

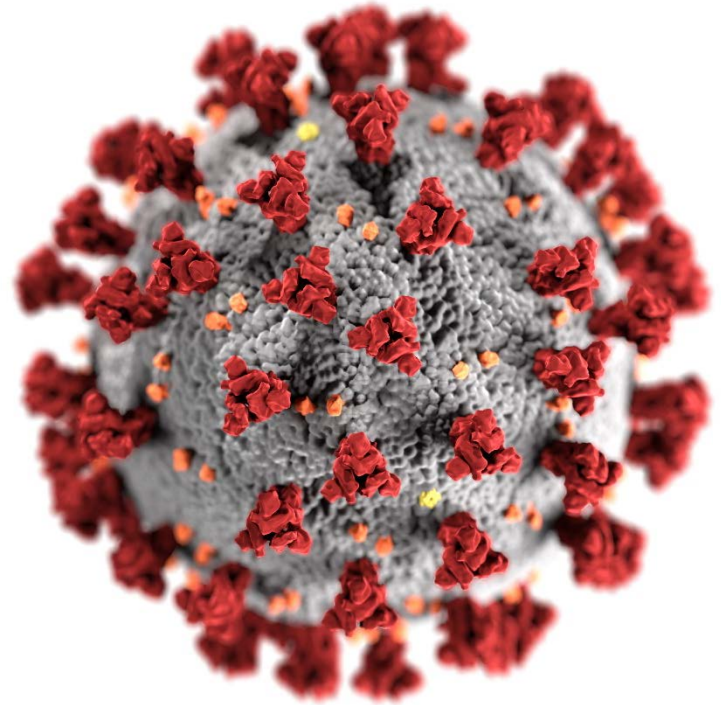
# Public genome repositories for SARS-CoV-2

## COVID-19 Genomic Epidemiology Toolkit: Module 3.5

Michael Weigand, PhD

Bioinformatician

Centers for Disease Control and Prevention



[cdc.gov/coronavirus](https://cdc.gov/coronavirus)

# Toolkit map

## Part 1: Introduction

- 1.1 What is genomic epidemiology?
- 1.2 The SARS-CoV-2 genome
- 1.3 How to read phylogenetic trees
- 1.4 Emerging variants of SARS-CoV-2

## Part 2: Case Studies

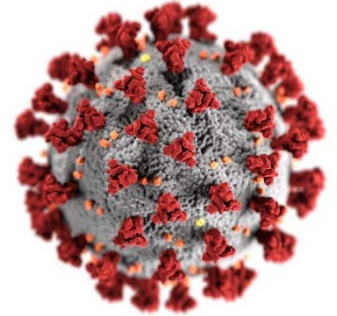
- 2.1 SARS-CoV-2 sequencing in Arizona
- 2.2 Healthcare cluster transmission
- 2.3 Community transmission
- 2.4 Superspreading event

## Part 3: Implementation

- 3.1 Getting started with Nextstrain
- 3.2 Getting started with MicrobeTrace
- 3.3 Phylogenetics with USHER
- 3.4 Walking through Nextstrain trees
- 3.5 Public genome repositories**



# Rationale for sequencing SARS-CoV-2



- State/local level
  - Supplement control efforts
  - Better understand epidemiology
- National level
  - Strain surveillance
  - Guide diagnostics, vaccine, and therapeutic development
- **Collecting sequences in public databases strengthens both**
  - Submission should be included in any SARS-CoV-2 sequencing workflow
  - More valuable if clinical, epidemiological metadata included

# Public repositories for SARS-CoV-2 genomic data

## 1. GISAID (*Global Initiative on Sharing All Influenza Data*)

[www.gisaid.org](http://www.gisaid.org)

- Facilitates rapid sharing of data from influenza (EpiFlu)
  - and SARS-CoV-2 (EpiCoV)

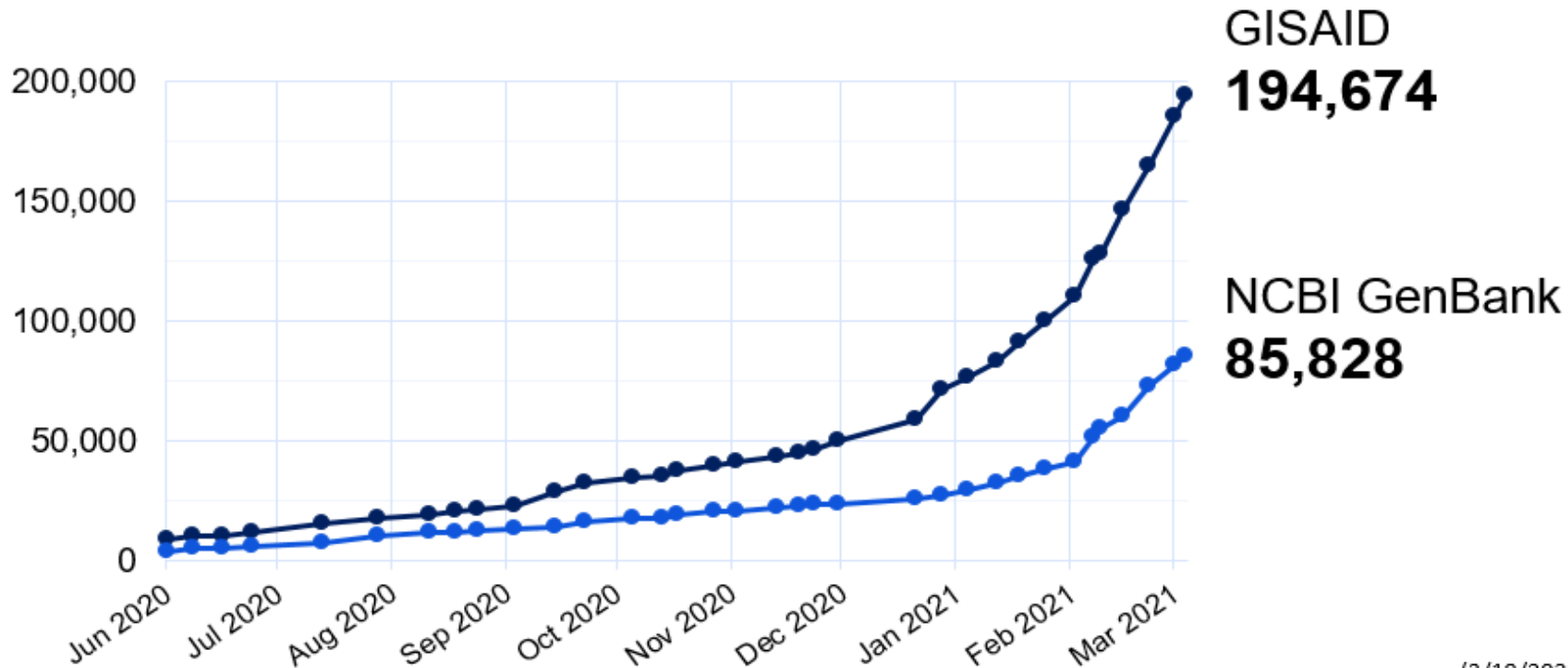
## 2. NCBI/NLM (*National Center for Biotechnology Information*)

[www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)

- US National Institutes of Health (NIH)
  - GenBank®, Sequence Read Archive (SRA), PubMed, BLAST®, etc.
  - Facilitates public access and rapid sharing of SARS-CoV-2 sequences
  - Integrates SARS-CoV-2 data with literature (PubMed/PMC) and the BLAST® databases
- 
- Different submission standards, organizations, primary uses
  - Both provide searchable collections of genomic data and epidemiologic metadata.

# SARS-CoV-2 submissions (GISAID + NCBI)

From US laboratories



GISAID  
**194,674**

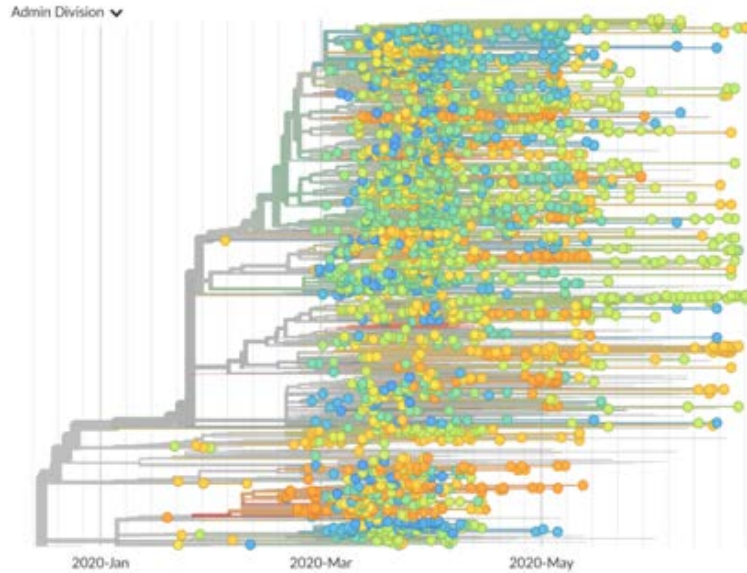
NCBI GenBank  
**85,828**

(3/19/2021)

(3/19/2021)

# GISAID data enables popular SARS-CoV-2 tools

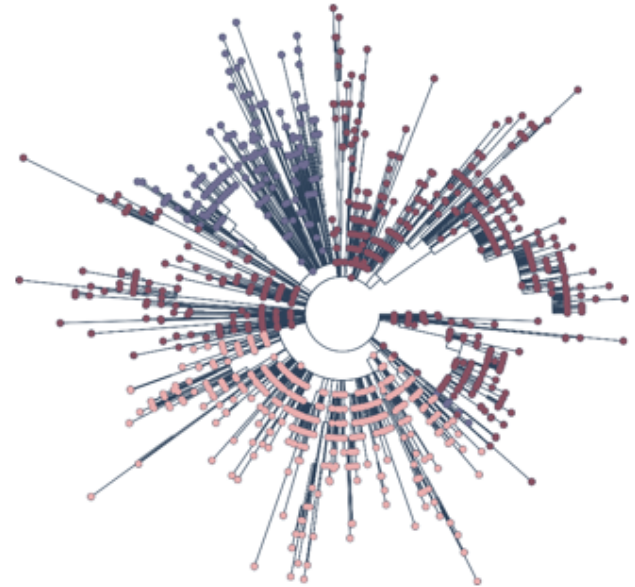
## Nextstrain



[nextstrain.org](https://nextstrain.org)

## PANGO lineages

Phylogenetic Assignment of Named Global Outbreak



[cov-lineages.org](https://cov-lineages.org)

© 2008 - 2021 | Terms of Use | Privacy Notice | Contact

You are logged in as Michael Weigand - [logout](#)

Registered Users EpiFlu™ EpiCoV™ My profile

EpiCoV™ Search Downloads Upload My Unreleased

## Pandemic coronavirus causing COVID-19

A previously unknown human coronavirus (hCoV-19) was first detected in late 2019 in patients in the City of Wuhan, who suffered from respiratory illnesses including atypical pneumonia, an illness that has become known as coronavirus disease (COVID-19). The coronavirus originated from an animal host and is closely related to the virus responsible for the Severe Acute Respiratory Syndrome coronavirus (SARS).

On 10. January 2020, the first virus genomes and associated data were publicly shared via GISAID. The World Health Organization announced on 11. March 2020 the first coronavirus pandemic. As the pandemic progresses, scientists from around the globe are tracking the virus and its genome sequences to ensure optimal virus diagnostic tests, to track and trace the ongoing outbreak and to identify potential intervention options. Several analyses to assist with these efforts are offered here, including sequence alignments, diagnostic primer and probe coordinates, 3D protein models, drug targets, phylogenetic trees and many more.

Search

### Analysis Update (2021-02-23)

 Full genome tree derived from all outbreak sequences	 Timecourse of clade distribution in collected sequences	 Regional clade distribution of new sequences	 Common primer check for high quality genomes	 Receptor binding surveillance for complete genomes	 Full genome tree of hCoV-19-related precursors
 Clade evolution in the first year	 Temporal and regional distribution of clades in the first year	 Spike comparison between pangolin, bat, human	 Spike comparison to SARS and bat precursor	 Highly conserved drug targets between hCoV-19 and SARS	 analysis update.pdf
 Official GISAID reference sequence	 Audacity	 Spike glycoprotein mutation surveillance	 BLAST		

Important note: In the [GISAID EpiCoV™ 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 EpiCoV™ Database Access Agreement](#) in respect of such data in the same manner as if were data relating to influenza viruses.

Sequence search

## Analyses

- Phylogeny
- Clade abundance
- Spike mutations

**Location**

(North America / USA / Georgia)

**Collection date**

(2021-01-01 to 2021-01-31)

**Substitutions**

(Spike\_E484K)

**Variants**

(GH/501Y.V2/B.1.351)

**CoVsurver**

**Download**

**Search**

Accession ID:  Virus name:   complete  high coverage  low coverage excl  w/Patient status  collection date compl

Location:  Host:

Collection:  to  Submission:  to

Side:  Lineage:  Substitutions:  Variants:

<input type="checkbox"/>	Virus name	Passage	Accession ID	Collection date	Submission	Length	Host	Location	Originating
<input type="checkbox"/>	hCoV-19/USA/FL-CDC-LC0021227/2021	Original	EPI_ISL_1298429	2021-03-02	2021-03-19	29,695	Human	North America / USA / Georgia	Laboratory
<input type="checkbox"/>	hCoV-19/USA/AL-CDC-LC0021173/2021	Original	EPI_ISL_1298428	2021-03-01	2021-03-19	29,694	Human	North America / USA / Georgia	Laboratory
<input type="checkbox"/>	hCoV-19/USA/FL-CDC-LC0021169/2021	Original	EPI_ISL_1298427	2021-03-01	2021-03-19	29,694	Human	North America / USA / Georgia	Laboratory
<input type="checkbox"/>	hCoV-19/USA/FL-CDC-LC0021166/2021	Original	EPI_ISL_1298426	2021-03-01	2021-03-19	29,694	Human	North America / USA / Georgia	Laboratory
<input type="checkbox"/>	hCoV-19/USA/TN-CDC-LC0021173/2021	Original	EPI_ISL_1298425	2021-03-01	2021-03-19	29,694	Human	North America / USA / Georgia	Laboratory
<input type="checkbox"/>	hCoV-19/USA/GA-CDC-LC0021169/2021	Original	EPI_ISL_1298424	2021-03-01	2021-03-19	29,694	Human	North America / USA / Georgia	Laboratory
<input type="checkbox"/>	hCoV-19/USA/TN-CDC-LC0021166/2021	Original	EPI_ISL_1298423	2021-03-01	2021-03-19	29,694	Human	North America / USA / Georgia	Laboratory
<input type="checkbox"/>	hCoV-19/USA/FL-CDC-LC0021161/2021	Original	EPI_ISL_1298422	2021-02-27	2021-03-19	29,694	Human	North America / USA / Georgia	Laboratory
<input type="checkbox"/>	hCoV-19/USA/GA-CDC-LC0021144/2021	Original	EPI_ISL_1298421	2021-02-28	2021-03-19	29,694	Human	North America / USA / Georgia	Laboratory
<input type="checkbox"/>	hCoV-19/USA/GA-CDC-LC0021138/2021	Original	EPI_ISL_1298420	2021-02-27	2021-03-19	29,694	Human	North America / USA / Georgia	Laboratory
<input type="checkbox"/>	hCoV-19/USA/TN-CDC-LC0021137/2021	Original	EPI_ISL_1298419	2021-02-27	2021-03-19	29,694	Human	North America / USA / Georgia	Laboratory
<input type="checkbox"/>	hCoV-19/USA/GA-CDC-LC0021113/2021	Original	EPI_ISL_1298418	2021-02-27	2021-03-19	29,694	Human	North America / USA / Georgia	Laboratory

Total: 45,927 viruses

<< < 1 2 3 4 5 > >>

Select  Analysis


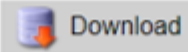


# GISAID: download

**Download**

Format

- Sequences (FASTA)
- Patient status metadata
- Sequencing technology metadata
- Acknowledgement (Supplemental table)

 Back 

## FASTA

```
>Sequence 1
AAAUGUUAUUCAUGCU
>Sequence 2
AAAUAUUACUCAUGCU
>Sequence 3
AAAUAUUACUCAUGCC
```

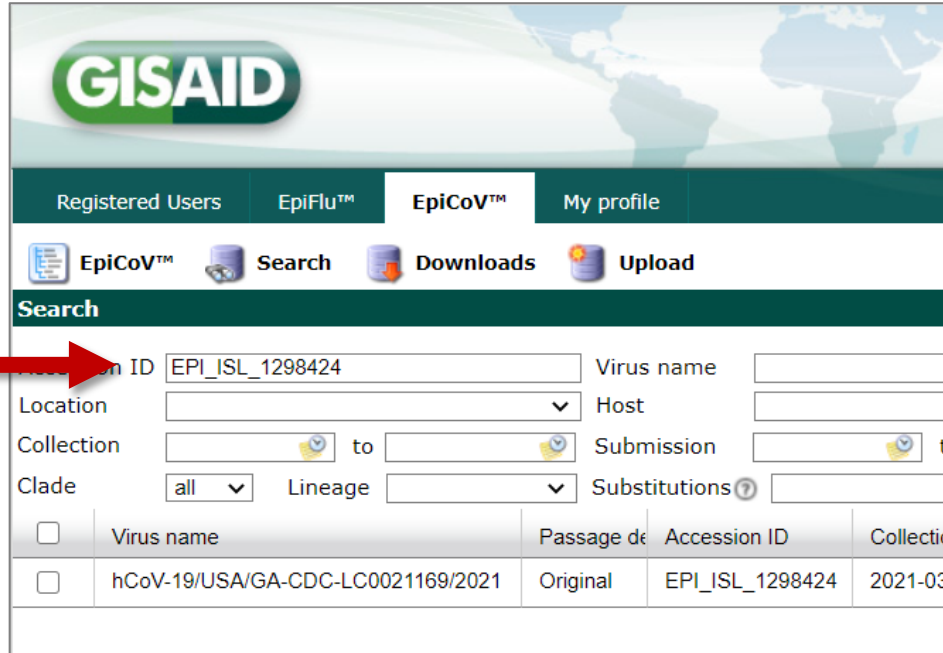
## Metadata tsv

Accession	Date	Location
Seq1	02/05/2021	USA/GA
Seq2	02/05/2021	USA/GA
Seq3	02/05/2021	USA/GA

GISAID Database Access Agreement

[www.gisaid.org/registration/terms-of-use/](http://www.gisaid.org/registration/terms-of-use/)

# GISAID: find individual records



The screenshot shows the GISAID search interface. At the top, there is a navigation bar with 'Registered Users', 'EpiFlu™', 'EpiCoV™', and 'My profile'. Below this is a secondary bar with 'EpiCoV™', 'Search', 'Downloads', and 'Upload'. The main search area has a 'Search' header and several input fields: 'Accession ID' (containing 'EPI\_ISL\_1298424'), 'Virus name', 'Location', 'Host', 'Collection' (with a date range), 'Submission' (with a date range), 'Clade' (set to 'all'), 'Lineage', and 'Substitutions'. Below the search fields is a table with columns: 'Virus name', 'Passage details', 'Accession ID', and 'Collection'. The first row of the table contains the following data: 'hCoV-19/USA/GA-CDC-LC0021169/2021', 'Original', 'EPI\_ISL\_1298424', and '2021-03-'. A red arrow points from a box labeled 'Accession ID' to the 'Accession ID' input field.

Accession ID

## Virus detail

Virus name:  
Accession ID:  
Type:  
Clade  
Pango Lineage  
AA Substitutions

## Variant

Passage details/history:

## Sample information

Collection date:  
Location:  
Host:  
Additional location information:  
Gender:  
Patient age:  
Patient status:  
Specimen source:  
Additional host information:  
Outbreak:  
Last vaccinated:  
Treatment:  
Sequencing technology:  
Assembly method:  
Coverage:  
Comment:

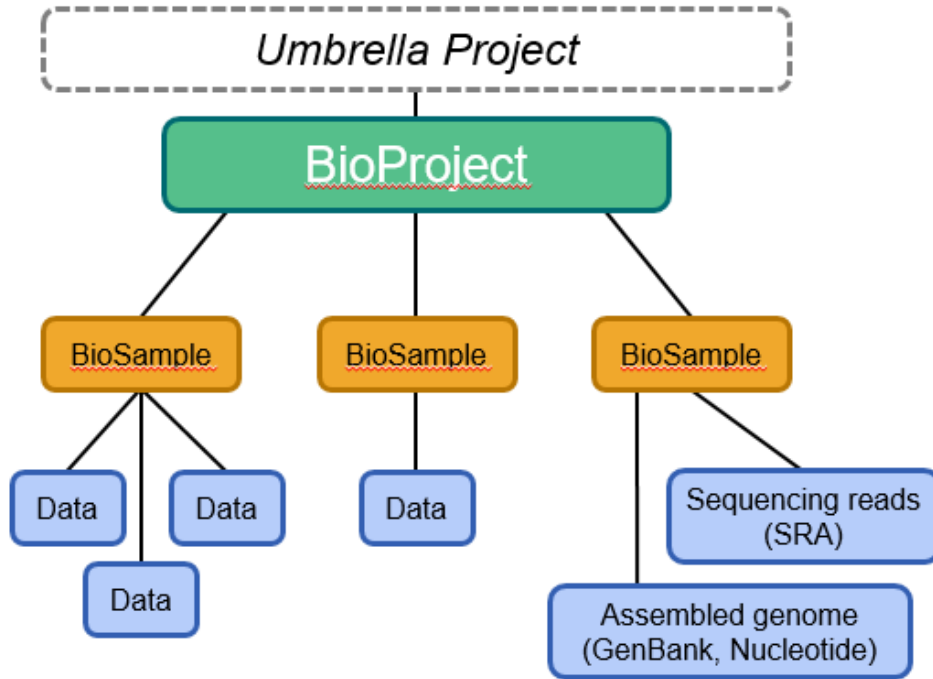
## Institute information

Originating lab:  
Address:  
Sample ID given by the  
originating laboratory:  
Submitting lab:

GISAID Database Access Agreement

[www.gisaid.org/registration/terms-of-use/](http://www.gisaid.org/registration/terms-of-use/)

# NCBI database organization



- BioProject – a collection of biological data for a single initiative
  - *Umbrella Project* – a collection of BioProjects
- BioSample – information about the physical specimen (metadata)
- Data – genomic datasets; e.g.
  - Raw sequencing reads
  - Assembled genome



## NCBI SARS-CoV-2 Resources

### Quick Navigation Guide

[Sequence Submission](#)

[Literature](#)

[Sequence-Related Resources](#)

[Clinical Resources](#)

[Other Websites](#)

### SARS-CoV-2 Data

**254,626**

[SRA runs](#)

**117,986**

[Nucleotide records](#)

**5,113**

[ClinicalTrials.gov](#)

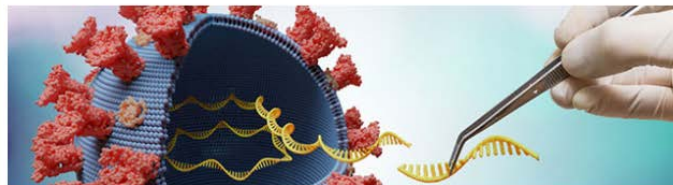
**115,033**

[PubMed](#)

**125,241**

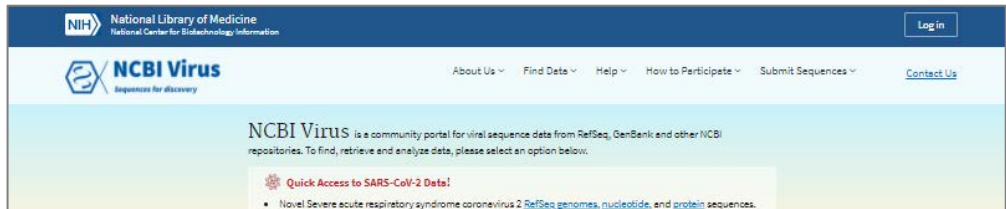
[PMC](#)

### Submit SARS-CoV-2 Sequences



Add assembled & raw read data to the growing public archive

[Submit Now](#)



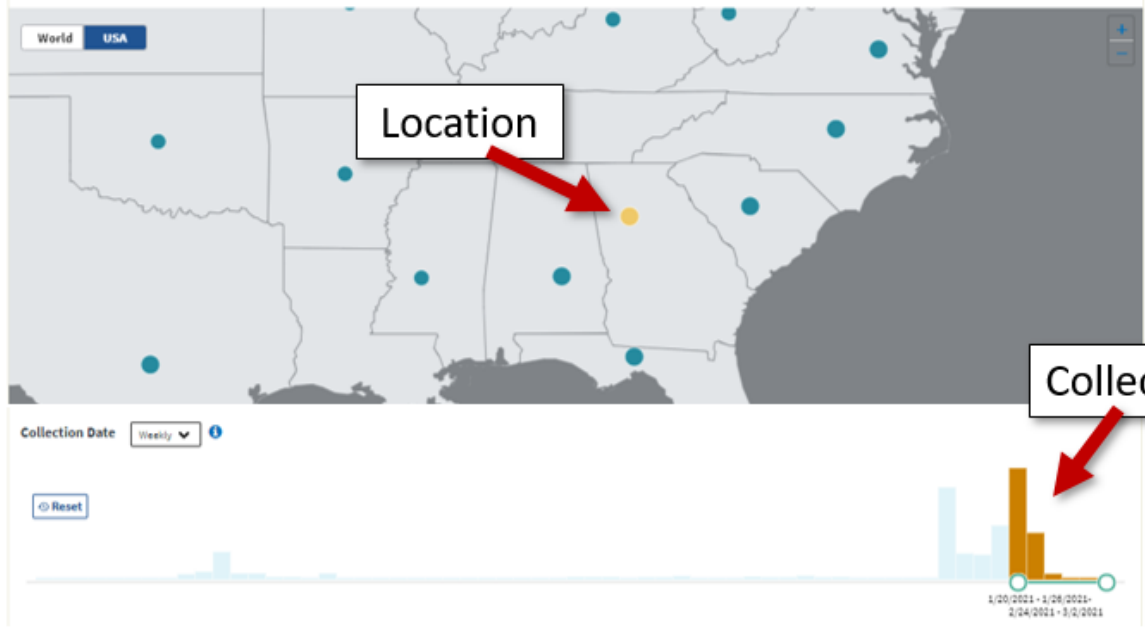
**SARS-CoV-2 Data Dashboard** Filters (0) View results, Analyze, or Download

<b>1</b> RefSeq Genomes	<b>1,317,831</b> All Proteins	<b>117,986</b> All Nucleotides	<b>38</b> RefSeq Proteins	<b>66,259</b> Complete Nucleotides
----------------------------	----------------------------------	-----------------------------------	------------------------------	---------------------------------------



**SARS-CoV-2 Data Dashboard** Filters (2) View results, Analyze, or Download View table of selection

0 RefSeq Genomes      4,690 All Proteins      403 All Nucleotides      0 RefSeq Proteins      69 Complete Nucleotides



Refine Results Reset

Virus +

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

Accession +

Sequence Length +

Sequence Type +

RefSeq Genome Completeness +

Nucleotide Completeness +

Proteins +

Provirus +

Geographic Region +

USA: GA x

Host +

Isolation Source +

Collection Date +

From Jan 20, 2021 x  
To Mar 2, 2021

Selected Results: 0

PubMed Download Align Build Phylogenetic Tree

Expand Table

Nucleotide (403)		Protein (4,690)		RefSeq Genome (0)		Select Columns	
<input type="checkbox"/>	Accession	Submitters	Release Date	Species	Molecule type	Length	Geographic Location
<input type="checkbox"/>	<a href="#">MW642833</a>	Cook,P.W., et al.	2021-02-21	Severe acute respiratory s...	ssRNA(+)	29713	USA: Georgia
<input type="checkbox"/>	<a href="#">MW642834</a>	Cook,P.W., et al.	2021-02-21	Severe acute respiratory s...	ssRNA(+)	29713	USA: Georgia
<input type="checkbox"/>	<a href="#">MW642850</a>	Cook,P.W., et al.	2021-02-21	Severe acute respiratory s...	ssRNA(+)	29713	USA: Georgia
<input type="checkbox"/>	<a href="#">MW642851</a>	Cook,P.W., et al.	2021-02-21	Severe acute respiratory s...	ssRNA(+)	29713	USA: Georgia
<input type="checkbox"/>	<a href="#">MW642852</a>	Cook,P.W., et al.	2021-02-21	Severe acute respiratory s...	ssRNA(+)	29713	USA: Georgia
<input type="checkbox"/>	<a href="#">MW642853</a>	Cook,P.W., et al.	2021-02-21	Severe acute respiratory s...	ssRNA(+)	29713	USA: Georgia
<input type="checkbox"/>	<a href="#">MW642854</a>	Cook,P.W., et al.	2021-02-21	Severe acute respiratory s...	ssRNA(+)	29713	USA: Georgia
<input type="checkbox"/>	<a href="#">MW642855</a>	Cook,P.W., et al.	2021-02-21	Severe acute respiratory s...	ssRNA(+)	29713	USA: Georgia
<input type="checkbox"/>	<a href="#">MW642856</a>	Cook,P.W., et al.	2021-02-21	Severe acute respiratory s...	ssRNA(+)	29696	USA: Georgia
<input type="checkbox"/>	<a href="#">MW642925</a>	Cook,P.W., et al.	2021-02-21	Severe acute respiratory s...	ssRNA(+)	29713	USA: Georgia
<input type="checkbox"/>	<a href="#">MW642934</a>	Cook,P.W., et al.	2021-02-21	Severe acute respiratory s...	ssRNA(+)	29713	USA: Georgia
<input type="checkbox"/>	<a href="#">MW642935</a>	Cook,P.W., et al.	2021-02-21	Severe acute respiratory s...	ssRNA(+)	29713	USA: Georgia
<input type="checkbox"/>	<a href="#">MW642936</a>	Cook,P.W., et al.	2021-02-21	Severe acute respiratory s...	ssRNA(+)	29713	USA: Georgia
<input type="checkbox"/>	<a href="#">MW642937</a>	Cook,P.W., et al.	2021-02-21	Severe acute respiratory s...	ssRNA(+)	29713	USA: Georgia
<input type="checkbox"/>	<a href="#">MW642938</a>	Cook,P.W., et al.	2021-02-21	Severe acute respiratory s...	ssRNA(+)	29713	USA: Georgia
<input type="checkbox"/>	<a href="#">MW642948</a>	Cook,P.W., et al.	2021-02-21	Severe acute respiratory s...	ssRNA(+)	29713	USA: Georgia

Location

Collection date

Download

Analyze

# NCBI download

Download Results ✕

Step 1 of 3: Select Data Type

<p>Sequence data (FASTA Format)</p> <p><input checked="" type="radio"/> Nucleotide</p> <p><input type="radio"/> Coding Region</p> <p><input type="radio"/> Protein</p>	<p>Accession List</p> <p><input type="radio"/> Nucleotide</p> <p><input type="radio"/> Protein</p> <p><input type="radio"/> Assembly</p>	<p>Current table view result</p> <p><input type="radio"/> CSV format</p> <p><input type="radio"/> XML format</p>
--	--	--

[Next](#)

## FASTA

```
>Sequence 1
AAAUGUUUAUCAUGCU
>Sequence 2
AAAUAUUACUCAUGCU
>Sequence 3
AAAUAUUACUCAUGCC
```

## List.txt

```
Sequence1
Sequence2
Sequence3
```

## Metadata.csv

Accession	Date	Location
Seq1	02/05/2021	USA/GA
Seq2	02/05/2021	USA/GA
Seq3	02/05/2021	USA/GA



**GISAID**



**NCBI**



## Supplement local analyses

Local  
data

**FASTA**

```
>Sequence 1  
AAAUGUUUUUCAUGCU  
>Sequence 2  
AAAUAUUACUCAUGCU  
>Sequence 3  
AAAUAUUACUCAUGCC  
>Sequence 4  
AAACGUUACUAAUGCU
```

**Nextstrain**

(Module 3.1)



**MicrobeTrace**

(Module 3.2)



**UShER**

(Module 3.3)



# Summary

- Public repositories facilitate organized, open data sharing
- Maximize utility of SARS-CoV-2 genome sequence data
  - Support resources like public Nextstrain builds
  - Query, download sequences to supplement local analyses or investigation

**GISAID**



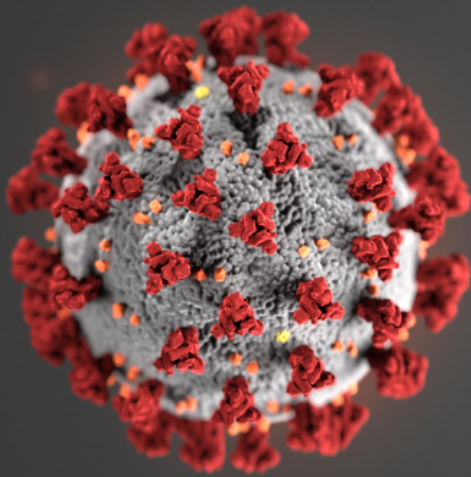
**National Library of Medicine**  
*National Center for Biotechnology Information*

- Pay it forward
  - Submitting your data to both repositories can help others!
  - Better with clinical, epidemiological metadata (PHA4GE)

# Learn more

- Other modules
  - Getting started with Nextstrain – Module 3.1
  - Getting started with MicrobeTrace – Module 3.2
  - Real-time phylogenetics with UShER – Module 3.3
- COVID-19 Genomic Epidemiology Toolkit
  - Find further reading
  - Complete a feedback survey
  - Subscribe to receive updates on new modules as they are released
  - [go.usa.gov/xAbMw](https://go.usa.gov/xAbMw)





For more information, contact CDC  
1-800-CDC-INFO (232-4636)  
TTY: 1-888-232-6348 [www.cdc.gov](http://www.cdc.gov)

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.



# FOR WEBSITE

- Further Reading:
- Resources:
  - Global Initiative on Sharing All Influenza Data (GISAID). [www.gisaid.org](http://www.gisaid.org)
  - National Center for Biotechnology Information (NCBI). [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)
  - NCBI SARS-CoV-2 resources. [www.ncbi.nlm.nih.gov/sars-cov-2/](http://www.ncbi.nlm.nih.gov/sars-cov-2/)
  - NCBI Virus SARS-CoV-2 data dashboard. [www.ncbi.nlm.nih.gov/labs/virus/vssi/#/sars-cov-2](http://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/sars-cov-2)
  - NCBI Umbrella project of US sequencing efforts. [www.ncbi.nlm.nih.gov/bioproject/PRJNA615625](http://www.ncbi.nlm.nih.gov/bioproject/PRJNA615625)
  - NCBI GenBank submission. <https://submit.ncbi.nlm.nih.gov/sarscov2/genbank>
  - NCBI SRA submission. <https://submit.ncbi.nlm.nih.gov/sarscov2/sra>
  
  - Public Health Alliance for Genomic Epidemiology (PHA4GE) resources. <https://pha4ge.org/resources/>
  - Public Health Alliance for Genomic Epidemiology (PHA4GE) protocols. <https://www.protocols.io/workspaces/pha4ge/publications>
  - SARS-CoV-2 GenBank submission protocol. <https://www.protocols.io/view/sars-cov-2-ncbi-assembly-submission-protocol-genba-bg2tjyen>
  - SARS-CoV-2 SRA submission protocol. <https://www.protocols.io/view/sars-cov-2-ncbi-submission-protocol-sra-biosample-bf7bjrin>
  - SARS-CoV-2 GISAID submission protocol. <https://www.protocols.io/view/sars-cov-2-gisaid-submission-protocol-bh98j99w>