



Fogarty International Center
Advancing Science for Global Health



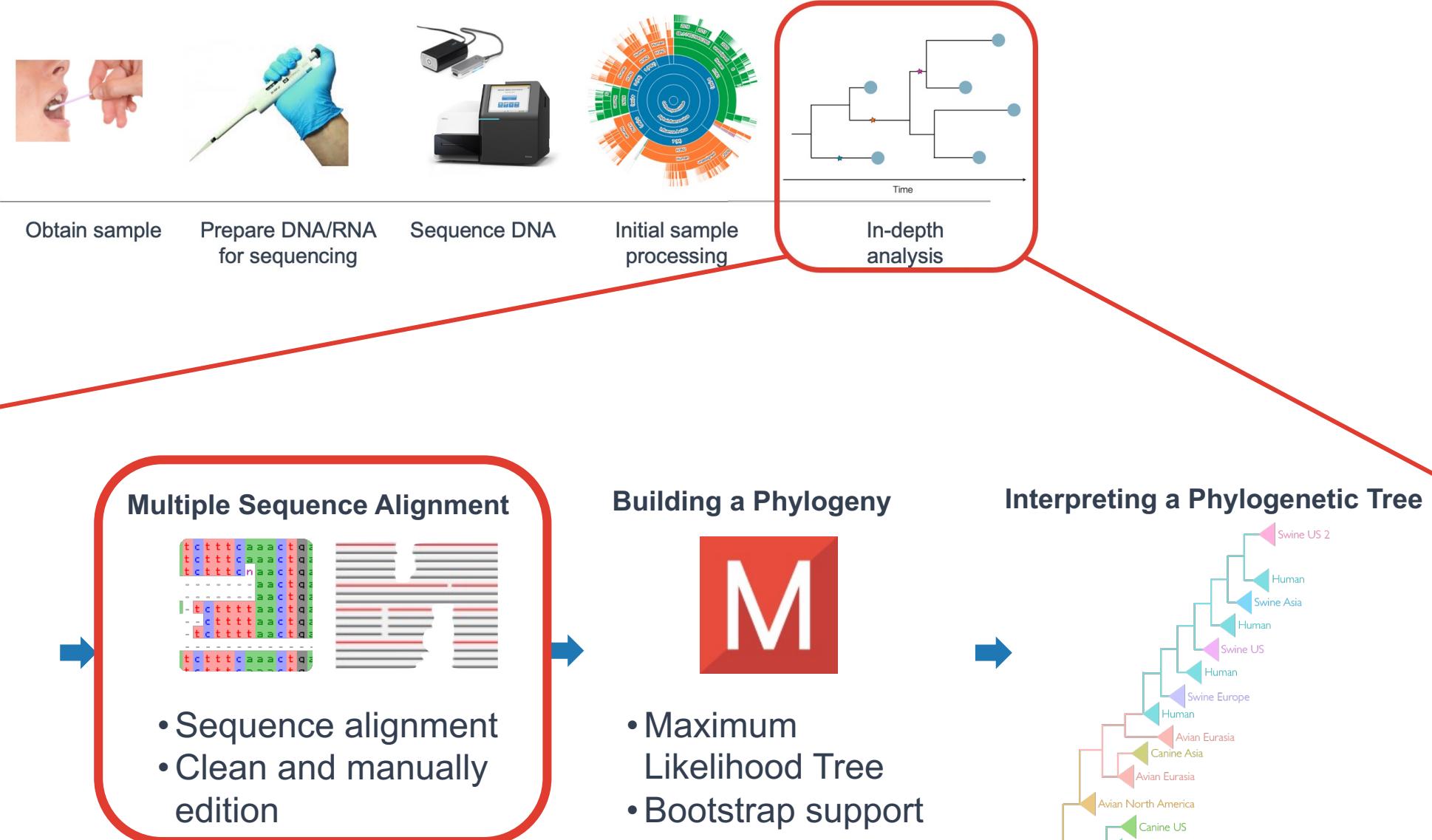
COVID-19 International Research Team

Bioinformatics towards phylogenetics

Navigating AliView

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

Phylogenetics protocol



How to View FASTA Files

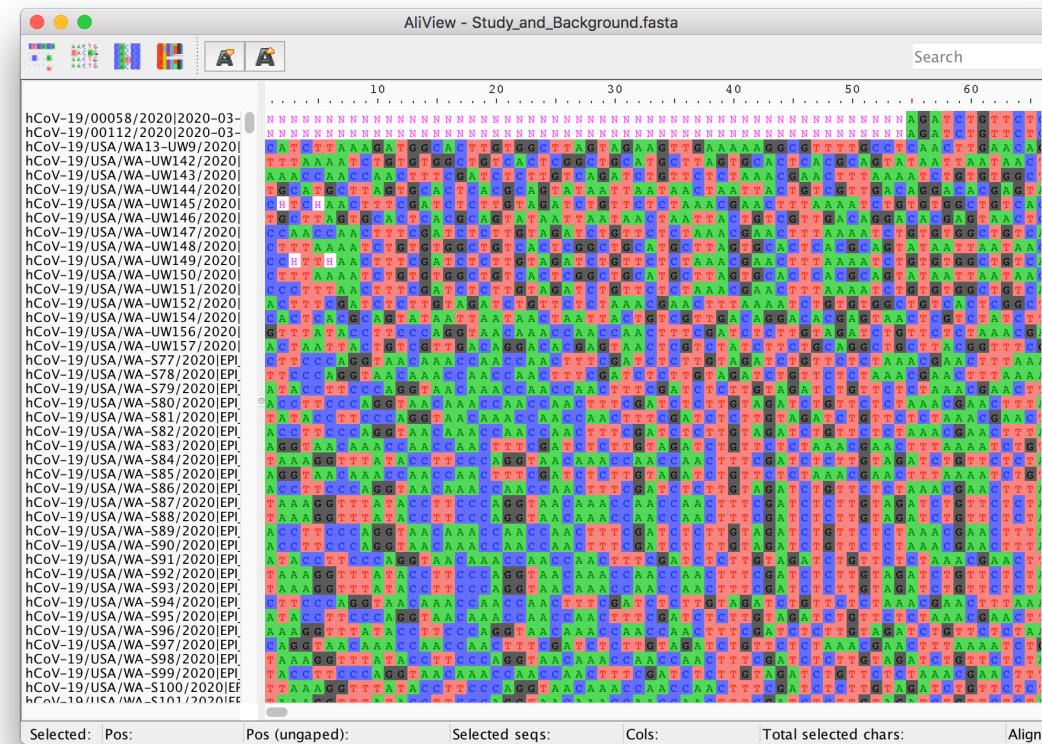


NCBI Resources ▾ How To ▾ Sign in to NCBI

Nucleotide Nucleotide Advanced Search Help

Nucleotide

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.



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

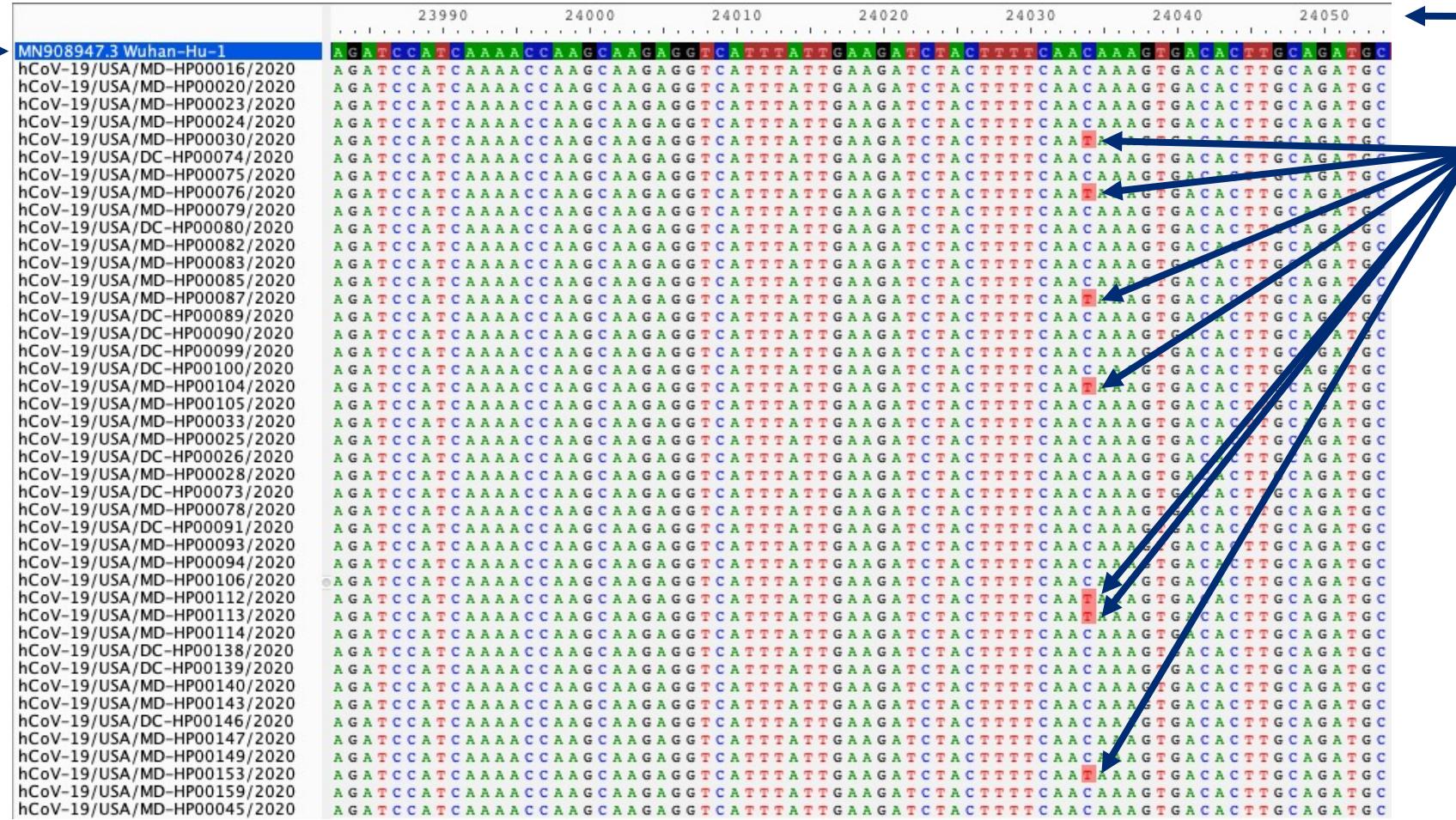
*The reference genome should always have a date **before** the earliest of your samples*

Multiple sequence alignments

Selecting an appropriate reference genome

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

Reference genome



Position along reference genome

Single nucleotide variants, compared to reference

Download MN908947.3 in Genbank

The screenshot shows three browser tabs, each displaying the NCBI Nucleotide page for the SARS-CoV-2 genome entry MN908947.3. The tabs are titled "Home - Nucleotide", "Severe acute resp...", and "Severe acute respiratory syndr...". Each tab has a prominent orange banner at the top stating "COVID-19 is an emerging, rapidly evolving situation." with links to "Public health information (CDC)", "Research information (NIH)", "SARS-CoV-2 data (NCBI)", and "Prevention and treatment information (HHS)".

Left Tab (Home - Nucleotide):

- Header:** Nucleotide
- Content:** Shows a sequence snippet: ACCCAGCACACAT TGTAGCTTACCTG GTTTGCTG...
- Links:** GenBank, Fasta, Graphics, Go to: dropdown.
- Information:** Locus: MN908947, Definition: Severe acute respiratory syndrome coronavirus 2 isolate Wu Hu-1, complete genome., Accession: MN908947, Version: MN908947.3, Keywords: ., Source: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), Organism: Severe acute respiratory syndrome coronavirus 2, Viruses; Riboviria; Nidovirales; Coronaviridae; Betacoronavirus; Sarbecovirus., Reference: 1 (bases 1 to 29903), Authors: Wu,F., Zhao,S., Yu,Tao,Z.W., Tian,J.H., Liu,Y., Wang,Q.M., JOURNAL: Nature 579 (7798) 265-269 (2020).

Middle Tab (Severe acute resp...):

- Header:** Nucleotide
- Content:** Shows a sequence snippet: ACCCAGCACACAT TGTAGCTTACCTG GTTTGCTG...
- Links:** GenBank, Fasta, Graphics, Go to: dropdown.
- Information:** Locus: MN908947, Definition: Severe acute respiratory syndrome coronavirus 2 isolate Wu Hu-1, complete genome., Accession: MN908947, Version: MN908947.3, Keywords: ., Source: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), Organism: Severe acute respiratory syndrome coronavirus 2, Viruses; Riboviria; Nidovirales; Coronaviridae; Betacoronavirus; Sarbecovirus., Reference: 1 (bases 1 to 29903), Authors: Wu,F., Zhao,S., Yu,B., Chen,Y.M., Wang,W., Song,Z.G., Hu,Y., Tao,Z.W., Tian,J.H., Pei,Y.Y., Yuan,M.L., Zhang,Y.L., Dai,F.H., Liu,Y., Wang,Q.M., Zheng,J.J., Xu,L., Holmes,E.C. and Zhang,Y.Z., JOURNAL: Nature 579 (7798) 265-269 (2020).

Right Tab (Severe acute respiratory syndr...):

- Header:** Nucleotide
- Content:** Shows a sequence snippet: ACCCAGCACACAT TGTAGCTTACCTG GTTTGCTG...
- Links:** GenBank, Fasta, Graphics, Go to: dropdown.
- Information:** Locus: MN908947, Definition: Severe acute respiratory syndrome coronavirus 2 isolate Wu Hu-1, complete genome., Accession: MN908947, Version: MN908947.3, Keywords: ., Source: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), Organism: Severe acute respiratory syndrome coronavirus 2, Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes; Nidovirales; Cornidovirinae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus., Reference: 1 (bases 1 to 29903), Authors: Wu,F., Zhao,S., Yu,B., Chen,Y.M., Wang,W., Song,Z.G., Hu,Y., Tao,Z.W., Tian,J.H., Pei,Y.Y., Yuan,M.L., Zhang,Y.L., Dai,F.H., Liu,Y., Wang,Q.M., Zheng,J.J., Xu,L., Holmes,E.C. and Zhang,Y.Z., JOURNAL: Nature 579 (7798) 265-269 (2020).

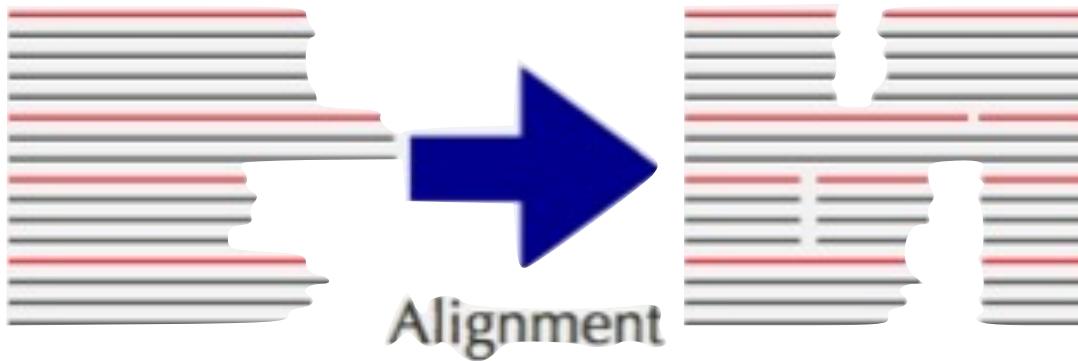
Common Right-Hand Side Elements:

- Send to:** dropdown set to "shown".
- Choose Destination:**
 - Complete Record
 - Coding Sequences
 - Gene Features
- File Options:** File, Clipboard, Collections, Analysis Tool.
- Download Options:** Download 1 item, Format: FASTA, Show GI: checkbox (unchecked), Create File button.
- Description:** Retrieve, view, and download SARS-CoV-2 coronavirus genomic and protein sequences.
- Related Information:** Assembly, Protein, PubMed.

Alignment software

Sequence Alignments

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



Align with MAFFT and a reference genome

MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences

Download version

Mac OS X
Windows
Linux
Source

Online version

Alignment
mafft --add
Merge
Phylogeny
Rough tree
Merits / limitations
Algorithms
Tips
Benchmarks
Feedback

MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences

Download version

Mac OS X
Windows
Linux
Source

Online version

Alignment
mafft --add
Merge
Phylogeny
Rough tree
Merits / limitations
Algorithms
Tips
Benchmarks
Feedback

Add new sequences to an existing alignment

https://mafft.cbrc.jp/alignment/server/add_fragments.html?frommanual

Allow unusual symbols (Selenocysteine "U", Inosine "i", non-alphabetical characters, etc.) [Help](#)

Allow unusual symbols (Selenocysteine "U", Inosine "i", non-alphabetical characters, etc.) [Help](#)

UPPERCASE / lowercase:

Same as input ↑ Did not work Oct/23 –. Fixed Oct/27.

Amino acid → UPPERCASE / Nucleotide → lowercase

Direction of nucleotide sequences:

Same as input ↑

Adjust direction according to the first sequence (accurate enough for most cases) [Beta](#)

Adjust direction according to the first sequence (only for highly divergent data; very slow) [Beta](#)

Output order:

Same as input ↑

Aligned

Sequence title:

Same as input ↑

Insert "New!" at the head of title of each new sequence

Job name (optional):

Workshop ↑ (basic Latin alphabet, number and space only)

Notify when finished (optional; recommended when submitting large data):

Email address: nidia.trovao@nih.gov ↑

Submit Reset

NIH Fogarty

10

Message

Delete Archive

mafft

M

Alignment
<https://mafft.cbrc.jp/alignment/>

If you have

MAFFT-FFT-NS-fragment Result

CLUSTAL format alignment by MAFFT (v7.411)

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 X

Multiple sequence alignment by X

[Clustal format](https://mafft.cbrc.jp/alignment/s) | [Fasta format](https://mafft.cbrc.jp/alignment/s) | [MAFFT result](https://mafft.cbrc.jp/alignment/s) | [View](#) | [Tree](#)

View

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

GUIDANCE2 computes the residue-wise confidence .

Refine dataset

Phylogenetic tree

(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 --any symbol --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 = .21032606397184R8FnjrzDXbTPuW0l6sMdHlsfnormal, 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

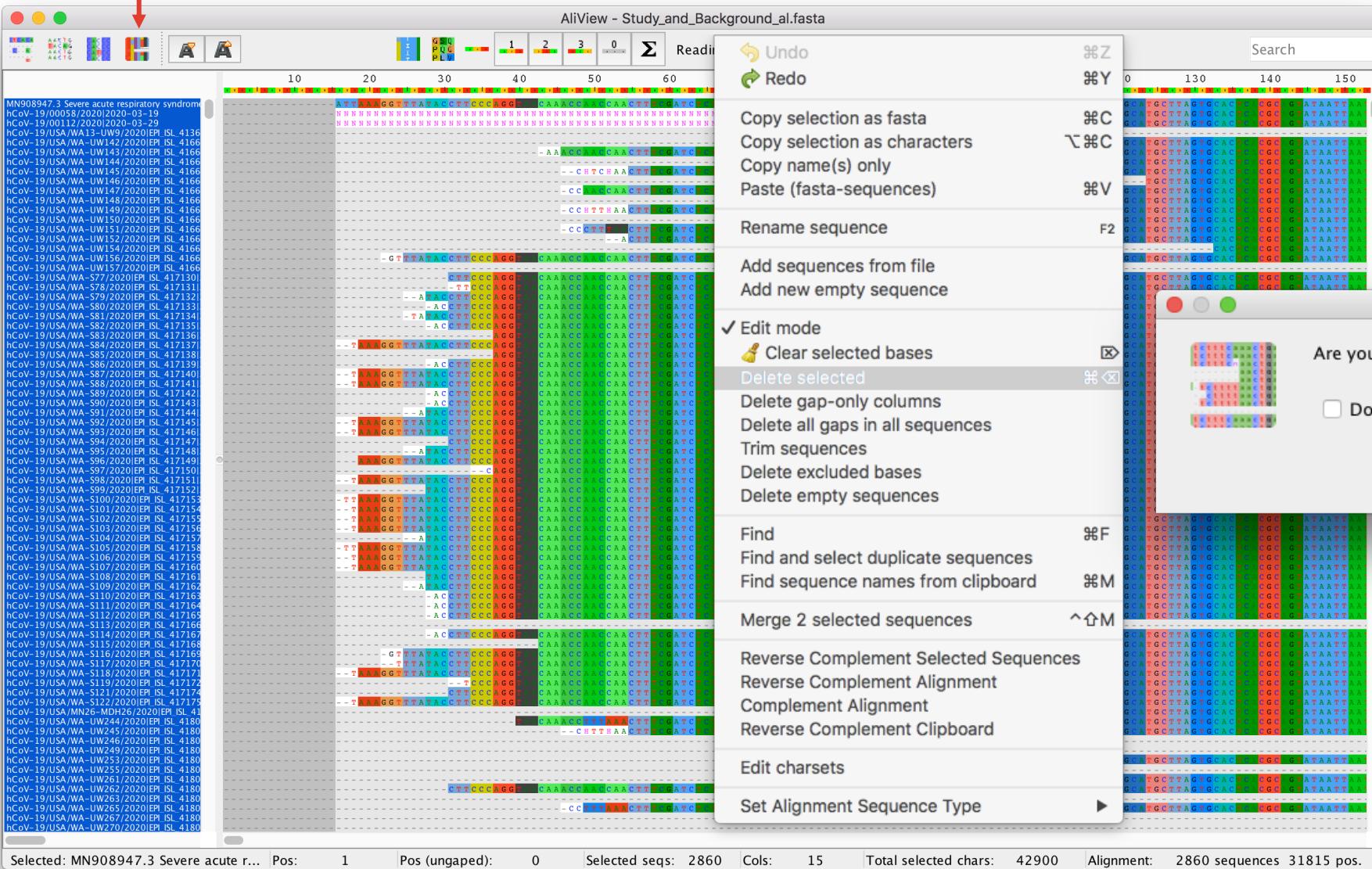


Sequence Alignments with Aliview



Hands-On

3/28/21



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.

View - *Study_and_Background_al.fasta

Undo Redo

Copy selection as fasta
Copy selection as characters
Copy name(s) only
Paste (fasta-sequences)

Rename sequence
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
Find and select duplicate sequences
Find sequence names from clipboard

Merge 2 selected sequences

Reverse Complement Selected Sequences
Reverse Complement Alignment
Complement Alignment
Reverse Complement Clipboard

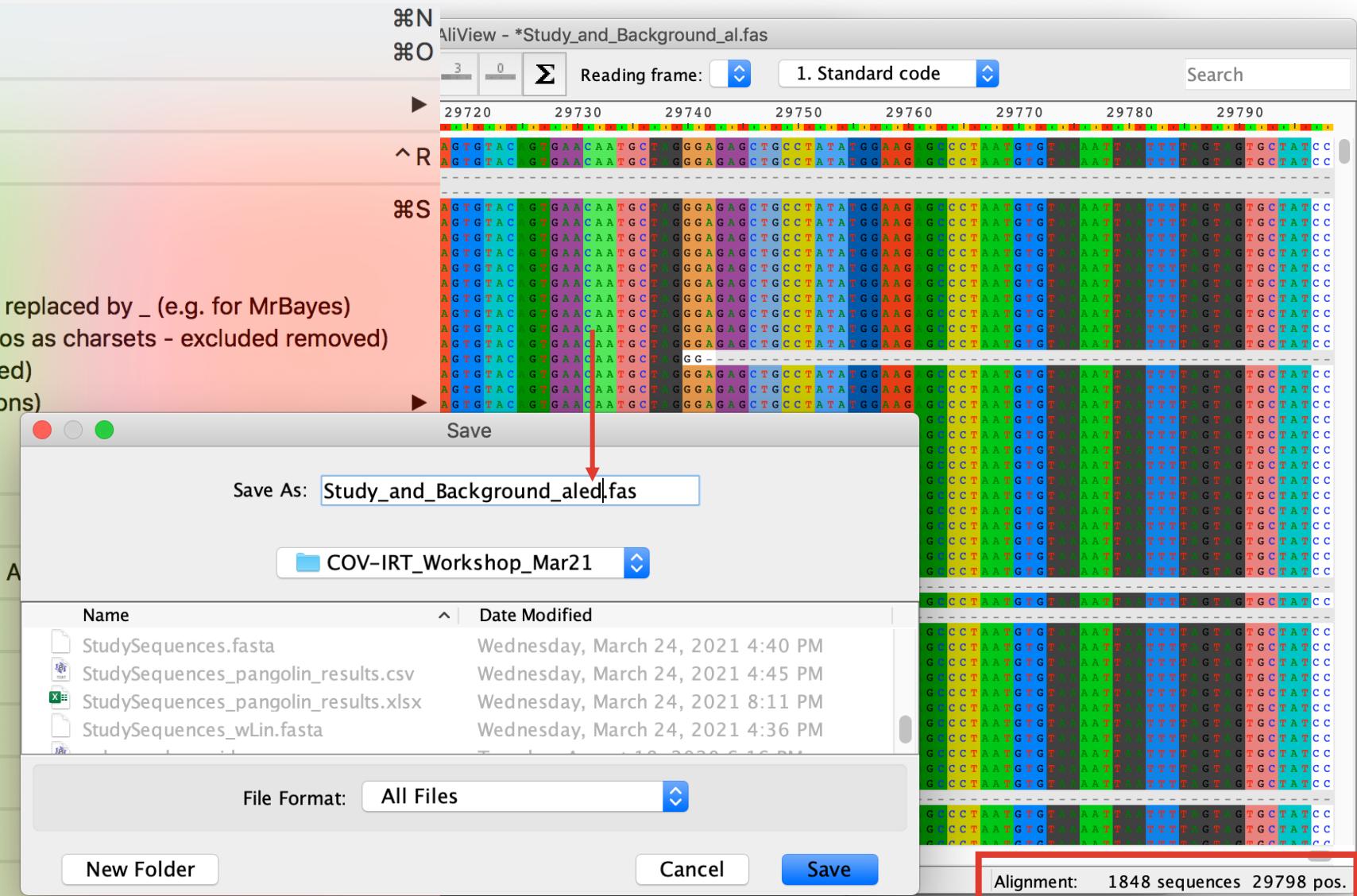
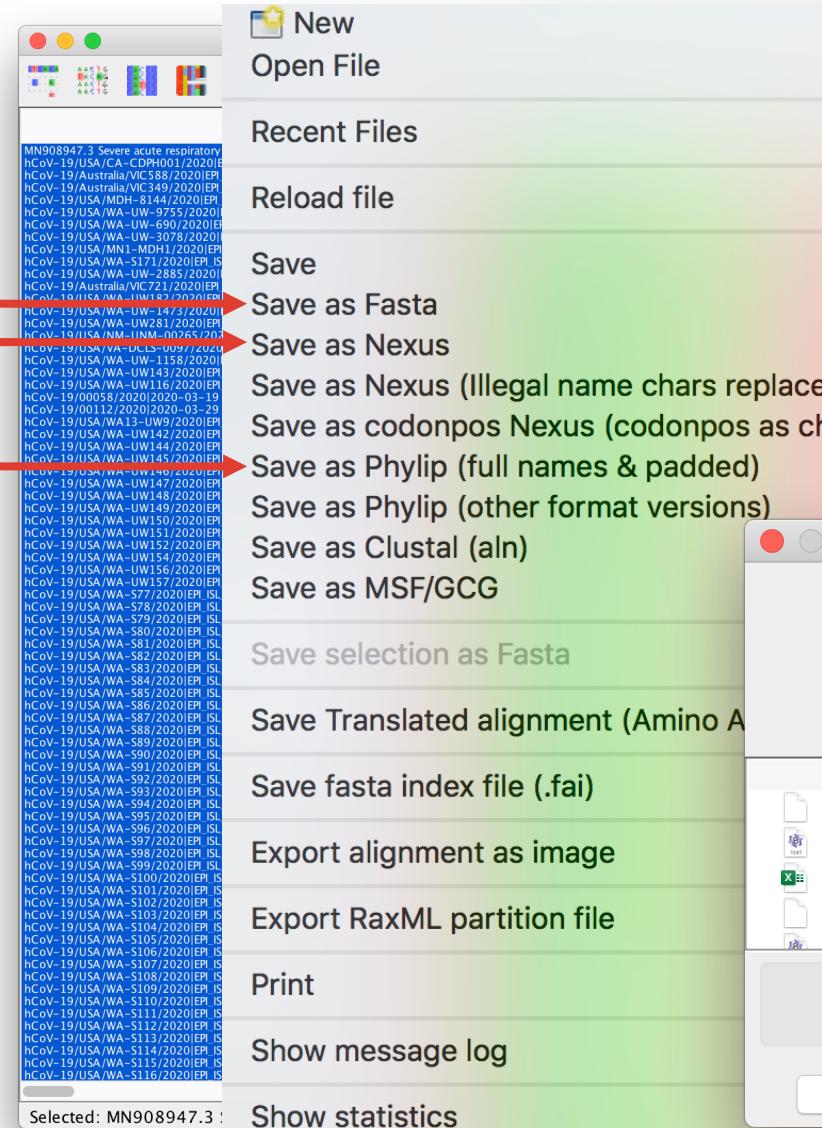
Edit charsets

Set Alignment Sequence Type

Reading frame: 1. Standard code

Are you sure you want to clear all selected positions?
 Don't show this message again (can be undone in Preferences)

Cancel OK



Subsampling the background dataset for computational efficiency

Large dataset → Clean and subsample

```
COV-IRT_Workshop_Mar21 — bash — 100x31
nida02074220-lt:COV-IRT_Workshop_Mar21 sequeiratrovant$ python3 fastatool.py
---Starting Fasta Modification Tool---

What operation (subsample,clean,tag,extension,review,genome): clean
*** Starting Fasta Clean ***
** Remove duplicate sequences, remove sparse columns, and remove sparse sequences **
Filepath for input (must be .fa (or .fas) or .csv format): Study_and_Background_aled.fas
INITIAL COUNT: 1848
Percent under max length for which columns will be cut (enter .5 for 50%): 0
REMOVED COLUMNS COUNT: 0
Percent under max length for which lines will be cut (enter .75 for 75%): 0
REMOVED LENGTH COUNT: 0
Example tag values:

0 >DC                               >hCoV-19/USA/TX-DSHS-2612/2021
1 MT646069                            EPI_ISL_983609
2 B.1.1.33                             B.1.1.33
3 USA                                  USA
4 2020-03-07                           2021-01-23
5 -----AGATCTGTTCT... -----AGATCTGTTCT...

Enter the tag numbers for which duplicates will be compared by (starts at 0, comma-separated): 3,4,5
REMOVED DUP COUNT: 125

Filepath for output (will create a new file if none exist): Study_and_Background_aled_noDups.fas

---Operation Done---
Perform another operation (y/n): n

---Script Done---
nida02074220-lt:COV-IRT_Workshop_Mar21 sequeiratrovant$
```

- Remove duplicates
(date, location, sequence)
 - -125 sequences
 - 1723 sequences

Large dataset → Clean and subsample

```
COV-IRT_Workshop_Mar21 — -bash — 100x33
[nida02074220-lt:COV-IRT_Workshop_Mar21 sequeiratrovant$ python3 subsample_covid.py

** WARNING: file subsamp_data.pic not found in /Volumes/NT5TB/NT1_23Dec17/Postdoc/MountSinaiNIH/Conferences/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

```
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 — 100x33
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



COVID-19 International Research Team

