Public Health Viral Genomics (Theiagen)

Release 2.0.0

Kevin G. Libuit

Apr 13, 2022

CONTENTS

1 Contents **1**

1.1 Public Health Viral Genomics 1

1.2 TheiaCoV Workflow Series 1

1.3 Mercury Workflow Series 38

1.4 License 54

CHAPTER

ONE

CONTENTS

1.1 Public Health Viral Genomics

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

1.1.1 Getting Started

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

1.1.2 Support

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

1.2 TheiaCoV Workflow Series

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

1.2.1 TheiaCoV Workflows for Genomic Characterization

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

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

All four TheiaCoV workflows for genomic characterization will generate a viral assembly by mapping input read data to a reference genome, removing primer reads from that alignment, and then calling the consensus assembly based on the primer-trimmed alignment. These consensus assemblies are then fed into the Pangolin and NextClade CLI tools for lineage and clade assignments.

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

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

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

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

TheiaCoV_Illumina_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 x 150 bp reads). Modifications to the optional parameter for trimmomatic_minlen may be required to accommodate for shorter read data, such as 2 x 75bp reads generated using a 150-cycle kit.

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

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

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

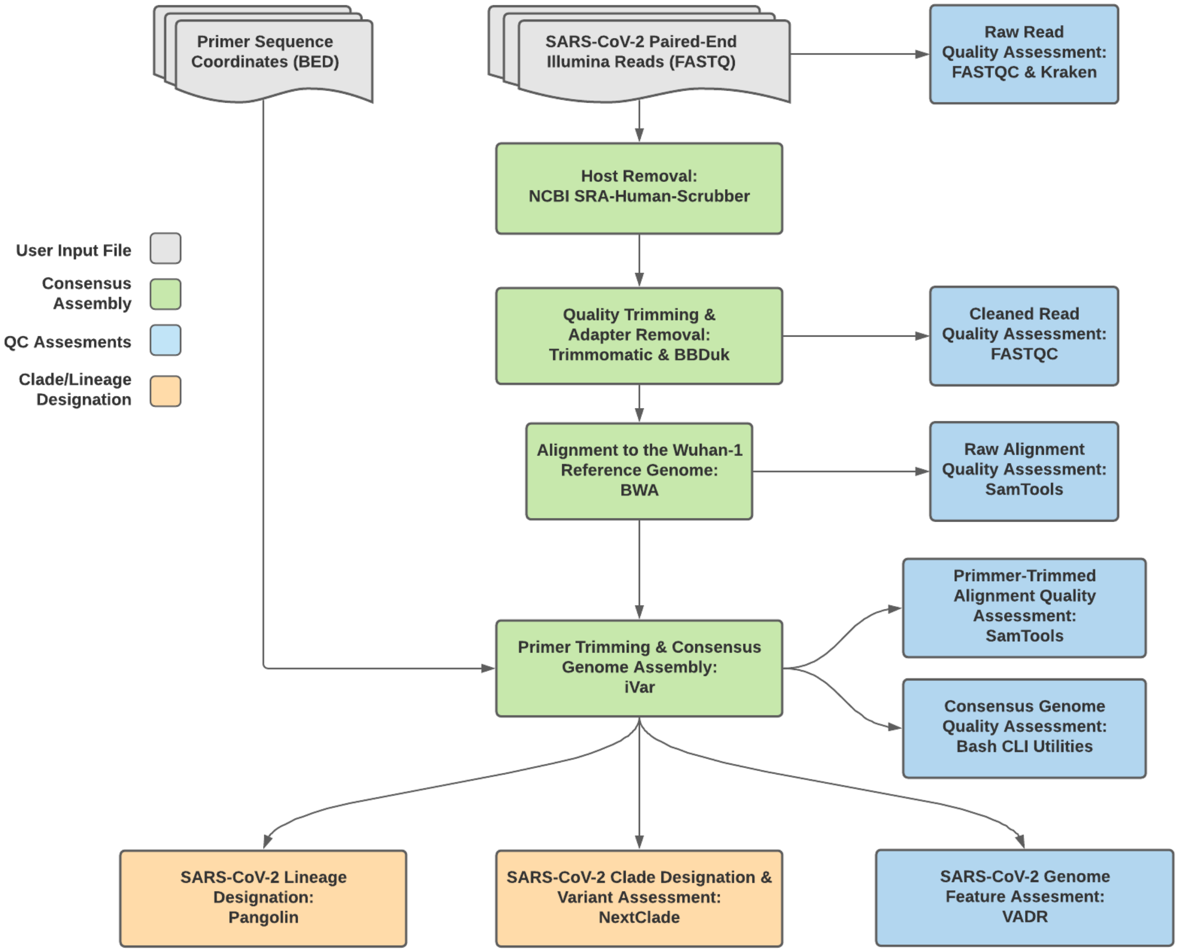
Primer Sequence Coordinates (BED)
SARS-CoV-2 Paired-End Illumina Reads (FASTQ)
Raw Read Quality Assessment: FASTQC & Kraken
Host Removal: NCBI SRA-Human-Scrubber
User Input File
Consensus Assembly
QC Assesments
Clade/Lineage Designation
Quality Trimming & Adapter Removal: Trimmomatic & BBDuk
Cleaned Read Quality Assessment: FASTQC
Alignment to the Wuhan-1 Reference Genome: BWA
Raw Alignment Quality Assessment: SamTools
Primmer-Trimmed Alignment Quality Assessment: SamTools
Primer Trimming & Consensus Genome Assembly: iVar
Consensus Genome Quality Assessment: Bash CLI Utilities
SARS-CoV-2 Lineage Designation: Pangolin
SARS-CoV-2 Clade Designation & Variant Assessment: NextClade
SARS-CoV-2 Genome Feature Assesment: VADR

Fig. 1: **TheiaCoV_Illumina_PE Data Workflow**

Required User Inputs

Download CSV: `TheiaCoV_Illumina_PE_required_inputs.csv`

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

Optional User Inputs

Download CSV: `TheiaCoV_Illumina_PE_optional_inputs.csv`

Task	Variable Name	Data Type	Description	Default
bwa	reference_genome	String	Path to the reference genome within the staphb/ivar:1.2.2_artic20200528 Docker container	/artic-ncov2019/primer_schemes/nCoV-2019/V3/nCoV-2019.reference.fasta
bwa	cpus	Int	CPU resources allocated to the BWA task runtime environment	6
consensus	char_unknown	String	Character to print in regions with less than minimum coverage for iVar consensus	N
consensus	count_orphans	Boolean	Do not skip anomalous read pairs in variant calling for SAMtools mpileup before running iVar consensus	TRUE

continues on next page

Table 1 – continued from previous page

Task	Variable Name	Data Type	Description	Default
consensus	disable_baq	Boolean	Disable read-pair overlap detection for SAMtools mpileup before running iVar consensus	TRUE
consensus	max_depth	Int	Maximum reads read at a position per input file for SAMtools mpileup before running iVar consensus	600000
consensus	min_bq	Int	Minimum mapping quality for an alignment to be used for SAMtools mpileup before running iVar consensus	0
consensus	min_depth	Int	Minimum read depth to call variants for iVar consensus	100
consensus	min_freq	Float	Minimum frequency threshold(0 - 1) to call variants for iVar consensus	0.6
consensus	min_qual	Int	Minimum quality threshold for sliding window to pass for iVar consensus	20
consensus	ref_genome	String	Path to the reference genome within the staphb/ivar:1.2.2_artic20200528 Docker container	/artic-ncov2019/primer_schemes/nCoV-2019/V3/nCoV-2019.reference.fasta

continues on next page

Table 1 – continued from previous page

Task	Variable Name	Data Type	Description	Default
consensus	ref_gff	String	Path to the general feature format of the reference genome within the staphb/ivar:1.2.2_artic20200528 Docker container	/reference/GCF_009858895.2_ASM985889v3_genomic.gff
nextclade_one_sample	docker	String	Docker tag used for running NextClade	nextstrain/nextclade:1.10.3
nextclade_output_parser_one_sample	docker	String	Docker tag used for parsing NextClade output	python:slim
pangolin3	docker	String	Docker tag used for running Pangolin	quay.io/staphb/3.1.20-pangolearn-2022-02-02
pangolin3	inference_engine	String	pangolin inference engine for lineage designations (usher or pangolarn)	usher
pangolin3	min_length	Int	Minimum query length allowed for pangolin to attempt assignment	10000
pangolin3	max_ambig	Float	Maximum proportion of Ns allowed for pangolin to attempt assignment	0.5
primer_trim	keep_noprimer_reads	Boolean	Include reads with no primers for iVar trim	True
read_QC_trim	bbduk_mem	Int	Memory allocated to the BBDuk VM	8
read_QC_trim	trimmomatic_minlen	Int	Specifies the minimum length of reads to be kept for Trimmomatic	25
read_QC_trim	trimmomatic_quality_trim_score	Int	Specifies the average quality required for Trimmomatic	30

continues on next page

Table 1 – continued from previous page

Task	Variable Name	Data Type	Description	Default
read_QC_trim	trimmomatic_window_size	Int	Specifies the number of bases to average across for Trimmomatic	4
theiacov_illumina_pe	nextclade_dataset_name	String	Nextclade organism dataset	sars-cov-2
theiacov_illumina_pe	nextclade_dataset_reference	String	Nextclade reference genome	MN908947
theiacov_illumina_pe	nextclade_dataset_tag	Nextclade dataset tag	2022-02-07T12:00:00Z	
theiacov_illumina_pe	seq_method	String	Description of the sequencing methodology used to generate the input read data	Illumina paired-end
vadr	docker	String	Docker tag used for running VADR	quay.io/staphb/1.4.1-models-1.3-2
vadr	maxlen	Int	Maximum length for the fasta-trim-terminal-ambigs.pl VADR script	30000
vadr	minlen	Int	Minimum length subsequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl VADR script	50
vadr	skip_length	Int	Minimum assembly length (unambiguous) to run vadr	10000
vadr	vadr_opts	String	Options for the v-annotate.pl VADR script	–glsearch -s -r –nomisc –mkey sarscov2 –alt_fail lowscore,fstukcnf,insertnn,deletinn –mdir /opt/vadr/vadr-models/
variant_call	count_orphans	Boolean	Do not skip anomalous read pairs in variant calling for SAMtools mpileup before running iVar variants	TRUE

continues on next page

Table 1 – continued from previous page

Task	Variable Name	Data Type	Description	Default
variant_call	disable_baq	Boolean	Disable read-pair overlap detection for SAMtools mpileup before running iVar variants	TRUE
variant_call	max_depth	Int	Maximum reads read at a position per input file for SAMtools mpileup before running iVar variants	600000
variant_call	min_bq	Int	Minimum mapping quality for an alignment to be used for SAMtools mpileup before running iVar variants	0
variant_call	min_depth	Int	Minimum read depth to call variants for iVar variants	100
variant_call	min_freq	Float	Minimum frequency threshold(0 - 1) to call variants for iVar variants	0.6
variant_call	min_qual	Int	Minimum quality threshold for sliding window to pass for iVar variants	20
variant_call	ref_gff	String	Path to the general feature format of the reference genome within the staphb/ivar:1.2.2_artic20200528 Docker container	/reference/GCF_009858895.2_ASM985889v3_genomic.gff
variant_call	ref_genome	String	Path to the reference genome within the staphb/ivar:1.2.2_artic20200528 Docker container	/artic-ncov2019/primer_schemes/nCoV-2019/V3/nCoV-2019.reference.fasta

continues on next page

Table 1 – continued from previous page

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

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
assembly_length_unambiguous	Int	Number of unambiguous basecalls within the SC2 consensus assembly
assembly_mean_coverage	Float	Mean sequencing depth throughout the conesnsus assembly generated after performing primer trimming–calculated using the SAMtools coverage command
assembly_method	String	Method employed to generate consensus assembly
auspice_json	File	Auspice-compatable JSON output generated from NextClade analysis 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 seqyclean filtering as determined by FastQC
fastqc_clean2	Int	Number of reverse reads after seqyclean 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 determined by FastQC
fastqc_raw2	Int	Number of reverse reads identified in the input fastq files as determined by FastQC
fastqc_raw_pairs	String	Number of paired reads identified in the input fastq files 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

continues on next page

Table 2 – continued from previous page

Output Name	Data Type	Description
ivar_variant_version	String	Version of iVar for running the iVar variants command
ivar_vcf	File	iVar tsv output converted to VCF format
ivar_version_consensus	String	Version of iVar for running the iVar consensus command
ivar_version_primtrim	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_dehosted	Float	Percent of human read data detected using the Kraken2 software after host removal
kraken_report	File	Full Kraken report
kraken_report_dehosted	File	Full Kraken report after host removal
kraken_sc2	Float	Percent of SARS-CoV-2 read data detected using the Kraken2 software
kraken_sc2_dehosted	Float	Percent of SARS-CoV-2 read data detected using the Kraken2 software 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
pangolin_assignment_version	String	Version of the pangolin software (e.g. PANGO or PUSHER) used 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
percent_reference_coverage	Float	Percent coverage of the reference genome after performing primer 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_percent	Float	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 submissionsamtools_version
samtools_version	String	Version of SAMtools used to sort and index the alignment file
samtools_version_consensus	String	Version of SAMtools used to create the pileup before running iVar consensus

continues on next page

Table 2 – continued from previous page

Output Name	Data Type	Description
samtools_version_primtrim	String	Version of SAMtools used to create the pileup before running iVar trim
samtools_version_stats	String	Version of SAMtools used to assess quality of read mapping
seq_platform	String	Description of the sequencing methodology used to generate the input read data
theiacov_illumina_pe_analysis_date	String	Date of analysis
theiacov_illumina_pe_version	String	Version of the Public Health Viral Genomics (PHVG) repository used
trimmomatic_version	String	Version of Trimmomatic 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

TheiaCoV_Illumina_SE

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

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

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

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

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

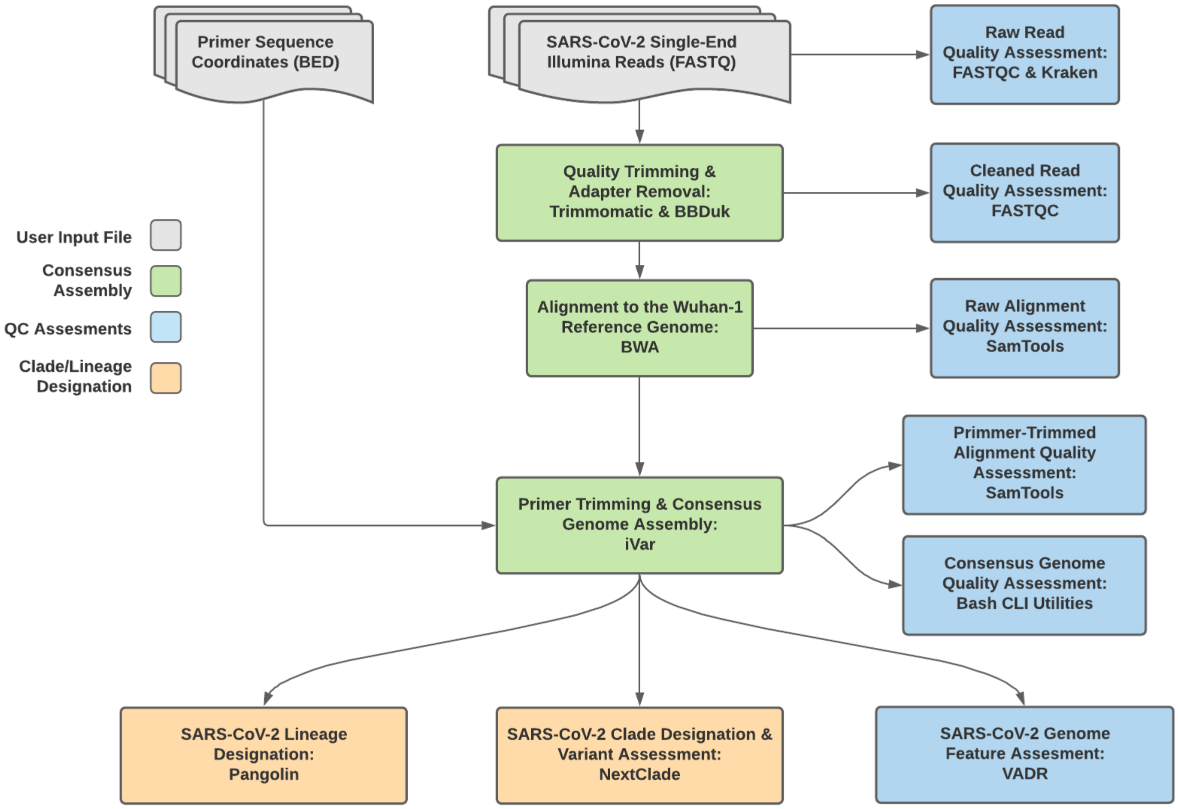
Primer Sequence Coordinates (BED)
SARS-CoV-2 Single-End Illumina Reads (FASTQ)
Raw Read Quality Assessment: FASTQC & Kraken
Quality Trimming & Adapter Removal: Trimmomatic & BBDuk
Cleaned Read Quality Assessment: FASTQC
User Input File
Consensus Assembly
QC Assesments
Clade/Lineage Designation
Alignment to the Wuhan-1 Reference Genome: BWA
Raw Alignment Quality Assessment: SamTools
Primmer-Trimmed Alignment Quality Assessment: SamTools
Primer Trimming & Consensus Genome Assembly: iVar
Consensus Genome Quality Assessment: Bash CLI Utilities
SARS-CoV-2 Lineage Designation: Pangolin
SARS-CoV-2 Clade Designation & Variant Assessment: NextClade
SARS-CoV-2 Genome Feature Assesment: VADR

Fig. 2: **TheiaCoV_Illumina_SE Data Workflow**

Required User Inputs

Download CSV: `TheiaCoV_Illumina_SE_required_inputs.csv`

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

Optional User Inputs

Download CSV: `TheiaCoV_Illumina_SE_optional_inputs.csv`

Task	Variable Name	Data Type	Description	Default
bwa	reference_genome	String	Path to the reference genome within the staphb/ivar:1.2.2_artic20200528 Docker container	/artic-ncov2019/primer_schemes/nCoV-2019/V3/nCoV-2019.reference.fasta
bwa	cpus	Int	CPU resources allocated to the BWA task runtime environment	6
bwa	read2	File	Optional input file for the Kraken task that is not applicable to this workflow	None
consensus	char_unknown	String	Character to print in regions with less than minimum coverage for iVar consensus	N
consensus	count_orphans	Boolean	Do not skip anomalous read pairs in variant calling for SAMtools mpileup before running iVar consensus	TRUE

continues on next page

Table 3 – continued from previous page

Task	Variable Name	Data Type	Description	Default
consensus	disable_baq	Boolean	Disable read-pair overlap detection for SAMtools mpileup before running iVar consensus	TRUE
consensus	max_depth	Int	Maximum reads read at a position per input file for SAMtools mpileup before running iVar consensus	600000
consensus	min_bq	Int	Minimum mapping quality for an alignment to be used for SAMtools mpileup before running iVar consensus	0
consensus	min_depth	Int	Minimum read depth to call variants for iVar consensus	100
consensus	min_freq	Float	Minimum frequency threshold(0 - 1) to call variants for iVar consensus	0.6
consensus	min_qual	Int	Minimum quality threshold for sliding window to pass for iVar consensus	20
consensus	ref_genome	String	Path to the reference genome within the staphb/ivar:1.2.2_artic20200528 Docker container	/artic-ncov2019/primer_schemes/nCoV-2019/V3/nCoV-2019.reference.fasta

continues on next page

Table 3 – continued from previous page

Task	Variable Name	Data Type	Description	Default
consensus	ref_gff	String	Path to the general feature format of the reference genome within the staphb/ivar:1.2.2_artic20200528 Docker container	/reference/GCF_009858895.2_ASM985889v3_genomic.gff
nextclade_one_sample	docker	String	Docker tag used for running NextClade	nextstrain/nextclade:1.10.3
nextclade_output_parser_one_sample	docker	String	Docker tag used for parsing NextClade output	python:slim
pangolin3	docker	String	Docker tag used for running Pangolin	quay.io/staphb/3.1.20-pangolearn-2022-02-02
pangolin3	inference_engine	String	pangolin inference engine for lineage designations (usher or pangolarn)	usher
pangolin3	min_length	Int	Minimum query length allowed for pangolin to attempt assignment	10000
pangolin3	max_ambig	Float	Maximum proportion of Ns allowed for pangolin to attempt assignment	0.5
primer_trim	keep_noprimer_reads	Boolean	Include reads with no primers for iVar trim	True
read_QC_trim	bbduk_mem	Int	Memory allocated to the BBDuk VM	8
read_QC_trim	trimmomatic_minlen	Int	Specifies the minimum length of reads to be kept for Trimmomatic	25
read_QC_trim	trimmomatic_quality_trim_score	Int	Specifies the average quality required for Trimmomatic	30

continues on next page

Table 3 – continued from previous page

Task	Variable Name	Data Type	Description	Default
read_QC_trim	trimmomatic_window_size	Int	Specifies the number of bases to average across for Trimmomatic	4
theiacov_illumina_se	nextclade_dataset_name	String	Nextclade organism dataset	sars-cov-2
theiacov_illumina_se	nextclade_dataset_reference	String	Nextclade reference genome	MN908947
theiacov_illumina_se	nextclade_dataset_tag	Nextclade dataset tag	2022-02-07T12:00:00Z	
theiacov_illumina_se	seq_method	String	Description of the sequencing methodology used to generate the input read data	Illumina paired-end
vadr	docker	String	Docker tag used for running VADR	quay.io/staphb/1.4.1-models-1.3-2
vadr	maxlen	Int	Maximum length for the fasta-trim-terminal-ambigs.pl VADR script	30000
vadr	minlen	Int	Minimum length subsequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl VADR script	50
vadr	skip_length	Int	Minimum assembly length (unambiguous) to run vadr	10000
vadr	vadr_opts	String	Options for the v-annotate.pl VADR script	–glsearch -s -r –nomisc –mkey sarscov2 –alt_fail lowscore,fstukcnf,insertnn,deletinn –mdir /opt/vadr/vadr-models/
variant_call	count_orphans	Boolean	Do not skip anomalous read pairs in variant calling for SAMtools mpileup before running iVar variants	TRUE

continues on next page

Table 3 – continued from previous page

Task	Variable Name	Data Type	Description	Default
variant_call	disable_baq	Boolean	Disable read-pair overlap detection for SAMtools mpileup before running iVar variants	TRUE
variant_call	max_depth	Int	Maximum reads read at a position per input file for SAMtools mpileup before running iVar variants	600000
variant_call	min_bq	Int	Minimum mapping quality for an alignment to be used for SAMtools mpileup before running iVar variants	0
variant_call	min_depth	Int	Minimum read depth to call variants for iVar variants	100
variant_call	min_freq	Float	Minimum frequency threshold(0 - 1) to call variants for iVar variants	0.6
variant_call	min_qual	Int	Minimum quality threshold for sliding window to pass for iVar variants	20
variant_call	ref_gff	String	Path to the general feature format of the reference genome within the staphb/ivar:1.2.2_artic20200528 Docker container	/reference/GCF_009858895.2_ASM985889v3_genomic.gff
variant_call	ref_genome	String	Path to the reference genome within the staphb/ivar:1.2.2_artic20200528 Docker container	/artic-ncov2019/primer_schemes/nCoV-2019/V3/nCoV-2019.reference.fasta

continues on next page

Table 3 – continued from previous page

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

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
assembly_length_unambiguous	Int	Number of unambiguous basecalls within the SC2 consensus assembly
assembly_mean_coverage	Float	Mean sequencing depth throughout the conesnsus assembly generated after performing primer trimming–calculated using the SAMtools coverage command
assembly_method	String	Method employed to generate consensus assembly
auspice_json	File	Auspice-compatable JSON output generated from NextClade analysis 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 seqyclean 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_consensus	String	Version of iVar for running the iVar consensus command
ivar_version_primtrim	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 software
kraken_version	String	Version of Kraken software used

continues on next page

Table 4 – continued from previous page

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
pangolin_assignment_version	String	Version of the pangolin software (e.g. PANGO or PUSHER) used 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
percent_reference_coverage	Float	Percent coverage of the reference genome after performing primer 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_percent	Float	Percent of read data with primers trimmed as deteremined by iVar trim
read1_clean	File	Forward read file after quality trimming and adapter removal
samtools_version	String	Version of SAMtools used to sort and index the alignment file
samtools_version_consensus	String	Version of SAMtools used to create the pileup before running iVar consensus
samtools_version_primtrim	String	Version of SAMtools used to create the pileup before running iVar trim
samtools_version_stats	String	Version of SAMtools used to assess quality of read mapping
seq_platform	String	Description of the sequencing methodology used to generate the input read data
theiacov_illumina_se_analysis_date	String	Date of analysis
theiacov_illumina_se_version	String	Version of the Public Health Viral Genomics (PHVG) repository used
trimmomatic_version	String	Version of Trimmomatic 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

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 avaialble 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.

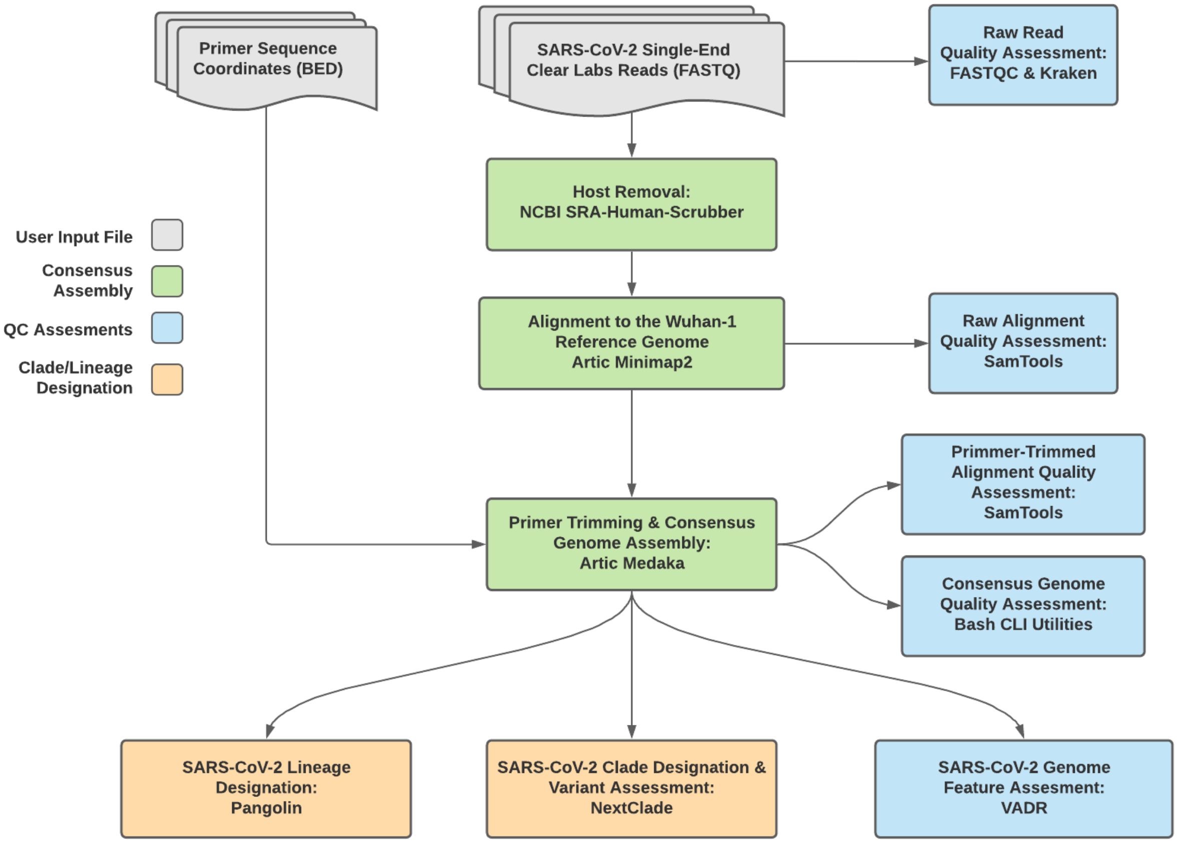
Primer Sequence Coordinates (BED)
SARS-CoV-2 Single-End Clear Labs Reads (FASTQ)
Raw Read Quality Assessment: FASTQC & Kraken
Host Removal: NCBI SRA-Human-Scrubber
User Input File
Consensus Assembly
QC Assesments
Clade/Lineage Designation
Alignment to the Wuhan-1 Reference Genome Artic Minimap2
Raw Alignment Quality Assessment: SamTools
Primmer-Trimmed Alignment Quality Assessment: SamTools
Primer Trimming & Consensus Genome Assembly: Artic Medaka
Consensus Genome Quality Assessment: Bash CLI Utilities
SARS-CoV-2 Lineage Designation: Pangolin
SARS-CoV-2 Clade Designation & Variant Assessment: NextClade
SARS-CoV-2 Genome Feature Assesment: VADR

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 Artric Medaka task runtime environment	8
consensus	docker	String	Docker tag used for running Medaka assemblyer	quay.io/staphb/artic-ncov2019:1.3.0-medaka-1.4.3
consensus	medaka_model	String	Model for consensus genome assembly via Medaka	r941_min_high_g360
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 sample being analyzed	Inferred from the input read file-fastqc_se_clean

continues on next page

Table 5 – continued from previous page

Task	Variable Name	Data Type	Description	Default
fastqc_se_raw	cpus	Int	CPU resources allocated to the FastQC task runtime environment for asessing raw read data	
fastqc_se_raw	read1_name	String	Name of the 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
kraken2_dehosted	kraken2_db	String	Path to the reference genome within the staphb/kraken2:2.0.8-beta_hv Docker container	/kraken2-db
kraken2_dehosted	read2	File	Optional input file for the Kraken task that is not applicable to this workflow	None
kraken2_raw	cpus	Int	CPU resources allocated to the Kraken task runtime environment for asessing raw read data	4
kraken2_raw	kraken2_db	String	Path to the reference genome within the staphb/kraken2:2.0.8-beta_hv Docker container	/kraken2-db
kraken2_raw	read2	File	Optional input file for the Kraken task that is not applicable to this workflow	None

continues on next page

Table 5 – continued from previous page

Task	Variable Name	Data Type	Description	Default
ncbi_scrub_se	docker	Docker tag used for running the NCBI SRA Human-Scruber tool	gcr.io/ncbi-sys-gcr-public-research/sra-human-scrubber@sha256:b7dba71079344daea4ea3363e1a67fa54edb7ec65459d03	
nextclade_one_sample	docker	String	Docker tag used for running NextClade	nextstrain/nextclade:1.10.3
nextclade_output_parser_one_sample	docker	String	Docker tag used for parsing NextClade output	python:slim
pangolin3	docker	String	Docker tag used for running Pangolin	quay.io/staphb/3.1.20-pangolearn-2022-02-02
pangolin3	inference_engine	String	pangolin inference engine for lineage designations (usher or pangolarn)	usher
pangolin3	min_length	Int	Minimum query length allowed for pangolin to attempt assignment	10000
pangolin3	max_ambig	Float	Maximum proportion of Ns allowed for pangolin to attempt assignment	0.5
theiacov_clearlabs	nextclade_dataset_name	String	Nextclade organism dataset	sars-cov-2
theiacov_clearlabs	nextclade_dataset_reference	String	Nextclade reference genome	MN908947
theiacov_clearlabs	nextclade_dataset_tag	Nextclade dataset tag	2022-02-07T12:00:00Z	
theiacov_clearlabs	normalise	Int	Value to normalize read counts	200
theiacov_clearlabs	seq_method	String	Description of the sequencing methodology used to generate the input read data	ONT via Clear Labs WGS
vadr	docker	String	Docker tag used for running VADR	quay.io/staphb/1.4.1-models-1.3-2

continues on next page

Table 5 – continued from previous page

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

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
assembly_length_unambiguous	Int	Number of unambiguous basecalls within the SC2 consensus assembly
assembly_mean_coverage	Float	Mean sequencing depth throughout the conesnsus assembly generated after performing primer trimming–calculated using the SAMtools coverage command
assembly_method	String	Method employed to generate consensus assembly

continues on next page

Table 6 – continued from previous page

Output Name	Data Type	Description
auspice_json	File	Auspice-compatable JSON output generated from NextClade analysis 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_dehosted	Float	Percent of human read data detected using the Kraken2 software after host removal
kraken_report	String	Full Kraken report
kraken_report_dehosted	File	Full Kraken report after host removal
kraken_sc2	Float	Percent of SARS-CoV-2 read data detected using the Kraken2 software
kraken_sc2_dehosted	Float	Percent of SARS-CoV-2 read data detected using the Kraken2 software 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
pangolin_assignment_version	String	Version of the pangolin software (e.g. PANGO or PUSHER) used 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
percent_reference_coverage	Float	Percent coverage of the reference genome after performing primer 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 input read data

continues on next page

Table 6 – continued from previous page

Output Name	Data Type	Description
theiacov_clearlabs_analysis_date	String	Date of analysis
theiacov_clearlabs_version	String	Version of the Public Health Viral Genomics (PHVG) repository 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
variants_from_ref_vcf	File	Number of variants relative to the reference genome

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 Aritc guppyplex command. These size-selected read data are aligned to the Wuhan-1 reference genome with minimap2 to generate a Binary Alignment Mapping (BAM) file. Primer sequences are then removed from the BAM file and a consensus assembly file is generated using the Artic medaka command. This assembly is then used to assign lineage and clade designations with Pangolin and NextClade. NCBI'S VADR tool is also employed to screen for potentially errant features (e.g. erroneous frame-shift mutations) in the consensus assembly.

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

Required User Inputs

Download CSV: `TheiaCoV_ONT_required_inputs.csv`

Task	Input Variable	Data Type	Description
theiacov_ont	demultiplexed_reads	File	Basecalled and demultiplexed ONT read 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

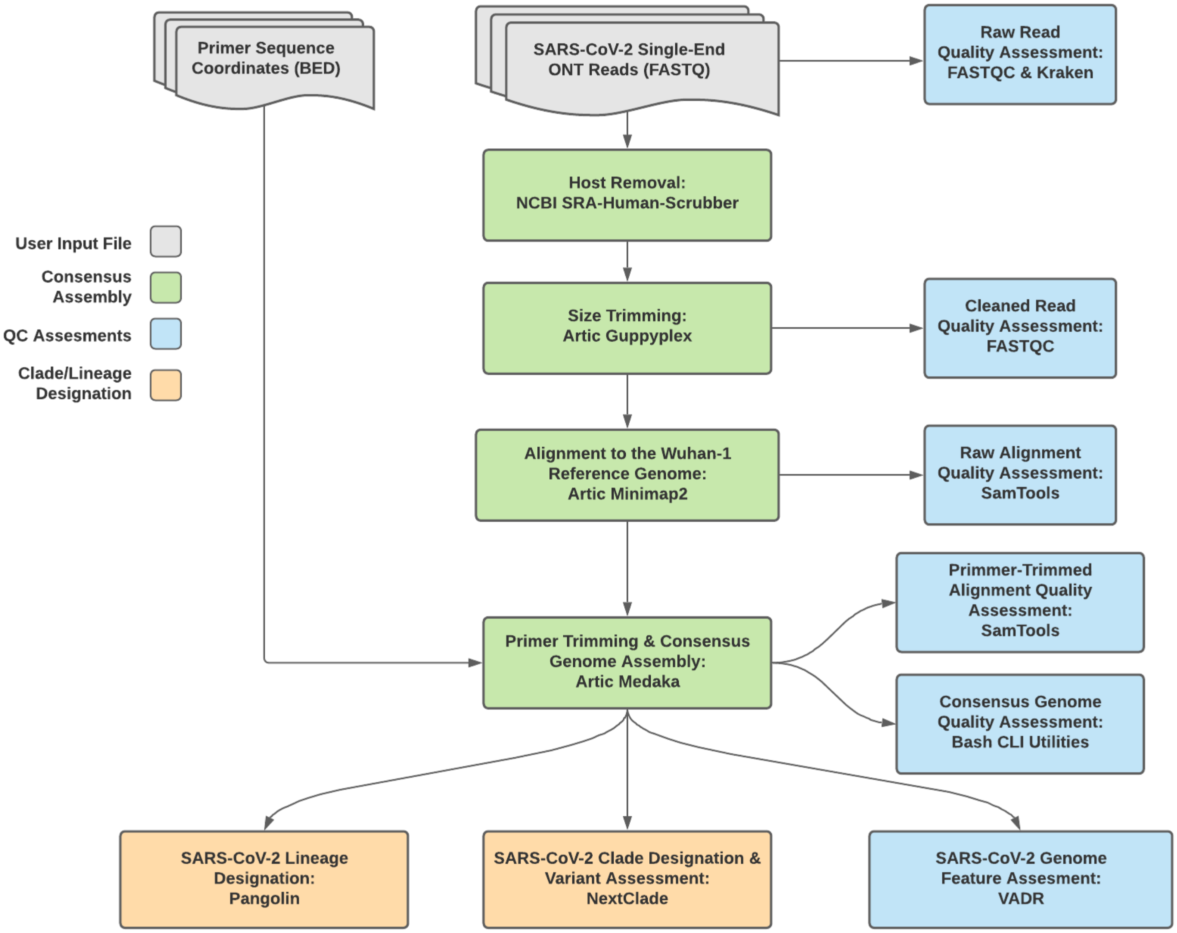
Primer Sequence Coordinates (BED)
SARS-CoV-2 Single-End ONT Reads (FASTQ)
Raw Read Quality Assessment: FASTQC & Kraken
Host Removal: NCBI SRA-Human-Scrubber
User Input File
Consensus Assembly
QC Assesments
Clade/Lineage Designation
Size Trimming: Artic Guppyplex
Cleaned Read Quality Assessment: FASTQC
Alignment to the Wuhan-1 Reference Genome: Artic Minimap2
Raw Alignment Quality Assessment: SamTools
Primmer-Trimmed Alignment Quality Assessment: SamTools
Primer Trimming & Consensus Genome Assembly: Artic Medaka
Consensus Genome Quality Assessment: Bash CLI Utilities
SARS-CoV-2 Lineage Designation: Pangolin
SARS-CoV-2 Clade Designation & Variant Assessment: NextClade
SARS-CoV-2 Genome Feature Assesment: VADR

Fig. 4: **TheiaCoV_ONT Data Workflow**

Optional User Inputs

Download CSV: `TheiaCoV_ONT_optional_inputs.csv`

Task	Variable Name	Data Type	Description	Default
consensus	cpu	Int	CPU resources allocated to the Artric Medaka task runtime environment	
consensus	docker	String	Docker tag used for running Medaka assemblyer	quay.io/staphb/artic-ncov2019-epi2me
consensus	medaka_model	String	Model for consensus 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

continues on next page

Table 7 – continued from previous page

Task	Variable Name	Data Type	Description	Default
kraken2_dehosted	kraken2_db	String	Path to the reference genome within the staphb/kraken2:2.0.8-beta_hv Docker container	/kraken2-db
kraken2_dehosted	read2	File	Optional input file for the Kraken task that is not applicable to this workflow	None
kraken2_raw	cpus	Int	CPU resources allocated to the Kraken task runtime environment for asessing raw read data	4
kraken2_raw	kraken2_db	String	Path to the reference genome within the staphb/kraken2:2.0.8-beta_hv Docker container	/kraken2-db
kraken2_raw	read2	File	Optional input file for the Kraken task that is not applicable to this workflow	None
ncbi_scrub_se	docker	Docker tag used for running the NCBI SRA Human-Scruber tool	gcr.io/ncbi-sys-gcr-public-research/sra-human-scrubber@sha256:b7dba71079344daea4ea3363e1a67fa54edb7ec65459d03	
nextclade_one_sample	docker	String	Docker tag used for running NextClade	nextstrain/nextclade:1.10.3
nextclade_output_parser_one_sample	docker	String	Docker tag used for parsing NextClade output	python:slim
pangolin3	docker	String	Docker tag used for running Pangolin	quay.io/staphb/3.1.20-pangolearn-2022-02-02
pangolin3	inference_engine	String	pangolin inference engine for lineage designations (usher or pangolarn)	usher

continues on next page

Table 7 – continued from previous page

Task	Variable Name	Data Type	Description	Default
pangolin3	min_length	Int	Minimum query length allowed for pangolin to attempt assignment	10000
pangolin3	max_ambig	Float	Maximum proportion of Ns allowed for pangolin to attempt assignment	0.5
read_filtering	cpu	Int	CPU resources allocated to the read filtering task (Artic guppypyled) runtime environment	8
read_filtering	max_length	Int	Maximum sequence length	700
read_filtering	min_length	Int	Minimum sequence length	400
read_filtering	run_prefix	String	Run name	artic_ncov2019
theiacov_ont	nextclade_dataset_name	String	Nextclade organism dataset	sars-cov-2
theiacov_ont	nextclade_dataset_reference	String	Nextclade reference genome	MN908947
theiacov_ont	nextclade_dataset_tag	Nextclade dataset tag	2022-02-07T12:00:00Z	
theiacov_ont	artic_primer_version	String	Version of the Artic PCR protocol used to generate input read data	V3
theiacov_ont	normalise	Int	Value to normalize read counts	200
theiacov_ont	seq_method	String	Description of the sequencing methodology used to generate the input read data	ONT
theiacov_ont	pangolin_docker_image	String	Docker tag used for running Pangolin	staphb/pangolin:2.4.2-pangolearn-2021-05-19
vadr	docker	String	Docker tag used for running VADR	quay.io/staphb/1.4.1-models-1.3-2

continues on next page

Table 7 – continued from previous page

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

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
assembly_length_unambiguous	Int	Number of unambiguous basecalls within the SC2 consensus assembly
assembly_mean_coverage	Float	Mean sequencing depth throughout the conesnsus assembly generated after performing primer trimming–calculated using the SAMtools coverage command

continues on next page

Table 8 – continued from previous page

Output Name	Data Type	Description
assembly_method	String	Method employed to generate consensus assembly
auspice_json	File	Auspice-compatable JSON output generated from NextClade analysis 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_dehosted	Float	Percent of human read data detected using the Kraken2 software after host removal
kraken_report	File	Full Kraken report
kraken_report_dehosted	File	Full Kraken report after host removal
kraken_sc2	Float	Percent of SARS-CoV-2 read data detected using the Kraken2 software
kraken_sc2_dehosted	Float	Percent of SARS-CoV-2 read data detected using the Kraken2 software 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
pangolin_assignment_version	String	Version of the pangolin software (e.g. PANGO or PUSHER) used 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
percent_reference_coverage	Float	Percent coverage of the reference genome after performing primer trimming; calculated as assembly_length_unambiguous / length of reference genome (SC2: 29,903) x 100
primer_bed_name	String	Name of the primer bed files used for primer trimming
pangolin_versions	String	All Pangolin software and database versions

continues on next page

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 input read data
theiacov_ont_analysis_date	String	Date of analysis
theiacov_ont_version	String	Version of the Public Health Viral Genomics (PHVG) repository 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
variants_from_ref_vcf	File	Number of variants relative to the reference genome

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.

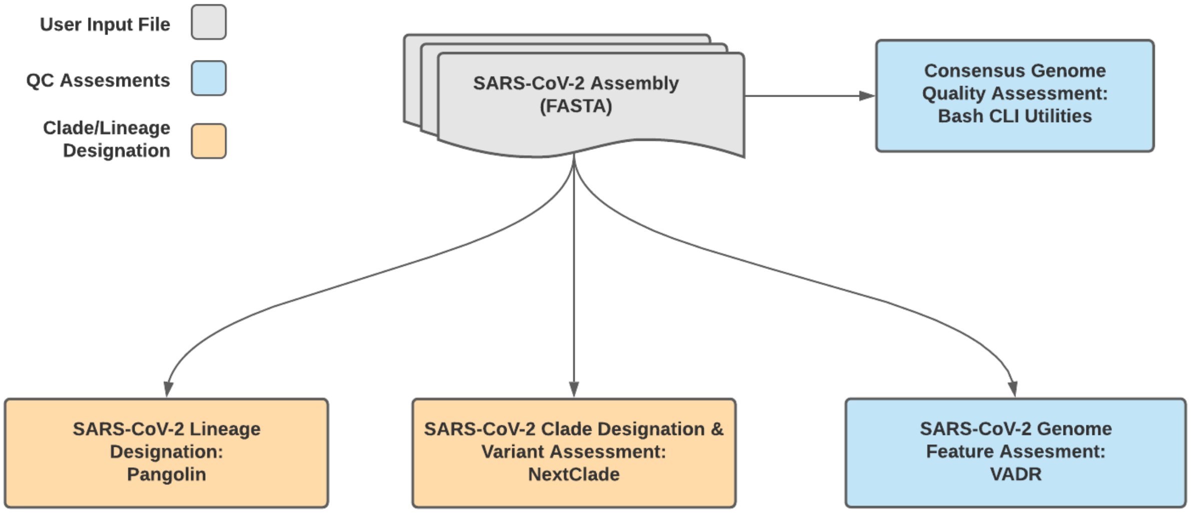
User Input File
QC Assesments
Clade/Lineage Designation
SARS-CoV-2 Assembly (FASTA)
Consensus Genome Quality Assessment: Bash CLI Utilities
SARS-CoV-2 Lineage Designation: Pangolin
SARS-CoV-2 Clade Designation & Variant Assessment: NextClade
SARS-CoV-2 Genome Feature Assesment: VADR

Fig. 5: **TheiaCoV_FASTA Data Workflow**

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

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

Required User Inputs

Download CSV: `TheiaCoV_FASTA_required_inputs.csv`

Task	Input Variable	Data Type	Description
theiacov_fasta	assembly_fasta	File	SARS-CoV-2 assemly file in fasta format
theiacov_fasta	input_assembly_method	String	Description of the method utilized to generate 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 utilized 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_sample	docker	String	Docker tag used for running NextClade	nextstrain/nextclade:1.10.3
nextclade_output_parser_one_sample	docker	String	Docker tag used for parsing NextClade output	python:slim
pangolin3	docker	String	Docker tag used for running Pangolin	quay.io/staphb/3.1.20-pangolearn-2022-02-02
pangolin3	inference_engine	String	pangolin inference engine for lineage designations (usher or pangolarn)	usher
pangolin3	max_ambig	Float	Maximum proportion of Ns allowed for pangolin to attempt assignment	0.5
pangolin3	min_length	Int	Minimum query length allowed for pangolin to attempt assignment	10000
titan_fasta	nextclade_dataset_name	String	Nextclade organism dataset	sars-cov-2
titan_fasta	nextclade_dataset_reference	String	Nextclade reference genome	MN908947
titan_fasta	nextclade_dataset_tag	Nextclade dataset tag	2022-02-07T12:00:00Z	
vadr	docker	String	Docker tag used for running VADR	quay.io/staphb/1.4.1-models-1.3-2
vadr	maxlen	Int	Maximum length for the fasta-trim-terminal-ambigs.pl VADR script	30000
vadr	minlen	Int	Minimum length subsequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl VADR script	50
vadr	skip_length	Int	Minimum assembly length (unambiguous) to run vadr	10000
vadr	vadr_opts	String	Options for the v-annotate.pl VADR script	–glsearch -s -r –nomisc –mkey sarscov2 –alt_fail lowscore,fstukcnf,insertnn,deletinn –mdir /opt/vadr/vadr-models/
version_capture	timezone	String	User time zone in valid	None

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

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

Download CSV: `TheiaCoV_Augur_Prep_required_inputs.csv`

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

TheiaCoV_Augur_Prep

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

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

Upon initiating a TheiaCoV_Augur_Prep run, input assembly/consensus files and associated metadata will be used to produce the array of assembly/consensus files and the array of metadata files to be used as inputs for the TheiaCoV_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 phylogeneic 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	sample_metadata_tsvs	Array[File]	An array of sample metadata files (TSV)
sarscov2_nextstrain	build_name	String	The name of the Augur build to be used in this analysis

1.3 Mercury Workflow Series

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

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

1.3.1 Mercury Workflows for Single-Sample Preparation

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

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

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

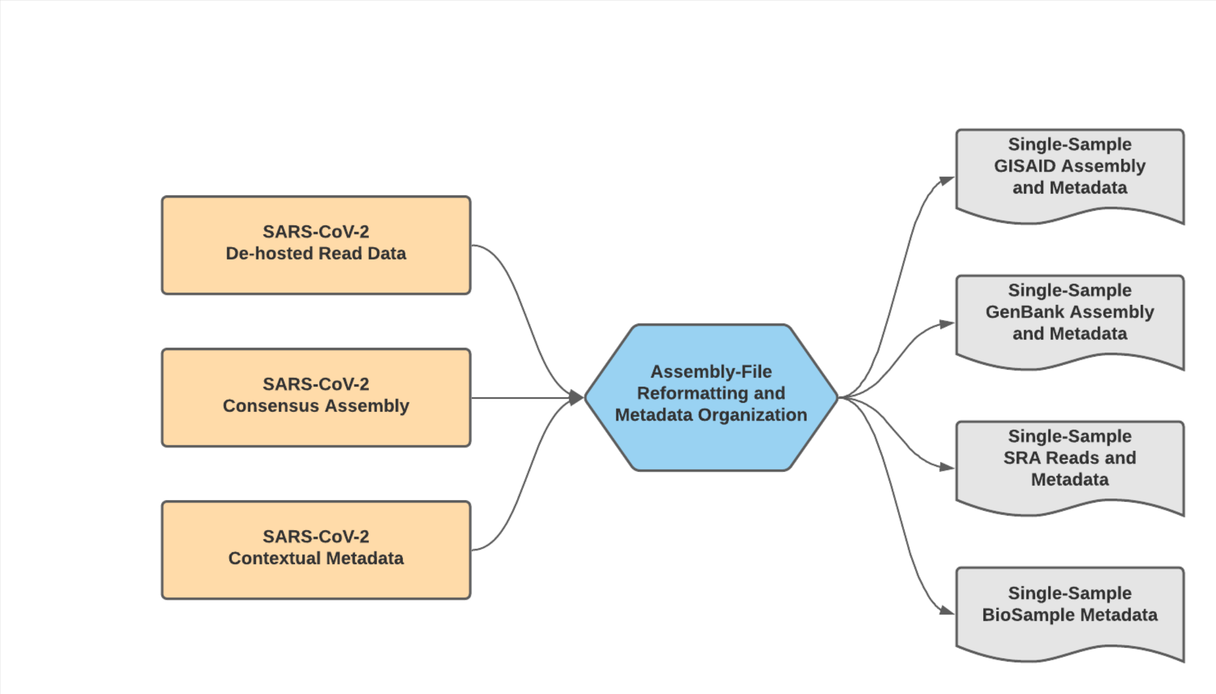

Fig. 6: **Mercury_Prep Data Workflow**

Mercury_PE_Prep

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

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

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

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

Required User Inputs

Download CSV: `Mercury_PE_Prep_required_inputs.csv`

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

Optional User Inputs

Download CSV: `Mercury_PE_Prep_optional_inputs.csv`

Task	Input Variable	Data Type	Description	Default
gisaid_prep_one_sample	specimen_source	String	Biologial source of the specimen, e.g. e.g. sputum, Alveolar lavage fluid, Oro-pharyngeal swab, Blood, Tracheal swab, Urine, Stool, Cloakal swab, Organ, Feces, Other	None
gisaid_prep_one_sample	mem_size_gb	Int	Memory allocated to the gisaid_prep_one_sample task	1
gisaid_prep_one_sample	disk_size	Int	Disk size allocated to the gisaid_prep_one_sample task	25
gisaid_prep_one_sample	patient_status	String	Status of the patient, e.g. Hospitalized, Released, Live, Deceased, unknown	unknown
gisaid_prep_one_sample	type	String	Organism typoe	betacoronovirus
gisaid_prep_one_sample	CPUs	Int	CPUs allocated to the gisaid_prep_one_sample task	None
gisaid_prep_one_sample	preemptible_tries	Int	Number of preemptible tries for the gisaid_prep_one_sample task	0
gisaid_prep_one_sample	outbreak	String	Outbreak associated with this submision, e.g. date, place, family cluster	None
gisaid_prep_one_sample	last_vaccinated	String	Date of last vaccine recieved	None

continues on next page

Table 9 – continued from previous page

Task	Input Variable	Data Type	Description	Default
gisaid_prep_one_sample	docker_image	String	Docker image utilized for the gisaid_prep_one_sample task	quay.io/theiagen/utility:1.1
gisaid_prep_one_sample	passage_details	String	Passage details of the sample being submitted, e.g. original, vero, etc	original
mercury_pe_prep	dehosting_method	String	Method utilized to dehost read data	NCBI Human Scrubber
mercury_pe_prep	filetype	String	File type of the read data being submitted to SRA	fastq
mercury_pe_prep	submitter_email	String	Email address of the submitter	None
mercury_pe_prep	purpose_of_sequencing	String	Reason that this sample was sequenced; for labs that are sequencing samples as part of a federal surveillance program "baseline surveillance" would be the most accurate entry for this field	None
mercury_pe_prep	library_layout	String	Layout of the sequenced library	paired
mercury_pe_prep	number_N_threshold	Int	Maximum number of ambiguous nucleotides in a sample to prepare submission files	5000
mercury_pe_prep	host_sci_name	String	Scientific name of the host organism	Homo sapiens
mercury_pe_prep	gisaid_accession	String	Accession number in GISAID	None
mercury_pe_prep	gisaid_organism	String	Orgiansm name as per GISAID submission	hCoV-19

continues on next page

Table 9 – continued from previous page

Task	Input Variable	Data Type	Description	Default
mercury_pe_prep	county	String	County the laboratory was collected in	None
mercury_pe_prep	amplicon_size	String	Average size of the amplicons sequenced	None
mercury_pe_prep	host	String	Common name of the host organism	Human
mercury_pe_prep	amplicon_primer_scheme	String	Name of the amplicon primer scheme utilized to generate the amplicons sequenced	None
mercury_pe_prep	biosample_accession	String	BioSample accession number	None
mercury_pe_prep	treatment	String	Treatment administered to the patient, e.g. drug name, dosage, etc.	None
mercury_pe_prep	patient_gender	String	Gender of the patient	unknown
mercury_pe_prep	purpose_of_sampling	String	Reason that the original specimen was taken, e.g. clinical diagnostics	None
mercury_pe_prep	patient_age	String	Age of the patient	unknown
ncbi_prep_one_sample	mem_size_gb	Int	Memory allocated to the ncbi_prep_one_sample task	1
ncbi_prep_one_sample	docker_image	String	Docker image utilized for the ncbi_prep_one_sample task	quay.io/staphb/vadr:1.3
ncbi_prep_one_sample	maxlen	Int	VADR –maxlen input utilized when trimming terminal ambiguous ends	30000
ncbi_prep_one_sample	preemptible_tries	Int	Number of preemptible tries for the ncbi_prep_one_sample task	0

continues on next page

Table 9 – continued from previous page

Task	Input Variable	Data Type	Description	Default
ncbi_prep_one_sample	CPUs	Int	CPUs allocated to the ncbi_prep_one_sample task	1
ncbi_prep_one_sample	minlen	Int	VADR –minen input utilized when trimming terminal ambiguous ends	50
ncbi_prep_one_sample	disk_size	Int	Disk size allocated the ncbi_prep_one_sample task	25
version_capture	timezone	String	User time zone in valid Unix TZ string (e.g. America/New_York)	None

Outputs

Download CSV: `Mercury_PE_Prep_default_outputs.csv`

Output Name	Data Type	Description
biosample_attributes	File	Sample metadata compiled and formatted to meet the BioSample submission requirements
genbank_assembly	File	Assembly file reformatted to meet the GenBank submission requirements
genbank_modifier	File	Sample metadata compiled and formatted to meet the GenBank submission requirements; will need to be manually modified to include BioSample accession numbers
gisaid_assembly	File	Assembly file reformatted to meet the GISAID submission requirements
gisaid_metadata	File	Metadata compiled and formatted to meet the GISAID submission requirements
mercury_pe_prep_analysis_date	String	Date of analysis
mercury_pe_prep_version	String	Version of the Public Health Viral Genomics (PHVG) repository 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	assembly_mean_coverage	Float	Mean sequencing depth throughout the conesnsus assembly
mercury_pe_prep	assembly_method	String	Method employed to generate the input assembbly file
mercury_pe_prep	authors	String	Authors associated with this submission
mercury_pe_prep	bioproject_accession	String	NCBI BioProject accession number
mercury_pe_prep	collecting_lab	String	Name of the laboratory that orginial laboratory that collected the sample
mercury_pe_prep	collecting_lab_address	String	Address of the laboratory that orginial laboratory that collected the sample
mercury_pe_prep	collection_date	String	Date on which the sample was collected
mercury_pe_prep	continent	String	Continent the sample was collected in
mercury_pe_prep	country	String	Country the sample was collected in
mercury_pe_prep	gisaid_submitter	String	GISAID username
mercury_pe_prep	host_disease	String	Host disease; for SARS-CoV-2 sequences from human samples, “COVID-19” would be the most accurate entry for this field
mercury_pe_prep	instrument_model	String	Model of the sequencing instrument utilized to generate the read data
mercury_pe_prep	isolation_source	String	Isolation source, i.e. clinical, animal, or environmental
mercury_pe_prep	library_id	String	Unique identifer for the sequenced library
mercury_pe_prep	library_selection	String	Selection methodology used to designate samples as eligible for sequencing, e.g., “PCR” for samples selected based on PCT Ct values
mercury_pe_prep	library_source	String	Source of the genomic material used to prepare the sequencing libraries
mercury_pe_prep	library_strategy	String	Library preparation strategy, e.g., “AMPLICON” for data generated from tiling PCR amplicons
mercury_pe_prep	number_N	Int	Number of fully ambiguous basecalls within the consensus assembly
mercury_pe_prep	organism	String	Name of the organism sequenced, e.g. “SARS-CoV-2”
mercury_pe_prep	reads_dehosted	File	Dehosted read files
mercury_pe_prep	seq_platform	String	Description of the sequencing methodology used to generate the input read data
mercury_pe_prep	state	String	State the sample was collected in
mercury_pe_prep	submission_id	String	Unique identfier for the sample utilized upon submission
mercury_pe_prep	submitting_lab	String	Name of the submitting laboratory
mercury_pe_prep	submitting_lab_address	String	Address of the submitting laboratory

Optional User Inputs

Download CSV: `Mercury_SE_Prep_optional_inputs.csv`

Task	Input Variable	Data Type	Description	Default
gisaid_prep_one_sample	specimen_source	String	Biologial source of the specimen, e.g. e.g. sputum, Alveolar lavage fluid, Oro-pharyngeal swab, Blood, Tracheal swab, Urine, Stool, Cloakal swab, Organ, Feces, Other	None
gisaid_prep_one_sample	mem_size_gb	Int	Memory allocated to the gisaid_prep_one_sample task	1
gisaid_prep_one_sample	disk_size	Int	Disk size allocated to the gisaid_prep_one_sample task	25
gisaid_prep_one_sample	patient_status	String	Status of the patient, e.g. Hospitalized, Released, Live, Deceased, unknown	unknown
gisaid_prep_one_sample	type	String	Organism typoe	betacoronovirus
gisaid_prep_one_sample	CPUs	Int	CPUs allocated to the gisaid_prep_one_sample task	None
gisaid_prep_one_sample	preemptible_tries	Int	Number of preemptible tries for the gisaid_prep_one_sample task	0
gisaid_prep_one_sample	outbreak	String	Outbreak associated with this submision, e.g. date, place, family cluster	None
gisaid_prep_one_sample	last_vaccinated	String	Date of last vaccine recieved	None

continues on next page

Table 10 – continued from previous page

Task	Input Variable	Data Type	Description	Default
gisaid_prep_one_sample	docker_image	String	Docker image utilized for the gisaid_prep_one_sample task	quay.io/theiagen/utility:1.1
gisaid_prep_one_sample	passage_details	String	Passage details of the sample being submitted, e.g. original, vero, etc	original
mercury_pe_prep	dehosting_method	String	Method utilized to dehost read data	NCBI Human Scrubber
mercury_pe_prep	filetype	String	File type of the read data being submitted to SRA	fastq
mercury_pe_prep	submitter_email	String	Email address of the submitter	None
mercury_pe_prep	purpose_of_sequencing	String	Reason that this sample was sequenced; for labs that are sequencing samples as part of a federal surveillance program “baseline surveillance” would be the most accurate entry for this field	None
mercury_pe_prep	library_layout	String	Layout of the sequenced library	paired
mercury_pe_prep	number_N_threshold	Int	Maximum number of ambiguous nucleotides in a sample to prepare submission files	5000
mercury_pe_prep	host_sci_name	String	Scientific name of the host organism	Homo sapiens
mercury_pe_prep	gisaid_accession	String	Accession number in GISAID	None
mercury_pe_prep	gisaid_organism	String	Orgiansm name as per GISAID submission	hCoV-19

continues on next page

Table 10 – continued from previous page

Task	Input Variable	Data Type	Description	Default
mercury_pe_prep	county	String	County the laboratory was collected in	None
mercury_pe_prep	amplicon_size	String	Average size of the amplicons sequenced	None
mercury_pe_prep	host	String	Common name of the host organism	Human
mercury_pe_prep	amplicon_primer_scheme	String	Name of the amplicon primer scheme utilized to generate the amplicons sequenced	None
mercury_pe_prep	biosample_accession	String	BioSample accession number	None
mercury_pe_prep	treatment	String	Treatment administered to the patient, e.g. drug name, dosage, etc.	None
mercury_pe_prep	patient_gender	String	Gender of the patient	unknown
mercury_pe_prep	purpose_of_sampling	String	Reason that the original specimen was taken, e.g. clinical diagnostics	None
mercury_pe_prep	patient_age	String	Age of the patient	unknown
ncbi_prep_one_sample	mem_size_gb	Int	Memory allocated to the ncbi_prep_one_sample task	1
ncbi_prep_one_sample	docker_image	String	Docker image utilized for the ncbi_prep_one_sample task	quay.io/staphb/vadr:1.3
ncbi_prep_one_sample	maxlen	Int	VADR –maxlen input utilized when trimming terminal ambiguous ends	30000
ncbi_prep_one_sample	preemptible_tries	Int	Number of preemptible tries for the ncbi_prep_one_sample task	0

continues on next page

Table 10 – continued from previous page

Task	Input Variable	Data Type	Description	Default
ncbi_prep_one_sample	CPUs	Int	CPUs allocated to the ncbi_prep_one_sample task	1
ncbi_prep_one_sample	minlen	Int	VADR –minen input utilized when trimming terminal ambiguous ends	50
ncbi_prep_one_sample	disk_size	Int	Disk size allocated the ncbi_prep_one_sample task	25
version_capture	timezone	String	User time zone in valid Unix TZ string (e.g. America/New_York)	None

Outputs

Download CSV: `Mercury_SE_Prep_default_outputs.csv`

Output Name	Data Type	Description
biosample_attributes	File	Sample metadata compiled and formatted to meet the BioSample submission requirements
genbank_assembly	File	Assembly file reformatted to meet the GenBank submission requirements
genbank_modifier	File	Sample metadata compiled and formatted to meet the GenBank submission requirements; will need to be manually modified to include BioSample accession numbers
gisaid_assembly	File	Assembly file reformatted to meet the GISAID submission requirements
gisaid_metadata	File	Metadata compiled and formatted to meet the GISAID submission requirements
mercury_pe_prep_analysis_date	String	Date of analysis
mercury_pe_prep_version	String	Version of the Public Health Viral Genomics (PHVG) repository 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.

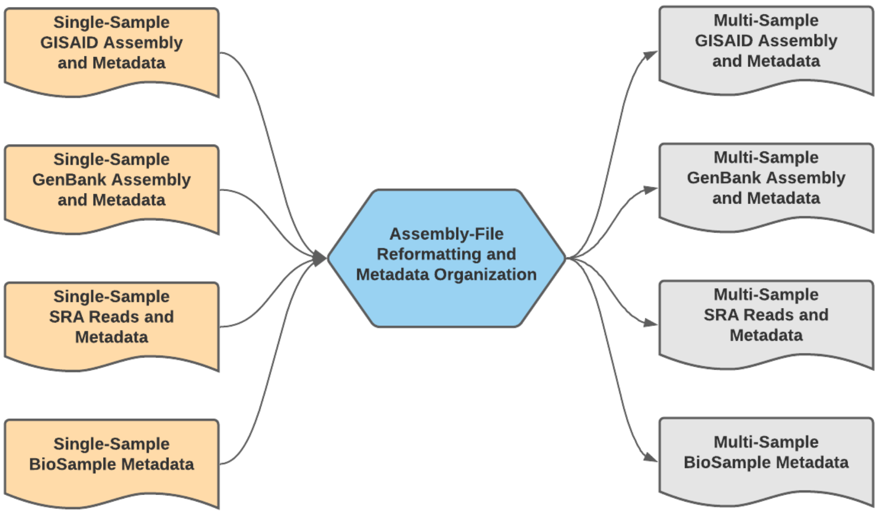
Single-Sample GISAID Assembly and Metadata
Single-Sample GenBank Assembly and Metadata
Single-Sample SRA Reads and Metadata
Single-Sample BioSample Metadata
Assembly-File Reformatting and Metadata Organization
Multi-Sample GISAID Assembly and Metadata
Multi-Sample GenBank Assembly and Metadata
Multi-Sample SRA Reads and Metadata
Multi-Sample BioSample Metadata

Fig. 7: **Mercury_Batch Data Workflow**

Mercury_Batch

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

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

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

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

Required User Inputs

Download CSV: `Mercury_Batch_required_inputs.csv`

Task	Input Variable	Data Type	Description
mercury_batch	biosample_attributes	Array[File]	Array of sample metadata filescompiled and formatted to meet the BioSample submission 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 modified to include BioSample accession numbers
mercury_batch	gisaid_assembly	Array[File]	Array of metadata files compiled and formatted to meet the GISAID submission requirements
mercury_batch	gisaid_metadata	Array[File]	Array of assembly files reformatted to meet the GISAID submission requirements
mercury_batch	samplename	Array[String]	Array of sample identifiers
mercury_batch	sra_metadata	Array[File]	Array of sample and read metadata files compiled and formatted to meet the SRA submission requirements
mercury_batch	sra_reads	Array[String]	Array of forward and reverse reads formatted for submission to SRA
mercury_batch	submission_id	Array[String]	Array of submission identifiers
mercury_batch	vadr_num_alerts	Array[String]	Array of VADR number of alerts

Optional User Inputs

Download CSV: `Mercury_Batch_optional_inputs.csv`

Task	Input Variable	Data Type	Description	Default
compile_biosamp_n_sra	docker_image	String	Docker image utilized for the compile_biosample_n_sra task	quay.io/theiagen/utility:1.1
compile_biosamp_n_sra	preemptible_tries	Int	Number of preemptible tries for the compile_biosample_n_sra task	0
genbank_compile	docker_image	String	Docker image utilized for the genbank_compile task	quay.io/theiagen/utility:1.1
genbank_compile	preemptible_tries	Int	Number of preemptible tries for the genbank_compile task	0
gisaid_compile	docker_image	String	Docker image utilized for the gisaid_compile task	quay.io/theiagen/utility:1.1
gisaid_compile	preemptible_tries	Int	Number of preemptible tries for the gisaid_compile task	0
mercury_batch	CPUs	Int	CPUs allocated for each task in the mercury_batch workflow	4
mercury_batch	disk_size	Int	Disk size allocated for each task in the mercury_batch workflow	100
mercury_batch	gcp_bucket	String	GCP bucket for SRA transfer	None
mercury_batch	mem_size_gb	Int	Memory allocated for each task in the mercury_batch workflow	8
mercury_batch	vadr_threshold	Int	Maximum number of VADR alerts for samples included in the batch submission files	0
version_capture	timezone	String	User time zone in valid Unix TZ string (e.g. America/New_York)	None

Outputs

Download CSV: `Mercury_Batch_default_outputs.csv`

Output Name	Data Type	Description
GenBank_batched_samples	File	File detailing all of the files bacthed for GenBank submission
GenBank_excluded_samples	File	File detailing all of the files excluded from the prepared submission 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
GISAID_batched_samples	File	File detailing all of the files bacthed for GenBank submission
GISAID_excluded_samples	File	File detailing all of the files excluded from the prepared submission files for GenBank
GISAID_metadata	File	Compiled metadata formatted for batch submissino to GISAID
mercury_batch_analysis_date	String	Date of analysis
mercury_batch_version	String	Version of the Public Health Viral Genomics (PHVG) repository 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