viral-ngs Documentation

Release v2.0.21.1

Broad Institute Viral Genomics

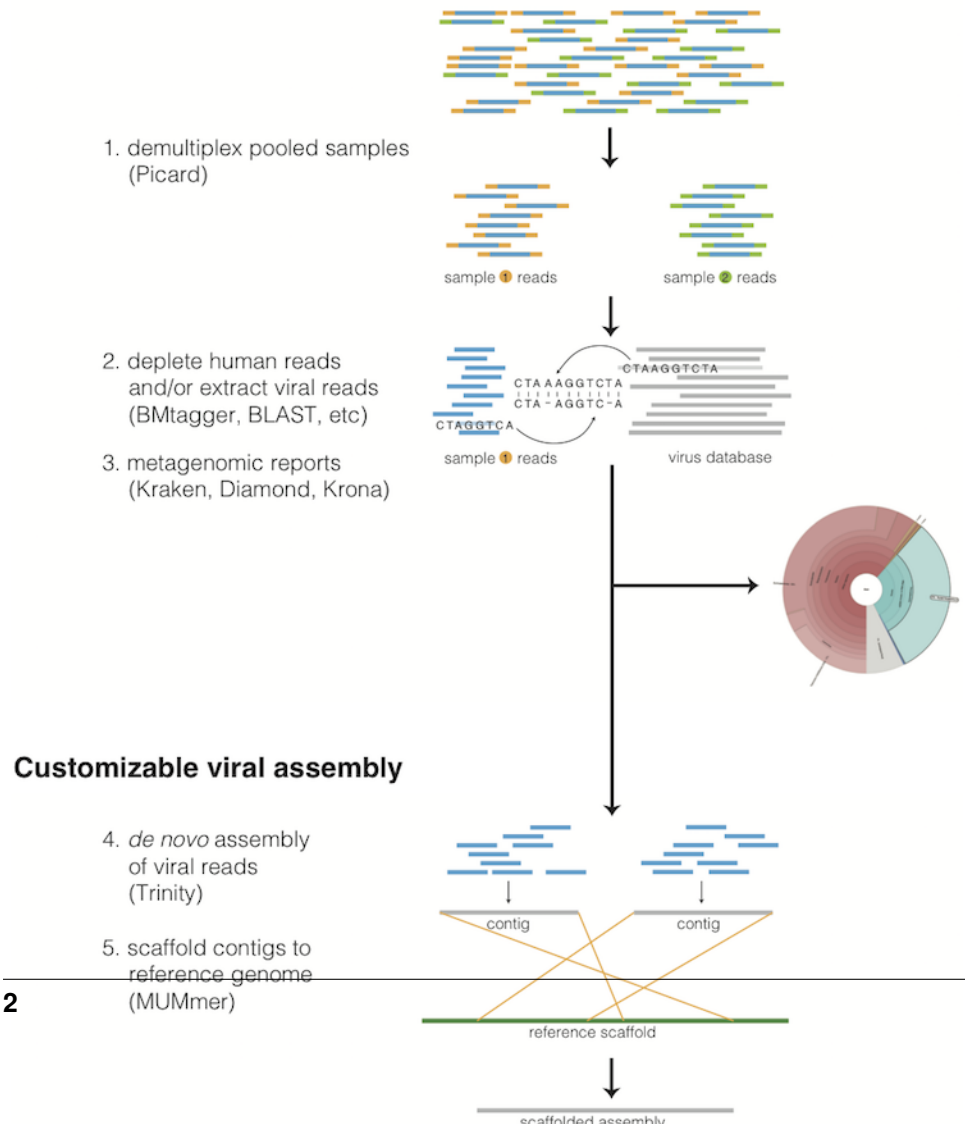
2020-05-03

Contents

CHAPTER 1

Contents

1.1 Description of the methods



1.1.1 Taxonomic read filtration

Human, contaminant, and duplicate read removal

The assembly pipeline begins by depleting paired-end reads from each sample of human and other contaminants using BMTAGGER and BLASTN, and removing PCR duplicates using M-Vicuna (a custom version of Vicuna).

Taxonomic selection

Reads are then filtered to to a genus-level database using LASTAL, quality-trimmed with Trimmomatic, and further deduplicated with PRINSEQ.

1.1.2 Viral genome analysis

Viral genome assembly

The filtered and trimmed reads are subsampled to at most 100,000 pairs. *de novo* assemby is performed using Trinity. SPAdes is also offered as an alternative *de novo* assembler. Reference-assisted assembly improvements follow (contig scaffolding, orienting, etc.) with MUMMER and MUSCLE or MAFFT. Gap2Seq is used to seal gaps between scaffolded *de novo* contigs with sequencing reads.

Each sample's reads are aligned to its *de novo* assembly using Novoalign and any remaining duplicates were removed using Picard MarkDuplicates. Variant positions in each assembly were identified using GATK IndelRealigner and UnifiedGenotyper on the read alignments. The assembly was refined to represent the major allele at each variant site, and any positions supported by fewer than three reads were changed to N.

This align-call-refine cycle is iterated twice, to minimize reference bias in the assembly.

Intrahost variant identification

Intrahost variants (iSNVs) were called from each sample's read alignments using V-Phaser2 and subjected to an initial set of filters: variant calls with fewer than five forward or reverse reads or more than a 10-fold strand bias were eliminated. iSNVs were also removed if there was more than a five-fold difference between the strand bias of the variant call and the strand bias of the reference call. Variant calls that passed these filters were additionally subjected to a 0.5% frequency filter. The final list of iSNVs contains only variant calls that passed all filters in two separate library preparations. These files infer 100% allele frequencies for all samples at an iSNV position where there was no intra-host variation within the sample, but a clear consensus call during assembly. Annotations are computed with snpEff.

1.1.3 Taxonomic read identification

Metagenomic classifiers include Kraken and Diamond. In each case, results are visualized with Krona.

1.2 Using the WDL pipelines

Rather than chaining together viral-ngs pipeline steps as series of tool commands called in isolation, it is possible to execute them as a complete automated pipeline, from processing raw sequencer output to creating files suitable for GenBank submission. This utilizes the Workflow Description Language, which is documented at: https://github.com/openwdl/wdl

There are various methods for executing these workflows on your infrastructure which are more thoroughly documented in our README.

1.3 WDL Workflows

Documentation for each workflow is provided here. Although there are many workflows that serve different functions, some of the primary workflows we use most often include:

- demux_plus (on every sequencing run)
- classify_krakenuniq (included in demux_plus)
- assemble_denovo (for most viruses)
- assemble_refbased (for less diverse viruses, such as those from single point source human outbreaks)
- build_augur_tree (for nextstrain-based visualization of phylogeny)

1.3.1 align_and_count_report

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.2 align_and_plot

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.3 assemble_denovo

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.4 assemble_refbased

Reference-based microbial consensus calling. Aligns short reads to a singular reference genome, calls a new consensus sequence, and emits: new assembly, reads aligned to provided reference, reads aligned to new assembly, various figures of merit, plots, and QC metrics. The user may provide unaligned reads spread across multiple input files and this workflow will parallelize alignment per input file before merging results prior to consensus calling.

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.5 bams_multiqc

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.6 build_augur_tree

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.7 classify_kaiju

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.8 classify_krakenuniq

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.9 contigs

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.10 coverage_table

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.11 demux_metag

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.12 demux_only

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.13 demux_plus

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.14 deplete_only

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.15 downsample

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.16 fastq_to_ubam

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.17 fetch_annotations

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.18 multi_Fetch_SRA_to_BAM

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.19 filter_classified_bam_to_taxa

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.20 genbank

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.21 isnvs_merge_to_vcf

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.22 isnvs_one_sample

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.23 mafft

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.24 mafft_and_trim

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.25 merge_bams

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.26 merge_metagenomics

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.27 merge_tar_chunks

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.28 multiqc_only

Inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.29 scaffold_and_refine

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)

1.3.30 trimal

Inputs

Required inputs

Other inputs

Generated using WDL AID (0.1.1)