

Document TG-SC2-SE, Version 4

Date: 5/1/2025 Workflow Versions 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 <u>here</u>.

## 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
TC TER 02	Uploading Local or SRA NGS Data & Creating a
<u>TG-TER-03</u>	Results Metadata Table in Terra
TC TER 04	Linking BaseSpace and Importing BaseSpace
<u>TG-TER-04</u>	Reads to Terra Workspace



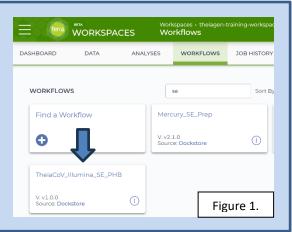
Document TG-SC2-SE, Version 4

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

## 4. PROCEDURE

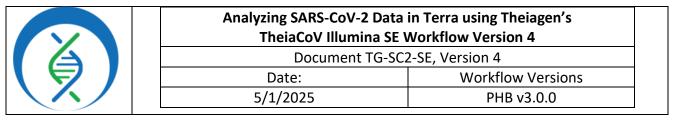
# 4.1 RUNNING THE THEIACOV WORKFLOW

- 1. Open Terra and navigate to the workflows tab within the workspace containing SC2 data
- Select the <u>TheiaCoV\_Illumina\_SE\_PHB</u> 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)
- Select the relevant data table under the select data table dropdown (Fig 2, c)
- 7. Click *select data* (Fig 2, d)



L	TheiaCoV_IIIumina_SE_PHB
	Version: v1.1.0 • 🗲 👌
L	Source: github.com/theiagen/public_health_bioinformatics/TheiaCoV_Illumina_SE_PHB:v1.1.0
L	Synopsis:
L	No documentation provided
L	O Run workflow with inputs defined by file paths
L	Run workflow(s) with inputs defined by data table
	Stel 2
	Sele C = TheiaCoV_Illumina V SELECT DATA G d
	Use call caching 1 Delete intermediate outputs 1 Use reference disks 1 Retry with m 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



	ose specific TheiaCoV_IIIumina_SEs to ose existing sets of TheiaCoV_IIIumina_				
Selec to pro	t TheiaCoV_Illumina_SEs	S rows AD selected	VANCED	SEARCH	Q
•	SE_id 1	nextclade_clade		pango_lineage	
age	00301353	20B		Unassigned	
ll (5) one	00301756	21K (Omicron)		BA.1.1	
	3000301757	21K (Omicron)		BA.1.20	
~	3000301758	21K (Omicron)		Unassigned	
~	3000301759	21K (Omicron)		BA.1.1	
	d TheiaCoV_Illumina_SEs will be		ina_SE	1 > » Items per page: 100 _set named:	~
Theia	CoV_IIIumina_SE_PHB_2023-07-31T	13-50-06			

- 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, <u>upload the TheiaCov input json file</u> on the inputs tab by navigating to the Key Resources Notion page titled <u>Docker Image and Reference Materials for SARS-CoV-2</u> <u>Genomic Characterization</u>
    - i. **NOTE**: TheiaCoV PHB v2.0.0+ workflows are not backwards compatible with older versions of Nextclade; <u>use Nextclade Dataset Tag</u> 2024-04-15—15-08-22Z or <u>newer</u>
    - ii. Expand the <u>TheiaCoV in PHB (v2.0.0 or higher</u>) section, followed by the <u>Terra.Bio Input</u>
       <u>JSONs for PHB v2.0.0 or higher</u>; click on the json file associated with the Illumina PE sequencing platform, <u>TheiaCoV Illumina SE PHB 2025-04-02.json</u> 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

		-	Data in Terra using T a SE Workflow Versi	-		
		Document T	G-SC2-SE, Version 4			
	Da	ite:	Workfl	ow Versions		
	5/1/	2025	PH	PHB v3.0.0		
SCRIPT •• INPUTS		NALYSIS	Download jsol og or click to upload jsor	CANCEL SAVE		
SCRIPT •• INPUTS Hide optional inputs Task name ↓		NALYSIS		Sear inputs SEARCH INPUTS		
SCRIPT •• INPUTS	• •• OUTPUTS •• RUN A	NALYSIS	Download jsol og or click to upload jsor			
SCRIPT •• INPUTS Hide optional inputs Task name ↓	0 •• OUTPUTS •• RUN A	Type	Download json ag or click to upload json	Sear inputs SEARCH INPUTS		
SCRIPT •• INPUTS Hide optional inputs Task name ↓ theiacov_Illumina_se	OUTPUTS •• RUN A  Variable read1_raw	NALYSIS Type File	Download json ag or click to upload json Attribute this.read1 this. this.TheiaCoV_Illumina_SE_id	Gear inputs SEARCH INPUTS		
SCRIPT •• INPUTS Hide optional inputs Task name ↓ theiacov_illumina_se theiacov_illumina_se	OUTPUTS ** RUN A      Variable     read1_raw     samplename	NALYSIS Type File String	Download json ag or click to upload json Attribute this.read1 this.	Gear inputs SEARCH INPUTS		

- 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
  - Labs using the Artic V4-1 will choose workspace.Artic\_V4-1\_primer\_bed; for other primer bed files, see <u>Docker Image and Reference Materials for SARS-CoV-2 Genomic Characterization</u> 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

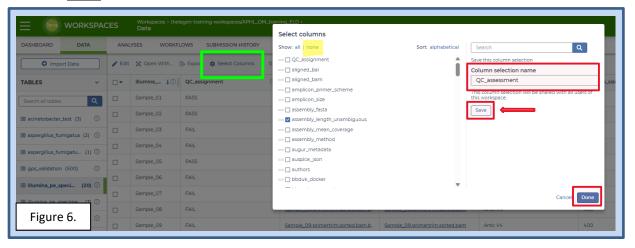
SCRIPT •• INPU	JTS •• OUTPUTS •• RUN ANA		
Output files will be saved to The Files / submission unique ID / the	eiacov_illumina_se / y v unique ID		
References to outputs will be writter Tables / TheiaCoV_IIIumina_SE			
Fill in the attributes below to add or	update columns in yo	Download json   Drag or (	click to upload json   Clear outputs SEARCH OUTPUTS
Fill in the attributes below to add or Task name ↓	update columns in yo	Download json   Drag or o	Click to upload json   Clear outputs SEARCH OUTPUTS
	· · · · ·		



Document TG-SC2-SE, Version 4			
Date: Workflow Versions			
5/1/2025	PHB v3.0.0		

## 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 <u>Select Columns</u> (Fig 6, green rectangle) and select <u>none</u> to deselect all output columns (Fig 6, yellow highlight)
- 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. <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 6, red rectangle)
  - b. Click done



- 4. <u>Optional</u>: 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

DASHBOARD DAT	ГА	NOTE		FLOWS JOB H	ISTORY			
		/ ED	ur .	- EXPORT	¢ SET	TINGS 0 rows selected		
TABLES	~	Add row. Add column		en_id ③ ↓	assem	oly_length_unambiguous 🛈	number_N	percent_reference_cove
- and a second second			1	_	29007		44.5	97
Search all tables		Edit selected	1- 02		2745	Add a new column		91.8
∃ TheiaCoV_Illumi (5)	0 0	Save selection			2392	Column name		80.01
🗐 illumina_pe_specim (			Sample_04		2329	QC_assignment Default value (optignal, while		77.91
REFERENCE DATA						Type:	entered for all rows)	
			Sample_05		2505	String	O am r O Boolean	83.77
OTHER DATA	~		Sample_06		8546	Value is a list		28.58
Workspace Data			Sample_07	$\sim$	2479	Enter a value		82.91
C Files			Sample_08	1	1994		CANCEL SAVE	66.69
	1	0	Sample_09		6397			21.39
Figure 7.			Sample_10		26482		3311	88.56



Document TG-SC2-SE, Version 4		
Date:	Workflow Versions	
5/1/2025	PHB v3.0.0	

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

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

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

<sup>&</sup>lt;sup>1</sup> 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



Document TG-SC2-SE, Version 4

# 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. <u>Optional</u>: 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 <u>nextclade\_clade</u>; result formats will use the following nomenclature: <u>21L (Omicron)</u> where:
    - i. 211 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: <u>pipeline-resources/docs/sc2-recombinants.md at main · pha4qe/pipeline-resources · GitHub</u>
- 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 <u>https://cov-lineages.org/lineage\_list.html</u>
- 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



Document TG-SC2-SE, Version 4

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

### 6. TROUBLESHOOTING

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

### 7. LIMITATIONS

- 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 <u>Theiagen Public Health Resources documentation</u> 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—15-08-222 or newer

### 8. REFERENCES

- Smith, E., Wright, S., & Libuit, K. (2022, June 28). *Identifying SARS-CoV-2 Recombinants*. Github. Retrieved June 16, 2023, from <u>https://github.com/pha4ge/pipeline-</u> <u>resources/blob/main/docs/sc2-recombinants.md#identifying-sars-cov-2-recombinants</u>
- 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
- 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). <u>https://doi.org/10.1099/mgen.0.001051</u>.
- 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

Analyzing SARS-CoV-2 Data TheiaCoV Illumina SE V	0 0		
Document TG-SC2-SE, Version 4			
Date:	Workflow Versions		
5/1/2025	PHB v3.0.0		

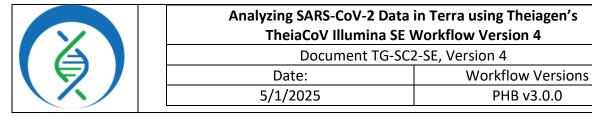
#### **10. APPENDICES**

### 10.1 IMPORTING THE THEIACOV\_ILLUMINA\_SE\_PHB WORKFLOW FROM DOCKSTORE

HBOARD DATA A	NALYSES WORKFLOWS	JOB HISTOR	Find a workflow Fig	gure 9
WORKFLOWS			Dockstore.org @         Terra Workflow Repository           A community repository of best practice workflows that offers integration with CitHub.         A repository of WDL workflows hosted in the platform.	ers private
Find a Workflow	Augur_PHB			
0	V. main Source: Dockstore	(	Curated collections from our community:           CATK Best Practices Id         Long Read Pipelines Id           WDL Analysis Research Pipelines Id         Viral Genomics Id	
Freyja_Dashboard_PHB	Freyja_FASTQ		Visit our documentation to learn how to import and configure your workflow, as well as how to save tim	1e and mor

- 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)

Organizations			
63 organizations			S
Galaxy The Galaxy Intergalactic Workflow ( This group exists to create and curate a best practice community workflows for (	egistry of Public heal	Cheiogen® GENOMICS Genomics th bioinformatics for pathogen e	* 24



the Organization
anization does not have a description
Denver, CO USA
)

5. To find the TheiaCoV\_Illumina\_SE\_PHB workflow in Windows environments, hold <u>Ctrl + F</u> and <u>search TheiaCoV\_Illumina\_SE</u>, then click on the link (Fig 12)



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

created: 321 o	days ago urce repository: 7 he	ours ago					
Info	Launch	Versions	Files	Tools	DAG	Metrics	Launch with DNAnexus
Source Code: TRS: <u>#workfl</u>	 <u>ow/github.com/thei</u>	agen/public_health_t for genomic characte ern.			-	ology of	Terra

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

	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	

Importing from Do	ockstore	Workflow Name
github.com/theiagen/pub PHB <sup>V. v1.0.0</sup>	lic_health_bioinformatics/TheiaCoV_Illumina_SE_	TheiaCoV_Illumina_SE_PHB
	not guarantee that the WDL and Docker image referenced by ge. We advise you to review the WDL before future runs.	Destination Workspace
1 version 1.0	A	Training_demo

### **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*

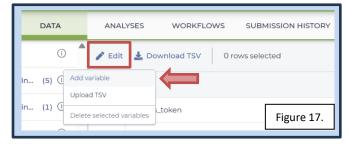


	WORKSPA	CES Workspaces > theiagen-training-workspaces/Theiagen_Kropp_Sandbox > Data
DASHBOARD	DATA	ANALYSES WORKFLOWS JOB HISTORY
🗏 sd_theiaprok	(103) 🔅	Files DPLO
sd_theiaprok_	set (2) 🔅	□ Name
🗏 sra_fetch (28)	(i)	Assemblies/
sra_fetch_set	(4) 🔅	Candida_albicans_ATCC_10231.fasta
Figure 16.	(5) 🕕	▷ FASTAS/



Document TG-SC2-SE, Version 4		
Date:	Workflow Versions	
5/1/2025	PHB v3.0.0	

- 3. Open the workspace data tab (Fig 15, circle)
- 4. Click *Edit* and *Add variable* in the top tool bar (Fig 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. <u>Optional</u>: 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

Key	Value	Description Figure 18.		
Artic_V4-1_primer_bed	V4-1_nCoV-2021.primer.bed			
Artic_V4_primer_bed	V4_nCoV-2021.primer.bed			
FreyjaLineageMetadata	<u>curated_lineages.json</u>	Taken from Freyja_Workflows Demo Data		
FreyjaUsherBarcodes	usher_barcodes.csv	Updated 8/3/23; taken from Freyja_Workflows Demo Data		
Freyja_ReferenceGenome	1 nCoV-2019.reference.fasta	MN908947.3		
Midnight_primer_bed	Midnight_Primers_SARS-CoV-2.scheme.bed			
SWIFT_primer_bed	gs://theiagen-public-files/terra/theia	Updated 2023-07-05 kk 🖉 😒		