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

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 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 https://www.bv-brc.org and click on “Viruses”

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
• Note the taxonomy expansion and the number of genomes selected
• Use the “FILTER” button to select the following criteria and click “APPLY”:
  o Genome Status = Complete
  o Segment = M
  o Host Common Name = human AND tick

• 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”
• 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.

• 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
• **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 - **CCHF_complete_M_human+tick_curated**)

Step 2. Phylogenetic analysis

• Go to Tools & Services and select Gene Tree service

![Gene Tree Service](image)

• 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**”
Monitor progress in the Jobs page

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
  o Host_group
  o Isolation_country
  o Strain
Adjust External Label Size

Gene Tree Service is being tested. For further explanation, please see the Gene Tree Service Quick Reference Guide and Tutorial.
• In Visualization:
  o Label color by isolation_country
  o 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 (MetaCATS, Documentation), 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:
  - Complete
  - M segment
  - 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
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
• 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.
- 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 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".
- 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.
17

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

**Step 3b) 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).

- 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).
At this point, you may download the aligned reads or use them for further analysis, as shown in “step 3c”.
Step 3c) 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.

```
<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Owner</th>
<th>Members</th>
<th>Created</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly_CCHFV_contigs.fasta</td>
<td>4.7 kB</td>
<td>me</td>
<td>Only me</td>
<td>3/22/22, 8:52 AM</td>
</tr>
<tr>
<td>Assembly_CCHFV_asm_report.html</td>
<td>17.5 kB</td>
<td>me</td>
<td>Only me</td>
<td>3/22/22, 8:52 AM</td>
</tr>
</tbody>
</table>
```

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

```
<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Owner</th>
<th>Members</th>
<th>Created</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly_CCHFV_contigs.fasta</td>
<td>4.7 kB</td>
<td>me</td>
<td>Only me</td>
<td>3/22/22, 8:52 AM</td>
</tr>
<tr>
<td>Assembly_CCHFV_asm_report.html</td>
<td>17.5 kB</td>
<td>me</td>
<td>Only me</td>
<td>3/22/22, 8:52 AM</td>
</tr>
</tbody>
</table>
```

**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.
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.

**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.
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!