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

Document TG-TER-03, Version 3	
Date:	Workflow Versions:
4/9/2024	PHB v1.3.0 and v2

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.

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)
5. For new data sets, click *create a new collection* (Figure 4)
   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.
7. When upload is complete, files will populate in the table below.

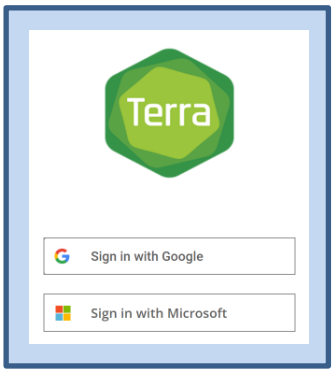


Figure 1

8. Continue to section 4.2 to upload sample metadata.

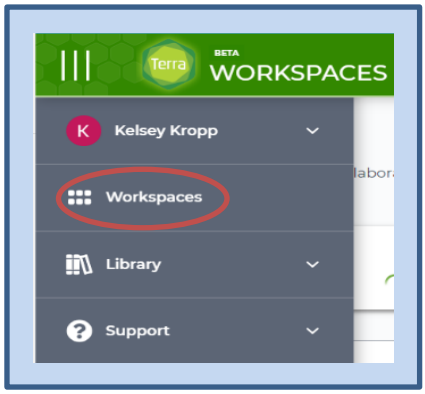


Figure 2

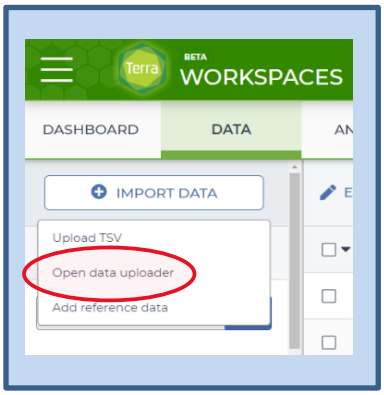


Figure 3

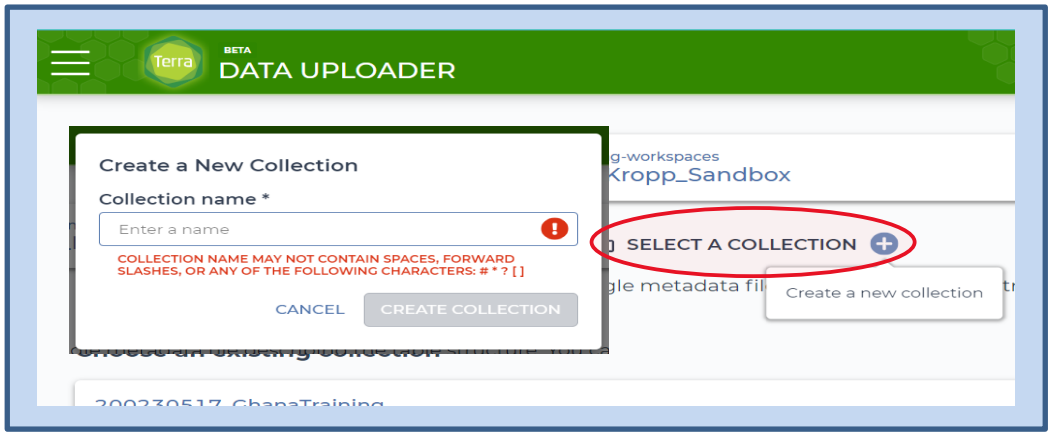


Figure 4

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

1. Once files have successfully uploaded, click *next* (Figure 5) and import the associated sample metadata file or drag and drop from file explorer (Figure 6).
2. Click *create table* (Figure 7).
   a. *See appendix 10.1 to create a sample metadata file.*
3. 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.
4. In the data tab, all read files and metadata are now populated for each sample in the illumina_pe_specimen table; open the data table by clicking on the table name in the sidebar (Figure 8).

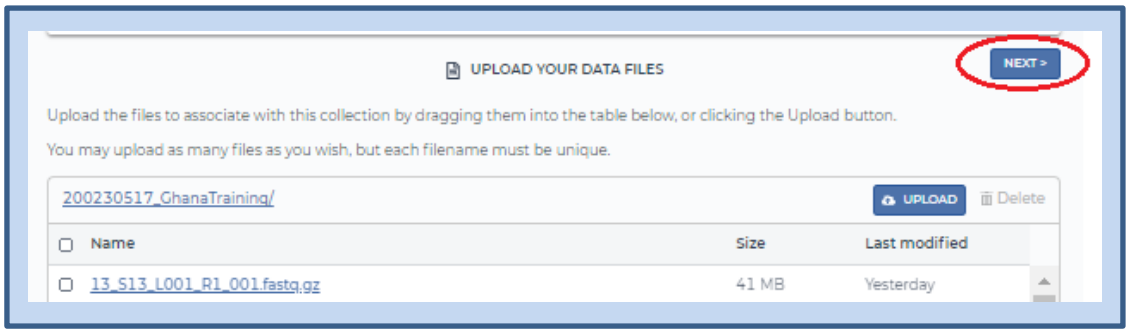


Figure 5

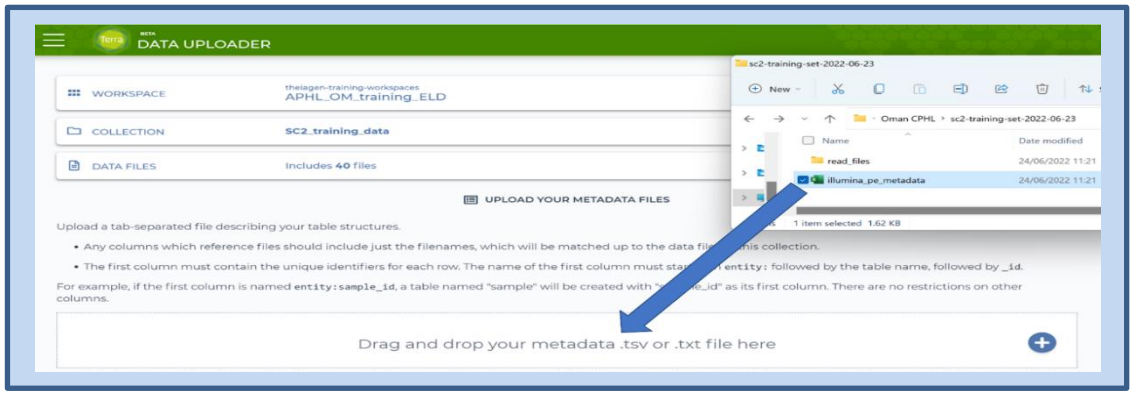


Figure 6

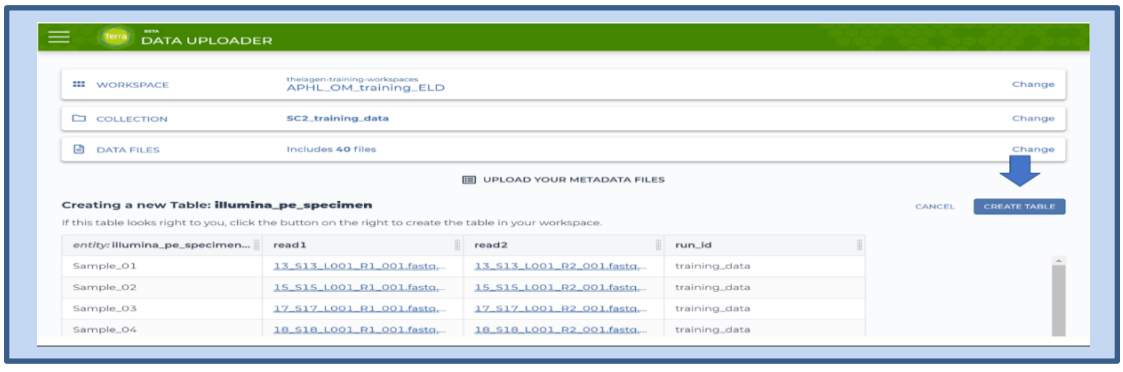


Figure 7

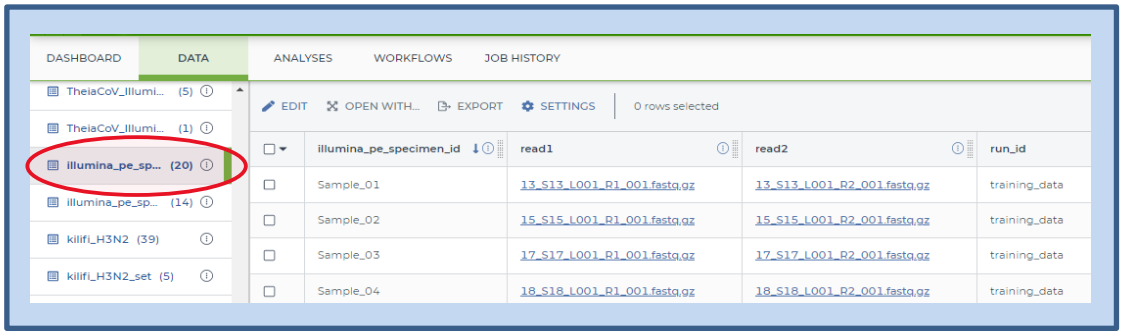


Figure 8

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 9).
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. *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. entity:[DATATABLE_NAME]_id*
4. All samples and metadata should now be populated in the corresponding Terra data table (Figure 8).

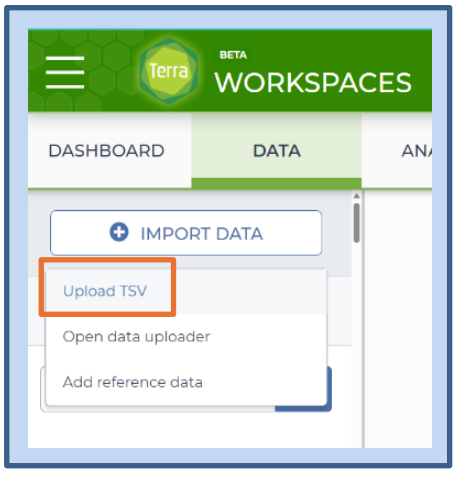


Figure 9

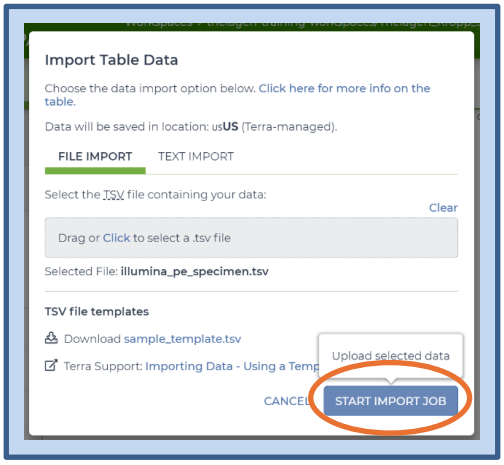


Figure 10

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 11).
3. Import a tsv file containing the table of SRA accession numbers for desired samples; *select* or *drag and drop* the file; click *start import job* (Figure 12).
   a. *A template tsv file can be downloaded from this pop-up; follow section 10.1 to create a metadata/tsv file; for SRA Fetch the tsv file does not need read1 or read2 columns* (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.*

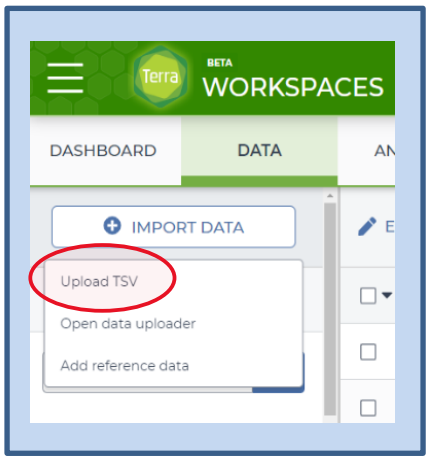


Figure 11

Figure 12

entity: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

7. Select the sample table to use under *select root entity type.*
   a. *Do not choose the "set" table version.*

8. Click *select data* and *mark the checkbox* 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 14).
10. In the outputs tab, click *use defaults* (Figure 15)and *save* the workflow.
11. Then click *run analysis*, enter any comments if desired, and click *launch.*

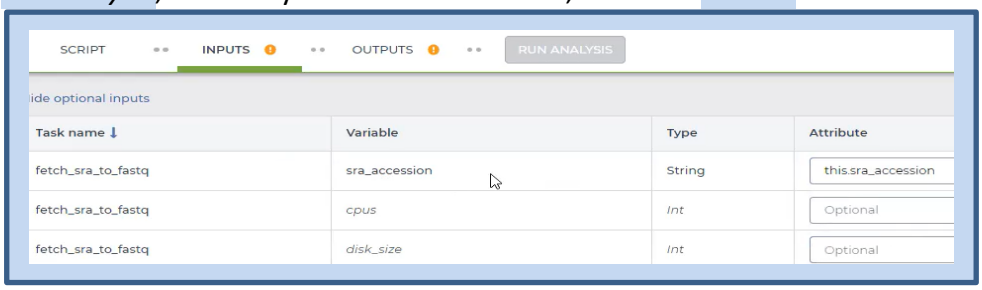


Figure 14

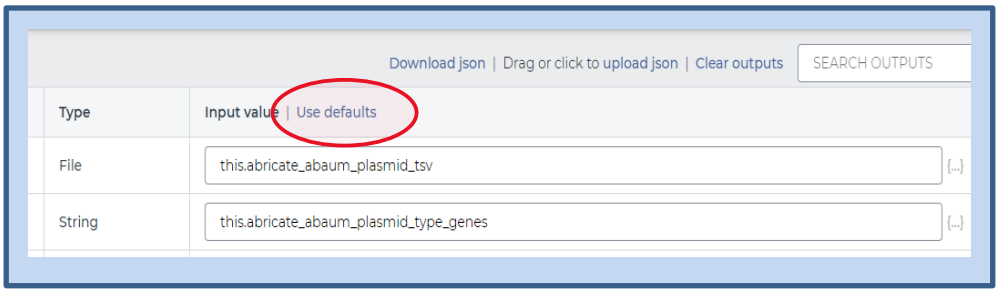


Figure 15

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. Theaigen's BaseSpace Fetch Setup (*https://theiagen.notion.site/BaseSpace_Fetch-34978656aa2d46ba82f2059434bd9369*)
   b. Document TG-TER-04, Version 1: BaseSpaceFetch_v1.docx

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 start with *entity:* and 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. Theiagen's BaseSpace Fetch Setup (initial setup)
2. Theiagen's BaseSpace Fetch (workflow resource page)

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

10. APPENDICES

10.1 CREATING A METADATA FILE (TSV FILE)

1. Open the downloadable tsv template located in the upload tsv pop-up window (Figure 16)
2. Cell A1 must contain the following text: entity:name_id where the prefix contains *entity:* and the suffix contains *_id* (Figure 17); do not include spaces.
3. The middle text should contain text to indicate the project name; for example, if running sequencing run 217, cell A1 may read *entity:SEQ217_id* or for all HAI organisms it may read *entity:HAI_id*

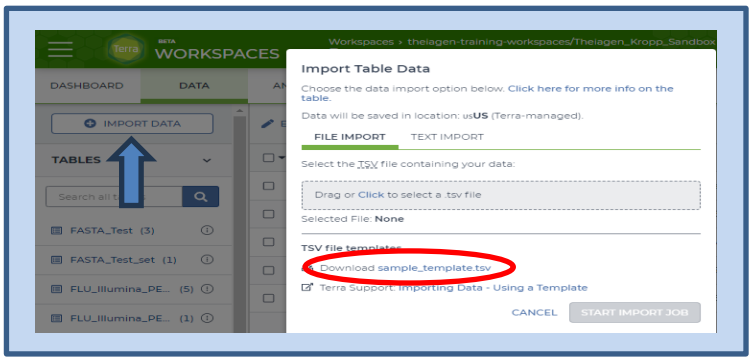


Figure 16

	A	B	C	D
1	entity:HAI_id	run_id	sample_matrix	county
2	2168435186	SEQ217	NP swab	Adams
3	2168435187	SEQ217	Buccal swab	Alameda
4	2168435188	SEQ217	Buccal swab	Tulare
5	2168435189	SEQ217	NP wash	Gilpin

Figure 17

4. Enter all *sample IDs* into column 1 below cell A1.
5. **<u>For paired end sequencing platforms like MiSeq:</u>**
   a. *<u>Optional</u>: Label column headers for desired metadata; do not include spaces:*

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

b. Optional: enter run_id in the first row of column D and the run_id for each sample.

c. Optional: add additional metadata columns and sample information, as needed.

6. For single end sequencing platforms like Illumina single-end sequencing, MinION, PacBio, etc:

a. Optional: Label column headers for desired metadata; do not include spaces:

Entity:HAI_id	run_id	sample_matrix	
2168435186	SEQ217	NP swab	
2168435187	SEQ217	Buccal swab	

b. Optional: enter run_id in the first row of column C and the run_id for each sample.

c. Optional: add additional metadata columns and sample information, as needed.

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.

9. For SRA uploads:

a. *Label the column header* for column 2 as *sra_accession*, or similar.

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

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 18).
2. In the pop-up window, click *dockstore* (Figure 19).
3. Workflows may be found through the search bar or by navigating through the organization if it is known.
   a. To find a Theiagen workflow, for example, click *organizations* (Figure 20).
      i. In the search bar type *Theiagen.*(Figure 21).
      ii. Click on the *logo*, *view*, or *# collections.* (Figure 22).
      iii. Click on the *collection* of interest to see all available workflows.
         1. Find and *open the workflow*; the workflow name is listed at the end of the file path.
         2. Click *Terra* to launch the workflow in Terra (Figure 23).
         3. Choose the *destination workspace* in the dropdown.
         4. Click *import* or *create a new workspace*. (Figure 24).

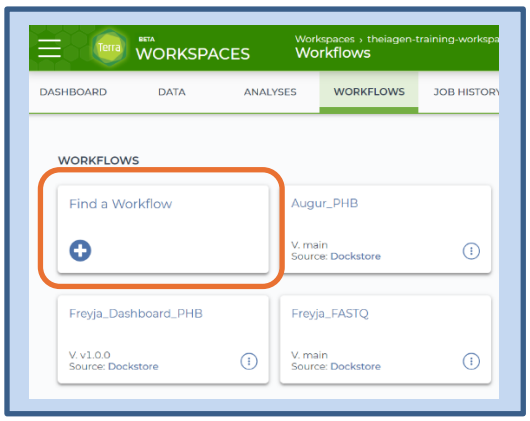


Figure 18

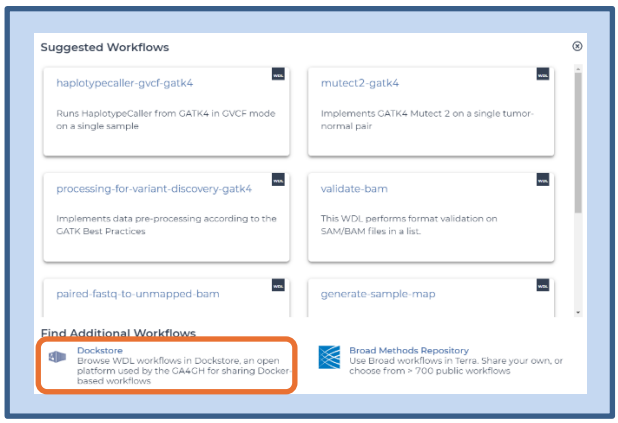


Figure 19

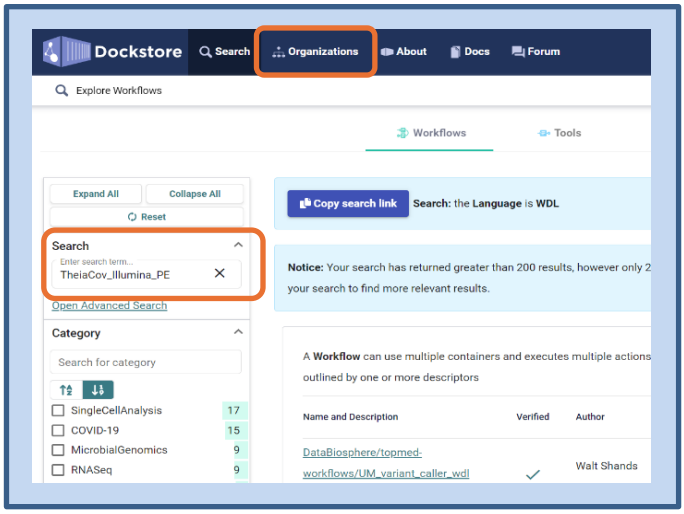


Figure 20

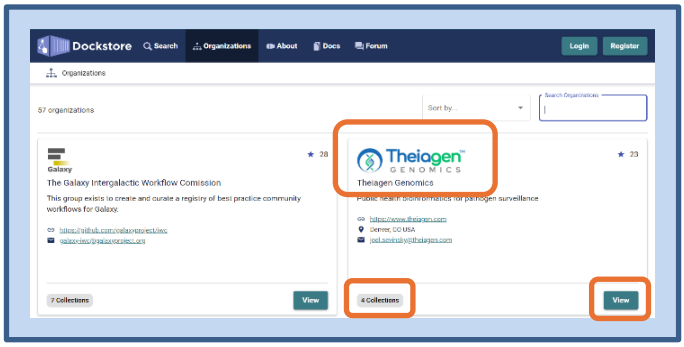


Figure 21

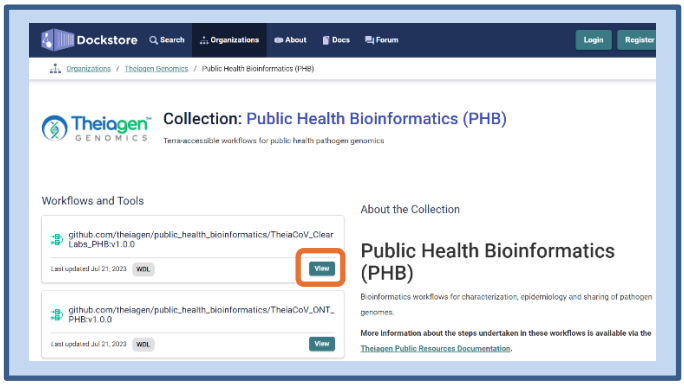


Figure 22

Figure 23

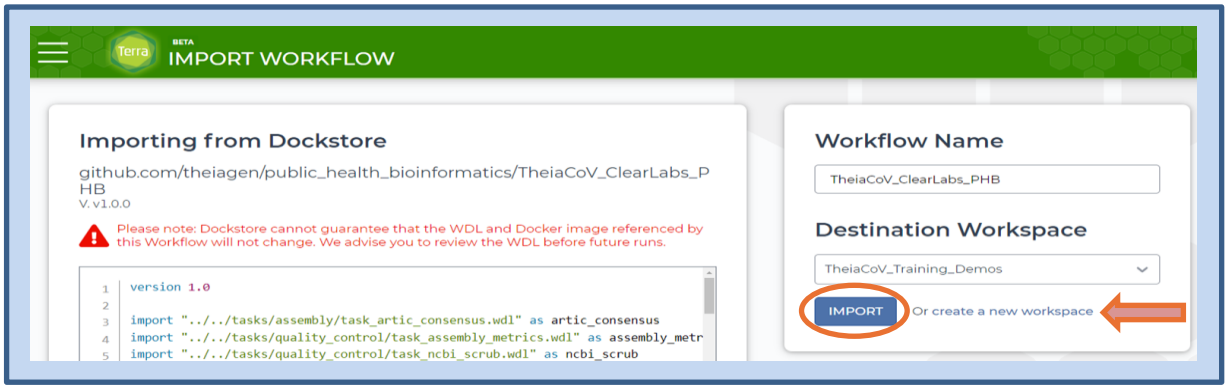


Figure 24

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, open the *Files* tab in the bottom left of the workspace (Figure 25).
   a. Click *upload.*
   b. Once the upload is complete, *right click* on the file name and click *copy link.*
3. Open the *workspace data* tab and click the *blue plus symbol* in the bottom right (Figure 25).
4. Click in the *key field* and *name the element* being added (Figure 26).
   a. E.g. to add the Artic V4-1 primer bed file, the key Artic_V4-1_primer_bed *may be used to specify its use with the Phoenix workflow.*
5. In the value field, choose *string* as the value type.
   a. Paste the file path copied above in step 2 or the desired file path.
      i. Docker images and tags for TheiaCoV workflows can be found on the Theiagen Public Health Resources page under key resources or at: *https://theiagen.notion.site/Docker-Image-and-Reference-Materials-for-SARS-CoV-2-Genomic-Characterization-98328c61f5cb4f77975f512b55d09108*
         1. Ensure the docker images and dataset tags are aligned with versions used for internal validation procedures or are re-verified before use.

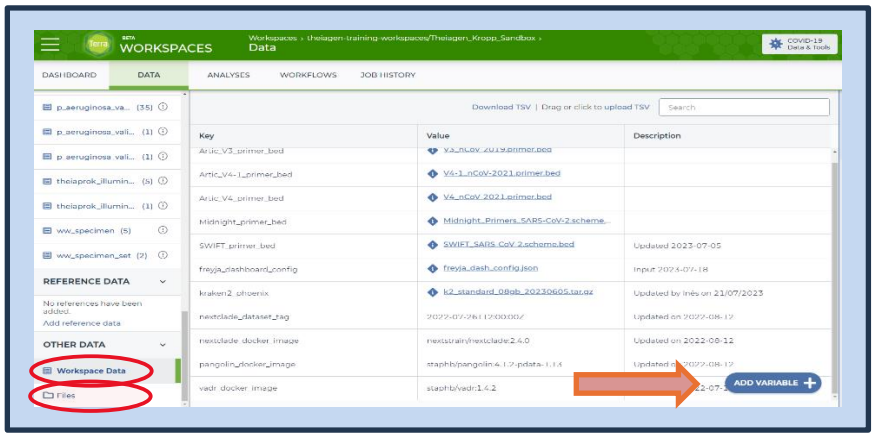
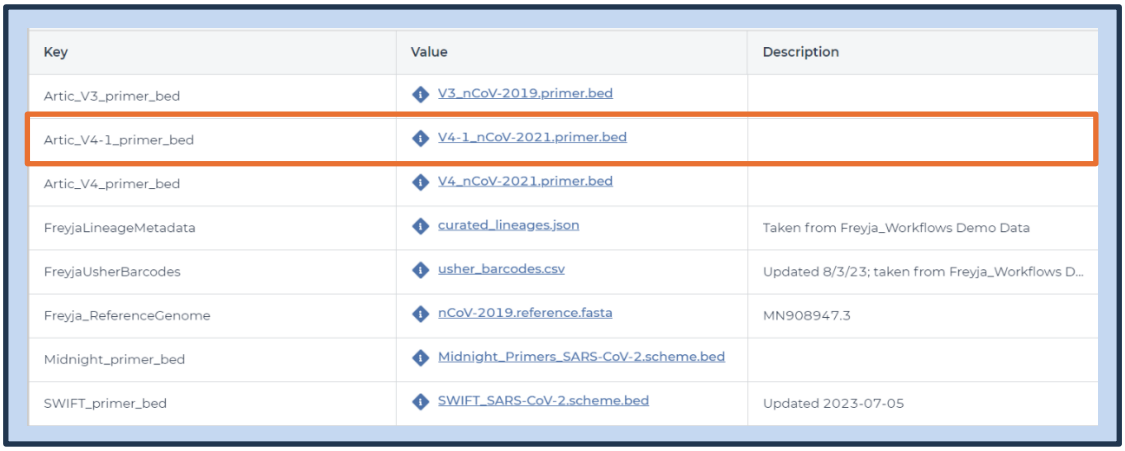


Figure 25



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 26