# Public Health Viral Genomics (Theiagen)

Release 1.4.3

Kevin G. Libuit

# **CONTENTS**

1 Contents				
	1.1	Public Health Viral Genomics	1	
	1.2	Titan Workflow Series	1	
		License	3	

**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 Titan Workflow Series

The Titan Workflow Series is a collection of WDL workflows developed for performing genomic characterization and genomic epidemiology of viral samples to support public health decision-making. As of today (May 4th, 2021) these workflows are specific to SARS-CoV-2 amplicon read data, but work is underway to allow for the analysis of other viral pathogens of concern.

#### 1.2.1 Titan 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 Titan Series includes four separate WDL workflows (Titan\_Illumina\_PE, Titan\_Illumina\_SE, Titan\_ClearLabs, and Titan\_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 Titan 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 Titan 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 series of introductory training videos that provide conceptual overviews of methodologies and walkthrough tutorials on how to utilize these Titan workflows through Terra are available on the Theiagen Genomics YouTube page:

#### Titan Illumina PE

The Titan\_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 Titan\_Illumina\_PE 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. 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.  $2 \times 150$  bp reads). Modifications to the optional parameter for trimmomatic\_minlen may be required to accommodate for shorter read data, such as  $2 \times 75$ bp reads generated using a 150-cycle kit.

Upon initiating a Titan\_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 samples SARS-CoV-2 lineage and clade types as outlined in the Titan Illumina PE data workflow below.

Consensus genome assembly with the Titan\_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 Titan\_Illumina\_PE are outlined below.

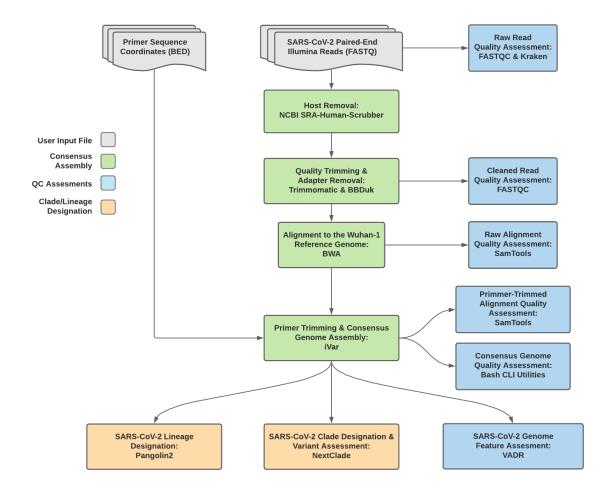


Fig. 1: Titan\_Illumina\_PE v1.4.4 Data Workflow

# **Required User Inputs**

Download CSV: Titan\_Illumina\_PE\_required\_inputs.csv

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

# **Optional User Inputs**

Download CSV: Titan\_Illumina\_PE\_optional\_inputs.csv

Task	Variable Name	Data Type	Description	Default
bedtools_cov	primer_bed	String	Path to the	/artic-
			primer sequence	ncov2019/primer_schemes/nCoV-
			coordinates of	2019/V3/nCoV-2019_amplicon.bed
			the PCR scheme	
			utilized in BED	
			file format	
bedtools_cov	fail_threshold	String	Minimum cov-	20x
			erage threshold	
			to determin	
			amplicon	
			sequencing	
			failture	
bwa	refer-	String	Path to the ref-	/artic-
	ence_genome		erence genome	ncov2019/primer_schemes/nCoV-
			within the	2019/V3/nCoV-
				ar <b>21029.2010528</b> ce.fasta
			Docker con-	
			tainer	
bwa	cpus	Int	CPU resources	6
			allocated to	
			the BWA task	
			runtime envi-	
			ronment	

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_ASM985889v3_genomic.gff
			ture format	
			of the refer-	
			ence genome	
			within the	
			staphb/ivar:1.2.2_	artic20200528
			Docker con-	
			tainer	
consensus	ref_genome	String	Path to the ref-	/artic-
		3	erence genome	ncov2019/primer_schemes/nCoV-
			within the	2019/V3/nCoV-
				ar <b>20:20120052:8</b> ce.fasta
			Docker con-	diditexituate c.idsta
			tainer	
consensus	min_qual	Int	Minimum qual-	20
COHSCHSUS	IIIII_quai	111t	ity threshold for	20
			sliding window	
			to pass for iVar	
			1 -	
		Tilant	consensus	0.6
consensus	min_freq	Float	Minimum	0.6
			frequency	
			threshold(0 -	
			1) to call vari-	
			ants for iVar	
			consensus	
consensus	min_depth	Int	Minimum read	10
			depth to call	
			variants for 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	max_depth	Int	Maximum reads	600000
			read at a posi-	
			tion per input	
			file for SAM-	
			tools mpileup	
			before running	
			iVar consensus	
			1 var consciisus	

Table 1 – continued from previous page

Task	Variable Name	Data Type	Description	Default
consensus	disable_baq	Boolean	Disable read-	TRUE
			pair overlap	
			detection for	
			SAMtools	
			mpileup before	
			running iVar	
			consensus	
consensus	count_orphans	Boolean	Do not skip	TRUE
	-		anomalous	
			read pairs in	
			variant calling	
			for SAMtools	
			mpileup before	
			running iVar	
			consensus	
consensus	char_unknown	String	Character to	N
	_		print in regions	
			with less than	
			minimum cov-	
			erage for iVar	
			consensus	
nextclade_one_sar	mpdet_sequence	File	Custom ref-	None
			erence se-	
			quence file for	
			NextClade	
nextclade_one_sar	nple_config_json	File	Custom QC	None
			configu-	
			raiton file	
			for NextClade	
nextclade_one_sar	nplæ_primers_csv	File	Custom PCR	None
			primers file for	
			NextClade	
nextclade_one_sar	n <b>ge</b> ne_annotations_	j <b>s</b> tile	Custom gene an-	None
			notation file for	
			NextClade	
nextclade_one_sar	m <b>øb</b> æker	String	_	neherlab/nextclade:0.14.2
			for running	
			NextClade	
nextclade_one_sar		File	Custom refer-	None
	pice_reference_tre	e_json	ence tree file for	
			NextClade	
pangolin3	infer-	String	pangolin infer-	usher
	ence_engine		ence engine for	
			lineage designa-	
			tions (usher or	
			pangolarn)	
pangolin3	min_length	Int	Minimum query	10000
	- 0 1			
	_ 2		length allowed	
	_ 2		for pangolin	
	_ 2		-	

Table 1 – continued from previous page

pangolin3 max_ambig Float Maximum proportion of Ns allowed for pangolin to attempt assignment primer_trim keep_noprimer_readboolean Include reads with no primers for iVar trim read_QC_trim trimmo- Int matic_window_size	Task	Variable Name	Data Type	Description	Default
primer_trim keep_noprimer_readboolean	pangolin3	max_ambig			0.5
primer_trim keep_noprimer_readboolean					
primer_trim keep_noprimer_readboolean				lowed for pan-	
primer_trim   keep_noprimer_realBoolean   linclude reads with no primers for iVar trim					
primer_trim keep_noprimer_readboolean linclude reads with no primers for iVar trim read_QC_trim trimmo- Int matic_window_size losses to average across for Trimmomatic puality_trim_score losses to average quality required for Trimmomatic puality_trim_score losses to average quality required for Trimmomatic losses losses to average quality required for Trimmomatic losses losses to average quality required for Trimmomatic losses los entre los ent					
read_QC_trim trimmo- matic_window_size  read_QC_trim trimmo- matic_quality_trim_score  read_QC_trim trimmo- matic_quality_trim_score  read_QC_trim trimmo- matic_minlen  Int Specifies the average quality required for Trim- momatic  read_QC_trim trimmo- matic_minlen  Int Specifies the average quality required for Trim- momatic  Description of the sequencing methodology used to generate the input read data  ti- ti- tian_illumina_pe  String Docker tag used for running Pangolin  vadr docker String Docker tag used for running Pangolin  vadr docker String Docker tag used for running VADR  vadr maxlen Int Maximum length for the fasta-trim-terminal- ambigs.pl  vadr minlen Int Minimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl	primer_trim	keep_noprimer_re	a <b>B</b> soolean		True
read_QC_trim trimmo-	-			with no primers	
matic_window_size  number of bases to average across for Trimmomatic  read_QC_trim  re					
matic_window_size  number of bases to average across for Trimmomatic  read_QC_trim  re	read_QC_trim	trimmo-	Int	Specifies the	4
read_QC_trim trimmo-		matic_window_si	ze	_	
read_QC_trim trimmo-				bases to aver-	
read_QC_trim trimmo-				age across for	
matic_quality_trim_score erage quality required for Trimmomatic  read_QC_trim trimmo-matic_minlen					
matic_quality_trim_score erage quality required for Trimmomatic  read_QC_trim trimmo-matic_minlen	read_QC_trim	trimmo-	Int	Specifies the av-	30
read_QC_trim trimmomatic  read_QC_trim trimmomatic  read_QC_trim trimmomatic  ti- ti- tan_illumina_pe  ti- tan_illumina_pe  vadr docker String Docker tag used for running VADR  vadr maxlen Int Maximum length for the fasta-trim-terminal-ambigs.pl VADR script  vadr minlen Int Minimum length of reads to be kept for Trimmomatic  String Description of the sequencing methodology used to generate the input read data  Docker tag used for running Pangolin  VADR  vadr Maximum length for the fasta-trim-terminal-ambigs.pl VADR script  vadr Minimum length sub-sequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl  vadr Minimum length sub-sequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl  vadr Minimum length sub-sequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl		matic_quality_trii	n_score	erage quality re-	
read_QC_trim trimmo- matic_minlen length of reads to be kept for Trimmomatic  ti- ti- ti- ti- ti- ti- ti- ti- ti- tan_illumina_pe  pan- golin_docker_image  vadr  docker  String  Docker tag used for running Pan- golin VADR  vadr  maxlen  Int  Maximum length for the fasta-trim-terminal- ambigs.pl VADR script  vadr  minlen  Int  Minimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl  valr  minlen  Int  Minimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl					
matic_minlen   minimum   length of reads to be kept for Trimmomatic    ti- ti- tian_illumina_pe   seq_method   String   Description of the sequencing methodology used to generate the input read data    ti- tan_illumina_pe   pan				momatic	
matic_minlen   minimum   length of reads to be kept for Trimmomatic    ti- ti- tian_illumina_pe   seq_method   String   Description of the sequencing methodology used to generate the input read data    ti- tian_illumina_pe   pan- golin_docker_image   String golin    vadr   docker   String   Docker tag used for running Pan- golin    vadr   maxlen   Int   Maximum   length for the fasta-trim- terminal- ambigs.pl   vADR    vadr   minlen   Int   Minimum   length sub- sequence to possibly replace   Ns for the fasta- trim-terminal- ambigs.pl    vadr   String   Docker tag used for running VADR    staphb/pangolin:2.4.2-pangolearn- 2021-05-19   2021-05-19   30000    staphb/vadr:1.2.1    staphb/vadr:1.2.	read_QC_trim	trimmo-	Int	Specifies the	75
ti- ti- tian_illumina_pe  seq_method  String  Description of the sequencing methodology used to generate the input read data  ti- tian_illumina_pe  pan- golin_docker_image  vadr  docker  String  Docker tag used for running Pan- golin  VADR  vadr  maxlen  Int  Maximum length for the fasta-trim-terminal- ambigs.pl  VADR script  vadr  minlen  Int  Minimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl  VADR  String  Docker tag used for running VADR  staphb/pangolin:2.4.2-pangolearn- 2021-05-19  staphb/vadr:1.2.1		matic_minlen		minimum	
ti- tan_illumina_pe seq_method string seq_method ology used to generate the input read data  ti- tan_illumina_pe seq_method string pan- golin_docker_image vadr  docker String Docker tag used for running Pan- golin VADR  vadr  maxlen  Int  Maximum length for the fasta-trim- terminal- ambigs.pl VADR script  vadr  minlen  Int  Minimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl  vabrum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl  vabrum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl				length of reads	
ti- tan_illumina_pe  seq_method  String  Description of the sequencing methodology used to generate the input read data  ti- ti- tan_illumina_pe  golin_docker_image  vadr  docker  String  Docker tag used for running Pangolin  golin  VADR  vadr  maxlen  Int  Maximum length for the fasta-trim-terminal-ambigs.pl  vADR  vadr  minlen  Int  Minimum length sub-sequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl  vafre (a)  String  Docker tag used for running VADR  Staphb/pangolin:2.4.2-pangolearn-2021-05-19  staphb/vadr:1.2.1  String  Staphb/vadr:1.2.1				to be kept for	
tan_illumina_pe  the sequencing methodology used to generate the input read data  ti- ti- tan_illumina_pe  yadr  docker  String  Docker tag used for running Pangolin  VADR  vadr  maxlen  Int  Maximum length for the fasta-trim-terminal-ambigs.pl  VADR script  vadr  minlen  Int  Minimum length sub-sequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl  VS for the fasta-trim-terminal-ambigs.pl  VS for the fasta-trim-terminal-ambigs.pl  VS for the fasta-trim-terminal-ambigs.pl				Trimmomatic	
ti- ti- tian_illumina_pe vadr  methodology used to generate the input read data  Docker tag used for running Pan- golin VADR  vadr  maxlen  Int  Maximum length for the fasta-trim- terminal- ambigs.pl VADR script  vadr  minlen  Int  Minimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl  vafr  Minimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl	ti-	seq_method	String	Description of	Illumina paired-end
ti- ti- tan_illumina_pe vadr  maxlen  minlen  lused to generate the input read data  staphb/pangolin:2.4.2-pangolearn- golin Docker tag used for running Pan- golin Docker tag used for running VADR  wadr  maxlen  lint  Maximum length for the fasta-trim- terminal- ambigs.pl VADR script  vadr  minlen  Int  Minimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl  wadr  minlen  lint  minlen  sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl	tan_illumina_pe			the sequencing	
ti- ti- tan_illumina_pe				methodology	
ti- ti- tan_illumina_pe vadr  docker  String golin_docker_image vadr  bocker tag used for running Pan- golin  Docker tag used for running VADR  staphb/pangolin:2.4.2-pangolearn- 2021-05-19  staphb/vadr:1.2.1  staphb/vadr:1.2.1  staphb/vadr:1.2.1  staphb/vadr:1.2.1  staphb/vadr:1.2.1  staphb/vadr:1.2.1  vadr  maxlen  Int  Maximum length for the fasta-trim- terminal- ambigs.pl  VADR script  vadr  minlen  Int  Minimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl				used to generate	
ti- tan_illumina_pe golin_docker_image golin_docker_image vadr  docker  String Docker tag used for running Pan- golin  Docker tag used for running VADR  vadr  maxlen  Int  Maximum length for the fasta-trim- terminal- ambigs.pl VADR script  vadr  minlen  Int  Minimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl  VADR  String  Docker tag used for running VADR  staphb/pangolin:2.4.2-pangolearn- 2021-05-19  staphb/vadr:1.2.1				the input read	
tan_illumina_pe golin_docker_image for running Pan- golin  vadr docker String Docker tag used for running VADR  vadr maxlen Int Maximum length for the fasta-trim- terminal- ambigs.pl VADR script  vadr minlen Int Minimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl  Viant sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl				data	
vadr docker String Docker tag used for running VADR  vadr maxlen Int Maximum length for the fasta-trim-terminal-ambigs.pl vadr minlen Int Minimum length subsequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl	ti-	pan-	String	Docker tag used	staphb/pangolin:2.4.2-pangolearn-
vadr docker String Docker tag used for running VADR  vadr maxlen Int Maximum length for the fasta-trimterminal-ambigs.pl VADR script  vadr minlen Int Minimum length subsequence to possibly replace Ns for the fasta-trimterminal-ambigs.pl	tan_illumina_pe	golin_docker_ima	ge	for running Pan-	2021-05-19
vadr maxlen Int Maximum length for the fasta-trim-terminal-ambigs.pl VADR script  vadr minlen Int Minimum length subsequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl				golin	
vadr maxlen Int Maximum length for the fasta-trim-terminal-ambigs.pl VADR script  vadr minlen Int Minimum length subsequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl	vadr	docker	String	Docker tag used	staphb/vadr:1.2.1
vadr maxlen Int Maximum length for the fasta-trimterminal-ambigs.pl VADR script  vadr minlen Int Minimum length subsequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl				for running	
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	
reaction fasta-trimterminal-ambigs.pl vadr  minlen  Int  Minimum length subsequence to possibly replace Ns for the fasta-trim-terminal-ambigs.pl	vadr	maxlen	Int	Maximum	30000
terminal- ambigs.pl VADR script  vadr minlen Int Minimum length sub- sequence to possibly replace Ns for the fasta- trim-terminal- ambigs.pl				length for the	
vadr minlen Int Minimum 50 length subsequence to possibly replace Ns for the fastatrim-terminal-ambigs.pl					
vadr minlen Int Minimum 50 length subsequence to possibly replace Ns for the fastatrim-terminal- ambigs.pl				terminal-	
vadr minlen Int Minimum 50 length subsequence to possibly replace Ns for the fastatrim-terminal- ambigs.pl				ambigs.pl	
length subsequence to possibly replace Ns for the fastatrim-terminal ambigs.pl				VADR script	
sequence to possibly replace Ns for the fastatrim-terminal-ambigs.pl	vadr	minlen	Int	Minimum	50
possibly replace Ns for the fasta- trim-terminal- ambigs.pl				length sub-	
Ns for the fasta- trim-terminal- ambigs.pl					
trim-terminal- ambigs.pl					
ambigs.pl					
				trim-terminal-	
VADR script					
VADA Script				VADR script	

Table 1 – continued from previous page

Task	Variable Name	Data Type	Description	Default
vadr	vadr_opts	String	Options for the v-annotate.pl VADR script	-glsearch -s -r -nomisc -mkey sarscov2 -alt_fail lows- core,fstukcnf,insertnn,deletinn -mdir /opt/vadr/vadr-models/
vadr	skip_length	Int	Minimum as- sembly length (unambiguous) to run vadr	10000
variant_call	ref_gff	String	Path to the general feature format of the reference genome within the staphb/ivar:1.2.2_Docker container	/refer- ence/GCF_009858895.2_ASM985889v3_genomic. artic20200528
variant_call	ref_genome	String	Path to the reference genome within the staphb/ivar:1.2.2_Docker container	/artic- ncov2019/primer_schemes/nCoV- 2019/V3/nCoV- ar <b>2029/2005/28</b> ce.fasta
variant_call	min_qual	Int	Minimum quality threshold for sliding window to pass for iVar variants	20
variant_call	min_freq	Float	Minimum frequency threshold(0 - 1) to call variants for iVar variants	0.6
variant_call	min_depth	Int	Minimum read depth to call variants for iVar variants	10
variant_call	min_bq	Int	Minimum mapping quality for an alignment to be used for SAMtools mpileup before running iVar variants	0

Table 1 – continued from previous page

Task	Variable Name	Data Type	Description	Default
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	disable_baq	Boolean	Disable read-	TRUE
			pair overlap	
			detection for	
			SAMtools	
			mpileup before	
			running iVar	
			variants	
variant_call	count_orphans	Boolean	Do not skip	TRUE
			anomalous	
			read pairs in	
			variant calling	
			for SAMtools	
			mpileup before running iVar	
			running iVar variants	
varsion conture	timezone	String	User time	None
version_capture	umezone	String	zone in valid	None
			Unix TZ string	
			(e.g. Amer-	
			ica/New_York)	
			ICM/I (CW_IOIK)	

# Outputs

Download CSV: Titan\_Illumina\_PE\_default\_outputs.csv

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

Table 2 – continued from previous page

Output Name	Data Type	Description
auspice_json	File	Auspice-compatable JSON output generated from NextClade analy-
auspice_json	riie	sis that includes the NextClade default samples for clade-typing and
		1
hh dult do altan	Ctuin a	the single sample placed on this tree  Docker image used to run BBDuk
bbduk_docker	String	
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)
dehosted_read1	File	Dehosted forward reads; suggested read file for SRA submission
dehosted_read2	File	Dehosted reverse reads; suggested read file for SRA submission
fastqc_clean_pairs	String	Number of paired reads after SeqyClean filtering as determined by FastQC
fastqc_clean1	Int	Number of forward reads after sequelean filtering as determined by FastQC
fastqc_clean2	Int	Number of reverse reads after sequelean filtering as determined by FastQC
fastqc_raw_pairs	String	Number of paired reads identified in the input fastq files as determined by FastQC
fastqc_raw1	Int	Number of forward reads identified in the input fastq files as deter-
-	mt	mined by FastQC
fastqc_raw2	Int	Number of reverse 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
ivar_variant_version	String	Version of iVar for running the iVar variants command
ivar_version_consenst		Version of iVar for running the iVar consensus command
ivar_version_primtrin		Version of iVar for running the iVar trim command
kraken_human	Float	Percent of human read data detected using the Kraken2 software
kraken_human_dehos	teHloat	Percent of human read data detected using the Kraken2 software after host removal
kraken_report	File	Full Kraken report
kraken_report_dehost		Full Kraken report after host removal
kraken_sc2	Float	Percent of SARS-CoV-2 read data detected using the Kraken2 soft-
_		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
··• ,		

Table 2 – continued from previous page

Output Name	Data Type	Description
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_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_version	String	Pangolin and PangoLEARN versions used
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_trimmed_read	<b>pdoat</b> 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
read2_clean	File	Reverse read file after quality trimming and adapter removal
samtools_version	String	Version of SAMtools used to sort and index the alignment file
sam-	String	Version of SAMtools used to create the pileup before running iVar
tools_version_consens	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 input read data
ti-	String	Date of analysis
tan_illumina_pe_anal	ysis_date	
ti-	String	Version of the Public Health Viral Genomics (PHVG) repository
tan_illumina_pe_versi	on	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

#### Titan Illumina SE

The Titan\_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 Titan\_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. 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 Titan\_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 samples SARS-CoV-2 lineage and clade types as outlined in the Titan\_Illumina\_PE data workflow below.

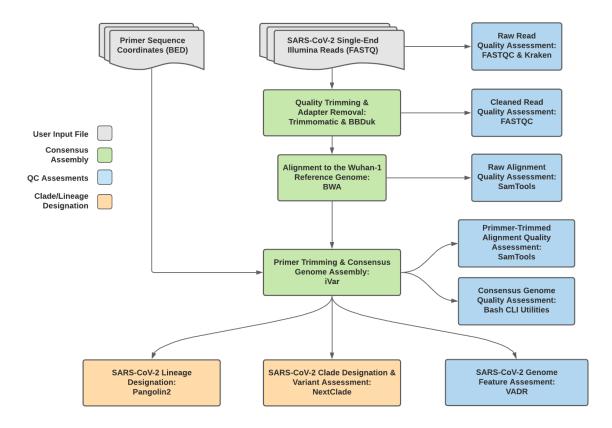


Fig. 2: Titan\_Illumina\_SE v1.4.4 Data Workflow

Consensus genome assembly with the Titan\_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 Titan\_Illumina\_SE are outlined below.

#### **Required User Inputs**

Download CSV: Titan\_Illumina\_SE\_required\_inputs.csv

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

# **Optional User Inputs**

Download CSV: Titan\_Illumina\_SE\_optional\_inputs.csv

1 1, 1	ata Type	Description	Default	
bedtools_cov   primer_bed   Stri	ring	Path to the	/artic-	
		primer sequence	ncov2019/primer_schemes/nCoV-	
		coordinates of	2019/V3/nCoV-2019_amplicon.bed	
	1	the PCR scheme		
	į	utilized in BED		
		file format		
bedtools_cov fail_threshold Stri	ring	Minimum cov-	20x	
		erage threshold		
		to determin		
		amplicon		
		sequencing		
		failture		
bwa refer- Stri	ring	Path to the ref-	/artic-	
ence_genome	-	erence genome	ncov2019/primer_schemes/nCoV-	
		within the	2019/V3/nCoV-	
		staphb/ivar:1.2.2_	ar <b>21029)20105218</b> ce.fasta	
		Docker con-		
	1	tainer		
bwa cpus Int		CPU resources	6	
	;	allocated to		
	1	the BWA task		
	!	runtime envi-		
		ronment		
bwa read2 File	I	Optional input	None	
		file for the bwa		
		task that is not		
		applicable to		
		this workflow		
consensus ref_gff Stri	ing	Path to the	/refer-	
		general fea-	ence/GCF_009858895.2_ASM98588	9v3_genomic.gff
	I	ture format		
	'	of the refer-		
	(	ence genome		
	,	within the		
		staphb/ivar:1.2.2_	artic20200528	
		Docker con-		
	1	tainer		
consensus ref_genome Stri	ing	Path to the ref-	/artic-	
		erence genome	ncov2019/primer_schemes/nCoV-	
		within the	2019/V3/nCoV-	
		staphb/ivar:1.2.2_	ar <b>210:120:120:165:2:8</b> ce.fasta	
		Docker con-		
		tainer		

Table 3 – continued from previous page

Task	Variable Name	Data Type	Description	Default
consensus	min_qual	Int	Minimum qual-	20
			ity threshold for	
			sliding window	
			to pass for iVar	
			consensus	
consensus	min_freq	Float	Minimum	0.6
	1		frequency	
			threshold(0 -	
			1) to call vari-	
			ants for iVar	
			consensus	
consensus	min_depth	Int	Minimum read	10
Comsensus	mm_depth	1110	depth to call	10
			variants for iVar	
			consensus	
consensus	min ha	Int	Minimum map-	0
Consensus	min_bq	1111	ping quality for	U
			an alignment	
			for SAMtools	
			mpileup before	
			running iVar	
	1 1	T .	consensus	600000
consensus	max_depth	Int	Maximum reads	600000
			read at a posi-	
			tion per input	
			file for SAM-	
			tools mpileup	
			before running	
			iVar consensus	
consensus	disable_baq	Boolean	Disable read-	TRUE
			pair overlap	
			detection for	
			SAMtools	
			mpileup before	
			running iVar	
			consensus	
consensus	count_orphans	Boolean	Do not skip	TRUE
			anomalous	
			read pairs in	
			variant calling	
			for SAMtools	
			mpileup before	
			running iVar	
			consensus	
consensus	char_unknown	String	Character to	N
			print in regions	
			with less than	
			minimum cov-	
			erage for iVar	
			consensus	
	•	•	•	continues on next nage

14

Table 3 – continued from previous page

Task	Variable Name	Data Type	Description	Default
nextclade_one_sa	mplet_sequence	File	Custom reference sequence file for NextClade	None
nextclade_one_sa	m <b>pl</b> e_config_json	File	Custom QC configuraiton file for NextClade	None
nextclade_one_sa	mpltr_primers_csv	File	Custom PCR primers file for NextClade	None
	mgene_annotations		Custom gene annotation file for NextClade	None
nextclade_one_sa		String	Docker tag used for running NextClade	neherlab/nextclade:0.14.2
nextclade_one_sa	pice_reference_tro	File e_json	Custom reference tree file for NextClade	None
pangolin3	infer- ence_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_re	a <b>&amp;</b> soolean	Include reads with no primers for iVar trim	True
read_QC_trim	trimmo- matic_window_si		Specifies the number of bases to average across for Trimmomatic	4
read_QC_trim	trimmo- matic_quality_trii	Int n_score	Specifies the average quality required for Trimmomatic	30

Table 3 – continued from previous page

· - ·		able 3 – continued		<del>-</del>
Task	Variable Name	Data Type	Description	Default
read_QC_trim	trimmo-	Int	Specifies the	25
	matic_minlen		minimum	
			length of reads	
		'	to be kept for	
			Trimmomatic	
ti-	seq_method	String	Description of	Illumina paired-end
tan_illumina_pe		'	the sequencing	
		'	methodology	
		'	used to generate	
		'	the input read	
		'	data	
ti-	pan-	String	Docker tag used	staphb/pangolin:2.4.2-pangolearn-
tan_illumina_pe	golin_docker_ima	ige '	for running Pan-	2021-05-19
		<u> </u>	golin	
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	vadr_opts	String	Options for the	-glsearch -s -r -nomisc
		'	v-annotate.pl	-mkey sarscov2 -alt_fail lows-
		'	VADR script	core,fstukcnf,insertnn,deletinn
		'	1	-mdir /opt/vadr/vadr-models/
vadr	skip_length	Int	Minimum as-	10000
	_	'	sembly length	
		'	(unambiguous)	
		'	to run vadr	
variant_call	ref_gff	String	Path to the	/refer-
_			general fea-	ence/GCF_009858895.2_ASM985889v3_genomic.gff
		'	ture format	
		'	of the refer-	
		'	ence genome	
		'	within the	
		'	staphb/ivar:1.2.2_	artic20200528
		'	Docker con-	
		'	tainer	
			-	continues on next page

Table 3 – continued from previous page

Task	Variable Name	Data Type	Description	Default
variant_call	ref_genome	String	Path to the ref-	/artic-
_			erence genome	ncov2019/primer_schemes/nCoV-
			within the	2019/V3/nCoV-
				ar <b>20c29.200528</b> ce.fasta
			Docker con-	
			tainer	
variant_call	min_qual	Int	Minimum qual-	20
variani_can	mm_quur	1111	ity threshold for	20
			sliding window	
			to pass for iVar	
			variants	
variant_call	min_freq	Float	Minimum	0.6
variant_can	mm_rreq	Tiout	frequency	0.0
			threshold(0 - 1)	
			to call variants	
			for iVar variants	
variant_call	min denth	Int	Minimum read	10
variani_Can	min_depth	1111	depth to call	10
			variants for iVar	
			variants	
	1	T		0
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	max_depth	Int	Maximum reads	600000
			read at a posi-	
			tion per input	
			file for SAM-	
			tools mpileup	
			before running	
			iVar variants	
variant_call	disable_baq	Boolean	Disable read-	TRUE
			pair overlap	
			detection for	
			SAMtools	
			mpileup before	
			running iVar	
			variants	
variant_call	count_orphans	Boolean	Do not skip	TRUE
<del>-</del>	I		anomalous	
			read pairs in	
			variant calling	
			for SAMtools	
			mpileup before	
			running iVar	
			variants	
			variants	continues on poyt page

Table 3 – continued from previous page

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

# Outputs

Download CSV: Titan\_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		bly
assem-	Float	Mean sequencing depth throughout the conesnsus assembly gener-
bly_mean_coverage		ated after performing primer trimming-calculated using the SAM-
		tools coverage command
assembly_method	String	Method employed to generate consensus assembly
auspice_json	File	Auspice-compatable JSON output generated from NextClade analy-
		sis that includes the NextClade default samples for clade-typing and
		the single sample placed on this tree
bbduk_docker	String	Docker image used to run BBDuk
bwa_version	String	Version of BWA used to map read data to the reference genome
consensus_flagstat	File	Output from the SAMtools flagstat command to assess quality of the
		alignment file (BAM)
consensus_stats	File	Output from the SAMtools stats command to assess quality of the
		alignment file (BAM)
fastqc_clean	Int	Number of reads after SeqyClean filtering as determined by FastQC
fastqc_raw	Int	Number of reads after sequelean filtering as determined by FastQC
fastqc_version	String	Version of the FastQC software used for read QC analysis
ivar_tsv	File	Variant descriptor file generated by iVar variants
ivar_variant_version	String	Version of iVar for running the iVar variants command
ivar_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
meanbaseq_trim	Float	Mean quality of the nucleotide basecalls aligned to the reference
		genome after primer trimming

Table 4 – continued from previous page

Output Name	Data Type	Description		
meanmapq_trim	Float	Mean quality of the mapped reads to the reference genome after		
meannapq_ann	Tiout	primer trimming		
nextclade_aa_dels	String	Amino-acid deletions as detected by NextClade		
nextclade_aa_subs	String	Amino-acid substitutions as detected by NextClade		
nextclade_clade	String	NextClade clade designation		
nextclade_json	File	NexClade output in JSON file format		
nextclade_tsv	File	NextClade output in TSV file format		
nextclade_version	String	Version of NextClade software used		
number_Degenerate	Int	Number of degenerate basecalls within the consensus assembly		
number_N	Int	Number of fully ambiguous basecalls within the consensus assembly		
number_Total	Int	Total number of nucleotides within the consensus assembly		
pango_lineage	String	Pango lineage as detremined by Pangolin		
pango_lineage_report		Full Pango lineage report generated by Pangolin		
pangolin_conflicts	String	Number of lineage conflicts as deteremed by Pangolin		
		Docker image used to run Pangolin		
pangolin_docker	String			
pangolin_notes	String	Lineage notes as deteremined by Pangolin		
pangolin_version	String	Pangolin and PangoLEARN versions used		
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_trimmed_read	_ <b>pdoae</b> nt	Percent of read data with primers trimmed as deteremined by iVar		
11 1	Tri1	trim		
read1_clean	File	Forward read file after quality trimming and adapter removal		
samtools_version	String	Version of SAMtools used to sort and index the alignment file		
sam-	String	Version of SAMtools used to create the pileup before running iVar		
tools_version_consens		consensus		
sam-	String	Version of SAMtools used to create the pileup before running iVar		
tools_version_primtri		trim		
sam-	String	Version of SAMtools used to assess quality of read mapping		
tools_version_stats				
seq_platform				
	String	Description of the sequencing methodology used to generate the in-		
ti-		put read data		
	String			
tan_illumina_se_analy	String vsis_date	put read data  Date of analysis		
tan_illumina_se_analy	String //sis_date String	put read data		
tan_illumina_se_analy	String vsis_date String on	put read data  Date of analysis  Version of the Public Health Viral Genomics (PHVG) repository used		
tan_illumina_se_analy ti- tan_illumina_se_versi trimmo-	String //sis_date String	put read data  Date of analysis  Version of the Public Health Viral Genomics (PHVG) repository		
tan_illumina_se_analy ti- tan_illumina_se_versi trimmo- matic_version	String vsis_date String on String	put read data Date of analysis  Version of the Public Health Viral Genomics (PHVG) repository used  Version of Trimmomatic used		
tan_illumina_se_analy ti- tan_illumina_se_versi trimmo- matic_version vadr_alerts_list	String vsis_date String on String File	put read data Date of analysis  Version of the Public Health Viral Genomics (PHVG) repository used  Version of Trimmomatic used  File containing all of the fatal alerts as determined by VADR		
tan_illumina_se_analy ti- tan_illumina_se_versi trimmo- matic_version	String vsis_date String on String	put read data Date of analysis  Version of the Public Health Viral Genomics (PHVG) repository used  Version of Trimmomatic used		

#### **Titan ClearLabs**

The Titan\_ClearLabs workflow was written to process ClearLabs WGS read data for SARS-CoV-2 Artic V3 amplicon sequencing.

Upon initiating a Titan\_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 samples SARS-CoV-2 lineage and clade types as outlined in the Titan\_ClearLabs data workflow below.

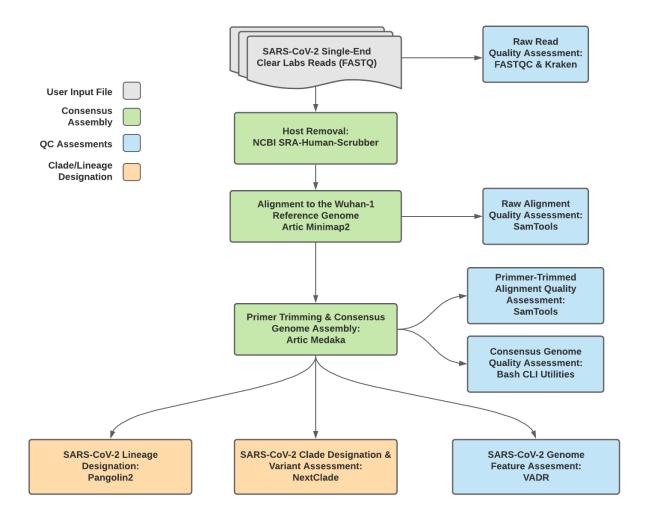


Fig. 3: Titan\_ClearLabs v1.4.4 Data Workflow

Consensus genome assembly with the Titan\_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* <a href="https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html">https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html</a>. 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 Titan\_ClearLabs workflow.

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

## **Required User Inputs**

Download CSV: Titan\_ClearLabs\_required\_inputs.csv

Task	Input Variable	Data Type	Description
titan_clearlabs	clear_lab_fastq	File	Clear Labs FASTQ read files
titan_clearlabs	samplename	String	Name of the sample being analyzed

## **Optional User Inputs**

Download CSV: Titan\_ClearLabs\_optional\_inputs.csv

Task	Variable Name	Data Type	Description	Default
bedtools_cov	primer_bed	String	Path to the	/artic-
			primer sequence	ncov2019/primer_schemes/nCoV-
			coordinates of	2019/V3/nCoV-2019_amplicon.bed
			the PCR scheme	
			utilized in BED	
hadead.	£a:1 4h	Ct:	file format	20
bedtools_cov	fail_threshold	String	Minimum cov-	20x
			erage threshold to determin	
			amplicon	
			sequencing	
			failture	
consensus	cpu	Int	CPU resources	8
201102110415	-P"		allocated to the	, The state of the
			Artric Medaka	
			task runtime	
			environment	
fastqc_se_raw	cpus	Int	CPU resources	
			allocated to	
			the FastQC	
			task runtime	
			environment for	
			asessing raw	
	14		read data	
fastqc_se_raw	read1_name	String	Name of the	Inferred from the input read file
			sample being	
langle and 2		T4	analyzed	4
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	
1 1 2	10	To the second se	container	
kraken2_raw	read2	File	Optional input	None
			file for the	
			Kraken task that	
			is not applicable to this workflow	
nextclade_one_sa	m <b>rile</b> t seguence	File	Custom ref-	None
nextende_one_sa	inpoor_sequence	1110	erence se-	110110
			quence file for	
			NextClade	
nextclade_one_sa	mple_config_json	File	Custom QC	None
			configu-	
			raiton file	
			for NextClade	
	m <b>p&amp;</b> _primers_csv	File	Custom PCR	None Chantar 1 Contanta
22			primers file for	Chapter 1. Contents
		1701	NextClade	
nextclade_one_sar	m <b>g&amp;n</b> e_annotations	_j <b>s</b> töte	Custom gene an-	None
			notation file for	

# Outputs

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

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
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	te <b>H</b> loat	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
		continues on next page

Table 5 – continued from previous page

Output Name	Data Type	Description	
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_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_version	String	Pangolin and PangoLEARN versions used	
per-	Float	Percent coverage of the reference genome after performing primer	
cent_reference_coverage		trimming; calculated as assembly_length_unambiguous / length of	
		reference genome (SC2: 29,903) x 100	
pool1_percent	Float	Percentage of aligned read data assocaited with the pool1 amplicons	
pool2_percent	Float	Percentage of aligned read data assocaited with the pool 2 amplicons	
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	
ti-	String	Date of analysis	
tan_clearlabs_analysis			
ti-	String	Version of the Public Health Viral Genomics (PHVG) repository	
tan_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			

#### **Titan ONT**

The Titan\_ONT workflow was written to process basecalled and demultiplexed Oxford Nanopore Technology (ONT) read data. IInput reads are assumed to be the product of sequencing Artic V3 tiled PCR-amplicons designed for the SARS-CoV-2 genome.

**Note:** As of May 2021, alternative primer schemes are not currently supported for the Titan\_ONT workflow, but active development us underway to allow for such analysis in the near future.

Upon initiating a Titan\_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 samples SARS-CoV-2 lineage and clade types as outlined in the Titan\_ONT data workflow below.

Consensus genome assembly with the Titan\_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 <a href="https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html">https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html</a>. 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.

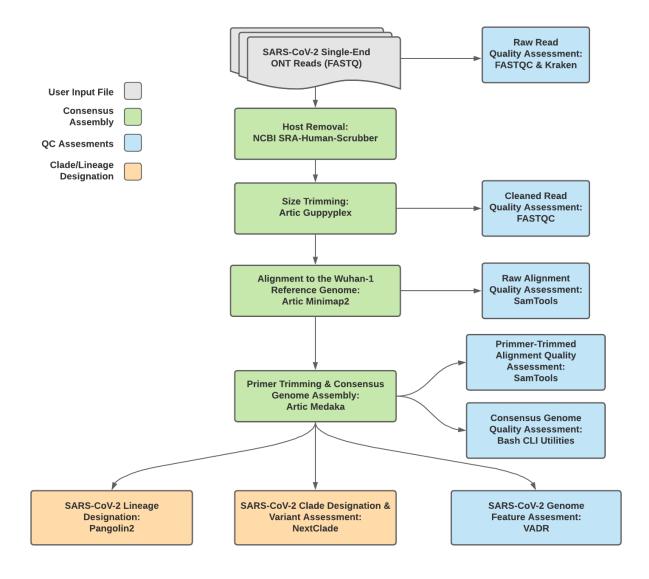


Fig. 4: Titan\_ONT v1.4.4 Data Workflow

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 Titan\_ONT are outlined below.

## **Required User Inputs**

Download CSV: Titan\_ONT\_required\_inputs.csv

Task	Input Variable	Data Type	Description
titan_ont	demulti-	File	Basecalled and demultiplexed ONT read
	plexed_reads		data (single FASTQ file per sample)
titan_ont	samplename	String	Name of the sample being analyzed

#### **Optional User Inputs**

Download CSV: Titan\_ONT\_optional\_inputs.csv

Task	Variable Name	Data Type	Description	Default
bedtools_cov	primer_bed	String	Path to the	/artic-
			primer sequence	ncov2019/primer_schemes/nCoV-
			coordinates of	2019/V3/nCoV-2019_amplicon.bed
			the PCR scheme	
			utilized in BED	
			file format	
bedtools_cov	fail_threshold	String	Minimum cov-	20x
			erage threshold	
			to determin	
			amplicon	
			sequencing	
			failture	
consensus	cpu	Int	CPU resources	8
			allocated to the	
			Artric Medaka	
			task runtime	
			environment	
fastqc_se_clean	cpus	Int	CPU resources	2
			allocated to	
			the FastQC	
			task runtime	
			environment	
			for assssing	
			size-selected	
			read data	

Table 6 – continued from previous page

Task	Variable Name	Data Type	Description	Default
fastqc_se_clean	read1_name	String	Name of the	Inferred from the input read file
1 – –	_		sample being	•
			analyzed	
fastqc_se_raw	cpus	Int	CPU resources	
rustqe_se_ru	Сриз	1111	allocated to	
			the FastQC	
			task runtime	
			environment for	
			_	
			asessing raw read data	
fastqc_se_raw	read1_name	String	Name of the	Inferred from the input read file
rasiqc_sc_raw	read1_name	String	sample being	Interred from the input read inc
1 1 0		Τ.,	analyzed	
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	
nextclade_one_sar	m <b>ple</b> t_sequence	File	Custom ref-	None
	1		erence se-	
			quence file for	
			NextClade	
nextclade_one_sar	n <b>pl</b> e config ison	File	Custom QC	None
	<u> </u>		configu-	
			raiton file	
			for NextClade	
nextclade_one_sar	n <b>pbr</b> primers csv	File	Custom PCR	None
	1r	-	primers file for	
			NextClade	
nextclade one sar	n <b>gen</b> e_annotations	iköle	Custom gene an-	None
		-)	notation file for	
			NextClade	
nextclade_one_sar	ndbeker	String	Docker tag used	neherlab/nextclade:0.14.2
nexterade_one_sar	iiput KCi	Sumg	for running	menerial/nextelade.0.14.2
			NextClade	
nextclade_one_sar	n enly w	File		Nana
	112DUKS-	File	Custom refer-	None
nextciade_one_sai			4 C1 - C	
nextcrade_one_sar	pice_reference_tre	ee_json	ence tree file for NextClade	

Table 6 – continued from previous page

pangolin3 inference_engine String pangolin inference_ence_engine lineage designations (usher or pangolarn)  pangolin3 min_length Int Minimum query length allowed for pangolin to attempt assignment  pangolin3 max_ambig Float Maximum proportion of Ns allowed for pangolin to attempt assignment  pangolin to attempt assignment  Pangolin to attempt assignment  CPL reserves Server assignment	
lineage designations (usher or pangolarn)  pangolin3 min_length Int Minimum query length allowed for pangolin to attempt assignment  pangolin3 max_ambig Float Maximum proportion of Ns allowed for pangolin to attempt assignment  pangolin4 max_ambig Float Maximum proportion of Ns allowed for pangolin to attempt assignment	
pangolin3 min_length Int Minimum query length allowed for pangolin to attempt assignment  pangolin3 max_ambig Float Maximum proportion of Ns allowed for pangolin to attempt assignment  pangolin to attempt assignment	
pangolin3 min_length Int Minimum query length allowed for pangolin to attempt assignment  pangolin3 max_ambig Float Maximum proportion of Ns allowed for pangolin to attempt assignment  pangolin to attempt assignment	
pangolin3 min_length Int Minimum query length allowed for pangolin to attempt assignment  pangolin3 max_ambig Float Maximum proportion of Ns allowed for pangolin to attempt assignment  pangolin to attempt assignment	
length allowed for pangolin to attempt assignment  pangolin3 max_ambig Float Maximum proportion of Ns allowed for pangolin to attempt assignment  Description of Ns allowed for pangolin to attempt assignment	
pangolin3 max_ambig Float Maximum proportion of Ns allowed for pangolin to attempt assignment	
pangolin3 max_ambig Float Maximum proportion of Ns allowed for pangolin to attempt assignment	
pangolin3 max_ambig Float Maximum proportion of Ns allowed for pangolin to attempt assignment	
pangolin3 max_ambig Float Maximum proportion of Ns allowed for pangolin to attempt assignment 0.5	
portion of Ns allowed for pangolin to attempt assignment	
lowed for pan- golin to attempt assignment	
golin to attempt assignment	
golin to attempt assignment	
and filtering and Let CDU	
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	
titan_ont ar- String Version of the V3	
tic_primer_version Artic PCR	
protocol used to	
generate input	
read data	
titan_ont normalise Int Value to nor- 200	
malize read	
counts	
titan_ont seq_method String Description of ONT	
the sequencing	
methodology	
used to generate	
the input read	
data	
titan_ont pan- String Docker tag used staphb/pangolin:2.4.2-pangole	arn-
golin_docker_image for running Pan- 2021-05-19	
golin	
vadr docker String Docker tag used staphb/vadr:1.2.1	
for running	
VADR	

28

Table 6 – continued from previous page

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

# Outputs

Download CSV: Titan\_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_unambiguous		bly	
assem-	Float	Mean sequencing depth throughout the conesnsus assembly gener-	
bly_mean_coverage		ated after performing primer trimming-calculated using the SAM-	
		tools coverage command	

Table 7 – continued from previous page

Output Name	Data Type	Description		
assembly_method	String	Method employed to generate consensus assembly		
auspice_json	File	Auspice-compatable JSON output generated from NextClade analy-		
		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 filtering and dehosting as determined by FastQC		
fastqc_raw	Int	Number of raw reads input reads as determined by FastQC		
fastqc_version	String	Version of the FastQC version used		
kraken_human	Float	Percent of human read data detected using the Kraken2 software		
kraken_human_dehos	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		
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		Full Pango lineage report generated by Pangolin		
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_version	String	Pangolin and PangoLEARN versions used		
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		
pool1_percent	Float	Percentage of aligned read data assocaited with the pool 1 amplicons		
pool2_percent	Float	Percentage of aligned read data associated with the pool 2 amplicons		
samtools_version	String	Version of SAMtools used to sort and index the alignment file		
Samtoons_version	541115	TOTOTOTIOT OF INTEGOTO GOOD TO BOTT GITG INGENT GITCHINGHT INC		

Table 7 – continued from previous page

Output Name	Data Type	Description	
seq_platform	String	Description of the sequencing methodology used to generate the in-	
		put read data	
ti-	String	Date of analysis	
tan_ont_analysis_date			
titan_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	
vari-	File	Number of variants relative to the reference genome	
ants_from_ref_vcf			

# 1.3 License

GNU Affero General Public License v3.0

1.3. License 31