Guidance Document for Generating a Nextclade Dataset

Principle/Purpose

This Guidance Document details the generation of a pathogen-specific dataset for use within Nextclade, a strain-typing tool for pathogen genomic sequences, to allow strain typing of that pathogen from genomic sequence data. A dataset for Hepatitis A (HAV) is used as an example.

Scope

This document describes the process of generating a Nextclade dataset for a pathogen of interest. It is intended for use by public health laboratories within the Northeast (NE) region through the NE Bioinformatics & Laboratory Training Lead and the NE Pathogen Genomic Centers of Excellence (NE PGCoE) sites but can be followed by anyone.

Materials Required

1) Computer
2) Internet access
3) Web browser (such as Google Chrome)
4) Terra.bio account and workspace
5) GitHub account
6) Command line with Perl
7) Sequence alignment software (such as MAFFT or Geneious)
8) Tree-building software (such as FastTree or IQ-TREE or Geneious)
9) Plain text editor (such as Notepad or Visual Studio Code)
10) Microsoft Excel
11) Download the helper scripts located here:
    - https://github.com/MASPHL-Bioinformatics/Northeast-BTL/tree/main/Nextclade_Dataset_Guidance/scripts

1. Introduction to Nextclade

1. As an example input fasta, download the ten most recent SARS-CoV-2 sequences uploaded to GenBank:
    a. Search for “SARS-CoV-2” in NCBI’s Taxonomy Browser (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi).
    b. On the right, click the link in the “Nucleotide” row and “Subtree links” column. This will take you to NCBI’s Nucleotide Browser.
    c. Click “Sort by Default order” and select “Date Released”.
    d. Select the ten most recent sequences.
    e. Click “Send to:”. Check “Complete Record”, destination “File,” and Format “FASTA”. Click “Create File.”
2. Navigate to https://clades.nextstrain.org
3. Drag your example input fasta into the box under “Provide sequence data”, or click “Select files” and select your example input fasta.
4. Under “Select reference dataset”, click “Select reference dataset”.
5. Select “SARS-CoV-2”.
6. Click “Run”.
7. You will find yourself in the “Results” tab. View your sequences’ assigned lineages in the “Clade” column.
8. Click the “Tree” tab. Here, you can view your new sequences in the context of existing sequences.

9. Under “Color By”, select “Node type”. Your input sequences will be highlighted, while the context sequences will not be.
10. Under “Color By”, select “Clade”. Your input sequences and the context sequences will be colored by assigned clade.
11. Click the “Export” tab.
12. nextclade.csv and nextclade.tsv are tables including the assigned clades of your input sequences, similar to what you saw in the “Results” tab.
13. nextclade.auspice.json is the tree you saw in the “Tree” tab. You can view this tree by downloading it and dragging it into https://auspice.us.

2. Retrieve Sequences and Metadata for Nextclade Dataset

1. Retrieve all sequences belonging to your pathogen species and their associated metadata from GenBank. (If preferred, another source can be used.)
   a. Search for your pathogen species in NCBI’s Taxonomy Browser (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi).
   b. On the right, click the link in the “Nucleotide” row and “Subtree links” column. This will take you to NCBI’s Nucleotide Browser.
   c. Download all sequences belonging to your pathogen species. Click “Send to:”. Check “Complete Record”, destination “File,” and Format “FASTA”. Click “Create File.”
   d. Download the metadata attached to all sequences belonging to your pathogen species. Click “Send to:”. Check “Complete Record”, destination “File,” and Format “GenBank”. Click “Create File.”
   e. Move and rename your files.
   f. Shorten sequence headers to accession numbers only. Enter in your terminal:
      ```
      perl shorten_headers_cut_at_first_space.pl [fasta file] >
      [fasta file]_shortened_headers.fasta
      ```
2. Convert your GenBank metadata file to a table. Enter in your terminal:
   ```
   perl retrieve_metadata_from_GenBank_file.pl [GenBank .gb file] >
   [GenBank .gb file]_metadata.txt
   ```
3. If desired for filtering, add the number of unambiguous bases to your metadata table. Enter in your terminal:
   ```
   perl add_number_unambiguous_bases_to_table.pl [fasta
   file]_shortened_headers.fasta [GenBank .gb file]_metadata.txt >
   [GenBank .gb file]_metadata.txt_with_number_unambiguous_bases.txt
   ```
4. Convert dates to YYYY-MM-DD format. Enter in your terminal:
   ```
   perl dates_in_columns_to_YYYY_MM_DD.pl [GenBank .gb
   file]_metadata.txt collection_date > [GenBank .gb
   file]_metadata.txt_date_formatted.txt
   ```

5. For sequences for which the clade is known, manually add known clades to your metadata table. Information about the known clade might be found in the “note” column of the automatically generated metadata table. In this example, we name this column “known_clade”.
6. Select sequences to include in the Nextclade dataset.
    a. Open your supplemented metadata table ([GenBank .gb file]_metadata.txt_with_number_unambiguous_bases.txt) in Excel.
    b. Highlight the first row. Click the Data tab, then click Filter.
    c. Click the triangle in the number_unambiguous_bases header to filter by number unambiguous bases. In our example, for whole-genome HAV sequences, we use a cutoff of 95% of the length of the genome (7,125 bases).
    d. Optionally, fill in the host where it is missing. Click the triangle in the host header, uncheck “(Select All)”, and select “(Blank)”. Examine the paper title associated with each sequence, listed in the title column. Where the paper title is not revealing, read the abstract of the paper by copying the title of the paper into https://pubmed.ncbi.nlm.nih.gov.
7. Filter down the metadata and sequence files to included sequences only.
    a. Filter your metadata table to included sequences. If filtering by host, for example, click the triangle in the host header, check and then uncheck “(Select All)”, and select Homo sapiens or other values describing human host.
    b. Save your filtered metadata table. Click the top left cell, command A to select the full table, and copy the table into a plain text file. Save your text file. (Clicking File, Save As, and choosing format “Tab delimited text (.txt)” will save the full table, not the filtered table.)
    c. Filter your sequences. Highlight the filtered sequence accession numbers in your table into a plain text file and save it. In this example, we name the file selected_sequences.txt. Enter in your terminal:
       ```
       perl retrieve_sequences_by_names_listed_in_file.pl [fasta
       file] [list of accession numbers] > [filtered fasta file]
       ```
8. Open your metadata table in a text editor and update the column titles. The first column, containing accession numbers, must be changed to “strain”. The column containing collection dates must be changed to “date”.

3a. Generate Phylogeny for Nextclade Dataset (Rooting on Outgroup)

1. Select the nearest outgroup. In our example, we will only include human host sequences, so we select the non-human host sequence nearest to the human host sequences.

   a. Filter your sequences to only potential outgroups. To view non-human host sequences, for example, click the triangle in the host header, check and then uncheck “(Select All)”, and select values describing non-human hosts.
   b. Highlight the filtered sequence accession numbers in your table, copy/paste them into a plain text file, and save your text file.
   c. Add the sequence accession number of an ingroup sequence (in our example, a human host sequence) to the top of your list.
   d. Retrieve the sequences of your potential outgroups and your one example ingroup. Enter in your terminal:

      ```
      perl retrieve_sequences_by_names_listed_in_file.pl [fasta
      file] [list of accession numbers] > [potential outgroups
      and one ingroup fasta file]
      ```
   e. Use MAFFT or another aligner to align your sequences. Enter in your terminal:

      ```
      mafft [potential outgroups and one ingroup fasta file] >
      [potential outgroups and one ingroup aligned fasta file]
      ```
   f. Identify the outgroup closest to your example ingroup sequence. Enter in your terminal:

      ```
      perl select_closest_outgroup.pl [potential outgroups and
      one ingroup aligned fasta file]
      ```
   g. In your metadata table opened in Excel, annotate your outgroup in a field you will be filtering by. In our example, we mark our selected outgroup’s host as “non-human outgroup”.

2. Use MAFFT or another aligner to align your sequences and your outgroup to the reference, using --add in order to avoid introducing gaps in the reference. Enter in your terminal:

   ```
   mafft --add [human host sequences and outgroup] --reorder --
   keeplength [reference fasta file] > [aligned human host sequences
   and outgroup]
   ```
3. Open your aligned file in a text editor application and manually remove your reference sequence.
4. Use FastTree, IQ-TREE, or another tree-building software to generate a phylogeny of your sequences and your outgroup. Enter in your terminal:

   ```
   FastTree -nt < [aligned human host sequences and outgroup] >
   [phylogeny of human host sequences and outgroup]
   ```
5. Use your outgroup to reroot your phylogeny, and then remove the outgroup from your tree. Enter in your terminal:

   ```
   Rscript reroot_tree_and_remove_outgroup.R [newick tree] [outgroup
   name] > [rerooted phylogeny without outgroup]
   ```

3b. Generate Phylogeny for Nextclade Dataset (Midpoint Rooting)

1. Use MAFFT or another aligner to align your sequences to the reference, using --add in order to avoid introducing gaps in the reference. Enter in your terminal:
   ```
   mafft --add [human host sequences] --reorder --keeplength
   [reference fasta file] > [aligned human host sequences]
   ```
2. Open your aligned file in a text editor application and manually remove your reference sequence.
3. Use FastTree, IQ-TREE, or another tree-building software to generate a phylogeny of your sequences. Enter in your terminal:
   ```
   FastTree -nt < [aligned sequences] > [phylogeny of sequences]
   ```
4. Midpoint root your tree. Enter in your terminal:
   ```
   Rscript midpoint_root_tree.R Rscript midpoint_root_tree.R [newick
   tree] > [midpoint rooted tree]
   ```

4. Import Phylogeny into Augur

1. Open your Terra workspace.
2. Add the augur_from_mltree workflow to your Terra workspace:
   a. Click “WORKFLOWS” (at the top of the window).
   b. Click “Find a Workflow +” (at the top left of the displayed workflows).
   c. Click “Dockstore” at the bottom of the pop-up.
   d. Search for augur_from_mltree.
   e. Click “broadinstitute/viral-pipelines/augur_from_mltree”.
   f. On the left, under “Launch with”, select “Terra”.
   g. Under “Destination Workspace”, select your workspace name.
3. Generate your auspice_config file. Copy/paste the following into a plain text editor, edit it to suit your own project, and save it:
   ```
   {
     "title": "Hepatitis A human host GenBank WGS",
     "maintainers": [
       {
         "name": "NE PGCoE",
         "url": "https://www.nepgcoe.org"
       }
     ],
     "panels": [
        "tree",
        "entropy",
        "frequencies"
   ```

```
  ],
  "colorings": [
    {
      "key": "known_clade",
      "title": "known_clade",
      "type": "categorical"
    },
    {
      "key": "geo_loc_name",
      "title": "Location",
      "type": "categorical"
    },
    {
      "key": "date",
      "title": "Date",
      "type": "temporal"
    }
  ]
}
```

4. Navigate to the RefSeq sequence(s) for your pathogen.
    a. Option A: from the RefSeq browser:
        i. In your web browser, navigate to https://www.ncbi.nlm.nih.gov/refseq/
        ii. In the textbox, enter your pathogen name. Click “Search”.
        iii. Select the item(s) corresponding to your pathogen. You can either check the checkboxes or click the links.
    b. Option B: from the taxonomy browser:
        i. In your web browser, navigate to https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi
        ii. In the textbox to the right of “Search for”, enter your pathogen name. Click “Go”.
        iii. Select your pathogen.
        iv. Click the first number to the right of “Datasets Genome” (at the top right of the window).
        v. For the assembly that says “NCBI RefSeq” under “Annotation”, click its link in the “Assembly” column.
        vi. Click “view RefSeq sequences” under “Assembly statistics”.
5. Download your ref_fasta file and your genbank_gb file.
    a. At the top right, click “Send to:”.
    b. Check “Complete Record”.
    c. Under “Choose Destination”, select “File”.
    d. Under “Format”, select “GenBank”.
    e. Click “Create File”.

   f. Save the file. This is your genbank_gb file.
   g. Click “Send to:” again, this time selecting “FASTA” under “Format”. Click “Create File”.
   h. Save the file. This is your ref_fasta file.
   i. Open your ref_fasta file and remove the description after the accession number in the header. The remaining header should be just the accession number.
6. Generate your msa_or_vcf alignment file.
   a. Copy your fasta file containing your sequences and your outgroup sequence. Open the new copy in a text editor and remove the outgroup sequence. Copy in the reference sequence from the ref_fasta file.
   b. Use MAFFT or another sequence aligner to align your ref_fasta sequence and your sequences (not including your outgroup sequence if you used one). Enter in your terminal:

```
mafft --add [sequences] --reorder --keeplength [ref_fasta
file] > [alignment to reference]
```

7. Note your raw_tree phylogeny file from the preceding steps.
8. Note your sample_metadata metadata table file from the preceding steps.
9. Upload your input files:
   a. In Terra, click “DASHBOARD” (at the top of the window).
   b. On the right, click “Open bucket in browser”.
   c. Click “CREATE FOLDER”.
   d. Name your new folder.
   e. Click “UPLOAD”. Click “Upload files”.
   f. Select the input files.
10. Run augur_from_mltree:
   a. Click “WORKFLOWS” (at the top of the window).
   b. Click “augur_from_mltree”. If “augur_from_mltree” is not available, click “Find a Workflow +”, click “Dockstore.org”, and search “augur_from_mltree”. Click “broadinstitute/viral-pipelines/augur_from_mltree”. Under “Launch with”, click “Terra”. Select your workspace and click “Import”.
   c. Select “Run workflow with inputs defined by file paths”.
   d. For “auspice_config”, enter the gs:// address of your auspice_config file, in double quotes. To find the gs:// address of your file, click on it in your bucket and copy the “gsutil URI” address.
   e. For “genbank_gb”, enter the gs:// address of your genbank_gb file, in double quotes.
   f. For “msa_or_vcf”, enter the gs:// address of the alignment of your human host sequences and non-human host outgroup aligned to the reference, in double quotes.
   g. For “raw_tree”, enter the gs:// address of your phylogeny, in double quotes.
   h. For “ref_fasta”, enter the gs:// address of your reference, in double quotes.

- i. For “root_sequence”, enter the gs:// address of your reference, in double quotes. (Nextclade requires the root and the reference to be the same sequence. Adding your reference as “root_sequence” here catalogues the mutations between the inferred root and your reference and adds the reference upstream of the inferred root, as the new root.)
- j. For “sample_metadata”, enter the gs:// address of your metadata table, in double quotes.
- k. For “generate_timetree”, enter false.
- l. For “keep_polytomies”, enter true.
- m. Click “Save” (at the top right).
- n. Click “Launch”.

11. Download your augur tree:
    - a. When your job is done, it will appear in green in the “SUBMISSION HISTORY” tab. Click your submission.
    - b. Click the first icon under “Links” (three horizontal lines with two checkboxes and a dot to the left of them). A new window will open. Sign in if prompted.
    - c. Click “OUTPUTS”.
    - d. Click the gs:// file link to the right of “auspice_input_json”.
    - e. Download the file by clicking the down arrow at the right of the file’s row. If a new window opens, save the file by clicking “File” → “Save Page As…”
    - f. Name your augur tree file. In this example, we name the file 2024_05_06_txid12092_human_host_and_outgroup_7125_bases_aligned.fasta _phylogeny_rerooted_auspice.txt
12. Open your augur tree:
    - a. In your browser, navigate to https://auspice.us
    - b. Drag your auspice phylogeny file onto the browser window or open it by clicking “Select files”.

5. Generate Clade Assignment File

1. Create a new plain text file to store your mutations.
2. In your plain text file, write the name of each clade preceded by “CLADE ”, one clade per line.
3. As before, open your augur tree in https://auspice.us
4. Color your nodes by the known clade by selecting “known_clade” under “Color By” (in the left panel).
5. For each clade:
    - a. Hover your mouse over the branch leading to all nodes known to be in the clade.
    - b. Hold down the shift key and click. You should see a pop-up with “Mutations observed on branch”.

   c. Scroll down to the bottom of the pop-up and click the link: “Click to copy all mutations to clipboard as TSV”.
   d. Paste the defining mutations in your plain text file under the clade’s name.
6. Verify that your mutations file looks correct. Here is an example mutations file:

```
CLADE III
nuc   unique     T12C, T176G, A333G, G618A, C638T, T640G
nuc   homoplasies     T112C, C180T, T546C
nuc   undeletions     -122A
nuc   gaps G20-, T204-
HAVgp1     unique     V63S
HAVgp1     homoplasies     K77R
HAVgp2     unique     V61S, S283A, D337E, S389H
HAVgp2     homoplasies     K75R, K1142R, I1385V

CLADE IIIA
nuc   unique     T24C, G535A, T845C, T921C, T1055C, T1145C, A1172T
nuc   homoplasies     T171C, A177G, C375T, C566T, T604C, G818A
nuc   reversionsToRoot G818A, G1076A, C1812T, G1865A
nuc   undeletions     -120T
HAVgp1     unique     S63P, V1123L, N1138H
HAVgp1     homoplasies     R520K, K525R, V533I, K1449R, E1804D
HAVgp2     unique     S61P, V1121L, N1136H, S1388N
HAVgp2     homoplasies     R518K, K523R, V531I, K1447R, E1802D
```

7. Convert your mutations file into your clades.tsv file. Enter in your terminal:

```
generate_clades_tsv_file.pl [file with copy/pasted defining
mutations for each clade] > [output clades.tsv file path]
```

6. Import Phylogeny into Augur With Clade Assignment File

1. Open your Terra workspace.
2. Upload your clades.tsv file to your workspace’s bucket as you previously uploaded files.
3. Fill in the inputs to augur_from_mltree as before.
4. Fill in your clade.tsv’s gs:// address in the optional clades_tsv field, in double quotes.
5. Click “Save” and “Run Analysis”.
6. When your job finishes, download your augur tree file as before.
7. Open your augur tree in https://auspice.us. You can now see an additional “Color By” option, “Clade”. Select it and view your clade-annotated tree.
8. Verify that the automatic clade assignment assigned the correct clades by toggling between “Clade” and “known_clade” in “Color By”.

7. Prepare Phylogeny for Nextclade

1. Remove the "known_clade" column from your metadata table. Select columns to include by entering the following into your terminal:

```
perl retrieve_subset_of_columns.pl [GenBank .gb
file]_metadata.txt_date_formatted.txt strain date authors title
geo_loc_name > [GenBank .gb
file]_metadata.txt_date_formatted.txt_selected_columns.txt
```

2. Remove the "known_clade" coloring from your auspice_config file. Here are the lines you should remove:

```
      {
       "key": "known_clade",
       "title": "known_clade",
       "type": "categorical"
     },
```

In our example, the resulting auspice_config file looks like this:

```
{
  "title": "Hepatitis A human host GenBank WGS",
  "maintainers": [
    {
      "name": "NE PGCoE",
      "url": "https://www.nepgcoe.org"
    }
  ],
  "panels": [
     "tree",
     "entropy",
     "frequencies"
  ],
  "colorings": [
    {
      "key": "geo_loc_name",
      "title": "Location",
      "type": "categorical"
    },
    {
      "key": "date",
      "title": "Date",
      "type": "temporal"
    }
  ]
}
```

3. Save your updated auspice_config file and your updated metadata table with new file names and upload them to your workspace's bucket as you previously uploaded files. (If an updated file has the same filepath as it did in a previous Terra run, Terra will not recognize it as a new file.)
4. Fill in the inputs to augur_from_mltree as before.
5. Update your auspice_config file's gs:// address and your sample_metadata table's gs:// address.
6. Click "Save" and "Run Analysis".
7. When your job finishes, download your augur tree file as before. This tree is your reference tree file for your Nextclade dataset.

8. Compose Nextclade Dataset

1. Create a new directory named data. Within that data directory, create a second new directory with a name that describes your Nextclade dataset, for example HAV-WGS.
2. Note your reference tree file from the previous step. Copy it to your new directory, and rename the reference tree file copy to tree.json.
3. Note your reference fasta file from the previous step. Copy it to your new directory and rename the copy to reference.fasta.
4. Create the empty files CHANGELOG.md and README.md in your new directory.
5. Navigate to the RefSeq sequence(s) for your pathogen.
    a. Option A: from the RefSeq browser:
        i. In your web browser, navigate to https://www.ncbi.nlm.nih.gov/refseq/
        ii. In the textbox, enter your pathogen name. Click "Search".
        iii. Select the item(s) corresponding to your pathogen. You can either check the checkboxes or click the links.
    b. Option B: from the taxonomy browser:
        i. In your web browser, navigate to https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi
        ii. In the textbox to the right of "Search for", enter your pathogen name. Click "Go".
        iii. Select your pathogen.
        iv. Click the first number to the right of "Datasets Genome" (at the top right of the window).
        v. For the assembly that says "NCBI RefSeq" under "Annotation", click its link in the "Assembly" column.
        vi. Click "view RefSeq sequences" under "Assembly statistics".
6. Download your genome_annotation.gff3 file to your new directory:
    a. At the top right, click "Send to:".
    b. Check "Complete Record".
    c. Check "File" under "Choose Destination".

   d. Under “Format”, select “GFF3”.
   e. Click “Create File”.
   f. Save the file as genome_annotation.gff3.
   g. In our particular example, for HAV, the gene, transcript, and exon (region is fine) from 1 to 7,478 need to be removed from the genome_annotation.gff3 file.

7. Create your pathogen.json file. Create an empty plain text file named pathogen.json in your new directory, copy/paste the below, and adjust the attributes section to fit your project.

```
{
  "schemaVersion": "3.0.0",
  "files": {
    "reference": "reference.fasta",
    "pathogenJson": "pathogen.json",
    "genomeAnnotation": "genome_annotation.gff3",
    "treeJson": "tree.json",
    "readme": "README.md",
    "changelog": "CHANGELOG.md"
  },
  "attributes": {
    "name": "Hepatitis A human host GenBank WGS",
    "reference name": "Hepatitis A virus",
    "reference accession": "NC_001489.1"
  }
}
```

9. Run Nextclade From the Command Line

1. Download the Nextclade executable file here: https://docs.nextstrain.org/projects/nextclade/en/stable/user/nextclade-cli/installation/standalone.html
2. Move the Nextclade executable file to the directory above your “data” directory. In this example, the directory above our “data” directory is named “HAV”.
3. Copy a fasta file containing sequences you would like to classify to the directory above your “data” directory. For testing, use your previously generated sequence file with all included sequences. In this example, we named this file 2024_05_06_txid12092_human_host_and_outgroup_7125_bases.fasta
4. In the directory above your “data” directory, create an output directory. In this example, we name our output directory output_WGS_on_WGS.
5. Within your “data” directory, create a new directory. Its name is the name of your Nextclade dataset. In this example, we name the directory (and the Nextclade dataset) “HAV-WGS”.

6. Move your Nextclade dataset files to your new Nextclade dataset directory.
7. Examine your file structure. It should look something like this example:
   ```
   HAV >
       data >
           HAV-WGS >
               CHANGELOG.md
               genome_annotation.gff3
               pathogen.json
               README.md
               reference.fasta
               tree.json
       example_whole_genome_sequences.fasta
       output_WGS_on_WGS >
   ```
8. In your terminal, navigate to the directory above your “data” directory using cd. In our example, the directory is called “HAV”. Enter in your terminal:
   ```
   cd HAV
   ```
9. Run Nextclade. Enter in your terminal:
   ```
   ./nextclade run \
      --input-dataset data/[dataset name] \
      --output-all=[output directory name] \
      [input sequences file name]
   ```
   for example:
   ```
   ./nextclade run \
      --input-dataset data/HAV-WGS \
      --output-all=output_WGS_on_WGS/ \
      example_whole_genome_sequences.fasta
   ```
10. Navigate to your output directory, in this example named “output_WGS_on_WGS”. Your output directory should now be populated.
11. Open the output file nextclade.auspice.json in https://auspice.us to view your reference dataset’s phylogeny with the input sequences added in. You can highlight the input sequences by selecting “Node type” under “Color By” at left. The input sequences will be highlighted in color, while the dataset sequences will not be highlighted.
12. Click “Clade” under “Color By” to view the clades assigned to each of your sequences.
13. Open the output file nextclade.tsv in Microsoft Excel. Note the columns “seqName”, which contains your input sequence names, and “clade”, which contains the clades assigned to your input sequences.
14. Some sequences may be too divergent from the reference for clade assignment. You can remedy this to some extent by adjusting the “min-seed-cover” parameter of nextclade. For example:
    ```
    ./nextclade run \
       --input-dataset data/HAV-WGS \
       --output-all=output_WGS_on_WGS/ \
    ```

```
    --min-seed-cover 0.20 \
    example_whole_genome_sequences.fasta
```

10. Run Nextclade From the Web User Interface

1. Upload your data sub-directory (in this example, HAV-WGS) to a public GitHub repository.
2. Compose your url by adding the url of your sub-directory on GitHub to https://clades.nextstrain.org?dataset-url=
   In this example, our url might be:
   https://clades.nextstrain.org?dataset-url=https://github.com/MASPHL-Bioinformatics/nextclade/tree/main/data/HAV-WGS
   You can view directions on and options for composing this url here: https://docs.nextstrain.org/projects/nextclade/en/stable/user/nextclade-web/url-parameters.html
3. Enter your url into the web browser. The selected reference dataset should now be your dataset.
4. Click “Select files” under “Provide sequence data” at left and select an unaligned sequence fasta file.
5. Click “Run”.
6. You should be on the “Results” tab. View the clades assigned to your input sequences.
7. Click the “Tree” tab in the top menu. This view shows your reference dataset’s phylogeny with the input sequences added in. You can highlight the input sequences by selecting “Node type” under “Color By” at left. The input sequences will be highlighted in color, while the dataset sequences will not be highlighted.
8. Click “Clade” under “Color By” to view the clades assigned to each of your sequences.

11. Submit Nextclade Dataset for Review

Follow the guide for contributing a Nextclade dataset under “Adding a new dataset” here: https://github.com/nextstrain/nextclade_data/blob/master/docs/dataset-curation-guide.md

Appendix: Genome Annotation Files

This guide uses annotation files in two formats:

- augur_from_mltree takes as input a .gb genome annotation file
- your Nextclade dataset includes an (optional) .gff3 genome annotation file

These two files, which you will be downloading from NCBI, should be concordant.

You may find that you want to modify the .gff3 genome annotation file that you include in your Nextclade dataset. (Perhaps you would like to add, remove, or modify certain regions of the genome.) If you do, you will need to first convert your modified .gff3 genome annotation file to a .gb genome annotation file, and then use your updated .gb genome annotation file as input to re-generate your tree using augur_from_mltree (Step 6 of this guide). You can easily convert your .gff3 annotation file to a .gb annotation file using Galaxy: https://usegalaxy.eu/?tool_id=gff_to_sequence (You will also need the corresponding fasta sequence you downloaded from NCBI.) Note that all feature annotations have to be multiples of 3, and that Nextclade treats CDS/genes as interchangeable but the preferred format for features is CDS.