



Sep 03, 2021

# SRA and Genbank BioSample-Linked Submission with Mercury\_Prep and Mercury\_Batch

Francis J Ambrosio<sup>1</sup><sup>1</sup>Theiagen Genomics

1 Works for me

 Share

This protocol is published without a DOI.

Theiagen

Francis Ambrosio

## ABSTRACT

Submitting sequencing data to public data repositories is a meaningful yet tedious procedure. Linking submissions between SRA and Genbank will enhance the value of both submissions to the public health community. The Mercury protocols offered by Theiagen Genomics allows users to efficiently and accurately produce all required inputs for SRA and Genbank submissions (the Mercury workflows also allow for GISAID submission, but that will not be covered in this protocol). This protocol provides a detailed procedure for submitting BioSample-linked sequencing data to SRA and Genbank.

## PROTOCOL CITATION

Francis J Ambrosio 2021. SRA and Genbank BioSample-Linked Submission with Mercury\_Prep and Mercury\_Batch. [protocols.io](#)  
<https://protocols.io/view/sra-and-genbank-biosample-linked-submission-with-m-bxwmpc6>



## LICENSE

This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

## CREATED

Sep 02, 2021

## LAST MODIFIED

Sep 03, 2021

## PROTOCOL INTEGER ID

52909

## Data Preparation

- 1 The Titan Genomic Characterization workflow must be run prior to submitting them to SRA and Genbank in order to prepare the data for submission. Please use the Titan workflow that is compatible with your sequencing data.



Illumina Paired-End



Illumina Single-End



Oxford Nanopore



Clear Labs



FASTA file

## Metadata Formatting

The Terra Metadata Formatter is an excel spreadsheet tool that will help you by collecting all required metadata for each 

- 2 of the sequencing data repositories and formatting this data into a Terra-uploadable data table.

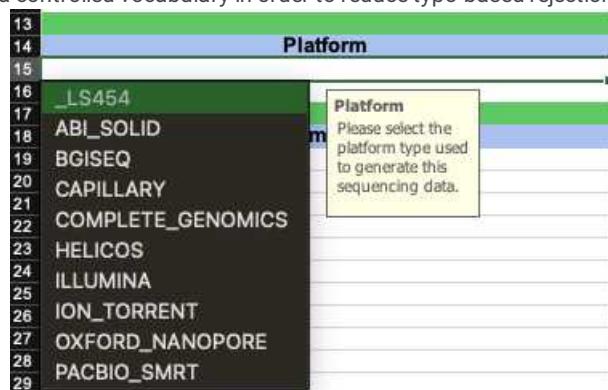
Terra Metadata Formatter

- 2.1 Download and open the Terra Metadata Formatter:

[https://storage.cloud.google.com/theiagen-public-files/terra/mercury-files/Terra\\_Metadata\\_Formatter.xlsx?authuser=0&\\_ga=2.56853090.-974392011.1628263637](https://storage.cloud.google.com/theiagen-public-files/terra/mercury-files/Terra_Metadata_Formatter.xlsx?authuser=0&_ga=2.56853090.-974392011.1628263637)

- 2.2 Enter the sample metadata into the 'User Input' tab of the Terra Metadata Formatter. The required<sup>15m</sup> fields are highlighted in blue. The optional fields are highlighted in grey. We recommend that you attempt to include as much data about your samples as is available at the time of submission, with particular emphasis on the fields of 'Purpose of Sampling' and 'Purpose of Sequencing', which will be used to correct for statistical biases in the data due to diversity of the sampling methodologies.

Note that some of the fields have dropdown menus. These have been implemented for fields that have a controlled vocabulary in order to reduce typo-based rejections from the various databases.



Platform Dropdown Menu

The **General Metadata** section consists of two required fields:

- Root Entity: This input will define the name of the Terra Data Table when this metadata is uploaded in subsequent steps.
- Submission ID Prefix: This input will be the prefix to the submission ID in the final NCBI submission files. Typical inputs are formatted as the state abbreviation and laboratory abbreviation separated by a hyphen.

The **Laboratory Data** section consists of eight required fields and one optional field:

- GISAID Submitter ID (required): the GISAID Submission ID in the final GISAID submission files (if you have already submitted these samples to GISAID then list the GISAID Submission ID that was used)
- Authors (required): the list of authors included in the final SRA, Genbank and GISAID submission files
- BioProject (required): the BioProject accession number used in the SRA and Genbank submissions
- State: the state of the Originating Laboratory
- Country (required): the country of the Originating Laboratory
- Continent (required): the continent of the Originating Laboratory
- Submitting Laboratory (required): the name of the Submitting Laboratory
- Submitting Laboratory Address (required): the address of the Submitting Laboratory
- Submitter Email (optional): The email associated with the NCBI account that will be used to submit to SRA and Genbank

The **Sequencing Run** section consists of five required fields and two optional fields:

- Platform (required): the sequencing Platform used to generate this sequencing data

- Instrument Model (required): the sequencing Instrument Model used to generate this sequencing data
- Library Strategy (required): the Library Strategy used to generate the sequencing libraries (if using Artic V3 or similar amplicon-based protocol then "AMPLICON" is the most accurate entry for this field.)
- Library Source (required): the material used as the Library Source in the generation of the sequencing libraries (if extracting viral RNA as starting material then "VIRAL RNA" is the most accurate entry for this field.)
- Library Selection (required): the tool used to select libraries to be sequenced
- Primer Scheme (optional): the Primer Scheme in the amplicon generation step of the library preparation
- Amplicon Size (optional): the average Amplicon Size of the Primer Scheme

The **Sample Metadata** section consists of nine required fields and nine optional fields:

- Samples (required): the unique ID of the Samples
- Submission ID Suffix (required): the second component of the Submission ID (this field can be the same as Samples)
- Library ID Suffix (required): this input is used to keep track of samples that have been sequenced more than once, or on multiple platforms (for the first or only sequencing submission for these samples it is recommended to use "01" for this field)
- Collection Date (required): the date the samples were originally collected
- Originating Lab(required): the laboratory where the samples were originally collected
- Originating Lab Address (required): the address of the laboratory where the samples were originally collected
- Organism (required): the target organism of the sequencing run (if sequencing SARS-CoV-2 the "SARS-CoV-2" is the most accurate entry for this field)
- Isolation Source (required): source of the sample (if sequencing samples that were collected as part of a diagnostic assay or or surveillance program from humans then "Clinical" would be the most accurate entry for this field)
- Host Disease (required): disease caused by the target Organism (if sequencing SARS-CoV-2 the "COVID-19" would be the most accurate entry for this field)
- Run ID (optional): the Run ID of the samples
- Patient Gender (optional): the gender of the individual from whom the sample was collected
- Patient Age (optional): the age of the individual from whom the sample was collected
- County (optional): the county from which the sample was collected
- BioSample Accession (optional): if the sample has already been registered with NCBI then include the BioSample here
- Specimen Processing (optional): sample processing steps such as transport media and extraction method can be included here
- Purpose of Sampling (optional): this input can be clinical diagnostics if the sample was taken as a human specimen for SARS-CoV-2 testing
- Purpose of Sequencing (optional): this input can be used to tag samples as Baseline Surveillance or Targeted Sampling (for detailed guidance on what entry is most accurate for your samples please see the APHL guidance document here: <https://www.aphl.org/programs/preparedness/Crisis-Management/Documents/Technical-Assistance-for-Categorizing-Baseline-Surveillance.pdf>)

#### **For Baseline Surveillance:**

1. Sampled randomly for genomic surveillance
2. Those not identified in a targeted sampling effort (targeted efforts defined below)
3. Sampled across targeted sequencing efforts to be representative of the community

#### **For Targeted Sequencing:**

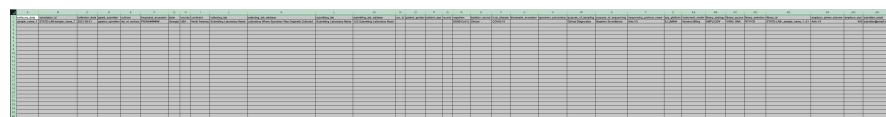
1. Sampled based on cluster/outbreak investigations
  2. Longitudinally or repeatedly sampled from the same individual
  3. Sampled based on pre-screening for a particular variant (e.g., S-gene target failure)
  4. Sampled for the purpose of vaccine escape studies
  5. Sampled based on travel history
  6. Sampled based on disease severity (i.e., targeted sequencing of cases resulting in hospitalization  
or death)
- Sequencing Protocol Name: if using a named sequencing protocol enter the name in this field

[Upload Metadata](#)

- 3.1 Once the sample metadata has been entered into the **User Input** tab of the Metadata Formatter click the 'Terra Data Table' tab:



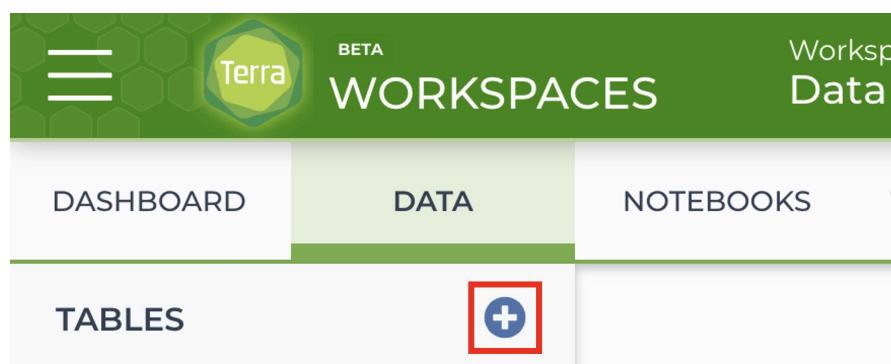
- 3.2 **Select** the whole sheet by hitting control+'a' on your keyboard.  
**Copy** the whole sheet by hitting control+'c' on your keyboard.



- 3.3 Log in and navigate to the **Data** tab in your workspace on Terra.bio:



- 3.4 Select the **plus button** in the blue circle to add a Terra Data Table:



- 3.5 Select the **Text Import** tab:

## Import Table Data

Choose the data import option below. Click here for more info on the table.

FILE IMPORT

TEXT IMPORT

Copy and paste tab separated data here:

Clear

```
entity:my_data_id      submission_id   collection_date gis
sample_name_1    STATE-LAB-sample_name_1 2021-08-31      gis
```

### TSV file templates

 Download sample\_template.tsv

 Terra Support: Importing Data - Using a Template

CANCEL

UPLOAD

3.6 Paste your metadata into the text input field:

## Import Table Data

Choose the data import option below. Click here for more info on the table.

FILE IMPORT

TEXT IMPORT

Copy and paste tab separated data here:

Clear

```
entity:my_data_id      submission_id   collection_date gis
sample_name_1    STATE-LAB-sample_name_1 2021-08-31      gis
```

### TSV file templates

 Download sample\_template.tsv

 Terra Support: Importing Data - Using a Template

CANCEL

UPLOAD

3.7 Click **UPLOAD**

## Import Table Data

Choose the data import option below. Click here for more info on the table.

FILE IMPORT

TEXT IMPORT

Copy and paste tab separated data here:

Clear

```
entity:my_data_id      submission_id   collection_date gis
sample_name_1           STATE-LAB-sample_name_1 2021-08-31      gis
```

### TSV file templates

Download sample\_template.tsv

Terra Support: Importing Data - Using a Template

CANCEL

UPLOAD

Mercury

4 Mercury Prep

15m

4.1 Select Mercury\_Prep\_SE or Mercury\_Prep\_PE from the **Workflows** tab in your Terra workspace:

Mercury\_PE\_Prep

V. kgl-pha4ge-meta-dev  
Source: Dockstore

Mercury\_SE\_Prep

V. main  
Source: Dockstore

Mercury Paired End and Mercury Single-End

4.2 Choose the appropriate Version and Root Entity, then click **Select Data**:

Back to list

Mercury\_Prep

Version v1.5.1

Source: [github.com/theiagen/public\\_health\\_viral\\_genomics/Mercury\\_Prep:v1.5.1](https://github.com/theiagen/public_health_viral_genomics/Mercury_Prep:v1.5.1)

Synopsis:

No documentation provided

Run workflow with inputs defined by file paths  
 Run workflow(s) with inputs defined by data table

Step 1

Select root entity type

my\_data

Step 2

SELECT DATA

No data selected

Use call caching

Delete intermediate outputs

i

Use reference disks

i

Retry with more memory

i

#### 4.3 Select the samples that you would like to prepare for submission:

2m

Select Data

Choose specific my\_data to process

Select my\_data to process

my_data_id	amplicon_primer_scheme	amplicon_size	authors	bioproject_accession	collecting_lab
sample_name_1	Arctic V3	400	list, of, authors	PRJNA#####	Submitting Laborator

Search

#### 4.4 Enter the input variables:

10m

SCRIPT    INPUTS    OUTPUTS    RUN ANALYSIS

Hide optional inputs

Download json | Drag or click to upload json  SEARCH INPUTS

Task name	Variable	Type	Attribute
mercury_pe_prep	assembly_method	String	this.assembly_method
mercury_pe_prep	Authors	String	this.Authors
mercury_pe_prep	bioproject_accession	String	this.bioproject_accession
mercury_pe_prep	collecting_lab	String	this.collecting_lab
mercury_pe_prep	collecting_lab_address	String	this.collecting_lab_address
mercury_pe_prep	collection_date	String	this.collection_date
mercury_pe_prep	gisaid_submitter	String	this.gisaid_submitter

SAVE CANCEL

#### 4.5 Select the default outputs:

SCRIPT    INPUTS    **OUTPUTS**    RUN ANALYSIS

Output files will be saved to  
Files / submission unique ID / mercury\_pe\_prep / workflow unique ID

References to outputs will be written to  
Tables / my\_data

Fill in the attributes below to add or update columns in your data table

Download json | Drag or click to upload json  SEARCH OUTPUTS

Task name	Variable	Type	Attribute   Use defaults
mercury_pe_prep	delID_assembly	File	this.delID_assembly
mercury_pe_prep	genbank_assembly	File	this.genbank_assembly
mercury_pe_prep	genbank_metadata	File	this.genbank_metadata
mercury_pe_prep	gisaid_assembly	File	this.gisaid_assembly
mercury_pe_prep	gisaid_metadata	File	this.gisaid_metadata
mercury_pe_prep	mercury_pe_prep_analysis_date	String	this.mercury_pe_prep_analysis_date
mercury_pe_prep	mercury_pe_prep_version	String	this.mercury_pe_prep_version

SAVE CANCEL

#### 4.6 Once the inputs and outputs have been defined, Save the workflow parameters:

SCRIPT    INPUTS    OUTPUTS    RUN ANALYSIS

Output files will be saved to  
Files / submission unique ID / mercury\_pe\_prep / workflow unique ID

References to outputs will be written to  
Tables / my\_data

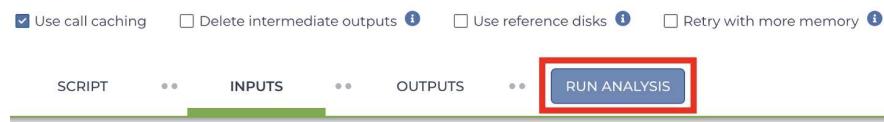
Fill in the attributes below to add or update columns in your data table

Download json | Drag or click to upload json  SEARCH OUTPUTS

Task name	Variable	Type	Attribute   Use defaults
mercury_pe_prep	delID_assembly	File	this.delID_assembly
mercury_pe_prep	genbank_assembly	File	this.genbank_assembly
mercury_pe_prep	genbank_metadata	File	this.genbank_metadata
mercury_pe_prep	gisaid_assembly	File	this.gisaid_assembly
mercury_pe_prep	gisaid_metadata	File	this.gisaid_metadata
mercury_pe_prep	mercury_pe_prep_analysis_date	String	this.mercury_pe_prep_analysis_date
mercury_pe_prep	mercury_pe_prep_version	String	this.mercury_pe_prep_version

SAVE CANCEL

#### 4.7 Click RUN ANALYSIS



Confirm and launch the analysis by clicking LAUNCH:

## Confirm launch

This analysis will be run by [Cromwell 67](#).

Output files will be saved as workspace data in:

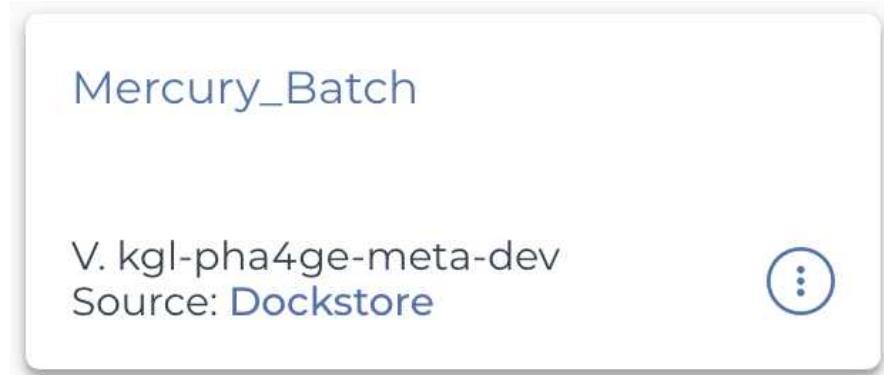
multi-region: US [i](#)

This will launch **1** analysis.



- 5 Mercury Batch 20m

- 5.1 Once Mercury Prep has successfully completed navigate to the Mercury Batch workflow:



- 5.2 Select the appropriate version of the workflow:



### 5.3 Select the SET LEVEL root entity type:

← Back to list

#### Mercury\_Batch

Version: v1.5.1

Source: [github.com/theiagen/public\\_health\\_viral\\_genomics/Mercury\\_Batch:v1.5.1](https://github.com/theiagen/public_health_viral_genomics/Mercury_Batch:v1.5.1)

Synopsis:  
No documentation provided

Run workflow with inputs defined by file paths  
 Run workflow(s) with inputs defined by data table

**Step 1**

Select root entity type: my\_data\_set

**Step 2**

**SELECT DATA** No data selected

Use call caching  Delete intermediate outputs i  Use reference disks i  Retry with more memory i

### 5.4 Click SELECT DATA:

← Back to list

#### Mercury\_Batch

Version: v1.5.1

Source: [github.com/theiagen/public\\_health\\_viral\\_genomics/Mercury\\_Batch:v1.5.1](https://github.com/theiagen/public_health_viral_genomics/Mercury_Batch:v1.5.1)

Synopsis:  
No documentation provided

Run workflow with inputs defined by file paths  
 Run workflow(s) with inputs defined by data table

**Step 1**

Select root entity type: my\_data\_set

**Step 2**

**SELECT DATA** No data selected

Use call caching  Delete intermediate outputs i  Use reference disks i  Retry with more memory i

### 5.5 Select the dataset of sample that you would like to batch for submission (Note: the dataset root entity is the plural form of the original root entity): 2m

#### Select Data

Create a new my\_data\_set from selected Mercury\_Devs  
 Choose specific my\_data\_sets to process

Select my\_data\_sets to process

my_data_set_id
<input checked="" type="checkbox"/> Mercury_PE_Prep_2021-08-25T23-52-31

### 5.6 Enter the INPUTS. The inputs for Mercury Batch will be entered at the Array Level. This means the notation will be formatted as this.data\_sets.{attribute}: 10m

SCRIPT    \* \* INPUTS    \* \* OUTPUTS    \* \* RUN ANALYSIS

SAVE CANCEL

Hide optional inputs Download json | Drag or click to upload json SEARCH INPUTS

Task name	Variable	Type	Attribute
mercury_batch	biosample_attributes	Array[File]	this.my_data.biosample_attributes

## 5.7 Enter the public GCP bucket to stage your data for the final submission to NCBI SRA:

mercury\_batch      gcp\_bucket      String      "gs://theiagen\_sra\_transfer"      [...]

Note: If you are using the GCP bucket submission method please ensure that you have write access to the public Theiagen GCP bucket for NCBI submission:  
"gs://theiagen\_sra\_transfer"

## 5.8 Select the default OUTPUTS:

SCRIPT    \*\*    INPUTS    \*\*    OUTPUTS    \*\*    RUN ANALYSIS

Output files will be saved to  
Files / submission unique ID / mercury\_batch / workflow unique ID

References to outputs will be written to  
Tables / my\_data.set

Fill in the attributes below to add or update columns in your data table

SAVE    CANCEL

Task name	Variable	Type	Attribute   Use defaults
mercury_batch	BioSample_attributes	File	this.BioSample_attributes
mercury_batch	GenBank_assembly	File	this.GenBank_assembly
mercury_batch	GenBank_batched_samples	File	this.GenBank_batched_samples
mercury_batch	GenBank_excluded_samples	File	this.GenBank_excluded_samples
mercury_batch	GenBank_modifier	File	this.GenBank_modifier

## 5.9 Once the inputs and outputs have been defined, Save the workflow parameters:

SCRIPT    \*\*    INPUTS    \*\*    OUTPUTS    \*\*    RUN ANALYSIS

Output files will be saved to  
Files / submission unique ID / mercury\_pe\_prep / workflow unique ID

References to outputs will be written to  
Tables / my\_data

Fill in the attributes below to add or update columns in your data table

SAVE    CANCEL

Task name	Variable	Type	Attribute   Use defaults
mercury_pe_prep	delD_assembly	File	this.delD_assembly
mercury_pe_prep	genbank_assembly	File	this.genbank_assembly
mercury_pe_prep	genbank_metadata	File	this.genbank_metadata
mercury_pe_prep	gisaid_assembly	File	this.gisaid_assembly
mercury_pe_prep	gisaid_metadata	File	this.gisaid_metadata
mercury_pe_prep	mercury_pe_prep_analysis.date	String	this.mercury_pe_prep_analysis.date
mercury_pe_prep	mercury_pe_prep_version	String	this.mercury_pe_prep_version

## 5.10 Click RUN ANALYSIS

Use call caching     Delete intermediate outputs     Use reference disks     Retry with more memory

SCRIPT    \*\*    INPUTS    \*\*    OUTPUTS    \*\*    RUN ANALYSIS

Confirm and launch the analysis by clicking LAUNCH:

## Confirm launch

This analysis will be run by [Cromwell 67](#).

Output files will be saved as workspace data in:

 multi-region: US [\*i\*](#)

This will launch **1** analysis.

CANCEL LAUNCH

- 5.11 Retrieve your submission files by navigating to the Terra Data Table containing the Mercury Batch outputs: 5m

		<a href="#">DOWNLOAD ALL ROWS</a>	<a href="#">COPY PAGE TO CLIPBOARD</a>	0 rows selected	Search <input type="text"/>	
<input type="checkbox"/>	Mercury_Dev_id	biosample_attributes	genbank_assembly	genbank_modifier	grnstd_assembly	grnstd_metadata
<input type="checkbox"/>	2000027963	CA-CDPH-2000027963.biosample_attributes	CA-CDPH-2000027963.genbank_assembly	CA-CDPH-2000027963.genbank_modifier	CA-CDPH-2000027963.grnstd_assembly	CA-CDPH-2000027963.grnstd_metadata

Click on the file names in blue:

		<a href="#">DOWNLOAD ALL ROWS</a>	<a href="#">COPY PAGE TO CLIPBOARD</a>	0 rows selected	Search <input type="text"/>	
<input type="checkbox"/>	Mercury_Dev_id	<a href="#">biosample_attributes</a>				
<input type="checkbox"/>	2000027963	<a href="#">CA-CDPH-2000027963_biosample_attributes.tsv</a>				

Download the files:

## File Details



### Filename

CA-CDPH-2000027963\_biosample\_attributes.tsv

### Preview

```
*sample_name      sample_title      bioproject_accession  
CA-CDPH-2000027963                      PRJNA750736      SARS-CoV
```

### File size

1.15 KB

[View this file in the Google Cloud Storage Browser](#)

[DOWNLOAD FOR < \\$0.01\\*](#)

### Terminal download command

```
gsutil cp gs://fc-6377040c-403f-416a-bfc
```



[More Information](#)

\* Estimated. Download cost may be higher in China or Australia.

[DONE](#)

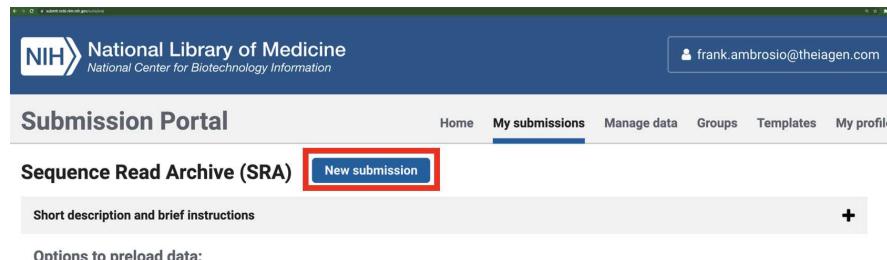
## SRA Submission

- 6 Submit your data to SRA (and simultaneously generate BioSample accession numbers for your samples) 20m

- 6.1 Navigate and Log in to the SRA Submission Portal:



- 6.2 Select New Submission:



- 6.3 Enter your submitter information, select your submission group, and enter the information for your organization: 1m

#### **Sequence Read Archive (SRA) submission: SUB10316559**

New

1 SUBMITTER > 2 GENERAL INFO > 3 SRA METADATA > 4 FILES > 5 REVIEW & SUBMIT

##### Submitter

★ First (given) name	Middle name	★ Last (family) name		
Frank	J	Ambrosio		
★ Email (primary)	Email (secondary)			
frank.ambrosio@theiagen.com				
<small>ⓘ At least one email should be from the organization's domain.</small>				
<b>Group for this submission</b>				
<input checked="" type="radio"/> No group (affiliation from my personal profile) <input type="radio"/> 4 members John Bell's shared submissions <input type="radio"/> 1 member Frank Ambrosio's shared submissions (edit group)				
★ Submitting organization	Submitting organization URL	★ Department		
Theiagen Genomics, LLC	https://theiagen.com/contact/	Bioinformatics		
Phone	Fax			
★ Street	★ City	★ State/Province	★ Postal code	★ Country
1745 Shea Center Drive	Highlands Ranch	CO	80129	USA

Continue

Update my contact information in profile

- 6.4 Enter your BioProject number: 1m

## Sequence Read Archive (SRA) submission: SUB10316559

New



### General Information

#### BioProject

BioProject describes the goal of your research effort.

**\* Did you already register a BioProject for this research, e.g. for the submission of the reads to SRA and/or of the genome to GenBank?**

Yes  No

**\* Existing BioProject**

PRJNAXXXXX

#### BioSample

The BioSample records the detailed biological and physical properties of the sample that was sequenced. A BioSample can be used in more than one BioProject since it should be used for all the data that were obtained from that sample. Usually SRA data sets are generated from more than one sample.

**\* Did you already register a BioSample for this sample, e.g. for the submission of the reads to SRA and/or of the genome to GenBank?**

Yes  No

#### Release date

Note: Release of BioProject or BioSample is also triggered by the release of linked data.

**\* When should this submission be released to the public?**

Release immediately following processing  
 Release on specified date or upon publication, whichever is first

**Continue**

- 6.5 Select 'No' if you do not already have BioSample accession numbers for your samples in order to generate them upon SRA submission: 1m

## Sequence Read Archive (SRA) submission: SUB10316559

New



### General Information

#### BioProject

BioProject describes the goal of your research effort.

**\* Did you already register a BioProject for this research, e.g. for the submission of the reads to SRA and/or of the genome to GenBank?**

Yes  No

**\* Existing BioProject**

PRJNAXXXXX

#### BioSample

The BioSample records the detailed biological and physical properties of the sample that was sequenced. A BioSample can be used in more than one BioProject since it should be used for all the data that were obtained from that sample. Usually SRA data sets are generated from more than one sample.

**By clicking "No," you indicate you do not have an existing BioSample to associate with this sequence data and will create the BioSample on one of the next steps.**

**\* Did you already register a BioSample for this research, e.g. for the submission of the reads to SRA and/or of the genome to GenBank?**

Yes  No

#### Release date

Note: Release of BioProject or BioSample is also triggered by the release of linked data.

**\* When should this submission be released to the public?**

- Release immediately following processing
- Release on specified date or upon publication, whichever is first

**Continue**

6.6 Select your Release Date (we recommend releasing your data immediately following processing):

1m

# Sequence Read Archive (SRA) submission: SUB10316559

New



## General Information

### BioProject

i BioProject describes the goal of your research effort.

**\* Did you already register a BioProject for this research, e.g. for the submission of the reads to SRA and/or of the genome to GenBank?**

Yes  No

**\* Existing BioProject**

PRJNAXXXXX

### BioSample

i The BioSample records the detailed biological and physical properties of the sample that was sequenced. A BioSample can be used in more than one BioProject since it should be used for all the data that were obtained from that sample. Usually SRA data sets are generated from more than one sample.

**By clicking 'No,' you indicate you do not have an existing BioSample to associate with this sequence data and will create the BioSample on one of the next steps.**

**\* Did you already register a BioSample for this research, e.g. for the submission of the reads to SRA and/or of the genome to GenBank?**

Yes  No

### Release date

i Note: Release of BioProject or BioSample is also triggered by the release of linked data.

**\* When should this submission be released to the public?**

- Release immediately following processing  
 Release on specified date or upon publication, whichever is first

**Continue**

- 6.7 Select the appropriate submission package (if you are submitting SARS-CoV-2 sequences extracted<sup>1m</sup> from a human specimen please select the SARS-CoV-2 clinical or host-associated package):

## Sequence Read Archive (SRA) submission: SUB10316559

Severe acute respiratory syndrome coronavirus 2 Genome sequencing, Sep 03 '21

1 SUBMITTER 2 GENERAL INFO 3 BIOSAMPLE TYPE 4 BIOSAMPLE ATTRIBUTES 5 SRA METADATA 6 FILES 7 REVIEW & SUBMIT

### Sample Type

\* Select the package that best describes your samples.

All packages Packages for MAG submitters Packages for metagenome submitters

(Optional) Filter packages by organism name

Enter the full scientific name of your samples, e.g., Escherichia coli

Reset and show all packages

To filter for relevant BioSample packages, enter the full scientific name of the organism of your samples.

- If your BioSamples are derived from a species not represented in NCBI's Taxonomy database, enter the genus-level name, e.g., *Escherichia*
- If your BioSamples are derived from more than one organism, enter the common species, genus, or family, e.g., *Enterobacteriaceae*
- If your BioSamples are metagenomic/environmental, or metagenome-assembled genomes (MAG), select the appropriate tab above
- For more information about organism names, see [Organism information](#).

NCBI packages [More...](#)

#### SARS-CoV-2: clinical or host-associated

Use for SARS-CoV-2 samples that are relevant to public health. Required attributes include those considered useful for the rapid analysis and trace back of SARS-CoV-2 cases.

#### SARS-CoV-2: wastewater surveillance

Use for SARS-CoV-2 wastewater surveillance samples that are relevant to public health. Required attributes include those considered useful for the rapid analysis and trace back of SARS-CoV-2 cases.

#### Pathogen

Use for pathogen samples that are relevant to public health. Required attributes include those considered useful for the rapid analysis and trace back of pathogens.

#### Microbe

Use for bacteria or other unicellular microbes when it is not

#### GSC MiGS packages for genomes, metagenomes, and marker sequences [More...](#)

#### MIGS Cultured Bacterial/Archaeal

Use for cultured bacterial or archaeal genomic sequences. Organism must have lineage [Bacteria](#) or [Archaea](#).

#### MIGS Eukaryotic

Use for eukaryotic genomic sequences. Organism must have lineage [Eukaryota](#).

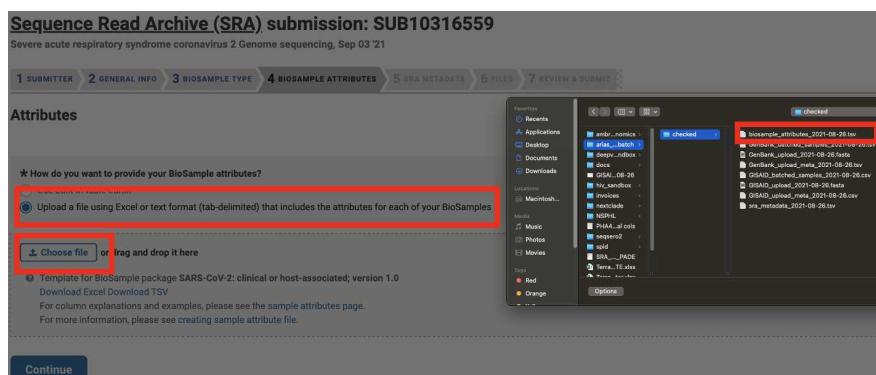
#### MIGS Viral

Use for virus genomic sequences. Organism must have lineage [Viruses](#).

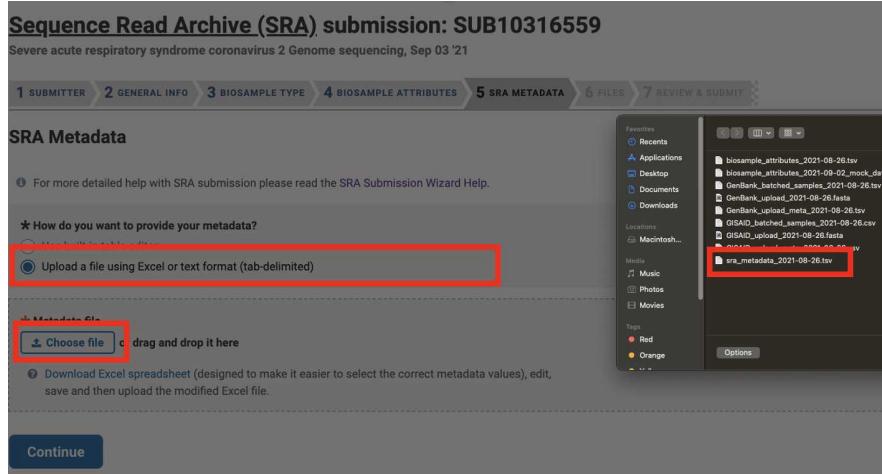
#### MIMAG Metagenome-assembled Genome

Use for metagenome-assembled genome sequences produced using computational binning tools that group sequences into individual organism genome assemblies starting from metagenomic data sets. Organism cannot contain the term 'metagenome'. Use the MIUVIG package for virus genomes.

- 6.8 Choose the 'Upload a file...' option and upload the BioSample attributes file downloaded in previous steps: 1m



- 6.9 Choose the 'Upload a file...' option and upload the SRA Metadata file downloaded in previous steps: 1m



- 6.10 Select the 'AWS or GCP bucket' option and enter the name of the public data bucket where your reads have been placed in the staging phase of the data submission procedure: 1m

**Sequence Read Archive (SRA) submission: SUB10316559**  
Severe acute respiratory syndrome coronavirus 2 Genome sequencing, Sep 03 '21

1 SUBMITTER 2 GENERAL INFO 3 BIOSAMPLE TYPE 4 BIOSAMPLE ATTRIBUTES 5 SRA METADATA 6 FILES 7 REVIEW & SUBMIT

### Files

- Each file must be listed in the SRA metadata table you uploaded. If you are uploading a **tar** archive, list each file name, not the archive name.
- Unique file names that do not contain any sensitive information should be used for all files. File names as submitted appear publicly when data is retrieved from the cloud.
- Files can be compressed using **gzip** or **bzip2**, and may be submitted in a **tar** archive, but archiving or compressing your files is not required. **Do not use zip!**

**\* How do you want to provide files for this submission?**

Web browser upload via HTTP or Aspera Connect plugin  
Do not use web browser HTTP upload if you are uploading files over 10 GB or more than 300 files.

FTP or Aspera Command Line file preload  
All files for a submission must be uploaded into a single folder.

AWS or GCP bucket

**\* Which cloud provider do you use to store these files?**

AWS  GCP

**\* Bucket name**  
theiagen\_sra\_transfer

Google Cloud Storage instructions

**Do not modify or move the files** you are submitting from this cloud bucket until you see the "SRA: Processed" status is present for the submission.

Autofinish submission

**Continue**

- 6.11 Review and Submit to complete your SRA submission! You will be able to download your BioSample accession numbers from the SRA submission portal as soon as they become available. 1m

- 7 Retrieve the BioSample accession numbers '.tsv' file from the SRA portal 1m

- 7.1 Navigate to the SRA Submission Portal (you should already be logged in)  
 Locate the Status column of the submissions table:

- 7.2 Click 'Download attributes file with BioSample accessions' for the SRA submission executed earlier in this protocol:

Genbank Submission

40m

5m

- 8 Add BioSample accession numbers to Genbank\_meta\_upload file

- 8.1 Open the attributes file downloaded from SRA containing the BioSample accession numbers

	A	B	C	D
1	accession	message	sample_name	sample_title
2	SAMN11111111	Provided	CA-CDPH-2000099999	
3	SAMN11111111	Successful	CA-CDPH-2000099999	
4	SAMN22222222	Provided	CA-CDPH-2000099998	
5	SAMN22222222	Successful	CA-CDPH-2000099998	
6	SAMN33333333	Provided	CA-CDPH-2000099997	
7	SAMN33333333	Successful	CA-CDPH-2000099997	
8	SAMN44444444	Provided	CA-CDPH-2000099996	
9	SAMN44444444	Successful	CA-CDPH-2000099996	
10	SAMN55555555	Provided	CA-CDPH-2000099995	
11	SAMN55555555	Successful	CA-CDPH-2000099995	
12				
13				

**8.2** Open the Genbank\_meta\_sra file downloaded from Terra (the output from Mercury Batch)

GenBank_upload_meta_2021-08-26				
Home	Insert	Draw	Page Layout	Formulas
G18	A	B	C	D
1	Sequence_ID	country	host	isolate
2	CA-CDPH-2000099999	USA	Homo sapiens	SARS-CoV-2/Human/USA/CA-CDPH-2000027968/2021
3	CA-CDPH-2000099998	USA	Homo sapiens	SARS-CoV-2/Human/USA/CA-CDPH-2000027969/2021
4	CA-CDPH-2000099997	USA	Homo sapiens	SARS-CoV-2/Human/USA/CA-CDPH-2000027970/2021
5	CA-CDPH-2000099996	USA	Homo sapiens	SARS-CoV-2/Human/USA/CA-CDPH-2000027971/2021
6	CA-CDPH-2000099995	USA	Homo sapiens	SARS-CoV-2/Human/USA/CA-CDPH-2000027972/2021
7				
8				

8.3 Use XLOOKUP to algorithmically add the BioSample accession numbers to the Genbank\_meta\_sra file. 5m

Use this formula:  
=XLOOKUP(A2,attributes.tsv!\$C:\$C,attributes.tsv!\$A:\$A)

Drag down the formula using the green square in the bottom right corner of the cell

You have successfully added your BioSample accession numbers to the Genbank\_meta\_upload file.

#### 8.4 Save the Genbank\_meta\_upload file (now including the BioSample accession numbers)

## 9 Genbank submission

## 9.1 Navigate to the Genbank submission portal

## Submission Portal

Submit to the world's largest public repository of biological and scientific information

Type a few words about the sequence data you are submitting and select an option to learn more. You can also browse submission information below.

### What do you want to submit?

Enter a few words about your sequence data.



Suggest tool

SARS-CoV-2

16S rRNA

genome

ITS

SRA

### 9.2 Select SARS-CoV-2

## Submission Portal

Submit to the world's largest public repository of biological and scientific information

Type a few words about the sequence data you are submitting and select an option to learn more. You can also browse submission information below.

### What do you want to submit?

Enter a few words about your sequence data.



Suggest tool

SARS-CoV-2

16S rRNA

genome

ITS

SRA

## GenBank

### 9.3 Click the submit button under the Genbank heading:

## Submit SARS-CoV-2 sequences

Add your SARS-CoV-2 sequence data to the growing public archive

Easily submit assembled & raw read SARS-CoV-2 data on the web or via XML upload for COVID-19 response.  
NCBI is here to help.

### GenBank

Submitted 2021-08-30

Submit assembled reads of SARS-CoV-2 with FASTA files and source metadata. Annotation for SARS-CoV-2 is not required.

Accessions in 2 hours (avg)

Learn more

Submit

### Sequence Read Archive (SRA)

Started 2021-09-03

Submit unassembled reads of SARS-CoV-2 with BioProject, BioSample, metadata and NGS files.

Accessions in 2 hours (avg)

Learn more

Submit

### 9.4 Select 'New submission':

## Submission Portal

GenBank

New submission



Note: Submit only **ribosomal RNA (rRNA)**, **rRNA-ITS**, **metazoan COX1**, **Influenza**, **Norovirus**, **Dengue** or **SARS-CoV-2** sequences here.

All other submission types should use one of the alternate submission tools (e.g. BankIt, tbl2asn, etc.)

- 9.5 Select 'SARS-CoV-2, Influenza, Norovirus, or Dengue virus' and 'SARS-CoV-2' to the questions '**What do your sequences contain?**' and '**Which virus?**', respectively.

## Submission Portal

**GenBank submission: SUB10317355**

New

1 SUBMISSION TYPE > 2 SUBMITTER > 3 SEQUENCING TECHNOLOGY > 4 SEQUENCES > 5 SEQUENCE PROCESSING > 6 SOURCE INFO > 7 SOURCE MODIFIERS > 8 REFERENCES > 9 REVIEW & SUBMIT >

### Submission Type

\* What do your sequences contain?

- rRNA or rRNA-ITS
- COX1 from metazoan mitochondria
- SARS-CoV-2, Influenza, Norovirus, or Dengue virus

\* Which virus?

- SARS-CoV-2
- Influenza virus
- Norovirus
- Dengue virus

### Review requirements for SARS-CoV-2 submissions

- If none of the options above describe your sequences, use BankIt to submit.

Submission title (Optional, not displayed in final records) [?](#)

Continue

**1 SUBMISSION TYPE****2 SUBMITTER****3 SEQUENCING TECHNOLOGY**

## Submission Type

**\* What do your sequences contain?**

- rRNA or rRNA-ITS [?](#)
- COX1 from metazoan mitochondria [?](#)
- SARS-CoV-2, Influenza, Norovirus, or Dengue virus [?](#)

**\* Which virus?**

- SARS-CoV-2

- Influenza virus
- Norovirus
- Dengue virus

## 9.6 Enter the required submitter information:

2m

National Library of Medicine  
National Center for Biotechnology Information

Submission Portal

**GenBank submission: SUB10317355**  
SARS-CoV-2

1 SUBMISSION TYPE 2 SUBMITTER 3 SEQUENCING TECHNOLOGY 4 SEQUENCES 5 SEQUENCE PROCESSING 6 SOURCE INFO 7 SOURCE MODIFIERS 8 REFERENCES 9 REVIEW & SUBMIT

**Submitter**

**Affiliation**  
 ⓘ The information you give here will be displayed in the final sequence records.  
For address details, provide the primary address where work was done to generate the data in this submission.

Group for this submission  
 0 members: No group  
 1 member: Frank Ambrosio's shared submissions (edit group)  
you  
 4 members: John Bell's shared submissions

**Submitting organization** **Department**  
Theigen Genomics, LLC Bioinformatics

**Street** **City** **State/Province** **Postal code** **Country**  
1745 Shea Center Drive Highlands Ranch CO 80129 USA

**Contact information**  
 ⓘ GenBank may use this information to contact you about your submission, it will not be displayed in the final sequence records.

**Email (primary)** **Email (secondary)**  ⓘ Please provide an alternate email address to ensure that messages are received

**First (given) name** **Middle name** **Last (family) name**  
Frank J Ambrosio

**Phone** **Fax**

**Continue**  Update my contact information in profile

## 9.7 Select the Sequencing Technology used to generate the sequencing data of which the Genbank assembly submissions are composed. Select 'Assembled sequences (...) as the assembly state:

1m

Illumina:

The screenshot shows the "Submission Portal" for GenBank submission (SUB10317355). The "SEQUENCING TECHNOLOGY" step is selected. Under "Method", the "Illumina" checkbox is checked and highlighted with a red box. Other options like Sanger, 454, Helicos, IonTorrent, Pacific Biosciences, SOLID, and Other are also listed but not checked.

**Sequencing Technology**

**Method**

\* What methods were used to obtain these sequence(s)?

Sanger dideoxy sequencing  
 454  
 Helicos  
 Illumina  
 IonTorrent  
 Pacific Biosciences  
 SOLID  
 Other

**Assembly state**

These sequences are:

Unassembled sequence reads  
 Assembled sequences (each sequence was assembled from two or more overlapping sequence reads)

**Assembly Information**

\* Assembly program: iVar  
\* Version or date: 1.3.1  
Delete  
Add another assembly program

**Continue**

Oxford Nanopore Technologies (and Clear Labs):

The screenshot shows the "Submission Portal" for GenBank submission (SUB10317355). The "SEQUENCING TECHNOLOGY" step is selected. Under "Method", the "Other" checkbox is checked and highlighted with a red box. A custom method "Oxford Nanopore Techn" is entered in the text field below it.

**Sequencing Technology**

**Method**

\* What methods were used to obtain these sequence(s)?

Sanger dideoxy sequencing  
 454  
 Helicos  
 Illumina  
 IonTorrent  
 Pacific Biosciences  
 SOLID  
 Other

\* Method  
Oxford Nanopore Techn

**Assembly state**

These sequences are:

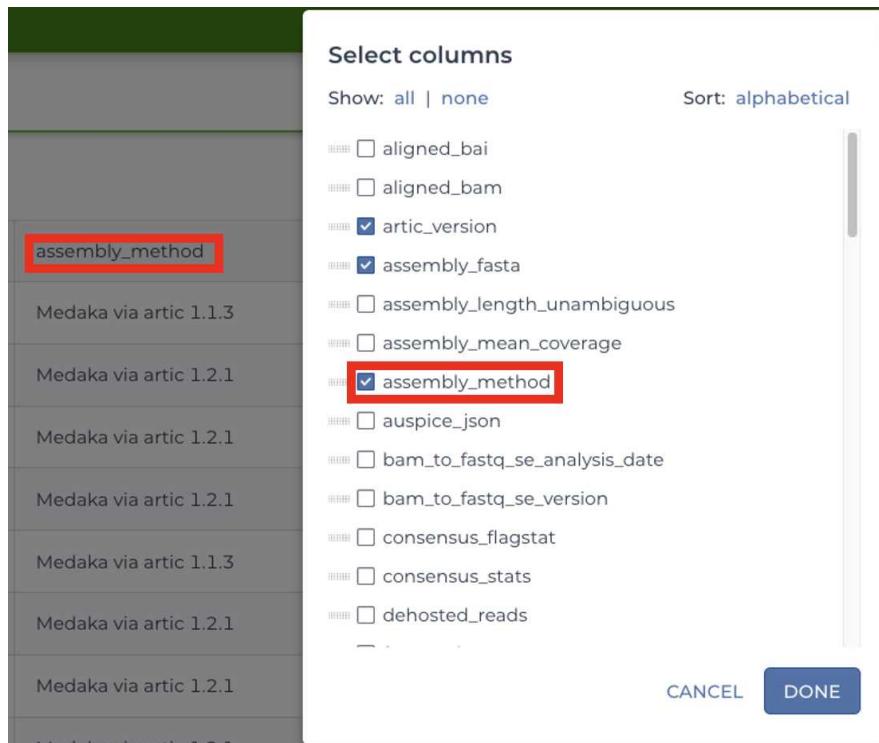
Unassembled sequence reads  
 Assembled sequences (each sequence was assembled from two or more overlapping sequence reads)

**Assembly Information**

\* Assembly program: Medaka via Artic 1.2.1  
\* Version or date: Artic 1.2.1  
Delete  
Add another assembly program

**Continue**

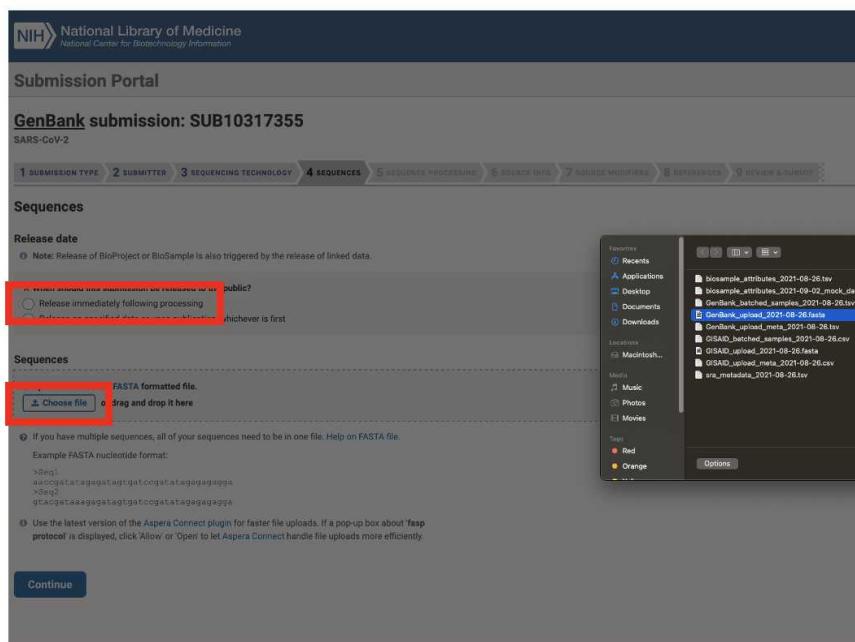
Note: the assembly method is a default output for the Titan Genomic Characterization workflow. The assembly software and version can be found in your Terra data table:



9.8

1m

Select 'Release immediately following processing' and upload the Genbank\_assembly.fasta file:



Submission Portal

**GenBank submission: SUB10317355**

SARS-CoV-2

1 SUBMISSION TYPE 2 SUBMITTER 3 SEQUENCING TECHNOLOGY 4 SEQUENCES 5 SEQUENCE PROCESSING 6 SOURCE INFO 7 SOURCE MODIFIERS 8 REFERENCES 9 REVIEW & SUBMIT

Sequences

Release date

Note: Release of BioProject or BioSample is also triggered by the release of linked data.

\* When should this submission be released to the public?

Release immediately following processing

Release on specified date or upon publication, whichever is first

Sequences

\* Upload a nucleotide FASTA formatted file. To upload a new file, you must delete your previous file.

GenBank\_upload\_2021-08-26.fasta

147.4 kB 2021-09-03 14:38

Delete

If you have multiple sequences, all of your sequences need to be in one file. Help on FASTA file.

Example FASTA nucleotide format:

```
>Seq1  
aaccgatatacgatagtgtatccgatatacgagagaggaa  
>Seq2  
gtacgataaagatcgatgtatccgatatacgagagaggaa
```

Use the latest version of the Aspera Connect plugin for faster file uploads. If a pop-up box about 'Aspera protocol' is displayed, click 'Allow' or 'Open' to let Aspera Connect handle file uploads more efficiently.

Continue

- 9.9 You will be asked to explain the strings of N's in your assemblies. The software used by the Titan Genomic Characterization Workflows estimates the length between sequenced regions using the Wuhan-1 reference genome for alignments: 1m

**What do the internal NNN's represent?**

- The nucleotide sequence(s) in your file contain strings of internal NNN's (length > 10). Please answer the question below and click Continue at the bottom of the page.

A region of estimated length between the sequenced regions based on an alignment to similar sequences or genome

**Release date**

Note: Release of BioProject or BioSample is also triggered by the release of linked data.

\* When should this submission be released to the public?

Release immediately following processing

Release on specified date or upon publication, whichever is first

**Sequences**

\* Upload a nucleotide FASTA formatted file. To upload a new file, you must delete your previous file.

SUB10317355\_GenBank\_upload\_2021-08-26.fasta 147.5 kB 2021-09-03 14:57

Delete

If you have multiple sequences, all of your sequences need to be in one file. Help on FASTA file.

Example FASTA nucleotide format:

```
>Seq1  
aaccgatatacgatagtgtatccgatatacgagagaggaa  
>Seq2  
gtacgataaagatcgatgtatccgatatacgagagaggaa
```

Use the latest version of the Aspera Connect plugin for faster file uploads.

Continue

We recommend selecting yes for the question 'During processing, should NCBI remove' 1m

## 9.10 sequences with errors and process the rest?!

**GenBank submission: SUB10317355**  
SARS-CoV-2

1 SUBMISSION TYPE > 2 SUBMITTER > 3 SEQUENCING TECHNOLOGY > 4 SEQUENCES > 5 SEQUENCE PROCESSING > 6 SOURCE INFO > 7 SOURCE MODIFIERS > 8 REFERENCES > 9 REVIEW & SUBMIT

**Sequence Processing**

**Option to automatically remove failed sequences**

If errors are found on sequences during processing, they will be removed from this submission and the successful sequences accessioned. You will receive a detailed report on these errors.

\* During processing, should NCBI remove sequences with errors and process the rest?

Yes  No

**Continue**

## 9.11 Indicate whether the source of your genomic material was an individual isolate:

1m

**GenBank submission: SUB10317355**  
SARS-CoV-2

1 SUBMISSION TYPE > 2 SUBMITTER > 3 SEQUENCING TECHNOLOGY > 4 SEQUENCES > 5 SEQUENCE PROCESSING > 6 SOURCE INFO > 7 SOURCE MODIFIERS > 8 REFERENCES > 9 REVIEW & SUBMIT

**Source Information**

The first few sequence IDs that we found are:  
CA-CDPH-200027968  
CA-CDPH-200027970  
CA-CDPH-200027970  
CA-CDPH-200027971  
CA-CDPH-200027972

\* Do your sequence IDs represent one of these?

Isolate  NONE of these

Values for these are typically alpha-numeric sample codes used in your laboratory to track individual samples. Select 'NONE of these' if it does not describe your sequence IDs or the sequence IDs contain more information than the descriptions of these fields.

**Continue**

Isolate - Individual isolate from which the sequence was obtained, typically an alphanumeric sample ID.

## 9.12 Upload the Genbank\_meta\_upload file downloaded from Terra in previous steps (Note: This should be the version with the BioSample accession numbers added in Step 8)

1m

**GenBank submission: SUB10317355**  
SARS-CoV-2

1 SUBMISSION TYPE > 2 SUBMITTER > 3 SEQUENCING TECHNOLOGY > 4 SEQUENCES > 5 SEQUENCE PROCESSING > 6 SOURCE INFO > 7 SOURCE MODIFIERS > 8 REFERENCES > 9 REVIEW & SUBMIT

**Source Modifiers**

For each sequence, GenBank requires the following source information:

- collection-date,
- country,
- host, and
- isolate.

Current source modifiers - what you have provided so far

More help: what is a source modifier, description of each modifier, how to provide source modifiers.

If you have already provided all the required information, you can press Continue to proceed.

\* How do you want to apply source modifiers?

Apply source modifiers by uploading a tab-delimited table

- Download source modifier template table.
- Edit the downloaded table in Microsoft Excel or another editor.
- Save the table as a tab-delimited text file.
- Upload your saved table file.

Choose file or drag and drop it here

Click Continue to validate the information and follow the instructions.

**Continue**

- 9.13 Enter authors to be publicly credited for the submission of this sequencing data. If there is a publication associated with this sequence data please enter the name of the publication as well as the authors listed on the publication: 1m

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

**Submission Portal**

**GenBank submission: SUB10317355**  
SARS-CoV-2

**References**

**Sequence authors**  
Who should be publicly credited as the submitter of this sequence data? Enter authors below. Drag and drop to reorder authors.

Sequence authors from your recent submissions (Optional)  
Ambrosio,F.J.

\* First (given) name MI \* Last (family) name Delete      Names will appear in your records as:  
Francis J Ambrosio Ambrosio, F.J.

Add another sequence author

**Reference**

References from your recent submissions (Optional)

\* Publication status  
 Unpublished  In-press  Published

\* Reference title  
N/A

\* Reference authors  
 Same as sequence authors  Specify authors

**Continue**

- 9.14 Review your submission information and click 'Submit' to complete the Genbank submission process!

**Submission Portal**

**GenBank submission: SUB10317355**  
SARS-CoV-2

**Review & Submit**

You have requested that your sequence data be released **immediately following processing**.

**Submitter**

Submitter Frank Ambrosio  
frank.ambrosio@theagen.com  
frankambrosio3@gmail.com

Institution Threagen Genomics, LLC

Department Bioinformatics

Street 1745 Shea Center Drive

City Highlands Ranch

State CO

Postal code 80129

Country USA

**Sequence authors**

Francis J. Ambrosio

**References**

Publication status unpublished

Reference title N/A

Authors same as sequence authors

**Sequencing Technology**

Methods Other: Oxford Nanopore Technologies

Assembly state assembled

Assembly Programs Medaka via Artic 1.2.1 (Artic 1.2.1)

**Submit**

- 9.15** Congratulations! You have submitted both read and assembly data to NCBI, linked by the BioSample accession number. This type of submission greatly enhances the statistical power of the data in public genomic repositories. Thank you for your contribution to public health!