Fogarty International Center
Advancing Science for Global Health

COV-IRT
COVID-19 International Research Team

Building a background dataset

Retrieve sequences from public databases

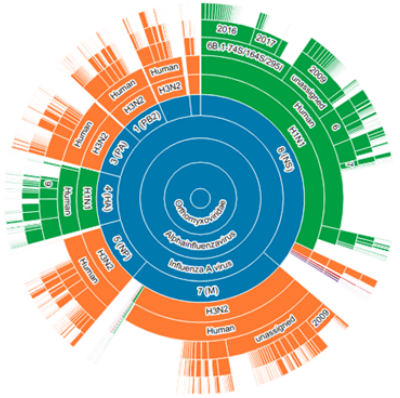
Nídia Trovão, PhD
Division of International Epidemiology and Population Studies
Fogarty International Center
National Institutes of Health

General data processing workflow

Obtain sample

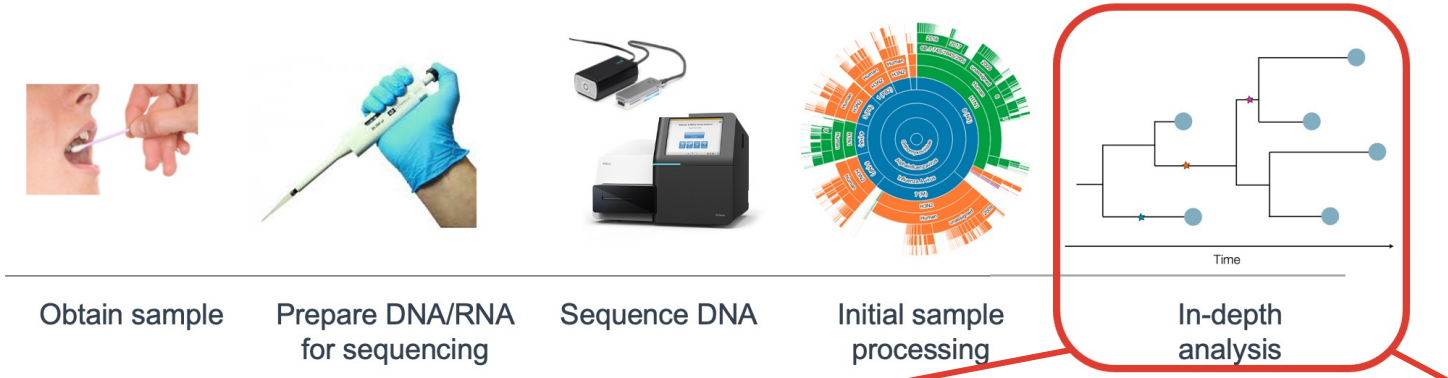
Prepare DNA/RNA for sequencing

Sequence DNA



Initial sample processing

Phylogenetics protocol



Obtain sample | Prepare DNA/RNA for sequencing | Sequence DNA | Initial sample processing | In-depth analysis

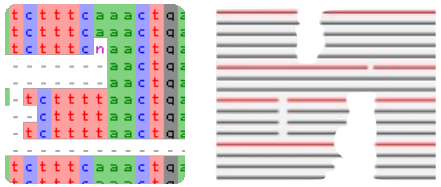
Background Dataset

NCBI Virus
Sequences for discovery

GISAID

- Download from public database

Multiple Sequence Alignment

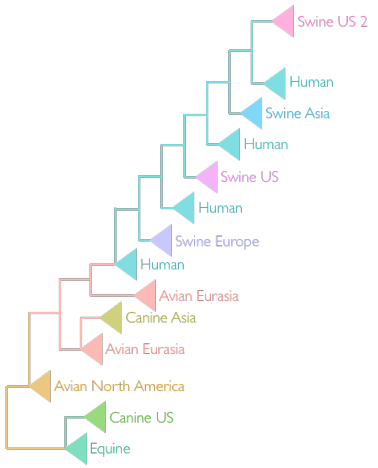


- Sequence alignment
- Clean and manually edition

Building a Phylogeny

- Maximum Likelihood Tree
- Bootstrap support

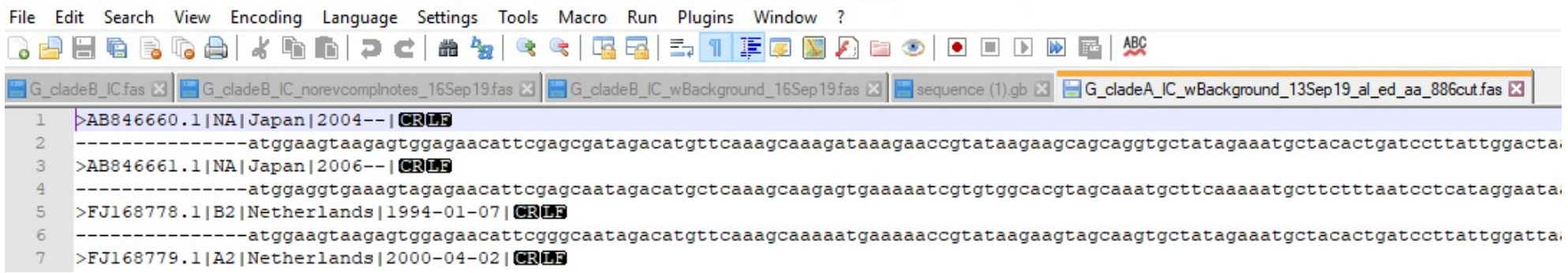
Interpreting a Phylogenetic Tree



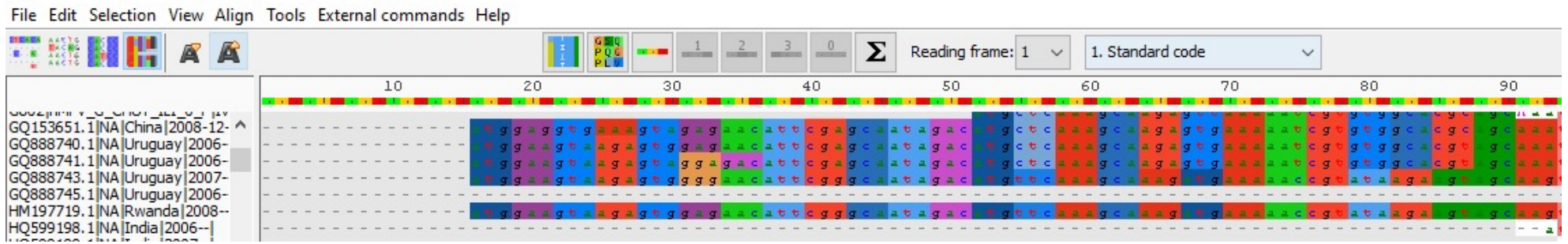
Alignment viewer and editing software

How to view Fasta (.fa, .fas, .fasta)

- In order to **view** Fasta-format files → text editor
  - Notepad ++ (Windows)
  - BBEdit (macOS only)

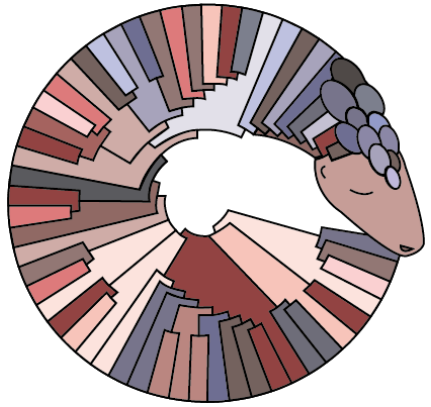


- In order to **edit** fasta-format files → AliView



Identify the study sequences' lineage

Phylogenetic **A**ssignment of **N**amed **G**lobal **O**utbreak **LIN**eages **PANGOLIN**



- Local

https://github.com/cov-lineages/pangolin

- Webserver

https://pangolin.cog-uk.io/

Nextclade
Clade assignment, mutation calling, and sequence quality checks

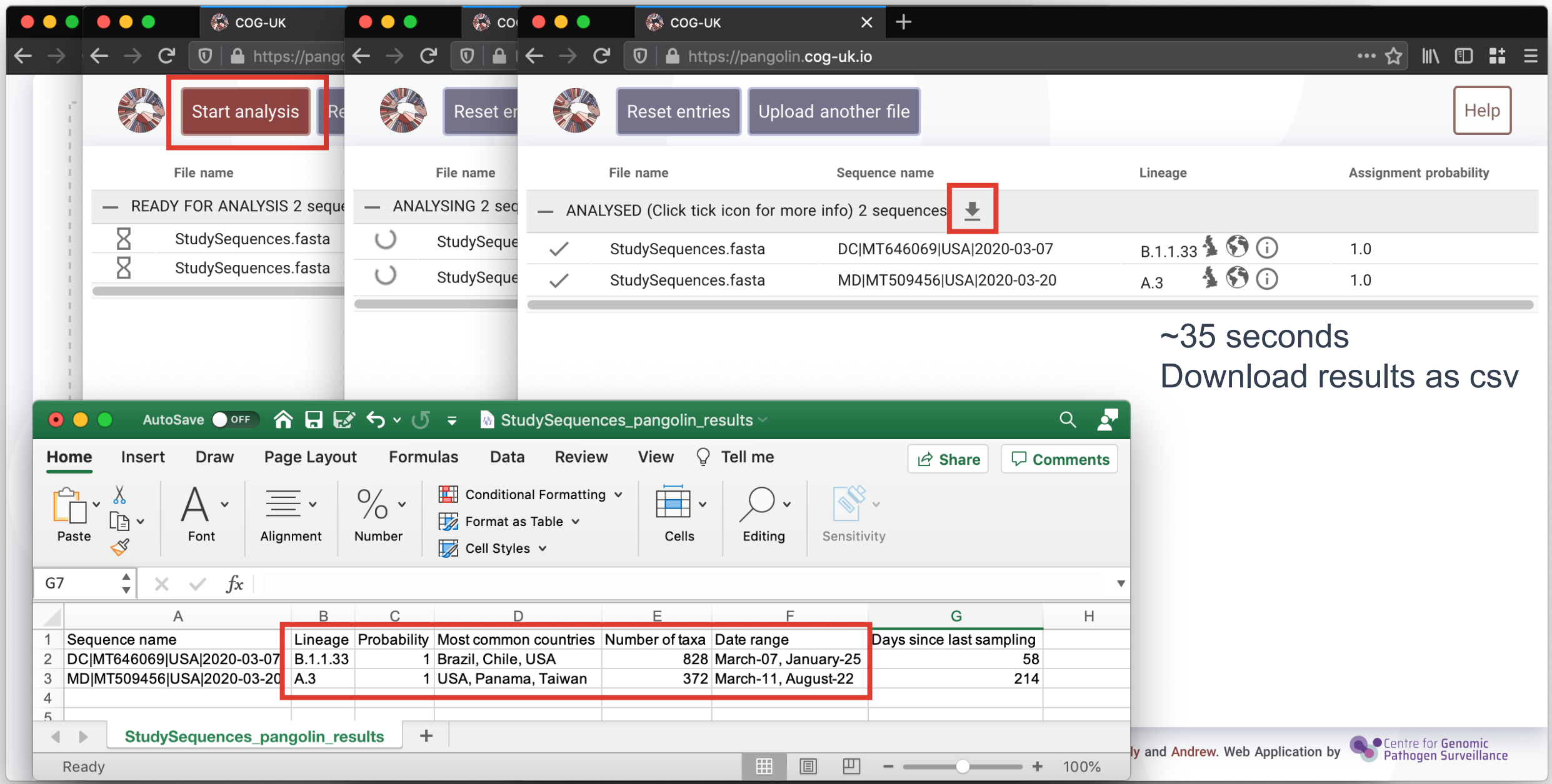
Nextclade beta

- Local

https://github.com/nextstrain/nextclade

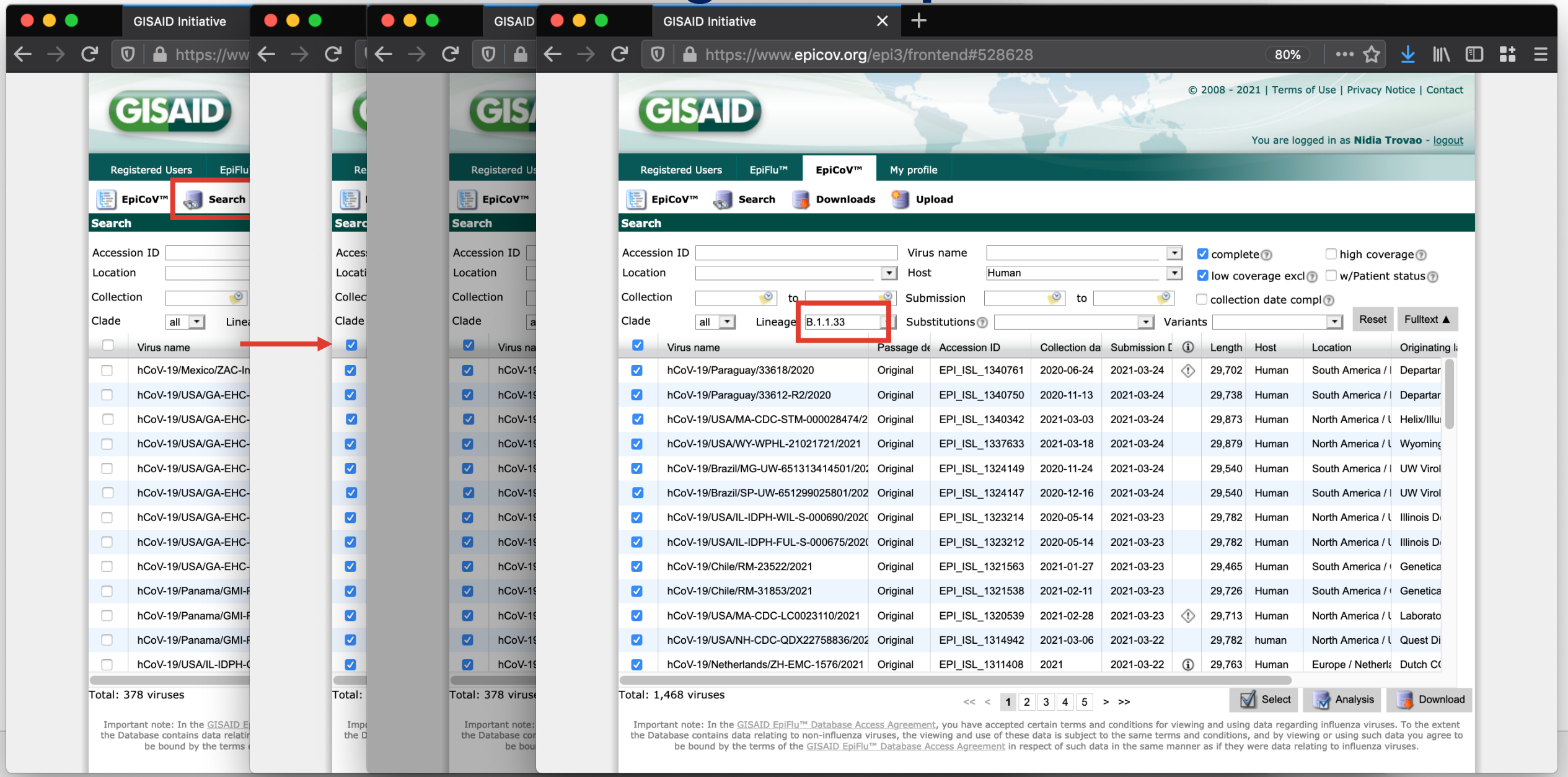
- Webserver

https://clades.nextstrain.org/


COG-UK
https://pangolin.cog-uk.io
Start analysis
Reset entries
Upload another file
Help
File name
Sequence name
Lineage
Assignment probability
READY FOR ANALYSIS 2 seque
ANALYSING 2 seq
ANALYSED (Click tick icon for more info) 2 sequences
StudySequences.fasta
DC|MT646069|USA|2020-03-07
B.1.1.33
1.0
MD|MT509456|USA|2020-03-20
A.3
1.0
~35 seconds
Download results as csv
AutoSave
StudySequences_pangolin_results
Home
Insert
Draw
Page Layout
Formulas
Data
Review
View
Tell me
Share
Comments
Paste
Font
Alignment
Number
Conditional Formatting
Format as Table
Cell Styles
Cells
Editing
Sensitivity
G7
Sequence name
Lineage
Probability
Most common countries
Number of taxa
Date range
Days since last sampling
DC|MT646069|USA|2020-03-07
B.1.1.33
1
Brazil, Chile, USA
828
March-07, January-25
58
MD|MT509456|USA|2020-03-20
A.3
1
USA, Panama, Taiwan
372
March-11, August-22
214
StudySequences_pangolin_results
Ready
100%
and Andrew. Web Application by
Centre for Genomic Pathogen Surveillance

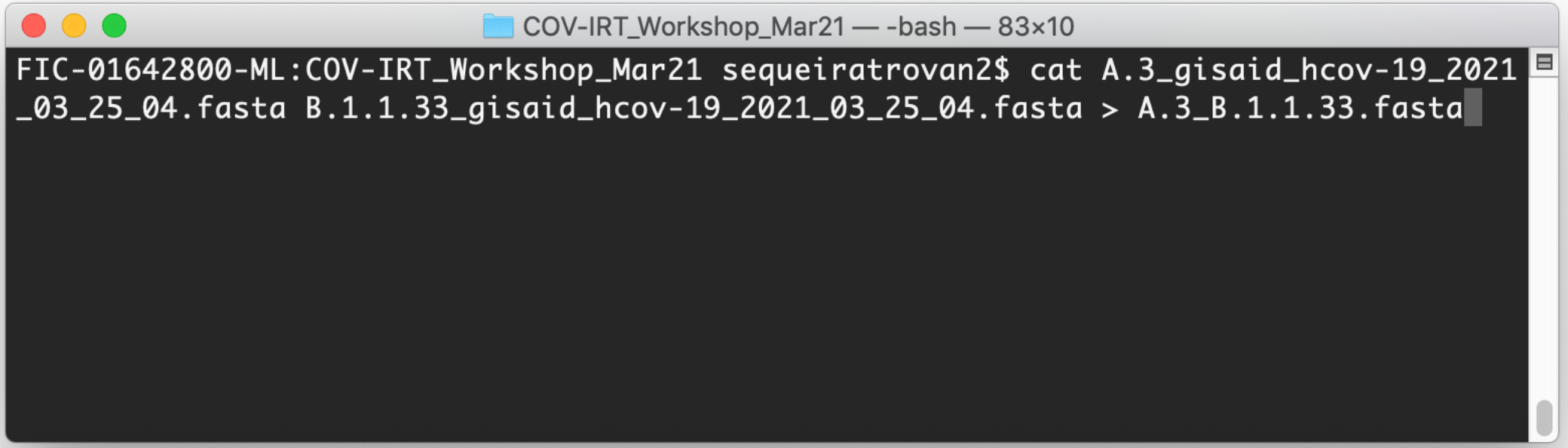
Accessing sequences on public genetic databases

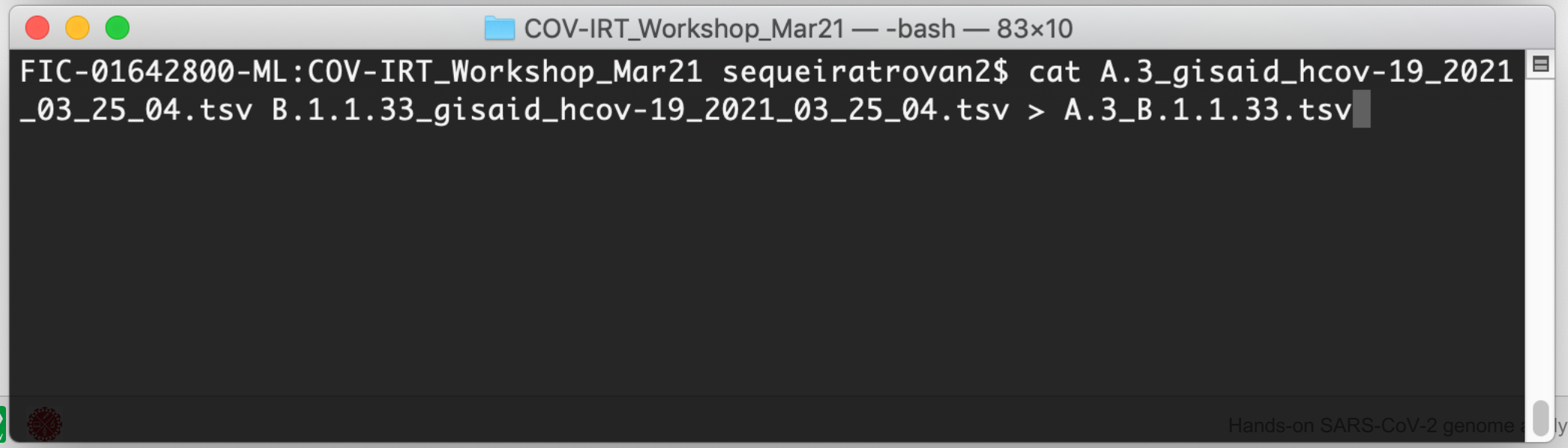
GISAID.ORG → Log in → EpiCOV → Browse



Assembling the background dataset

Concatenate A.3 and B.1.1.33 fasta and metadata files

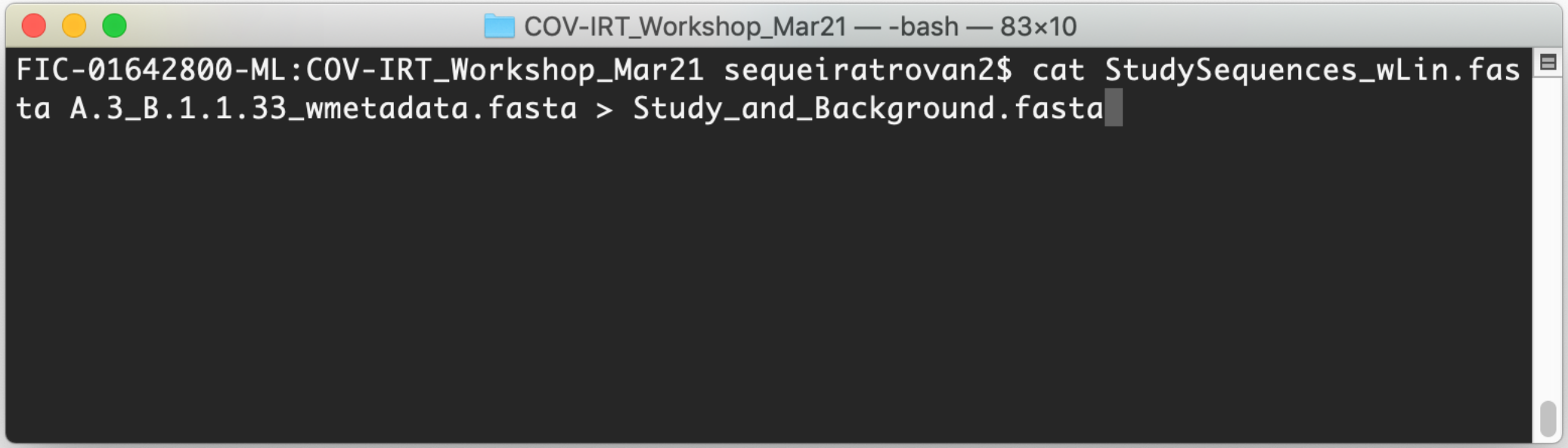




Or copy the sequences from one file to the other

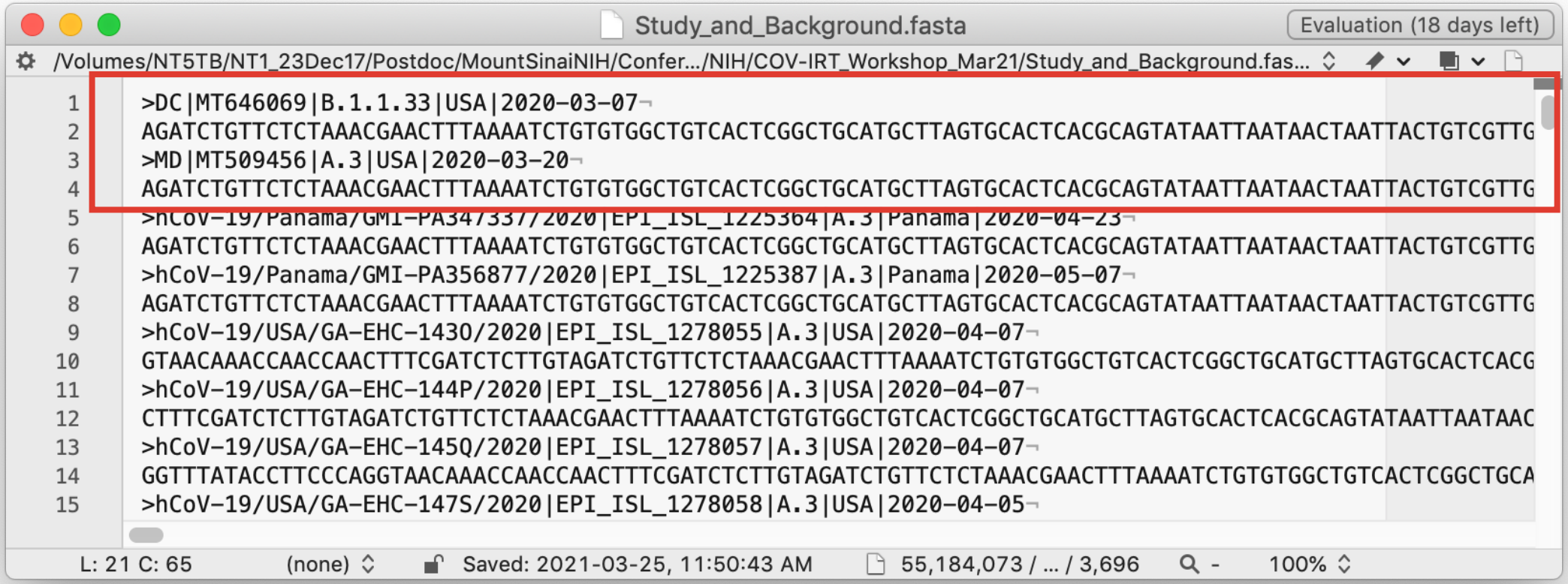
Add metadata to background (R, vlookup in Excel)

Concatenate background dataset with study sequences A.3_B.1.1.33_wmetadata.fasta + StudySequences_wLin.fasta



Or copy the sequences from one file to the other

Remove illegal characters for compatibility across phylogetics software



Replace with underscore _

Illegal characters

Parentheses ( ) Commas , Semi-colon ; Colon : Plus sign + Accents á, ê, í…


COV-IRT
COVID-19 International Research Team


NIH
Fogarty International Center
Advancing Science for Global Health