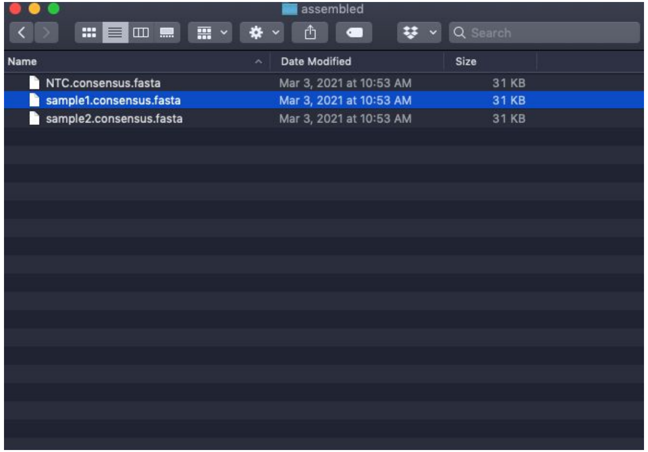
Agenda

- Viewing FASTA files
- Downloading publicly available genomes
- Concatenating consensus genomes
- Sequence Alignment
- Tree inference

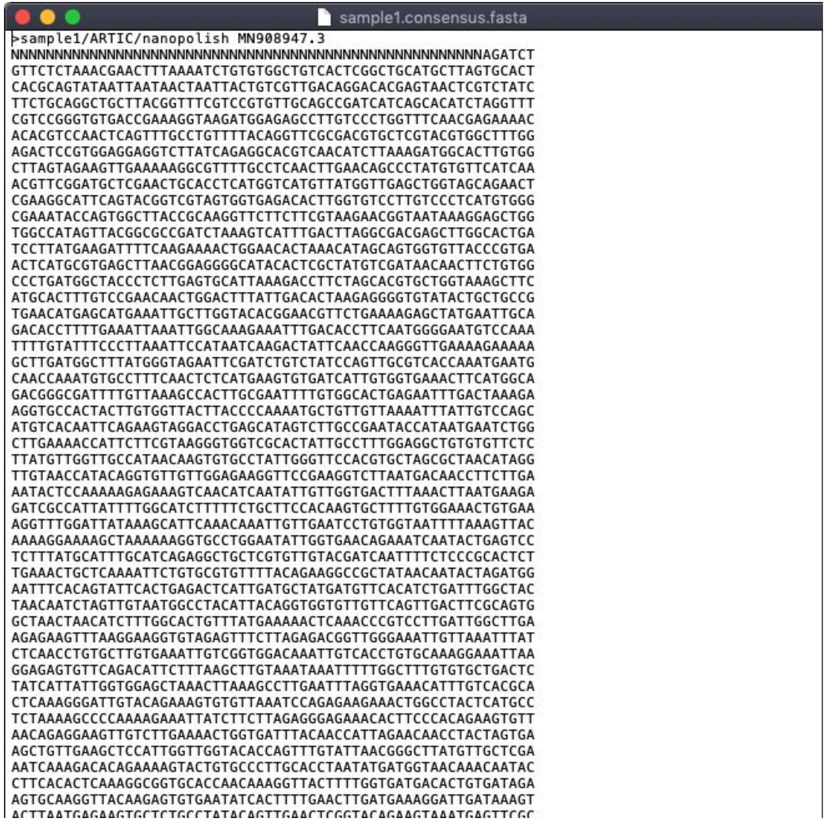
Viewing FASTA files

How to view FASTA files



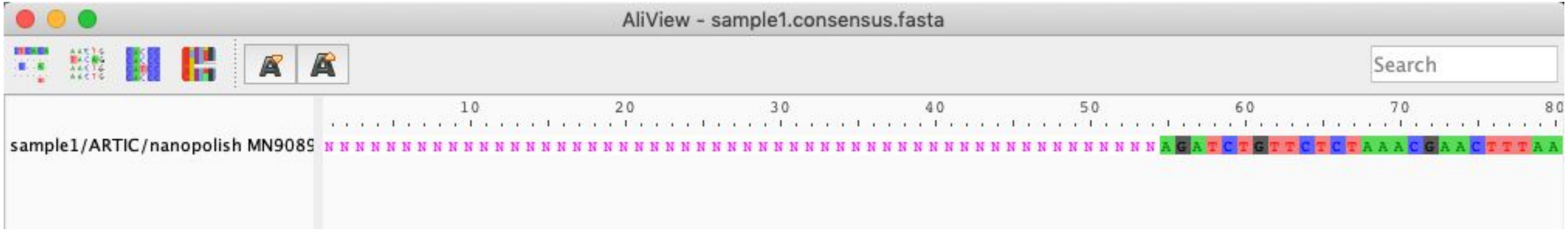
Name	Date Modified	Size
NTC.consensus.fasta	Mar 3, 2021 at 10:53 AM	31 KB
sample1.consensus.fasta	Mar 3, 2021 at 10:53 AM	31 KB
sample2.consensus.fasta	Mar 3, 2021 at 10:53 AM	31 KB

- Open in your favorite text editor:
    - TextEdit (Mac)
    - Notepad++ (Windows)
    - gedit (Ubuntu)



```
>sample1/ARTIC/nanopolish MN908947.3
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGATCT
GTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACT
CACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATC
TTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTT
CGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAAC
ACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGG
AGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAA
ACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACT
CGAAGGCATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGG
CGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGG
TGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGA
TCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGA
ACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTC
ATGCACTTTGTCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCG
TGAACATGAGCATGAAATTGCTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCA
GACACCTTTTGAAATTAAATTGGCAAAGAAATTTGACACCTTCAATGGGGAATGTCCAAA
TTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAACCAAGGGTTGAAAAGAAAAA
GCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCACCAAATGAATG
CAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCA
GACGGGCGATTTTGTTAAAGCCACTTGCGAATTTTGTGGCACTGAGAATTTGACTAAAGA
AGGTGCCACTACTTGTGGTTACTTACCCCAAAATGCTGTTGTTAAAATTTATTGTCCAGC
ATGTCACAATTCAGAAGTAGGACCTGAGCATAGTCTTGCCGAATACCATAATGAATCTGG
CTTGAAAACCATTCTTCGTAAGGGTGGTCGCACTATTGCCTTTGGAGGCTGTGTGTTCTC
TTATGTTGGTTGCCATAACAAGTGTGCCTATTGGGTTCCACGTGCTAGCGCTAACATAGG
TTGTAACCATACAGGTGTTGTTGGAGAAGGTTCCGAAGGTCTTAATGACAACCTTCTTGA
AATACTCCAAAAAGAGAAAGTCAACATCAATATTGTTGGTGACTTTAAACTTAATGAAGA
GATCGCCATTATTTTGGCATCTTTTTCTGCTTCCACAAGTGCTTTTGTGGAAACTGTGAA
AGGTTTGGATTATAAAGCATTCAAACAAATTGTTGAATCCTGTGGTAATTTTAAAGTTAC
AAAAGGAAAAGCTAAAAAAGGTGCCTGGAATATTGGTGAACAGAAATCAATACTGAGTCC
TCTTTATGCATTTGCATCAGAGGCTGCTCGTGTTGTACGATCAATTTTCTCCCGCACTCT
TGAAACTGCTCAAAATTCTGTGCGTGTTTTACAGAAGGCCGCTATAACAATACTAGATGG
AATTTCACAGTATTCACTGAGACTCATTGATGCTATGATGTTCACATCTGATTTGGCTAC
TAACAATCTAGTTGTAATGGCCTACATTACAGGTGGTGTTGTTCAGTTGACTTCGCAGTG
GCTAACTAACATCTTTGGCACTGTTTATGAAAAACTCAAACCCGTCCTTGATTGGCTTGA
AGAGAAGTTTAAGGAAGGTGTAGAGTTTCTTAGAGACGGTTGGGAAATTGTTAAATTTAT
CTCAACCTGTGCTTGTGAAATTGTCGGTGGACAAATTGTCACCTGTGCAAAGGAAATTAA
GGAGAGTGTTCAGACATTCTTTAAGCTTGTAAATAAATTTTTGGCTTTGTGTGCTGACTC
TATCATTATTGGTGGAGCTAAACTTAAAGCCTTGAATTTAGGTGAAACATTTGTCACGCA
CTCAAAGGGATTGTACAGAAAGTGTGTTAAATCCAGAGAAGAAACTGGCCTACTCATGCC
TCTAAAAGCCCCAAAAGAAATTATCTTCTTAGAGGGAGAAACACTTCCCACAGAAGTGTT
AACAGAGGAAGTTGTCTTGAAAACTGGTGATTTACAACCATTAGAACAACCTACTAGTGA
AGCTGTTGAAGCTCCATTGGTTGGTACACCAGTTTGTATTAACGGGCTTATGTTGCTCGA
AATCAAAGACACAGAAAAGTACTGTGCCCTTGCACCTAATATGATGGTAACAAACAATAC
CTTCACACTCAAAGGCGGTGCACCAACAAAGGTTACTTTTGGTGATGACACTGTGATAGA
AGTGCAAGGTTACAAGAGTGTGAATATCACTTTTGAACTTGATGAAAGGATTGATAAAGT
ACTTAATGAGAAGTGCTCTGCCTATACAGTTGAACTCGGTACAGAAGTAAATGAGTTCGC
```

How to view FASTA files



AliView - sample1.consensus.fasta

```
                           10        20        30        40        50        60        70        80
sample1/ARTIC/nanopolish MN9089 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGATCTGTTCTCTAAACGAACTTTAA
```

- You can also open FASTA files in specialized software for working with genomic data
- e.g., AliView: https://ormbunkar.se/aliview/ (free software)

- Or view directly on command line using less

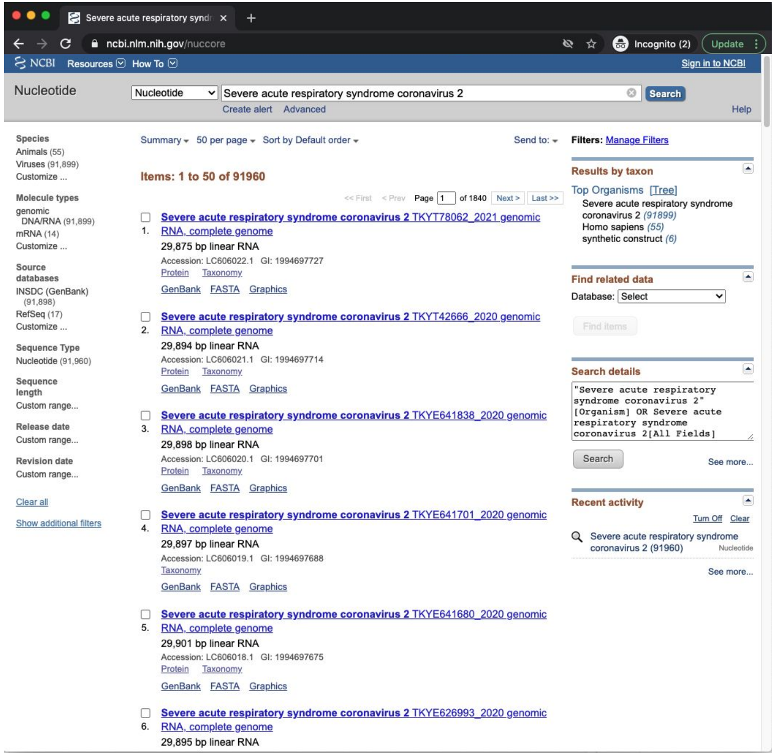
Downloading Publicly Available Data

Selecting background data for phylogenetics

- For pathogens with limited data, use all available sequences
- Selecting background data is not a trivial task
- Considerations include:
  - Geography
  - Time period
  - Hosts / sample source
  - Subsampling of highly similar sequences

There is a lack of broadly useful sampling guidance for phylogenetics

Downloading publicly available genomes



Nucleotide

Nucleotide: Severe acute respiratory syndrome coronavirus 2

Create alert Advanced

Species
Animals (55)
Viruses (91,899)
Customize ...

Molecule types
genomic DNA/RNA (91,899)
mRNA (14)
Customize ...

Source databases
INSDC (GenBank) (91,898)
RefSeq (17)
Customize ...

Sequence Type
Nucleotide (91,960)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

Clear all

Show additional filters

Summary 50 per page Sort by Default order

Items: 1 to 50 of 91960

<< First < Prev Page 1 of 1840 Next > Last >>

□ 1. **Severe acute respiratory syndrome coronavirus 2** TKYT78062_2021 genomic RNA, complete genome
29,875 bp linear RNA
Accession: LC606022.1 GI: 1994697727
Protein Taxonomy
GenBank FASTA Graphics

□ 2. **Severe acute respiratory syndrome coronavirus 2** TKYT42666_2020 genomic RNA, complete genome
29,894 bp linear RNA
Accession: LC606021.1 GI: 1994697714
Protein Taxonomy
GenBank FASTA Graphics

□ 3. **Severe acute respiratory syndrome coronavirus 2** TKYE641838_2020 genomic RNA, complete genome
29,898 bp linear RNA
Accession: LC606020.1 GI: 1994697701
Protein Taxonomy
GenBank FASTA Graphics

□ 4. **Severe acute respiratory syndrome coronavirus 2** TKYE641701_2020 genomic RNA, complete genome
29,897 bp linear RNA
Accession: LC606019.1 GI: 1994697688
Taxonomy
GenBank FASTA Graphics

□ 5. **Severe acute respiratory syndrome coronavirus 2** TKYE641680_2020 genomic RNA, complete genome
29,901 bp linear RNA
Accession: LC606018.1 GI: 1994697675
Protein Taxonomy
GenBank FASTA Graphics

□ 6. **Severe acute respiratory syndrome coronavirus 2** TKYE626993_2020 genomic RNA, complete genome
29,895 bp linear RNA

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
Severe acute respiratory syndrome coronavirus 2 *(91899)*
Homo sapiens *(55)*
synthetic construct *(6)*

Find related data

Database: Select

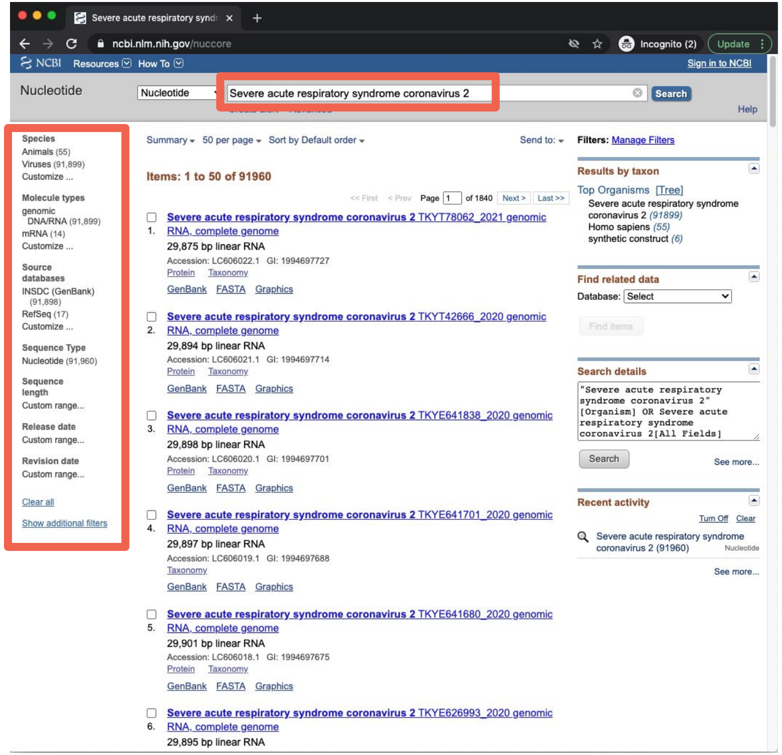
Search details

```
"Severe acute respiratory
syndrome coronavirus 2"
[Organism] OR Severe acute
respiratory syndrome
coronavirus 2[All Fields]
```

Recent activity

Severe acute respiratory syndrome coronavirus 2 (91960) Nucleotide

Downloading publicly available genomes



Nucleotide: Severe acute respiratory syndrome coronavirus 2

Species
Animals (55)
Viruses (91,899)
Customize ...

Molecule types
genomic DNA/RNA (91,899)
mRNA (14)
Customize ...

Source databases
INSDC (GenBank) (91,898)
RefSeq (17)
Customize ...

Sequence Type
Nucleotide (91,960)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

Clear all

Show additional filters

Summary ▾ 50 per page ▾ Sort by Default order ▾ Send to: ▾

Items: 1 to 50 of 91960

<< First < Prev Page 1 of 1840 Next > Last >>

1. □ **Severe acute respiratory syndrome coronavirus 2** TKYT78062_2021 genomic RNA, complete genome
29,875 bp linear RNA
Accession: LC606022.1 GI: 1994697727
Protein Taxonomy
GenBank FASTA Graphics

2. □ **Severe acute respiratory syndrome coronavirus 2** TKYT42666_2020 genomic RNA, complete genome
29,894 bp linear RNA
Accession: LC606021.1 GI: 1994697714
Protein Taxonomy
GenBank FASTA Graphics

3. □ **Severe acute respiratory syndrome coronavirus 2** TKYE641838_2020 genomic RNA, complete genome
29,898 bp linear RNA
Accession: LC606020.1 GI: 1994697701
Protein Taxonomy
GenBank FASTA Graphics

4. □ **Severe acute respiratory syndrome coronavirus 2** TKYE641701_2020 genomic RNA, complete genome
29,897 bp linear RNA
Accession: LC606019.1 GI: 1994697688
Taxonomy
GenBank FASTA Graphics

5. □ **Severe acute respiratory syndrome coronavirus 2** TKYE641680_2020 genomic RNA, complete genome
29,901 bp linear RNA
Accession: LC606018.1 GI: 1994697675
Protein Taxonomy
GenBank FASTA Graphics

6. □ **Severe acute respiratory syndrome coronavirus 2** TKYE626993_2020 genomic RNA, complete genome
29,895 bp linear RNA

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
Severe acute respiratory syndrome coronavirus 2 *(91899)*
Homo sapiens *(55)*
synthetic construct *(6)*

Find related data

Database: Select

Find items

Search details

```
"Severe acute respiratory
syndrome coronavirus 2"
[Organism] OR Severe acute
respiratory syndrome
coronavirus 2[All Fields]
```

Search See more...

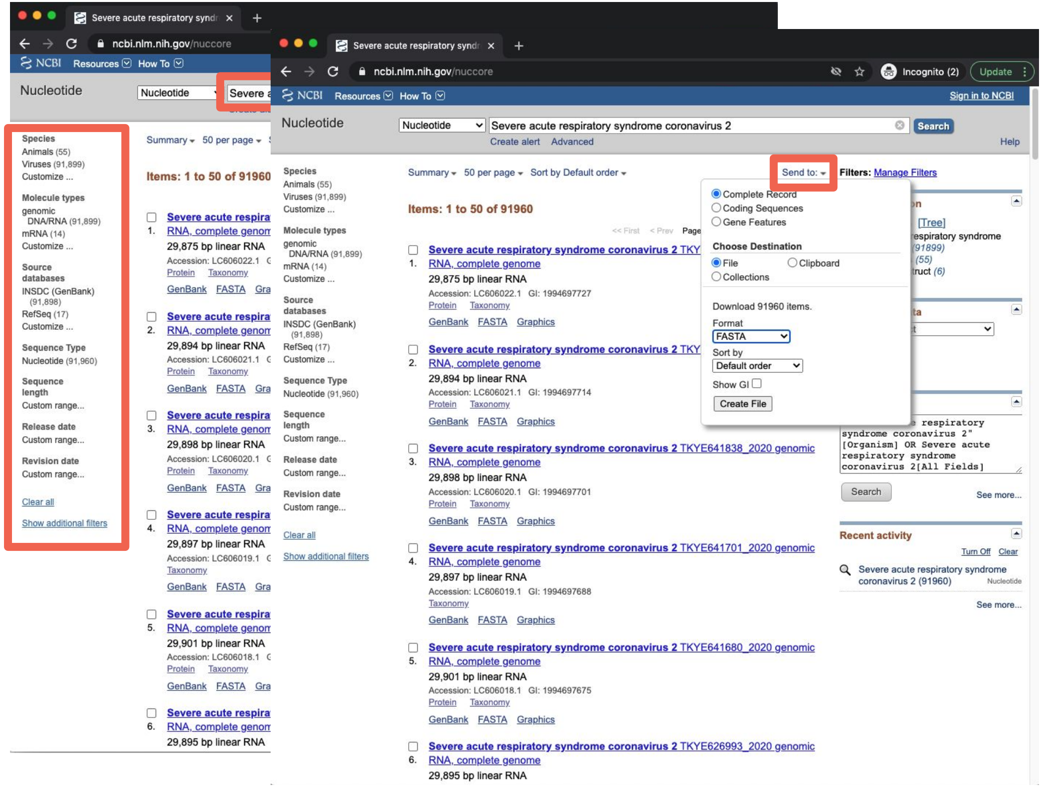
Recent activity

Turn Off Clear

Severe acute respiratory syndrome coronavirus 2 (91960) Nucleotide

See more...

Downloading publicly available genomes



Severe acute respiratory syndr ×

ncbi.nlm.nih.gov/nuccore

Incognito (2) Update

NCBI Resources How To Sign in to NCBI

Nucleotide | Nucleotide | Severe acute respiratory syndrome coronavirus 2 | Search

Create alert Advanced Help

Summary 50 per page Sort by Default order

Send to:

Filters: Manage Filters

- ● Complete Record
- ○ Coding Sequences
- ○ Gene Features

Choose Destination

- ● File ○ Clipboard
- ○ Collections

Download 91960 items.

Format: FASTA

Sort by: Default order

Show GI □

Create File

Species
Animals (55)
Viruses (91,899)
Customize ...

Molecule types
genomic
DNA/RNA (91,899)
mRNA (14)
Customize ...

Source databases
INSDC (GenBank) (91,898)
RefSeq (17)
Customize ...

Sequence Type
Nucleotide (91,960)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

Clear all

Show additional filters

Items: 1 to 50 of 91960

<< First < Prev Page

1. □ Severe acute respiratory syndrome coronavirus 2 TKY... RNA, complete genome
29,875 bp linear RNA
Accession: LC606022.1 GI: 1994697727
Protein Taxonomy
GenBank FASTA Graphics

2. □ Severe acute respiratory syndrome coronavirus 2 TKY... RNA, complete genome
29,894 bp linear RNA
Accession: LC606021.1 GI: 1994697714
Protein Taxonomy
GenBank FASTA Graphics

3. □ Severe acute respiratory syndrome coronavirus 2 TKYE641838_2020 genomic RNA, complete genome
29,898 bp linear RNA
Accession: LC606020.1 GI: 1994697701
Protein Taxonomy
GenBank FASTA Graphics

4. □ Severe acute respiratory syndrome coronavirus 2 TKYE641701_2020 genomic RNA, complete genome
29,897 bp linear RNA
Accession: LC606019.1 GI: 1994697688
Taxonomy
GenBank FASTA Graphics

5. □ Severe acute respiratory syndrome coronavirus 2 TKYE641680_2020 genomic RNA, complete genome
29,901 bp linear RNA
Accession: LC606018.1 GI: 1994697675
Protein Taxonomy
GenBank FASTA Graphics

6. □ Severe acute respiratory syndrome coronavirus 2 TKYE626993_2020 genomic RNA, complete genome
29,895 bp linear RNA

...on
[Tree]
...espiratory syndrome
(91899)
(55)
...truct (6)

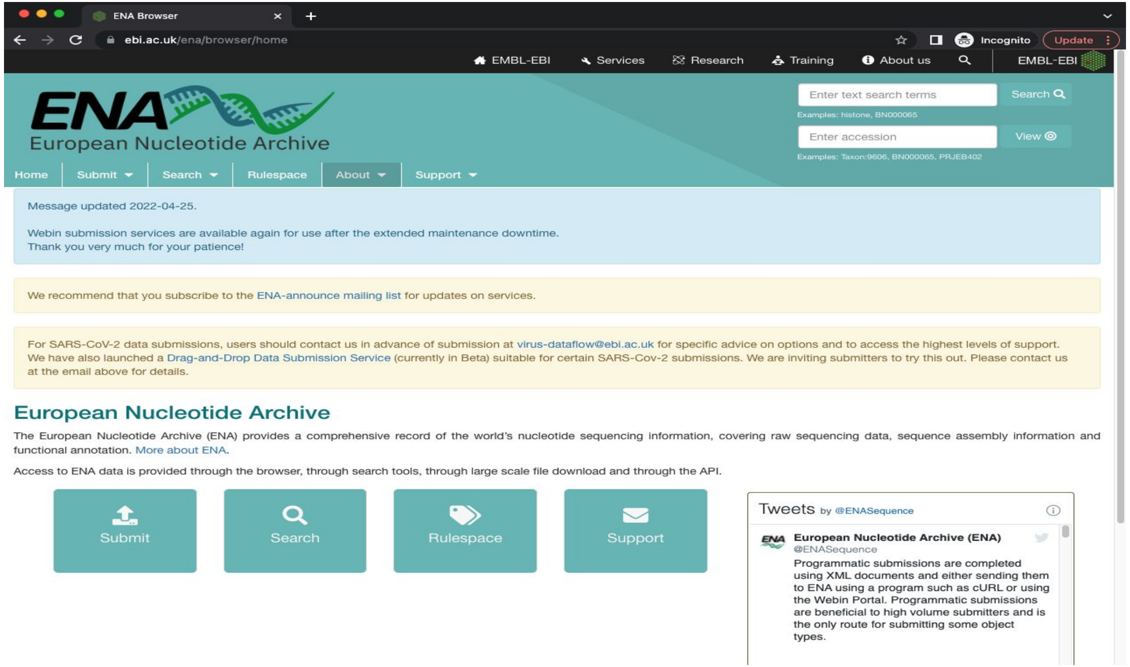
...ta

```
...e respiratory
syndrome coronavirus 2"
[Organism] OR Severe acute
respiratory syndrome
coronavirus 2[All Fields]
```

Search See more...

Recent activity
Turn Off Clear
Severe acute respiratory syndrome coronavirus 2 (91960) Nucleotide
See more...

Downloading publicly available reads



Message updated 2022-04-25.

Webin submission services are available again for use after the extended maintenance downtime.
Thank you very much for your patience!

We recommend that you subscribe to the ENA-announce mailing list for updates on services.

For SARS-CoV-2 data submissions, users should contact us in advance of submission at virus-dataflow@ebi.ac.uk for specific advice on options and to access the highest levels of support. We have also launched a Drag-and-Drop Data Submission Service (currently in Beta) suitable for certain SARS-Cov-2 submissions. We are inviting submitters to try this out. Please contact us at the email above for details.

European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. More about ENA.

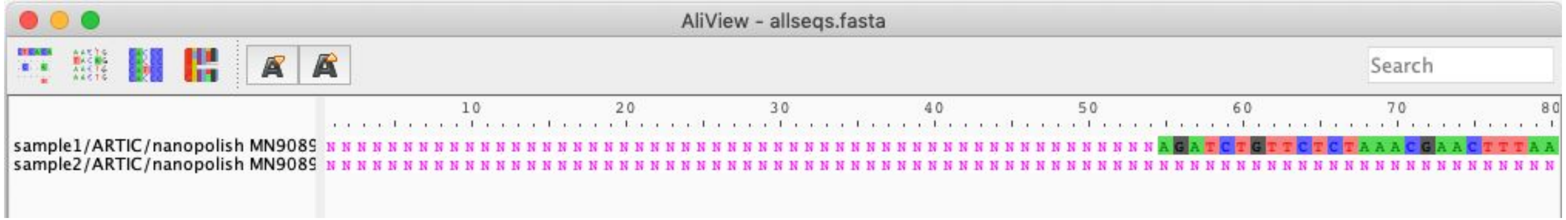
Access to ENA data is provided through the browser, through search tools, through large scale file download and through the API.

Tweets by @ENASequence

European Nucleotide Archive (ENA)
@ENASequence
Programmatic submissions are completed using XML documents and either sending them to ENA using a program such as cURL or using the Webin Portal. Programmatic submissions are beneficial to high volume submitters and is the only route for submitting some object types.

Concatenating consensus genomes

Concatenating consensus genomes

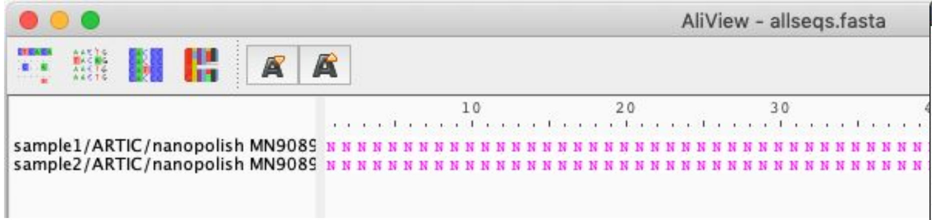


```
sample1/ARTIC/nanopolish MN9089 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGATCTGTTCTCTAAACGAACTTTAA
sample2/ARTIC/nanopolish MN9089 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

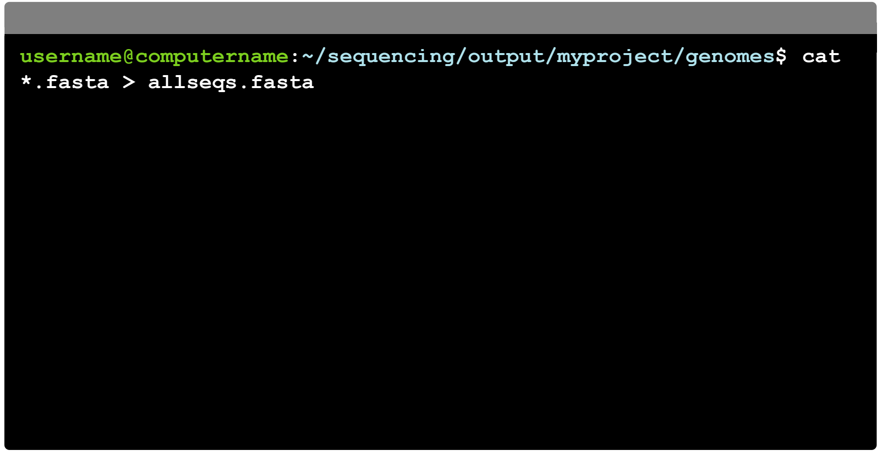
- Two methods for concatenating sequences:
  - Command line using cat
  - Copying and pasting in a text editor

- Both methods produce a multi-fasta
- FASTA file format: (.fa, .fas, .fasta); sometimes .mfa is used to indicate multi-fasta

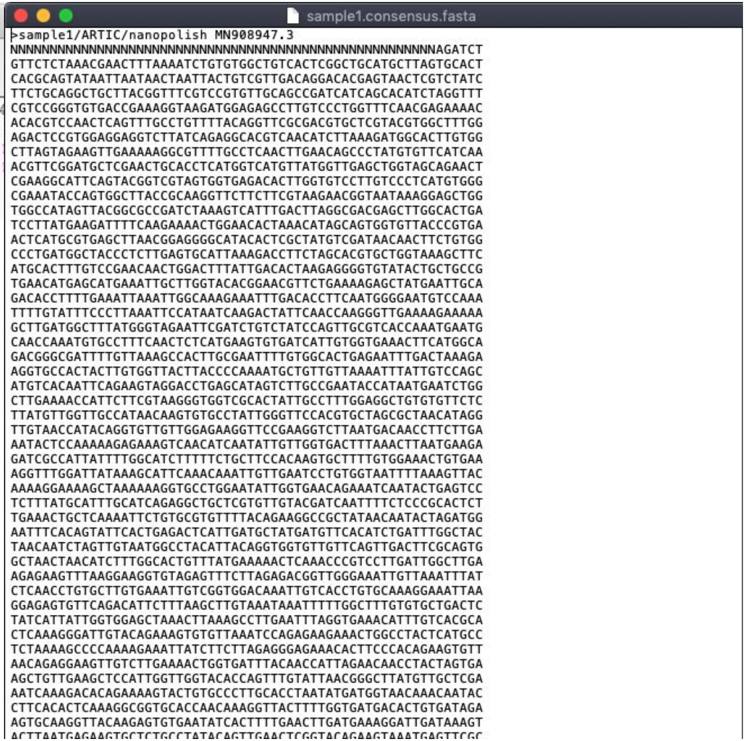
Concatenating consensus genomes



```
sample1/ARTIC/nanopolish MN9089 N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N
sample2/ARTIC/nanopolish MN9089 N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N N
```



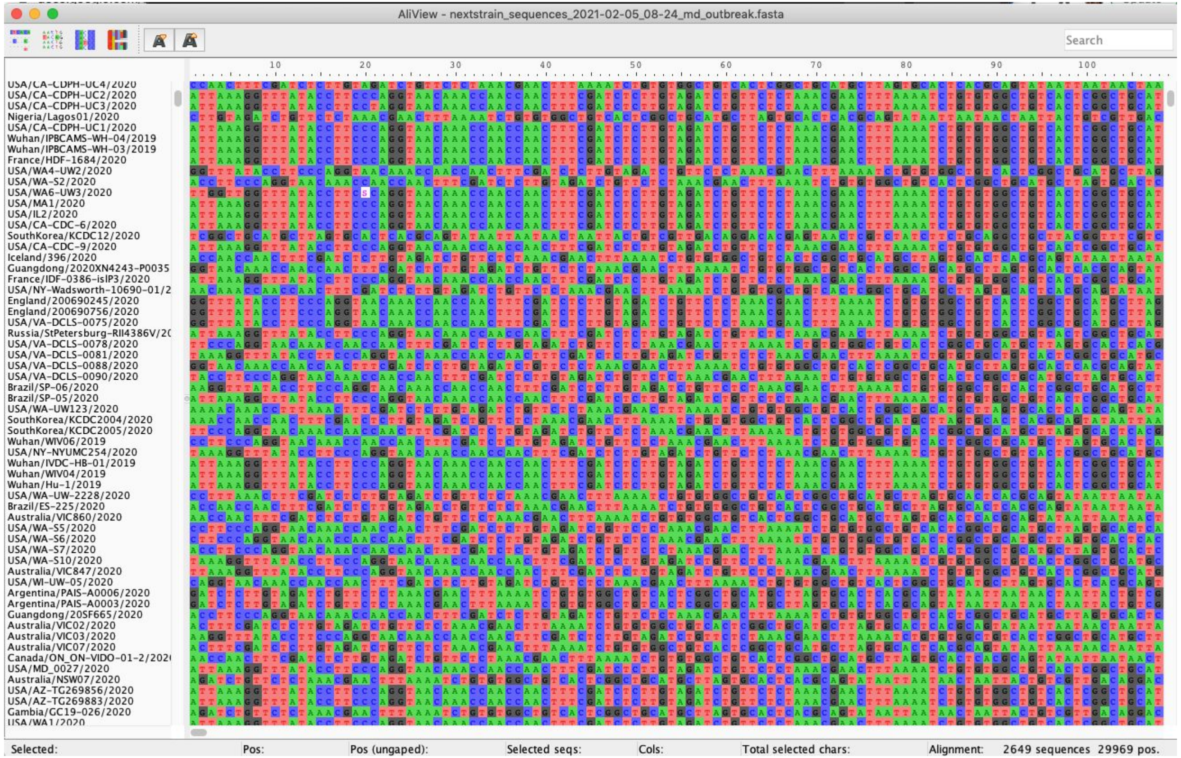
```
username@computername:~/sequencing/output/myproject/genomes$ cat
*.fasta > allseqs.fasta
```



```
>sample1/ARTIC/nanopolish MN908947.3
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGATCT
GTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACT
CACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATC
TTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTT
CGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAAC
ACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGG
AGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAA
ACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACT
CGAAGGCATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGG
CGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGG
TGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGA
TCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGA
ACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTC
ATGCACTTTGTCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCG
TGAACATGAGCATGAAATTGCTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCA
GACACCTTTTGAAATTAAATTGGCAAAGAAATTTGACACCTTCAATGGGGAATGTCCAAA
TTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAACCAAGGGTTGAAAAGAAAAA
GCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCACCAAATGAATG
CAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCA
GACGGGCGATTTTGTTAAAGCCACTTGCGAATTTTGTGGCACTGAGAATTTGACTAAAGA
AGGTGCCACTACTTGTGGTTACTTACCCCAAAATGCTGTTGTTAAAATTTATTGTCCAGC
ATGTCACAATTCAGAAGTAGGACCTGAGCATAGTCTTGCCGAATACCATAATGAATCTGG
CTTGAAAACCATTCTTCGTAAGGGTGGTCGCACTATTGCCTTTGGAGGCTGTGTGTTCTC
TTATGTTGGTTGCCATAACAAGTGTGCCTATTGGGTTCCACGTGCTAGCGCTAACATAGG
TTGTAACCATACAGGTGTTGTTGGAGAAGGTTCCGAAGGTCTTAATGACAACCTTCTTGA
AATACTCCAAAAAGAGAAAGTCAACATCAATATTGTTGGTGACTTTAAACTTAATGAAGA
GATCGCCATTATTTTGGCATCTTTTTCTGCTTCCACAAGTGCTTTTGTGGAAACTGTGAA
AGGTTTGGATTATAAAGCATTCAAACAAATTGTTGAATCCTGTGGTAATTTTAAAGTTAC
AAAAGGAAAAGCTAAAAAAGGTGCCTGGAATATTGGTGAACAGAAATCAATACTGAGTCC
TCTTTATGCATTTGCATCAGAGGCTGCTCGTGTTGTACGATCAATTTTCTCCCGCACTCT
TGAAACTGCTCAAAATTCTGTGCGTGTTTTACAGAAGGCCGCTATAACAATACTAGATGG
AATTTCACAGTATTCACTGAGACTCATTGATGCTATGATGTTCACATCTGATTTGGCTAC
TAACAATCTAGTTGTAATGGCCTACATTACAGGTGGTGTTGTTCAGTTGACTTCGCAGTG
GCTAACTAACATCTTTGGCACTGTTTATGAAAAACTCAAACCCGTCCTTGATTGGCTTGA
AGAGAAGTTTAAGGAAGGTGTAGAGTTTCTTAGAGACGGTTGGGAAATTGTTAAATTTAT
CTCAACCTGTGCTTGTGAAATTGTCGGTGGACAAATTGTCACCTGTGCAAAGGAAATTAA
GGAGAGTGTTCAGACATTCTTTAAGCTTGTAAATAAATTTTTGGCTTTGTGTGCTGACTC
TATCATTATTGGTGGAGCTAAACTTAAAGCCTTGAATTTAGGTGAAACATTTGTCACGCA
CTCAAAGGGATTGTACAGAAAGTGTGTTAAATCCAGAGAAGAAACTGGCCTACTCATGCC
TCTAAAAGCCCCAAAAGAAATTATCTTCTTAGAGGGAGAAACACTTCCCACAGAAGTGTT
AACAGAGGAAGTTGTCTTGAAAACTGGTGATTTACAACCATTAGAACAACCTACTAGTGA
AGCTGTTGAAGCTCCATTGGTTGGTACACCAGTTTGTATTAACGGGCTTATGTTGCTCGA
AATCAAAGACACAGAAAAGTACTGTGCCCTTGCACCTAATATGATGGTAACAAACAATAC
CTTCACACTCAAAGGCGGTGCACCAACAAAGGTTACTTTTGGTGATGACACTGTGATAGA
AGTGCAAGGTTACAAGAGTGTGAATATCACTTTTGAACTTGATGAAAGGATTGATAAAGT
ACTTAATGAGAAGTGCTCTGCCTATACAGTTGAACTCGGTACAGAAGTAAATGAGTTCGC
```

Sequence Alignment

Multiple sequence alignment

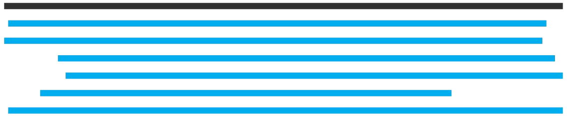


AliView - nextstrain_sequences_2021-02-05_08-24_md_outbreak.fasta

USA/CA-CDPH-UC4/2020
USA/CA-CDPH-UC2/2020
USA/CA-CDPH-UC3/2020
Nigeria/Lagos01/2020
USA/CA-CDPH-UC1/2020
Wuhan/IPBCAMS-WH-04/2019
Wuhan/IPBCAMS-WH-03/2019
France/HDF-1684/2020
USA/WA4-UW2/2020
USA/WA-S2/2020
USA/WA6-UW3/2020
USA/MA1/2020
USA/IL2/2020
USA/CA-CDC-6/2020
SouthKorea/KCDC12/2020
USA/CA-CDC-9/2020
Iceland/396/2020
Guangdong/2020XN4243-P0035
France/IDF-0386-islP3/2020
USA/NY-Wadsworth-10690-01/2
England/200690245/2020
England/200690756/2020
USA/VA-DCLS-0075/2020
Russia/StPetersburg-RII4386V/2(
USA/VA-DCLS-0078/2020
USA/VA-DCLS-0081/2020
USA/VA-DCLS-0088/2020
USA/VA-DCLS-0090/2020
Brazil/SP-06/2020
Brazil/SP-05/2020
USA/WA-UW123/2020
SouthKorea/KCDC2004/2020
SouthKorea/KCDC2005/2020
Wuhan/WIV06/2019
USA/NY-NYUMC254/2020
Wuhan/IVDC-HB-01/2019
Wuhan/WIV04/2019
Wuhan/Hu-1/2019
USA/WA-UW-2228/2020
Brazil/ES-225/2020
Australia/VIC860/2020
USA/WA-S5/2020
USA/WA-S6/2020
USA/WA-S7/2020
USA/WA-S10/2020
Australia/VIC847/2020
USA/WI-UW-05/2020
Argentina/PAIS-A0006/2020
Argentina/PAIS-A0003/2020
Guangdong/20SF665/2020
Australia/VIC02/2020
Australia/VIC03/2020
Australia/VIC07/2020
Canada/ON_ON-VIDO-01-2/202(
USA/MD_0027/2020
Australia/NSW07/2020
USA/AZ-TG269856/2020
USA/AZ-TG269883/2020
Gambia/GC19-026/2020
USA/WA1/2020

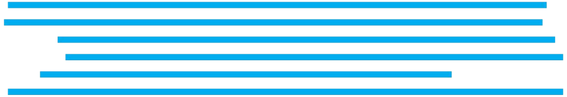
Selected: Pos: Pos (ungaped): Selected seqs: Cols: Total selected chars: Alignment: 2649 sequences 29969 pos.

Types of multiple sequence alignment



Alignment with a reference genome

- Faster and less computationally intensive
- Ensures consistent coordinates



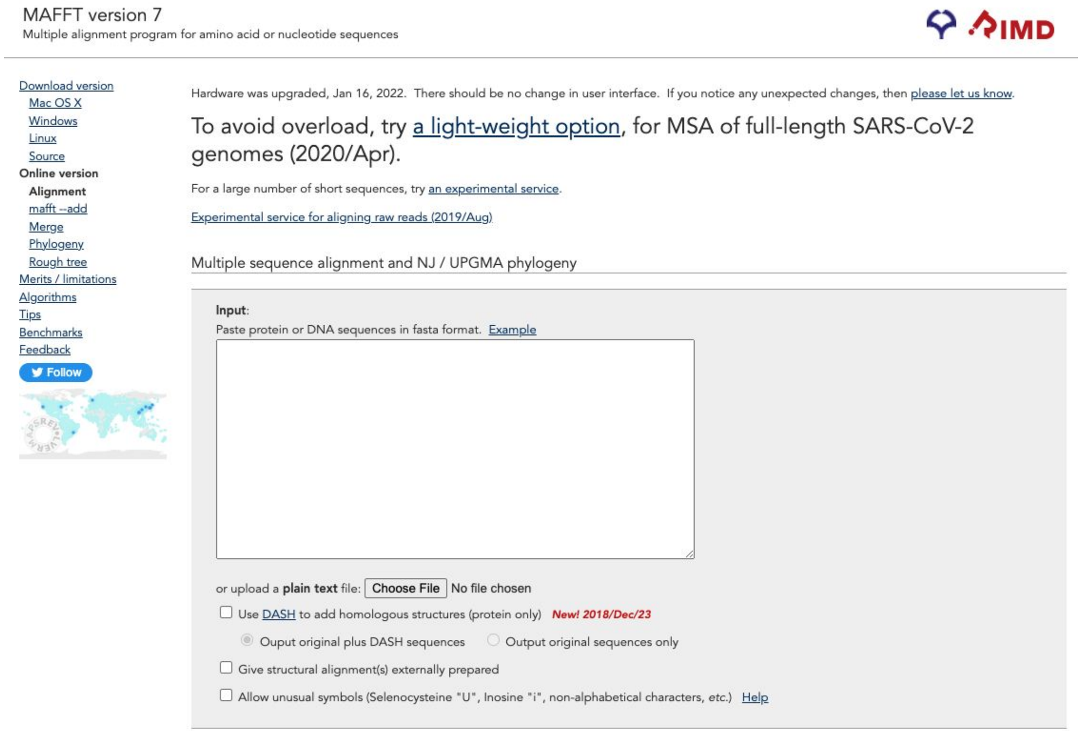
Alignment without a reference genome

Selecting an appropriate reference genome

- Use the same reference genome as others working on the same outbreak
  - e.g., Wuhan-Hu-1 for SARS-CoV-2
- Use the earliest sequence from the outbreak, if available
- Use a sequence from a prior outbreak in the same location
- Use NCBI RefSeq (https://www.ncbi.nlm.nih.gov/refseq/)

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

MAFFT Server



MAFFT version 7

Multiple alignment program for amino acid or nucleotide sequences

Hardware was upgraded, Jan 16, 2022. There should be no change in user interface. If you notice any unexpected changes, then please let us know.

To avoid overload, try a light-weight option, for MSA of full-length SARS-CoV-2 genomes (2020/Apr).

For a large number of short sequences, try an experimental service.

Experimental service for aligning raw reads (2019/Aug)

Multiple sequence alignment and NJ / UPGMA phylogeny

Input:

Paste protein or DNA sequences in fasta format. Example

or upload a **plain text** file: Choose File No file chosen

□ Use DASH to add homologous structures (protein only) *New! 2018/Dec/23*

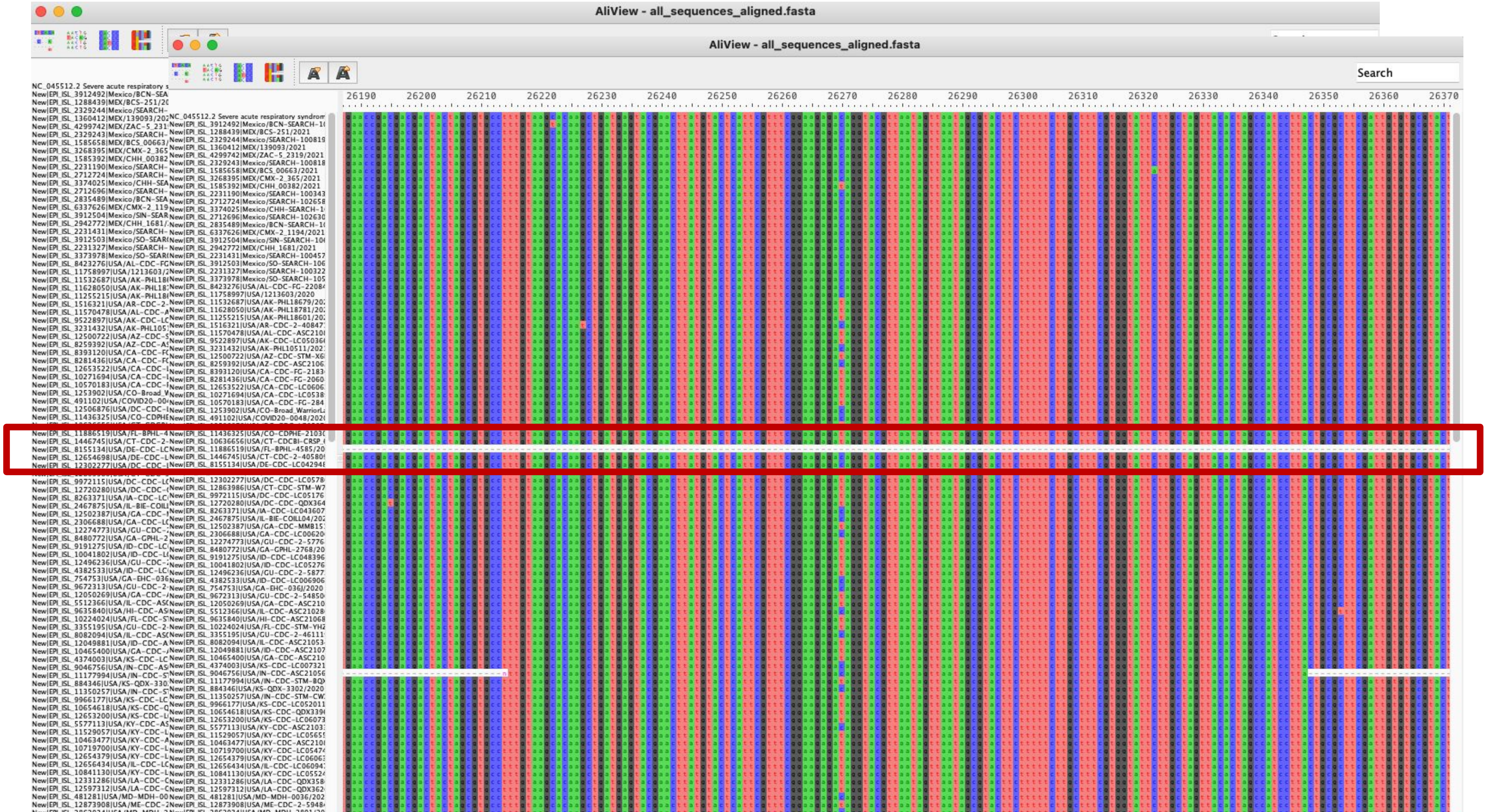
● Ouput original plus DASH sequences ○ Output original sequences only

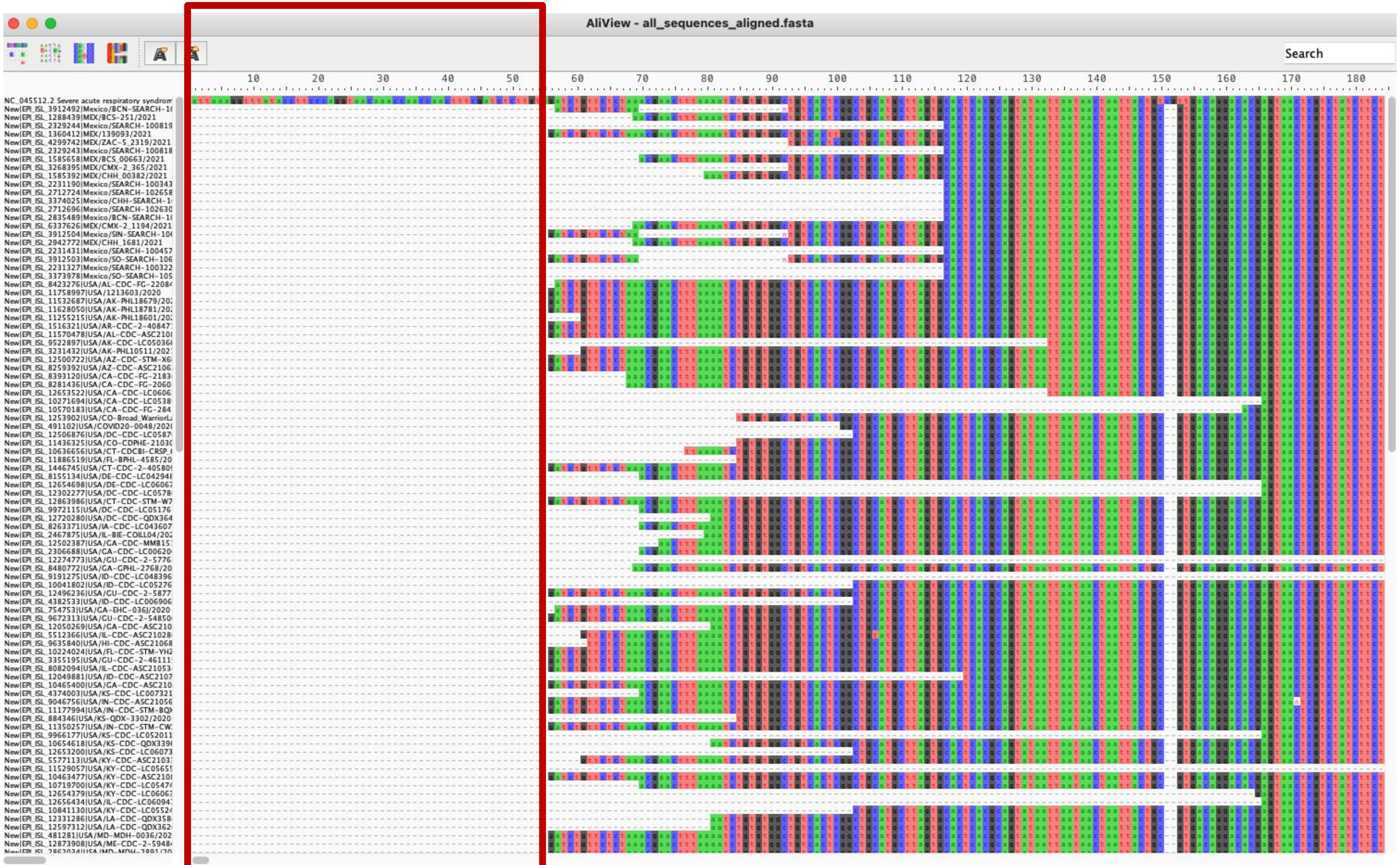
□ Give structural alignment(s) externally prepared

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

MAFFT Commandline

```
mafft --auto consensus.fasta > consensus_aln.fasta
```


AliView - all_sequences_aligned.fasta
AliView - all_sequences_aligned.fasta
Search
26190
26200
26210
26220
26230
26240
26250
26260
26270
26280
26290
26300
26310
26320
26330
26340
26350
26360
26370
NC_045512.2 Severe acute respiratory syndrom



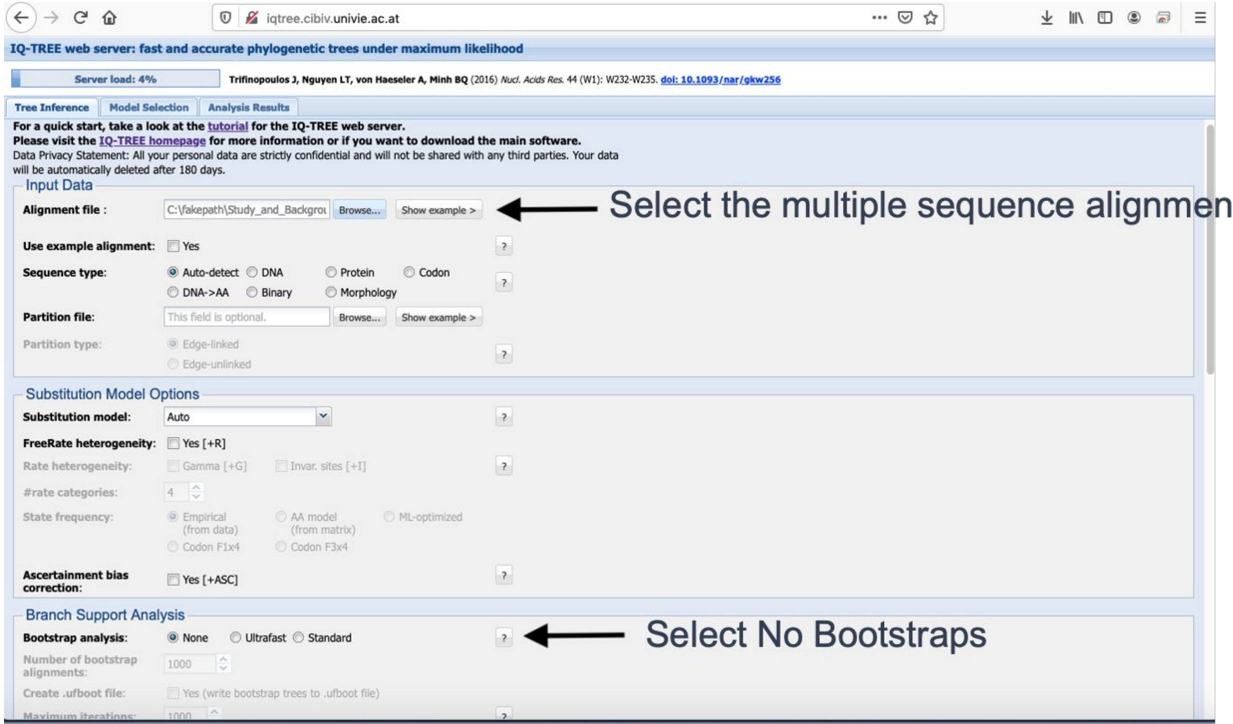
AliView - all_sequences_aligned.fasta

```
NC_045512.2 Severe acute respiratory syndrom
New|EPI_ISL_3912492|Mexico/BCN-SEARCH-1(
New|EPI_ISL_1288439|MEX/BCS-251/2021
New|EPI_ISL_2329244|Mexico/SEARCH-100819
New|EPI_ISL_1360412|MEX/139093/2021
New|EPI_ISL_4299742|MEX/ZAC-5_2319/2021
New|EPI_ISL_2329243|Mexico/SEARCH-100818
New|EPI_ISL_1585658|MEX/BCS_00663/2021
New|EPI_ISL_3268395|MEX/CMX-2_365/2021
New|EPI_ISL_1585392|MEX/CHH_00382/2021
New|EPI_ISL_2231190|Mexico/SEARCH-100343
New|EPI_ISL_2712724|Mexico/SEARCH-102658
New|EPI_ISL_3374025|Mexico/CHH-SEARCH-1
New|EPI_ISL_2712696|Mexico/SEARCH-102630
New|EPI_ISL_2835489|Mexico/BCN-SEARCH-1(
New|EPI_ISL_6337626|MEX/CMX-2_1194/2021
New|EPI_ISL_3912504|Mexico/SIN-SEARCH-10
New|EPI_ISL_2942772|MEX/CHH_1681/2021
New|EPI_ISL_2231431|Mexico/SEARCH-100457
New|EPI_ISL_3912503|Mexico/SO-SEARCH-106
New|EPI_ISL_2231327|Mexico/SEARCH-100322
New|EPI_ISL_3373978|Mexico/SO-SEARCH-105
New|EPI_ISL_8423276|USA/AL-CDC-FG-22084
New|EPI_ISL_11758997|USA/1213603/2020
New|EPI_ISL_11532687|USA/AK-PHL18679/202
New|EPI_ISL_11628050|USA/AK-PHL18781/202
New|EPI_ISL_11255215|USA/AK-PHL18601/202
New|EPI_ISL_1516321|USA/AR-CDC-2-408471
New|EPI_ISL_11570478|USA/AL-CDC-ASC210
New|EPI_ISL_9522897|USA/AK-CDC-LC05036
New|EPI_ISL_3231432|USA/AK-PHL10511/2021
New|EPI_ISL_12500722|USA/AZ-CDC-STM-X6
New|EPI_ISL_8259392|USA/AZ-CDC-ASC2106
New|EPI_ISL_8393120|USA/CA-CDC-FG-2183
New|EPI_ISL_8281436|USA/CA-CDC-FG-2060
New|EPI_ISL_12653522|USA/CA-CDC-LC0606
New|EPI_ISL_10271694|USA/CA-CDC-LC0538
New|EPI_ISL_10570183|USA/CA-CDC-FG-284
New|EPI_ISL_1253902|USA/CO-Broad_WarriorL
New|EPI_ISL_491102|USA/COVID20-0048/202(
New|EPI_ISL_12506876|USA/DC-CDC-LC0587
New|EPI_ISL_11436325|USA/CO-CDPHE-2103(
New|EPI_ISL_10636656|USA/CT-CDCBI-CRSP_(
New|EPI_ISL_11886519|USA/FL-BPHL-4585/20
New|EPI_ISL_1446745|USA/CT-CDC-2-40580
New|EPI_ISL_8155134|USA/DE-CDC-LC04294
New|EPI_ISL_12654698|USA/DE-CDC-LC0606
New|EPI_ISL_12302277|USA/DC-CDC-LC0578
New|EPI_ISL_12863986|USA/CT-CDC-STM-W7
New|EPI_ISL_9972115|USA/DC-CDC-LC05176
New|EPI_ISL_12720280|USA/DC-CDC-QDX364
New|EPI_ISL_8263371|USA/IA-CDC-LC043607
New|EPI_ISL_2467875|USA/IL-BIE-COILL04/202
New|EPI_ISL_12502387|USA/GA-CDC-MMB15
New|EPI_ISL_2306688|USA/GA-CDC-LC00620
New|EPI_ISL_12274773|USA/GU-CDC-2-5776
New|EPI_ISL_8480772|USA/GA-GPHL-2768/20
New|EPI_ISL_9191275|USA/ID-CDC-LC048396
New|EPI_ISL_10041802|USA/ID-CDC-LC05276
New|EPI_ISL_12496236|USA/GU-CDC-2-5877
New|EPI_ISL_4382533|USA/ID-CDC-LC006906
New|EPI_ISL_754753|USA/GA-EHC-036J/2020
New|EPI_ISL_9672313|USA/GU-CDC-2-54850
New|EPI_ISL_12050269|USA/GA-CDC-ASC210
New|EPI_ISL_5512366|USA/IL-CDC-ASC21028
New|EPI_ISL_9635840|USA/HI-CDC-ASC21068
New|EPI_ISL_10224024|USA/FL-CDC-STM-YH2
New|EPI_ISL_3355195|USA/GU-CDC-2-46111
New|EPI_ISL_8082094|USA/IL-CDC-ASC21053
New|EPI_ISL_12049881|USA/ID-CDC-ASC2107
New|EPI_ISL_10465400|USA/GA-CDC-ASC210
New|EPI_ISL_4374003|USA/KS-CDC-LC007321
New|EPI_ISL_9046756|USA/IN-CDC-ASC21056
New|EPI_ISL_11177994|USA/IN-CDC-STM-BQ
New|EPI_ISL_884346|USA/KS-QDX-3302/2020
New|EPI_ISL_11350257|USA/IN-CDC-STM-CW
New|EPI_ISL_9966177|USA/KS-CDC-LC052011
New|EPI_ISL_10654618|USA/KS-CDC-QDX339
New|EPI_ISL_12653200|USA/KS-CDC-LC06073
New|EPI_ISL_5577113|USA/KY-CDC-ASC2103
New|EPI_ISL_11529057|USA/KY-CDC-LC0565
New|EPI_ISL_10463477|USA/KY-CDC-ASC210
New|EPI_ISL_10719700|USA/KY-CDC-LC0547
New|EPI_ISL_12654379|USA/KY-CDC-LC0606
New|EPI_ISL_12656434|USA/IL-CDC-LC06094
New|EPI_ISL_10841130|USA/KY-CDC-LC0552
New|EPI_ISL_12331286|USA/LA-CDC-QDX358
New|EPI_ISL_12597312|USA/LA-CDC-QDX362
New|EPI_ISL_481281|USA/MD-MDH-0036/202
New|EPI_ISL_12873908|USA/ME-CDC-2-5948
New|EPI_ISL_2862024|USA/MD-MDH-2891/20
```

Phylogenetic Tree Inference

IQ-Tree Server

http://iqtree.cibiv.univie.ac.at/



IQ-TREE web server: fast and accurate phylogenetic trees under maximum likelihood

Server load: 4%

Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ (2016) *Nucl. Acids Res.* 44 (W1): W232-W235. doi: 10.1093/nar/gkw256

For a quick start, take a look at the tutorial for the IQ-TREE web server.
Please visit the IQ-TREE homepage for more information or if you want to download the main software.
Data Privacy Statement: All your personal data are strictly confidential and will not be shared with any third parties. Your data will be automatically deleted after 180 days.

Input Data

Alignment file : C:\fakepath\Study_and_Backgrou Browse... Show example >

Select the multiple sequence alignment file

Use example alignment: □ Yes

Sequence type: ● Auto-detect ○ DNA ○ Protein ○ Codon ○ DNA->AA ○ Binary ○ Morphology

Partition file: This field is optional. Browse... Show example >

Partition type: ● Edge-linked ○ Edge-unlinked

Substitution Model Options

Substitution model: Auto

FreeRate heterogeneity: □ Yes [+R]

Rate heterogeneity: □ Gamma [+G] □ Invar. sites [+I]

#rate categories: 4

State frequency: ● Empirical (from data) ○ AA model (from matrix) ○ ML-optimized ○ Codon F1x4 ○ Codon F3x4

Ascertainment bias correction: □ Yes [+ASC]

Branch Support Analysis

Bootstrap analysis: ● None ○ Ultrafast ○ Standard

Select No Bootstraps

Number of bootstrap alignments: 1000

Create .ufboot file: □ Yes (write bootstrap trees to .ufboot file)

Maximum iterations: 1000

IQ-Tree Commandline

```
iqtree2 -s consensus_aln.fasta -T AUTO -m TEST -B 1000
```

What questions do you have?