Fogarty International Center
Advancing Science for Global Health

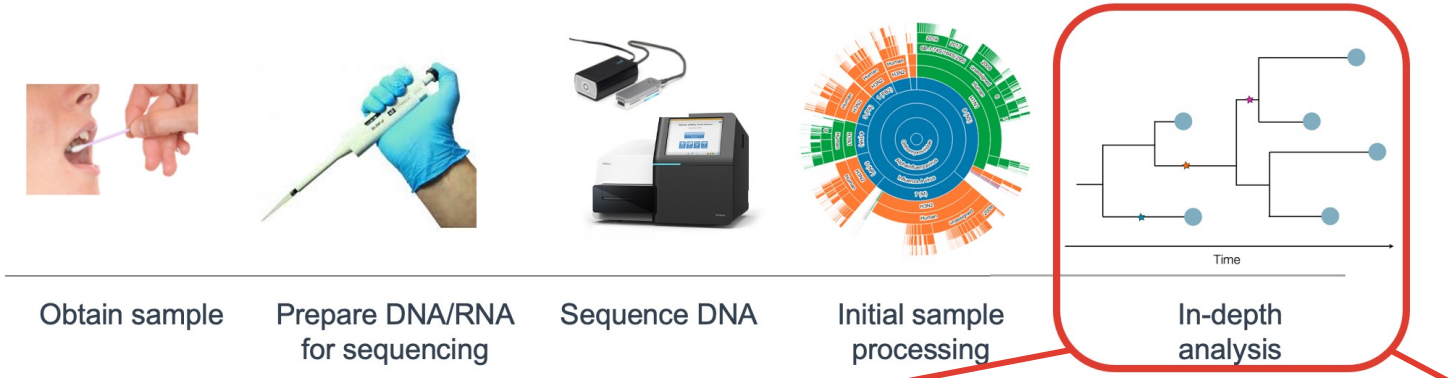
COV-IRT
COVID-19 International Research Team

Bioinformatics towards phylogenetics

Navigating AliView

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

Phylogenetics protocol



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

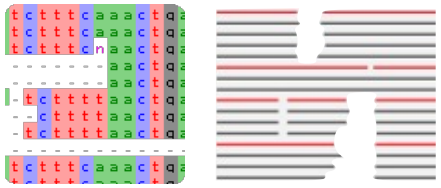
Background Dataset

NCBI Virus
Sequences for discovery

GISAID

- Download from public database

Multiple Sequence Alignment

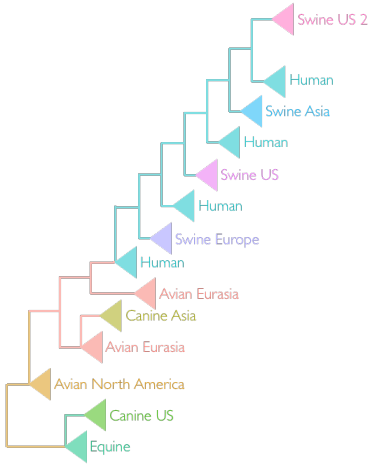


- Sequence alignment
- Clean and manually edition

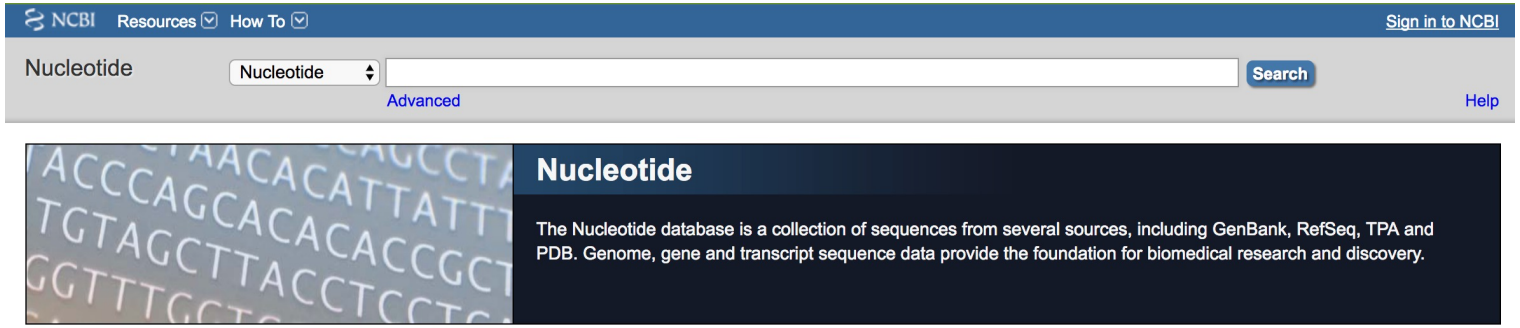
Building a Phylogeny

- Maximum Likelihood Tree
- Bootstrap support

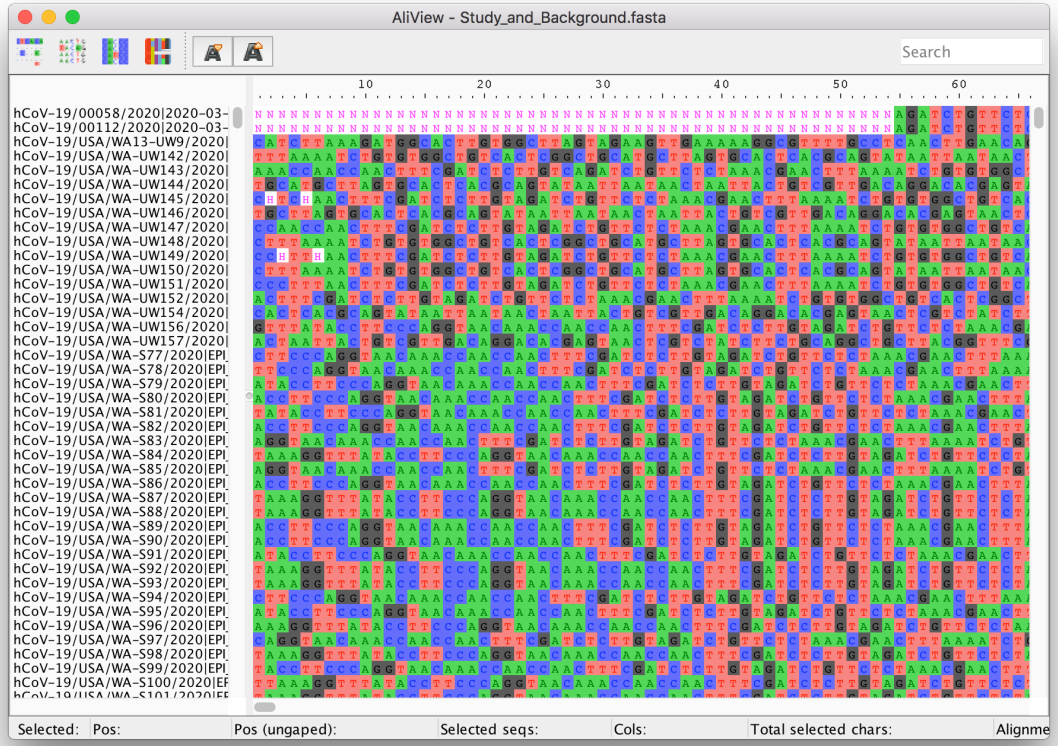
Interpreting a Phylogenetic Tree



How to View FASTA Files



Unaligned sequences



Selecting a reference genome

Multiple sequence alignments

Selecting an appropriate reference genome

Ways to choose a reference genome:

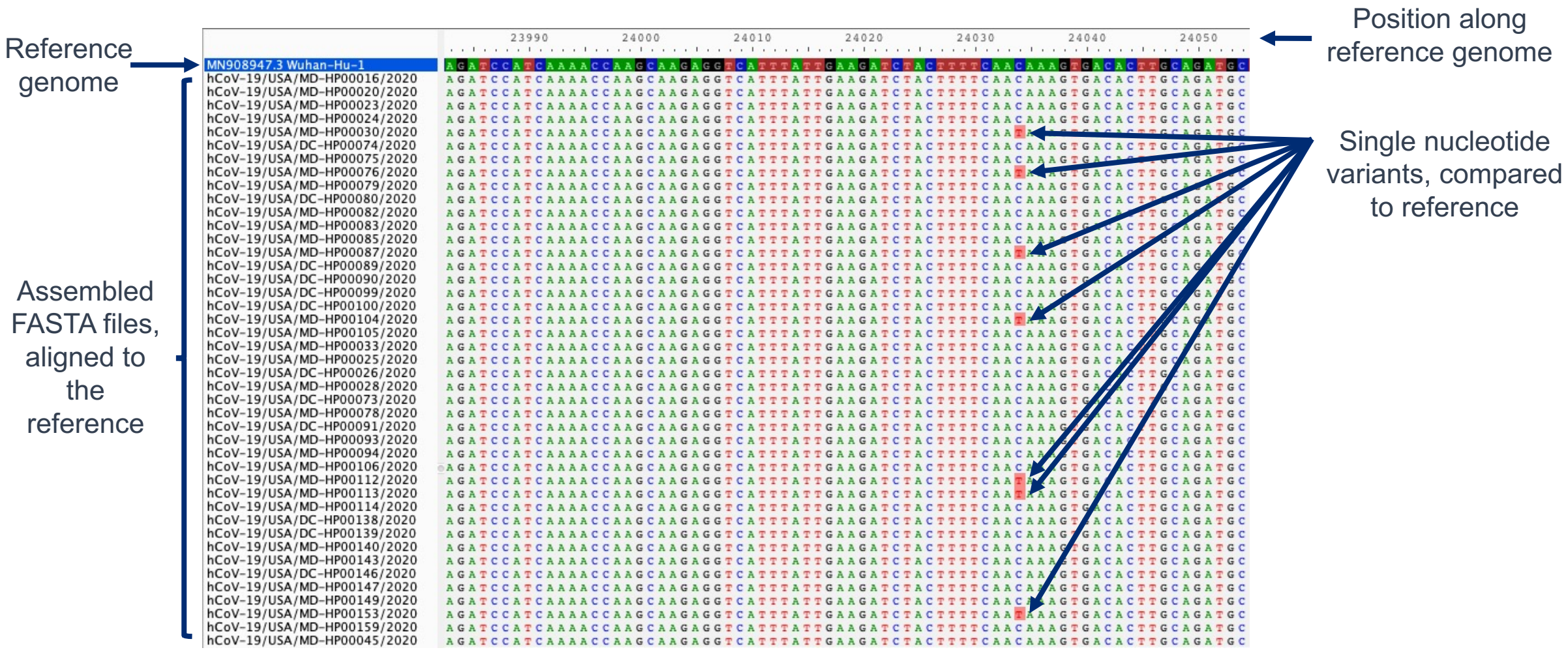
- If possible, use the same reference genome as others working on the same outbreak
  - E.g. Wuhan-Hu-1 for SARS-CoV-2 (accession: MN908947.3)
  - E.g. Makona-Kissidougou-C15 for Ebola virus (accession: KJ660346.2)
- Use the earliest sequence from the outbreak, if available
  - Wuhan-Hu-1 and Makona-Kissidougou-C15 are examples
- Use a sequence from a prior outbreak in the same location
- Search NCBI RefSeq (https://www.ncbi.nlm.nih.gov/refseq/) for an established reference

*The reference genome should always have a date **before** the earliest of your samples*

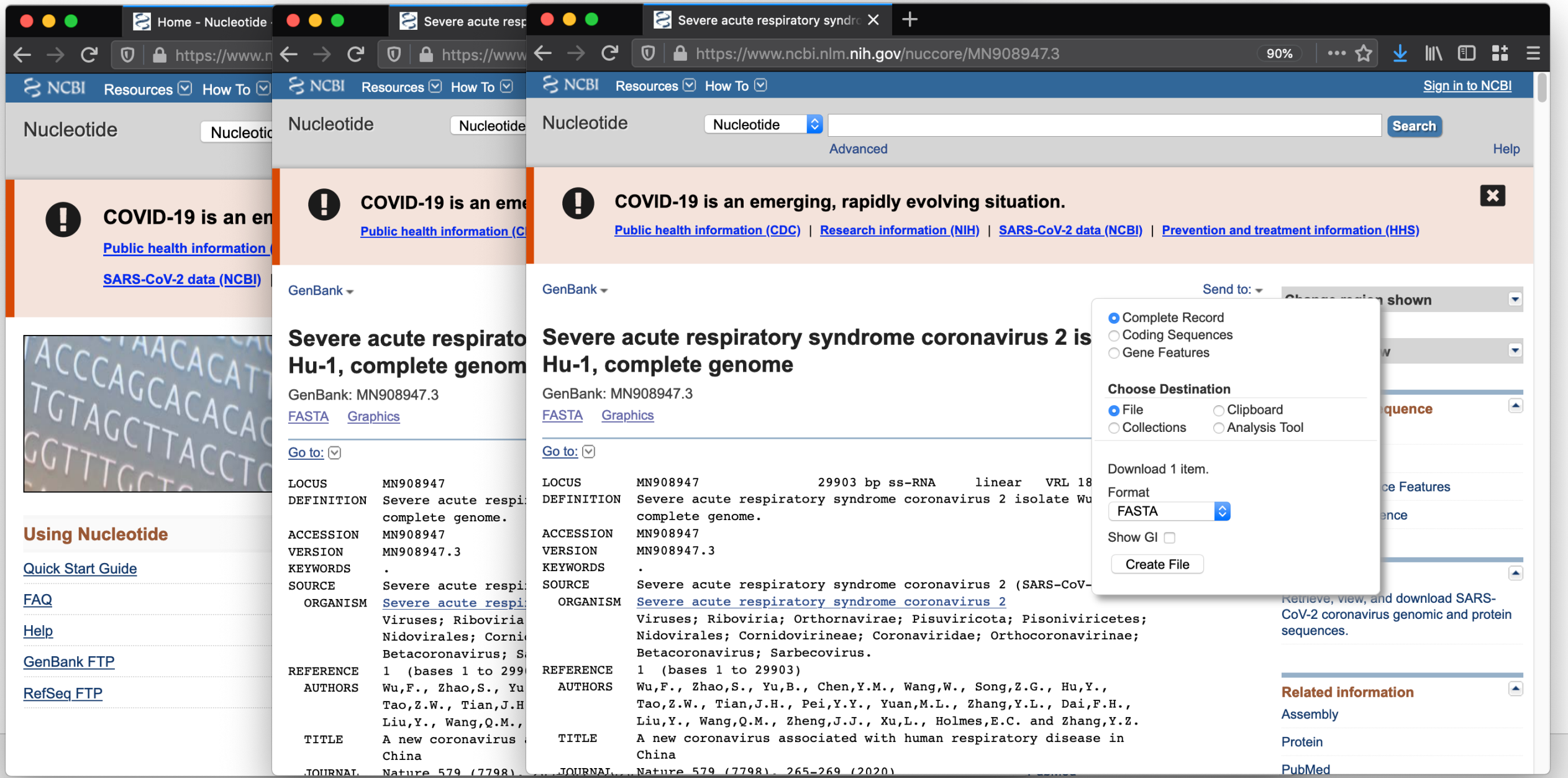
Multiple sequence alignments

Selecting an appropriate reference genome

- Aligning to a reference genome gives us a baseline to call variants against



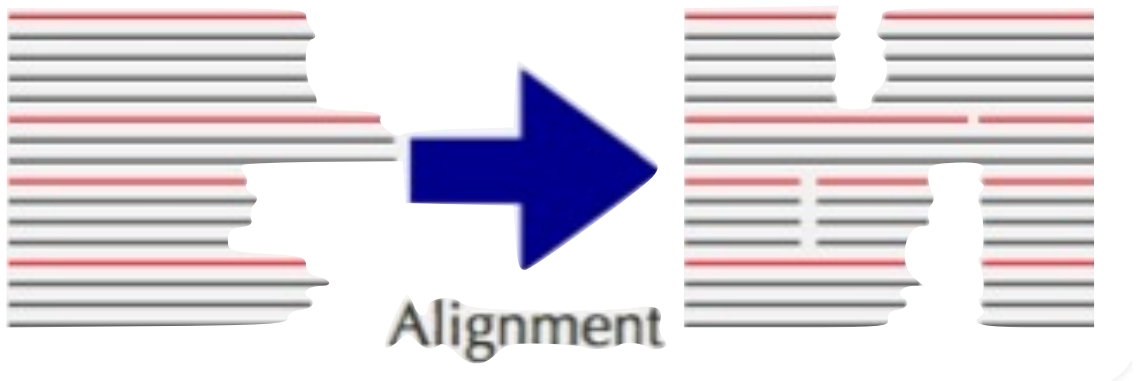
Download MN908947.3 in Genbank



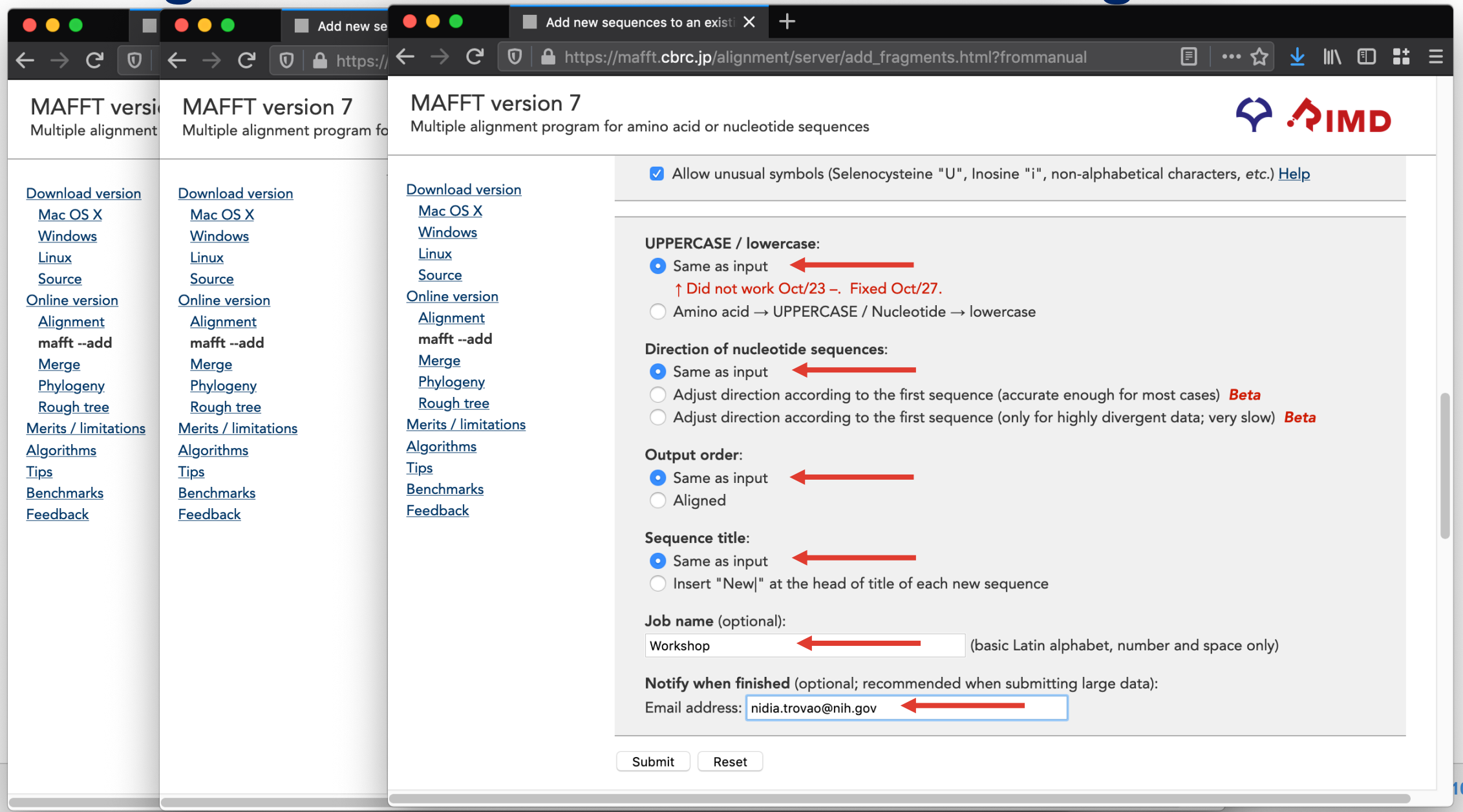
Alignment software

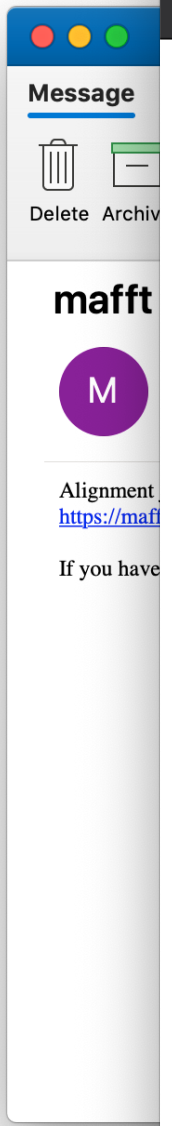
Sequence Alignments

- CLUSTALW
- CLUSTAL OMEGA
- DIALIGN-TX
- MAFFT
- MUSCLE
- POA
- Probalign
- Probcons
- T-Coffee
- BAliBASE



Align with MAFFT and a reference genome





Multiple sequence alignment by

https://mafft.cbrc.jp/alignment/s

Clustal format | Fasta format | MAFFT result | View | Tree

View

Reformat to GCG, PHYLIP, MSF, NEXUS, uppercase/

GUIDANCE2 computes the residue-wise confidence

Refine dataset

Phylogenetic tree

MAFFT-FFT-NS-fragment Result

```
CLUSTAL format alignment by MAFFT (v7.4

MN908947.3      ----------
DC|MT646069|B.1 ----------
MD|MT509456|A.3 ----------
hCoV-19/Panama/ ----------
hCoV-19/Panama/ ----------
hCoV-19/USA/GA- ----------
hCoV-19/USA/GA- ----------
hCoV-19/USA/GA- ----------
hCoV-19/USA/GA- ----------
hCoV-19/USA/GA- ----------
hCoV-19/USA/GA- ----------
hCoV-19/USA/GA- ----------
hCoV-19/USA/GA- ----------
hCoV-19/USA/GA- ----------
hCoV-19/Mexico/ ----------
hCoV-19/Austral ----------
hCoV-19/USA/NY- ----------
hCoV-19/Taiwan/ ----------
hCoV-19/Scotlan ----------

. . . . . .
```

Multiple sequence alignment by

https://mafft.cbrc.jp/alignment/server/spool/_out.21032606397184R8FnjqzDXbTP

```
hCoV-19/Taiwan/ ----------
hCoV-19/Scotlan ----------

. . . . . .
```

(Use the links below to download the full alignment.)

Clustal format (68.30MB), zipped file
Fasta format (54.25MB), zipped file

Method

FFT-NS-fragment (Not tested.)

```
% mafft --inputorder --anysymbol --maxambiguous 0.05 --addfragments fragments --auto input
```

References:

- Katoh *et al.* (2002) describes FFT-NS-1, FFT-NS-2 and FFT-NS-i.
- Katoh *et al.* (2005) describes G-INS-i, L-INS-i, E-INS-i and Mafft-homologs.
- Katoh and Toh (2008) describes Q-INS-i.
- Katoh and Frith (2012) describes *-fragment.
- Q-INS-i uses the McCaskill routine from the Vienna RNA package (Hofacker *et al.* 2003) and MXSCARNA (Tabei *et al.* 2008).
- Kuraku *et al.* (2013) outlines this web service.
- Katoh and Standley (2016) describes the VSM technique (`--unalignlevel x`).
- Yamada *et al.* (2016) explains how to build large MSAs.

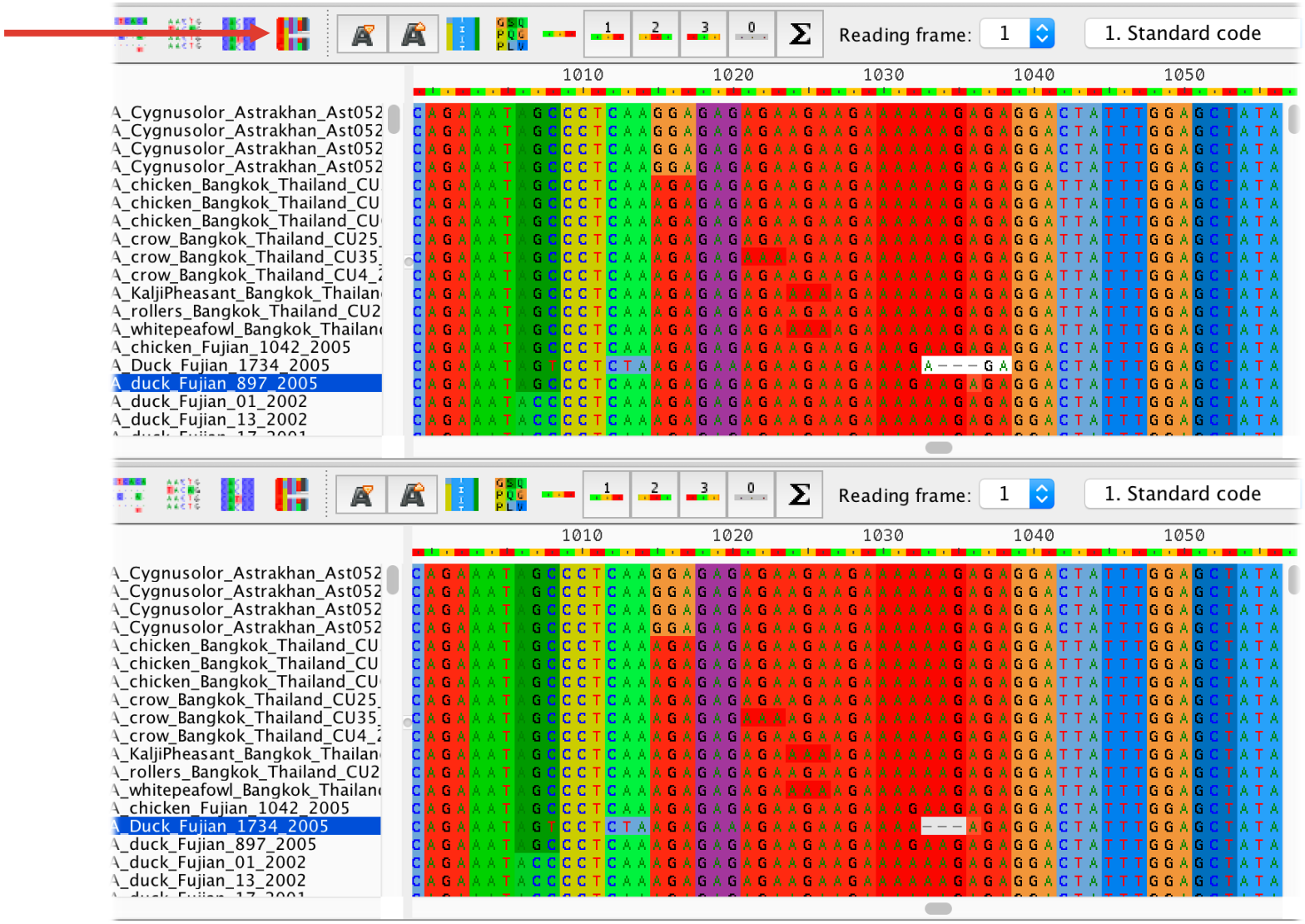
MAFFT home:
https://mafft.cbrc.jp/alignment/software/

Page Top ↑↑

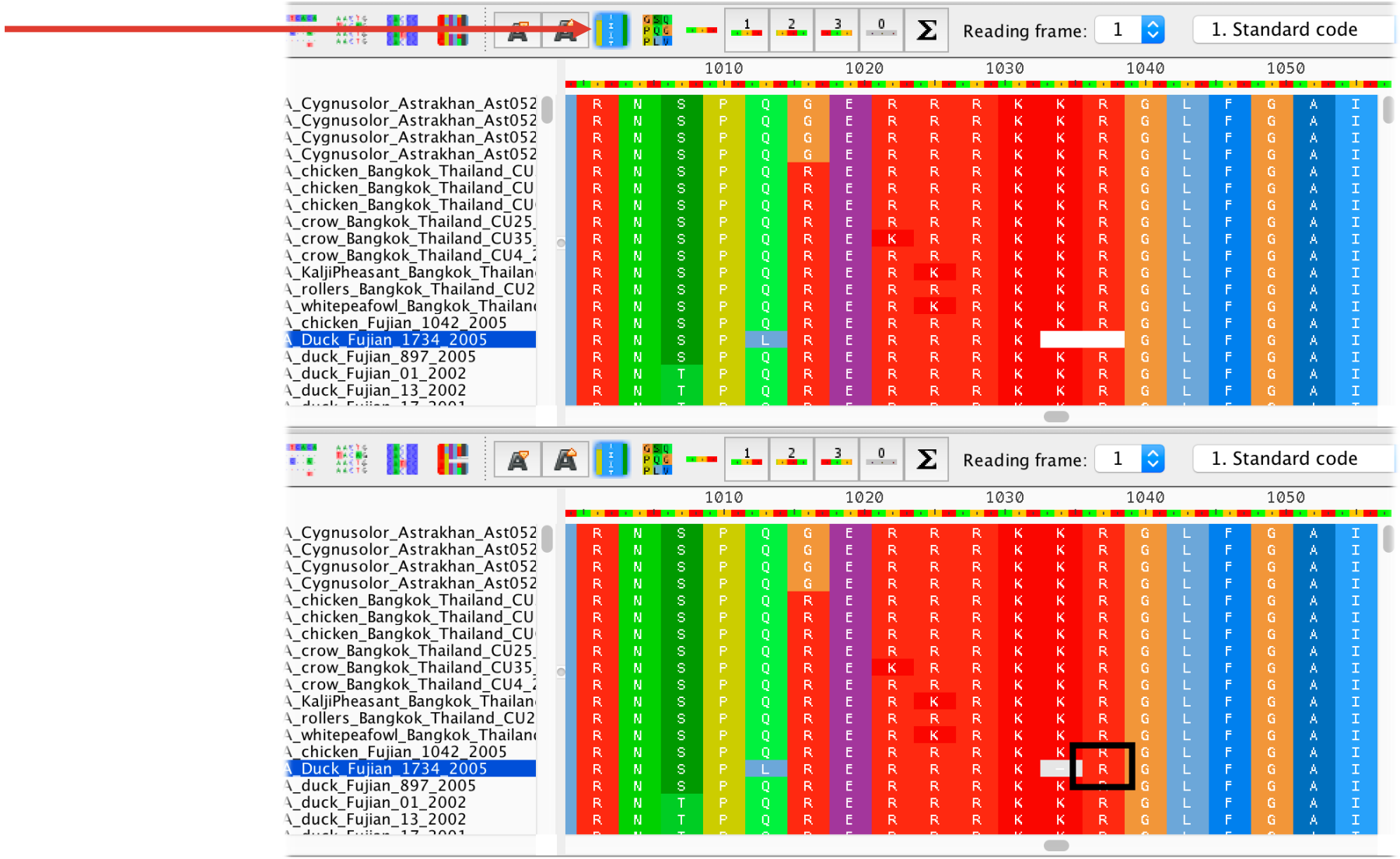
id = .21032606397184R8FnjqzDXbTPuW0l6sMdHlsfnormal, posted at Fri Mar 26 06:40:03 JST 2021
1849 nucleotide sequences, 29903-29903 sites input
This file will be removed after 96 hours.

Inspection and manual editing of the alignment

Sequence Alignments with Aliview

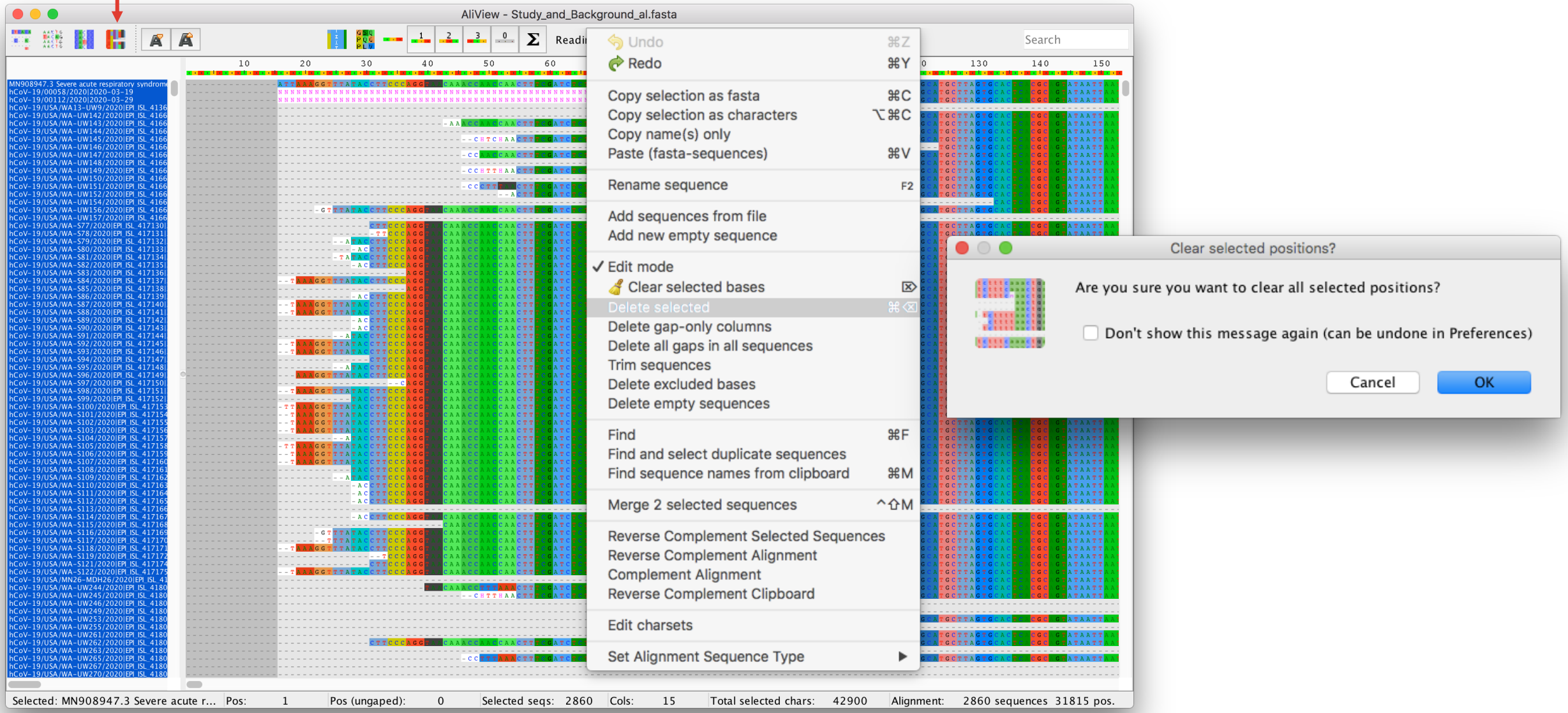


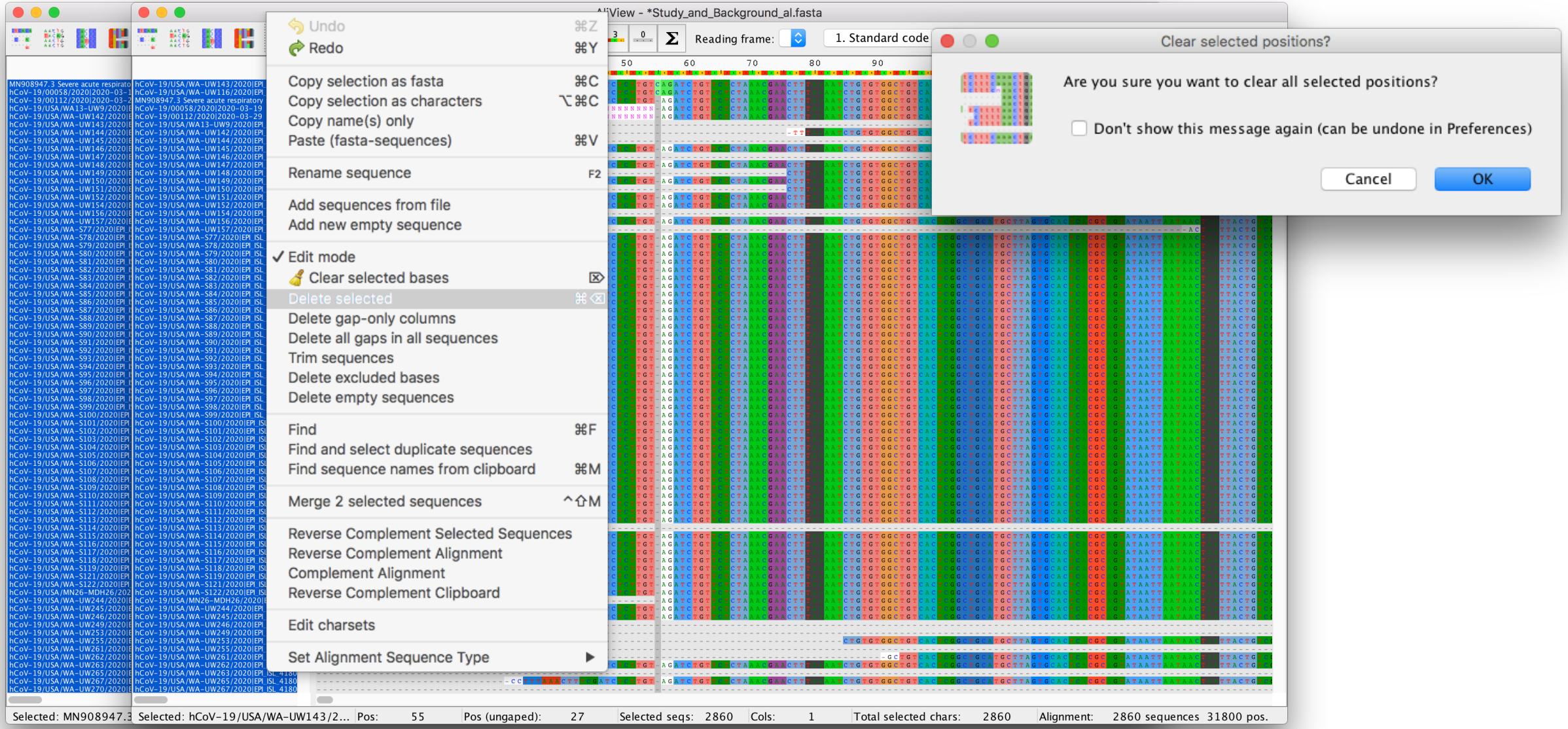
Sequence Alignments with Aliview

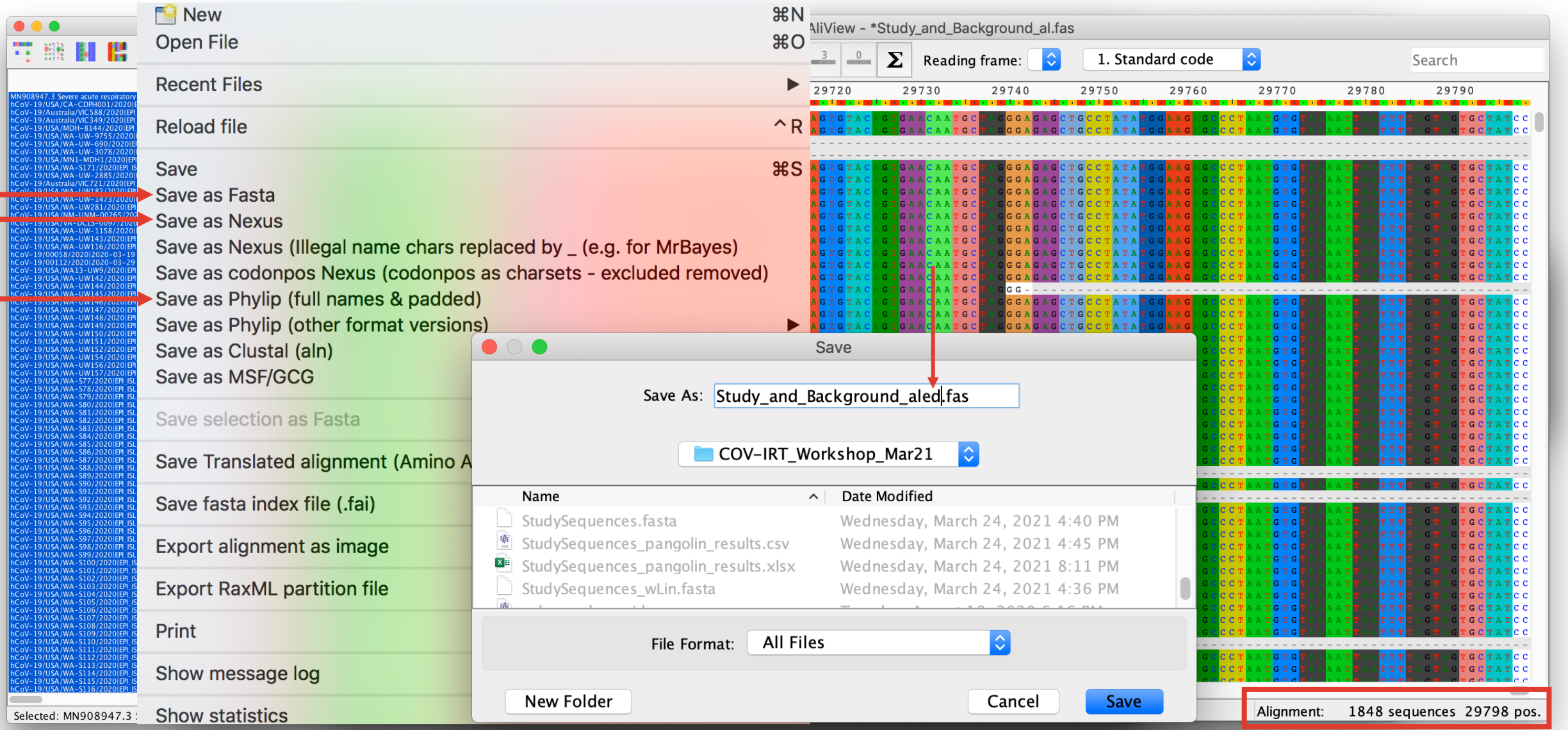


Sequence Alignments with Aliview

Hands-On

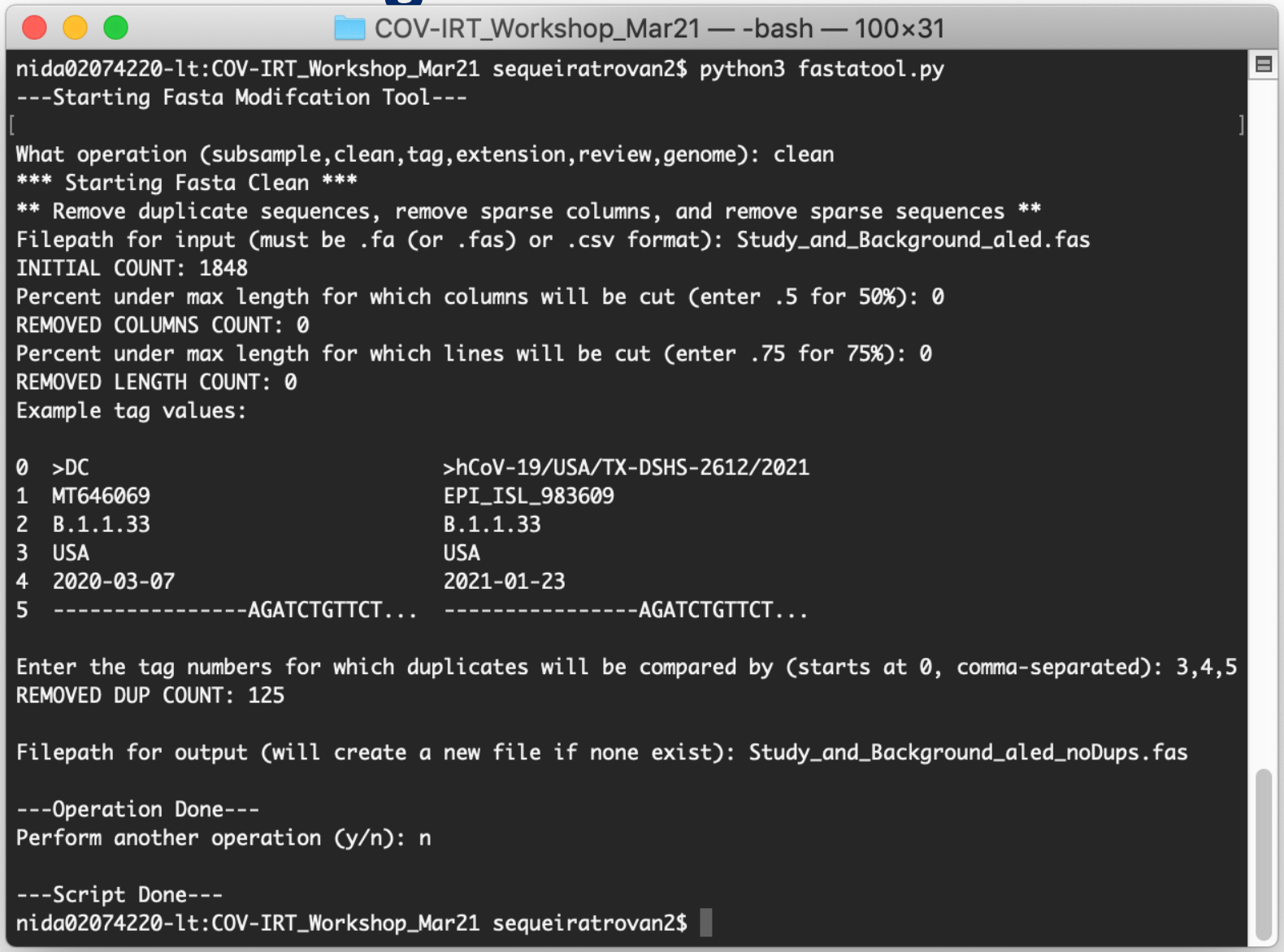

AliView - Study_and_Background_al.fasta
Search
Undo ⌘Z
Redo ⌘Y
Copy selection as fasta ⌘C
Copy selection as characters ⌥⌘C
Copy name(s) only
Paste (fasta-sequences) ⌘V
Rename sequence F2
Add sequences from file
Add new empty sequence
Edit mode
Clear selected bases
Delete selected
Delete gap-only columns
Delete all gaps in all sequences
Trim sequences
Delete excluded bases
Delete empty sequences
Find ⌘F
Find and select duplicate sequences
Find sequence names from clipboard ⌘M
Merge 2 selected sequences ^⇧M
Reverse Complement Selected Sequences
Reverse Complement Alignment
Complement Alignment
Reverse Complement Clipboard
Edit charsets
Set Alignment Sequence Type
Clear selected positions?
Are you sure you want to clear all selected positions?
Don't show this message again (can be undone in Preferences)
Cancel
OK
Selected: MN908947.3 Severe acute r... Pos: 1 Pos (ungaped): 0 Selected seqs: 2860 Cols: 15 Total selected chars: 42900 Alignment: 2860 sequences 31815 pos.


AliView - *Study_and_Background_al.fasta
Undo ⌘Z
Redo ⌘Y
Copy selection as fasta ⌘C
Copy selection as characters ⌥⌘C
Copy name(s) only
Paste (fasta-sequences) ⌘V
Rename sequence F2
Add sequences from file
Add new empty sequence
Edit mode
Clear selected bases
Delete selected
Delete gap-only columns
Delete all gaps in all sequences
Trim sequences
Delete excluded bases
Delete empty sequences
Find ⌘F
Find and select duplicate sequences
Find sequence names from clipboard ⌘M
Merge 2 selected sequences ^⇧M
Reverse Complement Selected Sequences
Reverse Complement Alignment
Complement Alignment
Reverse Complement Clipboard
Edit charsets
Set Alignment Sequence Type
Reading frame:
1. Standard code
Clear selected positions?
Are you sure you want to clear all selected positions?
Don't show this message again (can be undone in Preferences)
Cancel
OK
Selected: MN908947.3
Selected: hCoV-19/USA/WA-UW143/2...
Pos: 55
Pos (ungaped): 27
Selected seqs: 2860
Cols: 1
Total selected chars: 2860
Alignment: 2860 sequences 31800 pos.


New ⌘N
Open File ⌘O
Recent Files
Reload file ^R
Save ⌘S
Save as Fasta
Save as Nexus
Save as Nexus (Illegal name chars replaced by _ (e.g. for MrBayes)
Save as codonpos Nexus (codonpos as charsets - excluded removed)
Save as Phylip (full names & padded)
Save as Phylip (other format versions)
Save as Clustal (aln)
Save as MSF/GCG
Save selection as Fasta
Save Translated alignment (Amino A
Save fasta index file (.fai)
Export alignment as image
Export RaxML partition file
Print
Show message log
Show statistics
AliView - *Study_and_Background_al.fas
Reading frame:
1. Standard code
Search
Save
Save As: Study_and_Background_aled.fas
COV-IRT_Workshop_Mar21
Name
Date Modified
StudySequences.fasta
Wednesday, March 24, 2021 4:40 PM
StudySequences_pangolin_results.csv
Wednesday, March 24, 2021 4:45 PM
StudySequences_pangolin_results.xlsx
Wednesday, March 24, 2021 8:11 PM
StudySequences_wLin.fasta
Wednesday, March 24, 2021 4:36 PM
File Format: All Files
New Folder
Cancel
Save
Alignment: 1848 sequences 29798 pos.
Selected: MN908947.3

Subsampling the background dataset for computational efficiency

Large dataset → Clean and subsample



- Remove duplicates

(date, location, sequence)

→ -125 sequences

→ 1723 sequences

Large dataset → Clean and subsample

COV-IRT_Workshop_Mar21 — -bash — 100×33

```
nida02074220-lt:COV-IRT_Workshop_Mar21 sequeiratrovan2$ python3 subsample_covid.py

** WARNING: file subsamp_data.pic not found in /Volumes/NT5TB/NT1_23Dec17/Postdoc/MountSinaiNIH/Conf
erences/NIH/COV-IRT_Workshop_Mar21
** If the 'U' criterion is used, computations may be
** slower than they could be.
** Results will be correct nonetheless.


** Modified program for COVID: changed meanings of A and G criteria **

Enter FASTA-format file for input: Study_and_Background_aled_noDups.fas
Enter name for FASTA-format output file: Study_and_Background_aled_noDups_sub.fas
Enter delimiter, or return for '|':
Example tag values:

0  DC                           hCoV-19/Uruguay/UY-NYUMC857...
1  MT646069                     EPI_ISL_457953
2  B.1.1.33                     B.1.1.33
3  USA                          Uruguay
4  2020-03-07                   2020-03-23

Note: only the year will be used if the date field is included

Enter one or more field numbers separated by commas or whitespace, or S to load spreadsheet: 3,4

35 categories
Mean category size: 49.23
Minimum: 1
Maximum: 1053

Save a spreadsheet showing category sizes? y(es)/N(o) (RETURN for No): y
Enter name for output spreadsheet (CSV) file: Study_and_Background_aled_noDups.csv
```

- Subsample (date, location)

Large dataset → Clean and subsample

COV-IRT_Workshop_Mar21 — -bash — 100×33

```
Save a spreadsheet showing category sizes? y(es)/N(o) (RETURN for No): y
Enter name for output spreadsheet (CSV) file: Study_and_Background_aled_noDups.csv

Wrote CSV file Study_and_Background_aled_noDups.csv

Launch CSV file Study_and_Background_aled_noDups.csv now? y(es)/N(o) (RETURN for No): y

Launched CSV file

Total samples for some choices of samples per category:

   1  35                  60  401
   2  59                  70  435
   3  75                  80  465
   4  91                  90  495
   5  104                100  517
   6  117                200  717
   7  129                300  917
   8  141                400  1070
   9  151                500  1170
  10  161                600  1270
  20  228                700  1370
  30  278                800  1470
  40  321                900  1570
  50  361               1000  1670

Number to sample from each category: 30

278 sequences will be sampled

Save a spreadsheet showing category sizes and number sampled for editing and reloading? y(es)/N(o) (
RETURN for No): y
Enter name for output spreadsheet (CSV) file: Study_and_Background_aled_noDups_sub.csv
```

- Subsample (date, location)

→ 278 sequences

Large dataset → Clean and subsample

COV-IRT_Workshop_Mar21 — -bash — 100×33

```
Enter name for output spreadsheet (CSV) file: Study_and_Background_aled_noDups_sub.csv

Wrote CSV file Study_and_Background_aled_noDups_sub.csv

Launch CSV file Study_and_Background_aled_noDups_sub.csv now? y(es)/N(o) (RETURN for No): y

Launched CSV file

Specify the criteria for choosing within a category,
in order of importance.  The first criterion takes priority,
so the others only matter when candidates are "tied" for the
first.  Similarly, the second takes priority over all but
the first, and so on.  Randomization applies in all cases.

Specify one or more letters (either case), separated by whitespace, commas, or nothing:

N  No preferences.  Purely random choice.  Must be only letter specified.
U  Seek uniformity of date distribution.
   If used, this must come first in priorities,
   and the date (in effect, year) must be part of
   the category definition.
L  Maximize sequence length (includes internal gaps, but not leading and trailing gaps)
D  Completeness of date.  YMD > YM > Y
M  Presence of month; no preference for YMD over YM.
   It can be meaningful and useful to provide both D and M,
   provided that M comes first and something comes between, e.g., MLD.
G  Minimize number of internal gaps with lengths not divisible by 3 and less than 5
A  Minimize number of ambiguity characters (N, etc.).

Specify preferences for isolate selection in order of priority: UDGAL

Chose 30 (or all) from each category using fields [3, 4] and preferences UDGAL
278 sequences written to Study_and_Background_aled_noDups_sub.fas
```

- Subsample (date, location)

→278 sequences

→Preferences in case sequences from the same location and with the same collection date


COV-IRT
COVID-19 International Research Team


NIH
Fogarty International Center
Advancing Science for Global Health