Getting Started in Terra: Importing Reads, Metadata, Workflows, and More

Document TG-TER-03, Version 4

Date:	Workflow Versions:
5/1/2025	PHB v3.0.0

1. PURPOSE/SCOPE

To standardize the process of uploading next generation sequencing (NGS) data from local storage or the Sequencing Reads Archive (SRA) and creating and uploading a results metadata table using the online Terra platform for downstream Theiagen workflow analysis. Additional instructions are provided for importing workflows and adding workspace data elements and files. Acceptable NGS data types include Illumina, Oxford Nanopore Technology (ONT), ClearLabs, and FASTA file formats. Read the comprehensive workflow documentation here.

2. REQUIRED RESOURCES

- Computer
- Internet connection: at least 10 and 5Mbps for download and upload speeds, respectively
- Internet browser
    - Google Chrome, Firefox, or Edge
- Google account
- Terra account, linked to Google account.

IMPORTANT NOTES

Metadata column headers and workflow input text indicated in **gray** in this SOP are customizable; **black** is required text.

3. RELATED DOCUMENTS

Document Number	Document Name
TG-TER-04	Linking BaseSpace and Importing BaseSpace Reads to Terra Workspace

4. PROCEDURE

4.1 IMPORTING LOCAL RAW READS

1. *Sign in* to https://app.terra.bio/ using a Gmail account and Google Authentication (Figure 1)
2. Click on the *hamburger icon* in the top left and navigate to *Workspaces* (Figure 2)
3. *Open the workspace* designated for analysis
4. In the data tab, click *Import data* and select *Open data uploader* (Figure 3**Error! Reference source not found.**)
5. For new data sets, click *Create a New Collection* (Figure 4**Error! Reference source not found.**)
    a. *The following nomenclature may be useful: YYYYMMDD_# (do not include spaces)*
6. Click *Upload* or *drag and drop* raw sequencing reads into the data table

Figure 1

7. When upload is complete, files will populate on this page with *file names hyperlinked* to their corresponding Google Bucket location (Figure 5)
8. Continue to section 4.2 to upload sample metadata.

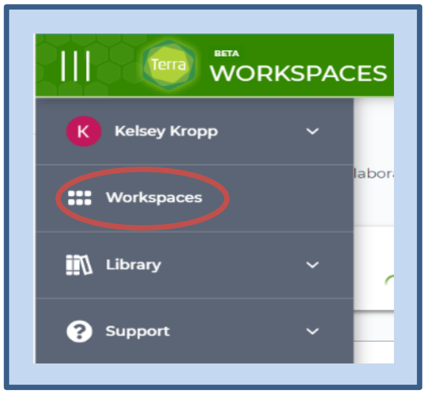


Figure 2

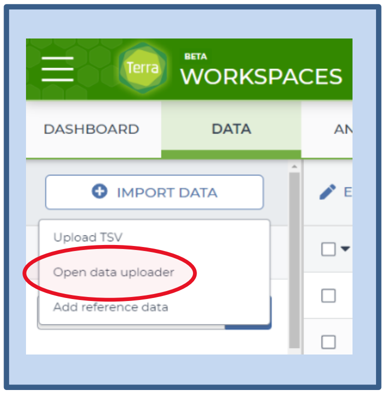
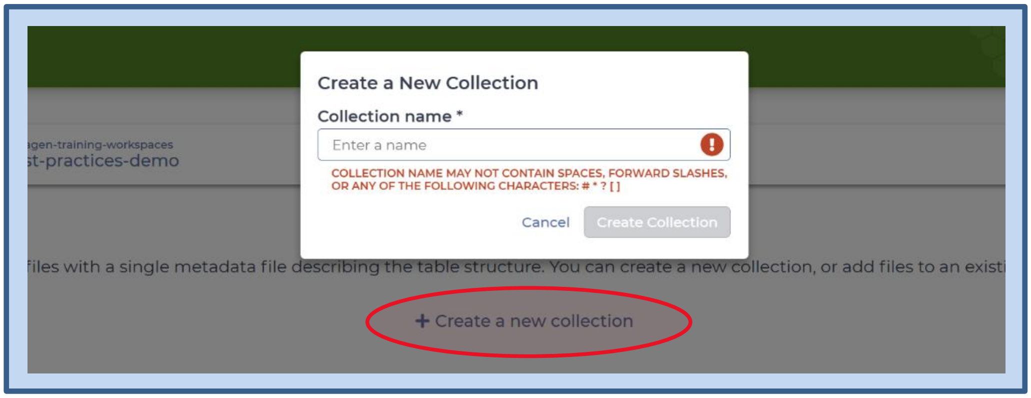


Figure 3


Create a New Collection

Collection name *

COLLECTION NAME MAY NOT CONTAIN SPACES, FORWARD SLASHES, OR ANY OF THE FOLLOWING CHARACTERS: # * ? [ ]

files with a single metadata file describing the table structure. You can create a new collection, or add files to an existi

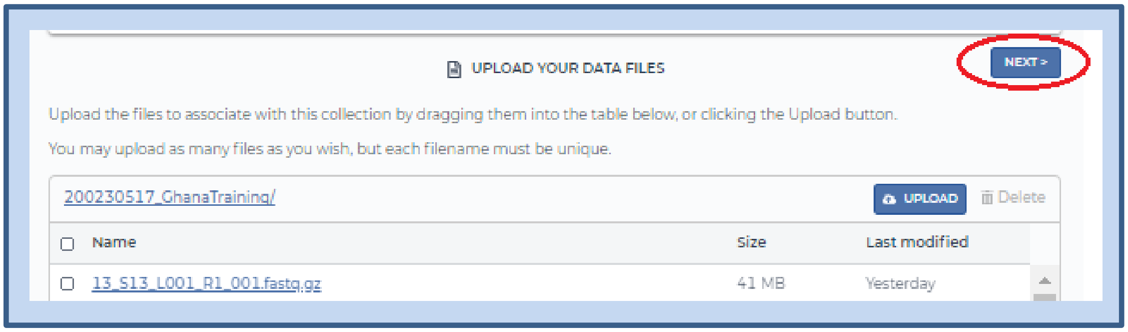
Figure 4

4.2 IMPORTING SAMPLE METADATA (TSV FILE) AFTER CREATING COLLECTION OF READ FILES

1. Once files have successfully uploaded, click *Next* (Figure 5)
2. Import the associated sample metadata file or drag and drop from file explorer (Figure)
   a. *See appendix 10.1 to create a sample metadata file*
   b. Alternatively, use the auto-generation option and carefully review

      i. When using this function, the collection name should be the desired table name
3. Click *create table* (Figure)

4. View the uploaded read data and metadata in the Terra data table by clicking *View the…table in the workspace* at the bottom of the screen
5. In the data tab, all read files and metadata are now populated for each sample in the associated table (e.g. *illumina_pe_specimen* table)
    a. Open the data table by clicking on the table name in the sidebar (Figure 8)



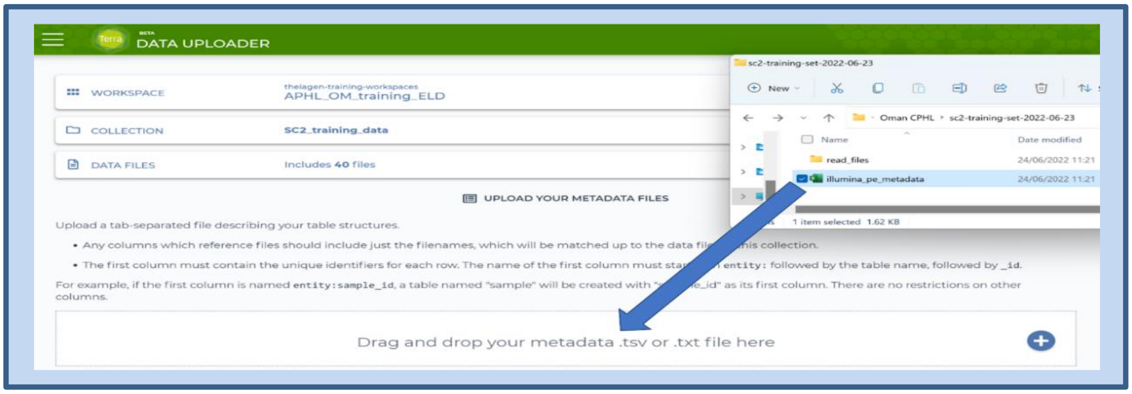
Upload the files to associate with this collection by dragging them into the table below, or clicking the Upload button.

You may upload as many files as you wish, but each filename must be unique.

200230517_GhanaTraining/

☐ Name	Size	Last modified
☐ 13_S13_L001_R1_001.fastq.gz	41 MB	Yesterday

Figure 5



WORKSPACE: theiagen-training-workspaces APHL_OM_training_ELD
COLLECTION: SC2_training_data
DATA FILES: Includes 40 files

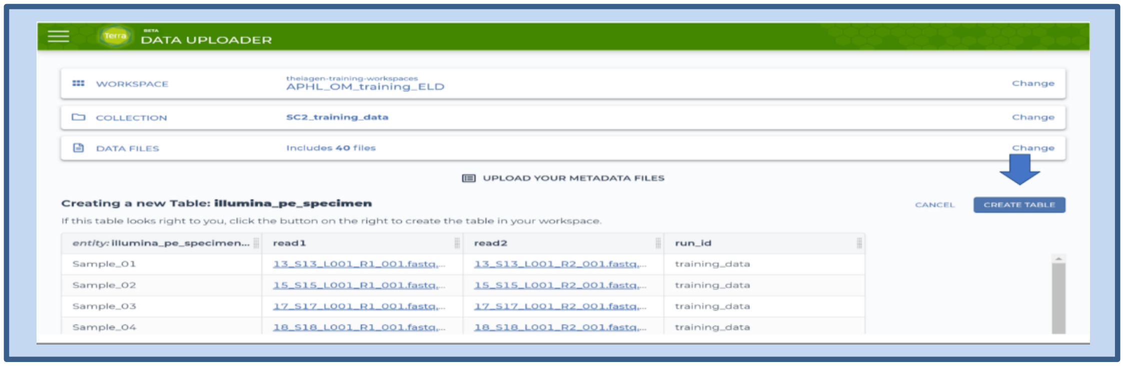
Upload a tab-separated file describing your table structures.

- Any columns which reference files should include just the filenames, which will be matched up to the data fil… his collection.
- The first column must contain the unique identifiers for each row. The name of the first column must sta… `entity:` followed by the table name, followed by `_id`.

For example, if the first column is named `entity:sample_id`, a table named "sample" will be created with "… e_id" as its first column. There are no restrictions on other columns.

Drag and drop your metadata .tsv or .txt file here

Figure 6



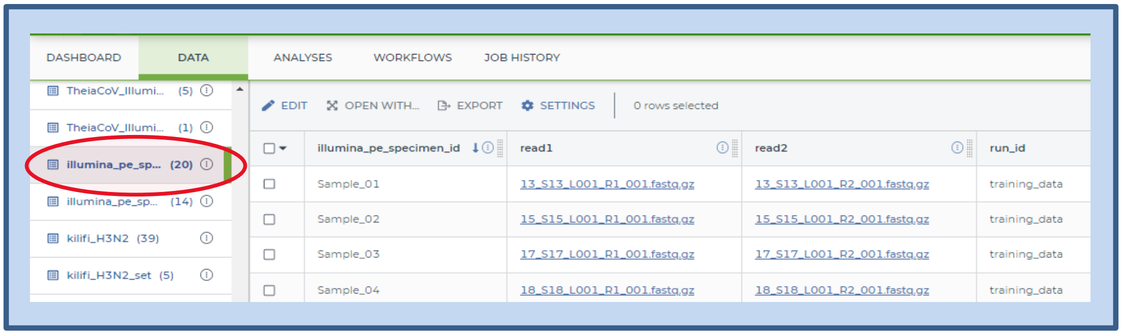
WORKSPACE: theiagen-training-workspaces APHL_OM_training_ELD
COLLECTION: SC2_training_data
DATA FILES: Includes 40 files

Creating a new Table: illumina_pe_specimen

If this table looks right to you, click the button on the right to create the table in your workspace.

entity: illumina_pe_specimen…	read1	read2	run_id
Sample_01	13_S13_L001_R1_001.fastq,…	13_S13_L001_R2_001.fastq,…	training_data
Sample_02	15_S15_L001_R1_001.fastq,…	15_S15_L001_R2_001.fastq,…	training_data
Sample_03	17_S17_L001_R1_001.fastq,…	17_S17_L001_R2_001.fastq,…	training_data
Sample_04	18_S18_L001_R1_001.fastq,…	18_S18_L001_R2_001.fastq,…	training_data

Figure 7



illumina_pe_specimen_id	read1	read2	run_id
Sample_01	13_S13_L001_R1_001.fastq.gz	13_S13_L001_R2_001.fastq.gz	training_data
Sample_02	15_S15_L001_R1_001.fastq.gz	15_S15_L001_R2_001.fastq.gz	training_data
Sample_03	17_S17_L001_R1_001.fastq.gz	17_S17_L001_R2_001.fastq.gz	training_data
Sample_04	18_S18_L001_R1_001.fastq.gz	18_S18_L001_R2_001.fastq.gz	training_data

Figure 8

4.3 IMPORTING SAMPLE METADATA (TSV FILE) WITHOUT CREATING A READ COLLECTION

1. To upload the metadata file to Terra, return to the *Terra workspace* containing data of interest
2. In the *data* tab, click *import data*, and *upload TSV* (Figure)
3. In the pop-up window, *drag and drop* the file in the gray box or *click to select* the metadata file, then click *start import job* (Figure 10)
    a. *See appendix 10.1 to create a sample metadata file*
    b. *If adding metadata to samples in an existing data table, ensure the data table name indicated in the TSV file in cell A1 contains the same data table name listed in Terra (e.g. [DATATABLE_NAME]_id)*
4. All samples and metadata should now be populated in the corresponding Terra data table (Figure 8)

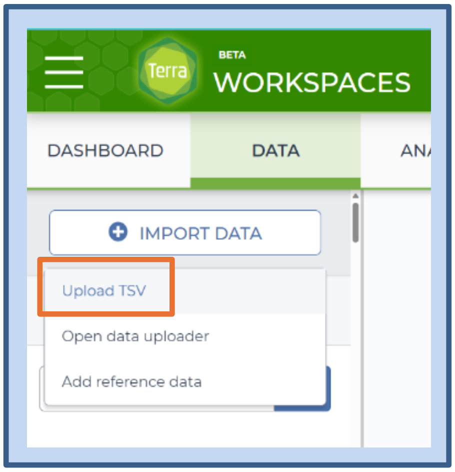
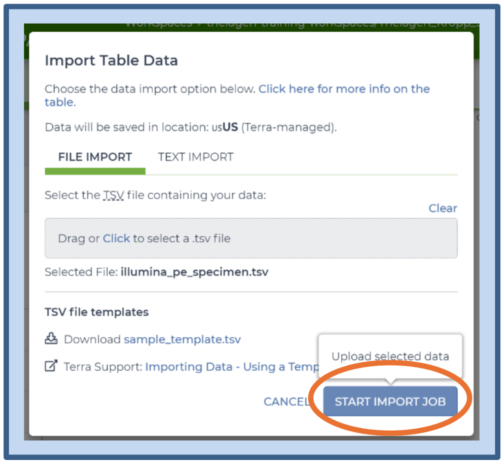


Figure 9


Import Table Data

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

Data will be saved in location: us**US** (Terra-managed).

Select the TSV file containing your data:

Selected File: **illumina_pe_specimen.tsv**

TSV file templates

Download sample_template.tsv

Terra Support: Importing Data - Using a Temp

Figure 10

4.4 IMPORTING RAW READS USING SRA_FETCH

1. Navigate to the Terra workspace that will be used to import reads
2. In the data tab, click on *import data* and *upload TSV* (Figure)
3. Import a TSV file containing the table of SRA accession numbers for desired samples: *select* or *drag and drop* the file, then click *start import job* (Figure 1)
   a. *A template TSV file can be downloaded from this pop-up; follow appendix 10.1 to create a metadata/TSV file (for SRA Fetch, the TSV file does not need read1 or read2 columns, only sra_accession)* (Figure 1)

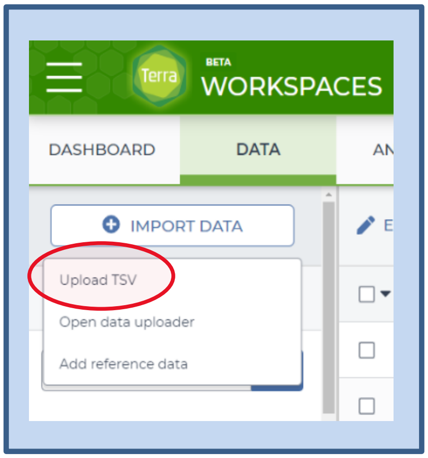


Figure 11

Import Table Data

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

Data will be saved in location: usUS (Terra-managed).

Select the TSV file containing your data:

Drag or Click to select a .tsv file

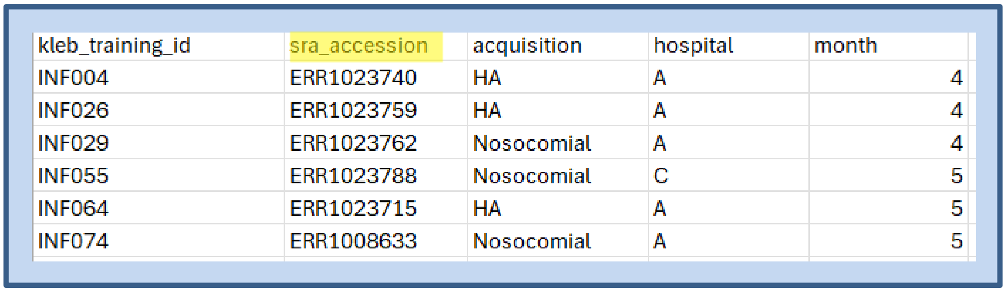
Selected File: None

TSV file templates

Download sample_template.tsv

Terra Support: Importing Data - Using a Template

Figure 42

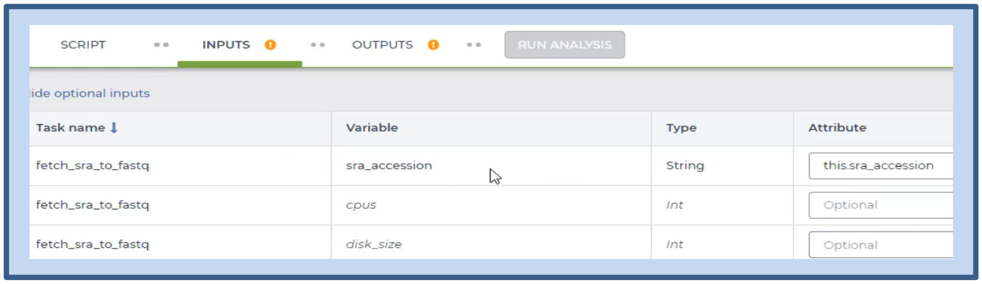


kleb_training_id	sra_accession	acquisition	hospital	month
INF004	ERR1023740	HA	A	4
INF026	ERR1023759	HA	A	4
INF029	ERR1023762	Nosocomial	A	4
INF055	ERR1023788	Nosocomial	C	5
INF064	ERR1023715	HA	A	5
INF074	ERR1008633	Nosocomial	A	5

Figure 13

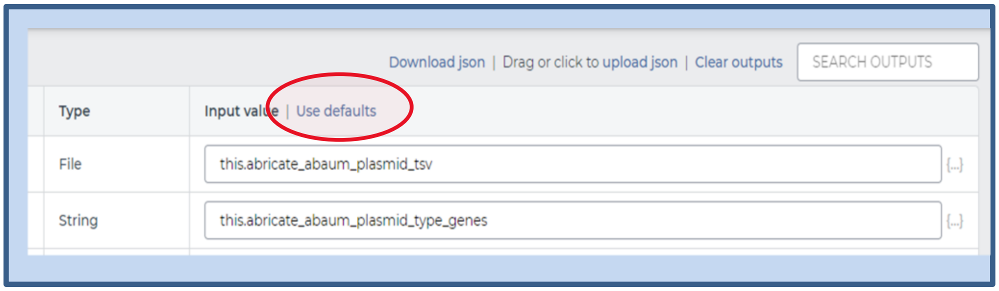
4. In the workflows tab, click *SRA_Fetch_PHB*
   a. Refer to appendix 10.2 for how to import a workflow into a workspace
5. Set the workflow *version* to the latest version, or the workflow version used for internal validations
6. Choose the second bullet to *run workflow(s) with inputs defined by data table*
7. Select the sample table to use under *select data table*
   a. *Do not choose the "set" table version*

8. Click *select data* and *mark the checkboxes* to specify which sample reads to import
9. Set the first input setting to *this.sra_accession*, where *sra_accession* is the TSV file column name containing SRA numbers (Figure 1)
10. In the outputs tab, click *use defaults* (Figure 1) and *save* the workflow
11. Then click *run analysis*, enter any comments if desired, and click *launch*
12. Fetched reads will be populated in the *read1* and *read2* columns of the data table



Task name	Variable	Type	Attribute
fetch_sra_to_fastq	sra_accession	String	this.sra_accession
fetch_sra_to_fastq	cpus	Int	Optional
fetch_sra_to_fastq	disk_size	Int	Optional

Figure 74



Type	Input value | Use defaults
File	this.abricate_abaum_plasmid_tsv
String	this.abricate_abaum_plasmid_type_genes

Figure 65

4.5 IMPORTING RAW READS FROM BASESPACE

1. Command line steps are required for the initial setup between BaseSpace and Terra. Refer to the following sites for details on initial setup and the import process:
    a. Theiagen's [BaseSpace Fetch Documentation]()
    b. Document TG-TER-04, Version 3

5. QUALITY RECORDS

- Raw read files
- Metadata results table
- Workspace elements and files

6. TROUBLESHOOTING

- Terra data table column headers become available as workflow inputs when running workflows, search for them in workflow input dropdowns using the prefix *this.* to filter
- For workspace data and files, search for them in workflow input dropdowns using the prefix *workspace.*
- If the first cell in the metadata table does not end with _*id*, an error message will prevent file import; adjust the metadata text in cell A1 and re-upload
- If any workflow, input, or output settings are entered incorrectly, the analysis will not run as expected; verify all settings are correct and re-launch analysis
- If analysis fails, navigate to the job history in the workspace and click on the job submission for details; for help resolving run failures, email support@theiagen.com

7. LIMITATIONS

N/A

8. REFERENCES

1. Libuit, Kevin G., Emma L. Doughty, James R. Otieno, Frank Ambrosio, Curtis J. Kapsak, Emily A. Smith, Sage M. Wright, et al. 2023. "Accelerating Bioinformatics Implementation in Public Health." Microbial Genomics 9 (7). https://doi.org/10.1099/mgen.0.001051
2. Theiagen Genomics Public Health Bioinformatics Workflow Documentation
3. Theiagen's BaseSpace_Fetch and SRA_Fetch workflow resource pages
4. Theiagen's Docker Image and Reference Materials for SARS-CoV-2
5. Theiagen's Public Health Resources

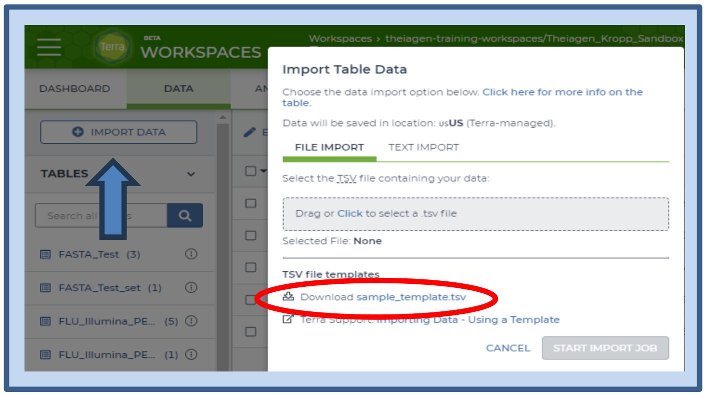
9. REVISION HISTORY

Revision	Version	Release Date
Document Creation	1	7/2023
Added internet speeds recommended for up/download, important notes, section 4.5, SRA metadata info in appendix 9.1, and appendix 9.3	2	9/2023
Formatting (Figure references and cross-references check), section 4.1 - 4.4 updates, inclusion of figure 15, appendix 10.1 updates	3	4/2024
Edits for version release, aligning with Terra interface, and formatting	4	5/2025

10. APPENDICES

10.1 CREATING A METADATA FILE (TSV FILE)

1. Open the downloadable TSV template located in the Import Data/Upload TSV pop-up window (Figure 6)



Import Table Data

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

Data will be saved in location: usUS (Terra-managed).

Select the TSV file containing your data:

Drag or Click to select a .tsv file

Selected File: **None**

TSV file templates

Download sample_template.tsv

Terra Support: Importing Data - Using a Template

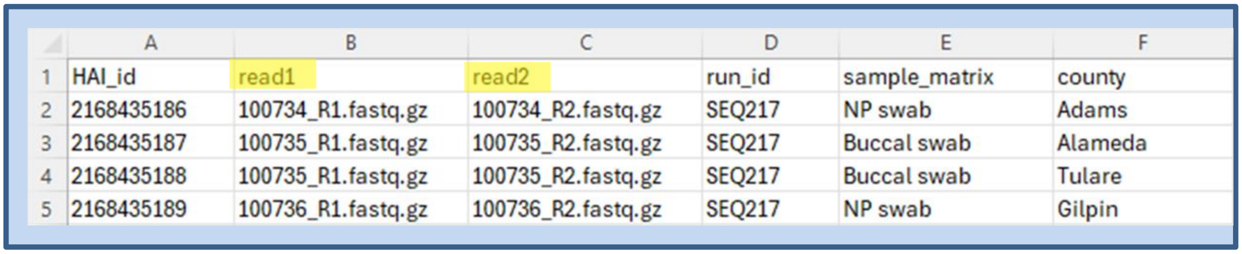
Figure 16

2. Cell A1 must contain the following text: *[name]_id* (Figure 17)
   - *a.* Do not include spaces or characters other than hyphens (-) or underscores (_)
   - *b.* The data table name should contain text to indicate the project name (e.g. Cell A1 may read *HAI_id* to create a table for all HAI organisms)
3. Enter all *sample IDs* into column 1 below cell A1
4. Label column headers for desired metadata (do not include spaces or special characters)
   - *a.* *Optional:* enter run_id as a column header and the run ID for each sample in the column
   - *b.* *Optional:* add additional metadata columns and sample information, as needed

HAI_id	run_id	sample_matrix	county
2168435186	SEQ217	NP swab	Adams
2168435187	SEQ217	Buccal swab	Alameda

Figure 17

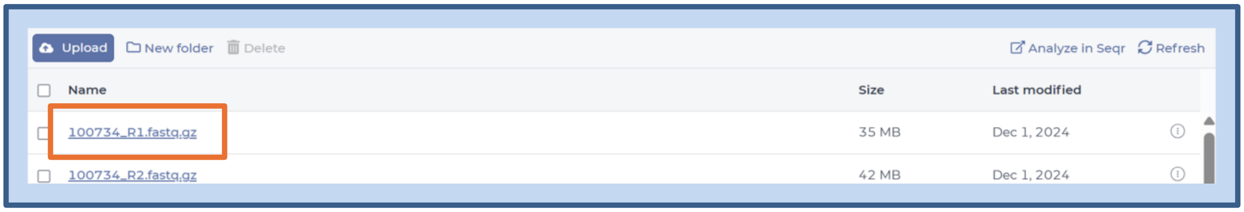
5. **For manually-uploaded reads** via the *Data Uploader*, include read columns (Figure 18)
   a. For **paired-end** (PE) sequencing: *read1* and *read2*



	A	B	C	D	E	F
1	HAI_id	read1	read2	run_id	sample_matrix	county
2	2168435186	100734_R1.fastq.gz	100734_R2.fastq.gz	SEQ217	NP swab	Adams
3	2168435187	100735_R1.fastq.gz	100735_R2.fastq.gz	SEQ217	Buccal swab	Alameda
4	2168435188	100735_R1.fastq.gz	100735_R2.fastq.gz	SEQ217	Buccal swab	Tulare
5	2168435189	100736_R1.fastq.gz	100736_R2.fastq.gz	SEQ217	NP swab	Gilpin

Figure 18

   b. For **single-end** (SE) sequencing: only *read1*
   c. Use the *file name* as the value for this column (e.g. 100734_R1.fastq.gz) (Figure 19)



Name	Size	Last modified
100734_R1.fastq.gz	35 MB	Dec 1, 2024
100734_R2.fastq.gz	42 MB	Dec 1, 2024

Figure 19

6. **For SRA uploads:**
   a. *Label the column header* for column 2 as *sra_accession*, or similar
   b. Enter SRA accessions as the values for this column (e.g. SRR# or ERR#)

HAI_id	sra_accession	hospital	month	year	age	run_id
2168435186	ERR1023740	D	4	2013	30-39	SEQ217
4831845358	SRR11445892	A	2	2014	80-89	SEQ217
8415835241	ERR4087740	C	6	2013	60-69	SEQ217
6846813545	SRR11842392	A	3	2013	80-89	SEQ217

7. In Excel, click *save as* and change the file type to *Text (Tab delimited).*
8. Refer to section 4.2 to import TSV files at the same time as creating new read collections (importing files locally) and section 4.3 to import TSV files without creating read collections

10.2 IMPORTING A WORKFLOW FROM DOCKSTORE

1. In the *Terra workspace* of interest, navigate to the *workflows* tab and click *find a workflow* (Figure 20)
2. In the pop-up window, click *Dockstore.org* (Figure 21)
3. Workflows may be found through the search bar or by navigating through the organization if it is known (Figure 22)
   a. To find a Theiagen workflow
      i. click *organizations* (Figure 2)
      ii. In the search bar type *Theiagen*
      iii. Click on the *logo*, *view*, or *# collections.* (Figure)
      iv. Click on the collection to see available workflows: *Public Health Bioinformatics (PHB)*
4. Find and *open the workflow* (workflow name listed at the end of the file path) (Figure 24)
5. Click *Terra* to launch the workflow in Terra (Figure)
6. Choose the *destination workspace* in the dropdown
7. Click *import* or *create a new workspace* (Figure 26)

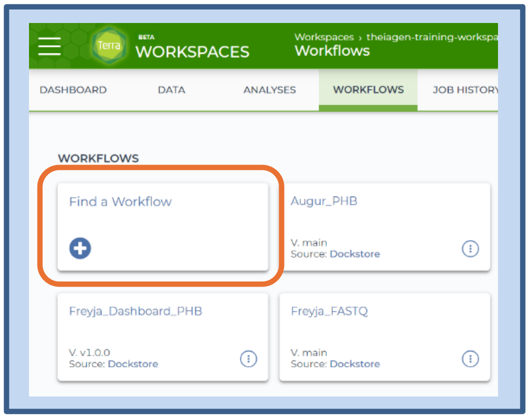
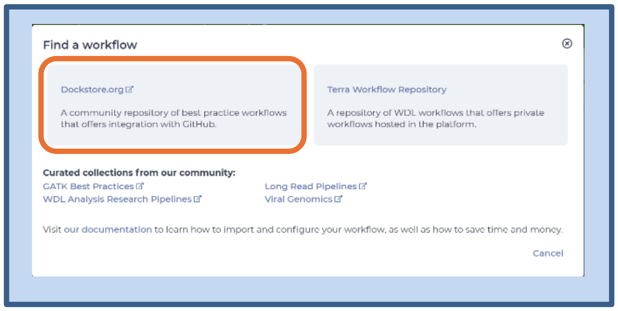


Figure 20



Dockstore.org

A community repository of best practice workflows that offers integration with GitHub.

Terra Workflow Repository

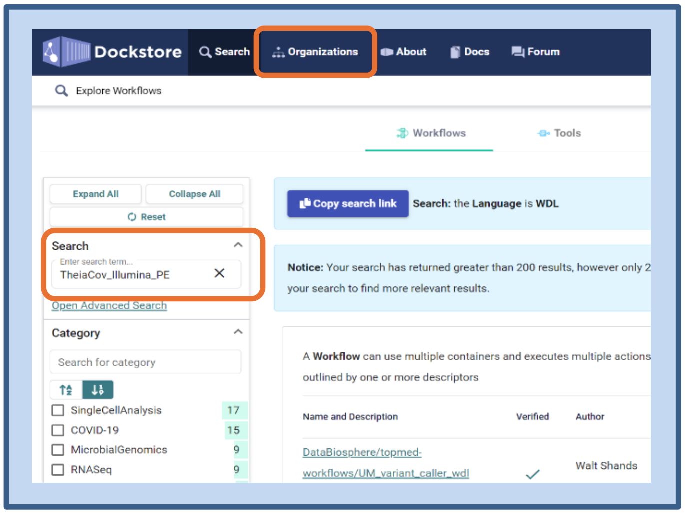
A repository of WDL workflows that offers private workflows hosted in the platform.

Curated collections from our community:

GATK Best Practices
WDL Analysis Research Pipelines
Long Read Pipelines
Viral Genomics

Visit our documentation to learn how to import and configure your workflow, as well as how to save time and money.

Figure 21



Search: the Language is WDL

Search: Enter search term... TheiaCov_Illumina_PE

Open Advanced Search

Notice: Your search has returned greater than 200 results, however only 2

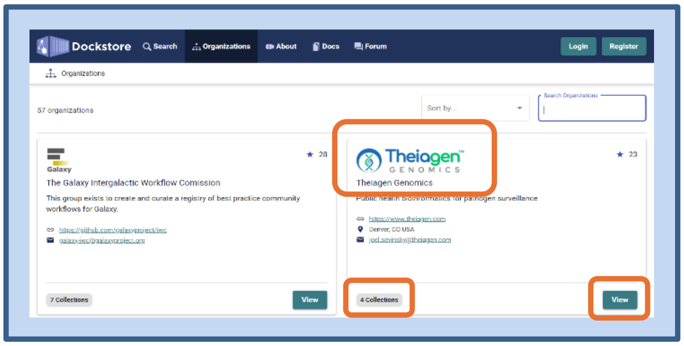
your search to find more relevant results.

A Workflow can use multiple containers and executes multiple actions

outlined by one or more descriptors

DataBiosphere/topmed-workflows/UM_variant_caller_wdl ✓ Walt Shands

Figure 22



57 organizations

The Galaxy Intergalactic Workflow Comission

This group exists to create and curate a registry of best practice community workflows for Galaxy.

https://github.com/galaxyproject/iwc

galaxy-iwc@galaxyproject.org

Theiagen Genomics

Public health bioinformatics for pathogen surveillance

https://www.theiagen.com

Denver, CO USA

joel.sevinsky@theiagen.com

Figure 23

Collection: Public Health Bioinformatics (PHB)

Terra-accessible workflows for public health pathogen genomics

Workflows and Tools

github.com/theiagen/public_health_bioinformatics/TheiaCoV_ClearLabs_PHB:v1.0.0

Last updated Jul 21, 2023

github.com/theiagen/public_health_bioinformatics/TheiaCoV_ONT_PHB:v1.0.0

Last updated Jul 21, 2023

About the Collection

Public Health Bioinformatics (PHB)

Bioinformatics workflows for characterization, epidemiology and sharing of pathogen genomes.

More information about the steps undertaken in these workflows is available via the Theiagen Public Resources Documentation.

Figure 24

github.com/theiagen/public_health_bioinformatics/TheiaCoV_ClearLabs_PHB:v1.0.0

Tag created: 38 days ago

Last update to source repository: 4 hours ago

Labels theiagen-phb

Workflow Information

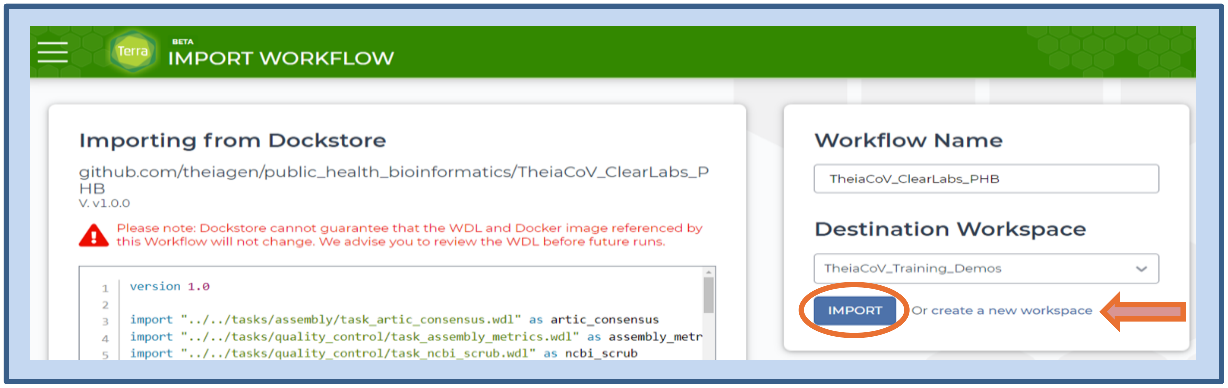
Source Code:...

TRS: #workflow/github.com/theiagen/public_health_bioinformatics/TheiaCoV_ClearLabs_PHB

Topic: Bioinformatics workflows for genomic characterization, submission preparation, and genomic epidemiology of pathogens of public health concern.

Checker Workflow: n/a

Figure 25



IMPORT WORKFLOW

Importing from Dockstore

github.com/theiagen/public_health_bioinformatics/TheiaCoV_ClearLabs_PHB
V. v1.0.0

Please note: Dockstore cannot guarantee that the WDL and Docker image referenced by this Workflow will not change. We advise you to review the WDL before future runs.

```
version 1.0

import "../../tasks/assembly/task_artic_consensus.wdl" as artic_consensus
import "../../tasks/quality_control/task_assembly_metrics.wdl" as assembly_metr
import "../../tasks/quality_control/task_ncbi_scrub.wdl" as ncbi_scrub
```

Workflow Name

TheiaCoV_ClearLabs_PHB

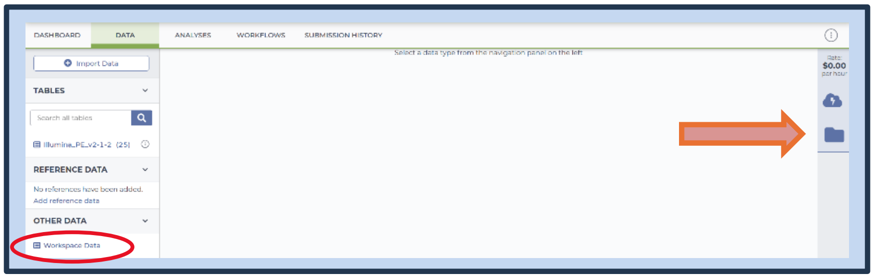
Destination Workspace

TheiaCoV_Training_Demos

Figure 26

10.3 ADDING AND UPDATING WORKSPACE DATA ELEMENTS AND FILES

1. Navigate to the *Terra workspace* where analyses will be run
2. To upload local files:
    a. Open the *workspace files* tab in the right-side panel of the workspace (Figure 7).
    b. Click *upload*
    c. Following upload, click the *clipboard* or *right click* on the file name to *copy link*
3. To add workspace variables:
    a. Click the *workspace data* tab at the bottom of the left-side panel (Figure 27)
    b. Click *Edit* in the top menu (Figure 28)
    c. Click *Add Variable* (Figure 8)
    d. Click in the *key field* and *name the element* being added (Figure 9)
    e. In the value field, choose *string* as the value type
    f. Paste the file path copied above in step 2, or enter the desired file path (see below)
        i. *E.g. to add the Artic V4-1 primer bed file, the key* Artic_V4-1_primer_bed *may be used with the file path as the value* (Figure 30)
4. Docker images and tags for TheiaCoV workflows should be included in workspace variables.
    a. Ensure the docker images and dataset tags are aligned with versions used for internal validation procedures or are re-verified before use.
    b. Available in the Docker Image and Reference Materials for SARS-CoV-2
    c. All resources can be found on the Theiagen Public Health Resources page



Select a data type from the navigation panel on the left

Figure 27

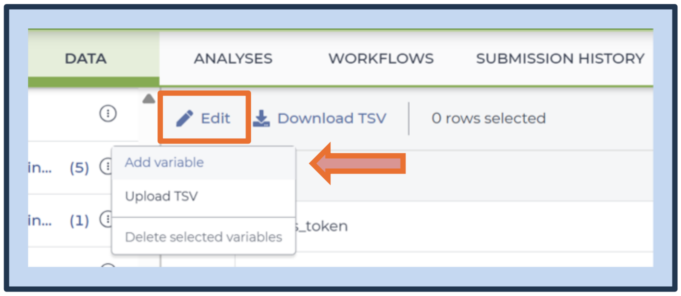
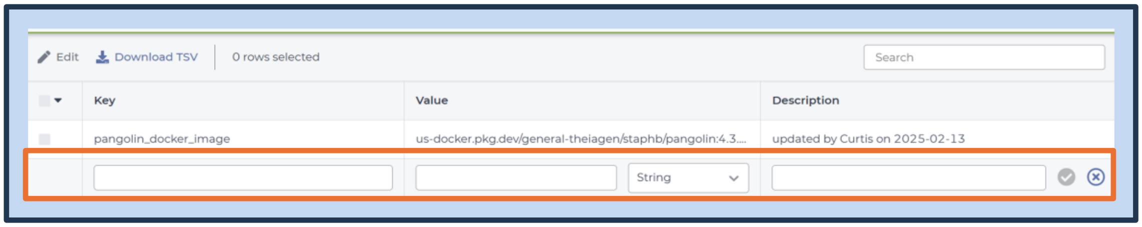
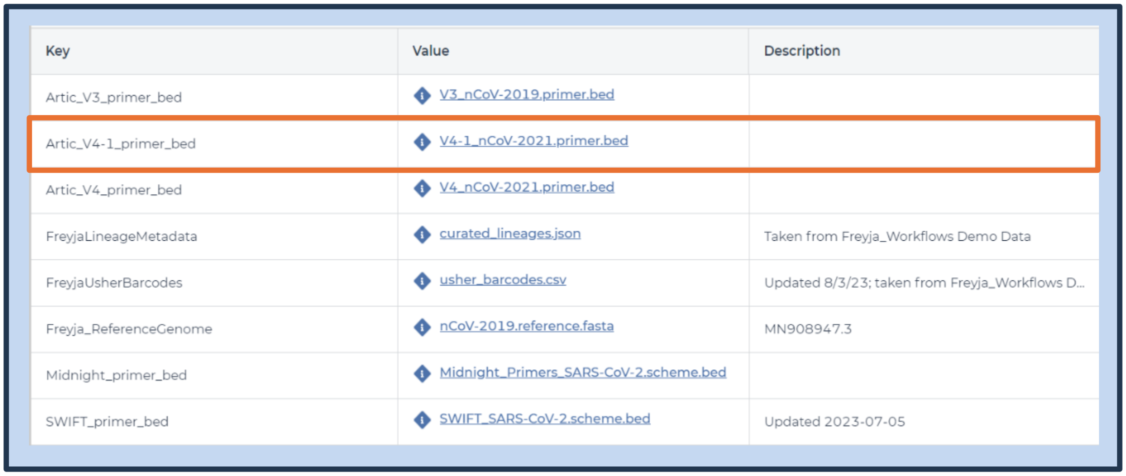


Figure 28



Key	Value	Description
pangolin_docker_image	us-docker.pkg.dev/general-theiagen/staphb/pangolin:4.3...	updated by Curtis on 2025-02-13

Figure 29



Key	Value	Description
Artic_V3_primer_bed	V3_nCoV-2019.primer.bed	
Artic_V4-1_primer_bed	V4-1_nCoV-2021.primer.bed	
Artic_V4_primer_bed	V4_nCoV-2021.primer.bed	
FreyjaLineageMetadata	curated_lineages.json	Taken from Freyja_Workflows Demo Data
FreyjaUsherBarcodes	usher_barcodes.csv	Updated 8/3/23; taken from Freyja_Workflows D...
Freyja_ReferenceGenome	nCoV-2019.reference.fasta	MN908947.3
Midnight_primer_bed	Midnight_Primers_SARS-CoV-2.scheme.bed	
SWIFT_primer_bed	SWIFT_SARS-CoV-2.scheme.bed	Updated 2023-07-05

Figure 30