Analyzing Viral Data in Terra using Theiagen's TheiaCoV ONT Workflow	
Document TG-SC2-ONT, Version 3	
Date:	Workflow Versions
5/8/2024	PHB 2

1. PURPOSE/SCOPE

To standardize the process of analyzing viral next generation sequence (NGS) data using Theiagen's TheiaCoV_ONT_PHB workflow in Terra to generate assemblies, assess quality control (QC) metrics, and determine Nextclade clade and Pangolin lineage assignments, when relevant. Acceptable data types include ONT raw read file format.

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
- ONT raw sequencing read files uploaded to Terra workspace, see TG-TER-03
- Theaigen's TheiaCoV_ONT_PHB workflow in Terra, see appendix 10.1

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

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_ONT_PHB* workflow (Fig 1)
3. *Uncheck call caching* (Fig 2)
4. Choose the latest version of *version 2*, 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)

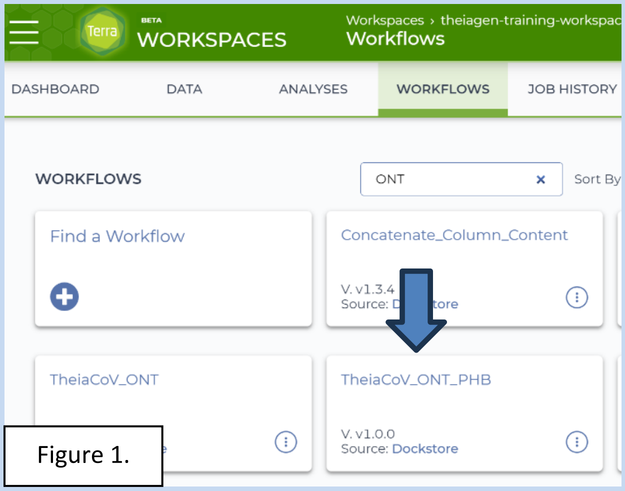
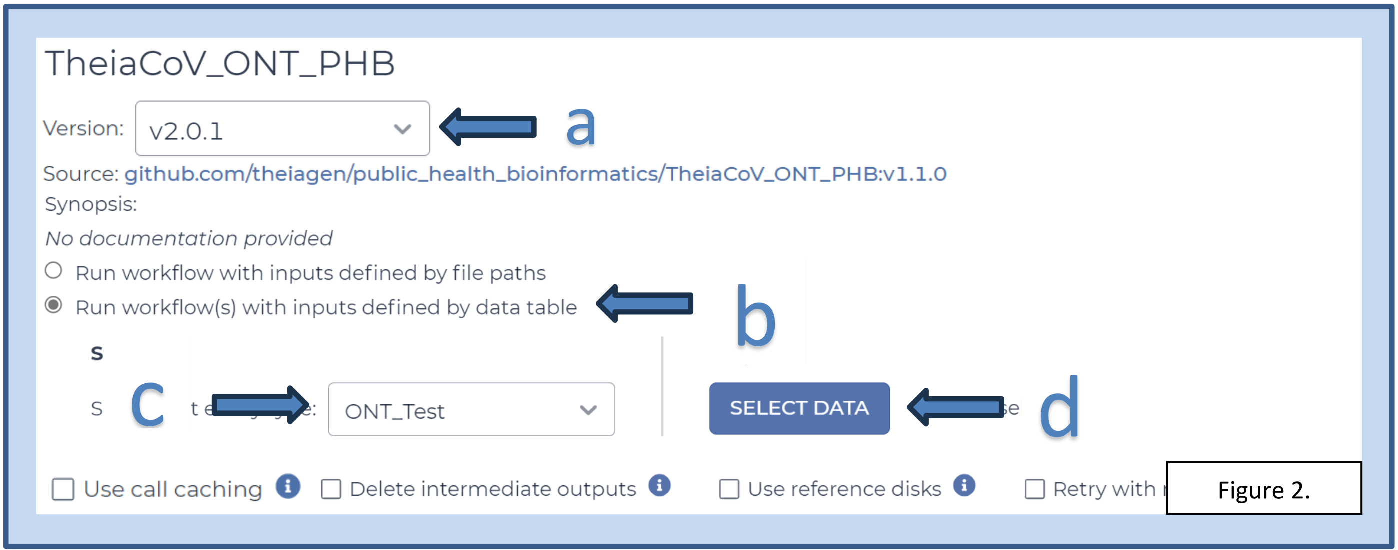


Figure 1.

6. Select the relevant data table under the select *root entity type* dropdown (Fig 2, c)
7. Click *select data* (Fig 4, d)



TheiaCoV_ONT_PHB

Version: v2.0.1

Source: github.com/theiagen/public_health_bioinformatics/TheiaCoV_ONT_PHB:v1.1.0

Synopsis:

No documentation provided

○ Run workflow with inputs defined by file paths

● Run workflow(s) with inputs defined by data table

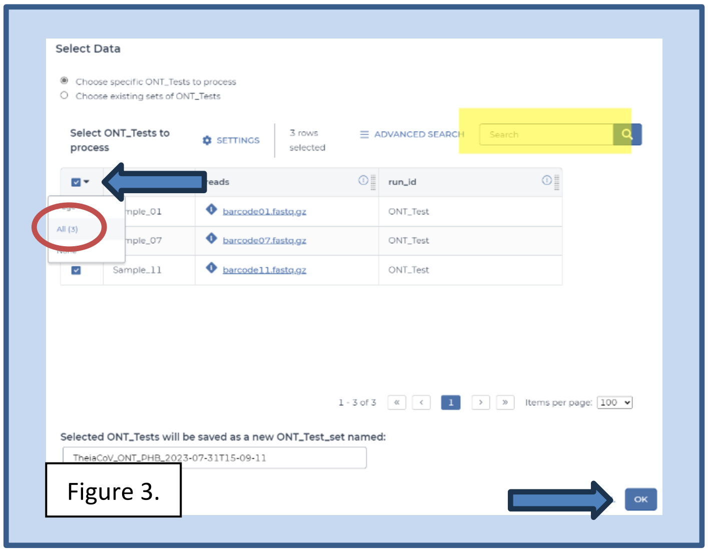
ONT_Test

SELECT DATA

☐ Use call caching ☐ Delete intermediate outputs ☐ Use reference disks ☐ Retry with

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*



Select Data

● Choose specific ONT_Tests to process

○ Choose existing sets of ONT_Tests

Select ONT_Tests to process

3 rows selected

		reads	run_id
☑	mple_01	barcode01.fastq.gz	ONT_Test
	mple_07	barcode07.fastq.gz	ONT_Test
☑	Sample_11	barcode11.fastq.gz	ONT_Test

1 - 3 of 3 Items per page: 100

Selected ONT_Tests will be saved as a new ONT_Test_set named:

TheiaCoV_ONT_PHB_2023-07-31T15-09-11

Figure 3.

9. To run TheiaCoV_ONT_PHB v2 for the first time or use 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
    a ***NOTE**: TheiaCoV PHB v2 workflows are not backwards compatible with older versions of Nextclade; use Nextclade Dataset Tag 2024-04-15—15-08-22Z or newer*
    b 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 ONT sequencing, TheiaCoV_ONT_PHB_2024-05-02.json, or newer

c *Right click* and *save* the file (text does not have to be selected to save properly)

d Return to the workflow in Terra, click *upload json* (Fig 4, red circle), *select* the saved json file, and click *open*

e *To run the workflow with previously saved dataset tags and docker images, no changes are needed*

10. To run TheiaCoV_ONT_PHB v2 for the first time or use 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

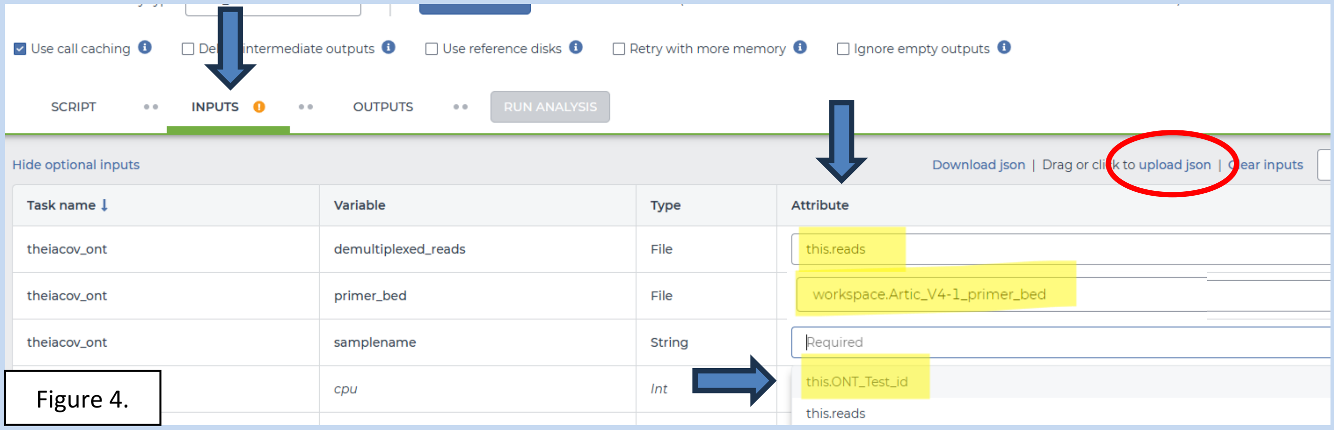
    a ***NOTE**: TheiaCoV PHB v2 workflows are not backwards compatible with older versions of Nextclade; use Nextclade Dataset Tag 2024-04-15—15-08-22Z or newer*

    b 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 ClearLabs platform, TheiaCoV_ClearLabs_PHB_2024-05-02.json, or newer

    c *Right click* and *save* the file (text does not have to be selected to save properly)

    d Return to the workflow in Terra, click *upload json* (Fig 4, red circle), *select* the saved json file, and click *open*

    e *To run the workflow with previously saved dataset tags and docker images, no changes are needed*



Task name	Variable	Type	Attribute
theiacov_ont	demultiplexed_reads	File	this.reads
theiacov_ont	primer_bed	File	workspace.Artic_V4-1_primer_bed
theiacov_ont	samplename	String	Required
	cpu	Int	

Figure 4.

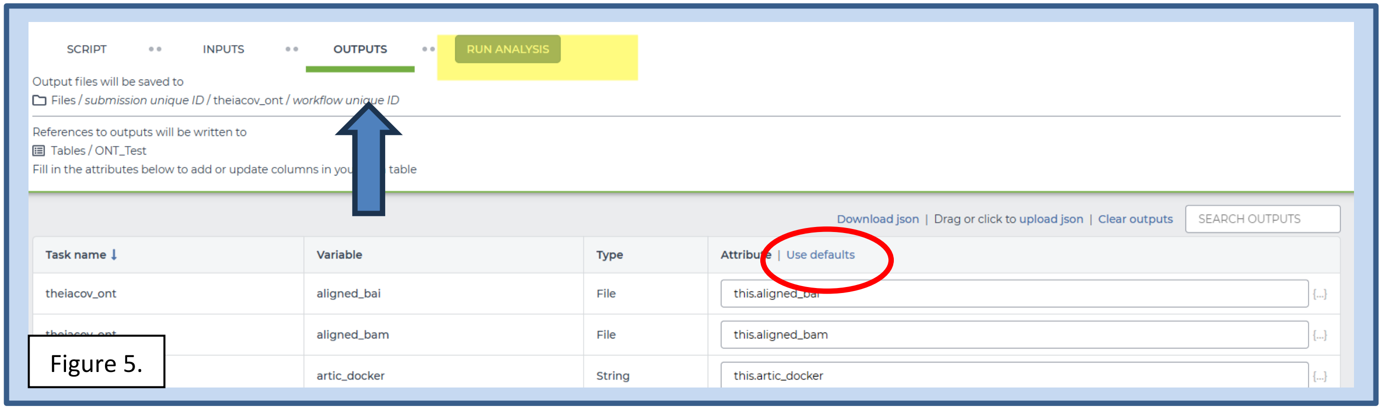
11. Set the first and third attributes in the table to *this.reads* and *this.ONT_Test_id* , respectively (Fig 4) where:

    a *this.ONT_Test_id* is the unique name of your data table in Terra

12. Manually choose the *primer_bed* file for the primer set used to sequence samples

   a 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
   b *To add workspace files for availability in input dropdowns, refer to appendix 10.2*

13. Specify outputs by clicking on the *outputs* tab and *use defaults* (Fig 5)
14. Click *save*
15. Launch the workflow by clicking *run analysis* (Fig 5); enter desired comments and click *launch*



Output files will be saved to
Files / *submission unique ID* / theiacov_ont / *workflow unique ID*

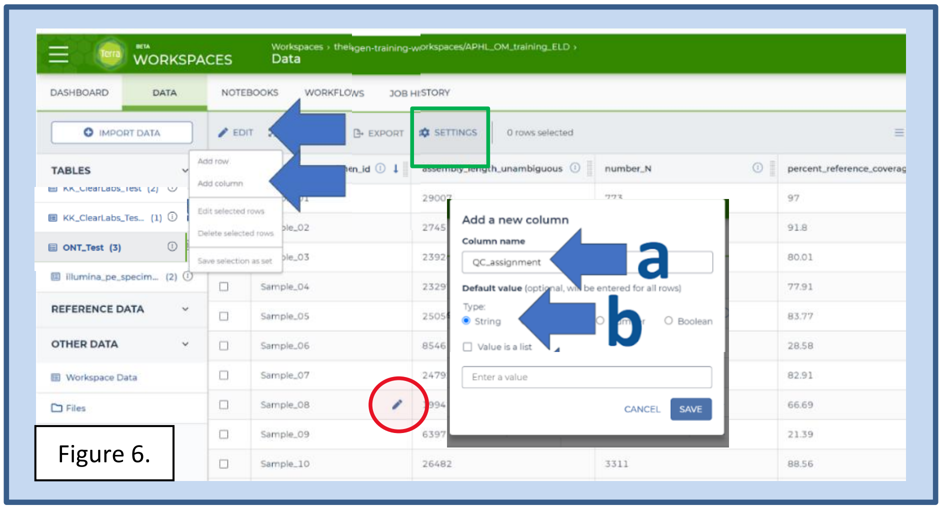
References to outputs will be written to
Tables / ONT_Test
Fill in the attributes below to add or update columns in you table

Task name ↓	Variable	Type	Attribute | Use defaults
theiacov_ont	aligned_bai	File	this.aligned_bai
theiacov_ont	aligned_bam	File	this.aligned_bam
	artic_docker	String	this.artic_docker

Figure 5.

4.2 QUALITY ASSESSMENT OF THEIACOV OUTPUTS

1. Navigate to the *data* tab of the workspace containing TheiaCoV data and open the pertinent data table
2. Click *settings* (Fig 6, green rectangle) and select *none* to deselect all output columns (Fig 7, yellow highlight)

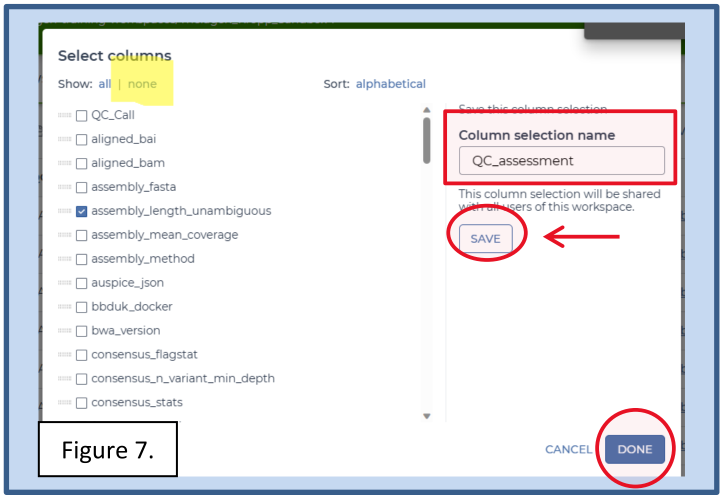


	assembly_length_unambiguous	number_N	percent_reference_coverage
	29007	773	97
ple_02	2745		91.8
ple_03	2392		80.01
Sample_04	2329		77.91
Sample_05	2505		83.77
Sample_06	8546		28.58
Sample_07	2479		82.91
Sample_08	994		66.69
Sample_09	6397		21.39
Sample_10	26482	3311	88.56

Add a new column
Column name: QC_assignment
Default value (optional, will be entered for all rows)
Type: ● String ○ Number ○ Boolean
☐ Value is a list
Enter a value
CANCEL SAVE

Figure 6.

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



☐ QC_Call
☐ aligned_bai
☐ aligned_bam
☐ assembly_fasta
☑ assembly_length_unambiguous
☐ assembly_mean_coverage
☐ assembly_method
☐ auspice_json
☐ bbduk_docker
☐ bwa_version
☐ consensus_flagstat
☐ consensus_n_variant_min_depth
☐ consensus_stats

Column selection name: QC_assessment

This column selection will be shared with all users of this workspace.

Figure 7.

a. <u>Optional</u>: 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 7, red rectangle)*

b. Click *done*

4. *<u>Optional</u>: add a column to record QC PASS/FAIL by clicking* edit, add a column *(Fig 6)*
   - *a. Name the new column (e.g. QC_Call); do not include any spaces*
   - *b. Set the value type as* string
   - c. *Click* save

Table 1. Guidance Thresholds for SC2 Genome Assembly QC

QC Metric	Guidance Threshold*[1]
Number N	<5kbp
Assembly length unambiguous	>24kbp
Percent reference coverage	>83%

5. Use table 1 to assess the quality of each sample's genome assembly (see next page) &/or lab-specific quality metrics

6. *<u>Optional</u>: notate in the QC_assessment field for each sample* PASS *or* FAIL *by* clicking the pencil icon *in the corresponding field (Fig 6, 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

---

[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 QC, 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_PE_PHB workflow is also compatible with the following pathogens: monkeypox virus (MPXV), human immunodeficiency virus (HIV), west nile virus (WNV), influenza virus, and respiratory syncytial viruses A and B (RSV). Refer to Theiagen Public Health Resources Notion 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

9. REVISION HISTORY

Revision	Version	Release Date
Document creation	1	7/2023
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	5/2024

10. APPENDICES

10.1 IMPORTING THE THEIACOV_ONT_PHB WORKFLOW FROM DOCKSTORE

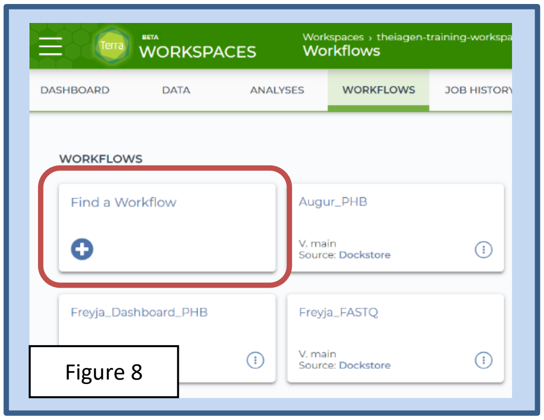


Figure 8

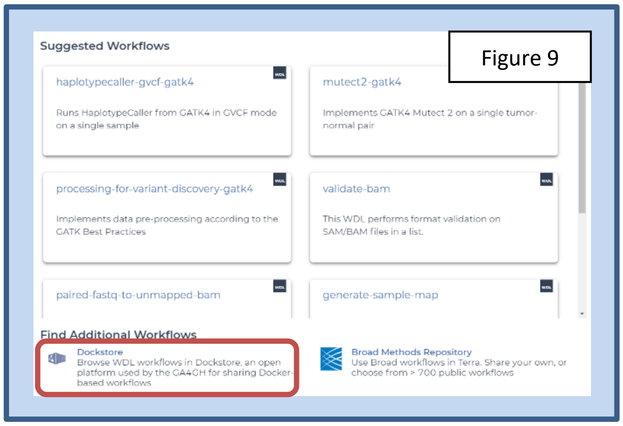


Figure 9

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)

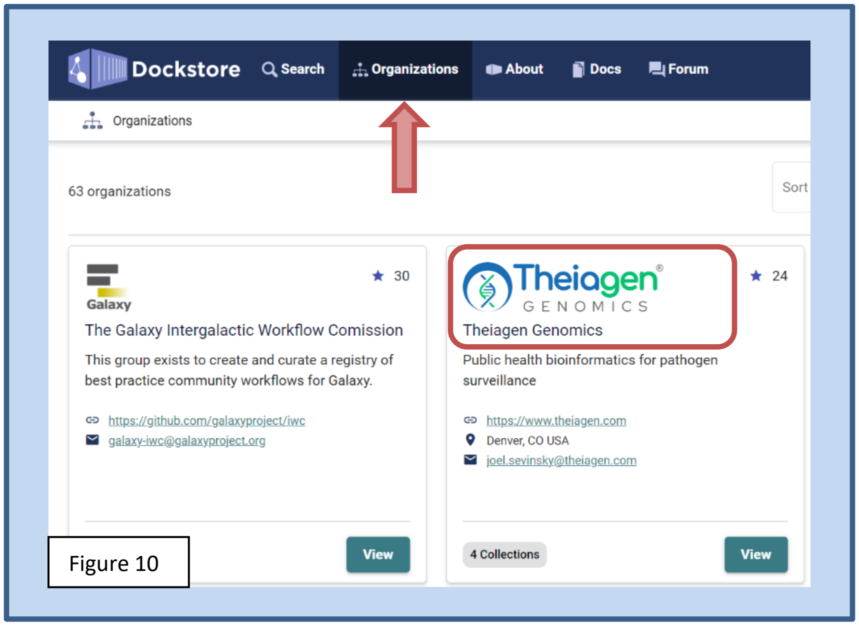
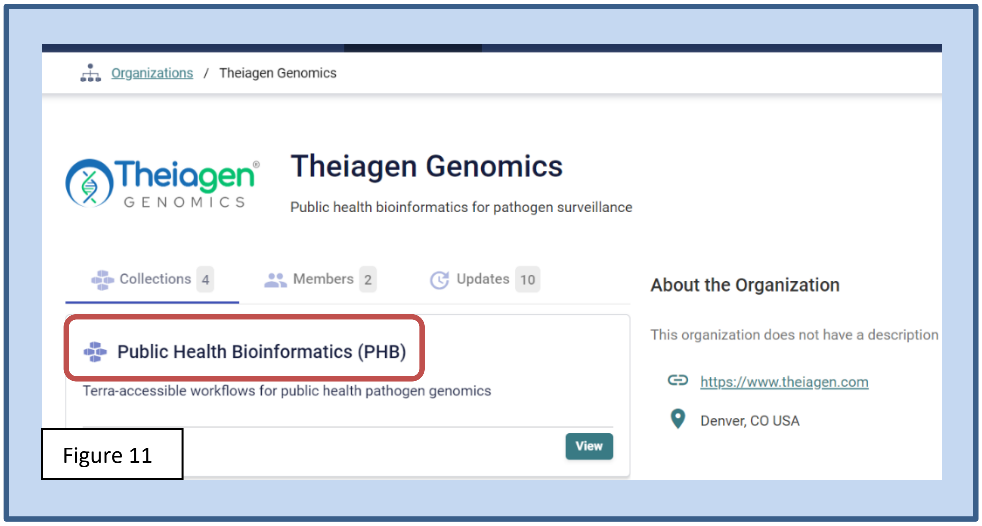


Figure 10



Theiagen Genomics

Public health bioinformatics for pathogen surveillance

Public Health Bioinformatics (PHB)

Terra-accessible workflows for public health pathogen genomics

About the Organization

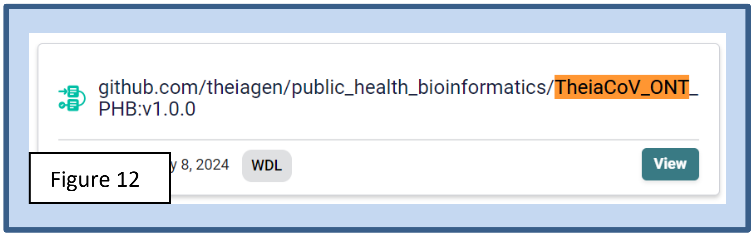
This organization does not have a description

https://www.theiagen.com

Denver, CO USA

Figure 11

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



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

Figure 12

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

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

Tag created: 330 days ago

Last update to source repository: 2 hours ago

Workflow Information

Source Code:...

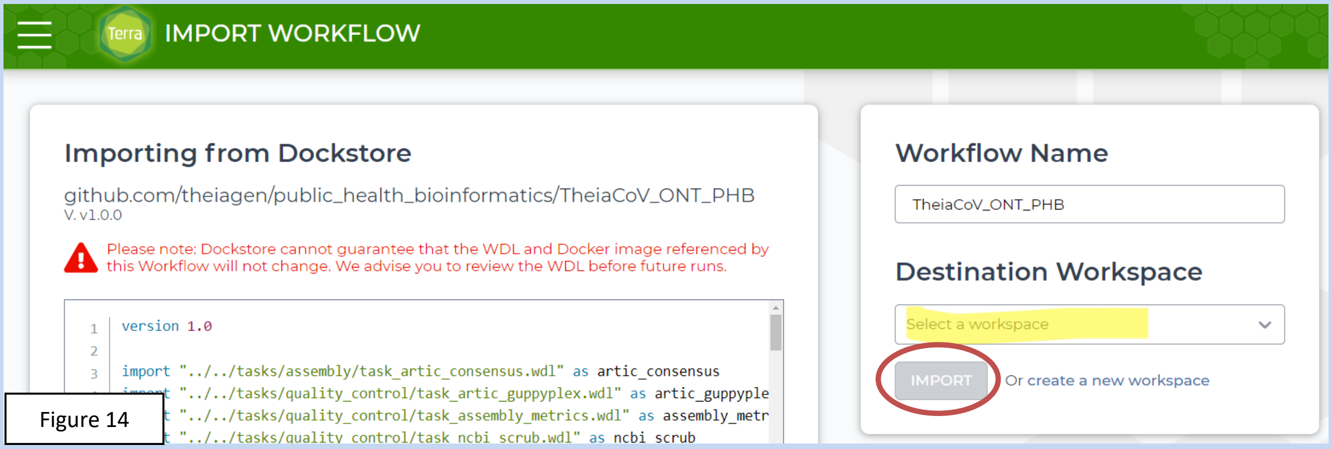
TRS: #workflow/github.com/theiagen/public_health_bioinformatics/TheiaCoV_ONT_PHB

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

ow: n/a

Figure 13

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



Importing from Dockstore

github.com/theiagen/public_health_bioinformatics/TheiaCoV_ONT_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_artic_guppyplex.wdl" as artic_guppyple
t "../../tasks/quality_control/task_assembly_metrics.wdl" as assembly_metr
t "../../tasks/quality_control/task_ncbi_scrub.wdl" as ncbi_scrub
```

Workflow Name

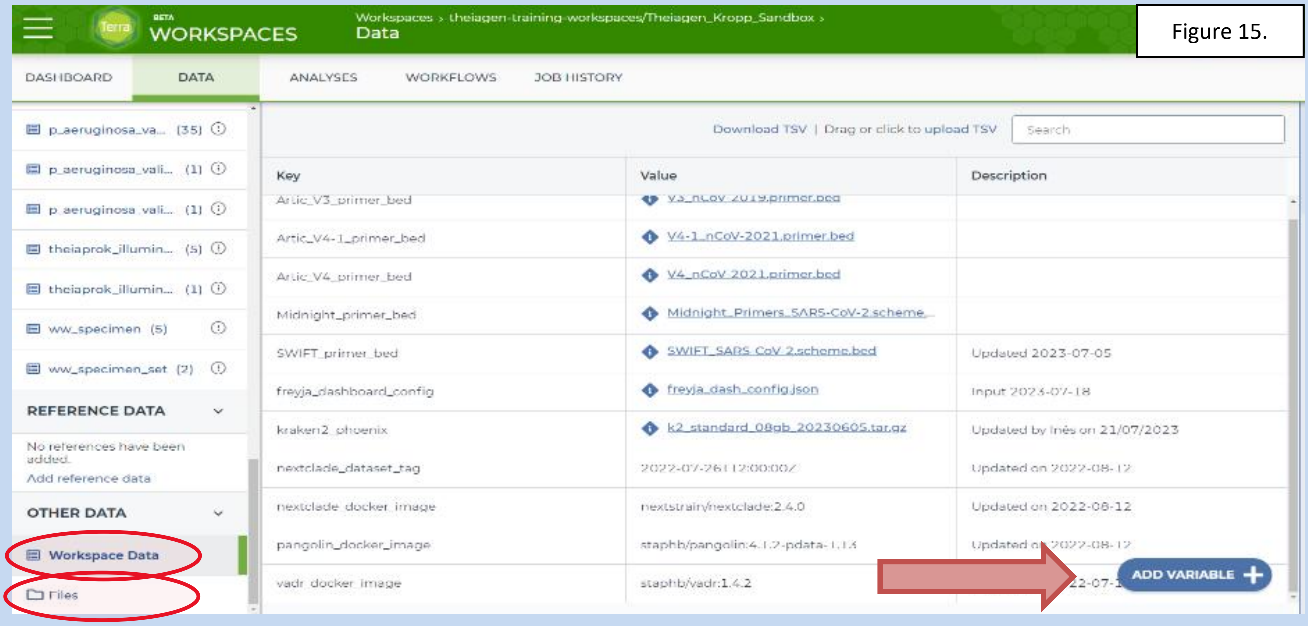
TheiaCoV_ONT_PHB

Destination Workspace

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 bottom left of the workspace (Fig 15)
    a. Click *upload* (Fig 16)
    b. Once the upload is complete, *right click* on the file name and click *copy link*



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	
Midnight_primer_bed	Midnight_Primers_SARS-CoV-2.scheme...	
SWIFT_primer_bed	SWIFT_SARS_CoV_2.scheme.bed	Updated 2023-07-05
freyja_dashboard_config	freyja_dash_config.json	Input 2023-07-18
kraken2_phoenix	k2_standard_08gb_20230605.tar.gz	Updated by Inès on 21/07/2023
nextclade_dataset_tag	2022-07-26T12:00:00Z	Updated on 2022-08-12
nextclade_docker_image	nextstrain/nextclade:2.4.0	Updated on 2022-08-12
pangolin_docker_image	staphb/pangolin:4.1.2-pdata-1.13	Updated on 2022-08-12
vadr_docker_image	staphb/vadr:1.4.2	22-07-1

Figure 15.

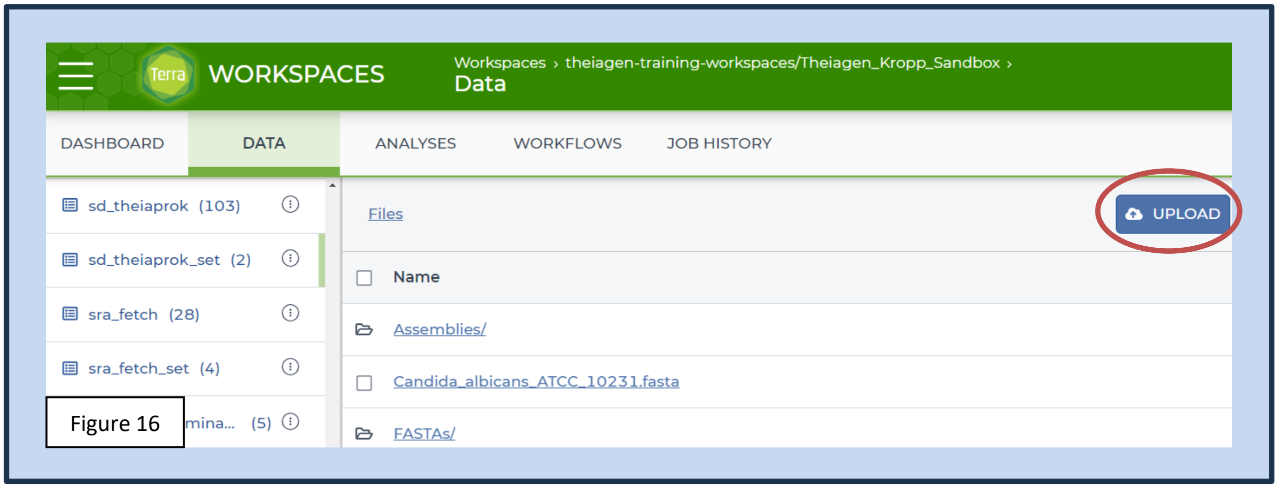


Figure 16

3. Open the *workspace data* tab (Fig 17) and click the *blue plus symbol* in the bottom right (Fig 17)
4. Click in the *key field* and *name the element* being added
   a. *E.g. to add the Artic v4-1 primer bed file, the key* Artic_v4-1_primer_bed *may be used*
5. 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*
6. *<u>Optional</u>: A description may be added to denote the date updated with staff initials*
7. 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*

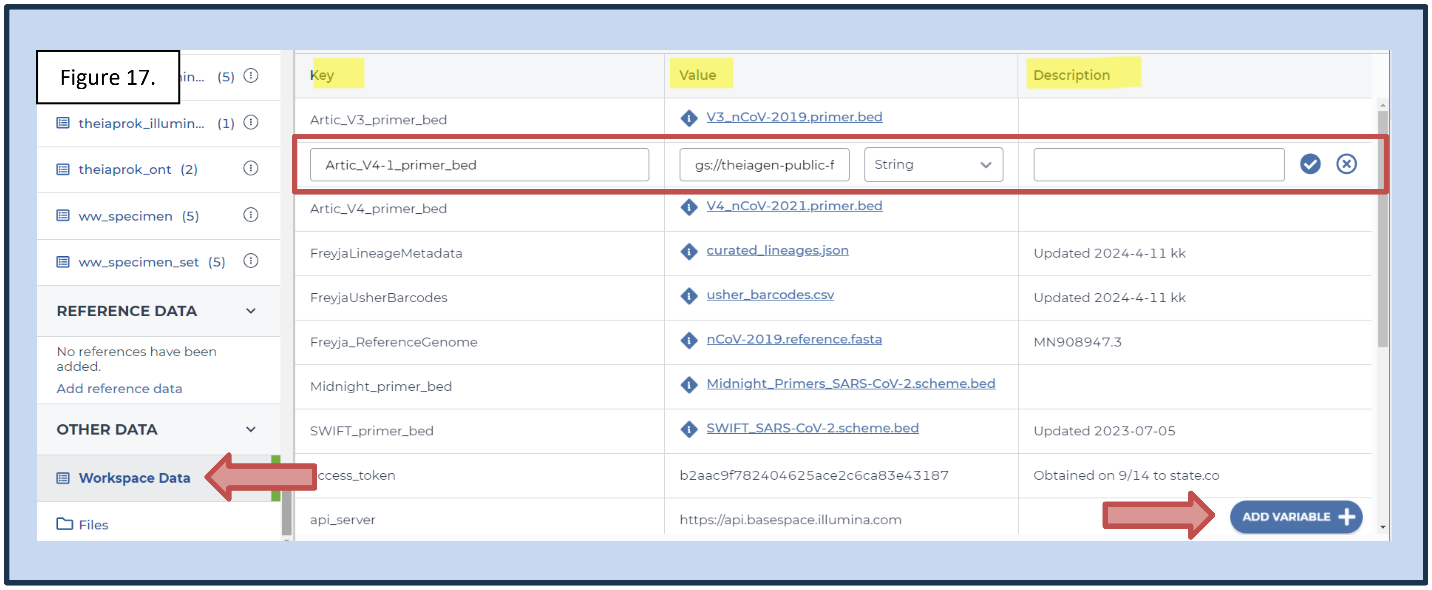


Figure 17.

Key	Value	Description
Artic_V3_primer_bed	V3_nCoV-2019.primer.bed	
Artic_V4-1_primer_bed	gs://theiagen-public-f	
Artic_V4_primer_bed	V4_nCoV-2021.primer.bed	
FreyjaLineageMetadata	curated_lineages.json	Updated 2024-4-11 kk
FreyjaUsherBarcodes	usher_barcodes.csv	Updated 2024-4-11 kk
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
ccess_token	b2aac9f782404625ace2c6ca83e43187	Obtained on 9/14 to state.co
api_server	https://api.basespace.illumina.com	