**South Africa Influenza Reference Dataset:**

* South African dataset from September 2021 - 5 August 2025: 890 sequences

**DOCUMENT**

* influenza\_rsa\_2021.fasta: FASTA file with South African sequences from 2021-2025 on GISAID
* influenza\_rsa\_2021\_w\_nextclade.tsv: GISAID metadata that I cleaned and included the Nextclade classifications
* influenza\_rsa\_2021.tsv: Sequence index that Augur requires to subsample
* influenza\_rsa\_2021\_w\_nextclade\_sub.tsv: Subsampled metadata
* influenza\_rsa\_2021\_sub.fasta: Subsampled FASTA

**CLASSIFY**

* - Done using Nextclade GUI

**INDEXING**

* Indexing sequences to subsample:

augur index --sequences influenza\_rsa\_2021.fasta --output influenza\_rsa\_2021.tsv

**SUBSAMPLING**

* augur filter --metadata influenza\_rsa\_2021\_w\_nextclade.tsv --sequences influenza\_rsa\_2021.fasta --sequence-index influenza\_rsa\_2021.tsv --metadata-id-columns name --group-by clade\_nextclade year month --sequences-per-group 5 --output-sequences influenza\_rsa\_2021\_subsampled.fasta --output-metadata influenza\_rsa\_2021\_w\_nextclade\_subsampled.tsv

Output from CLI after subsampling:

741 strains were dropped during filtering

741 were dropped because of subsampling criteria

149 strains passed all filters