



Fogarty International Center
Advancing Science for Global Health



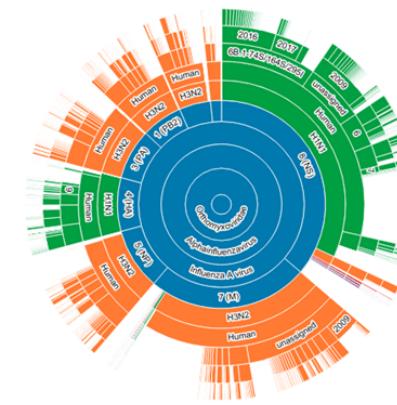
COVID-19 International Research Team

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



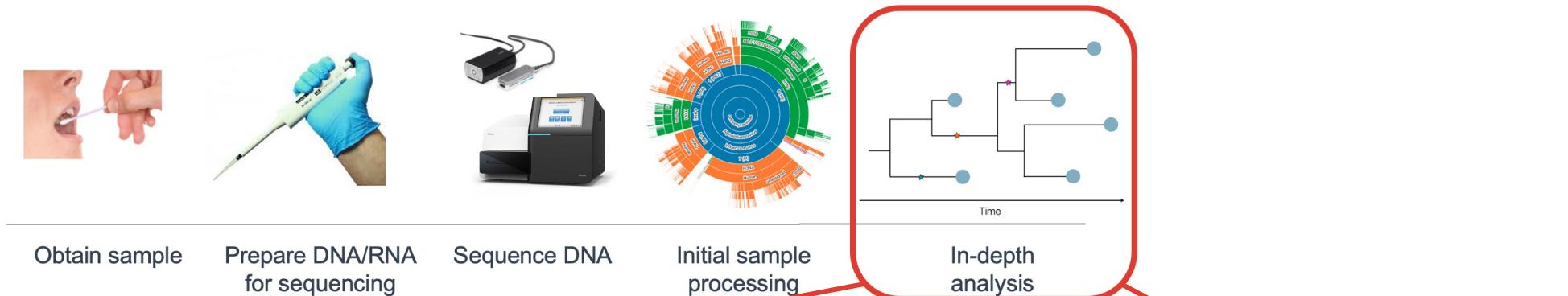
Obtain sample

Prepare DNA/RNA
for sequencing

Sequence DNA

Initial sample
processing

Phylogenetics protocol

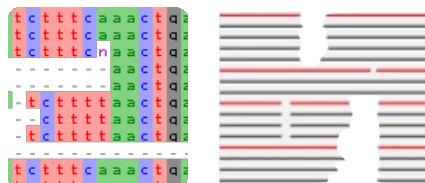


Background Dataset



- Download from public database

Multiple Sequence Alignment



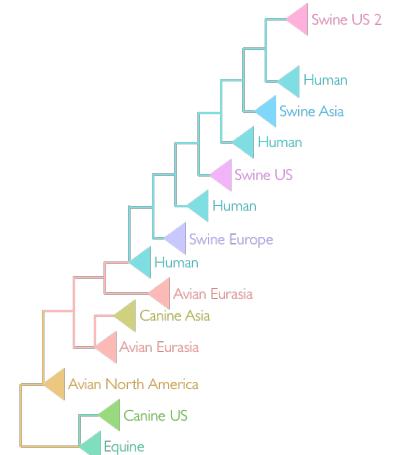
- Sequence alignment
- Clean and manually edition

Building a Phylogeny



- Maximum Likelihood Tree
- Bootstrap support

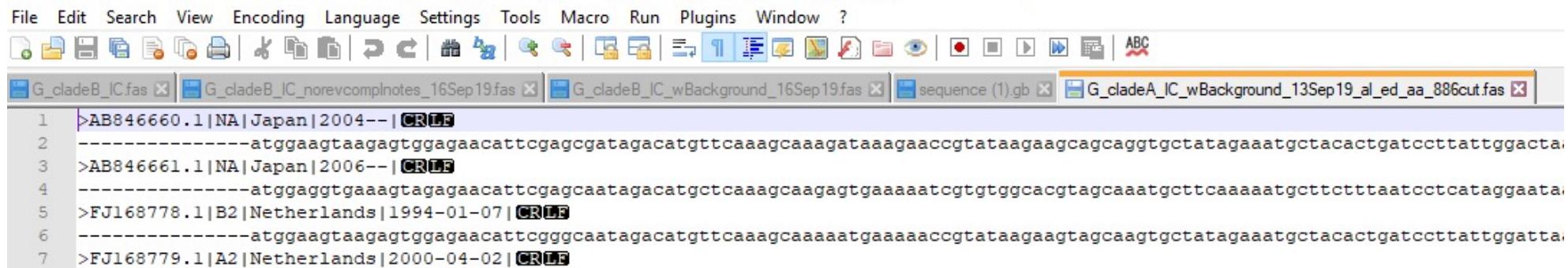
Interpreting a Phylogenetic Tree



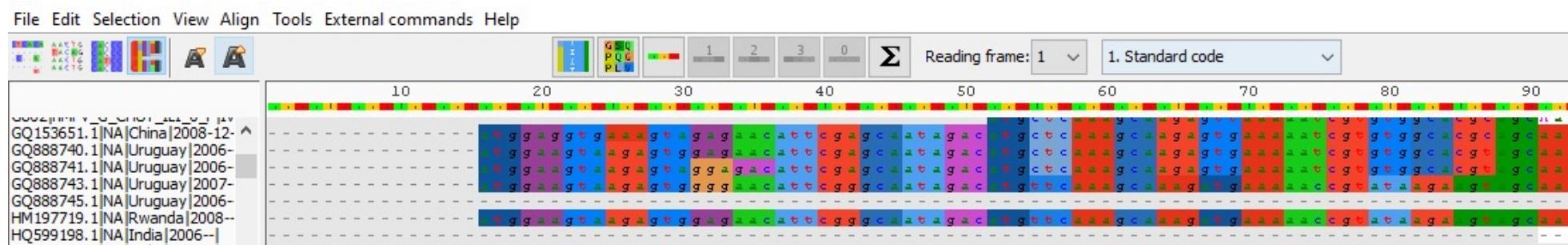
Alignment viewer and editing software

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

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

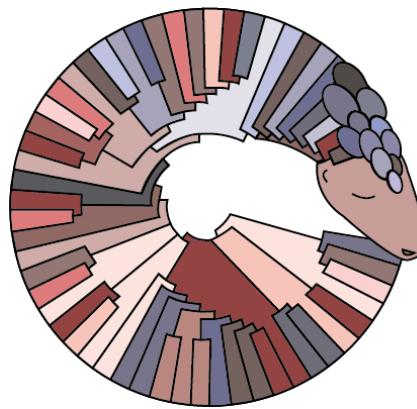


- In order to **edit** fasta-format files → AliView



Identify the study sequences' lineage

Phylogenetic Assignment of Named Global Outbreak LIneages **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/>

3/28/21

Three screenshots of a web application interface are shown side-by-side.

The first screenshot shows the initial state with two files selected:

File name
READY FOR ANALYSIS 2 sequences
StudySequences.fasta
StudySequences.fasta

The second screenshot shows the analysis progress:

File name	Sequence name	Lineage	Assignment probability
ANALYSING 2 sequences	ANALYSED (Click tick icon for more info) 2 sequences		
StudySequences.fasta	DC MT646069 USA 2020-03-07	B.1.1.33	1.0
StudySequences.fasta	MD MT509456 USA 2020-03-20	A.3	1.0

The third screenshot shows the results downloaded into a Microsoft Excel spreadsheet:

A	B	C	D	E	F	G	H
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	

~35 seconds
Download results as csv

Accessing sequences on public genetic databases

3/28/21

GISAID.ORG → Log in → EpiCOV → Browse

The screenshot shows three browser tabs illustrating the GISAID search process:

- Tab 1 (Left):** Shows the GISAID homepage with the "EpiCoV™" button highlighted by a red box.
- Tab 2 (Middle):** Shows the GISAID search results page for "EpiCoV™" with a red arrow pointing from the first tab to the "Search" button here.
- Tab 3 (Right):** Shows the detailed EpiCoV search interface. A red box highlights the "Lineage" dropdown set to "B.1.1.33".

EpiCoV™ Search Results (Tab 3):

Virus name	Passage date	Accession ID	Collection date	Submission date	Length	Host	Location	Originating lab
hCoV-19/Paraguay/33618/2020	Original	EPI_ISL_1340761	2020-06-24	2021-03-24	29,702	Human	South America / Paraguay	Departamento de Salud Pública
hCoV-19/Paraguay/33612-R2/2020	Original	EPI_ISL_1340750	2020-11-13	2021-03-24	29,738	Human	South America / Paraguay	Departamento de Salud Pública
hCoV-19/USA/MA-CDC-STM-000028474/2	Original	EPI_ISL_1340342	2021-03-03	2021-03-24	29,873	Human	North America / United States	Massachusetts Department of Public Health
hCoV-19/USA/WY-WPHL-21021721/2021	Original	EPI_ISL_1337633	2021-03-18	2021-03-24	29,879	Human	North America / Wyoming	Wyoming Department of Health
hCoV-19/Brazil/MG-UW-651313414501/2020	Original	EPI_ISL_1324149	2020-11-24	2021-03-24	29,540	Human	South America / Brazil	UW Virology
hCoV-19/Brazil/SP-UW-651299025801/2020	Original	EPI_ISL_1324147	2020-12-16	2021-03-24	29,540	Human	South America / Brazil	UW Virology
hCoV-19/USA/IL-IDPH-WIL-S-000690/2020	Original	EPI_ISL_1323214	2020-05-14	2021-03-23	29,782	Human	North America / Illinois	Illinois Department of Public Health
hCoV-19/USA/IL-IDPH-FUL-S-000675/2020	Original	EPI_ISL_1323212	2020-05-14	2021-03-23	29,782	Human	North America / Illinois	Illinois Department of Public Health
hCoV-19/Chile/RM-23522/2021	Original	EPI_ISL_1321563	2021-01-27	2021-03-23	29,465	Human	South America / Chile	Genetic Analysis
hCoV-19/Chile/RM-31853/2021	Original	EPI_ISL_1321538	2021-02-11	2021-03-23	29,726	Human	South America / Chile	Genetic Analysis
hCoV-19/USA/MA-CDC-LC0023110/2021	Original	EPI_ISL_1320539	2021-02-28	2021-03-23	29,713	Human	North America / Massachusetts	Laboratory
hCoV-19/USA/NH-CDC-QDX22758836/2020	Original	EPI_ISL_1314942	2021-03-06	2021-03-22	29,782	Human	North America / New Hampshire	Quest Diagnostics
hCoV-19/Netherlands/ZH-EMC-1576/2021	Original	EPI_ISL_1311408	2021	2021-03-22	29,763	Human	Europe / Netherlands	Dutch CDC

Total: 1,468 viruses

Important note: In the GISAID EpiFlu™ Database Access Agreement, you have accepted certain terms and conditions for viewing and using data regarding influenza viruses. To the extent the Database contains data relating to non-influenza viruses, the viewing and use of these data is subject to the same terms and conditions, and by viewing or using such data you agree to be bound by the terms of the GISAID EpiFlu™ Database Access Agreement in respect of such data in the same manner as if they were data relating to influenza viruses.

Assembling the background dataset

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

```
COV-IRT_Workshop_Mar21 — -bash — 83x10
FIC-01642800-ML:COV-IRT_Workshop_Mar21 sequeiratrovant$ cat A.3_gisaid_hcov-19_2021
_03_25_04.fasta B.1.1.33_gisaid_hcov-19_2021_03_25_04.fasta > A.3_B.1.1.33.fasta
```

Or copy the sequences from one file to the other

```
COV-IRT_Workshop_Mar21 — -bash — 83x10
FIC-01642800-ML:COV-IRT_Workshop_Mar21 sequeiratrovant$ cat A.3_gisaid_hcov-19_2021
_03_25_04.tsv B.1.1.33_gisaid_hcov-19_2021_03_25_04.tsv > A.3_B.1.1.33.tsv
```

Add metadata to background (R, vlookup in Excel)



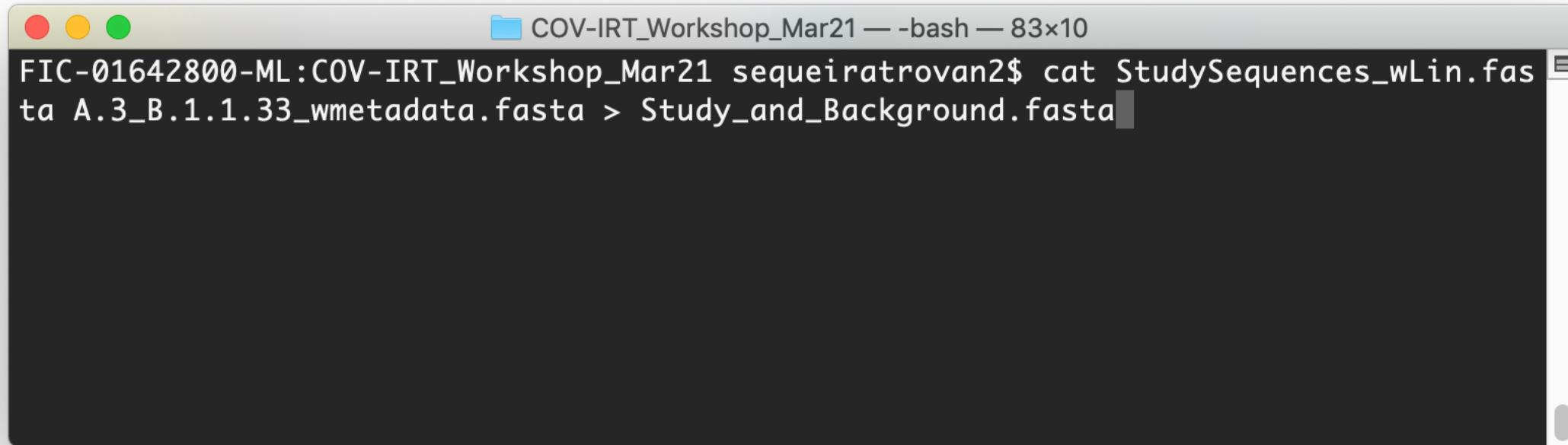
```
A.3_B.1.33.fasta
1 >hCoV-19/Panama/GMI-PA347337/2020|EPI_ISL_1225364|2020-04-23
2 AGATCTGTTCTCTAAACGAACTTAAAATCTGTGGCTGTCAGTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAAATAACTAATTACTGTCGTTG/
3 >hCoV-19/Panama/GMI-PA356877/2020|EPI_ISL_1225387|2020-05-07
4 AGATCTGTTCTCTAAACGAACTTAAAATCTGTGGCTGTCAGTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAAATAACTAATTACTGTCGTTG/
5 >hCoV-19/USA/GA-EHC-1430/2020|EPI_ISL_1278055|2020-04-07
6 GTAACAAACCAACCAACTTCGATCTCTTAGATCTGTTCTAAACGAACTTAAAATCTGTGGCTGTCAGTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAAATAACTAATTACTGTCGTTG/
7 >hCoV-19/USA/GA-EHC-144P/2020|EPI_ISL_1278056|2020-04-07
```



```
A.3_B.1.33_wmetadata.fasta
1 >hCoV-19/Panama/GMI-PA347337/2020|EPI_ISL_1225364|A.3|Panama|2020-04-23
2 AGATCTGTTCTCTAAACGAACTTAAAATCTGTGGCTGTCAGTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAAATAACTAATTACTGTCGTTG/
3 >hCoV-19/Panama/GMI-PA356877/2020|EPI_ISL_1225387|A.3|Panama|2020-05-07
4 AGATCTGTTCTCTAAACGAACTTAAAATCTGTGGCTGTCAGTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAAATAACTAATTACTGTCGTTG/
5 >hCoV-19/USA/GA-EHC-1430/2020|EPI_ISL_1278055|A.3|USA|2020-04-07
6 GTAACAAACCAACCAACTTCGATCTCTTAGATCTGTTCTAAACGAACTTAAAATCTGTGGCTGTCAGTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAAATAACTAATTACTGTCGTTG/
7 >hCoV-19/USA/GA-EHC-144P/2020|EPI_ISL_1278056|A.3|USA|2020-04-07
8 CTTTCGATCTCTTAGATCTGTTCTAAACGAACTTAAAATCTGTGGCTGTCAGTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAAATAAC-
9 >hCoV-19/USA/GA-EHC-145Q/2020|EPI_ISL_1278057|A.3|USA|2020-04-07
10 GGTTTATACCTTCCCAGGTAAACAAACCAACCAACTTCGATCTCTTAGATCTGTTCTAAACGAACTTAAAATCTGTGGCTGTCAGTCGGCTGCA-
11 >hCoV-19/USA/GA-EHC-147S/2020|EPI_ISL_1278058|A.3|USA|2020-04-05
12 GGTTTATACCTTCCCAGGTAAACAAACCAACCAACTTCGATCTCTTAGATCTGTTCTAAACGAACTTAAAATCTGTGGCTGTCAGTCGGCTGCA-
13 >hCoV-19/USA/GA-EHC-179Y/2020|EPI_ISL_1278059|A.3|USA|2020-04-18
14 ACCTTCCCAGGTAAACAAACCAACCAACTTCGATCTCTTAGATCTGTTCTAAACGAACTTAAAATCTGTGGCTGTCAGTCGGCTGCATGCTTAG-
15 >hCoV-19/USA/GA-EHC-202V/2020|EPI_ISL_1278062|A.3|USA|2020-04-21
16 GGTTTATACCTNCCCAGGTAAACAAACCAACCAACTTCGATCTCTTAGATCTGTTCTAAACGAACTTAAAATCTGTGGCTGTCAGTCGGCTGCA-
17 >hCoV-19/USA/GA-EHC-246N/2020|EPI_ISL_1278065|A.3|USA|2020-05-06
18 CTTGTAGATCTGTTCTAAACGAACTTAAAATCTGTGGCTGTCAGTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAAACTAATTACTGTCGTTG/
```

Concatenate background dataset with study sequences

A.3_B.1.1.33_wmetadata.fasta + StudySequences_wLin.fasta



```
FIC-01642800-ML:COV-IRT_Workshop_Mar21 sequeiratrovant2$ cat StudySequences_wLin.fasta A.3_B.1.1.33_wmetadata.fasta > Study_and_Background.fasta
```

Or copy the sequences from one file to the other

Remove illegal characters for compatibility across phylogenetics software

```

Study_and_Background.fasta
Evaluation (18 days left)
/Volumes/NT5TB/NT1_23Dec17/Postdoc/MountSinaiNIH/Confer.../NIH/COV-IRT_Workshop_Mar21/Study_and_Background.fas...
1 >DC |MT646069|B.1.1.33|USA|2020-03-07-
2 AGATCTGTTCTCTAAACGAACTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAACTAATTACTGTCGTTG
3 >MD |MT509456|A.3|USA|2020-03-20-
4 AGATCTGTTCTCTAAACGAACTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAACTAATTACTGTCGTTG
5 >hCoV-19/Panama/GMI-PA34/337/2020|EPI_ISL_1225364|A.3|Panama|2020-04-23-
6 AGATCTGTTCTCTAAACGAACTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAACTAATTACTGTCGTTG
7 >hCoV-19/Panama/GMI-PA356877/2020|EPI_ISL_1225387|A.3|Panama|2020-05-07-
8 AGATCTGTTCTCTAAACGAACTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAACTAATTACTGTCGTTG
9 >hCoV-19/USA/GA-EHC-1430/2020|EPI_ISL_1278055|A.3|USA|2020-04-07-
10 GTAACAAACCAACCAACTTTCGATCTTGTAGATCTGTTCTCTAAACGAACTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACG
11 >hCoV-19/USA/GA-EHC-144P/2020|EPI_ISL_1278056|A.3|USA|2020-04-07-
12 CTTTCGATCTTGTAGATCTGTTCTCTAAACGAACTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAAAC
13 >hCoV-19/USA/GA-EHC-145Q/2020|EPI_ISL_1278057|A.3|USA|2020-04-07-
14 GGTTTATACCTCCAGGTAAACAAACCAACCTTCGATCTTGTAGATCTGTTCTCTAAACGAACTTAAAATCTGTGTGGCTGTCACTCGGCTGCA
15 >hCoV-19/USA/GA-EHC-147S/2020|EPI_ISL_1278058|A.3|USA|2020-04-05-

```

L: 21 C: 65 (none) Saved: 2021-03-25, 11:50:43 AM 55,184,073 / ... / 3,696 100%

Illegal characters

Parentheses ()

Commas ,

Semi-colon ;

Colon :

Plus sign +

Accents á, ê, í...

Replace with underscore _



COVID-19 International Research Team

