## VEME 2022: BV-BRC Tutorial

## August 23, 2022

## Use Case 1: Finding a specific mutation reported in the literature

A recent paper has shown that "<u>A single mutation in Crimean-Congo hemorrhagic fever virus</u> <u>discovered in ticks impairs infectivity in human cells</u>". In this manuscript, the authors show that "R1116G", a point mutation in the glycoprotein precursor complex (GPC) protein contributes to 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:

## Search and Assemble dataset for exploration (Documentation: Genome/Protein Search)

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

Welcome to the B	esterial and Minal Disin formation D		IER
users is provided	acternal and VITAL Bioinformatics Resour l infectious diseases. <b>Learn more abou</b> in the sections at the bottom of this pag	rce Center (BV-BRC), an information syst <b>ut BV-BRC</b> . Introductory material for tra e.	em designed to support research ansitioning PATRIC and IRD/ViPR
This is the <b>Beta</b> T team.	<b>Version</b> of the website, designed to allow	w users to explore, try out features, and <b>p</b>	provide feedback to the BV-B
	Data Types - Find a gene genome mir	croarray, etc.	Q @ All term
SEARCH All	Data Types Trind a gene, genome, mit	, our ay, or o	
	BACTERIA AF	RCHAEA VIRUSES	EUKARYOTIC HOST
BROWSE ANALYZE D Upload and ana	BACTERIA AF	RCHAEA	EUKARYOTIC HOST
BROWSE ANALYZE D Upload and ana SEARCH	ATA IN BV-BRC .yze your data in the private workspace. ANALYZE	RCHAEA VIRUSES	EUKARYOTIC HOST
SEARCH AII BROWSE ANALYZE D Upload and ana SEARCH Taxa	BACTERIA AF BACTERIA AF ATA IN BV-BRC syze your data in the private workspace. ANALYZE Assembly	RCHAEA VIRUSES MANAGE DATA Access Private Workspace	EUKARYOTIC HOST BATCH ACCESS Command-Line Interface
SEARCH All BROWSE ANALYZE D Upload and ana SEARCH Taxa Genomes	BACTERIA AF BACTERIA AF ATA IN BV-BRC Lyze your data in the private workspace. ANALYZE Assembly Annotation	RCHAEA VIRUSES MANAGE DATA Access Private Workspace Save Search Results	EUKARYOTIC HOST BATCH ACCESS Command-Line Interface Data API
SEARCH All BROWSE ANALYZE D Upload and ana SEARCH Taxa Genomes Proteins	ATA IN BV-BRC ATA IN BV-BRC Lyze your data in the private workspace. ANALYZE Assembly Annotation BLAST	RCHAEA VIRUSES MANAGE DATA Access Private Workspace Save Search Results Upload Data	EUKARYOTIC HOST BATCH ACCESS Command-Line Interface Data API FTP
SEARCH All BROWSE ANALYZE D Upload and ana SEARCH Taxa Genomes Proteins Specialty Genes	ATA IN BV-BRC  Vyze your data in the private workspace.  ANALYZE  Assembly Annotation BLAST MSA	RCHAEA VIRUSES MANAGE DATA Access Private Workspace Save Search Results Upload Data Access Analysis Jobs	EUKARYOTIC HOST BATCH ACCESS Command-Line Interface Data API FTP

• This will take you to the Virus Overview Homepage.



- Note the different Virus Families and Features Viruses available on the right. Also note the Virus Data Summary on the left.
- Now click on "*Bunyavirales*"
- Select the Taxonomy tab and migrate through the Nairoviridae => Orthonairovirus => to select Crimean-Congo hemorrhagic fever orthonairovirus
- Select the "Genomes" view option in the green vertical bar

HC-BV	-BRC <sup>₿</sup>	TA ORG	ANISMS -	SEARCHES	TOOLS & SERVIO	CES -	All Data Ty	rpes 🔻 Find a g	gene, genom	e, microari	ay, etc	Q 🕲 All terms 👻 🤮
Taxor	n View es » Negarnavii	ricota » Ellio	oviricetes » Bu	unyavirales	( 52353 Genomes )							
Overview	Taxonomy	Strains	Genomes	Proteins	Protein Structures	Domains and Motifs	Epitopes	Experiments				
				,	lame			Rank	Genomes			
🖉 🔺 Bunyav	virales							order	52353		Crime	an-Congo hemorrhagic
Arena	aviridae							family	5235	HIDE	fever orti	ionairovirus
Cruliv	viridae							family	9	1	Taxon ID	1980519
Fimo	viridae							family	2557	GUIDE	Taxon	Crimean-Congo hemorrhagic fever ort
<ul> <li>Hanta</li> </ul>	aviridae							family	12789		Name	
► Leish	buviridae							family	14	SERVICES	Taxon Rank	species
<ul> <li>Mypo</li> </ul>	viridae							family	12		Other	CCHFV,Crimean-Congo haemorrhagio
A Nairo	viridae							family	5010	TAXON	Names	fever virus, Crimean-Congo hemorrhag
► Nor	wavirus							genus	2	OVERVIEW	Genetic	1
Orticity	honairovirus							genus	4967	G	Lineage	Viruses.Riboviria.Orthornavirae.Negar
► A	bu Hammad ortho	nairovirus						species	7	GENOMES	Names	Congo hemorrhagic fever orthonairovi
► A	bu Mina orthonaire	ovirus						species	4		Parent ID	1980517
► A	rtashat orthonairo	virus						species	13	FEATURES	Division	Viruses
► A <sup>2</sup>	valon orthonairovi	rus						species	10		Genomes	4104
→ c	him orthonairoviru	IS						species	7			
	ongoid orthonairo	virus						species	3			
V + C	rimean-Congo her	morrhagic feve	er orthonairoviru	S				species	4164			
► D	era Ghazi Khan o	rthonairovirus						species	9			
D ► D	ugbe orthonairovir	rus						species	103			
► E	rve orthonairovirus	S						species	7			
	stero Real orthona	airovirus						species	9			

- Note the taxonomy expansion and the number of genomes selected
- Use the "FILTER" button to select the following criteria and click "APPLY":
  - Genome Status = Complete
  - $\circ$  Segment = M
  - Host Common Name = human AND tick

	n View es » Negarnaviri	icota » Elli	oviricetes » E	Bunyavirales »	Nairovirida	e » Orthon	airovirus »	Crimean-Co	ngo hemor	rhagic fever	or	<b>thonairovirus</b> ( 4164 G	enomes)		
Overview	laxonomy	ottaina	Genomes	Froteina	Frotein G	uuctures	Domaina		Chitobea	CAPETITIET					
						Com	plete	м	Hun	nan OF Tick				T	
DOWNLOAD	KEYWORDS				ADV Search	GENOME_S	TATUS 🗙	SEGMENT X	HOST_CO	MMON_NAME X	:			HIDE	APPLY
Public Q	Genome Statu	<mark>اه ک د</mark>	gment Q	Reference G	enome Q	Collectio	n Year Q	Isolation C	ountry	c	۹.	Geographic Group Q	Host Common Name	م	Host Gr
true (143)	Complete (143	3) M	(143)	Reference (1)		1956 (2)	ſ	Afghanistan			$\Pi$	Africa (26)	Tick (40)		Human (
	Partial (931)	Ĺ (	138)			1958 (1)	U	Bulgaria (4)			U	Asia (69)	Human (101)		Tick (42)
		S	(157)			1965 (1)		China (7)				Europe (46)	Goat (2)		
						1967 (4)		Democratic	Republic of t	he Congo (1)			Sheep (1)		
						1968 (4)		Greece (1)					Small smooth bont-legge	ed tick (2)	
						1969 (3)		India (47)							
						1971 (1)		iran (1)							

- Selecting the "APPLY" option will result in the application of these filtering criteria across the subsequent tabs
- After clicking "APPLY", switch to the "Proteins" tab
- Note the Start and End positions and length of the GPC gene records
- Select all records and deselect those records annotated as "putative......"
- Select the "MSA" option in the green action bar on the right, and "Amino Acids"

L HC	Taxo	on View ses » Negarnav IMON NAME IS	iricota	a » Elliovii <mark>man"</mark> OR	ricetes » Bun HOST COMI	yavirales » MON NAMI	Nairovirida E <b>IS <sup>"</sup>Tick"</b>	e » Orthona ( 143 Geno	airovirus omes )	» Crime	ean-Cor	ngo hemo	rrhagic feve	er orthonairovirus » GENOME S	TATUS IS
C	verview	Taxonomy	Sti	rains (	Genomes	Proteins	Protein S	tructures	Domai	ins and I	Notifs	Epitopes	Experime	nts	
								CD	S	P/	ATRIC			T	6
DC	WNLOAD	KEYWORDS					ADV Search	FEATURE_T	YPE X	ANNOT	ATION 🗙			FILTERS	HIDE
0	Genome	Name	A	Accession	Feature Type	BRC ID	RefSeq Locus	Sag Sag	tart	End	Stran	d Leng (NA	th Gene ) Symbol	Product C	
✓	Crimean-	Congo hemorrhagi	c fe H	HM452306	CDS	fig 1980519	9.65€	9	93	5147	+	5055	GPC	Pre-glycoprotein polyprotein GP compl	
☑	Crimean-	Congo hemorrhagi	c fe D	Q211628	CDS	fig 1980519	0.592		16	5165	+	5094	GPC	Pre-glycoprotein polyprotein GP compl	
☑	Crimean-	Congo hemorrhagi	c fe D	DQ211626	CDS	fig 1980519	9.592	;	36	5180	+	5103	GPC	Pre-glycoprotein polyprotein GP compl	
☑	Crimean-	Congo hemorrhagi	c fe K	KX056060	CDS	fig 1980519	9.610	1	78	5144	+	5067	GPC	Pre-glycoprotein polyprotein GP compl	
☑	Crimean-	Congo hemorrhagi	c fe K	KX056057	CDS	fig 1980519	9.605	:	58	5109	+	5052	GPC	Pre-glycoprotein polyprotein GP compl	COPY
	Crimean-	Congo hemorrhagi	c fer K	XX056054	CDS	fig 1980519	9.60	:	52	5103	+	5052	GPC	Pre-glycoprotein polyprotein GP compl	
☑	Crimean-	Congo hemorrhagi	c fe K	KX056051	CDS	fig 1980519	9.60\$	9	93	5144	+	5052	GPC	Pre-glycoprotein polyprotein GP compl	FEATURES
	Crimean-	Congo hemorrhagi	c fe K	X238957	CDS	fig 1980519	9.612	1	92	763	+	672	GPC	putative Pre-glycoprotein polyprotein G	G
	Crimean-	Congo hemorrhagi	c fe K	X238957	CDS	fig 1980519	9.612	8	85	1683	+	799	GPC	putative Pre-glycoprotein polyprotein G	GENOMES
	Crimean-	Congo hemorrhagi	c fe K	X238957	CDS	fig 1980519	9.612	1	738	5146	+	3409	GPC	putative Pre-glycoprotein polyprotein G	>fasta ATCGCGG CTAGGAT
☑	Crimean-	Congo hemorrhagi	c fe N	MF547416	CDS	fig 1980519	9.648	1	93	5147	+	5055	GPC	Pre-glycoprotein polyprotein GP compl	FASTA
<	Crimean-	Congo hemorrhagi	c fe K	KY484026	CDS	fig 1980519	9.618	;	36	5180	+	5103	GPC	Pre-glycoprotein polyprotein GP compl	
✓	Crimean-	Congo hemorrhagi	c fe D	DQ211630	CDS	fig 1980519	0.592	1	93	5144	+	5052	GPC	Pre-glycoprotein polyprotein GP compl	MSA
	Crimean-	Congo hemorrhagi	c fe K	KY484029	CDS	fig 1980519	9.618	9	93	5144	+	5052	GPC	Pre-glycoprotein polyprotein GP compl	
	Crimean-	Congo hemorrhagi	c fe N	MG659726	CDS	fig 1980519	0.654	9	92	5146	+	5055	GPC	Pre-glycoprotein polyprotein GP compl	$\leftarrow$
✓	Crimean-	Congo hemorrhagi	c fe K	KX013445	CDS	fig 1980519	0.604	1	83	5134	+	5052	GPC	Pre-glycoprotein polyprotein GP compl	
✓	Crimean-	Congo hemorrhagi	c fe N	MN832722	CDS	fig 1980519	9.790	9	93	5147	+	5055	GPC	Pre-glycoprotein polyprotein GP compl	밑
$\checkmark$	Crimean-	Congo hemorrhagi	c fe K	KY484032	CDS	fig 1980519	9.615	9	93	5147	+	5055	GPC	Pre-glycoprotein polyprotein GP compl	GROUP

- Change "ID Type" in green bar to "Strain"
- Note sequences causing alignment gaps for potential removal
- Scroll to the appropriate area "LVSGRSES" consensus region (~1138) to view the region of interest.
- Search for the "Malko Tarnovo-BG2012-T1303" name to find the genome reported in the publication referenced above.

CBV-BRC <sup>BETA</sup> ORGANISMS	SEARCHES TOOLS & SERVICES	WORKSPACES	All Data Types 🛛 👻 Find a gene, genome, micro	roan
<ul> <li>MCL-19-1-1812</li> <li>MU1614467</li> <li>NIV1614467</li> <li>NIV1727150</li> <li>ArD8194</li> <li>UG3010</li> <li>UG3010</li> <li>UG3010</li> <li>VG3010</li> <li>Arb8194</li> <li>UG3010</li> <li>Makiwogo</li> <li>YL04057</li> <li>-&gt;9121M18</li> <li>Malko Tarnovo-BG2012-T1302</li> <li>Malko Tarnovo-BG2012-T1362</li> <li>Malko Tarnovo-BG2012-T1303</li> </ul>	E       H       R       G       N       K       I       L       V       S       R       S       E       S       I       M         E       H       R       G       N       K       I       L       V       S       R       S       E       S       I       M         E       H       R       G       N       I       L       V       S       R       S       E       S       I       M         E       H       K       G       N       I       L       V       S       R       S       E       S       I       M         E       H       K       G       N       K       I       L       V       S       R       S       E       S       I       M         E       H       R       G       N       K       I       V       S       R       S       S       I       M         E       H       R       G       N       K       I       V       S       R       S       S       I       M         E       H       R       G       N	L         E         R         T         G         I         S         W         S           L         E         R         T         G         I         S         W         S           L         E         R         T         G         I         S         W         S           L         E         E         R         T         G         I         S         W         S           L         E         E         R         T         G         I         S         W         C           L         E         E         R         T         G         I         S         W         S           L         E         E         R         T         G         I         S         W         S           L         E         E         R         T         G         V         S         W         N           L         E         E         R         T         G         M         S         W         N           L         E         E         R         T         G         M         S         W         N	L G V E D A S E S K T L T V S L G V E S K T L T V S L G V E S K T L T V S T V S T V S T V S	LS RS PE
Pentalofos-Greece-2015     →D-206     →ArD39554     •37-R-2013     •52-R-2014	E H R G N K I L V S G R S E S I M E H R G N K V L V S G R S E S I M E H R G N K I L V T G R S E S I M E H R G N K I L V T G R S E S I M E H R G N K I L V T G R S E S I M	L E E R T G I S W N L E E R T G I S W N L E E R T G I S W N L E E R T G V S W D L E E R T G V S W D	L G V E D A S E S R T L T I S L G V E D A S E S R T L T V S L G V E D A S E S R T L T V S L G V D A S E S K L L T V S L G V D A S E S K L L T V S L G V D A S E S K L L T V S	J D

- Note the R to G substitution responsible for impaired infectivity of human cells reported in the paper.
- Go to the Protein Structure tab, select 7A59 and "Structure" function in green bar



- Highlight LVSGRSES sequence to add ball and stick structure to ribbon view
- Find ARG at position 1105 in structure
- Note that it coordinates interactions between 4 different beta strands

## Use Case 2: Are there consistent amino acid difference in viruses isolated from different hosts or different geographic regions?

## Step 1. Sequence selection and quality control

- In Taxonomy tab, select Crimean-Congo hemorrhagic fever orthonairovirus as before
- Select the "Genomes" view option in the green vertical bar
- Select filtering criteria
  - Complete
  - M segment
  - Human and tick
- Select "APPLY" option above table
- Go to Proteins tab
- Select all records except the "putative....." proteins
- Select the "MSA" function in the green bar, and the "Amino Acids" option
- Note the problematic sequence in the middle of the alignment as candidates for removal

C	BV-BRC <sup>BETA</sup> ORGANISMS	- SEARCHES -	TOOLS & SERVICES	WORKSPACES -	All Data Types	- Find a gene, genome, microan
	> IbAr10200		<u>M H I</u>	SLMYAILCLOL	CGLGETH	GS-HNETR
	- 201643792		<mark>M H I</mark>	SLMYAILCLQL	CGLGETH	GS-HNETR
	Caceres 2014		M H I	SLMYAILCLQL	CGLGETH	GS-HNETR
	SPU41/84		MH I	SLMYAVLCLQL	CGLGETH	GS-HNGTR
	•K168 40		MHI	LMCAVLCLOL	YGLGGTH	GL HNGTE
	▶K168_125		M H I	SLMCAVLCLOL	YGLGGTH	GL-HNGTE COLORS
	• U2-2-002/U-6415		M H I	SLMCAVLCLQL	YGLGGTH	GS-HNGTE
	•K128_76	* * * * * * * *	M H I	SLMCAVLCLQL	YGLGGTH	GS-HNETE
	FK16116	* * * * * * * *	M H I	SLMCAVLCLQL	YGLGGTH	GS-HNGTE
	Hodzba		MH I	LMCAVLCLQL	YGLGGTH	GS-HNGTE DIVPE
	• SPU 130/89		MYT	PLICALLCLOL	WSLEGIY	GL SNKTO
- 1	SPU 383/87		M Y T	PLLCAILCLOL	WSLEGIY	GL-SNRTO
I	• SPU 44/08		M Y T	PLLCAILCLQL	WSLEGIY	GL-SNNTQ
I	• SPU 497/88		<mark>M Y T</mark>	PLLCAILCLQL	WSLEGIY	GL-SNKTQ
I	• SPU 18/88		M Y T	PLLCAILCLQL	WSLEGIY	GL-SNKTQ
I	HV-13		· · · · · · · · MYT	PLLCAILCLOV	WGLEGTH	GL-SNKTQ DWNLD
I	Hv13		MYT		WSLEGSH	GL SNKTO
I	• Oman		MHT	LLVYAVECLOL	WSPGGTR	GL - SNETO
I	- Oman_812056					
	•SPU97/85	* * * * * * * * *	MTVNTVCT	PLVCAVFCLQL	WNLGGTL	VP-TNRTQ
- <b>/</b> L	SPU 556/8/		- MTVNTVCT	PLVCAVFCLQL	WNLGGTL	VP-TNRTQ
	L HANM-18		MYT	PLVCAVICLOL	WNLGGTL	VS-TNRTQ
1	Gaib		MYI		Y G L G G A H	GI J NKTE
	•MCL-19-T-1929		VLTMSVCM	LLTNFILCOLF	WSGSGVT	S G G G T N
	MCL-19-T-1916		- VLTMSVCM	LLTNFILCQLF	WSGSGVT	S G G G T N
	MCL-19-T-1812		- V L T M S V C M	LLTNFILCQLF	WSGSGVT	S G G G T N
	NIV1614467		- VLTMSVCM	LLTNFILCQLF	WSGSGVT	S G G G T N
	• NIV1/2/150		- VLTMSVCM	LLTNFILCQLF	WSGSGVT	S G G G T N
	DAK8194		V L T M S V C M	LLTNFFFCQLL	WGGGGGVA	S G G R T N
	•UG3010	CALRATC	OHMFCLRLV	LLINLALCOLL	EGNDGVT	S · · A D E I N
	<b>UG3010</b>	CALRATC	OHMFCLRLV	LLINLALCOLL	EGNDGVT	S A D E I N
	• UG3010		MFCLRLV	LLINLALCOLL	EGNDGVT	S A D E I N
	• Nakiwogo				NDRVT	S A D G I N
	- 70121M18		- VLTMSVCM	LLINFILCHLL	WGGGGVT	G GVETN
	• Malko Tarnovo-BG2012-T1302		CHOYVONEL		WGGGGGVT	G GVETN
	Malko Tarnovo-BG2012-T1362		CMOYVGMEL	LLICTVLHKSP	OGVSANT	HLVSTSGN
	Malko Tarnovo-BG2012-T1303		CMQYVGMFL	LLICTVLHKSP	QGVSANT	HLVSTSGN
	<ul> <li>Pentalofos-Greece-2015</li> </ul>	N I M H	AIMSDHFTFL	LLICTVLLKSP	QGVSANT	HLVSSSGN
- N		V G	ETMLRYILYAI	LLASAILHQHL	YKVGADT	Q K P T V R
	- AID39554		CGMLCHIKHI	LLFCIILYHQQ	WNTGASN	TTISPTTN
- I	\$ 52-R-2014		MHT	LLVCFILYLOL	LGLGGAH	RO-SNATE
- I	•24-R-2012		M H T	LLVCFILYLQL	LGLGGAH	RQ-SNATE
- I	• 5180-R-2011		M H T	LLVCFILYLQL	LGLGGAH	RQ-SNATE
- I	• ROS/HUVLV-100		<mark>M H T</mark>	LLVCFILYLQL	LGLGGAH	RQ-SNATE
- 1	Kashmanov		M H T	LLVCFILYLQL	LGLGGAH	RQ - SNATG
- 1	Min		MHT MHT	LIVCFILYLQL	LGLGGAH	GO SNATE
- 1	Mamon		MHT	LLVCFILYLOL	LGLGGAH	GO - SNTTE
L	l∎ Saf		M H T	LLVCFILYLOL	LGLGGAH	GQ-SNATE
1	•75-ST-2010		M H T I	LLVCFILYLQL	LGPGGAH	GQ - PNATE
- 1	- \$45/4-51-2008		M H T I	LLVCFILYLQL	LGPGGAH	GQ - PNATE
	-4455-51-2000 -K229 243		M H T	LLVCFILYLQL	LGPGGAH	GQ - PNATE
	3809-ST-2007		M H T	LLVCFILYLQL	LGLGGAQ	GQ-SNATE
	- 01 D 2014		MHI	LUCFILILQL	LULUUAQ	UQ-SNAIB

• Remove the problematics sequences before further analysis (Note that in order to save time during the demo, the sequence curation has already been performed and the curated sequence records have been made available in the VEME public folder - <u>CCHF complete M human+tick curated</u>

## Step 2. Phylogenetic analysis

## • Go to Tools & Services and select Gene Tree service

C-BV-BRC <sup>BETA</sup>	ORGANISMS -	SEARCHES		TOOLS & SERVICES WORK	SPACES All Data Types
Taxon View Viruses » Negarnaviricota » HOST GROUP IS "Human" OR HO	Ellioviricetes »	Bunyavirales » "Tick" ( 143 G	Nair	Genomics Assembly <i>(B)</i> Annotation	Protein Tools MSA and SNP Analysis Gene Tree
Overview Taxonomy Strain	ns Genome	s Proteins	Pre	Comparative Systems	Proteome Comparison (B)
DOWNLOAD KEYWORDS			ADV	Analysis (B) BLAST	Metagenomics Metagenomic Read Mapping (B) Taxonomic Classification
Genome Name Acc	ession Feat Ty	ture BRC ID pe		SARS-COV-2 Genome Assembly and Annotation (V) Similar Genome Finder (B)	Metagenomic Binning
Crimean-Congo hemorrhagic fer KJ6	32807 CE	DS fig 1980519	9.748	Meta-CATS	Transcriptomics Expression Import
Crimean-Congo hemorrhagic fer KJ6	32810 CE	OS fig 1980519	9.748	Phylogenetic Tree (B)	RNA-Seq Analysis
Crimean-Congo hemorrhagic fe <sup>1</sup> KJ6	32814 CE	OS fig 1980519	9.74§	Primer Design	Utilities
Crimean-Congo hemorrhagic fe <sup>v</sup> KJ6	32806 CE	OS fig 1980519	9.748	Variation Analysis (B)	ID Mapper
Crimean-Congo hemorrhagic fer KJ6	32812 CE	DS fig 1980519	9.749	Tn-Seq Analysis (B)	Fastq Utilities
Crimean-Congo hemorrhagic fe <sup>1</sup> KJ6	32805 CE 32809 CE	DS fig 1980519 DS fig 1980519	9.747 9.748	SARS-CoV-2 Variant	
Crimean-Congo hemorrhagic fe <sup>,</sup> KJ6	32813 CE	OS fig 1980519	9.74§	Tracker (V)	

- In the Add/Select Genome Group option, go to the Public Workspace folder and then VEME folder and select the "*CCHF\_complete\_M\_human+tick\_curated*" Genome Group
- Click the + button to select
- Perform phylogenetic inferencing using RAXML and the HKY85 model of evolution
- Select "Output Folder" and assign an "Output Name"

Services



Gene Tree (1) Alignment Parameters () Choose fasta file or features for tree. TRIM ENDS OF ALIGNMENT THRESHOLD ONA O PROTEIN 0 -DNA/PROTEIN ALIGNED FASTA REMOVE GAPPY SEQUENCES THRESHOLD ↓<sup>A</sup> Optional Θ 0 -UNALIGNED GENE FASTA L<sup>A</sup> Optional Tree Parameters () FEATURE GROUP L<sup>A</sup> Optional -RAXML O PHYML O FASTTREE AND/OR SELECT GENOME GROUP MODEL ↓<sup>A</sup> CCHF\_complete\_M\_human+tick\_curate ▼ Θ HKY85 🔻 SELECTED FILE/FEATURE TABLE OUTPUT FOLDER CCHF\_complete\_M\_human+tick\_curated × - 5 ↓<sup>A</sup><sub>Z</sub> Genome Groups OUTPUT NAME CCHF\_complete\_M\_human+tick\_curated\_

The Gene Tree Service is being tested. For further explanation, please see the Gene Tree Service Quick Reference Guide and Tutorial.



Monitor progress in the Jobs page

🗙 Uploads 🛛 Jobs ≡ 0 🕑 2 🔗 21

Job Status Last updated: 5:11:21			All Services	🖺 0 queued 🤇	2	2 running v 21 c	ompleted <u>A</u> 3 failed
Status	ID	Service	Output Name	Submit	٠	Start	Completed
running	9221237	GeneTree	CCHF_complete_M_hu	u 8/11/22, 5:05 PM			
completed	9221041	MetaCATS	CCHF_complete_M_hu	u 8/11/22, 1:29 PM		8/11/22, 1:29 PM	8/11/22, 1:31 PM
completed	9220601	GeneTree	CCHF_complete_M_hu	8/11/22, 10:52 AM		8/11/22, 12:25 PM	8/11/22, 12:36 PM
completed	9220544	MetaCATS	CCHF_complete_M_hu	u 8/11/22, 9:33 AM		8/11/22, 9:34 AM	8/11/22, 9:34 AM
completed	9220081	GeneTree	CCHF_complete_M_hu	u 8/10/22, 3:56 PM		8/11/22, 12:10 PM	8/11/22, 12:27 PM
completed	9220080	MSA	CCHF_complete_M_hu	u 8/10/22, 3:52 PM		8/11/22, 1:55 AM	8/11/22, 1:56 AM
completed	9220072	MetaCATS	CCHF complete M hu	a 8/10/22, 3:07 PM		8/11/22, 1:55 AM	8/11/22, 1:56 AM

- When job is complete, select "View" function in green bar
- Select the .phyloxml file and the "View" function in the green bar
- Select the following Display Data options to adjust node labels
  - Host\_group
  - Isolation\_country
  - o Strain
- Adjust External Label Size

- In Visualization:
  - Label color by Isolation\_country
  - Node fill color by host common name
- Adjust color as desired



- Note that the hosts tend to be intermingled whereas the isolation countries match the phylogenetic structures
- Note the main segregation between isolates from India and Russia

## Step 3. Comparative genomic of protein sequences between geographic regions

To determine which amino acids differ between CCHF GPC protein sequences from Indian and Russian isolates, we will use the metaCATS service (<u>MetaCATS</u>, <u>Documentation</u>), which performs statistical analysis of each position in a multiple sequence alignment between selected groups of sequences.

- Use a similar workflow to select protein sequence based on the following criteria for India:
  - o Complete
  - o M segment
  - o Human and Tick
  - India and Russia
- Select "*APPLY*" option above table
- Go to Proteins tab, select GPC proteins
- View MSA Amino Acids
- Curate protein selections to remove problematics sequences (done in advance)
- Save as a Feature Group
- Go to Tools & Services and select Meta-CATS service
- Select "Output Folder" and "Output Name"
- Select Auto Grouping radio button
- Select Isolation Country for Metadata and select the following files from the Public Workspace/VEME/BV-BRC tutorial folder with the + sign:
  - CCHF\_complete\_M\_human+tick\_curated\_India\_Russia\_GPC
- Review group column in table
- Submit job

### Services Metadata-driven Comparative Analysis Tool (meta-CATS) () IN

The meta-CATS tool looks for positions that significantly differ between user-defined groups of sequences. However, biological biases due to covariation, codon biases, and differences in genotype, geography, time of isolation, or others may affect the robustness of the underlying statistical assumptions. For further explanation, please see Metadata-driven Comparative Analysis Tool (meta-CATS) Service Quick Reference Guide and Tutorial.

VALUE			
			0.05
<sup>1</sup> / <sub>2</sub> Feature Groups			-
By Isolation Country			
nput			
AUTO GROUPING O FEA	TURE GROUPS 🔘 ALIGNMENT FILE		
IETADATA			
Isolation Country	Y		
ELECT FEATURE GROUP			
" CCUE complete A	human tick ourstad India		
	1_human+tick_curated_India_	Russia_GPC	
Z       CCHF_complete_N         DNA       PROTEIN         ROUP NAMES	A_human+tick_curated_India_	Russia_GPC	
Ž       CCHF_complete_N         DNA       PROTEIN         \$ROUP NAMES	A_human+tick_curated_India_		0
Z       CCHF_complete_N         DNA       PROTEIN         BROUP NAMES       Delete Rows	A_human+tick_curated_India_	Russia_GPC	- E O
Z       CCHF_complete_N         DNA       PROTEIN         BROUP NAMES       Delete Rows         Main Scoups GRID       Main Scoups GRID	A_human+tick_curated_India_	Russia_GPC	
ZCCHF_complete_N         DNA       PROTEIN         BRC ID       Metada	A_human+tick_curated_India_	Russia_GPC s. Group	Geno <b>f</b> ID
ZCCHF_complete_N         DNA       PROTEIN         BRC ID       Metada         India       India	A_human+tick_curated_India_	Russia_GPC s. Group India	Genoria ID 198051
Z       CCHF_complete_N         DNA       PROTEIN         SROUP NAMES       Main and a standard stan	A_human+tick_curated_India_	Russia_GPC s. Group India India	Genor ID 198051 198051
ZCCHF_complete_N   DNA    PROTEIN   BRC ID   Metada   Ig 198051   India   Ig 198051   India   Ig 198051   India   Ig 198051   India	A_human+tick_curated_India_	Russia_GPC s. Group India India India India	Genoi ID 198051 198051 198051
ZCCHF_complete_N         DNA       PROTEIN         BRC ID       Metada         Ig 198051       India	A_human+tick_curated_India_	Russia_GPC s.  Group India India India India India India India	Genora ID 198051 198051 198051 198051
<ul> <li>Ż CCHF_complete_N</li> <li>DNA PROTEIN</li> <li>PROUP NAMES</li> <li>Delete Rows</li> <li>Matada</li> <li>Fig 198051 India</li> </ul>	A_human+tick_curated_India_	Russia_GPC s. Group India India India India India India India India India	<ul> <li>Genof</li> <li>ID</li> <li>198051</li> <li>198051</li> <li>198051</li> <li>198051</li> <li>198051</li> <li>198051</li> <li>198051</li> <li>198051</li> </ul>

- Monitor progress in the Jobs page
- When job is complete, select "View" function in green bar
- View the GPC\_India\_vs\_Russia\_metaCATS-chisqTable.tsv file
- Sort table by Chi-square value from largest to smallest

rscheuer / home / Feature Groups / .CCHF\_complete\_M\_human+tick\_curated\_India\_Russia\_GPC\_metaCATS\_v2 / CCHF\_complete\_M\_human+tick\_curated\_India\_Russia\_GPC\_metaCATS\_v2-chisqTable.tsv

First Row Conta	ains Column Headers						
Position	Chi-square_value <b>*</b>	P-value	Degrees_of_freedom	Fewer_5	India	Russia	
126	79.993974135199	4.26117356163036e-18	2	Y	42 R	1 N, 37 S	
153	79.993974135199	4.26117356163036e-18	2	Y	42 T	1 P, 37 S	
325	79.993974135199	4.26117356163036e-18	2	Y	42 G	1 N, 37 S	
369	79.993974135199	4.26117356163036e-18	2	Y	42 N	1 S, 37 T	
1610	79.993974135199	4.26117356163036e-18	2	Y	42 1	1 A, 37 V	
40	79.9939741346757	4.26117356274532e-18	2	Y	41 D, 1 N	38 T	
35	79.9939721391984	4.26117781428486e-18	2	Y	42 M	2 I, 36 T	
70	79.9939721391984	4.26117781428486e-18	2	Y	42 L	2 A, 36 T	
108	79.9939721391984	4.26117781428486e-18	2	Y	42 S	36 D, 2 N	
159	79.9939721391984	4.26117781428486e-18	2	Y	42 E	2 A, 36 T	
166	79.9939721391984	4.26117781428486e-18	2	Y	42 S	36 P, 2 T	
227	79.9939721391984	4.26117781428486e-18	2	Y	42 Q	36 M, 2 V	
298	79.9939721391984	4.26117781428486e-18	2	Y	42 T	36 D, 2 E	
107	79.9939721381098	4.26117781660424e-18	2	Y	40 A, 2 V	38 T	
24	79.9939714751559	4.26117922908684e-18	2	Y	42 H	3 P, 35 S	
208	79.9939714751559	4.26117922908684e-18	2	Y	42 L	3 P, 35 S	
876	79.9939714734553	4.26117923271e-18	2	Y	39 M, 3 T	38 V	
1483	79.9939711447461	4.26117993305444e-18	2	Y	42 D	4 N, 34 S	
216	79.9939709479988	4.26118035224225e-18	2	N	42 S	5 L, 33 P	
19	79.9939709449127	4.26118035881761e-18	2	N	5 E, 37 K	38 G	
392	79.9939709449127	4.26118035881761e-18	2	N	37 I, 5 V	38 T	
448	79.9939709449127	4.26118035881761e-18	2	N	37 A, 5 T	38 P	
367	79.9939708182263	4.26118062873445e-18	2	N	42 T	6 L, 32 S	

- Note positions that differ most between the two groups
- Go to Public VEME folder
- Select the CCHF\_complete\_M\_human+tick\_curated\_India\_Russia\_GPC Feature Broup
- Run MSA Amino Acids
- Change ID Type to isolation\_country
- View significant positions (e.g., 126)



## Use Case 3: 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 3a) 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.

Services

## Taxonomic Classification ()

The Taxonomic Classification Service computes taxonomic classification for read data. For further explanation, please see the Taxonomic Classification Service Quick Reference Guide and Tutorial.

Start With:   Read File  ASSEMBLED CONTIGS		
Input File (1) PAIRED READ LIBRARY	<ul> <li>Selected libraries 𝒞</li> <li>Place read files here using the second secon</li></ul>	ne arrow buttons.
↓2 READ FILE 1	SRR10769498	i ×
↓2 READ FILE 2		
SINGLE READ LIBRARY	<b>O</b>	
↓2 READ FILE - ►		
SRA RUN ACCESSION	$\bigcirc$	
SRR10769498		

- 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.

Parameters ()		
ALGORITHM		
Kraken2	*	
DATABASE		
All genomes	-	
SAVE CLASSIFIED SEQUENCES		
NO I YES		
SAVE UNCLASSIFIED SEQUENCES		
INO 🚫 YES		
OUTPUT FOLDER		
↓ <sup>A</sup> +TickTutorial		*
OUTPUT NAME		

- 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.

Status	ID	Service	Output Name	Submit	- :	Start	Completed
completed	6987745	Annotation	Nairoviridae Annotation	3/22/22, 8:57 AM		3/22/22, 8:57 AM	3/22/22, 9:01 AM
completed	6987740	Homology	Test2	3/22/22, 8:55 AM	:	3/22/22, 8:55 AM	3/22/22, 8:55 AM
completed	6987734	GenomeAssembly2	Assembly_CCHFV	3/22/22, 8:51 AM	:	3/22/22, 8:52 AM	3/22/22, 8:52 AM
completed	6987727	FastqUtils	CCHFV_Align	3/22/22, 8:48 AM		3/22/22, 8:48 AM	3/22/22, 8:55 AM
completed	6987720	TaxonomicClassificati	on SRR10769498_Tax	3/22/22, 8:45 AM	:	3/22/22, 8:45 AM	3/22/22, 8:52 AM

- A list of 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)



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	¢
t	Parent folder				-		
Ē	TaxonomicReport.html		19.1 kB	me	Only me	3/22/22, 8:52 AM	۱
	chart.html		2.5 MB	me	Only me	3/22/22, 8:52 AN	1
ľ	classified.fastq.gz		55.1 MB	me	Only me	3/22/22, 8:52 AN	1
	full_report.txt		2.0 MB	me	Only me	3/22/22, 8:52 AN	1
Ê	output.txt.gz		14.9 MB	me	Only me	3/22/22, 8:52 AN	1
	report.txt		427.8 kB	me	Only me	3/22/22, 8:52 AN	1

- 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	Р	2497569	Negarnaviricota
1.24	9386	0	P1	2497571	Polyploviricotina
1.24	9386	0	С	2497576	Ellioviricetes
1.24	9386	0	0	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 3b) Read QC and mapping (<u>Fastq Utilities</u>, <u>Documentation</u>)

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.

Primer Design
Variation Analysis (B)
Tn-Seq Analysis <i>(B)</i>

Utilities	
ID Mapper	
Fastq Utilities	

- 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.

Se	rvi	ce	15

Fastq Utilities () IN

The Fastq Utilities Service provides capability for aligning, measuring base call quality, and trimmiing fastq read files. For further explanation, please see the Fastq Utilities Service Quick Reference Guide and Tutorial.

Parameters ① OUTPUT FOLDER I12 +TickTutorial OUTPUT NAME Step2b_Fastq	Pipeline  Align FastQC Align Karget Genome Compo Hemorrhagic Fever Genome name.
Paired read library 1	Crimean-Congo hemorrhagic fever orthonairovirus IbAr10200 [1980519.4148] Crimean-Congo hemorrhagic fever orthonairovirus Semunya [1980519.5881] Crimean-Congo hemorrhagic fever orthonairovirus Iran/IR-T1-HAS/2014 [1980519.5895] Crimean-Congo hemorrhagic fever orthonairovirus Iran/IR-T2-HAN/2014 [1980519.5896] Crimean-Congo hemorrhagic fever orthonairovirus Iran/IR-T3-HSU/2015 [1980519.5897] Crimean-Congo hemorrhagic fever orthonairovirus Iran/IR-T4-HAS/2015 [1980519.5898] Crimean-Congo hemorrhagic fever orthonairovirus SPU128/81/7 [1980519.5883] Crimean-Congo hemorrhagic fever orthonairovirus SPU128/81/7 [1980519.5893]
Single read library	Crimean-Congo hemorrhagic fever orthonairovirus SPU4/81 [1980519.5894] Crimean-Congo hemorrhagic fever orthonairovirus Irap/IP. T5. HAN/2014/11080510.58001

 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).

<b>aniewiad1 / hor</b> FastqUtils Job F	ne / +TickTutorial / CCHFV_Align (8 items) Result	
Job ID	6987727	FastQC
Start time	3/22/22, 8:48 AM	Align
End time	3/22/22, 8:55 AM	DWNLD
Run time	6m9s	
Parameters		

- 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 🛓

<b><i>R</i></b> FastQC Report			Tue 22 Mar 2022 SRR10769498.fastq
Summary Basic Statistics Per base sequence quality	Basic Statistics	Value	
Per tile sequence quality	File type	Conventional base calls	
Por soquepeo quality secres	Encoding	Sanger / Illumina 1.9	
Per sequence quaity scores	Total Sequences	757773	
Per base sequence content	Sequences flagged as poor quality	0	
Per sequence GC content	Sequence length	160	
Per base N content	1GC	49	
Sequence Length Distribution Sequence Duplication Levels	Per base sequence qualit	<b>y</b> pres across all bases (Sanger /	Illumina 1.9 encoding)
	38 36 34 32 30 28 26 24 23		

- For the results of the "Align" pipeline, a summary of reads, read length, and base quality is displayed (note: only 0.5% of reads are mapped to the target CCHFV genome previously specified).

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



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

	Number	Percentage
MAPQ >= 30	3013.0	0.4
MAPQ < 30	407.0	0.1
MAPQ < 20	0.0	0.0
MAPQ < 10	287.0	0.0
MAPQ < 3	0.0	0.0
Unmapped	754066.0	99.5
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 3c**".

aniewiad1 / home / FastqUtils Job Res	/ +TickTutorial / CCHFV_Align (8 items) sult				VIEW	SHOV
Job ID	6987727					GUIDE
Start time	3/22/22, 8:48 AM					
End time	3/22/22, 8:55 AM					DWH
Run time	6m9s					Till
b Damanalana						DELET
<ul> <li>Parameters</li> </ul>						
• Parameters						Ø
Parameters						RENA
• Parameters						RENAR
• Parameters						RENA RENA COP
Name	•	Size	Owner	Members	Created	
Name f Parent folder	•	Size	Owner	Members	Created	
Name Parant folder SRR10769498.	*	Size 52.7 kB	Owner	Members - Only me	Created 3/22/22, 8:54 AM	
Name Paranteters Parant folder SRR10769498. SRR10769498.	▲ aligned.bam aligned.bam.bai	Size 52.7 kB 96 B	Owner me me	Members - Only me Only me	Created 3/22/22, 8:54 AM 3/22/22, 8:54 AM	
Name Paranteters Parent folder SRR10769498. SRR10769498.	aligned.bam aligned.bam.bai aligned.fq.gz	Size 52.7 kB 96 B 107.7 kB	Owner me me me	Members - Only me Only me	Created 3/22/22, 8:54 AM 3/22/22, 8:54 AM 3/22/22, 8:54 AM	
Name           Parent folder           SRR10769498.           SRR10769498.           SRR10769498.           SRR10769498.	aligned.bam .aligned.bam.bai .aligned.fq.gz .ali.bam.samstat.html	Size 52.7 kB 96 B 107.7 kB 232.2 kB	Owner me me me me	Members - Only me Only me Only me	Created 3/22/22, 8:54 AM 3/22/22, 8:54 AM 3/22/22, 8:54 AM 3/22/22, 8:55 AM	

## Step 3c) de novo sequence assembly (<u>Genome Assembly Service</u>, <u>Documentation</u>)

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

T	OOLS & SERVICES	WORKS
1	Genomics	
F	Assembly (B)	F.
	Annotation	

- 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.

#### Services

## Genome Assembly () IN

The Genome Assembly Service allows single or multiple assemblers to be invoked to compare results. The service attempts to select the best assembly. For further explanation, please see the Genome Assembly Service Quick Reference Guide and Tutorial..

Paired read library ()	Θ	Selected libraries	ns.
↓2 READ FILE 2 ↓2	- 1	S(SRR1076949gned.fq.gz)	i ×
ADVANCED -	(0)		
READ FILE ↓ <sup>A</sup> SRR10769498.aligned.fq.gz ADVANCED ▼			
SRA run accession	Θ		
SRR			

- 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.

Parameters ()
ASSEMBLY STRATEGY
Auto
Auto
Unicycler
SPAdes
Canu
metaSPAdes
plasmidSPAdes
MDA (single-cell)

- 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.

anie	aniewiad1 / home / +TickTutorial / Assembly_CCHFV (3 items)											
Geno	omeAssembly2 J	ob Result						HIDE				
Job I	D	6987734						GUIDE C				
Start	time	3/22/22, 8:52 AM						+				
End t	time	3/22/22, 8:52 AM						DWNLD				
Run t	time	23s						0				
► Pa	arameters							VIEW				
								匬				
								DELETE				
								Ø				
	Name		•	Size	Owner	Members	2					
ı t	Parent folder						Copy selected objects	COPY				
	Assembly_CCHFV	assembly_report.html		17.5 kB	me	Only me	3/22/22, 8:52 AM	<b>→</b>				
-	Assembly_CCHFV	_contigs.fasta		4.7 kB	me	Only me	3/22/22, 8:52 AM	MOVE				
	details				me	Only me	3/22/22, 8:52 AM					

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

Genoi	meAssembly2 Jo	b Result									
Job ID	Job ID 6987734										
Start t	time	3/22/22, 8:52 AM									
End ti	me	3/22/22, 8:52 AM					Download				
Run ti	me	23s									
▶ Par	rameters										
N	lame		•	Size	Owner	Members	Created O				
t P	Parent folder					-					
A 1	ssembly_CCHFV	_assembly_report.html		17.5 kB	me	Only me	3/22/22, 8:52 AM				
<b>≖</b> A	ssembly_CCHFV	_contigs.fasta		4.7 kB	me	Only me	3/22/22, 8:52 AM				
h d											
u	letails				me	Only me	3/22/22, 8:52 AM				

## Step 3d) 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.

#### TOOLS & SERVICES WORK

Genomics

Assembly (B) Annotation **Comprehensive Genome** Analysis (B) BLAST

- 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.

Services

## BLAST () IN

The BLAST service integrates the BLAST (Basic Local Aligment Search Tool) algorithms to perform searches against public or private genomes or other reference databases using DNA or protein sequence(s). For further explanation, please see BLAST Service Quick Reference Guide and Tutorial.

#### Search program ()

- BLASTN (nucleotide > nucleotide database)
- BLASTP (protein > protein database)
- BLASTX (translated nucleotide > protein database)
- tBLASTn (protein > translated nucleotide database)
- Query source

```
Enter sequence Select FASTA file Select feature group
```

>SRR10769498 aligned assembly contig 1 length 3405 coverage 59.1 normalized cov 0.96 AGTAGTCTAGGTCACAACCATCCCAGGACATCCAGGAGGTGTTGAAGTGTGGCTCAATTT TATGAATTAGATGTCCATTTACCTTATCCCCTTTTAATAAATTTCCGATGTGGTAGACCT GTAGGTCTCCTGGCACACCATGTGTGCAACTCTGCAACTTACACACTGTGCTTGCCGATA ACACCTTTTGCACATGCATCAGGTCAAAAAAACCTTCTTCGATCCTAGGATGCAGTGTGA TTATTTCAGGAGGGAGTTTTTGTTGGATGTTTCTTGGTTCTGACAGTGTGATGGTCACAG GACCTAAATTGAACCTTGTGCCCGCTTCAATCAAGCTACACTGCCTTTCCTGACTAGTAA GTTCTACACACACTATGGCCTCTGTCTTGATGTATTCAACTTTCCACTTGACAAACATAT AATCTGTAAAAAGGTCTTTCACATCTAATCCACAACAGGTGCAGCCAGTCCCTACACCCC

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"

Database Source	Database Type
Reference and representative genomes (virus)	Contigs (NT)
Reference and representative genomes (bacteria, archaea)	
<sup>O</sup> Reference and representative genomes (virus)	
Search within selected genome list	1 ( ) ( ) ( ) ( ) ( ) ( ) ( ) ( ) ( ) (
Search within selected genome group	
Search within selected feature group	
Search within a taxon	-
Search within 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	ldentit (%)	Query cover (%)	Subjec cover (%	Query Length	Subject Length	Score	E O value
٠	SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	NC_005300	segment	100	100.00	63.27	3405	5366	6219	0
٠	SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	KY484035	segment	100	100.00	63.27	3405	5366	6202	0
۲	SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	KX013451	segment	98	100.00	63.62	3405	5336	5864	0
٠	SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	MF287637	segment	97	100.00	63.25	3405	5368	5781	0
٠	SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	MF547416	segment	97	99.94	63.23	3405	5366	5633	0
٠	SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	KY484045	segment	96	100.00	63.28	3405	5365	5521	0
٠	SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	KY484042	segment	96	100.00	63.28	3405	5365	5499	0
٠	SRR10769498_aligned_asser	Crimean-Congo hemorrhagic fever ortho	KJ682814	segment	96	100.00	63.29	3405	5364	5493	0
٠	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 3e) 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 (1) III

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.

12 SRR10769498_aligned_assei	mbly_contigs.fasta 👻 🗖
ANNOTATION RECIPE	
Viruses	-
	TAXONOMY ID
Nairoviridae	1980415
MY LABEL	
Step2e_annotation	
OUTPUT NAME	
Nairoviridae Step2e_annotation	
OUTPUT FOLDER	

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

aniewiad1 / home / +TickTutorial / Nairoviridae Annotation (24 items)	VIEW	CDS CDS	BROWSER
GenomeAnnotation Job Result	View Ar	notated Ge	enome
	1		J

- 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.

Overview	Genome Browser	Proteins	Protein Structures	Domains and Motifs	Experiments	Interactions							
) Nairov	iridae Annotatio	on	Sedit (	Genomic Features					Jul :=	\$	īdī	2+	0
							PATRIC	•	RefSeq	Services	Add To Group	Share	GUIDE
Length: 449	92bp, Contigs: 2			CDS				2	0	External To	ools		
General Info										BEI Resourc	es		
Genome ID	1	1980415.2	25							Recent Pul	bMed Artic	les	
Genome Na	ame	Nairovirida	e Annotation							<ul> <li>No recent a</li> </ul>	articles found.		
Taxonomy In	ifo												
Taxon ID		1980415											
Superking	dom	Viruses											
Kingdom		Orthornavi	irae										
Phylum		Negarnavi	ricota										
Class		Ellioviricet	es										
Order		Bunyavira	les										
Family		Nairovirida	ie										

G Geno	ome View es » Negarnavirico	ta » Elliovirice	tes » Bun	yavirales » Na	iroviridae » <mark>N</mark>	airovirida	e Annotati	on			
Overview	Genome Brows	er Proteins	Prote	in Structures	Domains a	nd Motifs	Experimen	nts I	Interactions		
	KEYWORDS			AD	Search FEAT	CDS	PAT ANNOTA	TION X	t The prot	"Interaction ein interaction putational a	ns" tab shows a list of protein- tions, inferred using and laboratory methods.
Genome	Name	Accession	Feature Type	BRC ID	RefSeq Locus Tag	Start	End	Stra	nd Length (NA)	Gene Symbol	Product
Nairovirida	e Annotation	1980415.25.cor	CDS	fig 1980415.25.		1	3405	-	3401	GPC	putative Pre-glycoprotein polyprotein
Nairovirida	e Annotation	1980415.25.cor	CDS	fig 1980415.25.		1	951	+	951	GPC	putative Pre-glycoprotein polyprotein C

# THANKS FOR FOLLOWING ALONG, AND PLEASE CONTACT US WITH YOUR QUESTIONS AT <u>BV-BRC.ORG</u>!

<b>CBV-BRC</b> <sup>BEIA</sup> ORGANISMS SEARCHES TOOLS & SERVICES WORKSPACES HE	ELP ABOUT
<b>BACTERIAL AND VIRAL BIOINFORMATICS RESOURCE CE</b> Welcome to the Bacterial and Viral Bioinformatics Resource Center (BV-BRC), an information sy bacterial and viral infectious diseases. Learn more about BV-BRC. Introductory material for tran is provided in the sections at the bottom of this page. This is the <i>Beta Version</i> of the website, designed to allow users to explore, try out features, and page.	Quick Start Quick References Tutorials Common Tasks CLI Tutorial Webinars Instructional Videos Workshops Contact Us