Citrus Greening Solutions
A USDA NIFA Project

Annotation Tutorial: *Diaphorina citri* genome

Indian River State College, FL

Websites you will need

Basecamp: https://basecamp.com/

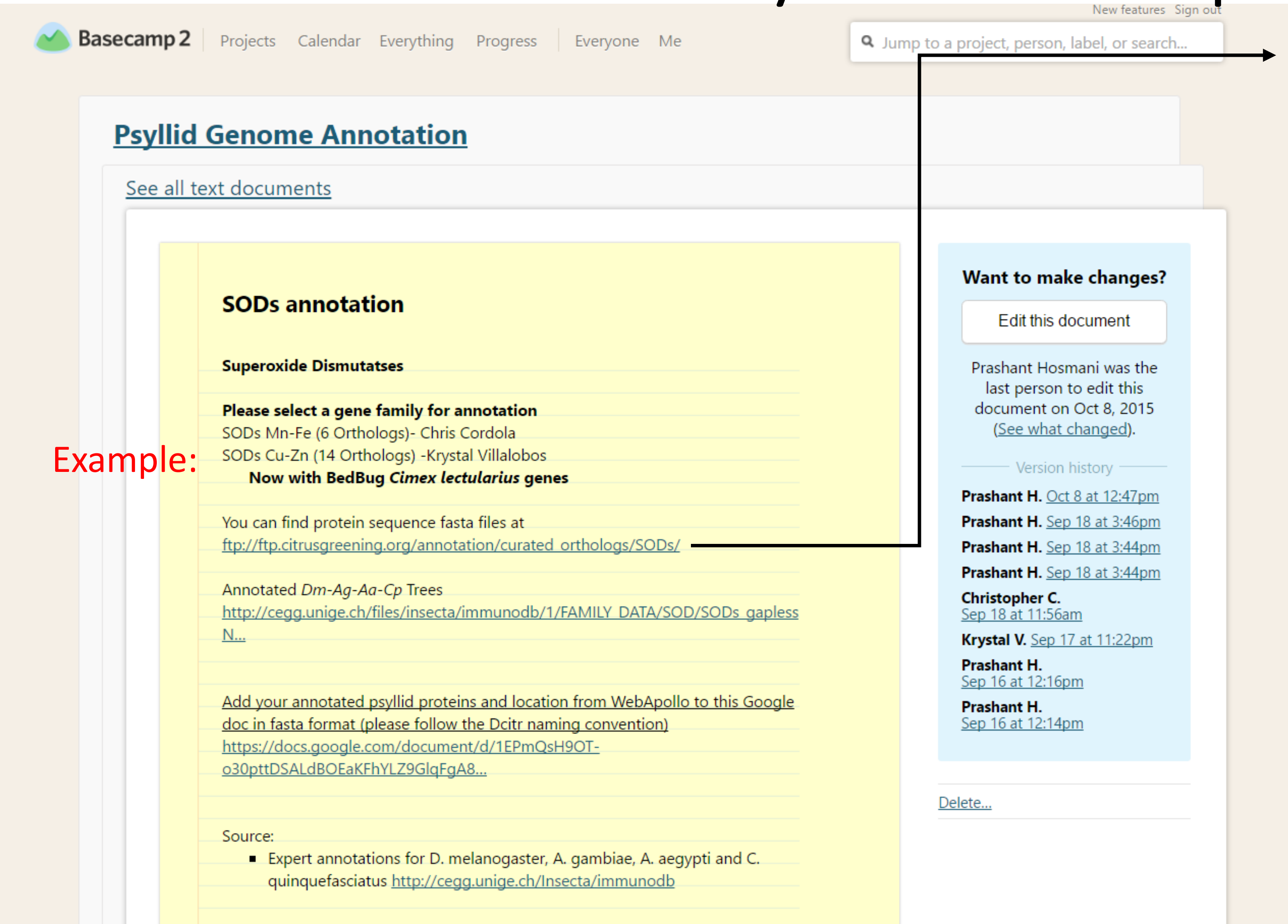
Apollo: https://apollo.nal.usda.gov/diacit/sequences

i5k Blast or hmmer: https://i5k.nal.usda.gov/webapp/blast/
https://i5k.nal.usda.gov/webapp/hmmer/

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

MCOT database: https://citrusgreening.org/tools/blast?db_id=27

Chose Gene or Gene family from Basecamp



Psyllid Genome Annotation

See all text documents

SODs annotation

Superoxide Dismutatses

Please select a gene family for annotation

Example:

SODs Mn-Fe (6 Orthologs)- Chris Cordola
SODs Cu-Zn (14 Orthologs) -Krystal Villalobos
Now with BedBug *Cimex lectularius* genes

You can find protein sequence fasta files at
ftp://ftp.citrusgreening.org/annotation/curated_orthologs/SODs/

Annotated *Dm-Ag-Aa-Cp* Trees
http://cegg.unige.ch/files/insecta/immunodb/1/FAMILY_DATA/SOD/SODs_gapless_N...

Add your annotated psyllid proteins and location from WebApollo to this Google doc in fasta format (please follow the Dcitr naming convention)
https://docs.google.com/document/d/1EPmQsH9OT-o30pttDSALdBOEaKFhYLZ9GlqFgA8...

Source:

- Expert annotations for D. melanogaster, A. gambiae, A. aegypti and C. quinquefasciatus http://cegg.unige.ch/Insecta/immunodb

Want to make changes?

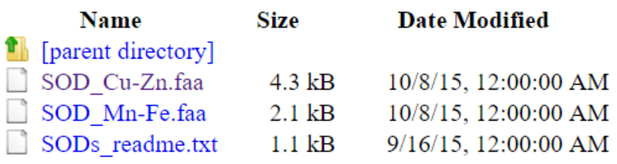
Edit this document

Prashant Hosmani was the last person to edit this document on Oct 8, 2015 (See what changed).

Version history

Prashant H. Oct 8 at 12:47pm
Prashant H. Sep 18 at 3:46pm
Prashant H. Sep 18 at 3:44pm
Prashant H. Sep 18 at 3:44pm
Christopher C. Sep 18 at 11:56am
Krystal V. Sep 17 at 11:22pm
Prashant H. Sep 16 at 12:16pm
Prashant H. Sep 16 at 12:14pm

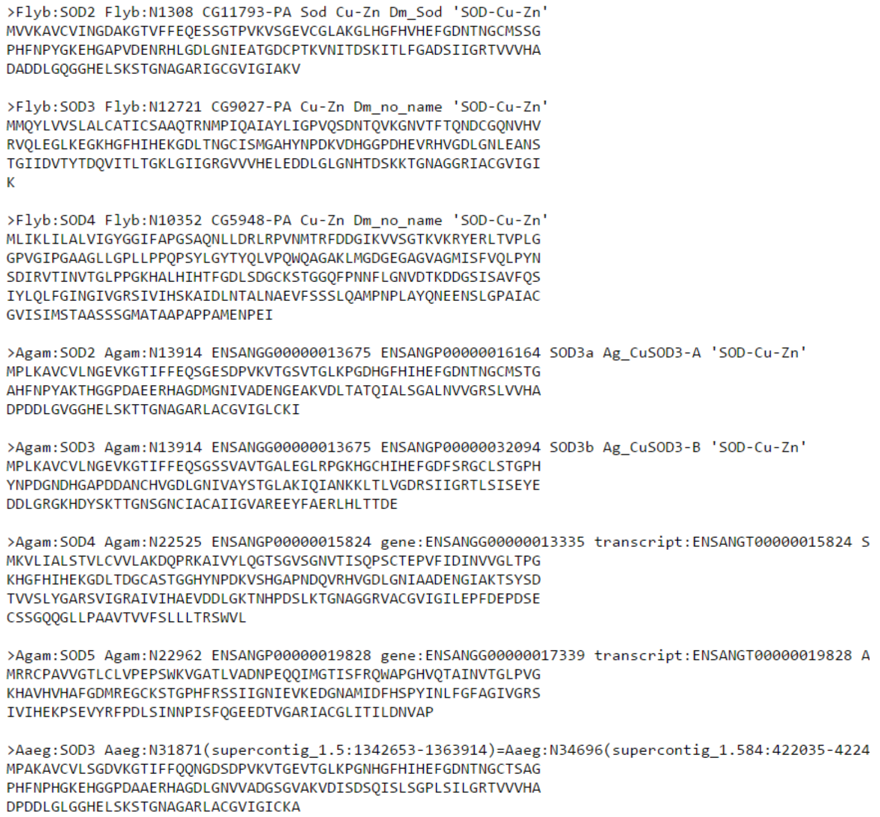
Delete...



Name	Size	Date Modified
[parent directory]		
SOD_Cu-Zn.faa	4.3 kB	10/8/15, 12:00:00 AM
SOD_Mn-Fe.faa	2.1 kB	10/8/15, 12:00:00 AM
SODs_readme.txt	1.1 kB	9/16/15, 12:00:00 AM

This page will take you to orthologous sequences & Gene family description

Orthologs



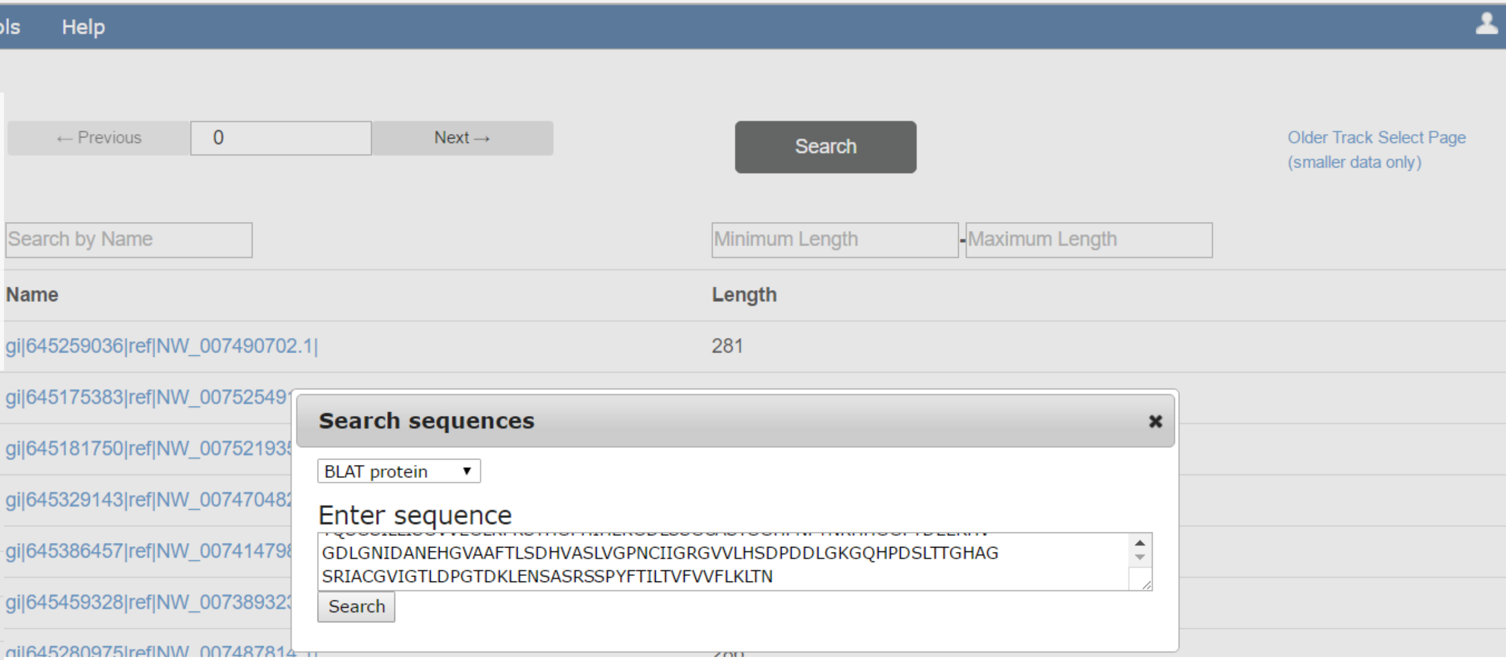
```
>Flyb:SOD2 Flyb:N1308 CG11793-PA Sod Cu-Zn Dm_Sod 'SOD-Cu-Zn'
MVVKAVCVINGDAKGTVFFEQESSGTPVKVSGEVCGLAKGLHGFHVHEFGDNTNGCMSSG
PHFNPYGKEHGAPVDENRHLGDLGNIEATGDCPTKVNITDSKITLFGADSIIGRTVVVHA
DADDLGQGGHELSKSTGNAGARIGCGVIGIAKV

>Flyb:SOD3 Flyb:N12721 CG9027-PA Cu-Zn Dm_no_name 'SOD-Cu-Zn'
MMQYLVVSLALCATICSAAQTRNMPIQAIAYLIGPVQSDNTQVKGNVTFTQNDCGQNVHV
RVQLEGLKEGKHGFHIHEKGDLTNGCISMGAHYNPDKVDHGGPDHEVRHVGDLGNLEANS
TGIIDVTYTDQVITLTGKLGIIGRGVVVHELEDDLGLGNHTDSKKTGNAGGRIACGVIGI
K

>Flyb:SOD4 Flyb:N10352 CG5948-PA Cu-Zn Dm_no_name 'SOD-Cu-Zn'
MLIKLILALVIGYGGIFAPGSAQNLLDRLRPVNMTRFDDGIKVVSGTKVKRYERLTVPLG
GPVGIPGAAGLLGPLLPPQPSYLGYTYQLVPQWQAGAKLMGDGEGAGVAGMISFVQLPYN
SDIRVTINVTGLPPGKHALHIHTFGDLSDGCKSTGGQFPNNFLGNVDTKDDGSISAVFQS
IYLQLFGINGIVGRSIVIHSKAIDLNTALNAEVFSSSLQAMPNPLAYQNEENSLGPAIAC
GVISIMSTAASSSGMATAAPAPPAMENPEI

>Agam:SOD2 Agam:N13914 ENSANGG00000013675 ENSANGP00000016164 SOD3a Ag_CuSOD3-A 'SOD-Cu-Zn'
MPLKAVCVLNGEVKGTIFFEQSGESDPVKVTGSVTGLKPGDHGFHIHEFGDNTNGCMSTG
AHFNPYAKTHGGPDAEERHAGDMGNIVADENGEAKVDLTATQIALSGALNVVGRSLVVHA
DPDDLGVGGHELSKTTGNAGARLACGVIGLCKI

>Agam:SOD3 Agam:N13914 ENSANGG00000013675 ENSANGP00000032094 SOD3b Ag_CuSOD3-B 'SOD-Cu-Zn'
MPLKAVCVLNGEVKGTIFFEQSGSSVAVTGALEGLRPGKHGCHIHEFGDFSRGCLSTGPH
YNPDGNDHGAPDDANCHVGDLGNIVAYSTGLAKIQIANKKLTLVGDRSIIGRTLSISEYE
DDLGRGKHDYSKTTGNSGNCIACAIIGVAREEYFAERLHLTTDE

>Agam:SOD4 Agam:N22525 ENSANGP00000015824 gene:ENSANGG00000013335 transcript:ENSANGT00000015824 S
MKVLIALSTVLCVVLAKDQPRKAIVYLQGTSGVSGNVTISQPSCTEPVFIDINVVGLTPG
KHGFHIHEKGDLTDGCASTGGHYNPDKVSHGAPNDQVRHVGDLGNIAADENGIAKTSYSD
TVVSLYGARSVIGRAIVIHAEVDDLGKTNHPDSLKTGNAGGRVACGVIGILEPFDEPDSE
CSSGQQGLLPAAVTVVFSLLLTRSWVL

>Agam:SOD5 Agam:N22962 ENSANGP00000019828 gene:ENSANGG00000017339 transcript:ENSANGT00000019828 A
MRRCPAVVGTLCLVPEPSWKVGATLVADNPEQQIMGTISFRQWAPGHVQTAINVTGLPVG
KHAVHVHAFGDMREGCKSTGPHFRSSIIGNIEVKEDGNAMIDFHSPYINLFGFAGIVGRS
IVIHEKPSEVYRFPDLSINNPISFQGEEDTVGARIACGLITILDNVAP

>Aaeg:SOD3 Aaeg:N31871(supercontig_1.5:1342653-1363914)=Aaeg:N34696(supercontig_1.584:422035-4224
MPAKAVCVLSGDVKGTIFFQQNGDSDPVKVTGEVTGLKPGNHGFHIHEFGDNTNGCTSAG
PHFNPHGKEHGGPDAAERHAGDLGNVVADGSGVAKVDISDSQISLSGPLSILGRTVVVHA
DPDDLGLGGHELSKSTGNAGARLACGVIGICKA
```

Test **ALL** orthologous sequences one by one in Apollo (Next slide)

“Blat” sequences in Apollo

Menu bar: Click Tools > Search Sequence

- Enter orthologous sequence (not in FASTA)
- Indicate nucleotide or protein
- Click Search



Name	Length
gi|645259036|ref|NW_007490702.1|	281
gi|645175383|ref|NW_00752549	
gi|645181750|ref|NW_00752193	
gi|645329143|ref|NW_00747048	
gi|645386457|ref|NW_00741479	
gi|645459328|ref|NW_00738932	

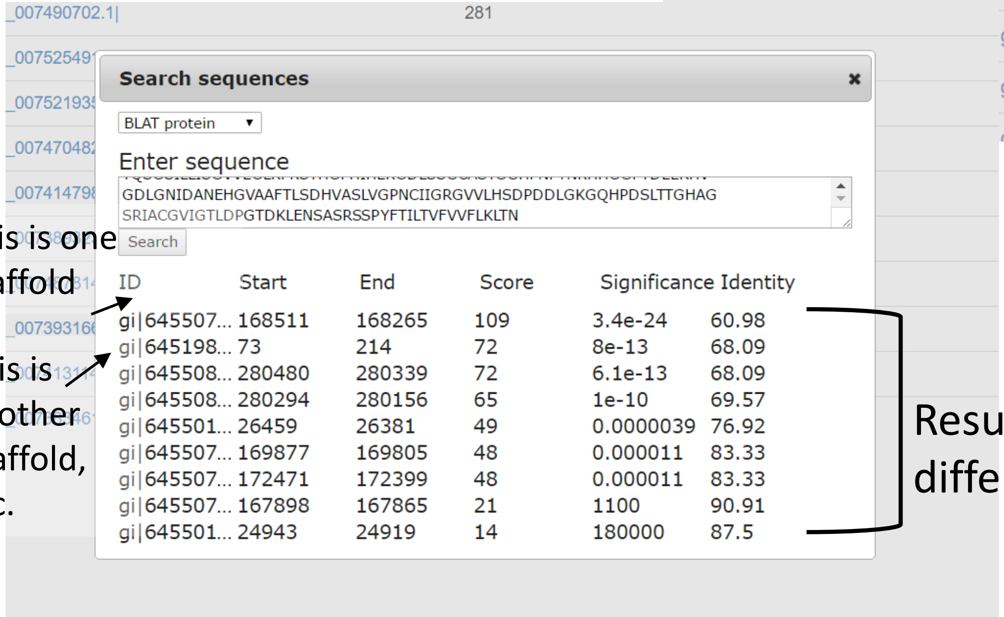
Search sequences ✖

BLAT protein ▾

Enter sequence

```
GDLGNIDANEHGVAAFTLSDHVASLVGPNCIIGRGVVLHSDPDDLGKGQHPDSLTTGHAG
SRIACGVIGTLDPGTDKLENSASRSSPYFTILTVFVVFLKLTN
```

Search


Search sequences ✖

BLAT protein ▾

Enter sequence

```
GDLGNIDANEHGVAAFTLSDHVASLVGPNCIIGRGVVLHSDPDDLGKGQHPDSLTTGHAG
SRIACGVIGTLDPGTDKLENSASRSSPYFTILTVFVVFLKLTN
```

Search

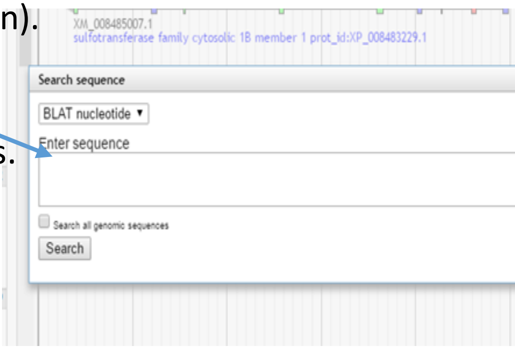
ID	Start	End	Score	Significance	Identity
gi|645507...	168511	168265	109	3.4e-24	60.98
gi|645198...	73	214	72	8e-13	68.09
gi|645508...	280480	280339	72	6.1e-13	68.09
gi|645508...	280294	280156	65	1e-10	69.57
gi|645501...	26459	26381	49	0.0000039	76.92
gi|645507...	169877	169805	48	0.000011	83.33
gi|645507...	172471	172399	48	0.000011	83.33
gi|645507...	167898	167865	21	1100	90.91
gi|645501...	24943	24919	14	180000	87.5

This is one scaffold

This is another scaffold, etc.

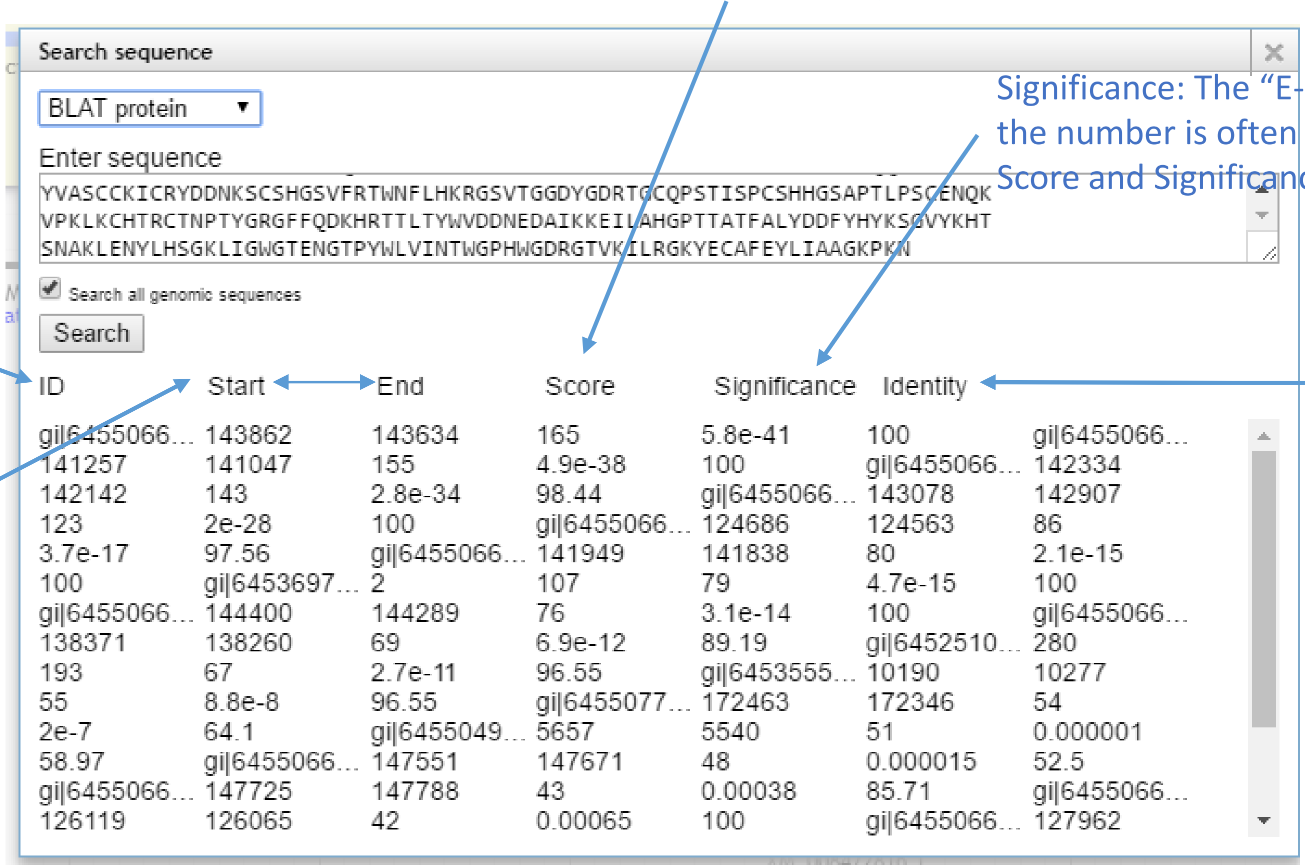
Results in different scaffolds

Change to BLAT protein if using a protein sequence (most often).
Be sure to check “search all genome sequences” if you want anything not already on your page.
Insert the letter sequence only. Do not use the > or any spaces.


XM_008485007.1
sulfotransferase family cytosolic 1B member 1 prot_id:XP_008483229.1
Search sequence
BLAT nucleotide ▾
Enter sequence
□ Search all genomic sequences
Search

- Look at all scaffolds
- Significance that is very low and Score that is high is ideal
- *Remember* Test all orthologs given from basecamp

Apollo Search Results



Enter sequence

```
YVASCCKICRYDDNKSCSHGSVFRTWNFLHKRGSVTGGDYGDRTGCQPSTISPCSHHGSAPTLPSCENQK
VPKLKCHTRCTNPTYGRGFFQDKHRTTLTYWVDDNEDAIKKEILAHGPTTATFALYDDFYHYKSGVYKHT
SNAKLENYLHSGKLIGWGTENGTPYWLVINTWGPHWGDRGTVKKILRGKYECAFEYLIAAGKPKN
```

☑ Search all genomic sequences

```
ID            Start         End           Score         Significance  Identity

gi|6455066... 143862        143634        165           5.8e-41       100           gi|6455066...
141257        141047        155           4.9e-38       100           gi|6455066... 142334
142142        143           2.8e-34       98.44         gi|6455066... 143078        142907
123           2e-28         100           gi|6455066... 124686        124563        86
3.7e-17       97.56         gi|6455066... 141949        141838        80            2.1e-15
100           gi|6453697... 2             107           79            4.7e-15       100
gi|6455066... 144400        144289        76            3.1e-14       100           gi|6455066...
138371        138260        69            6.9e-12       89.19         gi|6452510... 280
193           67            2.7e-11       96.55         gi|6453555... 10190         10277
55            8.8e-8        96.55         gi|6455077... 172463        172346        54
2e-7          64.1          gi|6455049... 5657          5540          51            0.000001
58.97         gi|6455066... 147551        147671        48            0.000015      52.5
gi|6455066... 147725        147788        43            0.00038       85.71         gi|6455066...
126119        126065        42            0.00065       100           gi|6455066... 127962
```

Score: A number that the BLAT results "scores" the hit. The higher is often better, but not always. There is a balance with score and Significance.

Significance: The "E-value" given by the BLAT results. The smaller the number is often better, but not always. There is a balance with Score and Significance.

ID: The scaffold location

Start or End : The number where the sequence starts or ends

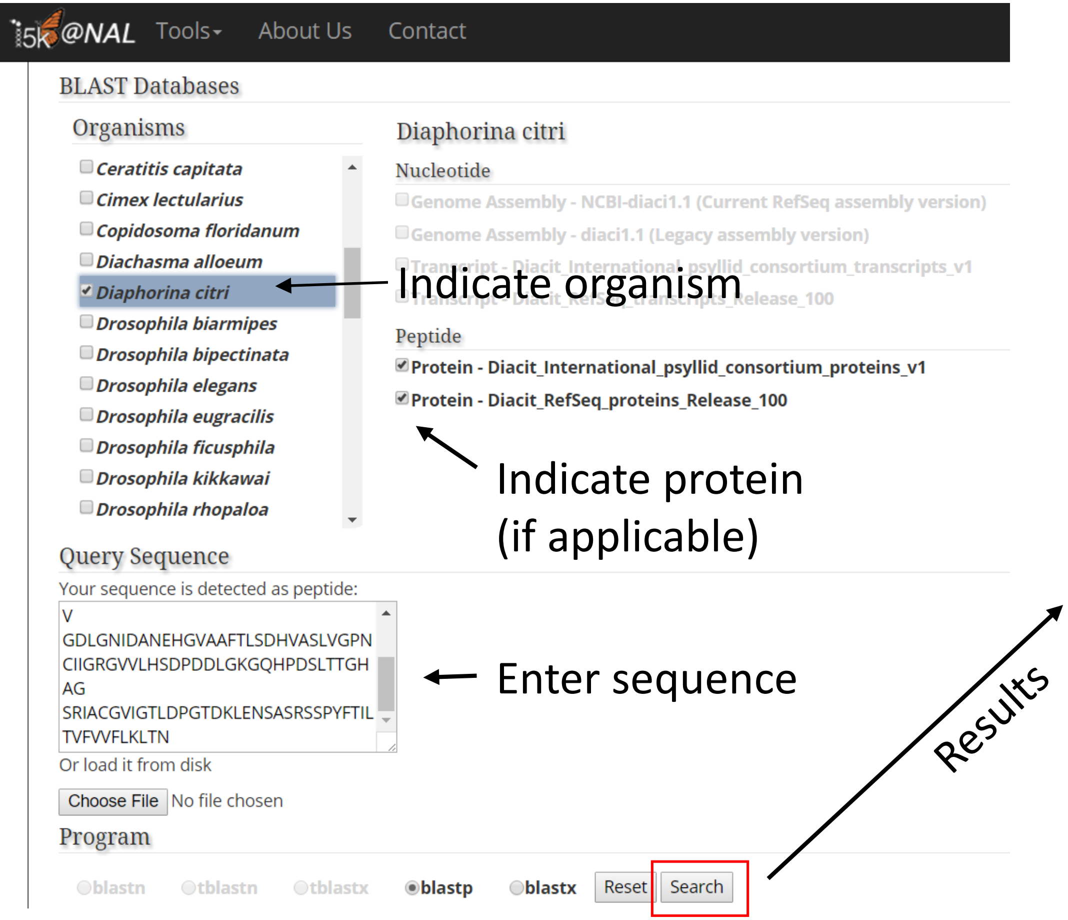
Identity: How well the sequence entered matches the location of the scaffold.

This example shows the columns out of place, this may happen, The ID section starts with gi.

Choose a scaffold. By clicking on it. You can then X out of the dialog box.

If Apollo says “No matches found”

Use i5k Blast



BLAST Databases

Organisms

- □ *Ceratitis capitata*
- □ *Cimex lectularius*
- □ *Copidosoma floridanum*
- □ *Diachasma alloeum*
- ☑ *Diaphorina citri*
- □ *Drosophila biarmipes*
- □ *Drosophila bipectinata*
- □ *Drosophila elegans*
- □ *Drosophila eugracilis*
- □ *Drosophila ficusphila*
- □ *Drosophila kikkawai*
- □ *Drosophila rhopaloa*

Diaphorina citri

Nucleotide

- □ Genome Assembly - NCBI-diaci1.1 (Current RefSeq assembly version)
- □ Genome Assembly - diaci1.1 (Legacy assembly version)
- □ Transcript - Diacit_International_psyllid_consortium_transcripts_v1
- □ Transcript - Diacit_RefSeq_transcripts_Release_100

Peptide

- ☑ Protein - Diacit_International_psyllid_consortium_proteins_v1
- ☑ Protein - Diacit_RefSeq_proteins_Release_100

Query Sequence

Your sequence is detected as peptide:

```
V
GDLGNIDANEHGVAAFTLSDHVASLVGPN
CIIGRGVVLHSDPDDLGKGQHPDSLTTGH
AG
SRIACGVIGTLDPGTDKLENSASRSSPYFTIL
TVFVVFLKLTN
```

Or load it from disk

Choose File No file chosen

Program

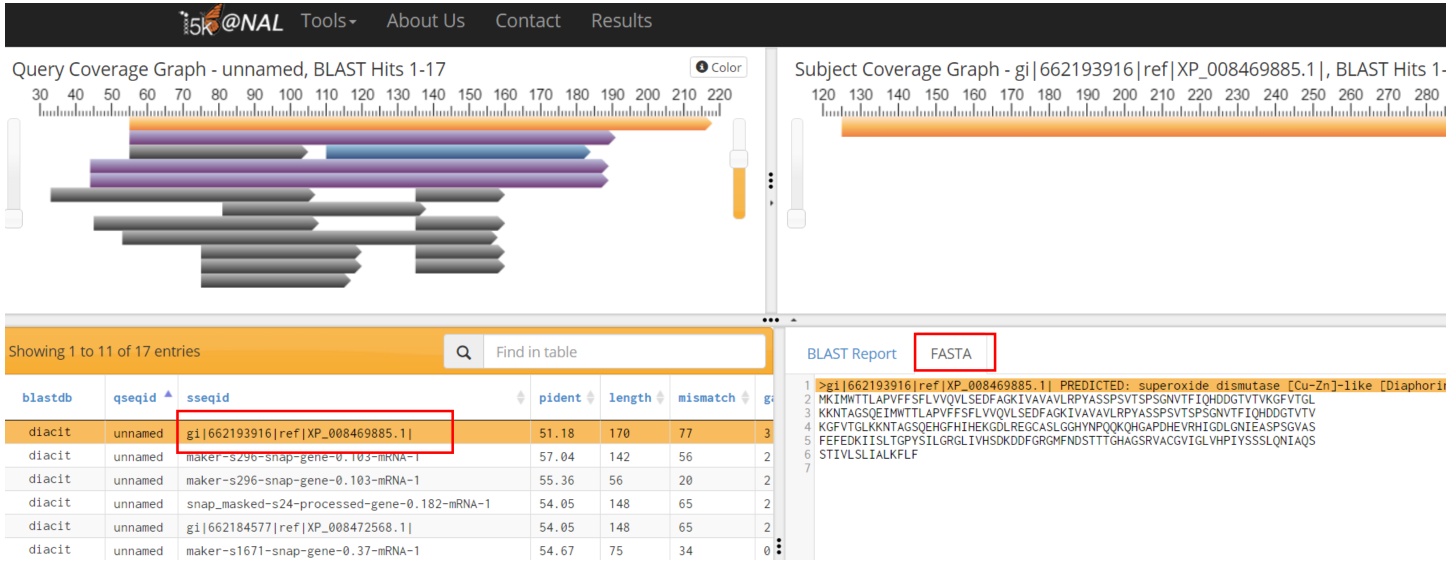
○ blastn ○ tblastn ○ tblastx ● blastp ○ blastx Reset Search

Indicate organism

Indicate protein (if applicable)

Enter sequence

Results



Showing 1 to 11 of 17 entries

blastdb	qseqid	sseqid	pident	length	mismatch	ga
diacit	unnamed	gi|662193916|ref|XP_008469885.1|	51.18	170	77	3
diacit	unnamed	maker-s296-snap-gene-0.103-mRNA-1	57.04	142	56	2
diacit	unnamed	maker-s296-snap-gene-0.103-mRNA-1	55.36	56	20	2
diacit	unnamed	snap_masked-s24-processed-gene-0.182-mRNA-1	54.05	148	65	2
diacit	unnamed	gi|662184577|ref|XP_008472568.1|	54.05	148	65	2
diacit	unnamed	maker-s1671-snap-gene-0.37-mRNA-1	54.67	75	34	0

```
>gi|662193916|ref|XP_008469885.1| PREDICTED: superoxide dismutase [Cu-Zn]-like [Diaphori
MKIMWTTLAPVFFSFLVVQVLSEDFAGKIVAVAVLRPYASSPSVTSPSGNVTFIQHDDGTVTVKGFVTGL
KKNTAGSQEIMWTTLAPVFFSFLVVQVLSEDFAGKIVAVAVLRPYASSPSVTSPSGNVTFIQHDDGTVTV
KGFVTGLKKNTAGSQEHGFHIHEKGDLREGCASLGGHYNPQQKQHGAPDHEVRHIGDLGNIEASPSGVAS
FEFEDKIISLTGPYSILGRGLIVHSDKDDFGRGMFNDSTTTGHAGSRVACGVIGLVHPIYSSSLQNIAQS
STIVLSLIALKFLF
```

- Chose best result(s) (Low e-value, best coverage, etc.)
- Copy sequence and Blat in Apollo.

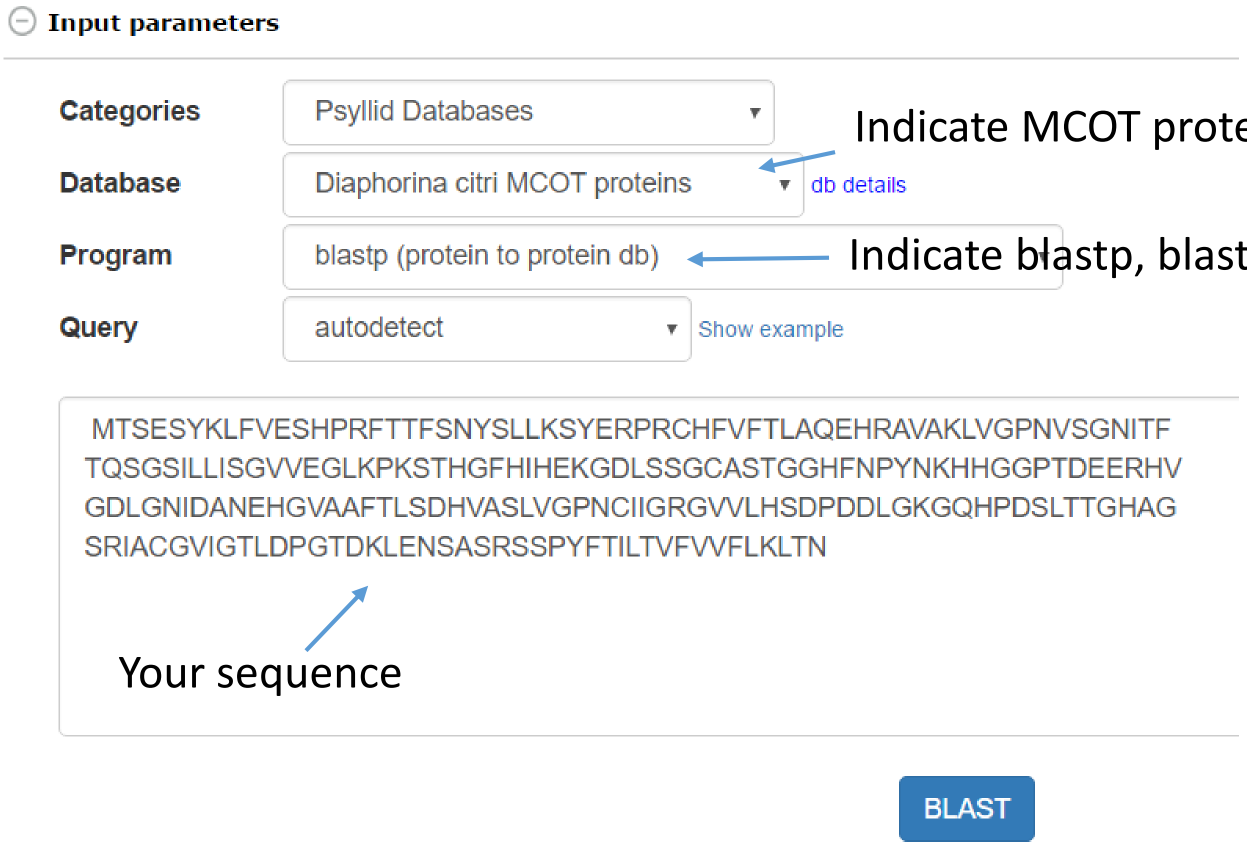
Apollo Blat vs i5k Blast

- Apollo Blatt is more “sensitive” than i5k Blast.
- i5k Blast will broaden the search.

If Apollo says “No matches found”

Search your sequence on *D. citri* MCOT database

https://citrusgreening.org/tools/blast?db_id=27 BLAST



Input parameters

Categories: Psyllid Databases

Database: Diaphorina citri MCOT proteins

Program: blastp (protein to protein db)

Query: autodetect

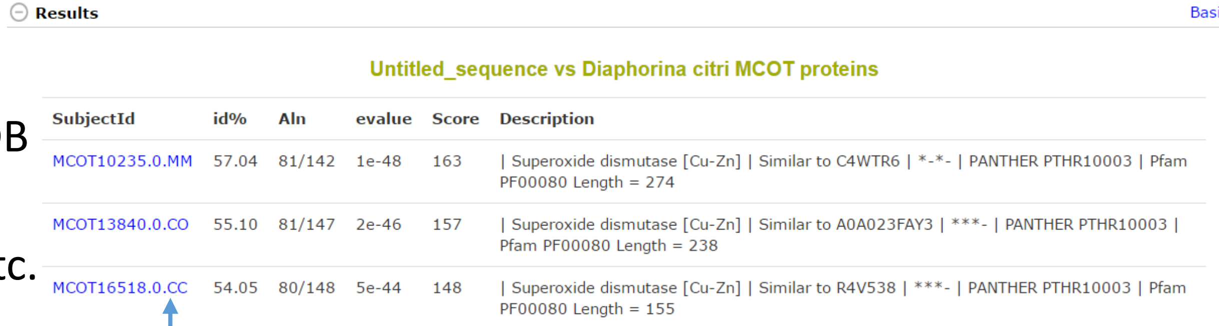
MTSESYKLFVESHPRFTTFSNYSLLKSYERPRCHFVFTLAQEHRAVAKLVGPNVSGNITF
TQSGSILLISGVVEGLKPKSTHGFHIHEKGDLSSGCASTGGHFNPYNKHHGGPTDEERHV
GDLGNIDANEHGVAAFTLSDHVASLVGPNCIIGRGVVLHSDPDDLGKGQHPDSLTTGHAG
SRIACGVIGTLDPGTDKLENSASRSSPYFTILTVFVVFLKLTN

Indicate MCOT protein DB

Indicate blastp, blastn, etc.

Your sequence

Results



Results

Untitled_sequence vs Diaphorina citri MCOT proteins

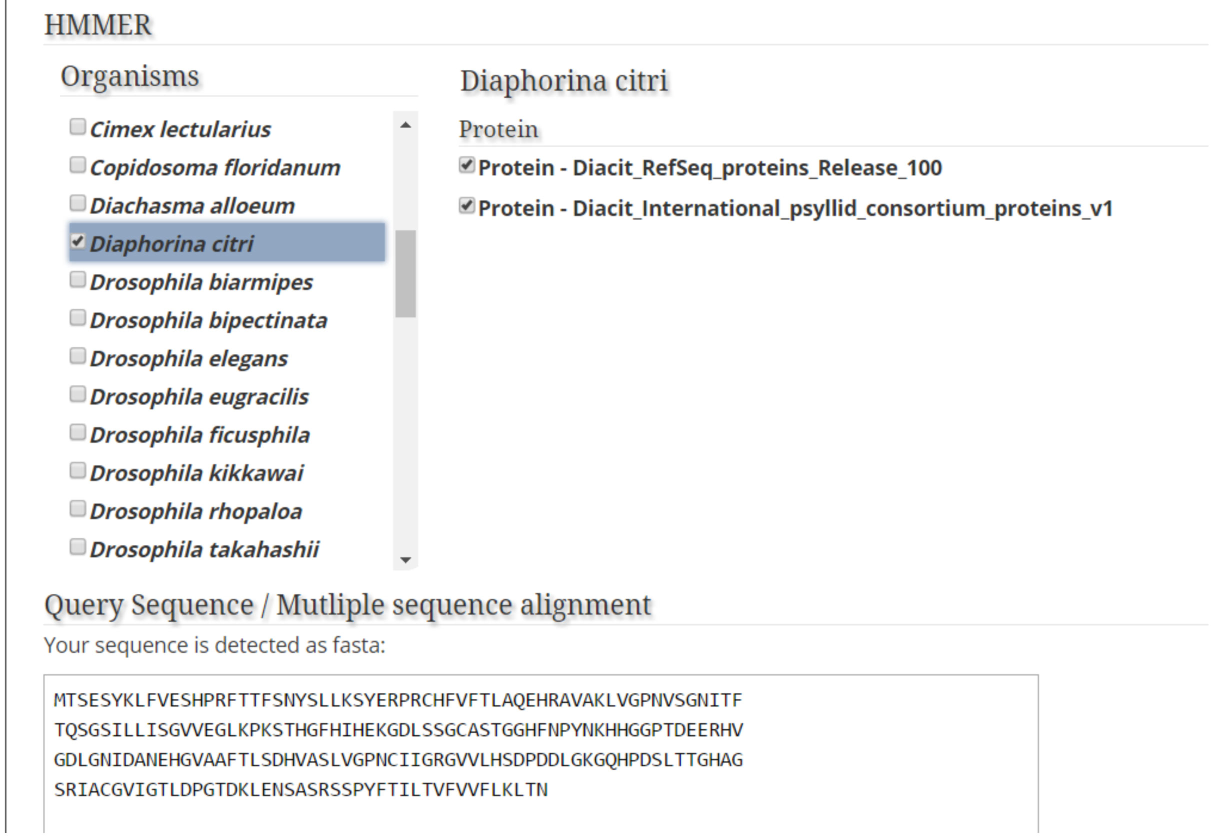
SubjectId	id%	Aln	evalue	Score	Description
MCOT10235.0.MM	57.04	81/142	1e-48	163	| Superoxide dismutase [Cu-Zn] | Similar to C4WTR6 | *-*- | PANTHER PTHR10003 | Pfam PF00080 Length = 274
MCOT13840.0.CO	55.10	81/147	2e-46	157	| Superoxide dismutase [Cu-Zn] | Similar to A0A023FAY3 | ***- | PANTHER PTHR10003 | Pfam PF00080 Length = 238
MCOT16518.0.CC	54.05	80/148	5e-44	148	| Superoxide dismutase [Cu-Zn] | Similar to R4V538 | ***- | PANTHER PTHR10003 | Pfam PF00080 Length = 155

Click for sequence, and blat on Apollo

If Apollo says “No matches found”

Search your sequence in i5k blast hmmer (hidden Markov model) created from a multiple sequence alignment of the ortholog proteins

https://i5k.nal.usda.gov/webapp/hmmer/



HMMER

Organisms

- □ *Cimex lectularius*
- □ *Copidosoma floridanum*
- □ *Diachasma alloeum*
- ☑ *Diaphorina citri*
- □ *Drosophila biarmipes*
- □ *Drosophila bipectinata*
- □ *Drosophila elegans*
- □ *Drosophila eugracilis*
- □ *Drosophila ficusphila*
- □ *Drosophila kikkawai*
- □ *Drosophila rhopaloa*
- □ *Drosophila takahashii*

Diaphorina citri

Protein

- ☑ Protein - Diacit_RefSeq_proteins_Release_100
- ☑ Protein - Diacit_International_psyllid_consortium_proteins_v1

Query Sequence / Mutliple sequence alignment

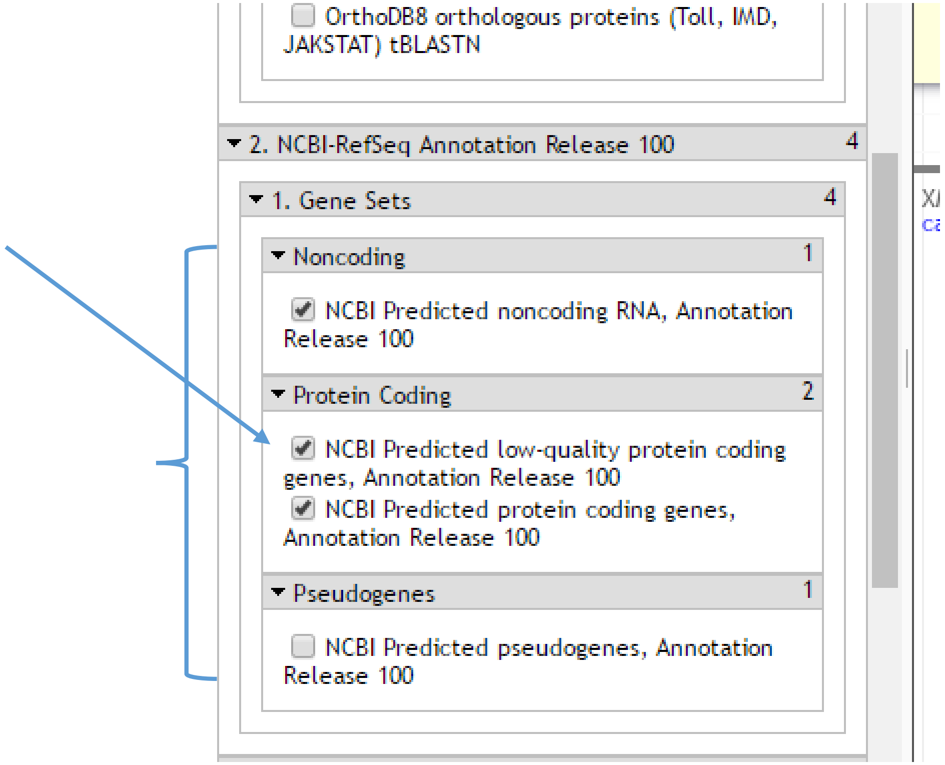
Your sequence is detected as fasta:

```
MTSESYKLFVESHPRFTTFSNYSLLKSYERPRCHFVFTLAQEHRAVAKLVGPNVSGNITF
TQSGSILLISGVVEGLKPKSTHGFHIHEKGDLSSGCASTGGHFNPYNKHHGGPTDEERHV
GDLGNIDANEHGVAAFTLSDHVASLVGPNCIIGRGVVLHSDPDDLGKGQHPDSLTTGHAG
SRIACGVIGTLDPGTDKLENSASRSSPYFTILTVFVVFLKLTN
```

Tutorial
https://i5k.nal.usda.gov/webapp/hmmer/manual/

On Apollo, once you have selected a scaffold.

Be sure that the items you want to view are chosen on the left side of the screen. For the psyllid, we start to base our models off of NCBI predictions. Be sure the following gene sets are chosen.



□ OrthoDB8 orthologous proteins (Toll, IMD, JAKSTAT) tBLASTN

2. NCBI-RefSeq Annotation Release 100 — 4

1. Gene Sets — 4

Noncoding — 1

☑ NCBI Predicted noncoding RNA, Annotation Release 100

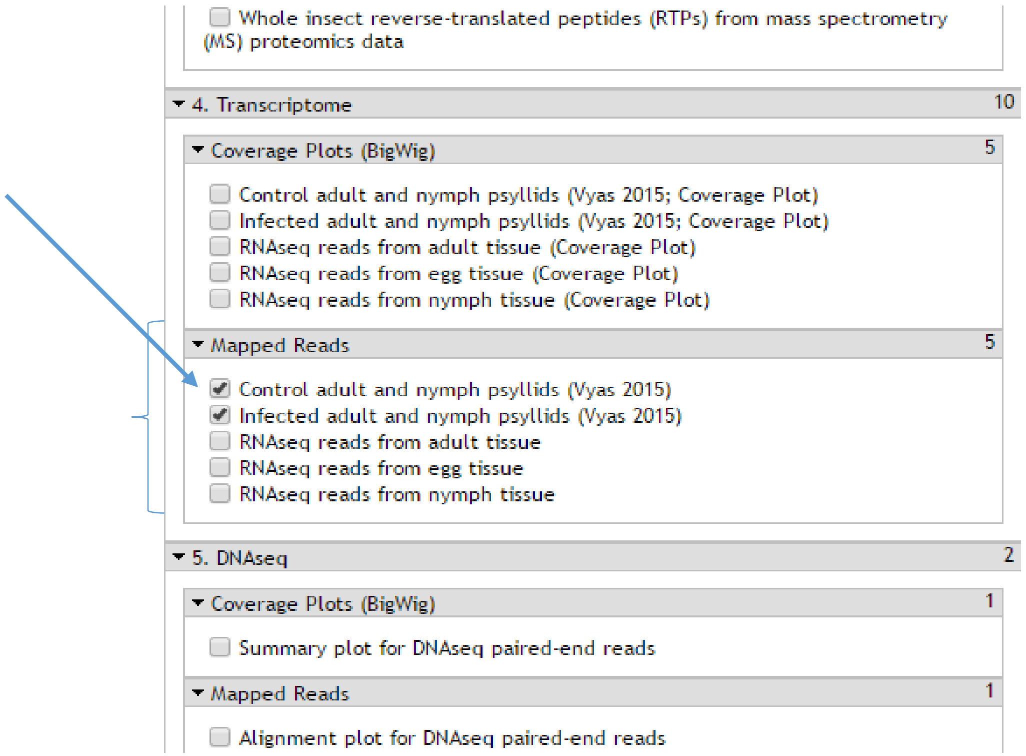
Protein Coding — 2

☑ NCBI Predicted low-quality protein coding genes, Annotation Release 100

☑ NCBI Predicted protein coding genes, Annotation Release 100

Pseudogenes — 1

□ NCBI Predicted pseudogenes, Annotation Release 100



□ Whole insect reverse-translated peptides (RTPs) from mass spectrometry (MS) proteomics data

4. Transcriptome — 10

Coverage Plots (BigWig) — 5

□ Control adult and nymph psyllids (Vyas 2015; Coverage Plot)

□ Infected adult and nymph psyllids (Vyas 2015; Coverage Plot)

□ RNAseq reads from adult tissue (Coverage Plot)

□ RNAseq reads from egg tissue (Coverage Plot)

□ RNAseq reads from nymph tissue (Coverage Plot)

Mapped Reads — 5

☑ Control adult and nymph psyllids (Vyas 2015)

☑ Infected adult and nymph psyllids (Vyas 2015)

□ RNAseq reads from adult tissue

□ RNAseq reads from egg tissue

□ RNAseq reads from nymph tissue

5. DNAseq — 2

Coverage Plots (BigWig) — 1

□ Summary plot for DNAseq paired-end reads

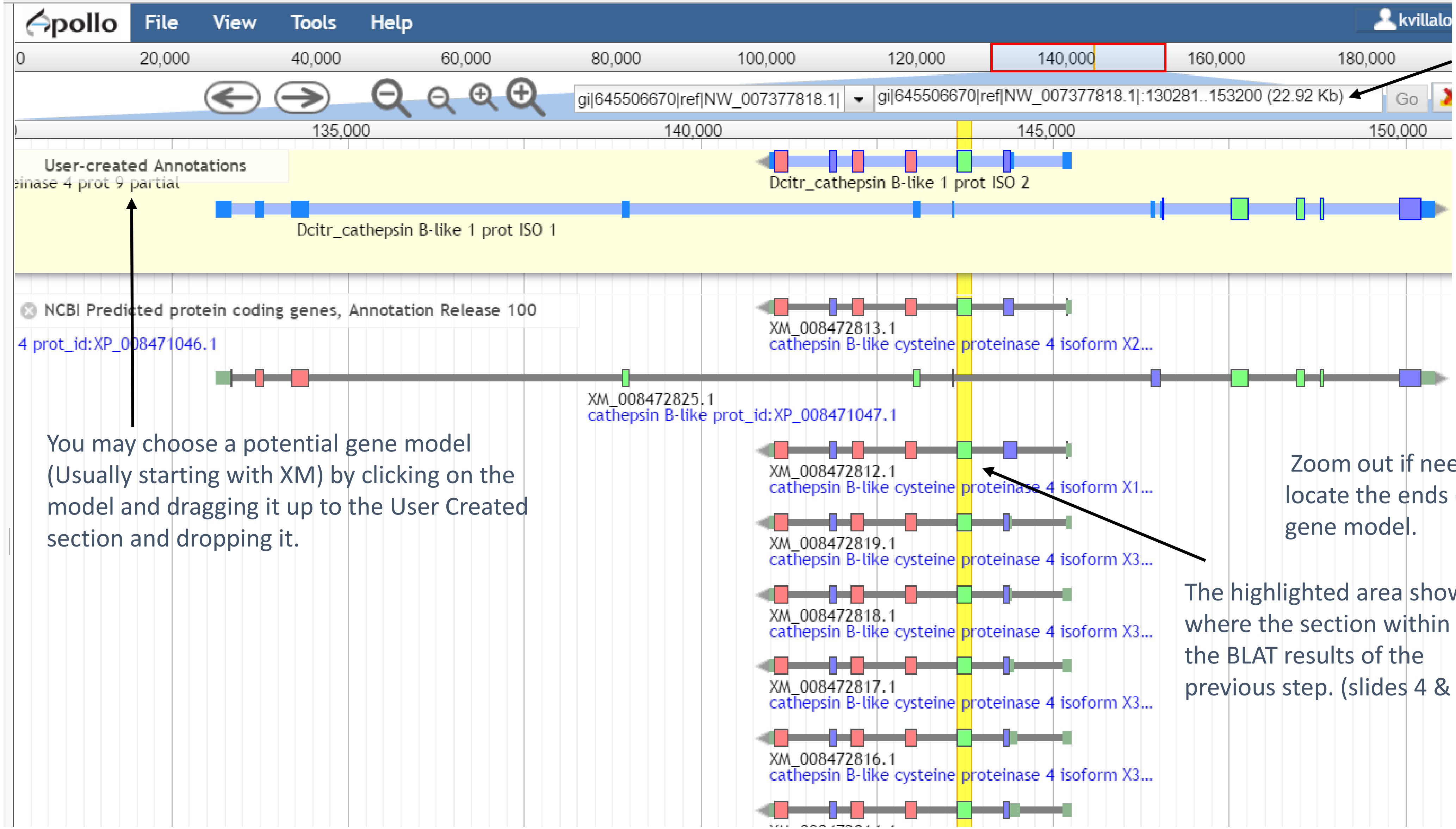
Mapped Reads — 1

□ Alignment plot for DNAseq paired-end reads

We will base our models off of RNAseq data, so be sure to choose some tracks under **Mapped Reads** as well.

Note: Be sure to check all RNAseq data, but it may overload the system if they are all turned on at once.

Predicted Gene Models


Apollo
File
View
Tools
Help
kvillalo
0
20,000
40,000
60,000
80,000
100,000
120,000
140,000
160,000
180,000
gi|645506670|ref|NW_007377818.1|
Go
135,000
140,000
145,000
150,000
User-created Annotations
einase 4 prot 9 partial
Dcitr_cathepsin B-like 1 prot ISO 2
Dcitr_cathepsin B-like 1 prot ISO 1
NCBI Predicted protein coding genes, Annotation Release 100
4 prot_id:XP_008471046.1
XM_008472813.1
cathepsin B-like cysteine proteinase 4 isoform X2...
XM_008472825.1
cathepsin B-like prot_id:XP_008471047.1
XM_008472812.1
cathepsin B-like cysteine proteinase 4 isoform X1...
XM_008472819.1
cathepsin B-like cysteine proteinase 4 isoform X3...
XM_008472818.1
cathepsin B-like cysteine proteinase 4 isoform X3...
XM_008472817.1
cathepsin B-like cysteine proteinase 4 isoform X3...
XM_008472816.1
cathepsin B-like cysteine proteinase 4 isoform X3...

gi|645506670|ref|NW_007377818.1|:130281..153200 (22.92 Kb)

Useful tip: You can use this as an "URL" in Apollo. As you need to look at all scaffolds (BLAT results), use this URL to organize all your potential gene models and to get back to them.

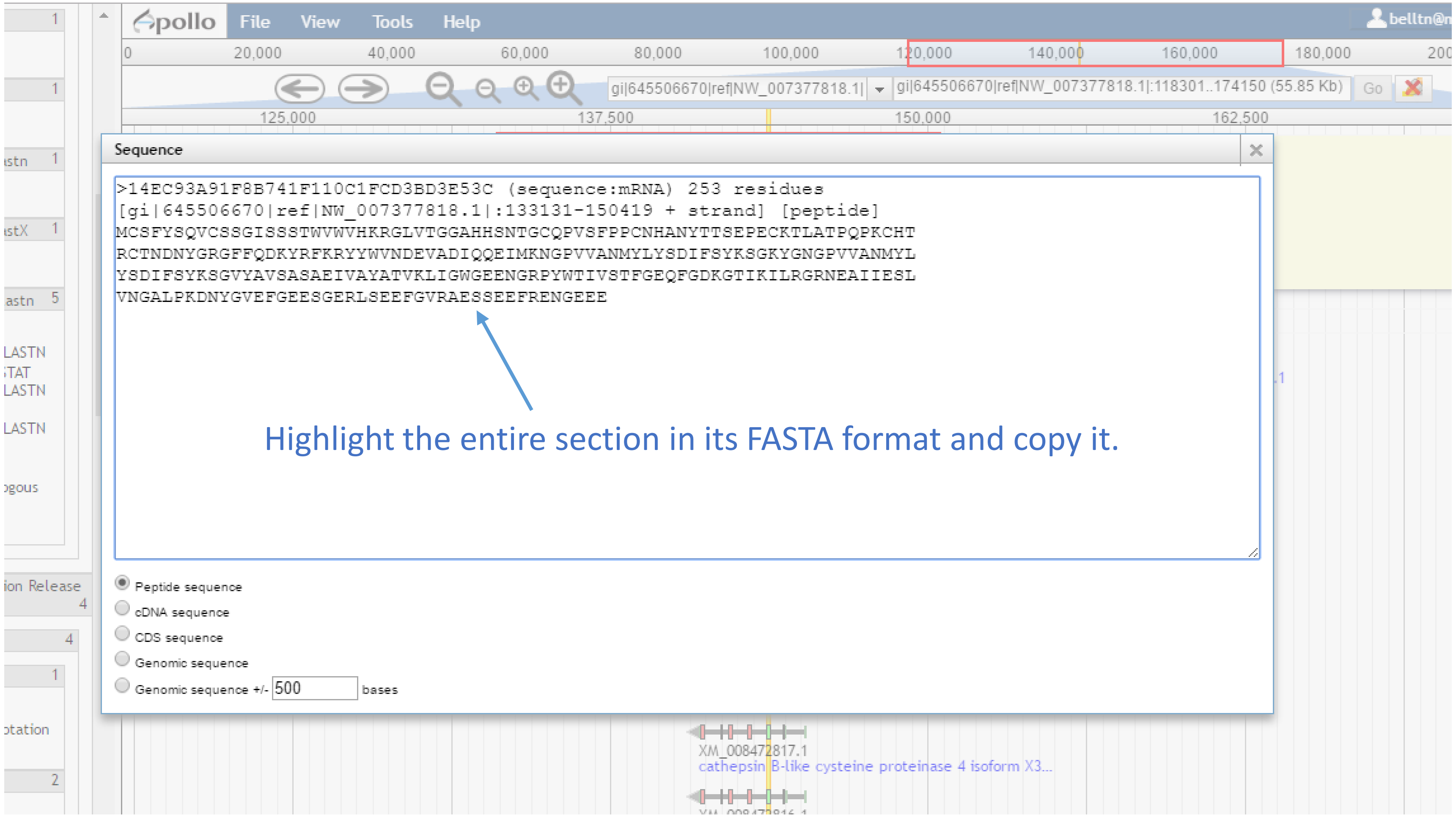
You may choose a potential gene model (Usually starting with XM) by clicking on the model and dragging it up to the User Created section and dropping it.

Zoom out if need be and locate the ends of a potential gene model.

The highlighted area shows where the section within the BLAT results of the previous step. (slides 4 & 5)

Check Gene Models

Right click on the model in the User Created section and choose “Get sequence”

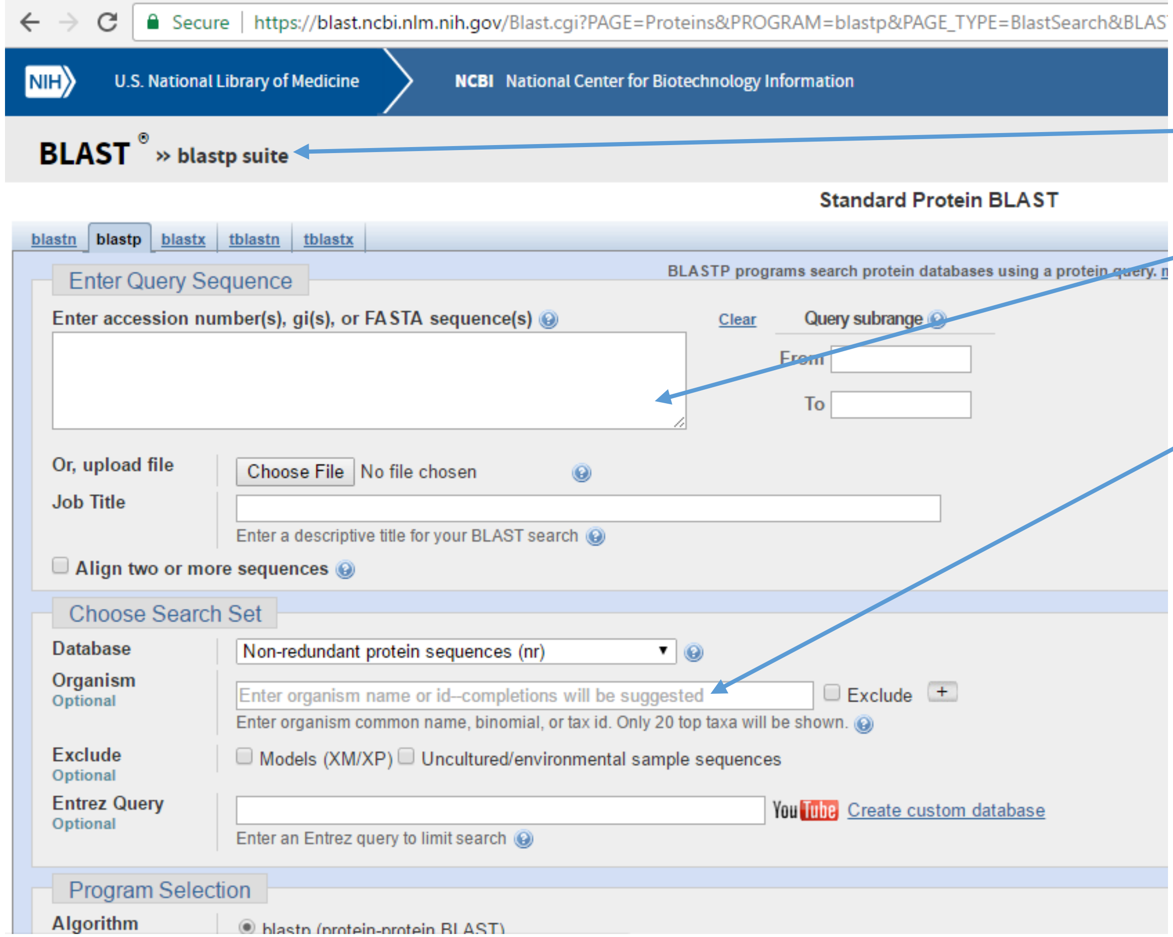


```
>14EC93A91F8B741F110C1FCD3BD3E53C (sequence:mRNA) 253 residues
[gi|645506670|ref|NW_007377818.1|:133131-150419 + strand] [peptide]
MCSFYSQVCSSGISSSTWVWVHKRGLVTGGAHHSNTGCQPVSFPPCNHANYTTSEPECKTLATPQPKCHT
RCTNDNYGRGFFQDKYRFKRYYWVNDEVADIQQEIMKNGPVVANMYLYSDIFSYKSGKYGNGPVVANMYL
YSDIFSYKSGVYAVSASAEIVAYATVKLIGWGEENGRPYWTIVSTFGEQFGDKGTIKILRGRNEAIIESL
VNGALPKDNYGVEFGEESGERLSEEFGVRAESSEEFRENGEEE
```

Highlight the entire section in its FASTA format and copy it.

● Peptide sequence
○ cDNA sequence
○ CDS sequence
○ Genomic sequence
○ Genomic sequence +/- 500 bases

Protein BLAST



Secure | https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&PAGE_TYPE=BlastSearch&BLAS

NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information

BLAST® » blastp suite

Standard Protein BLAST

blastn
blastp
blastx
tblastn
tblastx

BLASTP programs search protein databases using a protein query.

Enter Query Sequence

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

From

To

Or, upload file Choose File No file chosen

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

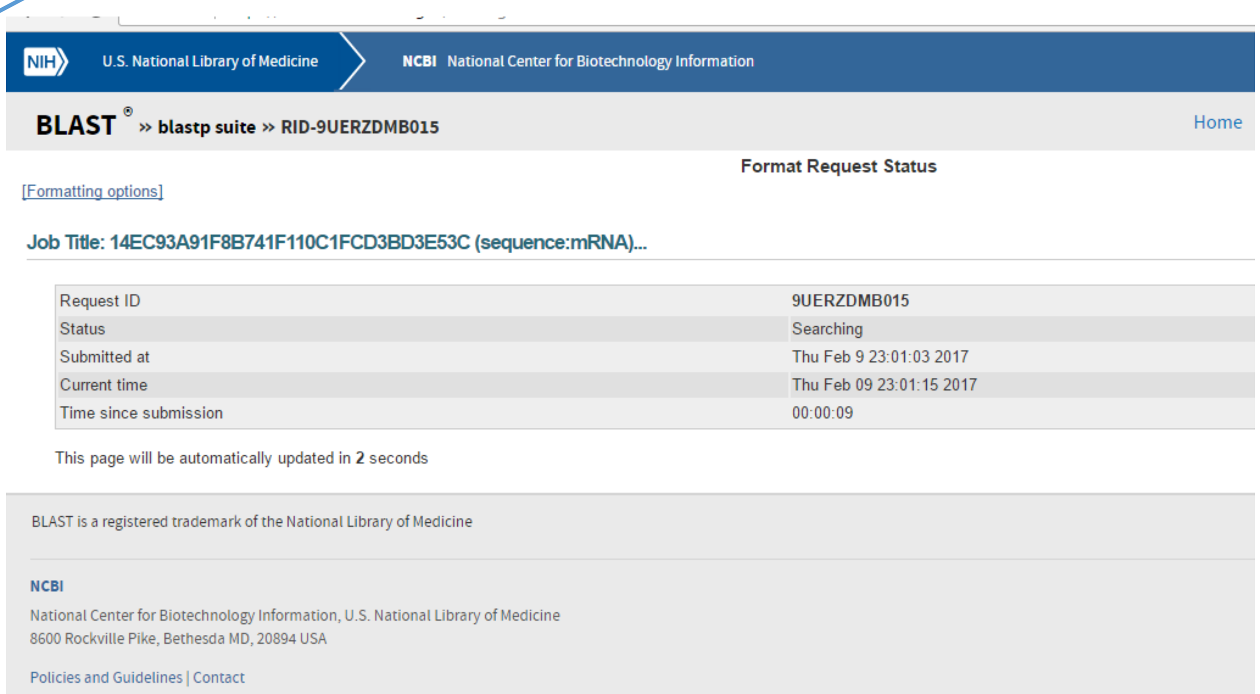
Program Selection

Algorithm ● blastp (protein-protein BLAST)

Use blastp for protein BLAST

Enter the FASTA sequence into the appropriate box, choose any selection / parameters. And press BLAST at the bottom.

Indicate “Insecta”. This will limited your search to insects only.



NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information

BLAST® » blastp suite » RID-9UERZDMB015 Home

Format Request Status

[Formatting options]

Job Title: 14EC93A91F8B741F110C1FCD3BD3E53C (sequence:mRNA)...

Request ID	9UERZDMB015
Status	Searching
Submitted at	Thu Feb 9 23:01:03 2017
Current time	Thu Feb 09 23:01:15 2017
Time since submission	00:00:09

This page will be automatically updated in **2** seconds

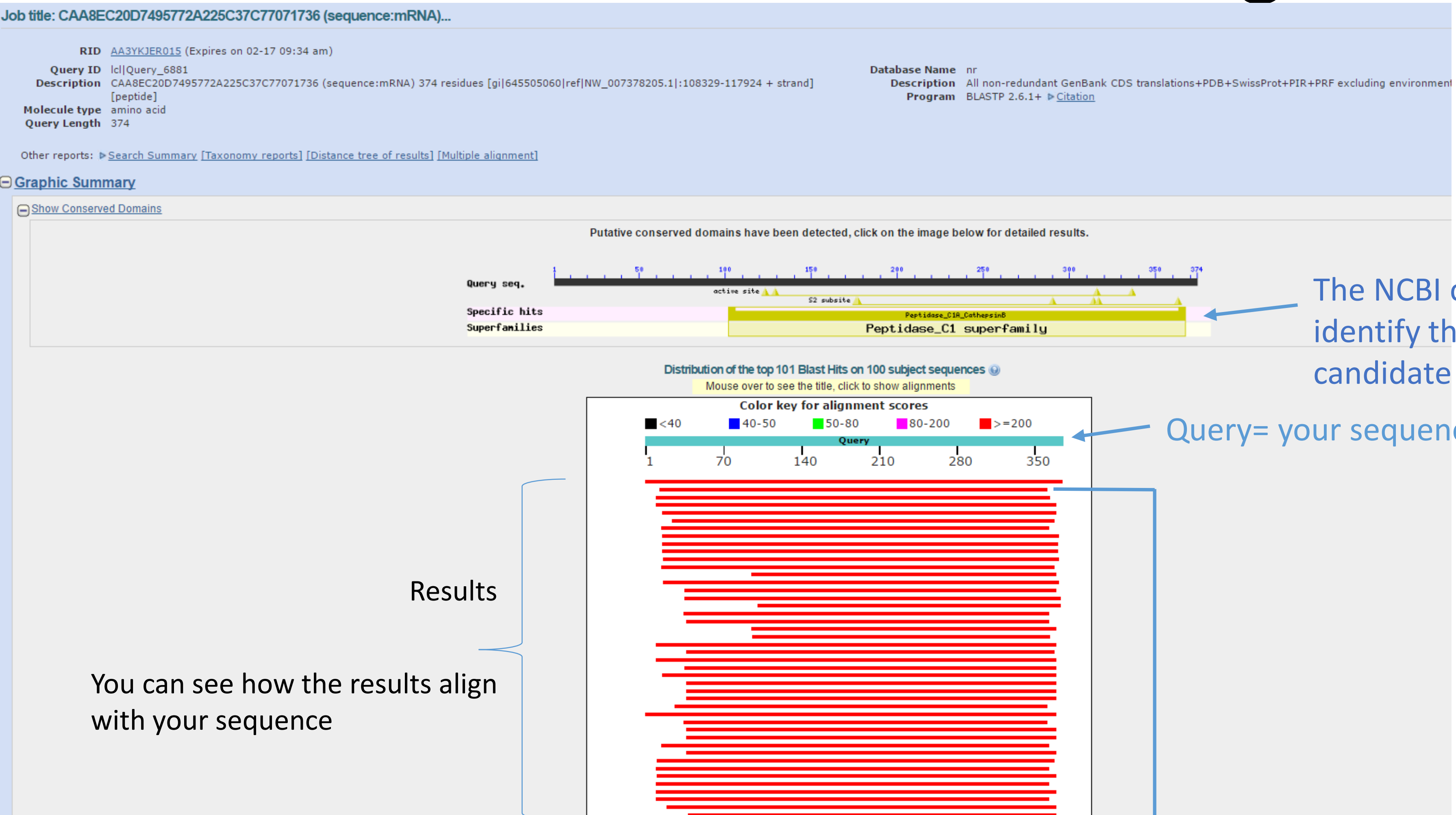
BLAST is a registered trademark of the National Library of Medicine

NCBI

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

Policies and Guidelines | Contact

The NCBI Results Page



Job title: CAA8EC20D7495772A225C37C77071736 (sequence:mRNA)...

RID AA3YKJER015 (Expires on 02-17 09:34 am)

Query ID lcl|Query_6881
Description CAA8EC20D7495772A225C37C77071736 (sequence:mRNA) 374 residues [gi|645505060|ref|NW_007378205.1|:108329-117924 + strand] [peptide]
Molecule type amino acid
Query Length 374

Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environment
Program BLASTP 2.6.1+ ▶Citation

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

⊖ Graphic Summary

⊖ Show Conserved Domains

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

Distribution of the top 101 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments

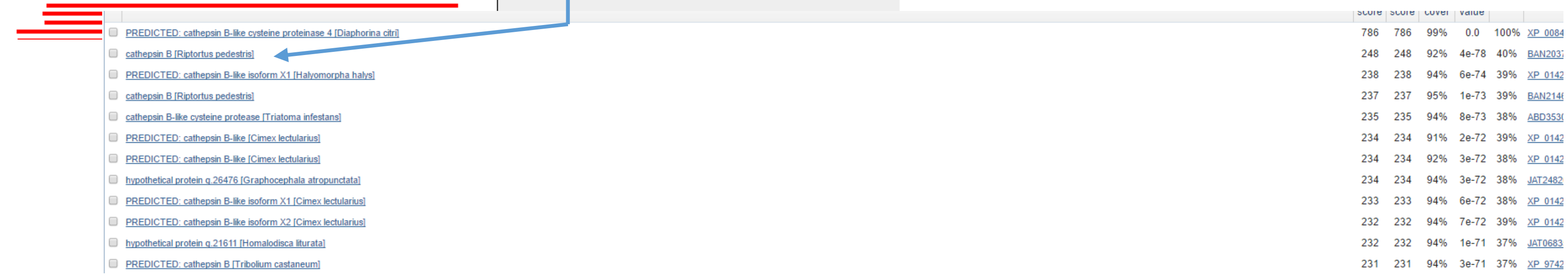
The top portion gives information on the searched sequence

The NCBI conserved domains database is used to identify the conserved domains in the orthologs and candidate genes.

Query= your sequence

Results

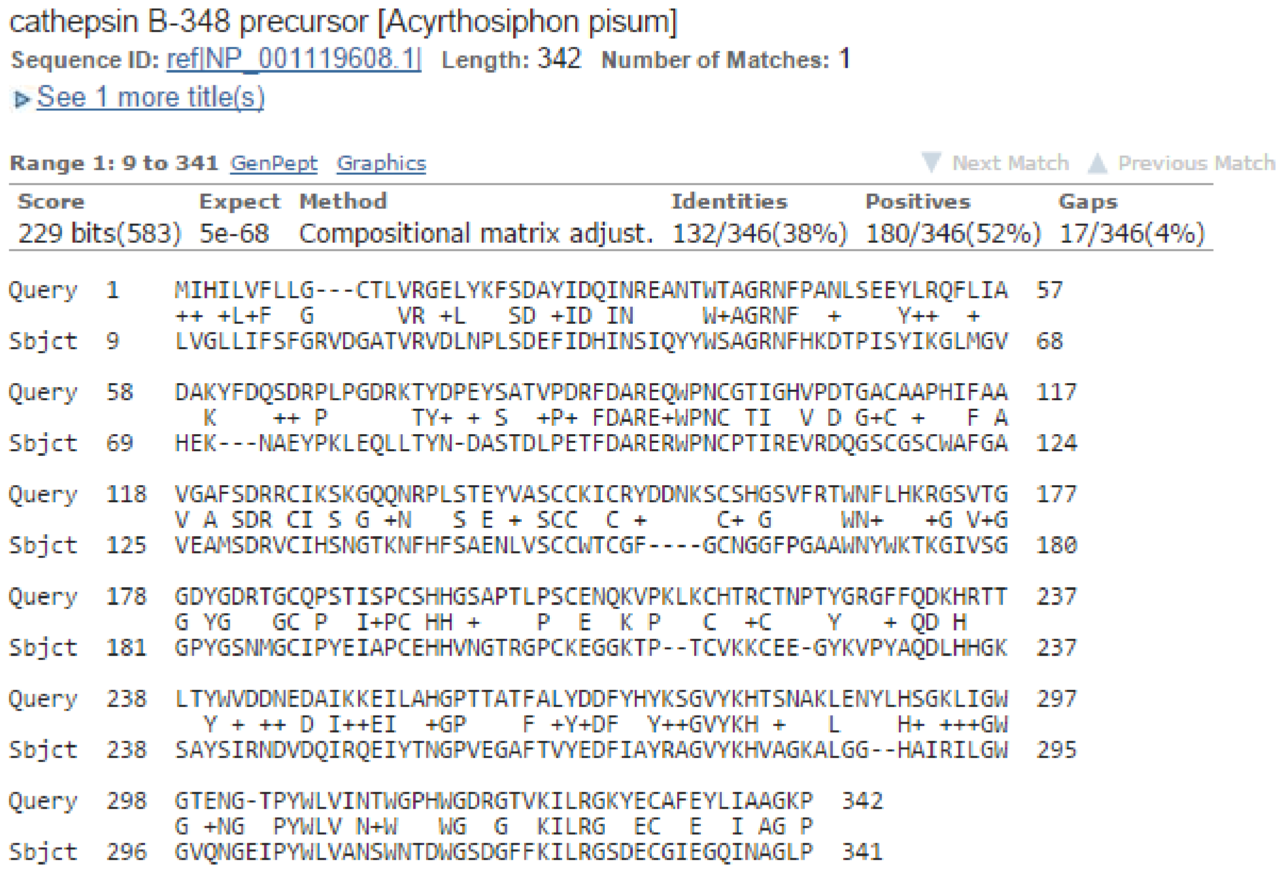
You can see how the results align with your sequence



	score	score	cover	value		
□ PREDICTED: cathepsin B-like cysteine proteinase 4 [Diaphorina citri]	786	786	99%	0.0	100%	XP_0084
□ cathepsin B [Riptortus pedestris]	248	248	92%	4e-78	40%	BAN203
□ PREDICTED: cathepsin B-like isoform X1 [Halyomorpha halys]	238	238	94%	6e-74	39%	XP_0142
□ cathepsin B [Riptortus pedestris]	237	237	95%	1e-73	39%	BAN214
□ cathepsin B-like cysteine protease [Triatoma infestans]	235	235	94%	8e-73	38%	ABD353
□ PREDICTED: cathepsin B-like [Cimex lectularius]	234	234	91%	2e-72	39%	XP_0142
□ PREDICTED: cathepsin B-like [Cimex lectularius]	234	234	92%	3e-72	38%	XP_0142
□ hypothetical protein g.26476 [Graphocephala atropunctata]	234	234	94%	3e-72	38%	JAT2482
□ PREDICTED: cathepsin B-like isoform X1 [Cimex lectularius]	233	233	94%	6e-72	38%	XP_0142
□ PREDICTED: cathepsin B-like isoform X2 [Cimex lectularius]	232	232	94%	7e-72	39%	XP_0142
□ hypothetical protein g.21611 [Homalodisca liturata]	232	232	94%	1e-71	37%	JAT0683
□ PREDICTED: cathepsin B [Tribolium castaneum]	231	231	94%	3e-71	37%	XP_9742

The NCBI Results Page cont'd

Pairwise Alignment



cathepsin B-348 precursor [Acyrthosiphon pisum]

Sequence ID: ref|NP_001119608.1| **Length:** 342 **Number of Matches:** 1

▸ See 1 more title(s)

Range 1: 9 to 341 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
229 bits(583)	5e-68	Compositional matrix adjust.	132/346(38%)	180/346(52%)	17/346(4%)

```
Query  1     MIHILVFLLG---CTLVRGELYKFSDAYIDQINREANTWTAGRNFPANLSEEYLRQFLIA  57
             ++ +L+F  G      VR +L   SD +ID IN     W+AGRNF  +    Y++  +
Sbjct  9     LVGLLIFSFGRVDGATVRVDLNPLSDEFIDHINSIQYYWSAGRNFHKDTPISYIKGLMGV  68

Query  58    DAKYFDQSDRPLPGDRKTYDPEYSATVPDRFDAREQWPNCGTIGHVPDTGACAAPHIFAA  117
               K    ++ P      TY+ + S  +P+ FDARE+WPNC TI  V D G+C +   F A
Sbjct  69    HEK---NAEYPKLEQLLTYN-DASTDLPETFDARERWPNCPTIREVRDQGSCGSCWAFGA  124

Query  118   VGAFSDRRCIKSKGQQNRPLSTEYVASCCKICRYDDNKSCSHGSVFRTWNFLHKRGSVTG  177
             V A SDR CI S G +N   S E + SCC  C +     C+ G     WN+   +G V+G
Sbjct  125   VEAMSDRVCIHSNGTKNFHFSAENLVSCCWTCGF----GCNGGFPGAAWNYWKTKGIVSG  180

Query  178   GDYGDRTGCQPSTISPCSHHGSAPTLPSCENQKVPKLKCHTRCTNPTYGRGFFQDKHRTT  237
             G YG   GC P  I+PC HH +    P  E  K P   C  +C    Y   + QD H
Sbjct  181   GPYGSNMGCIPYEIAPCEHHVNGTRGPCKEGGKTP--TCVKKCEE-GYKVPYAQDLHHGK  237

Query  238   LTYWVDDNEDAIKKEILAHGPTTATFALYDDFYHYKSGVYKHTSNAKLENYLHSGKLIGW  297
               Y + ++ D I++EI  +GP    F +Y+DF  Y++GVYKH +   L    H+ +++GW
Sbjct  238   SAYSIRNDVDQIRQEIYTNGPVEGAFTVYEDFIAYRAGVYKHVAGKALGG--HAIRILGW  295

Query  298   GTENG-TPYWLVINTWGPHWGDRGTVKILRGKYECAFEYLIAAGKP  342
             G +NG  PYWLV N+W   WG  G  KILRG  EC  E  I AG P
Sbjct  296   GVQNGEIPYWLVANSWNTDWGSDGFFKILRGSDECGIEGQINAGLP  341
```

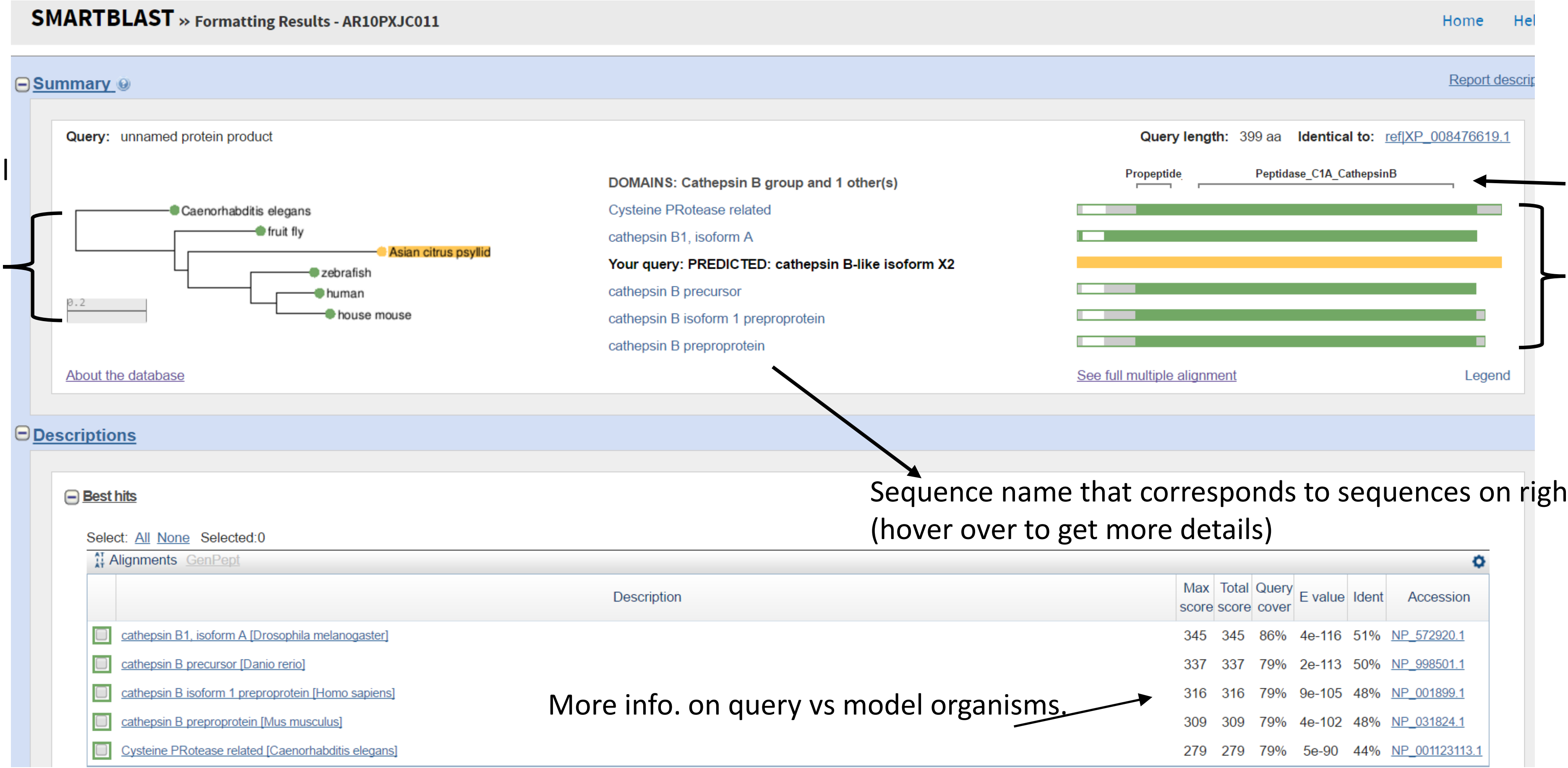
You can see how similar your gene model is to other similar genes in related organism(s), by clicking on each result (previous slide).

Example: The gene model is 38% identical and 52% similar to gene in *A. pisum*.

Or check gene model by using

NCBI Smart BLAST

https://blast.ncbi.nlm.nih.gov/smartblast/?LINK_LOC=BlastHomeLinkn



SMARTBLAST » Formatting Results - AR10PXJC011

Summary

Report descrip

Query: unnamed protein product

Query length: 399 aa **Identical to:** ref|XP_008476619.1

DOMAINS: Cathepsin B group and 1 other(s)

Cysteine PRotease related

cathepsin B1, isoform A

Your query: PREDICTED: cathepsin B-like isoform X2

cathepsin B precursor

cathepsin B isoform 1 preproprotein

cathepsin B preproprotein

About the database

See full multiple alignment

Legend

Descriptions

Best hits

Select: All None Selected:0

Alignments GenPept

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	cathepsin B1, isoform A [Drosophila melanogaster]	345	345	86%	4e-116	51%	NP_572920.1
□	cathepsin B precursor [Danio rerio]	337	337	79%	2e-113	50%	NP_998501.1
□	cathepsin B isoform 1 preproprotein [Homo sapiens]	316	316	79%	9e-105	48%	NP_001899.1
□	cathepsin B preproprotein [Mus musculus]	309	309	79%	4e-102	48%	NP_031824.1
□	Cysteine PRotease related [Caenorhabditis elegans]	279	279	79%	5e-90	44%	NP_001123113.1

Smart BLAST will provide a phylogenetic tree with model organisms.

Conserved domains within the sequences

Green: subject sequences

Yellow: your sequence (query)

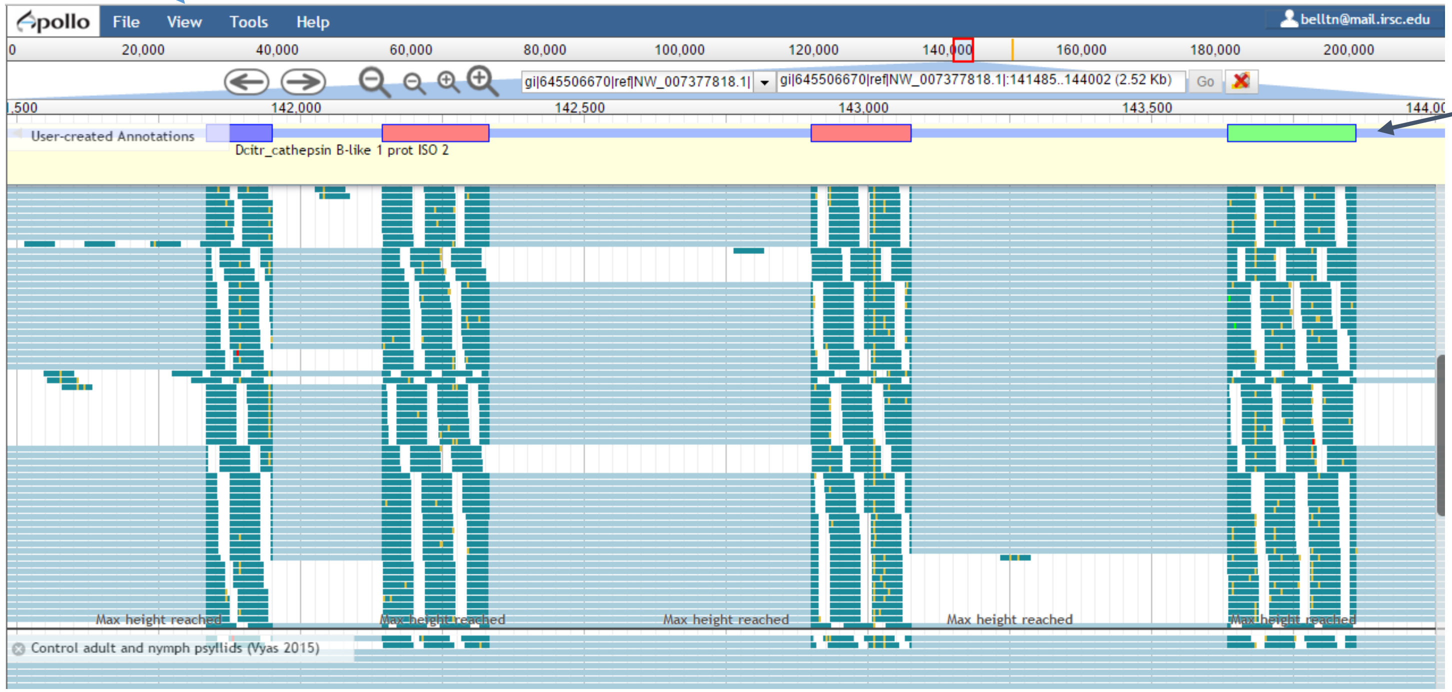
Sequence name that corresponds to sequences on right. (hover over to get more details)

More info. on query vs model organisms.

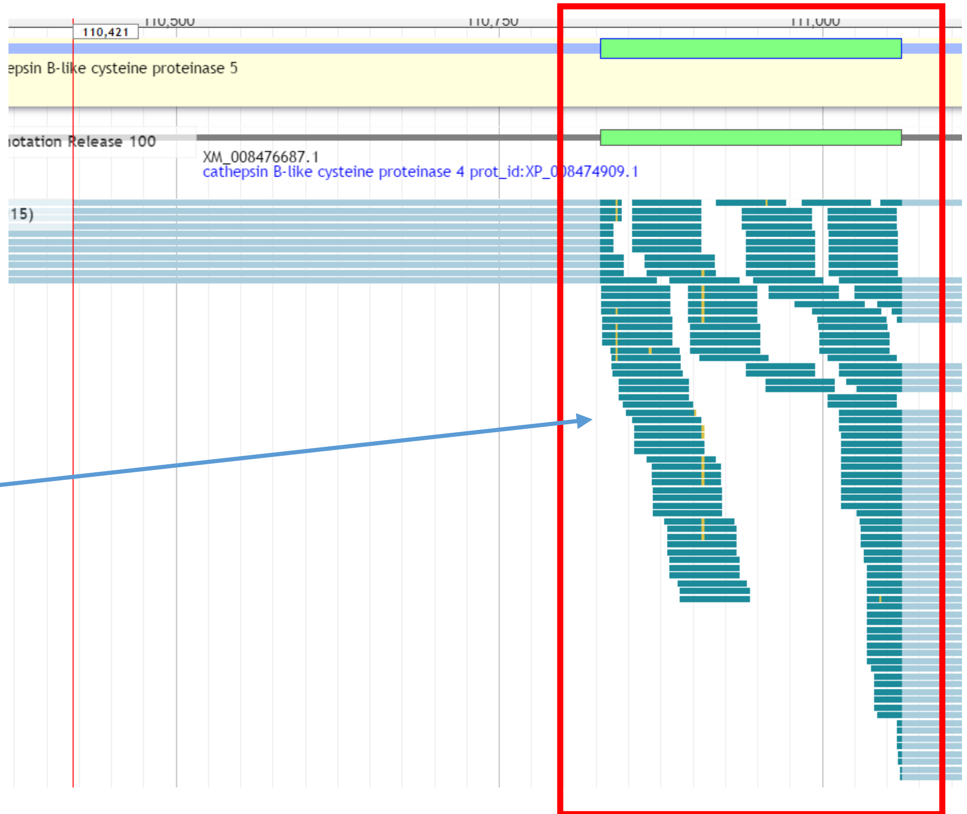
End of page would be BLASTp results.

RNASeq Data

To view gene model in different colors indicating the diff. frames: (As seen) Click View, Color by CDS frame.



The colored outlined boxes represent exons, while the thin blue lines connecting them represent introns. The blue boxes without an outline represent UTRs


110,421
110,500
110,750
111,000
epsin B-like cysteine proteinase 5
otation Release 100
XM_008476687.1
cathepsin B-like cysteine proteinase 4 prot_id:XP_008474909.1
15)

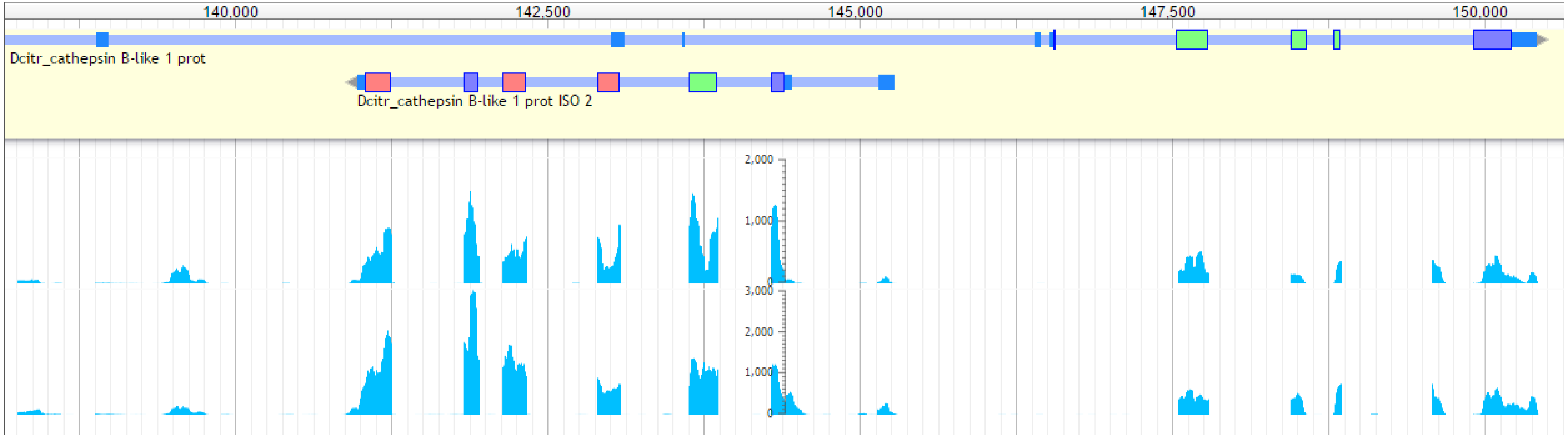
Look closely at the results and see how they line up with RNASeq data. You may have to zoom in close and look by sections.

Dark blue RNA seq reads indicate evidence for exons and light blue for introns.

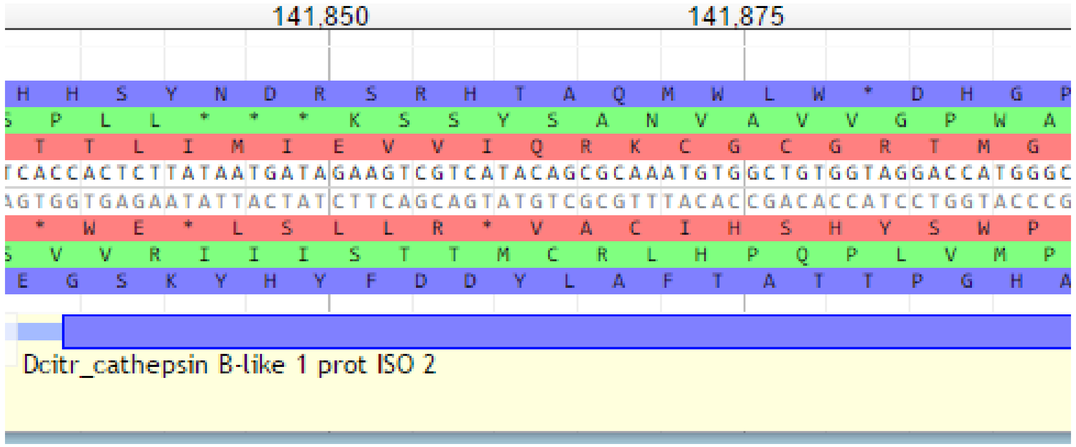
RNASeq Data

More details

You can also view the RNAseq Coverage Plots.
(On the left side on Apollo (Available Tracks))


140,000
142,500
145,000
147,500
150,000
Dcitr_cathepsin B-like 1 prot
Dcitr_cathepsin B-like 1 prot ISO 2
2,000
1,000
0
3,000
2,000
1,000
0

If you zoom in and see letters, this is the reference sequence, as shown here. You can view stops on the different frames by an *.


141,850
141,875
H H S Y N D R S R H T A Q M W L W * D H G P
S P L L * * * K S S Y S A N V A V V G P W A
T T L I M I E V V I Q R K C G C G R T M G
TCACCACTCTTATAATGATAGAAGTCGTCATACAGCGCAAATGTGGCTGTGGTAGGACCATGGGC
AGTGGTGAGAATATTACTATCTTCAGCAGTATGTCGCGTTTACACCGACACCATCCTGGTACCCG
* W E * L S L L R * V A C I H S H Y S W P
S V V R I I I S T T M C R L H P Q P L V M P
E G S K Y H Y F D D Y L A F T A T T P G H A
Dcitr_cathepsin B-like 1 prot ISO 2

Making Edits

Matching Data

Delete exons, or UTR (Untranslated region)

Merge exons, or gene models

Split

Duplicate

Make Intron

Set to Downstream Splice Donor

Set to Upstream Splice Donor

Set to Downstream Splice Acceptor

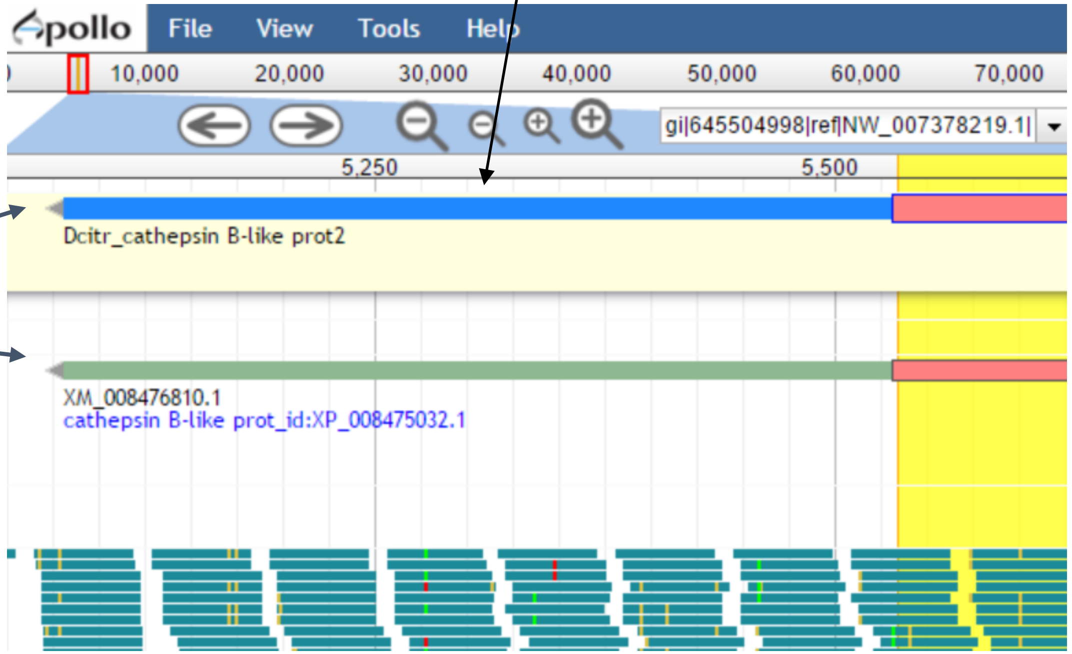
Set to Upstream Splice Acceptor

View History

UTR's are dark blue and are on the 5' and 3' ends of the gene model.

Example: This is the 3' UTR of the gene model

Indicates which direction (Upstream or Downstream) the sequence is reading. In this case = downstream (-) strand



Apollo
File
View
Tools
Help
10,000
20,000
30,000
40,000
50,000
60,000
70,000
gi|645504998|ref|NW_007378219.1|
5,250
5,500
Dcitr_cathepsin B-like prot2
XM_008476810.1
cathepsin B-like prot_id:XP_008475032.1

Important:

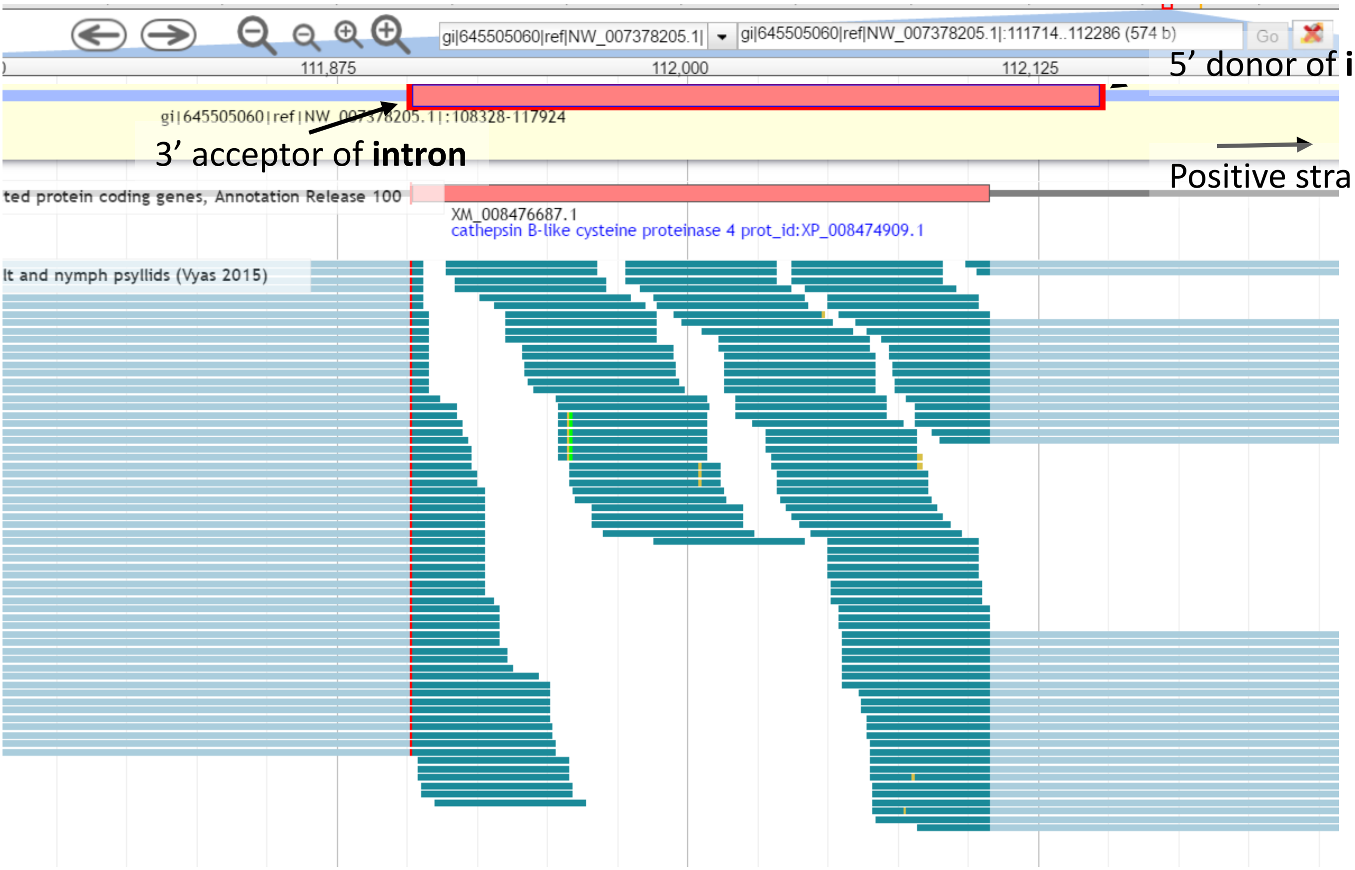
Splice sites:

Reverse (-) strand gene model:
3' acceptor of intron, GA
5' donor of intron, TG, CG
Positive (+) strand gene model:
3' acceptor of intron, AG
5' donor of intron, GT, GC

Make sure your gene model starts with ATG (M) and ends with TAG, TGA, or TAA (stop codons).

Fixing splice sites

Splice sites:
Reverse (-) strand gene model:
3’ acceptor of intron, GA
5’ donor of intron, TG, CG
Positive (+) strand gene model:
3’ acceptor of intron, AG
5’ donor of intron, GT, GC



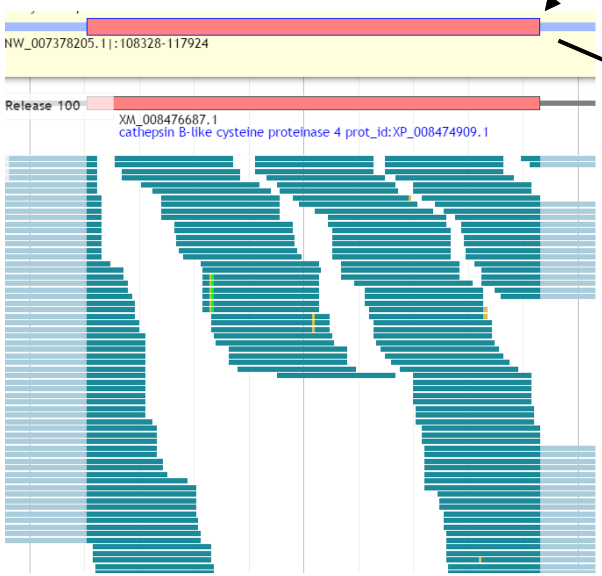
gi|645505060|ref|NW_007378205.1|
gi|645505060|ref|NW_007378205.1|:111714..112286 (574 b)
Go
111,875
112,000
112,125
gi|645505060|ref|NW_007378205.1|:108328-117924
ted protein coding genes, Annotation Release 100
XM_008476687.1
cathepsin B-like cysteine proteinase 4 prot_id:XP_008474909.1
lt and nymph psyllids (Vyas 2015)

5’ donor of **intron**

3’ acceptor of **intron**

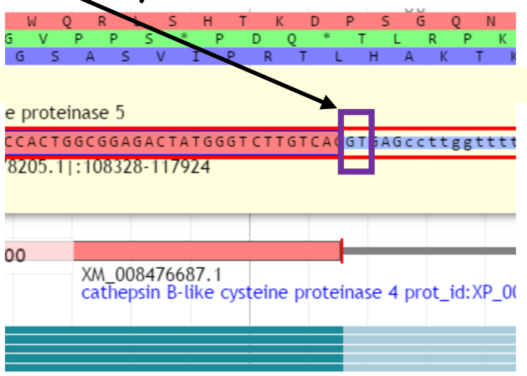
Positive strand

For this example, the donor would be moved upstream. (RNAseq evidence to do so)
Right click exon, Set to upstream splice donor.



NW_007378205.1|:108328-117924
Release 100
XM_008476687.1
cathepsin B-like cysteine proteinase 4 prot_id:XP_008474909.1

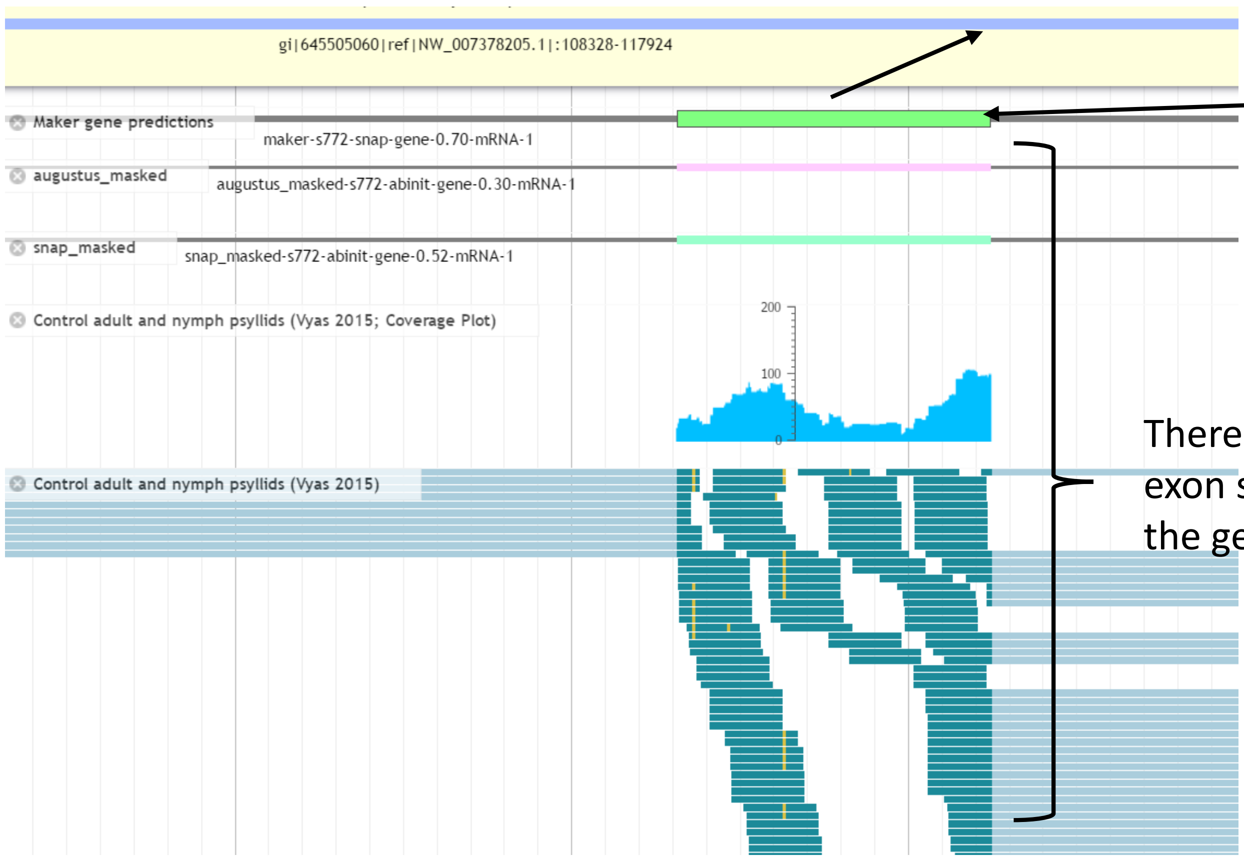
Zoom in



W Q R L S H T K D P S G Q N
G V P P S * P D Q * T L R P K
G S A S V I P R T L H A K T K
e proteinase 5
CCACTGGCGGAGACTATGGGTCTTGTCAGGTGAGccttggtttt
8205.1|:108328-117924
00
XM_008476687.1
cathepsin B-like cysteine proteinase 4 prot_id:XP_00

Adding an exon

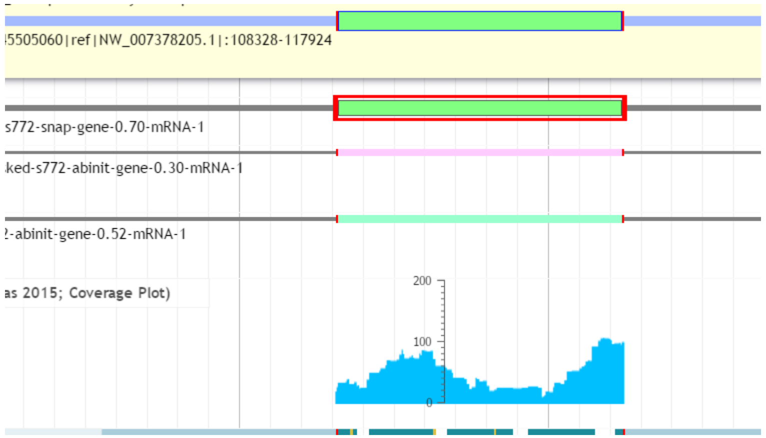
Other gene predictions can be used, not limited to, Maker, augustus, snap.


gi|645505060|ref|NW_007378205.1|:108328-117924
Maker gene predictions
maker-s772-snap-gene-0.70-mRNA-1
augustus_masked
augustus_masked-s772-abinit-gene-0.30-mRNA-1
snap_masked
snap_masked-s772-abinit-gene-0.52-mRNA-1
Control adult and nymph psyllids (Vyas 2015; Coverage Plot)
200
100
0
Control adult and nymph psyllids (Vyas 2015)

Click and drag exon on top of gene model and drop.

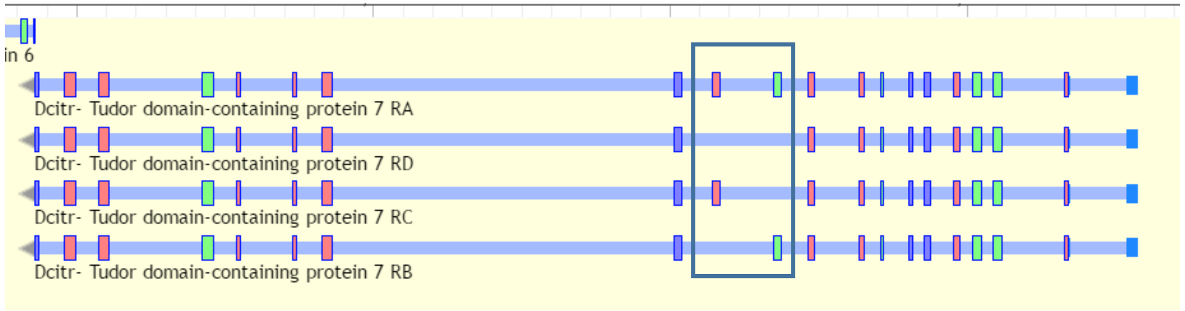
There is evidence that an exon should be added to the gene model

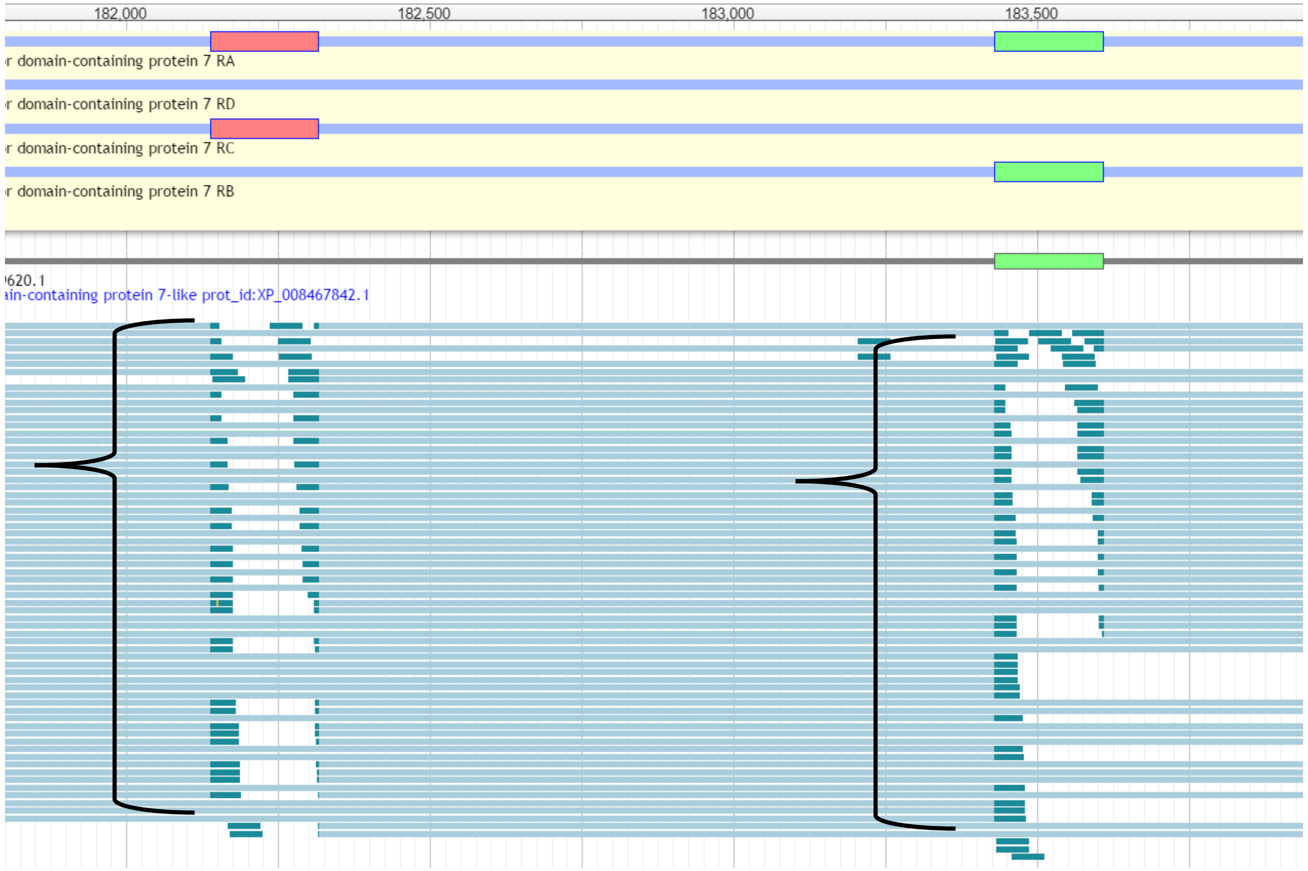
After


5505060|ref|NW_007378205.1|:108328-117924
s772-snap-gene-0.70-mRNA-1
ked-s772-abinit-gene-0.30-mRNA-1
2-abinit-gene-0.52-mRNA-1
as 2015; Coverage Plot)
200
100
0

Isoforms

When the gene can undergo alternative splicing


in 6
Dcitr- Tudor domain-containing protein 7 RA
Dcitr- Tudor domain-containing protein 7 RD
Dcitr- Tudor domain-containing protein 7 RC
Dcitr- Tudor domain-containing protein 7 RB

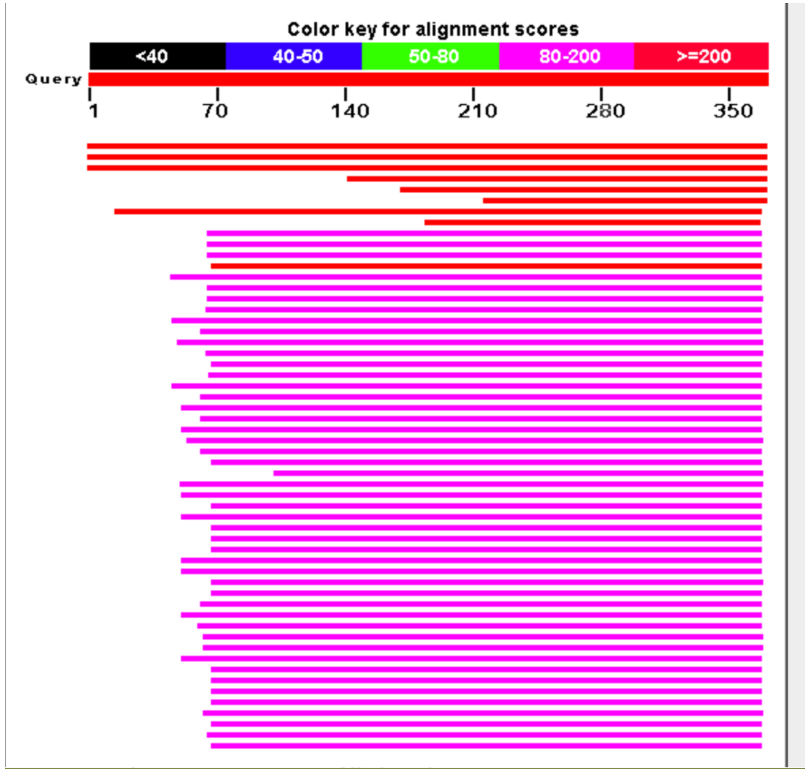

182,000
182,500
183,000
183,500
r domain-containing protein 7 RA
r domain-containing protein 7 RD
r domain-containing protein 7 RC
r domain-containing protein 7 RB
'620.1
ain-containing protein 7-like prot_id:XP_008467842.1

RNAseq reads indicate intron and exon evidence for the same part of the model.

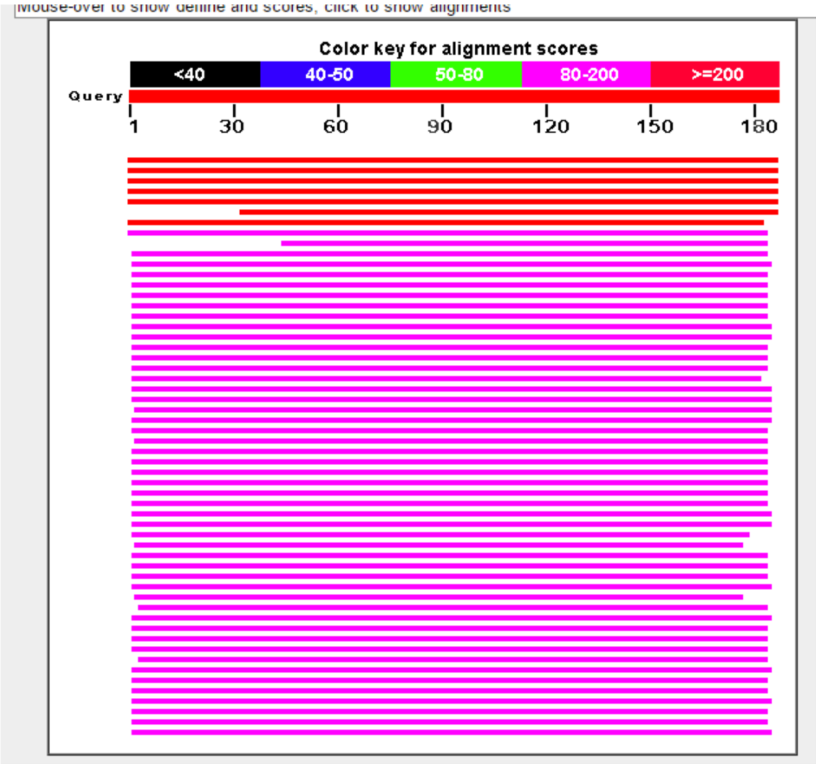
If there were many results containing isoforms when the gene model was initially blasted on NCBI, this is good evidence for isoforms as well.

If you have evidence for isoforms, right click your gene model in the “user-created annotations” space, and duplicate.

Edited Gene Models



Predicted Model Conserved domain results

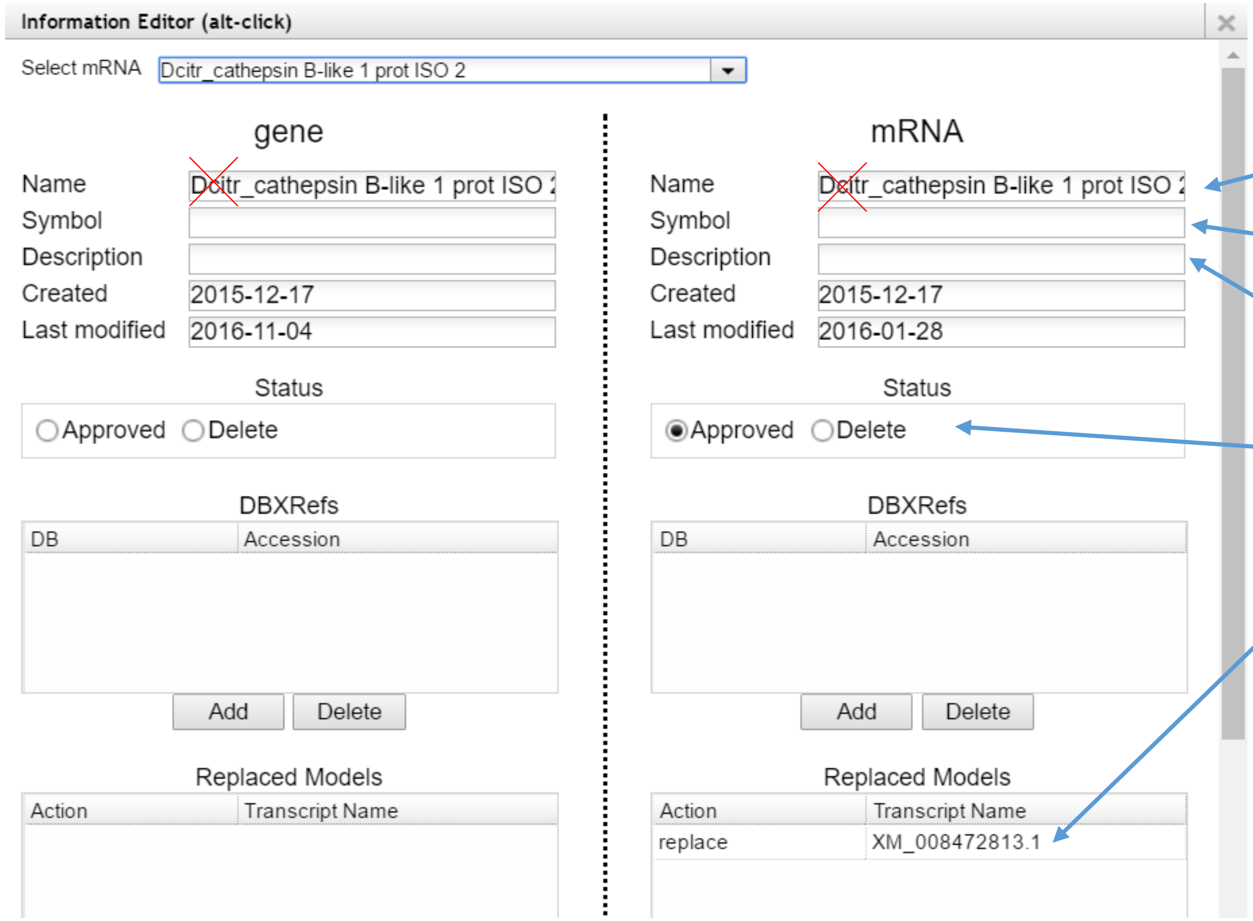

Mouse-over to show defline and scores, click to show alignments

Conserved domain results after matching your gene model with RNASeq data

After a change is made, check your results using the BLASTp or Smart BLAST again

Information editor

Right click on the model in the User Created section and choose “Edit Information”



Information Editor (alt-click)

Select mRNA: Dcitr_cathepsin B-like 1 prot ISO 2

gene

Name	Dcitr_cathepsin B-like 1 prot ISO 2
Symbol	
Description	
Created	2015-12-17
Last modified	2016-11-04

Status

○Approved ○Delete

DBXRefs

DB	Accession

Add Delete

Replaced Models

Action	Transcript Name

mRNA

Name	Dcitr_cathepsin B-like 1 prot ISO 2
Symbol	
Description	
Created	2015-12-17
Last modified	2016-01-28

Status

●Approved ○Delete

DBXRefs

DB	Accession

Add Delete

Replaced Models

Action	Transcript Name
replace	XM_008472813.1

Name of the gene. Do NOT use Dcitr

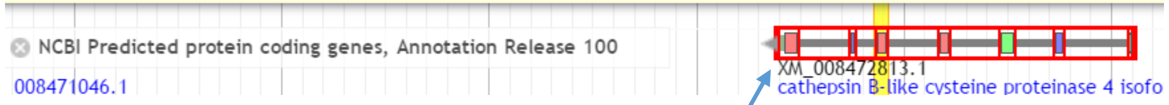
Symbol or abbreviation of the gene

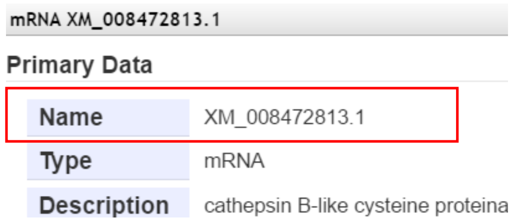
Description of the gene, usually function. Uniprot is a good source.

Click approved once all edits are made and your gene model is complete

You will need the name of the original model that your annotation will be replacing

Example:


NCBI Predicted protein coding genes, Annotation Release 100
008471046.1
XM_008472813.1
cathepsin B-like cysteine proteinase 4 isofo



mRNA XM_008472813.1

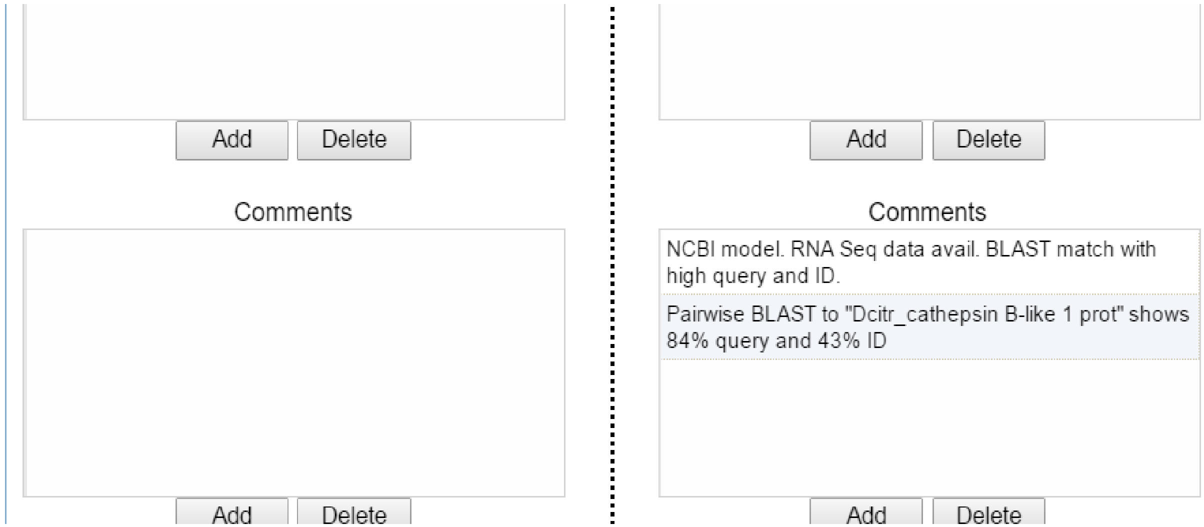
Primary Data

Name	XM_008472813.1
Type	mRNA
Description	cathepsin B-like cysteine proteina

Right click on model > View details

Fill in both sections: Gene and mRNA panels

Information editor cont'd

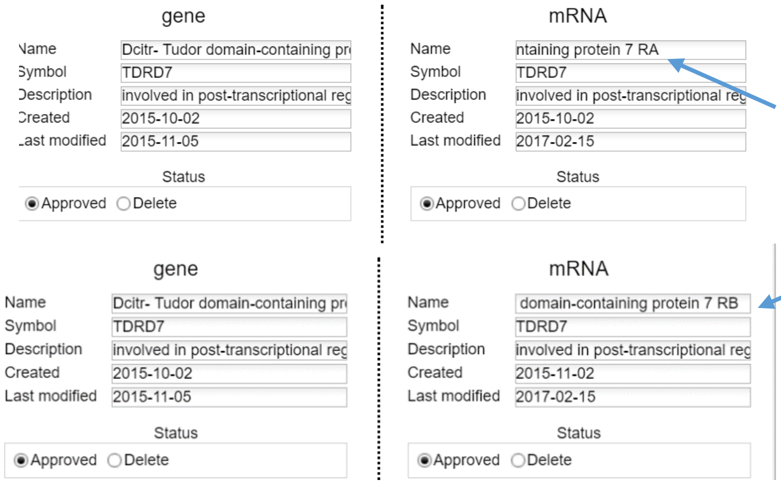


NCBI model. RNA Seq data avail. BLAST match with high query and ID.

Pairwise BLAST to "Dcitr_cathepsin B-like 1 prot" shows 84% query and 43% ID

Here you will enter comments about all the edits (splice sites moved, added exons, deleted exons, etc.) made to your gene model and other relevant info.

Naming isoforms:



gene

Name	Dcitr- Tudor domain-containing pr
Symbol	TDRD7
Description	involved in post-transcriptional reg
Created	2015-10-02
Last modified	2015-11-05

Status: ● Approved ○ Delete

mRNA

Name	ntaining protein 7 RA
Symbol	TDRD7
Description	involved in post-transcriptional reg
Created	2015-10-02
Last modified	2017-02-15

Status: ● Approved ○ Delete

gene

Name	Dcitr- Tudor domain-containing pr
Symbol	TDRD7
Description	involved in post-transcriptional reg
Created	2015-10-02
Last modified	2015-11-05

Status: ● Approved ○ Delete

mRNA

Name	domain-containing protein 7 RB
Symbol	TDRD7
Description	involved in post-transcriptional reg
Created	2015-11-02
Last modified	2017-02-15

Status: ● Approved ○ Delete

Naming convention for isoforms should have RA, RB, RC, etc. at the end for each isoform you have, respectively. (look at slide 20)

The gene section should be the same for all isoforms.

But the mRNA section should be different for all isoforms, comments should contain the edits made for the particular isoform clicked on.

Gene Family Report

Once your gene or gene family is annotated, the gene report should include:

- Introduction to the gene or gene family including the pathway (use literature)
- Methods of annotation and phylogenetic analysis
- Results and Discussion (use literature)
- References cited

Tables to include in report

Table with gene count in:

- *Drosophila melanogaster*
- *Anopheles gambiae*
- *Tribolium castaneum*
- *Apis mellifera*
- *Nasonia vitripennis*
- *Acyrthosiphon pisum*
- *Bemicia tabaci*

Table with blast match:

- Indicate organism and name of gene
- Query coverage
- % identity
- Bit score

Include legend for each table.

Comparative and Phylogenetic Analysis

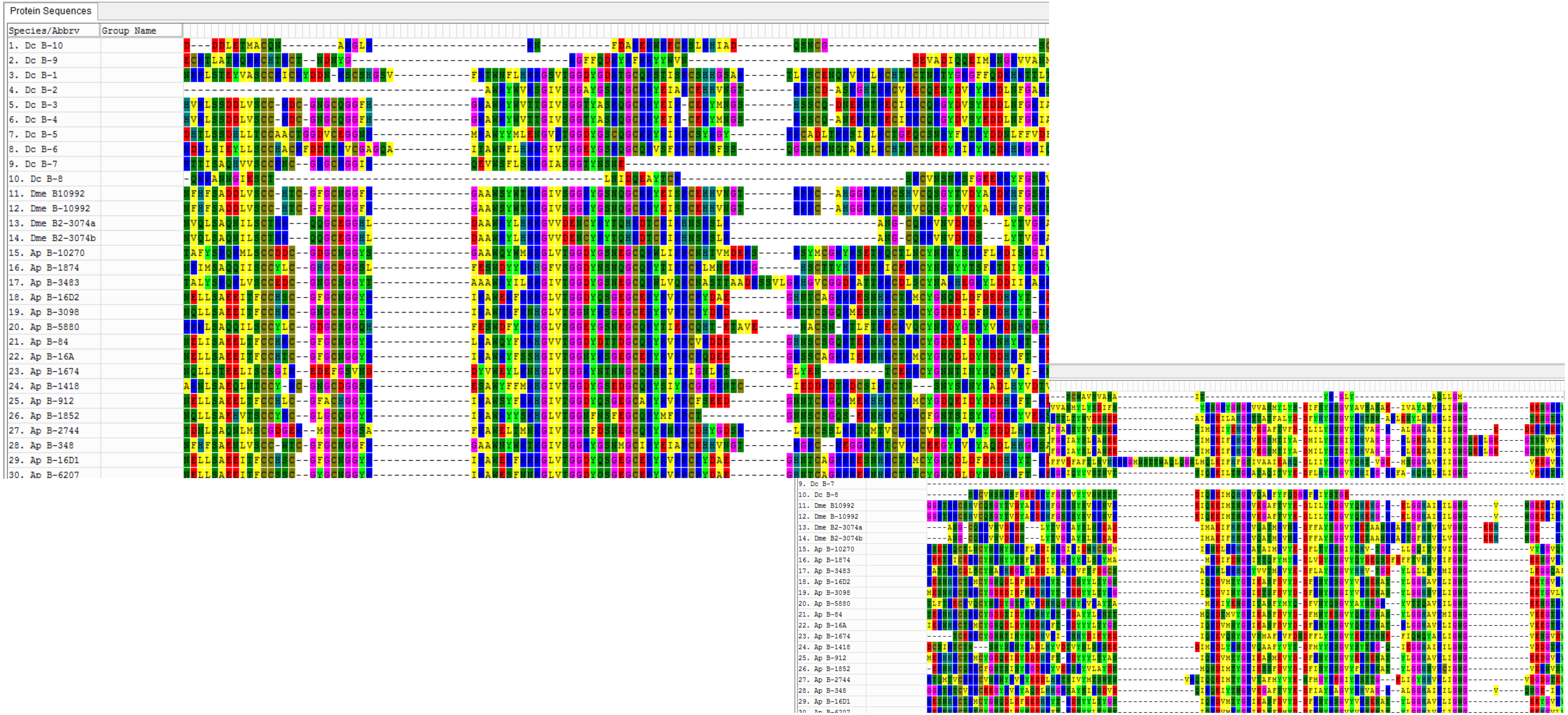
When your gene model(s) are completed, perform analysis in MEGA.

1. Construct tree with *D. citri* gene models only.
2. Construct tree with *D. citri* gene models and orthologs. Use related sequences from, but not limited to,
   - *Drosophila melanogaster*
   - *Anopheles gambiae*
   - *Tribolium castaneum*
   - *Apis mellifera*
   - *Nasonia vitripennis*
   - *Acyrthosiphon pisum*
   - *Bemicia tabaci*

Can find orthologous sequences in NCBI, Ensembl, Uniprot, etc.

Multiple Sequence Alignments (MSA)

Multiple sequence alignments should be generated using MUSCLE, tcoffee, or clustal to compare the ACP gene model to the query gene set. The final model should be refined in Apollo using homology, RNAseq and proteomics evidence tracks.



Protein Sequences

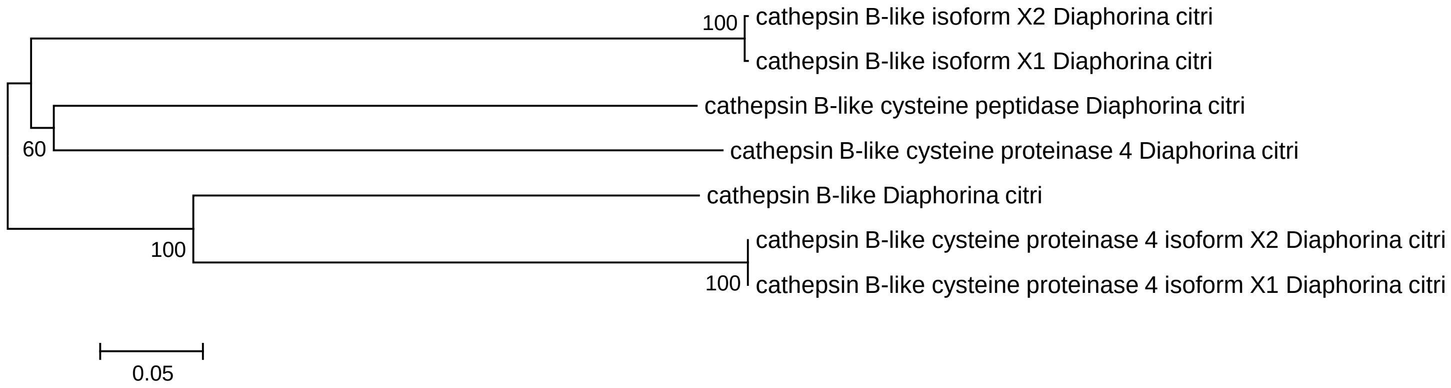
Species/Abbrv	Group Name
1. Dc B-10	
2. Dc B-9	
3. Dc B-1	
4. Dc B-2	
5. Dc B-3	
6. Dc B-4	
7. Dc B-5	
8. Dc B-6	
9. Dc B-7	
10. Dc B-8	
11. Dme B10992	
12. Dme B-10992	
13. Dme B2-3074a	
14. Dme B2-3074b	
15. Ap B-10270	
16. Ap B-1874	
17. Ap B-3483	
18. Ap B-16D2	
19. Ap B-3098	
20. Ap B-5880	
21. Ap B-84	
22. Ap B-16A	
23. Ap B-1674	
24. Ap B-1418	
25. Ap B-912	
26. Ap B-1852	
27. Ap B-2744	
28. Ap B-348	
29. Ap B-16D1	
30. Ap B-6207	

9. Dc B-7
10. Dc B-8
11. Dme B10992
12. Dme B-10992
13. Dme B2-3074a
14. Dme B2-3074b
15. Ap B-10270
16. Ap B-1874
17. Ap B-3483
18. Ap B-16D2
19. Ap B-3098
20. Ap B-5880
21. Ap B-84
22. Ap B-16A
23. Ap B-1674
24. Ap B-1418
25. Ap B-912
26. Ap B-1852
27. Ap B-2744
28. Ap B-348
29. Ap B-16D1
30. Ap B-6207

Tree of all *D. citri* gene models

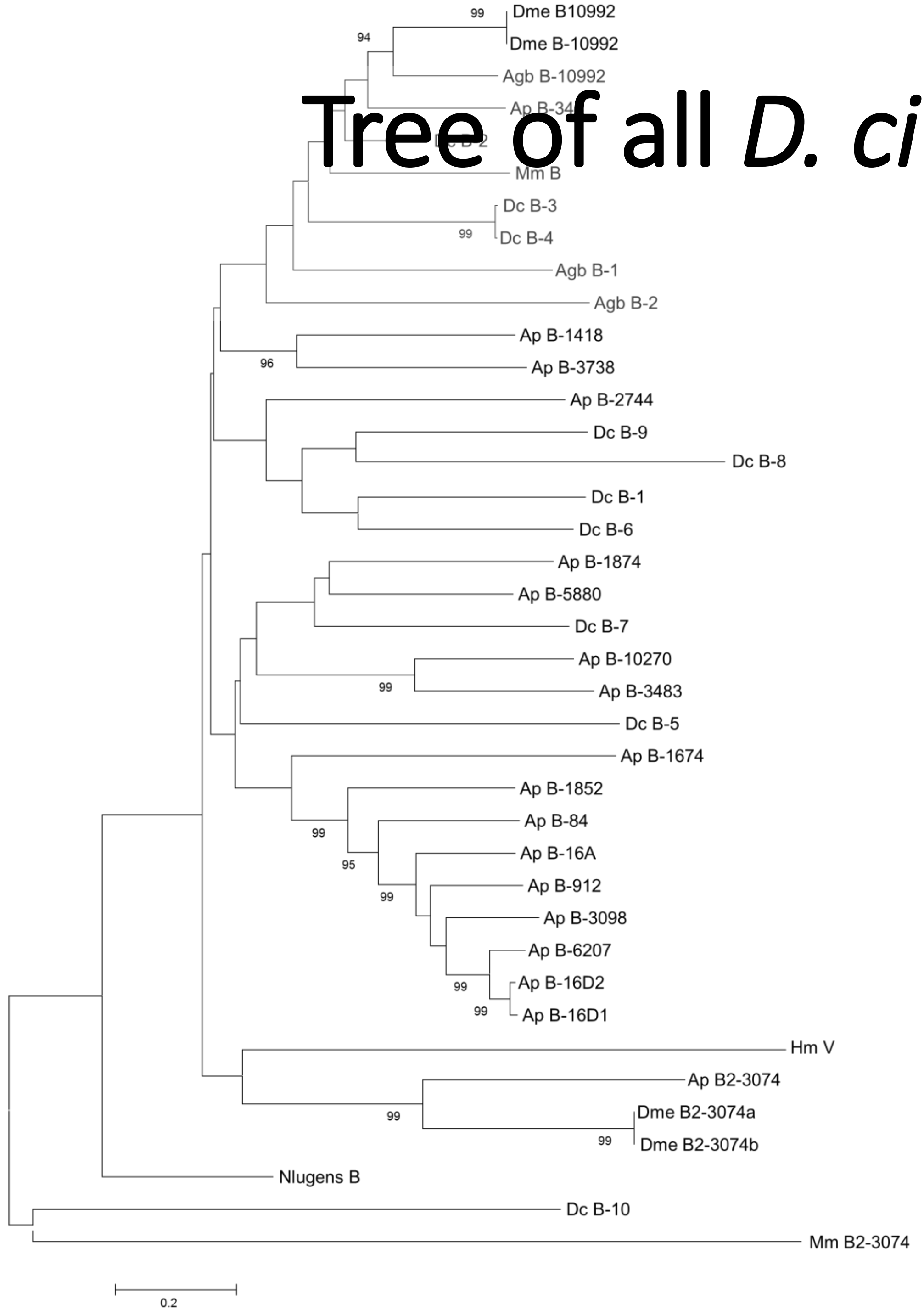
This would include all genes (paralogs) for the particular family annotated in the genome.

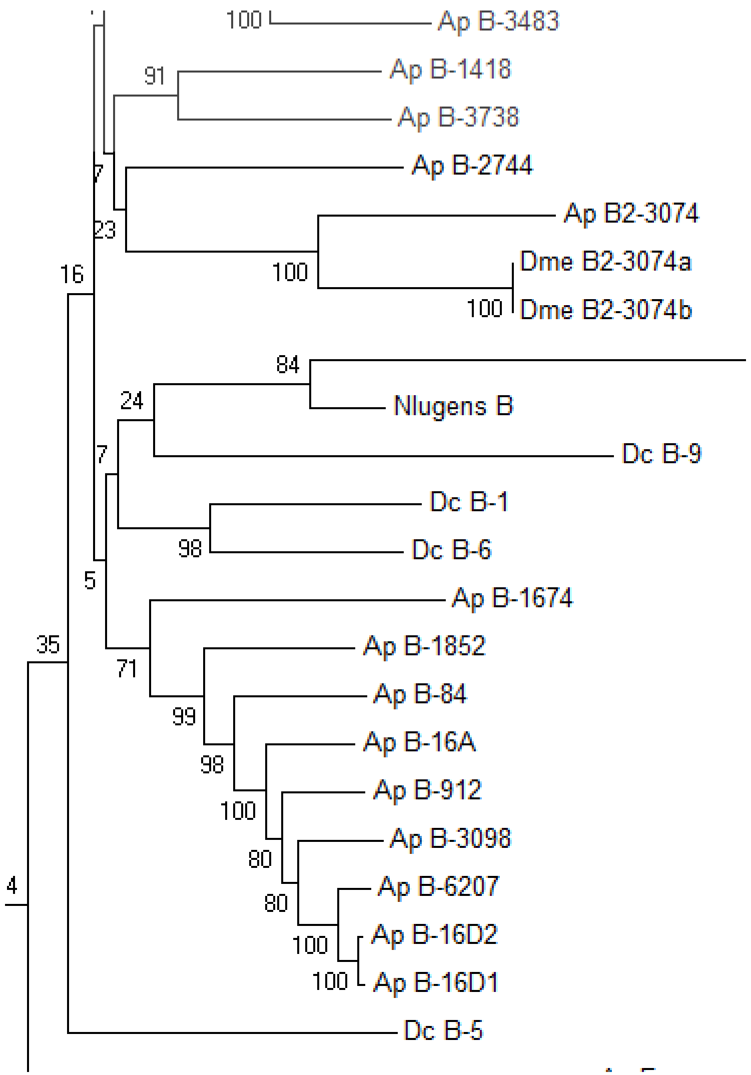
Example:


100
cathepsin B-like isoform X2 Diaphorina citri
cathepsin B-like isoform X1 Diaphorina citri
cathepsin B-like cysteine peptidase Diaphorina citri
60
cathepsin B-like cysteine proteinase 4 Diaphorina citri
cathepsin B-like Diaphorina citri
100
cathepsin B-like cysteine proteinase 4 isoform X2 Diaphorina citri
100
cathepsin B-like cysteine proteinase 4 isoform X1 Diaphorina citri
0.05

Tree should be included in report with legend.

Tree of all *D. citri* gene models and other orthologs





Tree should be included in report with legend

Citrus Greening Solutions
A USDA NIFA Project

Krystal Villalobos Ayala
Chris Cordola
Tracey Bell
Hannah Mann
Daniel DeAvila
Gabe DeAvila
Tom D’elia

Indian River State College, FL