Sep 03, 2021

SRA and Genbank BioSample-Linked Submission with Mercury_Prep and Mercury_Batch

Francis J Ambrosio[1]

[1]Theiagen Genomics

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 the 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-bxwmppc6

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 too that will help you by collecting all required metadata for each 15m

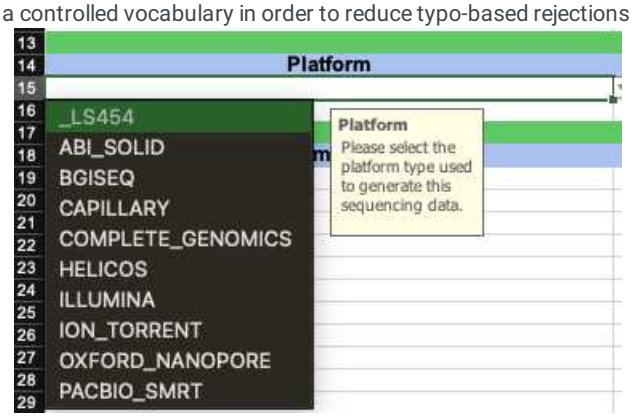
2 of the sequencing data repositories and formatting this data into a Terra-uploadable data table.

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

2.2 Enter the sample metadata into the 'User Input' tab of the Terra Metadata Formatter. The required 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. 15m

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

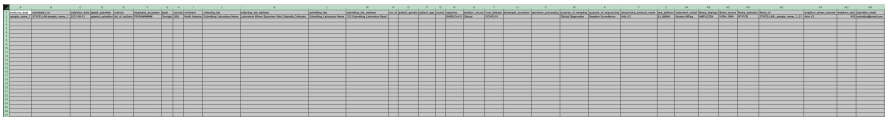
Upload Metadata

3 Upload the **Terra Data Table**

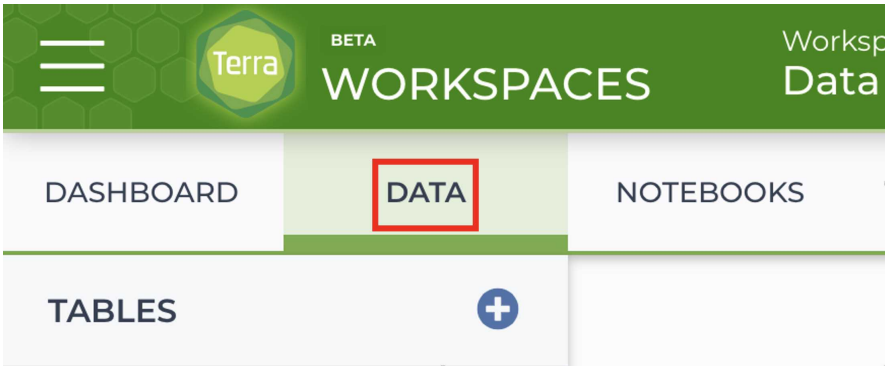
5m

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

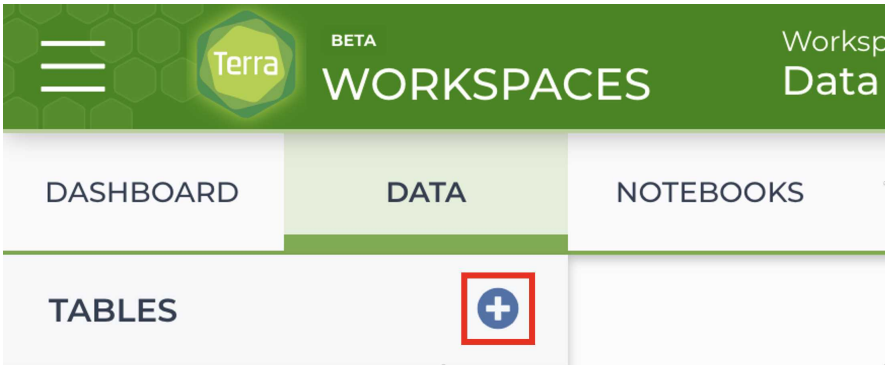
3.2 **Select** the whole sheet by hitting control+'a' on your keyboard.
Copy the whole sheet by hitting control+'c' on your keyboard.



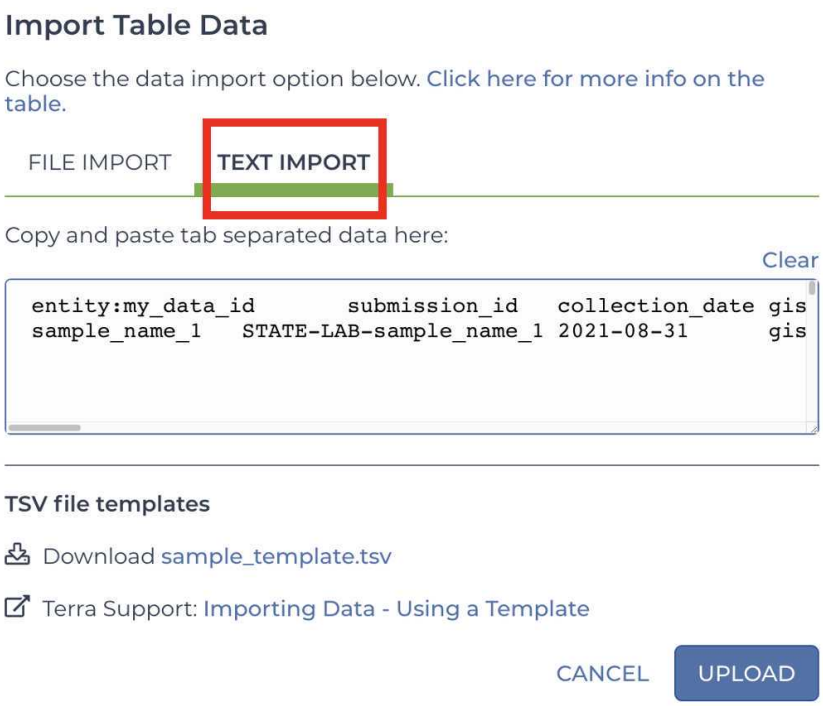
3.3 Log in and navigate to the **Data** tab in your workspace on Terra.bio:



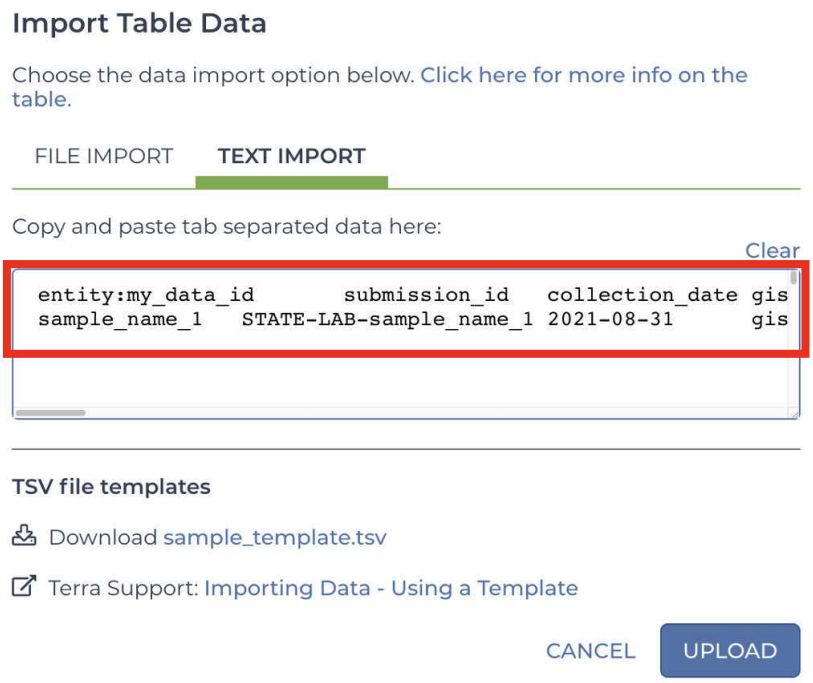
3.4 Select the **plus button** in the blue circle to add a Terra Data Table:



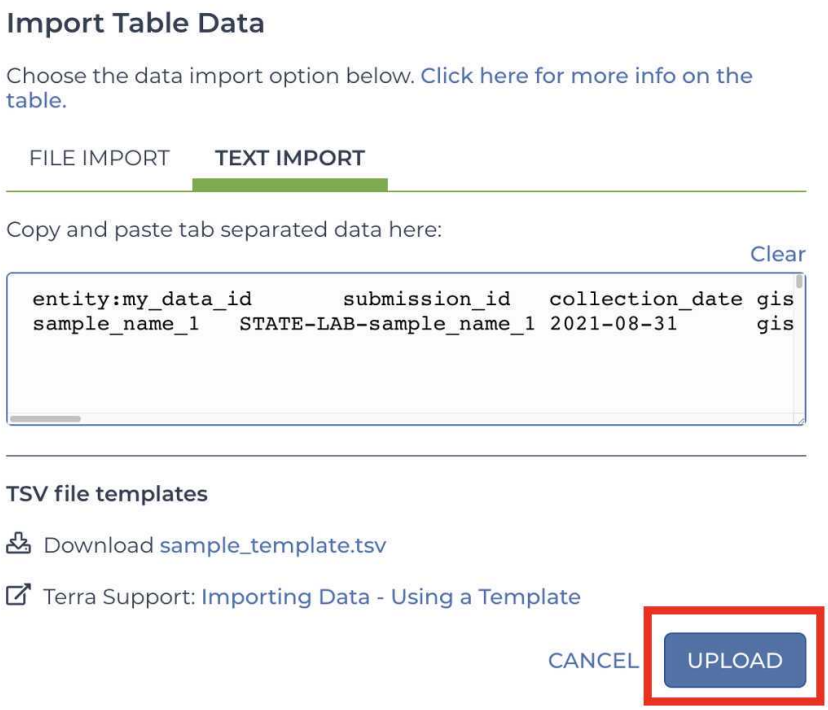
3.5 Select the **Text Import** tab:



3.6 **Paste** your metadata into the text input field:



3.7 Click **UPLOAD**



Mercury

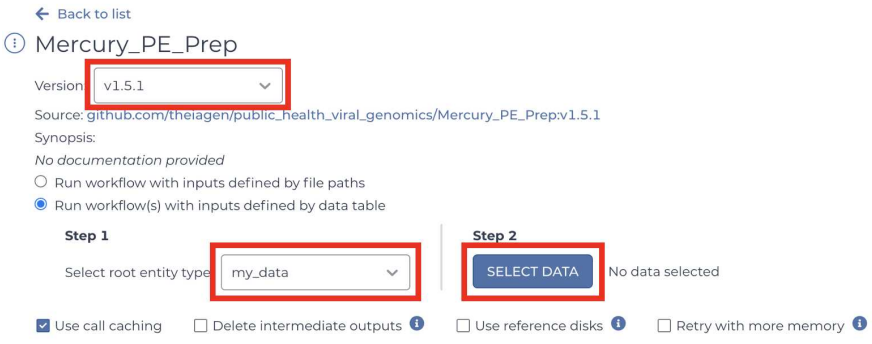
4 Mercury Prep

15m

4.1 Select Mercury_Prep_SE or Mercury_Prep_PE from the **Workflows** tab in your Terra workspace:

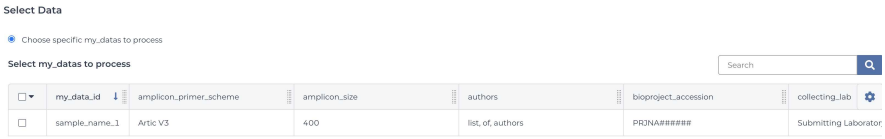
Mercury Paired End and Mercury Single-End

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



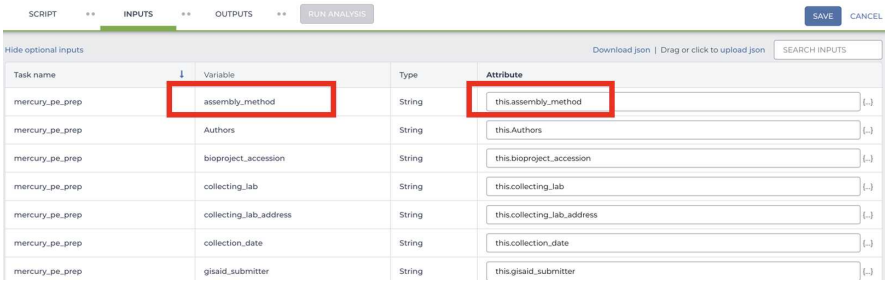
2m

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

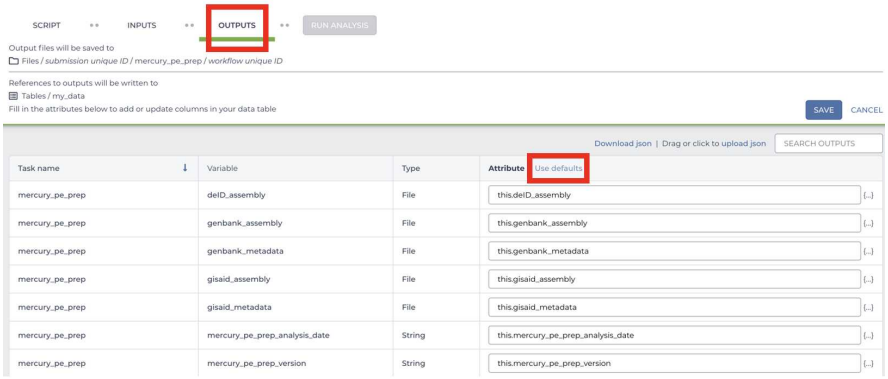


10m

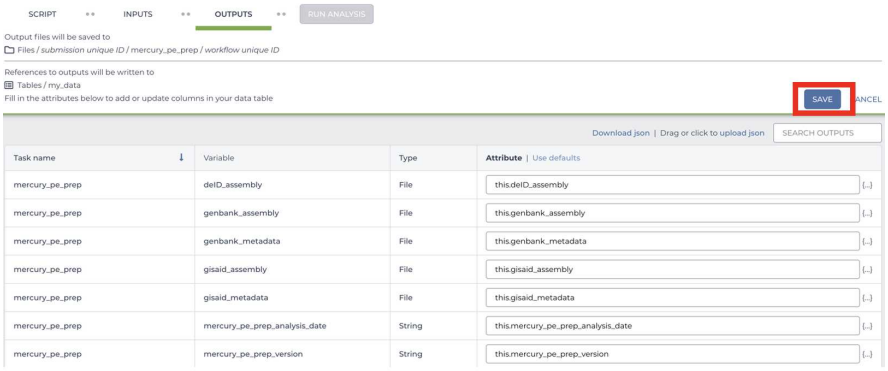
4.4 Enter the input variables:



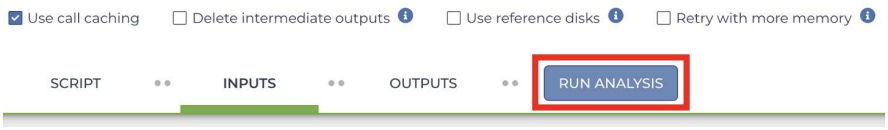
4.5 Select the default outputs:



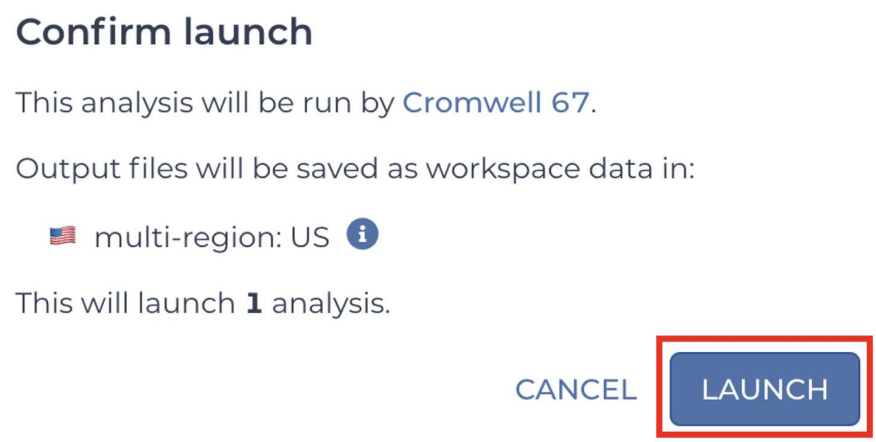
4.6 Once the inputs and outputs have been defined, Save the workflow parameters:



4.7 Click RUN ANALYSIS



Confirm and launch the analysis by clicking LAUNCH:

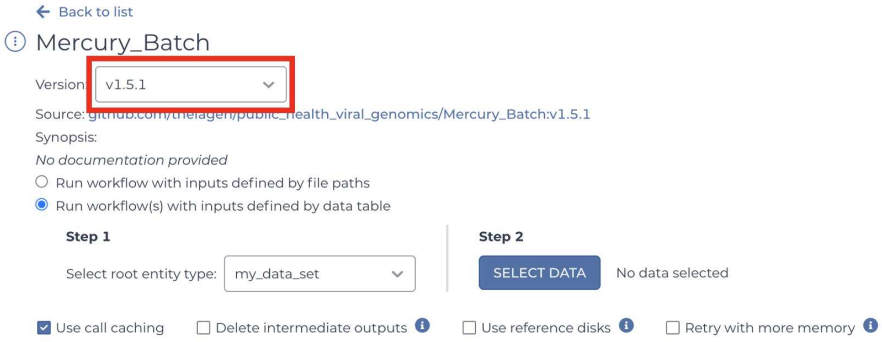


5 Mercury Batch

20m

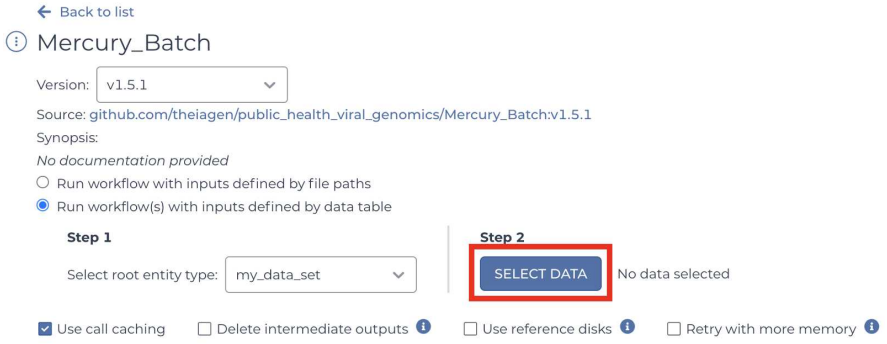
5.1 Once Mercury Prep has successfully completed navigate to the Mercury Batch workflow:

5.2 Select the appropriate version of the workflow:

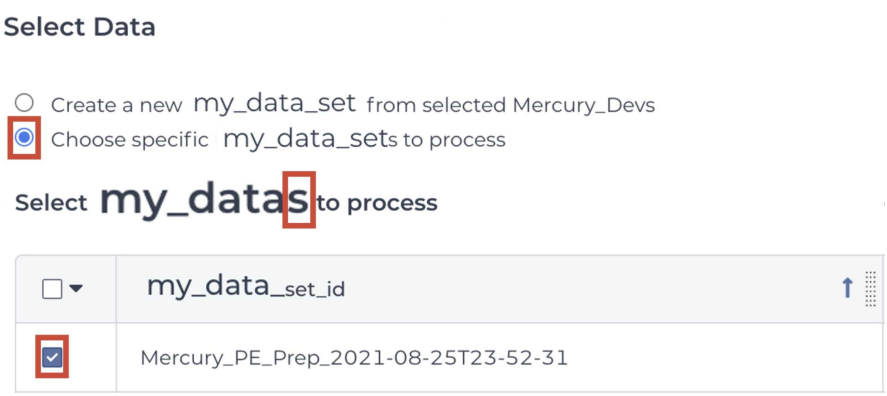


5.3 Select the SET LEVEL root entity type:

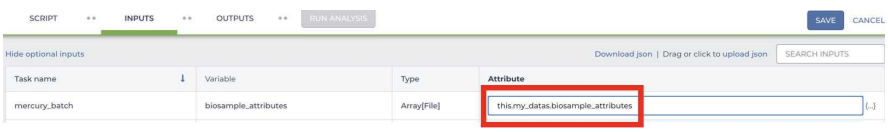
5.4 Click SELECT DATA:



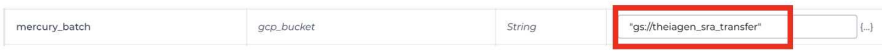
5.5 Select the dataset of sample that you would like to batch for submission (Note: the dataset root entity is the plural form of the original root entity): 2m



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

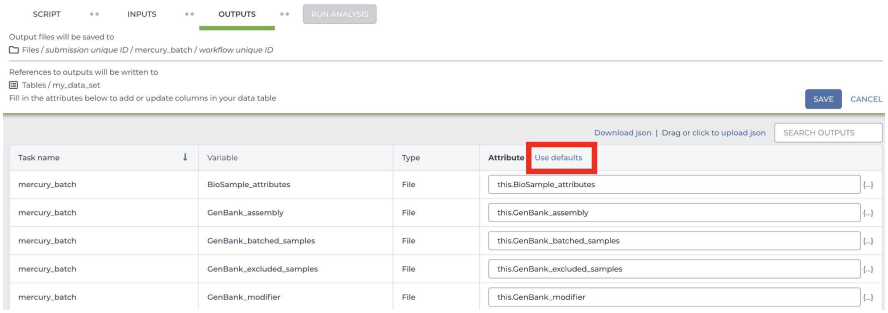


5.7 Enter the public GCP bucket to stage your data for the final submission to NCBI SRA:

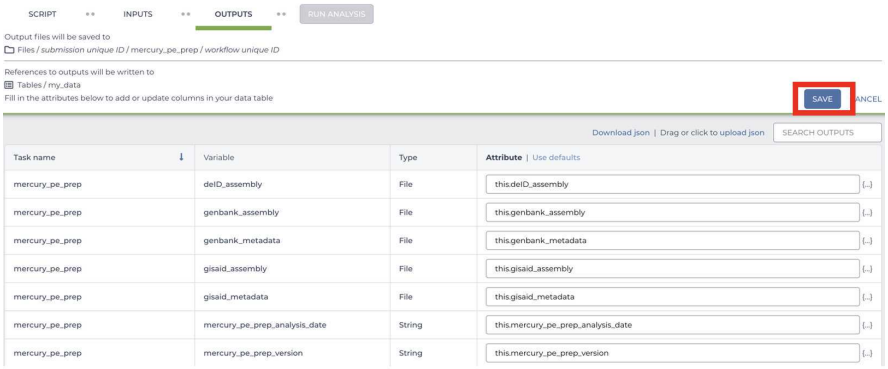


Note: If you are using the GCP bucket submission method please ensure that you have write access to the public Theiagen GCP bucket for NCBI submission:
"gs://theiagen_sra_transfer"

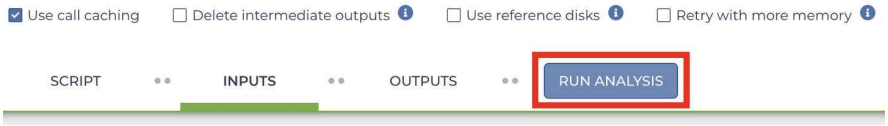
5.8 Select the default OUTPUTS:



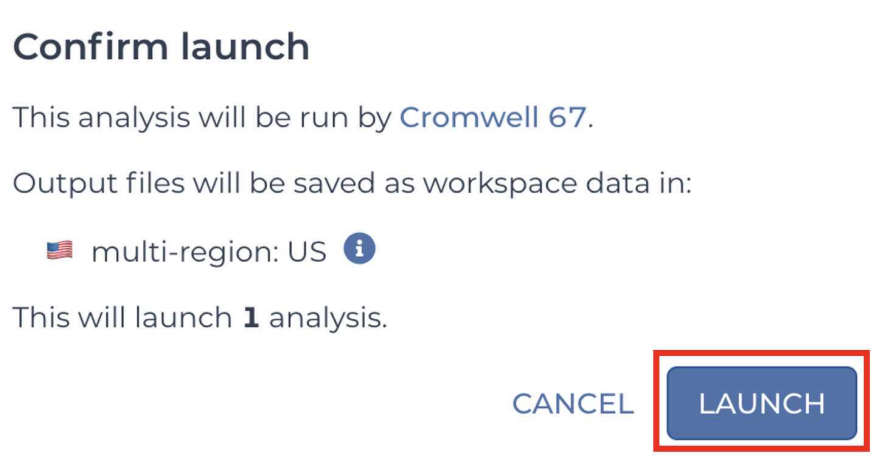
5.9 Once the inputs and outputs have been defined, Save the workflow parameters:



5.10 Click RUN ANALYSIS

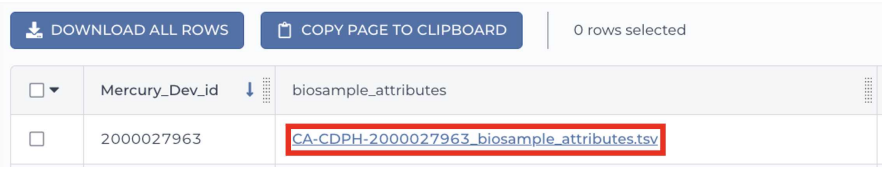


Confirm and launch the analysis by clicking LAUNCH:

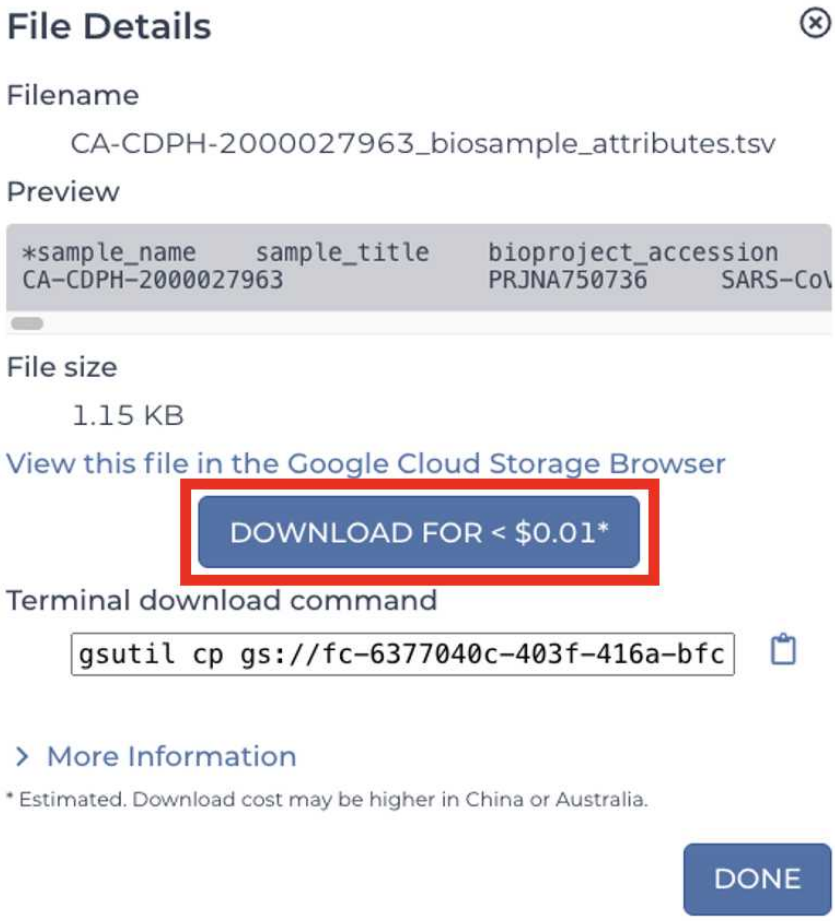


5.11 Retrieve your submission files by navigating to the Terra Data Table containing the Mercury Batch outputs: 5m

Click on the file names in blue:



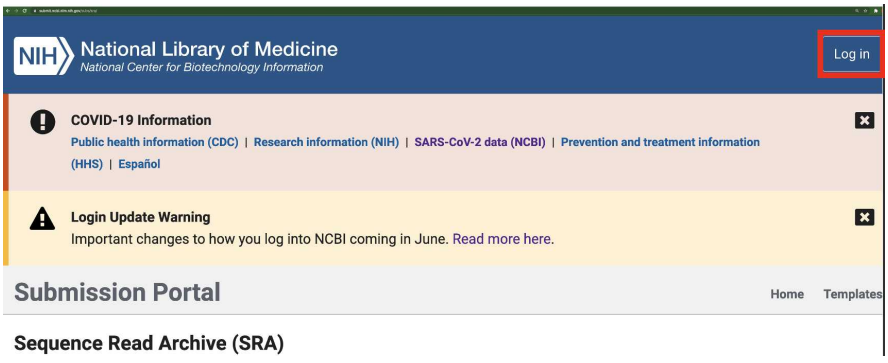
Download the files:



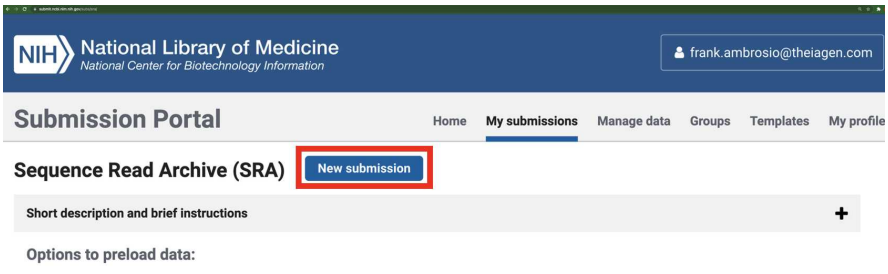
SRA Submission

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

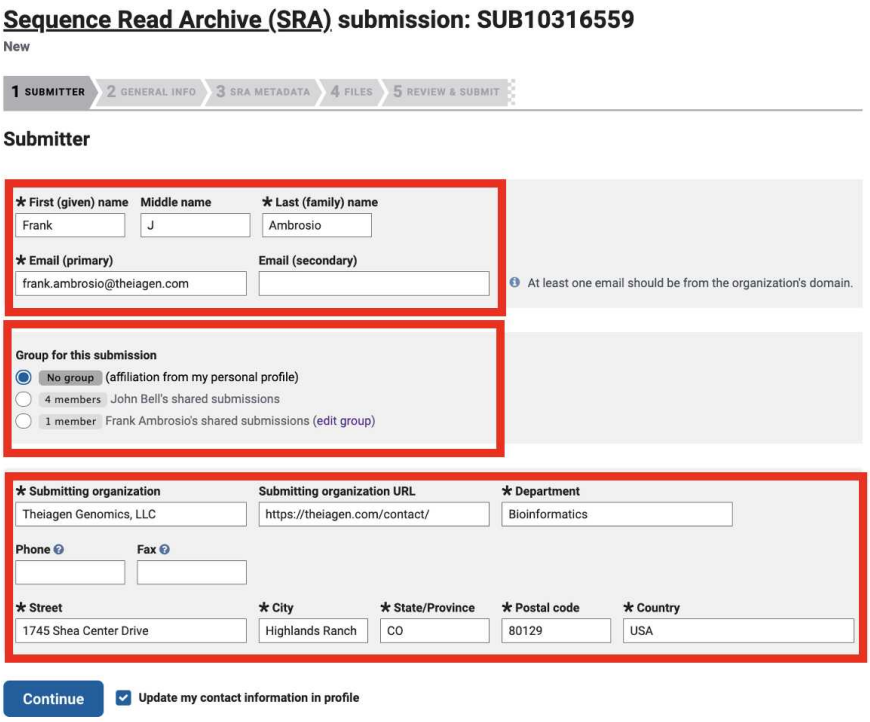
6.1 Navigate and Log in to the SRA Submission Portal:



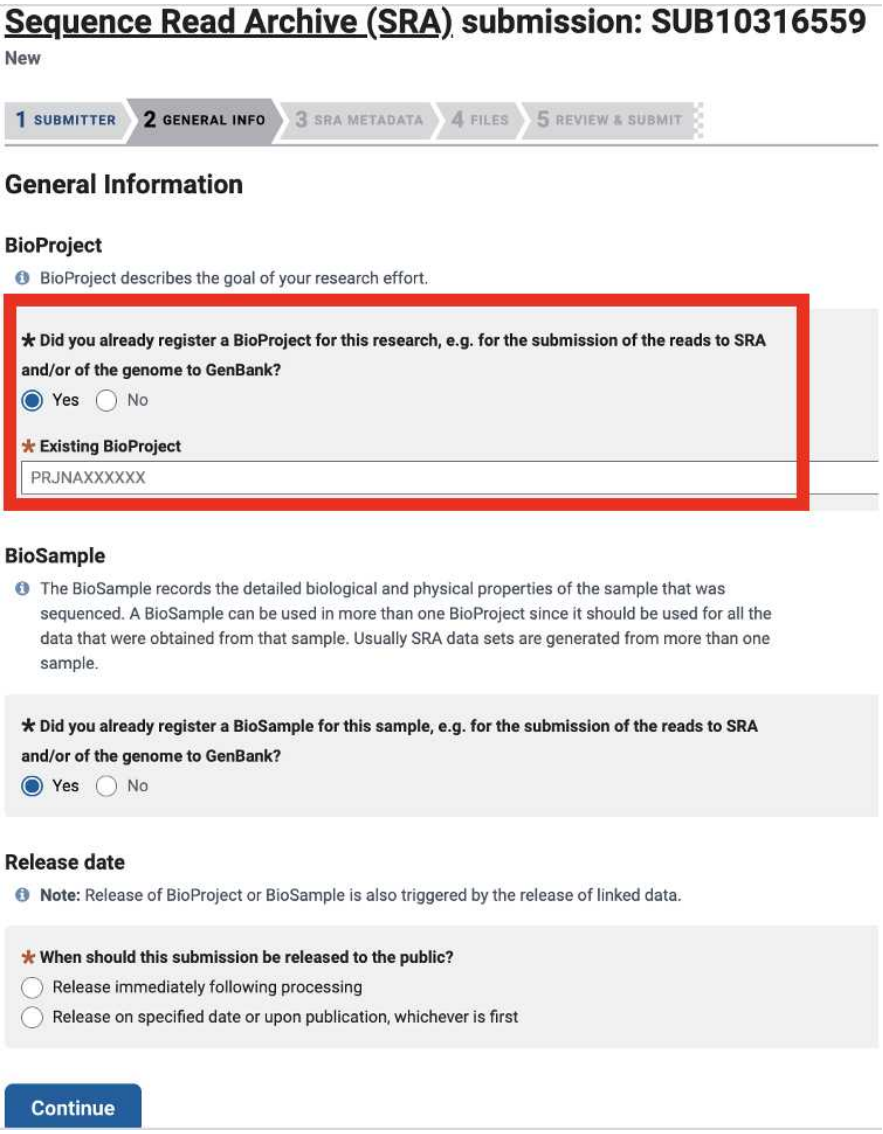
6.2 Select New Submission:



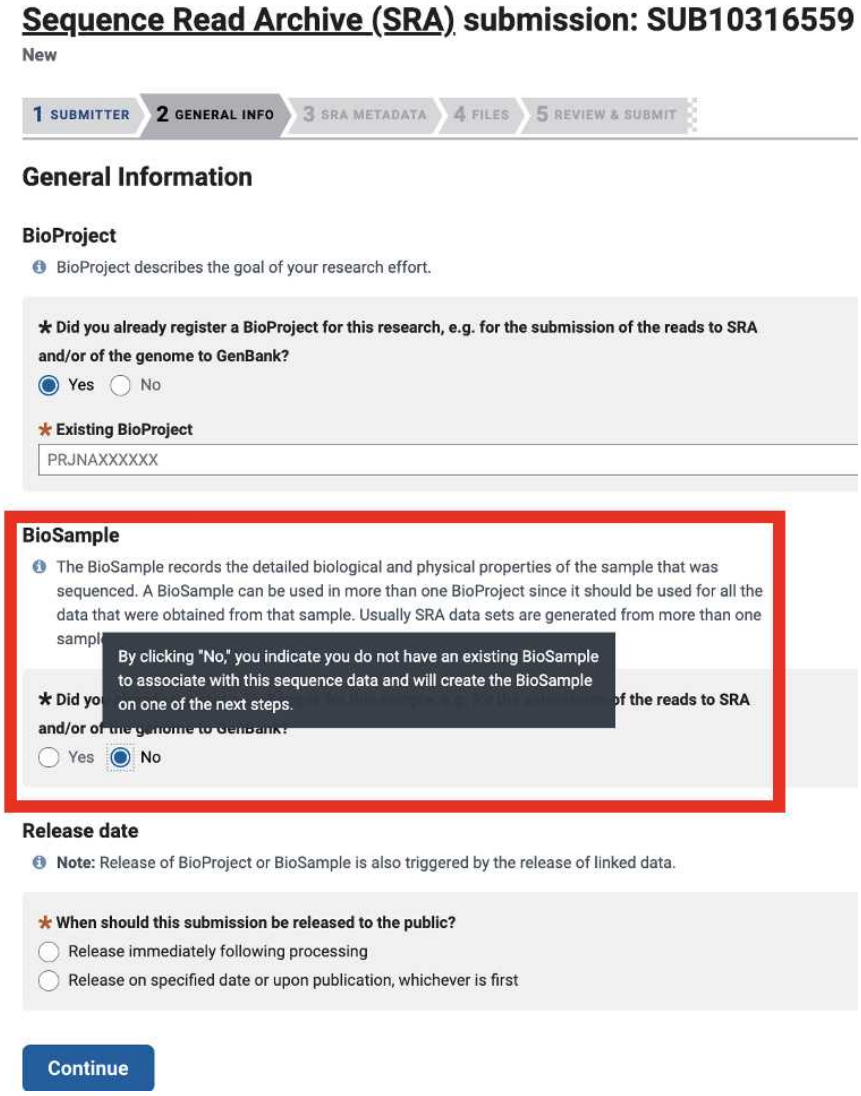
6.3 Enter your submitter information, select your submission group, and enter the information for your organization: 1m



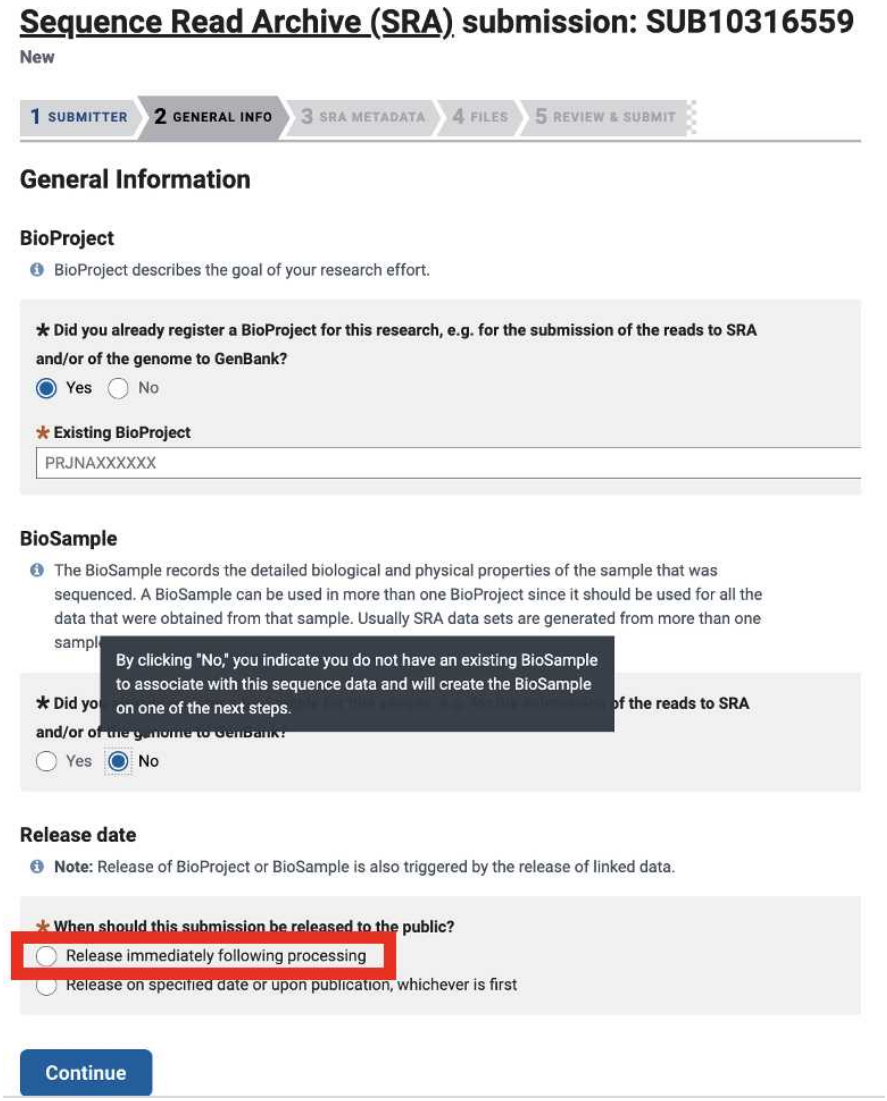
6.4 Enter your BioProject number: 1m



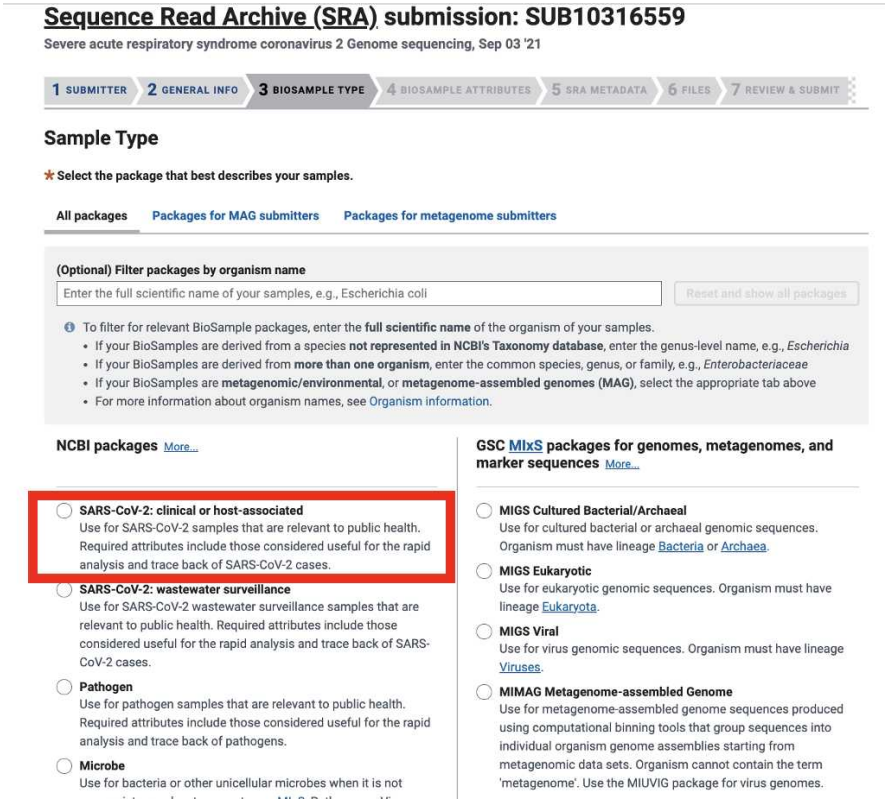
6.5 Select 'No' if you do not already have BioSample accession numbers for your samples in order to generate them upon SRA submission: 1m



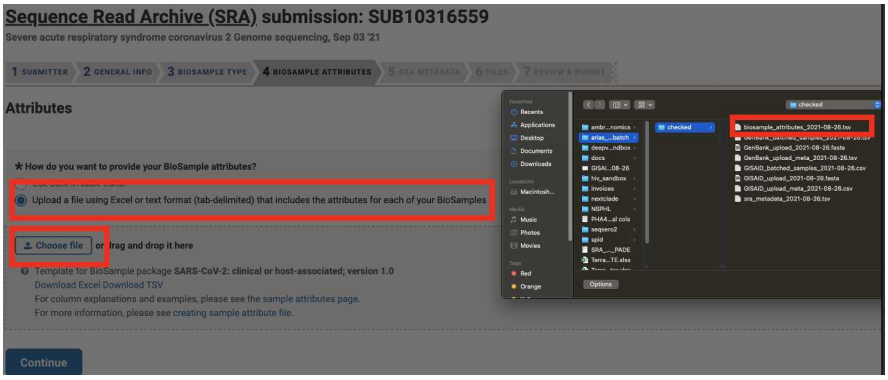
6.6 Select your Release Date (we recommend releasing your data immediately following processing: 1m



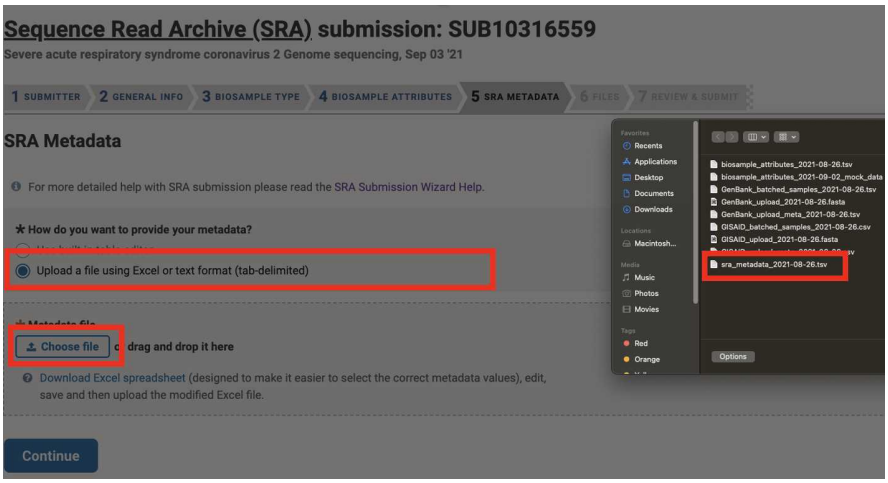
6.7 Select the appropriate submission package (if you are submitting SARS-CoV-2 sequences extracted from a human specimen please select the SARS-CoV-2 clinical or host-associated package): 1m



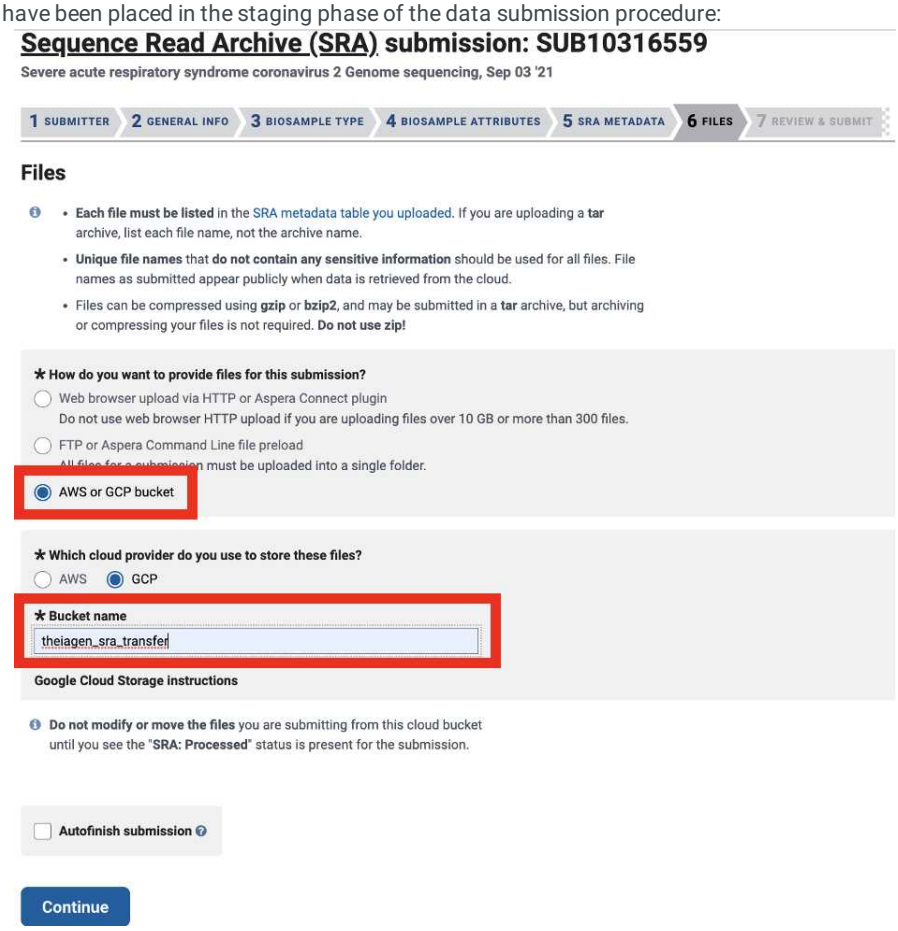
6.8 Choose the 'Upload a file...' option and upload the BioSample attributes file downloaded in previous steps: 1m



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



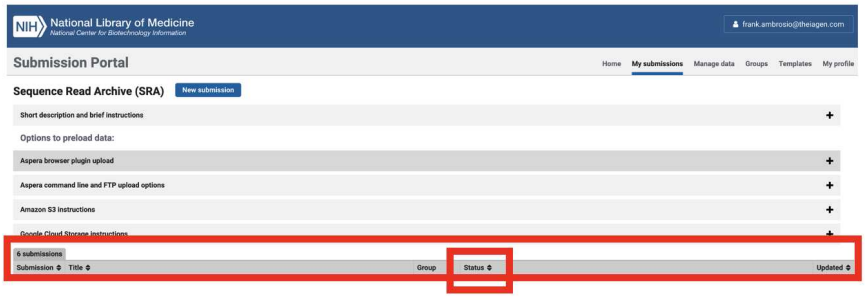
6.10 Select the 'AWS or GCP bucket' option and enter the name of the public data bucket where your reads have been placed in the staging phase of the data submission procedure: 1m



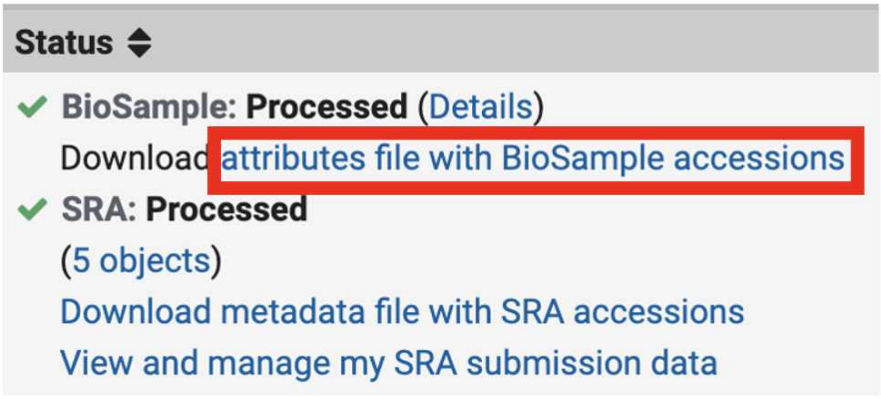
6.11 Review and Submit to complete your SRA submission! You will be able to download your BioSample accession numbers from the SRA submission portal as soon as they become available. 1m

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:



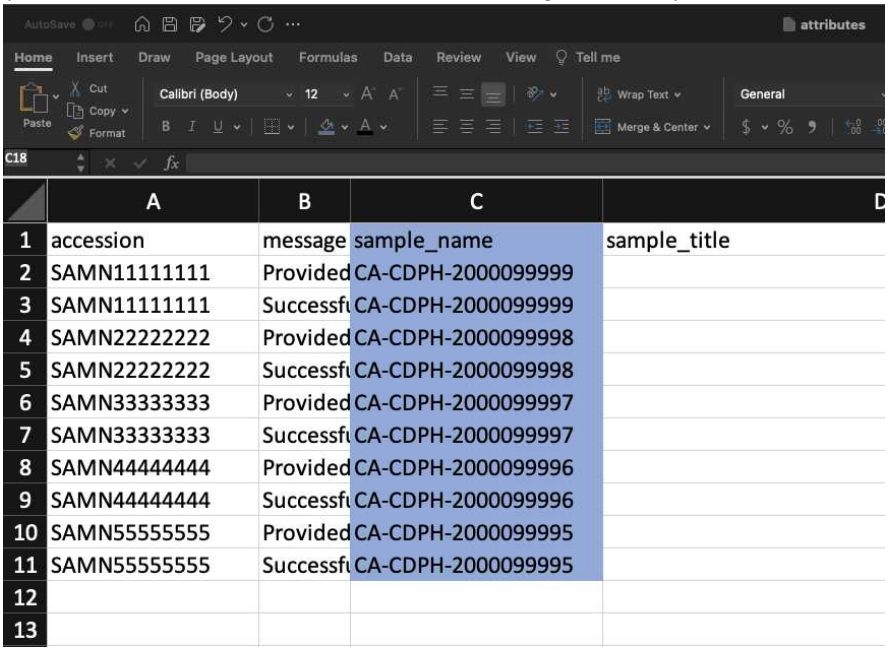
7.2 Click 'Download attributes file with BioSample accessions' for the SRA submission executed earlier in this protocol:



Genbank Submission 40m

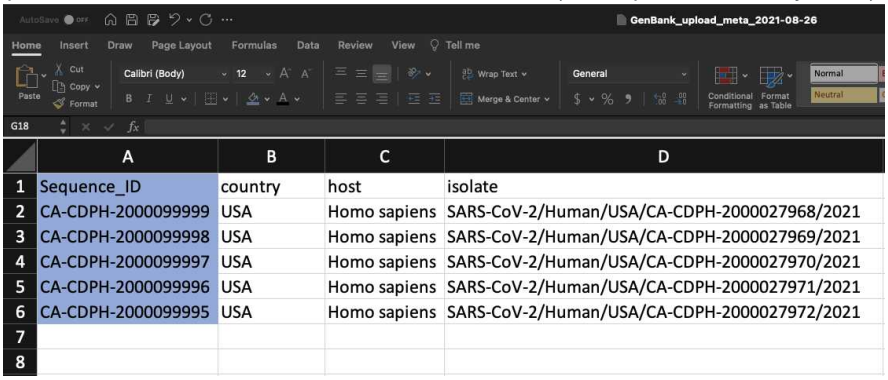
8 Add BioSample accession numbers to Genbank_meta_upload file 5m

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



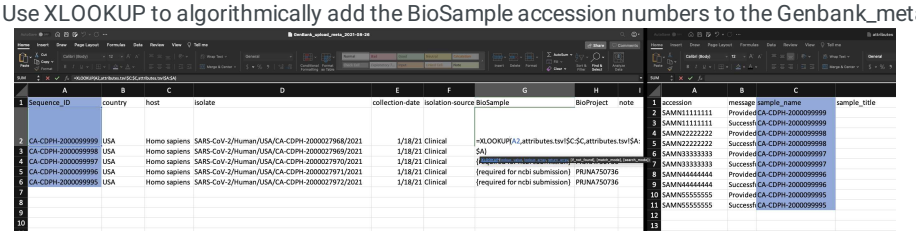
	A	B	C	D
1	accession	message	sample_name	sample_title
2	SAMN11111111	Provided	CA-CDPH-2000099999	
3	SAMN11111111	Successfu	CA-CDPH-2000099999	
4	SAMN22222222	Provided	CA-CDPH-2000099998	
5	SAMN22222222	Successfu	CA-CDPH-2000099998	
6	SAMN33333333	Provided	CA-CDPH-2000099997	
7	SAMN33333333	Successfu	CA-CDPH-2000099997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

8.2 Open the Genbank_meta_sra file downloaded from Terra (the output from Mercury Batch)

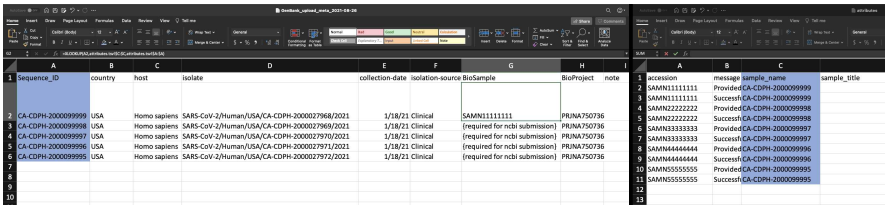


	A	B	C	D
1	Sequence_ID	country	host	isolate
2	CA-CDPH-2000099999	USA	Homo sapiens	SARS-CoV-2/Human/USA/CA-CDPH-2000027968/2021
3	CA-CDPH-2000099998	USA	Homo sapiens	SARS-CoV-2/Human/USA/CA-CDPH-2000027969/2021
4	CA-CDPH-2000099997	USA	Homo sapiens	SARS-CoV-2/Human/USA/CA-CDPH-2000027970/2021
5	CA-CDPH-2000099996	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				

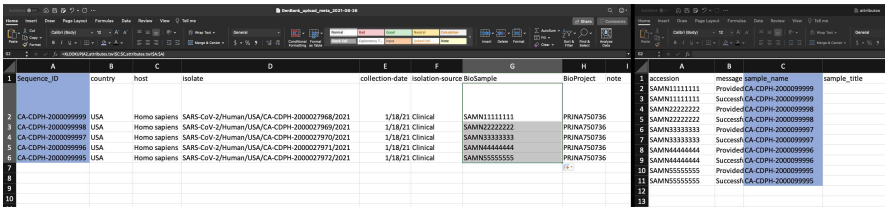
8.3 Use XLOOKUP to algorithmically add the BioSample accession numbers to the Genbank_meta_sra file. 5m



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

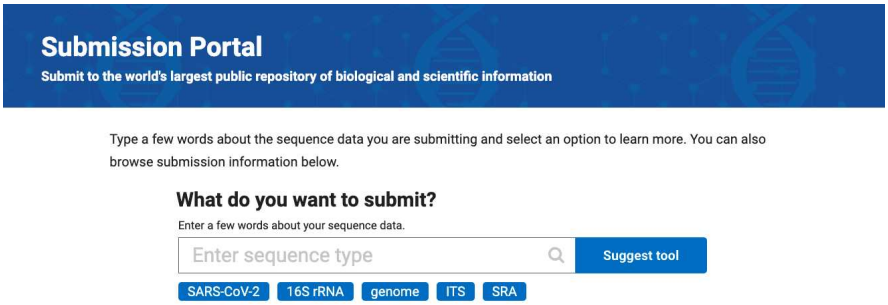


You have successfully added your BioSample accession numbers to the Genbank_meta_upload file

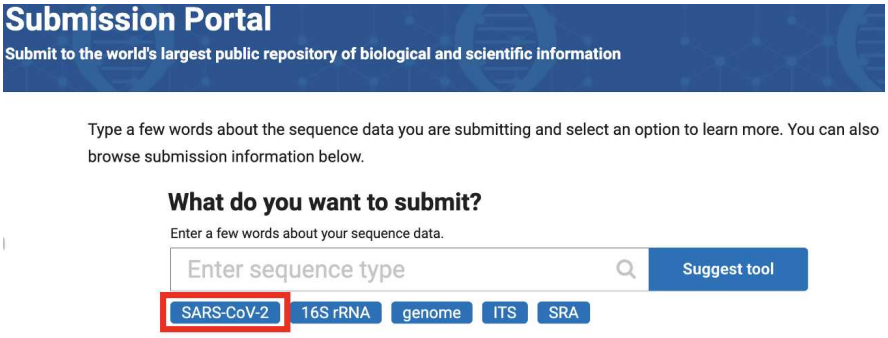
8.4 Save the Genbank_meta_upload file (now including the BioSample accession numbers)

9 Genbank submission 20m

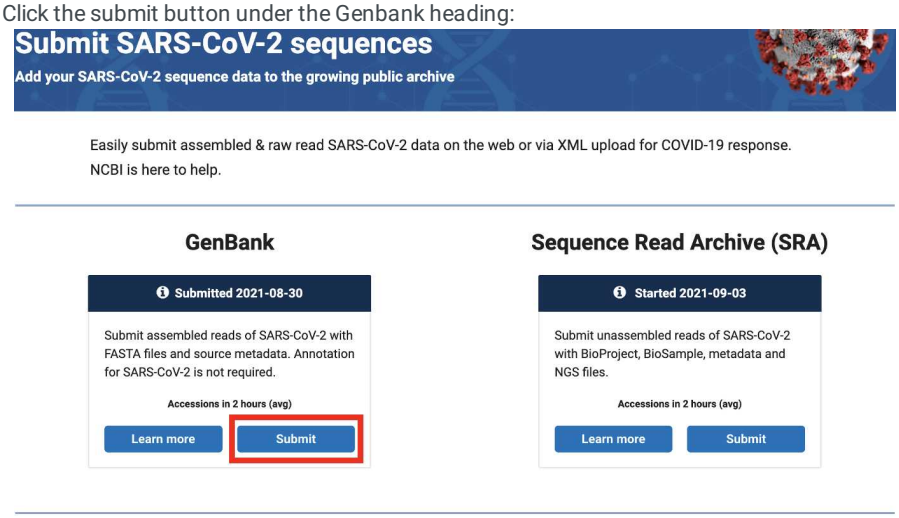
9.1 Navigate to the Genbank submission portal



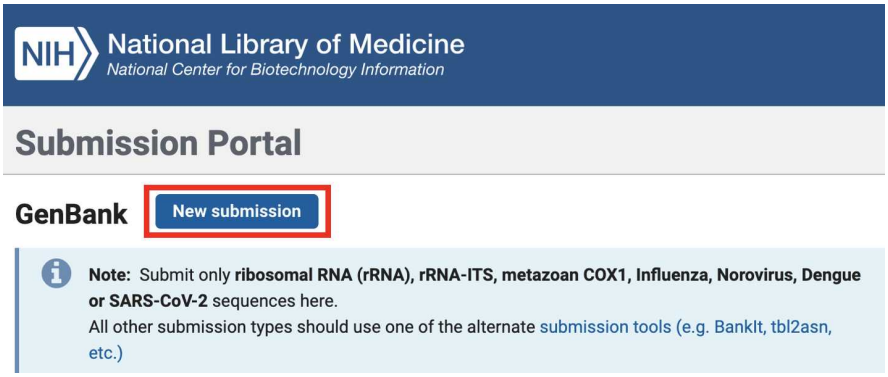
9.2 Select SARS-CoV-2



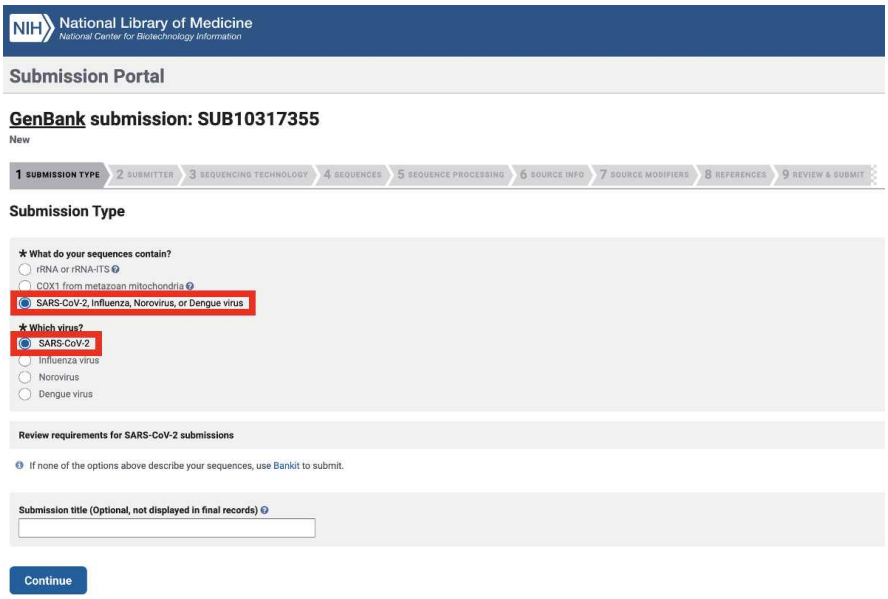
9.3 Click the submit button under the Genbank heading:

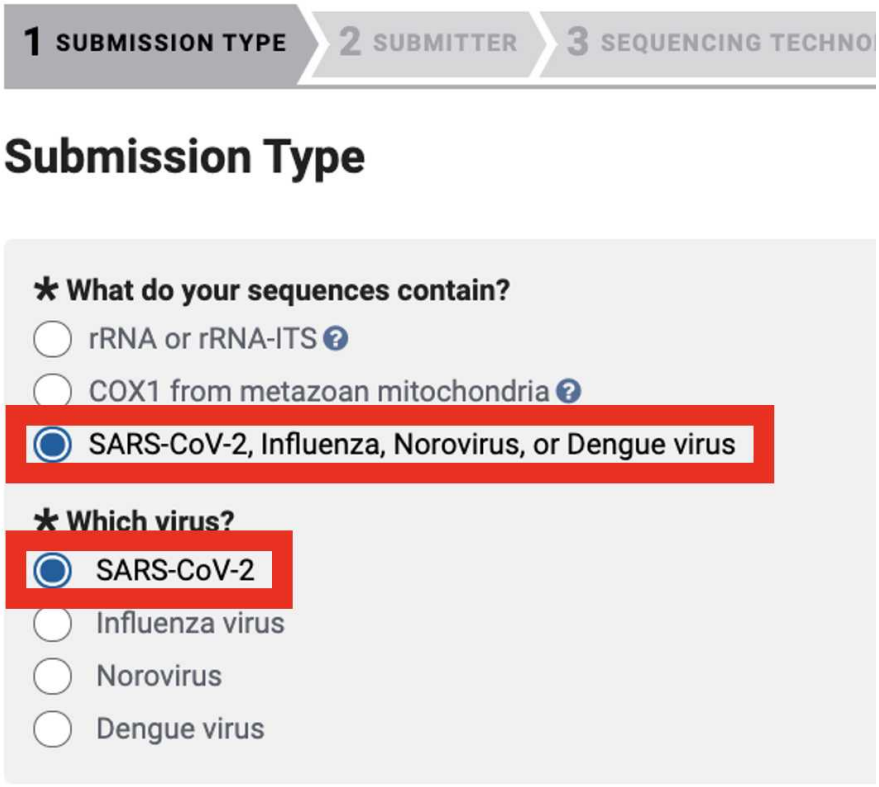


9.4 Select 'New submission':

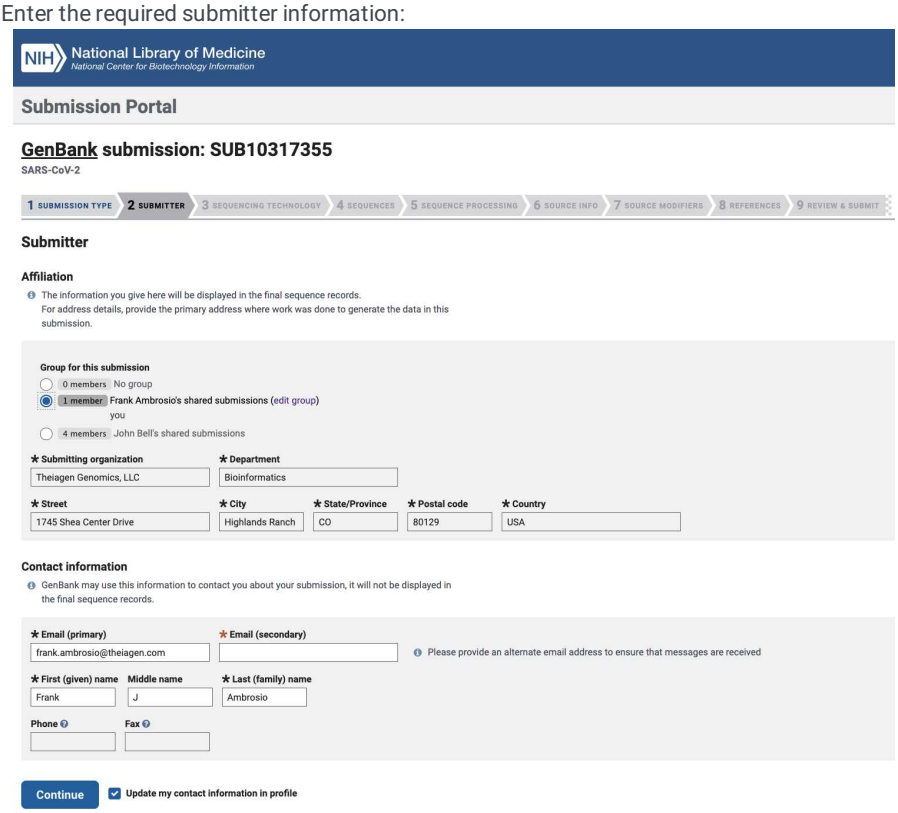


9.5 Select 'SARS-CoV-2, Influenza, Norovirus, or Dengue virus' and 'SARS-CoV-2' t the questions '**What do your sequences contain?**' and '**Which virus?**', respectively. 1m



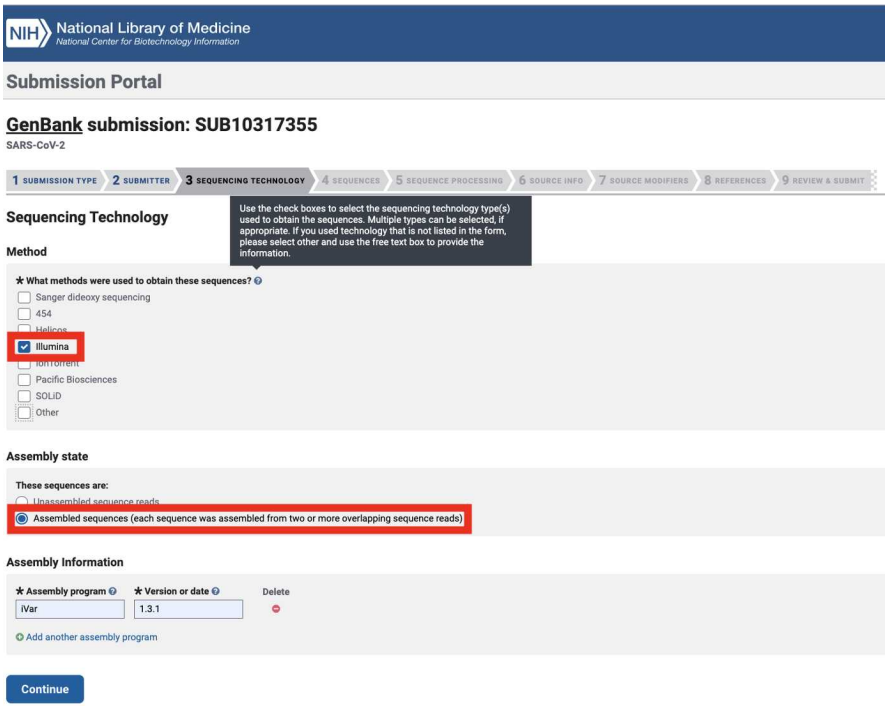


9.6 Enter the required submitter information: 2m

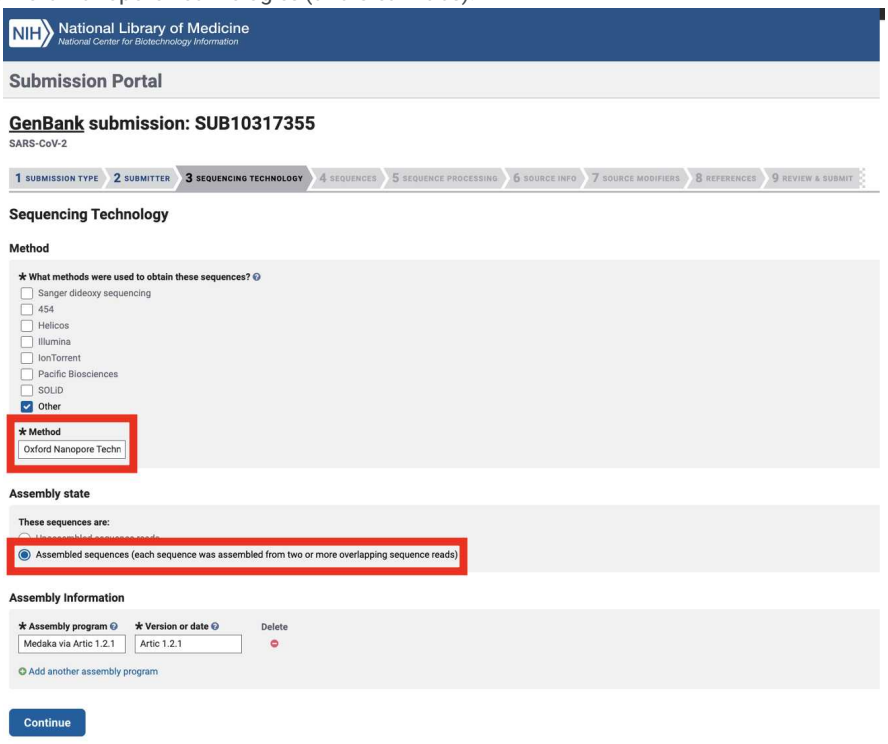


9.7 Select the Sequencing Technology used to generate the sequencing data of which the Genbank assembly submissions are composed. Select 'Assembled sequences (...)' as the assembly state: 1m

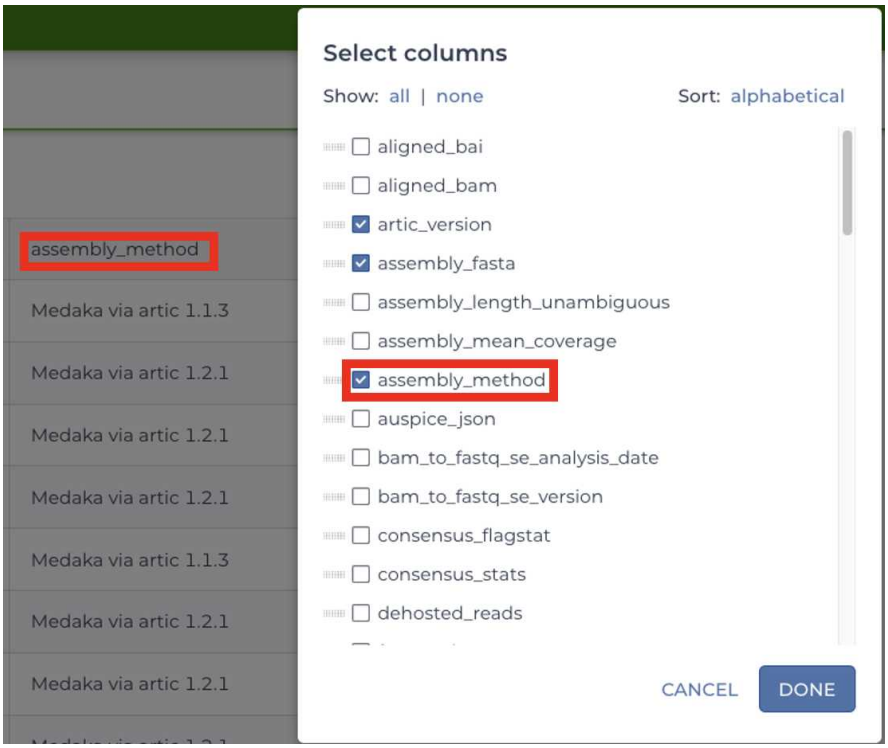
Illumina:



Oxford Nanopore Technologies (and Clear Labs):



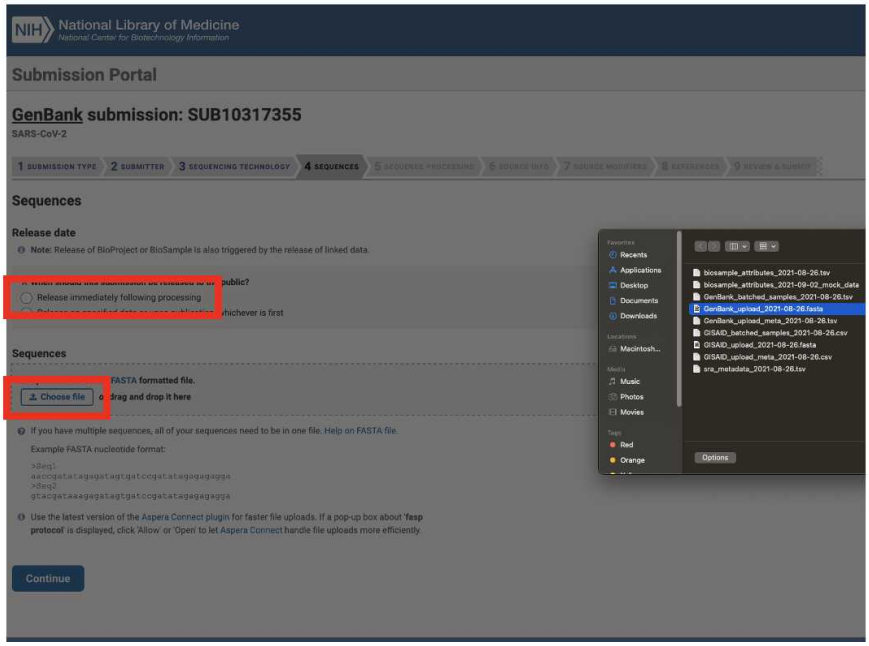
Note: the assembly method is a default output for the Titan Genomic Characterization workflow. The assembly software and version can be found in your Terra data table:

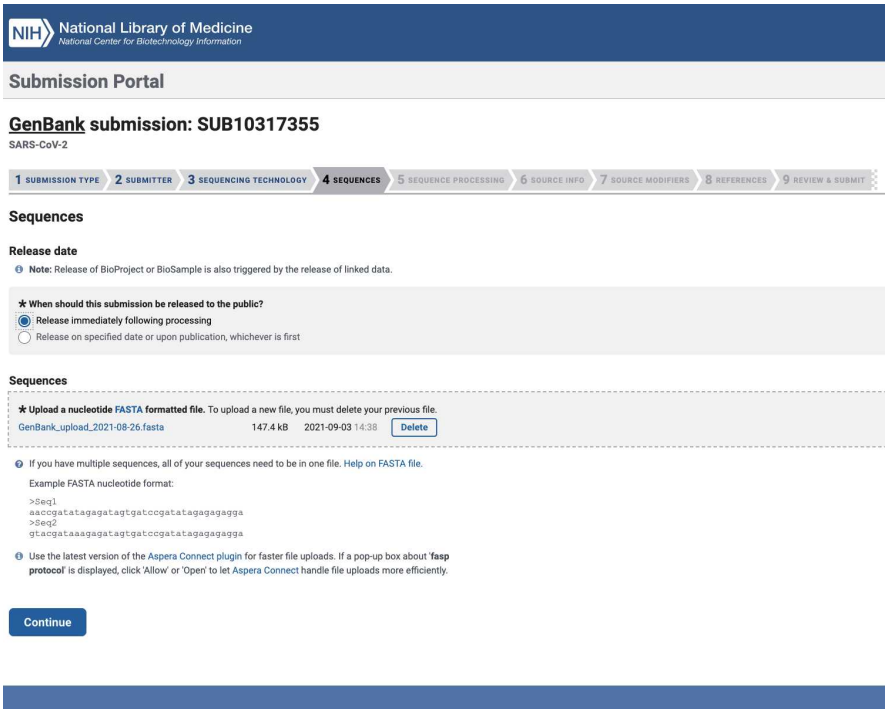


9.8

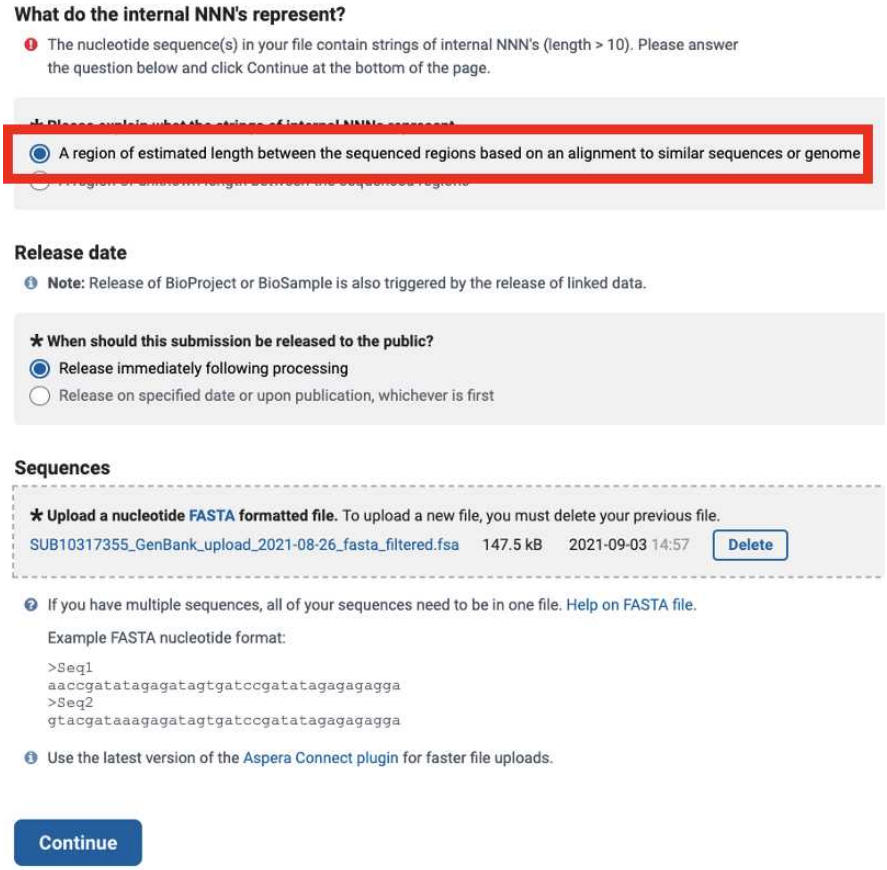
1m

Select 'Release immediately following processing' and upload the Genbank_assembly.fasta file:



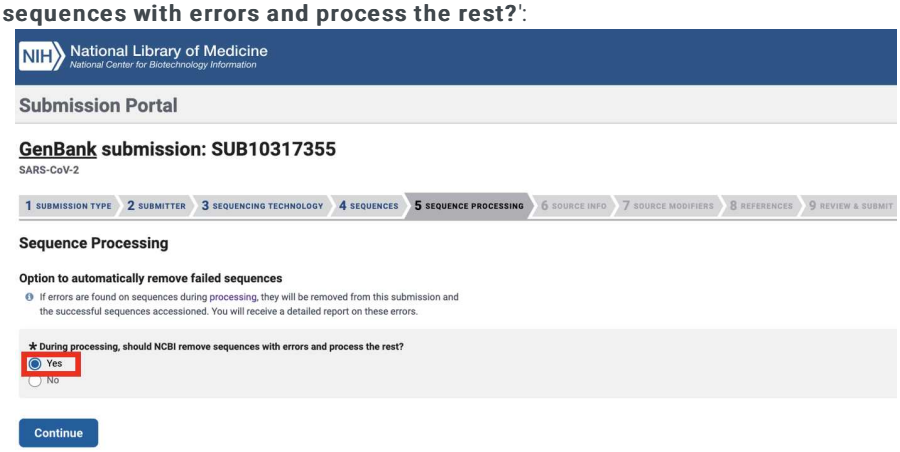


9.9 You will be asked to explain the strings of N's in your assemblies. The software used by the Titan Genomic Characterization Workflows estimates the length between sequenced regions using the Wuhan-1 reference genome for alignments: 1m

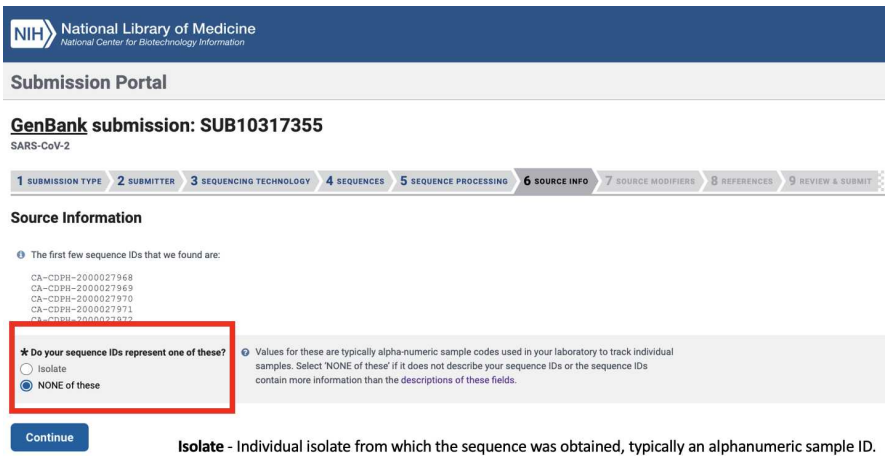


We recommend selecting yes for the question '**During processing, should NCBI remove** 1m

9.10 **sequences with errors and process the rest?'**:

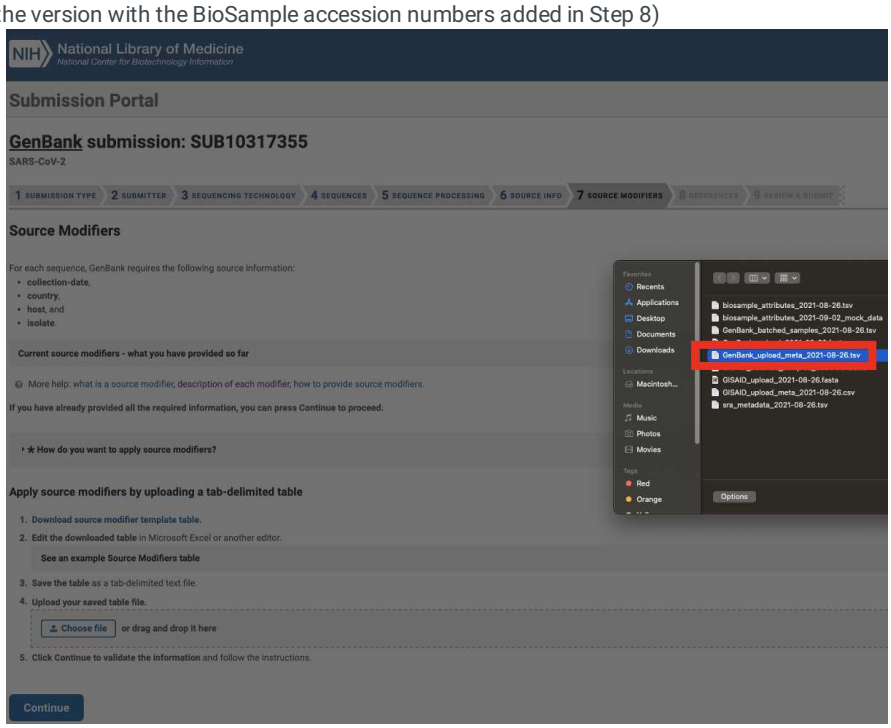


9.11 Indicate whether the source of your genomic material was an individual isolate: 1m

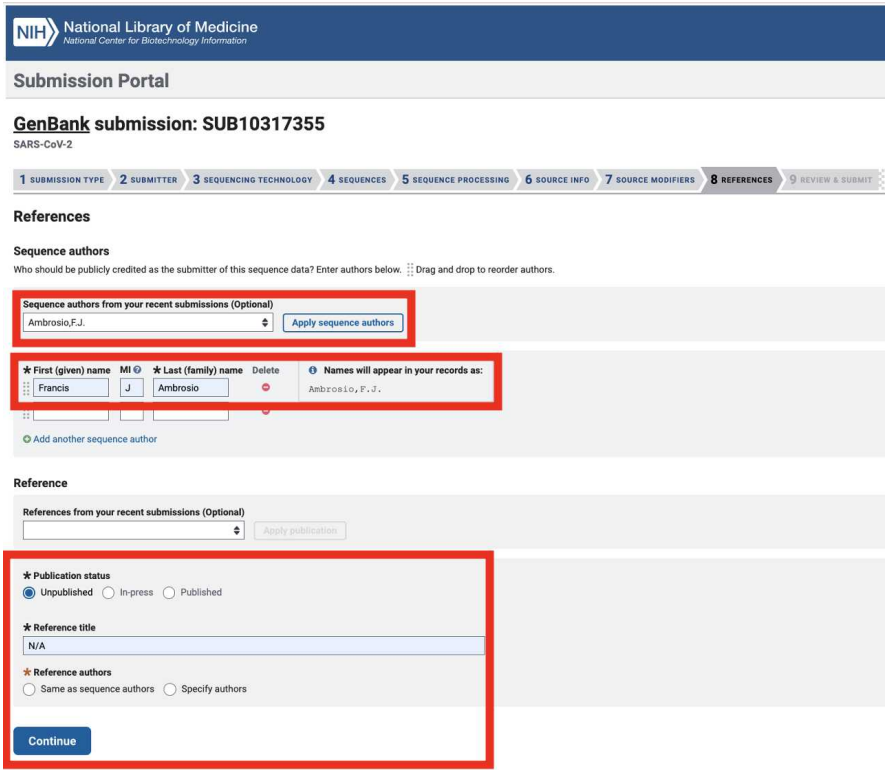


Isolate - Individual isolate from which the sequence was obtained, typically an alphanumeric sample ID.

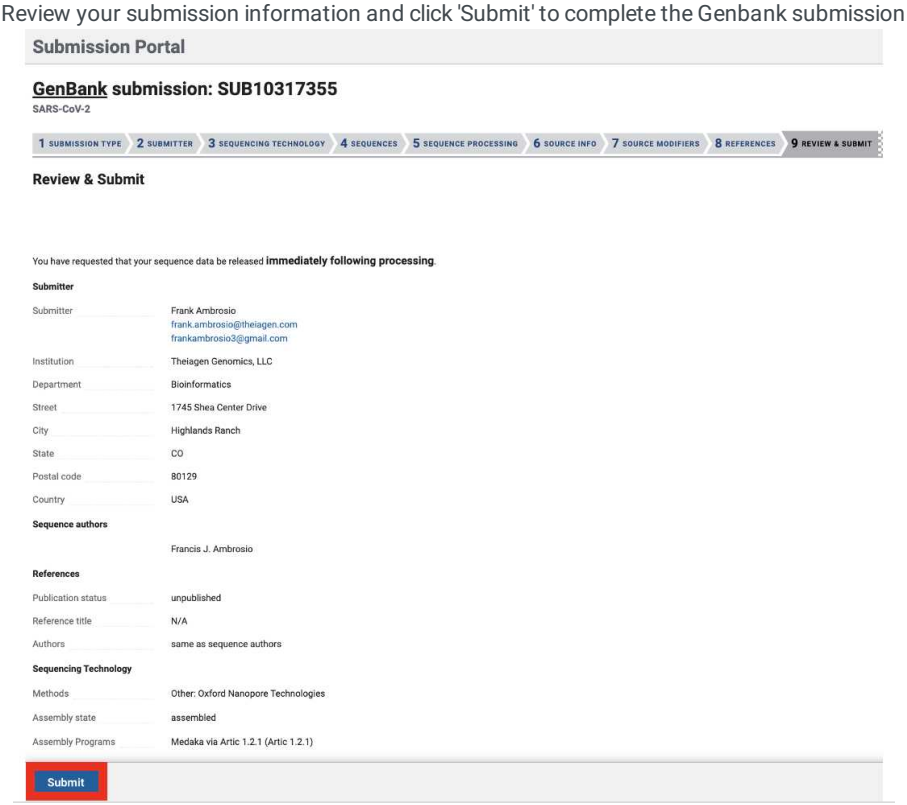
9.12 Upload the Genbank_meta_upload file downloaded from Terra in previous steps (Note: This should be the version with the BioSample accession numbers added in Step 8) 1m



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



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



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!