Ria Lamba

Bioinformatics Project Draft

Identification of Spider Latrotoxins Using Diamond and NCBI-BLAST

Video Tutorial: https://youtu.be/dEykHyQTfZ0

Datasets Used in Tutorial:

https://drive.google.com/drive/folders/1XrOYhCSlVW7oy5RVuTRS6YywRgAHPh1O?usp=sharing

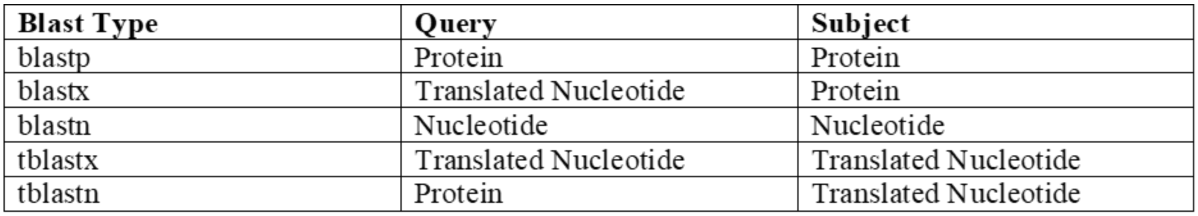
Ria Lamba

Bioinformatics Project Draft

Background

In this tutorial, we will be using Diamond and NCBI-BLAST on the command line to help us identify potential latrotoxin sequences in the Mediterranean Black Widow. We will be running these jobs on the HPCC UCR cluster, but the commands can be used on other servers. Diamond and NCBI-BLAST are both programs that run various BLAST searches. While both programs give the same BLAST output, they each have different search algorithms and capabilities, which is why one may need to be used over the other for specific situations. While Diamond is faster than NCBI-BLAST, it can only do two types of BLAST searches: BLASTx and BLASTp, whereas NCBI-BLAST can also do BLASTn, tBLASTx and tBLASTn. Both programs can be used to generate BLAST databases and BLAST outputs in different formats.

BLAST stands for Basic Local Alignment Search Tool (“BLAST: Basic Local Alignment Search Tool.”, n.d.). It is used to align and compare nucleotide or protein sequences to one another and to sequence databases. There are different types of blast searches that can be done for query sequences versus subject sequences.



Blast Type	**Query**	**Subject**
blastp	Protein	Protein
blastx	Translated Nucleotide	Protein
blastn	Nucleotide	Nucleotide
tblastx	Translated Nucleotide	Translated Nucleotide
tblastn	Protein	Translated Nucleotide

BLAST helps scientists get insight on sequence similarities that can be used to understand relationships between sequences among species and to identify gene families. Gene families are groups of similar genes formed by gene duplication. While one can do BLAST searches on the NCBI website, it is not an efficient way to BLAST multiple sequences, especially if you have a file with hundreds or thousands of sequences. Therefore, it is important to learn to run BLAST searches on the command line. If you connect to a high-performance computing cluster, you will have access to programs without having to download them on your own computer. You can also run bigger jobs with ease by submitting scripts with your commands, that then get sent to a queuing system to be processed. This tutorial shows how to submit BLAST jobs on a computing cluster using BASH scripts.

Dataset

Latrotoxins are a type of neurotoxin found in spiders of the genus *Latrodectus*. These spiders are incredibly poisonous and a bite from them can cause the illness latrodectism if enough venom is injected into the victim (Bonnet, 2004). The Mediterranean Black Widow (*Latrodectus tredecimguttatus*) is one of these toxic spiders, commonly found throughout the Mediterranean region as the name suggests. In this tutorial, we will perform BLAST and Diamond searches for latrotoxin-like sequences in the Mediterranean Black Widow, using the spider’s venom gland RNA sequences downloaded from NCBI’s Sequence Set Browser (bio sample SAMN02318955). We will also be using a FASTA file containing the first 320 amino acids of several previously described latrotoxins. This FASTA file was made by compiling the sequences of several latrotoxins collected and found on NCBI, followed by trimming after the first 320 amino acids. This FASTA file will be used to find potential latrotoxins in the Mediterranean Black Widow.

Ria Lamba

Bioinformatics Project Draft

First, we will do a BLASTx search using Diamond with the venom gland RNA sequences as the query against nr, the non-redundant protein database from NCBI. The BLASTx results will identify sequences that resemble the venom RNA sequences. Second, we will use NCBI-BLAST makeblastdb to make a BLASTable database out of the venom gland RNA sequences that will allow us to specifically search for latrotoxins; we will run a tBLASTn search with NCBI-BLAST using our FASTA file containing the first 320 amino acids of known latrotoxins as a query against these venom gland RNA sequences. The tBLASTn search will be repeated with the output format 0, a format that gives a visual of the alignments between each query and subject. The BLASTx and tBLASTn results give us different information that can be useful for identifying sequences of interest. The BLASTx results include protein sequences from nr with descriptive titles, that will help give us context as to what each venom gland sequence's function and role may be. The tBLASTn results show us the venom gland sequences that match to known latrotoxin sequences. We then will know which venom gland sequences are likely to be toxins.

Ria Lamba

Bioinformatics Project Draft

Instruction Manual Table of Contents

Instruction Manual

First open a terminal. In this tutorial, we will be using MobaXTerm on a Windows computer. However, you can access the cluster using any terminal you want. To log on to the cluster, type the server name into the search bar (for example, cluster.hpcc.ucr.edu). You will then be prompted to type in your username and then your password.

If you are using another terminal such as git-bash, type the following to log in:

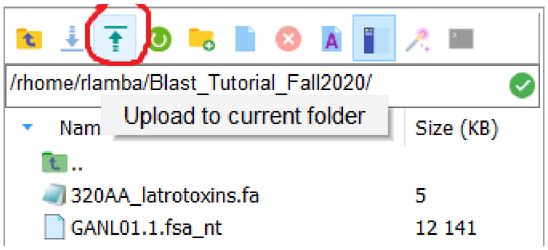
```
ssh -X rlamba@cluster.hpcc.ucr.edu
```

If it is your first time logging on to the cluster, you will be prompted to enter your password.

Make a directory for this tutorial and **cd** into it:

```
mkdir Blast_Tutorial_Fall2020
cd Blast_Tutorial_Fall2020
```

To upload files from your computer to the directory on the cluster, move over to the graphical interface on the left and click upload to current folder. Then select the files on your computer that you would like to upload.



/rhome/rlamba/Blast_Tutorial_Fall2020/

Nam	Size (KB)
..	
320AA_latrotoxins.fa	5
GANL01.1.fsa_nt	12 141

Alternatively, transfer files using the command line using **scp**:

```
scp -r <path_to_directory-in-pc> <username>@<host_name>:<path-to-file-on-cluster>
```

scp stands for “secure copy”. In <path_to_directory> type the path that the file you want to upload is located at in your computer. Then add your username on the cluster in <username>. Enter the name of the cluster you are connected to in <host_name>. In <path-to-file-in-cluster> type the path in the cluster that you want to upload the file to.

BLASTx Search With Diamond

On the cluster we are using, bioinformatics software we need are already installed and can be accessed via modules. To load DIAMOND on the cluster, type the following:

```
module load diamond
```

Next, we want to open a file to write our script, we will use the text editor **nano**:

```
nano tredecimguttatus_diamond_blastx.sh
```

Your script may vary depending on the cluster and queuing system you are using. Here is an example script:

```
#!/bin/bash -l

#SBATCH --nodes=1
#SBATCH --ntasks=6
#SBATCH --mem-per-cpu=6G
#SBATCH --time=1-00:00:00
#SBATCH --output=/rhome/rlamba/Blast_Tutorial_Fall2020/tredecimguttatus
_DIAMOND_BLASTx_runlog.out

module load diamond

diamond blastx \
-q /rhome/rlamba/Blast_Tutorial_Fall2020/GANL01.1.fsa_nt \
-d /rhome/rlamba/shared/DIAMOND_v0.9.24_nr_db_Feb2020.dmnd \
-o /rhome/rlamba/Blast_Tutorial_Fall2020/tredecimguttatus_nr_database.tsv \
-k 3 -f 6 qseqid sseqid evalue qframe qlen slen qstart qend sstart send length pident stitle
```

The first line in the script is called the sha-bang line. It is crucial that every script starts with this line, otherwise it cannot run. It is the path to the Bash Interpreter.

The next five lines that start with #SBATCH specify what resources are being requested for the job. In this case, we are asking for 1 node, 6 GB of CPU, and 1 day to be allocated for this job. This does not necessarily mean the job will take up all 6 GB or 1 whole day. It just means this is the maximum that will be given for the job. If you request too little resources, your job will immediately stop running once it runs out of resources. If you request a large amount of resources, your job may remain pending in the queue for a while. This is why one must be reasonable when requesting resources for a job.

Diamond Parameters Used In Script:

-q	Path to the file we are using for the query
-d	Path to the database we are blasting our query against (the subject)
-o	Path to where we want our output file to be when created + the name we want for the file.
-k	This parameter is optional. It specifies how many matches you want for each query sequence. If you do not use the -k parameter, you will get the default number of matches (25).

-f	This is the output format line, where you specify what output format you would like (in this case, we have 6 to specify output format 6) and what information you want in your output.
qseqid	Query sequence id
sseqid	Subject sequence id
evalue	E-value (Value that gives an idea of how good the match is. The lower the e-value, the better the match is.)
qframe	The frame that the query is translated in.
qlen	How long the alignment is for the query sequences (will be in bp since query is nucleotide sequences)
slen	How long the alignment is for the subject sequence (Will be in amino acids since the subject is protein sequences)
qstart	When the alignment starts for the query sequences.
qend	When the alignment ends for the query sequence.
sstart	When the alignment starts for the subject sequences.
send	When the alignment ends for the subject sequence.
length	Length of the alignment
pident	Percent of identical matches
stitle	The titles of the subject sequences. Note that these will be more descriptive when doing blast searches against databases like nr, which has been put together by ncbi.

Once you have finished your script, submit it to the queuing system:

```
sbatch tredecimguttatus_diamond_blastx.sh
```

Your job should be finished running in a few hours. You can check the status of jobs you have submitted by typing:

```
squeue -u rlamba
```

When you use **squeue -u** to check your job status, you should see something like this if it is running correctly:

```
JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
          2537826     intel tredecim   rlamba  R    2:40:12      1 i23
```

If you do not see any jobs running, that means that your job was submitted but there is an issue that caused it to not run. Check the *runlog.out* file to see what may have gone wrong and check your script for errors such as extra spaces.

Making a BLAST Database With NCBI-BLAST

To load NCBI-BLAST, type:

```
module load NCBI-BLAST/2.9.0+
```

To make a database of the *tredecimguttatus* sequences, type the following:

```
makeblastdb -dbtype nucl -in /rhome/rlamba/Blast_Tutorial_Fall2020/GANL01.1.fsa_nt -out
/rhome/rlamba/Blast_Tutorial_Fall2020/Tredecimguttatus_database_nucdb
```

Makeblastdb Parameters Used In Script:

dbtype	What type of database are you making. In this case, we type nucl for nucleotide. If we were making a protein database, we would type prot.
-in	Stands for input. Path to the file you want to make a database out of.
-out	Stands for output. Path to where you want the database to be when it is created + the name you want for your database.

The database should be done in approximately a second. It is a very fast process. Note that when you check your output files with **ls** you will see three files:

-Tredecimguttatus_database_nucdb.nsq
-Tredecimguttatus_database_nucdb.nin
-Tredecimguttatus_database_nucdb.nhr

These are all components of the database. To do a BLAST search against the database, simply refer to the database as *Tredecimguttatus_database_nucdb*.

tBLASTn With NCBI-BLAST

Open another text file using **nano**:

```
nano Tredecimguttatus_Latrotoxin_tblastn.sh
```

Here is an example script:

```
#!/bin/bash -l

#SBATCH --nodes=1
#SBATCH --ntasks=6
#SBATCH --mem-per-cpu=6G
#SBATCH --time=1-00:00:00
#SBATCH --output=/rhome/rlamba/Blast_Tutorial_Fall2020/tredecimguttatus
_Latrotoxin_tblastn_runlog.out


module load NCBI-BLAST/2.9.0+
tblastn \
-query /rhome/rlamba/Blast_Tutorial_Fall2020/320AA_latrotoxins.fa \
-out /rhome/rlamba/Blast_Tutorial_Fall2020/tredecimguttatus _Latrotoxin_tblastn.tsv \
-db /rhome/rlamba/Blast_Tutorial_Fall2020/Tredecimguttatus_database_nucdb \
-outfmt '0 qseqid sseqid evalue qframe qlen slen qstart qend sstart send length pident stitle' \
-max_target_seqs 5
```

NCBI-BLAST Parameters Used In Script:

-query	Path to the file we are using for the query
-out	Path to where we want our output file to be when created + the name we want for the file.
-db	Path to the database we are blasting our query against (the subject)
-max_target_seqs	This parameter is optional. It specifies how many matches you want for each query sequence. If you do not use the -max_target_seqs parameter, you will get the default number of matches (25).
-outfmt	This is the output format line, where you specify what output format you would like (in this case, we have 6 to specify output format 6) and what information you want in your output.
qseqid	Query sequence id
sseqid	Subject sequence id
evalue	E-value (Value that gives an idea of how good the match is. The lower the e-value, the better the match is.)
qframe	The frame that the query is translated in.
qlen	How long the alignment is for the query sequences (will be in amino acids since query is protein sequences)
slen	How long the alignment is for the subject sequence (Will be in bp since the subject is nucleotide sequences)
qstart	When the alignment starts for the query sequences.
qend	When the alignment ends for the query sequence.
sstart	When the alignment starts for the subject sequences.
send	When the alignment ends for the subject sequence.
length	Length of the alignment
pident	Percent of identical matches
stitle	The titles of the subject sequences. Note that these will be more descriptive when doing blast searches against databases like nr, which has been put together by ncbi.

tBLASTn In Output Format 0:

The BLASTx search and tBLASTn search we did previously were both in output format 6. We can easily change the output format. Different blast output formats show us the same information in different ways, giving us different insights. We will be repeating our last tBLASTn search, only this time we will ask the program to give us output in format 0 instead of format 6. While format 6 is a table with all the terms specified in the *-outfmt* line as headers, format 0 gives the same information specified, but also gives a visual of the alignments between each query and subject. See pages 13 and 14 to see screenshots of these two output formats.

Open another text file using **nano**:

```
nano Tredecimguttatus_Latrotoxin_tblastn.sh
```

Here is an example script:

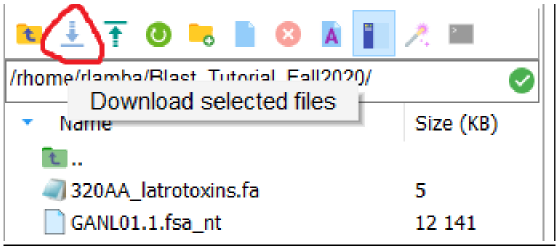
```
#!/bin/bash -l

#SBATCH --nodes=1
#SBATCH --ntasks=6
#SBATCH --mem-per-cpu=6G
#SBATCH --time=1-00:00:00
#SBATCH --output=/rhome/rlamba/Blast_Tutorial_Fall2020/tredecimguttatus_Latrotoxin_F0_tblastn_runlog.out
module load NCBI-BLAST/2.9.0+
tblastn \
-query /rhome/rlamba/Blast_Tutorial_Fall2020/320AA_latrotoxins.fa \
-out /rhome/rlamba/Blast_Tutorial_Fall2020/tredecimguttatus_Latrotoxin_F0_tblastn.tsv \
-db /rhome/rlamba/Blast_Tutorial_Fall2020/Tredecimguttatus_database_nucdb \
-outfmt '0 qseqid sseqid evalue qframe qlen slen qstart qend sstart send length pident stitle' \
-max_target_seqs 5
```

Notice that the script is the same as the script for our tBLASTn in format 6, only at the *-outfmt* line, there is a 0 in place of the 6. An F0 has also been added to the *runlog.out* file and the output file names in order to be able to distinguish them from the files for the previous tBLASTn search that was in format 6.

Downloading and Viewing Output:

If you are using MobaXTerm, you can download files from the cluster to your computer by clicking download selected files (blue downward arrow at top left).



/rhome/rlamba/Blast_Tutorial_Fall2020/

Name	Size (KB)
..	
320AA_latrotoxins.fa	5
GANL01.1.fsa_nt	12 141

You can also transfer files on the command line by typing:

```
scp -r <path-to-file-in-cluster> <username>@<host_name>:<path_to_directory-in-pc>
```

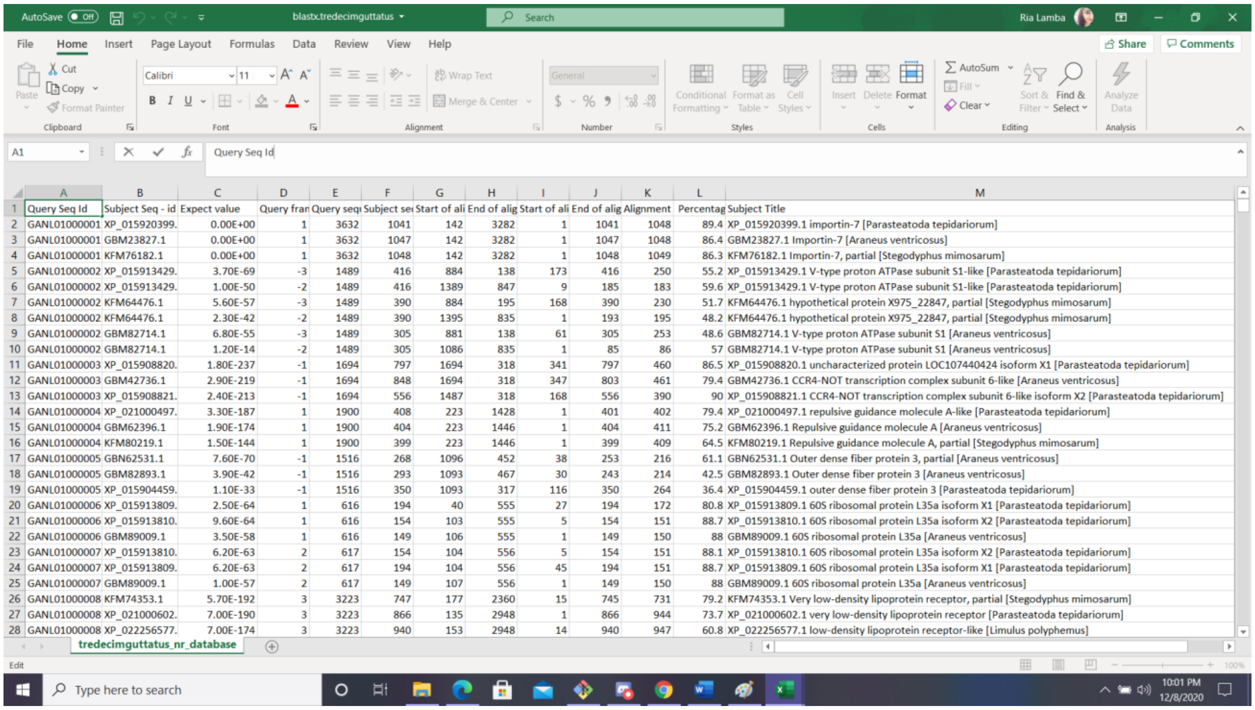
In <path-to-file-in-cluster> type the path that the file you want to download is located at in the cluster. Then add your username on the cluster in <username>. Enter the name of the cluster you are connected to in <host_name>. In <path_to_directory-in-pc> type the path to the directory you want the file to be downloaded to on your pc.

The output for these BLAST searches that are in format 6 are best viewed and analyzed in excel. You will notice several columns with various information. The order of the columns is the same as the order of the terms specified in the *-f/-outfmt* line. The output for the tblastn search in format 0 is best viewed and analyzed using notepad++ or a similar text editor. You will notice information for each query and subject match along with a visual of the sequence alignment between the two.

In this case we are interested in locating latrotoxins from these BLAST results.

-Open your BLASTx output with Excel.

Diamond BLASTx Output:



Query Seq Id

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Query Seq Id	Subject Seq - id	Expect value	Query fran	Query seq	Subject se	Start of ali	End of alig	Start of ali	End of alig	Alignment	Percentag	Subject Title
2	GANL01000001	XP_015920399.	0.00E+00	1	3632	1041	142	3282	1	1041	1048	89.4	XP_015920399.1 importin-7 [Parasteatoda tepidariorum]
3	GANL01000001	GBM23827.1	0.00E+00	1	3632	1047	142	3282	1	1047	1048	86.4	GBM23827.1 Importin-7 [Araneus ventricosus]
4	GANL01000001	KFM76182.1	0.00E+00	1	3632	1048	142	3282	1	1048	1049	86.3	KFM76182.1 Importin-7, partial [Stegodyphus mimosarum]
5	GANL01000002	XP_015913429.	3.70E-69	-3	1489	416	884	138	173	416	250	55.2	XP_015913429.1 V-type proton ATPase subunit S1-like [Parasteatoda tepidariorum]
6	GANL01000002	XP_015913429.	1.00E-50	-2	1489	416	1389	847	9	185	183	59.6	XP_015913429.1 V-type proton ATPase subunit S1-like [Parasteatoda tepidariorum]
7	GANL01000002	KFM64476.1	5.60E-57	-3	1489	390	884	195	168	390	230	51.7	KFM64476.1 hypothetical protein X975_22847, partial [Stegodyphus mimosarum]
8	GANL01000002	KFM64476.1	2.30E-42	-2	1489	390	1395	835	1	193	195	48.2	KFM64476.1 hypothetical protein X975_22847, partial [Stegodyphus mimosarum]
9	GANL01000002	GBM82714.1	6.80E-55	-3	1489	305	881	138	61	305	253	48.6	GBM82714.1 V-type proton ATPase subunit S1 [Araneus ventricosus]
10	GANL01000002	GBM82714.1	1.20E-14	-2	1489	305	1086	835	1	85	86	57	GBM82714.1 V-type proton ATPase subunit S1 [Araneus ventricosus]
11	GANL01000003	XP_015908820.	1.80E-237	-1	1694	797	1694	318	341	797	460	86.5	XP_015908820.1 uncharacterized protein LOC107440424 isoform X1 [Parasteatoda tepidariorum]
12	GANL01000003	GBM42736.1	2.90E-219	-1	1694	848	1694	318	347	803	461	79.4	GBM42736.1 CCR4-NOT transcription complex subunit 6-like [Araneus ventricosus]
13	GANL01000003	XP_015908821.	2.40E-213	-1	1694	556	1487	318	168	556	390	90	XP_015908821.1 CCR4-NOT transcription complex subunit 6-like isoform X2 [Parasteatoda tepidariorum]
14	GANL01000004	XP_021000497.	3.30E-187	1	1900	408	223	1428	1	401	402	79.4	XP_021000497.1 repulsive guidance molecule A-like [Parasteatoda tepidariorum]
15	GANL01000004	GBM62396.1	1.90E-174	1	1900	404	223	1446	1	404	411	75.2	GBM62396.1 Repulsive guidance molecule A [Araneus ventricosus]
16	GANL01000004	KFM80219.1	1.50E-144	1	1900	399	223	1446	1	399	409	64.5	KFM80219.1 Repulsive guidance molecule A, partial [Stegodyphus mimosarum]
17	GANL01000005	GBN62531.1	7.60E-70	-1	1516	268	1096	452	38	253	216	61.1	GBN62531.1 Outer dense fiber protein 3, partial [Araneus ventricosus]
18	GANL01000005	GBM82893.1	3.90E-42	-1	1516	293	1093	467	30	243	214	42.5	GBM82893.1 Outer dense fiber protein 3 [Araneus ventricosus]
19	GANL01000005	XP_015904459.	1.10E-33	-1	1516	350	1093	317	116	350	264	36.4	XP_015904459.1 outer dense fiber protein 3 [Parasteatoda tepidariorum]
20	GANL01000006	XP_015913809.	2.50E-64	1	616	194	40	555	27	194	172	80.8	XP_015913809.1 60S ribosomal protein L35a isoform X1 [Parasteatoda tepidariorum]
21	GANL01000006	XP_015913810.	9.60E-64	1	616	154	103	555	5	154	151	88.7	XP_015913810.1 60S ribosomal protein L35a isoform X2 [Parasteatoda tepidariorum]
22	GANL01000006	GBM89009.1	3.50E-58	1	616	149	106	555	1	149	150	88	GBM89009.1 60S ribosomal protein L35a [Araneus ventricosus]
23	GANL01000007	XP_015913810.	6.20E-63	2	617	154	104	556	5	154	151	88.1	XP_015913810.1 60S ribosomal protein L35a isoform X2 [Parasteatoda tepidariorum]
24	GANL01000007	XP_015913809.	6.20E-63	2	617	194	104	556	45	194	151	88.7	XP_015913809.1 60S ribosomal protein L35a isoform X1 [Parasteatoda tepidariorum]
25	GANL01000007	GBM89009.1	1.00E-57	2	617	149	107	556	1	149	150	88	GBM89009.1 60S ribosomal protein L35a [Araneus ventricosus]
26	GANL01000008	KFM74353.1	5.70E-192	3	3223	747	177	2360	15	745	731	79.2	KFM74353.1 Very low-density lipoprotein receptor, partial [Stegodyphus mimosarum]
27	GANL01000008	XP_021000602.	7.00E-190	3	3223	866	135	2948	1	866	944	73.7	XP_021000602.1 very low-density lipoprotein receptor [Parasteatoda tepidariorum]
28	GANL01000008	XP_022256577.	7.00E-174	3	3223	940	153	2948	14	940	947	60.8	XP_022256577.1 low-density lipoprotein receptor-like [Limulus polyphemus]

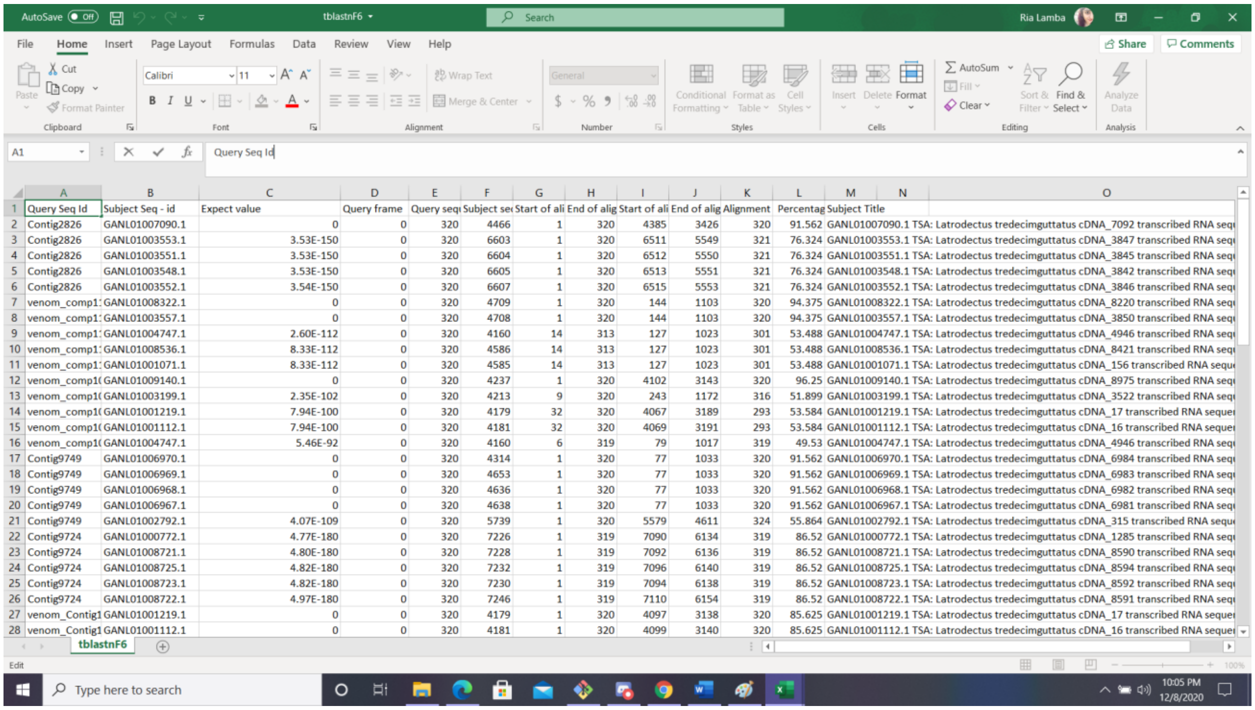
-Click find and select and type “latrotoxin”

-Click find all

Excel will then show you all the subject titles that contain the word "latrotoxin". You can then see which of the query sequences in the venom gland matched to sequences in NCBI that are labeled as latrotoxins. This can give you an idea of which sequences in the venom gland may be latrotoxins that you should look at more closely.

-Open your tBLASTn format 6 output with Excel

NCBI-BLAST tBLASTn Output In Format 6:



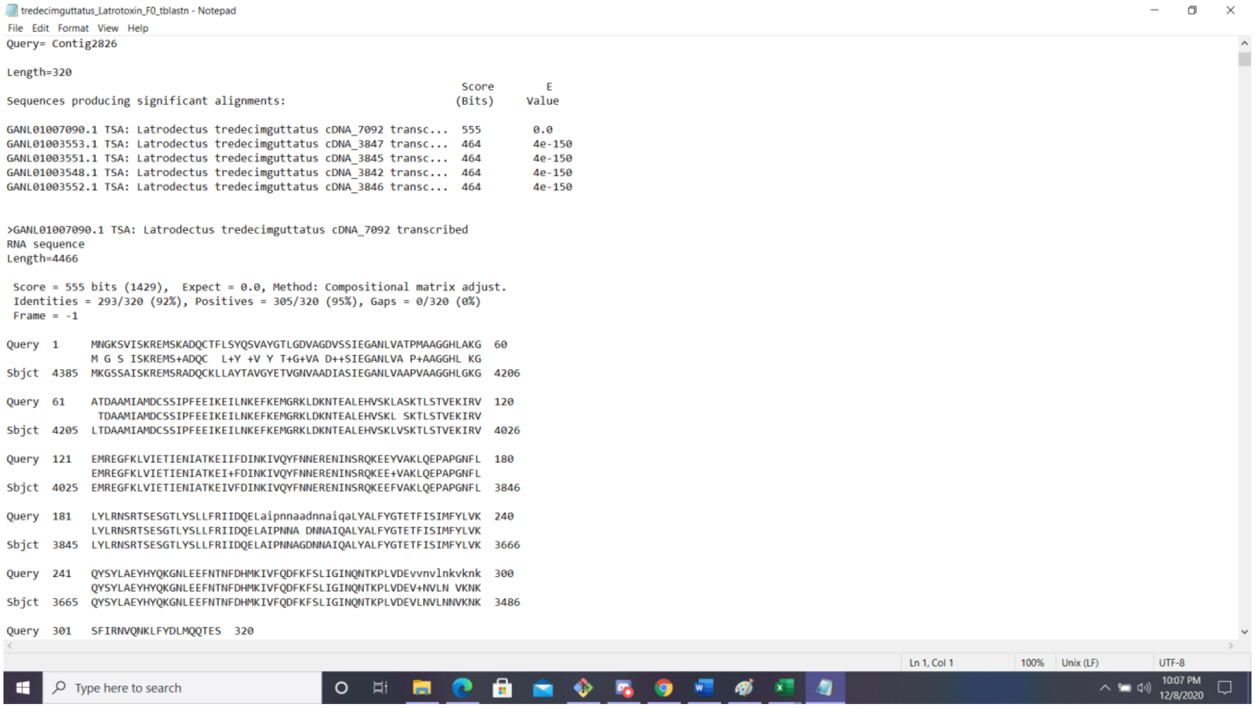
Query Seq Id

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Query Seq Id	Subject Seq - id	Expect value	Query frame	Query seq	Subject se	Start of ali	End of alig	Start of ali	End of alig	Alignment	Percentag	Subject Title		
2	Contig2826	GANL01007090.1	0	0	320	4466	1	320	4385	3426	320	91.562	GANL01007090.1 TSA: Latrodectus tredecimguttatus cDNA_7092 transcribed RNA seq		
3	Contig2826	GANL01003553.1	3.53E-150	0	320	6603	1	320	6511	5549	321	76.324	GANL01003553.1 TSA: Latrodectus tredecimguttatus cDNA_3847 transcribed RNA seq		
4	Contig2826	GANL01003551.1	3.53E-150	0	320	6604	1	320	6512	5550	321	76.324	GANL01003551.1 TSA: Latrodectus tredecimguttatus cDNA_3845 transcribed RNA seq		
5	Contig2826	GANL01003548.1	3.53E-150	0	320	6605	1	320	6513	5551	321	76.324	GANL01003548.1 TSA: Latrodectus tredecimguttatus cDNA_3842 transcribed RNA seq		
6	Contig2826	GANL01003552.1	3.54E-150	0	320	6607	1	320	6515	5553	321	76.324	GANL01003552.1 TSA: Latrodectus tredecimguttatus cDNA_3846 transcribed RNA seq		
7	venom_comp1	GANL01008322.1	0	0	320	4709	1	320	144	1103	320	94.375	GANL01008322.1 TSA: Latrodectus tredecimguttatus cDNA_8220 transcribed RNA seq		
8	venom_comp1	GANL01003557.1	0	0	320	4708	1	320	144	1103	320	94.375	GANL01003557.1 TSA: Latrodectus tredecimguttatus cDNA_3850 transcribed RNA seq		
9	venom_comp1	GANL01004747.1	2.60E-112	0	320	4160	14	313	127	1023	301	53.488	GANL01004747.1 TSA: Latrodectus tredecimguttatus cDNA_4946 transcribed RNA seq		
10	venom_comp1	GANL01008536.1	8.33E-112	0	320	4586	14	313	127	1023	301	53.488	GANL01008536.1 TSA: Latrodectus tredecimguttatus cDNA_8421 transcribed RNA seq		
11	venom_comp1	GANL01001071.1	8.33E-112	0	320	4585	14	313	127	1023	301	53.488	GANL01001071.1 TSA: Latrodectus tredecimguttatus cDNA_156 transcribed RNA seque		
12	venom_comp1(	GANL01009140.1	0	0	320	4237	1	320	4102	3143	320	96.25	GANL01009140.1 TSA: Latrodectus tredecimguttatus cDNA_8975 transcribed RNA seq		
13	venom_comp1(	GANL01003199.1	2.35E-102	0	320	4213	9	320	243	1172	316	51.899	GANL01003199.1 TSA: Latrodectus tredecimguttatus cDNA_3522 transcribed RNA seq		
14	venom_comp1(	GANL01001219.1	7.94E-100	0	320	4179	32	320	4067	3189	293	53.584	GANL01001219.1 TSA: Latrodectus tredecimguttatus cDNA_17 transcribed RNA sequer		
15	venom_comp1(	GANL01001112.1	7.94E-100	0	320	4181	32	320	4069	3191	293	53.584	GANL01001112.1 TSA: Latrodectus tredecimguttatus cDNA_16 transcribed RNA sequer		
16	venom_comp1(	GANL01004747.1	5.46E-92	0	320	4160	6	319	79	1017	319	49.53	GANL01004747.1 TSA: Latrodectus tredecimguttatus cDNA_4946 transcribed RNA seq		
17	Contig9749	GANL01006970.1	0	0	320	4314	1	320	77	1033	320	91.562	GANL01006970.1 TSA: Latrodectus tredecimguttatus cDNA_6984 transcribed RNA seq		
18	Contig9749	GANL01006969.1	0	0	320	4653	1	320	77	1033	320	91.562	GANL01006969.1 TSA: Latrodectus tredecimguttatus cDNA_6983 transcribed RNA seq		
19	Contig9749	GANL01006968.1	0	0	320	4636	1	320	77	1033	320	91.562	GANL01006968.1 TSA: Latrodectus tredecimguttatus cDNA_6982 transcribed RNA seq		
20	Contig9749	GANL01006967.1	0	0	320	4638	1	320	77	1033	320	91.562	GANL01006967.1 TSA: Latrodectus tredecimguttatus cDNA_6981 transcribed RNA seq		
21	Contig9749	GANL01002792.1	4.07E-109	0	320	5739	1	320	5579	4611	324	55.864	GANL01002792.1 TSA: Latrodectus tredecimguttatus cDNA_315 transcribed RNA seque		
22	Contig9724	GANL01000772.1	4.77E-180	0	320	7226	1	319	7090	6134	319	86.52	GANL01000772.1 TSA: Latrodectus tredecimguttatus cDNA_1285 transcribed RNA seq		
23	Contig9724	GANL01008721.1	4.80E-180	0	320	7228	1	319	7092	6136	319	86.52	GANL01008721.1 TSA: Latrodectus tredecimguttatus cDNA_8590 transcribed RNA seq		
24	Contig9724	GANL01008725.1	4.82E-180	0	320	7232	1	319	7096	6140	319	86.52	GANL01008725.1 TSA: Latrodectus tredecimguttatus cDNA_8594 transcribed RNA seq		
25	Contig9724	GANL01008723.1	4.82E-180	0	320	7230	1	319	7094	6138	319	86.52	GANL01008723.1 TSA: Latrodectus tredecimguttatus cDNA_8592 transcribed RNA seq		
26	Contig9724	GANL01008722.1	4.97E-180	0	320	7246	1	319	7110	6154	319	86.52	GANL01008722.1 TSA: Latrodectus tredecimguttatus cDNA_8591 transcribed RNA seq		
27	venom_Contig1	GANL01001219.1	0	0	320	4179	1	320	4097	3138	320	85.625	GANL01001219.1 TSA: Latrodectus tredecimguttatus cDNA_17 transcribed RNA sequer		
28	venom_Contig1	GANL01001112.1	0	0	320	4181	1	320	4099	3140	320	85.625	GANL01001112.1 TSA: Latrodectus tredecimguttatus cDNA_16 transcribed RNA sequer		

You can now see which sequences in the venom gland matched with our query file of latrotoxins. You can also take a look at the evalue to help you judge how good of a match there is between a certain query and subject. This also can give some insight into which sequences in the venom gland are latrotoxins.

If you want to view the alignments between the query and subject sequences, you can open and view your tBLASTn results in format 0.

NCBI-BLAST tBLASTn Output In Format 0:


tredecimguttatus_Latrotoxin_F0_tblastn - Notepad
File Edit Format View Help
Ln 1, Col 1
100%
Unix (LF)
UTF-8
Type here to search
10:07 PM
12/8/2020

```
Query= Contig2826

Length=320
                                                                      Score        E
Sequences producing significant alignments:                          (Bits)     Value

GANL01007090.1 TSA: Latrodectus tredecimguttatus cDNA_7092 transc...  555        0.0
GANL01003553.1 TSA: Latrodectus tredecimguttatus cDNA_3847 transc...  464        4e-150
GANL01003551.1 TSA: Latrodectus tredecimguttatus cDNA_3845 transc...  464        4e-150
GANL01003548.1 TSA: Latrodectus tredecimguttatus cDNA_3842 transc...  464        4e-150
GANL01003552.1 TSA: Latrodectus tredecimguttatus cDNA_3846 transc...  464        4e-150


>GANL01007090.1 TSA: Latrodectus tredecimguttatus cDNA_7092 transcribed
RNA sequence
Length=4466

 Score = 555 bits (1429),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 293/320 (92%), Positives = 305/320 (95%), Gaps = 0/320 (0%)
 Frame = -1

Query  1     MNGKSVISKREMSKADQCTFLSYQSVAYGTLGDVAGDVSSIEGANLVATPMAAGGHLAKG  60
             M G S ISKREMS+ADQC  L+Y +V Y T+G+VA D++SIEGANLVA P+AAGGHL KG
Sbjct  4385  MKGSSAISKREMSRADQCKLLAYTAVGYETVGNVAADIASIEGANLVAAPVAAGGHLGKG  4206

Query  61    ATDAAMIAMDCSSIPFEEIKEILNKEFKEMGRKLDKNTEALEHVSKLASKTLSTVEKIRV  120
              TDAAMIAMDCSSIPFEEIKEILNKEFKEMGRKLDKNTEALEHVSKL SKTLSTVEKIRV
Sbjct  4205  LTDAAMIAMDCSSIPFEEIKEILNKEFKEMGRKLDKNTEALEHVSKLVSKTLSTVEKIRV  4026

Query  121   EMREGFKLVIETIENIATKEIIFDINKIVQYFNNERENINSRQKEEYVAKLQEPAPGNFL  180
             EMREGFKLVIETIENIATKEI+FDINKIVQYFNNERENINSRQKEE+VAKLQEPAPGNFL
Sbjct  4025  EMREGFKLVIETIENIATKEIVFDINKIVQYFNNERENINSRQKEEFVAKLQEPAPGNFL  3846

Query  181   LYLRNSRTSESGTLYSLLFRIIDQELaipnnaadnnaiqaLYALFYGTETFISIMFYLVK  240
             LYLRNSRTSESGTLYSLLFRIIDQELAIPNNA DNNAIQALYALFYGTETFISIMFYLVK
Sbjct  3845  LYLRNSRTSESGTLYSLLFRIIDQELAIPNNAGDNNAIQALYALFYGTETFISIMFYLVK  3666

Query  241   QYSYLAEYHYQKGNLEEFNTNFDHMKIVFQDFKFSLIGINQNTKPLVDEvvnvlnkvknk  300
             QYSYLAEYHYQKGNLEEFNTNFDHMKIVFQDFKFSLIGINQNTKPLVDEV+NVLN VKNK
Sbjct  3665  QYSYLAEYHYQKGNLEEFNTNFDHMKIVFQDFKFSLIGINQNTKPLVDEVLNVLNNVKNK  3486

Query  301   SFIRNVQNKLFYDLMQQTES  320
```

grep Command to Pull Specific Rows:

Another trick you can use to find matches of interest in your BLAST output is to use the **grep** command while you are still logged on to the cluster. In this example, we will use the command to find any lines in our output file with the word “toxin” and output them to a new file:

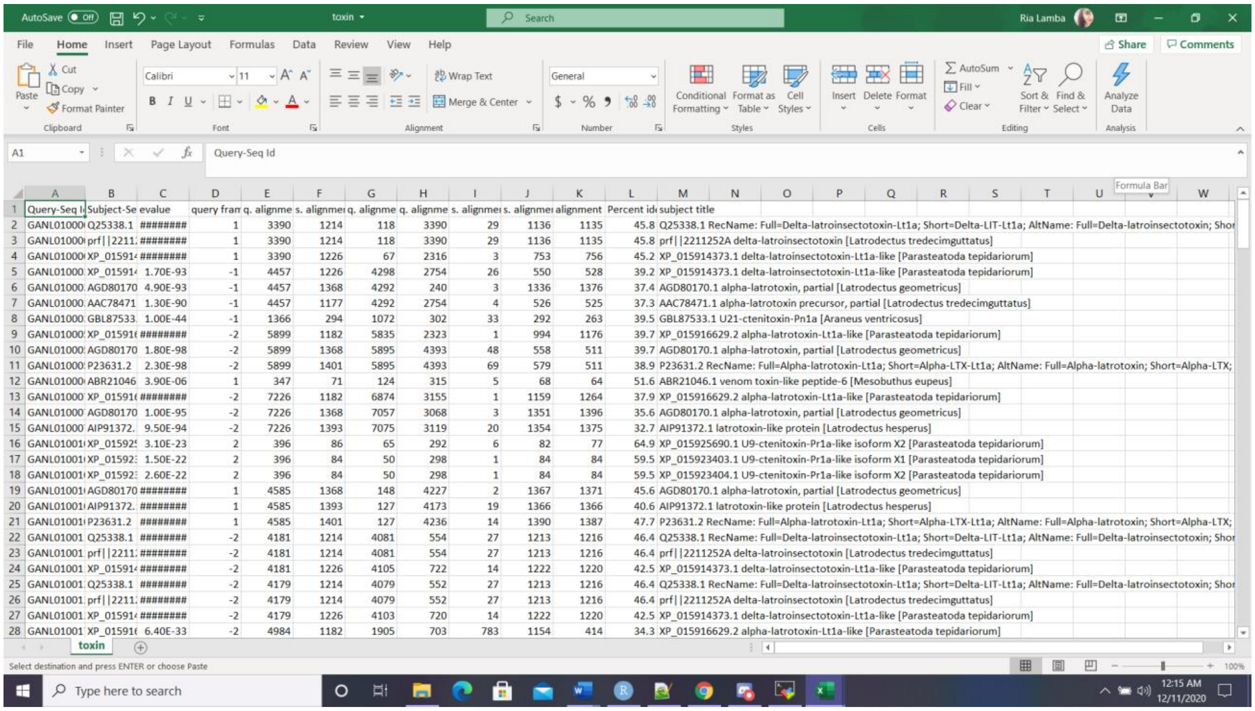
```
fgrep toxin tredecimguttatus_nr_database.tsv > toxin.tsv
```

The output file (toxin.tsv) will only contain rows that mention the word “toxin” in the line.

Ria Lamba

Bioinformatics Project Draft

Grep Output File:



Query-Seq Id

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Query-Seq I	Subject-Se	evalue	query fram	q. alignme	s. alignme	q. alignme	q. alignme	s. alignme	s. alignme	alignment	Percent id	subject title
2	GANL01000(	Q25338.1	########	1	3390	1214	118	3390	29	1136	1135	45.8	Q25338.1 RecName: Full=Delta-latroinsectotoxin-Lt1a; Short=Delta-LIT-Lt1a; AltName: Full=Delta-latroinsectotoxin; Shor
3	GANL01000(	prf||2211:	########	1	3390	1214	118	3390	29	1136	1135	45.8	prf||2211252A delta-latroinsectotoxin [Latrodectus tredecimguttatus]
4	GANL01000(	XP_01591‹	########	1	3390	1226	67	2316	3	753	756	45.2	XP_015914373.1 delta-latroinsectotoxin-Lt1a-like [Parasteatoda tepidariorum]
5	GANL01000:	XP_01591‹	1.70E-93	-1	4457	1226	4298	2754	26	550	528	39.2	XP_015914373.1 delta-latroinsectotoxin-Lt1a-like [Parasteatoda tepidariorum]
6	GANL01000:	AGD80170	4.90E-93	-1	4457	1368	4292	240	3	1336	1376	37.4	AGD80170.1 alpha-latrotoxin, partial [Latrodectus geometricus]
7	GANL01000:	AAC78471	1.30E-90	-1	4457	1177	4292	2754	4	526	525	37.3	AAC78471.1 alpha-latrotoxin precursor, partial [Latrodectus tredecimguttatus]
8	GANL01000:	GBL87533.	1.00E-44	-1	1366	294	1072	302	33	292	263	39.5	GBL87533.1 U21-ctenitoxin-Pn1a [Araneus ventricosus]
9	GANL01000:	XP_01591(	########	-2	5899	1182	5835	2323	1	994	1176	39.7	XP_015916629.2 alpha-latrotoxin-Lt1a-like [Parasteatoda tepidariorum]
10	GANL01000:	AGD80170	1.80E-98	-2	5899	1368	5895	4393	48	558	511	39.7	AGD80170.1 alpha-latrotoxin, partial [Latrodectus geometricus]
11	GANL01000:	P23631.2	2.30E-98	-2	5899	1401	5895	4393	69	579	511	38.9	P23631.2 RecName: Full=Alpha-latrotoxin-Lt1a; Short=Alpha-LTX-Lt1a; AltName: Full=Alpha-latrotoxin; Short=Alpha-LTX;
12	GANL01000(	ABR21046	3.90E-06	1	347	71	124	315	5	68	64	51.6	ABR21046.1 venom toxin-like peptide-6 [Mesobuthus eupeus]
13	GANL01000'	XP_01591(	########	-2	7226	1182	6874	3155	1	1159	1264	37.9	XP_015916629.2 alpha-latrotoxin-Lt1a-like [Parasteatoda tepidariorum]
14	GANL01000'	AGD80170	1.00E-95	-2	7226	1368	7057	3068	3	1351	1396	35.6	AGD80170.1 alpha-latrotoxin, partial [Latrodectus geometricus]
15	GANL01000'	AIP91372.	9.50E-94	-2	7226	1393	7075	3119	20	1354	1375	32.7	AIP91372.1 latrotoxin-like protein [Latrodectus hesperus]
16	GANL01001(	XP_01592!	3.10E-23	2	396	86	65	292	6	82	77	64.9	XP_015925690.1 U9-ctenitoxin-Pr1a-like isoform X2 [Parasteatoda tepidariorum]
17	GANL01001(	XP_01592:	1.50E-22	2	396	84	50	298	1	84	84	59.5	XP_015923403.1 U9-ctenitoxin-Pr1a-like isoform X1 [Parasteatoda tepidariorum]
18	GANL01001(	XP_01592:	2.60E-22	2	396	84	50	298	1	84	84	59.5	XP_015923404.1 U9-ctenitoxin-Pr1a-like isoform X2 [Parasteatoda tepidariorum]
19	GANL01001(	AGD80170	########	1	4585	1368	148	4227	2	1367	1371	45.6	AGD80170.1 alpha-latrotoxin, partial [Latrodectus geometricus]
20	GANL01001(	AIP91372.	########	1	4585	1393	127	4173	19	1366	1366	40.6	AIP91372.1 latrotoxin-like protein [Latrodectus hesperus]
21	GANL01001(	P23631.2	########	1	4585	1401	127	4236	14	1390	1387	47.7	P23631.2 RecName: Full=Alpha-latrotoxin-Lt1a; Short=Alpha-LTX-Lt1a; AltName: Full=Alpha-latrotoxin; Short=Alpha-LTX;
22	GANL01001	Q25338.1	########	-2	4181	1214	4081	554	27	1213	1216	46.4	Q25338.1 RecName: Full=Delta-latroinsectotoxin-Lt1a; Short=Delta-LIT-Lt1a; AltName: Full=Delta-latroinsectotoxin; Shor
23	GANL01001	prf||2211:	########	-2	4181	1214	4081	554	27	1213	1216	46.4	prf||2211252A delta-latroinsectotoxin [Latrodectus tredecimguttatus]
24	GANL01001	XP_01591‹	########	-2	4181	1226	4105	722	14	1222	1220	42.5	XP_015914373.1 delta-latroinsectotoxin-Lt1a-like [Parasteatoda tepidariorum]
25	GANL01001:	Q25338.1	########	-2	4179	1214	4079	552	27	1213	1216	46.4	Q25338.1 RecName: Full=Delta-latroinsectotoxin-Lt1a; Short=Delta-LIT-Lt1a; AltName: Full=Delta-latroinsectotoxin; Shor
26	GANL01001:	prf||2211:	########	-2	4179	1214	4079	552	27	1213	1216	46.4	prf||2211252A delta-latroinsectotoxin [Latrodectus tredecimguttatus]
27	GANL01001:	XP_01591‹	########	-2	4179	1226	4103	720	14	1222	1220	42.5	XP_015914373.1 delta-latroinsectotoxin-Lt1a-like [Parasteatoda tepidariorum]
28	GANL01001'	XP_01591(	6.40E-33	-2	4984	1182	1905	703	783	1154	414	34.3	XP_015916629.2 alpha-latrotoxin-Lt1a-like [Parasteatoda tepidariorum]

Conclusion:

The results of these different BLAST searches give a user varying information that can be put together to help form conclusions on the identity of sequences and how they are related to each other. The BLASTx search against nr shows us the protein sequences from the nr database that are most closely related to our venom gland sequences. The sequences in nr have descriptive titles that can help us identify the possible identity of our venom gland sequences. The tBLASTn search using the latrotoxin amino acids against the venom gland sequences shows us which venom gland sequences have similarity to latrotoxins. Both of these BLAST results can be compared and referenced to help understand the identity of the venom gland sequences as well as identify sequences of interest to look at more closely.

Works Cited

“BLAST: Basic Local Alignment Search Tool.” *National Center for Biotechnology Information*, U.S. National Library of Medicine, blast.ncbi.nlm.nih.gov/Blast.cgi.

Bonnet, M.S. “The Toxicology of Latrodectus Tredecimguttatus: the Mediterranean Black Widow Spider.” *Homeopathy*, No Longer Published by Elsevier, 9 Jan. 2004, www.sciencedirect.com/science/article/abs/pii/S1475491603001243.

“High-Performance Computing Center (HPCC).” *High-Performance Computing Center (HPCC) | HPCC @ UCR*, hpcc.ucr.edu/.

“Sequence Set Browser :: NCBI.” *National Center for Biotechnology Information*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/Traces/wgs/?view=TSA.

“Shebang.” *Shebang - Linux Shell Scripting Tutorial - A Beginner's Handbook*, bash.cyberciti.biz/guide/Shebang.