N X /

Document TG-SC2-PE, 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_IIIumina_PE_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 paired end (PE) 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

3.

Terra account, linked to Google account

RELATED DOCUMENTS

- Illumina PE raw read files uploaded to Terra workspace, see TG-TER-03 or TG-TER-04
- Theiagen's TheiaCoV_Illumina_PE_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.

Document NumberDocument NameTG-TER-03Getting Started in Terra: Importing Reads,
Metadata, Workflows, and MoreTG-TER-04Linking BaseSpace and Importing BaseSpace
Reads to Terra Workspace



Document TG-SC2-PE, Version 4

Date: 5/1/2025

Workflow Versions: PHB v3.0.0

4. PROCEDURE

4.1 RUNNING THE THEIACOV WORKFLOW

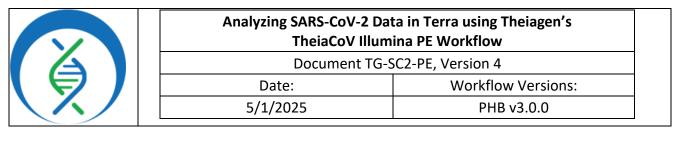
- Open Terra and navigate to the *workflows* tab within the workspace containing SC2 data
- Select the <u>TheiaCoV_Illumina_PE_PHB</u> workflow (Fig 1)
- 3. Uncheck call caching (Fig 2)
- 4. Choose the *latest version* of the workflow, or the version internally validated (Fig 2, a)
- 5. Select the second bullet to *run workflow(s) with inputs defined by data table* (Fig 2, b)

	ES		spaces → the rkflows	iagen-trai	ining-workspa
DASHBOARD DATA	ANALY	SES	WORKFLC	ows	JOB HISTORY
WORKFLOWS			pe		Sort B
Find a Workflow		Merc	ury_PE_Pre	p	
•		V. v2. Sourc	1.0 e: Dockstore		()
TheiaCoV_IIIumina_PE_PHB		Theia	aProk_Illum	ina_PE	
V. v1.0.0 Source: Dockstore	(i)	V. v1. Sourc	3.0 e: Dockstor	Figu	ure 1.

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

TheiaCoV_Illumina_PE_PHB
Version: v1.1.0 Version:
Source: github.com/theiagen/public_health_bioinformatics/TheiaCoV_Illumina_PE_PHB:v1.1.0
Synopsis:
No documentation provided
O Run workflow with inputs defined by file paths
Run workflow(s) with inputs defined by data table
Step 2
SELECT DATA
Use call caching 🚺 🗌 Delete intermediate outputs 🕕 🗌 Use reference disks 🕕 🗌 Retry with more 🛛 Figure 2.

- 8. In the pop-up window *select each sample* checkbox to include in the analysis (Fig 3)
 - a. Click the down arrow and select all to process all specimens (Fig 3)
 - b. Alternatively, 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* (Fig 3)



	O Choo	se specific illumina_s	pe_specimens to process imina_pe_specimens cimens to22	9 rows ≡ Advanced Search	Search
	proce		settings	nextclade_clade	D number_N
	age	mple_01	29007	22C (Omicron)	773
A	11 (20)	mple_02	27451	21L (Omicron)	2324
		Sample_03	23924	21L (Omicron)	5826
		Sample_04	23297	21L (Omicron)	6452
		Sample_05	25050	22C (Omicron)	4697
		Sample_06	8546	21L (Omicron)	20287
		Sample_07			
5	Theia	d illumina_pe_spe CoV_Illumina_PE_20 gure 3.	cimens will be saved as a new illun 25-04-11722-07-02	1 - 20 of 20 《 < 1	

- 9. Specify the desired *dataset tags* and *docker image* inputs
 - a. To run TheiaCoV_Illumina_PE_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—15-08-222 <u>or newer</u>
 - 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*
 - iii. Click on the json file associated with the Illumina PE sequencing platform, <u>TheiaCoV Illumina PE PHB 2025-04-02.json</u>, or newer
 - iv. *Right click* and *save* the file (text does not have to be selected to save properly)
 - v. 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
- 10. Set the first three attributes in the table manually to *this.read1* , *this.read2* , and

this.illumina_pe_specimen_id, respectively (Fig 4)

a. Where *illuming_pe_specimen* is the unique name of your data table in Terra

, .	0 0	
Document TG-	SC2-PE, Version 4	
Date:	Workflow Versions:	
5/1/2025	PHB v3.0.0	
	TheiaCoV Illum Document TG-S Date:	

Use call caching Delet		Retry with more memo NALYSIS	oy 0	CANCEL SA
Hide optional inputs			Download json Dra	ag or click to upload json SEARCH INPUTS
Task name ↓	Variable	Туре	Attribute	
theiacov_illumina_pe	primer_bed	File	workspace.Artic_V4_primer_bed	()
theiacov_illumina_pe	read1_raw	File	this.readl	()
theiacov_illumina_pe	read2_raw	File	this.read2	()
theiacov_illumina_pe	samplename	String	this.	{}
bwa	cpu	Int	thisillumina_pe_specimen_id	
consensus	char_unknown	String	thisread1 thisread2	
consensus	count_orphans	Boolean	this.run_id	
Figure 4.	disable_bag	Boolean	Optional	

- 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. Click run analysis (Fig 5), enter comments and select launch

SCRIPT •• INPUTS •	OUTPUTS ** RUN A	NALYSIS		1
Output files will be saved to	•			
Files / submission unique ID / theiacov_illum	ina_pe / wr w unique ID			
References to outputs will be written to Tables / illumina_pe_specimen Fill in the attributes below to add or update colu	ımns in you ta table	Download json Drag or cl	CANCEL SAV	e
Task name ↓	Variable	Туре	Attribute Use defaults	
Figure 5.	abricate_flu_database	String	Optional [



Document TG-SC2-PE, Version 4

Date:

5/1/2025

Workflow Versions: 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 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, 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

				Select columns		
DASHBOARD DATA	ANAL	YSES WORKF	LOWS SUBMISSION HISTORY	Show: all none	Sort: alphabetical	Search Q
Import Data	🖉 Edit	😒 Open With	🕒 Expor 🎄 Select Columns (QC_assignment	<u>^</u>	Save this column selection
				🔲 aligned_bai	I	Column selection name
ABLES V	□ •	illumina \downarrow 🕕	QC_assignment	🔲 aligned_bam	Ť	QC_assessment
				amplicon_primer_scheme		Inis column selection will be shared with all users of
Search all tables		Sample_01	PASS	: 🗋 amplicon_size		this workspace.
		Sample_02	PASS	: 🗋 assembly_fasta		Save
acinetobacter_test (3)				🔤 assembly_length_unambiguous		
aspergillus_fumigatus (2) (1)		Sample_03	FAIL	: 🗋 assembly_mean_coverage		
		Sample_04	FAIL	🗋 assembly_method		
) aspergillus_fumigatu (1) 🛈				augur_metadata		
gps_validation (500)		Sample_05	PASS	auspice_json		
gps_validation (500)		Sample_06	FAIL	authors		
) illumina_pe_speci (20) 🕕		barnpic_00		bbduk_docker		
		Sample_07	FAIL			Cancel Done
Lillumina pe specime (3) ①		Sample_08	FAIL			
Figure 6. (2) (3)		Sample_08	PAIL		Sumple_00.printer till.softed.butti	400

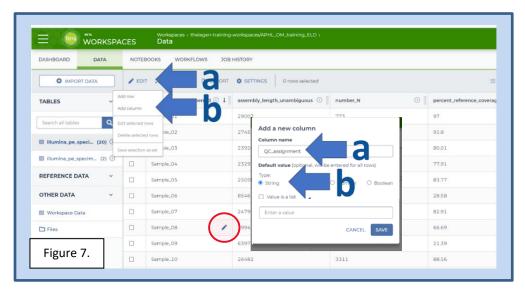
- 4. Optional: Add a column to record QC PASS/FAIL by clicking edit, add a column (Fig 7)
 - a. Name the new column (e.g. QC_Call); do not include any spaces
 - b. Set the value type as string
 - c. Click save



Document TG-SC2-PE, Version 4

Date: 5/1/2025 Workflow Versions:

PHB v3.0.0



- 5. Use table 1 below to assess the quality of each sample's genome assembly &/or lab-specific quality metrics
- 6. <u>Optional</u>: Notate in the QC_assessment field for each sample PASS or FAIL by <u>clicking the pencil</u> 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

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%

Table 1. Guidance thresholds for genome assembly QC

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



Document TG-SC2-PE, Version 4

PHB V3.0.0

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. Click Select Columns above the data table, select none (Fig 6)
- 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
 - b. Click *done*
- 5. 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>
- 6. 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>
- 7. 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



Document TG-SC2-PE, Version 4

Date: 5/1/2025 Workflow Versions:

PHB v3.0.0

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

- This SOP is written for the analysis of SC2 data; v2+ of the TheiaCoV Illumina PE PHB workflow 1. 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 documentation for organismspecific 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

REFERENCES 8.

- Smith, E., Wright, S., & Libuit, K. (2022, June 28). *Identifying SARS-CoV-2 Recombinants*. Github. 1. Retrieved June 16, 2023, from https://github.com/pha4ge/pipelineresources/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
- Libuit, Kevin G., Emma L. Doughty, James R. Otieno, Frank Ambrosio, Curtis J. Kapsak, Emily A. 3. 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**

Version	Release Date
1	7/2023
2	9/2023
3	4/2024
4	5/2025
-	

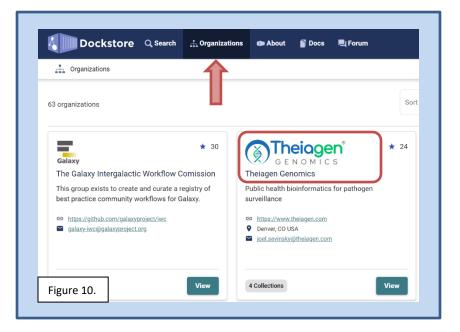
Analyzing SARS-CoV-2 Data in Terra using Theiagen's TheiaCoV Illumina PE Workflow		
Document TG-S	C2-PE, Version 4	
Date:	Workflow Versions:	
5/1/2025	PHB v3.0.0	
	TheiaCoV Illum Document TG-S Date:	TheiaCoV Illumina PE Workflow Document TG-SC2-PE, Version 4 Date: Workflow Versions:

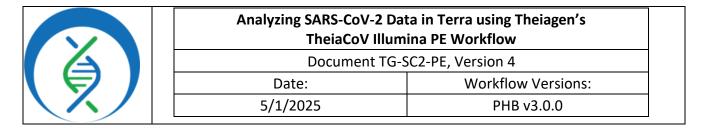
10. APPENDICES

10.1 IMPORTING THE THEIACOV_ILLUMINA_PE_PHB WORKFLOW FROM DOCKSTORE

SHBOARD DATA AI	NALYSES WORKFLOWS	JOB HISTOR)	Find a workflow Figu
WORKFLOWS			Dockstore.org II Terra Workflow Repository A community repository of best practice workflows A repository of WDL workflows that offers p that offers integration with CitHub.
Find a Workflow	Augur_PHB		
0	V. main Source: Dockstore	(1)	Curated collections from our community: GATK Best Practices of Long Read Pipelines of WDL Analysis Research Pipelines of Viral Genomics of
Freyja_Dashboard_PHB	Freyja_FASTQ		Visit our documentation to learn how to import and configure your workflow, as well as how to save time an

- 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)





Crganizations / Theiagen Genomics				
Theiagen Genomics GENOMICS Public health bioinformatics for pathogen surveillar	nce			
Collections 4 Members 2 C Updates 10	About the Organization			
Public Health Bioinformatics (PHB)	This organization does not have a description			
Terra-accessible workflows for public health pathogen genomics	<u>https://www.theiagen.com</u>			
Figure 11.	Denver, CO USA			

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

github.com/theiagen/public_health_bioinformatics/TheiaCoV_Illumi na_PE_PHB:v1.0.0			
Figure 12. 26, 2024 WDL View	github.c	com/theiagen/public_health_bioin PHB:v1.0.0	nformatics/ <mark>TheiaCoV_Illumi</mark>
0	Figure 12.	26, 2024 WDL	View

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

g created: 318 st update to so	days ago urce repository: 25 ı	minutes ago					
Info	Launch	Versions	Files	Tools	DAG	Metrics	Launch with DNAnexus
Source Code TRS: <u>#workfl</u>	 ow/github.com/thei	ag <u>en/public_health_b</u> for genomic characte				ology of	Terra

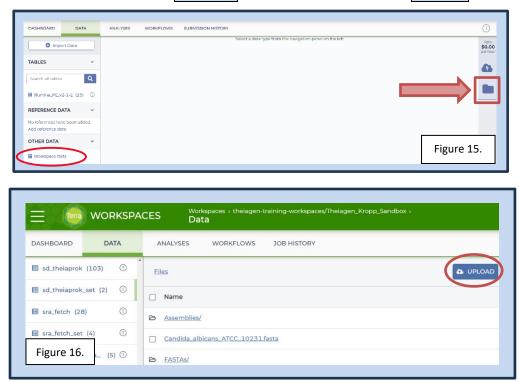
Analyzing SARS-CoV-2 Data in Terra using Theiagen's TheiaCoV Illumina PE Workflow		
Document TG-SC2-PE, Version 4		
Date:	Workflow Versions:	
5/1/2025	PHB v3.0.0	

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

Importing from Dockstore	Workflow Name
github.com/theiagen/public_health_bioinformatics/TheiaCoV_IIIumina_PE _PHB V. V1.0.0	TheiaCoV_Illumina_PE_PHB
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.	Destination Workspace
1 version 1.0 2 immont "/utilities/wf_read QC_trim_pe.wdl" as read gc	Training_demo

10.2 ADDING WORKSPACE DATA ELEMENTS

- 1. Navigate to the Terra workspace where analysis will be run
- 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*

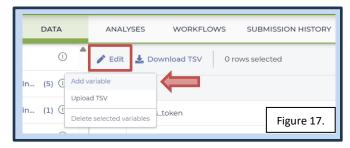




Document TG-SC2-PE, Version 4

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

- 3. Open the *workspace data* tab in the left-side panel (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. a. a. workspace 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	curated_lineages.json	Taken from Freyja_Workflows Demo Data
FreyjaUsherBarcodes	usher_barcodes.csv	Updated 8/3/23; taken from Freyja_Workflows Demo Data
Freyja_ReferenceGenome	nCoV-2019.reference.fasta	MN908947.3
Midnight_primer_bed	<u>Midnight_Primers_SARS-CoV-2.scheme.bed</u>	
SWIFT_primer_bed	gs://theiagen-public-files/terra/theia	Updated 2023-07-05 kk 🖉 😣