Finding *Schistocerca* genes in NCBI and making SnapGene files for cloning and alignments

What do you need?

1. NCBI website to find the genome and retrieve annotated GenBank files
   https://www.ncbi.nlm.nih.gov/

2. FlyBase or other server to find gene orthologs you want to find in *Schistocerca*
   https://flybase.org/

3. Some basic knowledge of BLAST
   https://blast.ncbi.nlm.nih.gov/Blast.cgi
   https://chanzuckerberg.zendesk.com/hc/en-us/articles/360050963352-A-Guide-to-BLAST

4. Gene visualization/manipulation software to look at/manipulate retrieved annotated Gene file

   SnapGene which has a free version called SnapGene Viewer
   https://www.snapgene.com/snapgene-viewer

   Benchling
   https://www.benchling.com/molecular-biology

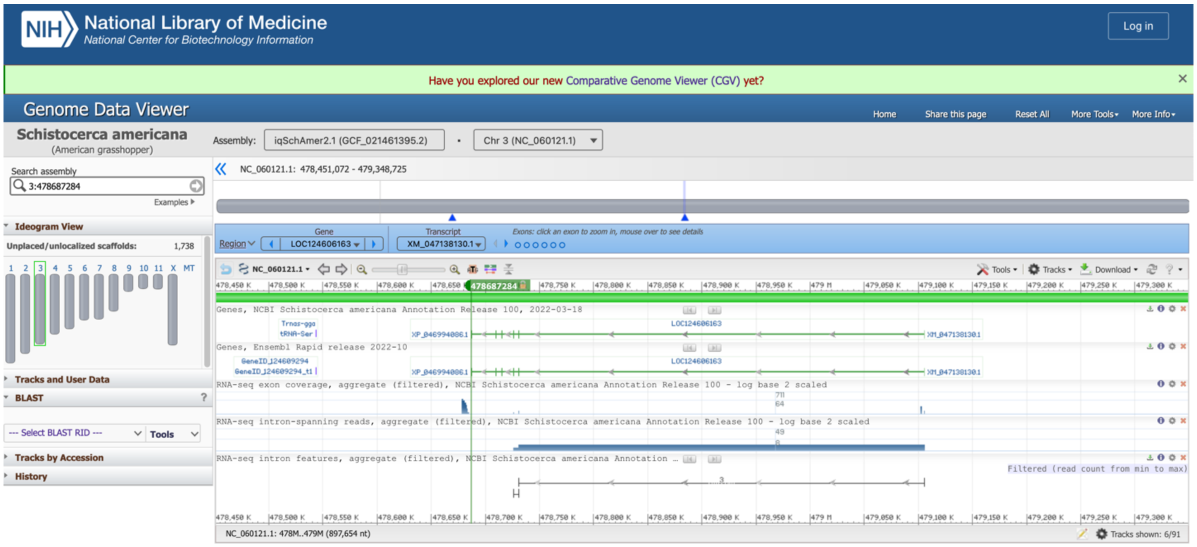
What will you be able to do once you learn this tutorial?

With experience, it will take ~10-15 min to retrieve an annotated GenBank file of your favorite gene and open it in SnapGene to further work with for Protein sequence alignments and building trees for example

With an annotated sequence from one *Schistocerca* species, you will find and retrieve the orthologous sequences in the other *Schistocerca* species in 5-10 min

> Note: Retrieving an annotated gene model from NCBI assumes that model is correct. All models are computer generated and may have errors. Once you do sequence comparison with the orthologous genes in different species errors may become apparent and these should be confirmed by the transcriptomic data. Monica's approach in Apollo will help verify and modify existing gene models.

Go to NCBI
Type in Schistocerca in the search bar
Go to the bottom left box, Genomes
Go to first line, Assembly (there should be 7)
Click on the genome you want to search, for example *Schistocerca americana* (iqSchAmer2.1)
On the top right of the page, under Access the data, click on Genome Data Viewer
This brings you to the page below:
https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_021461395.2



National Library of Medicine
National Center for Biotechnology Information

Have you explored our new Comparative Genome Viewer (CGV) yet?

Genome Data Viewer

Schistocerca americana
(American grasshopper)

Assembly: iqSchAmer2.1 (GCF_021461395.2) • Chr 3 (NC_060121.1)

Search assembly: 3:478687284

NC_060121.1: 478,451,072 - 479,348,725

Gene: LOC124606163
Transcript: XM_047138130.1

Exons: click an exon to zoom in, mouse over to see details

Unplaced/unlocalized scaffolds: 1,738

Genes, NCBI Schistocerca americana Annotation Release 100, 2022-03-18
Genes, Ensembl Rapid release 2022-10
RNA-seq exon coverage, aggregate (filtered), NCBI Schistocerca americana Annotation Release 100 - log base 2 scaled
RNA-seq intron-spanning reads, aggregate (filtered), NCBI Schistocerca americana Annotation Release 100 - log base 2 scaled
RNA-seq intron features, aggregate (filtered), NCBI Schistocerca americana Annotation ...
Filtered (read count from min to max)

NC_060121.1: 478M..479M (897,654 nt)
Tracks shown: 6/91

There are 3 ways you can find a gene

1. Use a search term in “Search assembly”
2. Blast the CDS of the fly ortholog
3. Blast the protein sequence of the fly ortholog

First method: In search assembly, you can type in a gene name, for example “hyperpolarization” (this will find 4 genes that are of the HCN family)

Click on the one or ones that you want to download in SnapGene and use for further analysis
The screen will jump to the gene diagram that you want to focus on.
Proceed to Step 3 in the procedure below to retrieve your annotated GenBank gene file

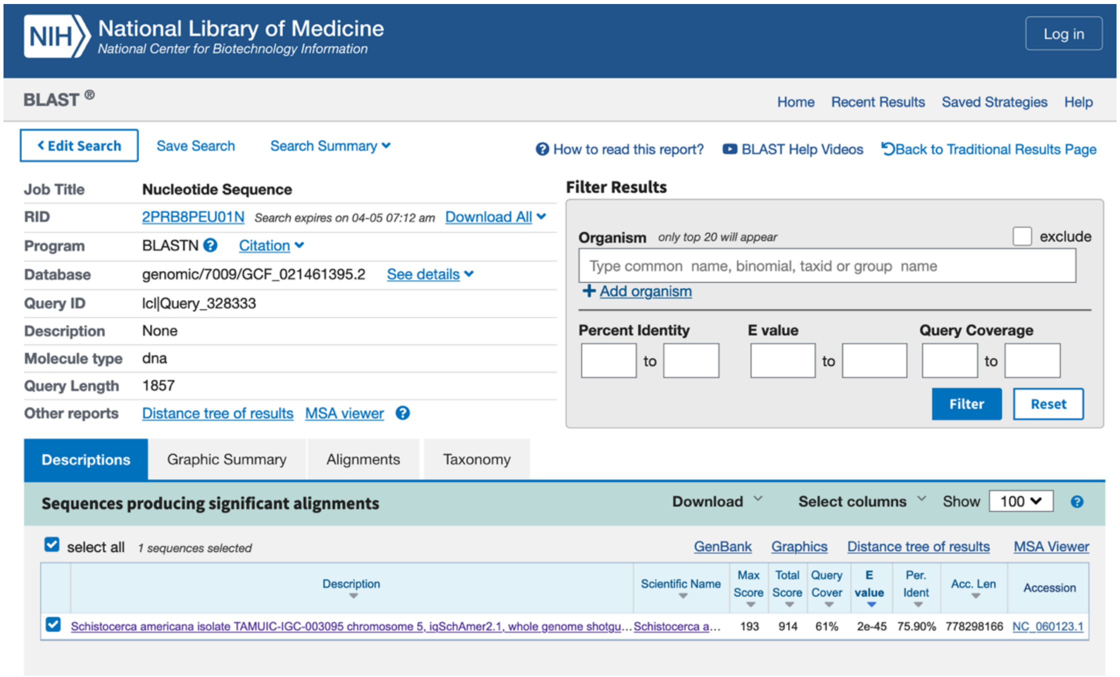
Second method: perform a BLAST search using the Drosophila ortholog CDS

Step 1: retrieve the CDS sequence from FLYBASE

Go to FlyBase
Type in the gene name in the Flybase Search window (in our example the fly gene is *ih*)
This brings up a page with gene hits or takes you directly to the gene page
In the gene page, go to the 2nd box labeled Genome Location
Under sequence tab you can click on the drop down menu with Gene Region
Select CDS and then click get sequence
This brings you to the page of the coding sequence
Copy the complete CDS of the *ih* gene

Step 2: Perform a BLAST search from within the Genome Data Viewer for Schistocerca

Go back to the Genome Data Viewer of *Schistocerca* to perform your Blast search with the fly CDS
Go to the Blast button below the chromosome diagram window and click on tools
Click New BLAST
The default BLAST is a blastn for nucleotide sequence
Paste the Drosophila CDS in the window
Optimize the blast to the lowest similarity “somewhat similar sequences” (*see below)
Click BLAST, the page will refresh until the search is finished



Job Title	Nucleotide Sequence
RID	2PRB8PEU01N *Search expires on 04-05 07:12 am* Download All
Program	BLASTN Citation
Database	genomic/7009/GCF_021461395.2 See details
Query ID	lcl|Query_328333
Description	None
Molecule type	dna
Query Length	1857
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* □ exclude
Type common name, binomial, taxid or group name
+ Add organism
Percent Identity ___ to ___ E value ___ to ___ Query Coverage ___ to ___
Filter Reset

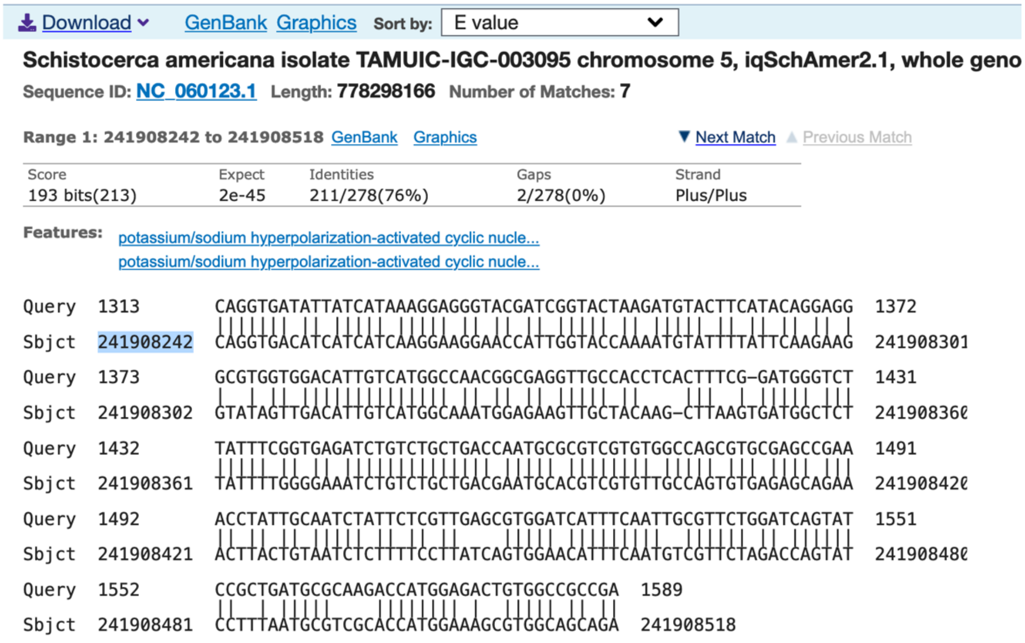
Sequences producing significant alignments

Show 100

☑ select all *1 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Schistocerca americana isolate TAMUIC-IGC-003095 chromosome 5, iqSchAmer2.1, whole genome shotgu...	Schistocerca a...	193	914	61%	2e-45	75.90%	778298166	NC_060123.1

This method returned 1 hit with a very good score on Chr 5 of the *Sa* genome (https://chanzuckerberg.zendesk.com/hc/en-us/articles/360050963352-A-Guide-to-BLAST)
Click on the hit, and you will find that the CDS shows high homology (~75%) to 7 pieces, which represent exons of the gene in the *Schistocerca americana* genome
Copy the coordinates of the gene in the first hit highlighted in blue below:



Download ⌄ GenBank Graphics Sort by: E value ⌄

Schistocerca americana isolate TAMUIC-IGC-003095 chromosome 5, iqSchAmer2.1, whole geno

Sequence ID: NC_060123.1 Length: 778298166 Number of Matches: 7

Range 1: 241908242 to 241908518 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
193 bits(213)	2e-45	211/278(76%)	2/278(0%)	Plus/Plus

Features: potassium/sodium hyperpolarization-activated cyclic nucle...
potassium/sodium hyperpolarization-activated cyclic nucle...

```
Query  1313       CAGGTGATATTATCATAAAGGAGGGTACGATCGGTACTAAGATGTACTTCATACAGGAGG  1372
                  ||||||| || ||||| ||||| || || || ||||| || ||||| || || || || |
Sbjct  241908242  CAGGTGACATCATCATCAAGGAAGGAACCATTGGTACCAAAATGTATTTTATTCAAGAAG  241908301

Query  1373       GCGTGGTGGACATTGTCATGGCCAACGGCGAGGTTGCCACCTCACTTTCG-GATGGGTCT  1431
                  |  | || |||||||||||||| || || || ||||| ||    |||  | ||||| |||
Sbjct  241908302  GTATAGTTGACATTGTCATGGCAAATGGAGAAGTTGCTACAAG-CTTAAGTGATGGCTCT  241908360

Query  1432       TATTTCGGTGAGATCTGTCTGCTGACCAATGCGCGTCGTGTGGCCAGCGTGCGAGCCGAA  1491
                  ||||| || || |||||||||||||| ||||| |||||||| ||||| ||| |||| |||
Sbjct  241908361  TATTTTGGGGAAATCTGTCTGCTGACGAATGCACGTCGTGTTGCCAGTGTGAGAGCAGAA  241908420

Query  1492       ACCTATTGCAATCTATTCTCGTTGAGCGTGGATCATTTCAATTGCGTTCTGGATCAGTAT  1551
                  || || || ||||| || || ||    ||||| |||||||||  |||||| || ||||||
Sbjct  241908421  ACTTACTGTAATCTCTTTTCCTTATCAGTGGAACATTTCAATGTCGTTCTAGACCAGTAT  241908480

Query  1552       CCGCTGATGCGCAAGACCATGGAGACTGTGGCCGCCGA  1589
                  ||  | |||||    |||||||| |  ||||| || ||
Sbjct  241908481  CCTTTAATGCGTCGCACCATGGAAAGCGTGGCAGCAGA  241908518
```

To locate the gene in the Genome Data Viewer, copy the coordinate into the Search Assembly box preceded by the chromosome number, so in this case type “5:241908242” and click the arrow.

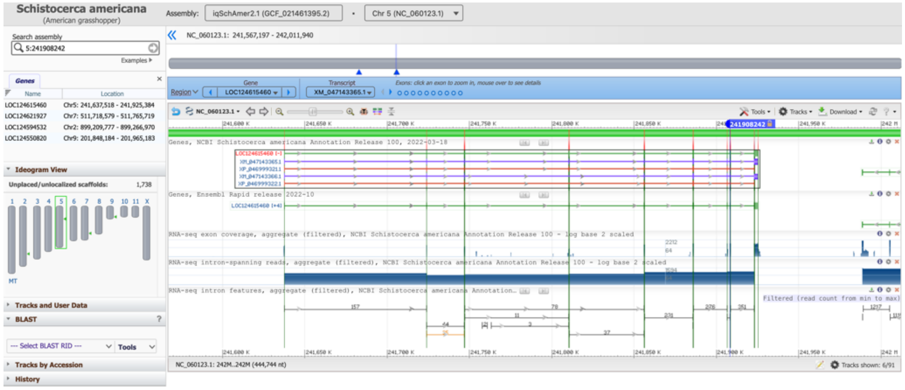
The screen to the right will jump to that specific base in the gene. Click the zoom out button on the menu bar above the chromosome until you see the whole gene diagram.

Step 3: Retrieve the GenBank file of your annotated gene

Click on the gene diagram

It will now add several colored lines to the diagram

1. Green represents the gene
2. Purple the mRNA
3. Red the protein



Schistocerca americana (American grasshopper)

Assembly: iqSchAmer2.1 (GCF_021461395.2) • Chr 5 (NC_060123.1)

Search assembly: 5:241908242

NC_060123.1: 241,567,197 - 242,011,940

Genes

Name	Location
LOC124615460	Chr5: 241,637,518 - 241,925,384
LOC124621927	Chr7: 511,718,579 - 511,765,719
LOC124594532	Chr2: 899,209,777 - 899,266,970
LOC124550820	Chr9: 201,848,184 - 201,965,183

Gene: LOC124615460 Transcript: XM_047143365.1

Exons: click an exon to zoom in, mouse over to see details

Genes, NCBI Schistocerca americana Annotation Release 100, 2022-03-18

Genes, Ensembl Rapid release 2022-10

RNA-seq exon coverage, aggregate (filtered), NCBI Schistocerca americana Annotation Release 100 - log base 2 scaled

RNA-seq intron-spanning reads, aggregate (filtered), NCBI Schistocerca americana Annotation Release 100 - log base 2 scaled

NC_060123.1: 242M..242M (444,744 nt)

If you put your cursor on a line it will show a drop down menu with the information on the gene/mRNA/protein

On the green line (the zoomed out gene diagram) go down on the drop down menu to retrieve the GenBank file of the gene

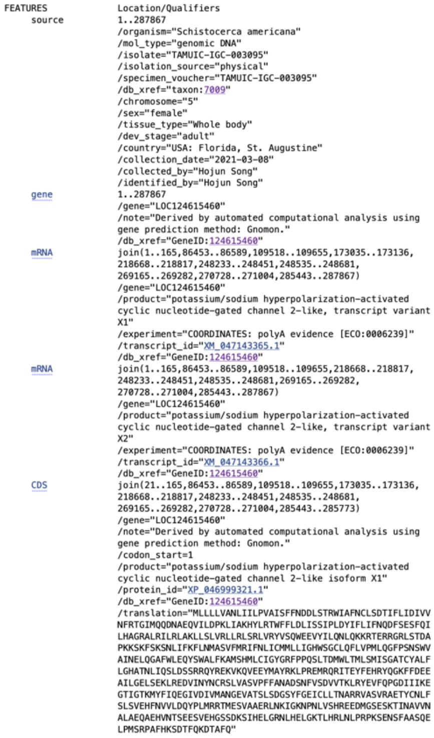
It is the last line in the drop down menu: GenBank Record: NC_060123.1

This is the GenBank file of the gene LOC124615460

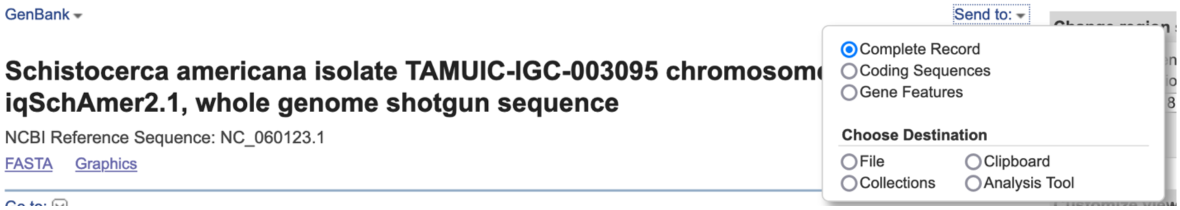
In this case this is one of 4 paralogs that encode a hyperpolarization cyclic nucleotide gated channel (this one is HCN2-like with the closest homology to *ih*)
Scroll down and you will find all the information on the Gene under FEATURES

- Source
- Gene
- mRNA
- CDS, this file automatically includes the full translation of the gene

When you click on mRNA it will show all the annotated exons of the gene in the sequence file to the right
To download this completely annotated gene file go to the top of the page to Send to and the drop down arrow
It will show a box shown below:



```
FEATURES             Location/Qualifiers
     source          1..287867
                     /organism="Schistocerca americana"
                     /mol_type="genomic DNA"
                     /isolate="TAMUIC-IGC-003095"
                     /isolation_source="physical"
                     /specimen_voucher="TAMUIC-IGC-003095"
                     /db_xref="taxon:7009"
                     /chromosome="5"
                     /sex="female"
                     /tissue_type="Whole body"
                     /dev_stage="adult"
                     /country="USA: Florida, St. Augustine"
                     /collection_date="2021-03-08"
                     /collected_by="Hojun Song"
                     /identified_by="Hojun Song"
     gene            1..287867
                     /gene="LOC124615460"
                     /note="Derived by automated computational analysis using
                     gene prediction method: Gnomon."
                     /db_xref="GeneID:124615460"
     mRNA            join(1..165,86453..86589,109518..109655,173035..173136,
                     218668..218817,248233..248451,248535..248681,
                     269165..269282,270728..271004,285443..287867)
                     /gene="LOC124615460"
                     /product="potassium/sodium hyperpolarization-activated
                     cyclic nucleotide-gated channel 2-like, transcript variant
                     X1"
                     /experiment="COORDINATES: polyA evidence [ECO:0006239]"
                     /transcript_id="XM_047143365.1"
                     /db_xref="GeneID:124615460"
     mRNA            join(1..165,86453..86589,109518..109655,218668..218817,
                     248233..248451,248535..248681,269165..269282,
                     270728..271004,285443..287867)
                     /gene="LOC124615460"
                     /product="potassium/sodium hyperpolarization-activated
                     cyclic nucleotide-gated channel 2-like, transcript variant
                     X2"
                     /experiment="COORDINATES: polyA evidence [ECO:0006239]"
                     /transcript_id="XM_047143366.1"
                     /db_xref="GeneID:124615460"
     CDS             join(21..165,86453..86589,109518..109655,173035..173136,
                     218668..218817,248233..248451,248535..248681,
                     269165..269282,270728..271004,285443..285773)
                     /gene="LOC124615460"
                     /note="Derived by automated computational analysis using
                     gene prediction method: Gnomon."
                     /codon_start=1
                     /product="potassium/sodium hyperpolarization-activated
                     cyclic nucleotide-gated channel 2-like isoform X1"
                     /protein_id="XP_046999321.1"
                     /db_xref="GeneID:124615460"
                     /translation="MLLLLVANLIILPVAISFFNDDLSTRWIAFNCLSDTIFLIDIVV
                     NFRTGIMQQDNAEQVILDPKLIAKHYLRTWFFLDLISSIPLDYIFLIFNQDFSESFQI
                     LHAGRALRILRLAKLLSLVRLLRLSRLVRYVSQWEEVYILQNLQKKRTERRGRLSTDA
                     PKKSKFSKSNLIFKFLNMASVFMRIFNLICMMLLIGHWSGCLQFLVPMLQGFPSNSWV
                     AINELQGAFWLEQYSWALFKAMSHMLCIGYGRFPPQSLTDMWLTMLSMISGATCYALF
                     LGHATNLIQSLDSSRRQYREKVKQVEEYMAYRKLPREMRQRITEYFEHRYQGKFFDEE
                     AILGELSEKLREDVINYNCRSLVASVPFFANADSNFVSDVVTKLRYEVFQPGDIIIKE
                     GTIGTKMYFIQEGIVDIVMANGEVATSLSDGSYFGEICLLTNARRVASVRAETYCNLF
                     SLSVEHFNVVLDQYPLMRRTMESVAAERLNKIGKNPNLVSHREEDMGSESKTINAVVN
                     ALAEQAEHVNTSEESVEHGSSDKSIHELGRNLHELGKTLHRLNLPRPKSENSFAASQE
                     LPMSRPAFHKSDTFQKDTAFQ"
```



GenBank ▾

Schistocerca americana isolate TAMUIC-IGC-003095 chromosome iqSchAmer2.1, whole genome shotgun sequence

NCBI Reference Sequence: NC_060123.1

FASTA Graphics

Send to: ▾

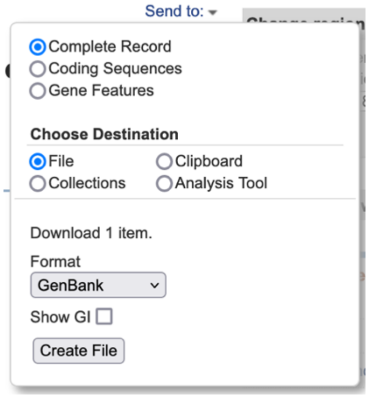
● Complete Record
○ Coding Sequences
○ Gene Features

Choose Destination

○ File ○ Clipboard
○ Collections ○ Analysis Tool

Click on File, this will further drop down the menu and ask for a format

Default is GenBank
Click on Create file
This creates your annotated GenBank file
When you open this file in SnapGene, you have a fully annotated Gene File



Send to: ▾

● Complete Record
○ Coding Sequences
○ Gene Features

Choose Destination

● File ○ Clipboard
○ Collections ○ Analysis Tool

Download 1 item.

Format
GenBank
Show GI □
Create File

It is very useful to also download the mRNA and Protein and make a Snap Gene file of all the splice variants

Store all the labeled files in separate folders per gene in an Annotation Folder

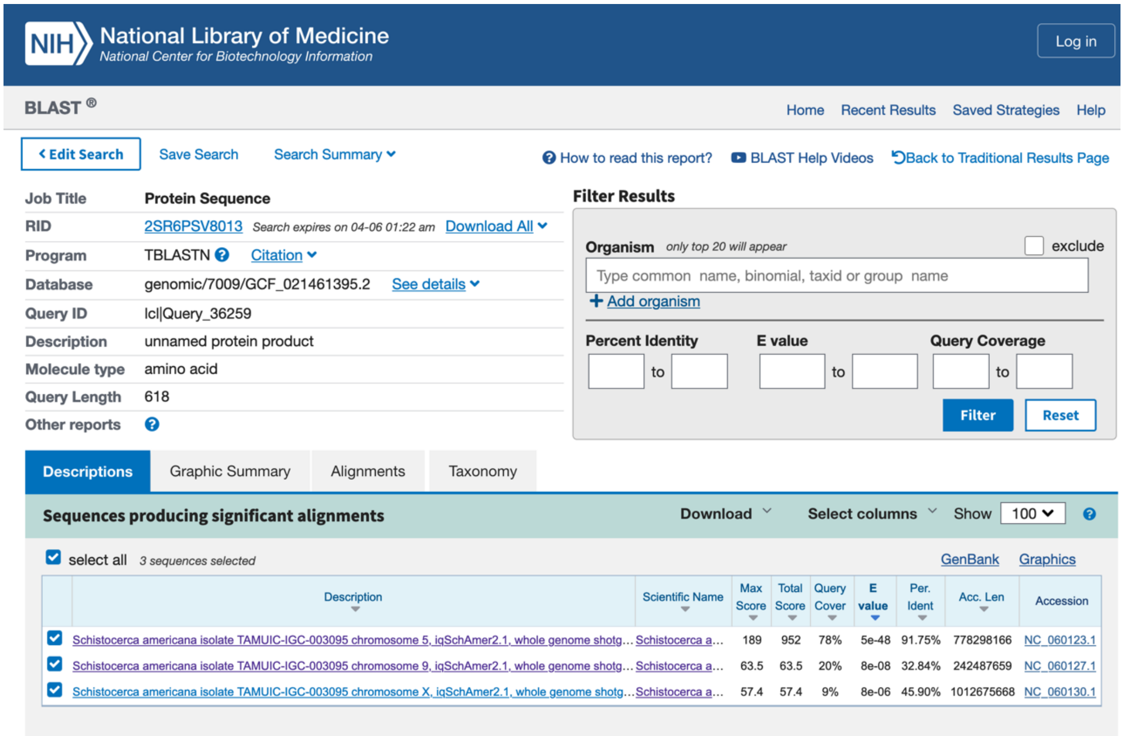
These files will be useful to use to Blast the same gene in different *Schistocerca* species and to perform sequence alignments or build trees

*When using a *Schistocerca* CDS sequence to blast in another *Schistocerca* species genome to find the ortholog, make sure you use the best setting for the homology search.

<u>Third method</u>: *Drosophila* ortholog protein BLAST search
Instead of retrieving the CDS of the fly ortholog in Flybase retrieve the protein sequence

Go to FlyBase
Type in the gene name in the Flybase Search window
This brings up a page with gene hits or takes you directly to the gene page
In the gene page, go to the 2nd box labeled Genome Location
Under sequence you can click on the drop down menu with Gene Region
Select Translations and then click get sequence
This brings you to the page of the protein sequence
Copy the complete protein sequence

Next, go back to the Genome Data Viewer of *Schistocerca* to perform your Blast search with the fly protein sequence
Go to the Blast button below the chromosome diagram window and click on tools
Click New BLAST
The default BLAST is a blastn for nucleotide sequence
Switch to Tblastn before pasting the sequence
Paste the *Drosophila* protein sequence in the window
Click BLAST, the page will refresh until the search is finished
This method returned 3 hits, the top one is the same one as the CDS Blast search



Job Title	Protein Sequence
RID	2SR6PSV8013 Search expires on 04-06 01:22 am Download All
Program	TBLASTN Citation
Database	genomic/7009/GCF_021461395.2 See details
Query ID	lcl|Query_36259
Description	unnamed protein product
Molecule type	amino acid
Query Length	618
Other reports	

Sequences producing significant alignments

☑ select all 3 sequences selected

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Schistocerca americana isolate TAMUIC-IGC-003095 chromosome 5, iqSchAmer2.1, whole genome shotg...	Schistocerca a...	189	952	78%	5e-48	91.75%	778298166	NC_060123.1
☑	Schistocerca americana isolate TAMUIC-IGC-003095 chromosome 9, iqSchAmer2.1, whole genome shotg...	Schistocerca a...	63.5	63.5	20%	8e-08	32.84%	242487659	NC_060127.1
☑	Schistocerca americana isolate TAMUIC-IGC-003095 chromosome X, iqSchAmer2.1, whole genome shotg...	Schistocerca a...	57.4	57.4	9%	8e-06	45.90%	1012675668	NC_060130.1

There are 8 matches in the first hit with sequence identities of ~95%
The other 2 hits only have a single stretch that matches with much lower sequence identity (30-45%)
To complete the retrieval of any of the hits, proceed as in method 2 to complete the process.

Perform sequence alignments

What do you need?

1. Gene visualization/manipulation software to look at/manipulate retrieved annotated Gene file

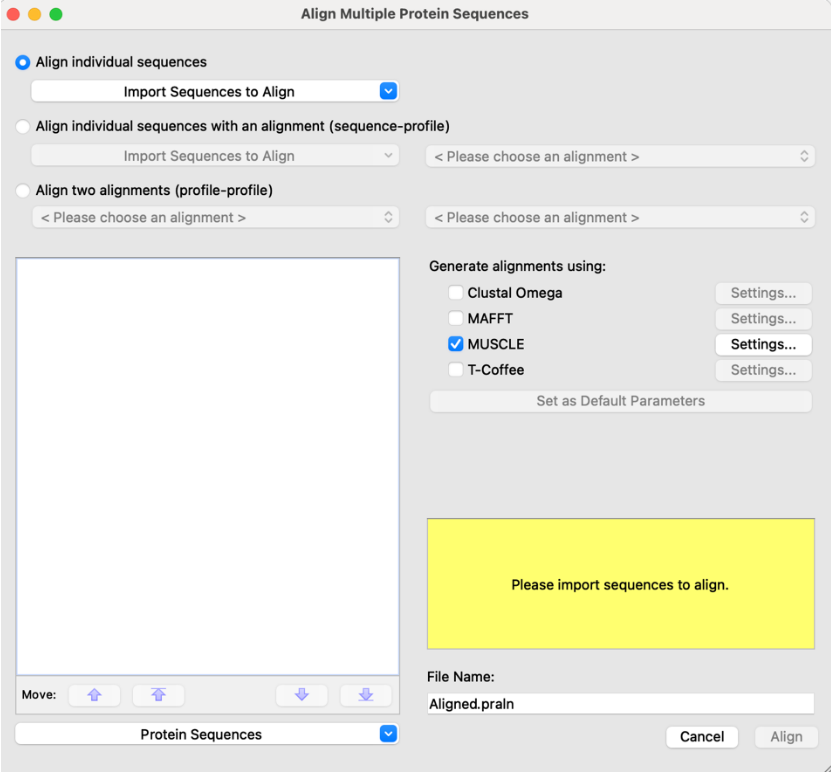
   SnapGene which has a free version called SnapGene Viewer
   https://www.snapgene.com/snapgene-viewer

   Benchling
   https://www.benchling.com/molecular-biology

2. Clustal Omega online access
   https://www.ebi.ac.uk/Tools/msa/clustalo/

SnapGene alignments

Go to SnapGene
Download all your protein sequences that you wish to align
Make a folder to add all your protein sequence files with the names of the species and protein
Go to Tools
In the drop down menu go to Align Sequences
When aligning more than 2 sequences, choose Align Multiple Protein Sequences
This opens a box I which you can import all your sequences to be aligned


Align Multiple Protein Sequences
Import Sequences to Align
Import Sequences to Align
< Please choose an alignment >
< Please choose an alignment >
< Please choose an alignment >
Settings...
Settings...
Settings...
Settings...
Set as Default Parameters
Move:
Protein Sequences
Cancel
Align

● Align individual sequences
○ Align individual sequences with an alignment (sequence-profile)
○ Align two alignments (profile-profile)

Generate alignments using:
□ Clustal Omega
□ MAFFT
☑ MUSCLE
□ T-Coffee

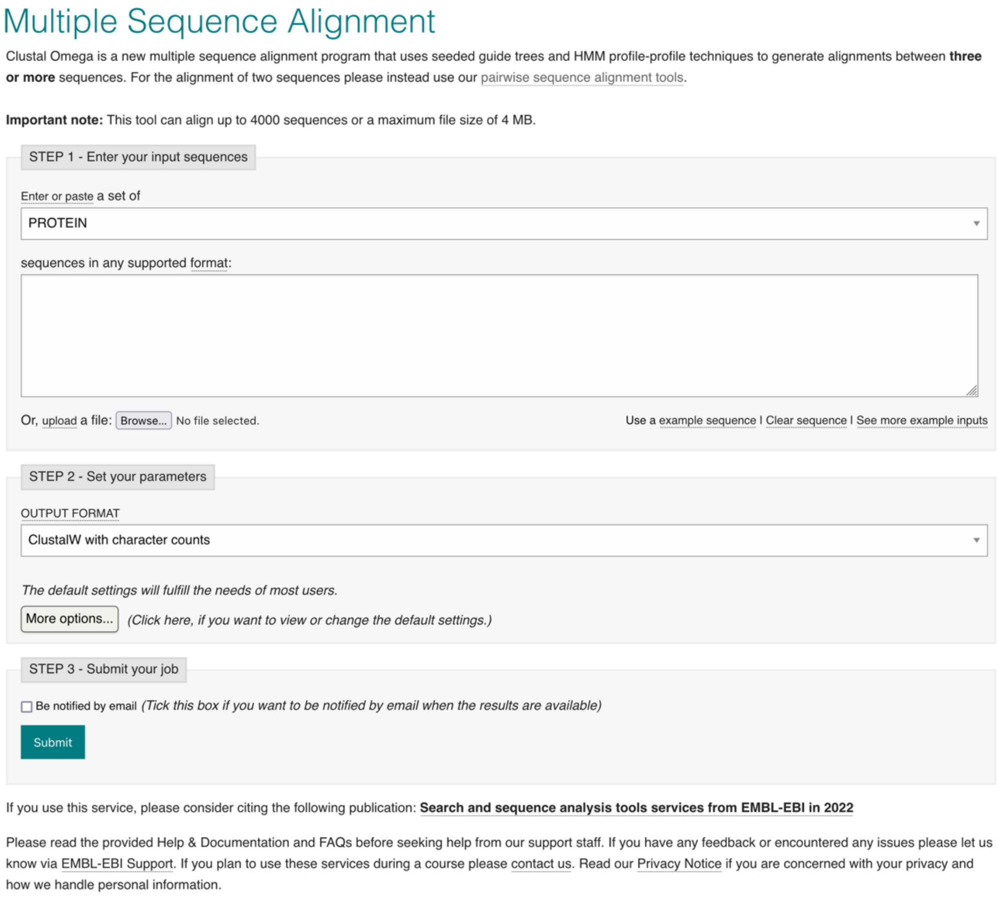
Please import sequences to align.

File Name: Aligned.praln

Import all the sequences from your folder
Select Clustal Omega and press align
This will show you an alignment of all the seqeuces
If you wish to omit a sequence, unclick it in the list of protein sequences
If you wish to change the output order, then use the up and down keys to move the sequences in the preferred order

Clustal Omega web alignments

Go to the Clustal Omega Website
Paste all your sequences you wish to align in the box
The easiest way to do this is to make a Word file with all your sequences
Make sure to start each sequence with >followed by a name and then paste the sequence on the next line
Repeat for all the sequences you wish to align



Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of
PROTEIN

sequences in any supported format:

Or, upload a file: Browse... No file selected.

Use a example sequence | Clear sequence | See more example inputs

STEP 2 - Set your parameters

OUTPUT FORMAT
ClustalW with character counts

The default settings will fulfill the needs of most users.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 3 - Submit your job

□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

If you use this service, please consider citing the following publication: **Search and sequence analysis tools services from EMBL-EBI in 2022**

Please read the provided Help & Documentation and FAQs before seeking help from our support staff. If you have any feedback or encountered any issues please let us know via EMBL-EBI Support. If you plan to use these services during a course please contact us. Read our Privacy Notice if you are concerned with your privacy and how we handle personal information.

Use default parameters
Use default output format
When you submit, you will have the option to receive the results by email or just wait until the alignment is done
You can download the alignment file
You can also make a phylogenetic tree and download that file