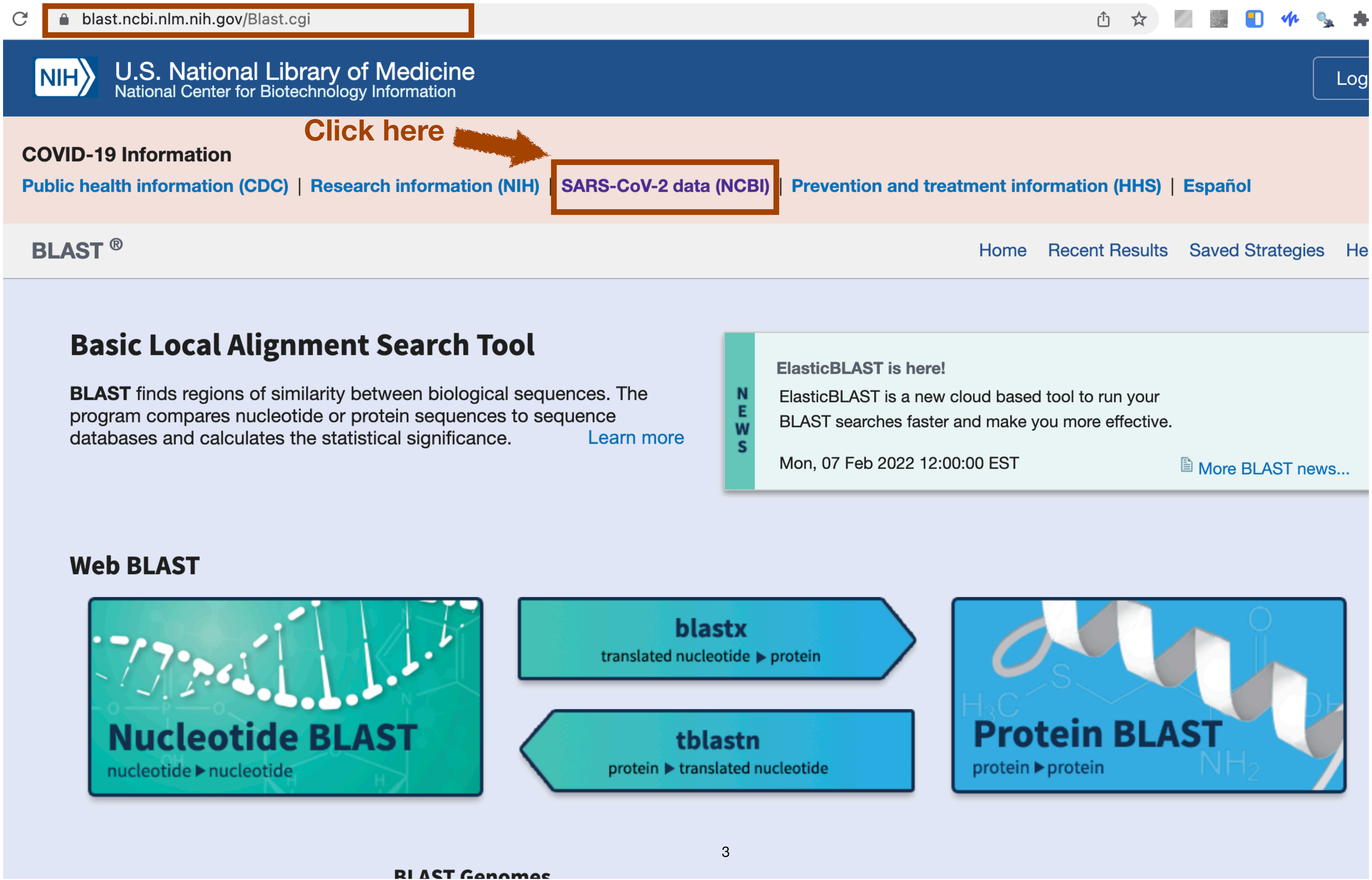
Workshop 1

Downloading SARS-COV2 genomic sequences from NCB

Natya Hans March 1st, 2022


blast.ncbi.nlm.nih.gov/Blast.cgi
NIH
U.S. National Library of Medicine
National Center for Biotechnology Information
Log
Click here
COVID-19 Information
Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español
BLAST ®
Home
Recent Results
Saved Strategies
He
Basic Local Alignment Search Tool
BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.
Learn more
NEWS
ElasticBLAST is here!
ElasticBLAST is a new cloud based tool to run your BLAST searches faster and make you more effective.
Mon, 07 Feb 2022 12:00:00 EST
More BLAST news...
Web BLAST
Nucleotide BLAST
nucleotide ▶ nucleotide
blastx
translated nucleotide ▶ protein
tblastn
protein ▶ translated nucleotide
Protein BLAST
protein ▶ protein
BLAST Genomes

NIH National Library of Medicine
National Center for Biotechnology Information

Search NCBI | Search

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Click here

Quick Navigation Guide

- Sequence Submission
- Literature
- Sequence-Related Resources
- Clinical Resources
- Other Websites

SARS-CoV-2 Data

3,099,725 SRA runs

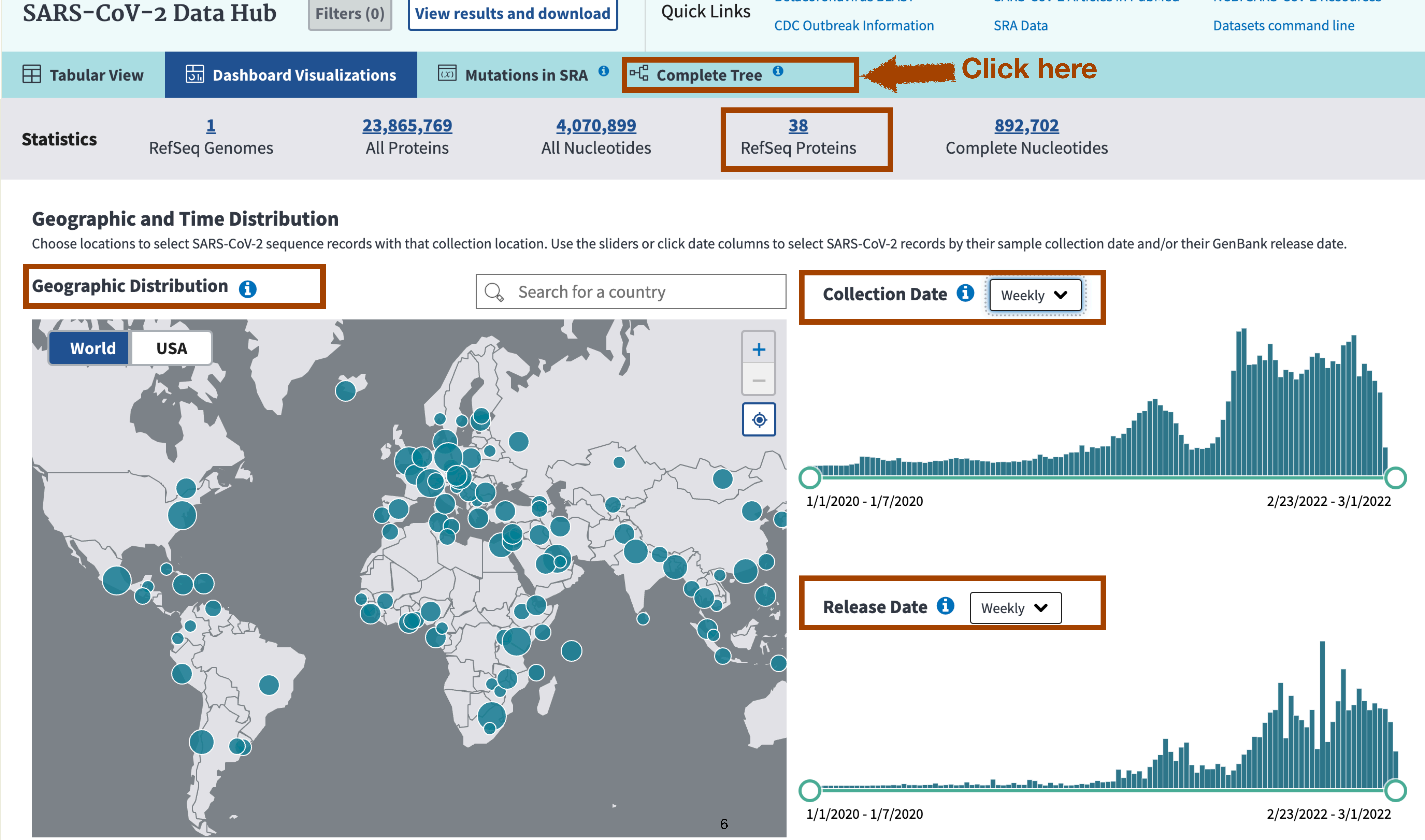
$\boxed{\textbf{4,070,899} \text{ Nucleotide records}}$

3,215 ClinicalTrials.gov

231,749 PubMed

288,388 PMC


NCBI Virus
Sequences for discovery
About Us
Find Data
Help
How to Participate
Submit Sequences
Contact Us
SARS-CoV-2 Data Hub
Download
Click here
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
Complete Tree
Selected Results: 0
Align
Build Phylogenetic Tree
Refine Results
Reset
Virus
Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), taxid:2697049
Accession
Sequence Length
Expand Table
Nucleotide (4,083,187)
Protein (24,012,480)
RefSeq Genome (1)
Select Columns
Accession
Submitters
Release Date
Pangolin
Isolate
Species
NC_045512
RefSeq
Wu,F., et al.
2020-01-13
B
Wuhan-Hu-1
Severe acute respirat
OM840138
Andrews,K....
2022-02-27
B.1.2
ID-UI-IIDS-U0847
Severe acute respirat
OM840139
Andrews,K....
2022-02-27
B.1.2
ID-UI-IIDS-U0850
Severe acute respirat
OM840140
Andrews,K....
2022-02-27
B.1
ID-UI-IIDS-U0852
Severe acute respirat


SARS-CoV-2 Data Hub
Filters (0)
View results and download
Quick Links
CDC Outbreak Information
SRA Data
Datasets command line
Tabular View
Dashboard Visualizations
Mutations in SRA
Complete Tree
Click here
Statistics
1
RefSeq Genomes
23,865,769
All Proteins
4,070,899
All Nucleotides
38
RefSeq Proteins
892,702
Complete Nucleotides
Geographic and Time Distribution
Choose locations to select SARS-CoV-2 sequence records with that collection location. Use the sliders or click date columns to select SARS-CoV-2 records by their sample collection date and/or their GenBank release date.
Geographic Distribution
Search for a country
World
USA
Collection Date
Weekly
1/1/2020 - 1/7/2020
2/23/2022 - 3/1/2022
Release Date
Weekly
1/1/2020 - 1/7/2020
2/23/2022 - 3/1/2022

NCBI Virus

Sequences for discovery

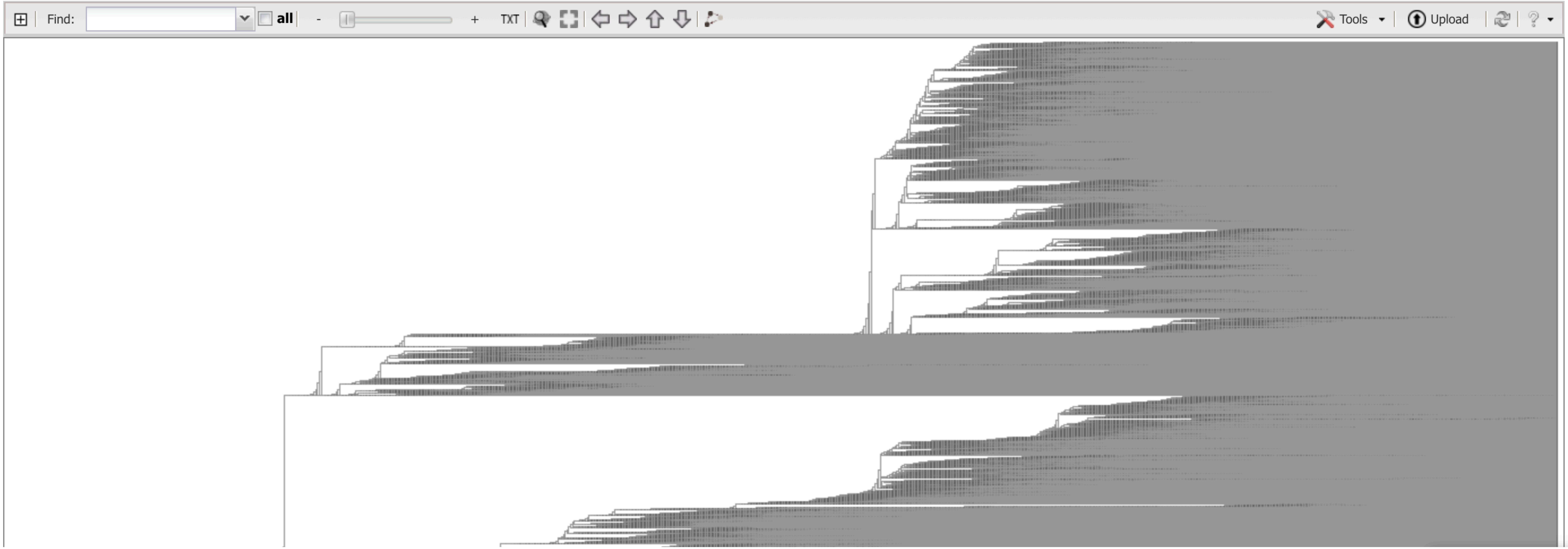
Click here

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

Tree of complete SARS-CoV-2 Sequences

Algorithm and parameters

Updated: Fri Feb 25 2022



Tree of complete SARS-CoV-2 Sequences

Algorithm and parameters

Our goal is to show the biological diversity of complete and nearly-complete natural SARS-CoV-2 sequences. All SARS-CoV-2 sequences which meet the following requirements are included:

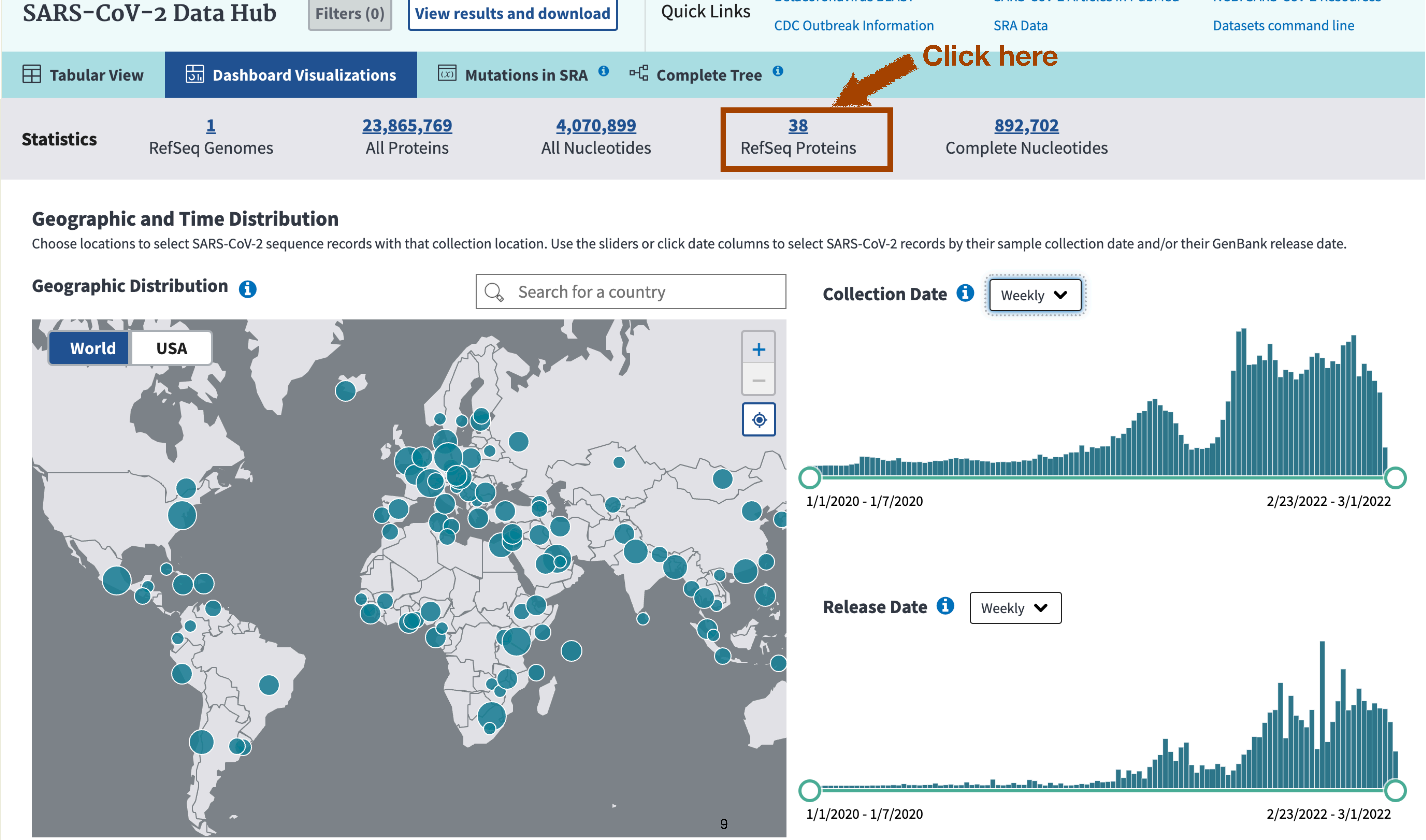
- Sequence length between 29600 and 31000 bp after trimming ends which do not align with RefSeq
- Must have collection date (year + month)
- The number of ambiguous nucleotides must be < 1% in the trimmed sequences

Each sequence is aligned to the RefSeq (NC_045512), and mutations such as SNPs, indels, and long substitutions are calculated.

Distances are computed between pairs of sequences as the number of different mutations, where non-ambiguous different mutations are counted as 1, and ambiguous vs. non-ambiguous different mutations are counted 0.1 (one sequence has a non-ambiguous mutation and another sequence has an ambiguous mutation in the same place).

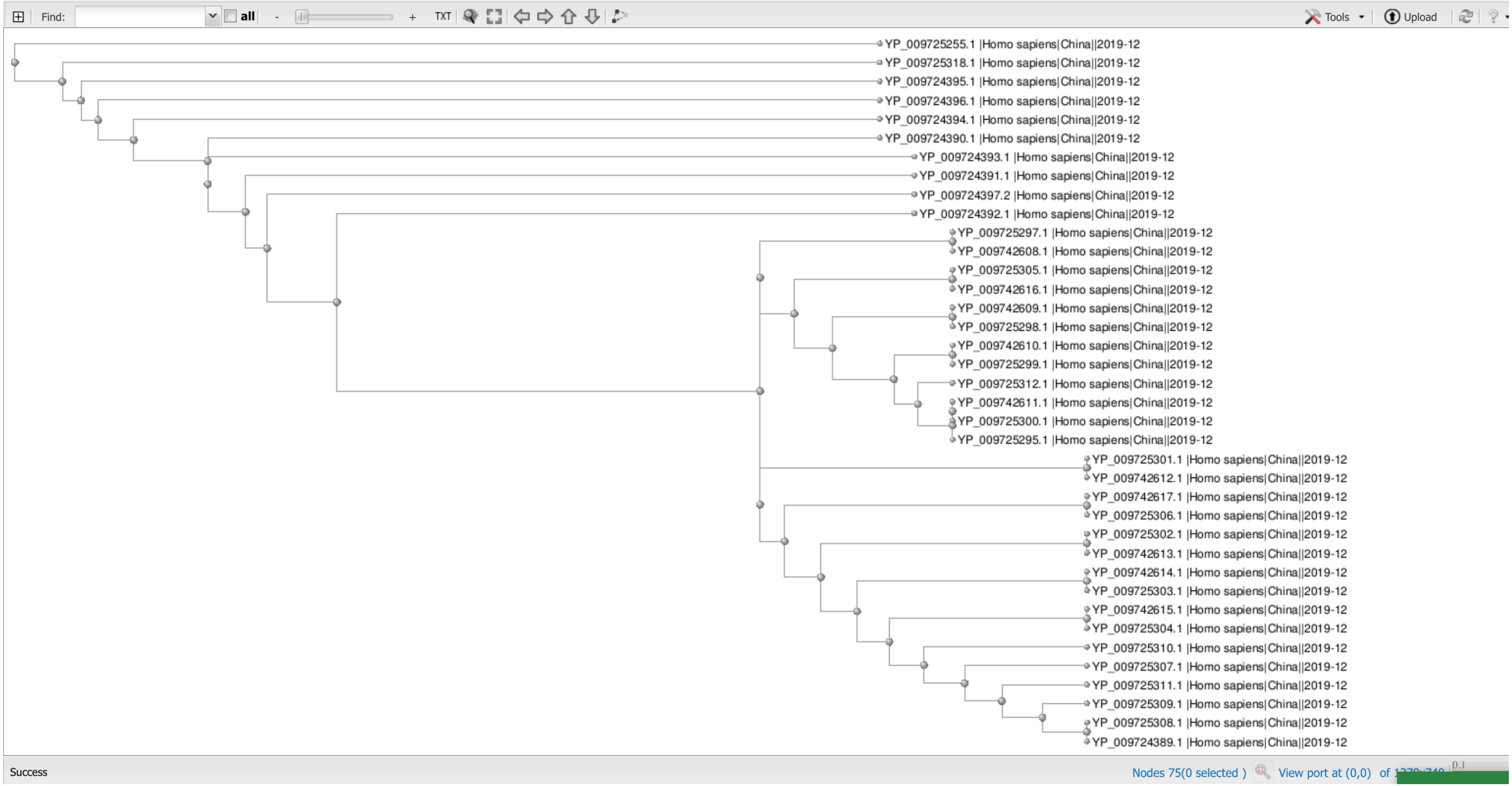
Any term in the leaf labels can be searched using the "**Find**" option.
The leaf labels are in this format: **accession.version|host|isolation source|geographic location|collection date|GenBank release date**,
for example "**MT370839.1|Homo sapiens|oronasopharynx|USA|2020-03-14|2020-04-23**".

Updated: Fri Feb 25 2022


SARS-CoV-2 Data Hub
Filters (0)
View results and download
Quick Links
CDC Outbreak Information
SRA Data
Datasets command line
Click here
Tabular View
Dashboard Visualizations
Mutations in SRA
Complete Tree
Statistics
1
RefSeq Genomes
23,865,769
All Proteins
4,070,899
All Nucleotides
38
RefSeq Proteins
892,702
Complete Nucleotides
Geographic and Time Distribution
Choose locations to select SARS-CoV-2 sequence records with that collection location. Use the sliders or click date columns to select SARS-CoV-2 records by their sample collection date and/or their GenBank release date.
Geographic Distribution
Search for a country
World
USA
Collection Date
Weekly
1/1/2020 - 1/7/2020
2/23/2022 - 3/1/2022
Release Date
Weekly
1/1/2020 - 1/7/2020
2/23/2022 - 3/1/2022

Then click here
SARS-CoV-2 Data Hub
Download
Quick Links
Betacoronavirus BLAST
CDC Outbreak Information
SARS-CoV-2 Articles in PubMed
SRA Data
SARS-CoV-2 Resources
Tabular View
Dashboard Visualizations
Mutations in SRA
Complete Tree
Selected Results: 38
Align
Build Phylogenetic Tree
Click here first
Refine Results
Reset
Virus
Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), taxid:2697049
Accession
Sequence Length
Ambiguous Characters
Sequence Type
RefSeq
RefSeq Genome Completeness
Nucleotide Completeness
Pango lineage
Random Sampling New!
Expand Table
Nucleotide (1)
Protein (38)
RefSeq Genome (1)
Select Columns
Accession Release Date Pangolin Isolate Species Length Protein Geo Location
YP_009742608 RefSeq 2020-03-30 B Wuhan-Hu-1 Severe acute respiratory s... 180 leader protein China
YP_009742609 RefSeq 2020-03-30 B Wuhan-Hu-1 Severe acute respiratory s... 638 nsp2 China
YP_009742610 RefSeq 2020-03-30 B Wuhan-Hu-1 Severe acute respiratory s... 1945 nsp3 China
YP_009742611 RefSeq 2020-03-30 B Wuhan-Hu-1 Severe acute respiratory s... 500 nsp4 China
YP_009742612 RefSeq 2020-03-30 B Wuhan-Hu-1 Severe acute respiratory s... 306 3C-like proteinase China
YP_009742613 RefSeq 2020-03-30 B Wuhan-Hu-1 Severe acute respiratory s... 290 nsp6 China
YP_009742614 RefSeq 2020-03-30 B Wuhan-Hu-1 Severe acute respiratory s... 83 nsp7 China
YP_009742615 RefSeq 2020-03-30 B Wuhan-Hu-1 Severe acute respiratory s... 198 nsp8 China
YP_009742616 RefSeq 2020-03-30 B Wuhan-Hu-1 Severe acute respiratory s... 113 nsp9 China
YP_009742617 RefSeq 2020-03-30 B Wuhan-Hu-1 Severe acute respiratory s... 139 nsp10 China
Feedback

Phylogenetic Tree




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
Complete Tree
Selected Results: 0
Align
Build Phylogenetic Tree
Click here
Refine Results
Reset
Virus
Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), taxid:2697049
Accession
Sequence Length
Expand Table
Nucleotide (4,083,187)
Protein (24,012,480)
RefSeq Genome (1)
Select Columns
Accession
Submitters
Release Date
Pangolin
Isolate
Species
NC_045512 RefSeq
Wu,F., et al.
2020-01-13
B
Wuhan-Hu-1
Severe acute respirat
OM840138
Andrews,K....
2022-02-27
B.1.2
ID-UI-IIDS-U0847
Severe acute respirat
OM840139
Andrews,K....
2022-02-27
B.1.2
ID-UI-IIDS-U0850
Severe acute respirat
OM840140
Andrews,K....
2022-02-27
B.1
ID-UI-IIDS-U0852
Severe acute respirat

Refine Results Reset

Virus +

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

Accession +

Sequence Length +

Ambiguous Characters +

Sequence Type +

RefSeq Genome Completeness +

Nucleotide Completeness +

Pango lineage +

Random Sampling New! +

Isolate +

Proteins +

Provirus +

Geographic Region +

Host +

Submitters +

Isolation Source +

Collection Date +

Release Date +

Environmental Source +

Lab Host +

Vaccine Strain +

Nucleotide Completeness −

- complete (899,433)
- partial (3,183,754)

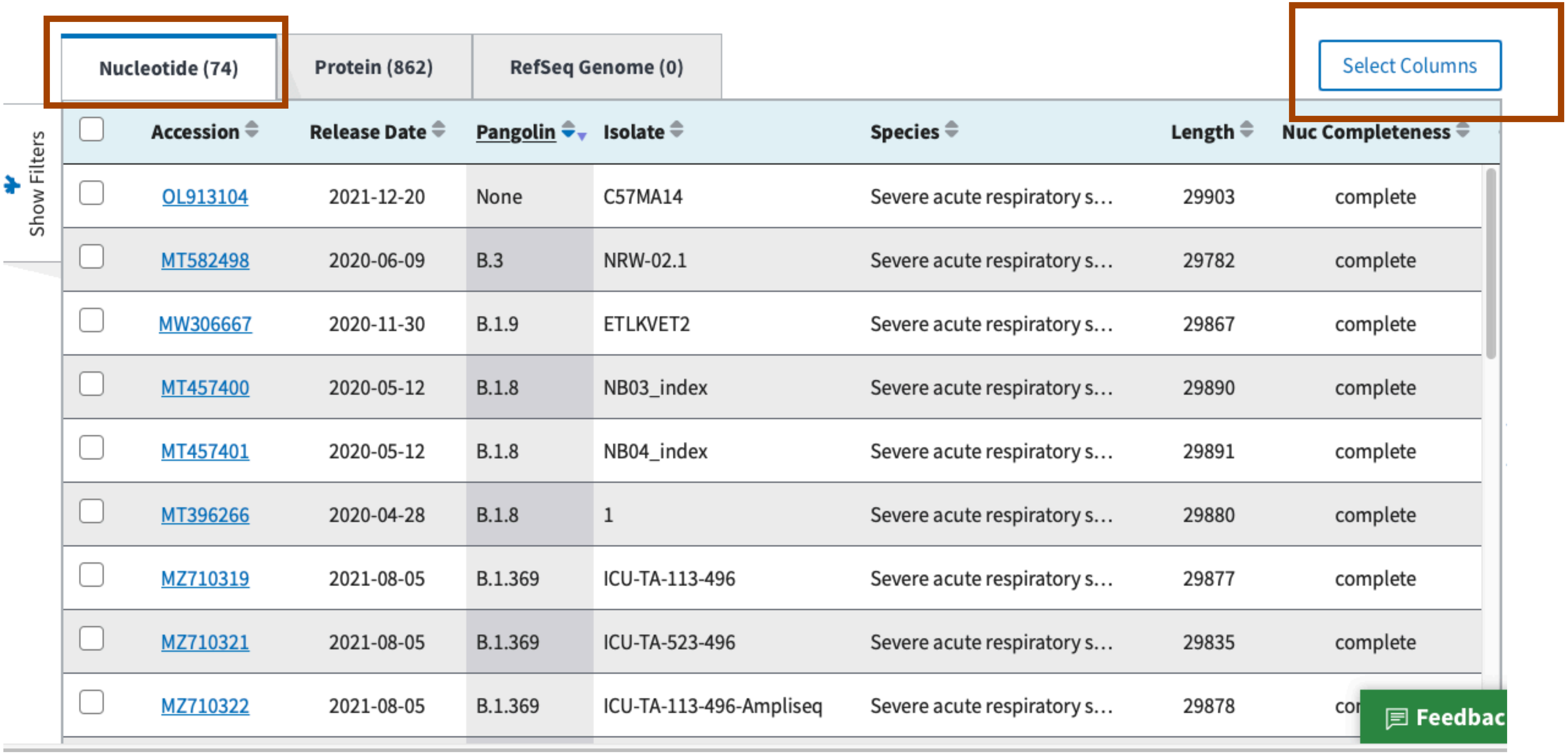
Random Sampling New! −

Filters sequences that were collected for a purpose of baseline surveillance.

- Include (899,433)
- Exclude (403,540)
- Only (495,893)

Isolation Source −

- feces (26)
- lung (74)
- lung, oronasopharynx (103)
- oronasopharynx (45,206)
- oronasopharynx, oronasopharynx (10)
- placenta (3)
- saliva, oronasopharynx (2,613)



Nucleotide (74) | Protein (862) | RefSeq Genome (0)

Select Columns

Show Filters

	Accession	Release Date	Pangolin	Isolate	Species	Length	Nuc Completeness
	OL913104	2021-12-20	None	C57MA14	Severe acute respiratory s...	29903	complete
	MT582498	2020-06-09	B.3	NRW-02.1	Severe acute respiratory s...	29782	complete
	MW306667	2020-11-30	B.1.9	ETLKVET2	Severe acute respiratory s...	29867	complete
	MT457400	2020-05-12	B.1.8	NB03_index	Severe acute respiratory s...	29890	complete
	MT457401	2020-05-12	B.1.8	NB04_index	Severe acute respiratory s...	29891	complete
	MT396266	2020-04-28	B.1.8	1	Severe acute respiratory s...	29880	complete
	MZ710319	2021-08-05	B.1.369	ICU-TA-113-496	Severe acute respiratory s...	29877	complete
	MZ710321	2021-08-05	B.1.369	ICU-TA-523-496	Severe acute respiratory s...	29835	complete
	MZ710322	2021-08-05	B.1.369	ICU-TA-113-496-Ampliseq	Severe acute respiratory s...	29878	co

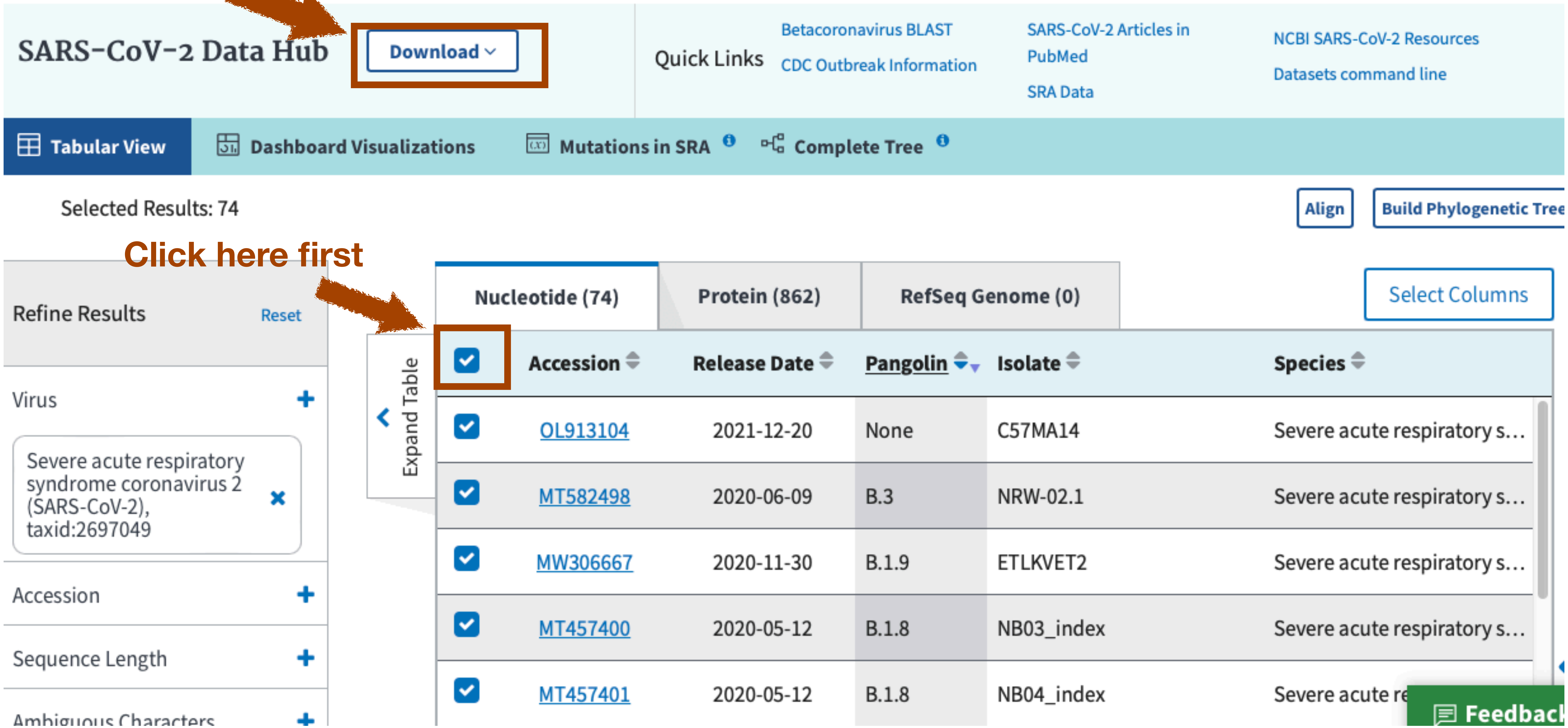
Select Columns

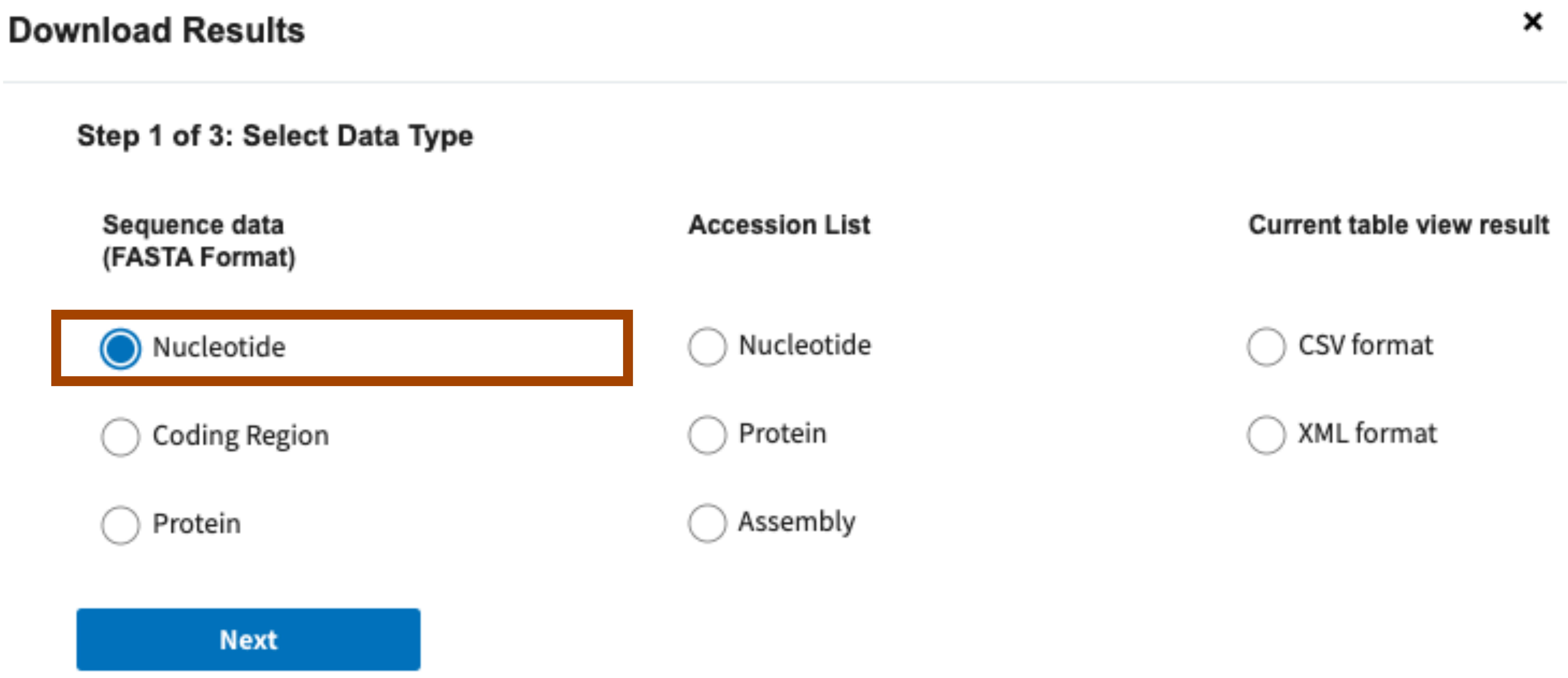
Add or Remove Columns From the Results Table

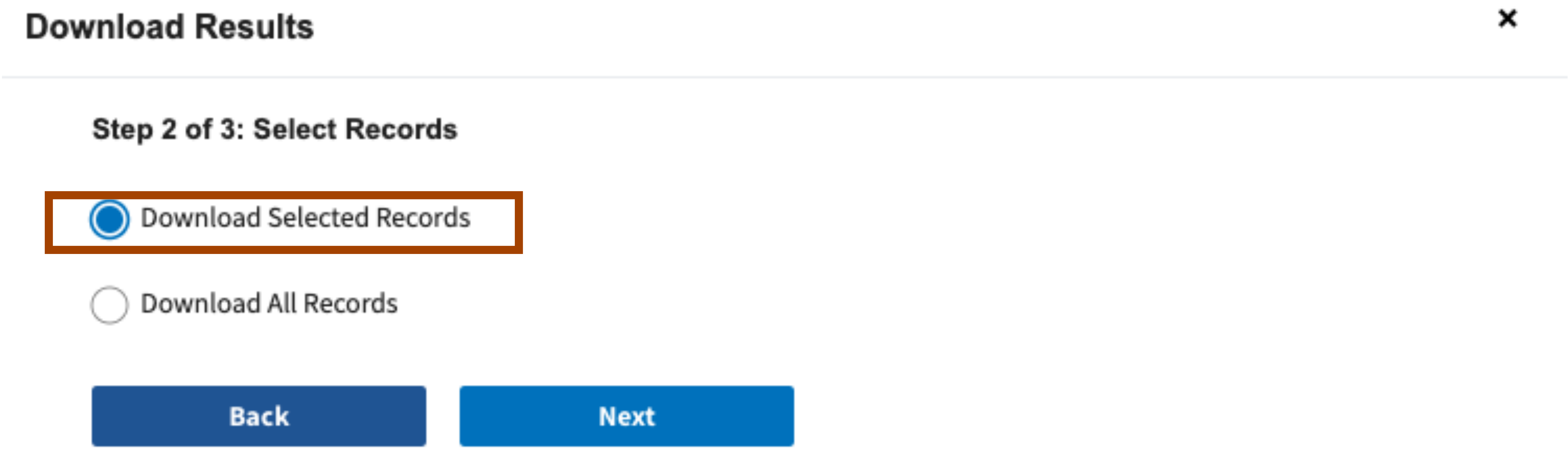
Columns to add:	**Displayed columns** (click to remove from the view):
+ SRA Accession	✖ Accession
+ Submitters	✖ Release Date
+ Genus	✖ Pangolin
+ Family	✖ Isolate New!
+ Molecule type	✖ Species
+ Sequence Type	✖ Length
+ Genotype	✖ Nuc Completeness
+ Segment	✖ Geo Location
+ Publications	✖ USA
+ Country New!	✖ Host
+ BioSample	✖ Isolation Source

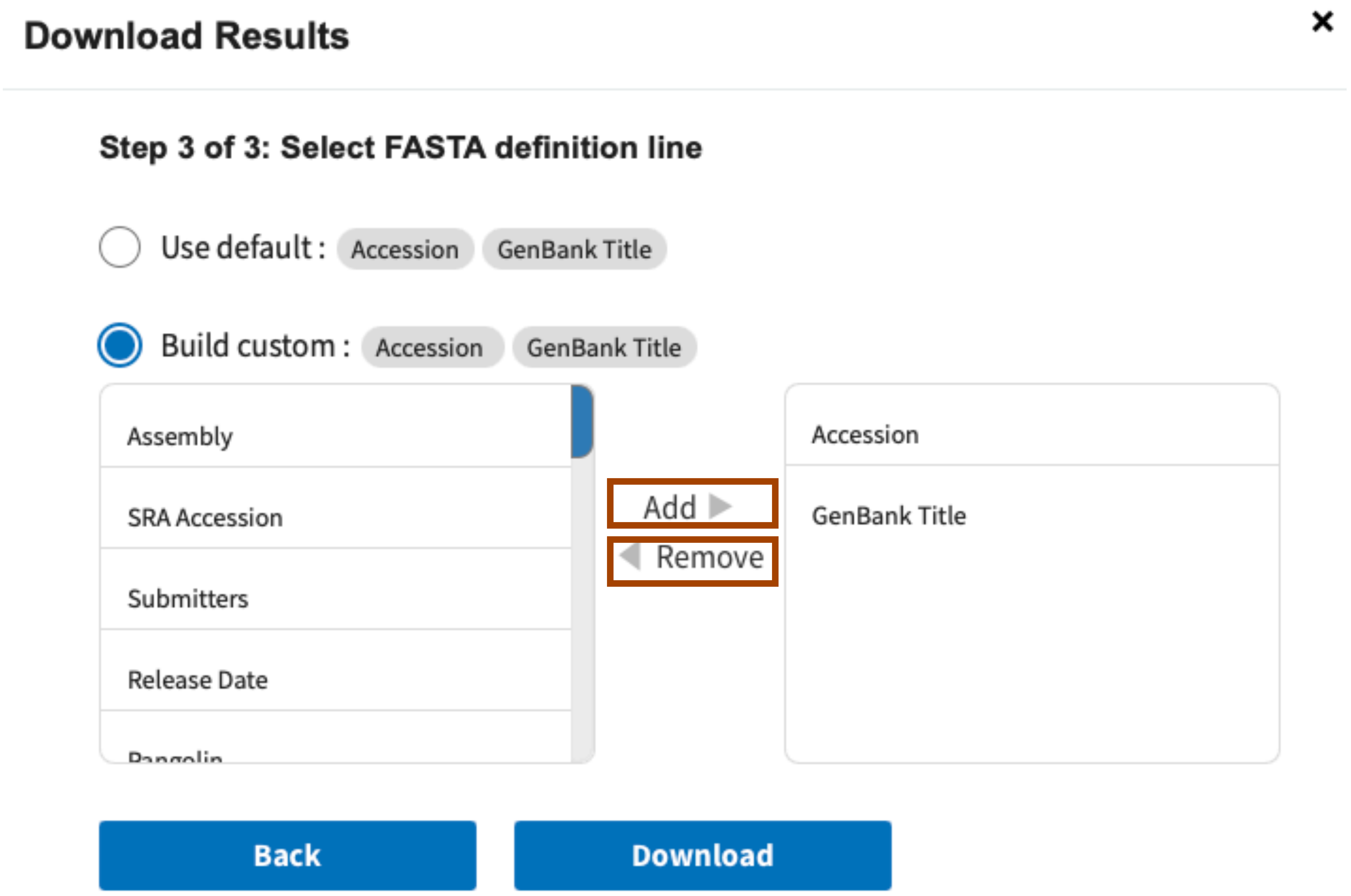
Then click here

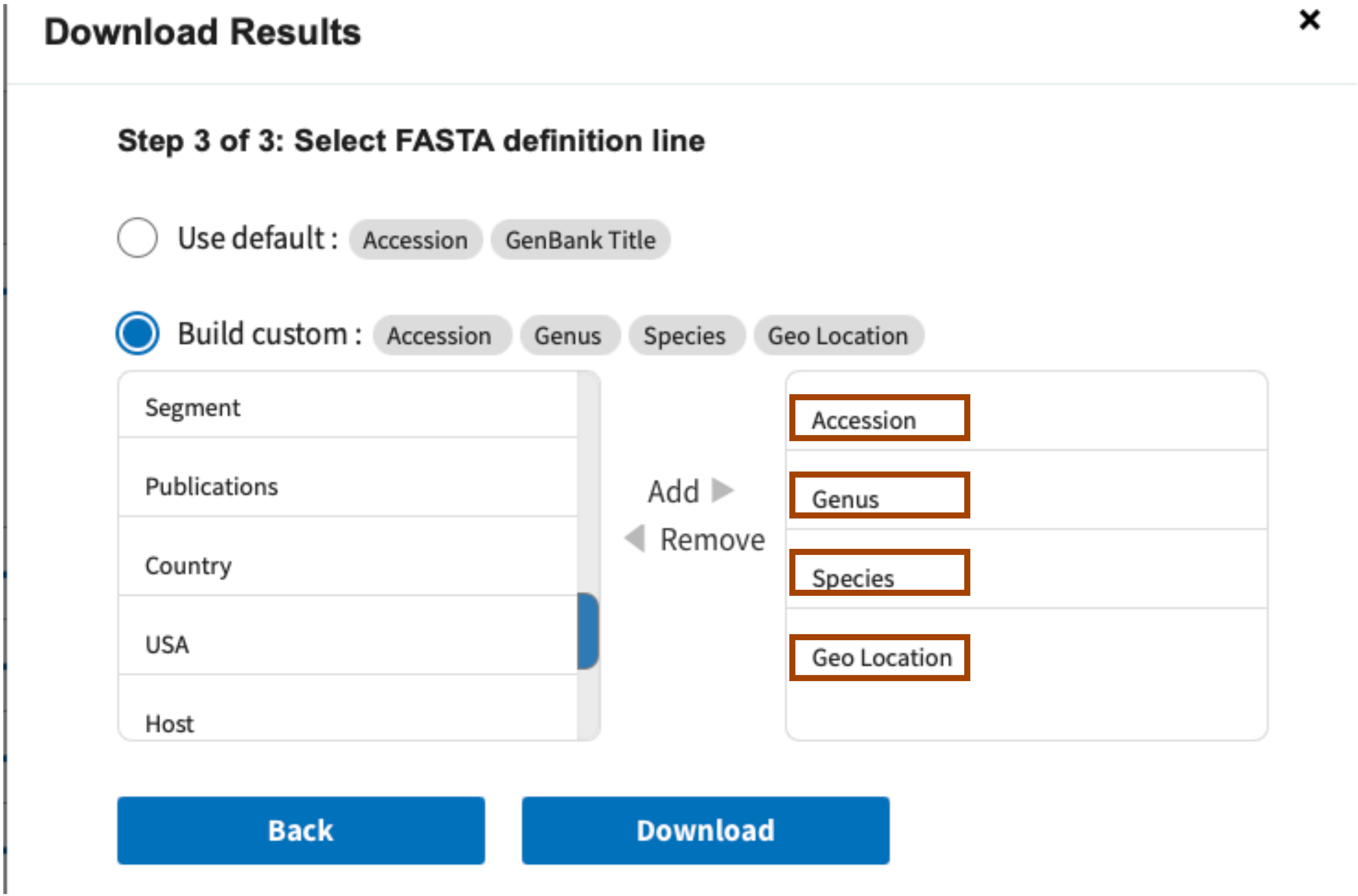
Click here first




Download Results
×
Step 1 of 3: Select Data Type
Sequence data (FASTA Format)
Accession List
Current table view result
Nucleotide
Nucleotide
CSV format
Coding Region
Protein
XML format
Protein
Assembly
Next


Download Results
×
Step 2 of 3: Select Records
Download Selected Records
Download All Records
Back
Next


Download Results
Step 3 of 3: Select FASTA definition line
Use default : Accession GenBank Title
Build custom : Accession GenBank Title
Assembly
SRA Accession
Submitters
Release Date
Add
Remove
Accession
GenBank Title
Back
Download


Download Results
Step 3 of 3: Select FASTA definition line
Use default : Accession GenBank Title
Build custom : Accession Genus Species Geo Location
Segment
Publications
Country
USA
Host
Add
Remove
Accession
Genus
Species
Geo Location
Back
Download

Downloads — vi sequences-2.fasta — 119×31

```
>OM062573.1 |Betacoronavirus|Severe acute respiratory syndrome-related coronavirus|China: Beijing
ATCAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCT
GTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACT
CACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCTTCTATC
TTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTT
TGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAAC
ACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGG
AGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAA
ACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACT
CGAAGGCATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGG
CGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGG
```