Public Health Viral Genomics (Theiagen)

Release 2.0.0

Kevin G. Libuit

Feb 16, 2022

CONTENTS

1 Contents					
	1.1	Public Health Viral Genomics			
	1.2	TheiaCoV Workflow Series			
	1.3	Mercury Workflow Series	3		
	1.4	License	5		

CHAPTER

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_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 preads 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_	ar210202010528ce.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	

Task	Variable Name	Data Type	Description	Default
consensus	disable bag	Boolean	Disable read	TRUE
consensus	uisable_baq	Doolean	pair overlap	INCL
			detection for	
			SAMtools	
			mpileup before	
			running iVar	
			consensus	
consensus	max depth	Int	Maximum reads	600000
	- 1		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	10
			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	
	6		consensus	
consensus	ref_genome	String	Path to the ref-	/artic-
			erence genome	ncov2019/primer_schemes/nCoV-
			within the	2017/ V 3/IICOV-
			Stapho/ivar:1.2.2	annersterederoce.rasta
			Docker con-	
			tainer	

Table 1 – 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_ASM98588	9v3_genomic.gff
			ture format		
			of the refer-		
			ence genome		
			within the		
			staphb/ivar:1.2.2_	artic20200528	
			Docker con-		
			tainer		
nextclade_one_sa	m øb æker	String	Docker tag used	neherlab/nextclade:0.14.2	
			for running		
			NextClade		
nextclade_output_	pdrxdrenone_sample	String	Docker tag	python:slim	
			used for pars-		
			ing NextClade		
			output		
pangolin3	docker	String	Docker tag used	staphb/pangolin:3.1.11-pangolearn-	
		-	for running Pan-	2021-08-24	
			golin		
pangolin3	infer-	String	pangolin infer-	usher	
1 0	ence engine	U	ence engine for		
	- 0		lineage designa-		
			tions (usher or		
			pangolarn)		
pangolin3	min length	Int	Minimum query	10000	
I O	_ * 8*		length allowed		
			for pangolin		
			to attempt		
			assignment		
pangolin3	max ambig	Float	Maximum pro-	0.5	
I O	8		portion of Ns al-		
			lowed for pan-		
			golin to attempt		
			assignment		
primer trim	keep noprimer re	aBroolean	Include reads	True	
r	<u>-</u> <u>-</u>		with no primers		
			for iVar trim		
read OC trim	bbduk mem	Int	Memory al-	8	
			located to the		
			BBDuk VM		
read OC trim	trimmo-	Int	Specifies the	25	
···· C ······	matic minlen		minimum		
			length of reads		
			to be kept for		
			Trimmomatic		
read OC trim	trimmo-	Int	Specifies the av-	30	
u	matic quality tri	n score	erage quality re-		
	quunty_un		auired for Trim-		
			momatic		
			monute		

Table 1 – continued from previous page

	iu		nom providuo pu	90
Task	Variable Name	Data Type	Description	Default
read_QC_trim	trimmo-	Int	Specifies the	4
	matic_window_siz	ze	number of	
			bases to aver-	
			age across for	
			Trimmomatic	
theia-	nextclade_dataset	r Stmie g	Nextclade or-	sars-cov-2
cov_illumina_pe			ganism dataset	
theia-	nextclade_dataset	r Sfeineg ce	Nextclade refer-	MN908947
cov_illumina_pe			ence genome	
theia-	nextclade_dataset	thextclade	2021-06-	
cov_illumina_pe		dataset tag	25T00:00:00Z	
theia-	seq_method	String	Description of	Illumina paired-end
cov_illumina_pe			the sequencing	
			methodology	
			used to generate	
			the input read	
		~ .	data	
vadr	docker	String	Docker tag used	staphb/vadr:1.2.1
			for running	
_	-	-	VADR	
vadr	maxlen	Int	Maximum	30000
			length for the	
			fasta-trim-	
			terminal-	
			ambigs.pl	
. 1.		Tut	VADR script	50
vadr	minien	Int	Minimum	50
			lengun sub-	
			sequence to	
			No for the fasta	
			trim terminal	
			ambigs nl	
			VADR script	
vadr	skin length	Int	Minimum as-	10000
vaar	skip_iengui	Int	sembly length	10000
			(unambiguous)	
			to run vadr	
vadr	vadr opts	String	Options for the	
, wor	·uui_opto	Sumg	v-annotate.pl	-mkey sarscov? -alt fail lows-
			VADR script	core.fstukcnf.insertnn.deletinn
			······	-mdir /opt/vadr/vadr-models/
variant call	count orphans	Boolean	Do not skip	TRUE
			anomalous	_
			read pairs in	
			variant calling	
			for SAMtools	
			mpileup before	
			running iVar	
			variants	

Table 1 – continued from previous page

Task	Variable Name	Data Type	Description	Default]
variant_call	disable_baq	Boolean	Disable read-	TRUE	-
	_ 1		pair overlap		
			detection for		
			SAMtools		
			mpileup before		
			running iVar		
			variants		
variant_call	max_depth	Int	Maximum reads	600000	-
	-		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	10	-
			depth to call		
			variants for iVar		
			variants		
variant_call	min_freq	Float	Minimum	0.6	
			frequency		
			threshold($0 - 1$)		
			to call variants		
			for iVar variants		_
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-	
			general fea-	ence/GCF_009858895.2_ASM98588	9v3_genomic.gff
			ture format		
			of the refer-		
			ence genome		
			within the		
			staphb/ivar:1.2.2	artic20200528	
			Docker con-		
vioniant11	nof government	String	Doth to the ref	lantia	-
variant_call	rei_genome	Suring	Pain to the ref-	/artic-	
			erence genome	ncov2019/primer_schemes/nCoV-	
			within the	2017/ V 3/1100V-	
				alle Collemente and College Co	
			toiner		
			tamer		

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)	

Table 1 – continued from previous page

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	ious	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 sequclean filtering as determined by
		FastQC
fastqc_clean2	Int	Number of reverse reads after sequclean 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

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_consense	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_human_dehos	teElloat	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	_ jFeloae 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
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		

Table	2 –	continued	from	previous	page
-------	-----	-----------	------	----------	------

TheiaCoV_IIIumina_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

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_	an210290296528ce.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	

Task	Variable Name	Data Type	Description	Default
consensus	disable_baq	Boolean	Disable read-	TRUE
	- 1		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 ba	Int	Minimum man	0
consensus	IIIII_0q		ning quality for	0
			an alignment	
			to be used	
			for SAMtools	
			noi SAMUOIS	
			mpneup belole	
			running rvar	
	min denth	Tert	Minimum and	10
consensus	min_depth	Int	Minimum read	10
			depth to call	
			variants for 1 var	
	Constant Constant	D 1	Consensus	0.6
consensus	min_freq	Float	Minimum	0.6
			trequency	
			threshold(0 -	
			1) to call vari-	
			ants for 1Var	
		.	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-
			staphb/ivar:1.2.2_	ar2102902010528ce.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-	
	_0	6	general fea-	ence/GCF 009858895.2 ASM98588	9v3 genomic.gff
			ture format		88
			of the refer-		
			ence genome		
			within the		
			staphb/ivar:1.2.2	artic20200528	
			Dealer con		
			Docker con-		
. 1 1	11 1	<u>0</u> , :			
nextclade_one_s	am p æker	String	Docker tag used	neherlab/nextclade:0.14.2	
			for running		
			NextClade		
nextclade_outpu	t_pdrxdre_one_sample	String	Docker tag	python:slim	
			used for pars-		
			ing NextClade		
			output		
pangolin3	docker	String	Docker tag used	staphb/pangolin:3.1.11-pangolearn-	
1 0		C C	for running Pan-	2021-08-24	
			golin		
pangolin3	infer-	String	pangolin infer-	usher	
pungonno	ence engine	Sung	ence engine for		
	enec_engine		lineage designa		
			tions (usher or		
			uons (usher or		
11. 0		T	pangolarn)	10000	
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	aBoolean	Include reads	True	
r			with no primers		
			for iVar trim		
read OC trim	bbduk mem	Int	Memory al-	8	
	bbduk_mem	1110	located to the	0	
1.00.43		T /		25	
read_QC_trim	trimmo-	Int	specifies the	23	
	matic_minlen		minimum		
			length of reads		
			to be kept for		
			Trimmomatic		
read_QC_trim	trimmo-	Int	Specifies the av-	30	
	matic_quality_trir	n_score	erage quality re-		
			quired for Trim-		
			momatic		
1				1	

Table 3 – continued from previous page	ge
--	----

read_QC_trimtrimmo- matic_window_sizeInt numberSpecifies the number of bases to aver age across for4theia- cov_illumina_senextclade_dataset nextclade_datasetfötning frimgceNextclade or ganism datasetsars-cov-2theia- cov_illumina_senextclade_dataset datasetföfningceNextclade refer- ence genomeMN908947theia- cov_illumina_senextclade_dataset datasetföfningceNextclade refer- ence genomeMN908947theia- cov_illumina_seseq_methodStringDescription of the sequencing methodology used to generate the input read dataIllumina paired-endvadrdockerStringDocker tag used for running VADRstaphb/vadr:1.2.1vadrmaxlenIntMaximum length for the fasta-trim- terminal- ambigs.pl30000vadrminlenIntMinimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl50vadrskip_lengthIntMinimum as- length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl10000	Task	Variable Name	Data Type	Description	Default
numbernumbernumberof bases to aver- age across for Trimmomatictheia- cov_illumina_senextclade_dataset_ifsfiringceNextclade or- ganism datasetsars-cov-2theia- cov_illumina_senextclade_dataset_ifsfiringceNextclade or- ence genomesars-cov-2theia- cov_illumina_senextclade_dataset_ifsfiringceNextclade or- ence genomeMN908947theia- cov_illumina_senextclade_dataset_ifsfiringceDescription of the sequencing methodology used to generate the input read dataIllumina paired-endvadrdockerStringDecker tag used for running VADRStoph/vadr:1.2.1vadrmaxlenIntMaximum length for the fasta-trim- terminal- ambigs.pl30000vadrminlenIntMinimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl50vadrskip_lengthIntMinimum maning.50vadrskip_lengthIntMinimum as- sembly length10000	read OC trim	trimmo-	Int	Specifies the	4
Internationbases to average across for Trimmomatictheia- cov_illumina_senextclade_datasetf&iming imingsars-cov-2theia- cov_illumina_senextclade_datasetf&fiming imingNextclade referer ence genomeMN908947theia- cov_illumina_senextclade_datasetfMining iming2021-06- 25700:00:00ZMN908947theia- cov_illumina_seseq_methodStringDescription of the sequencing methodology used to generate the input read dataIllumina paired-endvadrdockerStringDocker tag used for running VADRstaph/vadr:1.2.1vadrmaxlenIntMaximum length for the fasta-trim- terminal- ambigs.pl VADR script30000vadrminlenIntMinimum hength sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script50vadrskip_lengthIntMinimum as- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script10000		matic window size	ze	number of	
InterpretationIntAge across for Trimmomatictheia- cov_illumina_senextclade_dtaset_ iffcingccNextclade or ganism datasetsars-cov-2theia- cov_illumina_senextclade_dtaset_ iffcingccNextclade refer- ence genomeMN908947theia- cov_illumina_senextclade_dtaset_ iffcingccNextclade refer- ence genomeMN908947theia- cov_illumina_senextclade_dtaset_ iffcingccNextclade refer- ence genomeMN908947theia- cov_illumina_seseq_method201-06- dtaset tag25700:00:00ZIllumina paired-endtheia- cov_illumina_seseq_methodStringDescription of the sequencing methodology used to generate the input read dataIllumina paired-endvadrdockerStringDocker tag used for running VADRstaphb/vadr:1.2.1vadrmaxlenIntMaximum length for the fasta-trim- terminal- ambigs.pl30000vadrminlenIntMinimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl50vadrskip_lengthIntMinimum Minimum as- sembly length50				bases to aver-	
Image: constraint of the interm of the in				age across for	
theia- cov_illumina_senextclade_dataset førning Nextclade or- ganism datasetsars-cov-2theia- cov_illumina_senextclade_dataset føring ceNextclade refer- ence genomeMN908947theia- cov_illumina_senextclade_dataset føring ce2021-06- 25T00:00:00ZMN908947theia- cov_illumina_seseq_methodStringDescription of the sequencing used to generate the input read dataIllumina paired-endvadrdockerStringDocker tag used for running VADRstaphb/vadr:1.2.1vadrmaxlenIntMaximum length for the fata-t-trim- terminal- ambigs.pl30000vadrminlenIntMainimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl50vadrskip_lengthIntMinimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl50				Trimmomatic	
cov_illumina_seimatefextclade_datasetfeftingceganism datasettheia- cov_illumina_senextclade_datasettbgxtclade2021-06- 25700:00:00ZMN908947theia- cov_illumina_senextclade_datasettbgxtclade2021-06- the sequencing methodology used to generate the input read dataIllumina paired-endvadrdockerStringDescription of the sequencing methodology Used to generate the input read dataIllumina paired-endvadrdockerStringDocker tag used to generate the input read data30000vadrmaxlenIntMaximum length for the fasta-trim- terminal- ambigs.pl VADR script30000vadrminlenIntMinimum length sub- sequence to possibly replace trim-terminal- ambigs.pl VADR script50vadrskip_lengthIntMinimum length sub- sequence to possibly replace trim-terminal- ambigs.pl VADR script50vadrskip_lengthIntMinimum as- sembly length50	theia-	nextclade_dataset	r Strie g	Nextclade or-	sars-cov-2
theia- cov_illumina_senextclade_dataset_ibigxtcladeNextclade refer- ence genomeMN908947theia- cov_illumina_senextclade_dataset_ibigxtclade2021-06- 25T00:00:00ZIllumina paired-endtheia- cov_illumina_seseq_methodStringDescription of the sequencing methodology used to generateIllumina paired-endvadrdockerStringDocker tag used for running VADRstaphb/vadr:1.2.1vadrmaxlenIntMaximum length for the fasta-trim- terminal- ambigs.pl30000vadrminlenIntMinimum length sub- sequence to NS for the fasta- trim-terminal- ambigs.pl50vadrminlenIntMinimum length sub- sequence to NS for the fasta- trim-terminal- ambigs.pl50vadrminlenIntMinimum length sub- sequence to NS for the fasta- trim-terminal- ambigs.pl50vadrskip_lengthIntMinimum ambigs.pl NS for the fasta- trim-terminal- ambigs.pl50vadrskip_lengthIntMinimum Minimum as- sembly length50	cov_illumina_se			ganism dataset	
cov_illumina_seimate of the sequenciesence genomeence genometheia- cov_illumina_seimate of the sequencies methodology used to generate the input read dataIllumina paired-endvadrdockerStringDocker tag used for running VADRstaphb/vadr:1.2.1vadrmaxlenIntMaximum length for the fasta-trim- terminal- ambigs.pl30000vadrminlenIntMaximum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl50vadrminlenIntMinimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl50vadrskip_lengthIntMinimum as- sembly length50vadrskip_lengthIntMinimum as- sembly length10000	theia-	nextclade_dataset	r Sfeineg ce	Nextclade refer-	MN908947
theia- cov_illumina_senextclade_dataset_tkgextclade2021-06- 25700:00:00Ztheia- cov_illumina_seseq_methodStringDescription of the sequencing methodology used to generate the input read dataIllumina paired-endvadrdockerStringDocker tag used for running VADRstaphb/vadr:1.2.1vadrmaxlenIntMaximum length for the fasta-trim-l- ambigs.pl VADR script30000vadrminlenIntMaximum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script50vadrminlenIntMinimum length for the fasta-trim- terminal- ambigs.pl VADR script50vadrminlenIntMinimum length fasta- trim-terminal- ambigs.pl VADR script50vadrskip_lengthIntMinimum as- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script10000	cov_illumina_se			ence genome	
cov_illumina_seintdataset tag25T00:00:00Ztheia- cov_illumina_seseq_methodStringDescription of the sequencing methodology used to generate the input read dataIllumina paired-endvadrdockerStringDocker tag used for running VADRstaphb/vadr:1.2.1vadrmaxlenIntMaximum length for the fasta-trim- terminal- ambigs.pl30000vadrminlenIntMaximum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl50vadrminlenIntMinimum as- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl50vadrskip_lengthIntMinimum as- sembly length10000	theia-	nextclade_dataset	tagextclade	2021-06-	
theia- cov_illumina_seseq_methodStringDescription of the sequencing methodology used to generate the input read dataIllumina paired-endvadrdockerStringDocker tag used for vADRstaphb/vadr:1.2.1vadrmaxlenIntMaximum length for the fasta-trim- terminal- ambigs.pl VADR script30000vadrminlenIntMaximum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script50vadrskip_lengthIntMinimum ambigs.pl VADR script50vadrminlenIntMinimum ambigs.pl vADR script50vadrskip_lengthIntMinimum as- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script50vadrskip_lengthIntMinimum as- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script50	cov_illumina_se		dataset tag	25T00:00:00Z	
cov_illumina_sethe sequencing methodology used to generate the input read datathe sequencing methodology used to generate the input read datavadrdockerStringDocker tag used for running VADRstaphb/vadr:1.2.1vadrmaxlenIntMaximum length for the fasta-trim- terminal- ambles,pl30000vadrminlenIntMaximum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambles,pl50vadrminlenIntMinimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambles,pl50vadrskip_lengthIntMinimum as- sembly length50vadrskip_lengthIntMinimum as- sembly length10000	theia-	seq_method	String	Description of	Illumina paired-end
vadrdockerStringmethodology used to generate the input read datastaphb/vadr:1.2.1vadrdockerStringDocker tag used for running VADRstaphb/vadr:1.2.1vadrmaxlenIntMaximum length for the fasta-trim- terminal- ambigs.pl30000vadrminlenIntMaximum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl50vadrminlenIntMinimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl50vadrskip_lengthIntMinimum as- sembly length10000	cov_illumina_se			the sequencing	
vadrdockerStringDocker tag used for running VADRstaphb/vadr:1.2.1vadrmaxlenIntMaximum length for the fasta-trim- terminal- ambigs.pl VADR script30000vadrminlenIntMaximum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script50vadrminlenIntMinimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script50vadrminlenIntMinimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script50vadrskip_lengthIntMinimum as- sembly length10000				methodology	
vadrdockerStringDock rag used for running VADRstaphb/vadr:1.2.1vadrmaxlenIntMaximum length for the fasta-trim- terminal- ambigs.pl VADR script30000vadrminlenIntMaximum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script50vadrminlenIntMinimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script50vadrskip_lengthIntMinimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script10000				used to generate	
vadrdockerStringDocker tag used for running VADRstaphb/vadr:1.2.1vadrmaxlenIntMaximum length for the fasta-trim- terminal- ambigs.pl VADR script30000vadrminlenIntMaximum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script50vadrminlenIntMinimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script50vadrskip_lengthIntMinimum as- sembly length10000				the input read	
vadrdockerStringDocker tag used for running VADRstaphb/vadr:1.2.1vadrmaxlenIntMaximum length for the fasta-trim- terminal- ambigs.pl30000vadrminlenIntMaximum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl50vadrminlenIntMinimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl50vadrskip_lengthIntMinimum as- sembly length10000	1	1 1	Q. :	data	
vadr maxlen Int Maximum length for the fasta-trim- terminal- ambigs.pl 30000 vadr maxlen Int Maximum length for the fasta-trim- terminal- ambigs.pl 30000 vadr minlen Int Maximum length for the fasta-trim- terminal- ambigs.pl 50 vadr minlen Int Minimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl 50 vadr Kip_length Int Minimum as- sembly length 50	vadr	docker	String	Docker tag used	staphb/vadr:1.2.1
vadrmaxlenIntMaximum length for the fasta-trim- terminal- ambigs.pl VADR script30000vadrminlenIntMaximum length for the fasta-trim- terminal- ambigs.pl VADR script50vadrminlenIntMinimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script50vadrskip_lengthIntMinimum sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script50vadrskip_lengthIntMinimum as- sembly length10000				for running	
vadrmaxienintMaximumsooodlength for the fasta-trim- terminal- ambigs.pllength for the fasta-trim- terminal- ambigs.pllengthvadrminlenIntMinimum fasta-trim- terminal- lengthsooodvadrminlenIntMinimum lengthsooodvadrminlenIntMinimum lengthsooodvadrminlenIntMinimum lengthsequence to possibly replace Ns for the fasta- trim-terminal- ambigs.plsooodvadrskip_lengthIntMinimum as- sembly length10000			T ₁₁ 4	VADR	20000
vadrminlenIntMinimum length50vadrminlenIntMinimum length50vadrMinlenNa for the fasta- trim-terminal- ambigs.pl50vadrMinimum sequence50vadrNa for the fasta- trim-terminal- ambigs.pl10000vadrSkip_lengthIntMinimum as- sembly10000	vadr	maxien	Int	Maximum	30000
VadrminlenIntMinimum length50vadrminlenIntMinimum length50vadrMininumbiological possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script50vadrIntMinimum possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script50vadrIntMinimum possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script10000				facta trim	
vadrminlenIntMinimum length50vadrminlenIntMinimum length50vadrkkk				terminal	
vadrminlenIntMinimum length sequence trim-terminal- ambigs.pl50vadrkip_lengthIntMinimum length sequence trim-terminal- ambigs.pl VADR script50vadrkip_lengthIntMinimum sequence trim-terminal- ambigs.pl VADR script50vadrkip_lengthIntMinimum as- sembly length10000				ambigs pl	
vadrminlenIntMinimum length50lengthsub- sequenceto possibly replace Ns for the fasta- trim-terminal- ambigs.pl VADR script50vadrskip_lengthIntMinimum sequence trim-terminal- ambigs.pl VADR script10000				VADR script	
vadiinitialinitialinitialisolengthsub-lengthsub-sequencetopossibly replaceNs for the fasta-trim-terminal-ambigs.plVADR scriptvadrskip_lengthIntMinimum as- sembly length10000	vadr	minlen	Int	Minimum	50
vadr skip_length integer integer sequence to possibly replace Ns for the fasta- irim-terminal- ambigs.pl VADR script vadr skip_length Int Minimum as- sembly length	vuui	minen	m	length sub-	
vadr skip_length Int Minimum as- sembly length 10000				sequence to	
vadr skip_length Int Minimum as- sembly length 10000				possibly replace	
vadr skip_length Int Minimum as- sembly length 10000				Ns for the fasta-	
vadr skip_length Int Minimum as- sembly length 10000				trim-terminal-	
vadr skip_length Int Minimum as- sembly length 10000				ambigs.pl	
vadr skip_length Int Minimum as- sembly length 10000				VADR script	
sembly length	vadr	skip_length	Int	Minimum as-	10000
				sembly length	
(unambiguous)				(unambiguous)	
to run vadr				to run vadr	
vadr vadr_opts String Options for the –glsearch -s -r –nomisc	vadr	vadr_opts	String	Options for the	–glsearch -s -r –nomisc
v-annotate.pl –mkey sarscov2 –alt_fail lows-				v-annotate.pl	-mkey sarscov2 -alt_fail lows-
VADR script core,fstukcnf,insertnn,deletinn				VADR script	core,fstukcnf,insertnn,deletinn
mdir /opt/vadr/vadr-models/			D 1	D	-mdir /opt/vadr/vadr-models/
variant_call count_orphans Boolean Do not skip TRUE	variant_call	count_orphans	Boolean	Do not skip	TRUE
anomalous				anomalous	
read pairs in				read pairs in	
for SAMtoole				for SAMtoolo	
101 SAIVILOOIS mileun bafara				noi SAIVILOOIS	
rupping iVar				running iVor	
variants				variants	

|--|

Task	Variable Name	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 denth	Int	Maximum reads	600000	-
variant_can	max_depth		read at a posi		
			tion per input		
			file for SAM		
			tools mailsun		
			hoforo muning		
			iVan seriente		
	• 1	T	1 var 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	10	
			depth to call		
			variants for iVar		
			variants		
variant_call	min_freq	Float	Minimum	0.6	
			frequency		
			threshold($0 - 1$)		
			to call variants		
			for iVar variants		
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-	
			general fea-	ence/GCF_009858895.2_ASM98588	9v3_genomic.gff
			ture format		
			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-	1
			erence genome	ncov2019/primer_schemes/nCoV-	
			within the	2019/V3/nCoV-	
			staphb/ivar:1.2.2	ar2122920125218ce.fasta	
			Docker con-]	
			tainer		

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)	

Table 3 – continued from previous page

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 sequclean 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

Output Name	Data Type	Description
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	age	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	becae nt	Percent of read data with primers trimmed as determined 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_consen	sus	consensus
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-
1-1		put read data
theia-	String	Date of analysis
cov_illumina_se_anal	ysis_date	
theia-	String	Version of the Public Health Viral Genomics (PHVG) repository
cov_illumina_se_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_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 TheiaCoV_ClearLabs workflow.

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: TheiaCoV_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 allocated to the	8
			Artric Medaka	
			task runtime	
		-	environment	
consensus	docker	String	Docker tag	staphb/artic-ncov2019:1.3.0
			used for run-	
			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 analyzed	fastqc_se_clean

Task	Variable Name	Data Type	Description	Default
fastac se raw	cous	Int	CPU resources	
rastqe_se_raw	epus	IIIt	allocated to	
			the EastOC	
			the rasiQC	
			task runtime	
			environment for	
			asessing raw	
			read data	
fastqc_se_raw	read1_name	String	Name of the	Inferred from the input read file
			sample being	
			analyzed	
kraken2_dehosted	cpus	Int	CPU resources	4
	-		allocated to	
			the Kraken	
			task runtime	
			environment	
			for asessing	
			dehosted read	
			dete	
lungloon 2 dala asta d	1-mala - m2 - dh	Ctuin a	Uala Dath to the set	(lawshaan 2, dh
kraken2_denosied	kraken2_db	String	Path to the rel-	/kraken2-db
			erence genome	
			within the	
			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
	.1		allocated to	
			the Kraken	
			task runtime	
			anyironment for	
			assessing row	
			asessing law	
1-ma1-am2	1	Stuin a	Deth to the set	(levelser 2, db
krakenz_raw	kraken2_00	Sumg	raul to the ref-	/KIAKEIIZ-UU
			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	

Table 5	- continued	from	previous page	
---------	-------------	------	---------------	--

		Ia	ble 5 – continued	i from previous pa	ge	
	Task	Variable Name	Data Type	Description	Default	
Γ	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	b7dba71079344daea4ea3363e1a67fa5	4edb7ec65459d03
F	nextclade one sa	m db æker	String	Docker tag used	neherlah/nextclade:0.14.2	
	nextende_one_su	npiekoi	String	for running	henerius/hexterade.o.i i.2	
				NextClade		
-	nextclade output	ndrækerone sample	String	Docker tag	python:slim	
	nextende_output_	parate one_sample	String	used for pars	python.sinn	
				ing NextClade		
				ing hexiciade		
Ļ	1: 2	1 1	Q			
	pangolin3	docker	String	Docker tag used	staphb/pangolin:3.1.11-pangolearn-	
				for running Pan-	2021-08-24	
L				golin		
	pangolin3	infer-	String	pangolin infer-	usher	
		ence_engine		ence engine for		
				lineage designa-		
				tions (usher or		
				pangolarn)		
Γ	pangolin3	min_length	Int	Minimum query	10000	
	1 0	- 0		length allowed		
				for pangolin		
				to attempt		
				assignment		
F	nangolin3	max amhig	Float	Maximum pro-	0.5	
	pungonno	max_among	1 Iout	portion of Ns al-	0.5	
				lowed for pan-		
				golin to attempt		
				goini to attempt		
L	thaia	martalada datasat	et train a	Assignment Nextelade		
		nextenade_dataset		INEXICIALE 01-	sais-cov-2	
Ļ	cov_clearlabs	. 1 1 1	06:	ganism dataset	NO1000047	
	theia-	nextclade_dataset	nstenegice	Nextclade refer-	MN908947	
L	cov_clearlabs			ence genome		
	theia-	nextclade_dataset	tagextclade	2021-06-		
L	cov_clearlabs		dataset tag	25T00:00:00Z		
	theia-	normalise	Int	Value to nor-	200	
	cov_clearlabs			malize read		
l				counts		
ſ	theia-	seq_method	String	Description of	ONT via Clear Labs WGS	
	cov_clearlabs			the sequencing		
1				methodology		
				used to generate		
				the input read		
				data		
ŀ	vadr	docker	String	Docker tag used	staphb/yadr:1.2.1	
			8	for running		
1				VADR		
L.			1		1	

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)	

Table	5 – conti	nued from	previous page	Э
-------	-----------	-----------	---------------	---

Outputs

Download CSV: TheiaCoV_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
assembly_method	String	Method employed to generate consensus assembly

Output Name	Data Type	Description
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
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 dehosting as determined by FastQC
fastqc_raw	Int	Number of raw 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	teElloat	Percent of human read data detected using the Kraken2 software af-
		ter host removal
kraken_report	String	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 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
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

1able 0 - continued non 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		

Table 6 – continued from previous page

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 Artic 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

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

Download CSV: TheiaCoV_ONT_required_inputs.csv



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	staphb/artic-ncov2019:1.3.0
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 asessing 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

Taal	Variable Name				7
Task			Description		_
kraken2_dehosted	kraken2_db	String	Path to the ref-	/kraken2-db	
			erence genome		
			within the		
			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		
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	:b7dba71079344daea4ea3363e1a67fa5	4edb7ec65459d03
nextclade one sa	m øb æker	String	Docker tag used	neherlab/nextclade:0.14.2	-
	1	8	for running		
			NextClade		
nextclade output	pdrxckreione sample	String	Docker tag	python:slim	_
	-r <u>-</u>	~8	used for pars-	F. 5	
			ing NextClade		
			output		
pangolin3	docker	String	Docker tag used	staphb/pangolin:3.1.11-pangolearn-	-
r			for running Pan-	2021-08-24	
			golin		
pangolin3	infer-	String	pangolin infer-	usher	4
rangonno	ence engine	~	ence engine for		
	ence_engine		lineage designa-		
			tions (usher or		
			nangolarn)		
1	1		puisoiuii)	1	1

Table 7 – continued from previous page

Task	Variable Name	Data Type	Description	Default
pangolin3	min length	Int	Minimum query	10000
pungonno	inin_iongtii	IIIt	length allowed	10000
			for pangolin	
			to attempt	
			assignment	
nongolin?	max ambig	Float	Maximum pro	0.5
pangoinis	max_among	Float	nortion of Na al	0.3
			portion of the ar-	
			Ioweu Ioi pall-	
		Test	CDU	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	n Stmin g	Nextclade or-	sars-cov-2
			ganism dataset	
theiacov_ont	nextclade_dataset	rSfeinegce	Nextclade refer-	MN908947
			ence genome	
theiacov_ont	nextclade_dataset	_tagextclade	2021-06-	
		dataset tag	25T00:00:00Z	
theiacov_ont	ar-	String	Version of the	V3
	tic_primer_version	n	Artic PCR	
			protocol used to	
			generate input	
			read data	
theiacov ont	normalise	Int	Value to nor-	200
_			malize read	
			counts	
theiacov ont	sea method	String	Description of	ONT
		6	the sequencing	
			methodology	
			used to generate	
			the input read	
			data	
theiacov ont	pan-	String	Docker tag used	staphb/pangolin·2 4 2-pangolearn-
liencev_ont	olin docker ima	op	for running Pan-	2021-05-19
	50111_00CKCI_IIIIa	5~	oolin	
vadr	docker	String	Docker tag used	stanbh/yadr:1 2 1
vaui	UULKU	Sumg	for running	Supilo/ vaul.1.2.1
			VADK	

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)	

Table 7 – continued from previous page	Fable	7 – continue	d from	previous	page
--	--------------	--------------	--------	----------	------

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

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-
		sis that includes the NextClade default samples for clade-typing and
		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 filttering 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	teHloat	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
meanbased trim	Float	Mean quality of the nucleotide basecalls aligned to the reference
		genome after primer trimming
meanmapg 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 ison	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
pango_meugo_report	String	Version of the pargolin 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 covers	ige is a second s	trimming: calculated as assembly length unambiguous / length of
	-0-	reference genome (SC2: 29 903) x 100
primer bed name	String	Name of the primer bed files used for primer trimming
nangolin versions	String	All Pangolin software and database versions
Pungoini_versions	Sumg	

Table 8 – continued from previous 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	e	
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	d 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.





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 gener-
	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: TheiaCoV_FASTA_optional_inputs.csv

Task	Variable Name	Data Type	Description	Default
nextclade_one sa	møbæker	String	Docker tag used	neherlab/nextclade:0.14.2
	-	5	for running	
			NextClade	
nextclade output	pdrxdreione sample	String	Docker tag	python:slim
	-r <u>-</u>	~8	used for pars-	F)
			ing NextClade	
			output	
pangolin3	docker	String	Docker tag used	staphb/pangolin:3,1,11-pangolearn-
Pungonno		SumB	for running Pan-	2021-08-24
			golin	2021 00 21
nangolin3	infer-	String	nangolin infer-	usher
pungonno	ence engine	buing	ence engine for	
	ence_engine		lineage designa-	
			tions (usher or	
			nangolarn)	
pangolin3	max ambig	Float	Maximum pro	0.5
pangonno	max_among	Tioat	nortion of Ne al	0.5
			lowed for pap	
			golin to attempt	
nongolin?	min langth	Int	Assignment	10000
pangoinis	iiiii_ieiigui	IIIt	longth allowed	10000
			for mongolin	
			for pangoin	
			to attempt	
titan fasta		-Otain -	Assignment Neutolodo or	
titan_fasta	nextclade_dataset	_nsumeg	Nextclade or-	sars-cov-2
	. 1 1 1	06:	ganism dataset	ND1000047
titan_fasta	nextclade_dataset	_notemegace	Nextclade refer-	MN908947
			ence genome	
titan_fasta	nextclade_dataset	_tagextclade	2021-06-	
		dataset tag	25100:00:00Z	
vadr	docker	String	Docker tag used	staphb/vadr:1.2.1
			for running	
	-	_	VADR	
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-
1.2. TheiaCoV V	vorkflow Series		VADR script	core,fstukcnf,insertnn,deletinn 35
				-mdir /opt/vadr/vadr-models/
version_capture	timezone	String	User time	None
- 1		č	zone in valid	

Outputs

Download CSV: 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 separate 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.

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

Download CSV: TheiaCoV_Augur_Prep_required_inputs.csv

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

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 identifier 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		

Download CSV: Mercury_PE_Prep_required_inputs.csv

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,	
~	mam siza ah	Int	Mamamy alla	1
gi-	mple	1111	intention y ano-	1
salu_prep_one_sa	unpre		said prep one sa	mple
			task	inpic
σi-	disk size	Int	Disk size al-	25
said prep one sa	mple	int	located to the gi-	
suid_prop_one_st	, in pro		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	ample		cated to the gi-	
			said_prep_one_sa	mple
		T	task	
g1-	pre-	Int	Number of	U
said_prep_one_sa	amenteptible_tries		preemptible	
			tries for the gi-	male
			salu_prep_one_sa	Inpre
ai	outbreak	String	Outbreak as	None
said prep one st	outoreak	Sumg	sociated with	
sau_prep_one_sa	unpro		this submission	
			e g date place	
			family cluster	
gi-	last vaccinated	String	Date of last vac-	None
said_prep_one_sa	mple		cine recieved	

Task	Input Variable	Data Type	Description	Default
σi-	docker image	String	Docker im-	guay jo/theiagen/utility 1 1
said pren one sa	mple	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	age utilized	quarity and agent and general
suid_prop_one_su	inpie		for the gi-	
			said prep one sa	mple
			task	inpre
gi-	passage details	String	Passage de-	original
said prep one sa	mple	8	tails of the	6
_1 1	1		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	
			iederal surveil-	
			iance program	
			baselline	
			surveinance	
			would be the	
			entry for this	
			field	
mer-	library layout	String	Layout of the se-	paired
cury pe pren		8	quenced library	r 2
mer-	num-	Int	Maximum num-	5000
cury_pe_prep	ber_N_threshold		ber of ambigu-	
			ous nucleotides	
			in a sample to	
			prepare submis-	
			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	

Table 9 – continued from previous page

lask	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	· -	U	the amplicons	
			sequenced	
mer-	host	String	Common name	Human
curv pe prep	noor	Sumg	of the host or-	
eury_pe_prep			ganism	
mor	omnli	String	gainsin Nama of the am	Nona
mer-	ampii-	Sung		INOILE
cury_pe_prep	con_primer_scher	ne	plicon primer	
			scheme utilized	
			to generate	
			the amplicons	
			sequenced	
mer-	biosam-	String	BioSample ac-	None
cury_pe_prep	ple_accession		cession number	
mer-	treatment	String	Treatment ad-	None
cury pe prep		-	ministered to	
			the patient, e.g.	
			drug name.	
			dosage etc	
mer	natient gender	String	Gender of the	unknown
curv pe prep	patient_gender	String	nationt	
mer	DUE	String	Passon that the	None
	pui-	Sung	reason mat me	None
cury_pe_prep	pose_or_sampring	,	original speci-	
			men was taken,	
			e.g. clinical	
		~ .	diagnostics	
mer-	patient_age	String	Age of the pa-	unknown
cury_pe_prep			tient	
ncbi_prep_one_sa	mplæm_size_gb	Int	Memory al-	1
			located to the	
			ncbi_prep_one_sa	mple
			task	
ncbi_prep_one_sa	mpdeker_image	String	Docker image	quay.io/staphb/vadr:1.3
		-	utilized for the	
			ncbi prep one sa	mple
			task	1
nchi pren one sa	am m bexlen	Int	VADR _maxlen	30000
neor_prop_one_st	inputrien	IIIt	input utilized	
			when trim	
			ming tomin-1	
			ambiguess and	
1.	1	T /	amoiguous ends	
ncb1_prep_one_sa	ampne-	Int	number of	0
	emptible_tries		preemptible	
			tries for the	
			ncbi_prep_one_sa	mple
			task	

Table 9 – continued from previous page

Task	Input Variable	Data Type	Description	Default
ncbi_prep_one_sa	m pR Us	Int	CPUs al-	1
			located to the	
			ncbi_prep_one_sa	mple
			task	
ncbi_prep_one_sa	m pie len	Int	VADR –minen	50
			input utilized	
			when trim-	
			ming terminal	
			ambiguous ends	
ncbi_prep_one_sa	mplikak_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)	

Table 9	 – continued 	from	previous pa	ade
---------	---------------------------------	------	-------------	-----

Outputs

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 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	1	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-	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
		0	be the most accurate entry for this field
mercury_pe_prep	instrument_model	String	Model of the sequencing instrument utilized
	1	Cturing a	to generate the read data
mercury_pe_prep	isolation_source	String	Isolation source, i.e. clinical, animal, or en-
	1:1	<u>Stains</u>	Vironmental
mercury_pe_prep	library_10	String String	Selection methodale as used to designate
mercury_pe_prep	norary_selection	String	selection methodology used to designate
			"PCP" for samples selected based on PCT
			Ct values
mercury ne pren	library source	String	Source of the genomic material used to pre-
mereury_pe_prep	norary_source	String	pare the sequencing libraries
mercury pe prep	library strategy	String	Library preparation strategy e.g. "AMPLI-
mereury_pe_prep	norary_strategy	String	CON" for data generated from tiling PCR
			amplicons
mercury pe prep	number N	Int	Number of fully ambiguous basecalls within
merear j_pe_prep			the consensus assembly
mercury pe prep	organism	String	Name of the organism sequenced, e.g.
merear j_pe_prep	organismi	Sumg	"SARS-CoV-2"
mercury pe prep	reads dehosted	File	Dehosted read files
mercury pe prep	seg 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 identifier 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_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,	
			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	mple		patient, e.g.	
			Hospitalized,	
			Released, Live,	
			Deceased,	
•		<u></u>	unknown	1
g1-	type	String	Organism typoe	betacoronovirus
said_prep_one_sa	imple CDU:	T	CDUs slls	None
g1-	CPUS	Int	CPUS allo-	None
said_prep_one_sa	unple		called to the gi-	mala
			salu_prep_one_sa	mpre
gi	pro	Int	Lask Number of	0
said prep one sa	pro-		number of	0
said_prep_one_se	inqueptible_tries		tries for the gi-	
			said prep one sa	mple
			task	inpre
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	

Task	Input Variable	Data Type	Description	Default
gi-	docker_image	String	Docker im-	quay.io/theiagen/utility:1.1
said_prep_one_sa	mple	_	age utilized	
			for the gi-	
			said_prep_one_sa	mple
			task	
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		T .	quenced library	5000
mer-	num-	Int	Maximum num-	5000
cury_pe_prep	ber_N_threshold		ber of ambigu-	
			ous nucleotides	
			in a sample to	
			prepare submis-	
	1	Cturing a	sion files	
mer-	nost_sc1_name	String	of the heat	Homo sapiens
cury_pe_prep			of the nost or-	
		Stains	gamsm	None
mer-	gi-	String	Accession num-	INOILE
cury_pe_prep	salu_accession	Stains	Orgionary	hCoV 10
mer-	gisaid_organism	String	Orgiansm name	1100 - 19
cury_pe_prep			as per OISAID	
1	1		suomission	

Table 10 – continued from previous page

Task	Input Variable	Data Type	Description	Default
mer-	county	String	County the lab-	None
curv pe prep	county	Sums	oratory was col-	
eary_pe_prop			lected in	
mer-	amplicon size	String	Average size of	None
curv pe prep	umpneen_bille	Sumg	the amplicons	
eary_pe_prop			sequenced	
mer-	host	String	Common name	Human
curv pe prep	noor	Sums	of the host or-	
eary_pe_prep			ganism	
mer-	ampli-	String	Name of the am-	None
curv pe prep	con primer scher	me	plicon primer	
			scheme utilized	
			to generate	
			the amplicons	
			sequenced	
mer-	biosam-	String	BioSample ac-	None
curv pe prep	ple accession		cession number	
mer-	treatment	String	Treatment ad-	None
cury pe prep			ministered to	
<i>J</i> <u> </u>			the patient, e.g.	
			drug name,	
			dosage, etc.	
mer-	patient_gender	String	Gender of the	unknown
cury_pe_prep			patient	
mer-	pur-	String	Reason that the	None
cury_pe_prep	pose_of_sampling	3	original speci-	
			men was taken,	
			e.g. clinical	
			diagnostics	
mer-	patient_age	String	Age of the pa-	unknown
cury_pe_prep			tient	
ncbi_prep_one_sa	amplæm_size_gb	Int	Memory al-	1
			located to the	
			ncbi_prep_one_sa	mple
			task	
ncbi_prep_one_sa	am ple ker_image	String	Docker image	quay.io/staphb/vadr:1.3
			utilized for the	
			ncbi_prep_one_sa	mple
			task	
ncbi_prep_one_sa	amplæxlen	Int	VADR –maxlen	30000
			input utilized	
			when trim-	
			ming terminal	
			ambiguous ends	-
ncbi_prep_one_sa	anpprice-	Int	Number of	0
	emptible_tries		preemptible	
			tries for the	
			ncbi_prep_one_sa	mple
			task	

Table 10 – continued from previous page

Task	Input Variable	Data Type	Description	Default
ncbi_prep_one_sa	m f PRUs	Int	CPUs al-	1
			located to the	
			ncbi_prep_one_sa	mple
			task	
ncbi_prep_one_sa	m pie len	Int	VADR –minen	50
			input utilized	
			when trim-	
			ming terminal	
			ambiguous ends	
ncbi_prep_one_sa	mplikak_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)	

Table	10 – con	tinued from	previous	page
iabio	10 001		proviouo	pago

Outputs

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	1	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

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[File]	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

Download CSV: Mercury_Batch_required_inputs.csv

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_s	sra		age utilized	
			for the com-	
			pile_biosample_n	sra
			task	
com-	pre-	Int	Number of pre-	0
pile_biosamp_n_s	ræmptible_tries		emptible tries	
			for the com-	
			pile_biosample_n	sra
			task	
gen-	docker_image	String	Docker im-	quay.io/theiagen/utility:1.1
bank_compile			age utilized	
			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	
			task	
gisaid_compile	pre-	Int	Number of	0
	emptible_tries		preemptible	
			tries for the	
			gisaid_compile	
	CDU	Tut	task	
mercury_batch	CPUs	Int	CPUs allocated	4
			in the mor	
			iii uie iiiei-	
			cury_batch	
margury batch	dick size	Int	Disk size allo	100
mercury_batch	UISK_SIZE		Disk size and-	100
			task in the	
			mercury batch	
			workflow	
mercury batch	gen bucket	String	GCP bucket for	None
	Sep_sucket	Sums	SRA transfer	
mercury batch	mem size gb	Int	Memory allo-	8
			cated for each	
			task in the	
			mercury batch	
			workflow	
mercury_batch	vadr_threshold	Int	Maximum num-	0
			ber of VADR	
			alerts for sam-	
			ples included	
			in the batch	
			submission files	
version_capture	timezone	String	User time	None
1.0			zone in valid	
1.3. Mercury Wo	orktiow Series		Unix TZ string	53
			(e.g. Amer-	
			ica/New_York)	

Outputs

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