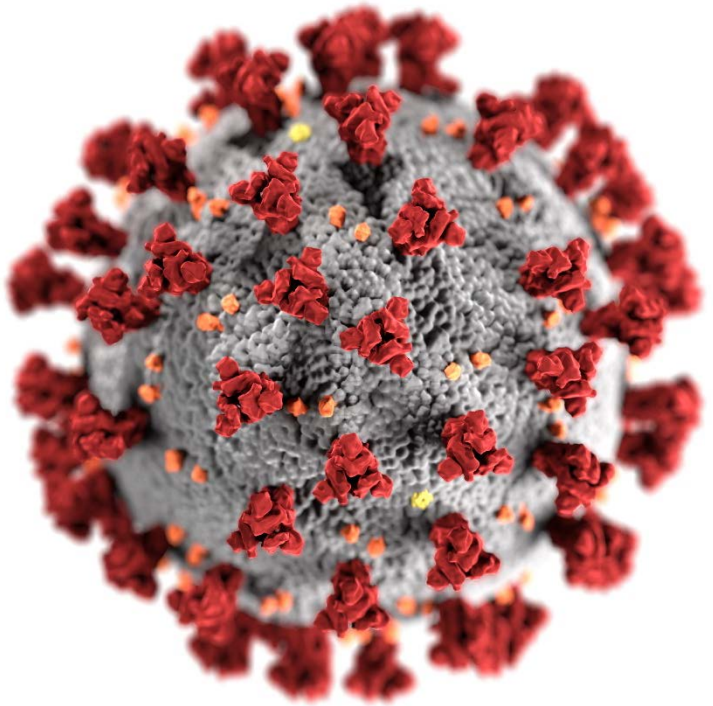
Public genome repositories for SARS-CoV-2

COVID-19 Genomic Epidemiology Toolkit: Module 3.5



Michael Weigand, PhD
Bioinformatician
Centers for Disease Control and Prevention

cdc.gov/coronavirus

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

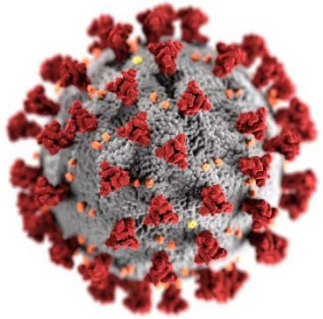
3.2 Getting started with MicrobeTrace

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

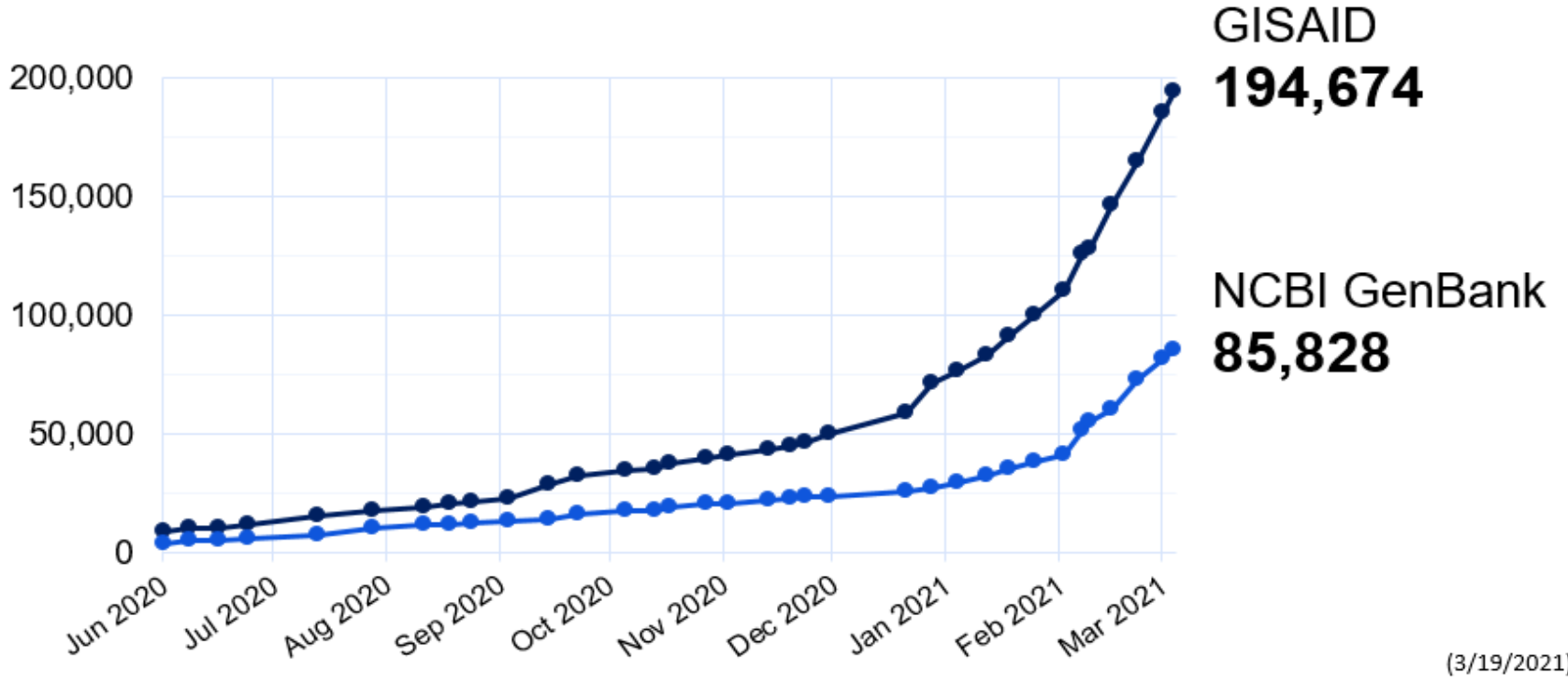
1. **GISAID** (*Global Initiative on Sharing All Influenza Data*) www.gisaid.org
   - Facilitates rapid sharing of data from influenza (EpiFlu)
     - and SARS-CoV-2 (EpiCoV)

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

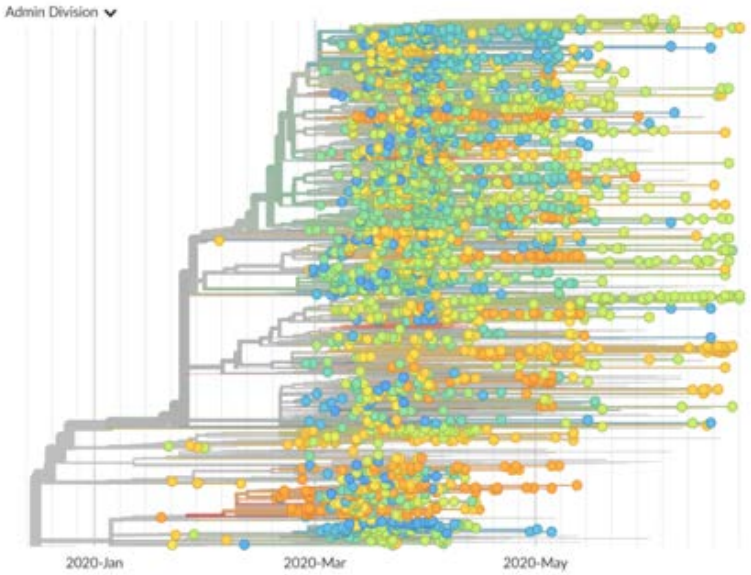
From US laboratories



(3/19/2021)

GISAID data enables popular SARS-CoV-2 tools

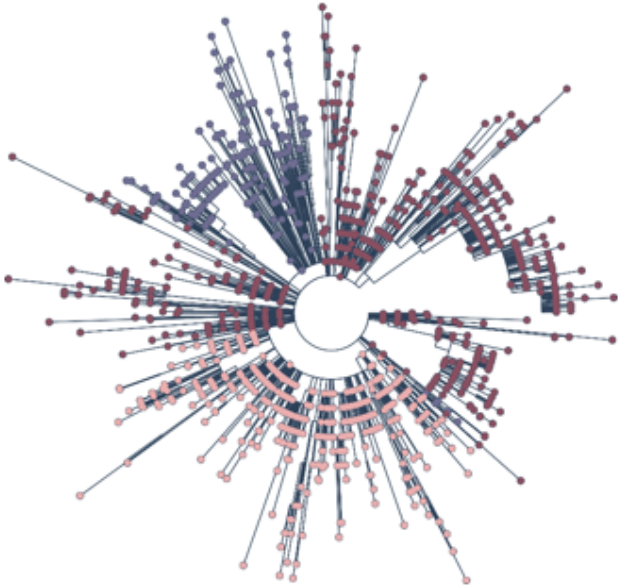
Nextstrain



nextstrain.org

PANGO lineages

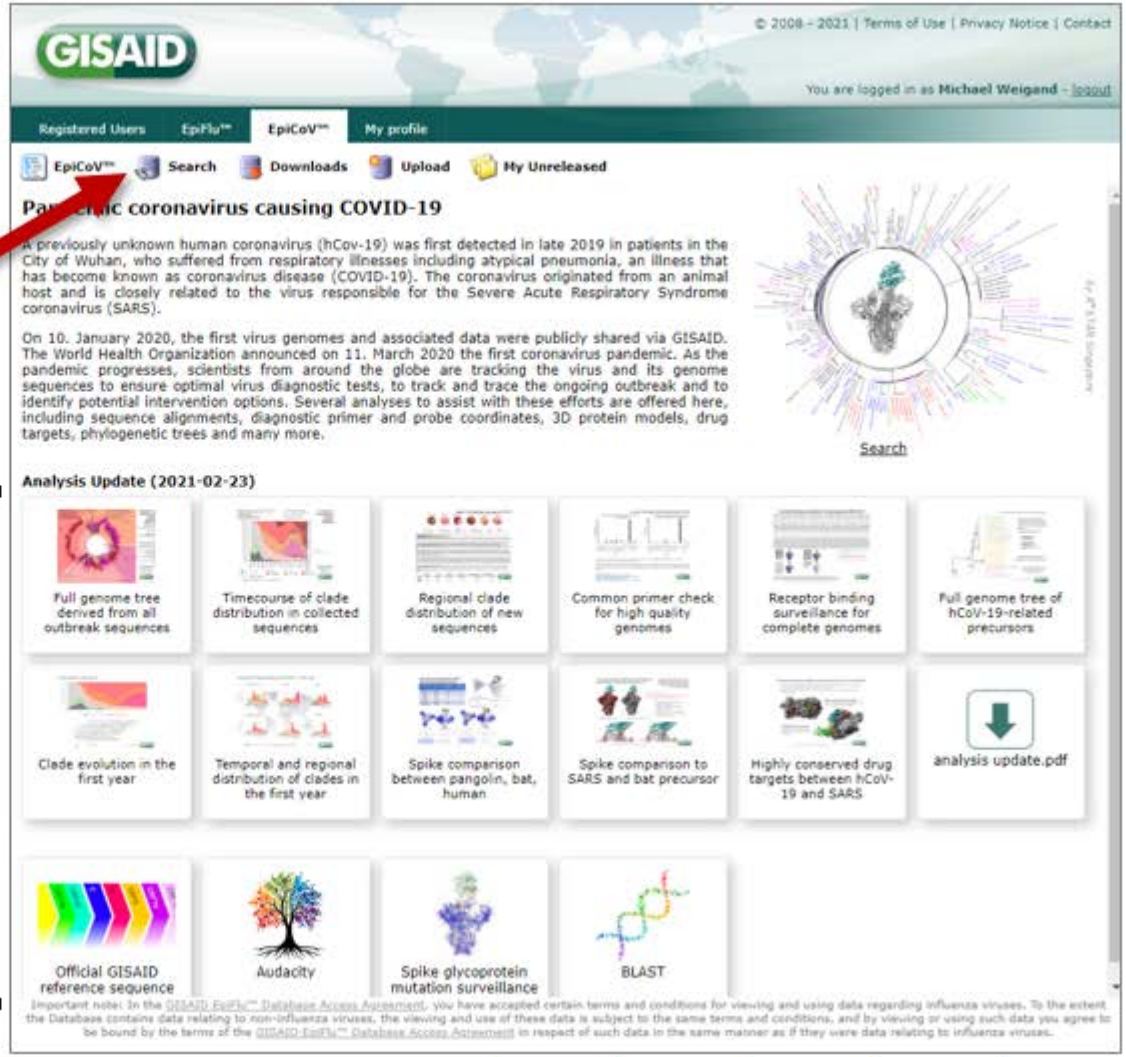
Phylogenetic Assignment of Named Global Outbreak



cov-lineages.org

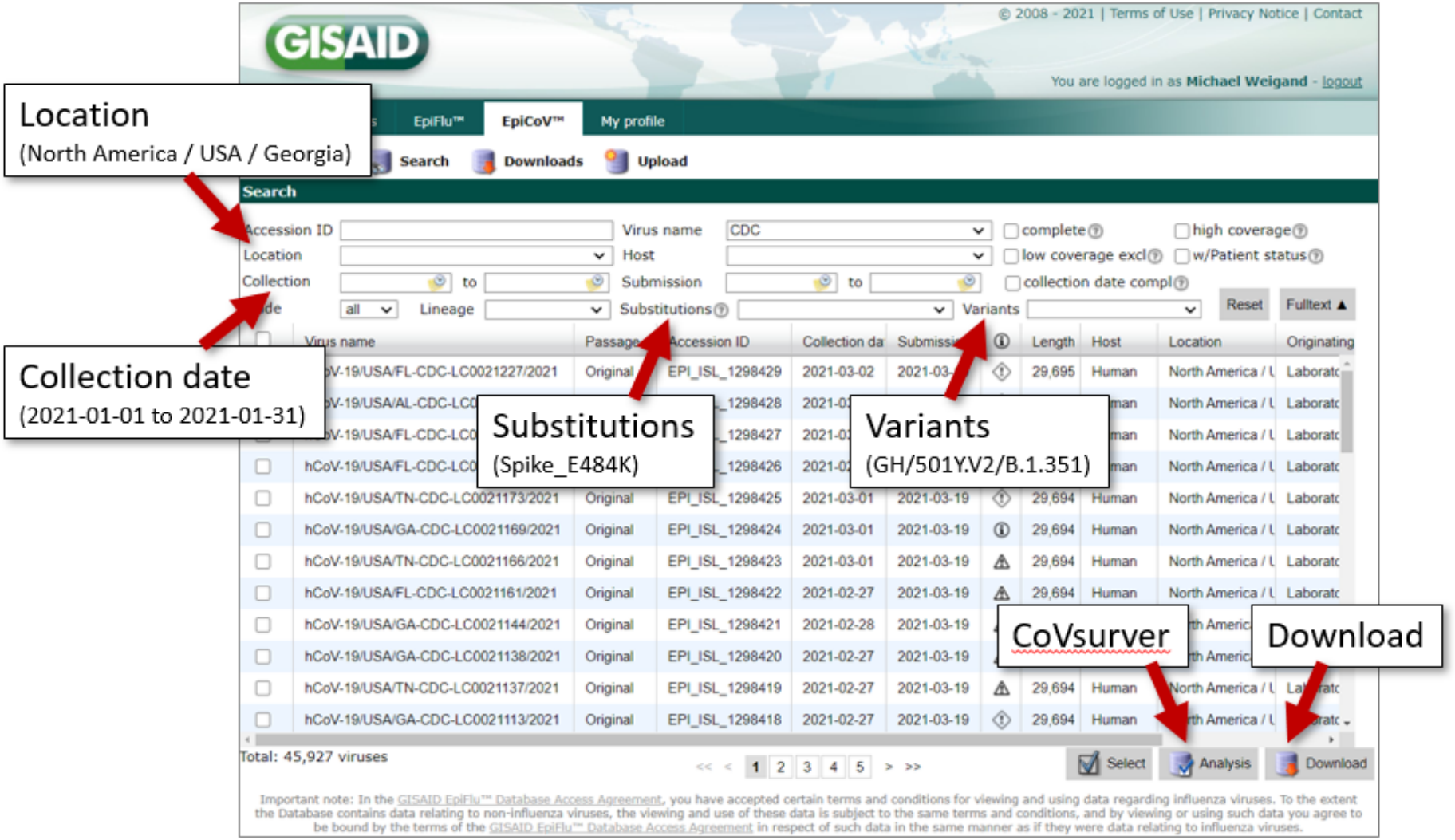
GISAID.org

Sequence search

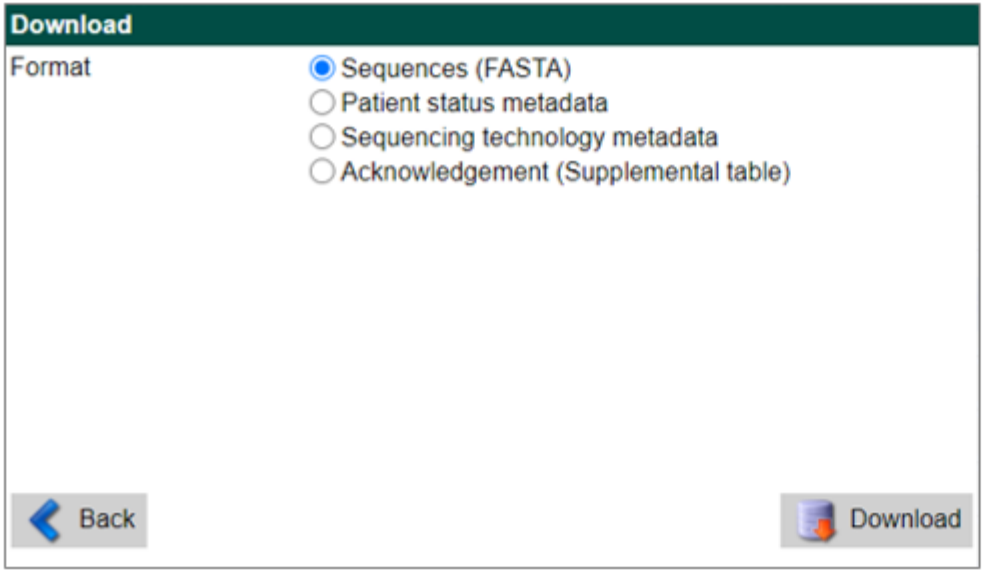


Analyses

- 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
GISAID
© 2008 - 2021 | Terms of Use | Privacy Notice | Contact
You are logged in as Michael Weigand - logout
EpiFlu™
EpiCoV™
My profile
Search
Downloads
Upload
Search
Accession ID
Virus name
CDC
complete
high coverage
Location
Host
low coverage excl
w/Patient status
Collection
to
Submission
to
collection date compl
Lineage
Substitutions
Variants
Reset
Fulltext
Virus name
Passage
Accession ID
Collection da
Submissi
Length
Host
Location
Originating
hCoV-19/USA/TN-CDC-LC0021173/2021
Original
EPI_ISL_1298425
2021-03-01
2021-03-19
29,694
Human
North America / U
hCoV-19/USA/GA-CDC-LC0021169/2021
Original
EPI_ISL_1298424
2021-03-01
2021-03-19
29,694
Human
North America / U
hCoV-19/USA/TN-CDC-LC0021166/2021
Original
EPI_ISL_1298423
2021-03-01
2021-03-19
29,694
Human
North America / U
hCoV-19/USA/FL-CDC-LC0021161/2021
Original
EPI_ISL_1298422
2021-02-27
2021-03-19
29,694
Human
North America / U
hCoV-19/USA/GA-CDC-LC0021144/2021
Original
EPI_ISL_1298421
2021-02-28
2021-03-19
hCoV-19/USA/GA-CDC-LC0021138/2021
Original
EPI_ISL_1298420
2021-02-27
2021-03-19
hCoV-19/USA/TN-CDC-LC0021137/2021
Original
EPI_ISL_1298419
2021-02-27
2021-03-19
29,694
Human
hCoV-19/USA/GA-CDC-LC0021113/2021
Original
EPI_ISL_1298418
2021-02-27
2021-03-19
29,694
Human
Total: 45,927 viruses
1 2 3 4 5
Select
Analysis
Download
Important note: In the GISAID EpiFlu™ 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 EpiFlu™ Database Access Agreement in respect of such data in the same manner as if they were data relating to influenza viruses.

GISAID: download



FASTA

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

Metadata.tsv

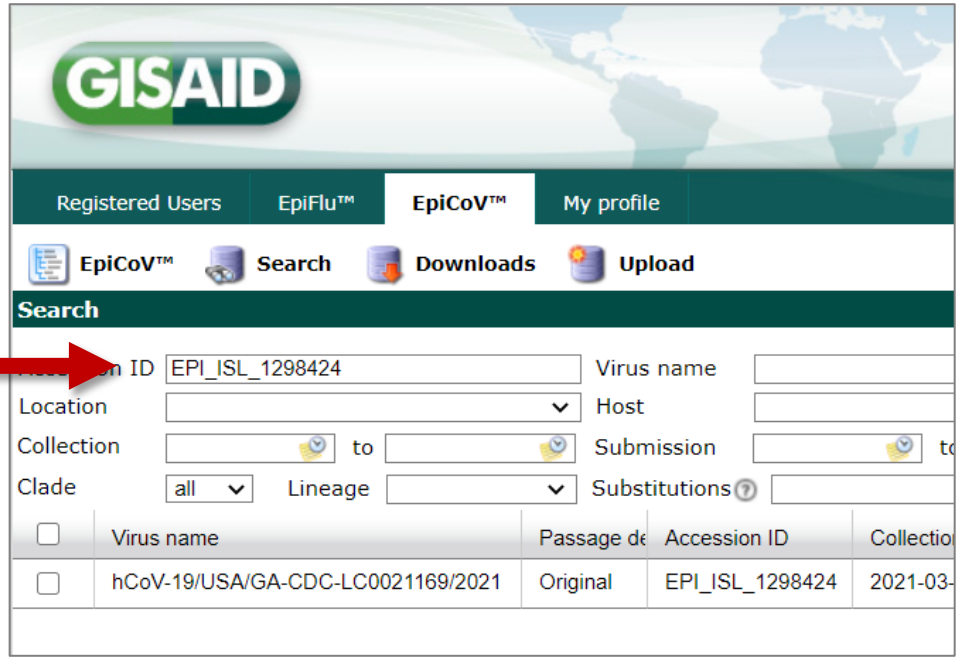
Accession	Date	Location
Seq1	02/05/2021	USA/GA
Seq2	02/05/2021	USA/GA
Seq3	02/05/2021	USA/GA

GISAID Database Access Agreement

www.gisaid.org/registration/terms-of-use/

GISAID: find individual records

Accession ID



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

Variant
Passage details/history:

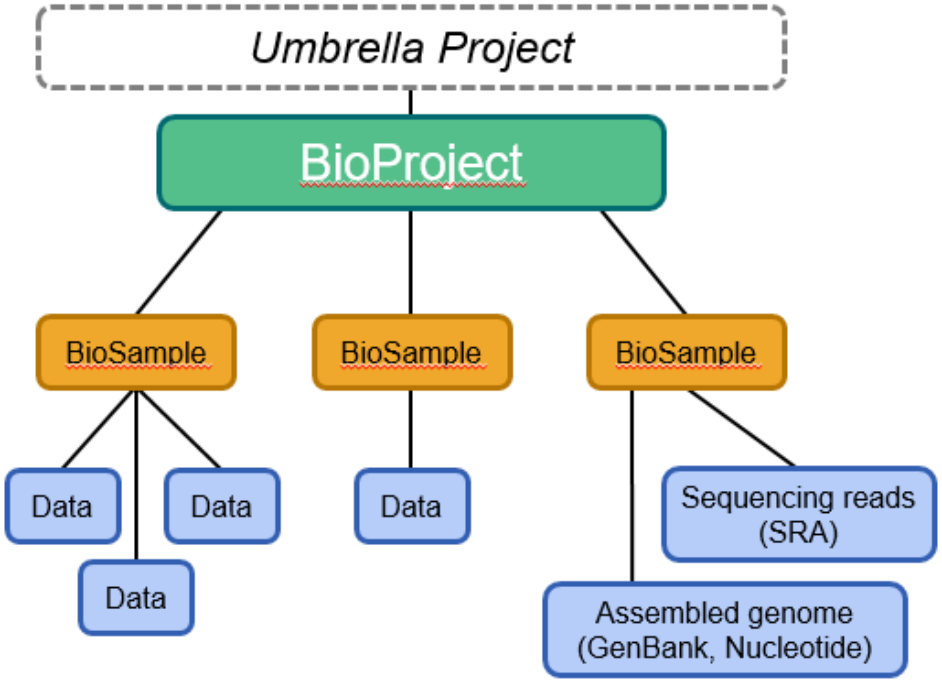
Sample information
Collection date:
Location:
Host:
Additional location information:
Gender:
Patient age:
Patient status:
Specimen source:
Additional host information:
Outbreak:
Last vaccinated:
Treatment:
Sequencing technology:
Assembly method:
Coverage:
Comment:

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

GISAID Database Access Agreement

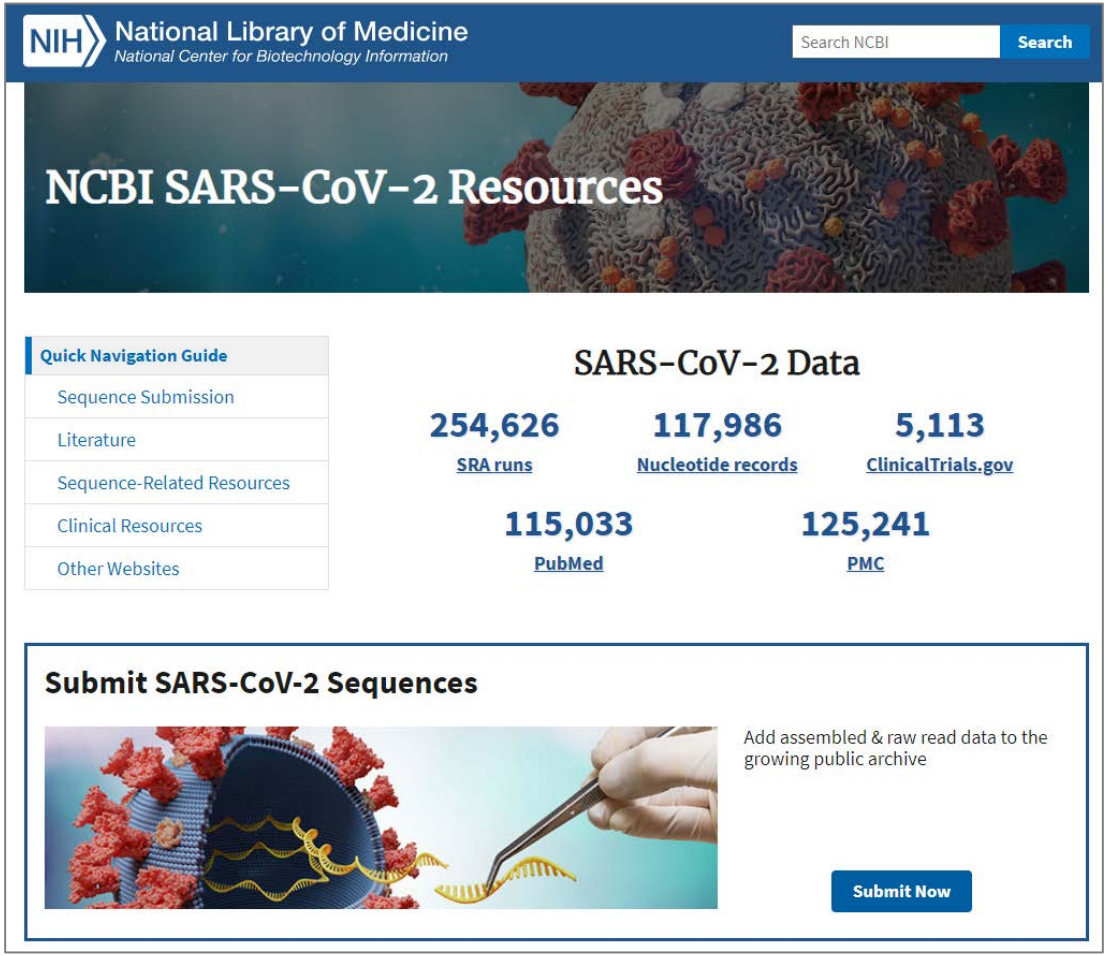
www.gisaid.org/registration/terms-of-use/

NCBI database organization



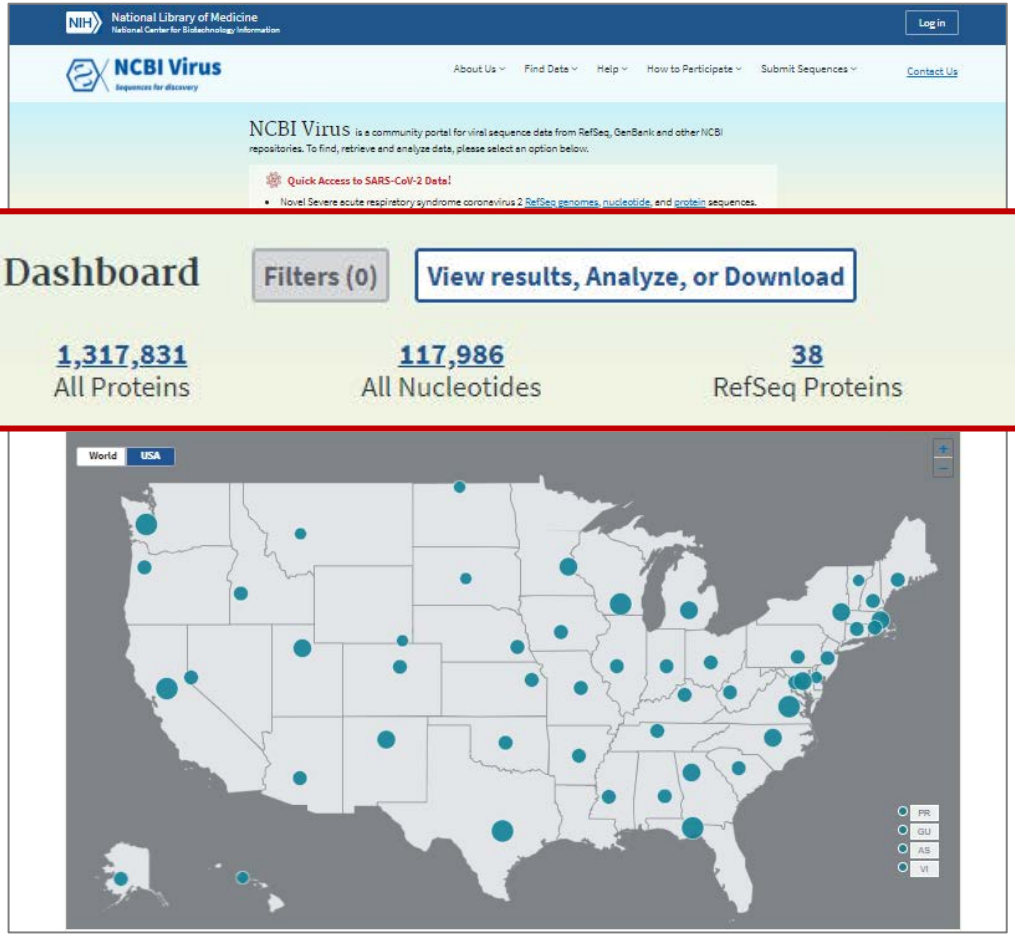
- BioProject – a collection of biological data for a single initiative
  - *Umbrella Project* – a collection of BioProjects
- BioSample – information about the physical specimen (metadata)
- Data – genomic datasets; e.g.
  - Raw sequencing reads
  - Assembled genome

www.ncbi.nlm.nih.gov/sars-cov-2/



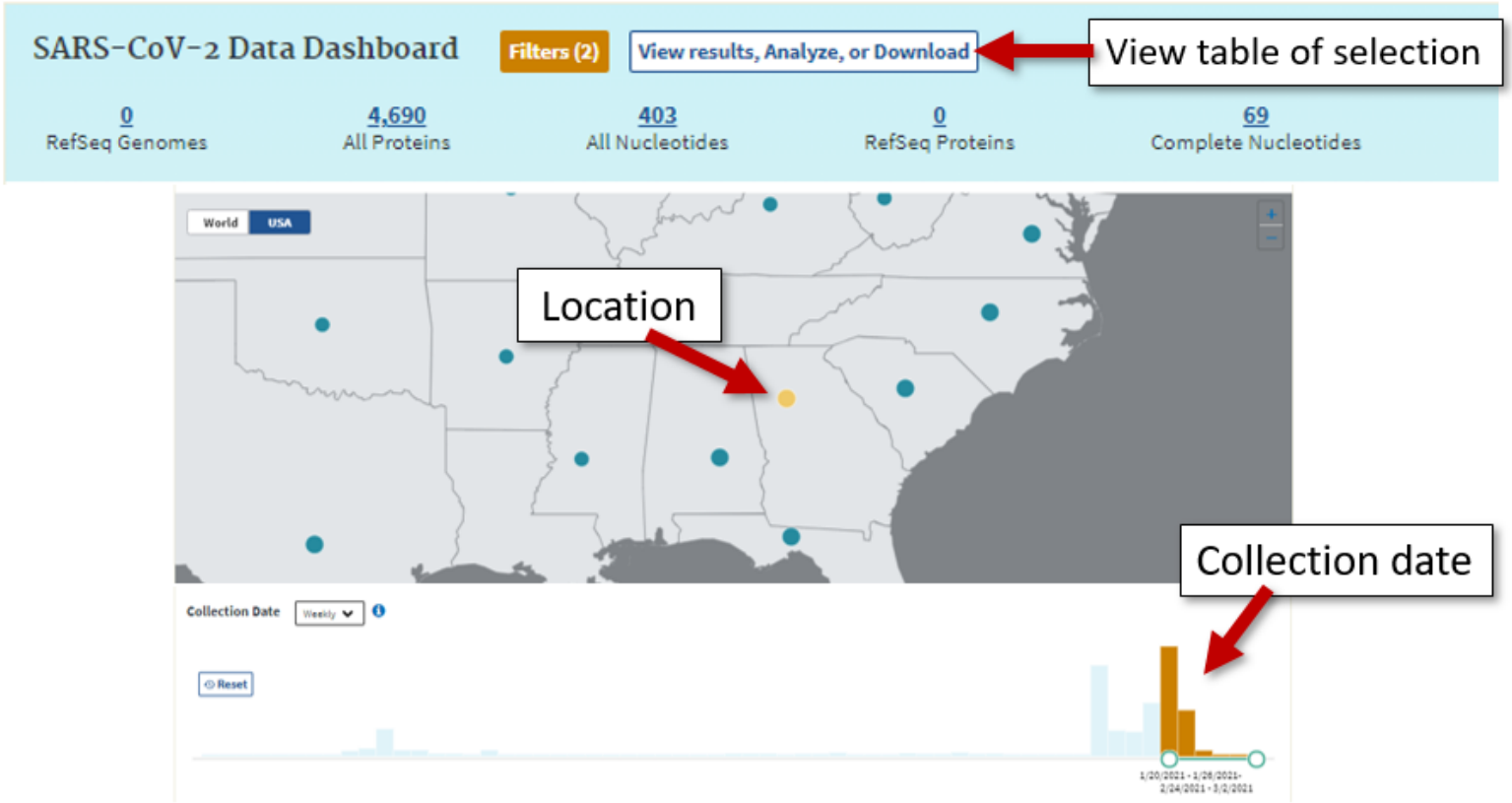
(3/22/2021)

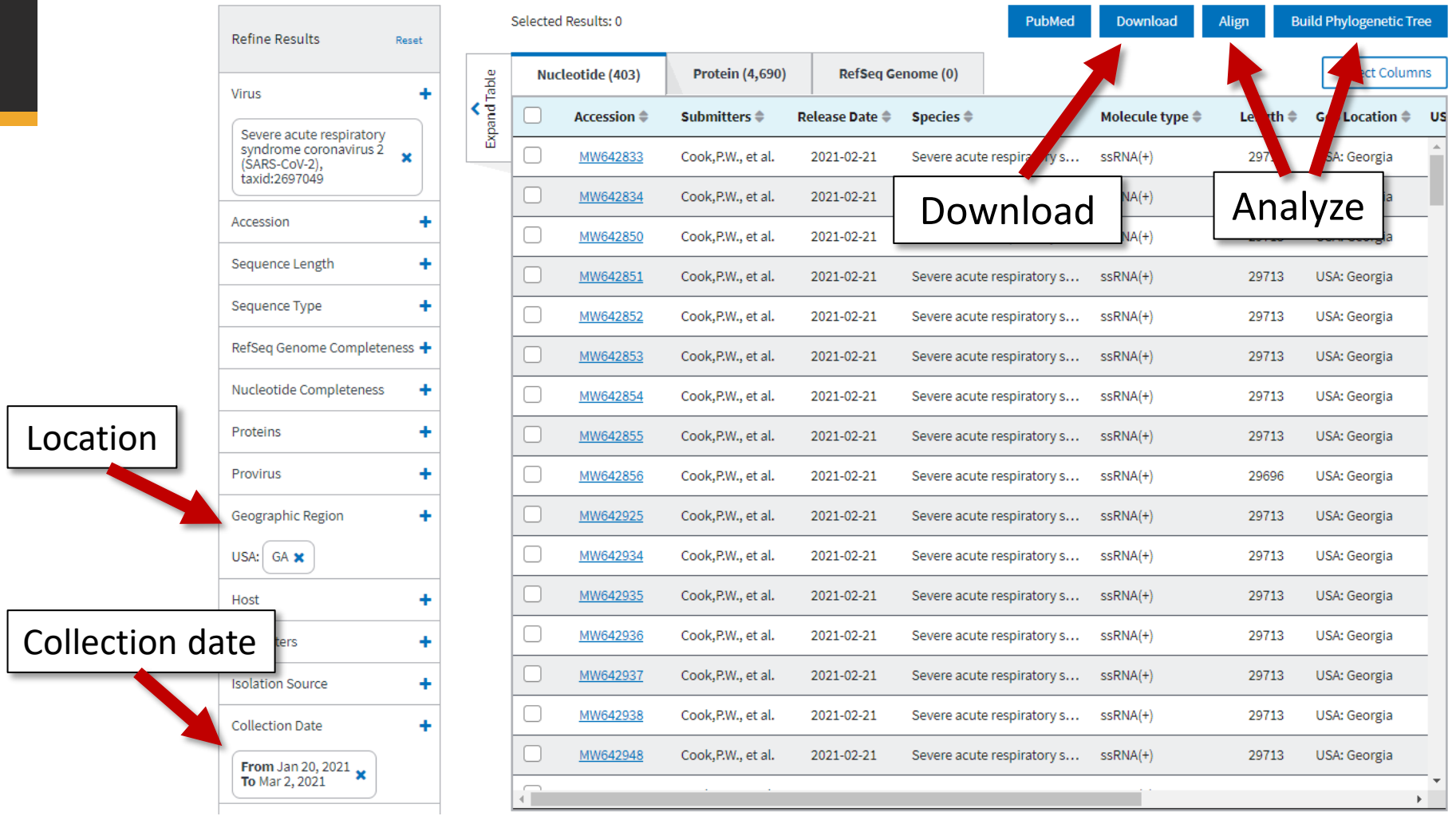
www.ncbi.nlm.nih.gov/labs/virus/vssi/#/sars-cov-2



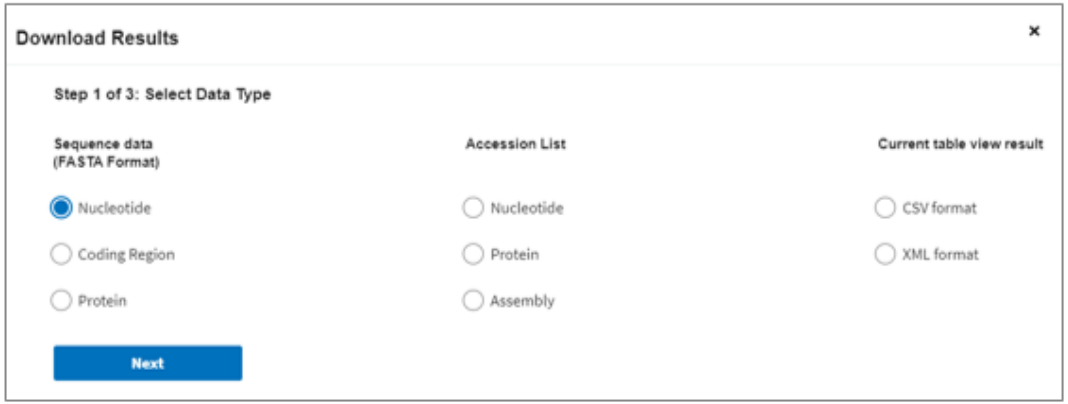
(3/22/2021)

www.ncbi.nlm.nih.gov/labs/virus/vssi/#/sars-cov-2




Refine Results
Reset
Virus
Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), taxid:2697049
Accession
Sequence Length
Sequence Type
RefSeq Genome Completeness
Nucleotide Completeness
Proteins
Provirus
Geographic Region
USA: GA
Host
Isolation Source
Collection Date
From Jan 20, 2021
To Mar 2, 2021
Location
Collection date
Selected Results: 0
PubMed
Download
Align
Build Phylogenetic Tree
Expand Table
Nucleotide (403)
Protein (4,690)
RefSeq Genome (0)
Accession
Submitters
Release Date
Species
Molecule type
Length
Location
MW642833 Cook,P.W., et al. 2021-02-21 ssRNA(+) USA: Georgia
MW642834 Cook,P.W., et al. 2021-02-21
MW642850 Cook,P.W., et al. 2021-02-21
MW642851 Cook,P.W., et al. 2021-02-21 Severe acute respiratory s... ssRNA(+) 29713 USA: Georgia
MW642852 Cook,P.W., et al. 2021-02-21 Severe acute respiratory s... ssRNA(+) 29713 USA: Georgia
MW642853 Cook,P.W., et al. 2021-02-21 Severe acute respiratory s... ssRNA(+) 29713 USA: Georgia
MW642854 Cook,P.W., et al. 2021-02-21 Severe acute respiratory s... ssRNA(+) 29713 USA: Georgia
MW642855 Cook,P.W., et al. 2021-02-21 Severe acute respiratory s... ssRNA(+) 29713 USA: Georgia
MW642856 Cook,P.W., et al. 2021-02-21 Severe acute respiratory s... ssRNA(+) 29696 USA: Georgia
MW642925 Cook,P.W., et al. 2021-02-21 Severe acute respiratory s... ssRNA(+) 29713 USA: Georgia
MW642934 Cook,P.W., et al. 2021-02-21 Severe acute respiratory s... ssRNA(+) 29713 USA: Georgia
MW642935 Cook,P.W., et al. 2021-02-21 Severe acute respiratory s... ssRNA(+) 29713 USA: Georgia
MW642936 Cook,P.W., et al. 2021-02-21 Severe acute respiratory s... ssRNA(+) 29713 USA: Georgia
MW642937 Cook,P.W., et al. 2021-02-21 Severe acute respiratory s... ssRNA(+) 29713 USA: Georgia
MW642938 Cook,P.W., et al. 2021-02-21 Severe acute respiratory s... ssRNA(+) 29713 USA: Georgia
MW642948 Cook,P.W., et al. 2021-02-21 Severe acute respiratory s... ssRNA(+) 29713 USA: Georgia
Download
Analyze

NCBI download



FASTA

```
>Sequence 1
AAAUGUUAUUCAUGCU
>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

Supplement local analyses

GISAID

NCBI

Local data

FASTA

```
>Sequence 1
AAAUGUUAUUCAUGCU
>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

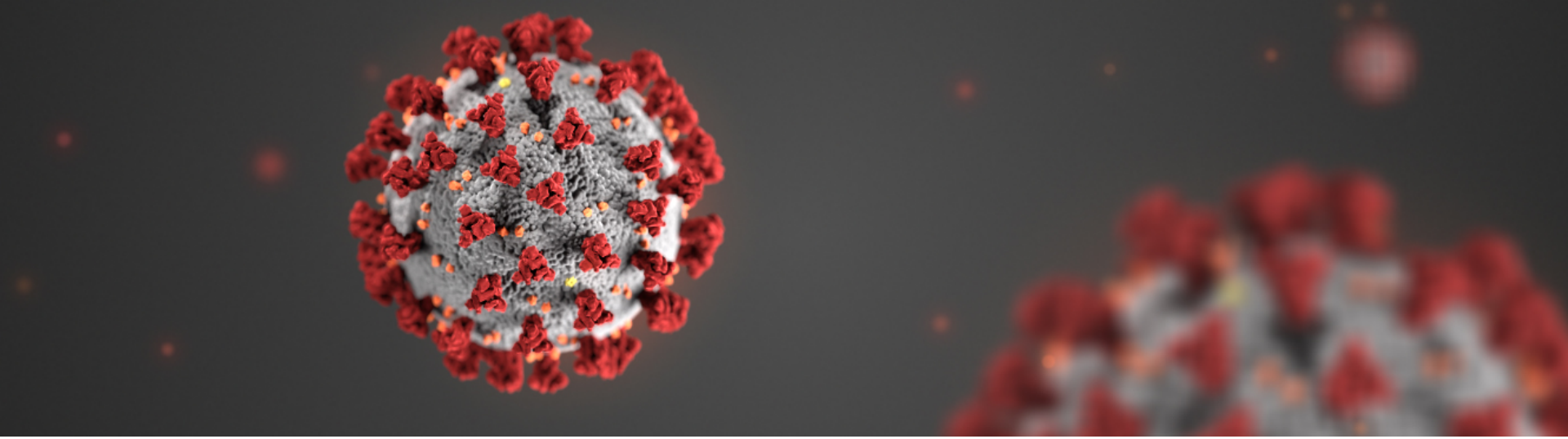
NIH National Library of Medicine
National Center for Biotechnology Information

- 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



For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 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
  - National Center for Biotechnology Information (NCBI). www.ncbi.nlm.nih.gov
  - NCBI SARS-CoV-2 resources. 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
  - NCBI Umbrella project of US sequencing efforts. 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-cov2-gisaid-submission-protocol-bh98j99w