Public Health Viral Genomics (Theiagen)

Release 2.0.0

Kevin G. Libuit

CONTENTS

1	Cont	Contents				
	1.1	Public Health Viral Genomics	1			
	1.2	TheiaCoV Workflow Series	1			
	1.3	Mercury Workflow Series	38			
	1 4	License	54			

CHAPTER

ONE

CONTENTS

1.1 Public Health Viral Genomics

The Theiagen Public Health Viral Genomics repository hosts a collection of WDL workflows for genomic characterization, submission preparation, and genomic epidemiology of the SARS-CoV-2 virus. While these workflows can be run locally or on an HPC system at the command-line with Cromwell or miniWDL, we strongly recommend use through Terra, a bioinformatics web application developed by the Broad Institute of MIT and Harvard in collaboration with Microsoft and Verily Life Sciences.

1.1.1 Getting Started

A series of introductory training videos that provide conceptual overviews of methodologies and walkthrough tutorials on how to utilize our WDL workflows through Terra are available on the Theiagen Genomics YouTube page:

1.1.2 Support

For questions or general support regarding the WDL workflows in this repository, please contact support@theiagen.com

1.2 TheiaCoV Workflow Series

The TheiaCoV Workflow Series is a collection of WDL workflows developed for performing genomic characterization and genomic epidemiology of SARS-CoV-2 samples to support public health decision-making.

1.2.1 TheiaCoV Workflows for Genomic Characterization

Genomic characterization, *i.e.* generating consensus assemblies (FASTA format) from next-generation sequencing (NGS) read data (FASTQ format) to assign samples with relevant nomenclature designation (e.g. PANGO lineage and NextClade clades) is an increasingly critical function to public health laboratories around the world.

The TheiaCoV Genomic Characterization Series includes four separate WDL workflows (TheiaCoV_Illumina_PE, TheiaCoV_Illumina_SE, TheiaCoV_ClearLabs, and TheiaCoV_ONT) that process NGS read data from four different sequencing approaches: Illumina paired-end, Illumina single-end, Clear Labs, and Oxford Nanopore Technology (ONT)) to generate consensus assemblies, produce relevant quality-control metrics for both the input read data and the generated assembly, and assign samples with a lineage and clade designation using Pangolin and NextClade, respectively.

All four TheiaCoV workflows for genomic characterization will generate a viral assembly by mapping input read data to a reference genome, removing primer reads from that alignment, and then calling the consensus assembly based on

the primer-trimmed alignment. These consensus assemblies are then fed into the Pangolin and NextClade CLI tools for lineage and clade assignments.

The major difference between each of these TheiaCoV Genomic Characterization workflows is in how the read mapping, primer trimming, and consensus genome calling is performed. More information on the technical details of these processes and information on how to utilize and apply these workflows for public health investigations is available below.

A fifth WDL workflow, TheiaCoV_FASTA, was added to take in assembled SC2 genomes, perform basic QC (e.g. number of Ns), and assign samples with a lineage and clade designation using Pangolin and NextClade, respectively.

A series of introductory training videos that provide conceptual overviews of methodologies and walkthrough tutorials on how to utilize these TheiaCoV workflows through Terra are available on the Theiagen Genomics YouTube page:

note Titan workflows in the video have since been renamed to TheiaCoV.

TheiaCoV_IIIumina_PE

The TheiaCoV_Illumina_PE workflow was written to process Illumina paired-end (PE) read data. Input reads are assumed to be the product of sequencing tiled PCR-amplicons designed for the SARS-CoV-2 genome. The most common read data analyzed by the TheiaCoV_Illumina_PE workflow are generated with the Artic V3 protocol. Alternative primer schemes such as the Qiaseq Primer Panel, the Swift Amplicon SARS-CoV-2 Panel and the Artic V4 Amplicon Sequencing Panel however, can also be analysed with this workflow since the primer sequence coordinates of the PCR scheme utilized must be provided along with the raw paired-end Illumina read data in BED and FASTQ file formats, respectively.

Note: By default, this workflow will assume that input reads were generated using a 300-cycle kit (i.e. 2×150 bp reads). Modifications to the optional parameter for trimmomatic_minlen may be required to accommodate for shorter read data, such as 2×75 bp reads generated using a 150-cycle kit.

Upon initiating a TheiaCoV_Illumina_PE job, the input primer scheme coordinates and raw paired-end Illumina read data provided for each sample will be processed to perform consensus genome assembly, infer the quality of both raw read data and the generated consensus genome, and assign SARS-CoV-2 lineage and clade types as outlined in the TheiaCoV_Illumina_PE data workflow below.

Consensus genome assembly with the TheiaCoV_Illumina_PE workflow is performed by first de-hosting read data with the NCBI SRA-Human-Scrubber tool then trimming low-quality reads with Trimmomatic and removing adapter sequences with BBDuk. These cleaned read data are then aligned to the Wuhan-1 reference genome with BWA to generate a Binary Alignment Mapping (BAM) file. Primer sequences are then removed from the BAM file using the iVar Trim sub-command. The iVar consensus sub-command is then utilized to generate a consensus assembly in FASTA format. This assembly is then used to assign lineage and clade designations with Pangolin and NextClade. NCBI'S VADR tool is also employed to screen for potentially errant features (e.g. erroneous frame-shift mutations) in the consensus assembly.

More information on required user inputs, optional user inputs, default tool parameters and the outputs generated by TheiaCoV_Illumina_PE are outlined below.

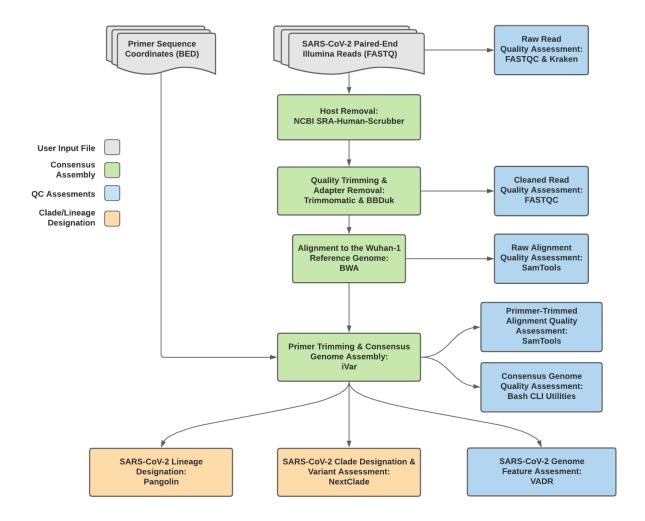


Fig. 1: TheiaCoV_Illumina_PE Data Workflow

Required User Inputs

Download CSV: TheiaCoV_Illumina_PE_required_inputs.csv

Task	Input Variable	Data Type	Description
theia-	primer_bed	File	Primer sequence coordinates of the PCR
cov_illumina_pe			scheme utilized in BED file format
theia-	read1_raw	File	Forward Illumina read in FASTQ file format
cov_illumina_pe			
theia-	read2_raw	File	Reverse Illumina read in FASTQ file format
cov_illumina_pe			
theia-	samplename	String	Name of the sample being analyzed
cov_illumina_pe			

Optional User Inputs

Download CSV: TheiaCoV_Illumina_PE_optional_inputs.csv

Task	Variable Name	Data Type	Description	Default
bwa	refer-	String	Path to the ref-	/artic-
	ence_genome		erence genome	ncov2019/primer_schemes/nCoV-
			within the	2019/V3/nCoV-
			staphb/ivar:1.2.2_	ar 20c20.2005∂s ce.fasta
			Docker con-	
			tainer	
bwa	cpus	Int	CPU resources	6
			allocated to	
			the BWA task	
			runtime envi-	
			ronment	
consensus	char_unknown	String	Character to	N
			print in regions	
			with less than	
			minimum cov-	
			erage for iVar	
			consensus	
consensus	count_orphans	Boolean	Do not skip	TRUE
			anomalous	
			read pairs in	
			variant calling	
			for SAMtools	
			mpileup before	
			running iVar	
			consensus	

Table 1 – continued from previous page

Task	Variable Name	Data Type	Description	Default
consensus	disable_baq	Boolean	Disable read-	TRUE
			pair overlap	
			detection for	
			SAMtools	
			mpileup before	
			running iVar	
			consensus	
consensus	max_depth	Int	Maximum reads	600000
			read at a posi-	
			tion per input	
			file for SAM-	
			tools mpileup	
			before running	
			iVar consensus	
consensus	min_bq	Int	Minimum map-	0
			ping quality for	
			an alignment	
			to be used	
			for SAMtools	
			mpileup before	
			running iVar	
			consensus	
consensus	min_depth	Int	Minimum read	100
			depth to call	
			variants for iVar	
			consensus	
consensus	min_freq	Float	Minimum	0.6
			frequency	
			threshold(0 -	
			1) to call vari-	
			ants for iVar	
			consensus	
consensus	min_qual	Int	Minimum qual-	20
			ity threshold for	
			sliding window	
			to pass for iVar	
			consensus	
consensus	ref_genome	String	Path to the ref-	/artic-
			erence genome	ncov2019/primer_schemes/nCoV-
			within the	2019/V3/nCoV-
				ar 21029)2010528 ce.fasta
			Docker con-	
			tainer	

Table 1 – continued from previous page

Task	Variable Name	ble 1 – continued Data Type	Description	Default
consensus	ref_gff	String	Path to the	/refer-
Consciisus	lei_gii	Jung	general fea-	ence/GCF_009858895.2_ASM985889v3_genomic.gff
ļ		I	ture format	elice/Oct007030073.2_ASIV1703007 v3_genomic.gn
ļ		I	of the refer-	
ļ		I		
ļ		I	ence genome within the	
ļ		I	staphb/ivar:1.2.2_	nut: 200200520
ļ		I		artic20200328
ļ		I		
t-1-40 one so	11	Ct	tainer	
nextclade_one_sa	mptæker	String	Docker tag used	nextstrain/nextclade:1.10.3
ļ		I	for running	
			NextClade	
nextclade_output_	_pdrxdre_one_sample	String	Docker tag	python:slim
ļ		I	used for pars-	
ļ		I	ing NextClade	
			output	
pangolin3	docker	String	Docker tag used	quay.io/staphb/3.1.20-pangolearn-
ļ		I	for running Pan-	2022-02-02
		<u> </u>	golin	
pangolin3	infer-	String	pangolin infer-	usher
ļ	ence_engine	I	ence engine for	
ļ		I	lineage designa-	
ļ		I	tions (usher or	
		<u></u> _	pangolarn)	
pangolin3	min_length	Int	Minimum query	10000
ļ		I	length allowed	
ļ		I	for pangolin	
ļ		I	to attempt	
ļ		I	assignment	
pangolin3	max_ambig	Float	Maximum pro-	0.5
-		I	portion of Ns al-	
ļ		I	lowed for pan-	
ļ		I	golin to attempt	
ļ		I	assignment	
primer_trim	keep_noprimer_re	a B soolean	Include reads	True
F			with no primers	
ļ		I	for iVar trim	
read_QC_trim	bbduk_mem	Int	Memory al-	8
10		1	located to the	
ļ		I	BBDuk VM	
read_QC_trim	trimmo-	Int	Specifies the	25
1044_20_41111	matic_minlen		minimum	
ļ	matic_mmen	I	length of reads	
ļ		I	to be kept for	
ļ		I	Trimmomatic	
read_QC_trim	trimmo-	Int	Specifies the av-	30
reau_QC_umii	matic_quality_trin		erage quality re-	30
ļ	manc_quanty_um	II_SCOIE	quired for Trim-	
ļ		I	momatic	
			Illomatic	continues on port page

Table 1 – continued from previous page

Task	Variable Name	Data Type	Description	Default
read_QC_trim	trimmo-	Int	Specifies the	4
read_Qe_triiii	matic_window_si		number of	T
	matic_window_si		bases to aver-	
			age across for	
			Trimmomatic	
theia-	nextclade_dataset	r Saturina cr	Nextclade or-	sars-cov-2
cov_illumina_pe	nexterade_dataset	Litering	ganism dataset	Sars-cov-2
theia-	nextclade_dataset	r@feinemce	Nextclade refer-	MN908947
cov_illumina_pe	ilexterade_dataset		ence genome	WIN900947
theia-	nextclade_dataset	tNevtclade	2022-02-	
cov_illumina_pe	ilexterade_dataset	dataset tag	07T12:00:00Z	
theia-	seq_method	String	Description of	Illumina paired-end
cov_illumina_pe	seq_memou	Sumg	the sequencing	mumma paneu-enu
cov_mumma_pc			methodology	
			used to generate	
			the input read	
			data	
vadr	docker	String	Docker tag used	quay.io/staphb/1.4.1-models-1.3-2
vaui	UOCKCI	Sumg	for running	quay.10/stap110/1.4.1-1110de1s-1.3-2
			VADR	
vadr	maxlen	Int	Maximum	30000
vaui	IIIaxicii	IIIt	length for the	30000
			fasta-trim-	
			terminal-	
			ambigs.pl	
			VADR script	
vadr	minlen	Int	Minimum	50
vaui	IIIIIICII	1111	length sub-	30
			sequence to	
			possibly replace	
			Ns for the fasta-	
			trim-terminal-	
			ambigs.pl	
			VADR script	
vadr	skip_length	Int	Minimum as-	10000
, au	Skip_iongui	1111	sembly length	10000
			(unambiguous)	
			to run vadr	
vadr	vadr_opts	String	Options for the	-glsearch -s -r -nomisc
	_ 1		v-annotate.pl	-mkey sarscov2 -alt_fail lows-
			VADR script	core,fstukcnf,insertnn,deletinn
			r ·	-mdir /opt/vadr/vadr-models/
variant_call	count_orphans	Boolean	Do not skip	TRUE
_	_ 1		anomalous	
			read pairs in	
			variant calling	
			for SAMtools	
			mpileup before	
			running iVar	
			variants	
	I.	1	1	continues on next need

Table 1 – continued from previous page

Task	Variable Name	e Data Type	Description	Default
variant_call	disable_baq	Boolean	Disable read-	TRUE
_			pair overlap	
			detection for	
			SAMtools	
			mpileup before	
			running iVar	
			variants	
variant_call	max_depth	Int	Maximum reads	600000
<u> </u>			read at a posi-	
			tion per input	
			file for SAM-	
			tools mpileup	
			before running	
			iVar variants	
variant_call	min_bq	Int	Minimum map-	0
			ping quality for	
			an alignment	
			to be used	
			for SAMtools	
			mpileup before	
			running iVar	
_			variants	
variant_call	min_depth	Int	Minimum read	100
			depth to call	
			variants for iVar	
11			variants	
variant_call	min_freq	Float	Minimum	0.6
			frequency	
			threshold(0 - 1)	
			to call variants	
• -4 -511		т.	for iVar variants	22
variant_call	min_qual	Int	Minimum qual-	20
			ity threshold for	
			sliding window to pass for iVar	
			variants	
variant_call	ref_gff	String	Path to the	/refer-
Variant_can	ici_gii	Sumg	general fea-	ence/GCF_009858895.2_ASM985889v3_genomic.g
			ture format	Clice/OCI _007030073.2_ASW1703007V3_genomic.g
			of the refer-	
			ence genome	
			within the	
			staphb/ivar:1.2.2_	artic20200528
			Docker con-	
			tainer	
variant_call	ref_genome	String	Path to the ref-	/artic-
· **	-6		erence genome	ncov2019/primer_schemes/nCoV-
			within the	2019/V3/nCoV-
				ar 210291206528 ce.fasta
			Docker con-	
			tainer	
				continues on next page

Table 1 – continued from previous page

Task	Variable Name	Data Type	Description	Default
version_capture	timezone	String	User time	None
			zone in valid	
			Unix TZ string	
			(e.g. Amer-	
			ica/New_York)	

Outputs

Download CSV: TheiaCoV_Illumina_PE_default_outputs.csv

Output Name	Data Type	Description
aligned_bai	File	Index companion file to the bam file generated during the consensus
		assembly process
aligned_bam	File	Primer-trimmed BAM file; generated during conensus assembly
		process
assembly_fasta	File	Consensus genome assembly
assem-	Int	Number of unambiguous basecalls within the SC2 consensus assem-
bly_length_unambigu	ous	bly
assem-	Float	Mean sequencing depth throughout the conesnsus assembly gener-
bly_mean_coverage		ated after performing primer trimming-calculated using the SAM-
		tools coverage command
assembly_method	String	Method employed to generate consensus assembly
auspice_json	File	Auspice-compatable JSON output generated from NextClade analy-
		sis that includes the NextClade default samples for clade-typing and
		the single sample placed on this tree
bbduk_docker	String	Docker image used to run BBDuk
bwa_version	String	Version of BWA used to map read data to the reference genome
consensus_flagstat	File	Output from the SAMtools flagstat command to assess quality of the
		alignment file (BAM)
consensus_stats	File	Output from the SAMtools stats command to assess quality of the
		alignment file (BAM)
fastqc_clean1	Int	Number of forward reads after sequelean filtering as determined by
		FastQC
fastqc_clean2	Int	Number of reverse reads after sequelean filtering as determined by
		FastQC
fastqc_clean_pairs	String	Number of paired reads after SeqyClean filtering as determined by
		FastQC
fastqc_raw1	Int	Number of forward reads identified in the input fastq files as deter-
		mined by FastQC
fastqc_raw2	Int	Number of reverse reads identified in the input fastq files as deter-
		mined by FastQC
fastqc_raw_pairs	String	Number of paired reads identified in the input fastq files as deter-
		mined by FastQC
fastqc_version	String	Version of the FastQC software used for read QC analysis
ivar_tsv	File	Variant descriptor file generated by iVar variants

Table 2 – continued from previous page

Output Name	Data Type	Description
ivar_variant_version	String	Version of iVar for running the iVar variants command
ivar_vcf	File	iVar tsv output converted to VCF format
ivar_version_consenst		Version of iVar for running the iVar consensus command
ivar_version_primtrin		Version of iVar for running the iVar trim command
kraken_human	Float	Percent of human read data detected using the Kraken2 software
kraken_human_dehos		Percent of human read data detected using the Kraken2 software af-
		ter host removal
kraken_report	File	Full Kraken report
kraken_report_dehost	edFile	Full Kraken report after host removal
kraken_sc2	Float	Percent of SARS-CoV-2 read data detected using the Kraken2 soft-
		ware
kraken_sc2_dehosted	Float	Percent of SARS-CoV-2 read data detected using the Kraken2 soft-
		ware after host removal
kraken_version	String	Version of Kraken software used
meanbaseq_trim	Float	Mean quality of the nucleotide basecalls aligned to the reference
_		genome after primer trimming
meanmapq_trim	Float	Mean quality of the mapped reads to the reference genome after
		primer trimming
nextclade_aa_dels	String	Amino-acid deletions as detected by NextClade
nextclade_aa_subs	String	Amino-acid substitutions as detected by NextClade
nextclade_clade	String	NextClade clade designation
nextclade_json	File	NexClade output in JSON file format
nextclade_tsv	File	NextClade output in TSV file format
nextclade_version	String	Version of NextClade software used
number_Degenerate	Int	Number of degenerate basecalls within the consensus assembly
number_N	Int	Number of fully ambiguous basecalls within the consensus assembly
number_Total	Int	Total number of nucleotides within the consensus assembly
pango_lineage	String	Pango lineage as detremined by Pangolin
pango_lineage_report	File	Full Pango lineage report generated by Pangolin
pan-	String	Version of the pangolin software (e.g. PANGO or PUSHER) used
golin_assignment_ver	sion	for lineage asignment
pangolin_conflicts	String	Number of lineage conflicts as deteremed by Pangolin
pangolin_docker	String	Docker image used to run Pangolin
pangolin_notes	String	Lineage notes as deteremined by Pangolin
pangolin_versions	String	All Pangolin software and database version
per-	Float	Percent coverage of the reference genome after performing primer
cent_reference_covera	ige	trimming; calculated as assembly_length_unambiguous / length of
		reference genome (SC2: 29,903) x 100
primer_bed_name	String	Name of the primer bed files used for primer trimming
primer_trimmed_read	_pdoaent	Percent of read data with primers trimmed as deteremined by iVar
		trim
read1_clean	File	Forward read file after quality trimming and adapter removal
read1_dehosted	File	Dehosted forward reads; suggested read file for SRA submission
read2_clean	File	Reverse read file after quality trimming and adapter removal
read2_dehosted	File	Dehosted reverse reads; suggested read file for SRA submissionsam-
		tools_version
samtools_version	String	Version of SAMtools used to sort and index the alignment file
sam-	String	Version of SAMtools used to create the pileup before running iVar
tools_version_consens	sus	consensus

Table 2 – continued from previous page

Output Name	Data Type	Description
sam-	String	Version of SAMtools used to create the pileup before running iVar
tools_version_primtri	m	trim
sam-	String	Version of SAMtools used to assess quality of read mapping
tools_version_stats		
seq_platform	String	Description of the sequencing methodology used to generate the in-
		put read data
theia-	String	Date of analysis
cov_illumina_pe_anal	ysis_date	
theia-	String	Version of the Public Health Viral Genomics (PHVG) repository
cov_illumina_pe_vers	ion	used
trimmo-	String	Version of Trimmomatic used
matic_version		
vadr_alerts_list	File	File containing all of the fatal alerts as determined by VADR
vadr_docker	String	Docker image used to run VADR
vadr_num_alerts	String	Number of fatal alerts as determined by VADR

TheiaCoV Illumina SE

The TheiaCoV_Illumina_SE workflow was written to process Illumina single-end (SE) read data. Input reads are assumed to be the product of sequencing tiled PCR-amplicons designed for the SARS-CoV-2 genome. The most common read data analyzed by the TheiaCoV_Illumina_SE workflow are generated with the Artic V3 protocol. Alternative primer schemes such as the Qiaseq Primer Panel, however, can also be analysed with this workflow since the primer sequence coordinates of the PCR scheme utilized must be provided along with the raw paired-end Illumina read data in BED and FASTQ file formats, respectively.

Note: By default, this workflow will assume that input reads were generated using a 35-cycle kit (i.e. 1 x 35 bp reads). Modifications to the optional parameter for trimmomatic_minlen may be required to accommodate for longer read data.

Upon initiating a TheiaCoV_Illumina_SE job, the input primer scheme coordinates and raw paired-end Illumina read data provided for each sample will be processed to perform consensus genome assembly, infer the quality of both raw read data and the generated consensus genome, and assign SARS-CoV-2 lineage and clade types as outlined in the TheiaCoV_Illumina_PE data workflow below.

Consensus genome assembly with the TheiaCoV_Illumina_SE workflow is performed by first trimming low-quality reads with Trimmomatic and removing adapter sequences with BBDuk. These cleaned read data are then aligned to the Wuhan-1 reference genome with BWA to generate a Binary Alignment Mapping (BAM) file. Primer sequences are then removed from the BAM file using the iVar Trim sub-command. The iVar consensus sub-command is then utilized to generate a consensus assembly in FASTA format. This assembly is then used to assign lineage and clade designations with Pangolin and NextClade. NCBI'S VADR tool is also employed to screen for potentially errant features (e.g. erroneous frame-shift mutations) in the consensus assembly.

More information on required user inputs, optional user inputs, default tool parameters and the outputs generated by TheiaCoV_Illumina_SE are outlined below.

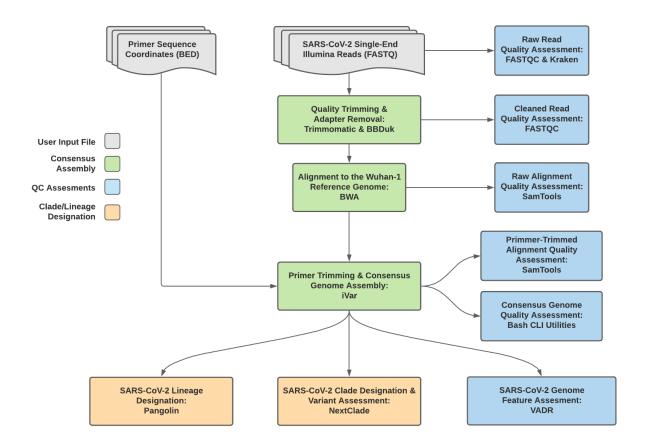


Fig. 2: TheiaCoV_Illumina_SE Data Workflow

12 Chapter 1. Contents

Required User Inputs

Download CSV: TheiaCoV_Illumina_SE_required_inputs.csv

Task	Input Variable	Data Type	Description
theia-	primer_bed	File	Primer sequence coordinates of the PCR
cov_illumina_pe			scheme utilized in BED file format
theia-	read1_raw	File	Single-end Illumina read in FASTQ file for-
cov_illumina_pe			mat
theia-	samplename	String	Name of the sample being analyzed
cov_illumina_pe			

Optional User Inputs

Download CSV: TheiaCoV_Illumina_SE_optional_inputs.csv

Task	Variable Name	Data Type	Description	Default
bwa	refer-	String	Path to the ref-	/artic-
	ence_genome		erence genome	ncov2019/primer_schemes/nCoV-
			within the	2019/V3/nCoV-
			staphb/ivar:1.2.2_	ar 20c20.200528 ce.fasta
			Docker con-	
			tainer	
bwa	cpus	Int	CPU resources	6
			allocated to	
			the BWA task	
			runtime envi-	
			ronment	
bwa	read2	File	Optional input	None
			file for the	
			Kraken task that	
			is not applicable	
			to this workflow	
consensus	char_unknown	String	Character to	N
			print in regions	
			with less than	
			minimum cov-	
			erage for iVar	
			consensus	
consensus	count_orphans	Boolean	Do not skip	TRUE
			anomalous	
			read pairs in	
			variant calling	
			for SAMtools	
			mpileup before	
			running iVar	
			consensus	

Table 3 – continued from previous page

Task	Variable Name		Description	Default
consensus	disable_baq	Boolean	Disable read-	TRUE
			pair overlap	
			detection for	
			SAMtools	
			mpileup before	
			running iVar	
			consensus	
consensus	max_depth	Int	Maximum reads	600000
			read at a posi-	
			tion per input	
			file for SAM-	
			tools mpileup	
			before running	
			iVar consensus	
consensus	min_bq	Int	Minimum map-	0
			ping quality for	
			an alignment	
			to be used	
			for SAMtools	
			mpileup before	
			running iVar	
			consensus	
consensus	min_depth	Int	Minimum read	100
			depth to call	
			variants for iVar	
			consensus	
consensus	min_freq	Float	Minimum	0.6
			frequency	
			threshold(0 -	
			1) to call vari-	
			ants for iVar	
			consensus	
consensus	min_qual	Int	Minimum qual-	20
			ity threshold for	
			sliding window	
			to pass for iVar	
			consensus	
consensus	ref_genome	String	Path to the ref-	/artic-
			erence genome	ncov2019/primer_schemes/nCoV-
			within the	2019/V3/nCoV-
				ar 210:29)2010:5:218 ce.fasta
			Docker con-	
			tainer	

Table 3 – continued from previous page

Task	Variable Name	Data Type	Description	Default
consensus	ref_gff	String	Path to the	/refer-
		-	general fea-	ence/GCF_009858895.2_ASM985889v3_genomic.gff
			ture format	
			of the refer-	
			ence genome	
			within the	
			staphb/ivar:1.2.2_	artic20200528
			Docker con-	
			tainer	
nextclade_one_sar	m øb æker	String	Docker tag used	nextstrain/nextclade:1.10.3
			for running	
			NextClade	
nextclade_output_	pdrxdreione_sample	String	Docker tag	python:slim
			used for pars-	
			ing NextClade	
			output	
pangolin3	docker	String	Docker tag used	quay.io/staphb/3.1.20-pangolearn-
			for running Pan-	2022-02-02
			golin	
pangolin3	infer-	String	pangolin infer-	usher
	ence_engine		ence engine for	
			lineage designa-	
			tions (usher or	
			pangolarn)	
pangolin3	min_length	Int	Minimum query	10000
			length allowed	
			for pangolin	
			to attempt	
			assignment	
pangolin3	max_ambig	Float	Maximum pro-	0.5
			portion of Ns al-	
			lowed for pan-	
			golin to attempt	
			assignment	
primer_trim	keep_noprimer_re	a B soolean	Include reads	True
			with no primers	
			for iVar trim	
read_QC_trim	bbduk_mem	Int	Memory al-	8
			located to the	
			BBDuk VM	
read_QC_trim	trimmo-	Int	Specifies the	25
	matic_minlen		minimum	
			length of reads	
			to be kept for	
			Trimmomatic	
read_QC_trim	trimmo-	Int	Specifies the av-	30
	matic_quality_trin	n_score	erage quality re-	
			quired for Trim-	
			momatic	
			шешине	

Table 3 – continued from previous page

read_QC_trim trimmo-matic_window_size Parallel	Task	Variable Name	Data Type	Description	Default
matic_window_size theia- theia- cov_illumina_se theia- cov_illumina_					
bases to average across for Trimmomatic strimmomatic stri			ze	-	
Trimmomatic Nextclade Attack String Nextclade Organism dataset				bases to aver-	
theia- cov_illumina_se theia- th				age across for	
theia- cov_illumina_se				Trimmomatic	
theia- cov_illumina_se theia- cov_illumina_paired-end thesequence the input read data theia- cov_illumina_paired-end thesequence the input read data theia- cov_illumina_paired-end thesequence the input read data theia- cov_illumina_se theia- cov_illumina_se theia- cov_illumina_se theia- cov_illumina_paired-end thesequence the input read data theia- cov_illumina_paired	theia-	nextclade_dataset	_r Strie g	Nextclade or-	sars-cov-2
theia- tov_illumina_se theia- cov_illumina_se	cov_illumina_se				
theia- cov_illumina_se the input read data Description of the sequencing methodology used to generate the input read data Description of the sequencing methodology used to generate the input read data 30000 Maximum length for the fasta-trim- terminal- ambigs_pl VADR script VADR script		nextclade_dataset	_n Steineg ce		MN908947
theia- th					
theia- cov_illumina_se seq_method String Description of the sequencing methodology used to generate the input read data vadr odocker String Docker tag used for running VADR VADR vadr maxlen Int Maximum length for the fasta-trim- terminal- ambigs.pl VADR script vadr minlen Int Minimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script vadr vadr skip_length vadr Vadr vadr vadr vadr vadr vadr vadr vadr ovadr vadr va		nextclade_dataset	-		
the sequencing methodology used to generate the input read data vadr docker String Docker tag used for running VADR vadr maxlen Int Maximum length for the fasta-trim-terminal-ambigs.pl VADR script vadr minlen Int Minimum length subsequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl VADR script vadr skip_length Int Minimum assembly length (unambiguous) to run vadr vadr vadr vadr_opts String Options for the v-annotate.pl VADR script vadr count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before the sequencing methodology used to generate the input read data data pocker tag used for running quay.io/staphb/1.4.1-models-1.3-2 for annotate place and the subsequence to possibly replace Ns for the fasta-trim-terminal-ambigs.place Ns for the					
vadr docker String Docker tag used for running VADR vadr maxlen Int Maximum length for the fasta-trim-terminal-ambigs.pl VADR script vadr minlen Int Minimum length sub-sequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl VADR script vadr skip_length Int Minimum as-sembly length (unambiguous) to run vadr vadr vadr opts String Options for the v-annotate.pl VADR script vadr count_orphans Boolean Do not skip anomalous read pairs in variant_calling for SAMtools mpileup before		seq_method	String	•	Illumina paired-end
vadr docker String Docker tag used for running VADR quay.io/staphb/1.4.1-models-1.3-2 vadr maxlen Int Maximum length for the fasta-trim-terminal-ambigs.pl VADR script 30000 vadr minlen Int Minimum length sub-sequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl VADR script 50 vadr skip_length Int Minimum assembly length (unambiguous) to run vadr 10000 vadr vadr_opts String Options for the vannotate.pl VADR script —glsearch -s -r -nomisc -mkey sarscov2 -alt_fail lowscore,fstukcnf,insertnn,deletinn -mdir /opt/vadr/vadr-models/ variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before TRUE	cov_illumina_se				
vadr docker String Docker tag used for running VADR quay.io/staphb/1.4.1-models-1.3-2 vadr maxlen Int Maximum length for the fasta-trim-terminal-ambigs.pl VADR script 30000 vadr minlen Int Minimum length sub-sequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl VADR script 50 vadr skip_length Int Minimum as-sembly length (unambiguous) to run vadr 10000 vadr vadr_opts String Options for the v-annotate.pl VADR script —mkey sarscov2 —alt_fail lows-core,fstukenf,insertnn,deletim —ndir /opt/vadr/vadr-models/ variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before TRUE					
vadr docker String Docker tag used for running VADR quay.io/staphb/1.4.1-models-1.3-2 vadr maxlen Int Maximum length for the fasta-trim-terminal-ambigs.pl VADR script 30000 vadr minlen Int Minimum length sub-sequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl VADR script 50 vadr skip_length Int Minimum assembly length (unambiguous) to run vadr (unambiguous) to run vadr 10000 vadr vadr_opts String Options for the vannotate.pl VADR script -glsearch -s -r -nomisc -mkey sarscov2 -alt_fail lowscore,fstukcnf,insertnn,deletinn -mdir /opt/vadr/vadr-models/ variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before TRUE					
vadr docker String Docker tag used for running VADR quay.io/staphb/1.4.1-models-1.3-2 vadr maxlen Int Maximum length for the fasta-trim-terminal-ambigs.pl VADR script 30000 vadr minlen Int Minimum length sub-sequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl VADR script 50 vadr skip_length Int Minimum assembly length (unambiguous) to run vadr un vadr 10000 vadr vadr_opts String Options for the v-annotate.pl VADR script -glsearch -s -r -nomisc -mkey sarscov2 -alt_fail lowscore, fstukenf, insertun, deletinn -mdir /opt/vadr/vadr-models/ variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before				-	
vadr maxlen Int Maximum length for the fasta-trim-terminal-ambigs.pl VADR script vadr minlen Int Minimum length subsequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl VADR script vadr skip_length Int Minimum assembly length (unambiguous) to run vadr vadr vadr vadr_opts String Options for the v-annotate.pl VADR script variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before	vodr	docker	String		guavio/stanhh/1 // 1 models 1 2 2
vadr maxlen Int Maximum length for the fasta-trim-terminal-ambigs.pl VADR script vadr minlen Int Minimum length subsequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl VADR script vadr skip_length Int Minimum assembly length (unambiguous) to run vadr (unambiguous) to run vadr vadr vadr_opts String Options for the v-annotate.pl VADR script vadr vadr_opts String Options for the v-annotate.pl VADR script variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before	vaui	docker	Sung	_	quay.10/stap110/1.4.1-1110de1s-1.3-2
vadr maxlen Int Maximum length for the fasta-trim-terminal-ambigs.pl VADR script vadr minlen Int Minimum length subsequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl VADR script vadr skip_length Int Minimum assembly length (unambiguous) to run vadr vadr vadr_opts String Options for the v-annotate.pl VADR script —glsearch -s -r -nomisc -mkey sarscov2 -alt_fail lowscore,fstukcnf,insertun,deletimn -mdir /opt/vadr/vadr-models/ variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before					
length for the fasta-trim-terminal-ambigs.pl VADR script vadr	vadr	maxlen	Int	*	30000
vadr minlen Int Minimum length subsequence to possibly replace Ns for the fastatrim-terminal-ambigs.pl VADR script vadr skip_length Int Minimum assembly length (unambiguous) to run vadr vadr vadr_opts String Options for the v-annotate.pl VADR script vADR script vADR script vanitariant signal of the va	vacar	maxicii	THC		30000
vadr minlen Int Minimum length subsequence to possibly replace Ns for the fastatrim-terminal-ambigs.pl VADR script vadr skip_length Int Minimum assembly length (unambiguous) to run vadr vadr vadr_opts String Options for the v-annotate.pl VADR script vADR script vADR script vannotate.pl VADR script vannotate.pl VADR script vannotate.pl VADR script core,fstukenf,insertnn,deletinn—mdir /opt/vadr/vadr-models/ variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before				-	
vadr minlen Int Minimum length subsequence to possibly replace Ns for the fastatrim-terminal-ambigs.pl VADR script vadr skip_length Int Minimum assembly length (unambiguous) to run vadr vadr vadr_opts String Options for the v-annotate.pl VADR script variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before VADR script 50 Valdr vadr_opts String Options for the v-annotate.pl core,fistukenf,insertun,deletinn -mdir /opt/vadr/vadr-models/ TRUE					
vadr minlen Int Minimum length subsequence to possibly replace Ns for the fastatrim-terminal-ambigs.pl VADR script vadr skip_length Int Minimum assembly length (unambiguous) to run vadr vadr vadr_opts String Options for the v-annotate.pl VADR script variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before VADR script 50 Valdr vadr_opts String Options for the v-annotate.pl core,fistukenf,insertun,deletinn -mdir /opt/vadr/vadr-models/ TRUE				ambigs.pl	
vadr minlen Int Minimum length subsequence to possibly replace Ns for the fastatrim-terminal ambigs.pl vADR script vadr skip_length Int Minimum assembly length (unambiguous) to run vadr vadr vadr_opts String Options for the v-annotate.pl VADR script VADR script variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before Minimum assembly length (unambiguous) to run vadr Options for the v-annotate.pl —glsearch -s -r —nomisc —mkey sarscov2 —alt_fail lowscore, fstukenf, insertun, deletinn —mdir /opt/vadr/vadr-models/ TRUE					
sequence to possibly replace Ns for the fastatrim-terminal-ambigs.pl VADR script vadr skip_length Int Minimum assembly length (unambiguous) to run vadr vadr vadr_opts String Options for the v-annotate.pl VADR script	vadr	minlen	Int		50
vadr skip_length Int Minimum assembly length (unambiguous) to run vadr vadr opts String Options for the v-annotate.pl VADR script vADR script vADR script vantat_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before 10000 possibly replace Ns for the fastatrim-trim-terminal-ambigs.pl vADR script count_orphans are possibly replace. Ns for the fastatrim-trim-terminal-ambigs.pl vADR script count_orphans are possibly replace. Ns for the fastatrim-trim-terminal-ambigs.pl vADR script count_orphans are possibly replace. Ns for the fastatrim-trim-terminal-ambigs.pl vADR script count_orphans are possibly replace. Ns for the fastatrim-terminal-ambigs.pl vADR script count_orphans are possible replace. Ns for the fastatrim-terminal-ambigs.pl vADR script count_orphans are possible replace. Ns for the fastatrim-terminal-ambigs.pl vADR script count_orphans are possible replace. Ns for the fastatrim-terminal-ambigs.pl vADR script count_orphans are possible replace. Ns for the fastatrim-terminal-ambigs.pl vADR script count_orphans are possible replace. No possible replace in the fastatrim-terminal-ambigs.pl vADR script count_orphans are possible replace. No possible replace in the fastatrim-terminal-ambigs.pl vADR script count_orphans are possible replace. No possible replace in the fastatrim-terminal-ambigs.pl vADR script count_orphans are possible replace. The fastatrim-terminal-ambigs.pl vADR script count_orphans are				length sub-	
vadr skip_length				sequence to	
trim-terminal- ambigs.pl VADR script vadr skip_length Int Minimum as- sembly length (unambiguous) to run vadr vadr vadr_opts String Options for the v-annotate.pl VADR script —glsearch -s -r -nomisc -mkey sarscov2 -alt_fail lows- core,fstukcnf,insertnn,deletinn -mdir /opt/vadr/vadr-models/ variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before					
vadr skip_length Int Minimum assembly length (unambiguous) to run vadr vadr vadr_opts String Options for the v-annotate.pl VADR script core,fstukenf,insertnn,deletinn -mdir /opt/vadr/vadr-models/ variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before 100000					
vadr skip_length					
vadr skip_length Int Minimum assembly length (unambiguous) to run vadr vadr vadr_opts String Options for the v-annotate.pl VADR script -glsearch -s -r -nomisc -mkey sarscov2 -alt_fail lowscore,fstukcnf,insertnn,deletinn -mdir /opt/vadr/vadr-models/ variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before					
sembly length (unambiguous) to run vadr vadr vadr_opts String Options for the v-annotate.pl VADR script Core,fstukcnf,insertnn,deletinn —mdir /opt/vadr/vadr-models/ variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before					
vadr vadr_opts String Options for the v-annotate.pl VADR script Core,fstukcnf,insertnn,deletinn —mdir /opt/vadr/vadr-models/ variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before	vadr	skip_length	Int		10000
vadr vadr_opts String Options for the v-annotate.pl VADR script core,fstukcnf,insertnn,deletinn —mdir /opt/vadr/vadr-models/ variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before				, .	
vadr vadr_opts String Options for the v-annotate.pl -glsearch -s -r -nomisc -mkey sarscov2 -alt_fail lows-core,fstukcnf,insertnn,deletinn -mdir /opt/vadr/vadr-models/ variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before					
v-annotate.pl VADR script variant_call count_orphans boolean v-annotate.pl VADR script core,fstukcnf,insertnn,deletinn -mdir /opt/vadr/vadr-models/ TRUE TRUE TRUE	vodr	vadr onto	String		alcagrah s r namica
variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before Core,fstukcnf,insertnn,deletinnmdir /opt/vadr/vadr-models/ TRUE TRUE	vaui	vaur_opts	Sumg	_	
variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before —mdir /opt/vadr/vadr-models/ TRUE					
variant_call count_orphans Boolean Do not skip anomalous read pairs in variant calling for SAMtools mpileup before				, I IDIC script	
anomalous read pairs in variant calling for SAMtools mpileup before	variant call	count orphans	Boolean	Do not skip	
read pairs in variant calling for SAMtools mpileup before					
variant calling for SAMtools mpileup before					
for SAMtools mpileup before					
running iVar				mpileup before	
				running iVar	
variants				variants	

Table 3 – continued from previous page

Task	Variable Name		Description	Default	I
variant_call	disable_baq	Boolean	Disable read-	TRUE	I
_			pair overlap		I
			detection for		I
			SAMtools		I
			mpileup before		I
			running iVar		
			variants		
variant_call	max_depth	Int	Maximum reads	600000	I
	-		read at a posi-		
			tion per input		I
			file for SAM-		
			tools mpileup		
			before running		I
			iVar variants		
variant_call	min_bq	Int	Minimum map-	0	
			ping quality for		I
			an alignment		
			to be used		
			for SAMtools		
			mpileup before		
			running iVar		
			variants		1
variant_call	min_depth	Int	Minimum read	100	
			depth to call		
			variants for iVar		
11	<u> </u>	-	variants		I
variant_call	min_freq	Float	Minimum	0.6	
			frequency		
			threshold(0 - 1)		
			to call variants		
:	1	T4	for iVar variants	20	
variant_call	min_qual	Int	Minimum qual-	20	
			ity threshold for		
			sliding window to pass for iVar		
			variants		
variant_call	ref_gff	String	Path to the	/refer-	
variant_can	ici_gii	Sumg	general fea-	ence/GCF_009858895.2_ASM98588	0v3 genomic aff
			ture format	CIIC/GCI _009838893.2_A3W198388	genomic.gn
			of the refer-		
			ence genome		
			within the		
			staphb/ivar:1.2.2_	artic20200528	
			Docker con-		
			tainer		
variant_call	ref_genome	String	Path to the ref-	/artic-	
			erence genome	ncov2019/primer_schemes/nCoV-	
			within the	2019/V3/nCoV-	
				ar 2029.200528 ce.fasta	
			Docker con-]	
			tainer		
				continues on next page	

Table 3 – continued from previous page

Task	Variable Name	Data Type	Description	Default
version_capture	timezone	String	User time	None
			zone in valid	
			Unix TZ string	
			(e.g. Amer-	
			ica/New_York)	

Outputs

Download CSV: TheiaCoV_Illumina_SE_default_outputs.csv

Output Name	Data Type	Description
aligned_bai	File	Index companion file to the bam file generated during the consensus
		assembly process
aligned_bam	File	Primer-trimmed BAM file; generated during conensus assembly
		process
assembly_fasta	File	Consensus genome assembly
assem-	Int	Number of unambiguous basecalls within the SC2 consensus assem-
bly_length_unambigu	ous	bly
assem-	Float	Mean sequencing depth throughout the conesnsus assembly gener-
bly_mean_coverage		ated after performing primer trimming-calculated using the SAM-
		tools coverage command
assembly_method	String	Method employed to generate consensus assembly
auspice_json	File	Auspice-compatable JSON output generated from NextClade analy-
		sis that includes the NextClade default samples for clade-typing and
		the single sample placed on this tree
bbduk_docker	String	Docker image used to run BBDuk
bwa_version	String	Version of BWA used to map read data to the reference genome
consensus_flagstat	File	Output from the SAMtools flagstat command to assess quality of the
		alignment file (BAM)
consensus_stats	File	Output from the SAMtools stats command to assess quality of the
		alignment file (BAM)
fastqc_clean	Int	Number of reads after SeqyClean filtering as determined by FastQC
fastqc_raw	Int	Number of reads after sequelean filtering as determined by FastQC
fastqc_version	String	Version of the FastQC software used for read QC analysis
ivar_tsv	File	Variant descriptor file generated by iVar variants
ivar_variant_version	String	Version of iVar for running the iVar variants command
ivar_vcf	File	iVar tsv output converted to VCF format
ivar_version_consens	usString	Version of iVar for running the iVar consensus command
ivar_version_primtrin	n String	Version of iVar for running the iVar trim command
kraken_human	Float	Percent of human read data detected using the Kraken2 software
kraken_report	String	Full Kraken report
kraken_sc2	Float	Percent of SARS-CoV-2 read data detected using the Kraken2 soft-
		ware
kraken_version	String	Version of Kraken software used

Table 4 – continued from previous page

Output Name	Data Type	Description		
meanbaseq_trim	Float	Mean quality of the nucleotide basecalls aligned to the reference		
1—		genome after primer trimming		
meanmapq_trim	Float	Mean quality of the mapped reads to the reference genome after		
		primer trimming		
nextclade_aa_dels	String	Amino-acid deletions as detected by NextClade		
nextclade_aa_subs	String	Amino-acid substitutions as detected by NextClade		
nextclade_clade	String	NextClade clade designation		
nextclade_json	File	NexClade output in JSON file format		
nextclade_tsv	File	NextClade output in TSV file format		
nextclade_version	String	Version of NextClade software used		
number_Degenerate	Int	Number of degenerate basecalls within the consensus assembly		
number_N	Int	Number of fully ambiguous basecalls within the consensus assembly		
number_Total	Int	Total number of nucleotides within the consensus assembly		
pango_lineage	String	Pango lineage as detremined by Pangolin		
pango_lineage_report	File	Full Pango lineage report generated by Pangolin		
pan-	String	Version of the pangolin software (e.g. PANGO or PUSHER) used		
golin_assignment_ver	sion	for lineage asignment		
pangolin_conflicts	String	Number of lineage conflicts as deteremed by Pangolin		
pangolin_docker	String	Docker image used to run Pangolin		
pangolin_notes	String	Lineage notes as deteremined by Pangolin		
pangolin_versions	String	All Pangolin software and database version		
per-	Float	Percent coverage of the reference genome after performing primer		
cent_reference_covera	ige	trimming; calculated as assembly_length_unambiguous / length of		
		reference genome (SC2: 29,903) x 100		
primer_bed_name	String	Name of the primer bed files used for primer trimming		
primer_trimmed_read	_ peoae nt	Percent of read data with primers trimmed as deteremined by iVar trim		
read1_clean	File	Forward read file after quality trimming and adapter removal		
samtools_version	String	Version of SAMtools used to sort and index the alignment file		
sam-	String	Version of SAMtools used to create the pileup before running iVar		
tools_version_consens	_	consensus		
sam-	String	Version of SAMtools used to create the pileup before running iVar		
tools_version_primtri	_	trim		
sam-	String	Version of SAMtools used to assess quality of read mapping		
tools_version_stats	-			
seq_platform	String	Description of the sequencing methodology used to generate the in-		
		put read data		
theia-	String	Date of analysis		
cov_illumina_se_anal				
theia-	String	Version of the Public Health Viral Genomics (PHVG) repository		
cov_illumina_se_vers		used		
trimmo-	String	Version of Trimmomatic used		
matic_version				
vadr_alerts_list	File	File containing all of the fatal alerts as determined by VADR		
vadr_docker	String	Docker image used to run VADR		
vadr_num_alerts	String	Number of fatal alerts as determined by VADR		

TheiaCoV ClearLabs

The TheiaCoV_ClearLabs workflow was written to process ClearLabs WGS read data for SARS-CoV-2 amplicon sequencing. Currently, Clear Labs sequencing is performed with the Artic V3 protocol. If alternative primer schemes such as the Qiaseq Primer Panel, the Swift Amplicon SARS-CoV-2 Panel and the Artic V4 Amplicon Sequencing Panel become available on the platform, these data can can also be analysed with this workflow since the primer sequence coordinates of the PCR scheme utilized must be provided along with the raw Clear Labs read data must be provided in BED and FASTQ file formats, respectively.

Upon initiating a TheiaCoV_ClearLabs run, input ClearLabs read data provided for each sample will be processed to perform consensus genome assembly, infer the quality of both raw read data and the generated consensus genome, and assign SARS-CoV-2 lineage and clade types as outlined in the TheiaCoV_ClearLabs data workflow below.

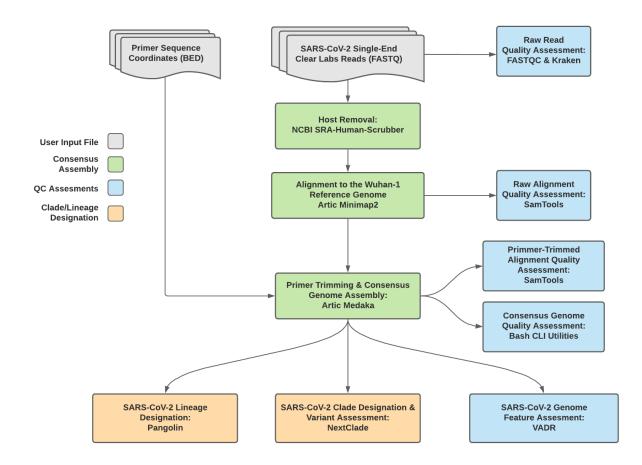


Fig. 3: TheiaCoV_ClearLabs Data Workflow

Consensus genome assembly with the TheiaCoV_ClearLabs workflow is performed by first de-hosting read data with the NCBI SRA-Human-Scrubber tool then following the *Artic nCoV-2019 novel coronavirs bioinformatics protocol* https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html. Briefly, input reads are aligned to the Wuhan-1 reference genome with minimap2 to generate a Binary Alignment Mapping (BAM) file. Primer sequences are then removed from the BAM file and a consensus assembly file is generated using the Artic medaka command. This assembly is then used to assign lineage and clade designations with Pangolin and NextClade. NCBI'S VADR tool is also employed to screen for potentially errant features (e.g. erroneous frame-shift mutations) in the consensus assembly.

Note: Read-trimming is performed on raw read data generated on the ClearLabs instrument and thus not a required

step in the Theia CoV_Clear Labs work flow.

More information on required user inputs, optional user inputs, default tool parameters and the outputs generated by TheiaCoV_CLearLabs are outlined below.

Required User Inputs

 $Download\ CSV: The \verb"iaCoV_ClearLabs_required_inputs.csv"$

Task	Input Variable	Data Type	Description
theiacov_clearlabs	clear_lab_fastq	File	Clear Labs FASTQ read files
theiacov_clearlabs	primer_bed	File	Primer sequence coordinates of the PCR scheme utilized in BED file format
theiacov_clearlabs	samplename	String	Name of the sample being analyzed

Optional User Inputs

Download CSV: TheiaCoV_ClearLabs_optional_inputs.csv

Task	Variable Name	Data Type	Description	Default
consensus	cpu	Int	CPU resources	8
			allocated to the	
			Artric Medaka	
			task runtime	
			environment	
consensus	docker	String	Docker tag	quay.io/staphb/artic-
			used for run-	ncov2019:1.3.0-medaka-1.4.3
			ning Medaka	
			assemblyer	
consensus	medaka_model	String	Model for con-	r941_min_high_g360
			sensus genome	
			assembly via	
			Medaka	
fastqc_se_clean	cpus	Int	CPU resources	
			allocated to	
			the FastQC	
			task runtime	
			environment for	
			asessing clean	
			read data	
fastqc_se_clean	read1_name	String	Name of the	Inferred from the input read file-
			sample being	fastqc_se_clean
			analyzed	agatinuag an navt naga

Table 5 – continued from previous page

Task	Variable Name	Data Type	Description	Default
fastqc_se_raw	cpus	Int	CPU resources	
	_		allocated to	
			the FastQC	
			task runtime	
			environment for	
			asessing raw	
			read data	
fastqc_se_raw	read1_name	String	Name of the	Inferred from the input read file
1	_		sample being	1
			analyzed	
kraken2_dehosted	cnus	Int	CPU resources	4
	- CP us	1111	allocated to	·
			the Kraken	
			task runtime	
			environment	
			for asessing	
			dehosted read	
			data	
kraken2_dehosted	Izralzan) dh	String	Path to the ref-	/kraken2-db
krakenz_denosted	Krakenz_do	Sumg		/Krakenz-do
			erence genome within the	
				0.0
			staphb/kraken2:2.	0.8-
			beta_hv Docker	
			container	
kraken2_dehosted	read2	File	Optional input	None
			file for the	
			Kraken task that	
			is not applicable	
			to this workflow	
kraken2_raw	cpus	Int	CPU resources	4
			allocated to	
			the Kraken	
			task runtime	
			environment for	
			asessing raw	
			read data	
kraken2_raw	kraken2_db	String	Path to the ref-	/kraken2-db
			erence genome	
			within the	
			staphb/kraken2:2.	0.8-
			beta_hv Docker	
			container	
kraken2_raw	read2	File	Optional input	None
_			file for the	
			Kraken task that	
			is not applicable	
			to this workflow	
			to this workhow	

22

Table 5 – continued from previous page

Task	Variable Name	Data Type	Description	Default	
ncbi_scrub_se	docker	Docker tag used	gcr.io/ncbi-		
		for running the	sys-gcr-public-		
		NCBI SRA	research/sra-		
		Human-Scruber	human-		
		tool	scrubber@sha256	:b7dba71079344daea4ea3363e1a67fa54edb7ec6	55459d0
nextclade_one_sar	m øbe ker	String	Docker tag used	nextstrain/nextclade:1.10.3	
		 	for running		
		l	NextClade		
nextclade_output_	pdrxdreione_sample	String	Docker tag	python:slim	
		 	used for pars-		
		 	ing NextClade		
			output		
pangolin3	docker	String	Docker tag used	quay.io/staphb/3.1.20-pangolearn-	
		 	for running Pan-	2022-02-02	
			golin		
pangolin3	infer-	String	pangolin infer-	usher	
	ence_engine	 	ence engine for		
		 	lineage designa-		
		 	tions (usher or		
11. 2			pangolarn)	10000	
pangolin3	min_length	Int	Minimum query	10000	
	ı	l I	length allowed		
	ı	I	for pangolin	1	
	ı	l I	to attempt		
2	···hia	T214	assignment Maximum pro	0.5	
pangolin3	max_ambig	Float	Maximum pro-	0.5	
	ı	l I	portion of Ns allowed for pan-		
	ı	l I	golin to attempt		
	ı	l I	assignment		
theia-	nextclade_dataset_	nString	Nextclade or-	sars-cov-2	
cov_clearlabs	Ilexterade_databet	_11911111115	ganism dataset	5813-667 2	
theia-	nextclade_dataset_	rafeinence	Nextclade refer-	MN908947	
cov_clearlabs		- 1012228 22	ence genome	I I I I I I I I I I I I I I I I I I I	
theia-	nextclade_dataset_	tNextclade	2022-02-		
cov_clearlabs		dataset tag	07T12:00:00Z		
theia-	normalise	Int	Value to nor-	200	
cov_clearlabs		 	malize read		
		 	counts		
theia-	seq_method	String	Description of	ONT via Clear Labs WGS	
cov_clearlabs		 	the sequencing		
		 	methodology		
		 	used to generate		
		 	the input read		
		·	data		
vadr	docker	String	Docker tag used	quay.io/staphb/1.4.1-models-1.3-2	
		 	for running		
		, I	VADR		

Table 5 – continued from previous page

Task	Variable Name	Data Type	Description	Default
vadr	maxlen	Int	Maximum	30000
			length for the	
			fasta-trim-	
			terminal-	
			ambigs.pl	
			VADR script	
vadr	minlen	Int	Minimum	50
			length sub-	
			sequence to	
			possibly replace	
			Ns for the fasta-	
			trim-terminal-	
			ambigs.pl	
			VADR script	
vadr	skip_length	Int	Minimum as-	10000
			sembly length	
			(unambiguous)	
			to run vadr	
vadr	vadr_opts	String	Options for the	-glsearch -s -r -nomisc
			v-annotate.pl	-mkey sarscov2 -alt_fail lows-
			VADR script	core,fstukcnf,insertnn,deletinn
				-mdir /opt/vadr/vadr-models/
version_capture	timezone	String	User time	None
			zone in valid	
			Unix TZ string	
			(e.g. Amer-	
			ica/New_York)	

Outputs

 $Download\ CSV: The \verb"iaCoV_ClearLabs_default_outputs.csv"$

Output Name	Data Type	Description
aligned_bai	File	Index companion file to the bam file generated during the consensus
		assembly process
aligned_bam	File	Primer-trimmed BAM file; generated during conensus assembly
		process
artic_version	String	Version of the Artic software utilized for read trimming and
		conesnsus genome assembly
assembly_fasta	File	Consensus genome assembly
assem-	Int	Number of unambiguous basecalls within the SC2 consensus assem-
bly_length_unambigu	ous	bly
assem-	Float	Mean sequencing depth throughout the conesnsus assembly gener-
bly_mean_coverage		ated after performing primer trimming-calculated using the SAM-
		tools coverage command
assembly_method	String	Method employed to generate consensus assembly

Table 6 – continued from previous page

Output Name	Data Type	- continued from previous page Description
auspice_json	File	Auspice-compatable JSON output generated from NextClade analy-
auspice_json	THE	sis that includes the NextClade default samples for clade-typing and
		the single sample placed on this tree
consensus_flagstat	File	Output from the SAMtools flagstat command to assess quality of the
consensus_nagstat	THE	alignment file (BAM)
consensus stats	File	Output from the SAMtools stats command to assess quality of the
consensus_stats	THE	alignment file (BAM)
dehosted reads	File	Dehosted reads; suggested read file for SRA submission
fastqc_clean	Int	Number of reads after dehosting as determined by FastQC
fastqc_raw	Int	Number of raw input reads as determined by FastQC
		Version of the FastQC version used
fastqc_version	String Float	
kraken_human		Percent of human read data detected using the Kraken2 software
kraken_human_dehos	temoat	Percent of human read data detected using the Kraken2 software af-
11	Ct	ter host removal
kraken_report	String	Full Kraken report
kraken_report_dehost		Full Kraken report after host removal
kraken_sc2	Float	Percent of SARS-CoV-2 read data detected using the Kraken2 soft-
1 1 2 11 1	T1 4	ware
kraken_sc2_dehosted	Float	Percent of SARS-CoV-2 read data detected using the Kraken2 soft-
	G	ware after host removal
kraken_version	String	Version of Kraken software used
meanbaseq_trim	Float	Mean quality of the nucleotide basecalls aligned to the reference
		genome after primer trimming
meanmapq_trim	Float	Mean quality of the mapped reads to the reference genome after
		primer trimming
nextclade_aa_dels	String	Amino-acid deletions as detected by NextClade
nextclade_aa_subs	String	Amino-acid substitutions as detected by NextClade
nextclade_clade	String	NextClade clade designation
nextclade_json	File	NexClade output in JSON file format
nextclade_tsv	File	NextClade output in TSV file format
nextclade_version	String	Version of NextClade software used
number_Degenerate	Int	Number of degenerate basecalls within the consensus assembly
number_N	Int	Number of fully ambiguous basecalls within the consensus assembly
number_Total	Int	Total number of nucleotides within the consensus assembly
pango_lineage	String	Pango lineage as detremined by Pangolin
pango_lineage_report	File	Full Pango lineage report generated by Pangolin
pan-	String	Version of the pangolin software (e.g. PANGO or PUSHER) used
golin_assignment_ver	sion	for lineage asignment
pangolin_conflicts	String	Number of lineage conflicts as deteremed by Pangolin
pangolin_docker	String	Docker image used to run Pangolin
pangolin_notes	String	Lineage notes as deteremined by Pangolin
pangolin_versions	String	All Pangolin software and database versions
per-	Float	Percent coverage of the reference genome after performing primer
cent_reference_covera		trimming; calculated as assembly_length_unambiguous / length of
		reference genome (SC2: 29,903) x 100
primer_bed_name	String	Name of the primer bed files used for primer trimming
reads_dehosted	File	De-hosted read files
samtools_version	String	Version of SAMtools used to sort and index the alignment file
santtools_version seq_platform	String	Description of the sequencing methodology used to generate the in-
boq_pianoriii	Sams	put read data
<u> </u>		put read data

Table 6 – continued from previous page	Table	6 - continued	from	previous page	,
--	-------	---------------	------	---------------	---

Output Name	Data Type	Description
theia-	String	Date of analysis
cov_clearlabs_analysi	s_date	
theia-	String	Version of the Public Health Viral Genomics (PHVG) repository
cov_clearlabs_version		used
vadr_alerts_list	File	File containing all of the fatal alerts as determined by VADR
vadr_docker	String	Docker image used to run VADR
vadr_num_alerts	String	Number of fatal alerts as determined by VADR
vari-	File	Number of variants relative to the reference genome
ants_from_ref_vcf		

TheiaCoV_ONT

The TheiaCoV_ONT workflow was written to process basecalled and demultiplexed Oxford Nanopore Technology (ONT) read data. The most common read data analyzed by the TheiaCoV_ONT workflow are generated with the Artic V3 protocol. Alternative primer schemes such as the Qiaseq Primer Panel, the Swift Amplicon SARS-CoV-2 Panel and the Artic V4 Amplicon Sequencing Panel however, can also be analysed with this workflow since the primer sequence coordinates of the PCR scheme utilized must be provided along with the raw paired-end Illumina read data in BED and FASTQ file formats, respectively.

Upon initiating a TheiaCoV_ONT run, input ONT read data provided for each sample will be processed to perform consensus genome assembly, infer the quality of both raw read data and the generated consensus genome, and assign SARS-CoV-2 lineage and clade types as outlined in the TheiaCoV_ONT data workflow below.

Consensus genome assembly with the TheiaCoV_ONT workflow is performed performed by first de-hosting read data with the NCBI SRA-Human-Scrubber tool then following then following *Artic nCoV-2019 novel coronavirs bioinformatics protocol* https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html. Briefly, input reads are filtered by size (min-length: 400bp; max-length: 700bp) with the Arite guppyplex command. These size-selected read data are aligned to the Wuhan-1 reference genome with minimap2 to generate a Binary Alignment Mapping (BAM) file. Primer sequences are then removed from the BAM file and a consensus assembly file is generated using the Artic medaka command. This assembly is then used to assign lineage and clade designations with Pangolin and NextClade. NCBI'S VADR tool is also employed to screen for potentially errant features (e.g. erroneous frame-shift mutations) in the consensus assembly.

More information on required user inputs, optional user inputs, default tool parameters and the outputs generated by TheiaCoV_ONT are outlined below.

Required User Inputs

 $Download \ CSV: \ Theia CoV_ONT_required_inputs.csv$

Task	Input Variable	Data Type	Description
theiacov_ont	demulti-	File	Basecalled and demultiplexed ONT read
	plexed_reads		data (single FASTQ file per sample)
theiacov_ont	primer_bed	File	Primer sequence coordinates of the PCR
			scheme utilized in BED file format
theiacov_ont	samplename	String	Name of the sample being analyzed

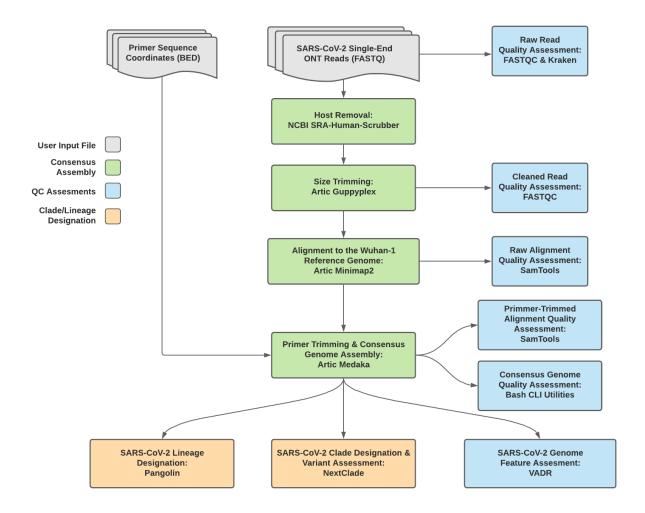


Fig. 4: TheiaCoV_ONT Data Workflow

Optional User Inputs

Download CSV: TheiaCoV_ONT_optional_inputs.csv

Task	Variable Name	Data Type	Description	Default
consensus	сри	Int	CPU resources allocated to the Artric Medaka task runtime environment	
consensus	docker	String	Docker tag used for run- ning Medaka assemblyer	quay.io/staphb/artic-ncov2019-epi2me
consensus	medaka_model	String	Model for con- sensus genome assembly via Medaka	r941_min_high_g360
fastqc_se_clean	cpus	Int	CPU resources allocated to the FastQC task runtime environment for assessing size-selected read data	2
fastqc_se_clean	read1_name	String	Name of the sample being analyzed	Inferred from the input read file
fastqc_se_raw	cpus	Int	CPU resources allocated to the FastQC task runtime environment for asessing raw read data	
fastqc_se_raw	read1_name	String	Name of the sample being analyzed	Inferred from the input read file
kraken2_dehosted	cpus	Int	CPU resources allocated to the Kraken task runtime environment for asessing dehosted read data	4

Table 7 – continued from previous page

Task	Variable Name	Data Type	Description	Default
kraken2_dehosted		String	Path to the ref-	/kraken2-db
Kitakon2_con	Kithon2_55	Julia	erence genome	/Riakenz do
		ı I	within the	
	J	l I	staphb/kraken2:2.	h 8-
		ı I	beta_hv Docker).0-
	J	l I	container	
kraken2_dehosted	read?	File	Optional input	None
Klakenz_denosiec	Teau2	Line	file for the	None
	J	l I	Kraken task that	
	J	ı I	is not applicable	
		ı I	to this workflow	
kraken2_raw	cpus	Int	CPU resources	4
Klakenz_raw	cpus	, IIII.	allocated to	4
		ı I	the Kraken	
	J	ı I	task runtime	
		ı I	environment for	
	J	ı I	asessing raw	
		ı I	read data	
kraken2_raw	kraken2_db	Ctin-a	Path to the ref-	/kraken2-db
Kraken2_raw	Kraken∠_uo	String		/kraken2-ab
	J	l I	erence genome within the	
	J	ı I		h.
		ı I	staphb/kraken2:2.	J.8-
		ı I	beta_hv Docker	
		<u> </u>	container	
kraken2_raw	read2	File	Optional input	None
	J	ı I	file for the	
	J	ı I	Kraken task that	
	J	ı I	is not applicable	
		<u> </u>	to this workflow	<u> </u>
ncbi_scrub_se	docker	Docker tag used	gcr.io/ncbi-	
	J	for running the	sys-gcr-public-	
	J	NCBI SRA	research/sra-	
		Human-Scruber	human-	
		tool		b7dba71079344daea4ea3363e1a67fa54edb7ec65459d
nextclade_one_san	n pb eker	String	Docker tag used	nextstrain/nextclade:1.10.3
		ı I	for running	
		<u> </u>	NextClade	
nextclade_output_j	pdrxdre_rone_sample	String	Docker tag	python:slim
	J	ı I	used for pars-	
	J	l I	ing NextClade	
		<u></u> !	output	
pangolin3	docker	String	Docker tag used	quay.io/staphb/3.1.20-pangolearn-
		ı I	for running Pan-	2022-02-02
		ı <u></u> !	golin	
pangolin3	infer-	String	pangolin infer-	usher
	ence_engine	l I	ence engine for	
		l I	lineage designa-	
	I	1	tions (usher or	
	1	· .	pangolarn)	

Table 7 – continued from previous page

Task	Variable Name	Data Type	Description	Default
pangolin3	min_length	Int	Minimum query	10000
1 0			length allowed	
			for pangolin	
			to attempt	
			assignment	
pangolin3	max_ambig	Float	Maximum pro-	0.5
pangomis	max_amoig	Tioat	portion of Ns al-	0.5
			lowed for pan-	
			golin to attempt	
1 C1(T4	assignment	0
read_filtering	cpu	Int	CPU resources	8
			allocated to the	
			read filtering	
			task (Artic gup-	
			pypled) runtime	
			environment	
read_filtering	max_length	Int	Maximum	700
			sequence length	
read_filtering	min_length	Int	Minimum	400
			sequence length	
read_filtering	run_prefix	String	Run name	artic_ncov2019
theiacov_ont	nextclade_dataset		Nextclade or-	sars-cov-2
_			ganism dataset	
theiacov_ont	nextclade_dataset	r Sfeine gice	Nextclade refer-	MN908947
_			ence genome	
theiacov_ont	nextclade_dataset	tMextclade	2022-02-	
_		dataset tag	07T12:00:00Z	
theiacov_ont	ar-	String	Version of the	V3
	tic_primer_version	_	Artic PCR	
	_rrr		protocol used to	
			generate input	
			read data	
theiacov_ont	normalise	Int	Value to nor-	200
theracov_ont	normanse	l IIIt	malize read	200
			counts	
theircov ont	seq_method	String	Description of	ONT
theiacov_ont	seq_memou	Sumg		OITI
			the sequencing	
			methodology	
			used to generate	
			the input read	
			data	
theiacov_ont	pan-	String	Docker tag used	staphb/pangolin:2.4.2-pangolearn-
	golin_docker_ima	ge	for running Pan-	2021-05-19
			golin	
vadr	docker	String	Docker tag used	quay.io/staphb/1.4.1-models-1.3-2
			for running	
			VADR	
				•

30

Table 7 – continued from previous page

Task	Variable Name	Data Type	Description	Default
vadr	maxlen	Int	Maximum	30000
			length for the	
			fasta-trim-	
			terminal-	
			ambigs.pl	
			VADR script	
vadr	minlen	Int	Minimum	50
			length sub-	
			sequence to	
			possibly replace	
			Ns for the fasta-	
			trim-terminal-	
			ambigs.pl	
			VADR script	
vadr	vadr_opts	String	Options for the	-glsearch -s -r -nomisc
			v-annotate.pl	-mkey sarscov2 -alt_fail lows-
			VADR script	core,fstukcnf,insertnn,deletinn
				-mdir /opt/vadr/vadr-models/
vadr	skip_length	Int	Minimum as-	10000
			sembly length	
			(unambiguous)	
			to run vadr	
version_capture	timezone	String	User time	None
			zone in valid	
			Unix TZ string	
			(e.g. Amer-	
			ica/New_York)	

Outputs

Download CSV: TheiaCoV_ONT_default_outputs.csv

Output Name	Data Type	Description
aligned_bai	File	Index companion file to the bam file generated during the consensus
		assembly process
aligned_bam	File	Primer-trimmed BAM file; generated during conensus assembly
		process
amp_coverage	File	Sequence coverage per amplicon
artic_version	String	Version of the Artic software utilized for read trimming and
		conesnsus genome assembly
assembly_fasta	File	Consensus genome assembly
assem-	Int	Number of unambiguous basecalls within the SC2 consensus assem-
bly_length_unambigu	ous	bly
assem-	Float	Mean sequencing depth throughout the conesnsus assembly gener-
bly_mean_coverage		ated after performing primer trimming-calculated using the SAM-
		tools coverage command

Table 8 – continued from previous page

Output Name	Data Type	Description
assembly_method	String	Method employed to generate consensus assembly
auspice_json	File	Auspice-compatable JSON output generated from NextClade analy-
auspice_json	riie	
		sis that includes the NextClade default samples for clade-typing and
1 1 1	Ct	the single sample placed on this tree
bedtools_version	String	bedtools version utilized when calculating amplicon read coverage
consensus_flagstat	File	Output from the SAMtools flagstat command to assess quality of the alignment file (BAM)
consensus_stats	File	Output from the SAMtools stats command to assess quality of the alignment file (BAM)
dehosted_reads	File	Dehosted reads; suggested read file for SRA submission
fastqc_clean	Int	Number of reads after size filtering and dehosting as determined by FastQC
fastqc_raw	Int	Number of raw reads input reads as determined by FastQC
fastqc_version	String	Version of the FastQC version used
kraken_human	Float	Percent of human read data detected using the Kraken2 software
kraken_human_dehos	te H loat	Percent of human read data detected using the Kraken2 software af-
		ter host removal
kraken_report	File	Full Kraken report
kraken_report_dehost	edFile	Full Kraken report after host removal
kraken_sc2	Float	Percent of SARS-CoV-2 read data detected using the Kraken2 soft-
		ware
kraken_sc2_dehosted	Float	Percent of SARS-CoV-2 read data detected using the Kraken2 soft-
		ware after host removal
kraken_version	String	Version of Kraken software used
meanbaseq_trim	Float	Mean quality of the nucleotide basecalls aligned to the reference
meanouseq_trim	11040	genome after primer trimming
meanmapq_trim	Float	Mean quality of the mapped reads to the reference genome after
meannapq_am	11040	primer trimming
nextclade_aa_dels	String	Amino-acid deletions as detected by NextClade
nextclade_aa_subs	String	Amino-acid substitutions as detected by NextClade
nextclade_clade	String	NextClade clade designation
nextclade_json	File	NexClade output in JSON file format
nextclade tsv	File	NextClade output in TSV file format
nextclade_version	String	Version of NextClade software used
number_Degenerate		
	Int	Number of degenerate basecalls within the consensus assembly
number_N	Int	Number of fully ambiguous basecalls within the consensus assembly
number_Total	Int	Total number of nucleotides within the consensus assembly
pango_lineage	String	Pango lineage as detremined by Pangolin
pango_lineage_report		Full Pango lineage report generated by Pangolin
pan-	String	Version of the pangolin software (e.g. PANGO or PUSHER) used
golin_assignment_ver		for lineage asignment
pangolin_conflicts	String	Number of lineage conflicts as deteremed by Pangolin
pangolin_docker	String	Docker image used to run Pangolin
pangolin_notes	String	Lineage notes as deteremined by Pangolin
pangolin_versions	String	All Pangolin software and database versions
per-	Float	Percent coverage of the reference genome after performing primer
cent_reference_covera	ige	trimming; calculated as assembly_length_unambiguous / length of
		reference genome (SC2: 29,903) x 100
primer_bed_name	String	Name of the primer bed files used for primer trimming
pangolin_versions	String	All Pangolin software and database versions

		oonmisses nom provides page
Output Name	Data Type	Description
reads_dehosted	File	De-hosted read files
samtools_version	String	Version of SAMtools used to sort and index the alignment file
seq_platform	String	Description of the sequencing methodology used to generate the in-
		put read data
theia-	String	Date of analysis
cov_ont_analysis_date		
theia-	String	Version of the Public Health Viral Genomics (PHVG) repository
cov_ont_version		used
vadr_alerts_list	File	File containing all of the fatal alerts as determined by VADR
vadr_docker	String	Docker image used to run VADR
vadr_num_alerts	String	Number of fatal alerts as determined by VADR
vari-	File	Number of variants relative to the reference genome
ants_from_ref_vcf		

Table 8 – continued from previous page

TheiaCoV FASTA

The TheiaCoV_FASTA workflow was written to process SARS-CoV-2 assembly files to infer the quality of the input assembly and assign SARS-CoV-2 lineage and clade types as outlined in the TheiaCoV_FASTA data workflow below.

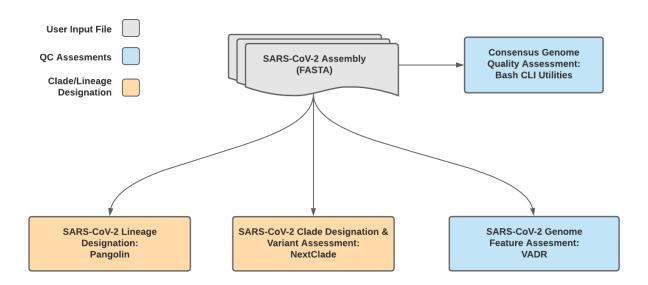


Fig. 5: TheiaCoV_FASTA Data Workflow

The quality of input SARS-CoV-2 genome assemblies are assessed by the TheiaCoV_FASTA workflow using a series of bash shell scripts. Input assemblies are then used to assign lineage and clade designations with Pangolin and NextClade. NCBI'S VADR tool is also employed to screen for potentially errant features (e.g. erroneous frame-shift mutations) in the consensus assembly.

More information on required user inputs, optional user inputs, default tool parameters and the outputs generated by TheiaCoV_FASTA are outlined below.

Required User Inputs

Download CSV: TheiaCoV_FASTA_required_inputs.csv

Task	Input Variable	Data Type	Description
theiacov_fasta	assembly_fasta	File	SARS-CoV-2 assemly file in fasta format
theiacov_fasta	in-	String Description of the method utilized to	
	put_assembly_method	1	ate the input assembly fasta file; if unknown
			"NA" will be accepted
theiacov_fasta	samplename	String	Name of the sample being analyzed
theiacov_fasta	seq_method	String	Description of the sequencing method uti-
			lized to generate the raw sequencing data; if
			unknown "NA" will be accepted

Optional User Inputs

 $Download\ CSV: \verb|TheiaCoV_FASTA_optional_inputs.csv|\\$

Task	Variable Name	Data Type	Description	Default
nextclade_one_sa	-	String	Docker tag used for running NextClade	nextstrain/nextclade:1.10.3
nextclade_output_	p drxdr<u>e</u>r one_sample	String	Docker tag used for pars- ing NextClade output	python:slim
pangolin3	docker	String	Docker tag used for running Pan- golin	quay.io/staphb/3.1.20-pangolearn- 2022-02-02
pangolin3	infer- ence_engine	String	pangolin infer- ence engine for lineage designa- tions (usher or pangolarn)	usher
pangolin3	max_ambig	Float	Maximum proportion of Ns allowed for pangolin to attempt assignment	0.5
pangolin3	min_length	Int	Minimum query length allowed for pangolin to attempt assignment	10000
titan_fasta	nextclade_dataset	_r Strin g	Nextclade or- ganism dataset	sars-cov-2
titan_fasta	nextclade_dataset	_r Sfeineg ce	Nextclade reference genome	MN908947
titan_fasta	nextclade_dataset	_talextclade dataset tag	2022-02- 07T12:00:00Z	
vadr	docker	String	Docker tag used for running VADR	quay.io/staphb/1.4.1-models-1.3-2
vadr	maxlen	Int	Maximum length for the fasta-trim- terminal- ambigs.pl VADR script	30000
vadr	minlen	Int	Minimum length subsequence to possibly replace Ns for the fastatrim-terminal-ambigs.pl VADR script	50
vadr	skip_length	Int	Minimum as- sembly length (unambiguous) to run vadr	10000
vadr	vadr_opts	String	Options for the v-annotate.pl	-glsearch -s -r -nomisc -mkey sarscov2 -alt_fail lows-
1.2. TheiaCoV V	orkflow Series		VADR script	core,fstukcnf,insertnn,deletinn —mdir /opt/vadr/vadr-models/
version_capture	timezone	String	User time zone in valid	None

 $Download\ CSV: \verb|TheiaCoV_FASTA_default_outputs.csv|\\$

1.2.2 TheiaCoV Workflows for Genomic Epidemiology

Genomic Epidemiology, i.e. generating phylogenetic trees from a set of consensus assemblies (FASTA format) to track the spread and evolution of viruses on a local, national or global scale, has been an important methodological approach in the effort to mitigate disease transmission.

The TheiaCoV Genomic Epidemiology Series contains two seperate WDL workflows (TheiaCoV_Augur_Prep and TheiaCoV_Augur_Run) that process a set of viral genomic assemblies to generate phylogenetic trees (JSON format) and metadata files which can be used to assign epidemiological data to each assembly for subsequent analyses.

The two TheiaCoV workflows for genomic epidemiology must be run sequentially to first prepare the data for phylogenetic analysis and second to generate the phylogenetic trees. More information on the technical details of these processes and information on how to utilize and apply these workflows for public health investigations is available below.

Download CSV: TheiaCoV_Augur_Prep_required_inputs.csv

Task	Input Variable	Data Type	Description
prep_augur_metadata	assembly	File	Assembly/consensus file (single FASTA file
			per sample)
prep_augur_metadata	collection_date	String	Collection date of the sample to be included
			in the analysis
prep_augur_metadata	iso_country	String	Country of the sample to be included in the
			analysis
prep_augur_metadata	iso_state	String	State of the sample to be included in the
			analysis
prep_augur_metadata	iso_continent	String	Continent of the sample to be included in the
			analysis
prep_augur_metadata	pango_lineage	String	Pango Lineage of the sample to be included
			in the analysis

TheiaCoV_Augur_Prep

The TheiaCoV_Augur_Prep workflow was written to process consensus assemblies (FASTA format) and the associated metadata in preparation for running the TheiaCoV_Augur_Run. Input assemblies should be of similar quality (percent reference coverage, number of ambiguous bases, etc.). Inputs with highly discordant quality metrics may result in inaccurate inference of genetic relatedness.

Note: There must be some sequence diversity in the input set of assemblies to be analyzed. As a rule of thumb, the smaller the input set, the more sequence diversity will be required to make any sort of genomic inference. If a small (~ 10) set of viral genomic assemblies is used as the input then it may be necessary to add one significantly divergent assembly.

Upon initiating a TheiaCoV_Augur_Prep run, input assembly/consensus files and associated metadata will be used to produce the array of assembly/consensus files and the array of metadata files to be used as inputs for the Theia-CoV_Augur_Run workflow.

Metadata files are prepared with the Augur_Prep workflow by using BASH commands to first de-identify, and then to parse the headers of the input assembly files.

Required User Inputs

Download CSV: TheiaCoV_Augur_Prep_required_inputs.csv

Task	Input Variable	Data Type	Description
prep_augur_metadata	assembly	File	Assembly/consensus file (single FASTA file
			per sample)
prep_augur_metadata	collection_date	String	Collection date of the sample to be included
			in the analysis
prep_augur_metadata	iso_country	String	Country of the sample to be included in the
			analysis
prep_augur_metadata	iso_state	String	State of the sample to be included in the
			analysis
prep_augur_metadata	iso_continent	String	Continent of the sample to be included in the
			analysis
prep_augur_metadata	pango_lineage	String	Pango Lineage of the sample to be included
			in the analysis

TheiaCoV_Augur_Run

The TheiaCoV_Augur_Run workflow was written to process an array of assembly/consensus files (FASTA format) and and array of sample metadata files (TSV format) using a modified version of The Broad Institute's sarscov2_nextstrain WDL workflow to create an Auspice JSON file; output from the modified sarscov2_nextstrain workflow will also be used to infer SNP distances and create a static PDF report.

Upon initiating a TheiaCoV_Augur_Run run, the input assembly/consensus file array and the associated metadata file array will be used to generate a JSON file that is compatible with phylogenetic tree building software. This JSON can then be used in Auspice or Nextstrain to view the phylogenetic tree. This phylogenetic tree can be used in genomic

epidemiological analysis to visualize the genetic relatedness of a set of samples. The associated metadata can then be used to add context to the phylogenetic visualization.

Required User Inputs

Download CSV: TheiaCoV_Augur_Run_required_inputs.csv

Task	Input Variable	Data Type	Description
sarscov2_nextstrain	assembly_fastas	Array[File]	An array of assembly/consensus files
			(FASTA)
sarscov2_nextstrain	sam-	Array[File]	An array of sample metadata files (TSV)
	ple_metadata_tsvs		
sarscov2_nextstrain	build_name	String	The name of the Augur build to be used in
			this analysis

1.3 Mercury Workflow Series

The Mercury workflow series was developed to allow users to efficiently and accurately prepare submission files for GISAID, SRA, and Genbank submissions as well as BioSample registration. As of today (November 11th, 2021) these workflows are specific to SARS-CoV-2 amplicon read data from clinical samples, but work is underway to allow for the submission preparation of other viral pathogens of concern.

These workflows were written to ingest and properly format all suggested metadata fields as per the Public Health Alliance for Genomic Epidemiology's SARS-CoV-2 Contextual Data Specifications.

1.3.1 Mercury Workflows for Single-Sample Preparation

Sharing of sample read and assembly data through internationally accessible databases allows insights to be drawn about how the virus is spreading and mutating across the globe; the more freely available these data are to international researchers and public health scientists, the stronger our decision making can be.

The Mercury workflows for single-sample preparation is made up of two separate WDL workflows, Mercury_SE_Prep & Mercury_PE_Prep, for preparing submission files to GISAID, SRA, and GenBank for single and paired-end read data, respectively. These two workflows will process read data, assembly files, and contextual metadata to prepare submission for samples individually—while these workflows can process multiple samples in a single run, the submission files prepared are for single-sample submission; for preparation of multiple samples (i.e. batch submission), please see details for the Mercury_Batch workflow below.

A series of introductory training videos that provide conceptual overviews of methodologies and walkthrough tutorials on how to utilize these Mercury workflows through Terra are available on the Theiagen Genomics YouTube page:

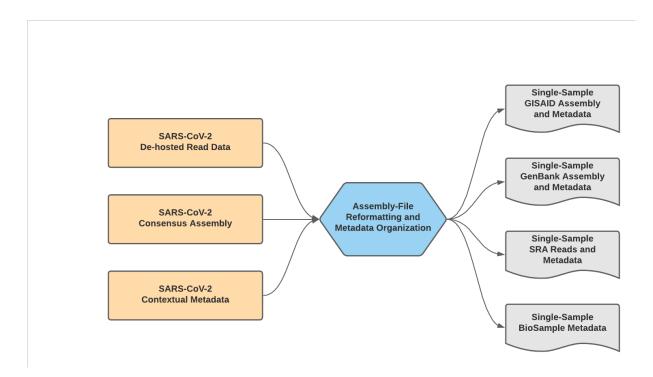


Fig. 6: Mercury_Prep Data Workflow

Mercury_PE_Prep

The Mercury_PE_Prep workflow was written to process paired-end read data, assembly files, and contextual metadata to prepare submission for samples individually.

Note: With default settings, this workflow will only prepare submission files for samples with assembly files containing less than 5,000 Ns. This quality threshold can be adjusted by modifying the number_N_threshold.

A step-by-step video tutorial for utilizing the Mercury_PE_Prep workflow has been made available on the Theiagen YouTube Page:

More information on required user inputs, optional user inputs, default tool parameters and the outputs generated by Mercury_PE_Prep are outlined below.

Required User Inputs

Download CSV: Mercury_PE_Prep_required_inputs.csv

Task	Input Variable	Data Type	Description
mercury_pe_prep	assembly_fasta	File	Consensus genome assembly
mercury_pe_prep	assem-	Float	Mean sequencing depth throughout the
	bly_mean_coverage		conesnsus assembly
mercury_pe_prep	assembly_method	String	Method employed to generate the input as-
			sembbly file
mercury_pe_prep	authors	String	Authors associated with this submission
mercury_pe_prep	biopro-	String	NCBI BioProject accession number
	ject_accession		
mercury_pe_prep	collecting_lab	String	Name of the laboratory that orginial labora-
			tory that collected the sample
mercury_pe_prep	collect-	String	Address of the laboratory that orginial labo-
	ing_lab_address		ratory that collected the sample
mercury_pe_prep	collection_date	String	Date on which the sample was collected
mercury_pe_prep	continent	String	Continent the sample was collected in
mercury_pe_prep	country	String	Country the sample was collected in
mercury_pe_prep	gisaid_submitter	String	GISAID username
mercury_pe_prep	host_disease	String	Host disease; for SARS-CoV-2 sequences
			from human samples, "COVID-19" would
			be the most accurate entry for this field
mercury_pe_prep	instrument_model	String	Model of the sequencing instrument utilized
			to generate the read data
mercury_pe_prep	isolation_source	String	Isolation source, i.e. clinical, animal, or en-
			vironmental
mercury_pe_prep	library_id	String	Unique identifer for the sequenced library
mercury_pe_prep	library_selection	String	Selection methodology used to designate
			samples as eligible for sequencing, e.g.,
			"PCR" for samples selected based on PCT
			Ct values
mercury_pe_prep	library_source	String	Source of the genomic material used to pre-
			pare the sequencing libraries
mercury_pe_prep	library_strategy	String	Library preparation strategy, e.g., "AMPLI-
			CON" for data generated from tiling PCR
			amplicons
mercury_pe_prep	number_N	Int	Number of fully ambiguous basecalls within
			the consensus assembly
mercury_pe_prep	organism	String	Name of the organism sequenced, e.g.
			"SARS-CoV-2"
mercury_pe_prep	read1_dehosted	File	Dehosted forward read file
mercury_pe_prep	read2_dehosted	File	Dehosted reverse read file
mercury_pe_prep	seq_platform	String	Description of the sequencing methodology
			used to generate the input read data
mercury_pe_prep	state	String	State the sample was collected in
mercury_pe_prep	submission_id	String	Unique identfier for the sample utilized upon
			submission
mercury_pe_prep	submitting_lab	String	Name of the submitting laboratory
mercury_pe_prep	submit-	String	Address of the submitting laboratory
	ting_lab_address		

Optional User Inputs

Download CSV: Mercury_PE_Prep_optional_inputs.csv

Task	Input Variable	Data Type	Description	Default
gi-	speci-	String	Biologial source	None
said_prep_one_sa	amplen_source		of the specimen,	
			e.g. e.g. spu-	
			tum, Alveolar	
			lavage fluid,	
			Oro-pharyngeal	
			swab, Blood,	
			Tracheal swab,	
			Urine, Stool,	
			Cloakal swab,	
			Organ, Feces,	
			Other	
gi-	mem_size_gb	Int	Memory allo-	1
said_prep_one_sa	mple		cated to the gi-	
			said_prep_one_sa	mple
			task	
gi-	disk_size	Int	Disk size al-	25
said_prep_one_sa	mple		located to the gi-	
			said_prep_one_sa	mple
			task	
gi-	patient_status	String	Status of the	unknown
said_prep_one_sa	ample		patient, e.g.	
			Hospitalized,	
			Released, Live,	
			Deceased,	
			unknown	
gi-	type	String	Organism typoe	betacoronovirus
said_prep_one_sa	ample			
gi-	CPUs	Int	CPUs allo-	None
said_prep_one_sa	mple		cated to the gi-	
			said_prep_one_sa	mple
			task	
gi-	pre-	Int	Number of	0
said_prep_one_sa	m phe ptible_tries		preemptible	
			tries for the gi-	
			said_prep_one_sa	mple
			task	
gi-	outbreak	String	Outbreak as-	None
said_prep_one_sa	mple		sociated with	
			this submision,	
			e.g. date, place,	
			family cluster	
gi-	last_vaccinated	String	Date of last vac-	None
said_prep_one_sa	mple		cine recieved	
				continues on next page

Table 9 – continued from previous page

Task	Input Variable	Data Type	Description	Default
gi-	docker_image	String	Docker im-	quay.io/theiagen/utility:1.1
said_prep_one_sa	_	Sums	age utilized	quaj.io/meiagon/amity.i.i
said_prep_one_sa	impic		for the gi-	
			said_prep_one_sa	mnle
			task	inpie
a:	magage 4.4.11.	Ctuin a		oui aimol
gi-	passage_details	String	Passage de-	original
said_prep_one_sa	mple		tails of the	
			sample being	
			submitted, e.g.	
			original, vero,	
			etc	
mer-	dehost-	String	Method utilized	NCBI Human Scrubber
cury_pe_prep	ing_method		to dehost read	
			data	
mer-	filetype	String	File type of the	fastq
cury_pe_prep			read data being	
			submitted to	
			SRA	
mer-	submitter_email	String	Email address of	None
cury_pe_prep			the submitter	
mer-	pur-	String	Reason that	None
cury_pe_prep	pose_of_sequenci	ng	this sample	
			was sequenced;	
			for labs that	
			are sequenc-	
			ing samples	
			as part of a	
			federal surveil-	
			lance program	
			"baseline	
			surveillance"	
			would be the	
			most accurate	
			entry for this	
			field	
mer-	library_layout	String	Layout of the se-	paired
cury_pe_prep		541115	quenced library	punou
mer-	num-	Int	Maximum num-	5000
cury_pe_prep	ber_N_threshold	int	ber of ambigu-	3000
cary_pc_prep	oci_iv_uiicsiiolu		ous nucleotides	
			in a sample to	
			prepare submis-	
			sion files	
	host soi	Ctuin a		Hama sonions
mer-	host_sci_name	String	Scientific name	Homo sapiens
cury_pe_prep			of the host or-	
		G. ·	ganism	X
mer-	gi-	String	Accession num-	None
cury_pe_prep	said_accession	G. I	ber in GISAID	10 110
mer-	gisaid_organism	String	Orgiansm name	hCoV-19
cury_pe_prep			as per GISAID	
1			submission	

Table 9 – continued from previous page

Task	Input Variable	Data Type	Description	Default
mer-	county	String	County the lab-	None
cury_pe_prep			oratory was col-	
			lected in	
mer-	amplicon_size	String	Average size of	None
cury_pe_prep			the amplicons	
			sequenced	
mer-	host	String	Common name	Human
cury_pe_prep			of the host or-	
			ganism	
mer-	ampli-	String	Name of the am-	None
cury_pe_prep	con_primer_schen	ne	plicon primer	
			scheme utilized	
			to generate	
			the amplicons	
		~ .	sequenced	
mer-	biosam-	String	BioSample ac-	None
cury_pe_prep	ple_accession	G	cession number	
mer-	treatment	String	Treatment ad-	None
cury_pe_prep			ministered to	
			the patient, e.g.	
			drug name,	
		0.1	dosage, etc.	
mer-	patient_gender	String	Gender of the	unknown
cury_pe_prep		G	patient	
mer-	pur-	String	Reason that the	None
cury_pe_prep	pose_of_sampling		original speci-	
			men was taken,	
			e.g. clinical	
mar	nationt age	String	diagnostics Age of the pa-	unknown
mer-	patient_age	Sumg	tient	ulikilowii
cury_pe_prep ncbi_prep_one_sa	mmbm siza ah	Int	Memory al-	1
iicoi_prep_one_sa	inpen_size_go	IIIt	located to the	
			ncbi_prep_one_sa	mnle
			task	imple
ncbi_prep_one_sa	mhleker image	String	Docker image	quay.io/staphb/vadr:1.3
neor_prep_one_sa	mpocker_image	Sumg	utilized for the	quay.10/stapho/vadi.1.3
			ncbi_prep_one_sa	mple
			task	
ncbi_prep_one_sa	mnkexlen	Int	VADR –maxlen	30000
	1	-	input utilized	
			when trim-	
			ming terminal	
			ambiguous ends	
ncbi_prep_one_sa	m pk e-	Int	Number of	0
_r -r	emptible_tries		preemptible	
	1		tries for the	
			ncbi_prep_one_sa	mple
			task	_
	1		l	continues on next nage

Table 9 – continued from previous page

Task	Input Variable	Data Type	Description	Default
ncbi_prep_one_sa	m ©R Us	Int	CPUs al-	1
			located to the	
			ncbi_prep_one_sa	mple
			task	
ncbi_prep_one_sa	m ple len	Int	VADR -minen	50
			input utilized	
			when trim-	
			ming terminal	
			ambiguous ends	
ncbi_prep_one_sa	m țik k_size	Int	Disk size	25
			allocated the	
			ncbi_prep_one_sa	mple
			task	
version_capture	timezone	String	User time	None
			zone in valid	
			Unix TZ string	
			(e.g. Amer-	
			ica/New_York)	

Download CSV: Mercury_PE_Prep_default_outputs.csv

Output Name	Data Type	Description
biosam-	File	Sample metadata compiled and formatted to meet the BioSample
ple_attributes		submission requirements
genbank_assembly	File	Assembly file reformatted to meet the GenBank submission requirements
genbank_modifier	File	Sample metadata compiled and formatted to meet the GenBank sub- mission requirements; will need to be manually modified to include BioSample accession numbers
gisaid_assembly	File	Assembly file reformatted to meet the GISAID submission requirements
gisaid_metadata	File	Metadata compiled and formatted to meet the GISAID submission requirements
mer-	String	Date of analysis
cury_pe_prep_analysi	s_date	
mer-	String	Version of the Public Health Viral Genomics (PHVG) repository
cury_pe_prep_version	h	used
sra_metadata	File	Sample and read metadata compiled and formatted to meet the SRA submission requirements
sra_read1	File	Forward read formatted for submission to SRA
sra_read2	File	Reverse read formatted for submission to SRA
sra_reads	File	Forward and reverse reads formatted for submission to SRA

Mercury_SE_Prep

The Mercury_SE_Prep workflow was written to process single-end read data, assembly files, and contextual metadata to prepare submission for samples individually.

Note: With default settings, this workflow will only prepare submission files for samples with assembly files containing less than 5,000 Ns. This quality threshold can be adjusted by modifying the number_N_threshold.

More information on required user inputs, optional user inputs, default tool parameters and the outputs generated by Mercury_SE_Prep are outlined below.

Required User Inputs

Download CSV: Mercury_SE_Prep_required_inputs.csv

Task	Input Variable	Data Type	Description
mercury_pe_prep	assembly_fasta	File	Consensus genome assembly
mercury_pe_prep	assem- bly_mean_coverage	Float	Mean sequencing depth throughout the conesnsus assembly
mercury_pe_prep	assembly_method	String	Method employed to generate the input assembly file
mercury_pe_prep	authors	String	Authors associated with this submission
mercury_pe_prep	biopro- ject_accession	String	NCBI BioProject accession number
mercury_pe_prep	collecting_lab	String	Name of the laboratory that orginial laboratory that collected the sample
mercury_pe_prep	collect- ing_lab_address	String	Address of the laboratory that orginial laboratory that collected the sample
mercury_pe_prep	collection_date	String	Date on which the sample was collected
mercury_pe_prep	continent	String	Continent the sample was collected in
mercury_pe_prep	country	String	Country the sample was collected in
mercury_pe_prep	gisaid_submitter	String	GISAID username
mercury_pe_prep	host_disease	String	Host disease; for SARS-CoV-2 sequences from human samples, "COVID-19" would be the most accurate entry for this field
mercury_pe_prep	instrument_model	String	Model of the sequencing instrument utilized to generate the read data
mercury_pe_prep	isolation_source	String	Isolation source, i.e. clinical, animal, or environmental
mercury_pe_prep	library_id	String	Unique identifer for the sequenced library
mercury_pe_prep	library_selection	String	Selection methodology used to designate samples as eligible for sequencing, e.g., "PCR" for samples selected based on PCT Ct values
mercury_pe_prep	library_source	String	Source of the genomic material used to prepare the sequencing libraries
mercury_pe_prep	library_strategy	String	Library preparation strategy, e.g., "AMPLI-CON" for data generated from tiling PCR amplicons
mercury_pe_prep	number_N	Int	Number of fully ambiguous basecalls within the consensus assembly
mercury_pe_prep	organism	String	Name of the organism sequenced, e.g. "SARS-CoV-2"
mercury_pe_prep	reads_dehosted	File	Dehosted read files
mercury_pe_prep	seq_platform	String	Description of the sequencing methodology used to generate the input read data
mercury_pe_prep	state	String	State the sample was collected in
mercury_pe_prep	submission_id	String	Unique identfier for the sample utilized upon submission
mercury_pe_prep	submitting_lab	String	Name of the submitting laboratory
mercury_pe_prep	submit- ting_lab_address	String	Address of the submitting laboratory

46

Optional User Inputs

Download CSV: Mercury_SE_Prep_optional_inputs.csv

Task	Input Variable	Data Type	Description	Default
gi-	speci-	String	Biologial source	None
said_prep_one_sa	mplen_source		of the specimen,	
			e.g. e.g. spu-	
			tum, Alveolar	
			lavage fluid,	
			Oro-pharyngeal	
			swab, Blood,	
			Tracheal swab,	
			Urine, Stool,	
			Cloakal swab,	
			Organ, Feces,	
. •		Total	Other	1
gi-	mem_size_gb	Int	Memory allo-	1
said_prep_one_sa	impie		cated to the gi- said_prep_one_sa	mnla
			task	impie
gi-	disk_size	Int	Disk size al-	25
said_prep_one_sa	_	l Inc	located to the gi-	23
sara_prop_one_se			said_prep_one_sa	mple
			task	r
gi-	patient_status	String	Status of the	unknown
said_prep_one_sa	-		patient, e.g.	
			Hospitalized,	
			Released, Live,	
			Deceased,	
			unknown	
gi-	type	String	Organism typoe	betacoronovirus
said_prep_one_sa	_			
gi-	CPUs	Int	CPUs allo-	None
said_prep_one_sa	imple		cated to the gi-	1.
			said_prep_one_sa task	mpie
gi	nro	Int	Number of	0
gi- said_prep_one_sa	pre-	1111	preemptible	U
said_prep_one_sa	uniquiquote_utes		tries for the gi-	
			said_prep_one_sa	mple
			task	r
gi-	outbreak	String	Outbreak as-	None
said_prep_one_sa			sociated with	
	•		this submission,	
			e.g. date, place,	
			family cluster	
gi-	last_vaccinated	String	Date of last vac-	None
said_prep_one_sa	mple		cine recieved	

Table 10 – continued from previous page

Task	Input Variable	Data Type	Description	Default
gi-	docker_image	String	Docker im-	quay.io/theiagen/utility:1.1
said_prep_one_s	_	Sumg	age utilized	quay.io/theragen/utility.i.i
said_prep_one_s	ampie		for the gi-	
			said_prep_one_sa	mnle
			task	inpic
~i	maganga dataila	Ctuina		ouisino!
gi-	passage_details	String	Passage de-	original
said_prep_one_s	ampie		tails of the	
			sample being	
			submitted, e.g.	
			original, vero,	
	1.1	G. I	etc	NGDIII
mer-	dehost-	String	Method utilized	NCBI Human Scrubber
cury_pe_prep	ing_method		to dehost read	
			data	
mer-	filetype	String	File type of the	fastq
cury_pe_prep			read data being	
			submitted to	
			SRA	
mer-	submitter_email	String	Email address of	None
cury_pe_prep			the submitter	
mer-	pur-	String	Reason that	None
cury_pe_prep	pose_of_sequenci	ng	this sample	
			was sequenced;	
			for labs that	
			are sequenc-	
			ing samples	
			as part of a	
			federal surveil-	
			lance program	
			"baseline	
			surveillance"	
			would be the	
			most accurate	
			entry for this	
			field	
mer-	library_layout	String	Layout of the se-	paired
	1101ai y_iayout	Sumg	quenced library	paned
cury_pe_prep mer-	num-	Int	Maximum num-	5000
	ber_N_threshold	1111	ber of ambigu-	3000
cury_pe_prep	DEI_IN_UITESHOIG		ous nucleotides	
			in a sample to	
			prepare submis-	
	1	G. ·	sion files	***
mer-	host_sci_name	String	Scientific name	Homo sapiens
cury_pe_prep			of the host or-	
			ganism	
mer-	gi-	String	Accession num-	None
cury_pe_prep	said_accession		ber in GISAID	
mer-	gisaid_organism	String	Orgiansm name	hCoV-19
cury_pe_prep			as per GISAID	
			submission	
				continues on payt page

Table 10 – continued from previous page

Task	Input Variable	Data Type	Description	Default
mer-	county	String	County the lab-	None
cury_pe_prep		C	oratory was col-	
7 -1 -1 1			lected in	
mer-	amplicon_size	String	Average size of	None
cury_pe_prep			the amplicons	
<i>•</i> – <i>•</i> – <i>•</i>			sequenced	
mer-	host	String	Common name	Human
cury_pe_prep			of the host or-	
			ganism	
mer-	ampli-	String	Name of the am-	None
cury_pe_prep	con_primer_schen	ne	plicon primer	
			scheme utilized	
			to generate	
			the amplicons	
			sequenced	
mer-	biosam-	String	BioSample ac-	None
cury_pe_prep	ple_accession	C	cession number	
mer-	treatment	String	Treatment ad-	None
cury_pe_prep		<u> </u>	ministered to	
7-1 -1 1			the patient, e.g.	
			drug name,	
			dosage, etc.	
mer-	patient_gender	String	Gender of the	unknown
cury_pe_prep	_8	8	patient	
mer-	pur-	String	Reason that the	None
cury_pe_prep	pose_of_sampling	•	original speci-	
7-1 -1 1			men was taken,	
			e.g. clinical	
			diagnostics	
mer-	patient_age	String	Age of the pa-	unknown
cury_pe_prep		•	tient	
ncbi_prep_one_sa	mplem_size_gb	Int	Memory al-	1
			located to the	
			ncbi_prep_one_sa	mple
			task	_
ncbi_prep_one_sa	m hle ker_image	String	Docker image	quay.io/staphb/vadr:1.3
1		<u>-</u>	utilized for the	
			ncbi_prep_one_sa	mple
			task	
ncbi_prep_one_sa	mplexlen	Int	VADR -maxlen	30000
			input utilized	
			when trim-	
			ming terminal	
			ambiguous ends	
ncbi_prep_one_sa	mprke-	Int	Number of	0
	emptible_tries		preemptible	
	• -		tries for the	
			ncbi_prep_one_sa	mple
			task	
	1		·	continues on poyt page

Table 10 – continued from previous page

Task	Input Variable	Data Type	Description	Default
ncbi_prep_one_sa	m ©R Us	Int	CPUs al-	1
			located to the	
			ncbi_prep_one_sa	mple
			task	
ncbi_prep_one_sa	m ple len	Int	VADR -minen	50
			input utilized	
			when trim-	
			ming terminal	
			ambiguous ends	
ncbi_prep_one_sa	mdpikak_size	Int	Disk size	25
			allocated the	
			ncbi_prep_one_sa	mple
			task	
version_capture	timezone	String	User time	None
			zone in valid	
			Unix TZ string	
			(e.g. Amer-	
			ica/New_York)	

Download CSV: Mercury_SE_Prep_default_outputs.csv

Output Name	Data Type	Description	
biosam-	File	Sample metadata compiled and formatted to meet the BioSample	
ple_attributes		submission requirements	
genbank_assembly	File	Assembly file reformatted to meet the GenBank submission require-	
		ments	
genbank_modifier	File	Sample metadata compiled and formatted to meet the GenBank sub-	
		mission requirements; will need to be manually modified to include	
		BioSample accession numbers	
gisaid_assembly	File	Assembly file reformatted to meet the GISAID submission require-	
		ments	
gisaid_metadata	File	Metadata compiled and formatted to meet the GISAID submission	
		requirements	
mer-	String	Date of analysis	
cury_pe_prep_analysi	s_date		
mer-	String	Version of the Public Health Viral Genomics (PHVG) repository	
cury_pe_prep_version	h	used	
sra_metadata	File	Sample and read metadata compiled and formatted to meet the SRA	
		submission requirements	
sra_reads	File	Forward and reverse reads formatted for submission to SRA	

1.3.2 Mercury Workflows for Multiple-Sample (Batch) Preparation

We have made a single WDL workflow for multiple-sample (batch) preparation: Mercury_Batch.

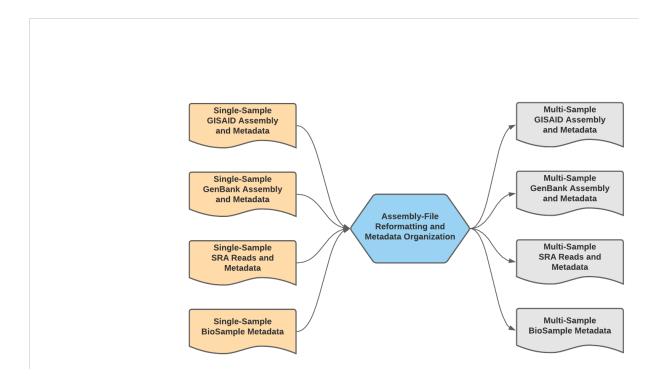


Fig. 7: Mercury_Batch Data Workflow

Mercury_Batch

The Mercury_Batch workflow was written to process the output submission files from Mercury_PE_Prep or Mercury_SE_Prep and combine them to enable GISAID, SRA, and Genbank batch submission as well as batch BioSample registration. To avoid issues with NCBI GenBank rejections, the Mercury_Batch workflow will remove any sample with raised VADR alerts from the prepared batch submission files.

Note: With default settings, this workflow will remove samples any sample with one or more raised VADR alerts. This screening threshold can be adjusted by modifying the vadr_threshold.

A step-by-step video tutorial for utilizing the Mercury_Batch workflow has been made available on the Theiagen YouTube Page:

More information on required user inputs, optional user inputs, default tool parameters and the outputs generated by Mercury_Batch are outlined below.

Required User Inputs

Download CSV: Mercury_Batch_required_inputs.csv

Task	Input Variable	Data Type	Description
mercury_batch	biosam-	Array[File]	Array of sample metadata filescompiled and
	ple_attributes		formatted to meet the BioSample submis-
			sion requirements
mercury_batch	genbank_assembly	Array[File]	Array of assembly files reformatted to meet
			the GenBank submission requirements
mercury_batch	genbank_modifier	Array[File]	Array of sample metadata files compiled and
			formatted to meet the GenBank submission
			requirements; will need to be manually mod-
			ified to include BioSample accession num-
			bers
mercury_batch	gisaid_assembly	Array[File]	Array of metadata files compiled and for-
			matted to meet the GISAID submission re-
			quirements
mercury_batch	gisaid_metadata	Array[File]	Array of assembly files reformatted to meet
			the GISAID submission requirements
mercury_batch	samplename	Array[String]	Array of sample identifiers
mercury_batch	sra_metadata	Array[File]	Array of sample and read metadata files
			compiled and formatted to meet the SRA
			submission requirements
mercury_batch	sra_reads	Array[String]	Array of forward and reverse reads formatted
			for submission to SRA
mercury_batch	submission_id	Array[String]	Array of submission identifiers
mercury_batch	vadr_num_alerts	Array[String]	Array of VADR number of alerts

Optional User Inputs

Download CSV: Mercury_Batch_optional_inputs.csv

Task	Input Variable	Data Type	Description	Default
com-	docker_image	String	Docker im-	quay.io/theiagen/utility:1.1
pile_biosamp_n_	sra		age utilized	
			for the com-	awa.
			pile_biosample_n task	Sra
com	nra	Int	Number of pre-	0
com- pile_biosamp_n_	pre-	IIII	emptible tries	O
pric_orosamp_n_	sia inpublic_uies		for the com-	
			pile_biosample_n	sra
			task	
gen-	docker_image	String	Docker im-	quay.io/theiagen/utility:1.1
bank_compile		8	age utilized	1,
_ 1			for the gen-	
			bank_compile	
			task	
gen-	pre-	Int	Number of	0
bank_compile	emptible_tries		preemptible	
			tries for the gen-	
			bank_compile	
			task	
gisaid_compile	docker_image	String	Docker image	quay.io/theiagen/utility:1.1
			utilized for the	
			gisaid_compile	
1 1		T .	task	0
gisaid_compile	pre-	Int	Number of	0
	emptible_tries		preemptible	
			tries for the gisaid_compile	
			task	
mercury_batch	CPUs	Int	CPUs allocated	4
mercury_batem	CIUS	IIIt	for each task	7
			in the mer-	
			cury_batch	
			workflow	
mercury_batch	disk_size	Int	Disk size allo-	100
•			cated for each	
			task in the	
			mercury_batch	
			workflow	
mercury_batch	gcp_bucket	String	GCP bucket for	None
			SRA transfer	
mercury_batch	mem_size_gb	Int	Memory allo-	8
			cated for each	
			task in the	
			mercury_batch workflow	
marazer hatah	wode throchold	Int	Maximum num-	0
mercury_batch	vadr_threshold	Int	ber of VADR	U
			alerts for sam-	
			ples included	
			in the batch	
			submission files	
version_capture	timezone	String	User time	None
			zone in valid	
1.3. Mercury Wo	orkflow Series		Unix TZ string	53
			(e.g. Amer-	
			ica/New_York)	

Download CSV: Mercury_Batch_default_outputs.csv

Output Name	Data Type	Description		
Gen-	File	File detailing all of the files bacthed for GenBank submission		
Bank_batched_sample	es			
Gen-	File	File detailing all of the files excluded from the prepared submission		
Bank_excluded_samp	les	files for GenBank		
GenBank_modifier	File	Compiled matadata formatted for batch submissinon to GenBank		
GISAID_assembly	File	Concatenated assemly file for batch submission to GenBank		
GI-	File	File detailing all of the files bacthed for GenBank submission		
SAID_batched_sample	es			
GI-	File	File detailing all of the files excluded from the prepared submission		
SAID_excluded_samples		files for GenBank		
GISAID_metadata	File	Compiled metadata formatted for batch submissino to GISAID		
mer-	String	Date of analysis		
cury_batch_analysis_	date			
mer-	String	Version of the Public Health Viral Genomics (PHVG) repository		
cury_batch_version		used		
SRA_gcp_bucket	String	GCP bucket location for SRA read transfer		
SRA_metadata	File	Compiled metadata formatted for batch submissino to SRA		
SRA_zipped_reads	File	All reads prepared for SRA submission (empty file is GCP bucket		
		location was provided for SRA read transfer)		

1.4 License

GNU Affero General Public License v3.0