



# The COMBAT-TB Workbench: Making Powerful *Mycobacterium tuberculosis* Bioinformatics Accessible

 Peter van Heusden,<sup>a</sup> Ziphozakhe Mashologu,<sup>a</sup> Thoba Lose,<sup>a</sup> Robin Warren,<sup>b</sup>  Alan Christoffels<sup>a,c</sup>

<sup>a</sup>South African Medical Research Council Bioinformatics Unit, South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa

<sup>b</sup>DSI-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

<sup>c</sup>Africa Centres for Disease Control and Prevention, African Union Headquarters, Addis Ababa, Ethiopia

**ABSTRACT** Whole-genome sequencing (WGS) is a powerful method for detecting drug resistance, genetic diversity, and transmission dynamics of *Mycobacterium tuberculosis*. Implementation of WGS in public health microbiology laboratories is impeded by a lack of user-friendly, automated, and semiautomated pipelines. We present the COMBAT-TB Workbench, a modular, easy-to-install application that provides a web-based environment for *Mycobacterium tuberculosis* bioinformatics. The COMBAT-TB Workbench is built using two main software components: the IRIDA platform for its web-based user interface and data management capabilities and the Galaxy bioinformatics workflow platform for workflow execution. These components are combined into a single easy-to-install application using Docker container technology. We implemented two workflows, for *M. tuberculosis* sample analysis and phylogeny, in Galaxy. Building our workflows involved updating some Galaxy tools (Trimmomatic, snippy, and snp-sites) and writing new Galaxy tools (snp-dists, TB-Profiler, tb\_variant\_filter, and TB Variant Report). The irida-wf-ga2xml tool was updated to be able to work with recent versions of Galaxy and was further developed into IRIDA plugins for both workflows. In the case of the *M. tuberculosis* sample analysis, an interface was added to update the metadata stored for each sequence sample with results gleaned from the Galaxy workflow output. Data can be loaded into the COMBAT-TB Workbench via the web interface or via the command line IRIDA uploader tool. The COMBAT-TB Workbench application deploys IRIDA, the COMBAT-TB IRIDA plugins, the MariaDB database, and Galaxy using Docker containers (<https://github.com/COMBAT-TB/irida-galaxy-deploy>).

**IMPORTANCE** While the reduction in the cost of WGS is making sequencing more affordable in lower- and middle-income countries (LMICs), public health laboratories in these countries seldom have access to bioinformaticians and system support engineers adept at using the Linux command line and complex bioinformatics software. The COMBAT-TB Workbench provides an open-source, modular, easy-to-deploy and -use environment for managing and analyzing *M. tuberculosis* WGS data and thereby makes WGS usable in practice in the LMIC context.

**KEYWORDS** *Mycobacterium tuberculosis*, bioinformatics, multidrug resistance

Tuberculosis (TB) was until recently the world's deadliest infectious disease, infecting an estimated 10 million people in 2019 and killing 1.4 million people (1). Whole-genome sequencing (WGS) of *Mycobacterium tuberculosis*, the bacterium that causes TB, is increasingly being used, at least in high-income countries (2), for species and lineage identification, drug resistance profiling, and outbreak investigation. Increasing the use of WGS in low- and middle-income countries (LMICs) requires reducing the cost of both

**Editor** Sarah E. F. D'Orazio, University of Kentucky

**Copyright** © 2022 van Heusden et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

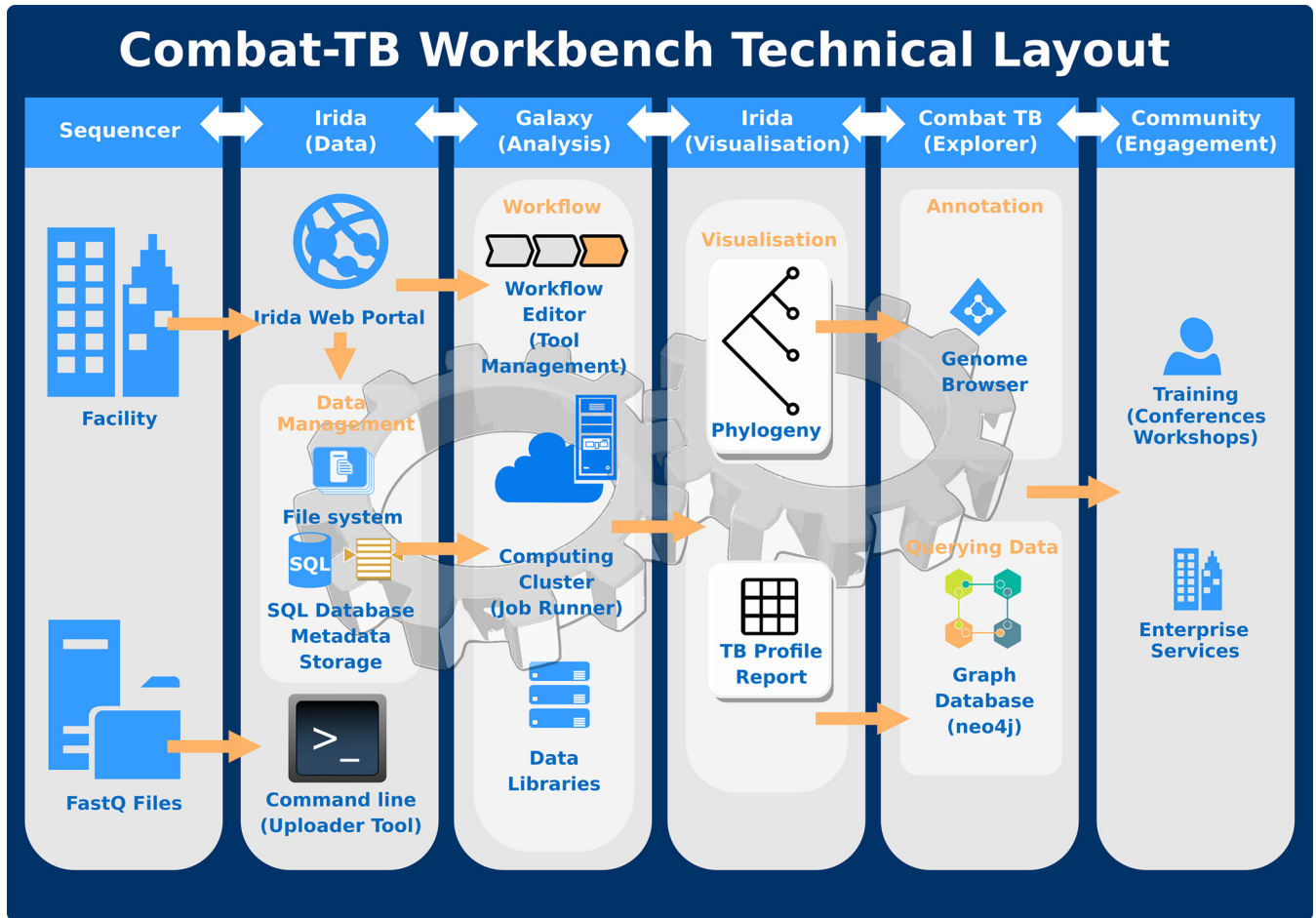
Address correspondence to Alan Christoffels, [alan@sanbi.ac.za](mailto:alan@sanbi.ac.za).

The authors declare no conflict of interest.

**Received** 18 December 2021

**Accepted** 5 January 2022

**Published** 9 February 2022



**FIG 1** COMBAT-TB Workbench technical layout. Data can be loaded into the COMBAT-TB Workbench via the web interface (e.g., storage at a sequencing facility) or via the command line IRIDA uploader tool. Sequence data are stored in IRIDA on disk, and metadata are stored in a MariaDB database. Sequence data are shared between IRIDA and the Galaxy analysis platform. The workbench uses Galaxy for its bioinformatics workflow composition and execution.

sequencing and bioinformatics analysis of sequencing results and reducing the time and effort involved in going from sequence to analysis results (3).

This cost takes the form of equipment, consumables, and expertise. The command line tools whose use is prevalent in bioinformatics (4) require skills not readily accessible outside specialist labs. On the other hand, web-based tools such as TB-Profiler (5) and the NIAID TB Portals (6) are restricted to the analyses provided by their authors and often lack features for bulk analysis. Finally, platforms like Galaxy (7), while customizable, do not provide a data management and analysis user interface specific for common *M. tuberculosis* analysis tasks.

To address these deficiencies, the computational bacterial analytical toolkit for Tuberculosis research (COMBAT-TB) was developed. The COMBAT-TB Workbench (downloadable from <https://github.com/COMBAT-TB/irida-galaxy-deploy>) represents a modular and accessible open-source workbench for storing and analyzing *M. tuberculosis* and other microbial WGS data.

## RESULTS

**Design and implementation.** The COMBAT-TB Workbench is built using two main software components: the Integrated Rapid Infectious Disease Analysis (IRIDA) platform (8) for its web-based user interface and data management capabilities and the Galaxy bioinformatics workflow platform for workflow execution (Fig. 1). These components are combined into a single easy-to-install application using Docker container technology.

**Data management and user interface.** The IRIDA platform, a project of the Public Health Agency of Canada (PHAC-NML), is a web application written in Java that provides a

user-friendly web interface for sequencing data and metadata storage, workflow execution, and result visualization. We adopted IRIDA as the basis of the COMBAT-TB Workbench because of its integration with the Galaxy platform, its proven track record in public health bioinformatics (PHAC-NML has analyzed over 200,000 pathogen samples using IRIDA [T. Matthews, personal communication]), and the fact that it is open source.

IRIDA stores sequence samples on disk and sample metadata in a MariaDB database (10). Sequence data are shared between IRIDA and the Galaxy analysis platform (reducing data duplication).

**Scientific workflows and IRIDA plugins.** The Workbench uses Galaxy for its bioinformatics workflow composition and execution. Two workflows, for *M. tuberculosis* sample analysis and phylogeny, were implemented in Galaxy. Building workflows in Galaxy involves connecting Galaxy tools to construct an analysis workflow where the Galaxy tools (also known as tool wrappers) themselves connect command line bioinformatics tools to the Galaxy framework. Building our workflows involved updating some Galaxy tools (Trimmomatic [11], snippy [12], and snp-sites [13]) and writing new Galaxy tools (snp-dists [14], TB-Profler [5], tb\_variant\_filter [15], and TB Variant Report [16]). In addition to the work on Galaxy tools, the command line tb\_variant\_filter and TB Variant Report tools were created as part of the COMBAT-TB project.

The irida-wf-ga2xml tool (17) was updated to be able to work with recent versions of Galaxy, and it was used to build IRIDA plugin skeletons. These plugin skeletons were further developed into IRIDA plugins for both workflows, in the case of the *M. tuberculosis* sample analysis involving the addition of an interface between the Galaxy workflow output and the metadata stored for each sequence sample. The *M. tuberculosis* Sample Report and *M. tuberculosis* Phylogeny plugins are hosted in Github repositories (<https://github.com/COMBAT-TB/irida-plugin-tb-sample-report> and <https://github.com/COMBAT-TB/irida-plugin-tb-phylogeny>, respectively) and deployed into IRIDA as part of the COMBAT-TB Workbench deployment process.

**Deployment.** The COMBAT-TB Workbench application (<https://github.com/COMBAT-TB/irida-galaxy-deploy>) deploys IRIDA, the COMBAT-TB IRIDA plugins, the MariaDB database IRIDA uses for metadata storage, and Galaxy using Docker containers. The Docker containers are orchestrated using docker-compose (19). This allows the entire Workbench to be installed by users without advanced Linux systems administration knowledge and hides the complexity of the underlying software from the user.

**Integration with external data storage.** Data can be loaded into the COMBAT-TB Workbench via the web interface or via the command line IRIDA uploader tool. This allows data from external storage (for example, the storage of a sequencing facility) to be loaded into the workbench in bulk.

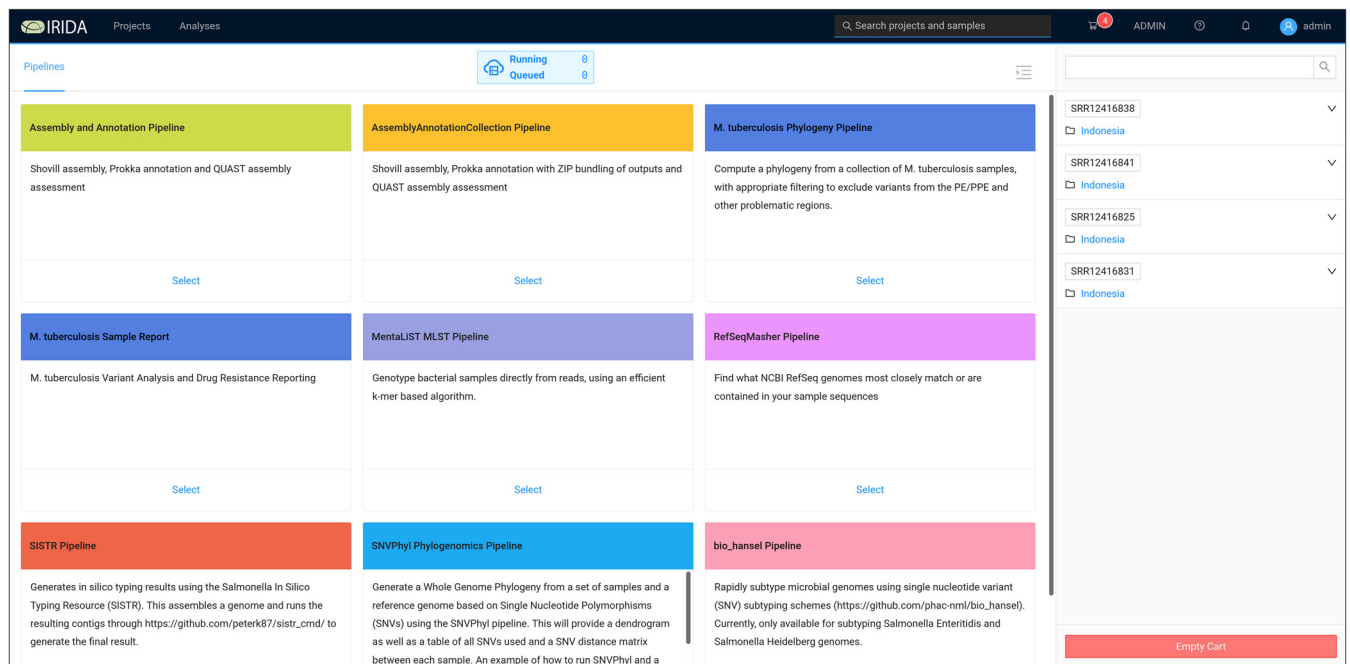
**Comparison to other similar open-source software.** At the time of writing we found two systems, Innuendo (20) and the IRIDA project (8), that were comparable to the COMBAT-TB Workbench (Table 1).

Innuendo is a web interface to sequence storage and Nextflow (21) workflow execution aimed at analysis of foodborne pathogens. It is oriented around common tasks in the foodborne pathogen surveillance terrain, such as molecular typing of pathogens. The platform is strongly tied to the workflows of a foodborne pathogen surveillance lab, and adding additional species to the analysis system requires modifying the underlying database and adding a whole-genome multilocus sequence typing (wgMLST) scheme. As such, Innuendo addresses a different challenge to the one the COMBAT-TB project tackles.

IRIDA is the official bioinformatics platform for public health genomics within the Public Health Agency of Canada. It has been in use since 2016 by Canada's provincial and national public health laboratories for genomic investigations of foodborne disease outbreaks, as part of PulseNet Canada's foodborne disease surveillance activities. While it is more flexible than Innuendo, as it allows deployment of a wide variety of Galaxy workflows and is not species specific, it is complex to deploy, with the installation guide assuming knowledge of deployment of both Galaxy and the Tomcat Java Servlet

**TABLE 1** Comparison of COMBAT-TB Workbench with other NGS analysis pipelines

Feature	COMBAT-TB	Galaxy	Innuendo	IRIDA
Version	1.0	21.05	NA	21.05
Latest commit	September 2021	October 2021	April 2018	September 2021
Workflow	Workflow plugins combine user interface specification, Galaxy workflow files, and modules for updating metadata	Galaxy workflow files	Nextflow workflows	As per first column
<i>M. tuberculosis</i> workflow	Yes	Yes	No	No
Resume if stopped	No	Yes	No	No
Reuse existing runs for expanded analysis	No	Yes	No	No
Build-in high-performance computing cluster and cloud capability	Yes	Yes	Yes	Yes
Batch upload	Yes	Yes	No	Yes
Single-sample processing from command line	No	No	No	No
Per sample metadata storage	Yes	No	Yes	Yes
Metadata upload and download	Yes	No	Yes	Yes
Install from container	Yes	Yes	Yes	No
Documentation	Yes (in progress: <a href="https://docs.combat-tb-workbench.readthedocs.io/en/latest/">https://docs.combat-tb-workbench.readthedocs.io/en/latest/</a> )	Yes ( <a href="https://docs.galaxyproject.org/">https://docs.galaxyproject.org/</a> )	Yes ( <a href="https://innuendo.readthedocs.io/en/latest/">https://innuendo.readthedocs.io/en/latest/</a> )	Yes ( <a href="https://phac-nml.github.io/irida-documentation/">https://phac-nml.github.io/irida-documentation/</a> )
Github repository	<a href="https://github.com/COMBAT-TB/irida-galaxy">https://github.com/COMBAT-TB/irida-galaxy</a> -deploy	<a href="https://github.com/galaxyproject/galaxy/">https://github.com/galaxyproject/galaxy/</a>	<a href="https://github.com/theInnuendoProject/INNUENDO">https://github.com/theInnuendoProject/INNUENDO</a>	<a href="https://github.com/phac-nml/irida">https://github.com/phac-nml/irida</a>



**FIG 2** COMBAT-TB Workbench workflow selection screen. The workbench user interface is organized around projects and analyses. Samples are selected from a project and added to a cart (not shown here). The user sees the workflow selection screen after selecting a cart. One of the 9 workflows can be selected, and the corresponding workflow parameters will be set automatically for analysis. Sequences stored in the cart are displayed on the far right.

system (22). The COMBAT-TB Workbench, in contrast, is straightforward to deploy with a single Linux command.

Innuendo and IRIDA both support storing and modifying metadata for samples. While the Galaxy platform provides a flexible platform for web-based bioinformatics, it lacks similar features for organizing samples together with their metadata and forces users to maintain metadata separately (perhaps in a spreadsheet). The COMBAT TB Workbench builds on the IRIDA support for metadata storage.

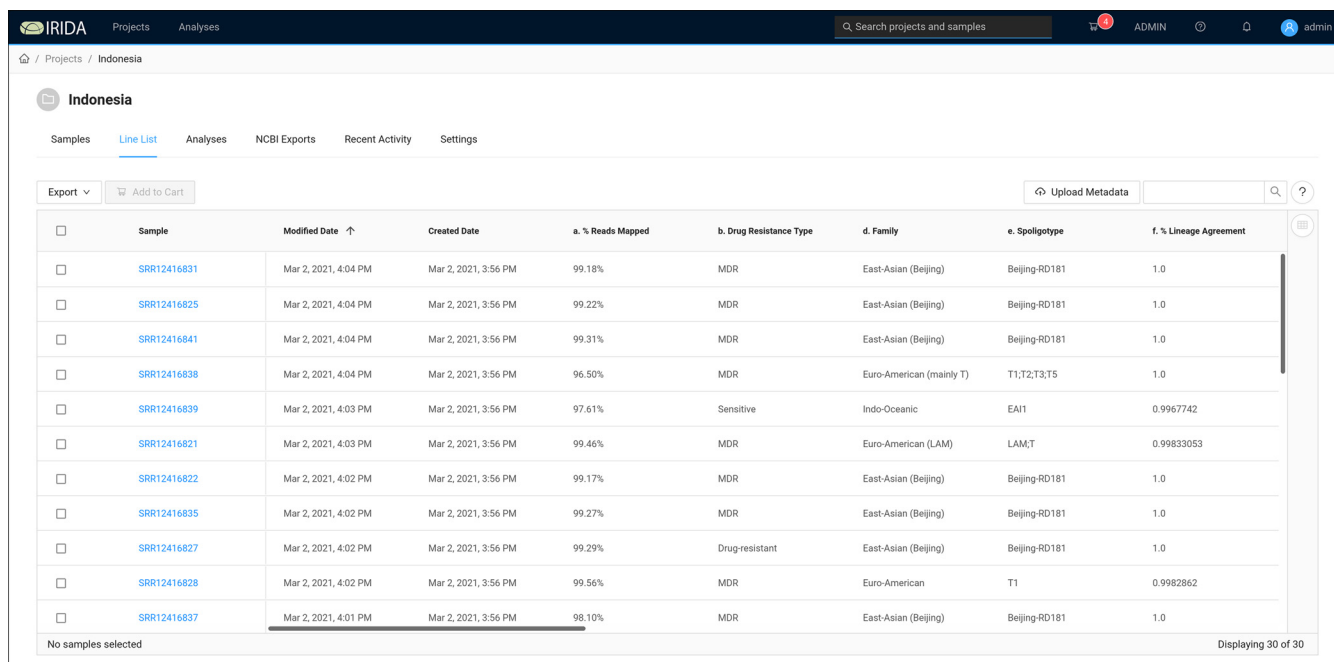
**Use case.** For the purpose of these analyses, we installed the COMBAT-TB Workbench on a virtual machine with 8 virtual CPUs, 32 GB RAM, and 3000 GB hard disk space, running Ubuntu 18.04 with Docker version 19.03.14 and docker-compose version 1.27.4.

The COMBAT-TB Workbench user interface is organized around projects and analyses. Projects store sequence samples, are associated with a reference SNV genome, and allow controlling sharing and access to samples. Samples can be selected from a project and added to a cart. Once samples are in the cart, selecting the cart displays a workflow selection screen from which analyses can be started (Fig. 2).

Selecting one of the workflows allows the workflow parameters to be set and the analysis to be started. Workflows can be either per-sample, in which case a workflow instance is started for each sample, or multisample (for example, phylogenies) in which case the samples are analyzed as a group. Once a workflow has been executed, it creates a new entry visible via the Analyses interface. This interface allows for monitoring the status of workflow execution and visualization of workflow analysis results.

**Analysis of data from MDR *M. tuberculosis* in Indonesia.** Tania et al. (23) collected *M. tuberculosis* samples from 30 patients with confirmed pulmonary tuberculosis treated in four hospitals in the western region of the Indonesian island of Java. Phenotypic drug susceptibility testing (DST) was performed on the cultured samples, and DNA was isolated and sequenced. Thirty samples were retrieved from EBI European Nucleotide Archive (ENA) and uploaded to the COMBAT-TB Workbench using the *irida-uploader* (24) command line tool. Per-sample quality control was performed automatically (using FastQC [25]) after sample uploading. Uploading and quality control took 9 and 3 min, respectively.

The inferred ancestral *M. tuberculosis* reference genome produced by Comas et al. (26) was downloaded from Zenodo (27) and uploaded to the COMBAT-TB Workbench



Sample	Modified Date ↑	Created Date	a. % Reads Mapped	b. Drug Resistance Type	d. Family	e. Spoligotype	f. % Lineage Agreement
SRR12416831	Mar 2, 2021, 4:04 PM	Mar 2, 2021, 3:56 PM	99.18%	MDR	East-Asian (Beijing)	Beijing-RD181	1.0
SRR12416825	Mar 2, 2021, 4:04 PM	Mar 2, 2021, 3:56 PM	99.22%	MDR	East-Asian (Beijing)	Beijing-RD181	1.0
SRR12416841	Mar 2, 2021, 4:04 PM	Mar 2, 2021, 3:56 PM	99.31%	MDR	East-Asian (Beijing)	Beijing-RD181	1.0
SRR12416838	Mar 2, 2021, 4:04 PM	Mar 2, 2021, 3:56 PM	96.50%	MDR	Euro-American (mainly T)	T1;T2;T3;T5	1.0
SRR12416839	Mar 2, 2021, 4:03 PM	Mar 2, 2021, 3:56 PM	97.61%	Sensitive	Indo-Oceanic	EAI1	0.9967742
SRR12416821	Mar 2, 2021, 4:03 PM	Mar 2, 2021, 3:56 PM	99.46%	MDR	Euro-American (LAM)	LAM;T	0.99833053
SRR12416822	Mar 2, 2021, 4:02 PM	Mar 2, 2021, 3:56 PM	99.17%	MDR	East-Asian (Beijing)	Beijing-RD181	1.0
SRR12416835	Mar 2, 2021, 4:02 PM	Mar 2, 2021, 3:56 PM	99.27%	MDR	East-Asian (Beijing)	Beijing-RD181	1.0
SRR12416827	Mar 2, 2021, 4:02 PM	Mar 2, 2021, 3:56 PM	99.29%	Drug-resistant	East-Asian (Beijing)	Beijing-RD181	1.0
SRR12416828	Mar 2, 2021, 4:02 PM	Mar 2, 2021, 3:56 PM	99.56%	MDR	Euro-American	T1	0.9982862
SRR12416837	Mar 2, 2021, 4:01 PM	Mar 2, 2021, 3:56 PM	98.10%	MDR	East-Asian (Beijing)	Beijing-RD181	1.0

**FIG 3** Sample line list. The sample line list holds information on samples and metadata created by the workbench itself (creation and modification date), metadata computed by the *M. tuberculosis* Sample Report plugin (% Reads Mapped, Drug Resistance Type [phenotypes, MDR or sensitive or mono-resistant], Family, Spoligotype, % Lineage Agreement). Columns not seen in the figure are the genotypes corresponding to the phenotypic information.

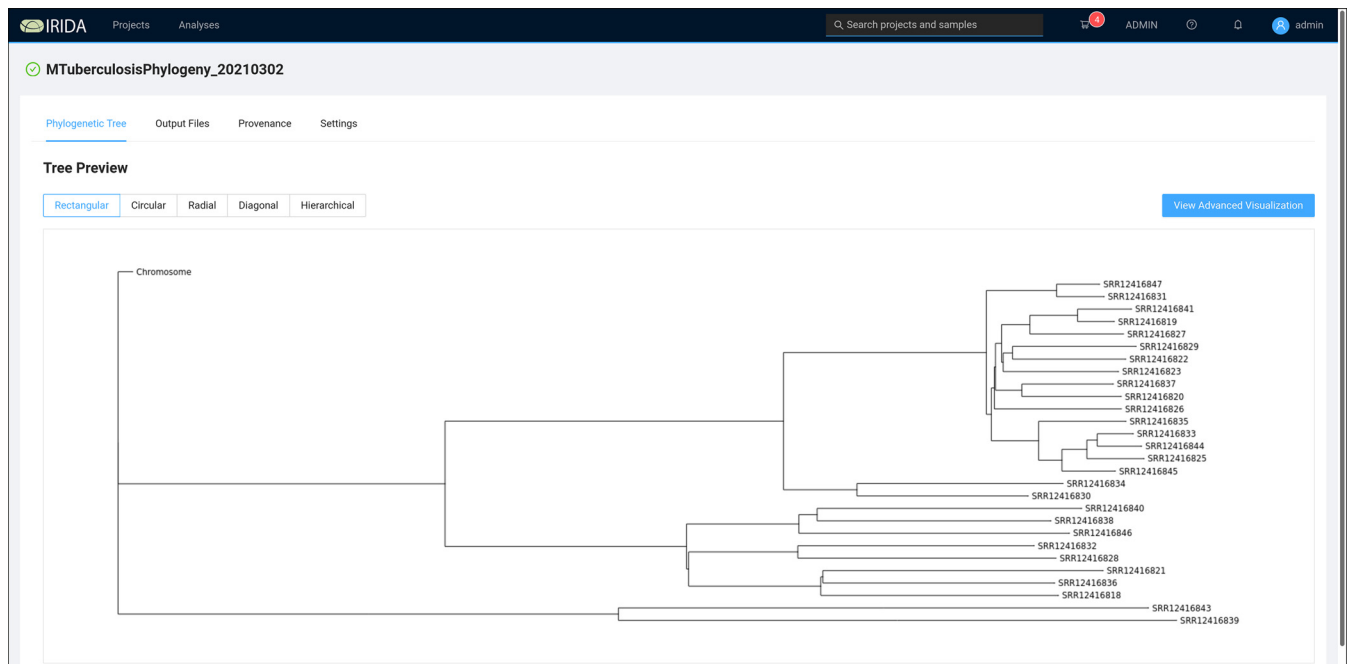
as the default reference genome for the sequencing project. While the H37Rv reference genome can be used, we used the *M. tuberculosis* inferred ancestral reference, as Goig et al. (28) showed previously that this genome is equidistant, in terms of sequence variants, from all known *M. tuberculosis* lineages and thus provides a superior reference to H37Rv (NC\_000962.3) for variant calling, especially if that variant calling is going to be used for phylogeny construction.

**Per-sample reporting.** The *M. tuberculosis* Sample Report pipeline, to generate drug resistance phenotype prediction and lineage assignment (see Materials and Methods), was run on all 30 samples. Detailed output from this pipeline is made available in a per-sample analysis report (Fig. S1) that includes a full report on variants identified in the sample (annotated using information from the COMBAT-TB NeoDB [29]), drug resistance prediction (from TB-Profler), and quality control information on read mapping.

Upon completion of the Sample Report pipeline, the output (as described in Fig. S1)—namely, drug resistance prediction, drug resistance-associated variants, and assigned lineage—was added to the metadata that were originally associated with each sample (Fig. 3).

Examining the read mapping outputs shows that only 9.63% and 1.47% of reads from samples SRR12416824 and SRR12416842, respectively, mapped against the *M. tuberculosis* genome. Further investigation with kraken2 (30) using the standard database from 14 April 2020 showed that the majority (63.55%) of reads from SRR12416824 were classified as belonging to the *Mycobacterium avium* complex (MAC) and the majority of reads (75.39%) from SRR12416842 were classified as the nontubercular mycobacterium *Mycobacterium fortuitum*. These samples were thus excluded and not used in subsequent analyses. This step provides a useful interactive space for users to consider the origin (sequence identity) of the reads before submitting it to a phylogenetic analysis pipeline.

The results were broadly concordant with those found by Tania et al. (23). Some small differences are likely a result of different versions of the TB-Profler software used. In our own analysis, updating the TB-Profler version from 2.8.4 to version 3.0.6 reduced the discordance between streptomycin resistance predicted by the COMBAT-TB Workbench and that predicted by MGIT by two samples. This illustrates the improvement of drug resistance prediction from WGS data over time. As noted above,



**FIG 4** COMBAT TB Workbench phylogeny viewer. Newick format trees created by the *M. tuberculosis* Phylogeny Pipeline are visualized in the phylogeny viewer available on the analysis results page for each Phylogeny Pipeline analysis output.

the COMBAT-TB Workbench pipeline uses snippy for mapping and variant calling and applies TB-Profiler only to the results of the mapping step. Our results are, however, wholly concordant with running the complete TB-Profiler pipeline.

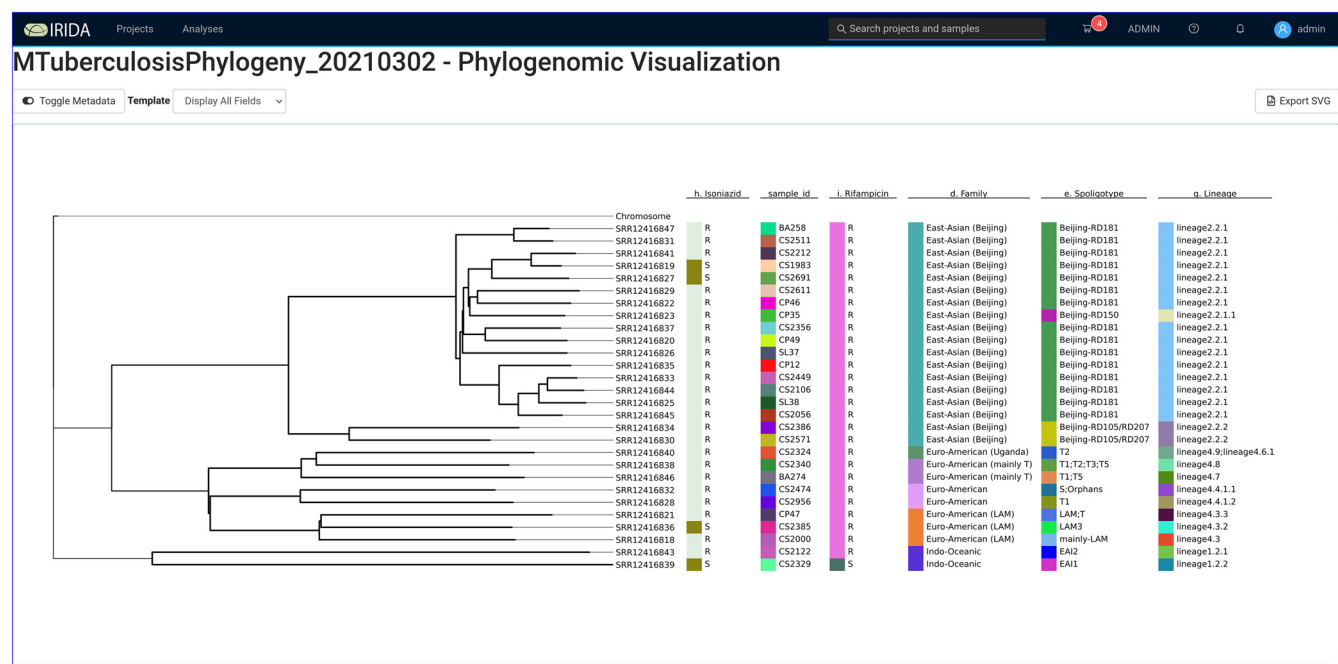
**Phylogeny on all samples.** The 28 samples that previously passed quality control were selected in the web interface and submitted to the *M. tuberculosis* Phylogeny Pipeline (<https://github.com/COMBAT-TB/irida-plugin-tb-phylogeny>). This pipeline computes a maximum-likelihood phylogeny using the single nucleotide variants (SNVs) identified in each sample (Fig. 4).

The IRIDA Advanced Visualization view (Fig. 5) allows metadata from the sequence analysis project's sequencing store to be associated with tips (i.e., samples) in the phylogeny view. In addition to the metadata computed by the *M. tuberculosis* Sample Report pipeline, an Excel spreadsheet was generated associating each sample ID with the sample IDs and hospital collection sites identified in the paper by Tania et al. (23). This spreadsheet was used to load additional metadata into the IRIDA project.

The distance between hospital sites varied from 10 km (Mampang Prapatan Hospital, Jakarta, Indonesia, to Cempaka Putih Islamic Hospital, Jakarta, Indonesia) to 107 km (Drajat Prawiranegara Hospital, Serang, Indonesia, to M. Goenawan Partowidigdo Pulmonary Hospital, Bogor, Indonesia). When data were visualized this way, it was apparent that there was no clear relationship between phylogenetic relationship (and thus lineage) and hospital site, illustrating that the outbreaks occurring in the region were circulating in areas broader than hospital catchment areas.

**Analysis of run times.** The run times of the steps in the analyses are listed in Table 2. Galaxy scheduled the execution of analysis steps in parallel when data dependencies allowed. The similarities of the run times between the sample processing and the phylogeny pipelines are to some extent due to the fact that IRIDA currently starts all analyses from sequence reads rather than from assembled genomes. This limitation is currently being addressed by the IRIDA development community (31).

**Analysis of data from *M. tuberculosis* in Spain.** Xu et al. (32) collected *M. tuberculosis* samples from 117 patients and analyzed them to understand dynamics of TB transmission in the Valencia Region, in Spain. While the key element of their analysis is identifying individual transmission patterns, their work also provides a convenient larger data set to examine the performance of the COMBAT-TB Workbench.



**FIG 5** Advanced phylogeny visualisation with metadata columns. The advanced phylogeny visualisation allows information from the metadata tables associated with samples to be combined with a phylogenetic tree and associated with colored bars, thereby displaying information about samples alongside the corresponding tips in a phylogeny.

Uploading of the 117 samples took 21 min, and per-sample analysis took a total of 6 h. A phylogeny was computed using the samples. Phylogeny generation took just over 8 h, and visualization with the advanced phylogeny viewer revealed clustering by *M. tuberculosis* lineage (Fig. S2), as expected. Unfortunately, Xu et al. (32) did not include the output of their phylogeny construction in their results, so a direct comparison of the computed phylogenetic trees is not possible.

### DISCUSSION

The COMBAT-TB Workbench makes routine *M. tuberculosis* WGS sample storage and bioinformatics analysis accessible in an extensible framework. The IRIDA platform on which it is built allows pipelines to be added as needed, and the use of Docker container technology means that installation on a machine (with the required Docker and docker-compose software) is a straightforward process not requiring advanced system administration skills.

Access to this project that is based on open-source best practices reduces the analytical cost of WGS, increasing opportunities for *M. tuberculosis* WGS deployment in low- and middle-income countries. As the capabilities of the underlying IRIDA and Galaxy platforms continue to evolve, the COMBAT-TB Workbench will naturally acquire additional features in addition to the features being added by the authors.

**TABLE 2** Runtime of analyses in COMBAT TB Workbench

Data set	Processing step	Running time
Tania et al. (23) (30 samples, 25 GB)	Upload	9 min
	Sample processing	3 h 23 min
	Phylogeny	3 h 43 min
Xu et al. (32) (117 samples, 42 GB)	Upload	21 min
	Sample processing	6 h
	Phylogeny	8 h 4 min



## MATERIALS AND METHODS

**Implementation of the COMBAT-TB Workbench.** IRIDA and Galaxy are both server environments, and IRIDA relies on a MariaDB database. The COMBAT-TB Workbench executes each of these servers in Docker containers, and the execution of the COMBAT-TB Workbench as a whole is orchestrated using docker-compose. The COMBAT-TB Workbench is deployed on any machine that runs Docker and docker-compose by fetching its code from Github (<https://github.com/COMBAT-TB/irida-galaxy-deploy>) and running a single command (“docker-compose up –build –d”). Updates to the workbench are similarly applied using a single command. The COMBAT-TB Workbench updates its IRIDA plugins from their repositories on Github on startup.

**Implementation of the *M. tuberculosis* Sample Report pipeline.** The TB sample report workflow starts by running Trimmomatic (v. 0.38.1) with the Sliding Window trimmer, truncating reads when the average quality within a 4-base window drops below a quality score of 30, followed by the minimum-length trimmer, which discards reads shorter than 20 bases. Only reads which remain in read pairs after quality trimming are used in subsequent analyses.

After quality trimming, sequence reads are mapped to a user supplied reference genome. The workflow requires a genome with the same coordinate scheme as the *M. tuberculosis* H37Rv reference (RefSeq NC000962). Mapping and variant calling are done using snippy (v. 4.4.5), a Perl-based pipeline that combines the bwa-mem (v. 0.7.17) (33) mapper and the freebayes (v. 1.3.2) (34) variant caller. Snippy has been shown (35) to produce good-quality variant calls in *M. tuberculosis* when run with default parameters. Samtools (v. 1.9) (36) flagstat is run to provide an overview of statistics from the mapping process.

After variant calling, variants are annotated with SnpEff (v. 4.3) (37) using the H37Rv reference genome annotation. Variants are then filtered using tb\_variant\_filter (v. 0.1.3). While snippy performs quality-based filtering of the variants it predicts, this tool offers a variety of filtering options commonly used in *M. tuberculosis* variant filtering. In our workflow, it filters out variants in the PE/PPE gene regions (38), in the repetitive and insertion section regions identified by UVP (39) and those with lower than 30 supporting reads or within 5 bp of an indel. These filtering options, like all tool options in the workflow, can optionally be altered by the user.

In parallel to the SnpEff annotation and variant filtering steps, the mapped reads are provided to TBProfiler (v. 2.8.4), which performs its own variant calling and lineage and drug resistance prediction.

Finally the filtered, annotated variants and the TBProfiler results are fed to tb\_vcf\_report (v. 0.1.7), which produces a report further annotated with information from the COMBAT-TB eXplorer database (29) in both text and HTML formats.

The final reports provided to the user are the variant reports from tb\_vcf\_report, text and JSON format reports from TBProfiler, variants in VCF format from SnpEff, and mapping statistics from samtools flagstats. These reports include both user-readable and raw data suitable for further downstream analysis. The metadata stored in the sample line list are updated with mapping percentage, *M. tuberculosis* lineage, spoligotype information, and drug resistance information.

**Implementation of the *M. tuberculosis* Phylogeny Pipeline.** The workflow in the phylogeny module starts with quality filtering of samples using fastp (v. 0.19.5) (40) with default settings. The filtered reads are then aligned to the user provided reference using snippy and predicted variants are filtered with tb\_variant\_filter as described above. In addition, only SNVs are retained, as the phylogeny software used in the workflow cannot extract meaningful information from indels.

For each sample, the identified sequence variants are inserted into the reference genome, yielding one sequence per sample, of the same length as the reference but with variants from the sample inserted. These are concatenated into a multiple-sequence alignment (MSA), which is used as input to snp\_dists (v. 0.6.3). Variant and constant sites are identified using snp\_sites (v. 2.5.1), and the multiple-sequence alignment is filtered to retain only variant sites. A phylogeny is then built using IQ-TREE (v. 1.5.5.1) (41–43).

The final report includes the SNP distance matrix and (Newick format) tree. The tree is also visualized in a phylogeny viewer and can also be displayed with associated sample metadata.

The SNP distance matrix, the phylogeny and the filtered variants (in VCF format) are presented to the user as output.

**Data availability.** DNA from cultured samples was deposited in the NCBI Sequence Read Archive (SRA) with BioProject accession number [PRJNA633244](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA633244).

## SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

**FIG S1**, TIF file, 2.4 MB.

**FIG S2**, TIF file, 1.8 MB.

## ACKNOWLEDGMENTS

This work was supported by The South African Research Chairs Initiatives of the Department of Science and Technology and National Research Foundation of South Africa grant UID 64751 and by the South African Medical Research Council flagship program MRC-RFA-UFSP-01-2013/COMBAT-TB.

## REFERENCES

1. World Health Organization. 2020. Global tuberculosis report 2020. World Health Organization, Geneva, Switzerland.
2. Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, Farhat MR, Guthrie JL, Laukens K, Miotto P, Ofori-Anyinam B, Dreyer V,

- Supply P, Suresh A, Utpatel C, van Soolingen D, Zhou Y, Ashton PM, Brites D, Cabibbe AM, de Jong BC, de Vos M, Menardo F, Gagneux S, Gao Q, Heupink TH, Liu Q, Loiseau C, Rigouts L, Rodwell TC, Tagliani E, Walker TM, Warren RM, Zhao Y, Zignol M, Schito M, Gardy J, Cirillo DM, Niemann S, Comas I, Van Rie A. 2019. Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues. *Nat Rev Microbiol* 17:533–545. <https://doi.org/10.1038/s41579-019-0214-5>.
3. Mahomed S, Naidoo K, Dookie N, Padayatchi N. 2017. Whole genome sequencing for the management of drug-resistant TB in low income high TB burden settings: challenges and implications. *Tuberculosis (Edinb)* 107:137–143. <https://doi.org/10.1016/j.tube.2017.09.005>.
  4. Joppich M, Zimmer R. 2019. From command-line bioinformatics to bio-GUI. *PeerJ* 7:e8111. <https://doi.org/10.7717/peerj.8111>.
  5. Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, O'Grady J, McNeerney R, Hibberd ML, Viveiros M, Huggett JF, Clark TG. 2019. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med* 11:41. <https://doi.org/10.1186/s13073-019-0650-x>.
  6. Rosenthal A, Gabrielian A, Engle E, Hurt DE, Alexandru S, Crudu V, Sergueev E, Kirichenko V, Lapitskii V, Snezhko E, Kovalev V, Astrovko A, Skrahina A, Taaffe J, Harris M, Long A, Wollenberg K, Akhundova I, Ismayilova S, Skrahin A, Mammadbayov E, Gadirova H, Abuzarov R, Seyfadinova M, Avaliani Z, Strambu I, Zaharia D, Muntean A, Ghita E, Bogdan M, Mindru R, Spinu V, Sora A, Ene C, Vashakidze S, Shubladze N, Nanava U, Tuzikov A, Tartakovskiy M. 2017. The TB portals: an open-access, web-based platform for global drug-resistant-tuberculosis data sharing and analysis. *J Clin Microbiol* 55:3267–3282. <https://doi.org/10.1128/JCM.01013-17>.
  7. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltmann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 46:W537–W544. <https://doi.org/10.1093/nar/gky379>.
  8. Matthews TC, Bristow FR, Griffiths EJ, Petkau A, Adam J, Dooley D, Kruczkiewicz P, Curatcha J, Cabral J, Fornika D, Winsor GL, Courtot M, Bertelli C, Roudgar A, Feijao P, Mabon P, Enns E, Thiessen J, Keddy A, Isaac-Renton J, Gardy JL, Tang P, Consortium TI, Carrico JA, Chindelevitch L, Chauve C, Graham MR, McArthur AG, Taboada EN, Beiko RG, Brinkman FS, Hsiao WW, Domselaar GV. 2018. The Integrated Rapid Infectious Disease Analysis (IRIDA) Platform. *bioRxiv* 381830.
  9. Reference deleted.
  10. MariaDB Foundation. 2021. MariaDB server: the open source relational database. [MariaDB.org](https://mariadb.org).
  11. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>.
  12. Seemann T. 2020. snippy: rapid haploid variant calling and core genome alignment. <https://github.com/tseemann/snippy>.
  13. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SRY. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* 2:e000056.
  14. Seemann T. 2021. snp-dists: convert a FASTA alignment to SNP distance matrix. <https://github.com/tseemann/snp-dists>.
  15. van Heusden P. 2021. tb\_variant\_filter: a tool for filtering VCF files (relative to *M. tuberculosis* H37Rv). Python. [https://github.com/COMBAT-TB/tb\\_variant\\_filter](https://github.com/COMBAT-TB/tb_variant_filter).
  16. Lose T, van Heusden P. 2021. tbvcfreport: generate an interactive HTML-based report from *M. tuberculosis* SnpEff annotated VCF(s). Python. <https://github.com/COMBAT-TB/tbvcfreport>.
  17. Thiessen J, Fornika D, Kruczkiewicz P, Petkau A, van Heusden P. 2021. irida-wf-ga2xml: create an IRIDA workflow from a Galaxy workflow file. <https://github.com/phac-nml/irida-wf-ga2xml>.
  18. Reference deleted.
  19. Docker. 2021. Overview of Docker Compose. Docker, Palo Alto, CA. <https://docs.docker.com/compose/>.
  20. Llarena A-K, Ribeiro-Gonçalves BF, Silva DN, Halkilahti J, Machado MP, Silva MSD, Jaakkonen A, Isidro J, Hämäläinen C, Joenperä J, Borges V, Viera L, Gomes JP, Correia C, Lunden J, Laukkanen-Ninios R, Fredriksson-Ahomaa M, Bikandi J, Millan RS, Martínez-Ballesteros I, Laorden L, Mäesaar M, Grantinalevina L, Hilbert F, Garaizar J, Oleastro M, Nevas M, Salmenlinna S, Hakkinen M, Carriço JA, Rossi M. 2018. INNUENDO: a cross-sectoral platform for the integration of genomics in the surveillance of food-borne pathogens. *EFSA Support Publ* 15:1498E.
  21. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. 2017. Nextflow enables reproducible computational workflows. *Nat Biotechnol* 35:316–319. <https://doi.org/10.1038/nbt.3820>.
  22. Apache Software Foundation. 2021. Apache Tomcat. The Apache Software Foundation, Wilmington, DE. <https://tomcat.apache.org/>.
  23. Tania T, Sudarmono P, Kusumawati RL, Rukmana A, Pratama WA, Regmi SM, Kaewprasert O, Chairprasert A, Chongsuivatwong V, Faksri K. 2020. Whole-genome sequencing analysis of multidrug-resistant *Mycobacterium tuberculosis* from Java, Indonesia. *J Med Microbiol* 69:1013–1019. <https://doi.org/10.1099/jmm.0.001221>.
  24. Thiessen J, Hole D, Kruczkiewicz P, van Heusden P, Matthews T. 2021. IRIDA Uploader: sequence file uploader for IRIDA. <https://github.com/phac-nml/irida-uploader>.
  25. Andrews S. 2019. FastQC: a quality control tool for high throughput sequence data. Babraham Institute, Cambridge, UK. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
  26. Comas I, Chakravarti J, Small PM, Galagan J, Niemann S, Kremer K, Ernst JD, Gagneux S. 2010. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet* 42:498–503. <https://doi.org/10.1038/ng.590>.
  27. Comas I. 2019. Genome of the inferred most recent common ancestor of the *Mycobacterium tuberculosis* complex. <https://zenodo.org/record/3497110>.
  28. Goig GA, Blanco S, Garcia-Basteiro A, Comas I. 2018. Pervasive contaminations in sequencing experiments are a major source of false genetic variability: a meta-analysis. *bioRxiv* <https://doi.org/10.1101/403824>.
  29. Lose T, van Heusden P, Christoffels A. 2020. COMBAT-TB-NeoDB: fostering tuberculosis research through integrative analysis using graph database technologies. *Bioinformatics* 36:982–983.
  30. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol* 20:257. <https://doi.org/10.1186/s13059-019-1891-0>.
  31. IRIDA Project. 2018. Add ability to use the results of an assembly workflow/assembled genome as input to other workflows. Issue 57—phac-nml/irida. <https://github.com/phac-nml/irida/issues/57>.
  32. Xu Y, Cancino-Muñoz I, Torres-Puente M, Villamayor LM, Borrás R, Borrás-Mañez M, Bosque M, Camarena JJ, Colomer-Roig E, Colomina J, Escribano I, Esparcia-Rodríguez O, Gil-Brusola A, Gimeno C, Gimeno-Gascón A, Gomila-Sard B, González-Granda D, Gonzalo-Jiménez N, Guna-Serrano MR, López-Hontangas JL, Martín-González C, Moreno-Muñoz R, Navarro D, Navarro M, Orta N, Pérez E, Prat J, Rodríguez JC, Ruiz-García MM, Vanaelochia H, Colijn C, Comas I. 2019. High-resolution mapping of tuberculosis transmission: whole genome sequencing and phylogenetic modelling of a cohort from Valencia Region, Spain. *PLoS Med* 16:e1002961. <https://doi.org/10.1371/journal.pmed.1002961>.
  33. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv* 13033997 Q-Bio.
  34. Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *ArXiv* 12073907 Q-Bio.
  35. Bush SJ, Foster D, Eyre DW, Clark EL, De Maio N, Shaw LP, Stoesser N, Peto TEA, Crook DW, Walker AS. 2020. Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines. *GigaScience* 9:giaa007. <https://doi.org/10.1093/gigascience/giaa007>.
  36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>.
  37. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms. *Fly (Austin)* 6:80–92. <https://doi.org/10.4161/fly.19695>.
  38. Fishbein S, van Wyk N, Warren RM, Sampson SL. 2015. Phylogeny to function: PE/PPE protein evolution and impact on *Mycobacterium tuberculosis* pathogenicity. *Mol Microbiol* 96:901–916. <https://doi.org/10.1111/mmi.12981>.
  39. Ezewudo M, Borens A, Chiner-Oms Á, Miotto P, Chindelevitch L, Starks AM, Hanna D, Liwski R, Zignol M, Gilpin C, Niemann S, Kohl TA, Warren RM, Crook D, Gagneux S, Hoffner S, Rodrigues C, Comas I, Engelthaler DM, Alland D, Rigouts L, Lange C, Dheda K, Hasan R, McNeerney R, Cirillo DM, Schito M, Rodwell TC, Posey J. 2018. Integrating standardized whole genome sequence analysis with a global *Mycobacterium tuberculosis* antibiotic resistance knowledgebase. *Sci Rep* 8:15382. <https://doi.org/10.1038/s41598-018-33731-1>.

40. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>.
41. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 37:1530–1534. <https://doi.org/10.1093/molbev/msaa015>.
42. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol* 35:518–522. <https://doi.org/10.1093/molbev/msx281>.
43. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 14:587–589. <https://doi.org/10.1038/nmeth.4285>.