

2 ▼

Oct 19, 2021

# Mercury Sequence and Sample Metadata Prep for Submission Workflow on the Terra Platform V.2

Jill V Hagey<sup>1</sup>, Kevin Libuit<sup>2</sup>, Lingzi Xiaoli<sup>1</sup>,  
Technical Outreach and Assistance for States Team<sup>1</sup>

<sup>1</sup>Centers for Disease Control and Prevention; <sup>2</sup>Theiagen Genomics



protocol .



Jill Hagey  
Centers for Disease Control and Prevention

The opinions expressed here do not necessarily reflect the opinions of the Centers for Disease Control and Prevention or the institutions with which the authors are affiliated. The protocol content here is under development and is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to [protocols.io](https://protocols.io) is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with [protocols.io](https://protocols.io), can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

Public health laboratories are encouraged to submit sequencing data for SARS-CoV-2 to multiple public data repositories to ensure broad access by the research community and public health institutions for rapid, genomics-driven public health response to disease outbreaks. There are three primary public repositories for depositing SARS-CoV-2 sequencing data:

1. Global initiative on sharing all influenza data (GISAID) EpiCov,
2. National Center for Biotechnology Information (NCBI) GenBank, and
3. NCBI SRA (sequence read archive).

Laboratories are encouraged to submit consensus assemblies to both GISAID and GenBank, as well as submit raw (filtered) sequencing reads to NCBI SRA. [NCBI](#) is a division of the National Library of Medicine within the US government's National Institutes of Health and organizes a breadth of biomedical resources including publicly accessible genome sequences, annotations, analyses, and derived data. [GISAID](#) was created as a framework for openly sharing influenza (and now SARS-CoV-2) sequence data, while maintaining strict governance over the data use and attribution to sequence submitters.

The Mercury workflows are designed to aid submission to public repositories as explained in Theiagen's YouTube series:

Mercury Overview: <https://www.youtube.com/watch?v=h8YASVckOrw>  
Mercury Tutorial Part 1: [https://www.youtube.com/watch?v=nFJT\\_QEk25s](https://www.youtube.com/watch?v=nFJT_QEk25s)  
Mercury Tutorial Part 2: <https://www.youtube.com/watch?v=mtBXbbT3vPg>

**Titan Mercury Workflows:** This workflow will take in reads, consensus sequence assemblies, and sample metadata and package all of those inputs into repository submittable files for submission to GISAID and GenBank (SRA coming soon). This protocol first starts with preparing SE (single-end) or PE (paired-end) reads for submission and then moves to prepare all the samples into a batch for bulk submission as you likely have many samples to submit.

**Mercury SE/PE Prep** - prepares assembly and sample metadata files for samples individually from either single-end or paired-end sequencing runs.

- Will only generate prepped files for assemblies with <5,000 Ns

**Mercury Batch** - concatenates assemblies (fasta files) and combines metadata sheets for batch submissions

- Will only batch samples with 0 VADR alerts

For technical assistance, please contact: **TOAST@cdc.gov**

Jill V Hagey, Kevin Libuit, Lingzi Xiaoli, Technical Outreach and Assistance for States Team 2021. Mercury Sequence and Sample Metadata Prep for Submission Workflow on the Terra Platform. **protocols.io**  
<https://protocols.io/view/mercury-sequence-and-sample-metadata-prep-for-subm-by74pzqw>  
Technical Outreach and Assistance for States Team



Illumina, Sequencing, MiSeq, iSeq, MiniSeq, NextSeq, NovaSeq, Paired-End, Next Generation Sequencing, NGS, SARS-CoV-2, Covid, Pangolin, Short-Read, Coronavirus, Genomics, Genetics, Virology, Molecular Biology, sequence submission, GISAID, GenBank, SRA

protocol ,

Oct 19, 2021

Oct 19, 2021

Oct 19, 2021



Jill Hagey

Centers for Disease Control and Prevention

54236

:

The opinions expressed here do not necessarily reflect the opinions of the Centers for Disease Control and Prevention or the institutions with which the authors are affiliated. The protocol content here is under development and is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to [protocols.io](https://protocols.io) is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with [protocols.io](https://protocols.io), can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

## Where to Begin

- 1 **This workflow is intended for users of Titan workflows for SARS-CoV-2 strain characterization on the Terra platform and expects the following protocols have already been completed:**

- Terra and Google Cloud Accounts
- Samples uploaded
- Output from the Titan workflow

**This protocol is intended only for submitting sequences derived from human clinical specimens.**

Please refer to the following documentation for submission criteria and minimum quality control thresholds.

GenBank Submission Criteria: [About GenBank Submission \(nih.gov\)](https://www.ncbi.nlm.nih.gov/genbank/submission/)

GISAID Submission Criteria: [Gisaaid inclusion criteria.pdf](#)

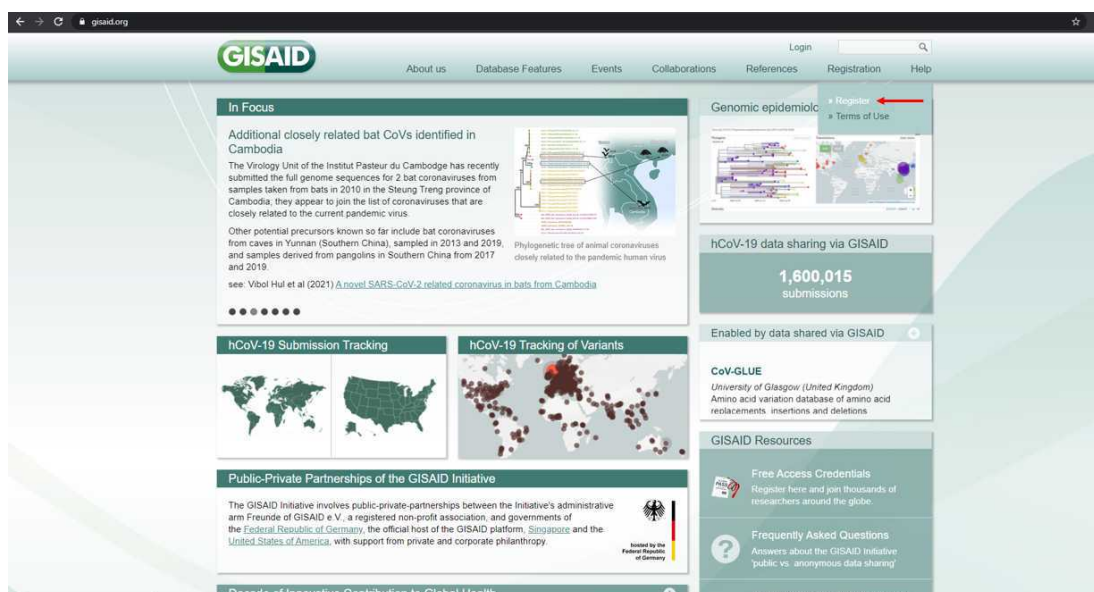
The Association of Public Health Laboratories (APHL) has also summarized these differences here: [Recommendations for SARS-CoV-2 Sequence Data Quality & Reporting](https://www.aphl.org/Portals/0/Documents/2020-04-01_SARS-CoV-2_Sequence_Data_Quality_Reporting.pdf).

## GISAID Registration

- 2 Submitting to GISAID EpiCoV requires registration. To create an account, navigate to the GISAID webpage and follow the steps below to register.

<https://www.gisaid.org/>

Hover over the 'Registration' tab near the top right corner and click 'Register' in the drop down menu:



GISAID home page. Click 'Register' in the drop down menu

Read the terms of use and then click 'Register' to continue.

The screenshot shows the GISAID website's registration process. The header includes the GISAID logo and navigation links: About us, Database Features, Events, Collaborations, References, Registration, and Help. A search bar is also present. On the left, a sidebar menu shows 'Registration' with sub-links for 'Register' and 'Terms of Use'. The main content area is titled 'The Signup Process' and features a three-step diagram: 1. SUBMIT (Applicant provides basic information and agrees to the Database Access Agreement), 2. REVIEW (The information is reviewed both automatically and manually to confirm the applicant's identity), and 3. APPROVE (Applicant receives access credentials, or if necessary a request to assist in the confirmation of identity). Below the diagram, a 'Read carefully:' section contains a paragraph about the Database Access Agreement (DAA) and a list of four rules. The rules are: 1. Acknowledge data contributors (Originating and Submitting Laboratories). 2. Do not attach restrictions on data made available through GISAID. 3. Do not distribute GISAID data outside the GISAID community. 4. Collaborate appropriately with the Originating Laboratory. Below the rules, there are two paragraphs of advice: one about completing the registration form to avoid losing priority, and another about using institutional email addresses for faster review. At the bottom, a recommendation for browsers (Google Chrome, Mozilla Firefox) is given. A red rectangular box highlights the 'Register' button at the bottom right of the main content area.

**GISAID** Login

About us Database Features Events Collaborations References Registration Help

Registration  
» Register  
» Terms of Use

### The Signup Process

- 1 SUBMIT**  
Applicant provides basic information and agrees to the Database Access Agreement
- 2 REVIEW**  
The information is reviewed both automatically and manually to confirm the applicant's identity
- 3 APPROVE**  
Applicant receives access credentials, or if necessary a request to assist in the confirmation of identity

**Read carefully:**

To receive your personal access credentials to the GISAID platform you must first positively identify yourself and agree to the terms of the [Database Access Agreement \(DAA\)](#) which call on all users to support the underlying principles in GISAID that facilitate the sharing of genetic sequence and related data, while recognizing the contributions and interests of data providers and users, including:

1. You must acknowledge data contributors, i.e. the Originating Laboratory where the clinical specimen or virus isolate was first obtained and the Submitting Laboratory where sequence data have been generated and submitted to a GISAID Database ([see sample of acknowledgement table](#)).
2. You may not attach restrictions on the data made available through a GISAID Database, such as including in a patent application any fraction of the genetic sequence data obtained from GISAID, to ensure unlimited access to the data.
3. You may not distribute GISAID data outside the GISAID community, such as by releasing genetic sequences obtained in GISAID in any publication, transferring the data to colleagues that are not registered users, or offering GISAID data on a server accessible by others who are not duly registered with GISAID.
4. You need to collaborate where appropriate with the Originating Laboratory responsible for obtaining the specimens.

Please complete the registration form and help GISAID protect the use of your identity and the integrity of its user base.

To avoid losing your priority in line for access credentials, you are strongly advised to properly complete the registration form. Incomplete or misspelled names or numbers require additional work for the reviewer, thereby potentially delaying other registration request.

Using webmail addresses for your identification, such as Gmail, Hotmail, Yahoo Mail, QQmail or 126 and 163 Mail, are likely to lower your priority or cause significant delays in reviewing your request. You are encouraged to provide your institutional email address to expedite this process.

We recommend using the following browsers: Google Chrome; Mozilla Firefox.

**Register**

GISAID registration overview and terms of use page.

Fill out the information on the registration page, scroll to the bottom to accept the DATABASE ACCESS AGREEMENT, and click 'Register' to continue.

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

**Registration**

Use your company or research institute's email! Web emails such as Gmail or Yahoo will lead to significant delays in processing.

Institution\*

Department

Street 1\*

Postal code

City\*

Location\*

State/province

Telephone\*

Fax

Mobile

E-Mail\*

**GISAID EPIFLU™ DATABASE ACCESS AGREEMENT**

Effective: March 16, 2011

**WHEREAS** Freunde von GISAID e.V. ("GISAID") maintains a global database for influenza gene sequences along with associated data, including virological, clinical, epidemiological and demographic information (if available) for all influenza viruses, including but not limited to H5N1 sequences, (the "GISAID EpiFlu™ Database") for the purpose of facilitating the sharing, research and investigation of such sequences and associated data.

**NOW, therefore**, this Database Access Agreement (the "Agreement") is entered into by and between the undersigned ("You") and GISAID.

- 1. Access to the GISAID EpiFlu™ Database, Data.** Access to, and use of, the GISAID EpiFlu™ Database and Data, as defined herein, is governed by this Agreement. By accessing or otherwise using the GISAID EpiFlu™ Database, whether as a provider or user of Data, You accept and agree to be bound by the terms of this Agreement. For purposes of this Agreement, the term "Data" means any and all (i) sequence data and other associated data and information contained in the GISAID EpiFlu™ Database pertaining to influenza viruses, (ii) any annotations, corrections, updates, modifications, improvements, derivatives or other enhancements to any such data contained in the GISAID EpiFlu™ Database, and (iii) any safety information relevant to use of the data or to regulatory approval of vaccines or other therapies that embody or utilize the data contained in the GISAID EpiFlu™ Database.
- 2. License Terms.** You are hereby granted a non-exclusive, worldwide, royalty-free, non-transferable and revocable license to access and use the GISAID EpiFlu™ Database and Data.

You can also choose to view [the PDF version](#) in a new window.

I accept the DAA ☐

**Register**

© 2008 - 2021 | Terms of Use | Privacy Notice

GISAID registration form

You will be notified by email when your account is set up and ready to use. This might take a day or two.

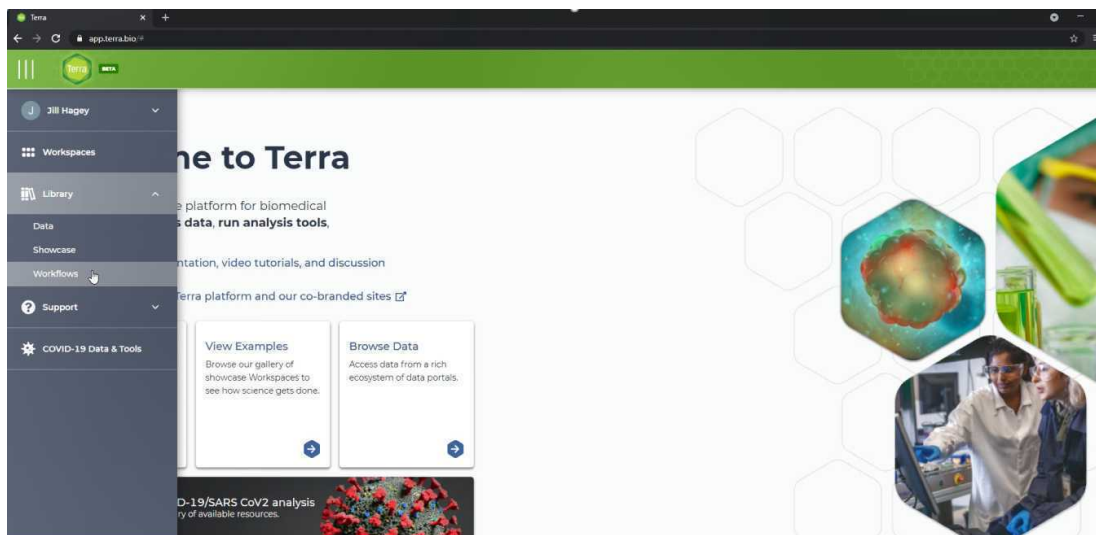
## Import Mercury workflows from Dockstore

### 3 Importing the Mercury Workflow from Dockstore to the User Workspace

We will first walk through steps to import the Mercury workflow, and at the end there is a video showing the full process. You should already have a workspace created from when you ran the Titan workflow.

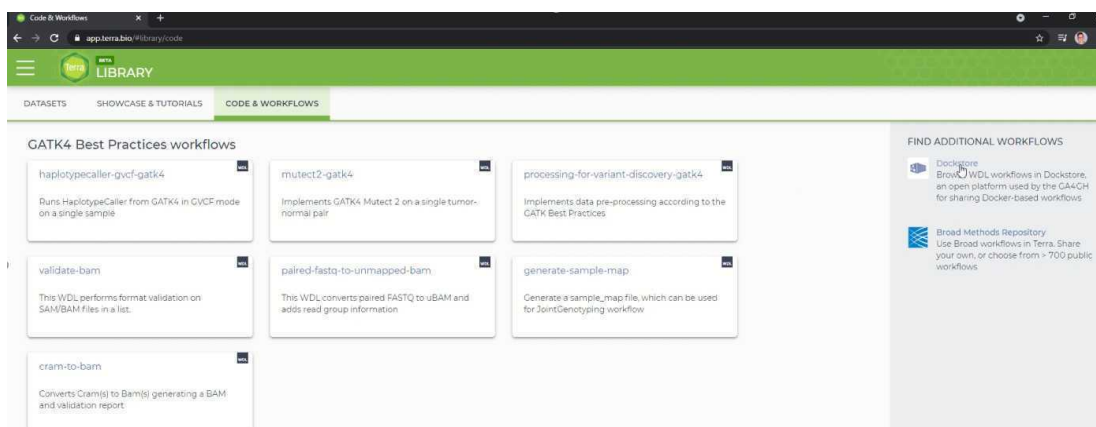
The script for the Titan workflow is located in the Dockstore repository and has to be imported into the user Terra Workspace. There are two ways to get to the Workspace page. The first method starts by clicking on the three parallel lines in the top left-hand corner, followed by clicking the 'Library' tab, and finally click the 'Workflows' button.





The library tab on the Terra Platform.

In the 'Workflows' panel, under 'Find Additional Workflows' click on the 'Dockstore' link

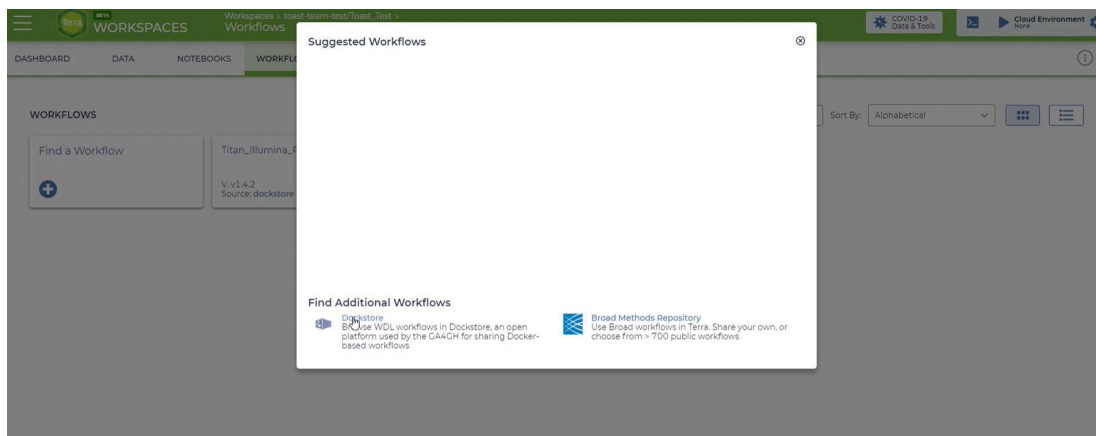


The 'Code and Workflows' page in the 'Library' panel.

Alternatively, from within your workspace click on the "Workflows" tab, then click on the blue "+" inside the "Find a workflow" panel.



Terra workflow space

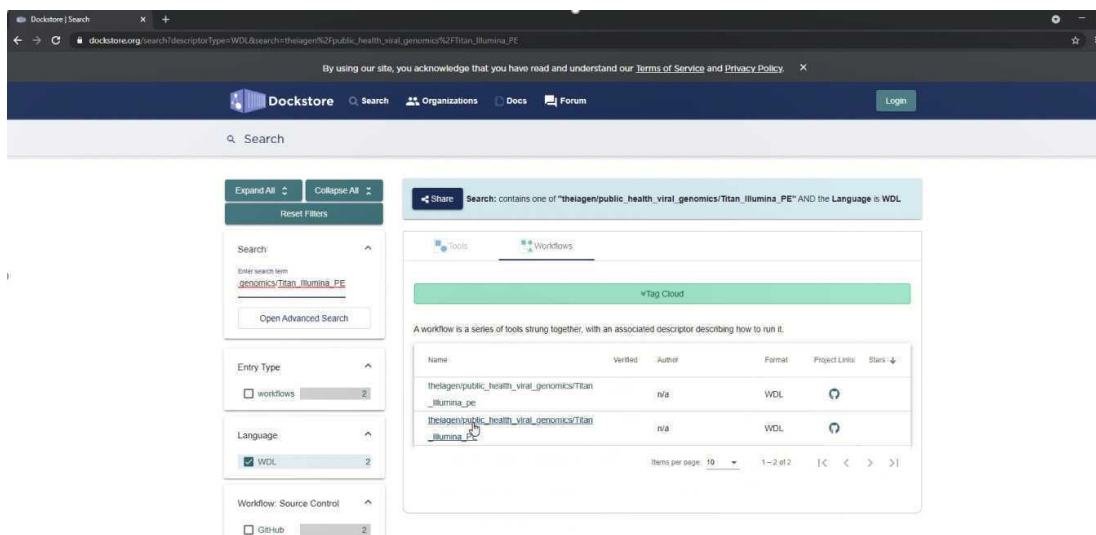


Find a workflow pop up window

On the left side of the Dockstore page, enter 'theiagen/public\_health\_viral\_genomics/' in the search bar. There are three Mercury workflows to choose from:

1. **Mercury\_PE\_Prep** - preparing sequences from paired-end sequencing runs
2. **Mercury\_SE\_Prep** - preparing sequences from single-end sequencing runs
3. **Mercury\_Batch** - combining sequences and metadata from multiple samples for submission

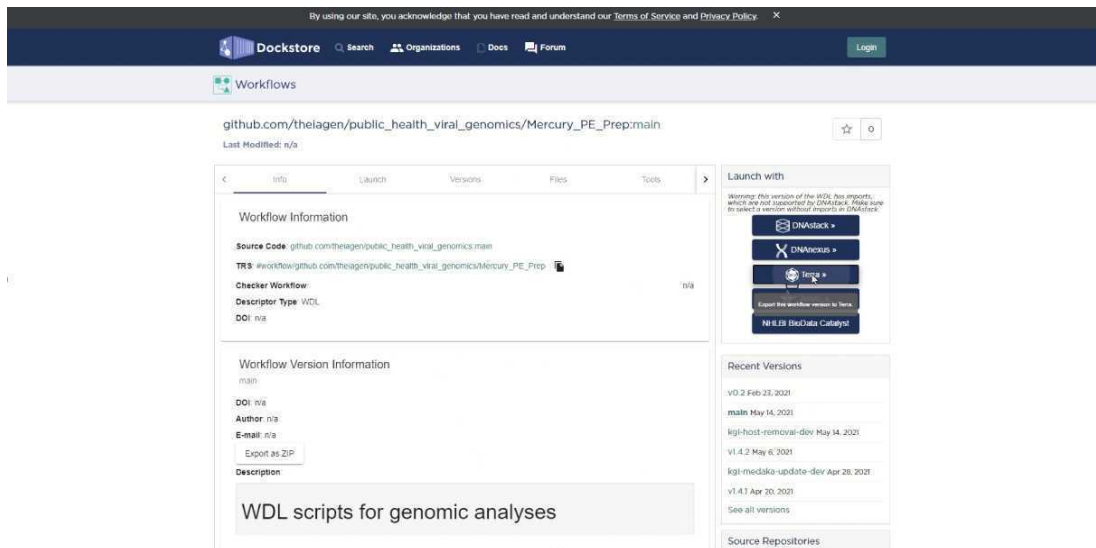
Select either PE or SE based on your situation. The examples below assume that you have several samples to prep using the **Mercury\_PE\_Prep** workflow.



Search results for 'theiagen/public\_health\_viral\_genomics/Mercury\_PE\_Prep' on the Dockstore page.

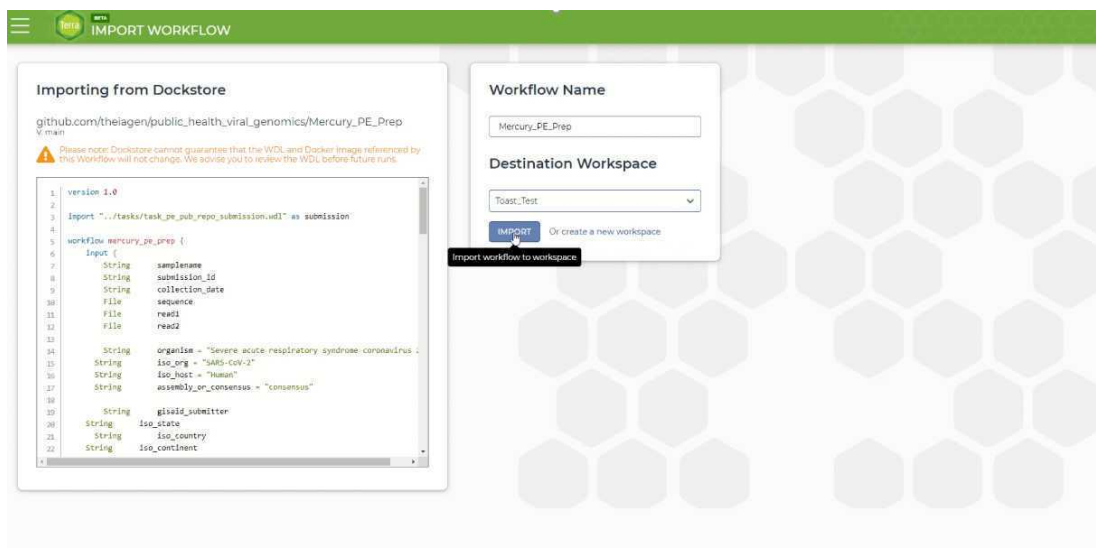
Click the '**theiagen/public\_health\_viral\_genomics/Mercury\_PE\_Prep**' link, which will take you to a new page.





The Theiagen Mercury PE Prep workflow page on the Dockstore website.

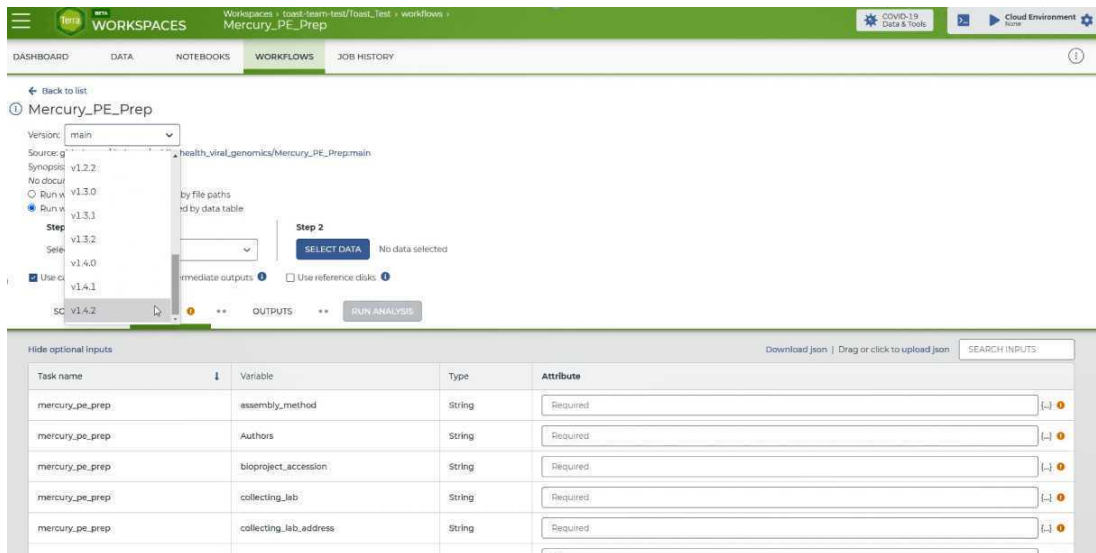
On the right-hand side of the page under the 'Launch with' window, click the 'Terra' button. It should bring you back to the Terra platform within the 'Import Workflow' page. Since you already have a workspace created, you can simply import it by selecting the workspace from the "Destination Workspace" drop down menu and clicking import.



The Terra 'Import Workflow' page. The Mercury PE Prep workflow code is shown on the left.

After clicking the import button you should be automatically directed to the Mercury workflow panel in the new workspace page that was just created.

**Make sure you pick the latest version (or an older stable version you want to use) from the drop down version menu before continuing. The "main" and "dev" versions are under active development and are not stable!**



The Mercury workflow panel within the newly created Terra workspace

Here is a video of the whole process:

Using the same method, import the Mercury\_Batch workflow, which will be run later in step 7.

## Metadata Table Preparation and Upload

4

Download [Terra\\_Metadata\\_Formatter.xlsx](#), which looks like this:

TERRA ENTITY									
(Sample Submission Prefix)									
LABORATORY METADATA									
Submitter	Authors	BioProject	State	Country	Continent	Originating Lab	Originating Lab Address	Submitting Laboratory	Submitting Laboratory
(OSAD Submitter's Username)	(Comma-Separated List of Authors)	(BioProject ID)	(State)	(Country)	(Continent)	(Originating Laboratory)	(Originating Laboratory Address)	(Submitting Laboratory)	(Submitting Laboratory)
OPTIONAL SAMPLE METADATA									
Samples	Submission ID	Collection Date	Run ID	Gender	Patient Age	County			
(Sample Identifier)	(Submission Identifier (without prefix))	(YYYY-MM-DD)	(Sequencing Run ID)	(Gender)	(Patient Age)	(County)			

Terra Metadata Formatter

This Excel file has two sheets: User Input and Terra Data Table. Users fill out the User Input sheet and the Terra Data Table sheet will populate with all the information necessary to run the Mercury workflow. After you finish, we will upload the Terra Data Table directly to Terra.

The following fields are in the User Input sheet (blue columns are required and orange are optional):

- **Terra entity** - this is the same 'root entity' as your data table in Terra (in our example its "entity:sample\_id") or the title of column one. So we will type "sample\_id" here.
- **Submission\_ID Prefix** - we recommend using the adopted convention of "state-

institution-accessionnumber" format. Using this, all samples will have the same prefix containing your state and institution. So for our example it would be GA-CDC-. **Don't forget the trailing hyphen!**

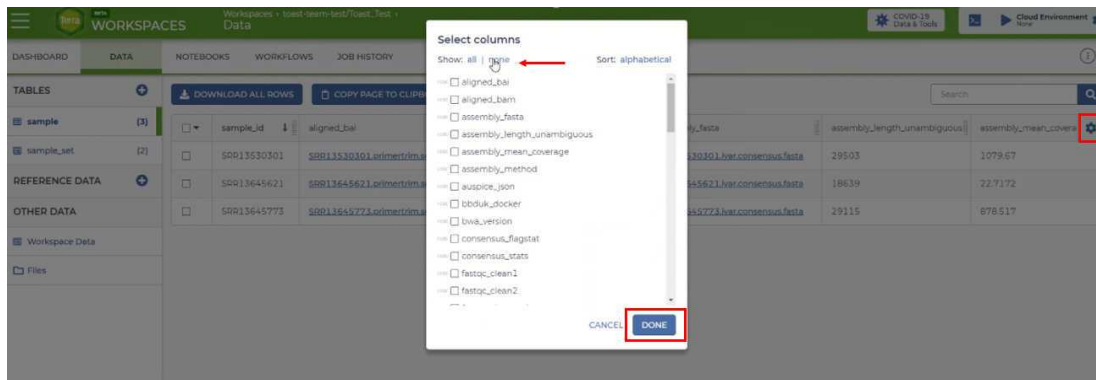
- **Submitter** - This is your GISAID user name (what you use to log into the platform). Not your actual name!
- **Authors** - Who contributed to the sequencing and analysis of this data.
- **BioProject** - an organizing tool at NCBI that pulls together different kinds of data submitted across multiple NCBI databases. Each BioProject has a unique URL, providing a home page with a title, description, links to lab websites, publications, funding resources associated with a particular project, along with links to the deposited data. A basic data BioProject holds actual sequence data, assemblies, and their associated metadata. BioProjects provide a framework for large-scale research efforts to connect submissions of various types (genome, transcriptome, etc.) and to various data repositories. (BioProject example: [PRJNA706724](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA706724)).

If your lab has not yet registered a BioProject, follow Step 3 in the [SARS-CoV-2 NCBI submission protocol: SRA, BioSample, and BioProject](#) before proceeding. If your lab does not have an NCBI submission user group established yet, its best to do that now (see step 1 of the above protocol).

- **State, Country, Continent, Originating Lab and Originating Lab address** are all information about the location of where the sample was taken from.
- **Sample Identifier** - use the file you downloaded to copy and paste the name of the samples you want to submit.
- **Submission ID** - you can use the same identifier that you used for the sample internally (this column would be the same as the Sample Identifier column) or if your lab uses secondary decoding then use that scheme here.

The GISAID EpiCoV Public Access repository is based on existing submission processes and data structures for large-scale influenza surveillance (GISAID EpiFlu). As such, submitters to EpiCoV will discover that several of the required metadata submission fields may be problematic. Location, gender and patient age are required fields, and several of them likely constitute personally-identifiable information (which is why they are listed as optional in the Terra Metadata Formatter). While these fields cannot be left blank for submission to GISAID this workflow will simply enter these fields as "unknown" if those fields are left blank in the Formatter. This will allow successful submission.

You can get the submission ID to copy and paste to your excel sheet by unselecting all columns other than the sample\_id by clicking the blue "gear" in the far right top of the data table. A pop up window will show up select "none" then click "Done".



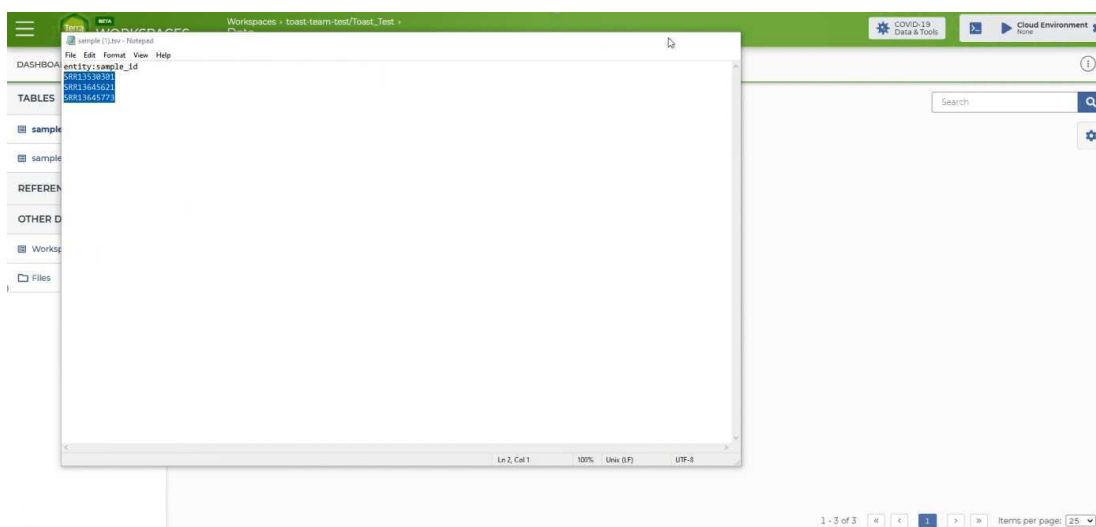
Pop up window for selection columns you want to see.

Select all the rows by clicking the top check box at the top of the rows in the far left corner of the data table. A little blue icon with three dots will show up (red arrow). Click it and a drop down menu will appear. Click "Download as a TSV".



Downloading rows as a tsv file

Open the downloaded file and copy just the rows with samples, but **NOT THE COLUMN HEADER**. This can be directly copied into the "Samples" column of the "User Input" sheet in your Terra Metadata Formatter file.



Copying sample IDs

- **Collection Date** - use the format YYYY-MM-DD
- **Run ID, Sex, Patient Age and Country** are all optional sample metadata fields. They can be left blank, just remember to remove the place holder text first. It is ok to have some of this data for some samples and not others.

Once you have this information filled out, check the Terra Data Table sheet to make sure everything is entered correctly. Here is an example with random data.

TERRA ENTITY		Submission ID Prefix		LABORATORY METADATA						
sample_id	GA-CDC-	Submitter	Authors	BioProject	State	Country	Continent	Originating Lab	Originating Lab Address	Sub
SRP13530301		Jill Hagey		B001	GA	USA	North America	CDC TOAST Team Lab	1600 Clifton Rd Atlanta GA 30333	CDC TO
SAMPLE METADATA		Submission ID	Collection Date	Run ID	Gender	Patient Age	Country			
		001	2021-01-01							
		002	2021-01-02							
		003	2021-01-03							

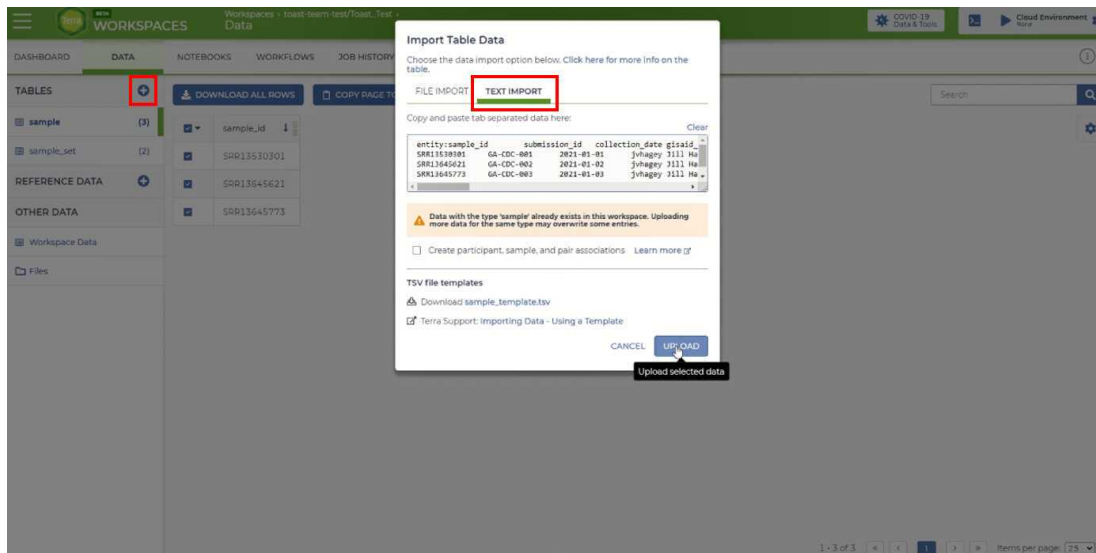
Some example metadata

Now you are ready to upload your metadata table! First, copy everything in the "Terra Data Table" to your computer's clipboard.

entity_sample_id	submission_id	collection_date	grand_submitter	authors	biospecimen_accession	iso_state
SRP13530301	GA-CDC-001	2021-01-01	jhagey	Jill Hagey	B001	GA
SRP13645621	GA-CDC-002	2021-01-02	jhagey	Jill Hagey	B001	GA
SRP13645773	GA-CDC-003	2021-01-03	jhagey	Jill Hagey	B001	GA

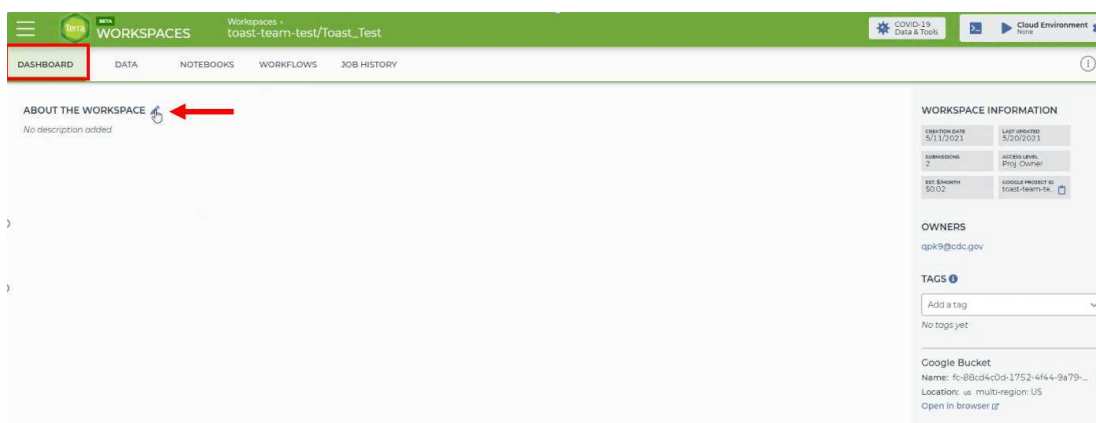
Copy populated information in the "Terra Data Table" sheet

Return to the Terra webpage "Data" tab in your workspace. Click the blue "+" button in the "Tables" tab on the left hand side of the webpage. A pop up window will appear in which you will click the "Text Import" tab. Now paste the contents of the Terra Data Table sheet into the window. Click "Upload" to finish the process. Now you should have extra columns of metadata for each of your samples.



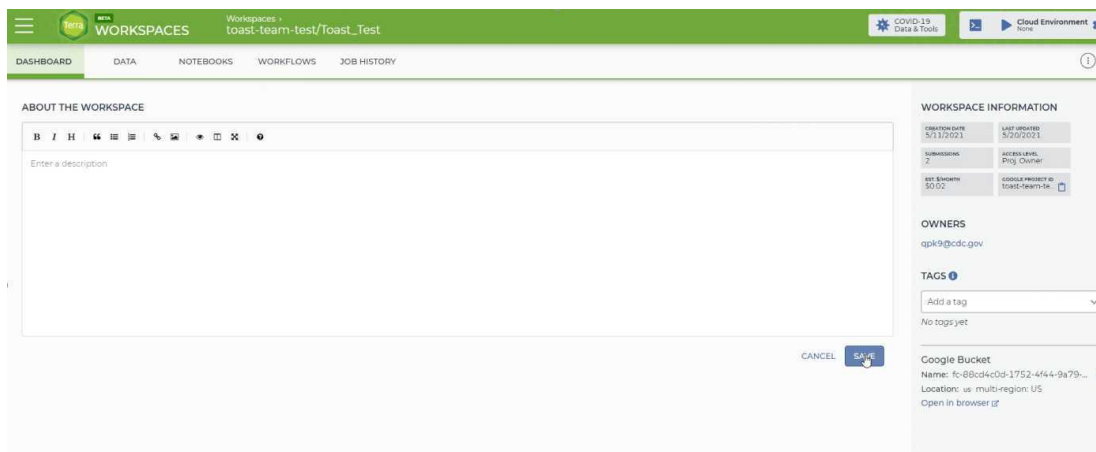
Uploading metadata

As of 5/20/2021, Terra was having some database deadlock issues with data table imports, so you might not see your import data shown right away. As a quick workaround, click on the "DASHBOARD" tab and make SOME edit to your workspace. The easiest way to do this is to click the blue "pencil" icon as if you are going to write a description of your workspace then just click "SAVE".



Click the pencil to edit the workspace description.





Save to cause a refresh of the workspace.

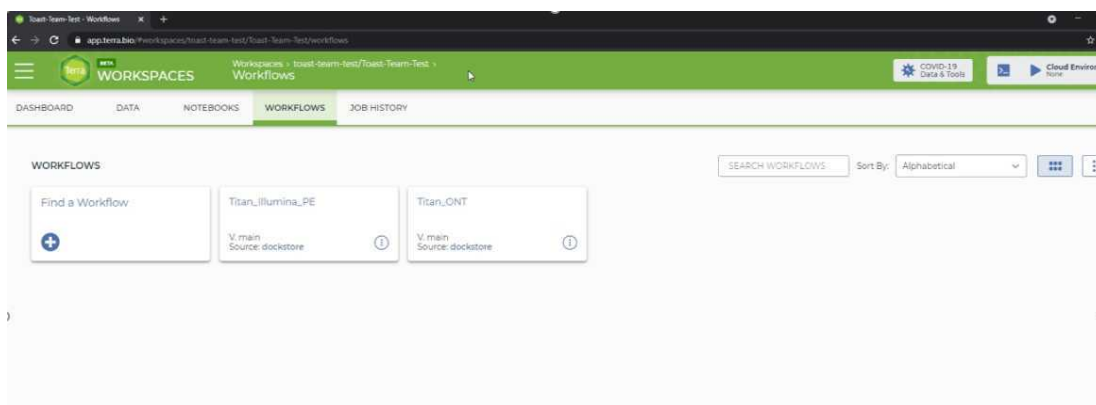
This will trigger a refresh of the workspace metadata, and you should see your metadata if you go back to the "DATA" tab and click on "Sample".

Here is a video showing the steps:

\_\_\_\_\_

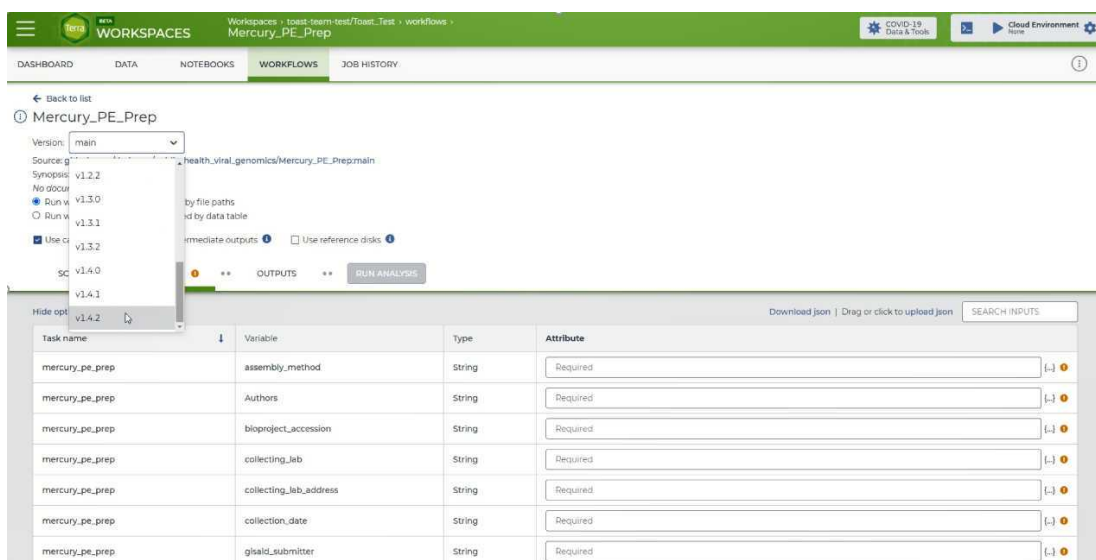
## Running the Mercury Prep Workflow

- 5 To run the Mercury workflow, click on the 'Workflows' panel within your workspace. It should bring you to the following page:



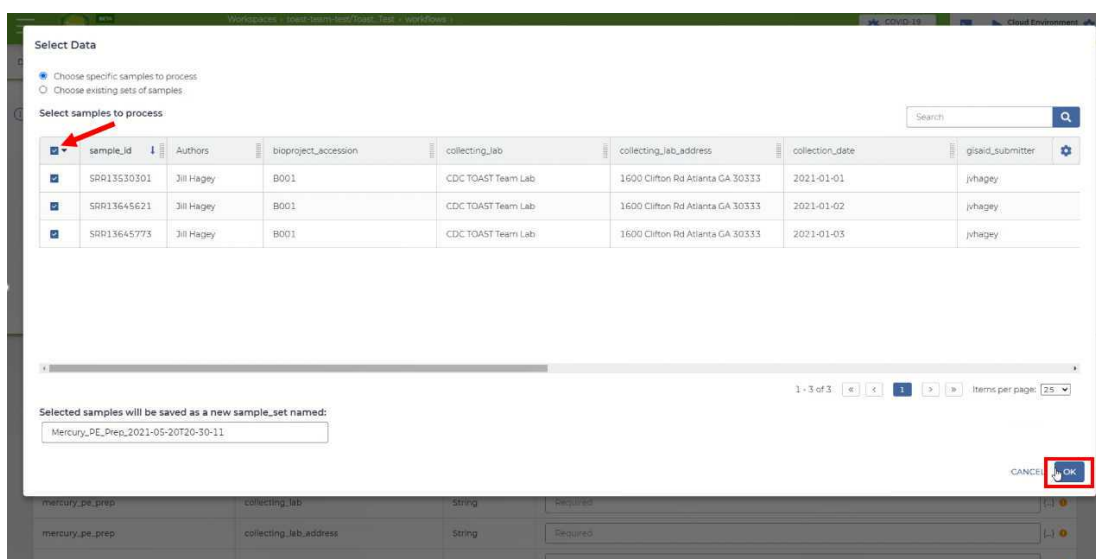
The workflows panel in the newly created Terra workspace.

Click on the '**Mercury\_PE\_Prep**' tile and it will take you to a new page. **Double check that you are using the latest version of the workflow.** Or if you want a specific version pick that one.



The Mercury PE Prep workflow page. Check version you are using.

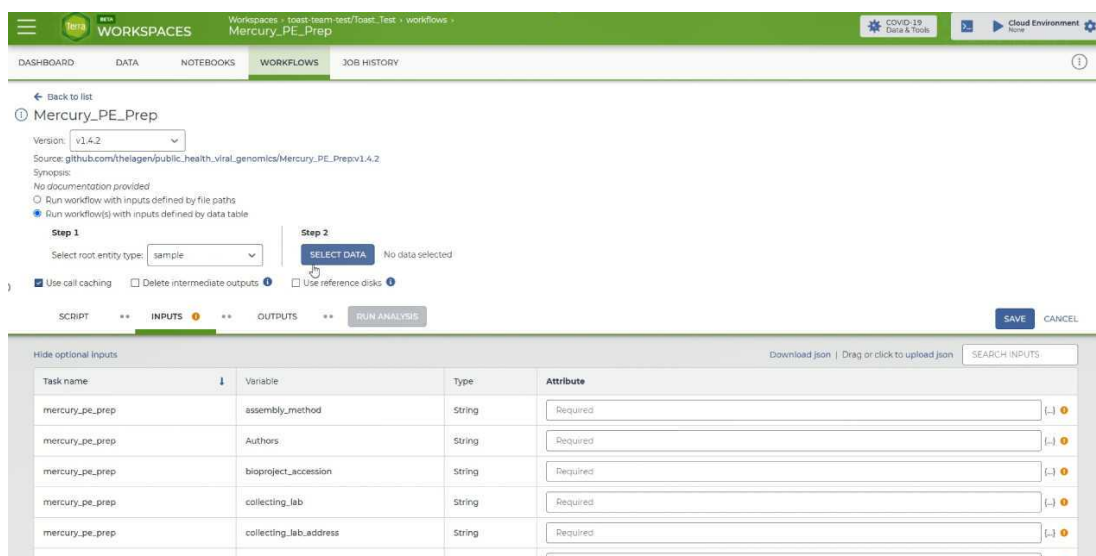
Click the 'Run workflow(s) with inputs defined by data table' option button and then in the 'Select root entity type' pull-down menu, select 'sample' or whichever entity type you specified in the sample table. Click the blue 'select data' button and select the samples you want to run or click the check box at the very top left to select all of them to be run.



Checking the box to have all sample be run through the Mercury PE Prep workflow.

Make sure "Use call caching" is checked and then select the root entity type for the data you wish to analyze (here it's sample).

**NOTE:** Call caching allows Terra to identify and skip jobs that have been run previously; this option is by default enabled to avoid unnecessary compute costs. More information on Terra call caching, including examples of when you may want to disable this feature, is available through the [Terra Support Documentation](https://support.terra.bio/docs/en/latest/call-caching/).



The Mercury PE Prep workflow input panel after defining inputs by the sample data table

The first several rows in the 'Inputs' tab should each have a yellow caution icon. **These rows represent variables that have to be provided by the user or were generated by the Titan workflow (this.assembly\_fasta).** This is the information that we just imported or was generated by the Titan workflow you ran.

For each row, click on the 'Attribute' text box and then click on the corresponding input as follows:

Variable	Attribute (AKA input)
assembly_method	this.assembly_method
Authors	this.Authors
bioproject_accession	this.bioproject_accession
collecting_lab	this.collecting_lab
collecting_lab_address	this.collecting_lab_address
collection_date	this.collection_date
gidaid_submitter	this.gidaid_submitter
iso_continent	this.iso_continent
iso_state	this.iso_state
<b>read_1</b>	<b>this.read1_clean</b>
<b>read_2</b>	<b>this.read2_clean</b>
<b>samplename</b>	<b>this.sample_id</b>
seq_platform	this.seq_platform
<b>sequence</b>	<b>this.assembly_fasta</b>
subLab_address	this.subLab_address
submission_id	this.submission_id
submitting_lab	this.submitting_lab

Required inputs for the workflow. The ones in bold are different than the normal "this.variable" notation

**Note that read\_1 and read\_2 are the RAW reads that you input into Titan; you might have named these something other than Forward\_Read and Reverse\_Read. Check the Excel file you used to create your collection. If you added any of the optional metadata then add the following attributes for these columns in the Terra Metadata Formatter.**

Variable	Attribute (AKA input)
gender	this.gender
iso_county	this.iso_county
patient_age	this.patient_age

Optional inputs to add to the workflow

mercury_pe_prep	mercury_pe_prep	String	mercury_pe_prep
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
mercury_pe_prep	iso_continent	String	this.iso_continent
mercury_pe_prep	iso_country	String	this.iso_country
mercury_pe_prep	iso_state	String	this.iso_state
mercury_pe_prep	read1	File	this.read1_clean
mercury_pe_prep	read2	File	this.read2_clean
mercury_pe_prep	sample_name	String	this.sample_id
mercury_pe_prep	seq_platform	String	this.seq_platform
mercury_pe_prep	sequence	File	this.assembly_fastq
mercury_pe_prep	subLab_address	String	this.subLab_address
mercury_pe_prep	submission_id	String	this.sub
mercury_pe_prep	submitting_lab	String	this.subLab_address
deidentify	CPUs	Int	this.submission_id
deidentify	disk_size	Int	this.submitting_lab
			Optional

Adding inputs into input panel

Next, click the output tab and click "use defaults" and it will autofill in the names of the output files according to the sample names. **If you forget to do this you won't have easily accessible results!**

Mercury\_PE\_Prep

Version: v1.4.2

Source: [github.com/helagen/public\\_health\\_viral\\_genomics/Mercury\\_PE\\_Prep](https://github.com/helagen/public_health_viral_genomics/Mercury_PE_Prep)

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 entry type: sample

Step 2: SELECT DATA 3 selected samples (will create a new sample\_set named 'Mercury\_PE\_Prep\_2021-05-20T20-30-11')

☒ Use call caching ☐ Delete intermediate outputs ☐ Use reference disks

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 / sample

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	Required	<input type="checkbox"/>
mercury_pe_prep	genbank_assembly	File	Required	<input type="checkbox"/>
mercury_pe_prep	genbank_metadata	File	Required	<input type="checkbox"/>
mercury_pe_prep	gisaid_assembly	File	Required	<input type="checkbox"/>
mercury_pe_prep	gisaid_metadata	File	Required	<input type="checkbox"/>

Setting output names by clicking "use defaults"

Click the 'Save' button on the top right-hand side of the page. The yellow caution icons should disappear. Now that you have saved the workflow the next time you go to run it this will already be done!

You are now ready to run the Mercury PE Prep workflow!

Click on the 'Run Analysis' button to the right of the 'Outputs' tab. A popup window should appear titled 'Confirm launch'. If the 'Run Analysis' button is greyed out, you need to save your recent changes by clicking the 'Save' button.

WORKSPACES

DASHBOARD DATA NOTEBOOKS WORKFLOWS JOB HISTORY

Mercury\_PE\_Prep

Version: v1.4.2

Source: [github.com/helagen/public\\_health\\_viral\\_genomics/Mercury\\_PE\\_Prep](https://github.com/helagen/public_health_viral_genomics/Mercury_PE_Prep)

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 entry type: sample

Step 2: SELECT DATA 3 selected samples (will create a new sample\_set named 'Mercury\_PE\_Prep\_2021-05-20T20-30-11')

☒ Use call caching ☐ Delete intermediate outputs ☐ Use reference disks

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

Wide 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

The 'Confirm launch' popup window

Clicking the 'Launch' button should bring you to the 'Job History' panel where each sample will be queued for the Mercury PE Prep analysis.

Data Entity	Last Changed	Status	Run Cost	Messages	Workflow ID	Links
SRP13530301 (sample)	May 20, 2021, 4:32 PM	Queued	N/A			
SRP13645621 (sample)	May 20, 2021, 4:32 PM	Queued	N/A			
SRP13645773 (sample)	May 20, 2021, 4:32 PM	Queued	N/A			

The 'Job History' page showing each sample queued for the Mercury PE Prep run.

The status will change from Queued to Submitted to Running. After the workflow has finished, the status column will show up as successful or failed.

Video of entire process:

## Understanding the Mercury Prep Output

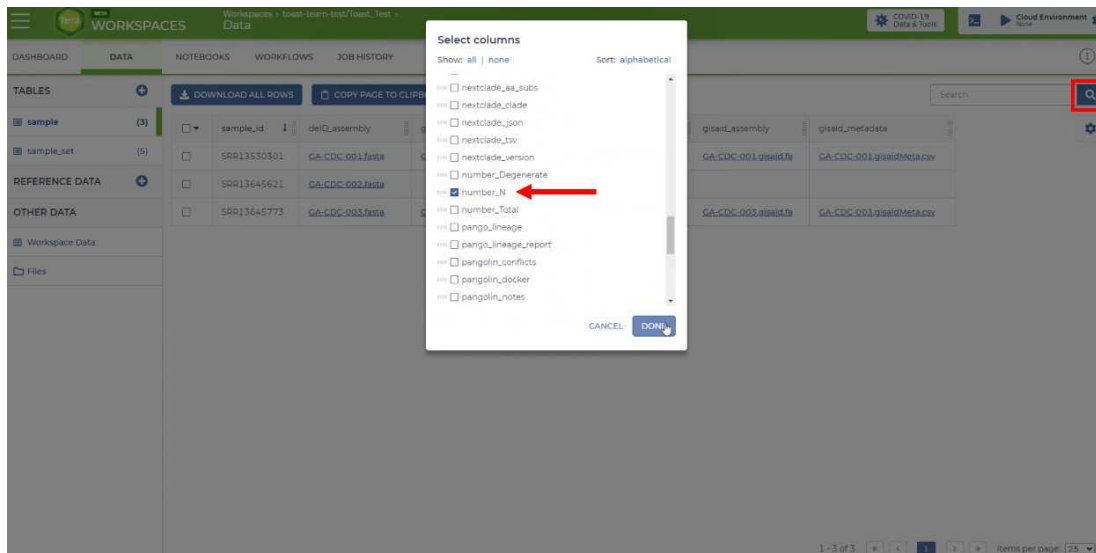
- If data are missing for a sample, it's likely because the consensus assembly included too many ambiguous bases (Ns) and exceeded the threshold for submission to GISAID (>5,000), therefore the Mercury workflow produced no data for the sample. Similarly, the consensus assembly may include too few unambiguous bases, below than the 25,000 bp threshold. This information would be reported in 'assembly\_length\_unambiguous' column highlighted below.

sample_id	deID_assembly	genbank_assembly	genbank_metadata	gisaid_assembly	gisaid_metadata	number_N
SRP13530301	GA-CDC-001.fasta	GA-CDC-001.genbank.fasta	GA-CDC-001.genbankMeta.csv	GA-CDC-001.gisaid.fasta	GA-CDC-001.gisaidMeta.csv	290
SRP13645621	GA-CDC-002.fasta	GA-CDC-002.genbank.fasta	GA-CDC-002.genbankMeta.csv	GA-CDC-002.gisaid.fasta	GA-CDC-002.gisaidMeta.csv	11210
SRP13645773	GA-CDC-003.fasta	GA-CDC-003.genbank.fasta	GA-CDC-003.genbankMeta.csv	GA-CDC-003.gisaid.fasta	GA-CDC-003.gisaidMeta.csv	692

Output of Mercury PE Prep

If you are not seeing "Number\_N" column then click the "gear" icon in top row on the right. Select the "Number\_N" then click "Done".



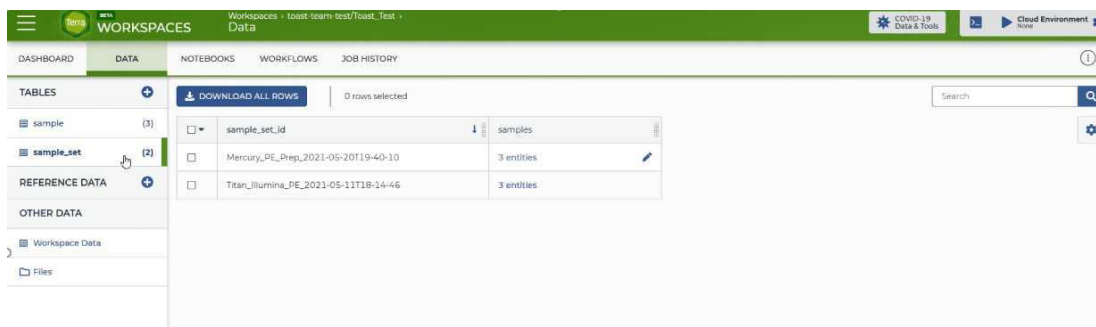


Checking Number\_N box to see the output from the Titan workflow

## Running the Mercury Batch Workflow

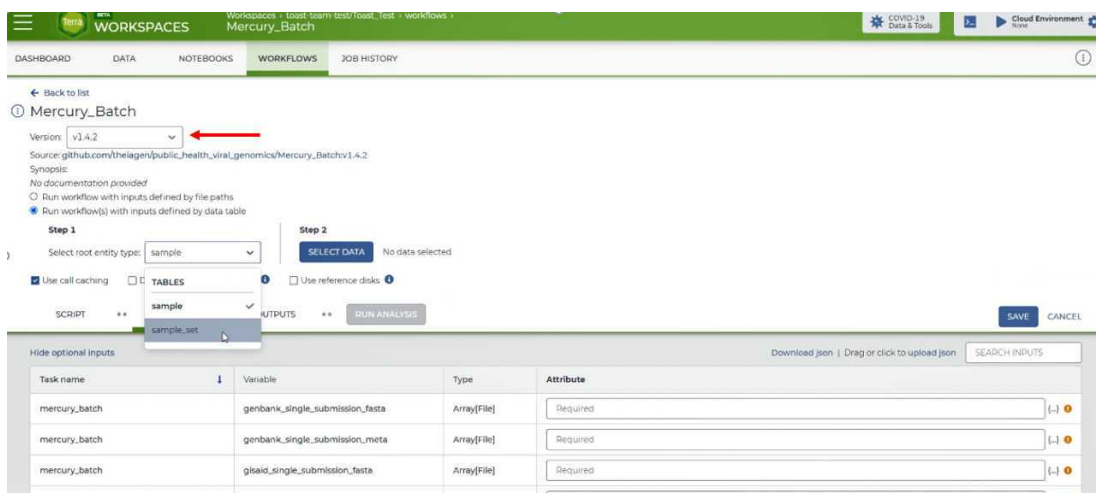
- You should have already imported the **Mercury\_Batch** workflow in step 3; if not do that now. The outputs of the Mercury PE Prep or Mercury SE Prep workflow will be used in the Mercury\_Batch workflow.

Unlike previous runs of Titan/Mercury, here we will use the root entity type "entity:sample\_set". Previously, we worked with a sample data table where each row had information for one sample. If you click on the "DATA" tab under the "TABLES" section you will see there is also a data table that contains sample\_set. In a sample set data table, each row represents a set of samples.



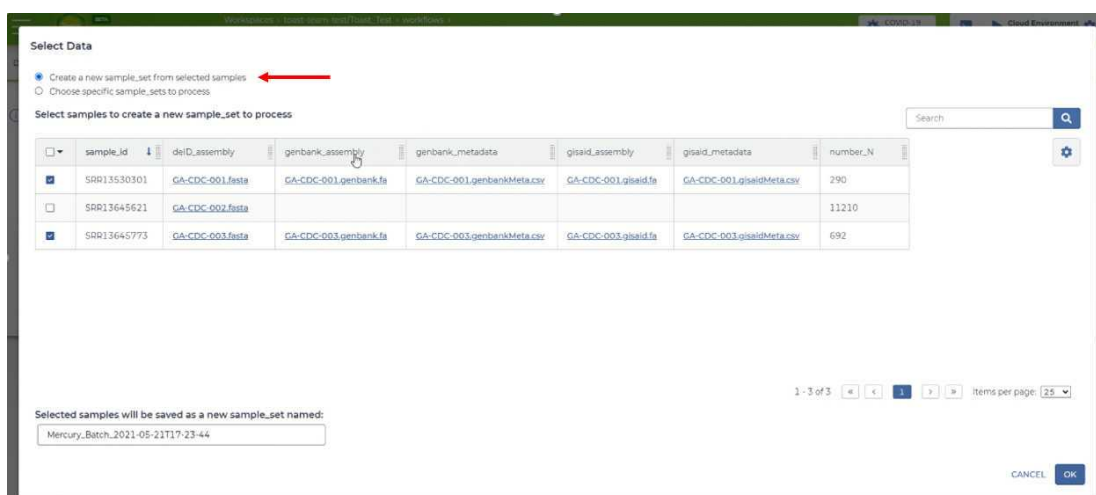
sample\_set data

Open the workflow to run the Mercury\_Batch analysis, select the latest version and click "Run workflow(s) with inputs defined by data table" as before. **Select the root entity type as "Sample\_Set" NOT "Sample\_id".**



Selecting sample\_set

Since we are using the "Sample\_Set," we have to be explicit about which set we want to analyze. To do this click "SELECT DATA," then select the appropriate samples by choosing "Create new set from selected samples" and select the samples that have output. DO NOT INCLUDE SAMPLES THAT HAVE NO OUTPUT; the pipeline will fail since there is no output to process for some samples. A quick way to do this is to select all the samples then sort the rows by "genbank\_assembly" by clicking that column header. This will organize all the samples with no output together so have you can easily deselect them, leaving only samples with output selected. You can also use the search bar to sort for particular samples. For example, if you included run\_id as part of the metadata you can search for that to easily access samples to select.



Picking Sample to run

**NOTE: If you have more than 25 samples, "select all" will only select those samples listed on the first page.** You will have to go to the next page to make sure everything you want is selected.

Now we will define the inputs for the Mercury\_Batch workflow. This is different than our previous run, so READ CAREFULLY. We have 6 required inputs for this workflow. Fill in the

following for each variable:

Variable	Attribute (AKA input)
genbank_single_submission_fasta	this.samples.genbank_assembly
genbank_single_submission_meta	this.samples.genbank_metadata
gisaid_single_submission_fasta	this.samples.gisaid_assembly
gisaid_single_submission_meta	this.samples.gisaid_metadata
samplename	this.samples.sample_id
vadr_num_alerts	this.samples.vadr_num_alerts

Required inputs for the Mercury\_Batch workflow

This notation is a bit confusing so we will break it down. Here "this." represents the "sample\_set", so for the first row we are telling the computer that for "this" (sample\_set) there is a "genbank\_assembly" element for each sample in the set and that should be the input here.

**Make sure you write this.samples.xxxx NOT this.sample.xxxx**

At the end it should look like this:

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: sample\_set

Step 2  
SELECT DATA 1 sample\_set containing 2 samples (will create a new sample\_set named "Mercury\_Batch\_2021-05-25T19:39:49")

☒ Use call caching ☐ Delete intermediate outputs ☐ Use reference disks

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

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

Task name	Variable	Type	Attribute
mercury_batch	genbank_single_submission_fasta	Array[File]	this.samples.genbank_assembly
mercury_batch	genbank_single_submission_meta	Array[File]	this.samples.genbank_metadata
mercury_batch	gisaid_single_submission_fasta	Array[File]	this.samples.gisaid_assembly
mercury_batch	gisaid_single_submission_meta	Array[File]	this.samples.gisaid_metadata
mercury_batch	samplename	Array[String]	this.samples.sample_id
mercury_batch	vadr_num_alerts	Array[Int]	this.samples.vadr_num_alerts
genbank_compile	CPUs	int	Optional

Mercury batch workflow inputs

We are inputting the number of VADR alerts to this workflow to exclude samples with VADR alerts to expedite the sharing of hassle free submission. If wanted, you can change the threshold of the number of VADR alerts by changing the optional parameter "vadr\_threshold".

Like we did before, go to the "OUTPUTS" tab and choose "Use defaults". Click "SAVE" then "RUN ANALYSIS". Once we have a successful run we can have a look at the outputs.

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

**Step 1**  
 Select root entity type: sample\_set

**Step 2**  
 SELECT DATA 1 sample\_set containing 2 samples (will create a new sample\_set named "Mercury\_Batch\_2021-05-21T17:23-44")

☒ Use call caching ☐ Delete intermediate outputs ☐ Use reference disks

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 / sample\_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	GenBank_batched_samples	File	this.GenBank_batched_samples
mercury_batch	GenBank_excluded_samples	File	this.GenBank_excluded_samples
mercury_batch	GenBank_upload_fasta	File	this.GenBank_upload_fasta
mercury_batch	GenBank_upload_meta	File	this.GenBank_upload_meta
mercury_batch	GISAID_batched_samples	File	this.GISAID_batched_samples
mercury_batch	GISAID_excluded_samples	File	this.GISAID_excluded_samples

Download json | Drag or click to upload json SEARCH OUTPUTS

Setting up outputs

If your "RUN ANALYSIS" button is greyed out, like it is in the image, you need to save the workflow first.

Video of the whole process:

## Understanding the Mercury Batch Output

- 8 First, check that there were no failures in your run by clicking on the "JOB HISTORY" tab and selecting the run. You should see something like this:

WORKSPACES toast-team-test/toast\_Test Job History

DASHBOARD DATA NOTEBOOKS WORKFLOWS **JOB HISTORY**

Job History Submission 21d3408c-9cdc-42c0-9e81-a160a0644bc8

**Workflow Statuses**  
 Succeeded: 1

**Workflow Configuration**  
 toast-team-test/Mercury\_Batch

**Submitted by**  
 gpk9@cdc.gov  
 May 21, 2021, 1:25 PM

**Total Run Cost**  
 N/A

**Data Entity**  
 Mercury\_Batch\_2021-05-21T17-23-44  
 sample\_set

**Submission ID**  
 21d3408c-9cdc-42c0-9e81-a160a0644bc8

**Call Caching**  
 Enabled

**Delete Intermediate Outputs**  
 Disabled

**Use Reference Disks**  
 Disabled

Search Completion status

Data Entity	Last Changed	Status	Run Cost	Messages	Workflow ID	Links
Mercury_Batch_2021-05-21T17-23-44 (sample_set)	May 21, 2021, 1:30 PM	Succeeded	N/A		3f19448-cd76-42c7-8109-56143223c764	📄 📧 📧

Status "Succeeded"

The outputs of this workflow that are used for submission are:

1. **GenBank\_upload\_fasta** - Concatenated assembly file for GenBank
2. **GenBank\_upload\_meta** - Combined metadata file for GenBank
3. **GISAID\_upload\_fasta** - Concatenated assembly file for GISAID
4. **GISAID\_upload\_meta** - Combined metadata file for GISAID

There are also 2 tsv files that contain the names that were batched and excluded from the

analysis.

sample_set_id	GenBank_batched_samples	GenBank_excluded_samples	GenBank_upload_fasta	GenBank_upload_meta	GISAID_batched_sample
Mercury_Batch_2021-05-21T17:23:44	GenBank_batched_samples.tsv	GenBank_excluded_samples.tsv	GenBank_upload_fasta	GenBank_upload_meta.csv	GISAID_batched_samples.tsv
Mercury_PE_Prep_2021-05-20T19:40:10					
Titan_Illumina_PE_2021-05-11T18:14:46					

Output of Mercury Batch workflow

The GenBank upload files (fasta and meta) as well as the GISAID fasta file are completely compatible for upload!

GenBank_upload_fasta	GenBank_upload_meta	GISAID_batched_samples	GISAID_excluded_samples	GISAID_upload_fasta	GISAID_upload_meta	samples
GenBank_upload_fasta	GenBank_upload_meta.csv	GISAID_batched_samples.tsv	GISAID_excluded_samples.tsv	GISAID_upload_fasta	GISAID_upload_meta.csv	2 entities
						3 entities
						3 entities

Output fasta and metadata files for upload to repositories

**However, GISAID is VERY specific in how to format and upload its metadata. So the GISAID\_upload\_meta will need to be transferred to the GISAID metadata template first.**

## GISAID Metadata Upload

- 9 Navigate to the GISAID homepage and login.

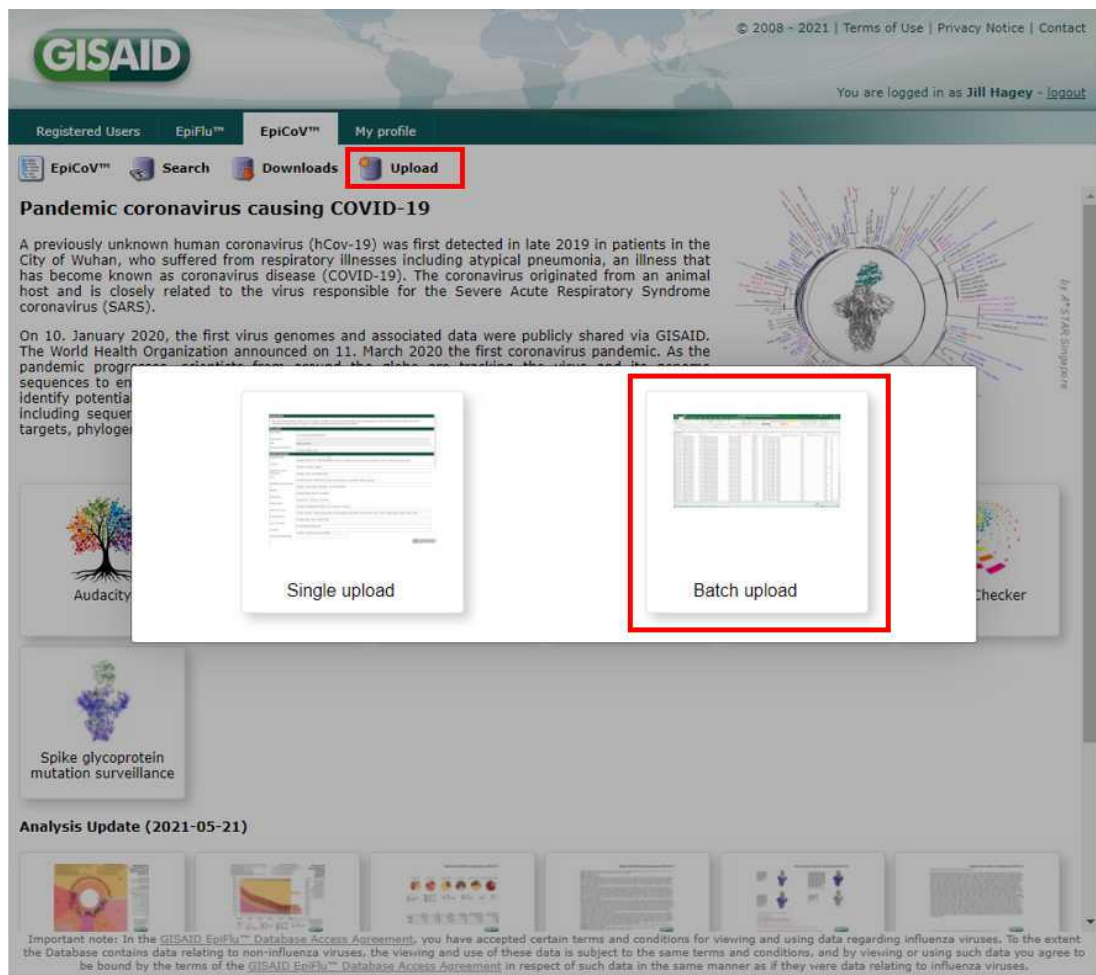
<https://www.gisaid.org/>

GISAID home page.

Login page

Click the 'Upload' button at the top of the page. Then select the 'Batch upload' option.





Upload pop-up window

On GISAID's batch upload page, click on the "Download Instructions and Template" button found at the bottom left of the page.

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

You are logged in as **Jill Hagey** - [logout](#)

Registered Users EpiFlu™ EpiCoV™ My profile

EpiCoV™ Search Downloads Upload

### GISAID hCoV-19 Batch Upload

Upload genetic sequence as single FASTA-File and metadata, available clinical and epidemiological data, geographical as well as species-specific data as XLS or CSV. Data will be reviewed by a curator prior to release. An email confirmation will be issued upon release.

Metadata as Excel or CSV\*

max size: 5M  No file chosen

Sequences as FASTA\*

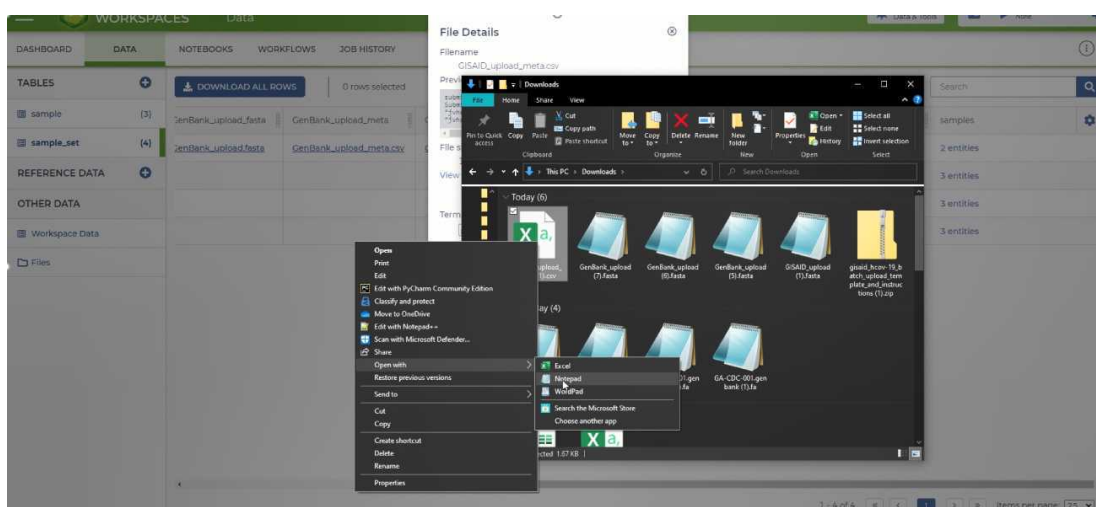
max size: 32M  No file chosen

Report

Important note: In the GISAID EpiFlu™ Database Access Agreement, you have accepted certain terms and conditions for viewing and using data regarding influenza viruses. To the extent the Database contains data relating to non-influenza viruses, the viewing and use of these data is subject to the same terms and conditions, and by viewing or using such data you agree to be bound by the terms of the GISAID EpiFlu™ Database Access Agreement in respect of such data in the same manner as if they were data relating to influenza viruses.

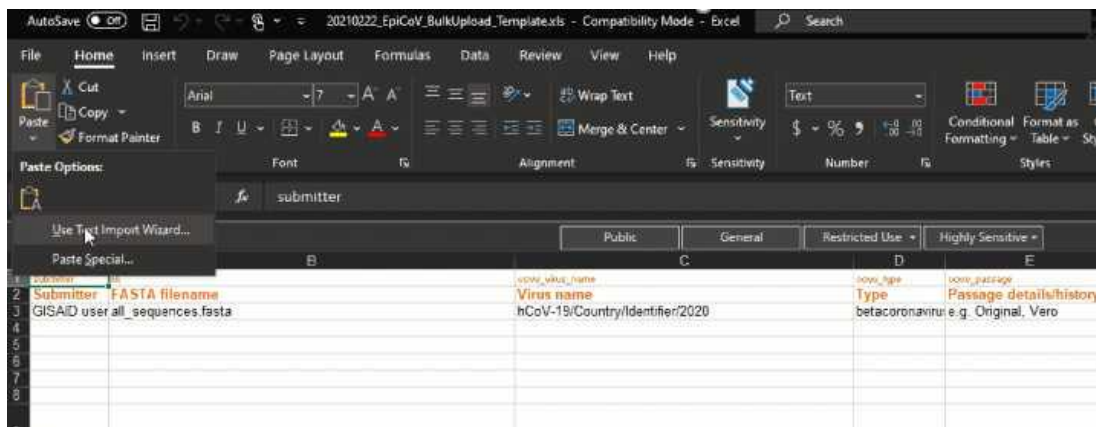
## Download Instruction and Template

On the Terra data page, download the GISAID\_upload\_meta file by clicking on it and then clicking "Download". Navigate to where the file downloaded and right click on it, chose "Open with" and then open it with a basic text editor (Notepad or TextEdit). **DO NOT OPEN WITH EXCEL!!! Excel will reformat the dates and this will lead to a bunch of chaos!!**



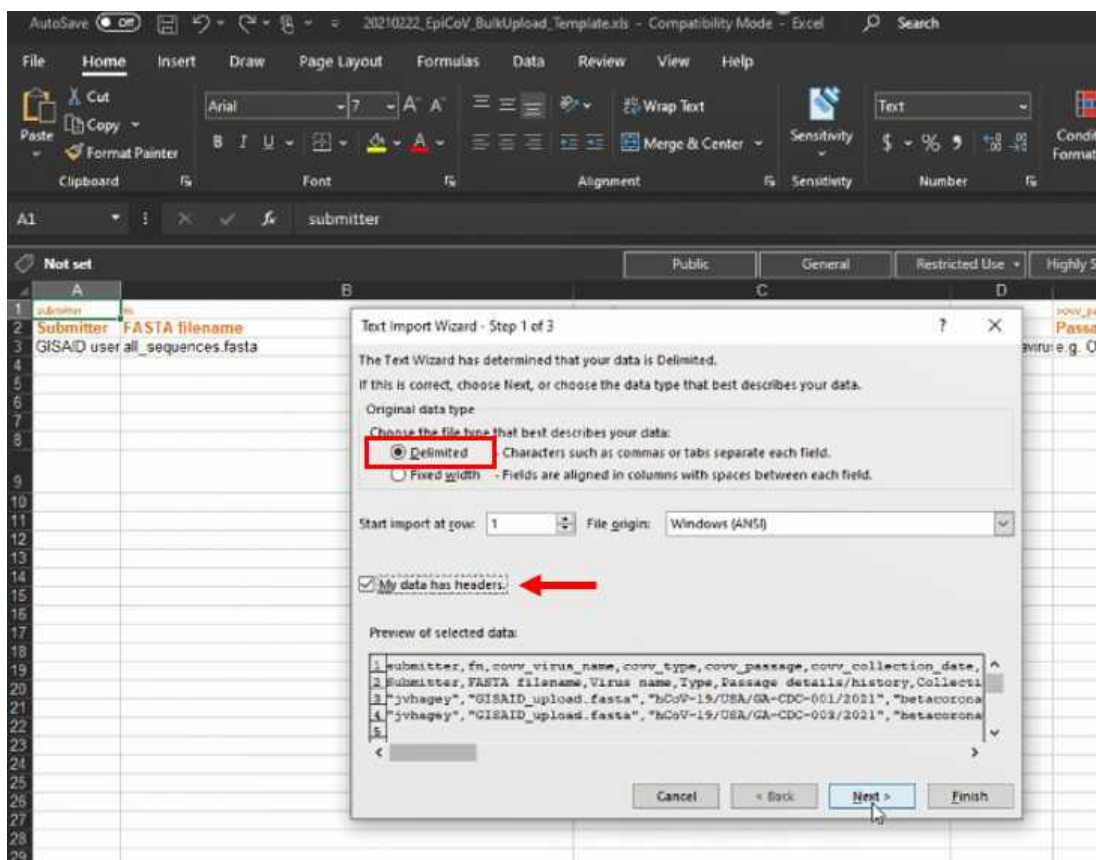
Open GISAID\_upload\_meta.csv file in notepad (or another text editor).

Now we will copy the information out of the GISAID\_upload\_meta sheet into the GISAID template on the SECOND SHEET. Copy everything in your opened GISAID upload file opened in the text editor. Return to the open GISAID template and in the SECOND sheet click the "A1" cell and then click on the paste drop down menu and choose "Text Import Wizard".

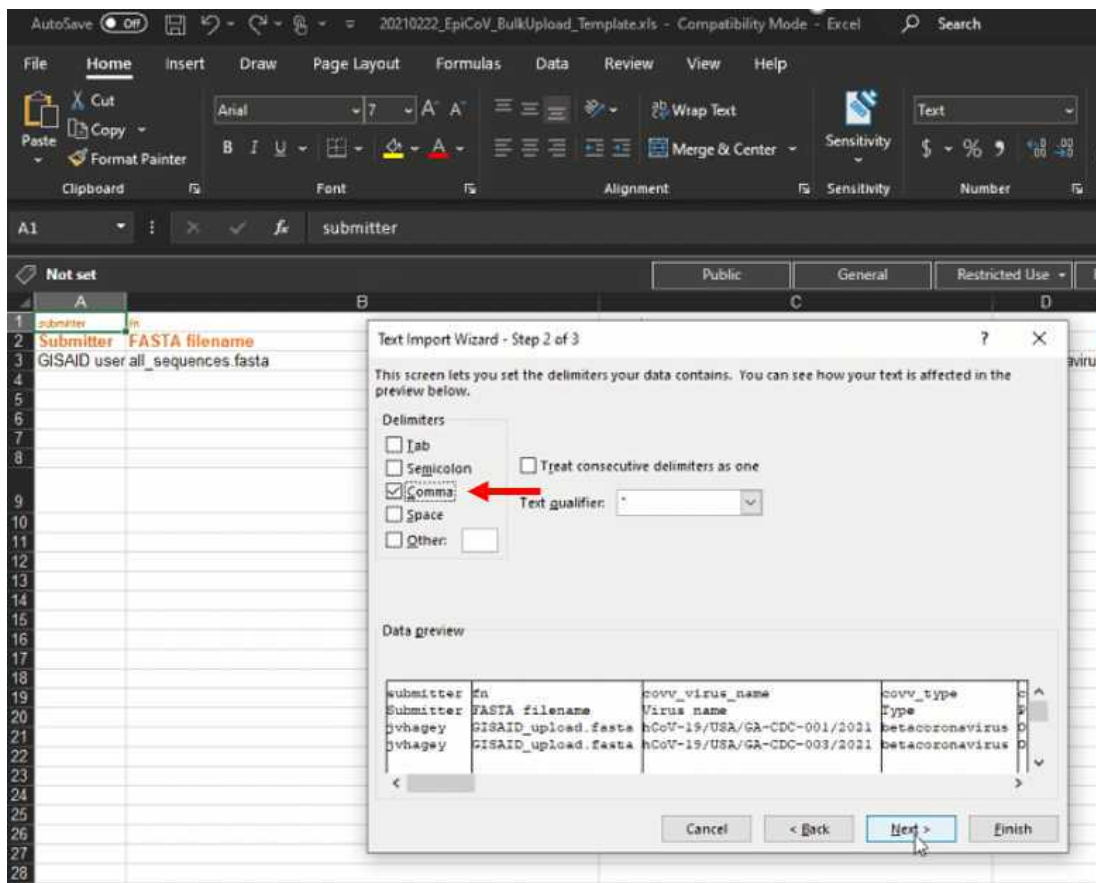


Paste with text import wizard.

Make sure the "Delimited" box is checked and click "Next". Select only the "Comma" delimiter and click "Next".

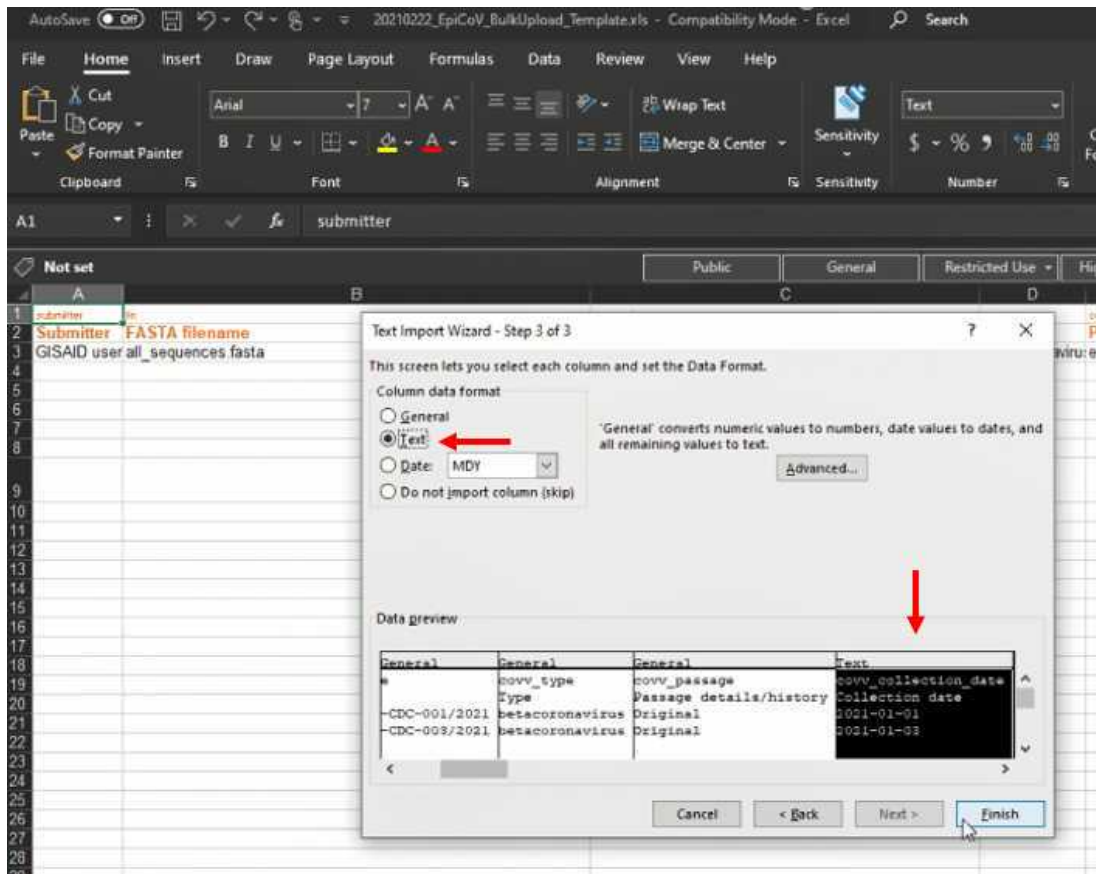


Step 1 import wizard



Step 2 of import wizard

Scroll to the "Collection Date" column and click it. While the "Collection Date" column is selected **make sure the "Column data format" is set to Text and NOT GENERAL**. Click "Finish" then "Ok". Save it and you can upload it directly to the GISAID page.



Step 3 of import wizard

Save the file and you can upload it directly to the GISAID page.

Upload to GISAID page

Here is a video of the whole process:

---

For more information you can check the [GISAID submission protocol](#).

## Submitting to SRA

- 10 Once you have submitted to GISAID, follow the [NCBI submission to SRA, BioSample, and BioProject protocol](#) to submit raw (filtered) sequencing reads to SRA.

During this step you will be asked to create a NCBI user group. **If you are not listed as an owner on the BioProject/BioSample(s), you will be unable to properly link GenBank submissions to those existing records.**

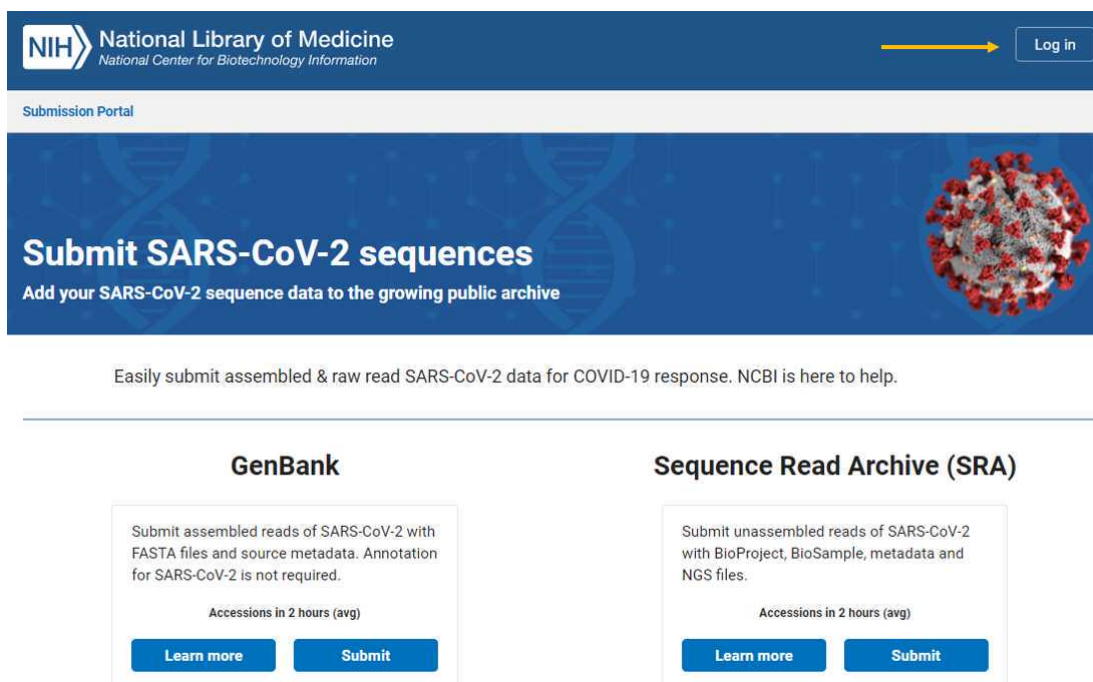
An overview of NCBI's SARS-CoV-2 submission process and the required metadata is provided here:

<https://www.protocols.io/view/overview-of-ncbi-39-s-sars-cov-2-submission-proces->



## Submitting to Genbank

- 11 Navigate to the [SARS-CoV-2 landing page for NCBI](#) and login. If you have not yet created an account, please follow the [NCBI submission to SRA, BioSample, and BioProject protocol](#) to do so before proceeding.



NCBI SARS-CoV-2 landing page before login

Once logged in, your home screen will have additional tabs at the top. Click on the GenBank submit button.

NCBI SARS-CoV-2 landing page after login

There is a lot of text on this page, but the SARS-CoV-2 sequence information at the bottom is the only thing that pertains to us here. Click the "New Submission" button at the top of the page.

New submission page with SARS-CoV-2 sequence information

If you click on "requirements and sequence processing steps" you will find the following information about sequence submission requirements.

**GenBank Submission Portal Wizards**

The GenBank Submission Portal currently supports the following submission types only:

- Prokaryotic 16S ribosomal RNA, 23S ribosomal RNA, and 16S/23S ribosomal RNA intergenic spacer region
- Eukaryotic nuclear rRNA-ITS region, small and large subunit ribosomal RNA, and internal transcribed spacer 1 and 2
- Eukaryotic organelle (mitochondrial or chloroplast) small and large subunit ribosomal RNA
- Mitochondrial (multicellular animal) Mitochondrial COXI (cytochrome oxidase subunit 1)
- Influenza A, B or C sequences
- Norovirus sequences
- Dengue virus sequences
- Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) sequences

**Common information** **Requirements for specific data types** **Frequently Asked Questions (FAQs)**

**Automatic sequence processing**

The sequences in your FASTA file will be automatically processed in Submission Portal to:

- trim terminal NNNs and ambiguous sequence ends
- remove low quality sequences with >50% ambiguous nucleotides
- check for vector and foreign contamination and:
  - trim terminal vector (strong and moderate matches)
  - remove sequences with internal vector
  - remove sequences that are entirely vector
- remove sequences below the minimal sequence length acceptable for GenBank (policy)
- remove sequences longer than the expected length for a given submission type
- Chimeric sequences are identified and may be removed
- Problematic rRNA-ITS sequences are identified. Misassembled, chimeric, or otherwise problematic rRNA-ITS sequences may be removed.

A detailed report of how your sequences were processed is available on the Review & Submit page of your submission. We will perform further processing after you complete your submission. If errors are detected, we will provide you with an error report with further details. After all errors have been resolved, you will receive an email with your final processed records. Your final processed record will also be posted in the Submission Portal.

**Sequence preparation:**

Prepare a FASTA file of quality checked sequences. Remove vector, low quality sequence and questionable data from your sequences **before** submitting.

**Source information for Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) sequences (the information provided will be used to construct a properly formatted ICTV isolate designation):**

- Unique isolate designation; the sample\_ID used in your laboratory
- Complete collection\_date, including month and day if known; this is the date that the virus sample was collected in the field; must be in the ISO format: "YYYY-MM-DD". For example: 2020-03-25.
- Country where virus was collected; See [INSDC country list](#) for allowed names and format. Information following the colon in the format should be presented in larger to smaller order, i.e. country, state, city. For example: USA: Maryland. During processing, the country specified before the colon will be automatically converted to the three letter country ISO code for use in construction of the ICTV-formatted isolate.
- Host organism; common or scientific name of the host animal from which the virus was located. This information will be used in the ICTV isolate.
- Isolation-source; the physical environment where the virus was collected. For example: nasopharyngeal swab.
- Optional: BioProject, BioSample and SRA accession numbers can be provided to link assembled sequences to previously submitted read data. This optional information can be added via "add field", "add column" or included in the tab-delimited table on source modifiers. These accessions must be owned by the same group as the submitting group and obtained ahead of your GenBank submission. It is recommended that the "SARS-CoV-2: clinical or host-associated" BioSample package is used.

## Sequence submission requirements for SARS-CoV-2 consensus sequences

Rest assured the Titan and Mercury protocols are compliant with these requirements, but this information is good to know as it can help guide your lab in QC cutoffs in your protocol or determining what kind of information you want to add to your metadata.

After clicking "New Submission" you will be directed to a new page. Under "\*What do your sequences contain?" Select "SARS-CoV-2, Influenza, Norovirus, or Dengue virus". Next select "SARS-CoV-2" under "Which virus?". Add an optional submission title if you want and click "continue".

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

**Submission Portal** Home My submissions Templates My profile

**GenBank submission: SUB9741150** Delete submission

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** Required fields are marked with \*

\* What do your sequences contain?

☐ rRNA or rRNA-ITS

☐ COXI 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**

## Step 1 of GenBank submission

The submitter info should auto populate, but if not, enter appropriate info. Here, "submitter" is the name of the person, or user group, who is physically doing the submissions, not a supervisor or PI. **This must be the same person or group that submitted the associated BioSamples and BioProject.** Select the appropriate submission group name for your laboratory and check the contact information.

If you do not have a submission group available to click, see Steps 1.2-1.3 in the [SRA submission protocol](#) to establish a new one for your laboratory, or to add your name to a group already established for your lab.

NIH National Library of Medicine  
National Center for Biotechnology Information

Submission Portal

GenBank submission: SUB9741237

SARS-CoV-2

1 SUBMISSION TYPE 2 SUBMITTER 3 SEQUENCING TECHNOLOGY 4 SEQUENCES 5 SEQUENCE PROCESSING 6 SOURCE MODIFIERS 7 REFERENCES 8 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.

\* Submitting organization \* Department \* Street \* City \* State/Province \* Postal code \* Country

**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

Phone ID Fax ID

Continue ☒ Update my contact information in profile

submitter page.

Click "Continue" when done. The new step requires information about the sequencing and bioinformatic platforms.

You will most likely have used either Illumina or Nanopore/ONT (Oxford Nanopore Technology). Select which is appropriate for your data, for Nanopore select "other" and enter "ONT" in the field. Select "Assembled sequences (each sequence was assembled from two or more overlapping sequence reads)" then enter the assembly information. You can find this information in the assembly\_method column of the Titan output on Terra.

sample	aligned_bam	assembly_length_unambiguous	assembly_mean_coverage	assembly_method	consensus_flagstat
sample	rimetrtrim.sorted.b-	SRR13530301.rimetrtrim.sorted.b-	29503	1079.67	SRR13530301.flagstat.txt
sample_set	rimetrtrim.sorted.b-	SRR13645621.rimetrtrim.sorted.b-	18639	22.7172	SRR13645621.flagstat.txt
OTHER DATA	rimetrtrim.sorted.b-	SRR13645773.rimetrtrim.sorted.b-	29115	878.517	SRR13645773.flagstat.txt

Assembly method column on Terra

Alternatively, you can find the assembly information in your "GISAID\_upload\_meta.csv" file under the "Assembly method" column.

Submitter	Submitter FASTA file	Virus name	Passage	Collection location	Additional host	Additional sampling	Gender	Patient age	Patient status	Specimen outbreak	Last vaccine treatment	Sequencing technology	Assembly method	BWA Version	Var version	Coverage	Originator address	Sample ID suffix	CDC TOAS 1600 Clifton Rd Atlanta GA	CDC TOAS 1600 Clifton Rd Atlanta GA
1																				
2																				
3																				
4																				
5																				
6																				
7																				
8																				
9																				
10																				
11																				
12																				
13																				
14																				
15																				
16																				
17																				
18																				
19																				
20																				
21																				
22																				
23																				
24																				
25																				
26																				
27																				
28																				
29																				
30																				
31																				
32																				
33																				
34																				
35																				
36																				
37																				
38																				
39																				
40																				
41																				
42																				
43																				
44																				
45																				
46																				
47																				
48																				
49																				
50																				
51																				
52																				
53																				
54																				
55																				
56																				
57																				
58																				
59																				
60																				
61																				
62																				
63																				
64																				
65																				
66																				
67																				
68																				
69																				
70																				
71																				
72																				
73																				
74																				
75																				
76																				
77																				
78																				
79																				
80																				
81																				
82																				
83																				
84																				
85																				
86																				
87																				
88																				
89																				
90																				
91																				
92																				
93																				
94																				
95																				
96																				
97																				
98																				
99																				
100																				

Assembly method in GISAID\_upload\_meta.csv

The completed form should look similar to the example below:

GenBank submission: SUB9741237

Sequencing Technology

Method

What methods were used to obtain these sequences?

☐ Sanger

☐ Illumina

☐ PacBio

☐ Nanopore

☐ Other

☐ None

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: BWA

Version or date: 0.7.17-r1188

Continue

Fill out sequencing and bioinformatic information.

Click "Continue" when done. Under "Release date" Click "Release immediately following processing" for routine surveillance isolates. Your lab might also have a policy describing the timetable for data release.

Drag-and-drop the "GenBank\_upload.fasta" file downloaded from the Mercury workflow into the "Sequences" box.

After clicking "Continue", your sequences will be checked for strings of NNNs and ambiguous bases. If it finds one of these it will return an error message. If you get a warning about NNNs click "A region of estimated length between the sequenced regions based on an alignment to similar sequences or genome" and then "Continue". Further details are given in the [Genbank submission protocol](#).

NIH National Library of Medicine  
National Center for Biotechnology Information

Submission Portal

GenBank submission: SUB9741237  
SARS-CoV-2

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

Sequences

Warning: Found one or more string of NNN's (length > 10):

Sequence ID
GA-CDC-001
GA-CDC-003

What do the internal NNN's represent?

☒ Please explain what the strings of internal NNNs represent

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

☐ A region of unknown length between the sequenced regions

NNNs warning

In the next tab, you will be asked how you want NCBI to handle sequences that fail to be processed. Pick how you would like to proceed and click "Continue"

NIH National Library of Medicine  
National Center for Biotechnology Information

Submission Portal

GenBank submission: SUB9741237  
SARS-CoV-2

1 SUBMISSION TYPE 2 SUBMITTER 3 SEQUENCING TECHNOLOGY 4 SEQUENCES 5 SEQUENCE PROCESSING 6 SOURCE MODIFIERS 7 REFERENCES 8 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

An example of the sequence processing tab

On the next tab, under "How do you want to apply source modifiers?" select "Upload a tab-delimited table" and then drag and drop the "GenBank\_upload\_meta.csv" file from the Mercury pipeline.

Once you have uploaded your metadata table you will get a warning about not submitting "isolation-source" missing from your data table. If you have this information, you can click "Under an editable table" under the "How do you want to apply source modifiers" header. Once you select this an editable table will appear.



NIH National Library of Medicine  
National Center for Biotechnology Information

Submission Portal

GenBank submission: SUB9741237

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

**Source Modifiers**

**Warning:** You have not included an isolation-source for your sequences. Examples include nasopharyngeal swab or respiratory swab. If this information is unavailable, press Continue.

Please review the data and submit updated source modifiers if necessary, or press Continue.

For each sequence, GenBank requires the following source information:

- collection-date
- country
- host
- isolate

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

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

How do you want to apply source modifiers?

- Use an editable table
- Upload a tab-delimited table (template file provided)

Apply source modifiers by editing a table

Sequence ID	country	host	isolate	collection date	isolation source
GA-CDC-001	USA	Human	GA-CDC-001	2021-01-01	
GA-CDC-002	USA	Human	GA-CDC-002	2021-01-02	

Examples: nasopharynx, lung, nasal swab

Continue

editable table.

If the answer for "isolation-source" is identical, you can type the answer into the first cell then hover your mouse over the lower right corner of that cell until a large "+" appears. Drag to the bottom cell to copy the same text into all the cells as you would in Excel. Click "Continue" when done.

Now you should be on step 7 "References". Enter the author(s) information and select if these sequences are for a article that is unpublished, in press or published. This page will require at least 1 author entered to continue. Click "Continue" when done.

On the last page, review that your data is correct and hit submit when ready. Review the submission and Genbank record preview, **paying close attention to correct linkage of BioProject and BioSample**, plus any other metadata submitted.

To proceed please review your submission, make any necessary changes using the tabs/steps above, then click on the Submit button below.

## GenBank Record Preview

Why is some information missing/different in this GenBank record preview? +

```

LOCUS      CA-IGI-3157          29782 bp    DNA        linear   VRL 11-FEB-2021
DEFINITION Severe acute respiratory syndrome coronavirus 2 isolate
            SARS-CoV-2/human/USA/CA-IGI-3157/2021.
ACCESSION
VERSION
DBLINK     BioProject: PRJNA186035
            BioSample: SAMN06175309
KEYWORDS
SOURCE     Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)
ORGANISM   Severe acute respiratory syndrome coronavirus 2
            Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes;
            Nidovirales; Coronidovirineae; Coronaviridae; Orthocoronavirinae;
            Betacoronavirus; Sarbecovirus.
REFERENCE  1 (bases 1 to 29782)
AUTHORS    Timme,R.E.
TITLE      Direct Submission
JOURNAL    Submitted (11-FEB-2021) CFSAN-ORS-DM-MMSB, FDA Center for Food
            Safety and Applied Nutrition, 5100 Paint Branch Parkway, College
            Park, MD 20740, USA
COMMENT    ##Assembly-Data-START##
            Assembly Method      :: ARTIC-nCoV-bioinformaticsSOP v. 1.1.0
            Sequencing Technology :: MinION
            ##Assembly-Data-END##
FEATURES   Location/Qualifiers
            source                1..29782
                                   /organism="Severe acute respiratory syndrome coronavirus
                                   2"
                                   /mol_type="genomic DNA"
                                   /isolate="SARS-CoV-2/human/USA/CA-IGI-3157/2021"
                                   /isolation_source="clinical; See additional sample source
                                   fields for further information (i.e. anatomical material,
                                   anatomical part, body product, environmental material,
                                   environmental site, collection device, collection method)"
                                   /host="Homo sapiens"
                                   /db_xref="taxon:2697049"
                                   /country="USA:California"
                                   /collection_date="2021-02-08"
                                   /note="GISAID accession: EPI_ISL_123457"
BASE COUNT 8817 a 5426 c 5786 g 9491 t 262 others
ORIGIN
            1 agatctgttc tctaaacgaa ctttaaaatc tgtgtggctg tcaactggct gcatgcttag
            61 tgcactcag cagtataatt aataactaat tactgtcgtt gacaggacac gactaactcg
            121 tctatcttct gcaggctgct tacggtttcg tccgtgttgc agccgatcat cagcacatct
  
```

Double check these fields to make sure they are correct before submitting.

For more details see the [Genbank submission protocol](#) written by NCBI.

## 11.1 Checking the status of your submissions

The status of your submission can be tracked under the "My Submissions" tab:  
<https://submit.ncbi.nlm.nih.gov/subs/>

GenBank accessions will be listed here, under "GenBank: Processed" and

available for download in the "AccessionReport.tsv" file.

Submission ↕	Title ↕	App ↕	Status ↕	Updated ↕
SUB7560154	SARS-CoV-2	GenBank	✓ GenBank: <b>Processed</b> MT683386-MT683418 3 files: <ul style="list-style-type: none"><li>• AccessionReport.tsv</li><li>• flatfile.zip</li><li>• email.txt</li></ul>	Jul 01

Sequences with no annotation issues will be listed as Processed.

**Submissions with annotation discrepancies will be marked as Error and a "Fix" button will appear.** A report is emailed to you and listed on the submissions page with the detailed issues to be resolved. If the data are incorrect, click the "Fix" button and you will return to the sequences page of your submission to upload a corrected file.

**If you have evidence that the discrepancy is due to a naturally occurring mutation, send an email to [gb-admin@ncbi.nlm.nih.gov](mailto:gb-admin@ncbi.nlm.nih.gov) with the SUB number and supporting evidence.**

## 11.2

### Data stewardship and curation

It is a good idea to develop an internal method for storing and tracking your GenBank accessions as these are required for making future updates to your records.

For updates to your GenBank records follow the NCBI Curation Protocol hosted by GenomeTrakr:

<https://www.protocols.io/view/ncbi-data-curation-protocol-bacaiase>