Varpipe_wgs 1.0.2 pipeline documentation

DISCLAIMER

The Laboratory Branch (LB) of the Division of Tuberculosis Elimination developed this bioinformatic pipeline for analyzing whole genome sequencing data generated on Illumina platforms. This is not a controlled document. The performance characteristics as generated at Centers for Disease Control and Prevention (CDC) are specific to the version as written. These documents are provided by LB solely as an example for how this test performed within LB. The recipient testing laboratory is responsible for generating validation or verification data as applicable to establish performance characteristics as required by the testing laboratory's policies, applicable regulations, and quality system standards. These data are only for the sample and specimen types and conditions described in this procedure. Tests or protocols may include hazardous reagents or biological agents. No indemnification for any loss, claim, damage, or liability is provided for the party receiving an assay or protocol. Use of trade names and commercial sources are for identification only and do not constitute endorsement by the Public Health Service, the United States Department of Health and Human Services, or the Centers for Disease Control and Prevention

1.0 Pipeline Software

This is a bioinformatic pipeline developed to analyze *Mycobacterium tuberculosis* whole genome sequencing (WGS) data generated on Illumina NGS platforms. The pipeline incorporates several open-sourced tools and custom python scripts to accept fastq files, map the reads against *Mycobacterium tuberculosis* H37Rv (NC_000962.3) reference genome, identify variants in the sample genome and provide a coverage report (list of tools and custom scripts in Supplementary tables 1 & 4). The results are reported in the standard variant call file (VCF) format as well as a printable PDF file format.

1.1 Description
Name: Varpipe_wgs
Version: v1.0.2
Stability: Stable production version, developed locally by SME

1.2 Pipeline Execution
Installation:
The pipeline software is locally installed on the biolinux.biotech.cdc.gov server at the CDC Scicomp unit.
Analysis runs:
Samples can be analyzed individually or in batches through the pipeline depending on the Bash user command presented at the Linux interface for the software. The pipeline achieves running multiple analysis sequentially via a Bash wrapper script included in the software.

1.3 Run-time Error tracking

The pipeline tracks all run-time errors and warnings and details those in the analysis run log. The user has access to the run logs at the end of each analysis, and if the analysis is terminated before completion, the user would need to re-run the analysis from the beginning.

1.4 Pipeline Versioning

The source code for the pipeline and details on the dependencies are managed and version tracked on Gitlab. The details of the pipeline including the current version as well as the history of previous versions and updates can be retrieved and are available on this CDC GitLab location: https://git.biotech.cdc.gov/krt7/varpipe_wgs

1.5 File Naming Convention

The initial input fastq files and all the subsequent intermediary and results files all have three unique identifiers within their names to ensure the files are always correctly tracked. The identifiers include the unique sample id, the date the sample was sequenced, and the identification for the instrument used for sequencing. The naming convention for a sample file will look this: <Sample id>-<Run date>-<instrument id>-<file type suffix>.

1.6 Pipeline Third-Party Tools Versions and Integrations

Decontamination of reads

This step removes contaminating reads from unwanted species in the sample file. Reads that map exclusively to the contaminant genomes but not to the reference genomes are excluded from the analysis. The goal is to ensure that variants from contaminating species do not confound the interpretation of the pipeline analysis results.

Tool: Clockwork version 0.11.3
Input file: sample fastq reads (sample_R*001.fastq.gz)
Output file: decontaminated sample fastq reads
Requirement: Customized reference genome database (contaminant genomes [human, HIV, *Mycobacterium avium, Mycobacterium chimaera, Mycobacterium species VKM, Mycobacterium species* QIA 37, *Mycobacterium fortuitum, Mycobacterium abscessus, Mycobacterium chelonae*]; reference genome [*Mycobacterium tuberculosis* NC_000962.3)

Trimming of reads

This step trims the leading and trailing sequence of reads and reads with low quality nucleotide base calls

Tool: Trimmomatic version 0.39
Input file: decontaminated sample fastq reads
Output file: trimmed fastq reads
Parameters: minimum base quality: Q15
minimum read length: 40 nucleotide base pairs,
leading base pairs to trim: 3

trailing base pairs to trim: 3

Mapping & Alignment file Refinement

The sequence reads are mapped to the *Mycobacterium tuberculosis* (NC_000962.3) reference genome in this step. The mapped reads in an alignment file are also refined, sorted, and indexed to enable easy processing by downstream tools.

Tools: BWA MEM version 0.7.17; Piccard tools Version 2.26.10
Input file: trimmed fastq reads
Output file: sorted and indexed BAM file
Requirements: H37Rv reference genome fasta file (NC_000962.3, https://www.ncbi.nlm.nih.gov/nuccore/NC_000962.3)
Parameters: Default tool settings

Variant calling & filtering

Variants in the alignment file are identified in this step, and the output variant file is filtered based on set criteria to remove spurious calls.

Tool: GATK Version 4.2.4.0
Command: `Mutect2`
Input file: sorted and indexed BAM file
Output file: VCF files (sample_full_raw.vcf and sample_DR_loci_raw.vcf)
Parameters: Max-MNP-Dist: 2
Recommended: BED File containing genome coordinates of interest (intervals.bed) (supplementary table 2)

Command: `FilterMutectCalls`
Input file: VCF files (sample_full_raw.vcf and sample_DR_loci_raw.vcf)
Output file: filtered VCF file (sample_full_filtered.vcf and sample_DR_loci_filtered.vcf)
Parameters: minimum-reads-per-strand: 1
minimum-median-read-position:10
microbial mode: True
minimum-allele-fraction: 1%

Variant Annotation:

In this step, the VCF file is annotated to give a sense of the consequence of the variants identified in the genome. The annotated VCF file is also rendered more human readable using a few custom scripts.

Tool: SnpEff Version 4.3
Description: This tool performs the functional annotation of the raw VCF
Input file: filtered VCF file (sample_full_filtered.vcf and sample_DR_loci_filtered.vcf)

Output file: raw annotated VCF file (sample_full_raw_annotation.txt and sample_DR_loci_raw_annotation.txt)

Tool: create_annotation.py
Description: This script generates a detailed annotated file
Input file: raw annotated VCF file (sample_full_raw_annotation.txt, sample_DR_loci_raw_annotation.txt)
Output file: annotated text file (sample_full_annotation.txt, sample_DR_loci_annotation.txt)

Tool: parse_annotation.py
Description: This script parses the raw annotated file, selecting only records of variants with allele frequency of a minimum of 5%. It also presents the annotation in an easily readable tab-delimited format in the final annotation text file.
Input file: raw annotated VCF file (sample_full_raw_annotation.txt, sample_DR_loci_raw_annotation.txt)
Output file: Final annotation text file (sample_full_Final_annotation.txt, sample_DR_loci_Final_annotation.txt)

Coverage Analysis

In this step, the depth and percentage of coverage for the target regions in the sample is estimated. Summary statistics on the total and number and proportion of mapped reads in the sample is also calculated.

Tools: Samtools Version 1.11, BEDTools version 2.18.2, target_coverage_estimator_parser.py

Together this group of tools generate the target region coverage report that indicates the depth and percent of coverage across regions of interest.

Command: `Samtools depth`
Input file: sorted and indexed BAM file
Output file: Samtools depth coverage file

Command: `BEDTools coverage`
Input files: sorted and indexed BAM file, BED file of target regions (amp_bed.txt)
Output file: BEDTools coverage output file

Command: `target_coverage_estimator.py`
Input files: Samtools depth coverage output file, BEDTools coverage output file
Output file: target region coverage text file (sample_target_region_coverage.txt)

Command: `genome_coverage_estimator.py`

Input files: Samtools depth coverage output file, BEDTools coverage output file
Output file: genome region coverage text file (sample_genome_region_coverage.txt)

Tools: Samtools Version 1.11, stats_estimator.py
Description: These tools generate summary statistics on the sample analyzed.
Commands: `Samtools view -c`
Input files: BAM files with unmapped reads, BAM file with unmapped reads filtered
Output files: mapped reads count text file, total reads count text file
Command: `stats_estimator.py`
Input file: target region coverage text file, mapped reads count text file, total reads count text file
Output file: stats text file (sample_stats.txt)

Summary Report

The final PDF report that includes a summary of the final annotation file, the stats file and the target region coverage file is generated here.

Tools: create_report.py, pdf_print.py
Description: This script generates the final PDF report from the pipeline analysis
Input files: stats text file, target region coverage text file and Final annotation text file
Output file: summary.txt, summary report PDF file (sample_report.pdf).

Drug Susceptibility Variant Interpretation

Development of Variant Interpretation Look-up Table (Supplementary Table 3)

A list of variants with well characterized association with resistance to isoniazid (INH), rifampicin (RIF), pyrazinamide (PZA), ethambutol (EMB), and fluoroquinolone (FQ) was created using data from the 2021 WHO catalogue of *Mycobacterium tuberculosis* complex mutations [3] (Catalogue of mutations in *Mycobacterium tuberculosis* complex and their association with drug resistance (who.int) associated with drug resistance and from CDC Molecular Detection of Drug Resistance Service in Division of TB Elimination Laboratory Branch (service implemented 2009). Variants listed in the WHO catalogue [3] as WHO: 1) Assoc w EMB resistance (R) were characterized as "EMB-R", WHO: 1) Assoc w INH R were characterized as "INH-R", WHO: 1) Assoc w levofloxacin (LEV) R were characterized as "FQ-R", WHO: 1) Assoc w moxifloxacin (MXF) R were characterized as "FQ-R", WHO: 1) Assoc w PZA R, were characterized as "PZA-R", WHO: 1) Assoc w RIF R were characterized as "RIF-R", WHO: 5) Not Assoc w R were characterized as susceptible or "S", WHO: 4) Not Assoc w R - Interim were considered unknown in absence of additional data, WHO: 2) Assoc w PZA R – Interim were characterized as "PZA-R", WHO: 2) Assoc w EMB R – Interim were considered "EMB-unknown (U)" in absence of MDDR data, WHO: 2) Assoc w MXF R – Interim were characterized as "FQ-R", WHO: 2) Assoc w INH R – Interim were considered "INH-U" in absence of MDDR data, and WHO: 2) Assoc w RIF R – Interim were considered "RIF-U" in absence of MDDR data. MDDR data or data collected from published research performed in

our laboratory was used to characterize variants not listed in the WHO catalog. The look up table will be updated yearly.

Interpretation of variants in loci associated with isoniazid resistance

fabG1 NC_000962.3:1673440-1674183 (+ strand)
DRLociBed: 1673409-1674052 (-30 to codon 204)
Report: only c.-17,c.-15, c.-8 and c.609 (amino acid 203)
Interpret: Reportable variants are initially interpreted as “INH-U”. The interpretation of variants found in the Interpretation Look-up Table (Supplementary table 3) is updated as indicated in the table.

katG NC_000962.3:2153889-2156111 (- strand)
DRLociBed: 2153888-2156112 (coding region all)
Report: [any "Non-synonymous"] OR ["synonymous" at codon 1]
Interpret: Reportable variants are initially interpreted as “INH-U”. The interpretation of variants annotated as loss of function including loss of start codon variants, premature stop codon variants and frameshift variants is updated to “INH-R”. The interpretation of variants found in the Interpretation Look-up Table (Supplementary table 3) is updated as indicated in the table.

Interpretation of variants in loci associated with rifampicin resistance

rpoB NC_000962.3:759807-763325 (+ strand)
DRLoicBed: 760307-761286 (coding region 167-493)
Report: ["Non-synonymous" at codon 170] OR ["all" at codons 426 to 452] OR ["Non-synonymous" at codon 491]
Interpret: Reportable variants are initially interpreted as “RIF-U”. The interpretation of synonymous variants is updated to “S”. The interpretation of variants annotated as in-frame insertion or deletion is updated to “RIF-R”. The interpretation of variants found in the Interpretation Look-up Table (Supplementary table 3) is updated as indicated in the table.
Note: Synonymous variants are included because they may be detected by commercially available molecular assays and interpreted as false resistance by these tests.

Interpretation of variants in loci associated with pyrazinamide resistance

pncA NC_000962.3:2288681-2289241 (- strand)
DRLociBed: 2288676-2289272 (-30 bp and coding region all)
Report: [all "Non-coding"] OR [all "Non-synonymous"] OR ["synonymous" at codon 1]
Interpret: Reportable variants are initially interpreted as “PZA-U”. The interpretation of variants annotated as loss of function including loss of start codon variants, premature stop codon variants and frameshift variants is updated to “PZA-R”. The interpretation of variants

found in the Interpretation Look-up Table (Supplementary table 3) is updated as indicated in the table.

Interpretation of variants in loci associated with ethambutol resistance

embB NC_000962.3:4246514-4249810 (+ strand)
DRLociBed: 4246513-4249811 (exclude 4246524-4246586, 4248314-4248329, 4249653-4249692)
Report: “all Non-synonymous"
Interpret: Reportable variants are initially interpreted as “EMB-U”. The interpretation of variants found in the Interpretation Look-up Table (Supplementary table 3) is updated as indicated in the table.
Note: Regions of the *embB* locus are excluded due to the detection of large numbers of sequencing artifacts in these regions.

Interpretation of variants in loci associated with fluoroquinolone resistance

gyrB NC_000962.3:5240-7267 (+ strand)
DRLociBed: 6571-6762 (coding region 446-507)
Report: "all Non-synonymous"
Interpret: Reportable variants are initially interpreted as “FQ-U”. The interpretation of variants found in the Interpretation Look-up Table (Supplementary table 3) is updated as indicated in the table.

gyrA NC_000962.3:7302-9818 (+ strand)
DRLociBed: 7360-7583 (coding region 20-94)
Report: "all Non-synonymous (codons 88-94)
Interpret: Reportable variants are initially interpreted as “FQ-U”. The interpretation of variants found in the Interpretation Look-up Table (Supplementary table 3) is updated as indicated in the table.

Final prediction of antimicrobial resistance

An isolate of *M. tuberculosis* complex is reported as resistant to isoniazid based on the presence of any variant interpreted as “INH-R”, is reported as susceptible to isoniazid based on the absence of any variants interpreted as “INH-R” or “INH-U” and is reported as isoniazid susceptibility unknown based on the presence of variants interpreted as “INH-U” in the absence of variants interpreted as “INH-R”.

An isolate of *M. tuberculosis* complex is reported as resistant to rifampicin based on the presence of any variant interpreted as “RIF-R”, is reported as susceptible to rifampicin based on the absence of any variants interpreted as “RIF-R” or “RIF-U” and is reported as rifampicin susceptibility unknown based on the presence of variants interpreted as “RIF-U” in the absence of variants interpreted as “RIF-R”.

An isolate of *M. tuberculosis* complex is reported as resistant to pyrazinamide based on the presence of any variant interpreted as “PZA-R”, is reported as susceptible to pyrazinamide based on the absence of any variants interpreted as “PZA-R” or “PZA-U” and is reported as pyrazinamide susceptibility unknown based on the presence of variants interpreted as “PZA-U” in the absence of variants interpreted as “PZA-R”.

An isolate of *M. tuberculosis* complex is reported as resistant to ethambutol based on the presence of any variant interpreted as “EMB-R”, is reported as susceptible to ethambutol based on the absence of any variants interpreted as “EMB-R” or “EMB-U” and is reported as ethambutol susceptibility unknown based on the presence of variants interpreted as “EMB-U” in the absence of variants interpreted as “EMB-R”.

An isolate of *M. tuberculosis* complex is reported as resistant to fluoroquinolone based on the presence of any variant interpreted as “FQ-R”, is reported as susceptible to fluoroquinolone based on the absence of any variants interpreted as “FQ-R” or “FQ-U” and is reported as fluoroquinolone susceptibility unknown based on the presence of variants interpreted as “FQ-U” in the absence of variants interpreted as “FQ-R”.

Tools: interprete.py
Input files: summary.txt, reportable.txt, DR_loci_annotation.txt, structural_variants.txt, target_coverage.txt

1.7 Metadata Capture

Metadata on each sample is captured in the log file that is a part of the result output for each individual sample. The log file lists details on the user id, the input command, and parameters for each of the tools in the pipeline, the different steps implemented in the analysis, the input filenames, the names of all the intermediary files created and the output file names and location. The log file also has all the runtime messages and warnings from the third-party tools implemented in the pipeline.
In addition to the log file, the stats text file which is included in the results output also captures metadata information about the name and version of the pipeline used, the date and time of the analysis, and quality metrics for the sample.

1.8 Output files of varpipe_wgs

The list of output files and a brief description are included in the table below.

Varpipe_wgs v1.0.2 output files	**Description**
Log.txt	
Stat.txt	Percent mapped reads, average genome coverage depth, pecent reference genome covered, pipeline version, analysis date
Lineage.txt	Lineage name

Region_coverage.txt	Average depth and percent region coverage for each open reading frame and intragenic region in the reference genome
Structural_variant.txt	Description of large deletions in pncA or furA..katG
full_raw_annotation.txt	Unfiltered, annotated vcf with header
DR_loci_raw_annotation.txt	
full_annotation.txt	Unfiltered, annotated vcf formatted for databasing
DR_loci_annotation.txt	
full_final_annotation.txt	Filtered annotated vcf (pass filter, >5% allelic frequency)
DR_loci_final_annotation.txt	
Summary.txt (report.pdf)	Sample statistics, loci coverage, reported variants and interpretation
Interpretation.txt	Effect of variants on antibiotic resistance

References

1. Shea J, Halse TA, Lapierre P, Shudt M, Kohlerschmidt D, Van Roey P, Limberger R, Taylor J, Escuyer V, Musser KA; Comprehensive Whole Genome Sequencing and reporting of Drug Resistance Profiles on Clinical cases of *Mycobacterium tuberculosis* in New York State; *J Clin Microbiol* 2017 Jun;55(6):1871-1882
2. Campbell P, Morlock G, Sikes D, Dalton T, Metchock B, Starks A, Hooks D, Cowan S, Plikaytis B, Posey J; Molecular Detection of Mutations Associated with First and Second-Line Drug Resistance Compared with Conventional Drug Susceptibility Testing of *Mycobacterium tuberculosis; Antimicrobial Agents and Chemotherapy* May 2011;55(5):2032-2041

3. Walker TM, Miotto P, Köser CU, Fowler PW, Knaggs J, Iqbal Z, Hunt M, Chindelevitch L, Farhat M, Cirillo DM, Comas I, Posey J, Omar SV, Peto TE, Suresh A, Uplekar S, Laurent S, Colman RE, Nathanson CM, Zignol M, Walker AS; CRyPTIC Consortium; Seq&Treat Consortium, Crook DW, Ismail N, Rodwell TC. The 2021 WHO catalogue of *Mycobacterium tuberculosis* complex mutations associated with drug resistance: A genotypic analysis. Lancet Microbe. 2022 Apr;3(4):e265-e273. doi: 10.1016/S2666-5247(21)00301-3. PMID: 35373160; PMCID: PMC7612554.