Document TG-SC2-CL, Version 3

Date: 5/7/2024 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 ClearLabs PHB workflow in Terra to generate assemblies, quality control (QC) metrics, and determine Nextclade clade and Pangolin lineage assignments. Acceptable data types include ClearLabs 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
- ClearLabs raw sequencing read files uploaded to Terra workspace, see TG-TER-03 or TG-TER-04
- Theiagen's TheiaCoV ClearLabs 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
IG-IER-05	Results Metadata Table in Terra
TG-TER-04	Linking BaseSpace and Importing BaseSpace
IG-IER-04	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 (Fig 1)

DASHBOARD DATA	ANALYSES	WORKFLOWS	JOB HISTOR	Y	
WORKFLOWS		labs	× Sor	t By: Alphabet	~
Find a Workflow	Thei	aCoV_ClearLabs		TheiaCoV_ClearLabs_	РНВ

		ata in Terra using Theiagen's os Workflow Version 1	
	Document TG-	-SC2-CL, Version 3	
	Date:	Workflow Versions	
\mathbf{X}	5/7/2024	PHB v2	
		·	

2. Select the *TheiaCoV_ClearLabs_PHB* (Fig 1)

TheiaCoV_ClearLabs_PHB
Version: v2.0.1
Source: github.com/theiagen/public_health_bioinformatics/TheiaCoV_ClearLabs_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 Impute table
S
s C t e KK_ClearLabs_Test V SELECT DATA
🗌 Use call caching 🟮 📄 Delete intermediate outputs 🕄 📄 Use reference disks 🕄 📄 Retry with mo 🛛 Figure 2.

- 3. Uncheck call caching (Fig 2)
- 4. Choose the latest version of *version 2*, or the version used for internal validation (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 *root entity type* dropdown (Fig 2, c)
- 7. Click select data (Fig 2, d)
- 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
 - 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

Choose s		55			
Compose e	e of KK_ClearLabs_Tests				
oce:	set Ki a s_Tests selected Search Search Q				
₩ ¥ 8	KK_ClearLabs_Test_id	QC_Call ①	assembly_length_unambiguous	ass	
Page	-00003868	FAIL	1061	1.7	
All (2)	-00003870	FAIL	715	1.3	
None					
		1-2 of 2 « <	1 > > Iten ge: 100 •	•	
Selected Kk	<_ClearLabs_Tests will be save	d as a new KK_ClearLabs_Test_set na	med:		
TheiaCoV_	ClearLabs_PHB_2023-07-31T13-5	4-41	N		



Document TG-SC2-CL, Version 3

Date:	Workflow Versions
5/7/2024	PHB v2

- To run TheiaCoV_ClearLabs_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>
 - a **NOTE**: TheiaCoV PHB v2.0.0 workflows are not backwards compatible with older versions of Nextclade; <u>use Nextclade Dataset Tag 2024-04-15–15-08-222 or newer</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 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
- 10. Set the first three attributes in the table to *this.reads*, *workspace.Artic_V4-1_primer_bed*, and *this.KK_ClearLabs_Test_id*, respectively (Fig 4) where:
 - a. *this.KK_ClearLabs_Test_id* is the unique name of your data table in Terra

Use call caching 0	termediate outputs		emory 1 Ignore empty outputs
SCRIPT •• INPUTS	•• OUTPUTS •• RUN ANALY	SIS	CA
Hide optional inputs			Download json Drag or ock to upload json dear inputs SEARCH
Task name ↓	Variable	Туре	Attribute
theiacov_clearlabs	clear_lab_fastq	File	this.reads
theiacov_clearlabs	primer_bed	File	workspace.Artic_V4-1_primer_bed
theiacov_clearlabs	samplename	String	Required
Figure 4.	сри	Int	this.KK_ClearLabs_Test_id

- 11. 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

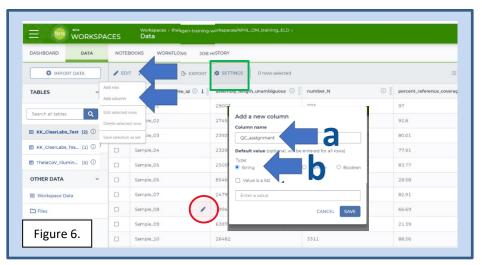
	Analyzing SARS-CoV-2 Data in Terra using Theiagen's TheiaCoV ClearLabs Workflow Version 1		
	Document TG-SC2-CL, Version 3		
	Date:	Workflow Versions	
\mathbf{X}	5/7/2024	PHB v2	
		•	

SCRIPT •• INPUTS ••	OUTPUTS •• RUN ANALYSIS			
Output files will be saved to				
Files / submission unique ID / theiacov_clearlab	s / workflow unique ID			
References to outputs will be written to 圖 Tables / KK_ClearLabs_Test Fill in the attributes below to add or update column	ns in yo a table			
			Download json Drag or click to upload json Clear outputs SEARCH OUTPUTS	
Task name 🌡	Variable	Туре	Attrib te Use defaults	
theiacov_clearlabs	aligned_bai	File	this.aligned_bai	{}
Figure 5.	aligned_bam	File	this.aligned_bam	{}
	artia daaleer	Chrise	this action classical	٦]

- 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

4.2 QUALITY ASSESSMENT OF THEIACOV OUTPUTS

- Navigate to the *data* tab of the workspace containing SC2 data and open the pertinent data table
- 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
- 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



Document TG-SC2-CL, Version 3

Workflow Versions

5/7/2024

Date:

PHB v2

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

Show: all none	Sort: alphabetical	
QC_Call	î.	Save this column selection
📖 🗌 aligned_bai		Column selection name
📖 🗌 aligned_bam		QC_assessment
assembly_fasta		This column selection will be shared
🔤 assembly_length_unambiguous		with all users of this workspace.
assembly_mean_coverage		SAVE -
assembly_method		
auspice_json		
bbduk_docker		
bwa_version		
📖 🗌 consensus_flagstat		
consensus_n_variant_min_depth		
consensus_stats		
	v	

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

Table 1. Guidance thresholds for genome assembly QC

- 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 5), 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
- 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. 211 indicates the sample clade and

¹ Metrics and thresholds 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-CL, Version 3

Date:	Workflow Versions
5/7/2024	PHB v2

- 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 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.
- 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 or newer



Document TG-SC2-CL, Version 3

Date: 5/7/2024 Workflow Versions PHB v2

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/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

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	5/2024

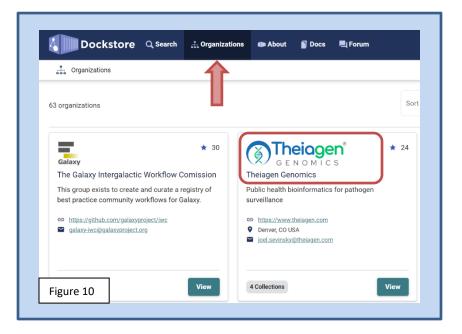
	Data in Terra using Theiagen's abs Workflow Version 1	
Document T	G-SC2-CL, Version 3	
Date: Workflow Versions		
5/7/2024	PHB v2	
		•

10. APPENDICES

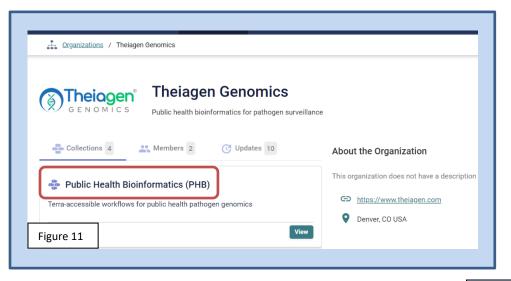
IMPORTING THE THEIACOV_CLEARLABS_PHB WORKFLOW FROM DOCKSTORE

		haplotypecaller-gvcf-gatk4	Figure 9
	VALYSES WORKFLOWS JOB HISTORY	Runs HaplotypeCaller from GATK4 in GVCF mode on a single sample	Implements GATK4 Mutect 2 on a single tumor- normal pair
Find a Workflow	Augur_PHB	processing-for-variant-discovery-gatk4	validate-bam This WDL performs format validation on
•	V. main Source: Dockstore	GATK Best Practices	SAM/BAM files in a list.
Freyja_Dashboard_PHB	Freyja_FASTQ	paired-fastq-to-unmapped-bam	generate-sample-map
Figure 8	V. main Source: Dockstore	Dockstore Browse WDL workflows in Dockstore, an open platform used by the CA4CH for sharing Docker-	Broad Methods Repository Use Broad workflows in Terra. Share your own choose from > 700 public workflows

- 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)



		ata in Terra using Theiagen's os Workflow Version 1	
	Document TG-	Document TG-SC2-CL, Version 3	
	Date:	Workflow Versions	
\mathbf{X}	5/7/2024	PHB v2	



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

github.com/theiagen/public_health_bioinformatics/T	heiaCoV_Clear
Figure 12 97, 2024 WDL	View

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

Tag created: 329 day	ys ago						
Last update to sourc	e repository: 3 ho	ours ago					
Labels theiagen-phb							
Info	Launch	Versions	Files	Tools	DAG	Metrics	Launch with
	<i>c</i>						DNAnexus
Workflow Ir	nformation						Terra
Source Code:							
TRS: <u>#workflow</u>	/github.com/theia	agen/public_health_b	bioinformatics/The	eiaCoV_ClearLabs_F	<u>рнв</u>		
Topic: Bioinform	natics workflows f	for genomic characte	rization, submissi	ion preparation, and	genomic epidemic	ology of	TANVIL 1

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

, .	Data in Terra using Theiagen's abs Workflow Version 1
Document T	G-SC2-CL, Version 3
Date:	Workflow Versions
5/7/2024	PHB v2

Importing from Dockstore	Workflow Name
github.com/theiagen/public_health_bioinformatics/TheiaCoV_ClearLabs_P HB V.v10.0	TheiaCoV_ClearLabs_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
<pre>uversion 1.0</pre>	Select a workspace

10.2 ADDING WORKSPACE DATA ELEMENTS

- 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 JO	BHISTORY	
🛿 p_aeruginosa_va (35) 🛈	*	Download TSV Drag or click to up	oad TSV Search
p_aeruginosa_vali (1) 🛈	Кеу	Value	Description
🗄 p aeruginosa vali (1) 🛈	Artic_V3_primer_bed	APTUCOA SOLADUMOUDED	
🛾 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	
ww_specimen (5)	Midnight_primer_bed	Midnight_Primers_5ARS-CoV-2 scheme	
ww_specimen_set (2)	SWIFT_primer_bed	SWIFT_SARS_CoV_2.scheme.bed	Updated 2023-07-05
	freyja_dashboard_config	frevia_dash_config.json	Input 2023-07-18
REFERENCE DATA ~	kraken2 ohoenix	k2_standard_08pb_20230605.tar.gz	Updated by Ines on 21/07/2023
lo references have been dided. \dd reference data	nextclade_dataset_tag	2025-07-59115:00:00X	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 +

	Analyzing SARS-CoV-2 Data in Terra using Theiagen's TheiaCoV ClearLabs Workflow Version 1		
	Document TG-	Document TG-SC2-CL, Version 3	
	Date:	Workflow Versions	
\mathbf{X}	5/7/2024 PHB v2		

	CES Workspaces > theiagen-training-workspaces/Theiagen_Kropp_Sandbox > Data
DASHBOARD DATA	ANALYSES WORKFLOWS JOB HISTORY
🗉 sd_theiaprok (103) 🤅	Files DPLOAD
🔳 sd_theiaprok_set (2) 🔅	□ Name
sra_fetch (28)	Assemblies/
■ sra_fetch_set (4) ③	Candida_albicans_ATCC_10231.fasta
Figure 16 mina (5) 🔅	➢ FASTAS/

- 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
I 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)	()	Artic_V4_primer_bed	V4_nCoV-2021,primer.bed	
🗏 ww_specimen_set (5)	1	FreyjaLineageMetadata	to <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_ReferenceGenome	• nCoV-2019.reference.fasta	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	