

Document TG-SC2-PE, Version 3

Date: 4/25/2024 Workflow Versions: PHB v2

1. PURPOSE/SCOPE

To standardize the process of analyzing SARS-COV-2 (SC2) next generation sequencing (NGS) data using Theiagen's TheiaCoV_Illumina_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.

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. RELATED DOCUMENTS

- Terra account, linked to Google account
- 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-03Uploading Local or SRA NGS Data & Creating a
Results Metadata Table in TerraTG-TER-04Linking BaseSpace and Importing BaseSpace
Reads to Terra Workspace



Document TG-SC2-PE, Version 3

Date: Workflow Versions:

4/25/2024

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 *version 2*, or the version internally validated (Fig 2, a)
- 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 root entity type dropdown (Fig 2, c)
- 7. Click *select data* (Fig 2, d)

	ES		(spaces > theiag rkflows	gen-training	-workspa
ASHBOARD DATA	ANAL	/SES	WORKFLOW	VS JOB	HISTORY
WORKFLOWS			pe		Sort B
Find a Workflow		Merc	ury_PE_Prep		
•		V. v2.: Sourc	1.0 e: Dockstore		:
TheiaCoV_IIIumina_PE_PHB		Theia	aProk_Illumin	a_PE	
V. v1.0.0 Source: Dockstore	()	V. v1.: Sourc	3.0 e: Dockstore	Figur	e 1.
				ingui	ε <u>1</u> .

TheiaCoV_IIIumina_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 Imputs Impu
Step 2
C kilifi_H3N2 V 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
 - 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
- 9. Specify the desired *dataset tags* and *docker image* inputs

	a in Terra using Theiagen's ina PE Workflow			
Document TG-S	Document TG-SC2-PE, Version 3			
Date:	Workflow Versions:			
4/25/2024	PHB v2			

Selec proce	t SD_TheiaPro	oks to 🌼 SETTINGS	100 rows selected	⊟ ADVANCED SEARCH	Searc	h C	
•		ectedTaxonID	()	gambit_predicted_taxon	()	ts_mlst_pubmlst_scher	ne
Page	QP1	Escherichia coli		Escherichia coli		ecoli_achtman_4	
All (103) None	IQP1	Escherichia coli		Escherichia coli		ecoli_achtman_4	
~	DHQP1	Escherichia coli		Escherichia coli		ecoli_achtman_4	
~	DHQP1	Escherichia coli		Escherichia coli		ecoli_achtman_4	
×	DHQP1	Escherichia coli		Escherichia coli		ecoli_achtman_4	
~	ERR10	Listeria monocytogenes		Listeria monocytogenes		listeria_2	
		oks will be saved as a ne PE_PHB_2023-07-31T13-45			>	ltems per page: 100 🔹]

- To run TheiaCoV_Illumina_PE_PHB v2.0.0 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 <u>Docker Image and Reference Materials for SARS-CoV-2 Genomic</u> <u>Characterization</u>
 - 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, <u>TheiaCoV Illumina PE PHB 2024-04-19.json</u>, or newer
 - ii. *Right click* and *save* the file (text does not have to be selected to save properly)
 - iii. 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
- 10. Set the first three attributes in the table manually to <u>this.read1</u>, <u>this.read2</u>, and <u>this.illumina_pe_specimen_id</u>, respectively (Fig 4):
 - a. Where *illuming_pe_specimen* is the unique name of your data table in Terra

	a in Terra using Theiagen's ina PE Workflow	
Document TG-S	C2-PE, Version 3	
Date:	Workflow Versions:	
4/25/2024	PHB v2	

Select root entity type: illumina		ks ① □ Retry with more men	hory 0	
SCRIPT ** INPUTS	• • OUTPUTS •• RUN	ANALYSIS		CANCEL SAV
Hide optional inputs			Down	load json Drag or click to upload json SEARCH INPUTS
Task name 🖡	Variable	Туре	Attribute	\mathbf{i}
theiacov_illumina_pe	primer_bed	File	workspace.Artic_V4_primer_bed	()
theiacov_illumina_pe	read1_raw	File	thisread1	() (
theiacov_illumina_pe	read2_raw	File	this read2	الم) 🔁
theiacov_illumina_pe	samplename	String	this.	{} 《
bwa	cpu	Int	this.illumina_pe_specimen_id	(
consensus	char_unknown	String	this.read1 this.read2	[]
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 ANALYSIS							
Output files will be saved to Files / submission unique ID / theiacov_illuminity	a_pe / www.unique ID							
References to outputs will be written to 国 Tables / illumina_pe_specimen Fill in the attributes below to add or update colum								
	Down	oad json Drag or click t	o upload json Clear outputs	SEARCH OUTPUTS				
Task name 🌡	Variable	Туре	Attribute Use defaults					
Figure 5.	abricate_flu_database	String	Optional	{}				



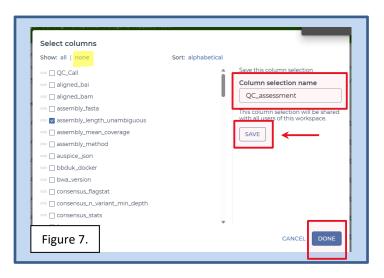
Document TG-SC2-PE, Version 3

Date: 4/25/2024 Workflow Versions: PHB v2

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 settings (Fig 6, green rectangle) and select none to deselect all output columns (Fig 7, 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. <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

DASHBOARD DATA	NOTE	BOOKS WORKFLOWS JO	B HISTORY		
	/ EDI		🌣 SETT	INCS 0 rows selected	
TABLES ~	Add row Add column	en_id ① ↓	assemt	ly_length_unambiguous 🕐 📗 number_N	percent_reference_cover
Search all tables Q	Edit selected	rows	29007	222	97
	Delete selecte	ed rows ble_02	2745	Add a new column	91.8
🖩 illumina_pe_speci (20) 🤅	Save selection	asset ble_03	2392	QC_assignment	80.01
🗐 illumina_pe_specim (2) 🛈	2	Sample_04	2329	Default value (optional, while entered for all rows)	77.91
REFERENCE DATA ~		Sample_05	2505	Type: String Orm O Boolean	83.77
OTHER DATA ~		Sample_06	8546	Value is a list	28.58
Workspace Data		Sample_07	2479:	Enter a value	82.91
🗅 Files		Sample_08	1994	CANCEL SAVE	66.69
		Sample_09	6397		21.39





Document TG-SC2-PE, Version 3

Date: Workflow Versions:

Р

PHB v2

4. <u>Optional</u>: Add a column to record QC PASS/FAIL by clicking edit, add a column (Fig 6)

4/25/2024

- a. Name the new column (e.g. QC_Call); do not include any spaces
- b. Set the value type as string
- c. Click save
- 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 clicking the pencil icon in the corresponding field (Fig 6, red circle)
- 7. For samples that pass the guidance thresholds, proceed to section 4.4
 - 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.	Gui	dance	thres	hold	s for	genome	assembly	QC

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

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 <u>save this column selection</u> field, naming it (e.g. SC2_Results), and clicking <u>save</u>
 - b. Click done

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 3

Date:Workflow Versions:4/25/2024PHB v2

- 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: pipeline-resources/docs/sc2-recombinants.md at main · pha4qe/pipeline-resources · GitHub
- 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

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_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.



Document TG-SC2-PE, Version 3

Date: 4/25/2024 Workflow Versions: PHB v2

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

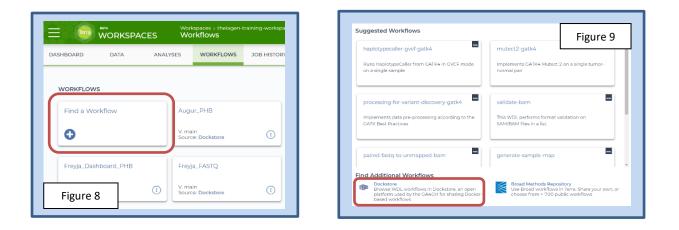
9. **REVISION HISTORY**

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

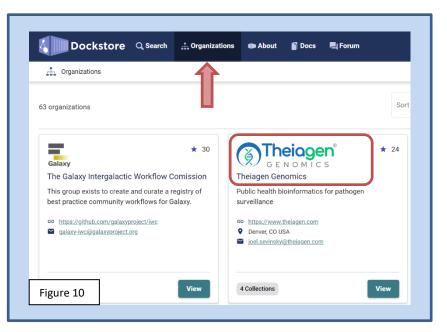
Analyzing SARS-CoV-2 Dat TheiaCoV Illumi	0 0			
Document TG-SC2-PE, Version 3				
Date:	Workflow Versions:			
4/25/2024	PHB v2			

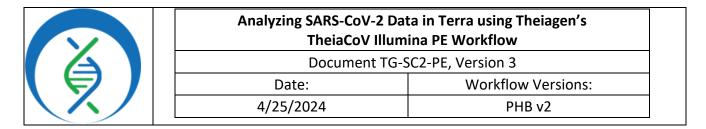
10. APPENDICES

10.1 IMPORTING THE THEIACOV_ILLUMINA_PE_PHB WORKFLOW FROM DOCKSTORE



- 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 surveillan	ice
Collections 4 Ambers 2 C Updates 10	About the Organization
Public Health Bioinformatics (PHB)	This organization does not have a description CD <u>https://www.theiagen.com</u>
Figure 11	Oenver, 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_bioinform na_PE_PHB:v1.0.0	atics/ <mark>TheiaCoV_IIIIumi</mark>
Figure 12 26, 2024 WDL	View

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

ast update to so	arce repository: 25	minutes ago					
Info	Launch	Versions	Files	Tools	DAG	Metrics	Launch with
Workflow	Information						DNAnexus
							Terra
Source Code: TRS: #workflow/github.com/theiagen/public_health_bioinformatics/TheiaCoV_Illumina_PE_PHB					eLwazi		
		iagen/public_health_t	ioinformatics/The	iaCoV_Illumina_PE	РНВ 🖪		W MATCHING
TRS: <u>#workfl</u>	<u>ow/github.com/thei</u>	iagen/public_health_t for genomic characte			-	blogy of	AnVIL

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

	Data in Terra using Theiagen's mina PE Workflow
Document TG-SC2-PE, Version 3	
Date:	Workflow Versions:
4/25/2024	PHB v2

Importing fro	m Dockstore	Workflow Name
github.com/theiag _PHB _{V.v1.0.0}	en/public_health_bioinformatics/TheiaCoV_Illumina_PE	TheiaCoV_Illumina_PE_PHB
Please note: Docks this Workflow will	tore cannot guarantee that the WDL and Docker image referenced by not change. 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
- 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*

ASHBOARD DATA	ANALYSES WORKFLOWS JC	BHISTORY	
🛿 p_aeruginosa_va (35) 🛈		Download TSV Drag or click to up	oad TSV Search
p_aeruginosa_vali (1) 🛈	Key	Value	Description
🗄 p aeruginosa vali (1) 🛈	Artic_V3_primer_bed	4 APTUCOA SATA bullet bed	
🛾 theiaprok_illumin (5) 🛈	Artic_V4-1_primer_bed	V4-1_nCoV-2021,primer.bed	
theiaprok_illumin (1) ①	Artic_V4_primer_bed	V4_nCoV_2021.primer.bed	
E ww_specimen (5)	Midnight_primer_bed	Midnight_Primers_SARS-CoV-2.scheme_	
ww_specimen_set (2)	SWIFT_primer_bed	SWIFT_SABS_CoV_2.scheme.bed	Updated 2023-07-05
	freyja_dashboard_config	freyja_dash_config.json	Input 2023-07-18
REFERENCE DATA V	kraken2 ohoenix	k2_standard_08pb_2023060S.tar.gz	Updated by Ines on 21/07/2023
lo references have been dided. Idd reference data	nextclade_dataset_tag	2022-07-26112:00:002	Updated on 2022-08-12
THER DATA ~	nextclade docker image	nextstrain/nextclade:2.4.0	Updated on 2022-08-12
Workspace Data	pangolin_docker_image	staphb/pangolin:4.1.2-pdata-1.1.6	Updated of 2022-08-12
	vadr docker image	staphb/vadr:1.4.2	22-07-1 ADD VARIABLE +

	ta in Terra using Theiagen's ina PE Workflow	
Document TG-SC2-PE, Version 3		
Date:	Workflow Versions:	
4/25/2024	PHB v2	

	CES Workspaces > theiagen-training-workspaces/Theiagen_Kropp_Sandbox > Data
DASHBOARD DATA	ANALYSES WORKFLOWS JOB HISTORY
🗏 sd_theiaprok (103)	
🗉 sd_theiaprok_set (2) 🔅	Name
🗉 sra_fetch (28)	Assemblies/
■ sra_fetch_set (4)	Candida_albicans_ATCC_10231.fasta
Figure 16 mina (5) 🕀	EASTAS/

- 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. (5)	()	Key	Value	Description
🗉 theiaprok_illumin (1)	()	Artic_V3_primer_bed	V3_nCoV-2019,primer.bed	
🗉 theiaprok_ont (2)	()	Artic_V4-1_primer_bed	gs://theiagen-public-f String ~	
🗉 ww_specimen (5)	1	Artic_V4_primer_bed	V4_nCoV-2021.primer.bed	
ww_specimen_set (5)	1	FreyjaLineageMetadata	<u>curated_lineages.json</u>	Updated 2024-4-11 kk
REFERENCE DATA	~	FreyjaUsherBarcodes	usher_barcodes.csv	Updated 2024-4-11 kk
No references have been		Freyja_ReferenceCenome	<u>nCoV-2019.reference.fasta</u>	MN908947.3
added. Add reference data		Midnight_primer_bed	Midnight_Primers_SARS-CoV-2.scheme.bed	
OTHER DATA	~	SWIFT_primer_bed	SWIFT_SARS-CoV-2.scheme.bed	Updated 2023-07-05
🗉 Workspace Data 🧹		ccess_token	b2aac9f782404625ace2c6ca83e43187	Obtained on 9/14 to state.co
C Files	- 1	api_server	https://api.basespace.illumina.com	ADD VARIABLE +