# COVID-19 Genomic Data Analysis

2023.07.17

Chih-Ting Yang

Dr. Hsin-Chou Yang's Lab

# Outline

- Introduction
- Data download
- Sequence alignment
- Data Analysis
- Supplementary
- Further Analysis
- Discussion

# Introduction

# Motivation

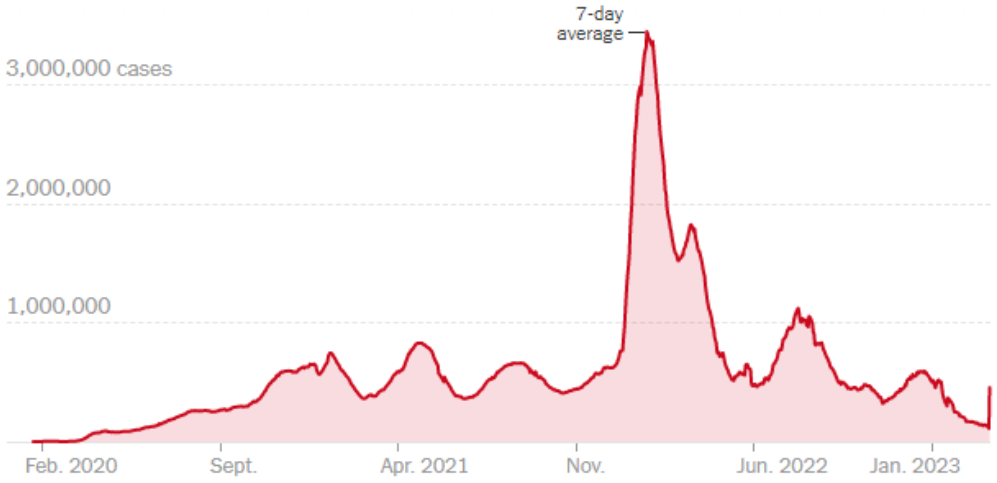
## Coronavirus World Map: Tracking the Global Outbreak

Updated March 10, 2023

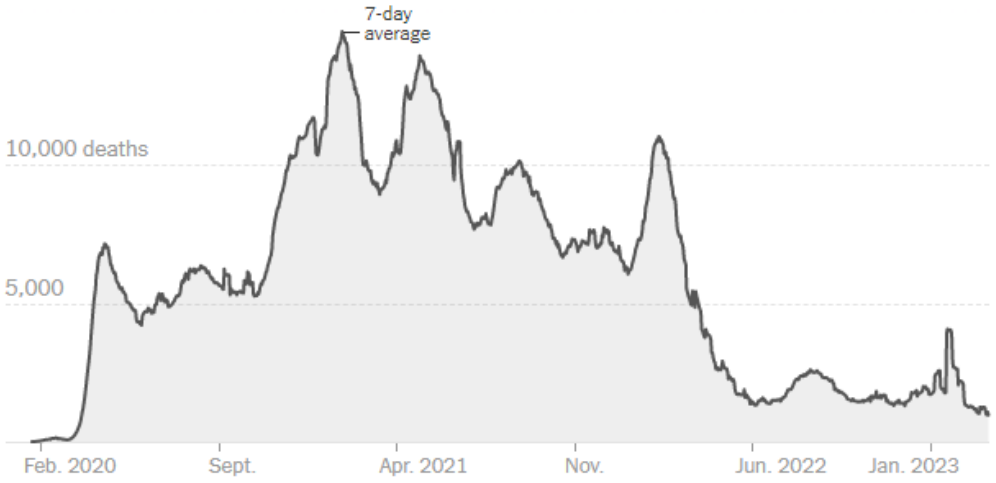
This page was archived on March 10 as global data on cases and deaths is no longer reported by our data source for all countries except the United States.

All time | Last 90 days

### New reported cases by day

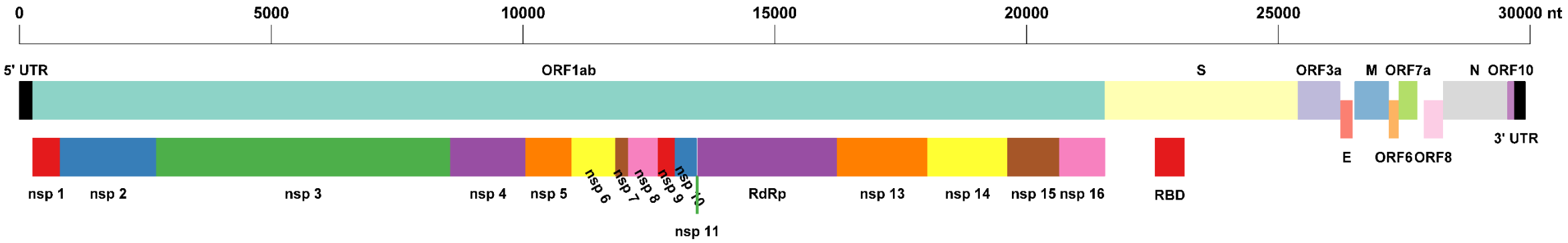


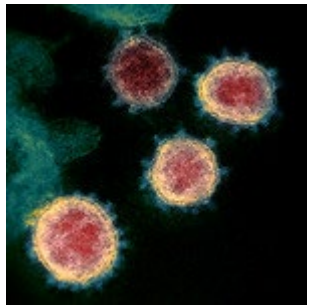

### New reported deaths by day




https://www.nytimes.com/interactive/2021/world/covid-cases.html

# Reference genome (GISAID: EPI_ISL_402124, GenBank: MN908947)



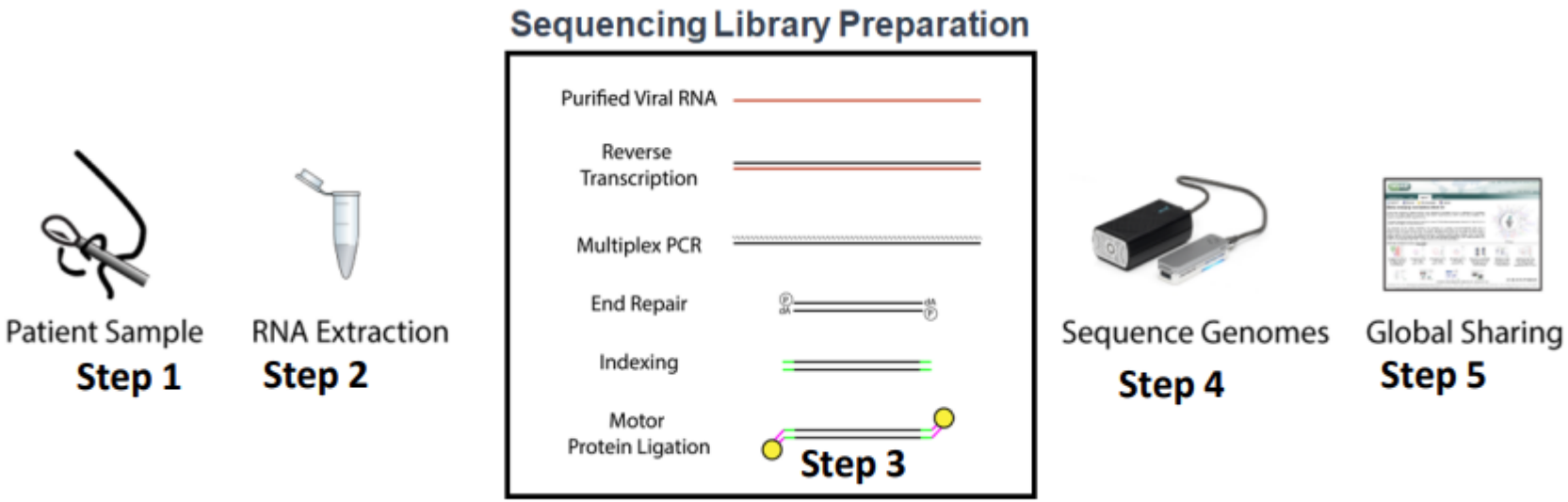



This transmission electron microscope image shows SARS-CoV-2—also known as 2019-nCoV, the virus that causes COVID-19—isolated from a patient in the U.S. Virus particles are shown emerging from the surface of cells cultured in the lab. The spikes on the outer edge of the virus particles give coronaviruses their name, crown-like. Image captured and colorized at NIAID's Rocky Mountain Laboratories (RML) in Hamilton, Montana. Credit: NIAID

https://www.flickr.com/photos/niaid/albums/72157712914621487

https://www.ncbi.nlm.nih.gov/nuccore/MN908947

# RNA whole-genome sequencing (WGS)



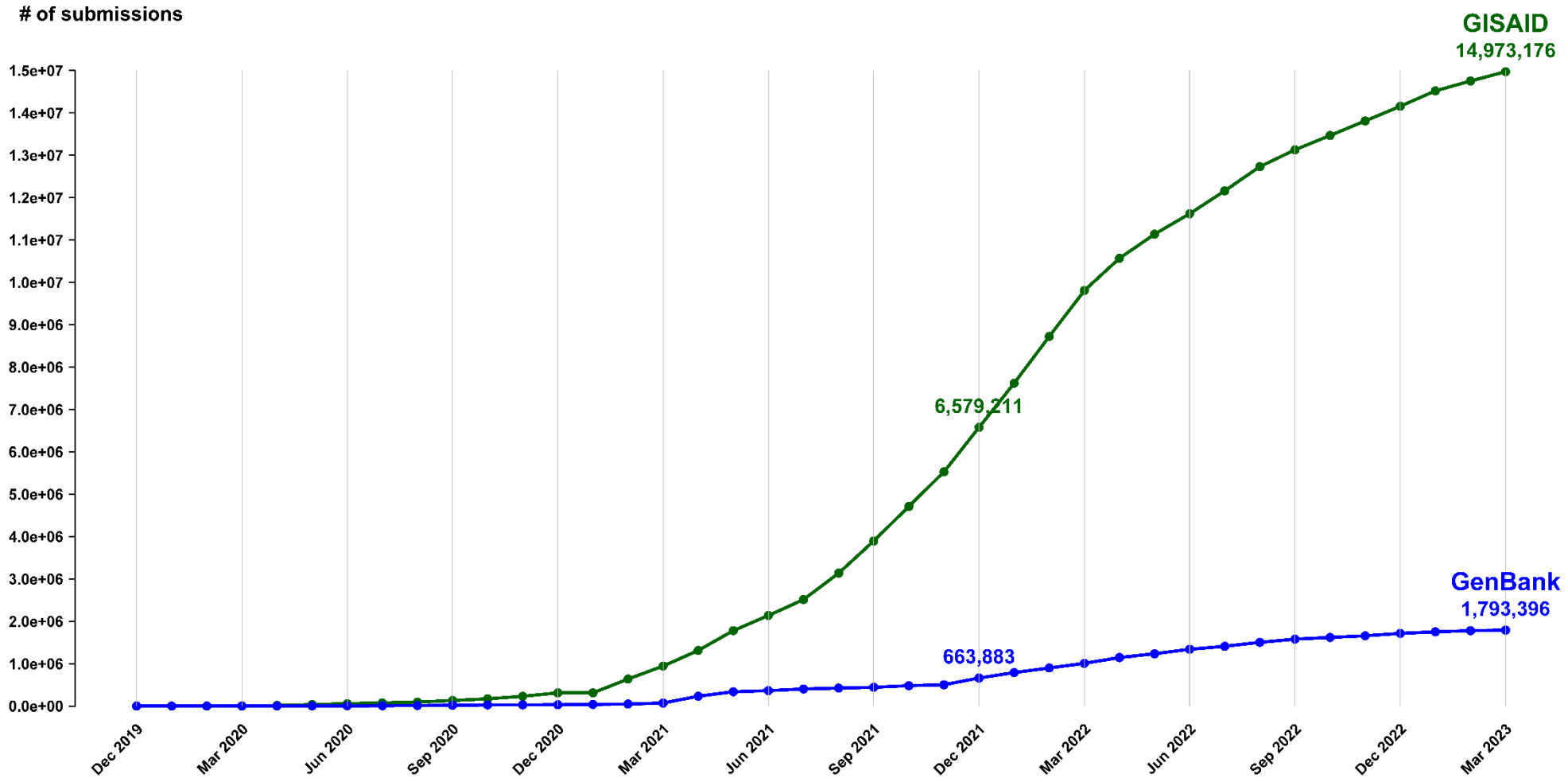

Public databases for viral genomic data

1. **GISAID** (Global Initiative on Shared All Influenza Data)
   https://gisaid.org/

2. **NCBI** (National Center for Biotechnology Information)
   https://www.ncbi.nlm.nih.gov/

https://www.cdc.gov/amd/training/covid-toolkit/module1-1.html

# SARS-CoV-2 submissions (complete human genome)




Data Last update: 2023/04/06

# Download Data from GenBank

https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/

National Library of Medicine
National Center for Biotechnology Information

Log in

NCBI Virus
Sequences for discovery

About Us | Find Data | Help | How to Participate | Submit Sequences | Contact Us

**Quick Access to SARS-CoV-2 Data!**

- Novel Severe acute respiratory syndrome coronavirus 2 RefSeq genomes, nucleotide, and protein sequences.
- View our new SARS-CoV-2 interactive dashboard.
- How to submit SARS-CoV-2 sequences.
- Visit our new SARS-CoV-2 Variants Overview New!.

NCBI Virus is a community portal for viral sequence data from RefSeq, GenBank and other NCBI repositories. To find, retrieve and analyze data, please select an option below.

**Search by sequence**
Use the NCBI BLAST™ tool to find similar viral nucleotide and protein sequences.

**Search by virus**
Use virus name or taxid to find viral nucleotide and protein sequences.

National Library of Medicine
National Center for Biotechnology Information

Log in

NCBI Virus
Sequences for discovery

About Us | Find Data | Help | How to Participate | Submit Sequences | Contact Us

**Search by sequence**
Use the NCBI BLAST™ tool to find similar viral nucleotide and protein sequences.

**Search by virus**
Use virus name or taxid to find viral nucleotide and protein sequences.

**Search by virus name or taxonomy**

Begin typing a virus name, viral taxonomy group, or taxid to select from a list of suggestions.

sars-cov-2

Severe acute respiratory syndrome coronavirus 2, taxid:2697049
Severe acute respiratory syndrome coronavirus 2 (**SARS-CoV-2**), taxid:2697049

All viruses | Human viruses | Bacteriophages | New sequences (past one month) | **Up-to-date SARS-CoV-2**

# Download data -1

Refine Results Reset

Virus +

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), taxid:2697049 ×

Accession +

Sequence Length +

Ambiguous Characters +

Sequence Type +

GenBank ×

RefSeq Genome Completeness +

Nucleotide Completeness +

complete ×

Pango lineage +

Random Sampling +

Isolate +

Proteins +

Provirus +

Geographic Region +

Host +

Homo sapiens (human), taxid:9606 ×

Submitters +

Isolation Source +

Collection Date +

Release Date +

Genome Molecule Type +

Environmental Source +

Lab Host +

Vaccine Strain +

Nucleotide (1,790,227) | Protein (16,597,251) | RefSeq Genome (0)

Select Columns

| Accession | Submitters | Organization | Release Date | Species | Length | Nuc Completeness | Geo Location | Host | Collection Date |
|---|---|---|---|---|---|---|---|---|---|
| OQ618228 | Schilling,W. | Mahidol-Oxford Tropical M... | 2023-03-14 | Severe acute respiratory sy... | 29701 | complete | Thailand: Bangkok | Homo sapiens | 2021-10-14 |
| OQ618229 | Schilling,W. | Mahidol-Oxford Tropical M... | 2023-03-14 | Severe acute respiratory sy... | 29701 | complete | Thailand: Bangkok | Homo sapiens | 2021-10-14 |
| OQ618230 | Schilling,W. | Mahidol-Oxford Tropical M... | 2023-03-14 | Severe acute respiratory sy... | 29684 | complete | Thailand: Bangkok | Homo sapiens | 2022-01-06 |
| OQ618231 | Schilling,W. | Mahidol-Oxford Tropical M... | 2023-03-14 | Severe acute respiratory sy... | 29684 | complete | Thailand: Samut Prakan | Homo sapiens | 2022-01-05 |
| OQ618293 | Schilling,W. | Mahidol-Oxford Tropical M... | 2023-03-14 | Severe acute respiratory sy... | 29684 | complete | Thailand: Samut Prakan | Homo sapiens | 2022-01-05 |
| OQ618294 | Schilling,W. | Mahidol-Oxford Tropical M... | 2023-03-14 | Severe acute respiratory sy... | 29684 | complete | Thailand: Samut Prakan | Homo sapiens | 2022-01-06 |
| OQ618295 | Schilling,W. | Mahidol-Oxford Tropical M... | 2023-03-14 | Severe acute respiratory sy... | 29684 | complete | Thailand: Samut Prakan | Homo sapiens | 2022-01-06 |
| OQ618296 | Schilling,W. | Mahidol-Oxford Tropical M... | 2023-03-14 | Severe acute respiratory sy... | 29684 | complete | Thailand: Samut Prakan | Homo sapiens | 2022-02-10 |
| OQ618297 | Schilling,W. | Mahidol-Oxford Tropical M... | 2023-03-14 | Severe acute respiratory sy... | 29684 | complete | Thailand: Samut Prakan | Homo sapiens | 2022-02-25 |
| OQ618298 | Schilling,W. | Mahidol-Oxford Tropical M... | 2023-03-14 | Severe acute respiratory sy... | 29684 | complete | Thailand: Samut Prakan | Homo sapiens | 2022-03-04 |
| OQ618300 | Schilling,W. | Mahidol-Oxford Tropical M... | 2023-03-14 | Severe acute respiratory sy... | 29701 | complete | Thailand: Samut Prakan | Homo sapiens | 2021-12-04 |
| OQ618301 | Schilling,W. | Mahidol-Oxford Tropical M... | 2023-03-14 | Severe acute respiratory sy... | 29684 | complete | Thailand: Samut Prakan | Homo sapiens | 2022-01-06 |
| OQ618302 | Schilling,W. | Mahidol-Oxford Tropical M... | 2023-03-14 | Severe acute respiratory sy... | 29684 | complete | Thailand: Samut Prakan | Homo sapiens | 2022-01-06 |
| OQ618303 | Schilling,W. | Mahidol-Oxford Tropical M... | 2023-03-14 | Severe acute respiratory sy... | 29684 | complete | Thailand: Samut Prakan | Homo sapiens | 2022-01-08 |
| OQ618304 | Schilling,W. | Mahidol-Oxford Tropical M... | 2023-03-14 | Severe acute respiratory sy... | 29684 | complete | Thailand: Samut Prakan | Homo sapiens | 2022-01-12 |

# Download data -2

This is an NCBI Labs Experiment. Learn more.

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

NCBI Virus
Sequences for discovery

About Us | Find Data | Help | How to Participate | Submit Sequences | Contact Us

SARS-CoV-2 Data Hub | Download

Quick Links: Betacoronavirus BLAST | CDC Outbreak Information | SARS-CoV-2 Articles in PubMed | SRA Data | NCBI SARS-CoV-2 Resources | Datasets command line

Tabular View | Dashboard Visualizations | Mutations in SRA | Variants Overview New!

Selected Results: 200 | Align | Build Phylogenetic Tree

New! Submitters' Information available

Refine Results | Reset

Virus
Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), taxid:2697049

Accession

Sequence Length

Ambiguous Characters

Nucleotide (1,790,227) | Protein (16,597,251) | RefSeq Genome (0)

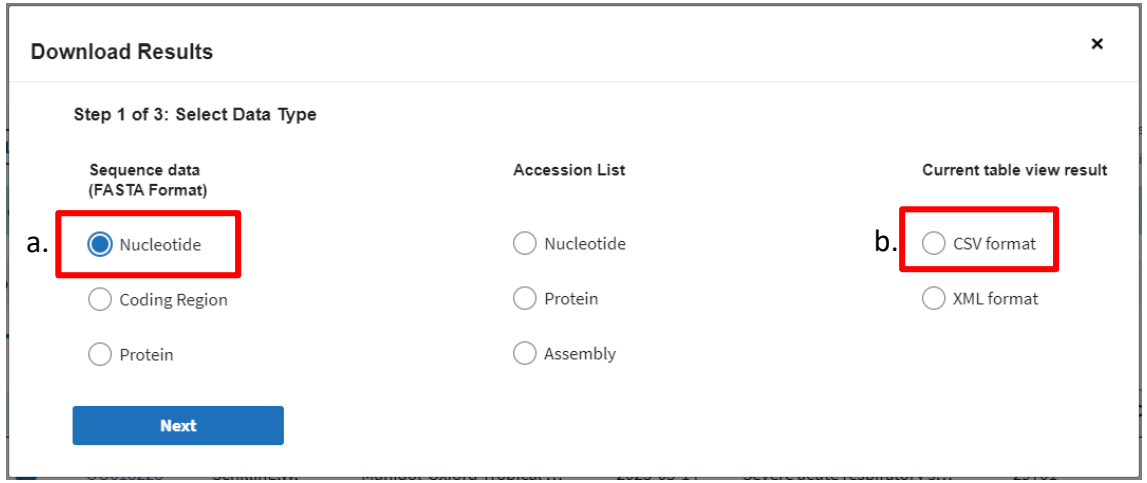
Select Columns

Expand Table

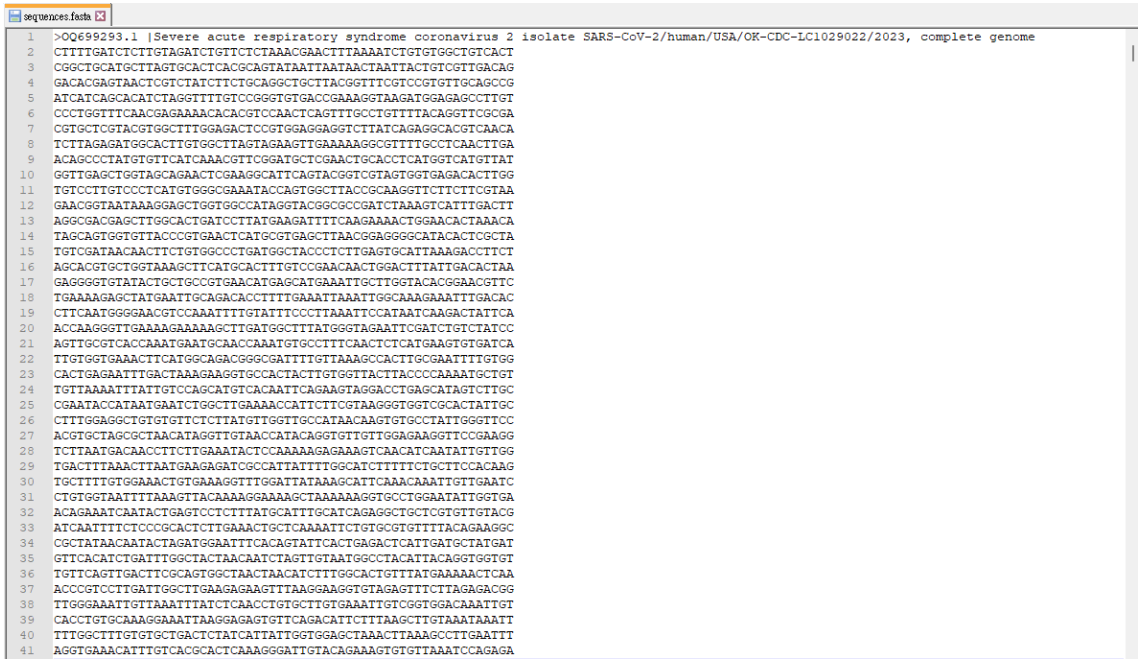
| Accession | Submitters | Organization | Release Date | Species | Length | Nuc Completeness | Geo Location |
|---|---|---|---|---|---|---|---|
| OQ618228 | Schilling,W. | Mahidol-Oxford Tropical ... | 2023-03-14 | Severe acute respiratory s... | 29701 | complete | Thailand: Bangkok |
| OQ618229 | Schilling,W. | Mahidol-Oxford Tropical ... | 2023-03-14 | Severe acute respiratory s... | 29701 | complete | Thailand: Bangkok |
| OQ618230 | Schilling,W. | Mahidol-Oxford Tropical ... | 2023-03-14 | Severe acute respiratory s... | 29684 | complete | Thailand: Bangkok |
| OQ618231 | Schilling,W. | Mahidol-Oxford Tropical ... | 2023-03-14 | Severe acute respiratory s... | 29684 | complete | Thailand: Samut Prakar |
| OQ618293 | Schilling,W. | Mahidol-Oxford Tropical ... | 2023-03-14 | Severe acute respiratory s... | 29684 | complete | Thailand: Samut Prakar |
| OQ618294 | Schilling,W. | Mahidol-Oxford Tropical ... | 2023-03-14 | Severe acute respiratory s... | 29684 | complete | Thailand: Samut Prakar |

Feedback

# Download data -3




(a) Nucleotide Sequences (FASTA)




(b) Metadata (CSV)

| Accession | Organism_Name | SRA_Accession | Submitters | Organization | Org_location | Release_Date | Isolate | Molecule_type | Length | Geo_Location | Country | Host | Isolation_Source | Collection_Date | BioSample | GenBank_Title |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| OQ699293 | Severe acute respiratory syndrome coronavirus 2 | SRR23985578 | . | Centers for Disease Control and Prevention, Respiratory Viruses Branch, Division of Viral Diseases | USA | 2023-03-28T00:00:00Z | OK-CDC-LC1029022 | ssRNA(+) | 29724 | USA: Oklahoma | USA | Homo sapiens | oronasopharynx | 2023/3/6 | SAMN33943417 | Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/OK-CDC-LC1029022/2023, complete genome |
| OQ699298 | Severe acute respiratory syndrome coronavirus 2 | SRR23985469 | . | Centers for Disease Control and Prevention, Respiratory Viruses Branch, Division of Viral Diseases | USA | 2023-03-28T00:00:00Z | NM-CDC-LC1029047 | ssRNA(+) | 29684 | USA: New Mexico | USA | Homo sapiens | oronasopharynx | 2023/3/6 | SAMN33943435 | Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/NM-CDC-LC1029047/2023, complete genome |

# Sequence Alignment

# Sequence Alignment

Goal: the process of comparing and matching sequences of nucleotides (DNA or RNA) from a sample to a reference genome or another set of sequences

Tool: Nextclade ([https://clades.nextstrain.org/](https://clades.nextstrain.org/))

**Raw sequence**

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Seq1** | A | A | T | N | N | G | C | C | A | G | T | A | G | G | A |
| **Seq2** | A | T | T | C | A | G | C | T | A | G | G | A | | | |
| **Seq3** | A | A | T | C | A | G | C | T | T | G | G | A | | | |
| **Seq4** | A | T | T | C | N | N | C | T | T | | | | | | |
| **Ref.** | A | A | T | C | A | G | C | T | A | G | G | A | | | |

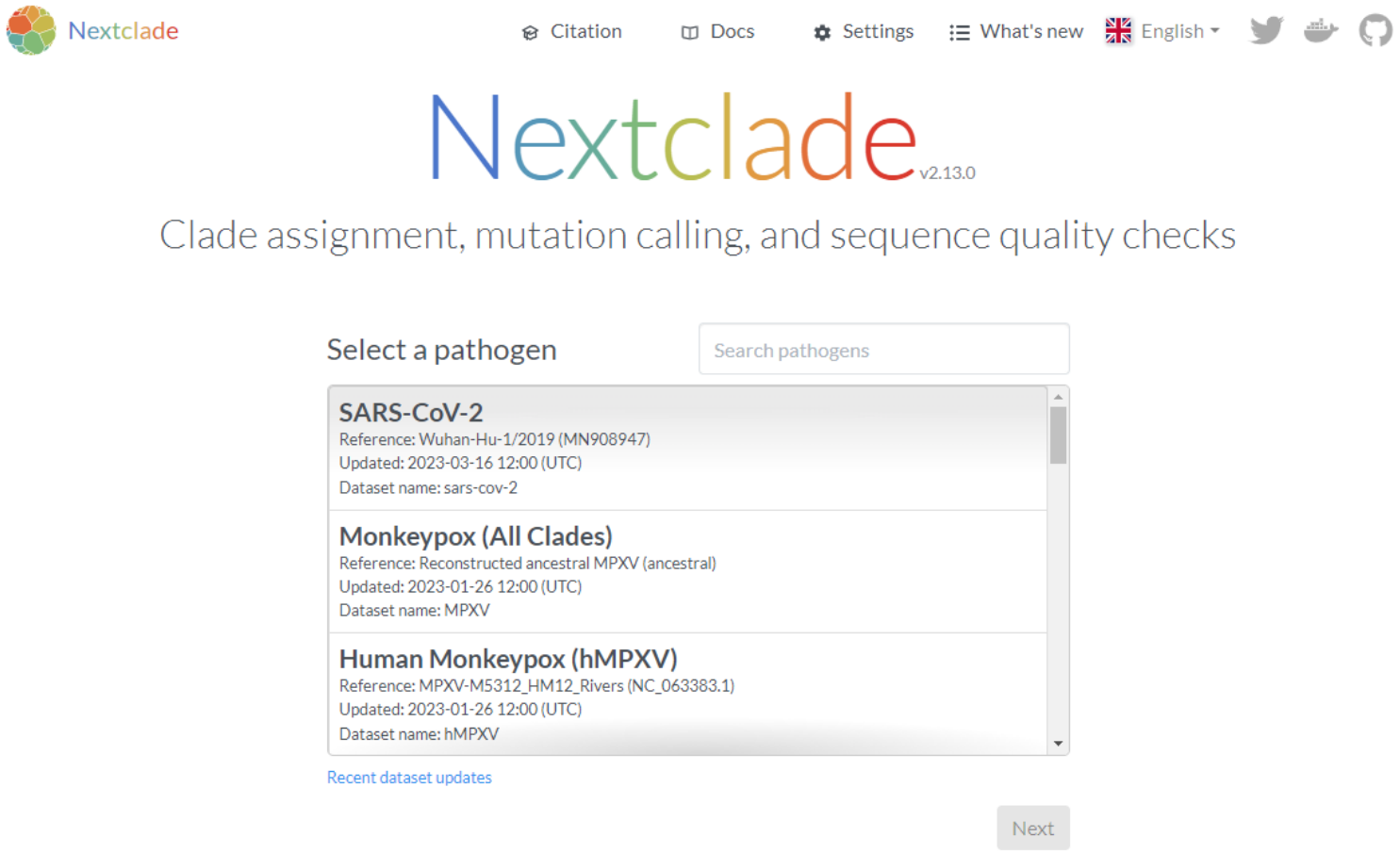
**MSA without Reference**

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Seq1** | A | A | T | N | N | G | C | C | A | G | T | A | G | G | A |
| **Seq2** | A | T | T | C | A | G | C | T | A | - | - | - | G | G | A |
| **Seq3** | A | A | T | C | A | G | C | T | T | - | - | | G | G | A |
| **Seq4** | A | T | T | C | N | N | C | T | T | - | - | - | - | - | - |
| **Ref.** | A | A | T | C | A | G | C | T | A | - | - | - | G | G | A |

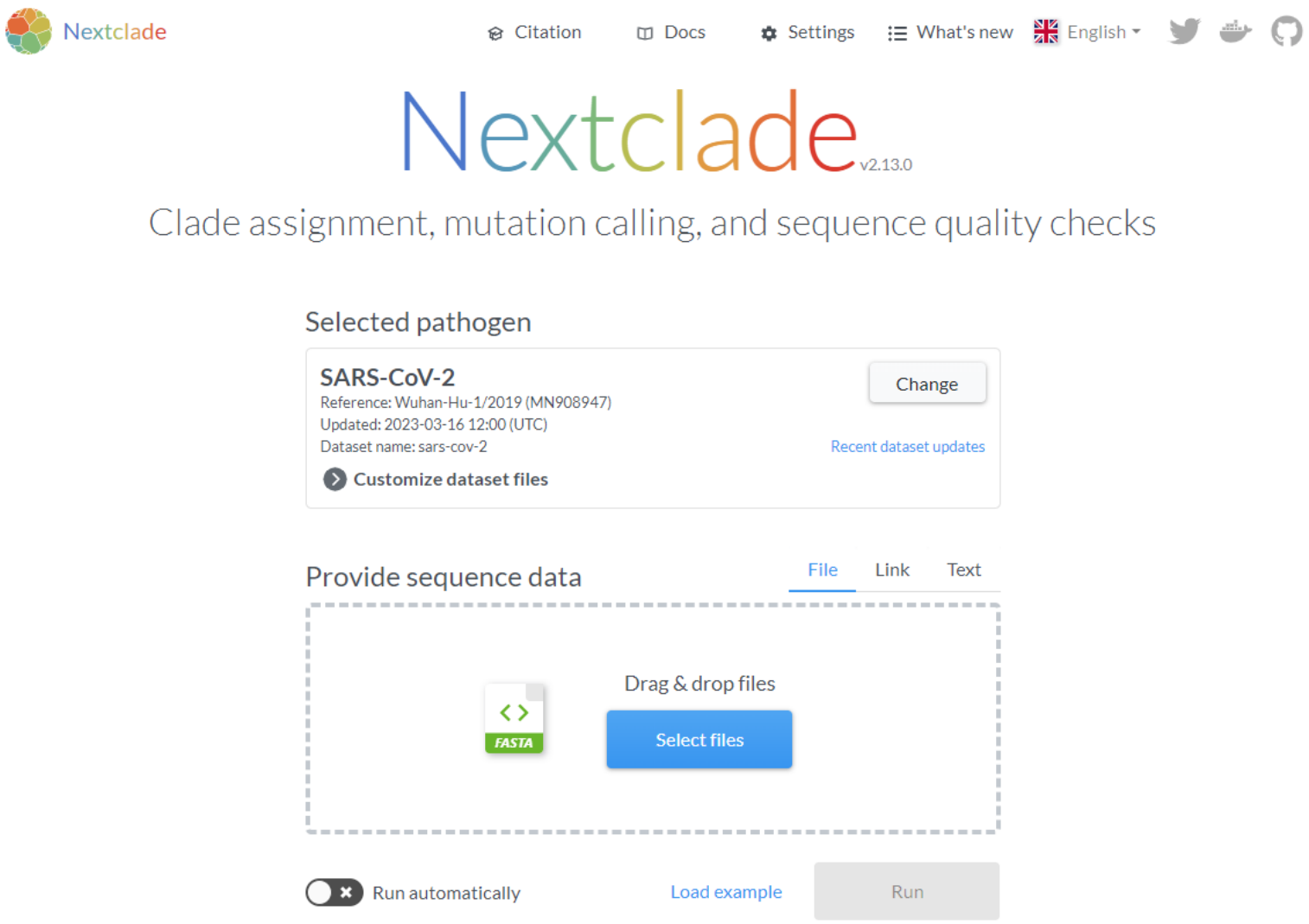
**MSA with Reference**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Seq1** | A | A | T | N | N | G | C | C | A | G | G | A |
| **Seq2** | A | T | T | C | A | G | C | T | A | G | G | A |
| **Seq3** | A | A | T | C | A | G | C | T | T | G | G | A |
| **Seq4** | A | T | T | C | N | N | C | T | T | - | - | - |
| **Ref.** | A | A | T | C | A | G | C | T | A | G | G | A |

# Nextclade – Getting started -1



# Nextclade – Getting started -2



# Nextclade - Result page

Download results

Nextclade

Citation Docs Settings What's new English

Back

Done. Total sequences: 122. Succeeded: 122

| # | i | Sequence name | QC | Clade | Pango lineage (Nextclade) | Unaliased | Mut. | non-ACGTN | Ns | Cov. | Gaps | Ins. | FS | SC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | OP958840 | N M P C F S | 23A | XBB.1.5 | XBB.1.5 | 92 | 0 | 112 | 99.3% | 56 | 0 | 0 | 0 |
| 1 | 1 | OP971202 | N M P C F S | 23A | XBB.1.5.13 | XBB.1.5.13 | 92 | 0 | 0 | 99.3% | 30 | 0 | 0 | 0 (1) |
| 2 | 2 | OP955488 | N M P C F S | 23A | XBB.1.5 | XBB.1.5 | 90 | 0 | 0 | 99.4% | 56 | 0 | 0 | 0 (1) |
| 3 | 3 | OP971186 | N M P C F S | 23A | XBB.1.5.13 | XBB.1.5.13 | 92 | 0 | 2 | 99.2% | 30 | 0 | 0 | 0 (1) |
| 4 | 4 | OP955486 | N M P C F S | 23A | XBB.1.5 | XBB.1.5 | 91 | 1 | 0 | 99.3% | 56 | 0 | 0 | 0 (1) |
| 5 | 5 | OX345943 | N M P C F S | 22F | XBB.3 | XBB.3 | 92 | 0 | 90 | 99.7% | 56 | 0 | 0 | 0 |
| 6 | 6 | OX339384 | N M P C F S | 22E | BQ.1.1 | BA.5.3.1.1.1.1 | 76 | 1 | 296 | 99.0% | 59 | 0 | 0 | 0 |
| 7 | 7 | OP523232 | N M P C F S | 21L | BA.2.3.20 | BA.2.3.20 | 88 | 0 | 0 | 99.6% | 53 | 0 | 0 | 0 |
| 8 | 8 | OX245426 | N M P C F S | 22B | BE.1.1 | BA.5.3.1.1.1 | 72 | 0 | 126 | 99.6% | 59 | 0 | 0 | 0 |
| 9 | 9 | OP333110 | N M P C F S | 21L | BH.1 | BA.2.38.3.1 | 78 | 0 | 0 | 99.6% | 62 | 0 | 0 | 0 |
| 10 | 10 | OP332458 | N M P C F S | 21L | BA.2.10.4 | BA.2.10.4 | 77 | 0 | 0 | 99.6% | 71 | 0 | 0 | 0 |
| 11 | 11 | OP334332 | N M P C F S | 22D | BM.1.1.1 | BA.2.75.3.1.1. | 87 | 0 | 0 | 99.6% | 53 | 0 | 0 | 0 |
| 12 | 12 | OP339227 | N M P C F S | 22D | BA.2.75.2 | BA.2.75.2 | 86 | 0 | 0 | 99.6% | 53 | 0 | 0 | 0 |
| 13 | 13 | ON895103 | N M P C F S | 22D | BA.2.75 | BA.2.75 | 82 | 1 | 0 | 99.6% | 53 | 0 | 0 | 0 |
| 14 | 14 | ON895548 | N M P C F S | 22D | BA.2.75 | BA.2.75 | 83 | 1 | 0 | 99.4% | 53 | 0 | 0 | 0 |
| 15 | 15 | ON537316 | N M P C F S | 22C | BA.2.12.1 | BA.2.12.1 | 74 | 0 | 0 | 99.2% | 53 | 0 | 0 | 0 |
| 16 | 16 | ON544943 | N M P C F S | 21L | BA.2 | BA.2 | 68 | 0 | 0 | 98.3% | 27 | 0 | 0 | 0 |
| 17 | 17 | ON629031 | N M P C F S | 21L | BA.2 | BA.2 | 57 | 0 | 2851 | 90.0% | 27 | 0 | 0 | 0 |
| 18 | 18 | ON626380 | N M P C F S | 21K | BA.1.1 | BA.1.1 | 64 | 0 | 0 | 98.8% | 39 | 9 | 0 | 0 |
| 19 | 19 | ON627541 | N M P C F S | 22B | BA.5.2 | BA.5.2 | 66 | 0 | 209 | 99.2% | 59 | 0 | 0 | 0 |
| 20 | 20 | ON627543 | N M P C F S | 22B | BA.5.2 | BA.5.2 | 69 | 0 | 204 | 99.0% | 59 | 0 | 0 | 0 |
| 21 | 21 | ON396327 | N M P C F S | 22A | BA.4.1 | BA.4.1 | 72 | 0 | 0 | 99.6% | 68 | 0 | 0 | 0 |

Gene S

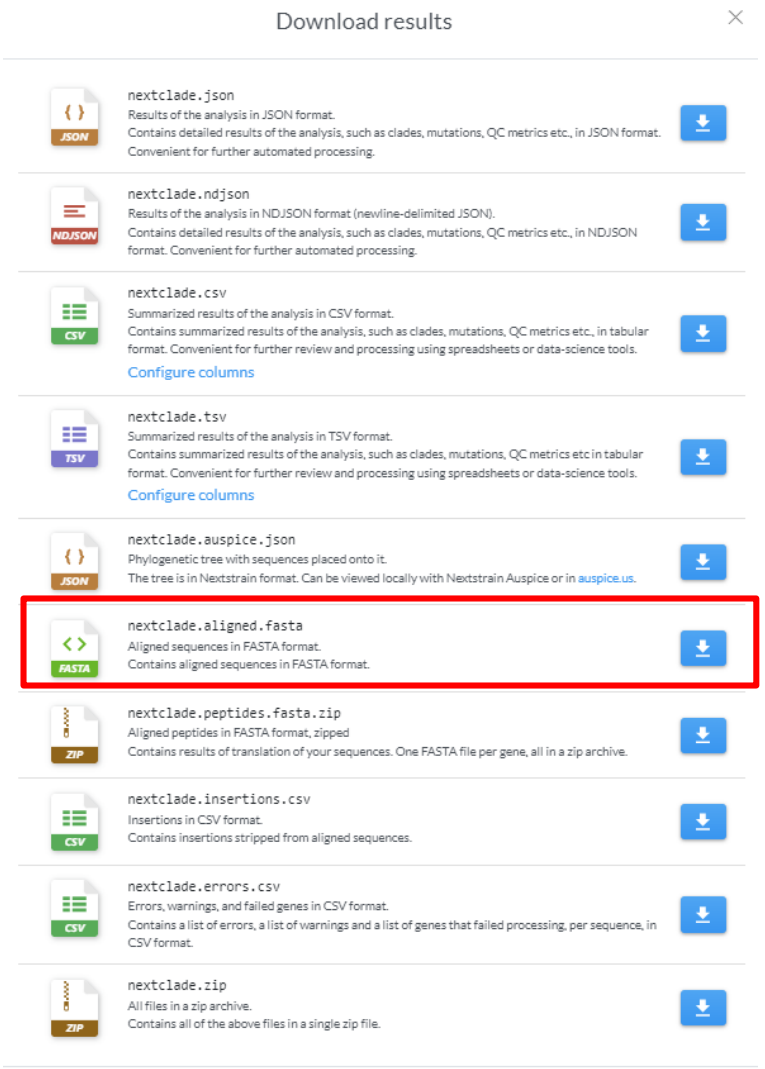
Genome annotation ?

200 400 600 800 1000 1200

Nextclade (c) 2020-2023 Nextstrain developers

version 2.13.0 (commit: 0bab6f0, branch: release)

# Nextclade – Download results



Download results

`nextclade.json`
Results of the analysis in JSON format.
Contains detailed results of the analysis, such as clades, mutations, QC metrics etc., in JSON format. Convenient for further automated processing.

`nextclade.ndjson`
Results of the analysis in NDJSON format (newline-delimited JSON).
Contains detailed results of the analysis, such as clades, mutations, QC metrics etc., in NDJSON format. Convenient for further automated processing.

`nextclade.csv`
Summarized results of the analysis in CSV format.
Contains summarized results of the analysis, such as clades, mutations, QC metrics etc., in tabular format. Convenient for further review and processing using spreadsheets or data-science tools.
Configure columns

`nextclade.tsv`
Summarized results of the analysis in TSV format.
Contains summarized results of the analysis, such as clades, mutations, QC metrics etc in tabular format. Convenient for further review and processing using spreadsheets or data-science tools.
Configure columns

`nextclade.auspice.json`
Phylogenetic tree with sequences placed onto it.
The tree is in Nextstrain format. Can be viewed locally with Nextstrain Auspice or in auspice.us.

`nextclade.aligned.fasta`
Aligned sequences in FASTA format.
Contains aligned sequences in FASTA format.

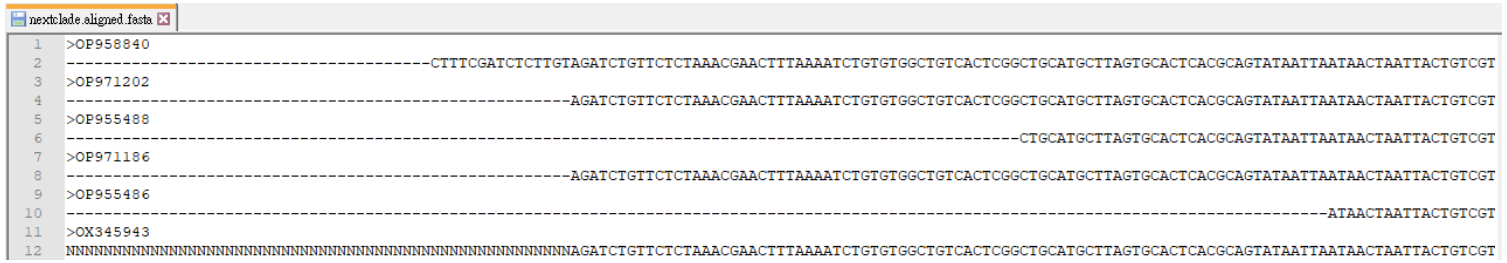
`nextclade.peptides.fasta.zip`
Aligned peptides in FASTA format, zipped
Contains results of translation of your sequences. One FASTA file per gene, all in a zip archive.

`nextclade.insertions.csv`
Insertions in CSV format.
Contains insertions stripped from aligned sequences.

`nextclade.errors.csv`
Errors, warnings, and failed genes in CSV format.
Contains a list of errors, a list of warnings and a list of genes that failed processing, per sequence, in CSV format.

`nextclade.zip`
All files in a zip archive.
Contains all of the above files in a single zip file.

Nextclade aligned (FASTA)



```
>OP958840
---------------------------------------CTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTCGT
>OP971202
------------------------------------------------------AGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTCGT
>OP955488
------------------------------------------------------------------------------------------------------CTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTCGT
>OP971186
------------------------------------------------------AGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTCGT
>OP955486
---------------------------------------------------------------------------------------------------------------------------------------ATAACTAATTACTGTCGT
>OX345943
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTCGT
```

more details in https://docs.nextstrain.org/projects/nextclade/en/stable/user/output-files.html

# Data Analysis in R

# Material

- Data
  - Meatadata_n1932.csv
  - n1932_29903.fas
  - MN908947.fas
  - region.csv
- Code
  - COVID19 Genomic data analysis.R

# Distance/ Similarity

- Continuous: Covariance, Euclidean distance, Kendall's tao, Pearson's correlation coefficient, Speakman's rank,
- Binary: Hamman, Jaccard, Phi, Rao, Rogers, Simple match, Sneath, Yule

# Agglomerative Hierarchical Clustering

- Agglomerative hierarchical clustering is a popular method used in cluster analysis and data mining to group similar items or data points into clusters.

1. **Initialization**: At the beginning, each data point is considered as a separate cluster.
2. **Compute Pairwise Similarities or Distances Merge Similar Clusters**: The algorithm identifies the two closest clusters based on the similarity or distance measure and merges them into a new larger cluster. This process is repeated iteratively, and at each step, the algorithm updates the similarity or distance matrix to reflect the newly formed clusters.
3. **Update Similarity Matrix**: Different linkage methods can be used to determine the distance between two clusters:
   - Single Linkage:
   $$d(C_i, C_j) = \min_{a \in C_i, b \in C_j} d(a, b)$$
   - Complete Linkage:
   $$d(C_i, C_j) = \max_{a \in C_i, b \in C_j} d(a, b)$$
   - Average Linkage:
   $$d(C_i, C_j) = \sum_{a \in C_i, b \in C_j} \frac{d(a, b)}{|C_i||C_j|}$$
4. **Stopping Criterion**: The process continues until a stopping criterion is met. This criterion can be a specific number of desired clusters or a threshold value for the distance between clusters. At this point, the algorithm stops, and the final dendrogram represents the hierarchical clustering.


23

# Agglomerative Hierarchical Clustering -- Complete Linkage

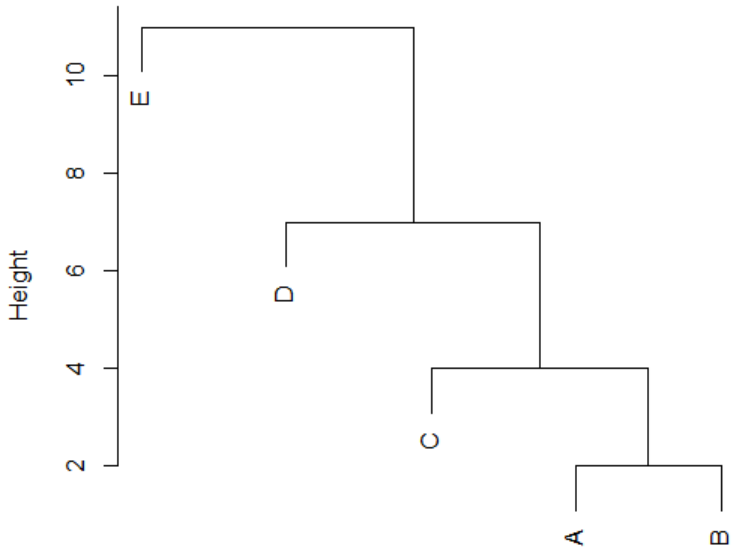
1. **Initialization**
2. **Compute Pairwise Similarities or Distances Merge Similar Clusters**
3. **Update Similarity Matrix**
4. **Stopping Criterion**

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 0 | | | | |
| B | 2 | 0 | | | |
| C | 3 | 4 | 0 | | |
| D | 5 | 6 | 7 | 0 | |
| E | 8 | 9 | 10 | 11 | 0 |

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 0 | | | | |
| B | 2 | 0 | | | |
| C | 3 | 4 | 0 | | |
| D | 5 | 6 | 7 | 0 | |
| E | 8 | 9 | 10 | 11 | 0 |

| | (A,B) | C | D | E |
|---|---|---|---|---|
| (A,B) | 0 | | | |
| C | 4 | 0 | | |
| D | 6 | 7 | 0 | |
| E | 9 | 10 | 11 | 0 |

**Cluster Dendrogram**




hclust (*, "complete")

# Further Analysis

1. Do descriptive statistics
2. Add metadata (Collection Date, Country, ……) and gene region as covariates to the analysis.
3. Visualize the results.
4. Clustering column (position).
5. Download current sequencing data and reanalysis.
6. ……

# Discussion