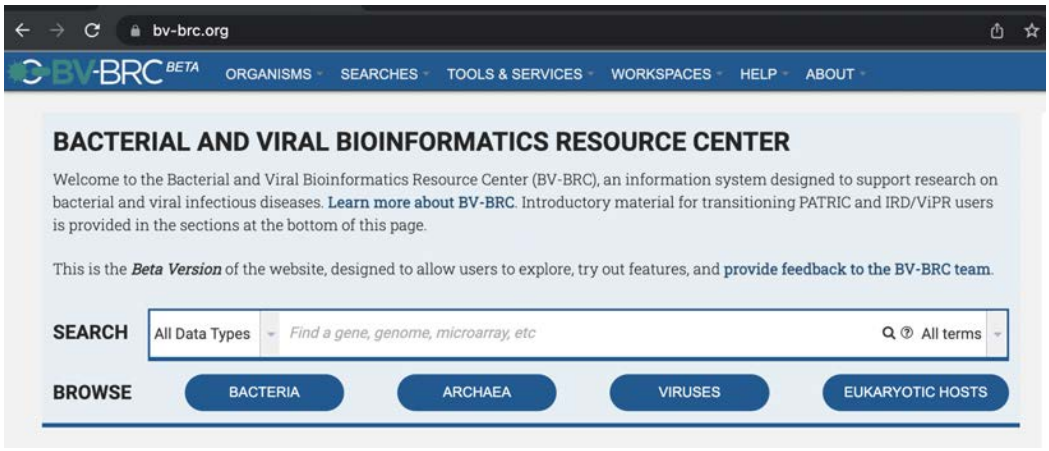
March 22, 2022

Tutorial for Tick Borne Pathogens Webinar: Bunyavirales and Ticks

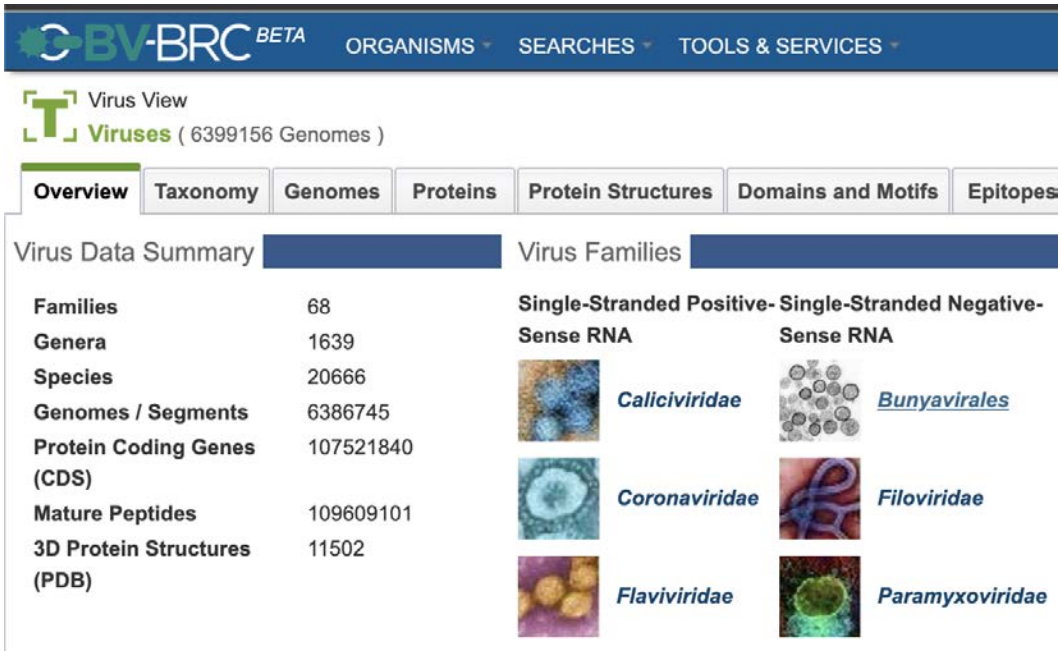
Use Case 1: A recent paper has shown that "A single mutation in Crimean-Congo hemorrhagic fever virus discovered in ticks impairs infectivity in human cells". In this manuscript, the authors show that "R1116G", a point mutation in the glycoprotein precursor complex (GPC) protein determines host tropism of CCHFV (human versus tick). For this use case, we will compare CCHFV GPC proteins from human and tick hosts using the following methods:

Step 1a) Search and Assemble dataset (Documentation: Genome/Protein Search)

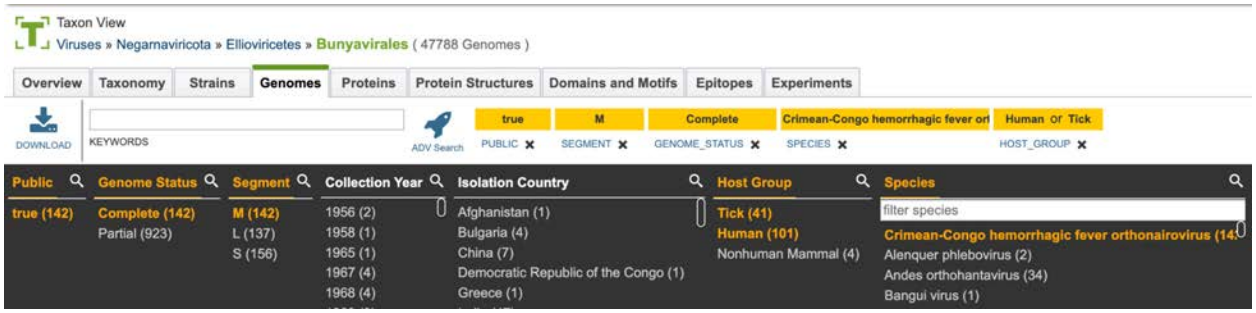
- Navigate to **https://www.bv-brc.org** and click on "Viruses"



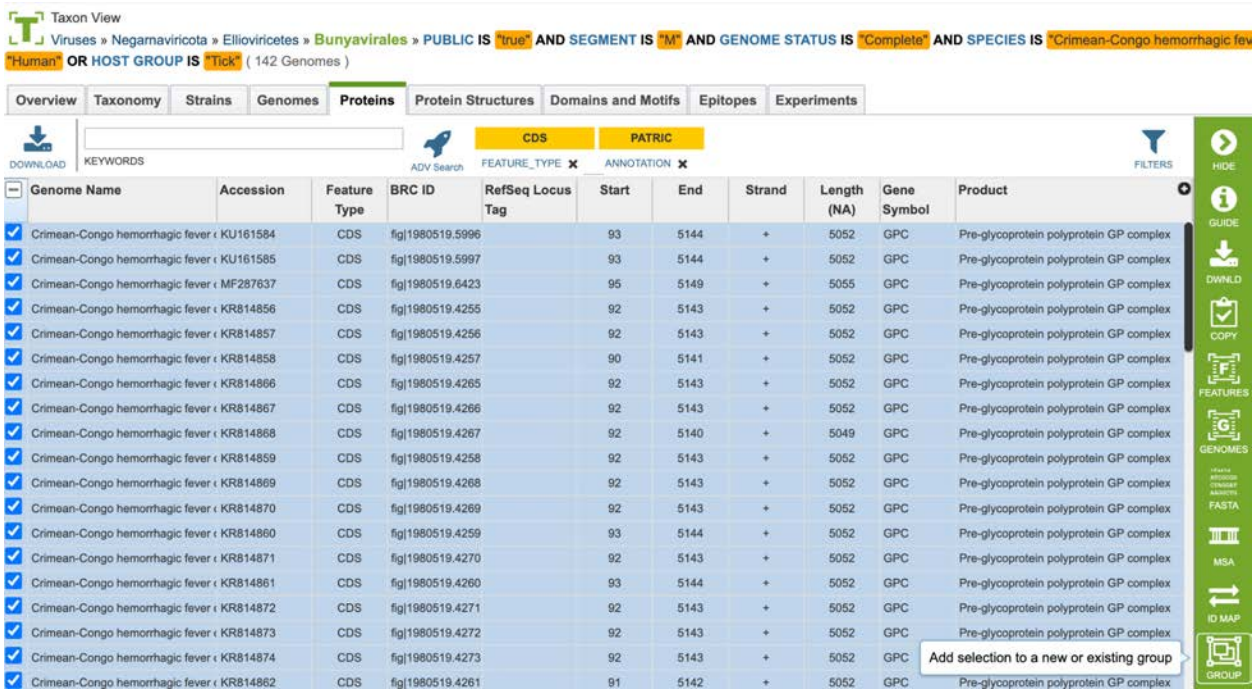
- This will take you to the Virus Overview Homepage. Now click on "Bunyavirales"



- Navigate to the genomes tab, use the “FILTER” button to select the following criteria and click “APPLY”:
  Public = True
  Genome Status = Complete
  Segment = M
  Host group =
  Species = Crimean-Congo Hemorrhagic Fever Virus



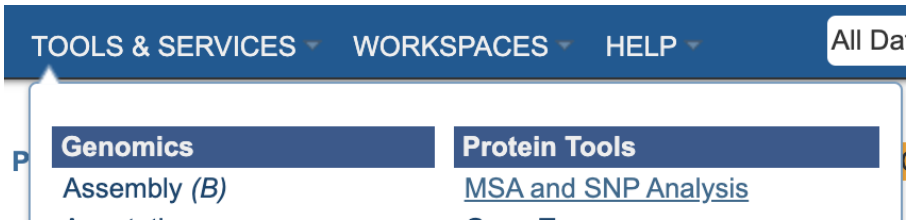
- After clicking “APPLY”, switch to the “Proteins” tab, select all GPC proteins, and select the “GROUP” option in the green action bar on the right, to create a “Protein Feature Group”.



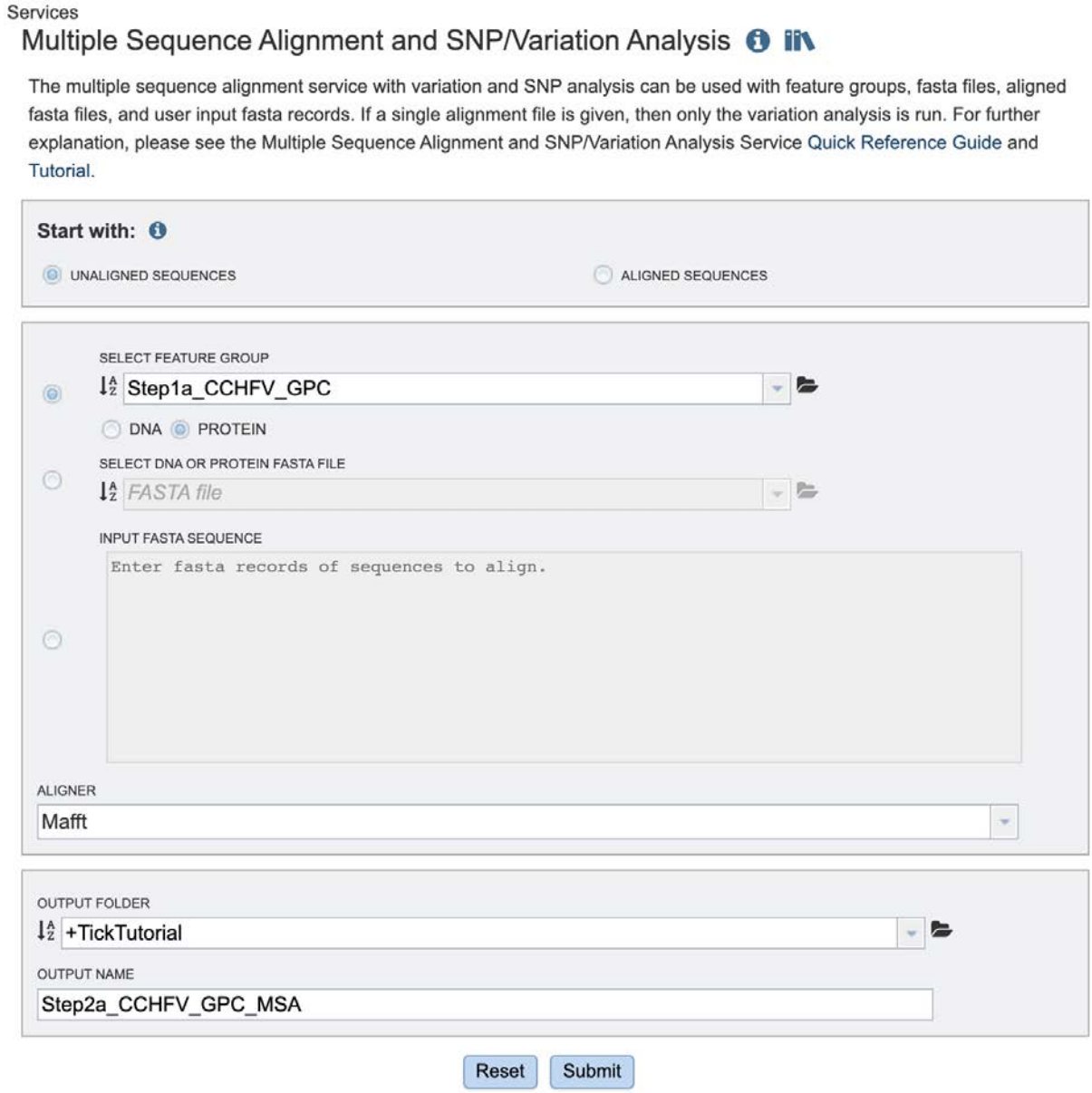
- You can also create a MSA directly from here using the MSA button in the green action bar on the right, making sure to select “Amino Acid”. *Note: not recommended for large data sets.*

Step 1b) Create MSA to verify and look for other genomes (MSA and SNP/Variation Analysis, Documentation)

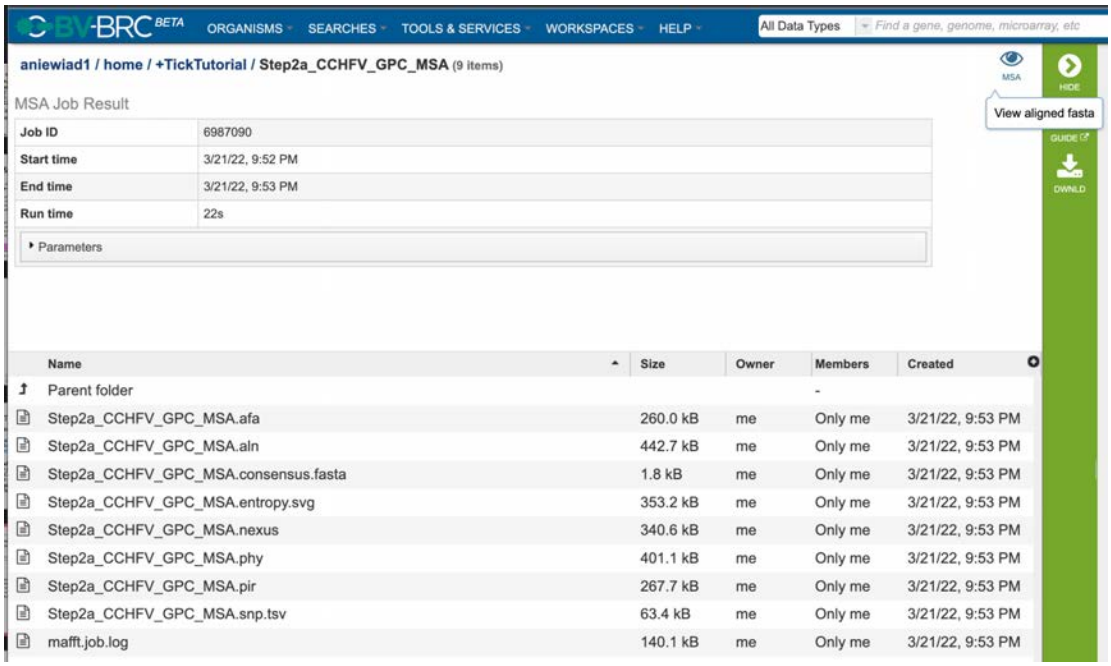
- To perform a multi-sequence alignment, navigate to the “TOOLS & SERVICES” tab, and select “MSA and SNP Analysis”.



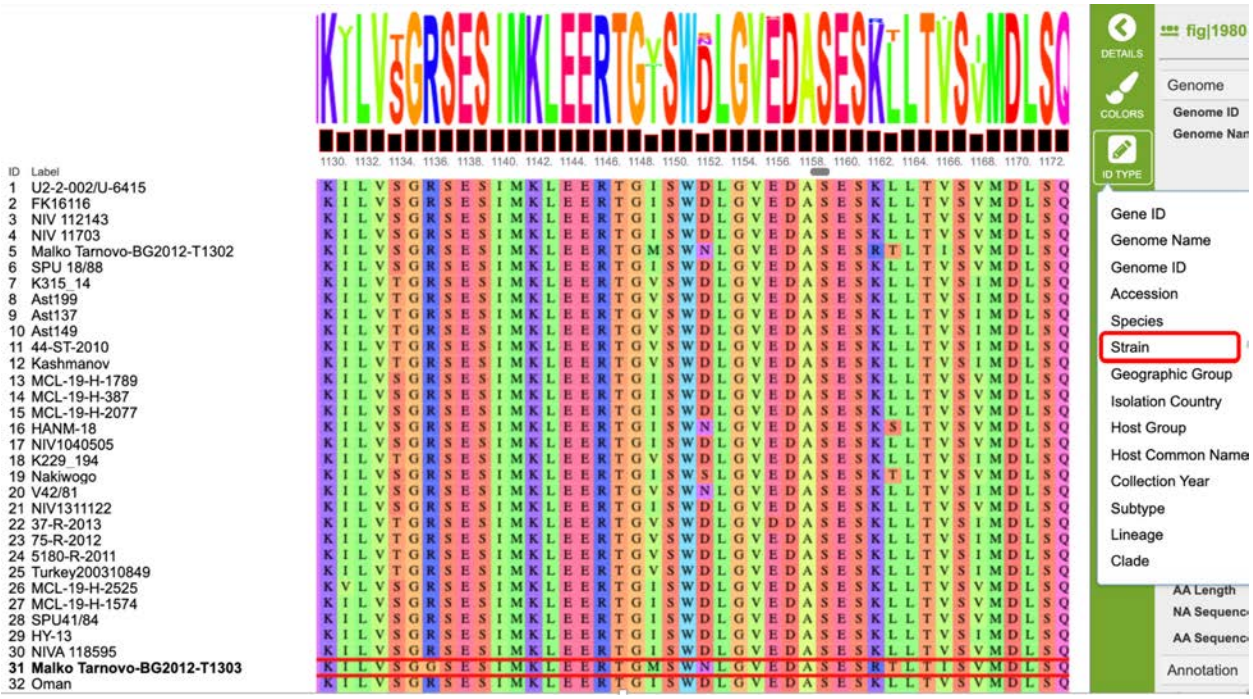
- Select the appropriate “Feature Group” saved in the previous step.
- Select “Unaligned Sequences”, “Protein” and “Mafft” options.
- Specify appropriate “Output Folder” and “Output Name”, and click submit.



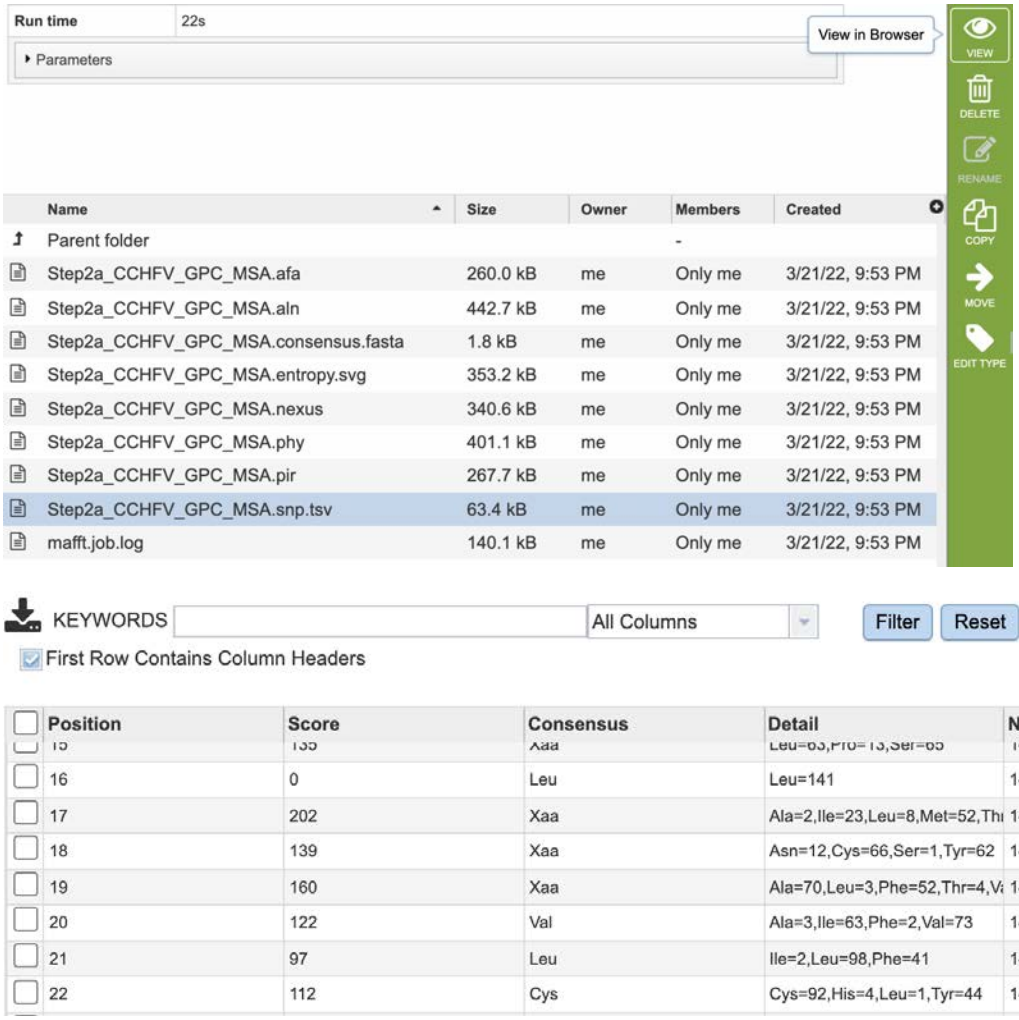
- Navigate to the job results and click on the appropriate job. Click the “eye” icon to view the aligned sequences.



- Scroll to the appropriate area “LVSGRSES” consensus region (~1134) to view the region of interest. Use the “ID TYPE” button shown below to select “Strain” name as the genome ID, and search for the “Malko Tarnovo-BG2012-T1303” name to find the genome reported in the publication referenced above.



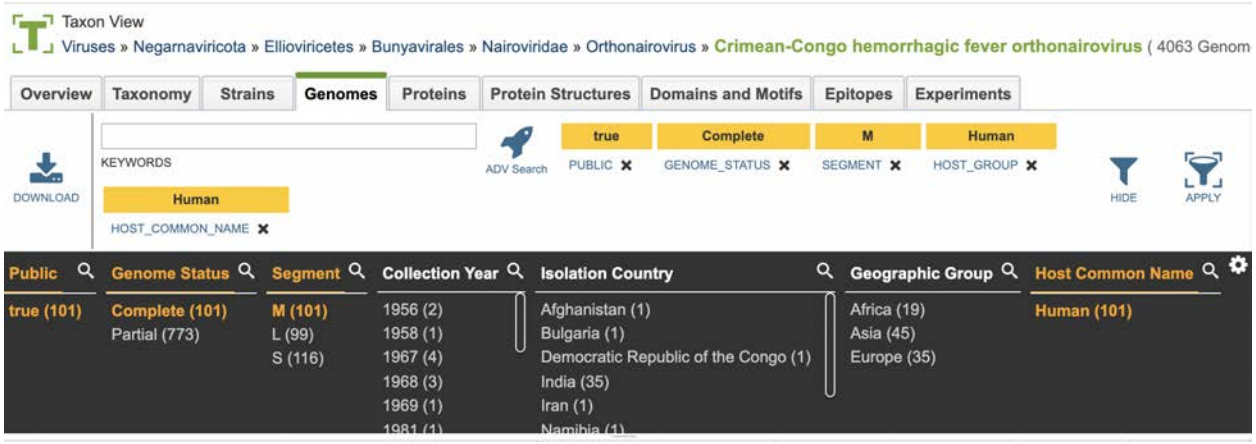
- Explore “ID Type” and “Color” schemes as desired, or return to the file list to select and view the SNP variation report (snp.tsv file extension) for a detailed view of variation in the MSA.



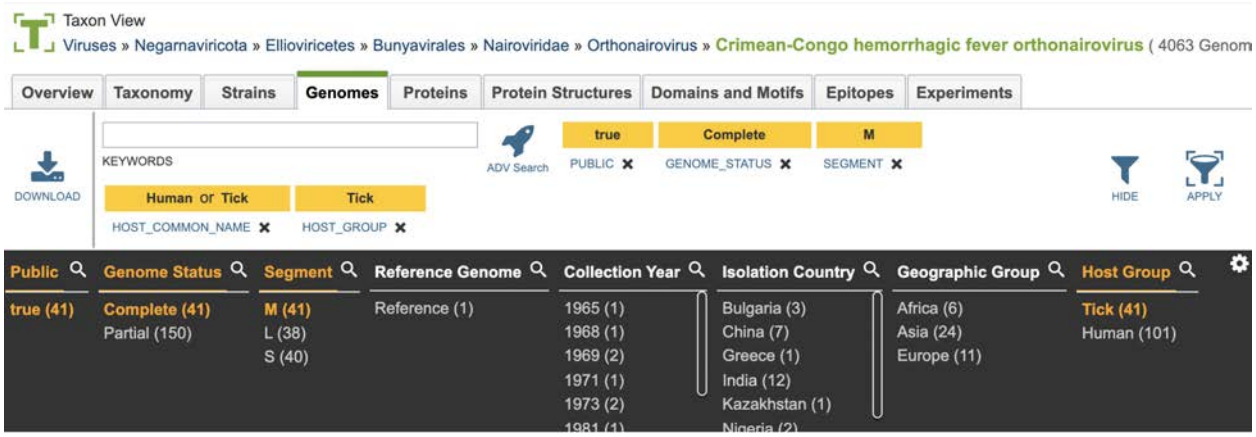
Step 1c) Use MetaCats to search for other tick specific sites (MetaCATS, Documentation). While our MSA only identified one strain with the published mutation relevant to tick host tropism, users may wish to use our Meta-CATS tool to search for other sites that may occur at a higher frequency in CCHFV strains isolated from tick hosts rather than human hosts. (Note this tool can be used to look for positions that significantly differ between user-defined groups of sequences, or groups based on database metadata such as genotype, geography, or time of isolation).

- Returning to the “Proteins” tab, use the “Host Group” Filter to create separate “Tick” and “Human” feature groups for GPC.

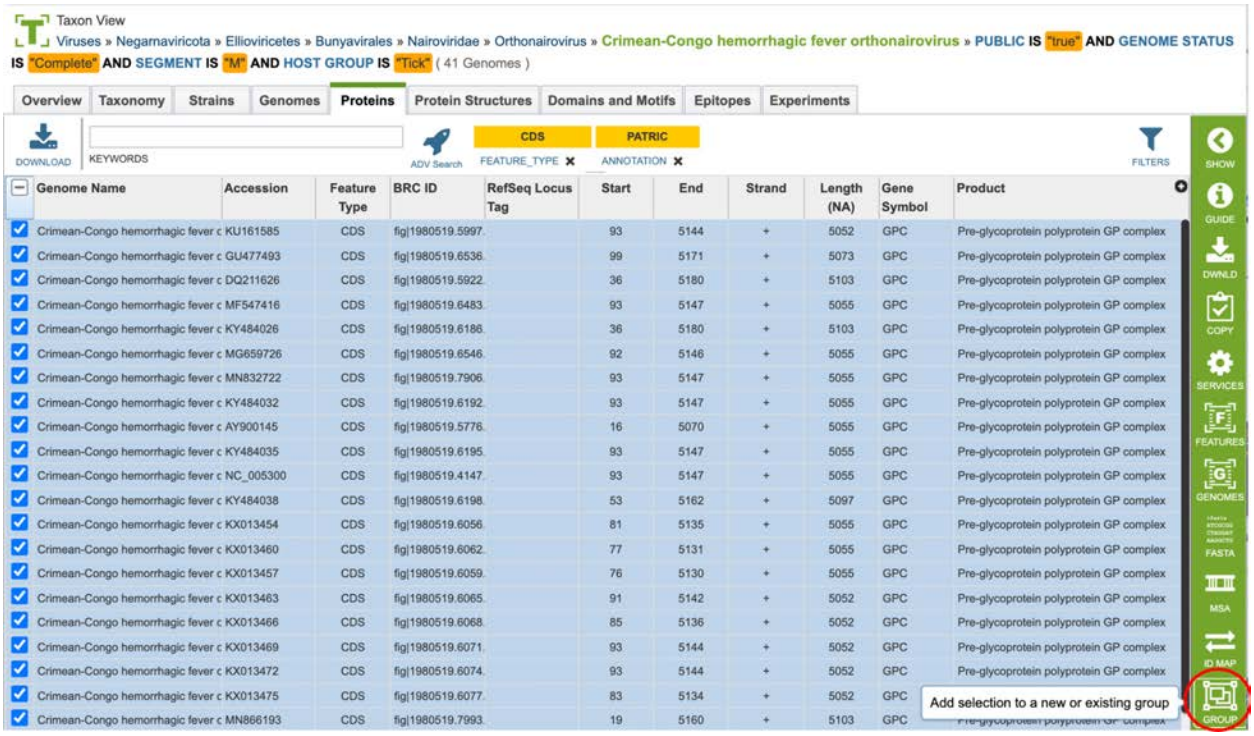
  (101 results for CCHFV M segments isolated from humans with the below criteria)

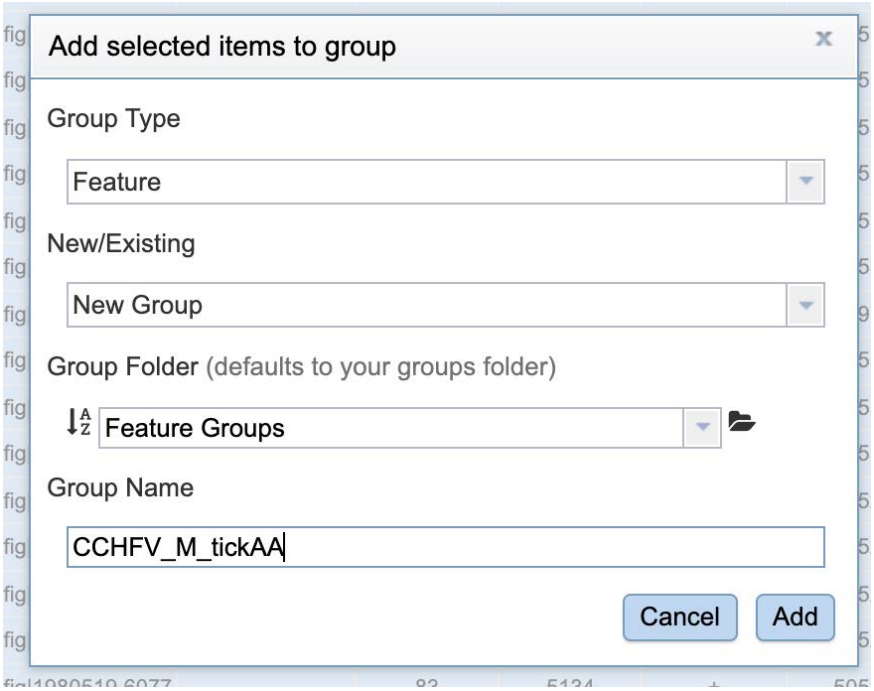


(41 results for CCHFV M segments isolated from ticks with the below criteria)

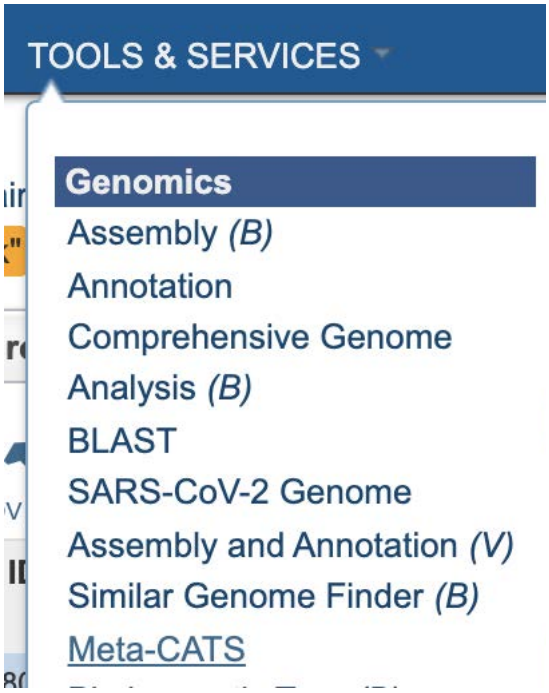


- Click apply, switch to the proteins tab and save the tick and human GPC protein datasets as feature groups named “CCHFV_M_tickAA” and “CCHFV_M_humanAA” respectively.

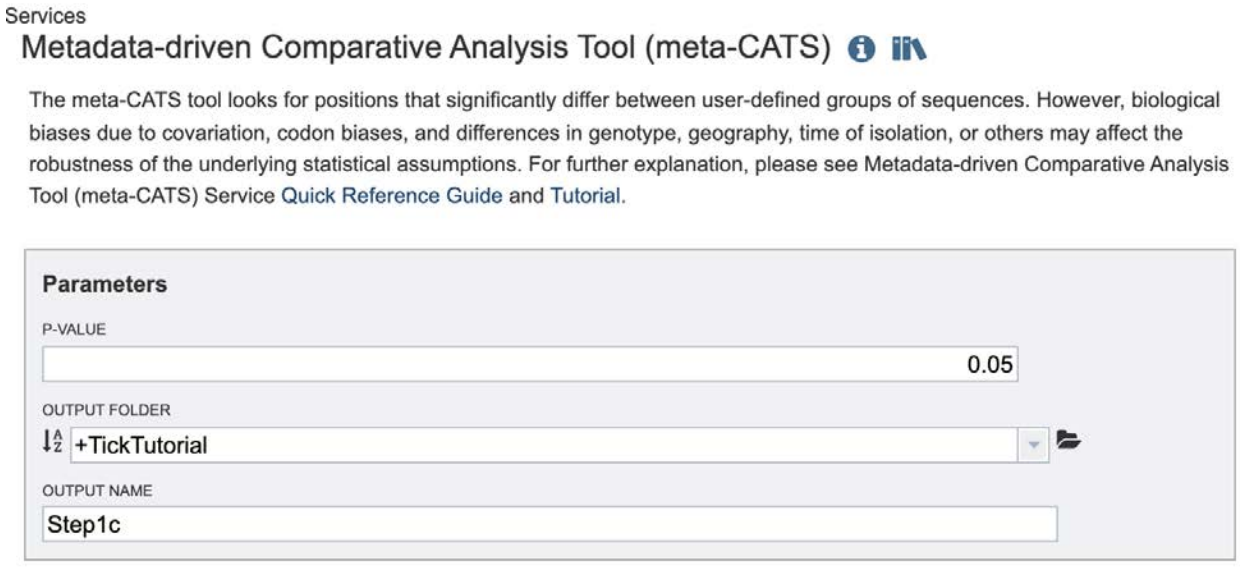




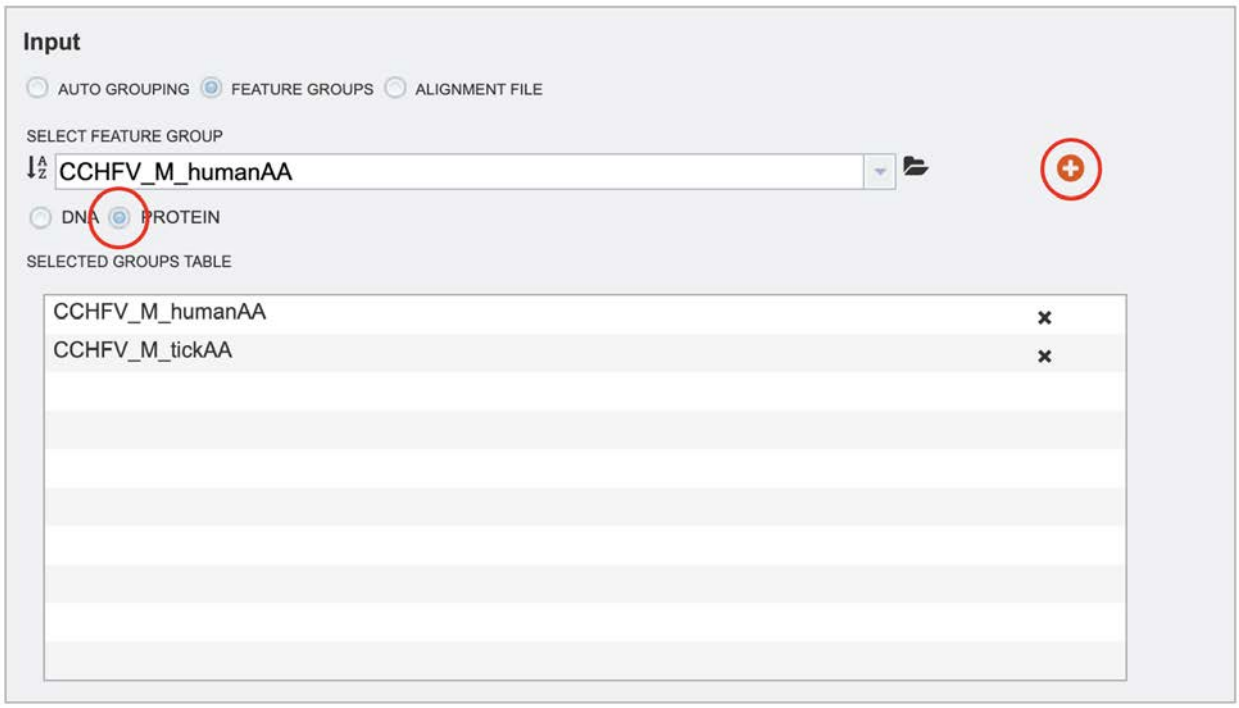
- Navigate to the MetaCATS tool underneath the “Genomics” header in the “TOOLS & SERVICES” tab.



- Specify the desired Output folders and names.

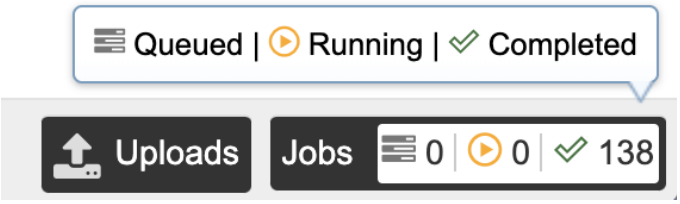


- Select “feature groups”: “CCHFV_M_tickAA” and “CCHFV_M_humanAA”, and click the “+” button to add them to the “selected groups table” (see below).
- Confirm “Protein” is selected, and click submit to launch the job.



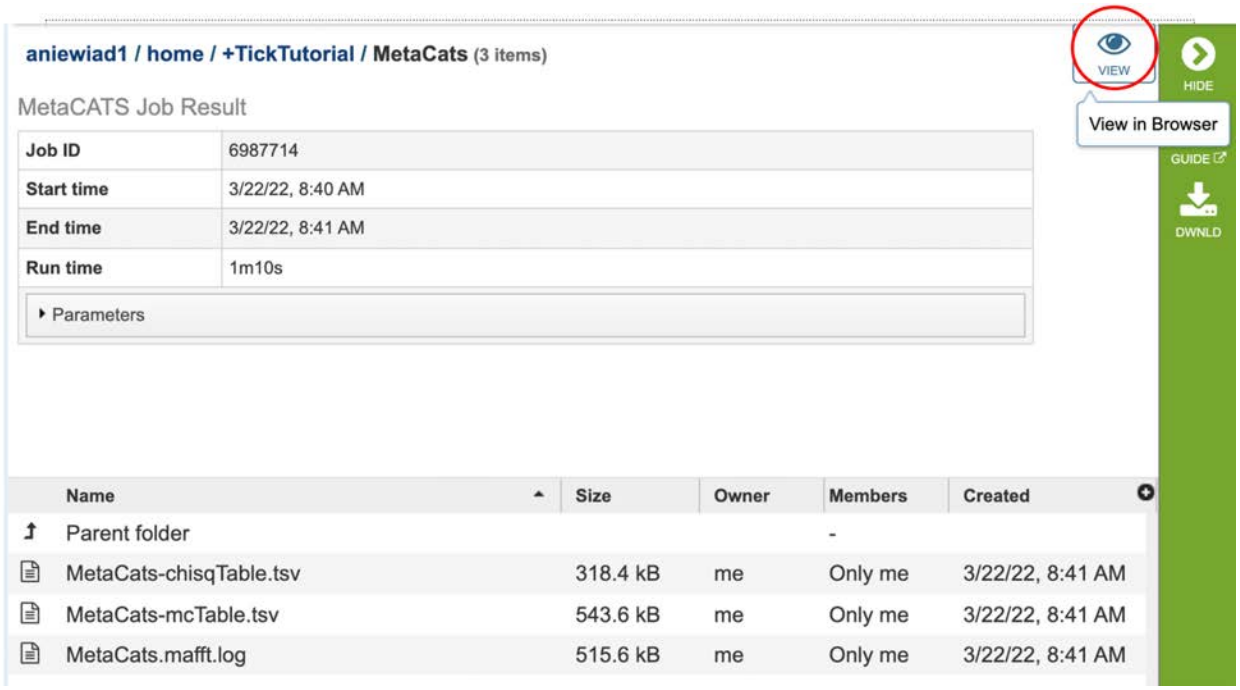
Reset Submit

- To view results, click on the job status box in the bottom right corner of the webpage..

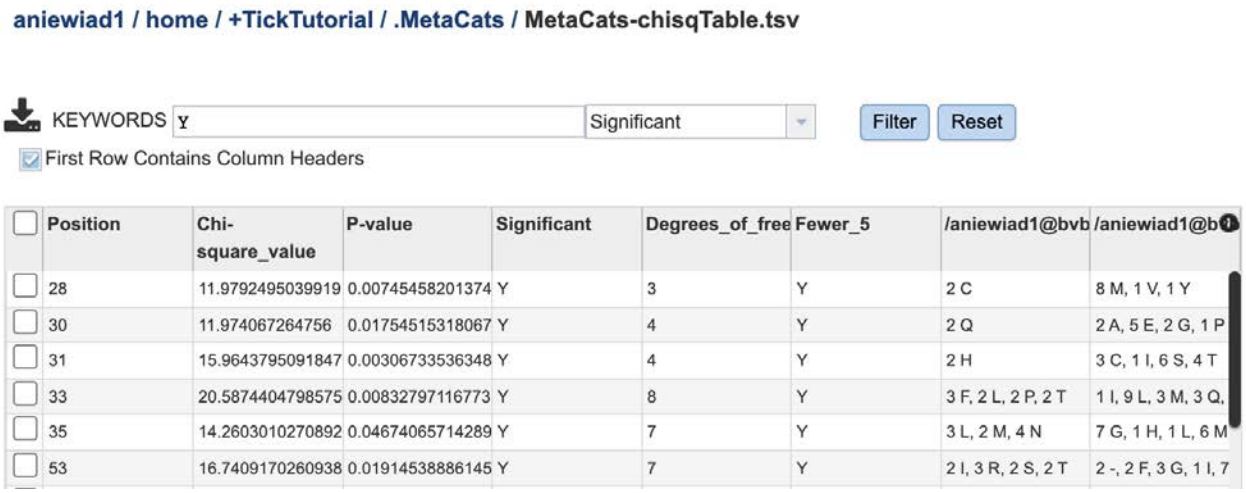


- Navigate to your job, and click on “View” to view results.

- A list output files is provided for the user (see explanations in the documentation). You may explore these individually, or use the eye-shaped “View” icon (red circle below) to navigate to view a list of sites that differ significantly between proteins isolated from tick versus human hosts.



- Results can further be filtered to only display statistically significant results, using "Keywords" filters as shown below.



Use Case 2: Isolation and characterization of an "unknown" Nairovirus.

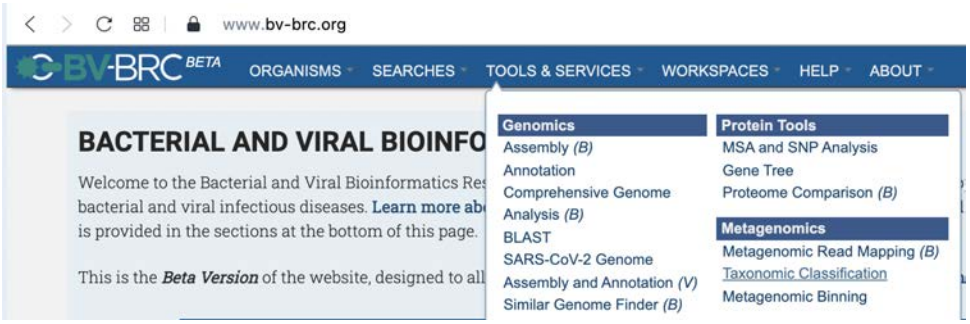
Often, researchers or clinicians encounter patients with symptoms of hemorrhagic fever but with unknown etiology. In this case, diagnostic measures may include whole genome sequencing of a patient sample, to try to detect the causative infectious disease agent.

For the purposes of this exercise, we will pretend that the following raw reads deposited in the Sequence Read Archive (SRA) database, are from just such a sample. Run number: **SRR10769498**

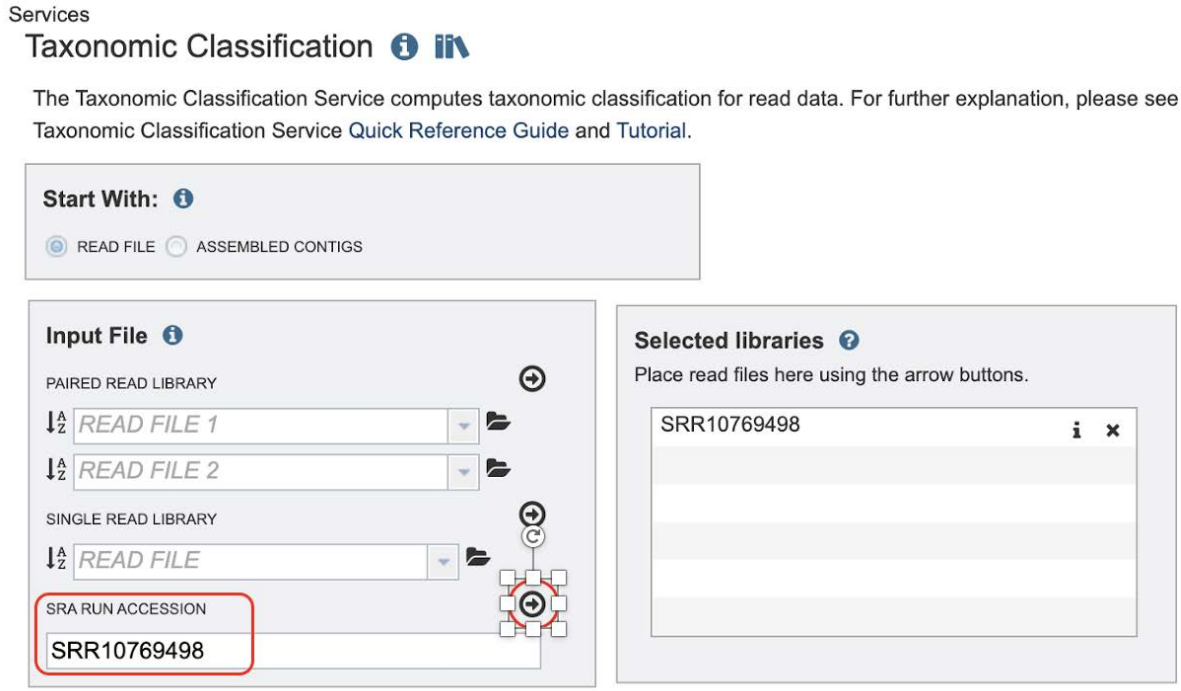
Step 2a) SRA file analysis (Taxonomic classifier, Documentation)

In order to assess read content in this sample, we will use taxonomic classification.

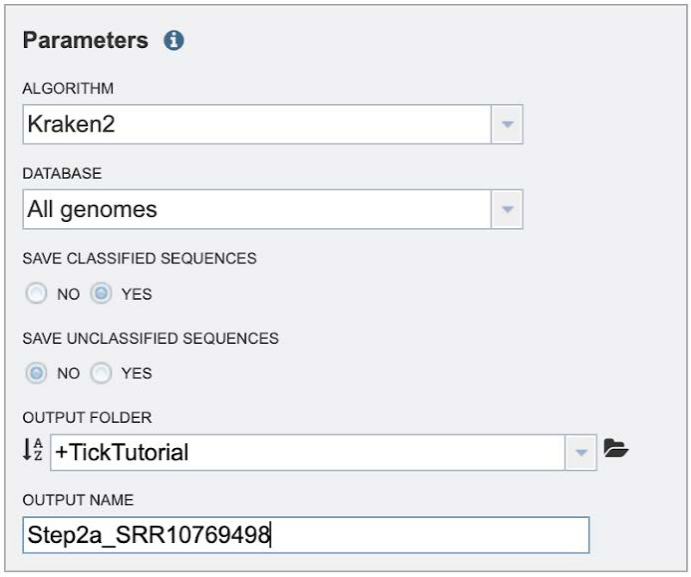
- Navigate to the “Taxonomic Classification” tool underneath the “Metagenomics” header in the “TOOLS & SERVICES” tab.



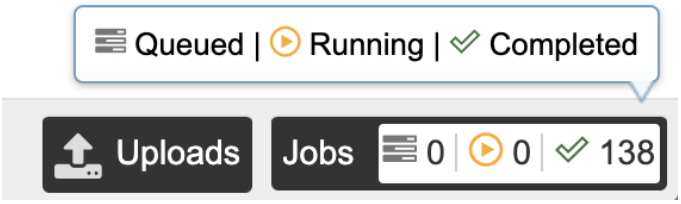
- Input the above SRA run number (SRR10769498) into the appropriate box (red rectangle below), then click the indicated arrow (red circle below) to move the dataset to the “Selected Libraries” box.



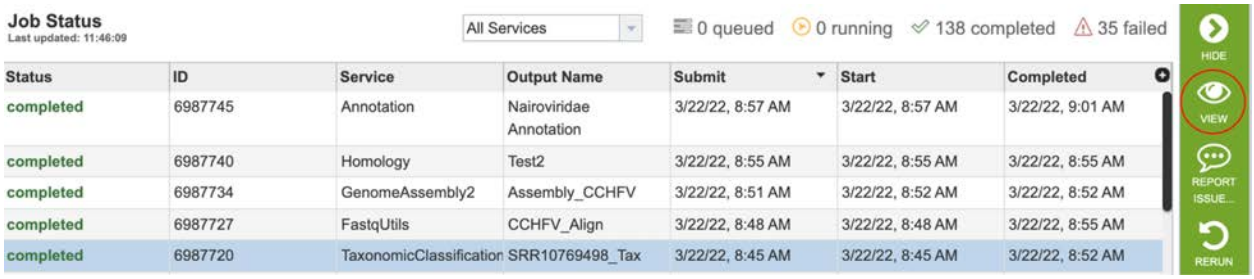
- Select the appropriate parameters as shown below, specifying your desired “output folder” and “output name”. Once selected, the “submit” button can be clicked to launch the job.



- To view results, click on the job status box in the bottom right corner of the webpage.



- Navigate to your job, and click on “View” to view results.



- A list output files is provided for the user (see explanations in the documentation). You may explore these individually, or use the eye-shaped “View” icon (red circle below) to navigate to the “Taxonomic Report”.

aniewiad1 / home / +TickTutorial / SRR10769498_Tax (6 items)

VIEW

TaxonomicClassification Job Result

Job ID	6987720
Start time	3/22/22, 8:45 AM
End time	3/22/22, 8:52 AM
Run time	7m15s

▸ Parameters

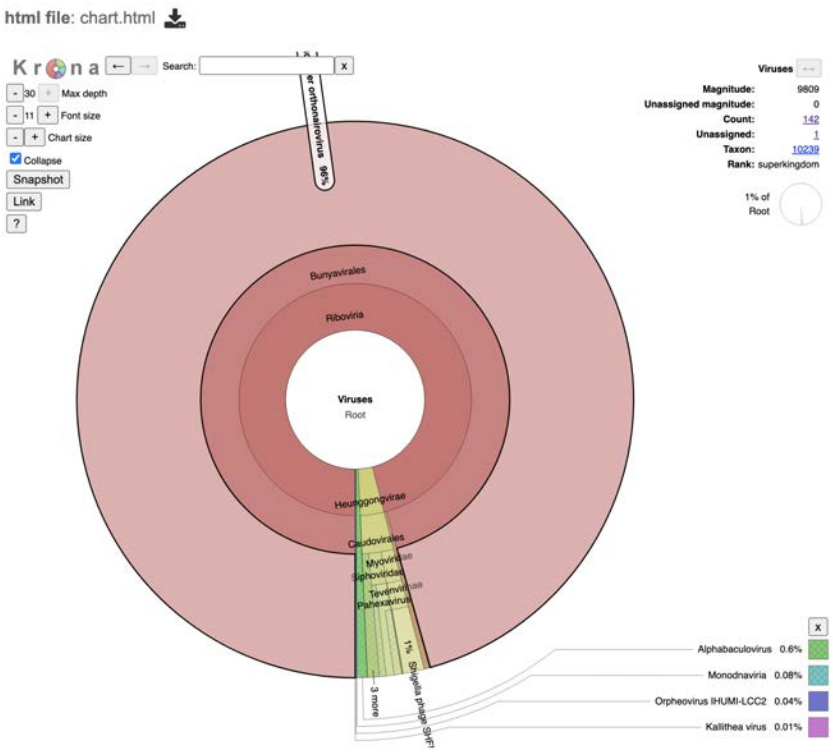
Name	Size	Owner	Members	Created
Parent folder			-	
TaxonomicReport.html	19.1 kB	me	Only me	3/22/22, 8:52 AM
chart.html	2.5 MB	me	Only me	3/22/22, 8:52 AM
classified.fastq.gz	55.1 MB	me	Only me	3/22/22, 8:52 AM
full_report.txt	2.0 MB	me	Only me	3/22/22, 8:52 AM
output.txt.gz	14.9 MB	me	Only me	3/22/22, 8:52 AM
report.txt	427.8 kB	me	Only me	3/22/22, 8:52 AM

- Results can be viewed either in a “Table format” or as an “Interactive chart” (see below).

html file: TaxonomicReport.html

In addition, the output file output.txt.gz contains information about each input sequence. Documentation on this format is available here.

Pct Coverage	Frags in Clade	Frags in Taxon	Rank	NCBI Taxon ID	Scientific Name
1.29	9809	0	D	10239	Viruses
1.24	9418	0	D1	2559587	Riboviria
1.24	9386	0	K	2732396	Orthornavirae
1.24	9386	0	P	2497569	Negarnaviricota
1.24	9386	0	P1	2497571	Polyploviricotina
1.24	9386	0	C	2497576	Ellioviricetes
1.24	9386	0	O	1980410	Bunyavirales
1.24	9378	0	F	1980415	Nairoviridae
1.24	9378	0	G	1980517	Orthonairovirus
1.24	9378	9378	S	1980519	Crimean-Congo hemorrhagic fever orthonairovirus



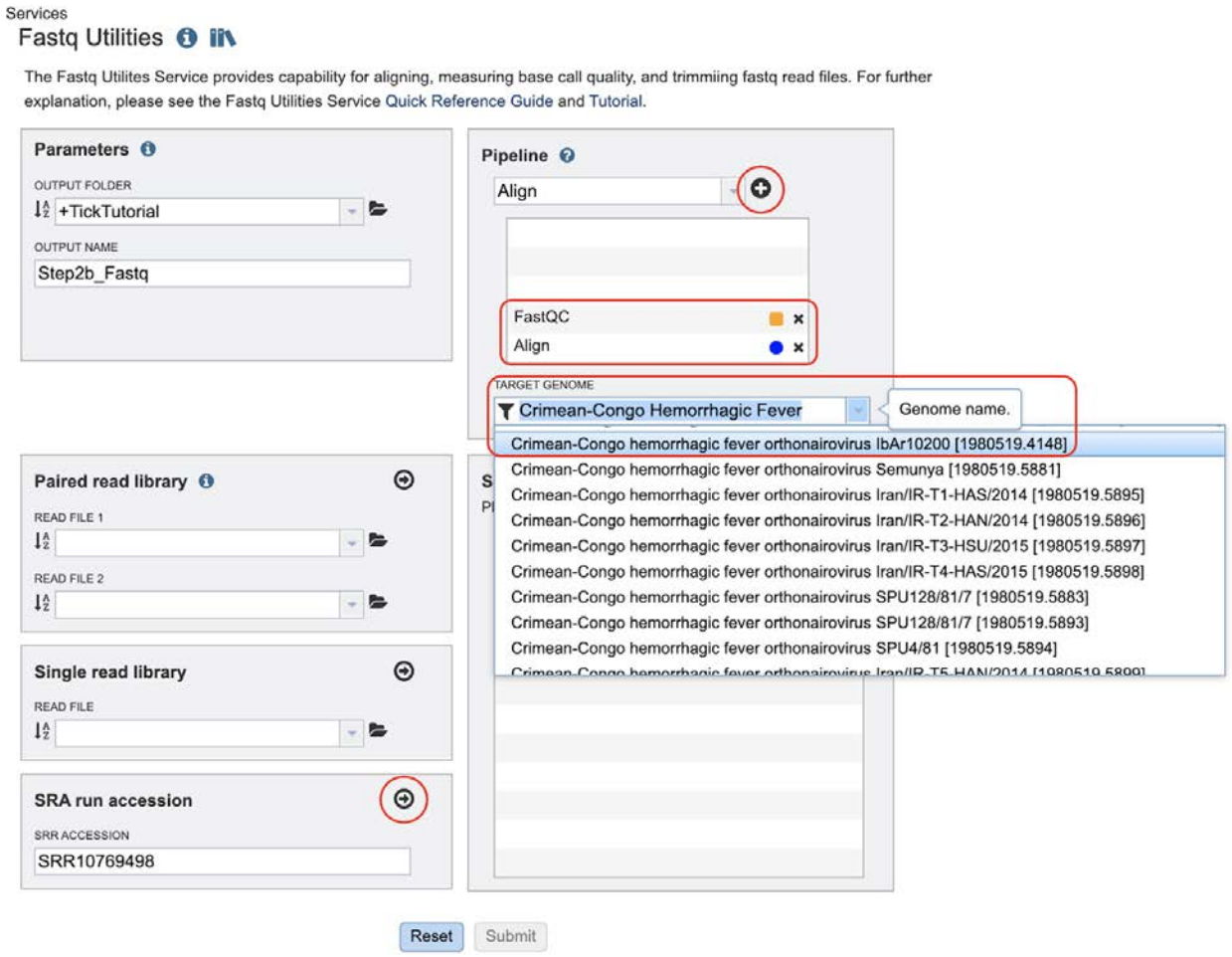
- Results for this SRA run number indicate the presence of Crimean-Congo Hemorrhagic Fever Virus reads.

Step 2b) Read QC and mapping (Fastq Utilities, Documentation)

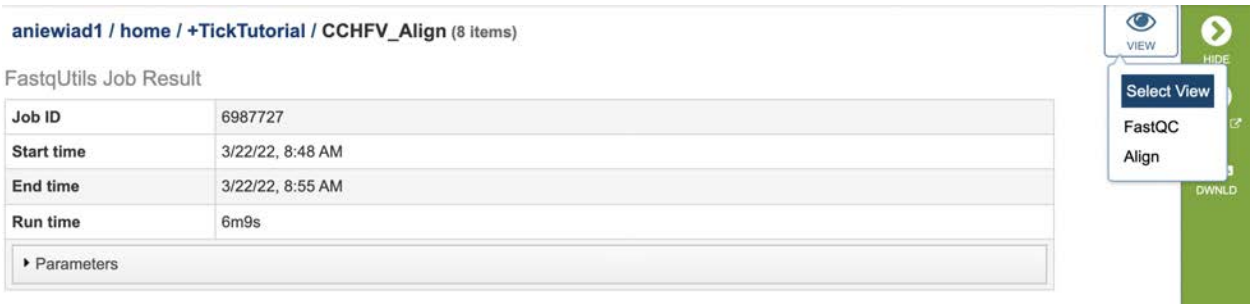
After detection of a virus of interest in a sample, users may want to assemble their reads into viral contigs for further analysis. While reference based sequencing is not yet available for viral sequences (*pipeline coming soon!*), users can utilize a combination of our “Fastq Utilities” and *de novo* sequence “Assembly” services to extract and assemble their viral reads (see below).

- Navigate to the “Fastq Utilities” tool underneath the “Utilities” header in the “TOOLS & SERVICES” tab.

- Enter the SRA run accession number from the previous step (don’t forget to press the arrow to transfer the data to the “Selected libraries” box as in the previous step!)
- Specify your desired “output folder” and “output name”.
- Select the appropriate analysis pipelines, “FastQC” and “Align”, as shown below, and click the “+” button to add these services.
- Select the appropriate “Target Genome”, in this case, CCHFV.
- Once selected, the “submit” button can be clicked to launch the job.



- Once your job has completed, and you have selected the appropriate job from the list, you can view the results of either the “FastQC” or “Align” pipelines (See eye view icon below).



- The FastQC results summarize several quality control metrics for the sequence reads as shown below (for more information on each of these, see linked service documentation above).

html file: SRR10769498_fastqc.html

FastQC Report

Tue 22 Mar 2022
SRR10769498.fastq

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	SRR10769498.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	757773
Sequences flagged as poor quality	0
Sequence length	160
%GC	49

Per base sequence quality

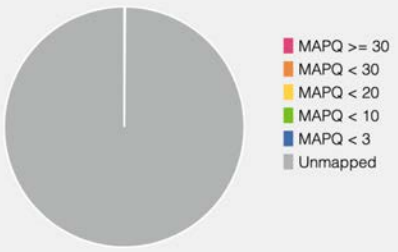
- For the results of the “Align” pipeline, a summary of reads, read length, and base quality is displayed (note: only 0.1% of reads are mapped to the target CCHFV genome previously specified).

html file: SRR10769498.all.bam.samstat.html

SRR10769498.all.bam

757773 reads, size:63437778 bytes, created 2022-03-22 15:50:26

Mapping stats:

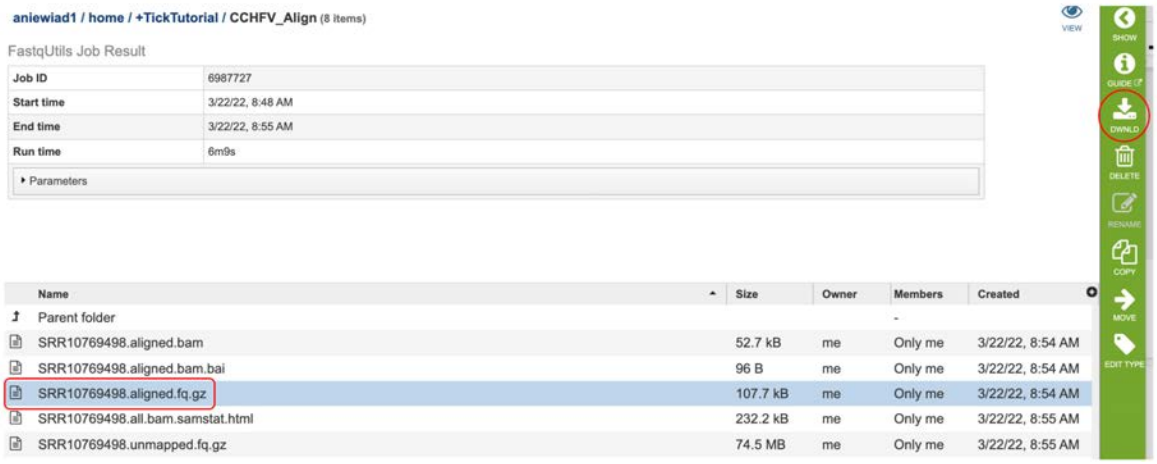


Number of alignments in various mapping quality (MAPQ) intervals and number of unmapped sequences.

	Number	Percentage
MAPQ >= 30	761.0	0.1
MAPQ < 30	108.0	0.0
MAPQ < 20	0.0	0.0
MAPQ < 10	82.0	0.0
MAPQ < 3	0.0	0.0
Unmapped	756822.0	99.9
Total	757773.0	100.0

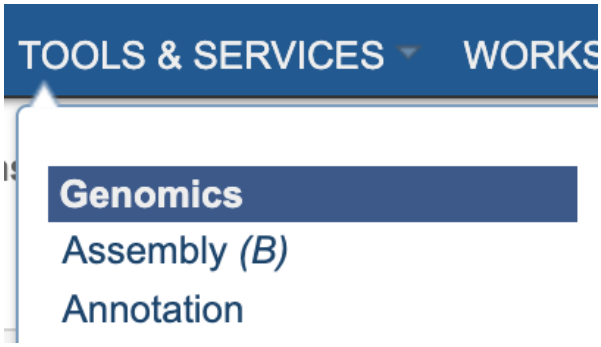
Number of alignments in various mapping quality (MAPQ) intervals and number of unmapped sequences.

- At this point, you may download the aligned reads or use them for further analysis, as shown in **“step 2c”**.

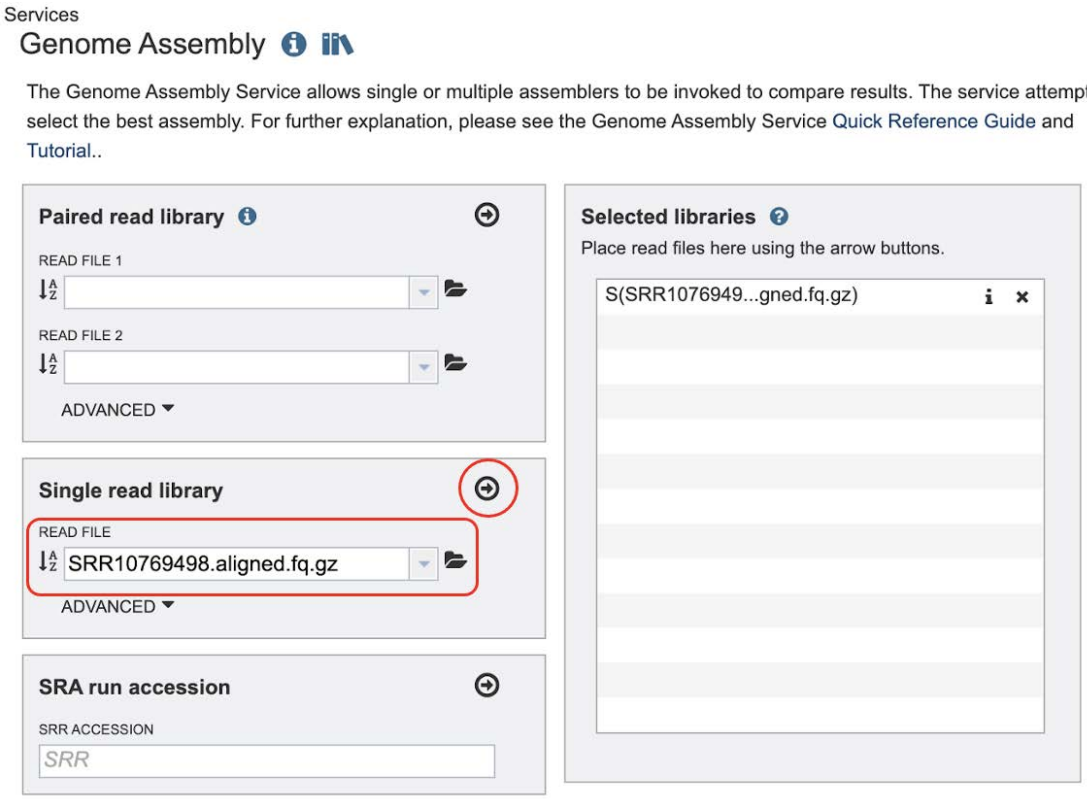


Step 2c) de novo sequence assembly (Genome Assembly Service, Documentation)

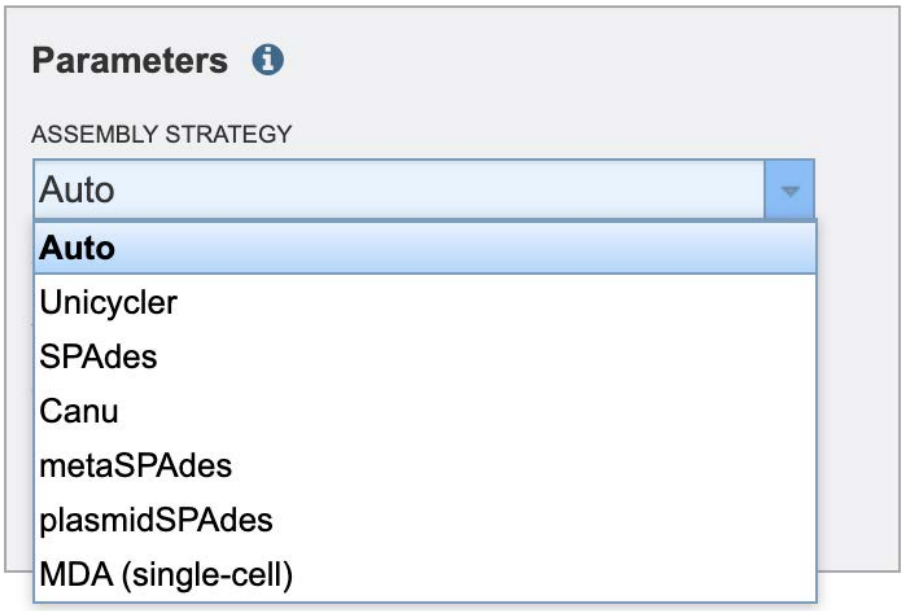
- To assemble the aligned reads from the previous step, navigate to the “Assembly” tool beneath the “Genomics” header underneath the “Tools & Services” tab.



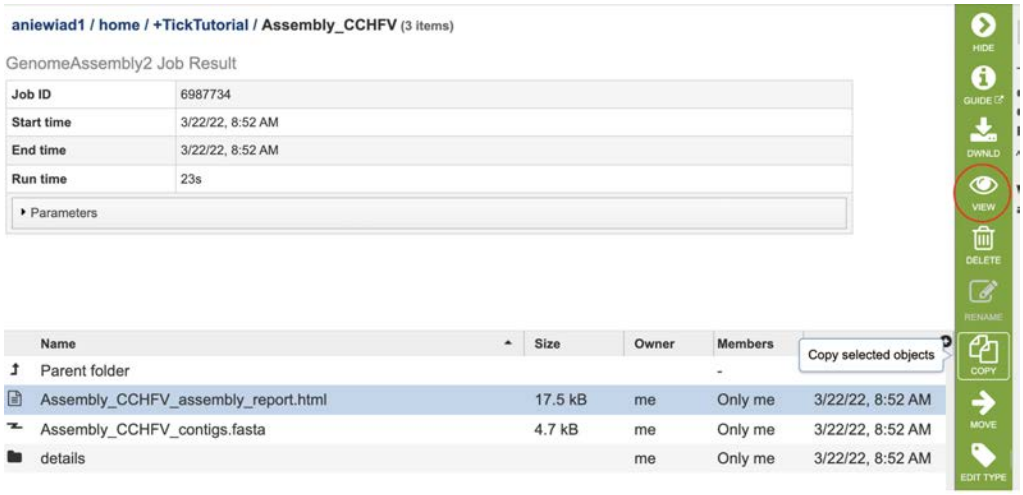
- Select or upload as appropriate the “SRR10769498.aligned.fq.gz” from the previous step, under Single Read Library and click the arrow to move it to the “Selected Libraries” box.



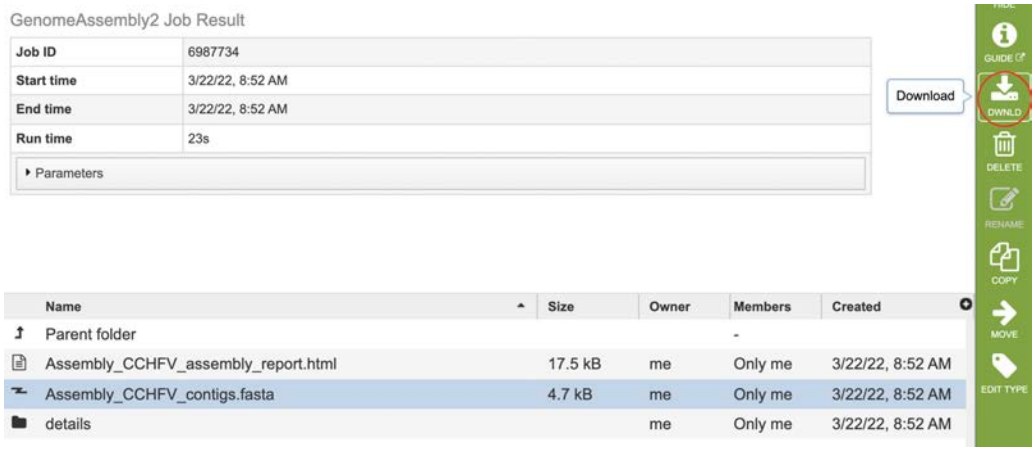
- Specify your desired “output folder” and “output name”.
- You may select your desired assembly strategy, or for the purposes of this exercise leave the “Auto” option selected.



- Once all of the appropriate criteria have been specified, the “submit” button can be clicked to launch the job.
- After completion, users can view an assembly report as shown below.



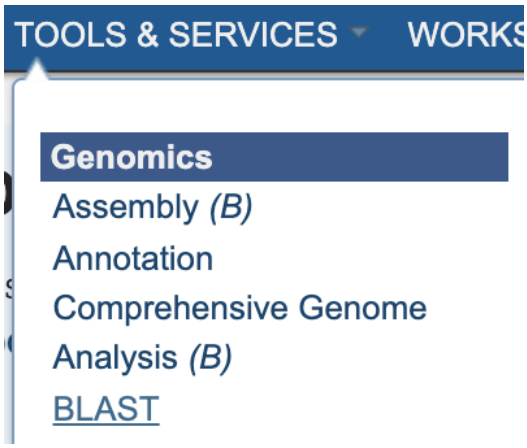
- Alternatively, users can download assembled contigs for further downstream analysis as shown below.



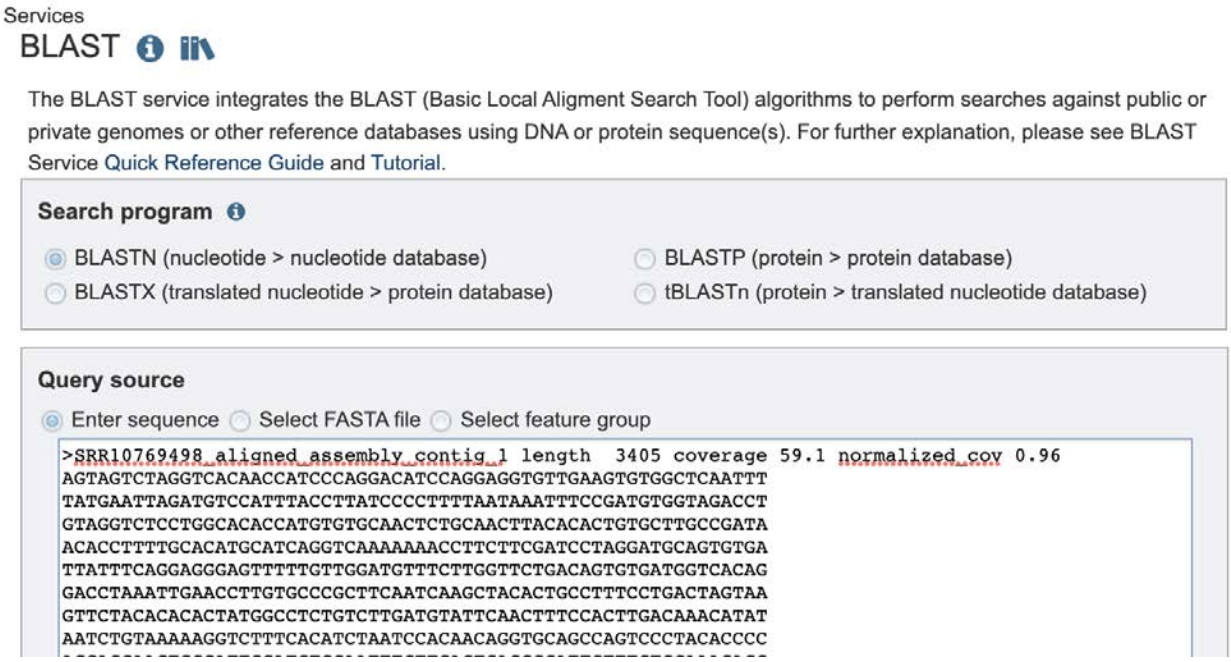
Step 2d) Blast against viral database (BLAST, Documentation)

Next we will utilize the BLAST service to search the BV-BRC databases for the genomes most similar to our assembled contigs.

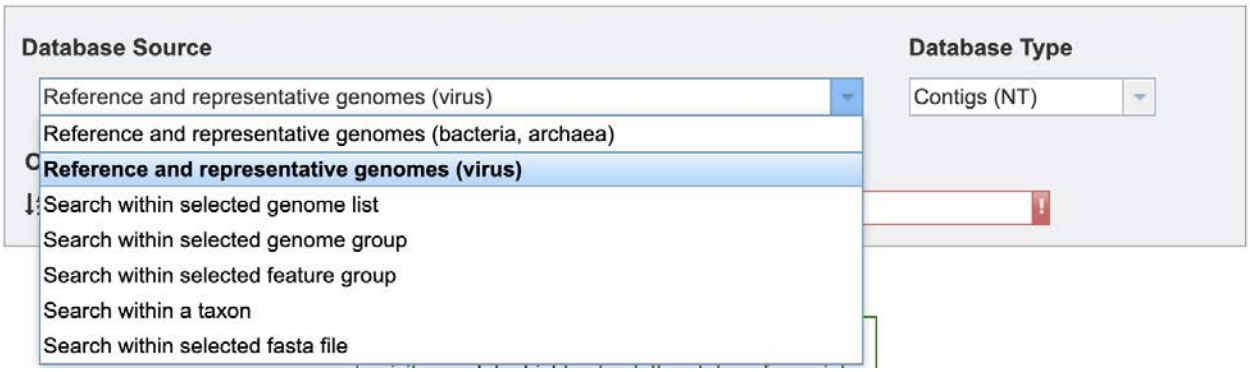
- Navigate to the “BLAST” tool underneath the “Genomics” header in the “TOOLS & SERVICES” tab.



- Select the “BLASTN” program (given that our contigs are nucleotide sequences).
- Next we can input our query either by copying and pasting contigs downloaded from the previous step (as shown below), or by directly selecting the fasta file from your workspace.



- Next, select your desired query database. Appropriate options for this query include:
  “Reference and representative genomes (virus)”
  “Search within a genome group *(searches within a user-compiled dataset of viral genomes)*
  “Search within a taxon”
  “Search within a selected fasta file”



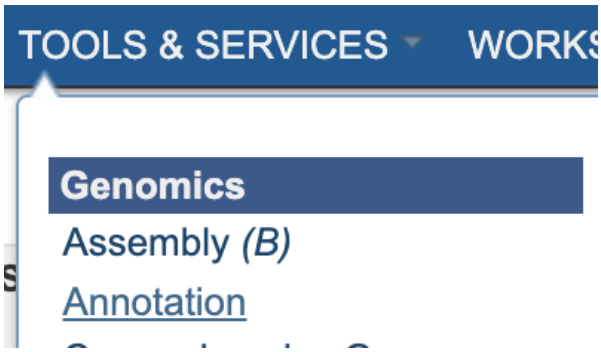
- In the example below, I have selected a saved “Genome group” that I previously compiled, containing all the complete Nairovirus M segments that are in the BV-BRC. Results are displayed as shown below.

Query ID	Genome	Subject ID	Product	Identit (%)	Query cover (%)	Subjec cover (%	Query Length	Subject Length	Score	E value
SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	NC_005300	segment	100	100.00	63.27	3405	5366	6219	0
SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	KY484035	segment	100	100.00	63.27	3405	5366	6202	0
SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	KX013451	segment	98	100.00	63.62	3405	5336	5864	0
SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	MF287637	segment	97	100.00	63.25	3405	5368	5781	0
SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	MF547416	segment	97	99.94	63.23	3405	5366	5633	0
SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	KY484045	segment	96	100.00	63.28	3405	5365	5521	0
SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	KY484042	segment	96	100.00	63.28	3405	5365	5499	0
SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	KJ682814	segment	96	100.00	63.29	3405	5364	5493	0
SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	AY900141	segment	96	100.00	67.16	3405	5055	5486	0
SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	KJ682813	segment	96	100.00	63.29	3405	5364	5454	0

Step 2e) Annotate my genome (Genome Annotation Service, Documentation)

Now that we have our assembled contigs, we can further characterize our viral genomes by annotating the proteins they code for. For this, we will utilize the “Genome annotation service”.

- Navigate to the “Annotation” tool underneath the “Genomics” header in the “TOOLS & SERVICES” tab.



- Upload or select your fasta formatted contig file as appropriate
- Select the desired annotation recipe; in this case “Viruses”
- Enter the appropriate Taxon name for annotation; in this case I have selected the entire *Nairoviridae* family, however users can also select CCHFV.
- Specify the appropriate output folders and names, and click “Annotate” to launch the job.

Services

Genome Annotation

The Genome Annotation Service provides annotation of genomic featuers using the RAST tool kit (RASTtk) for bacteria and VIGOR4 for viruses. The service accepts a FASTA formatted contig file and an annotation recipe based on taxonomy to provide an annotated genome. For further explanation, please see the Genome Annotation Service Quick Reference Guide and Tutorial.

Parameters

CONTIGS: SRR10769498_aligned_assembly_contigs.fasta

ANNOTATION RECIPE: Viruses

TAXONOMY NAME: Nairoviridae

TAXONOMY ID: 1980415

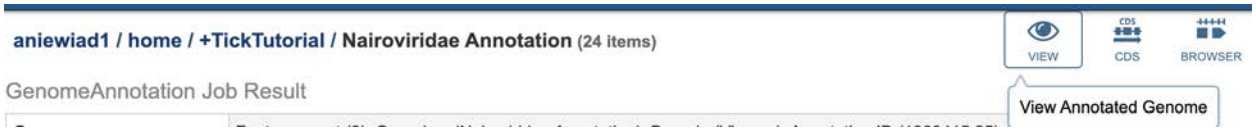
MY LABEL: Step2e_annotation

OUTPUT NAME: Nairoviridae Step2e_annotation

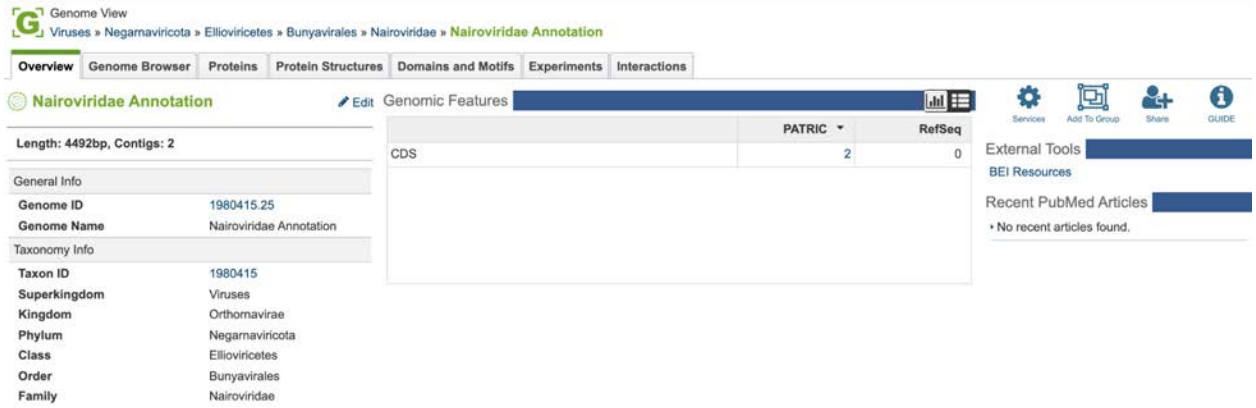
OUTPUT FOLDER: +TickTutorial

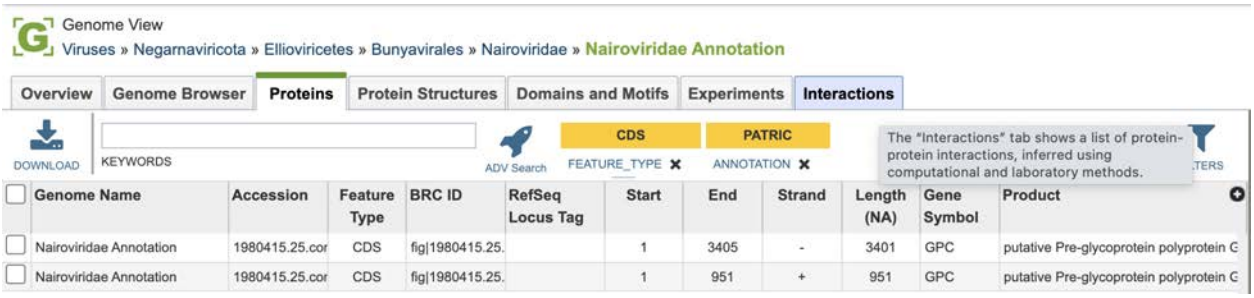
Reset | Annotate

- Results can be viewed in a variety of ways, including “Genome View”, “CDS view”, as well as in the “Genome Browser”..



- Examples of the genome view and the protein list are shown below. Given that our sample only contained fragments of the genome, viewing it in the Genome Browser is not appropriate.





THANKS FOR FOLLOWING ALONG, AND PLEASE CONTACT US WITH YOUR QUESTIONS AT [BV-BRC.ORG](BV-BRC.ORG)!