| | | |
|---|---|---|
| | **Analyzing SARS-CoV-2 Data in Terra using Theiagen's TheiaCoV Illumina SE Workflow Version 4** | |
| | Document TG-SC2-SE, Version 4 | |
| | Date: | Workflow Versions |
| | 5/1/2025 | PHB v3.0.0 |

## 1. PURPOSE/SCOPE

To standardize the process of analyzing SARS-COV-2 (SC2) next generation sequencing (NGS) data using Theiagen's TheiaCoV_Illumina_SE_PHB workflow in Terra to generate assemblies, quality control (QC) metrics, and determine Nextclade clade and Pangolin lineage assignments. Acceptable data types include Illumina's single end (SE) raw read file format. Read the 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
- Illumina SE raw read files uploaded to Terra workspace, see TG-TER-03 or TG-TER-04
- Theiagen's TheiaCoV_Illumina_SE_PHB workflow in Terra, see appendix 10.2

IMPORTANT NOTES

- Metadata *column headers* and *workflow input* text indicated in *gray* in this SOP are customizable; *black* is required text
- 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
- Filter for workspace data and files in workflow input dropdowns using the prefix *workspace.*

## 3. RELATED DOCUMENTS

| Document Number | Document Name |
|---|---|
| TG-TER-03 | Uploading Local or SRA NGS Data & Creating a Results Metadata Table in Terra |
| TG-TER-04 | Linking BaseSpace and Importing BaseSpace Reads to Terra Workspace |

## 4. PROCEDURE

### 4.1 RUNNING THE THEIACOV WORKFLOW

1. Open Terra and navigate to the *workflows* tab within the workspace containing SC2 data
2. Select the *TheiaCoV_Illumina_SE_PHB* workflow (Fig 1)
3. *Uncheck call caching* (Fig 2)
4. Choose the *latest version* of the workflow, or the version internally validated (Fig 2, a)
5. Select the second bullet to *run workflow(s) with inputs defined by data table* (Fig 2, b)
6. Select the relevant data table under the *select data table* dropdown (Fig 2, c)
7. Click *select data* (Fig 2, d)

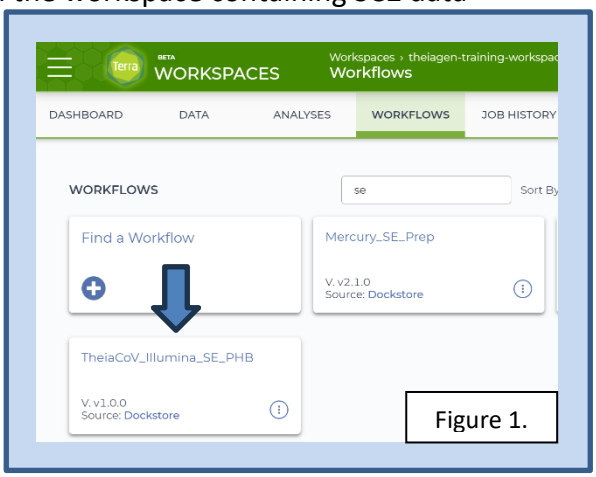


Figure 1.

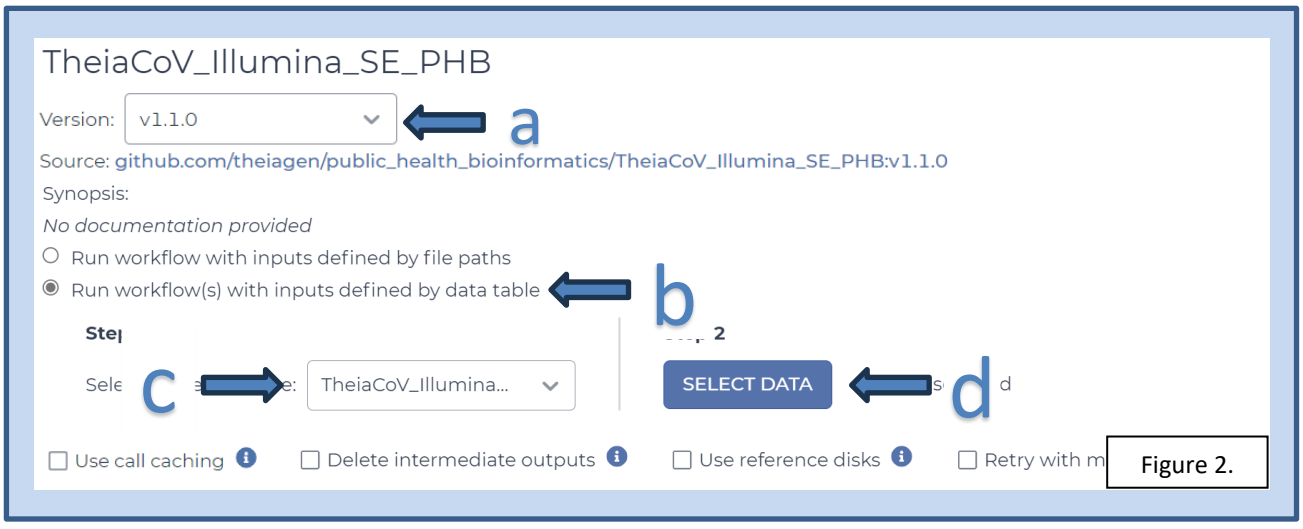


Figure 2.

8. In the pop-up window *select the checkbox* for each sample to be included in the analysis (Fig 3)
   a. *Click the down arrow and select all to process all specimens*
   b. *Additionally, a subset of samples may be chosen using the search bar to filter before selecting the checkbox at the top to only select samples matching the search criteria*
   c. Scroll to the bottom and click *ok*

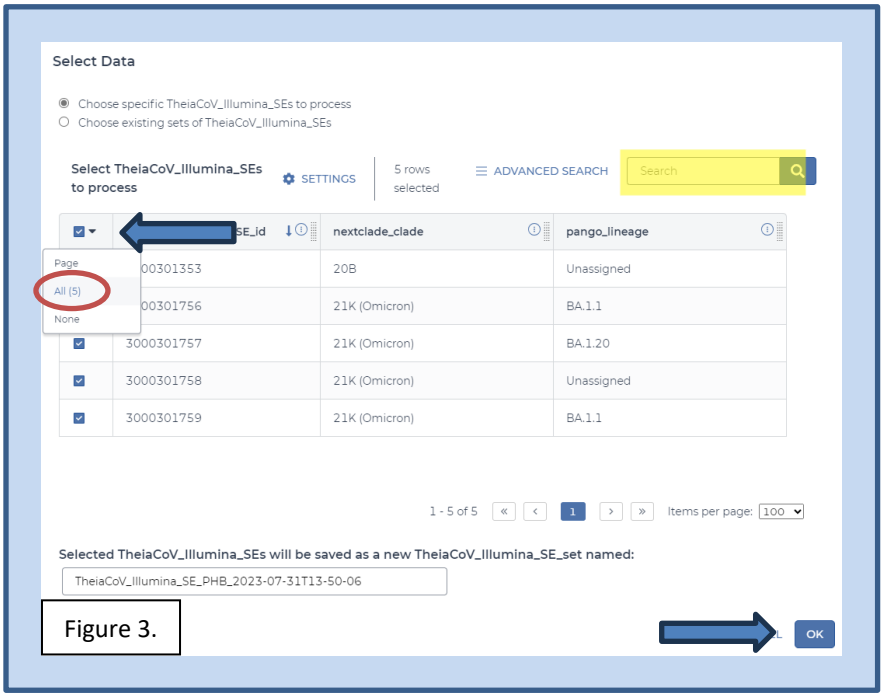

Figure 3.

9. Specify the desired *dataset tags* and *docker image* inputs
   a. To run TheiaCoV_Illumina_SE_PHB for the first time or configure with the newest dataset tags and docker images, *upload the TheiaCov input json file* on the inputs tab by navigating to the Key Resources Notion page titled Docker Image and Reference Materials for SARS-CoV-2 Genomic Characterization
      i. ***NOTE**: TheiaCoV PHB v2.0.0+ workflows are not backwards compatible with older versions of Nextclade; use Nextclade Dataset Tag 2024-04-15—15-08-22Z or newer*
      ii. Expand the *TheiaCoV in PHB (v2.0.0 or higher)* section, followed by the *Terra.Bio Input JSONs for PHB v2.0.0 or higher*; click on the json file associated with the Illumina PE sequencing platform, TheiaCoV_Illumina_SE_PHB_2025-04-02.json or newer
      iii. *Right click* and *save* the file (text does not have to be selected to save properly)
      iv. Return to the workflow in Terra, click *upload json* (Fig 4, red circle), *select* the saved json file, and click *open*
   b. *To run the workflow with previously saved dataset tags and docker images, no changes are needed*
   c. *To add docker images and dataset tags as workspace files for availability in input dropdowns, refer to appendix 10.2*

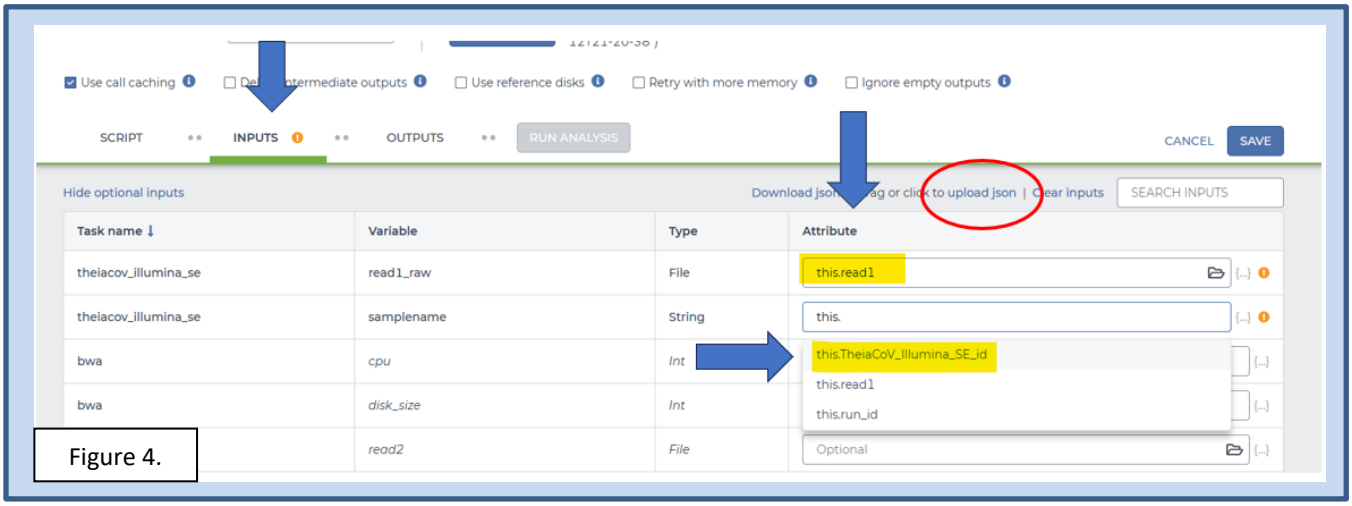

Figure 4.

10. Set the first two attributes in the table to *this.read1* and *this.TheiaCoV_Illumina_SE_id* (Fig 4)
    a. Where *TheiaCoV_Illumina_SE* is the unique name of your data table in Terra
11. Manually choose the *primer_bed* file for the primer set used to sequence samples
    a. *Ctrl + F* and search for *bed* to highlight this field in Windows environments
    b. Labs using the Artic V4-1 will choose *workspace.Artic_V4-1_primer_bed*; for other primer bed files, see Docker Image and Reference Materials for SARS-CoV-2 Genomic Characterization for available primer bed files
        i. *To add workspace files for availability in input dropdowns, refer to appendix 10.2*
12. Specify outputs by clicking on the *outputs* tab and *use defaults* (Fig 5)
13. Click *save*
14. Launch the workflow by clicking *run analysis* (Fig 5); enter desired comments and click *launch*

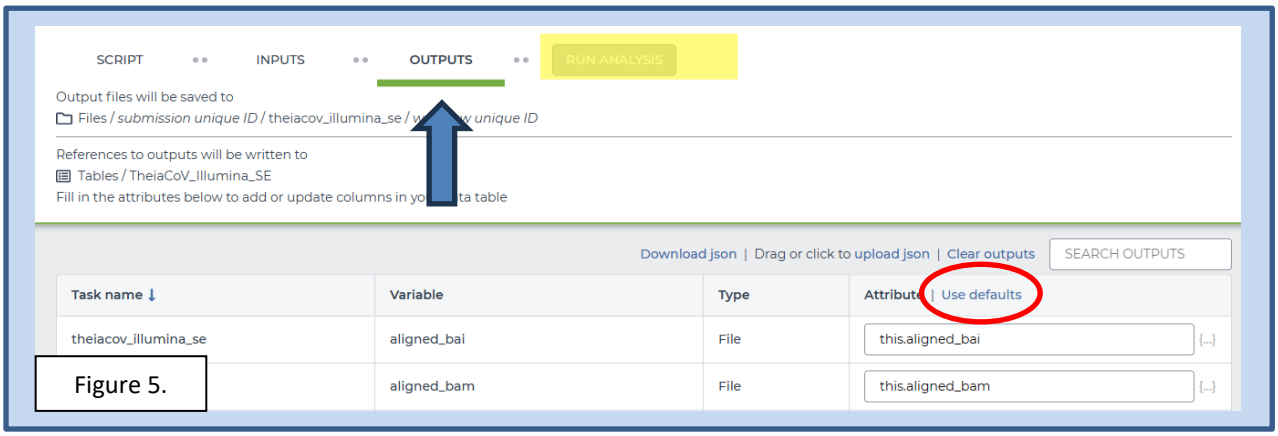

Figure 5.

## 4.2 QUALITY ASSESSMENT OF THEIACOV OUTPUTS

1. Navigate to the *data* tab of the workspace containing SC2 data and open the pertinent data table
2. Click *Select Columns* (Fig 6, green rectangle) and select *none* to deselect all output columns (Fig 6, yellow highlight)
3. To simplify the table, select the three following outputs that will be used to make a QC assessment: *assembly_length_unambiguous*, *Number_N*, and *percent_reference_coverage*
   a. *Optional: Save this selection by clicking in the* save this column selection *field and* naming it *(e.g. QC_assessment); do not include any spaces in the name (Fig 6, red rectangle)*
   b. Click *done*

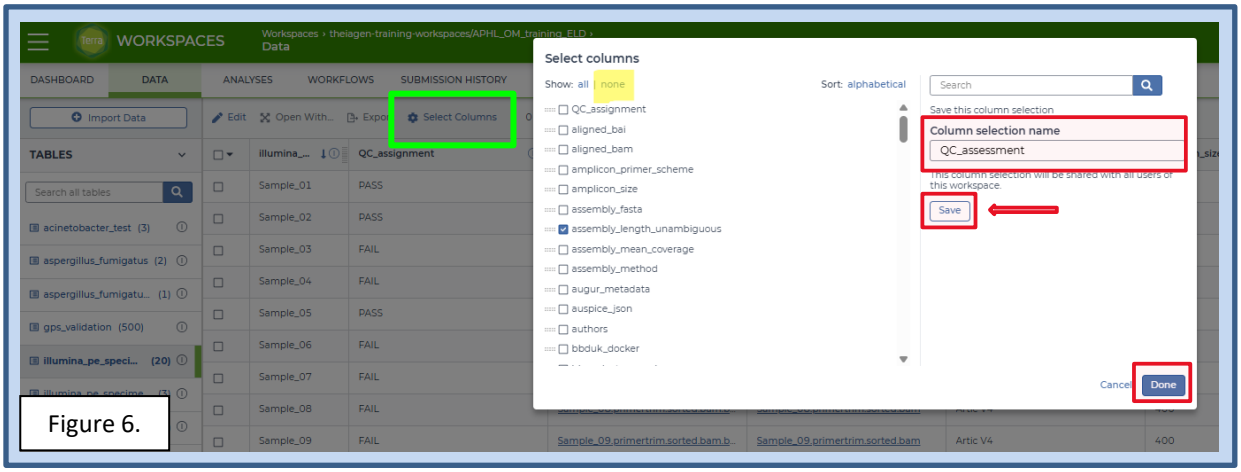


Figure 6.

4. *Optional: Add a column to record QC PASS/FAIL by clicking edit, add a column (Fig 7)*
   *a. Name the new column (e.g. QC_Call); do not include any spaces*
   *b. Set the value type as string*
   *c. Click save*

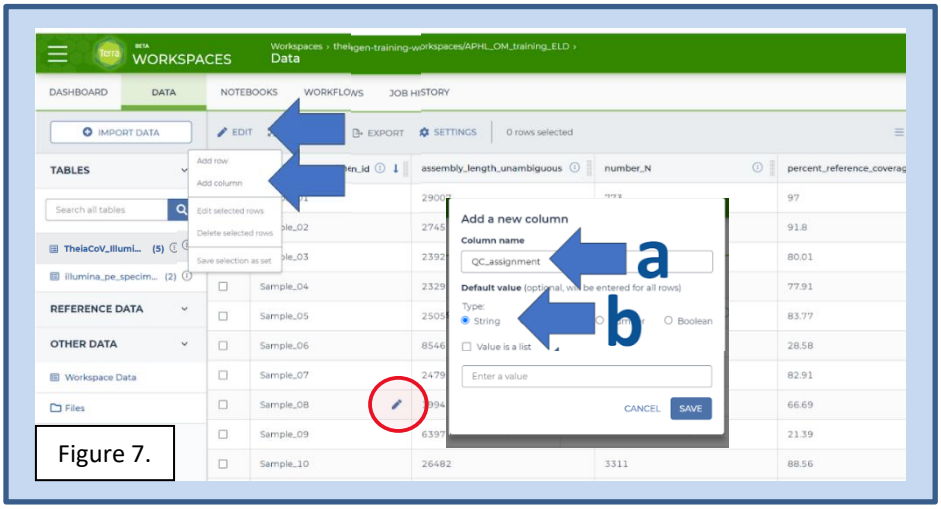


Figure 7.

5. Use table 1 to assess the quality of each sample's genome assembly (see next page) &/or lab-specific quality metrics
6. *Optional: notate in the QC_assessment field for each sample* PASS *or* FAIL *by* clicking the pencil icon *in the corresponding field (Fig 7, red circle)*
7. For samples that pass the guidance thresholds, proceed to section 4.3
   a. For samples that do not pass guidance thresholds, resequence
      i. Samples not meeting guidance thresholds indicated here may proceed to analysis at the discretion of the laboratory

**Table 1. Guidance thresholds for genome assembly QC**

| QC Metric | Data Table Column | Guidance Threshold*[1] |
|---|---|---|
| Number N | *number_N* | <5kbp |
| Assembly length unambiguous | *assembly_length_unambiguous* | >24kbp |
| Percent reference coverage | *percent_reference_coverage* | >83% |

[1] Metrics and thresholds are presented for guidance only as there are currently no standard assembly metric requirements; internal validation procedures will ultimately define acceptable assembly QC parameters

### 4.3 DETERMINING SARS-CoV-2 CLADES, LINEAGES, AND WHO VARIANTS OF CONCERN (VoC)

1. Navigate to the *data* tab of the Terra workspace containing SC2 data of interest
2. *Open the data table* by clicking on the name of the data table in the left sidebar
3. View *settings* above the data table (Fig 6), select *none* (Fig 7)
4. Select the following columns: *nextclade_clade* and *pango_lineage*
   a. *Optional: save this column group for future use by clicking the* save this column selection *field,* naming it *(e.g. SC2_Results), and clicking* save
5. Click *done*
6. Determine the Nextclade clade for each sample
   a. In the data table, find the column titled *nextclade_clade*; result formats will use the following nomenclature: *21L (Omicron)* where:
      i. *21L* indicates the sample clade and
      ii. In parentheses, *(Omicron)*, contains the WHO variant of concern classification
         1. *Not every sample will belong to a WHO classification*
   b. *Samples indicated as recombinant may indicate a case where multiple strains have combined during viral replication producing a new lineage*
   c. *More information on SARS-CoV-2 recombinants can be found at the following Github site: pipeline-resources/docs/sc2-recombinants.md at main · pha4ge/pipeline-resources · GitHub*
7. Identify the Pangolin lineage for each sample
   a. In the data table, find the column titled *pango_lineage*; nomenclature will be similar to the following: B.1.167
   b. *For more information on each of the lineages, visit https://cov-lineages.org/lineage_list.html*
8. Follow lab-specific resulting and reporting procedures, as applicable

## 5. QUALITY RECORDS

- Raw read files
- Workflow version and input parameters
- Reference sequence, if applicable
   a. SC2: Wu, F., et al. (2020). Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome. NC_045512.2. [FASTA Genome Assembly]. NCBI. https://www.ncbi.nlm.nih.gov/nuccore/1798174254.
- Sample read, assembly, and result-specific QC metrics
- All workflow outputs relevant to results, including tool and database versions

## 6. TROUBLESHOOTING

- Consult with internal staff familiar with this procedure or contact support@theiagen.com for troubleshooting inquiries
- For document edit requests, contact support@theiagen.com

## 7. LIMITATIONS

1. This SOP is written for the analysis of SC2 data; v2+ of the TheiaCoV_Illumina_SE_PHB workflow is also compatible with the following pathogens: monkeypox virus (MPXV) and west nile virus (WNV). Refer to Theiagen Public Health Resources documentation for organism-specific parameters and details.
2. *TheiaCoV PHB v2.0.0+ workflows are not backwards compatible with older versions of Nextclade; use Nextclade Dataset Tag 2024-04-15—15-08-22Z or newer*

## 8. REFERENCES

1. Smith, E., Wright, S., & Libuit, K. (2022, June 28). *Identifying SARS-CoV-2 Recombinants*. Github. Retrieved June 16, 2023, from https://github.com/pha4ge/pipeline-resources/blob/main/docs/sc2-recombinants.md#identifying-sars-cov-2-recombinants
2. O'Toole, Áine et al. "Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2 with grinch." *Wellcome open research* vol. 6 121. 17 Sep. 2021, doi:10.12688/wellcomeopenres.16661.2
3. 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.
4. Theiagen Genomics Public Health Bioinformatics Workflow Documentation.

## 9. REVISION HISTORY

| Revision | Version | Release Date |
|---|---|---|
| Document creation | 1 | 7/2023 |
| Added TG-TER-04 reference, uncheck call caching, updated input json, figures, and formatting | 2 | 9/2023 |
| Removed section 4.1 for creating a metadata tsv file (refer to TG-TER-03 and TG-TER-04 for details); updated quality records and limitations sections; added primer bed file upload instructions; added appendices 10.1 and 10.2 | 3 | 4/2024 |
| Updates for version release, aligning with Terra interface, formatting | 4 | 5/2025 |
| | | |

# 10. APPENDICES

## 10.1 IMPORTING THE THEIACOV_ILLUMINA_SE_PHB WORKFLOW FROM DOCKSTORE

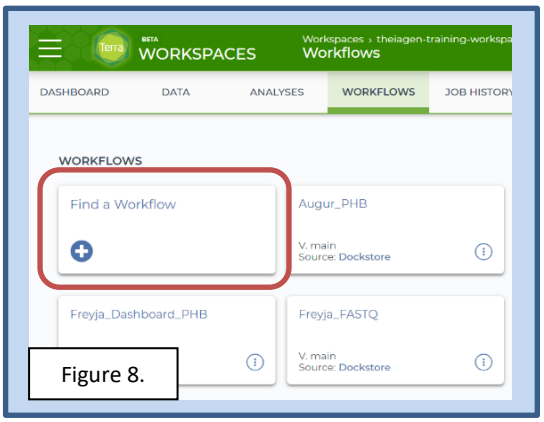

Figure 8.

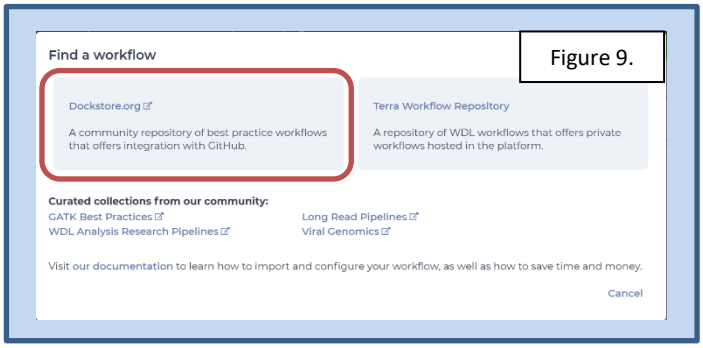

Figure 9.

1. In the *Terra workspace* of interest, open the *workflows* tab and click *find a workflow* (Fig 8)
2. In the pop-up window, click *dockstore* (Fig 9)
3. In the top banner click *Organizations*; then click *Theiagen Genomics* (Fig 10)
4. Open the Public Health Bioinformatics (PHB) collection (Fig 11)

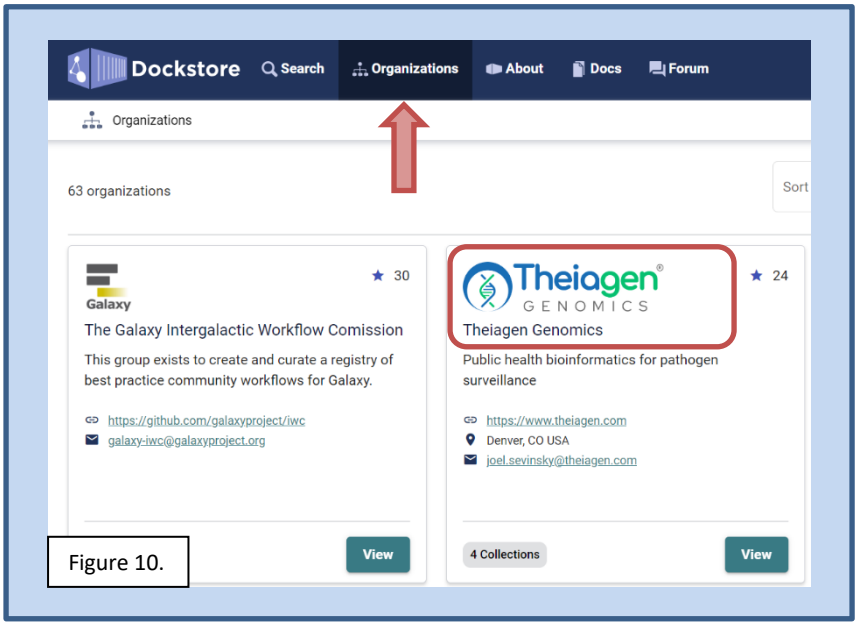

Figure 10.

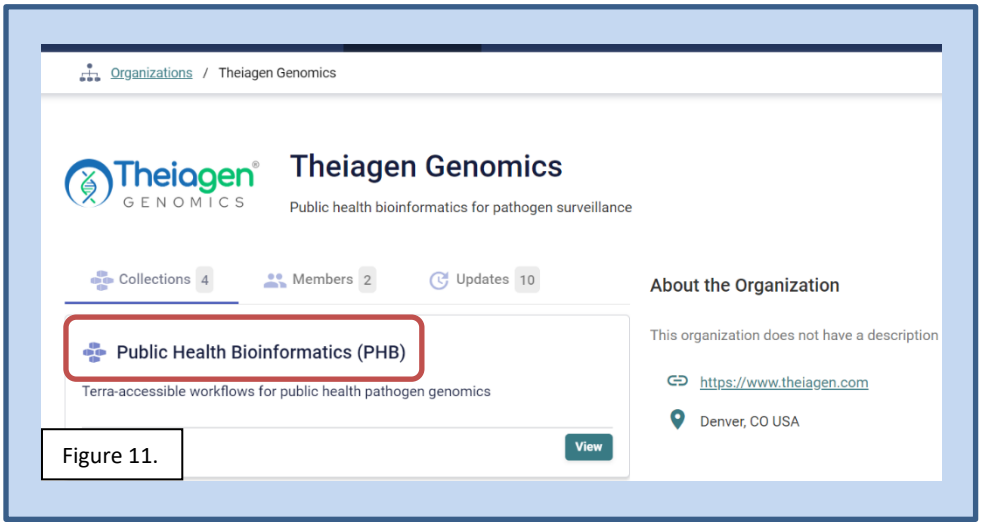

Figure 11.

5. To find the TheiaCoV_Illumina_SE_PHB workflow in Windows environments, hold *Ctrl + F* and *search TheiaCoV_Illumina_SE*, then click on the link (Fig 12)

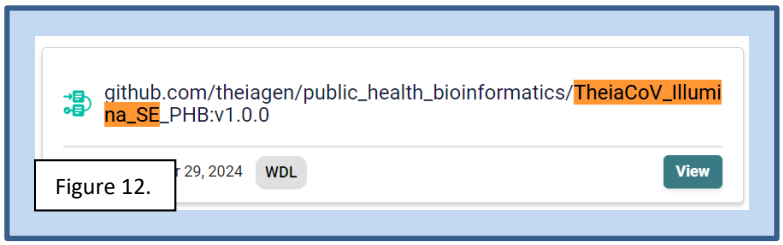

Figure 12.

6. Click *Terra* to launch the workflow in Terra (Fig 13)

Figure 13.

7. Choose the *destination workspace* in the dropdown and click *import* (Fig 14)

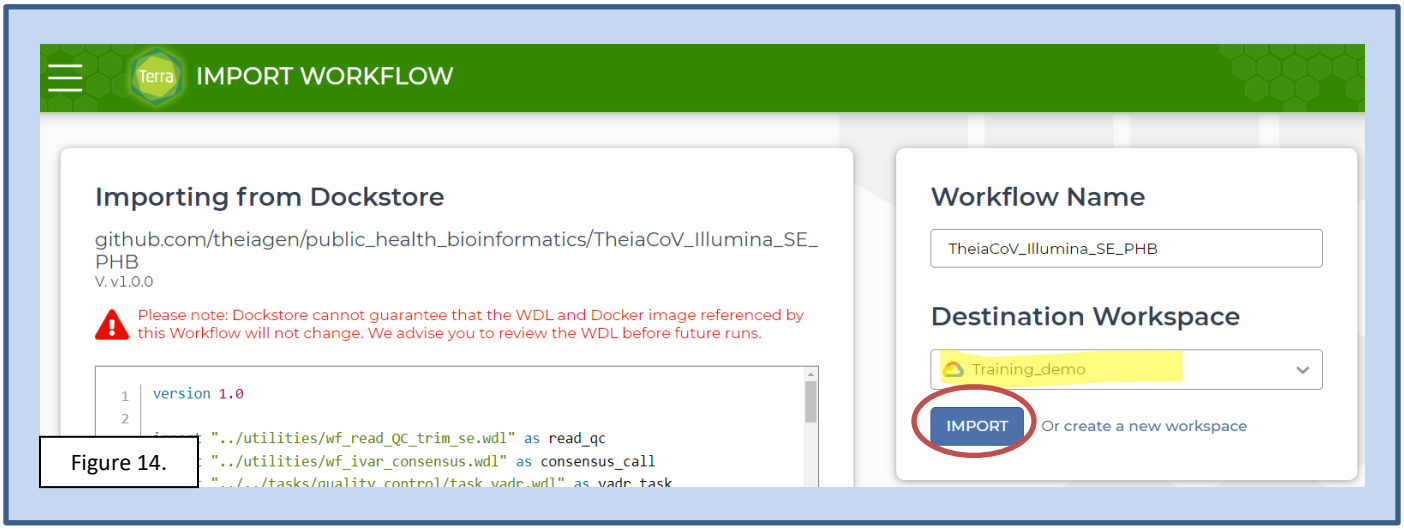


Figure 14.

## 10.2 ADDING WORKSPACE DATA ELEMENTS

1. Navigate to the *Terra workspace* where analysis will be run
2. To upload local files, open the *Files* tab in the right-side panel of the workspace (Fig 15, box)
   a. Click *upload* (Fig 16)
   b. Once the upload is complete, *right click* on the file name and click *copy link*

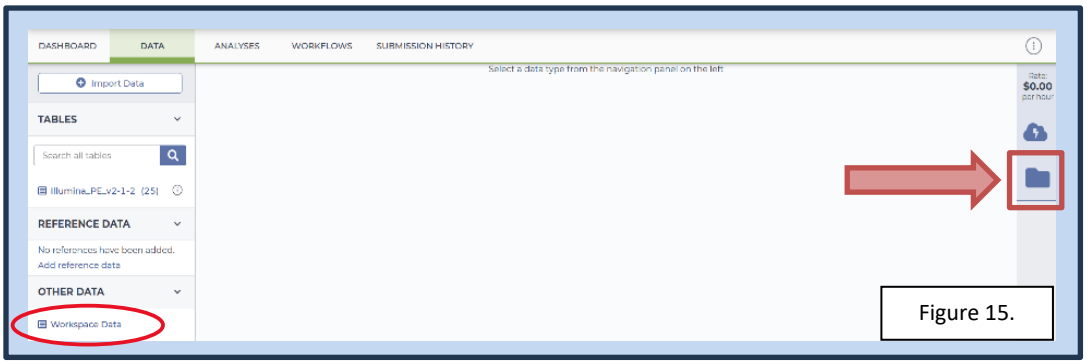


Figure 15.

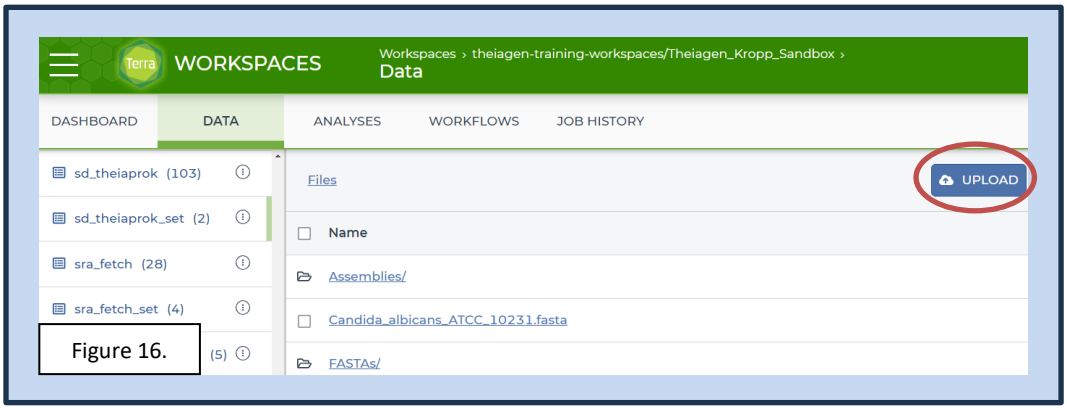


Figure 16.

3. Open the *workspace data* tab (Fig 15, circle)
4. Click *Edit* and *Add variable* in the top tool bar (Fig 17)

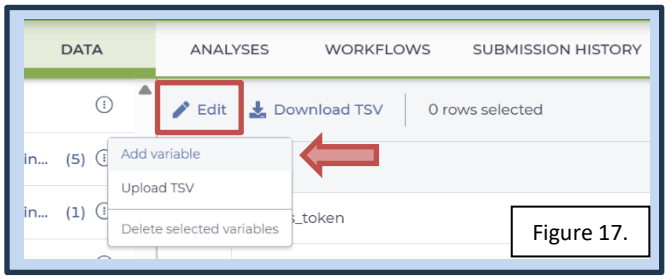


Figure 17.

5. Click in the *key field* and *name the element* being added (Fig 18)
   a. *E.g. to add the Artic v4-1 primer bed file, the key* Artic_v4-1_primer_bed *may be used*
6. In the value field, choose *string* as the value type
   a. *Paste the file path*; the value should start with *gs://*
   b. ***NOTE**: For other string elements like dataset tags and docker images* paste the ID value
      i. *E.g. for the nextclade docker image, add* nextstrain/nextclade:2.14.0
      ii. *Always ensure the docker images and dataset tags are aligned with versions used for internal validation procedures*
7. *Optional: A description may be added to denote the date updated with staff initials*
8. Click the blue check mark on the right-hand side of the variable to save it
   a. The variable will now be available as a workflow input which can be found by typing the prefix *workspace.* plus the key name *artic_v4-1_primer_bed*
      i. e.g. *workspace.artic_v4-1_primer_bed*

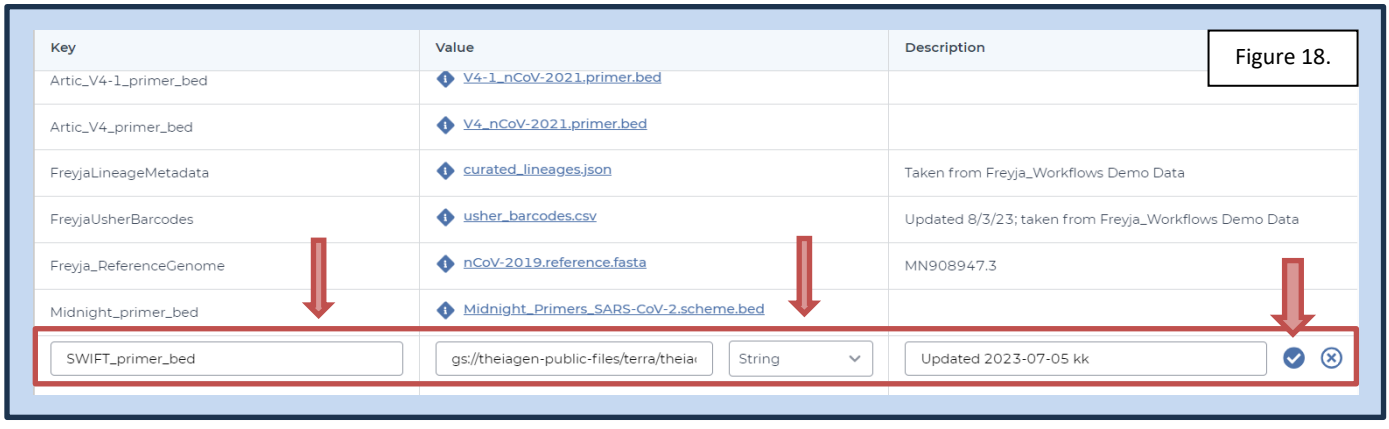


Figure 18.