: hide	SGP Genes: hide
SIB Genes: hide	TransMap...: hide	UCSC Alt Events: hide	UniProt: hide		

Phenotype and Literature — refresh

UCSC Genome Browser

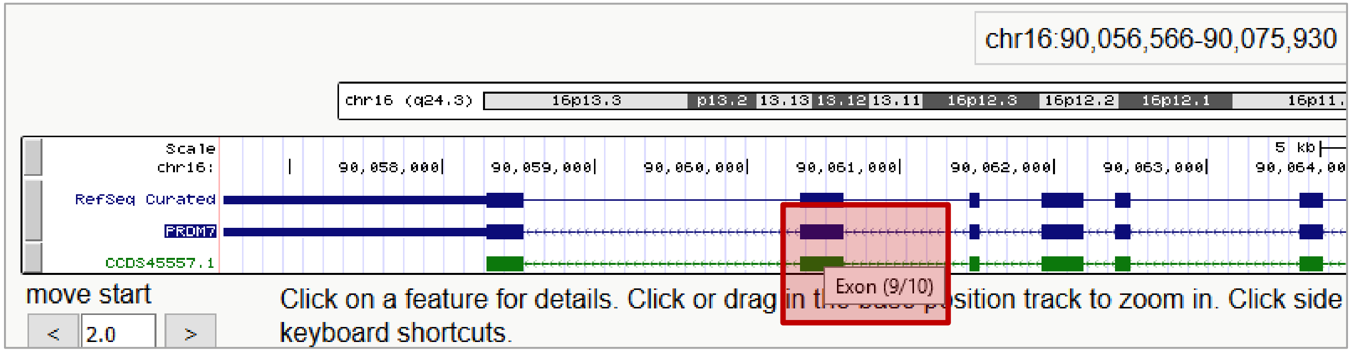


UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

chr16:90,056,566-90,075,930 19,365 bp.

enter position, gene symbol, HGVS or search terms

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts.



chr16:90,056,566-90,075,930

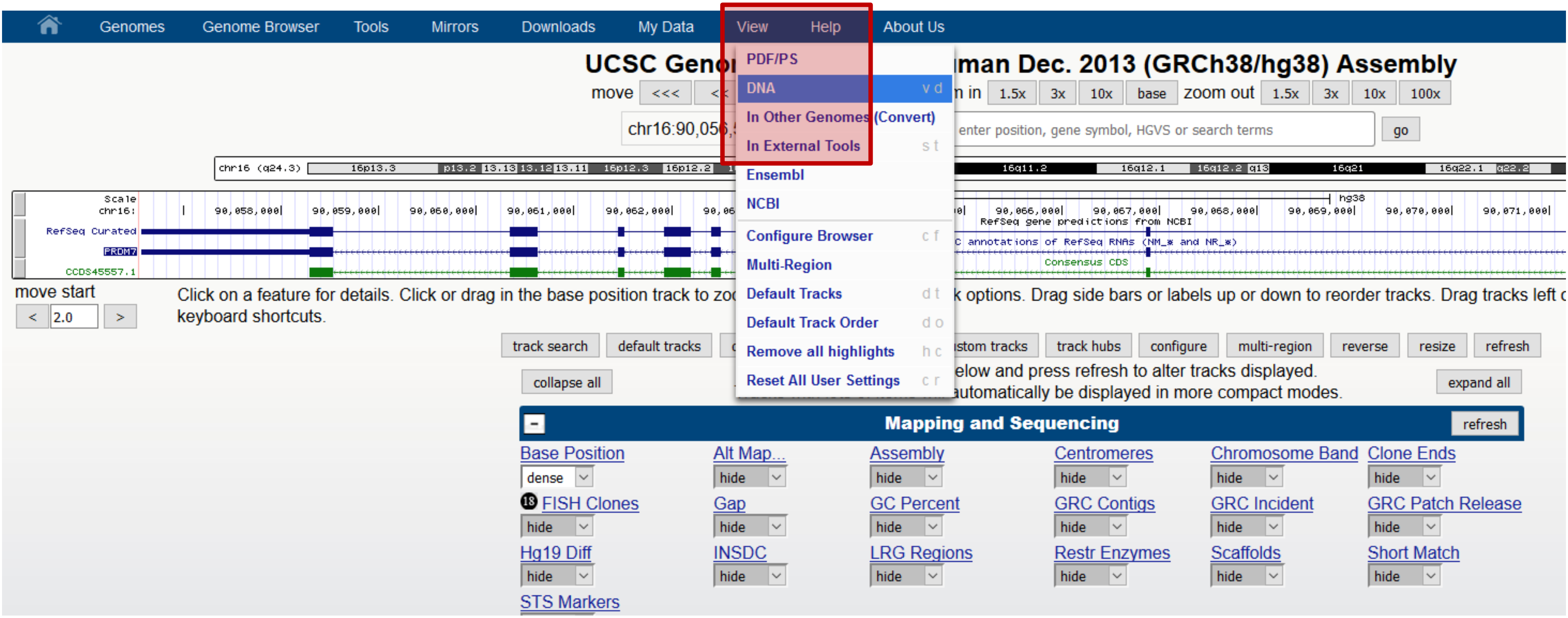
Click on a feature for details. Click or drag in the base position track to zoom in. Click side keyboard shortcuts.

You can see your selected tracks

Boxes = Exons; Lines = Introns

Navigate through the map by **moving** and **zooming**.

UCSC Genome Browser



UCSC Geno… man Dec. 2013 (GRCh38/hg38) Assembly

chr16:90,056,…

enter position, gene symbol, HGVS or search terms

Click on a feature for details. Click or drag in the base position track to zoo… k options. Drag side bars or labels up or down to reorder tracks. Drag tracks left o… keyboard shortcuts.

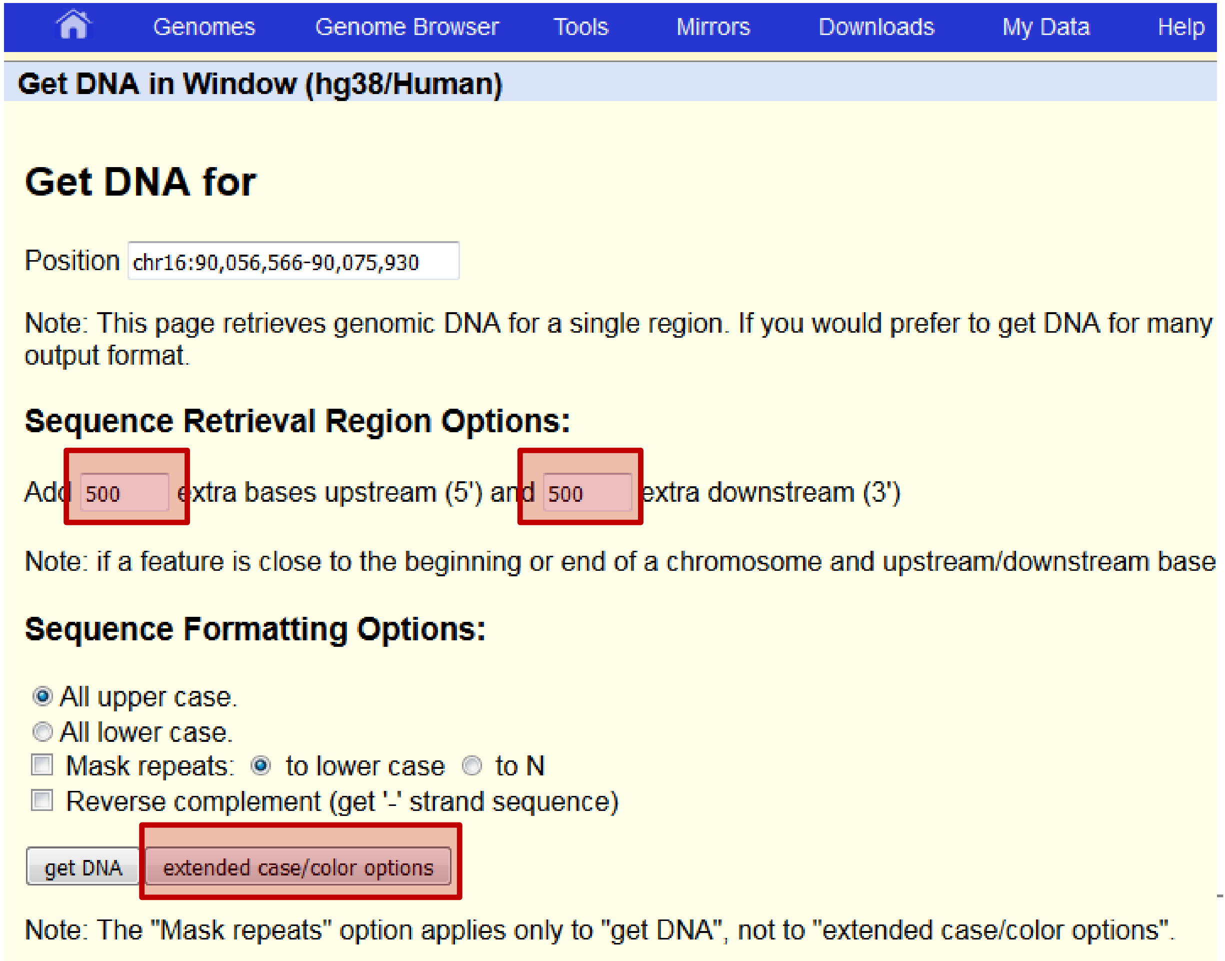
…elow and press refresh to alter tracks displayed. …automatically be displayed in more compact modes.

Base Position: dense
Alt Map…: hide
Assembly: hide
Centromeres: hide
Chromosome Band: hide
Clone Ends: hide
FISH Clones: hide
Gap: hide
GC Percent: hide
GRC Contigs: hide
GRC Incident: hide
GRC Patch Release: hide
Hg19 Diff: hide
INSDC: hide
LRG Regions: hide
Restr Enzymes: hide
Scaffolds: hide
Short Match: hide
STS Markers

If you would like to get the DNA sequence click on **View → DNA**

UCSC Genome Browser

- You can add bases up- and/or downstream of the DNA: e.g. 500bp
- You can highlight the '**tracks**' you selected by '**extended case/color options**'



Get DNA in Window (hg38/Human)

Get DNA for

Position chr16:90,056,566-90,075,930

Note: This page retrieves genomic DNA for a single region. If you would prefer to get DNA for many output format.

Sequence Retrieval Region Options:

Add 500 extra bases upstream (5') and 500 extra downstream (3')

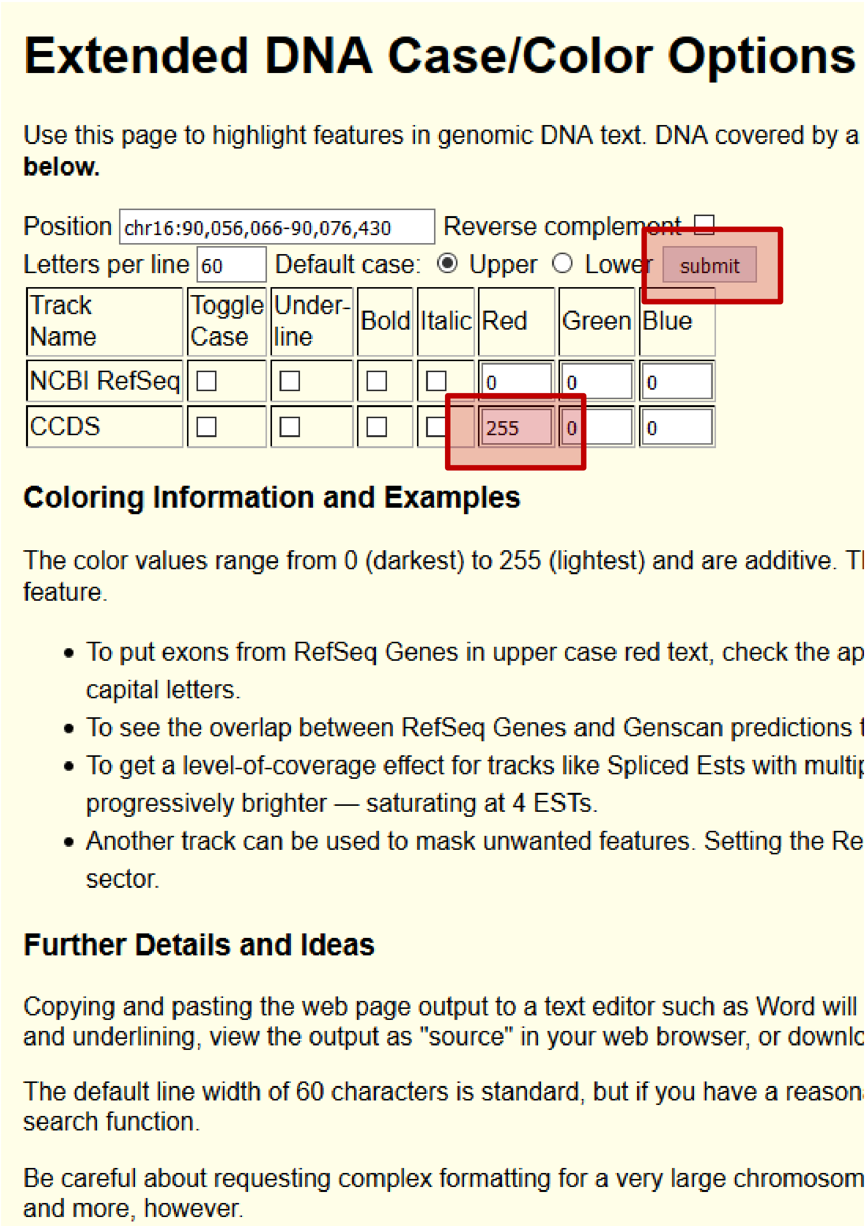
Note: if a feature is close to the beginning or end of a chromosome and upstream/downstream base

Sequence Formatting Options:

● All upper case.
○ All lower case.
□ Mask repeats: ● to lower case ○ to N
□ Reverse complement (get '-' strand sequence)

Note: The "Mask repeats" option applies only to "get DNA", not to "extended case/color options".

UCSC Genome Browser



Extended DNA Case/Color Options

Use this page to highlight features in genomic DNA text. DNA covered by a
below.

Position chr16:90,056,066-90,076,430 Reverse complement □
Letters per line 60 Default case: ● Upper ○ Lower submit

Track Name	Toggle Case	Under-line	Bold	Italic	Red	Green	Blue
NCBI RefSeq	□	□	□	□	0	0	0
CCDS	□	□	□	□	255	0	0

Coloring Information and Examples

The color values range from 0 (darkest) to 255 (lightest) and are additive. Tl
feature.

- To put exons from RefSeq Genes in upper case red text, check the ap capital letters.
- To see the overlap between RefSeq Genes and Genscan predictions t
- To get a level-of-coverage effect for tracks like Spliced Ests with multip progressively brighter — saturating at 4 ESTs.
- Another track can be used to mask unwanted features. Setting the Re sector.

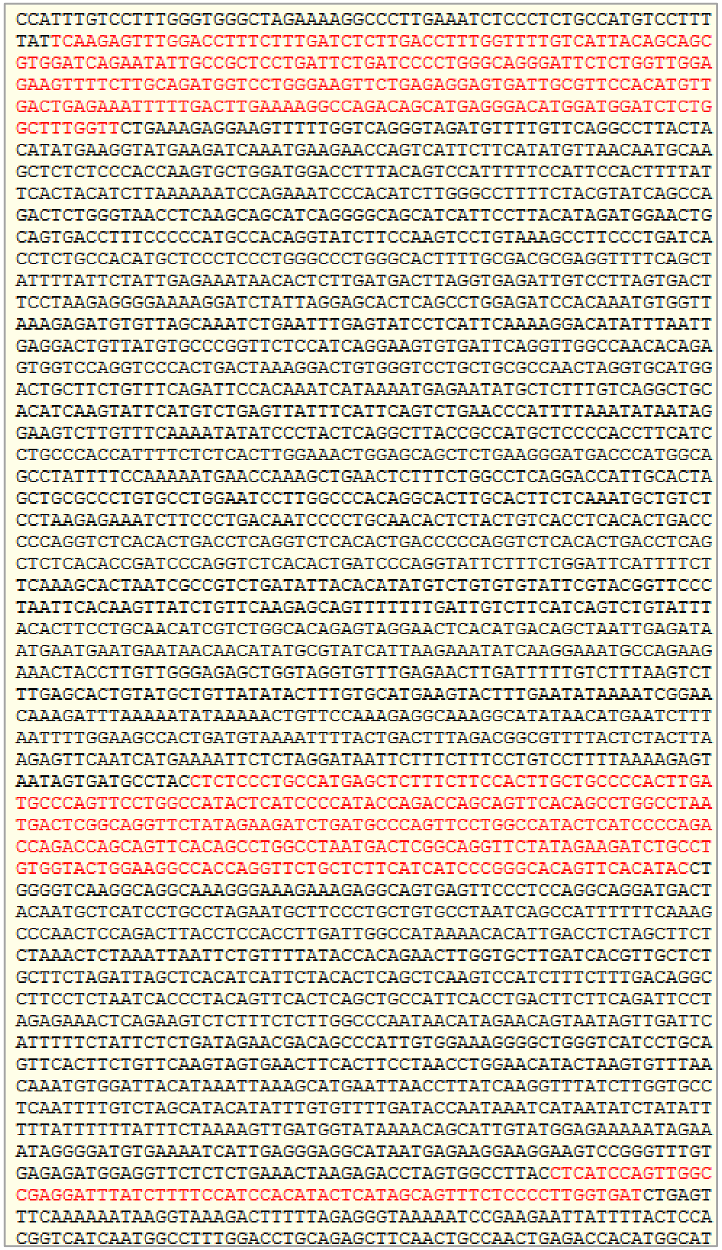
Further Details and Ideas

Copying and pasting the web page output to a text editor such as Word will
and underlining, view the output as "source" in your web browser, or downlo

The default line width of 60 characters is standard, but if you have a reason
search function.

Be careful about requesting complex formatting for a very large chromosom
and more, however.

UCSC Genome Browser



```
CCATTTGTCCTTTGGGTGGGCTAGAAAAGGCCCTTGAAATCTCCCTCTGCCATGTCCTTT
TATTCAAGAGTTTGGACCTTTCTTTGATCTCTTGACCTTTGGTTTTGTCATTACAGCAGC
GTGGATCAGAATATTGCCGCTCCTGATTCTGATCCCCTGGGCAGGGATTCTCTGGTTGGA
GAAGTTTTCTTGCAGATGGTCCTGGGAAGTTCTGAGAGGAGTGATTGCGTTCCACATGTT
GACTGAGAAATTTTTGACTTGAAAAGGCCAGACAGCATGAGGGACATGGATGGATCTCTG
GCTTTGGTTCTGAAAGAGGAAGTTTTTGGTCAGGGTAGATGTTTTGTTCAGGCCTTACTA
CATATGAAGGTATGAAGATCAAATGAAGAACCAGTCATTCTTCATATGTTAACAATGCAA
GCTCTCTCCCACCAAGTGCTGGATGGACCTTTACAGTCCATTTTTCCATTCCACTTTTAT
TCACTACATCTTAAAAAATCCAGAAATCCCACATCTTGGGCCTTTTCTACGTATCAGCCA
GACTCTGGGTAACCTCAAGCAGCATCAGGGGCAGCATCATTCCTTACATAGATGGAACTG
CAGTGACCTTTCCCCCATGCCACAGGTATCTTCCAAGTCCTGTAAAGCCTTCCCTGATCA
CCTCTGCCACATGCTCCCTCCCTGGGCCCTGGGCACTTTTGCGACGCGAGGTTTTCAGCT
ATTTTATTCTATTGAGAAATAACACTCTTGATGACTTAGGTGAGATTGTCCTTAGTGACT
TCCTAAGAGGGGAAAAGGATCTATTAGGAGCACTCAGCCTGGAGATCCACAAATGTGGTT
AAAGAGATGTGTTAGCAAATCTGAATTTGAGTATCCTCATTCAAAAGGACATATTTAATT
GAGGACTGTTATGTGCCCGGTTCTCCATCAGGAAGTGTGATTCAGGTTGGCCAACACAGA
GTGGTCCAGGTCCCACTGACTAAAGGACTGTGGGTCCTGCTGCGCCAACTAGGTGCATGG
ACTGCTTCTGTTTCAGATTCCACAAATCATAAAATGAGAATATGCTCTTTGTCAGGCTGC
ACATCAAGTATTCATGTCTGAGTTATTTCATTCAGTCTGAACCCATTTTAAATATAATAG
GAAGTCTTGTTTCAAAATATATCCCTACTCAGGCTTACCGCCATGCTCCCCACCTTCATC
CTGCCCACCATTTTCTCTCACTTGGAAACTGGAGCAGCTCTGAAGGGATGACCCATGGCA
GCCTATTTTCCAAAAATGAACCAAAGCTGAACTCTTTCTGGCCTCAGGACCATTGCACTA
GCTGCGCCCTGTGCCTGGAATCCTTGGCCCACAGGCACTTGCACTTCTCAAATGCTGTCT
CCTAAGAGAAATCTTCCCTGACAATCCCCTGCAACACTCTACTGTCACCTCACACTGACC
CCCAGGTCTCACACTGACCTCAGGTCTCACACTGACCCCCAGGTCTCACACTGACCTCAG
CTCTCACACCGATCCCAGGTCTCACACTGATCCCAGGTATTCTTTCTGGATTCATTTTCT
TCAAAGCACTAATCGCCGTCTGATATTACACATATGTCTGTGTGTATTCGTACGGTTCCC
TAATTCACAAGTTATCTGTTCAAGAGCAGTTTTTTTGATTGTCTTCATCAGTCTGTATTT
ACACTTCCTGCAACATCGTCTGGCACAGAGTAGGAACTCACATGACAGCTAATTGAGATA
ATGAATGAATGAATAACAACATATGCGTATCATTAAGAAATATCAAGGAAATGCCAGAAG
AAACTACCTTGTTGGGAGAGCTGGTAGGTGTTTGAGAACTTGATTTTTGTCTTTAAGTCT
TTGAGCACTGTATGCTGTTATATACTTTGTGCATGAAGTACTTTGAATATAAAATCGGAA
CAAAGATTTAAAAATATAAAAACTGTTCCAAAGAGGCAAAGGCATATAACATGAATCTTT
AATTTTGGAAGCCACTGATGTAAAATTTTACTGACTTTAGACGGCGTTTTACTCTACTTA
AGAGTTCAATCATGAAAATTCTCTAGGATAATTCTTTCTTTCCTGTCCTTTTAAAAGAGT
AATAGTGATGCCTACCTCTCCCTGCCATGAGCTCTTTCTTCCACTTGCTGCCCCACTTGA
TGCCCAGTTCCTGGCCATACTCATCCCCATACCAGACCAGCAGTTCACAGCCTGGCCTAA
TGACTCGGCAGGTTCTATAGAAGATCTGATGCCCAGTTCCTGGCCATACTCATCCCCAGA
CCAGACCAGCAGTTCACAGCCTGGCCTAATGACTCGGCAGGTTCTATAGAAGATCTGCCT
GTGGTACTGGAAGGCCACCAGGTTCTGCTCTTCATCATCCCGGGCACAGTTCACATACCT
GGGGTCAAGGCAGGCAAAGGGAAAGAAAGAGGCAGTGAGTTCCCTCCAGGCAGGATGACT
ACAATGCTCATCCTGCCTAGAATGCTTCCCTGCTGTGCCTAATCAGCCATTTTTTCAAAG
CCCAACTCCAGACTTACCTCCACCTTGATTGGCCATAAAACACATTGACCTCTAGCTTCT
CTAAACTCTAAATTAATTCTGTTTTATACCACAGAACTTGGTGCTTGATCACGTTGCTCT
GCTTCTAGATTAGCTCACATCATTCTACACTCAGCTCAAGTCCATCTTTCTTTGACAGGC
CTTCCTCTAATCACCCTACAGTTCACTCAGCTGCCATTCACCTGACTTCTTCAGATTCCT
AGAGAAACTCAGAAGTCTCTTTCTCTTGGCCCAATAACATAGAACAGTAATAGTTGATTC
ATTTTTCTATTCTCTGATAGAACGACAGCCCATTGTGGAAAGGGGCTGGGTCATCCTGCA
GTTCACTTCTGTTCAAGTAGTGAACTTCACTTCCTAACCTGGAACATACTAAGTGTTTAA
CAAATGTGGATTACATAAATTAAAGCATGAATTAACCTTATCAAGGTTTATCTTGGTGCC
TCAATTTTGTCTAGCATACATATTTGTGTTTTGATACCAATAAATCATAATATCTATATT
TTTATTTTTTATTTCTAAAAGTTGATGGTATAAAACAGCATTGTATGGAGAAAAATAGAA
ATAGGGGATGTGAAAATCATTGAGGGAGGCATAATGAGAAGGAAGGAAGTCCGGGTTTGT
GAGAGATGGAGGTTCTCTCTGAAACTAAGAGACCTAGTGGCCTTACCTCATCCAGTTGGC
CGAGGATTTATCTTTTCCATCCACATACTCATAGCAGTTTCTCCCCTTGGTGATCTGAGT
TTCAAAAAATAAGGTAAAGACTTTTTAGAGGGTAAAAATCCGAAGAATTATTTTACTCCA
CGGTCATCAATGGCCTTTGGACCTGCAGAGCTTCAACTGCCAACTGAGACCACATGGCAT
```

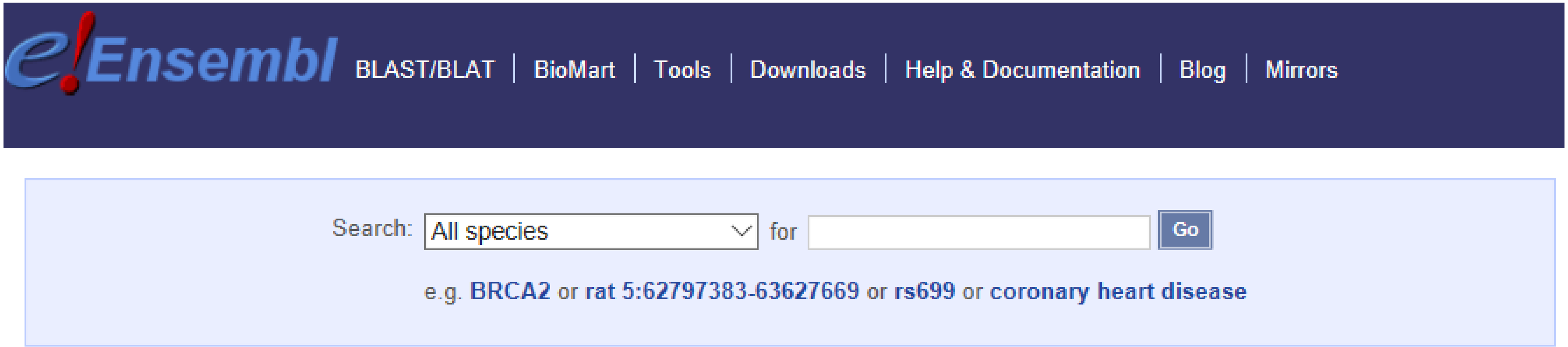
CCDS is highlighted in red.

Theresa Schwarz

Overview

- ❑ **Genome browsers**
  - ➢ **UCSC**
  - ➢ **ENSEMBL**
- ❑ **BLAST**
  - ➢ **Pairwise alignments**
  - ➢ **Database alignments**
  - ➢ **Primer-BLAST**

ENSEMBL Genome Browser



Search: All species for

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

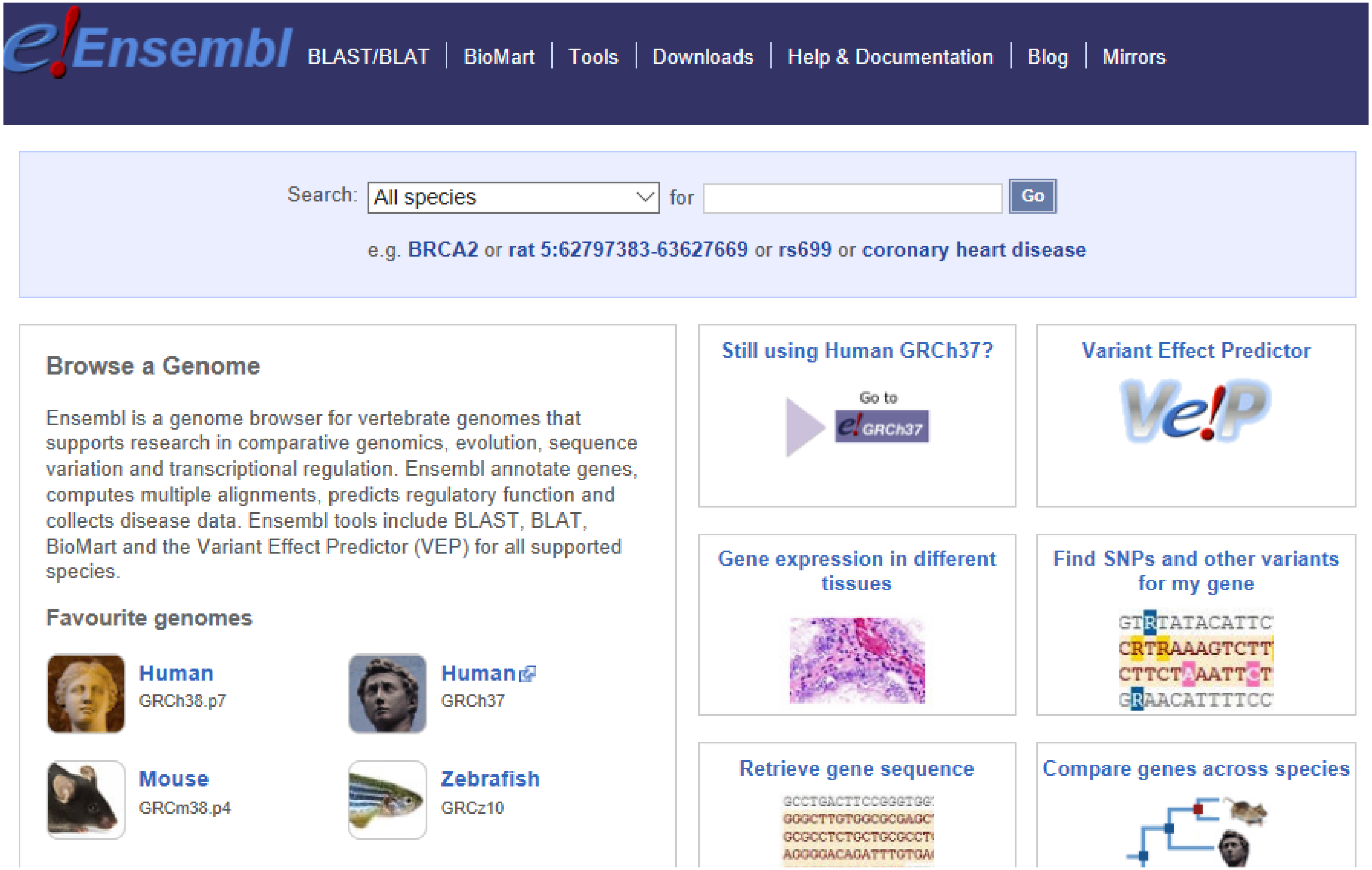
- **ENSEMBL** is a project between EMBL-EBI (European Bioinformatics Institute) and the Wellcome Trust Sanger Institute.

- This Genome Browser provides information about **eukaryotic genomes.**

- You can find an accurate description of **protein coding genes, promoters, exons, introns, transcripts, …**

ENSEMBL Genome Browser

http://www.ensembl.org/


e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Search: All species for Go

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

Browse a Genome

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

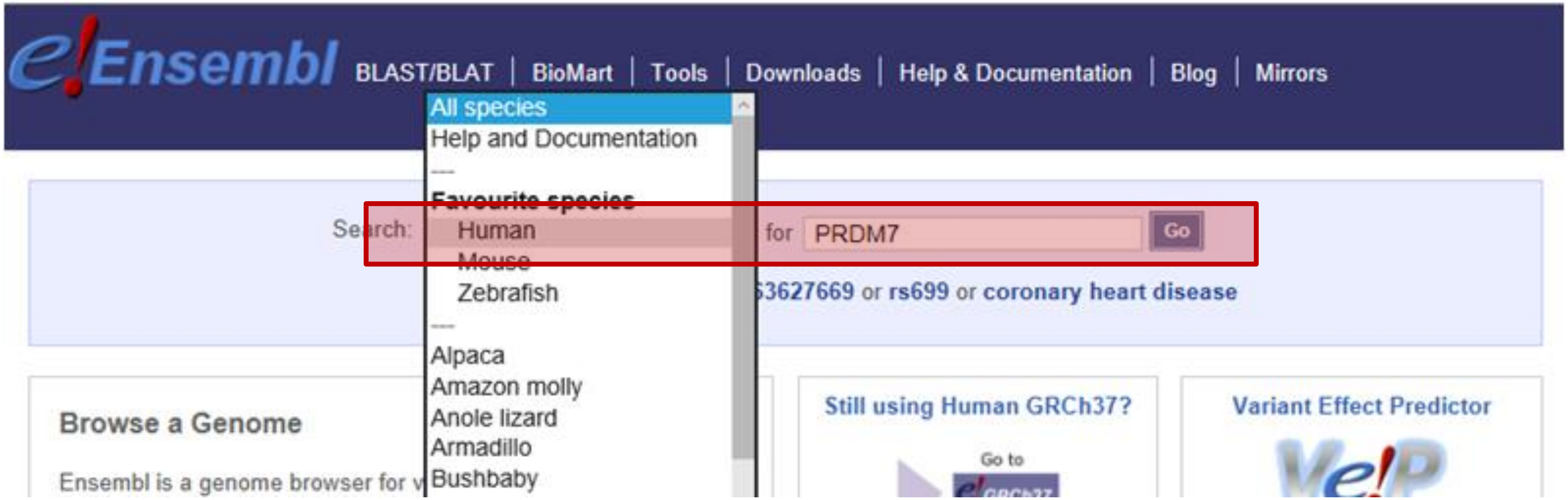
Favourite genomes

Human GRCh38.p7	Human GRCh37
Mouse GRCm38.p4	Zebrafish GRCz10

Still using Human GRCh37?
Go to
e! GRCh37
Variant Effect Predictor
Ve!P
Gene expression in different tissues
Find SNPs and other variants for my gene
Retrieve gene sequence
Compare genes across species

Search again for human PRDM7

ENSEMBL



Search: Human for PRDM7

3627669 or rs699 or coronary heart disease

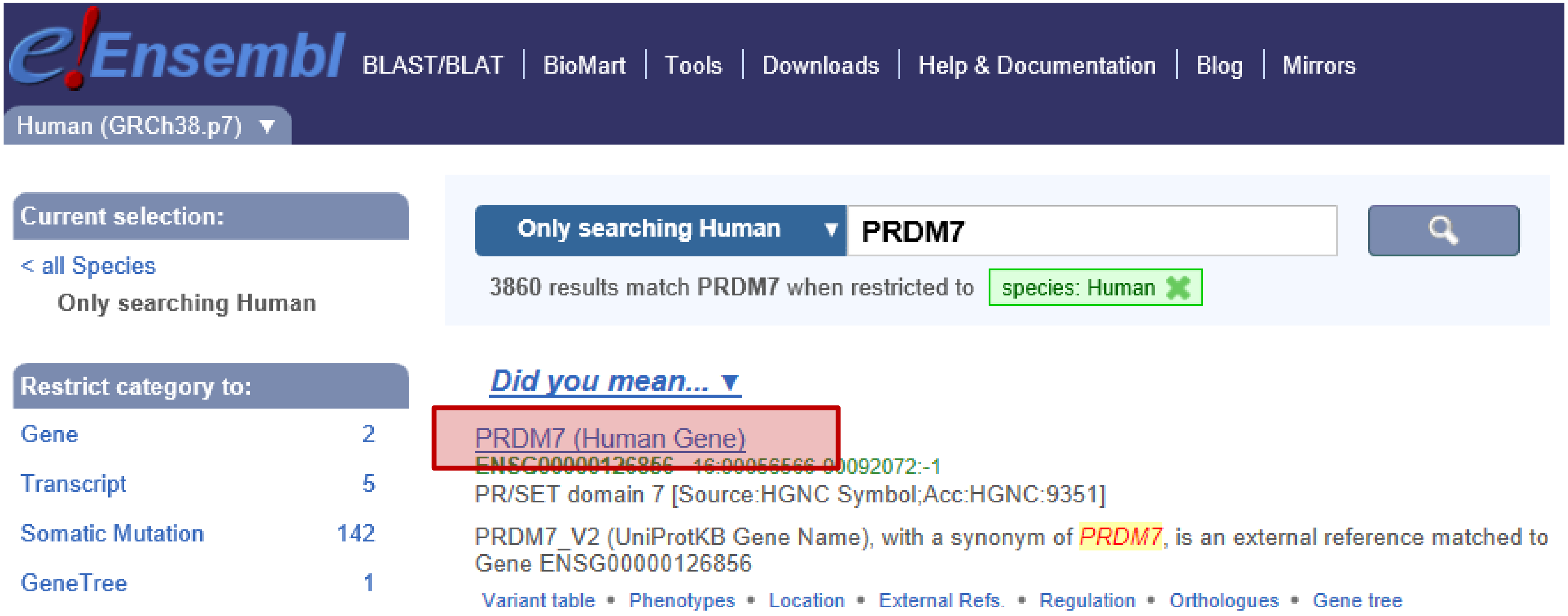
Browse a Genome

Ensembl is a genome browser for v

Still using Human GRCh37?

Go to

Variant Effect Predictor


e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Human (GRCh38.p7) ▼

Current selection:

< all Species

Only searching Human

Restrict category to:

Gene	2
Transcript	5
Somatic Mutation	142
GeneTree	1

Only searching Human ▼ PRDM7

3860 results match PRDM7 when restricted to species: Human ✖

Did you mean... ▼

PRDM7 (Human Gene)

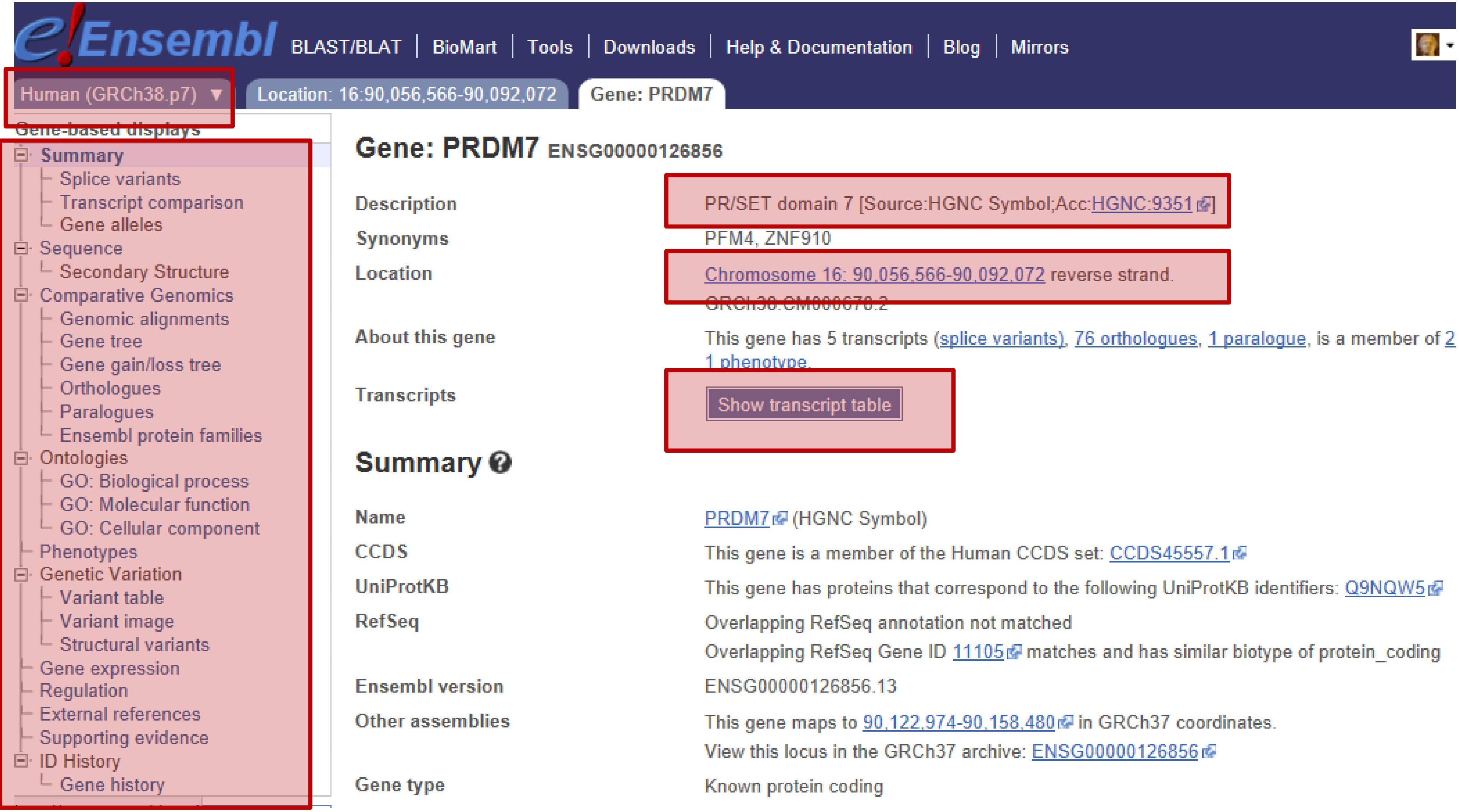
ENSG00000126856 16:90056566-90092072:-1

PR/SET domain 7 [Source:HGNC Symbol;Acc:HGNC:9351]

PRDM7_V2 (UniProtKB Gene Name), with a synonym of *PRDM7*, is an external reference matched to Gene ENSG00000126856

Variant table • Phenotypes • Location • External Refs. • Regulation • Orthologues • Gene tree

ENSEMBL



Gene: PRDM7 ENSG00000126856

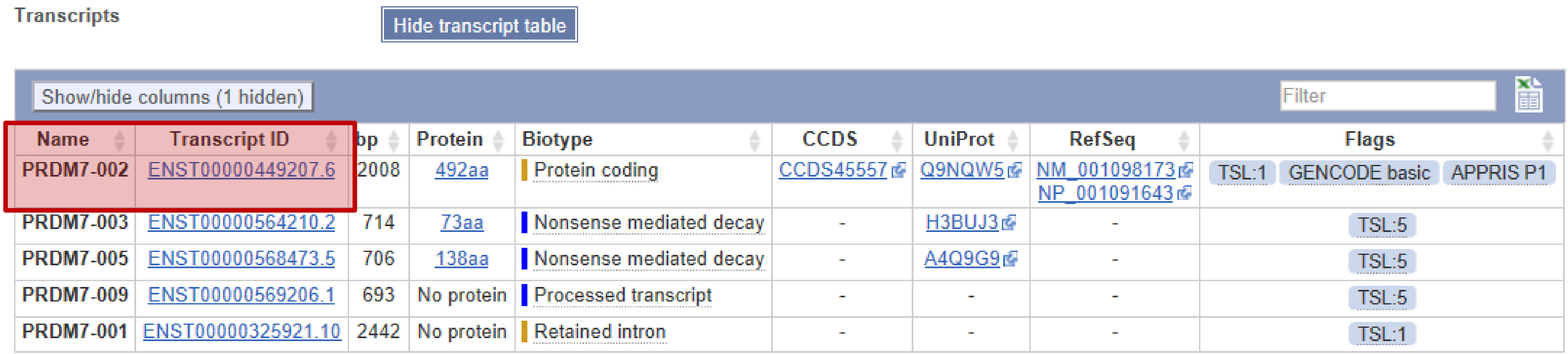
Description	PR/SET domain 7 [Source:HGNC Symbol;Acc:HGNC:9351]
Synonyms	PFM4, ZNF910
Location	Chromosome 16: 90,056,566-90,092,072 reverse strand. GRCh38:CM000678.2
About this gene	This gene has 5 transcripts (splice variants), 76 orthologues, 1 paralogue, is a member of 2 1 phenotype.
Transcripts	Show transcript table

Summary

Name	PRDM7 (HGNC Symbol)
CCDS	This gene is a member of the Human CCDS set: CCDS45557.1
UniProtKB	This gene has proteins that correspond to the following UniProtKB identifiers: Q9NQW5
RefSeq	Overlapping RefSeq annotation not matched Overlapping RefSeq Gene ID 11105 matches and has similar biotype of protein_coding
Ensembl version	ENSG00000126856.13
Other assemblies	This gene maps to 90,122,974-90,158,480 in GRCh37 coordinates. View this locus in the GRCh37 archive: ENSG00000126856
Gene type	Known protein coding

→ Click on **Show transcript table**

ENSEMBL



Transcripts

Hide transcript table

Show/hide columns (1 hidden)

Filter

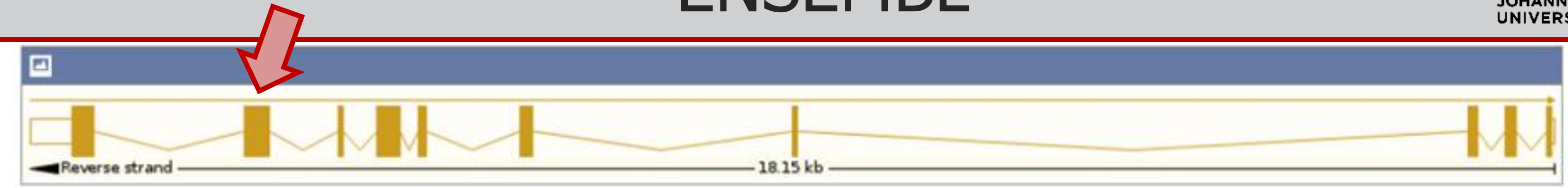
Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt	RefSeq	Flags
PRDM7-002	ENST00000449207.6	2008	492aa	Protein coding	CCDS45557	Q9NQW5	NM_001098173 NP_001091643	TSL:1 GENCODE basic APPRIS P1
PRDM7-003	ENST00000564210.2	714	73aa	Nonsense mediated decay	-	H3BUJ3	-	TSL:5
PRDM7-005	ENST00000568473.5	706	138aa	Nonsense mediated decay	-	A4Q9G9	-	TSL:5
PRDM7-009	ENST00000569206.1	693	No protein	Processed transcript	-	-	-	TSL:5
PRDM7-001	ENST00000325921.10	2442	No protein	Retained intron	-	-	-	TSL:1

Here you can see all transcript versions and some links to other databases like CCDS, UniProt.

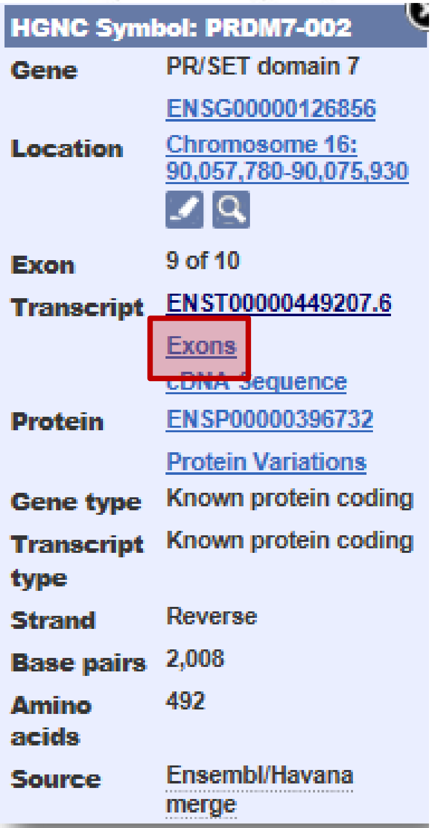
For the first transcript you can also get the NCBI's **RefSeq** sequences for nucleotide and protein.

→ **Select the first entry and scroll down!**

ENSEMBL



Statistics	Exons: 10, Coding exons: 10, Transcript length: 2,008 bps, Translation length: 492 residues
CCDS	This transcript is a member of the Human CCDS set: CCDS45557
Uniprot	This transcript corresponds to the following Uniprot identifiers: Q9NQW5
Transcript Support Level (TSL)	TSL:1
Ensembl version	ENST00000449207.6
Type	Known protein coding
Annotation Method	Transcript where the Ensembl genebuild transcript and the Vega manual annotation have the same sequence, for every base pair. See article.
Alternative transcripts	This transcript corresponds to the following database identifiers: Havana transcript: OTTHUMT00000420560
GENCODE basic gene	This transcript is a member of the Gencode basic gene set.



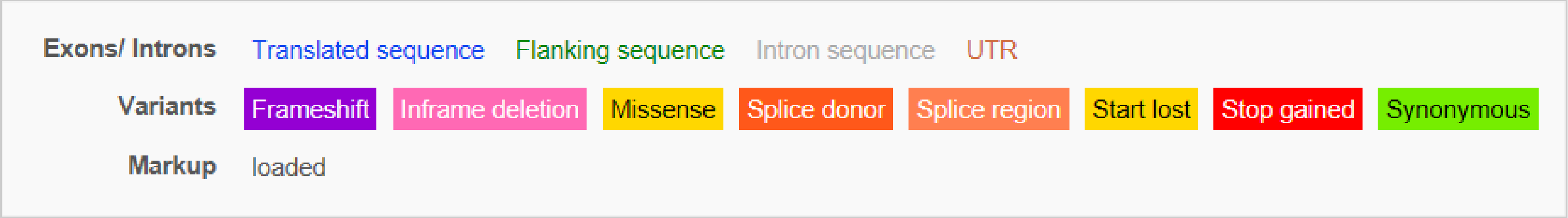
HGNC Symbol: PRDM7-002

Gene	PR/SET domain 7 ENSG00000126856
Location	Chromosome 16: 90,057,780-90,075,930
Exon	9 of 10
Transcript	ENST00000449207.6 Exons cDNA Sequence
Protein	ENSP00000396732 Protein Variations
Gene type	Known protein coding
Transcript type	Known protein coding
Strand	Reverse
Base pairs	2,008
Amino acids	492
Source	Ensembl/Havana merge

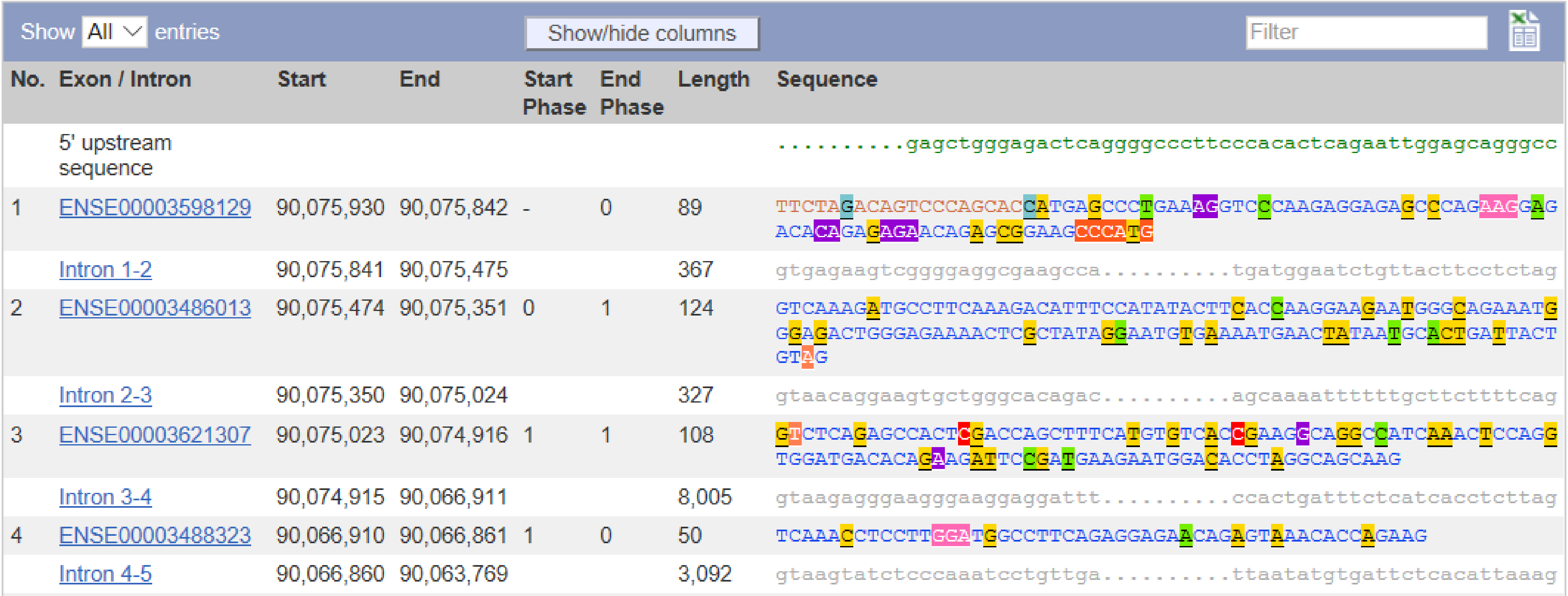
→ **To get more information about exons and introns, click on a box and select 'Exons'**

ENSEMBL

Here you can obtain sequence information on **exons** and **introns.**



Exons/ Introns	Translated sequence · Flanking sequence · Intron sequence · UTR
Variants	Frameshift · Inframe deletion · Missense · Splice donor · Splice region · Start lost · Stop gained · Synonymous
Markup	loaded

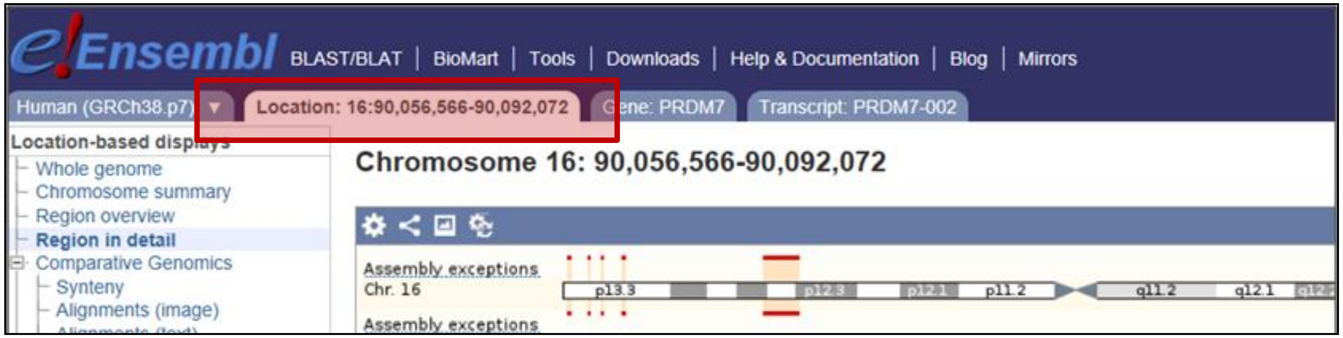


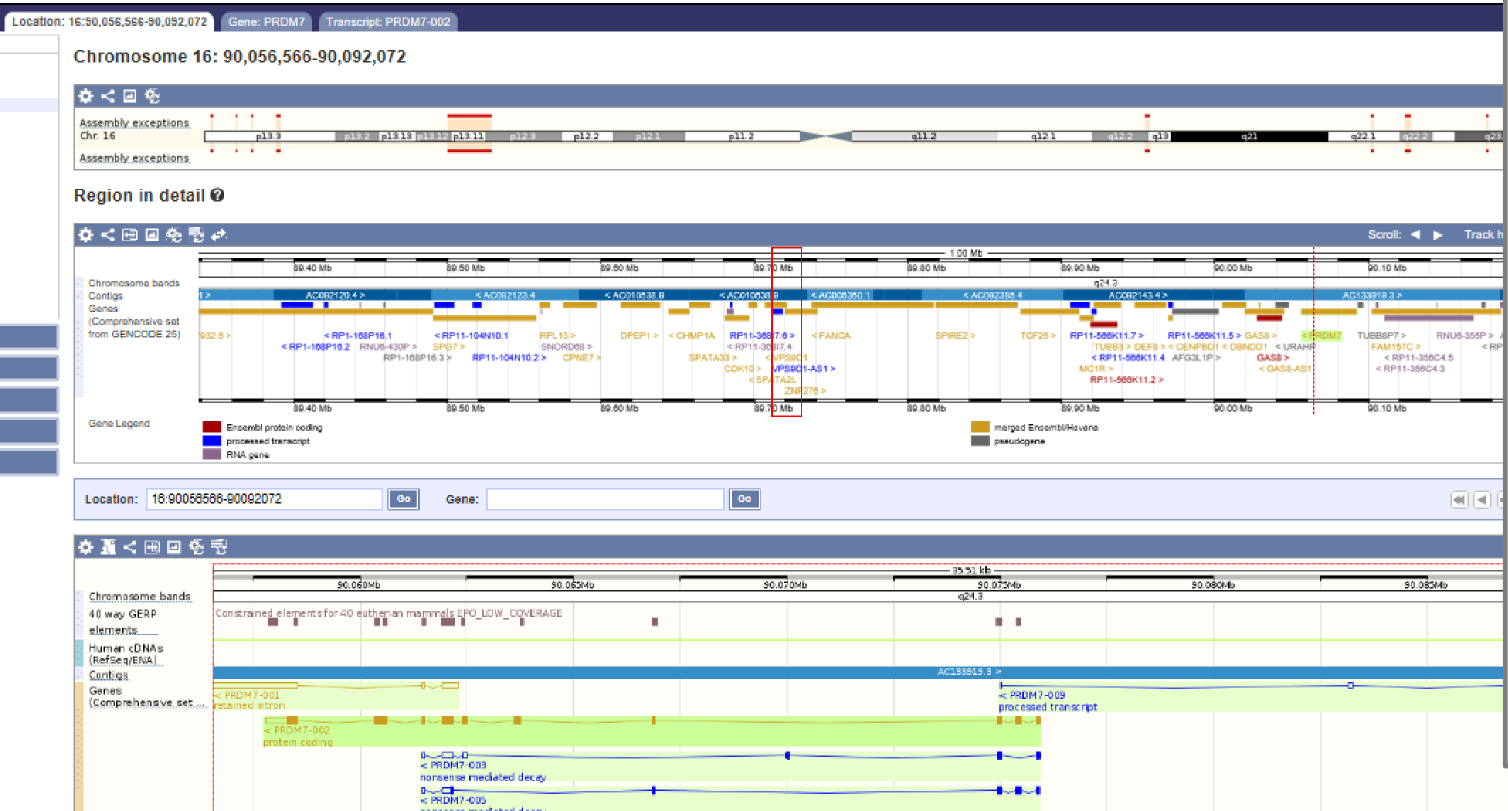
Show All entries · Show/hide columns · Filter

No.	Exon / Intron	Start	End	Start Phase	End Phase	Length	Sequence
	5' upstream sequence						..........gagctgggagactcaggggcccttcccacactcagaattggagcagggcc
1	ENSE00003598129	90,075,930	90,075,842	-	0	89	TTCTAGACAGTCCCAGCACCATGAGCCCTGAAAGGTCCCAAGAGGAGAGCCCAGAAGGAGACACAGAGAGAGAACAGAGCGGAAGCCCATG
	Intron 1-2	90,075,841	90,075,475			367	gtgagaagtcggggaggcgaagcca..........tgatggaatctgttacttcctctag
2	ENSE00003486013	90,075,474	90,075,351	0	1	124	GTCAAAGATGCCTTCAAAGACATTTCCATATACTTCACCAAGGAAGAATGGGCAGAAATGGGAGACTGGGAGAAAACTCGCTATAGGAATGTGAAAATGAACTATAATGCACTGATTACTGTAG
	Intron 2-3	90,075,350	90,075,024			327	gtaacaggaagtgctgggcacagac..........agcaaaatttttgcttcttttcag
3	ENSE00003621307	90,075,023	90,074,916	1	1	108	GTCTCAGAGCCACTCGACCAGCTTTCATGTGTCACCGAAGGCAGGCCATCAAACTCCAGGTGGATGACACAGAAGATTCCGATGAAGAATGGACACCTAGGCAGCAAG
	Intron 3-4	90,074,915	90,066,911			8,005	gtaagagggaagggaaggaggattt..........ccactgatttctcatcacctcttag
4	ENSE00003488323	90,066,910	90,066,861	1	0	50	TCAAACCTCCTTGGATGGCCTTCAGAGGAGAACAGAGTAAACACCAGAAG
	Intron 4-5	90,066,860	90,063,769			3,092	gtaagtatctcccaaatcctgttga..........ttaatatgtgattctcacattaaag

ENSEMBL

- Go back and select **Location**.





You can view:

the chromosome

a specific region of interest

a detailed view with certain tracks

Overview

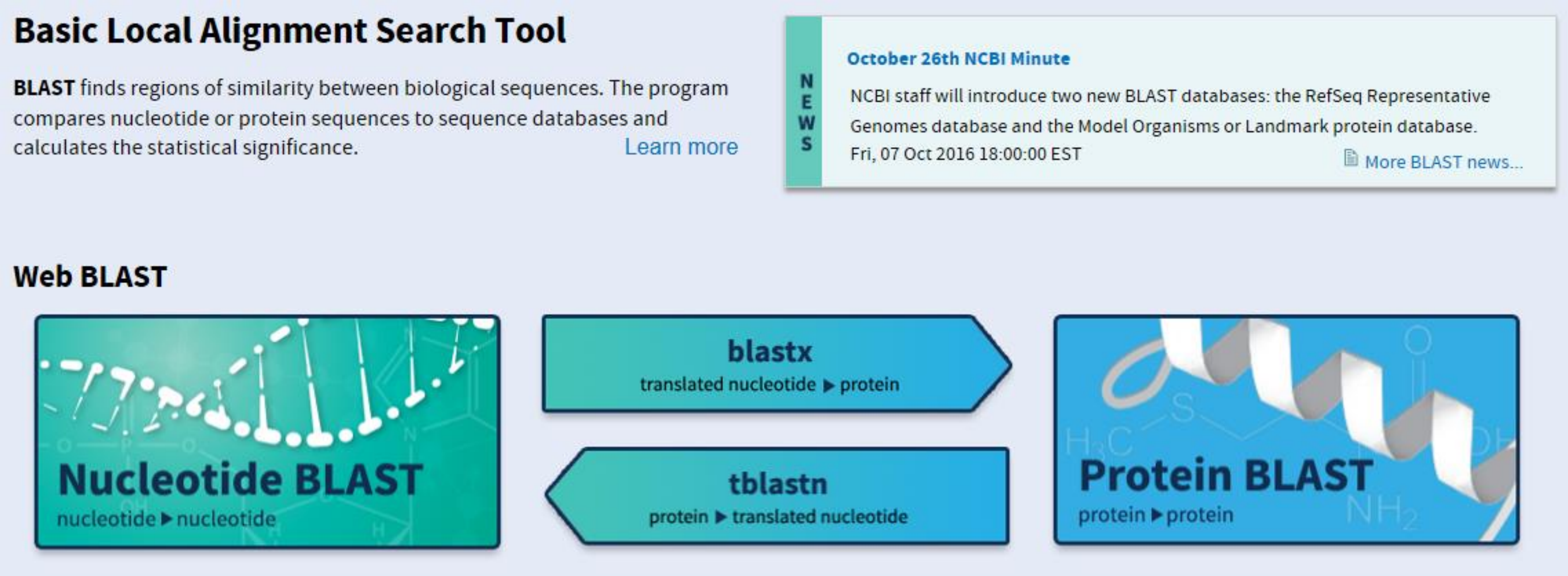
- ❑ **Genome browsers**
  - ➢ **UCSC**
  - ➢ **ENSEMBL**

- ❑ **BLAST**
  - ➢ **Pairwise alignments**
  - ➢ **Database alignments**
  - ➢ **Primer-BLAST**

BLAST

- **BLAST** = Basic Local Alignment Search Tool

- **BLAST** is a nice tool to compare biological sequences and to find regions of identity or similarity.



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

October 26th NCBI Minute

NCBI staff will introduce two new BLAST databases: the RefSeq Representative Genomes database and the Model Organisms or Landmark protein database.

Fri, 07 Oct 2016 18:00:00 EST

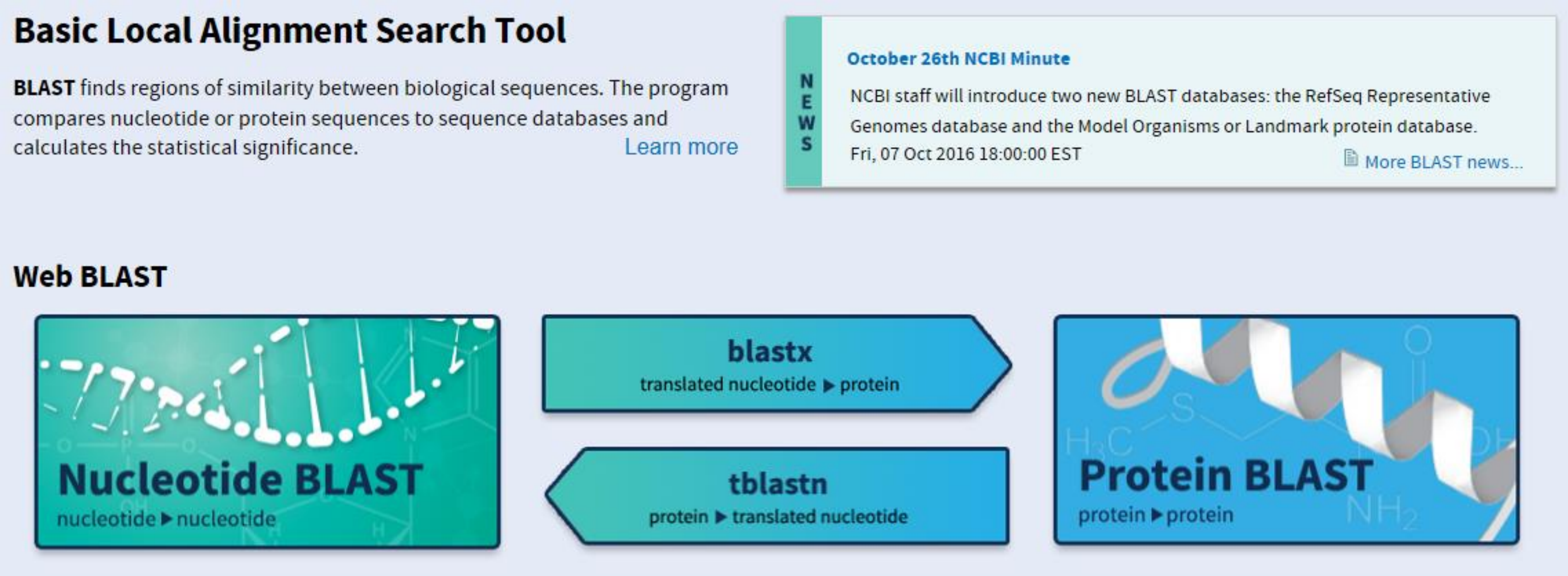
More BLAST news...

Web BLAST

Theresa Schwarz

BLAST

- **Pairwise alignment:** process of lining up two sequences to achieve maximal levels of identity
- **Database alignments:** input sequence is aligned to similar sequences of an entire database



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

October 26th NCBI Minute

NCBI staff will introduce two new BLAST databases: the RefSeq Representative Genomes database and the Model Organisms or Landmark protein database.

Fri, 07 Oct 2016 18:00:00 EST

More BLAST news...

Web BLAST

Overview

- ❑ **Genome browsers**
  - ➢ **UCSC**
  - ➢ **ENSEMBL**

- ❑ **BLAST**
  - ➢ **Pairwise alignments**
  - ➢ **Database alignments**
  - ➢ **Primer-BLAST**

Pairwise alignment

- Where do you get the highest level of identity when comparing two sequences?

```
glu glu ala gly glu asp asp glu       ➢ 1
asp gly ala glu asp glu asn asn

glu glu ala gly glu asp asp glu       ➢ 2
    asp gly ala glu asp glu asn asn

glu glu ala gly glu asp asp glu       ➢ 3
        asp gly ala glu asp glu asn asn
```

Pairwise alignment

- Aims (a few examples):
  - assess the **degree of similarity** of 2 sequences
  - search for **conservation** (e.g. protein domains or sequence motifs)
  - find **functionally or structurally** related proteins
  - assess the **possibility of homology**

Homolog-Paralog-Ortholog

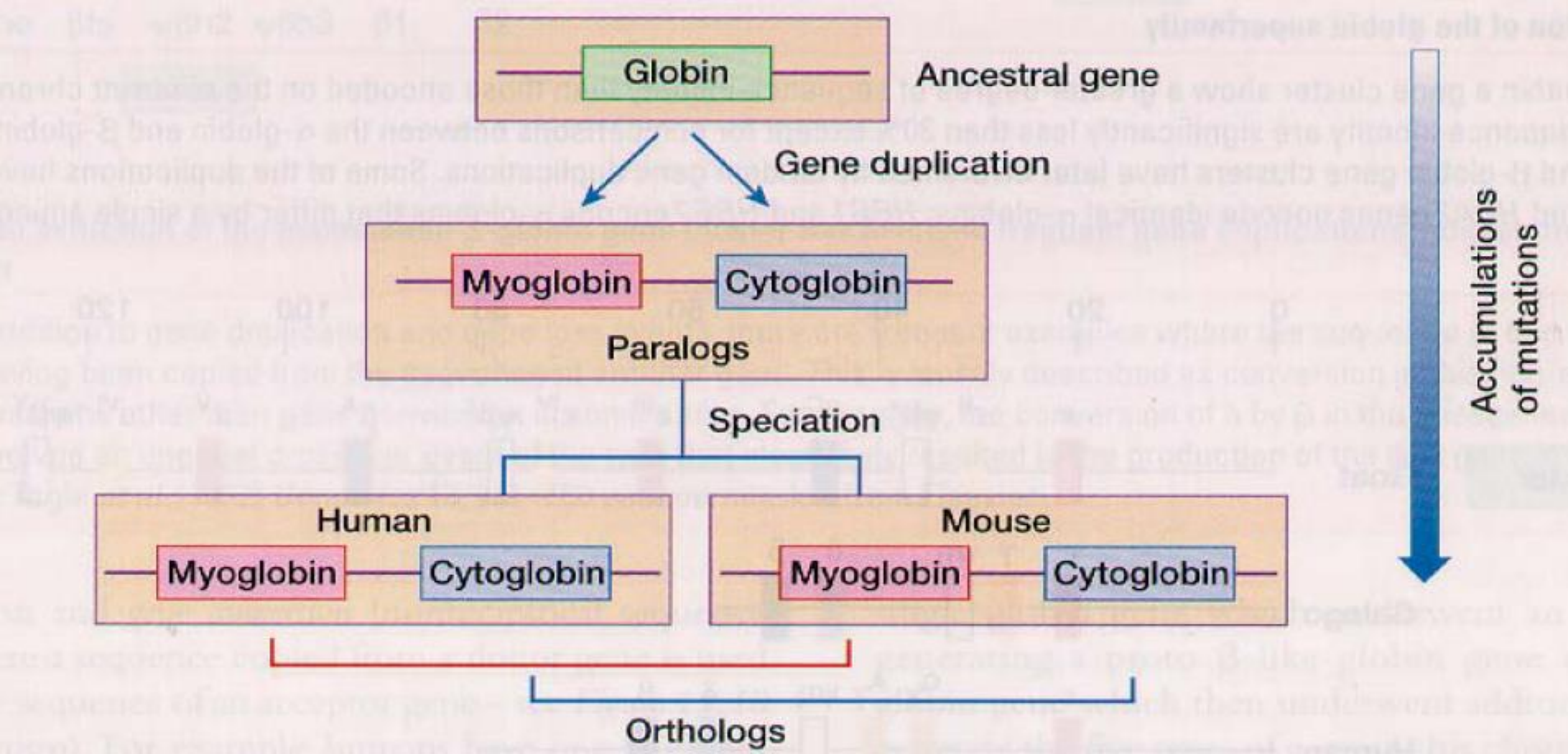
When are two genes/proteins **homologous**, **paralogous** or **orthologous**?

- **Homologs** are related genes that descended from a common ancestral gene. Two genes can be separated by the event of **speciation** (see ortholog) or **gene duplication** (see paralog).

- **Paralogs** are related genes in the same species that have been separated by a duplication event within a genome. Paralogs mostly evolve new functions.

- **Orthologs** are related genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution.

Homolog-Paralog-Ortholog



Example for paralogs

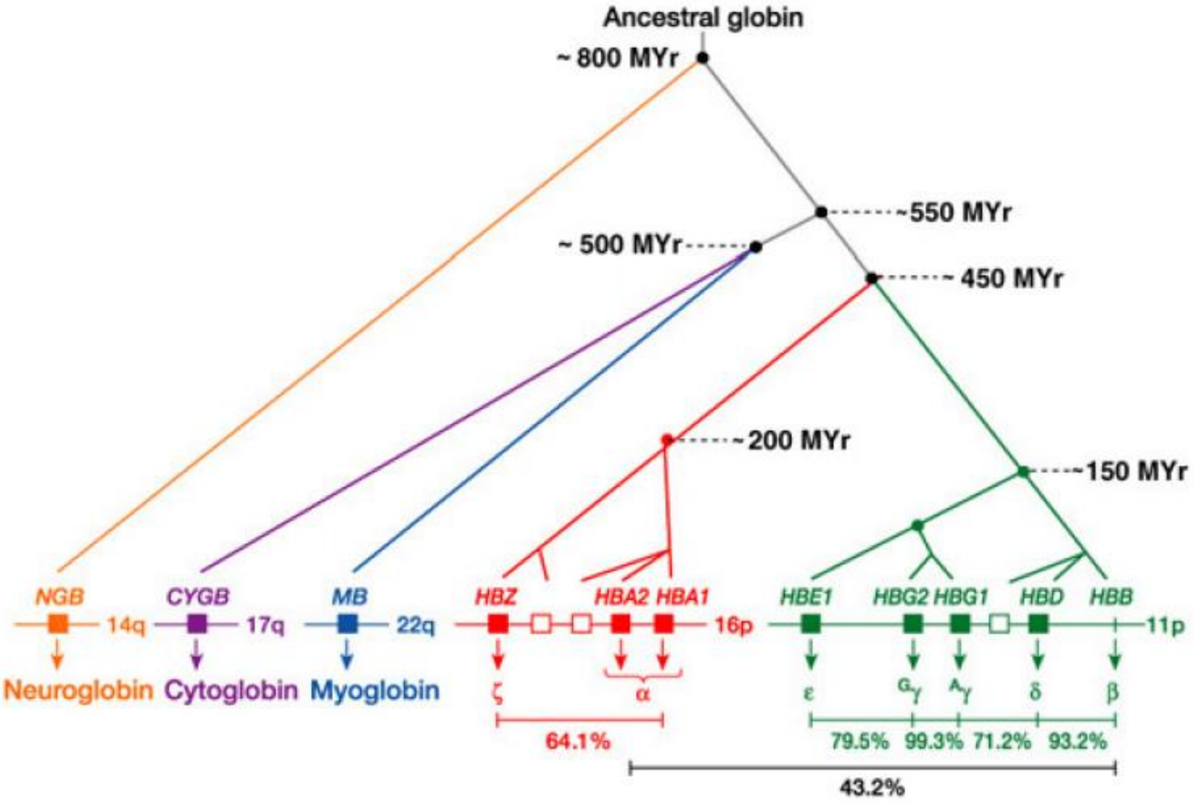


Figure 12-4 Human Molecular Genetics, 3/e. (© Garland Science 2004)

Hemoglobin	tetramer	oxygen transport in blood
Myoglobin	monomer	oxygen transport in muscle
Neuroglobin	monomer	oxygen transport in CNS
Cytoglobin	monomer	oxygen transfer blood-brain

Example for orthologs

- Neuroglobin [Homo sapiens]
- Neuroglobin [Mus musculus]

- Alignment of those two neuroglobins
  - on level of amino acid sequence

    → **92% identities → difference in 13 amino acids**
  - on level of nucleotide sequence

    → **79% identities → difference in 193 bases**

    (divided by 3 = ~64 aa)

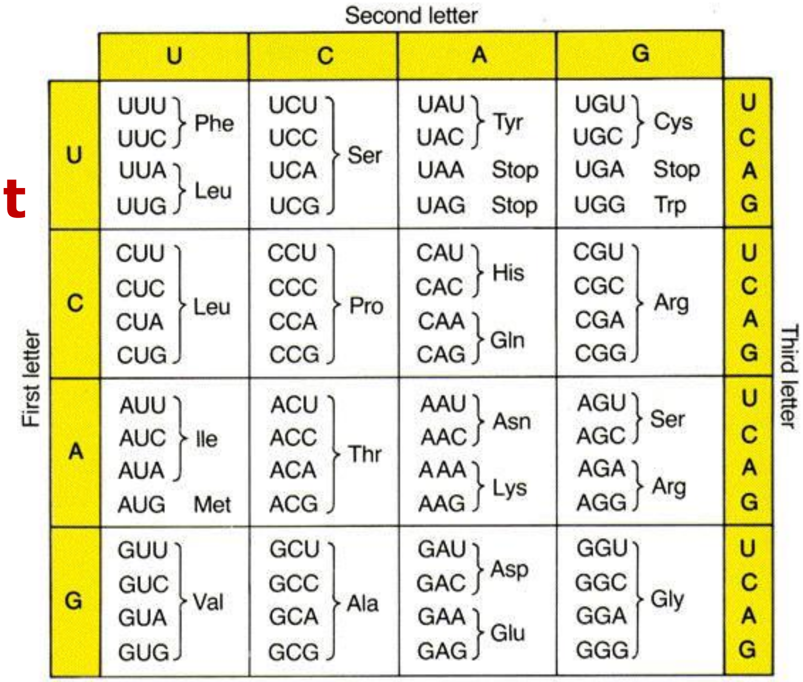
Pairwise alignment – finding homologs

- When you want to find homologs or conserved domains the **amino acid sequence is much more informative** than the nucleotide sequence !

- Because: the genetic code is **redundant**

  (codons are degenerate: changes in the 3rd position often do not change the aa)

  3rd position = "wobble base"



	Second letter				
First letter	U	C	A	G	Third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC	UCC	UAC	UGC	C
	UUA Leu	UCA	UAA Stop	UGA Stop	A
	UUG	UCG	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA Gln	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA	AAA Lys	AGA Arg	A
	AUG Met	ACG	AAG	AGG	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC	GCC	GAC	GGC	C
	GUA	GCA	GAA Glu	GGA	A
	GUG	GCG	GAG	GGG	G

- Because: more characters in proteins: 20 amino acids vs. 4 bases

Pairwise alignment – finding homologs

- **Protein alignments** are used
  - to find common ancestors million or billion of years ago (Amino acid sequences offer a longer **“look-back” time**)
  - DNA sequences can be translated into protein and then used in pairwise alignments to find homologs

Pairwise alignment – nucleotide sequences

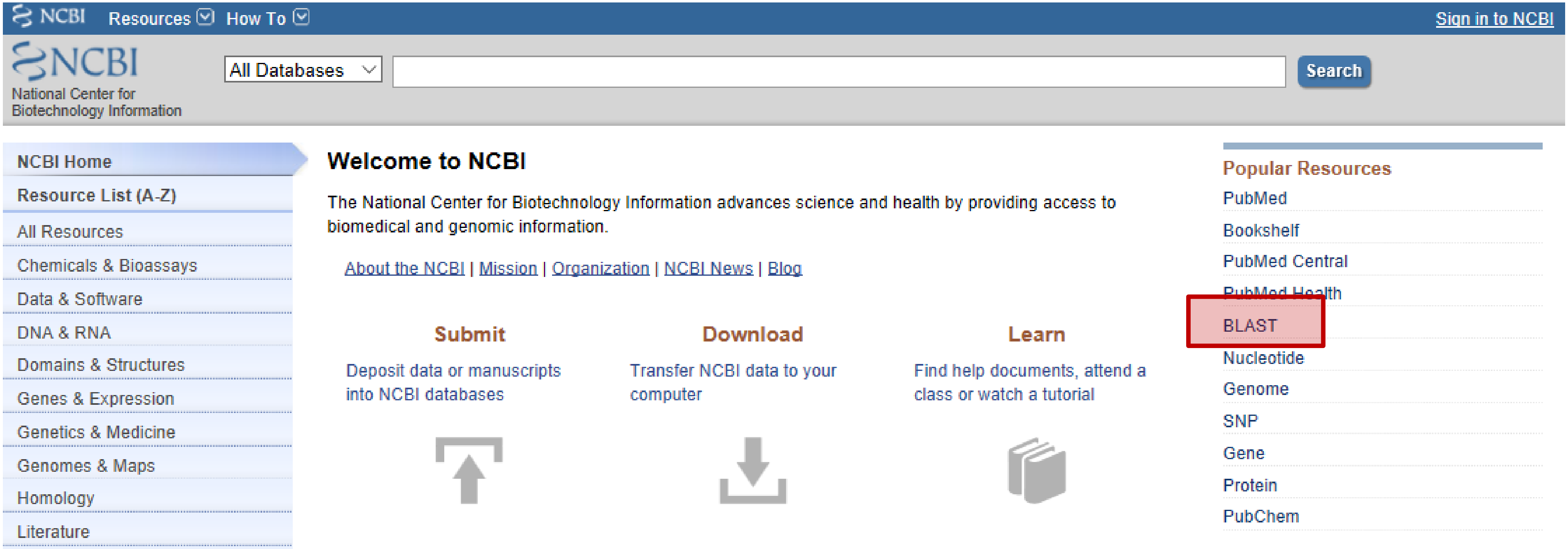
- **DNA alignments** are used
  - to study DNA **polymorphisms** (SNPs, insertions, deletions, microsatellites,…)
  - to study **non-coding** regions of DNA
  - to confirm the **identity of a cDNA**

```
Query: 181 catcaactacaactccaaagacacccttacacccactaggatatcaacaaacctacccac 240
           |||||||| |||| |||||| ||||| | |||||||||||||||||||||||||||||||
Sbjct: 189 catcaactgcaaccccaaagccacccct-cacccactaggatatcaacaaacctacccac 247
```

Pairwise alignment - BLAST

- You can do a pairwise alignment with nucleotide and protein sequences using BLAST – a tool of NCBI.

- *http://www.ncbi.nlm.nih.gov/*



Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News | Blog

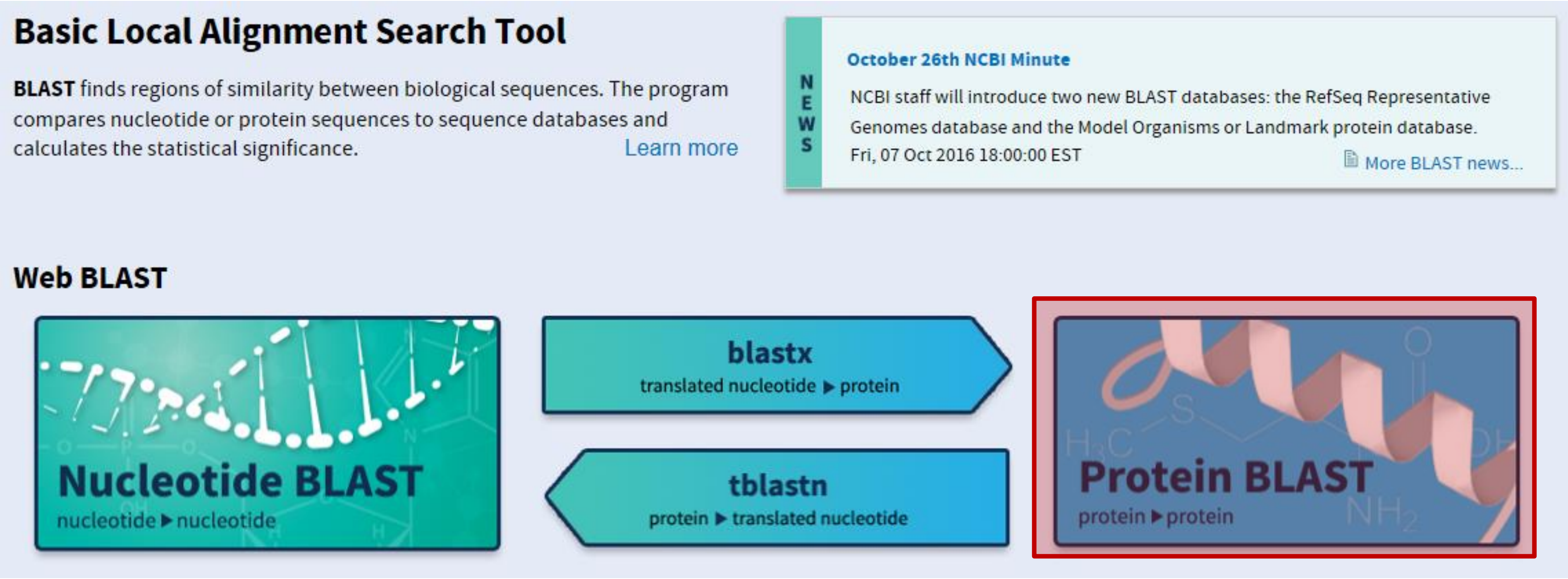
Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Pairwise alignment - BLAST

→ Let‘s compare **Hemoglobin** and **Myoglobin** by using **Protein BLAST**.



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

October 26th NCBI Minute

NCBI staff will introduce two new BLAST databases: the RefSeq Representative Genomes database and the Model Organisms or Landmark protein database.

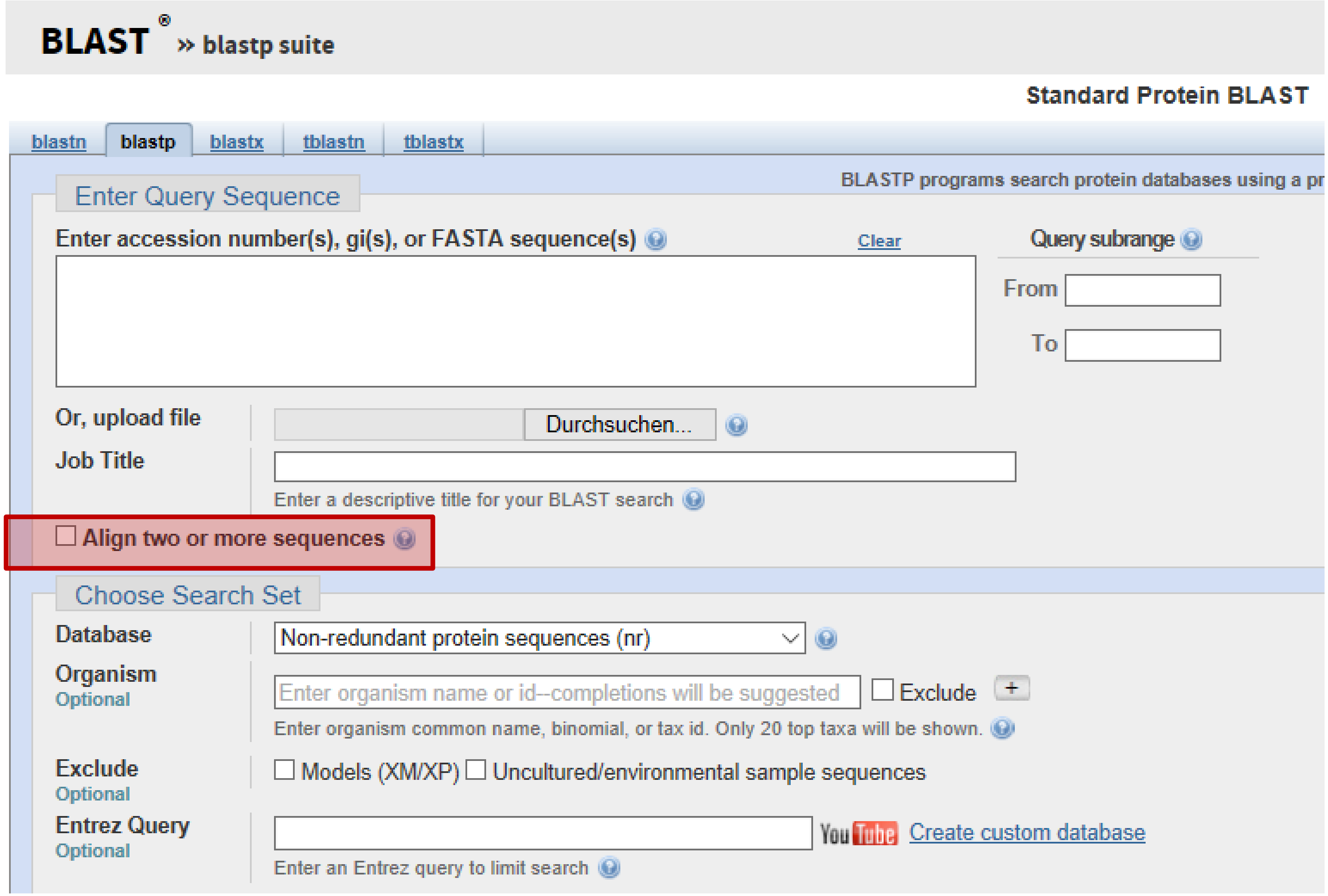
Fri, 07 Oct 2016 18:00:00 EST

More BLAST news...

Web BLAST

Pairwise alignment - BLAST

- Select **Align two or more sequences**



BLAST® » blastp suite

Standard Protein BLAST

BLASTP programs search protein databases using a pr

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Query subrange

From

To

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism
Optional

Enter organism name or id--completions will be suggested □ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude
Optional

□ Models (XM/XP) □ Uncultured/environmental sample sequences

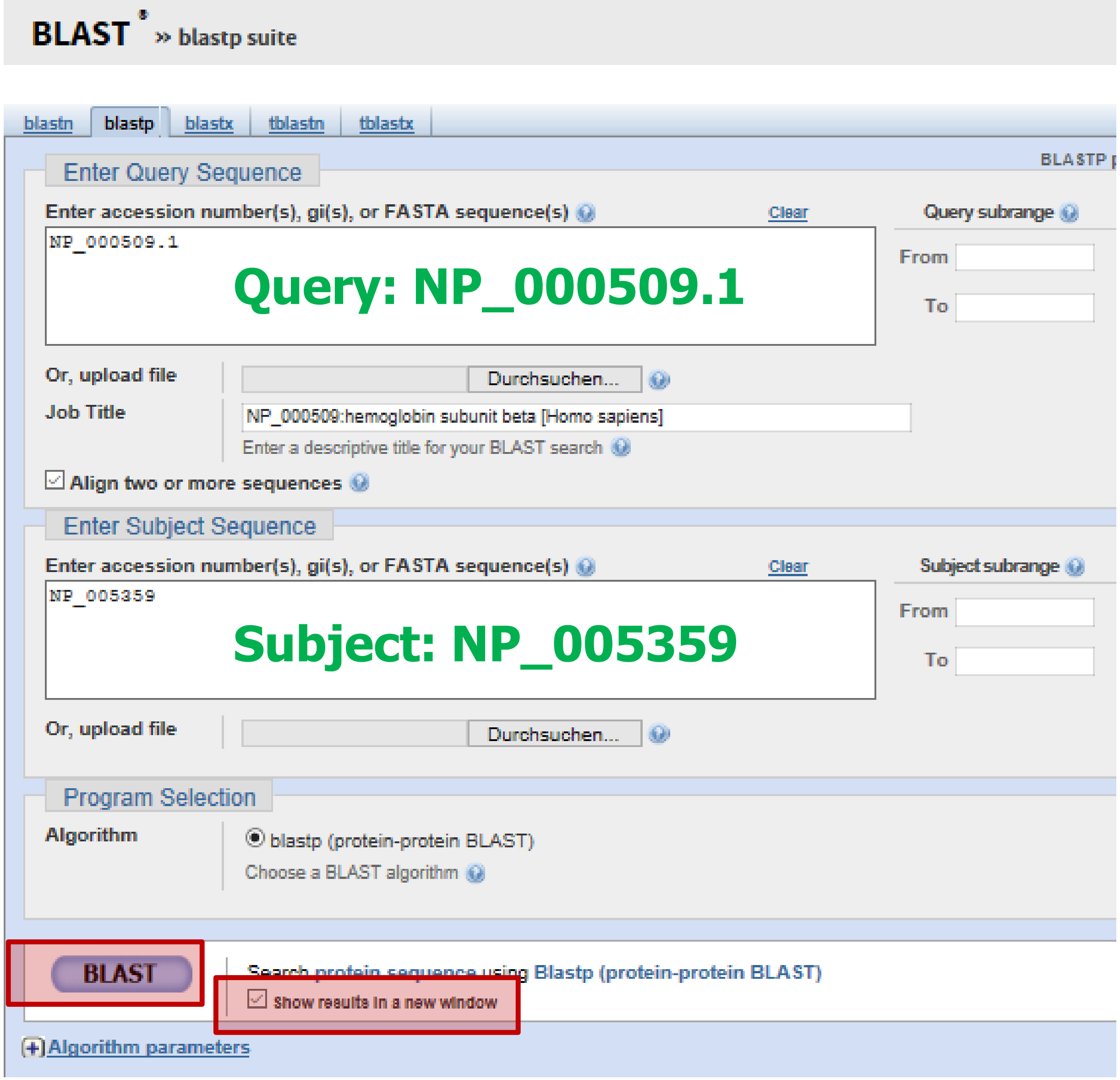
Entrez Query
Optional

Create custom database

Enter an Entrez query to limit search

Theresa Schwarz

Pairwise alignment - BLAST



BLAST® » blastp suite

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s): NP_000509.1

Query: NP_000509.1

Query subrange: From ___ To ___

Or, upload file

Job Title: NP_000509:hemoglobin subunit beta [Homo sapiens]

Enter a descriptive title for your BLAST search

☑ Align two or more sequences

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s): NP_005359

Subject: NP_005359

Subject subrange: From ___ To ___

Or, upload file

Program Selection

Algorithm: ● blastp (protein-protein BLAST)

Choose a BLAST algorithm

Search protein sequence using Blastp (protein-protein BLAST)

☑ Show results in a new window

Enter **Accession number** or sequence in **FASTA format** of the two proteins you want to compare:

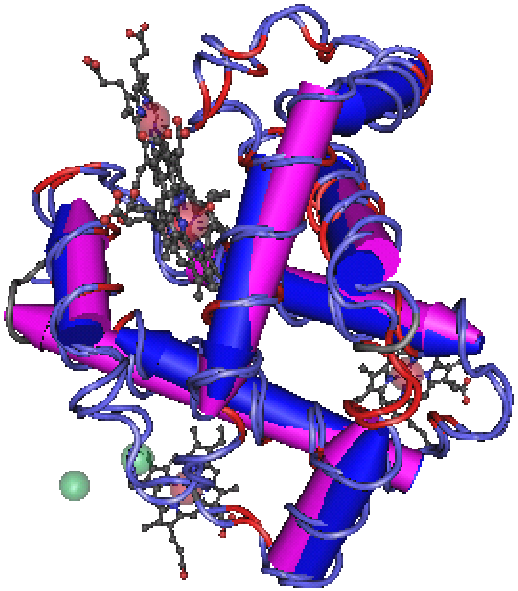
Hemoglobin: **NP_000509.1**

Myoglobin: **NP_005359**

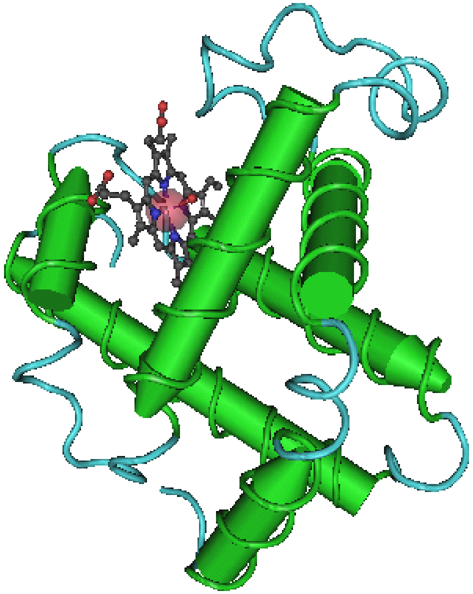
Pairwise alignment - BLAST

Do they look similar ?

Human
Hemoglobin subunit beta
(NP_000509.1)



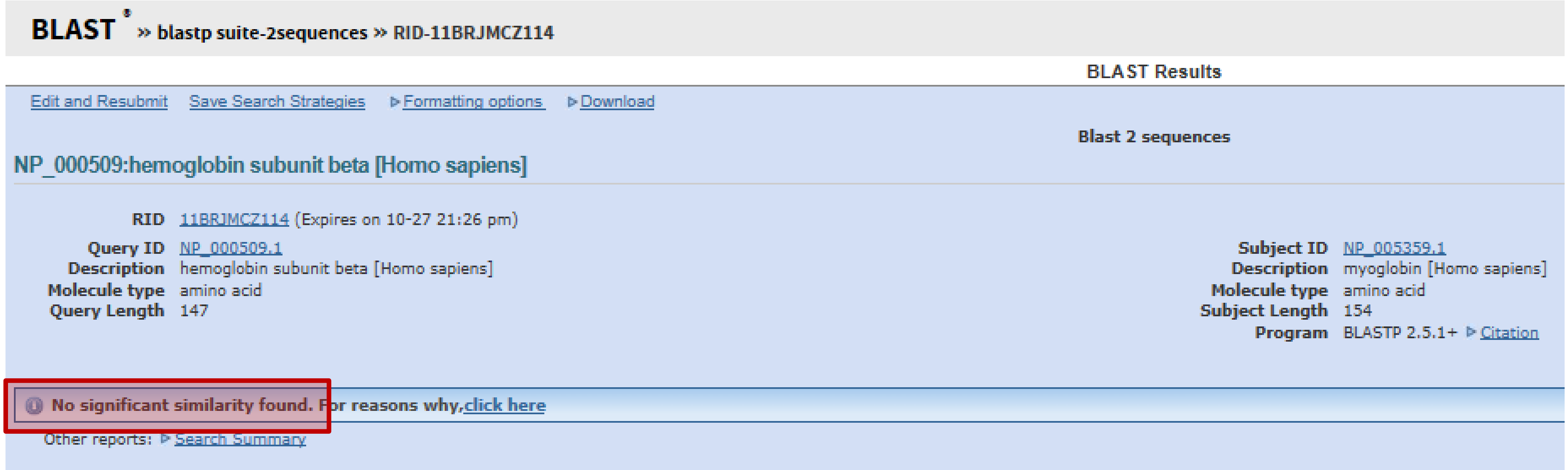
Human
Myoglobin
(NP_005359)



Theresa Schwarz

Pairwise alignment - BLAST

- **No significant similarity found!**



BLAST » blastp suite-2sequences » RID-11BRJMCZ114

BLAST Results

Edit and Resubmit   Save Search Strategies   ▷ Formatting options   ▷ Download

Blast 2 sequences

NP_000509:hemoglobin subunit beta [Homo sapiens]

RID 11BRJMCZ114 (Expires on 10-27 21:26 pm)

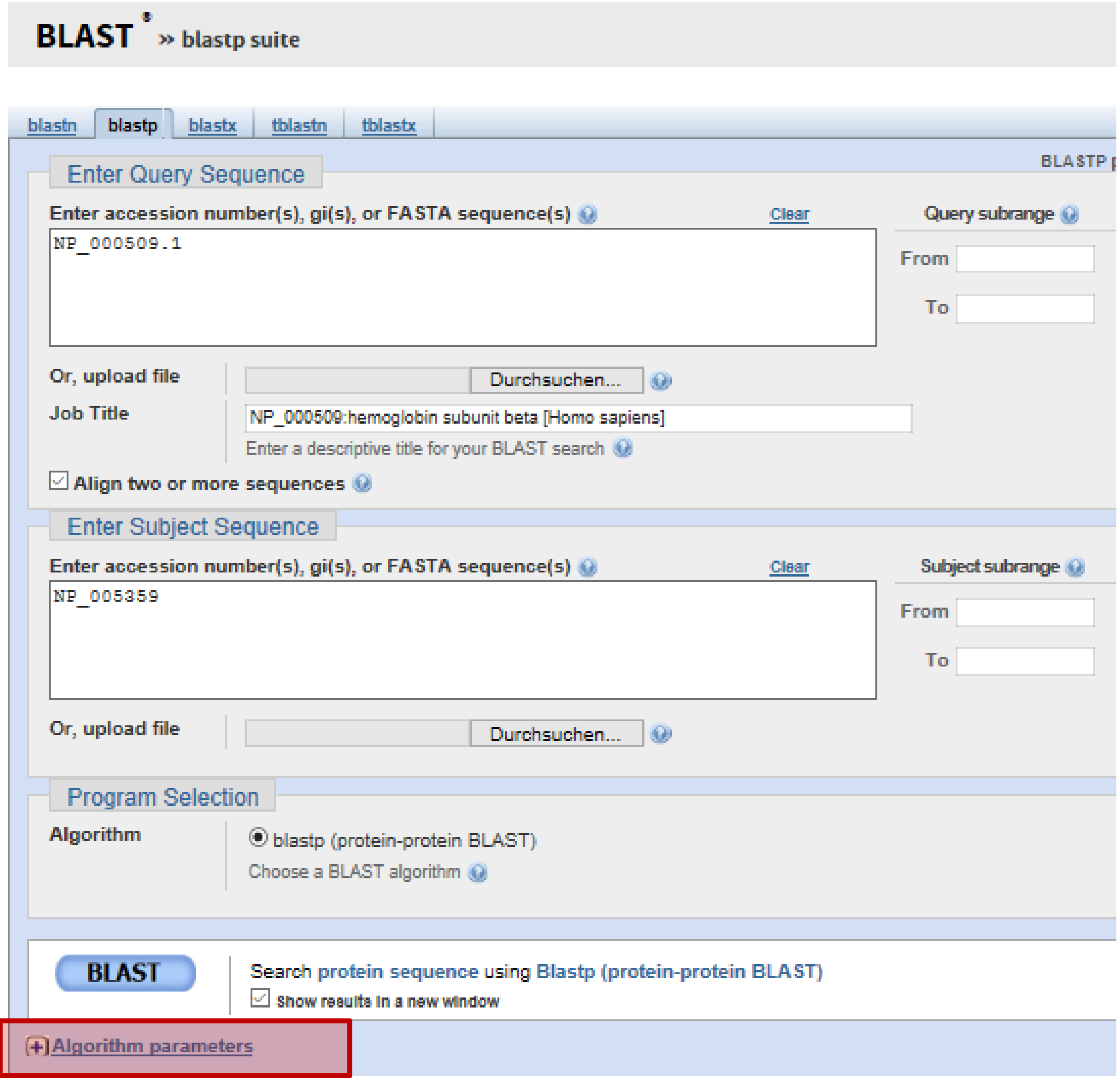
Query ID	NP_000509.1	**Subject ID**	NP_005359.1
Description	hemoglobin subunit beta [Homo sapiens]	**Description**	myoglobin [Homo sapiens]
Molecule type	amino acid	**Molecule type**	amino acid
Query Length	147	**Subject Length**	154
		Program	BLASTP 2.5.1+ ▷ Citation

No significant similarity found. For reasons why, click here

Other reports: ▷ Search Summary

- **We can go back and change some parameters**

Pairwise alignment - BLAST



BLAST » blastp suite

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
NP_000509.1
```

Query subrange
From
To

Or, upload file Durchsuchen...

Job Title: NP_000509:hemoglobin subunit beta [Homo sapiens]

Enter a descriptive title for your BLAST search

☑ Align two or more sequences

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
NP_005359
```

Subject subrange
From
To

Or, upload file Durchsuchen...

Program Selection

Algorithm: ● blastp (protein-protein BLAST)

Choose a BLAST algorithm

BLAST — Search protein sequence using Blastp (protein-protein BLAST)

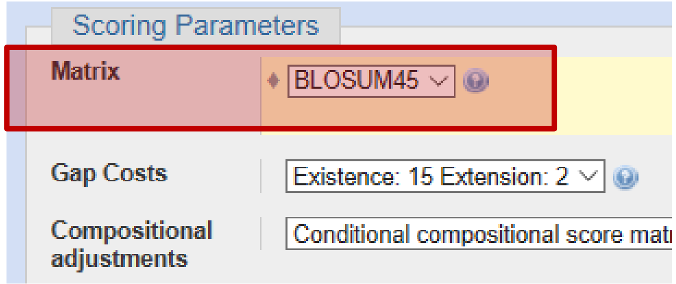
☑ Show results in a new window

[+] Algorithm parameters

- Select **Algorithm parameters** (and scroll down)

- Change the Matrix (Scoring Parameters) from BLOSUM62 to **BLOSUM45**

Repeat the BLAST search!



Scoring Parameters

Matrix: BLOSUM45

Gap Costs: Existence: 15 Extension: 2

Compositional adjustments: Conditional compositional score mat

Scoring Matrices

There are two kinds of sequence alignments using different matrices:

GLOBAL alignment algorithm

- Needleman and Wunsch (1970)

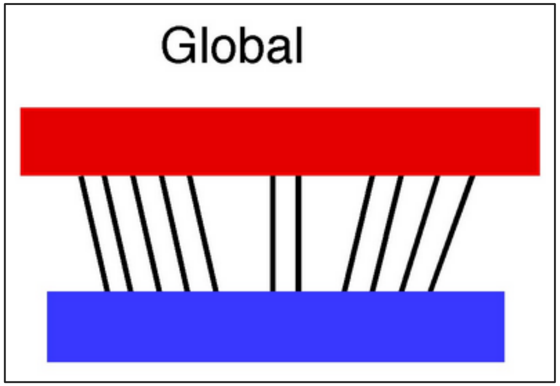
LOCAL alignment algorithm

- Smith and Waterman (1981)

Scoring Matrices

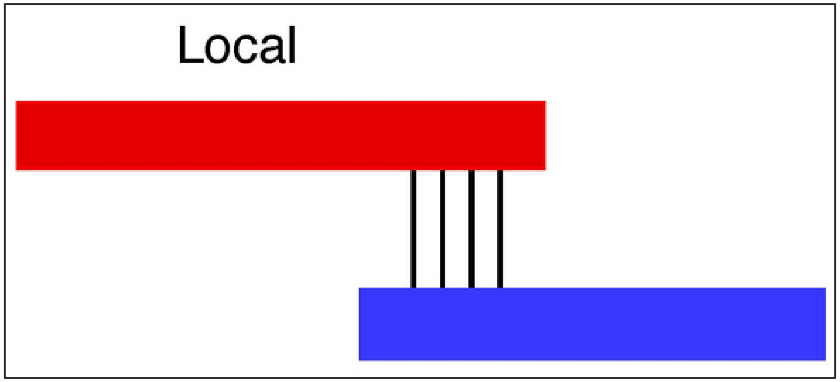
GLOBAL alignment extends from one end of each sequence to the other.

→ **PAM** matrices



LOCAL alignment finds optimally matching regions within two sequences (“subsequences”).

→ **BLOSUM** matrices



BLOSUM Matrices

- BLOSUM matrices are based on **LOCAL** alignments.
- BLOSUM stands for **blocks substitution matrix**.
- BLOSUM**80** is a matrix to compare sequences with similarities of **>80%**.
- BLOSUM**62** is a matrix to compare sequences with similarities of **>62%**.
- BLOSUM**45** is a matrix to compare sequences with similarities of **>45%**.

The higher – the better !

PAM Matrices

- PAM matrices are based on **GLOBAL** alignments.

- PAM stands for **point accepted mutation**.

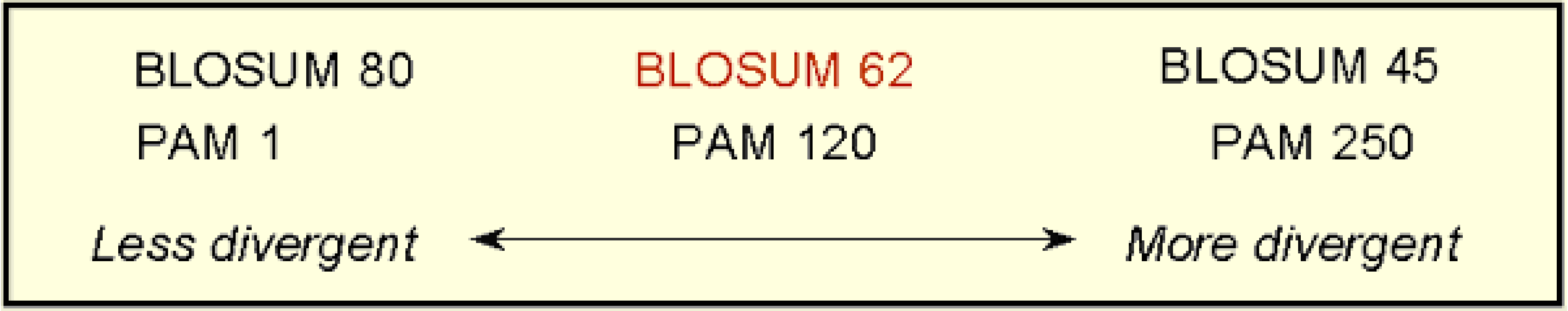
- PAM matrices are used to assess the **relatedness** of two proteins

- PAM matrices work like this:
  How many differences are allowed per 100 amino acids?

- PAM1	**1 difference per 100 amino acids**
- PAM10.7	**10 differences per 100 amino acids**
- PAM80	**50 differences per 100 amino acids**
- PAM250	**80 differences per 100 amino acids**

The lower – the better !

SUMMARY - Matrices



BLOSUM 80 — BLOSUM 62 — BLOSUM 45

PAM 1 — PAM 120 — PAM 250

Less divergent ← → *More divergent*

Closely related

Distantly related

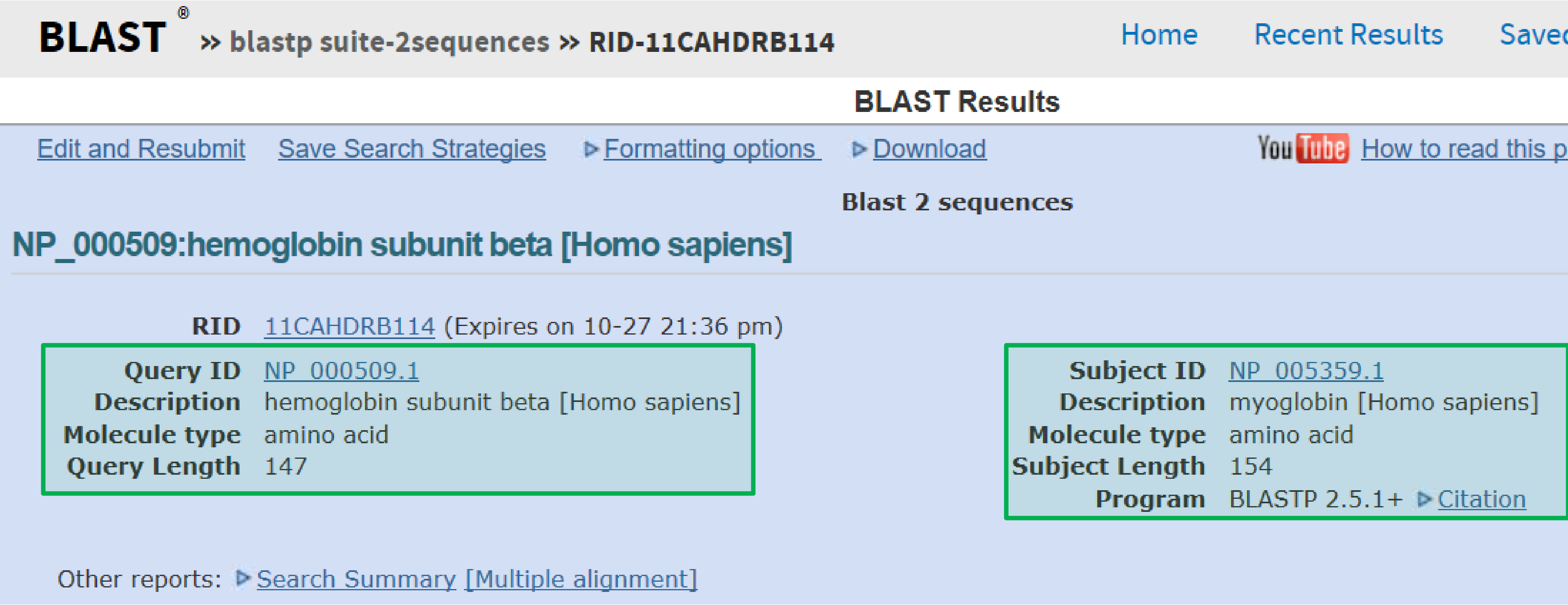
rat vs. mouse globin

rat vs. bacterial globin

By choosing the matrix you can chose **which part** of the sequence should be used (global, local) and **how stringent** the alignment should be done.

Pairwise alignment - BLAST

- When using a different **Scoring Matrix** an alignment is possible



BLAST® » blastp suite-2sequences » RID-11CAHDRB114

BLAST Results

Blast 2 sequences

NP_000509:hemoglobin subunit beta [Homo sapiens]

RID 11CAHDRB114 (Expires on 10-27 21:36 pm)

Query ID	NP_000509.1
Description	hemoglobin subunit beta [Homo sapiens]
Molecule type	amino acid
Query Length	147

Subject ID	NP_005359.1
Description	myoglobin [Homo sapiens]
Molecule type	amino acid
Subject Length	154
Program	BLASTP 2.5.1+ Citation

Other reports: Search Summary [Multiple alignment]

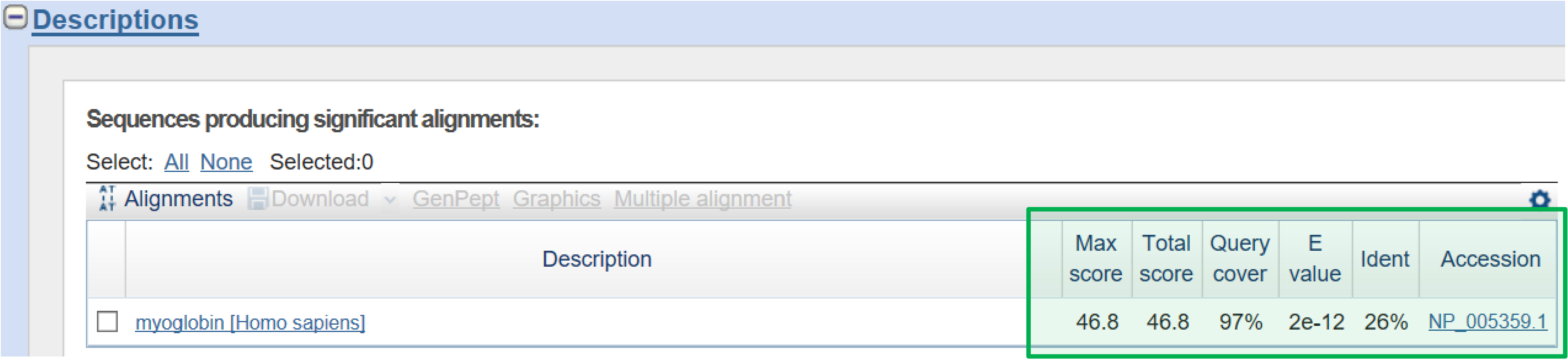
- Scroll down to get more information about your results

Pairwise alignment - BLAST

How does your BLAST result look like?

General information:

- Score
- Query coverage
- Expected (E) value
- Ident
- Accession



Descriptions

Sequences producing significant alignments:

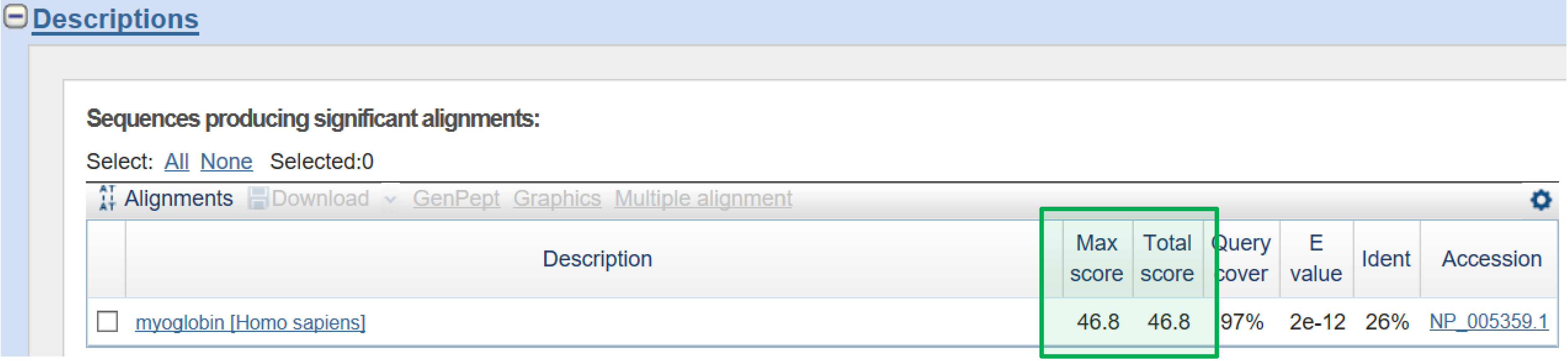
Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	myoglobin [Homo sapiens]	46.8	46.8	97%	2e-12	26%	NP_005359.1

Pairwise alignment - BLAST

How does your BLAST result look like?

- **Score**:
  - a measure for the **quality** of the alignment
  - it is calculated by the scoring matrix and reflects the **degree of similarity**
  - Max score: Score of single best aligned sequence
  - Total score: Sum of scores of all aligned sequences
  - The higher the better!



Descriptions

Sequences producing significant alignments:

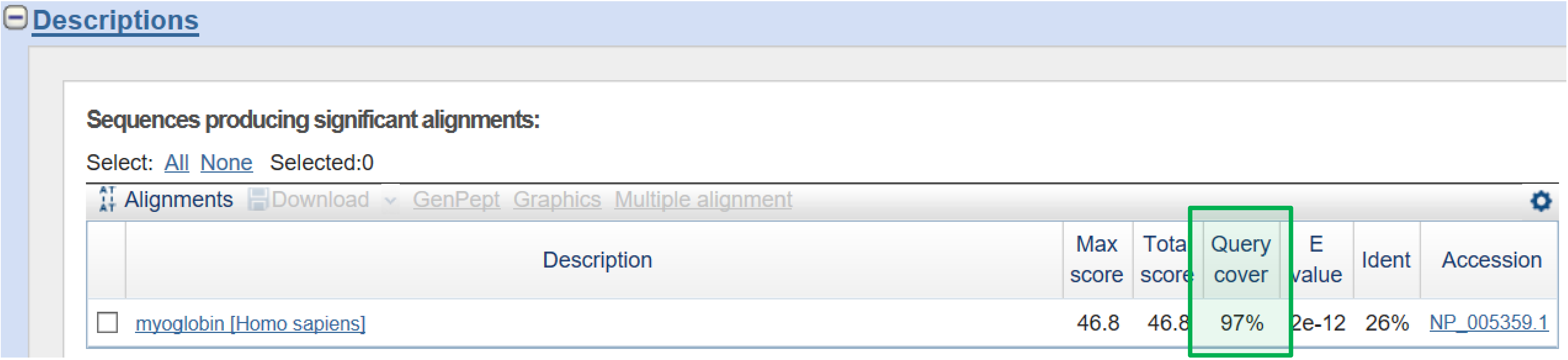
Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	myoglobin [Homo sapiens]	46.8	46.8	97%	2e-12	26%	NP_005359.1

Pairwise alignment - BLAST

How does your BLAST result look like?

- **Query coverage**:
  - ➢ Information on **how much of a sequence** is used for the alignment
  - ➢ Always check the query coverage to see whether the alignment is meaningful
  - ➢ The higher the better!



Descriptions

Sequences producing significant alignments:

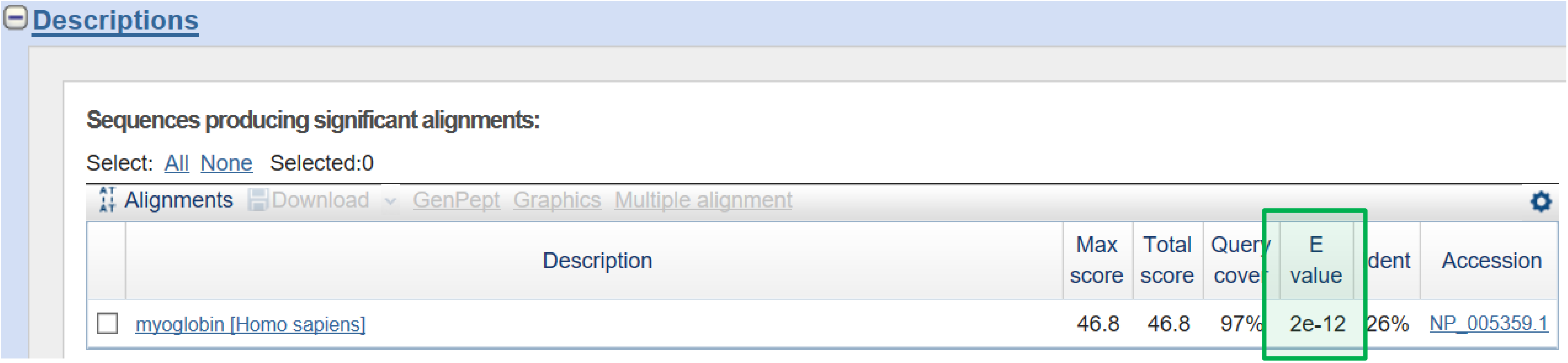
Select: All None Selected:0

	Description	Max score	Tota score	Query cover	E value	Ident	Accession
□	myoglobin [Homo sapiens]	46.8	46.8	97%	2e-12	26%	NP_005359.1

Pairwise alignment - BLAST

How does your BLAST result look like?

- **Expected (E) value**:
  - Represents the **significance** of a result
  - Probability of a random alignment
  - The lower the E-value the more significant
  - The lower the better!



Descriptions

Sequences producing significant alignments:

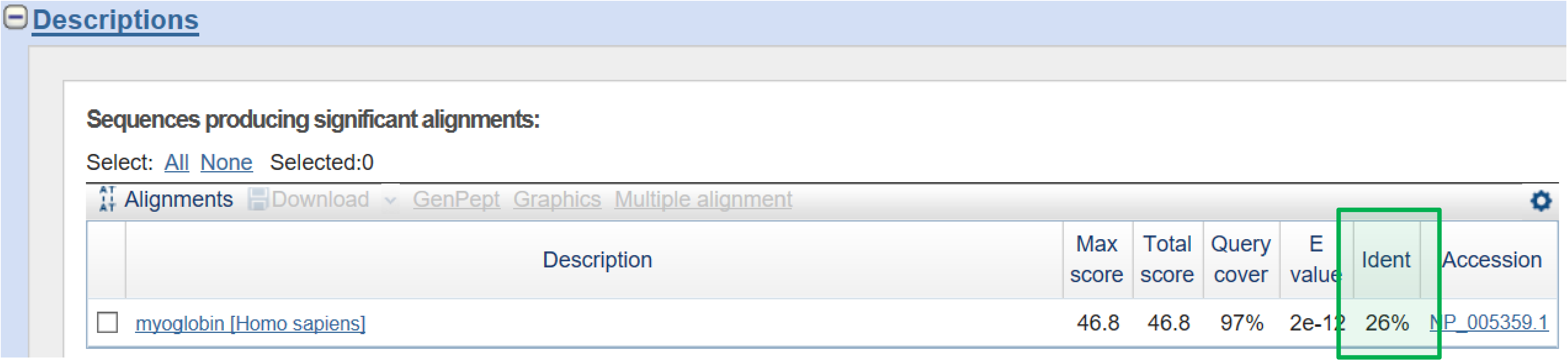
Select: All None Selected:0

	Description	Max score	Total score	Query cove	E value	dent	Accession
□	myoglobin [Homo sapiens]	46.8	46.8	97%	2e-12	26%	NP_005359.1

Pairwise alignment - BLAST

How does your BLAST result look like?

- **Ident**:
  - ➢ Shows how many amino acids of the two sequences match perfectly



Descriptions

Sequences producing significant alignments:

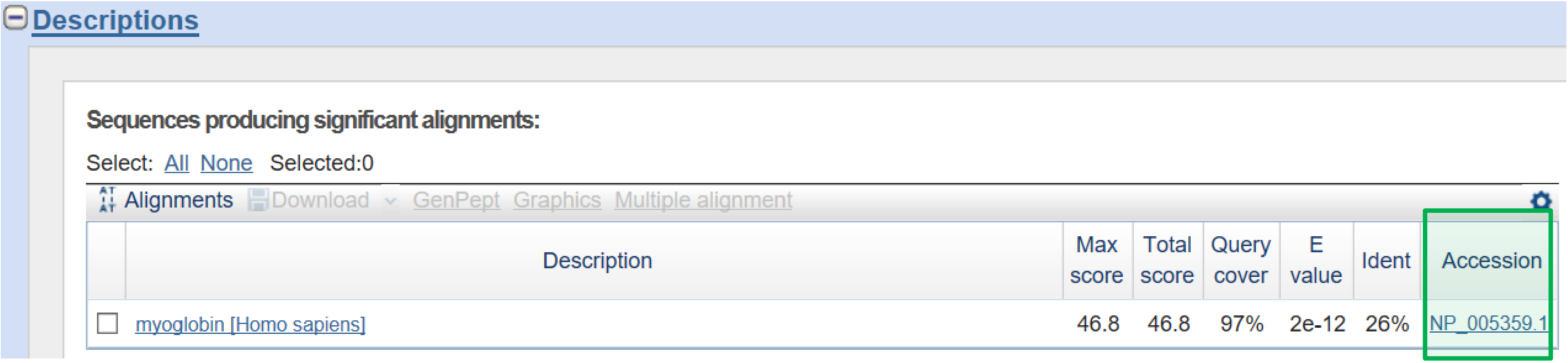
Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	myoglobin [Homo sapiens]	46.8	46.8	97%	2e-12	26%	NP_005359.1

Pairwise alignment - BLAST

How does your BLAST result look like?

- **Accession**:
  - ➢ Protein accession number is directly linked to myoglobin entry in protein database



Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

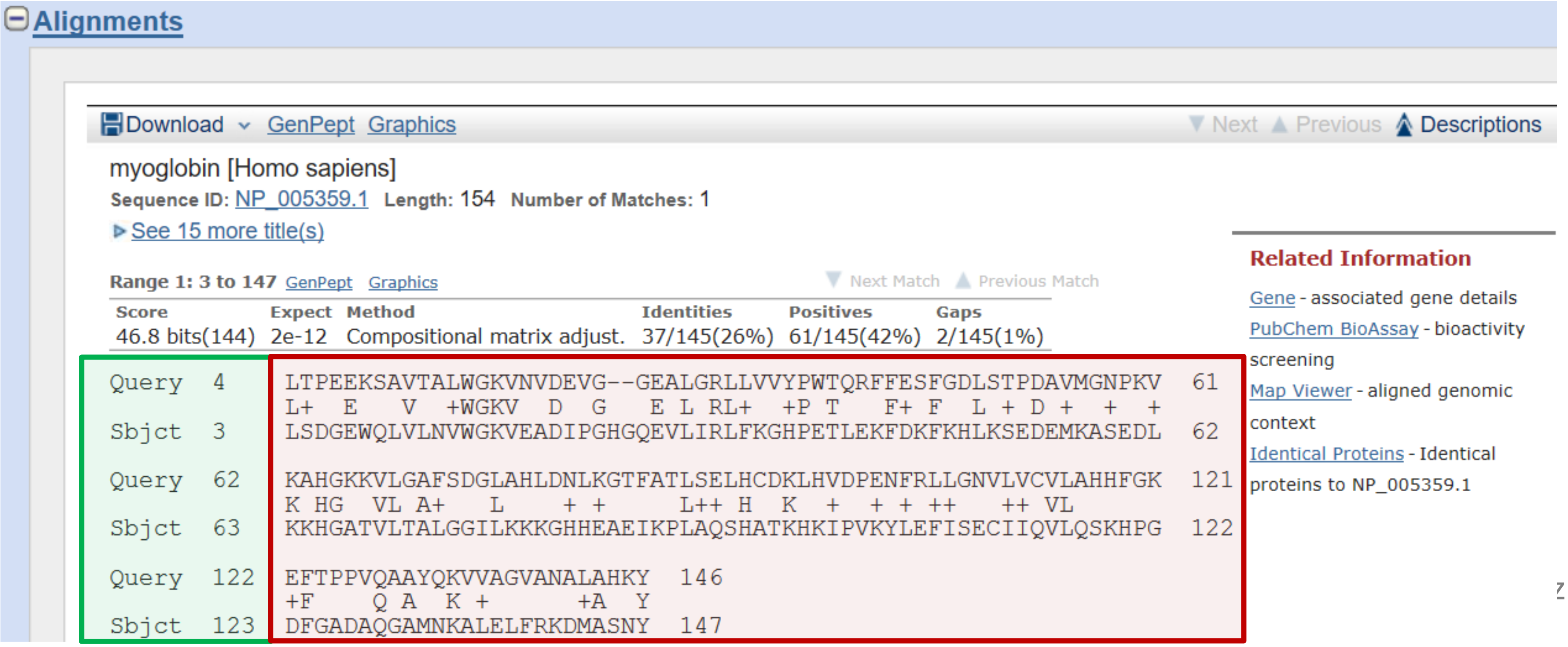
	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	myoglobin [Homo sapiens]	46.8	46.8	97%	2e-12	26%	NP_005359.1

Pairwise alignment - BLAST

Scroll down to **Alignments** for **details**.

- **Query** = **Hemoglobin**
- **Sbjct** = **Myoglobin**

- Sequence of alignment **hemoglobin vs. myglobin**



myoglobin [Homo sapiens]

Sequence ID: NP_005359.1 **Length:** 154 **Number of Matches:** 1

▶ See 15 more title(s)

Range 1: 3 to 147 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
46.8 bits(144)	2e-12	Compositional matrix adjust.	37/145(26%)	61/145(42%)	2/145(1%)

```
Query  4    LTPEEKSAVTALWGKVNVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
            L+  E   V  +WGKV  D  G   E L RL+  +P T   F+ F  L + D +  +  +
Sbjct  3    LSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDL  62

Query  62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
            K HG  VL A+   L    +  +     L++ H  K  +  + + ++   ++ VL
Sbjct  63   KKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPG  122

Query  122  EFTPPVQAAYQKVVAGVANALAHKY  146
            +F    Q A  K +      +A  Y
Sbjct  123  DFGADAQGAMNKALELFRKDMASNY  147
```

Related Information

Gene - associated gene details

PubChem BioAssay - bioactivity screening

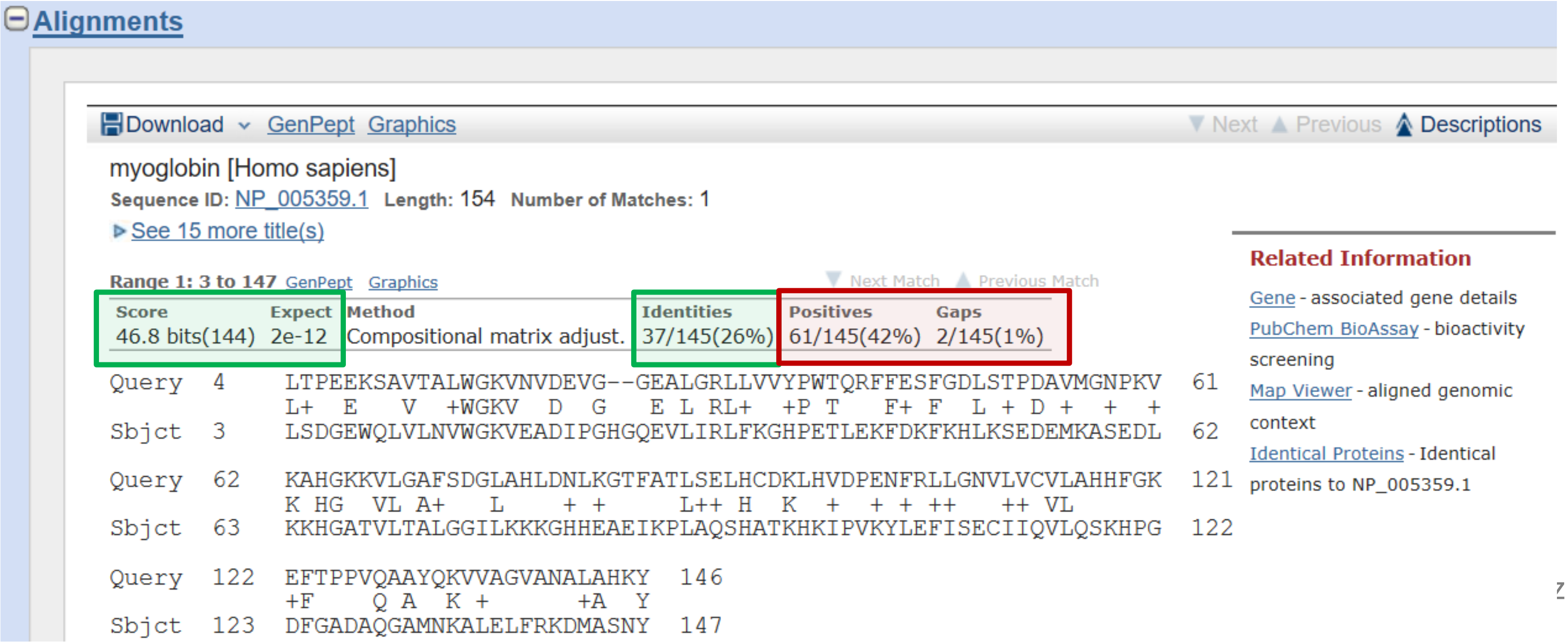
Map Viewer - aligned genomic context

Identical Proteins - Identical proteins to NP_005359.1

Pairwise alignment - BLAST

- You can see again the '**Score**', '**E-value**' and '**Identities**'
- Total number of aligned amino acids: 145

- Positives
- Gaps



Alignments

Download GenPept Graphics ▼ Next ▲ Previous Descriptions

myoglobin [Homo sapiens]

Sequence ID: NP_005359.1 **Length:** 154 **Number of Matches:** 1

▶ See 15 more title(s)

Range 1: 3 to 147 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
46.8 bits(144)	2e-12	Compositional matrix adjust.	37/145(26%)	61/145(42%)	2/145(1%)

```
Query  4    LTPEEKSAVTALWGKVNVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
            L+  E   V  +WGKV  D  G   E L RL+  +P T   F+ F  L + D +  +  +
Sbjct  3    LSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDL  62

Query  62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
            K HG  VL A+   L    + +     L++ H  K  +  + + ++   ++ VL
Sbjct  63   KKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPG  122

Query  122  EFTPPVQAAYQKVVAGVANALAHKY  146
            +F    Q A  K +      +A  Y
Sbjct  123  DFGADAQGAMNKALELFRKDMASNY  147
```

Related Information

Gene - associated gene details

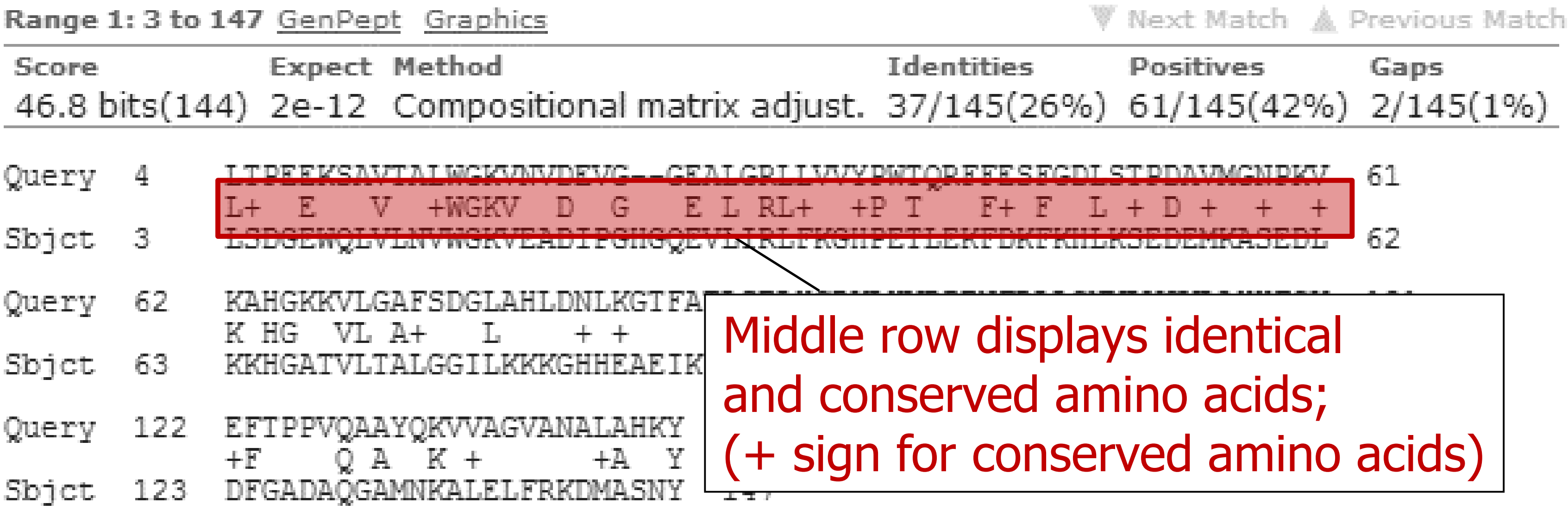
PubChem BioAssay - bioactivity screening

Map Viewer - aligned genomic context

Identical Proteins - Identical proteins to NP_005359.1

Pairwise alignment - BLAST

What is the difference between Identities & Positives?



Range 1: 3 to 147 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
46.8 bits(144)	2e-12	Compositional matrix adjust.	37/145(26%)	61/145(42%)	2/145(1%)

```
Query  4    LTPEEKSAVTALWGKVNVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
            L+  E   V  +WGKV  D  G   E L RL+  +P T   F+ F  L + D +   +  +
Sbjct  3    LSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDL  62

Query  62   KAHGKKVLGAFSDGLAHLDNLKGTFA
            K HG  VL A+    L     + +
Sbjct  63   KKHGATVLTALGGILKKKGHHEAEIK

Query  122  EFTPPVQAAYQKVVAGVANALAHKY
             +F     Q A  K +        +A  Y
Sbjct  123  DFGADAQGAMNKALELFRKDMASNY
```

Middle row displays identical and conserved amino acids;
(+ sign for conserved amino acids)

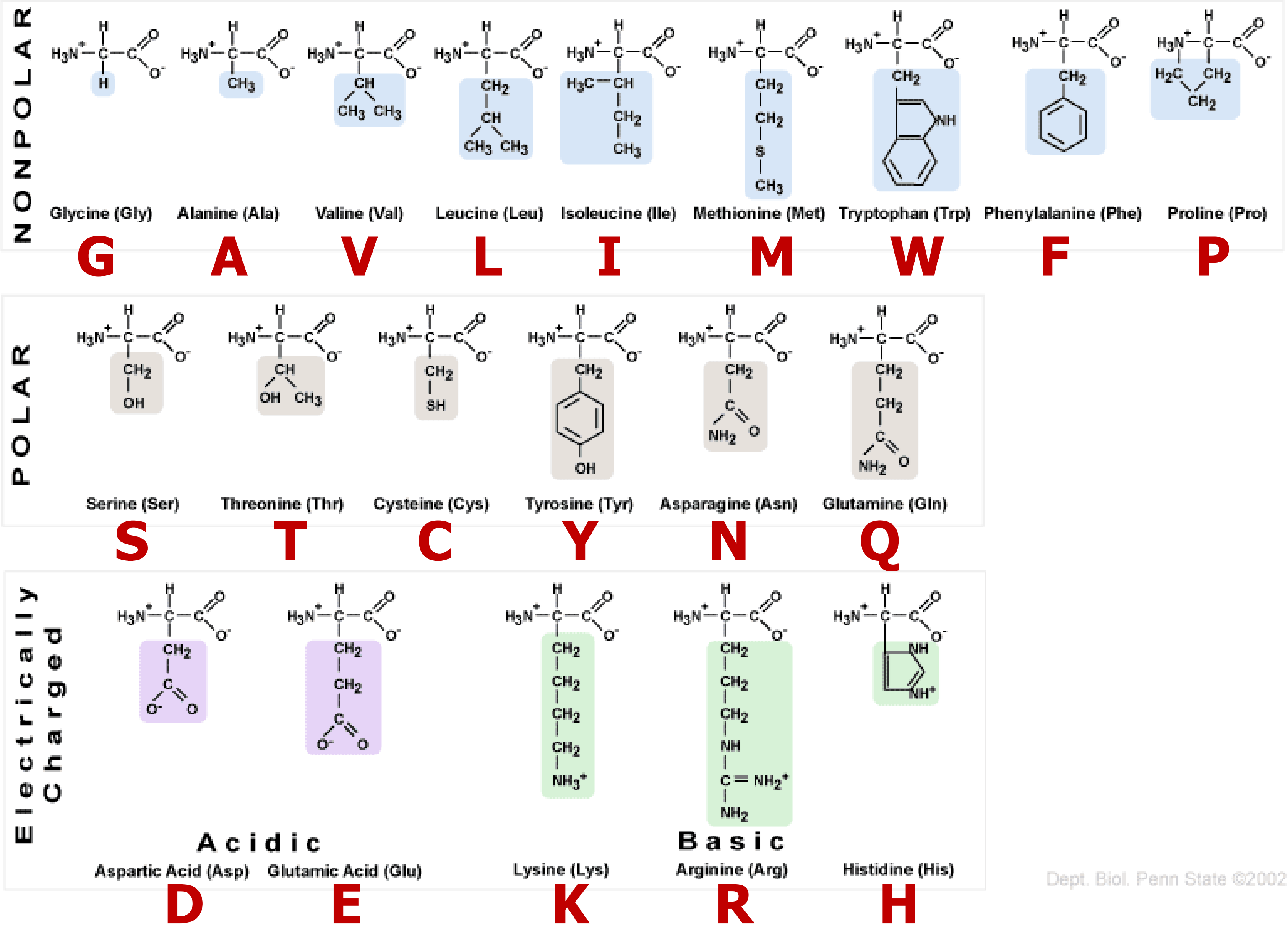
Identities: amino acids that are **identical** at a specific position in the two sequences

Similarity/Conservation: amino acids at a specific position in the two sequences are **not identical**, however they share the same chemical properties (see amino acid classification on the next slide) = they are **similar**

Positives: The **sum** of identical and similar amino acids.

Theresa Schwarz

Classification of amino acids

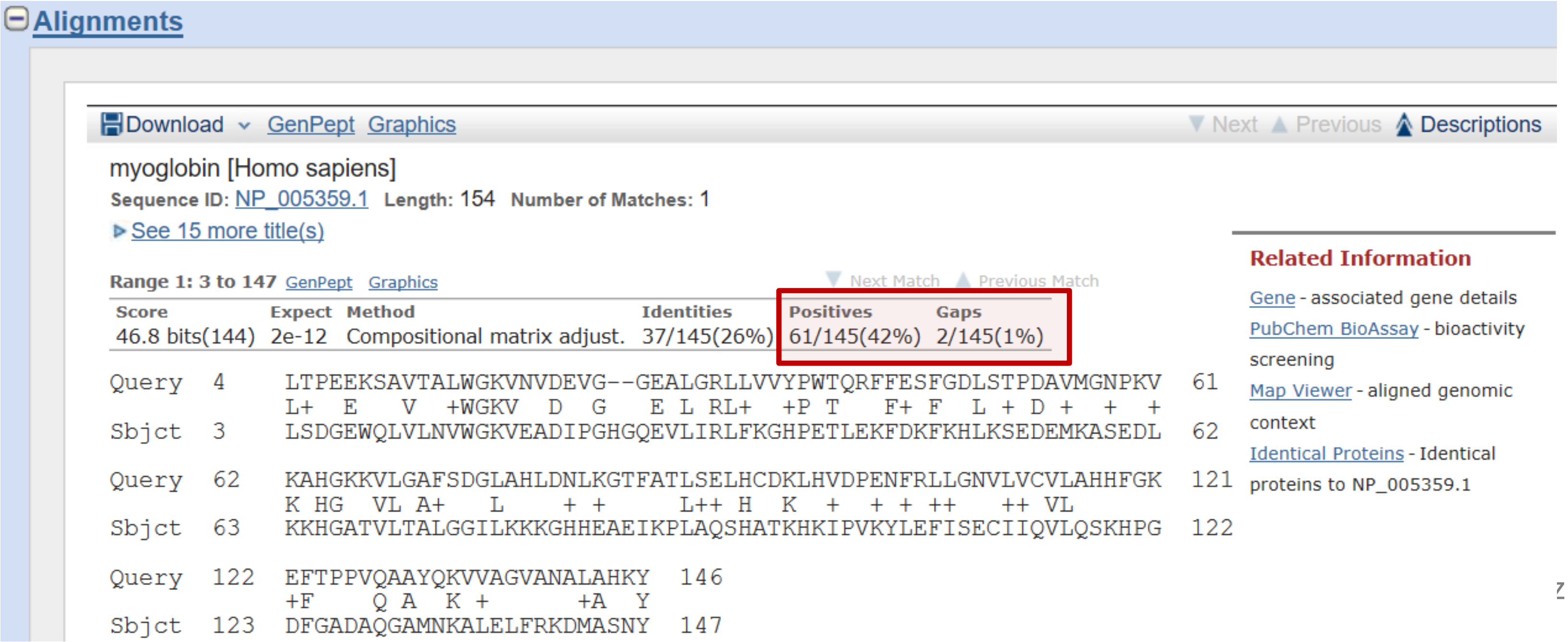


Dept. Biol. Penn State ©2002

Pairwise alignment - BLAST

- You can see again the '**Score**', '**E-value**' and '**Identities**'
- The alignment ranges from aa 3-147 = 145 (relative to the Sbjct)

- **Positives**: aligned amino acids which are either **ident** or **similar**
- **Gaps**: NO alignment with any amino acid. Signed with a ' **–** '


⊟ Alignments
Download ⌄ GenPept Graphics
▼ Next ▲ Previous ▲ Descriptions
▶ See 15 more title(s)
GenPept Graphics
▼ Next Match ▲ Previous Match

myoglobin [Homo sapiens]

Sequence ID: NP_005359.1 Length: 154 Number of Matches: 1

Range 1: 3 to 147

Score	Expect	Method	Identities	Positives	Gaps
46.8 bits(144)	2e-12	Compositional matrix adjust.	37/145(26%)	61/145(42%)	2/145(1%)

```
Query  4    LTPEEKSAVTALWGKVNVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
            L+  E   V  +WGKV  D  G   E L RL+  +P T   F+ F  L + D +  +  +
Sbjct  3    LSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDL  62

Query  62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
            K HG  VL A+   L    + +     L++ H  K  +  + + ++   ++ VL
Sbjct  63   KKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPG  122

Query  122  EFTPPVQAAYQKVVAGVANALAHKY  146
            +F    Q A  K +      +A  Y
Sbjct  123  DFGADAQGAMNKALELFRKDMASNY  147
```

Related Information

Gene - associated gene details

PubChem BioAssay - bioactivity screening

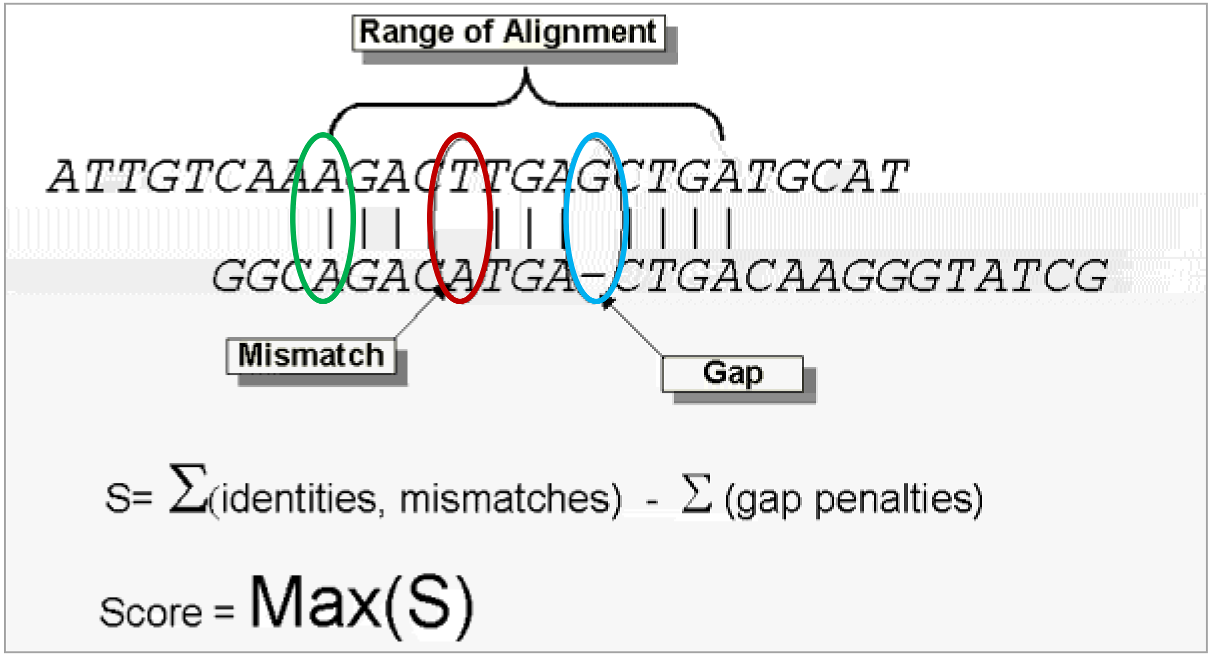
Map Viewer - aligned genomic context

Identical Proteins - Identical proteins to NP_005359.1

Pairwise alignment - BLAST

How is the score calculated?

The score is a sum of match, mismatch and gap.



$$S = \sum(\text{identities, mismatches}) - \sum(\text{gap penalties})$$

$$\text{Score} = \text{Max}(S)$$

Pairwise alignment – Summary

- Choose two sequences
- Select an **algorithm** that generates a score
- This algorithm can be used for global or local alignments
- **Score** reflects degree of similarity (quality control)
- **E-value** tells you the significance of the alignment
- Check whether your results are meaningful → **Query coverage**

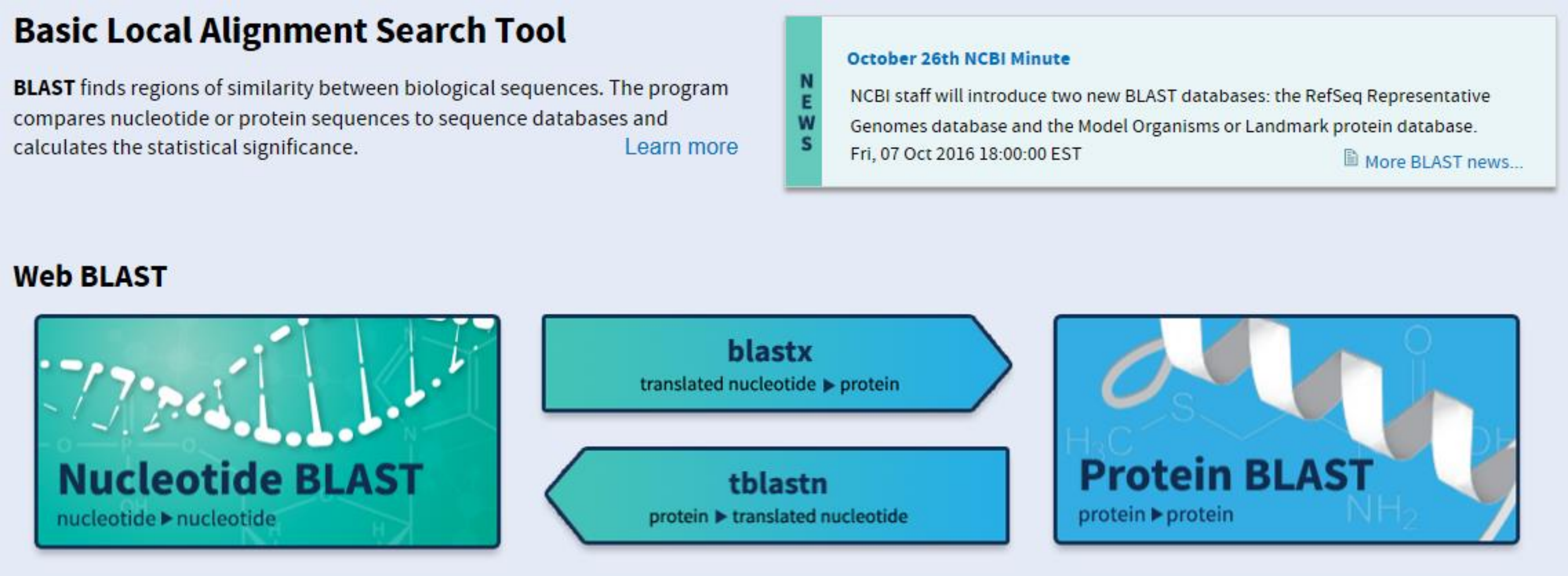
Overview

- ❑ Genome browsers
  - ➢ UCSC
  - ➢ ENSEMBL

- ❑ **BLAST**
  - ➢ Pairwise alignments
  - ➢ **Database alignments**
  - ➢ Primer-BLAST

BLAST

- Despite pairwise alignment you can also align a sequence against an entire database



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

October 26th NCBI Minute

NCBI staff will introduce two new BLAST databases: the RefSeq Representative Genomes database and the Model Organisms or Landmark protein database.

Fri, 07 Oct 2016 18:00:00 EST

More BLAST news...

Web BLAST

BLAST

1. Choose the BLAST program

2. Choose sequence (query)

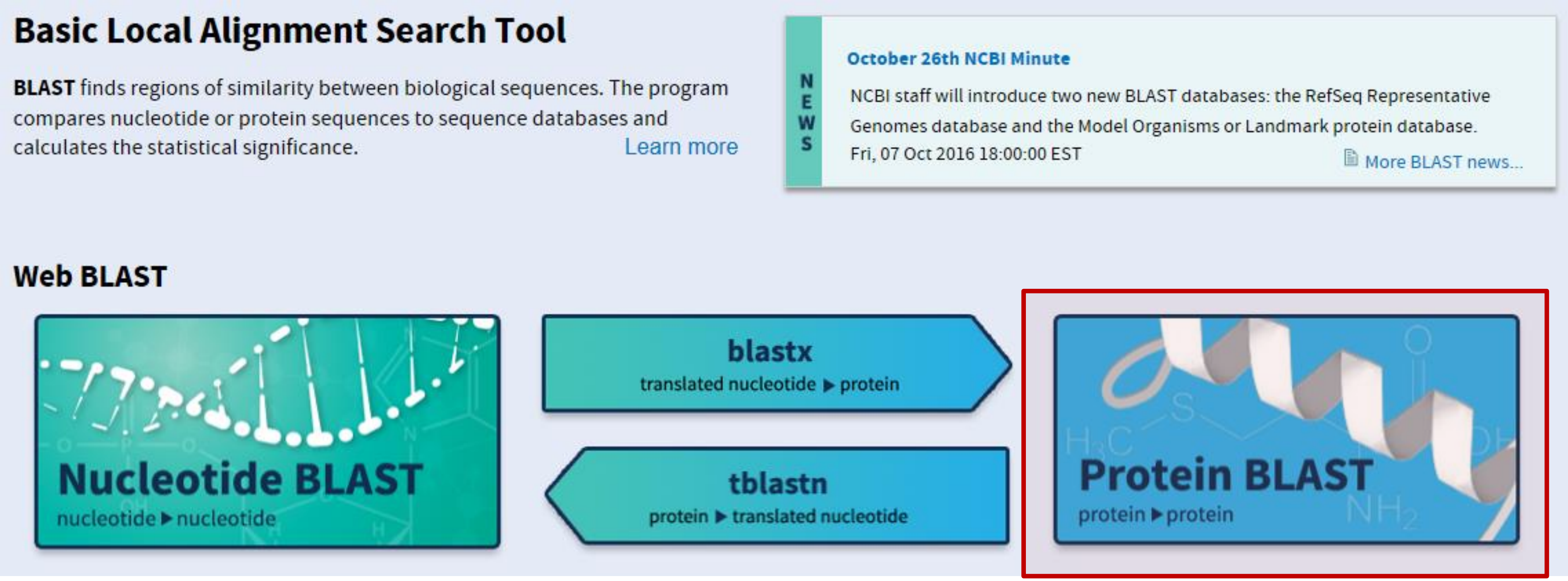
3. Choose the database to search

4. Choose optional parameters

Then click “BLAST”

1. Choose the BLAST program

- BLAST hemoglobin **NP_000509.1** against a very general protein database
- Therefore, first select **Protein BLAST** again.



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

October 26th NCBI Minute

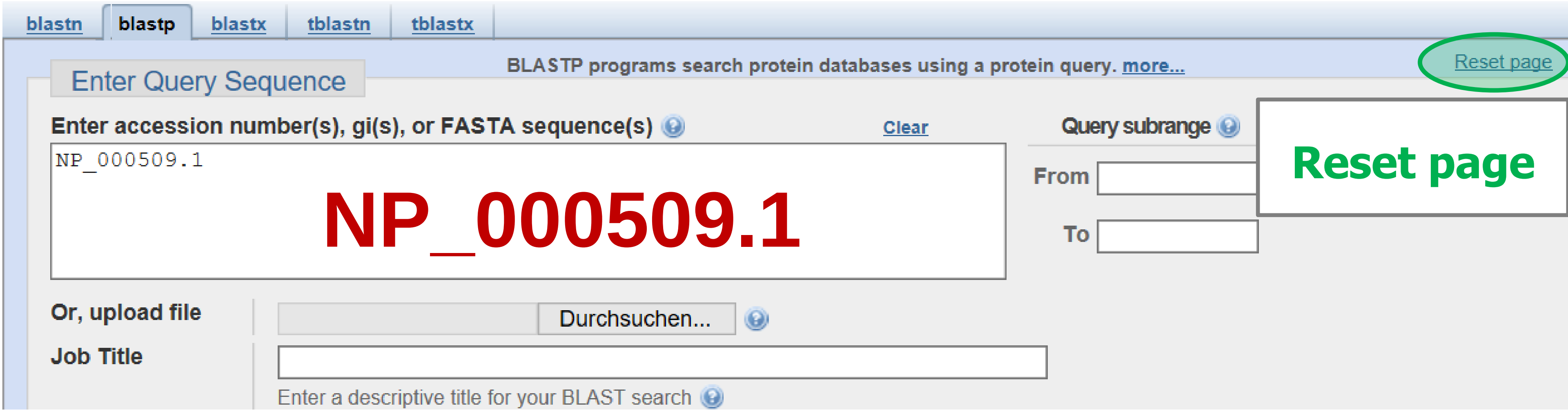
NCBI staff will introduce two new BLAST databases: the RefSeq Representative Genomes database and the Model Organisms or Landmark protein database.

Fri, 07 Oct 2016 18:00:00 EST

More BLAST news...

Web BLAST

2. Choose sequence (query)



BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

NP_000509.1

Query subrange

From

To

Or, upload file

Job Title

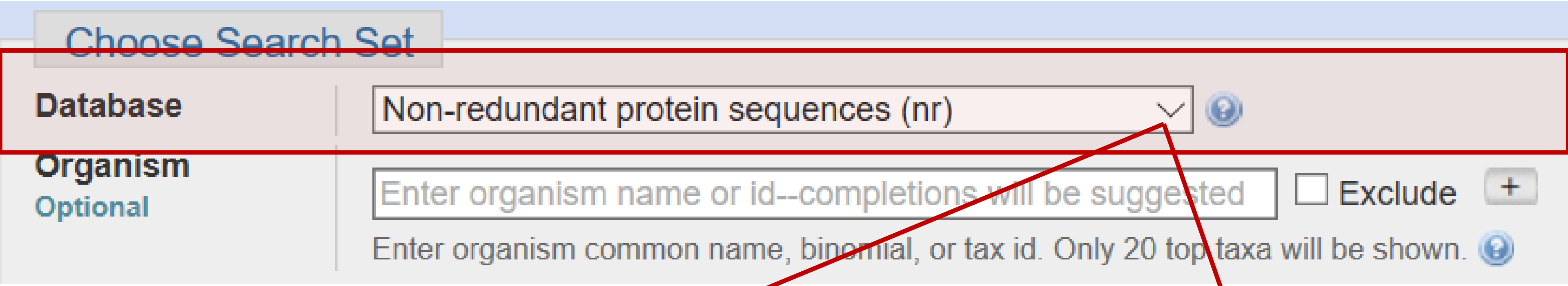
Enter a descriptive title for your BLAST search

NP_000509.1

Reset page

Sequence can be entered in FASTA format or as accession number

3. Choose the database



Choose Search Set

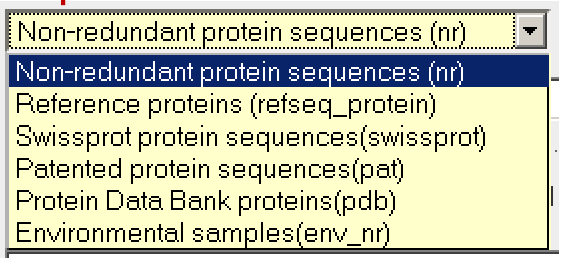
Database: Non-redundant protein sequences (nr)

Organism
Optional

Enter organism name or id--completions will be suggested □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

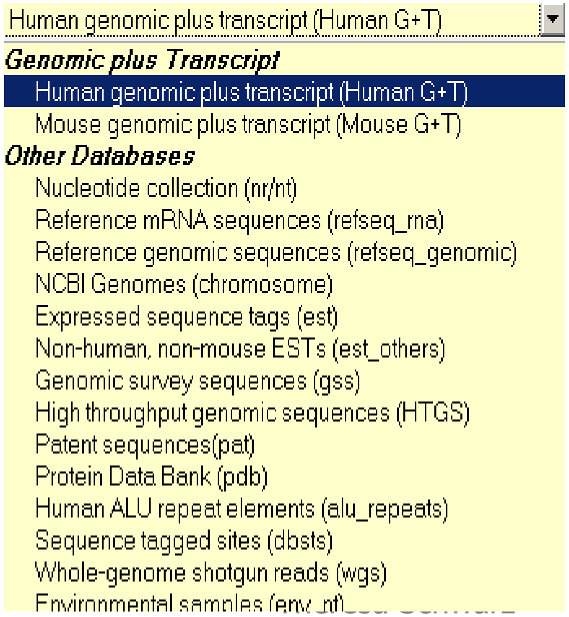
protein databases



Non-redundant protein sequences (nr)

- Non-redundant protein sequences (nr)
- Reference proteins (refseq_protein)
- Swissprot protein sequences(swissprot)
- Patented protein sequences(pat)
- Protein Data Bank proteins(pdb)
- Environmental samples(env_nr)

nucleotide databases



Human genomic plus transcript (Human G+T)

Genomic plus Transcript

- Human genomic plus transcript (Human G+T)
- Mouse genomic plus transcript (Mouse G+T)

Other Databases

- Nucleotide collection (nr/nt)
- Reference mRNA sequences (refseq_rna)
- Reference genomic sequences (refseq_genomic)
- NCBI Genomes (chromosome)
- Expressed sequence tags (est)
- Non-human, non-mouse ESTs (est_others)
- Genomic survey sequences (gss)
- High throughput genomic sequences (HTGS)
- Patent sequences(pat)
- Protein Data Bank (pdb)
- Human ALU repeat elements (alu_repeats)
- Sequence tagged sites (dbsts)
- Whole-genome shotgun reads (wgs)
- Environmental samples (env_nt)

nr = non-redundant (most general database)

4. Choose optional parameters

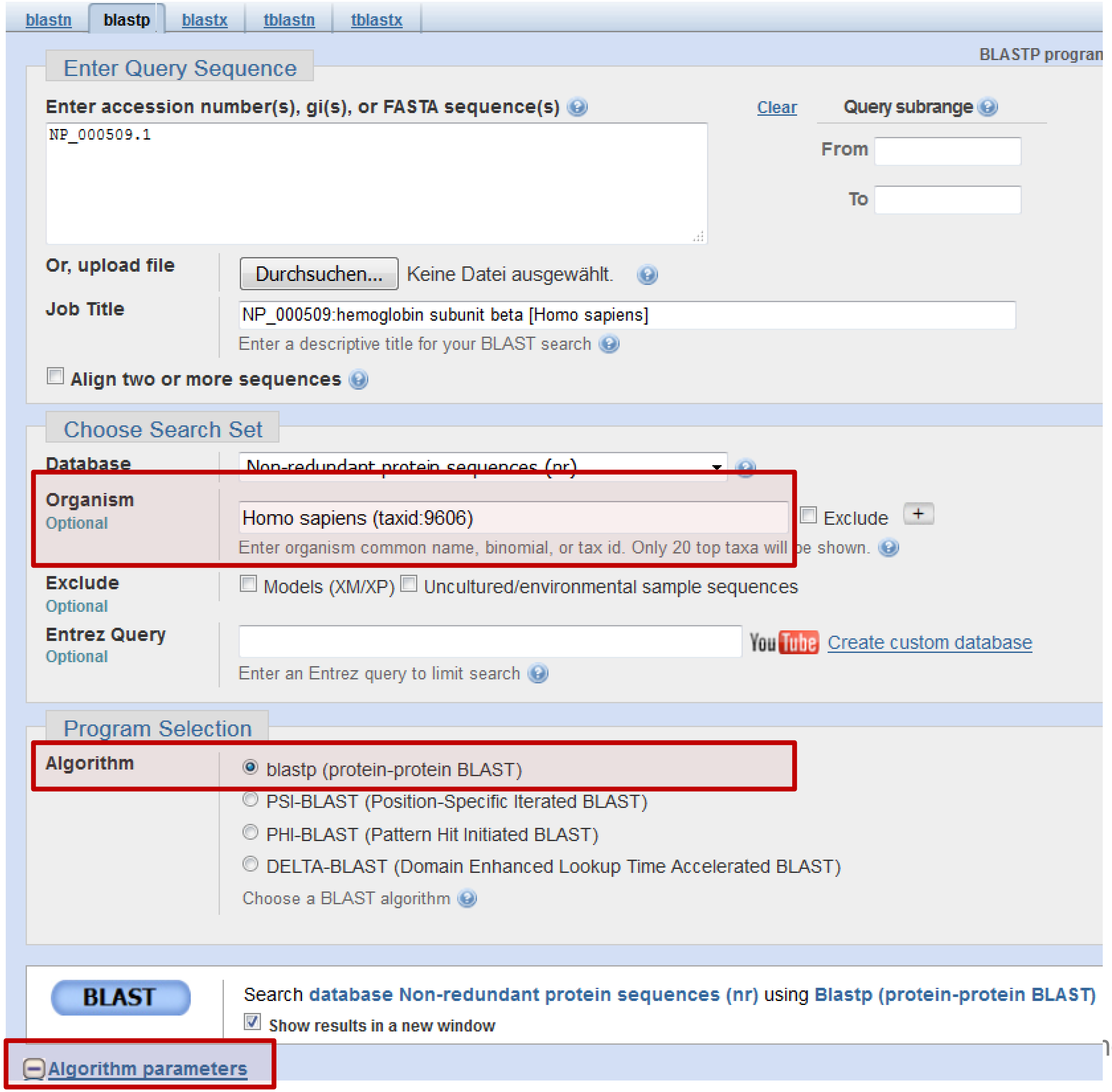
You can...

- choose the organism to search
- turn filtering on/off
- change the substitution matrix
- change the expect (e) value
- change the word size
- change the output format

4. Choose optional parameters

Organism

Algorithm



blastn | blastp | blastx | tblastn | tblastx

BLASTP progran

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

```
NP_000509.1
```

From

To

Or, upload file: Durchsuchen... Keine Datei ausgewählt.

Job Title: NP_000509:hemoglobin subunit beta [Homo sapiens]

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism
Optional: Homo sapiens (taxid:9606) □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude
Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query
Optional:

Create custom database

Enter an Entrez query to limit search

Program Selection

Algorithm:

● blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)

○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

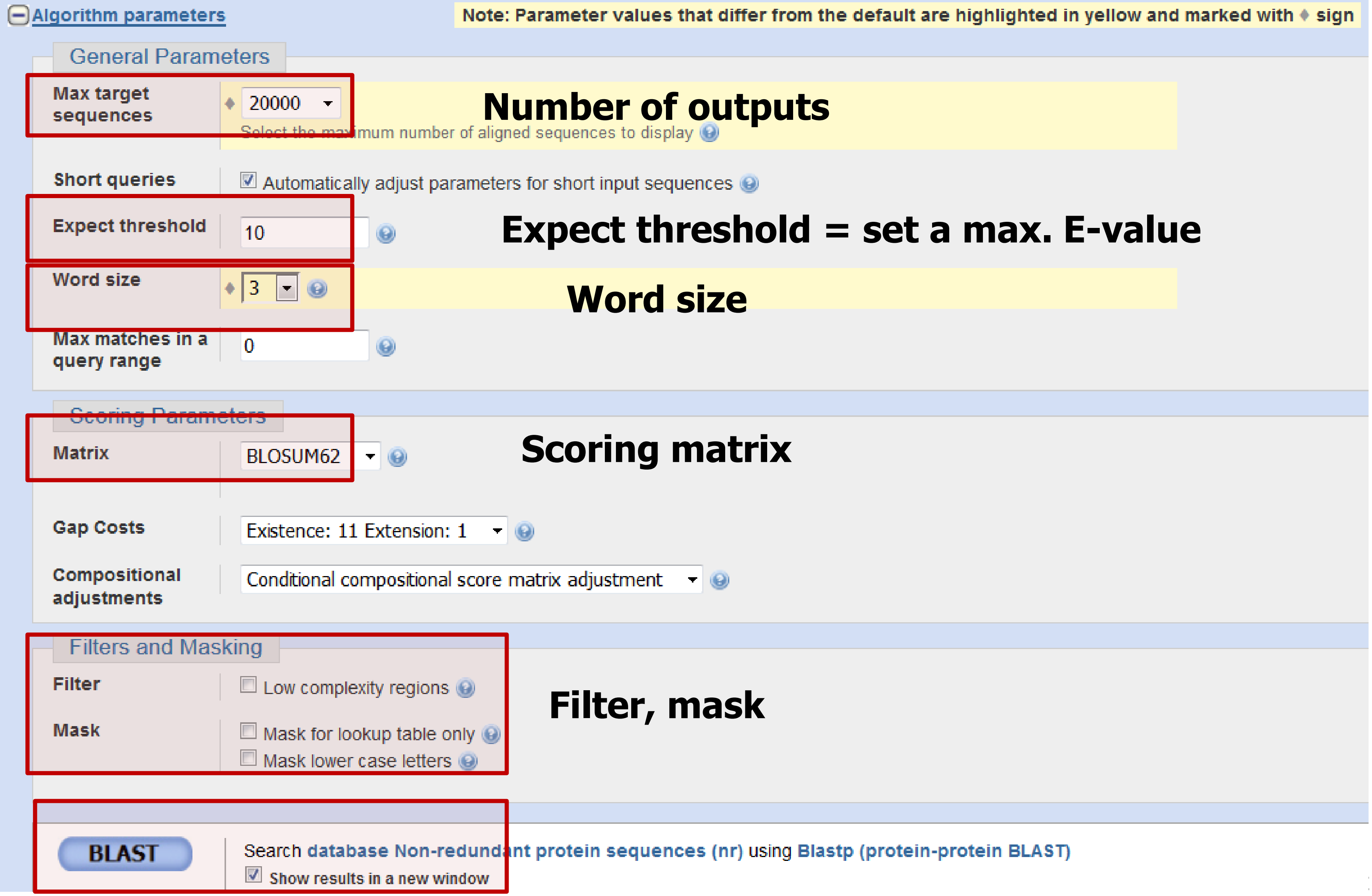
BLAST

Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

☑ Show results in a new window

Algorithm parameters

4. Choose optional parameters



⊟ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

General Parameters

Max target sequences: ♦ 20000 ▾
Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: ♦ 3 ▾

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62 ▾

Gap Costs: Existence: 11 Extension: 1 ▾

Compositional adjustments: Conditional compositional score matrix adjustment ▾

Filters and Masking

Filter: □ Low complexity regions

Mask: □ Mask for lookup table only
□ Mask lower case letters

BLAST

Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
☑ Show results in a new window

Number of outputs

Expect threshold = set a max. E-value

Word size

Scoring matrix

Filter, mask

BLAST – word size

1. The query sequence is cut into pieces (“**words**”)
   e.g. one piece consists of **3** amino acids when the word size is **3**

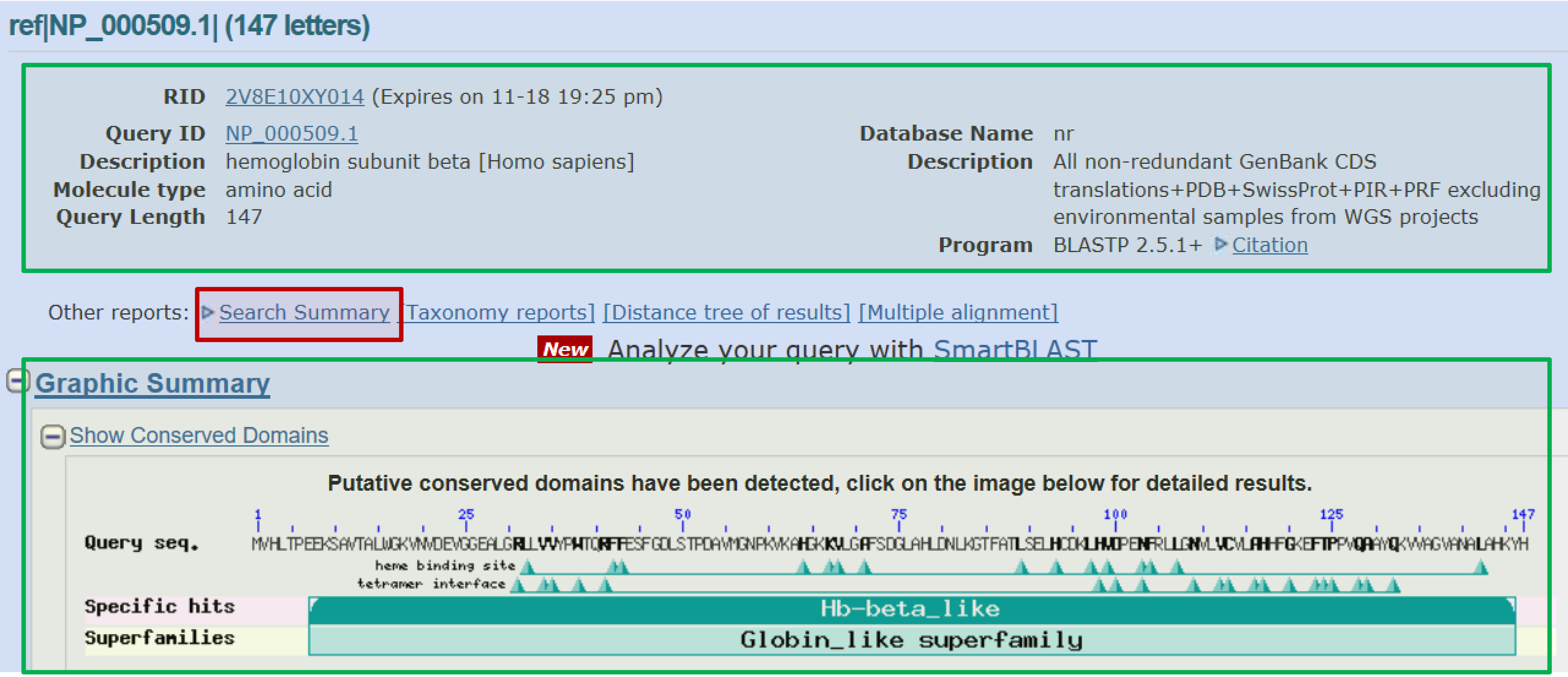
2. The BLAST algorithm uses those “words” to find similar regions in sequences which are present in the chosen database

3. The default word size = 3 amino acids or 11 nucleic acids

Word size = 3

```
SWVSQA = Query
SWV
 WVS
  VSQ
   SQA
```

BLAST Results



ref|NP_000509.1| (147 letters)

RID 2V8E10XY014 (Expires on 11-18 19:25 pm)

Query ID NP_000509.1

Description hemoglobin subunit beta [Homo sapiens]

Molecule type amino acid

Query Length 147

Database Name nr

Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects

Program BLASTP 2.5.1+ ▶ Citation

Other reports: ▶ Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]

New Analyze your query with SmartBLAST

Graphic Summary

Show Conserved Domains

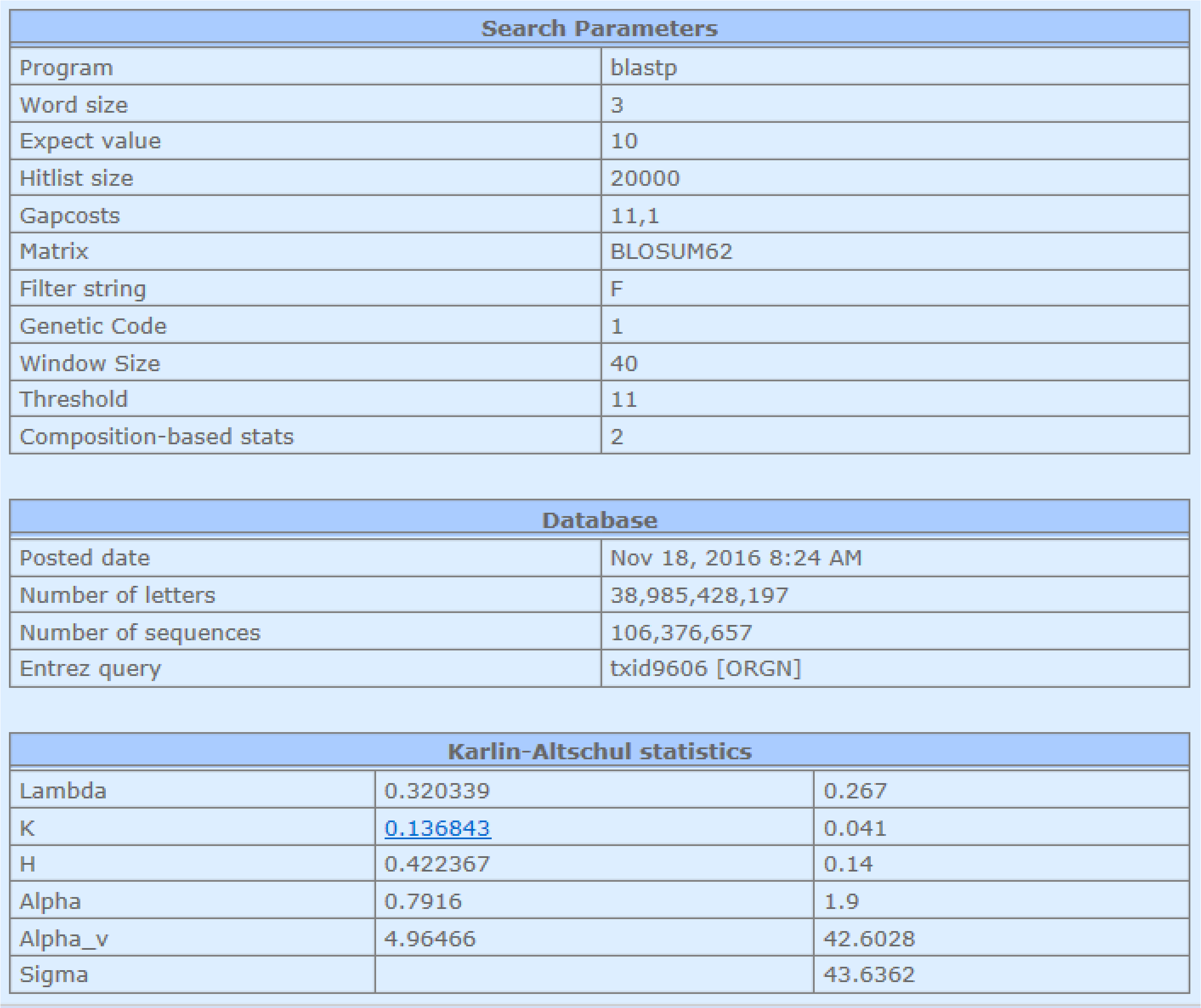
Putative conserved domains have been detected, click on the image below for detailed results.

1
25
50
75
100
125
147
Query seq.
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
heme binding site
tetramer interface
Specific hits
Hb-beta_like
Superfamilies
Globin_like superfamily

You can find:

- **general information** about your search
- **Graphic Summary** with the **Conserved Domains** of your Query sequence
- click on **search summary** to see all the chosen parameters

BLAST Results



Search Parameters	
Program	blastp
Word size	3
Expect value	10
Hitlist size	20000
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	11
Composition-based stats	2

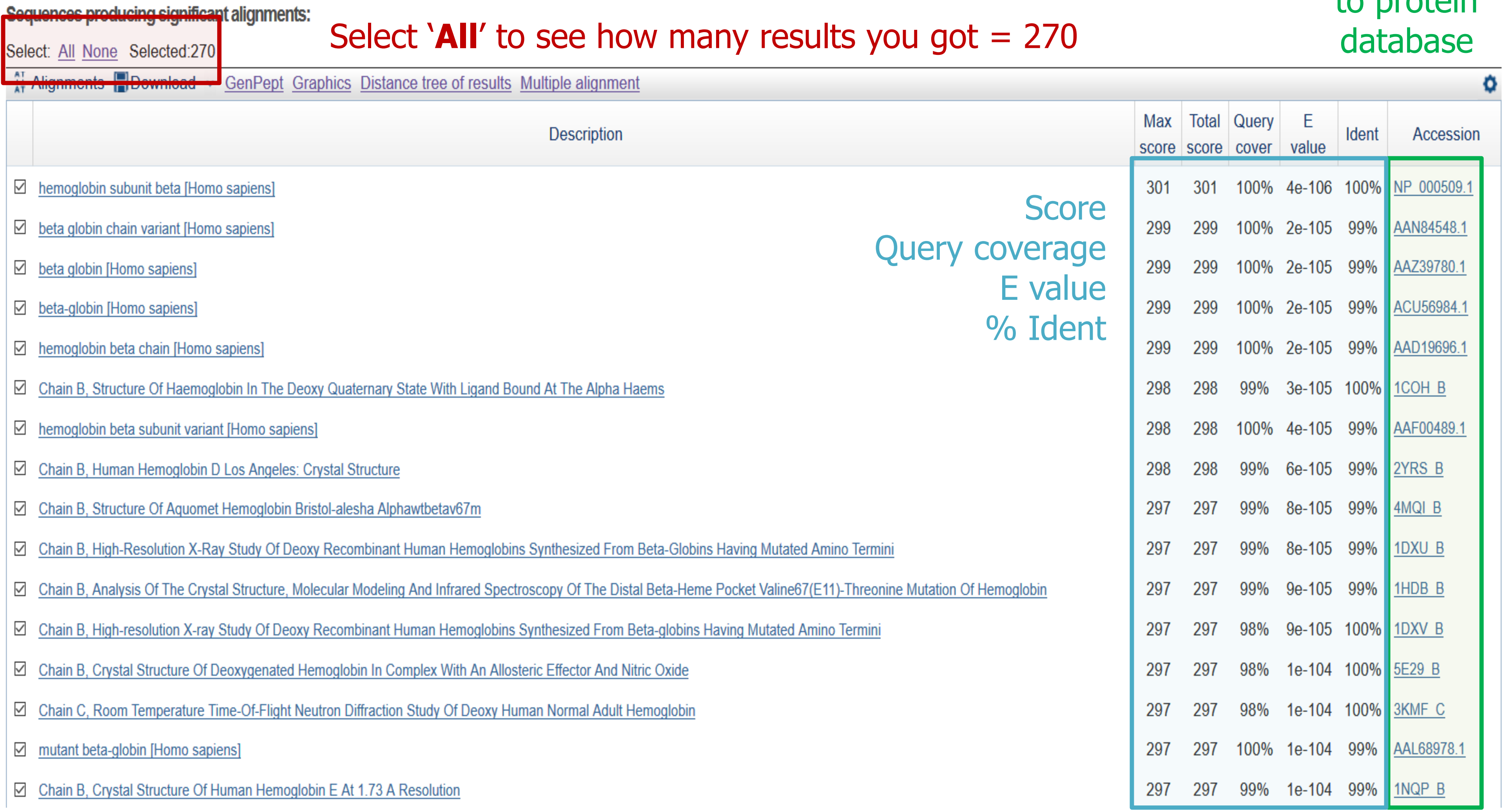
Database	
Posted date	Nov 18, 2016 8:24 AM
Number of letters	38,985,428,197
Number of sequences	106,376,657
Entrez query	txid9606 [ORGN]

Karlin-Altschul statistics		
Lambda	0.320339	0.267
K	0.136843	0.041
H	0.422367	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

Schwarz

BLAST Results

- Scroll down to **Descriptions** to get an overview of all the results



Select '**All**' to see how many results you got = 270

Direct links to protein database

Score
Query coverage
E value
% Ident

Sequences producing significant alignments:

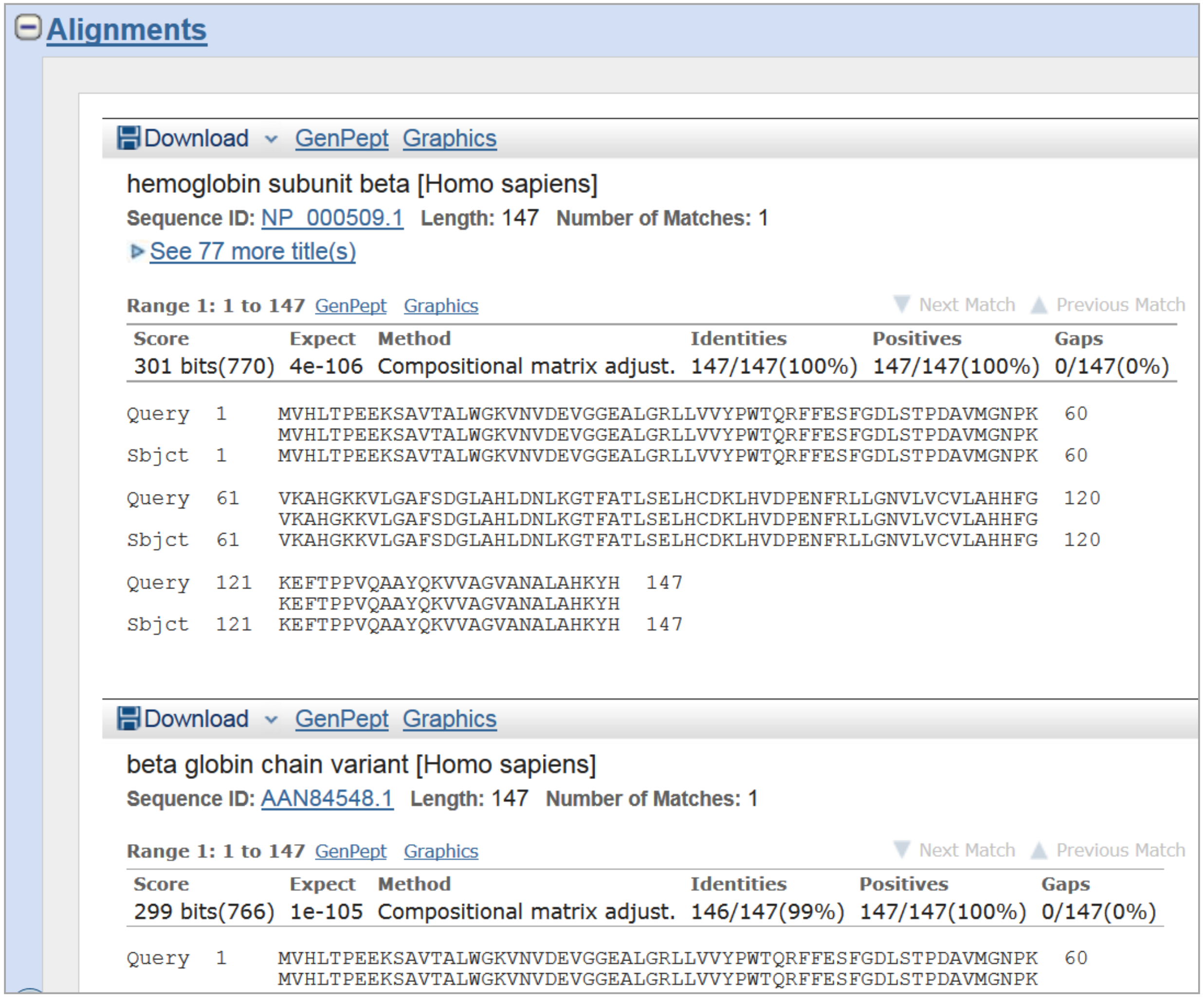
Select: All None Selected:270

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☑	hemoglobin subunit beta [Homo sapiens]	301	301	100%	4e-106	100%	NP_000509.1
☑	beta globin chain variant [Homo sapiens]	299	299	100%	2e-105	99%	AAN84548.1
☑	beta globin [Homo sapiens]	299	299	100%	2e-105	99%	AAZ39780.1
☑	beta-globin [Homo sapiens]	299	299	100%	2e-105	99%	ACU56984.1
☑	hemoglobin beta chain [Homo sapiens]	299	299	100%	2e-105	99%	AAD19696.1
☑	Chain B, Structure Of Haemoglobin In The Deoxy Quaternary State With Ligand Bound At The Alpha Haems	298	298	99%	3e-105	100%	1COH_B
☑	hemoglobin beta subunit variant [Homo sapiens]	298	298	100%	4e-105	99%	AAF00489.1
☑	Chain B, Human Hemoglobin D Los Angeles: Crystal Structure	298	298	99%	6e-105	99%	2YRS_B
☑	Chain B, Structure Of Aquomet Hemoglobin Bristol-alesha Alphawtbetav67m	297	297	99%	8e-105	99%	4MQI_B
☑	Chain B, High-Resolution X-Ray Study Of Deoxy Recombinant Human Hemoglobins Synthesized From Beta-Globins Having Mutated Amino Termini	297	297	99%	8e-105	99%	1DXU_B
☑	Chain B, Analysis Of The Crystal Structure, Molecular Modeling And Infrared Spectroscopy Of The Distal Beta-Heme Pocket Valine67(E11)-Threonine Mutation Of Hemoglobin	297	297	99%	9e-105	99%	1HDB_B
☑	Chain B, High-resolution X-ray Study Of Deoxy Recombinant Human Hemoglobins Synthesized From Beta-globins Having Mutated Amino Termini	297	297	98%	9e-105	100%	1DXV_B
☑	Chain B, Crystal Structure Of Deoxygenated Hemoglobin In Complex With An Allosteric Effector And Nitric Oxide	297	297	98%	1e-104	100%	5E29_B
☑	Chain C, Room Temperature Time-Of-Flight Neutron Diffraction Study Of Deoxy Human Normal Adult Hemoglobin	297	297	98%	1e-104	100%	3KMF_C
☑	mutant beta-globin [Homo sapiens]	297	297	100%	1e-104	99%	AAL68978.1
☑	Chain B, Crystal Structure Of Human Hemoglobin E At 1.73 A Resolution	297	297	99%	1e-104	99%	1NQP_B

BLAST Results

- Scroll down to **Alignments** to get **details** for single alignments



Alignments

Download GenPept Graphics

hemoglobin subunit beta [Homo sapiens]

Sequence ID: NP_000509.1 **Length:** 147 **Number of Matches:** 1

See 77 more title(s)

Range 1: 1 to 147 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
301 bits(770)	4e-106	Compositional matrix adjust.	147/147(100%)	147/147(100%)	0/147(0%)

```
Query  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
            MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
Sbjct  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60

Query  61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120
            VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
Sbjct  61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120

Query  121  KEFTPPVQAAYQKVVAGVANALAHKYH  147
            KEFTPPVQAAYQKVVAGVANALAHKYH
Sbjct  121  KEFTPPVQAAYQKVVAGVANALAHKYH  147
```

Download GenPept Graphics

beta globin chain variant [Homo sapiens]

Sequence ID: AAN84548.1 **Length:** 147 **Number of Matches:** 1

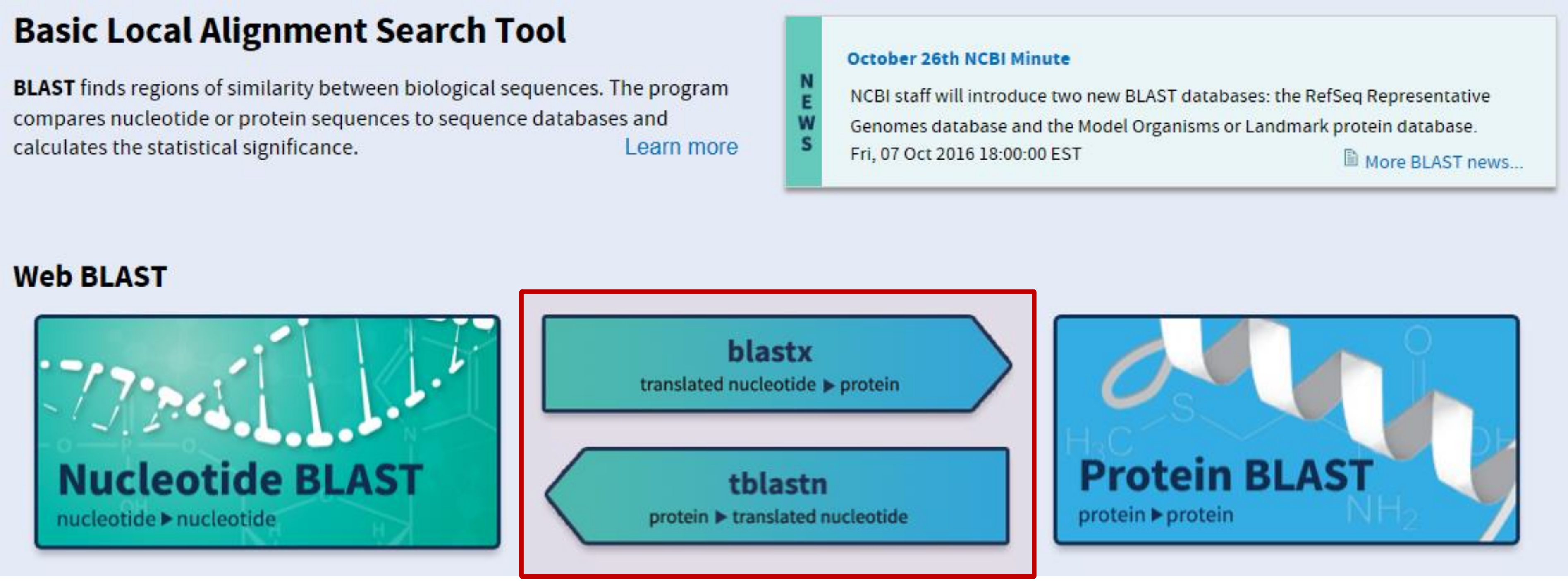
Range 1: 1 to 147 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
299 bits(766)	1e-105	Compositional matrix adjust.	146/147(99%)	147/147(100%)	0/147(0%)

```
Query  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
            MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
```

BLAST programs

- What other BLAST programs can we use?



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

October 26th NCBI Minute

NCBI staff will introduce two new BLAST databases: the RefSeq Representative Genomes database and the Model Organisms or Landmark protein database.

Fri, 07 Oct 2016 18:00:00 EST

More BLAST news...

Web BLAST

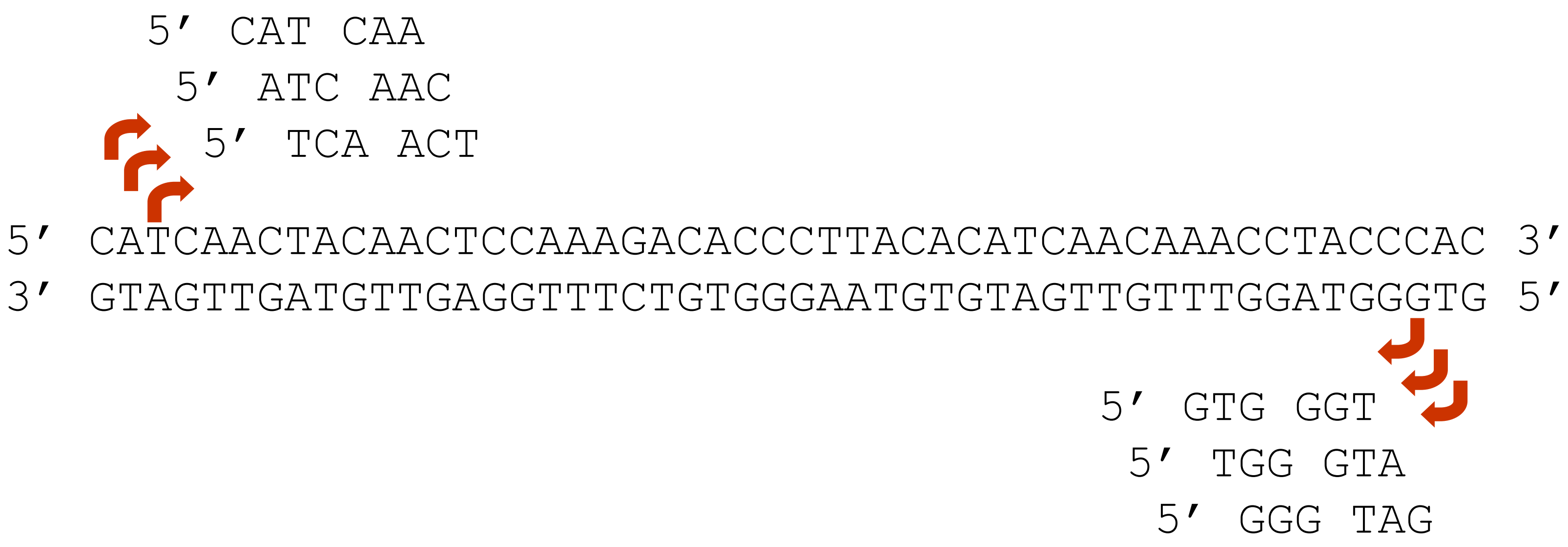
- Nucleotide BLAST = blastn
- Protein BLAST = blastp
- **blastx**
- **tblastn**

BLAST programs

Program	**Input**		**Database**
blastn	nt	→	nt
blastp	protein	→	protein
blastx	nt	→	protein
tblastn	protein	→	nt

BLAST programs - blastx

- When using blastx the input is a **nucleotide sequence**
- Then the program translates this sequence into a **protein sequence**
- Since the program does not know where the translation starts there are 6 possibilities



```
      5′ CAT CAA
       5′ ATC AAC
        5′ TCA ACT

5′ CATCAACTACAACTCCAAAGACACCCTTACACATCAACAAACCTACCCAC 3′
3′ GTAGTTGATGTTGAGGTTTCTGTGGGAATGTGTAGTTGTTTGGATGGGTG 5′

                                   5′ GTG GGT
                                    5′ TGG GTA
                                     5′ GGG TAG
```

BLAST programs

Program	Input	Database	
blastn	nt	nt	
blastp	protein	protein	
blastx	nt	protein	Search protein database using a translated nt query
tblastn	protein	nt	Search translated nt database using a protein query

Overview

❑ **Genome browsers**

- ➢ **UCSC**
- ➢ **ENSEMBL**

❑ **BLAST**

- ➢ **Pairwise alignments**
- ➢ **Database alignments**
- ➢ **Primer-BLAST**

Primer-BLAST at NCBI

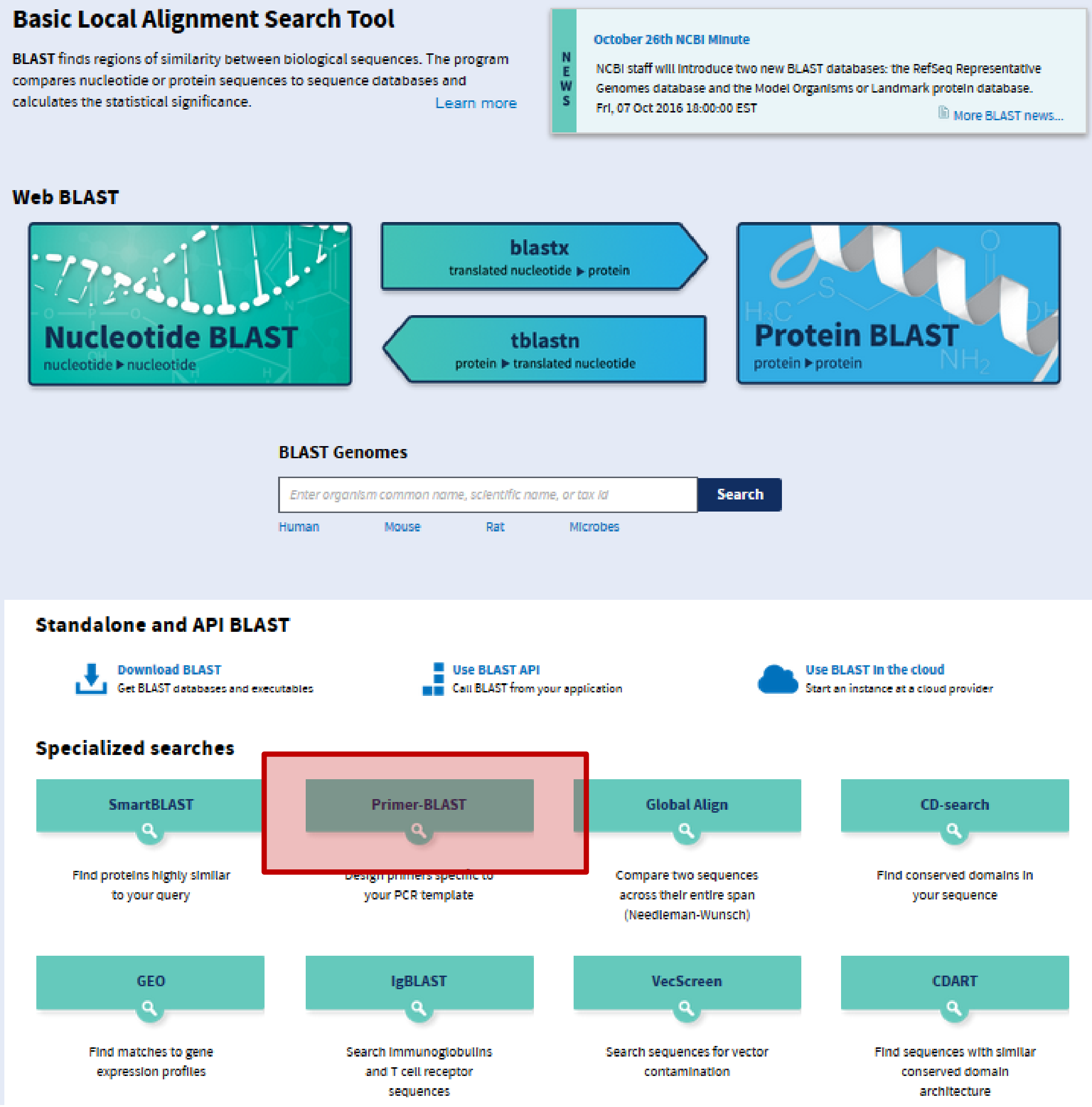
With **Primer-BLAST** you can check your primers that you designed to use them in a PCR

- do they amplify the desired product
- do they also bind to other regions in the given template

Go to BLAST (NCBI)

https://blast.ncbi.nlm.nih.gov/Blast.cgi

Under '**Specialized searches**' you can find **Primer-BLAST**



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

October 26th NCBI Minute

NCBI staff will introduce two new BLAST databases: the RefSeq Representative Genomes database and the Model Organisms or Landmark protein database.

Fri, 07 Oct 2016 18:00:00 EST

More BLAST news...

Web BLAST

Nucleotide BLAST
nucleotide ▸ nucleotide

blastx
translated nucleotide ▸ protein

tblastn
protein ▸ translated nucleotide

Protein BLAST
protein ▸ protein

BLAST Genomes

Enter organism common name, scientific name, or tax id | Search

Human | Mouse | Rat | Microbes

Standalone and API BLAST

Download BLAST
Get BLAST databases and executables

Use BLAST API
Call BLAST from your application

Use BLAST in the cloud
Start an instance at a cloud provider

Specialized searches

SmartBLAST
Find proteins highly similar to your query

Primer-BLAST
Design primers specific to your PCR template

Global Align
Compare two sequences across their entire span (Needleman-Wunsch)

CD-search
Find conserved domains in your sequence

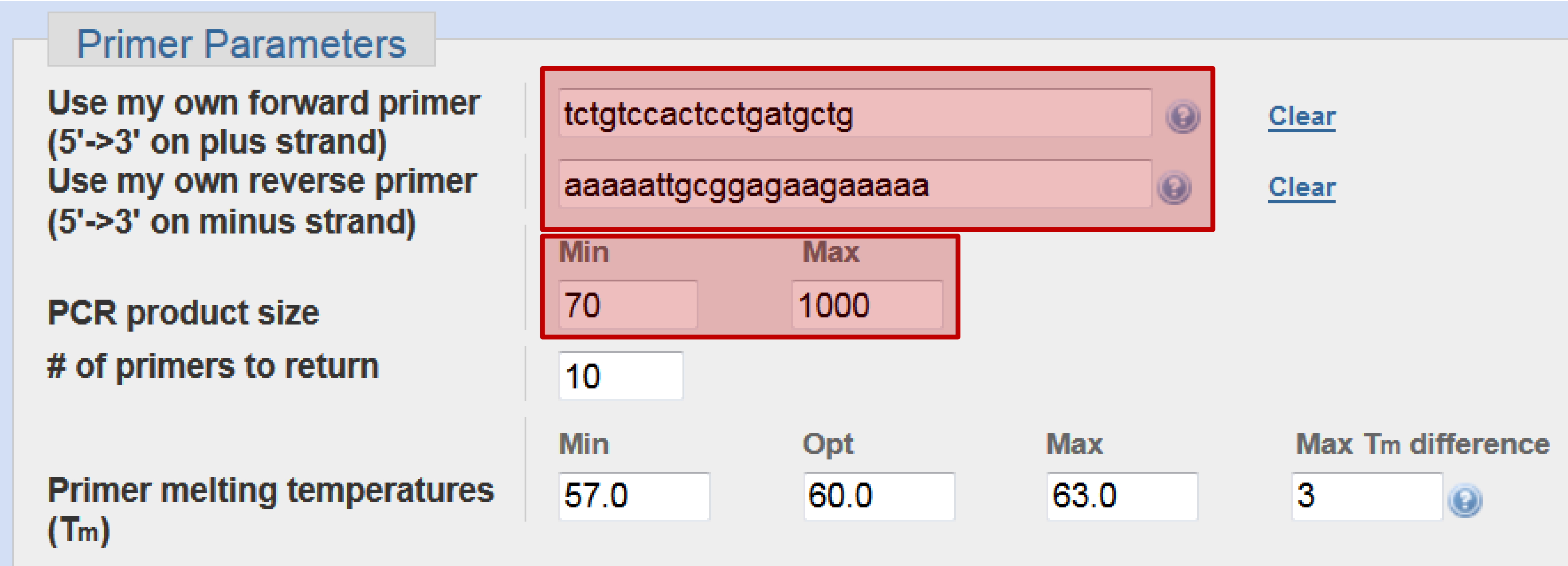
GEO
Find matches to gene expression profiles

IgBLAST
Search immunoglobulins and T cell receptor sequences

VecScreen
Search sequences for vector contamination

CDART
Find sequences with similar conserved domain architecture

Primer-BLAST at NCBI



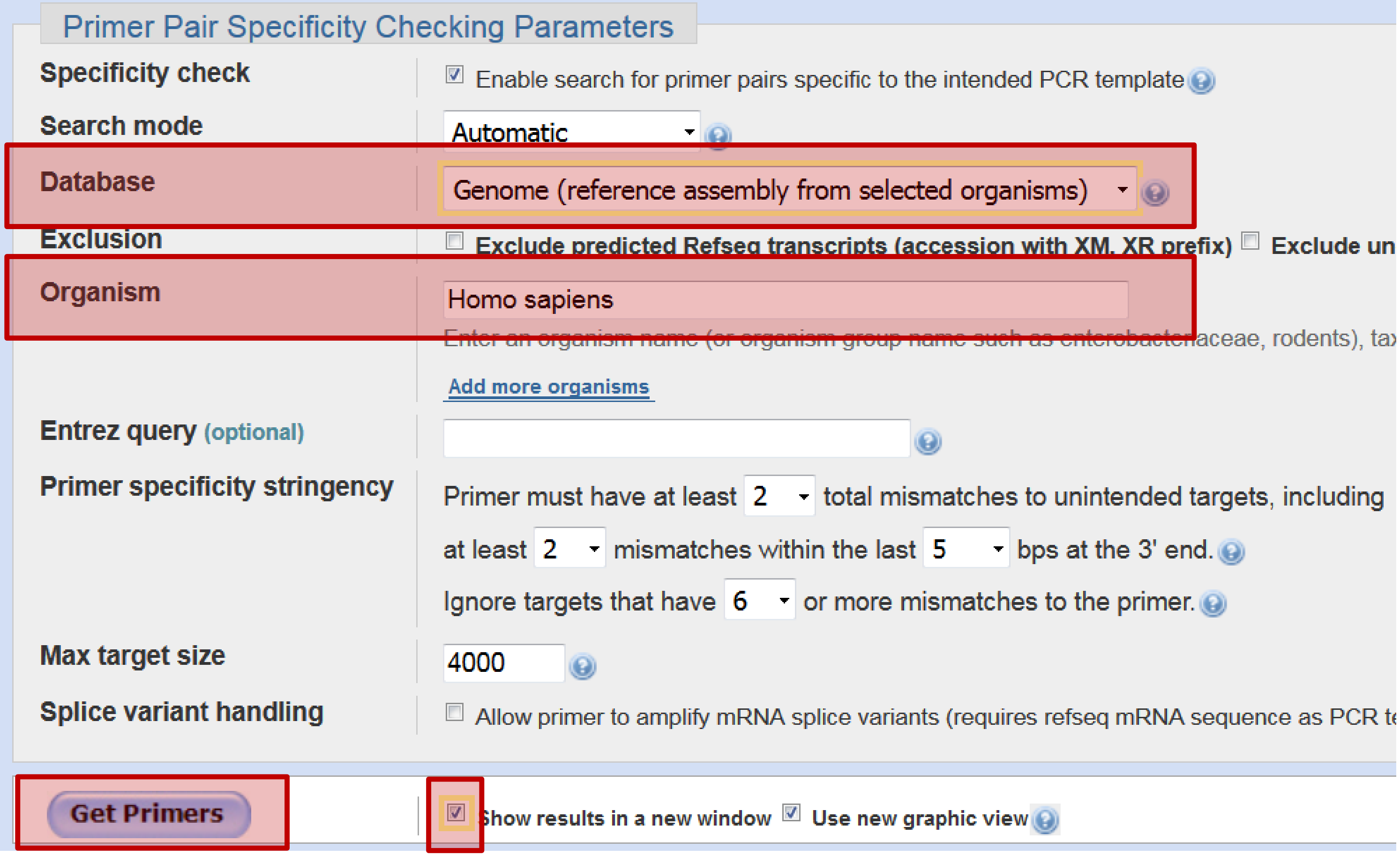
Primer Parameters

Use my own forward primer (5'->3' on plus strand)	tctgtccactcctgatgctg				Clear
Use my own reverse primer (5'->3' on minus strand)	aaaaattgcggagaagaaaaa				Clear
	Min	Max			
PCR product size	70	1000			
# of primers to return	10				
	Min	Opt	Max	Max Tm difference	
Primer melting temperatures (Tm)	57.0	60.0	63.0	3	

- Primers designed to amplify **hemoglobin subunit beta** of **401bp**
- Enter the **primer sequences** (You can download the sequences from MOODLE)
- You can select a **minimum** and **maximum product size** …

Theresa Schwarz

Primer-BLAST at NCBI



Primer Pair Specificity Checking Parameters

Specificity check ☑ Enable search for primer pairs specific to the intended PCR template

Search mode: Automatic

Database: Genome (reference assembly from selected organisms)

Exclusion ☐ Exclude predicted Refseq transcripts (accession with XM, XR prefix) ☐ Exclude un

Organism: Homo sapiens

Enter an organism name (or organism group name such as enterobacteriaceae, rodents), tax

Add more organisms

Entrez query (optional)

Primer specificity stringency: Primer must have at least 2 total mismatches to unintended targets, including at least 2 mismatches within the last 5 bps at the 3' end. Ignore targets that have 6 or more mismatches to the primer.

Max target size: 4000

Splice variant handling ☐ Allow primer to amplify mRNA splice variants (requires refseq mRNA sequence as PCR te

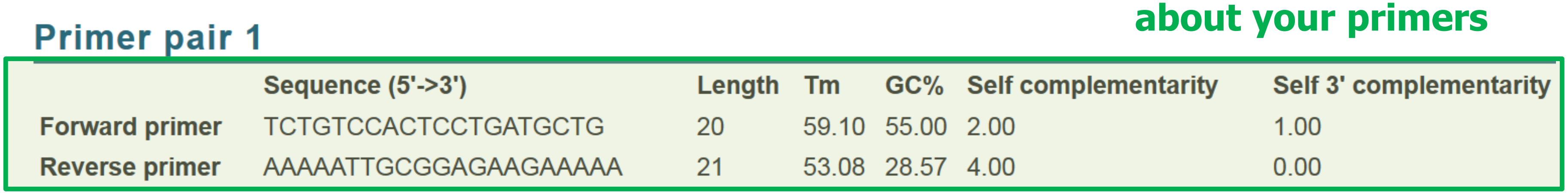
☑ Show results in a new window ☑ Use new graphic view

- As template I planned to use **human gDNA**
- Select the **database** and **organism** you want to check
- **Get Primers**

Primer-BLAST at NCBI

Output

General information about your primers

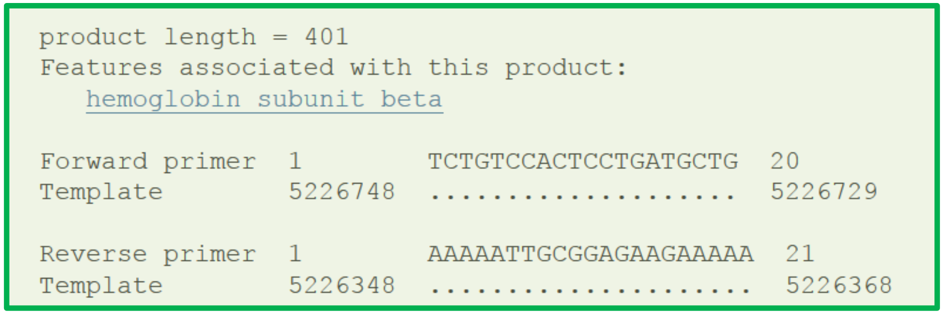


Primer pair 1

	Sequence (5'->3')	Length	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	TCTGTCCACTCCTGATGCTG	20	59.10	55.00	2.00	1.00
Reverse primer	AAAAATTGCGGAGAAGAAAAA	21	53.08	28.57	4.00	0.00

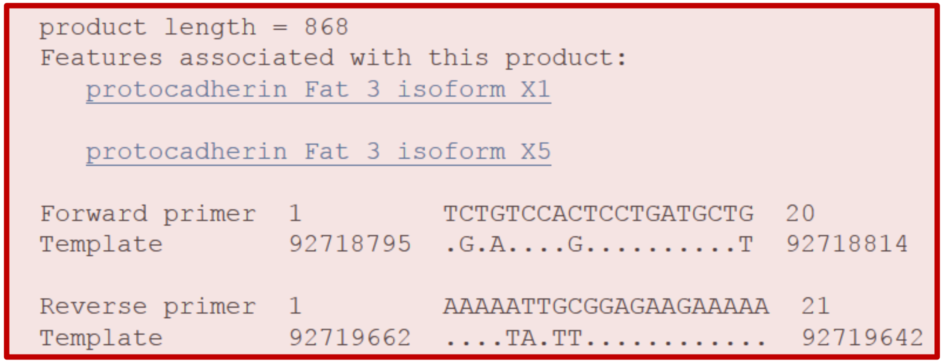
Products on target templates

>NC_000011.10 Homo sapiens chromosome 11, GRCh38.p7 Primary Assembly



```
product length = 401
Features associated with this product:
   hemoglobin subunit beta

Forward primer  1         TCTGTCCACTCCTGATGCTG  20
Template        5226748   ....................  5226729

Reverse primer  1         AAAAATTGCGGAGAAGAAAAA  21
Template        5226348   .....................  5226368
```

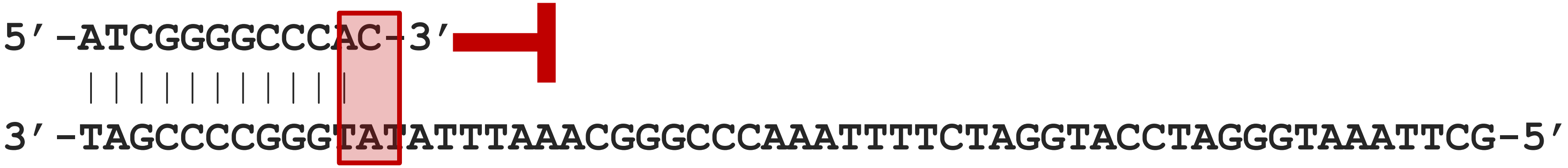
Product length of the amplicon = **401bp**
The product is **hemoglobin subunit beta**
Is this the region we expected? → **YES**



```
product length = 868
Features associated with this product:
   protocadherin Fat 3 isoform X1

   protocadherin Fat 3 isoform X5

Forward primer  1          TCTGTCCACTCCTGATGCTG  20
Template        92718795   .G.A....G.........T   92718814

Reverse primer  1          AAAAATTGCGGAGAAGAAAAA  21
Template        92719662   ....TA.TT............  92719642
```

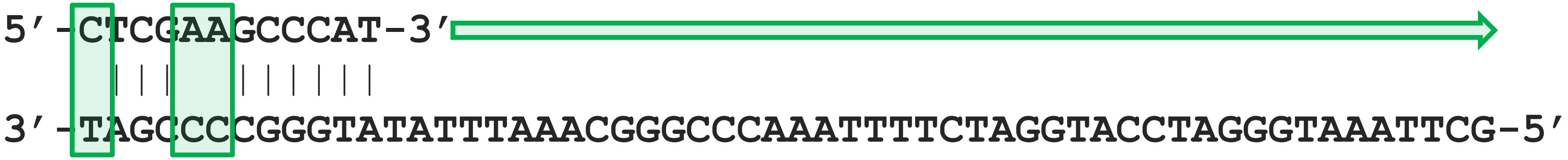
Do we want to amplify this region? → **NO**
The primers can also anneal to a different region in our genome. However, there are some mismatches in the annealing sequence.
Is this now a problem for our PCR?
Do we have to design new primers?

Mismatches in primer sequence



```
5′-ATCGGGGCCCAC-3′
     | | | | | | | | | | |
3′-TAGCCCCGGGTATATTTAAACGGGCCCAAATTTTCTAGGTACCTAGGGTAAATTCG-5′
```

- In most cases, a mismatch at the 3′-end of the primer (where the polymerase attaches the next nucleotide) impairs the elongation. (Because the primer-template complex is destabilized at a crucial position)

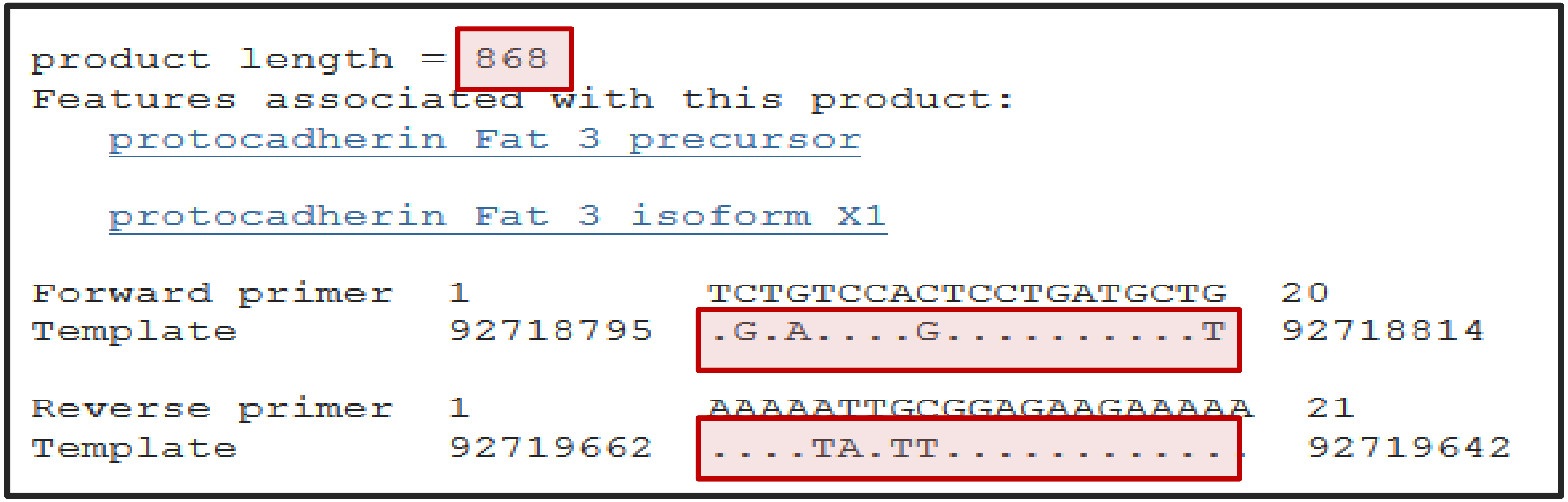


```
5′-CTCGAAGCCCAT-3′
       | | |     | | | | | |
3′-TAGCCCCGGGTATATTTAAACGGGCCCAAATTTTCTAGGTACCTAGGGTAAATTCG-5′
```

- In most cases, one or more mismatches in the middle or the 5′-end of the primer do not affect the binding of the polymerase and the DNA can be amplified.

Primer-BLAST at NCBI

What kind of **mismatches** do we have in our second BLAST-result?



```
product length = 868
Features associated with this product:
   protocadherin Fat 3 precursor

   protocadherin Fat 3 isoform X1

Forward primer  1         TCTGTCCACTCCTGATGCTG  20
Template        92718795  .G.A....G..........T  92718814

Reverse primer  1         AAAAATTGCGGAGAAGAAAAA  21
Template        92719662  ....TA.TT............  92719642
```

- In most cases, a mismatch at the 3’-end of the primer impairs the elongation.
- In most cases, one or more mismatches in the middle or the 5’-end of the primer do NOT affect PCR.

→ It is unlikely that the primers will amplify the WRONG product, therefore we don’t have to design new ones.

- **QUESTIONS?**

- Please, download **Exercises #3** from MOODLE and upload until next Monday 8:00 a.m.

GOOD LUCK!