Title: The COMBAT-TB Workbench: Making powerful TB bioinformatics accessible

Running title: COMBAT-TB Workbench - *M.tuberculosis* analysis

Peter van Heusden[a*], Ziphozahe Mashologu[a], Thoba Lose[a], Robin Warren[b], Alan Christoffels[a,c*]

[a]South African Medical Research Council Bioinformatics Unit, South African National Bioinformatics Institute, University of the Western Cape, South Africa

[b]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

[c]Africa Centres for Disease Control and Prevention, African Union Headquarters, Addis Ababa W21K19, Ethiopia

[*] Corresponding authors: pvh@sanbi.ac.za, alan@sanbi.ac.za

1

2

Abstract (250 word limit)

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 semi-automated 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 (150 word limit, non-technical explanation)

While the reduction in cost of WGS is making sequencing more affordable in Lower and Middle Income Countries, 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 analysing *M. tuberculosis* WGS data and thereby makes WGS usable in practice in the LMIC context.

3

Introduction

Tuberculosis (TB) 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 Whole Genome Sequencing in low and middle income countries (LMICs) requires reducing the cost of both 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 of 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 customisable 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 analysing *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. These components are combined into a single easy to install application using Docker container technology.

Data management & 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 results visualisation. 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 analysed over 200,000 pathogen samples using IRIDA (44)) and the fact that it is open source.

5

IRIDA stores sequence samples on disk and sample metadata in a MariaDB database (9). Sequence data is 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 (10), snippy (11) and snp-sites (12)) and writing new Galaxy tools (snp-dists (13), TB-Profiler (5), tb_variant_filter (14) and TB Variant Report (15)). 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 (16) 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 COMBAT-TB Workbench deployment process.

Deployment

The COMBAT-TB Workbench application (available at: 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 (17). The Docker containers are orchestrated using docker-compose (18). 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.

7

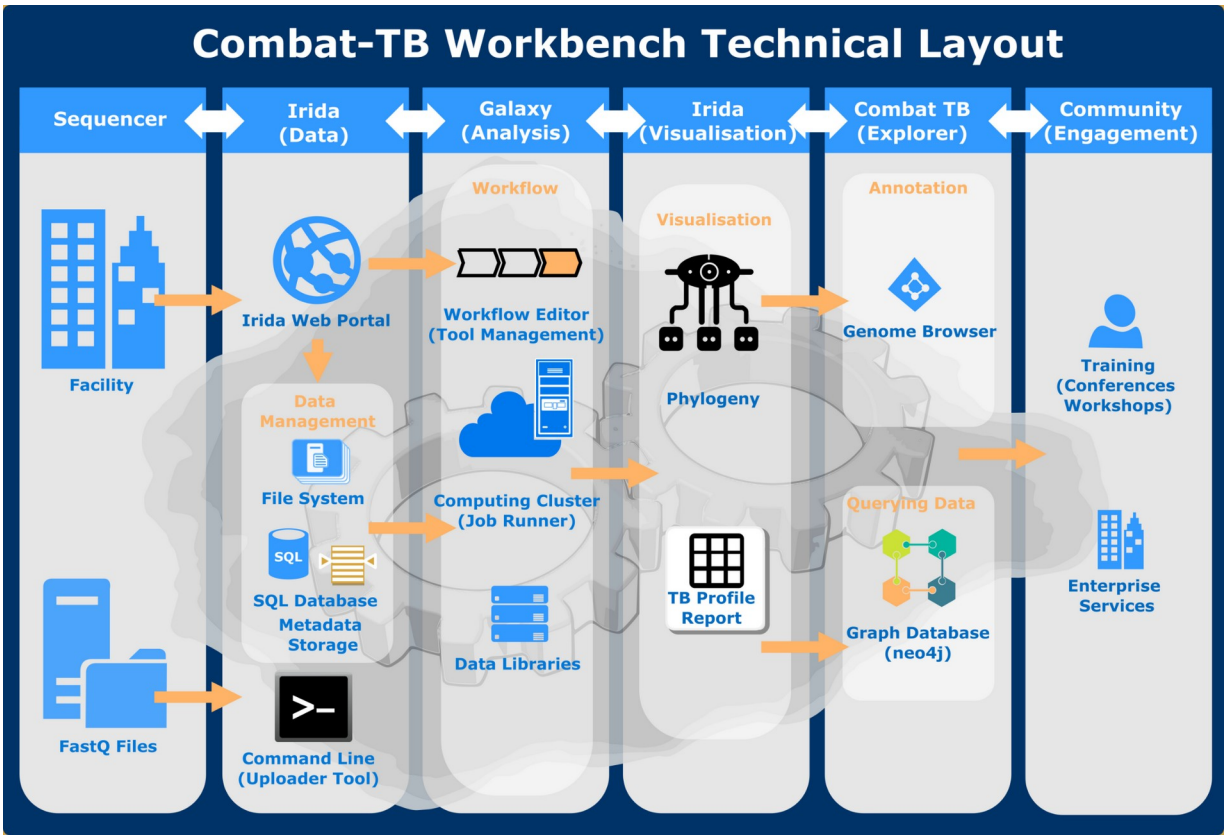


Figure 1. COMBAT-TB Workbench Technical Layout

Comparison to other similar open source software

At the time of writing we found two systems, Innuendo (19) and the IRIDA project (8) that were comparable to the COMBAT-TB Workbench.

Innuendo is a web interface to sequence storage and Nextflow (20) workflow execution aimed at analysis of food-borne pathogens. It is oriented around common tasks in the food-borne pathogens surveillance terrain such as molecular typing of pathogens. The platform is strongly tied to the workflows of a food-borne pathogen surveillance lab, and adding additional species to the analysis system requires modifying the underlying

8

124 database and adding a wgMLST scheme. As such, Innuendo is addressing a different
125 challenge to the one the COMBAT-TB project is tackling.
126
127 IRIDA is the official bioinformatics platform for public health genomics within the Public
128 Health Agency of Canada. It has been in use since 2016 by Canada's provincial and
129 national public health laboratories for genomic investigations of foodborne disease
130 outbreaks, as part of PulseNet Canada's foodborne disease surveillance activities.
131 While it is more flexible than Innuendo, as it allows deployment of a wide variety of
132 Galaxy workflows and is not species specific, it is complex to deploy, with the
133 installation guide assuming knowledge of deployment of both Galaxy and the Tomcat
134 Java Servlet system (21). The COMBAT-TB Workbench, by contrast, is straightforward
135 to deploy with a single Linux command.
136
137

138 Use Case

139 For the purpose of these analyses we installed the COMBAT-TB Workbench on a virtual
140 machine with 8 virtual CPUs, 32 GB RAM and 3000 GB hard disk space, running
141 Ubuntu 18.04 with Docker version 19.03.14 and docker-compose version 1.27.4.
142
143 The COMBAT-TB Workbench user interface is organised around Projects and
144 Analyses. Projects store sequence samples, are associated with a reference genome
145 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 (Figure 2):

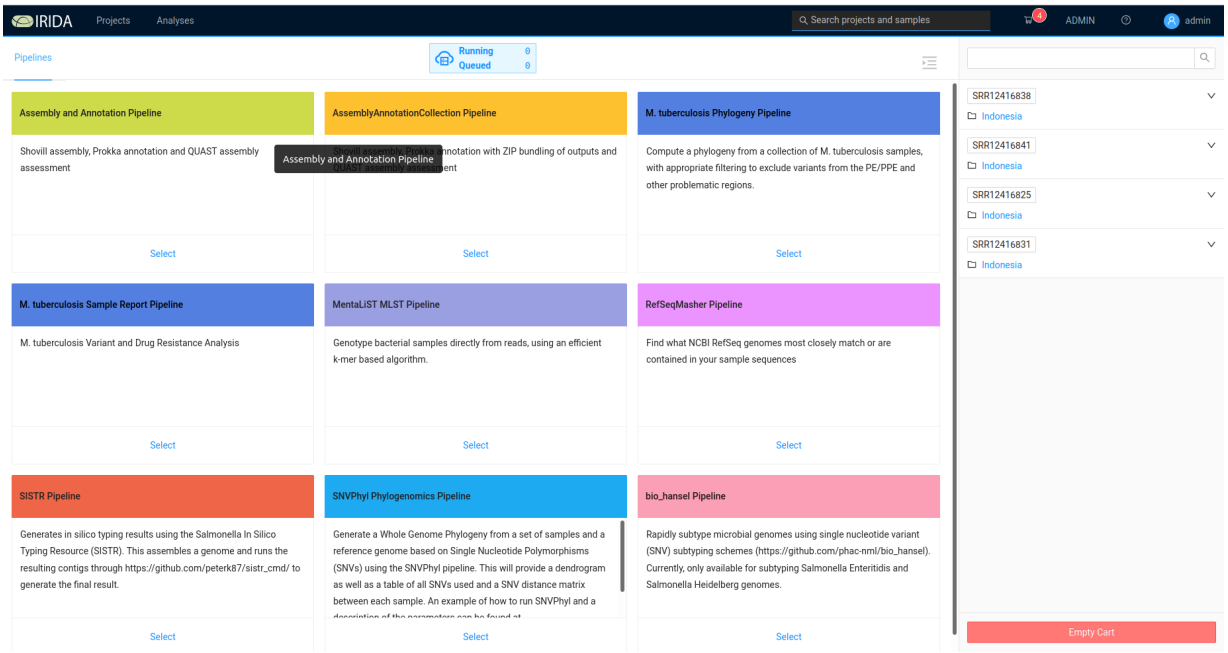


Figure 2. COMBAT-TB Workbench: Workflow selection screen

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

Analysis of data from MDR *M. tuberculosis* in Indonesia

10

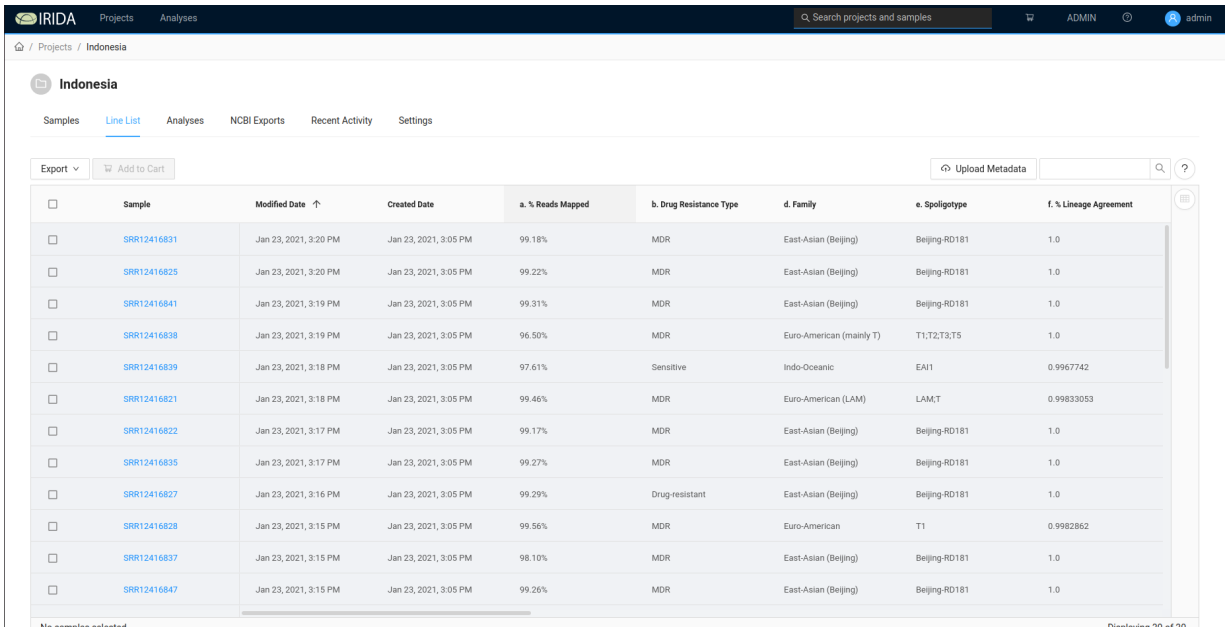
159 Tania et al (22) collected *M. tuberculosis* samples from 30 patients with confirmed
160 pulmonary tuberculosis treated in four hospitals in the Western region of the Indonesian
161 island of Java. Phenotypic drug susceptibility testing (DST) was performed on the
162 cultured samples and DNA was isolated, sequenced and deposited in the NCBI
163 Sequence Read Archive (SRA) with BioProject access number PRJNA633244. Thirty
164 samples were retrieved from EBI European Nucleotide Archive (ENA) and uploaded to
165 the COMBAT-TB Workbench using the irida-uploader (23) command line tool. Per
166 sample quality control was performed automatically (using FastQC(24)) after sample
167 upload. Uploading and quality control took 9 and 3 minutes respectively.
168
169 The inferred ancestral *M. tuberculosis* reference genome produced by Comas et al (25)
170 was downloaded from Zenodo (26) and uploaded to the COMBAT-TB Workbench as
171 the default reference genome for the sequencing project. While the H37Rv reference
172 genome can be used, we used the *M. tuberculosis* inferred ancestral reference as Goig
173 et al (27) have previously shown that this genome is equidistant, in terms of sequence
174 variants, from all known *M. tuberculosis* lineages and thus provides a superior reference
175 to H37Rv (NC_000962.3) for variant calling, especially if that variant calling is going to
176 be used for phylogeny construction.
177

178 *Per sample reporting*

179

180 The *M. tuberculosis* Sample Report pipeline (see Materials and Methods) was run on
181 all 30 samples. Detailed output from this pipeline is made available in a per-sample

182 analysis report (Supplementary Figure 1) that includes a full report on variants identified
183 in the sample (annotated using information from the COMBAT-TB NeoDB (28)), drug
184 resistance prediction (from TB-Profiler) and quality control information on read mapping.
185
186 The pipeline updated the metadata associated with each sample, with the result that
187 after pipeline completion each sample was annotated with percentage of reads mapped
188 to genome, drug resistance prediction, drug resistance associated variants and lineage
189 assigned (Figure 3).
190
191



Sample	Modified Date	Created Date	a. % Reads Mapped	b. Drug Resistance Type	d. Family	e. Spoligotype	f. % Lineage Agreement
SRR12416831	Jan 23, 2021, 3:20 PM	Jan 23, 2021, 3:05 PM	99.18%	MDR	East-Asian (Beijing)	Beijing-RD181	1.0
SRR12416825	Jan 23, 2021, 3:20 PM	Jan 23, 2021, 3:05 PM	99.22%	MDR	East-Asian (Beijing)	Beijing-RD181	1.0
SRR12416841	Jan 23, 2021, 3:19 PM	Jan 23, 2021, 3:05 PM	99.31%	MDR	East-Asian (Beijing)	Beijing-RD181	1.0
SRR12416838	Jan 23, 2021, 3:19 PM	Jan 23, 2021, 3:05 PM	96.50%	MDR	Euro-American (mainly T)	T1;T2;T3;T5	1.0
SRR12416839	Jan 23, 2021, 3:18 PM	Jan 23, 2021, 3:05 PM	97.61%	Sensitive	Indo-Oceanic	EAI1	0.9967742
SRR12416821	Jan 23, 2021, 3:18 PM	Jan 23, 2021, 3:05 PM	99.46%	MDR	Euro-American (LAM)	LAM;T	0.99833053
SRR12416822	Jan 23, 2021, 3:17 PM	Jan 23, 2021, 3:05 PM	99.17%	MDR	East-Asian (Beijing)	Beijing-RD181	1.0
SRR12416835	Jan 23, 2021, 3:17 PM	Jan 23, 2021, 3:05 PM	99.27%	MDR	East-Asian (Beijing)	Beijing-RD181	1.0
SRR12416827	Jan 23, 2021, 3:16 PM	Jan 23, 2021, 3:05 PM	99.29%	Drug-resistant	East-Asian (Beijing)	Beijing-RD181	1.0
SRR12416828	Jan 23, 2021, 3:15 PM	Jan 23, 2021, 3:05 PM	99.56%	MDR	Euro-American	T1	0.9982862
SRR12416837	Jan 23, 2021, 3:15 PM	Jan 23, 2021, 3:05 PM	98.10%	MDR	East-Asian (Beijing)	Beijing-RD181	1.0
SRR12416847	Jan 23, 2021, 3:15 PM	Jan 23, 2021, 3:05 PM	99.26%	MDR	East-Asian (Beijing)	Beijing-RD181	1.0

Figure 3: Sample Line List

192
193 Examining the read mapping outputs shows that only 9.63% and 1.47% of reads from
194 samples SRR12416824 and SRR12416842 respectively mapped against the *M.*

195 *tuberculosis* genome. Further investigation with kraken2 (29) using the Standard
196 database from 14 April 2020 showed that the majority (63.55%) of reads from
197 SRR12416824 were classified as belonging to the *Mycobacterium avium* complex
198 (MAC) and the majority of reads (75.39%) from SRR12416842 were classified as the
199 non-tubercular mycobacterium *Mycolicibacterium fortuitum*. These samples were thus
200 excluded and not used in subsequent analyses.
201
202 As shown in Table 1, the results were broadly concordant with those found by Tania et
203 al. Some small differences are likely a result of different versions of the TB-Profiler
204 software used. Without our own analysis, updating the TB-Profiler version from 2.8.4 to
205 version 3.0.6 reduced the discordance between streptomycin resistance predicted by
206 the COMBAT-TB Workbench and that predicted by MGIT by 2 samples. This illustrates
207 the evolution of drug resistance prediction from WGS data. As noted previously, the
208 COMBAT-TB Workbench pipeline uses snippy for mapping and variant calling and
209 applies TB-Profiler only to the results of the mapping step. Our results are, however,
210 wholly concordant with running the complete TB-Profiler pipeline.
211

Drug	**Concordant MGIT**	**Discordant MGIT Resistant**	**Discordant MGIT Susceptible**	**Concordant Tania et al TB-Profiler**	**Discordant Tania et al Resistant**
Isoniazid	24/28 (85.71%)	2	2	23/28 (82.14%)	1

Rifampicin	27/28 (96.43%)	0	1	27/28 (96.43%)	0
Ethambutol	17/28 (60.71%)	0	11	18/28 (64.29%)	1
Streptomycin	25/28 (85.71%)	2	1	23/28 (82.14%)	2
Table 1: Drug Resistance predictions from COMBAT TB Workbench & Tania et al					

212

213 The lower degree of concordance found between WGS and MGIT resistance testing for
214 ethambutol can be explained by the presence of variants impacting amino acid 306 of
215 the embB gene in 7 out of the 11 samples that were reported as resistant by WGS but
216 susceptible by MGIT. These variants (M306V in 2 and M306I in 5) lead to a change in
217 the MIC of ethambutol (30).

218

219 *Phylogeny on all samples*

220

221 The 28 samples that previously passed quality control were selected in the web
222 interface and submitted to the *M. tuberculosis* phylogeny pipeline
223 (https://github.com/COMBAT-TB/irida-plugin-tb-phylogeny). This pipeline computes a

Maximum Likelihood Phylogeny using the single nucleotide variants identified in each sample (Figure 4).

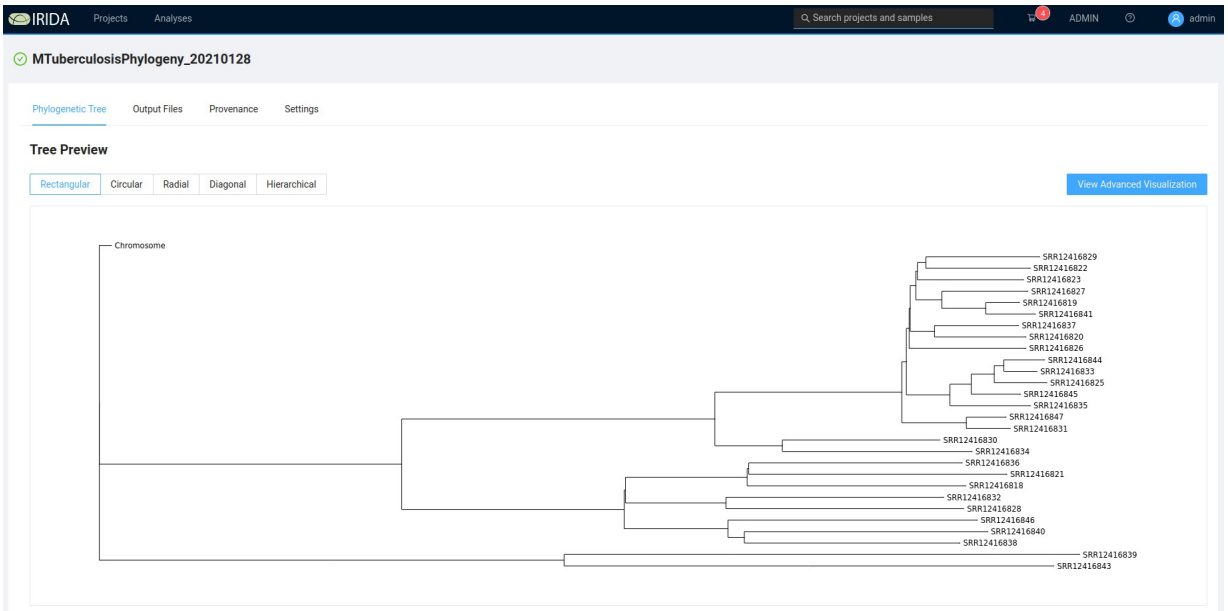


Figure 4: COMBAT TB Workbench phylogeny viewer

The IRIDA Advanced Visualization view (Figure 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 Tania et al. paper. This spreadsheet was used to load additional metadata into the IRIDA project.

The distance between hospital sites varied from 10km (Mampang Prapatan Hospital, Jakarta to Cempaka Putih Islamic Hospital, Jakarta) to 107km (Drajat Prawiranegara

15

238 Hospital, Serang to M. Goenawan Partowidigdo Pulmonary Hospital, Bogor). When
239 visualised this way it was apparently that there was no clear relationship between
240 phylogenetic relationship (and thus lineage) and hospital site, illustrating that the
241 outbreaks occurring in the region were circulating in areas broader than hospital
242 catchment areas.
243

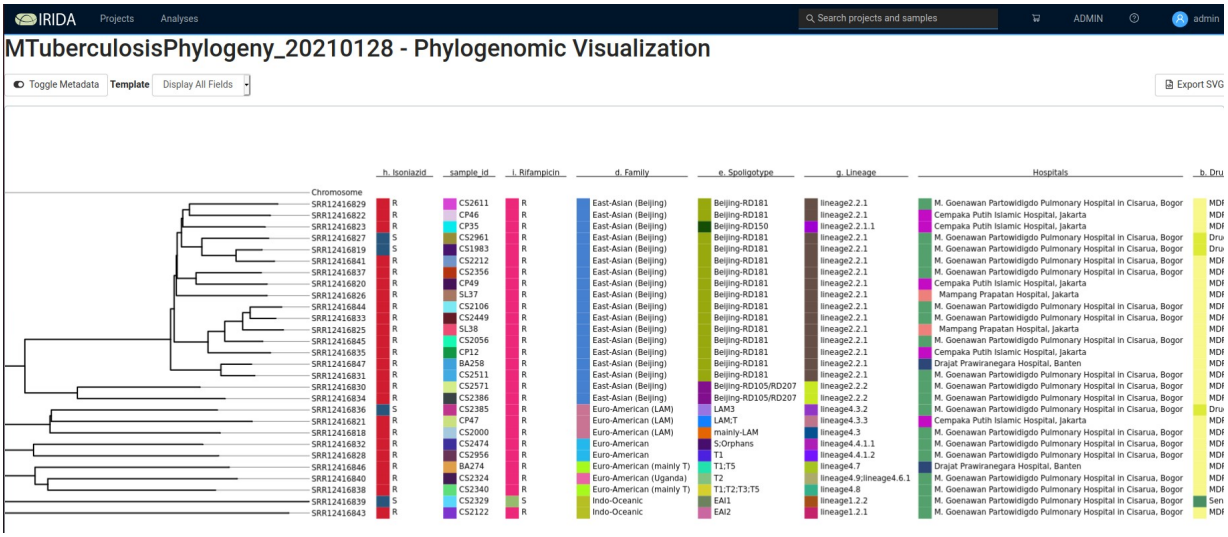


Figure 5: Advanced Phylogeny Visualisation with Metadata columns
244

245 *Analysis of runtimes*

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

Dataset	Processing Step	Running Time
Tania et al (30 samples, 25GB)	Upload	9 minutes
	Sample processing	3 hours 23 minutes
	Phylogeny	3 hours 43 minutes
Xu et al (117 samples, 42GB)	Upload	21 minutes
	Sample processing	6 hours
	Phylogeny	8 hours 4 minutes

Table 2: Runtime of analyses in COMBAT TB Workbench

Analysis of data from *M. tuberculosis* in Spain

Xu et al (32) collected *M. tuberculosis* samples from 117 patients and analysed 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 dataset to examine the performance of the COMBAT-TB Workbench

Uploading of the 117 samples took 21 minutes. Per sample analysis of the 117 samples took a total of 6 hours. A phylogeny was computed using the samples. Phylogeny generation took just over 8 hours and visualisation with the advanced phylogeny viewer revealed clustering by *M. tuberculosis* lineage (Figure 7) as expected. Unfortunately Xu

17

et al did not include the output of their phylogeny construction in their results, so a direct comparison of the computed phylogenetic trees is not possible.

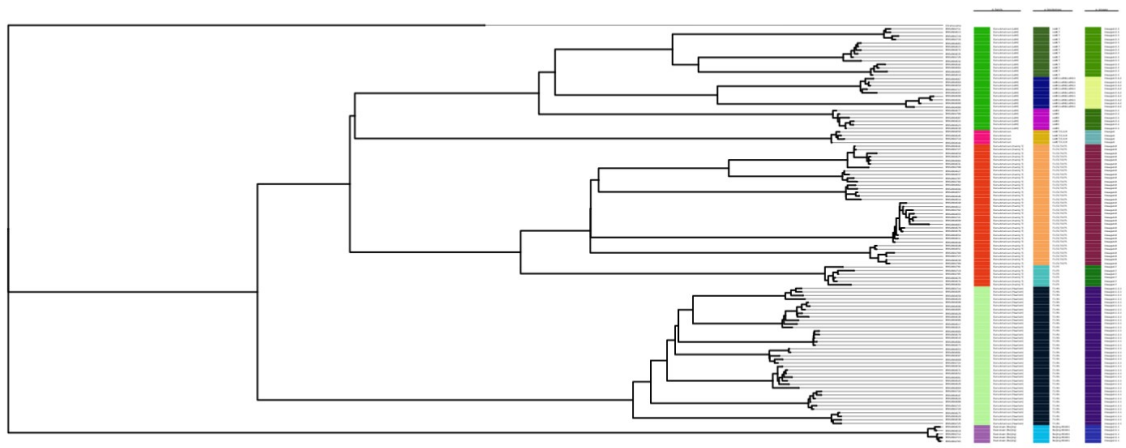


Figure 7: Advanced Phylogeny viewer with Xu et al data (metadata shows lineage)

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 systems administration skills.

18

By making this project, developed using open source best practices, available the cost of use of WGS can be reduced, 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.

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 four base window drops below quality score 30, followed by the Minimum length trimmer that

299 discards reads shorter than 20 bases. Only reads which remain in read pairs after
300 quality trimming are used in subsequent analysis.
301
302 After quality trimming, sequence reads are mapped to a user supplied reference
303 genome. The workflow requires a genome with the same coordinate scheme as the *M.*
304 *tuberculosis* H37Rv reference (RefSeq NC000962). Mapping and variant calling is done
305 using snippy (v. 4.4.5), a Perl-based pipeline that combines the bwa-mem (v. 0.7.17)
306 (33) mapper and the freebayes (v. 1.3.2) (34) variant caller. Snippy has been shown
307 (35) to produce good quality variant calls in *M. tuberculosis* when run with default
308 parameters. Samtools (v. 1.9) (36) flagstat is run to provide an overview of statistics
309 from the mapping process.
310
311 After variant calling, variants are annotated with SnpEff (v. 4.3) (37) using the H37Rv
312 reference genome annotation. Variants are then filtered using tb_variant_filter (v. 0.1.3).
313 While snippy performs quality based filtering of the variants it predicts, this tool offers a
314 variety of filtering options commonly used in *M. tuberculosis* variant filtering. In our
315 workflow it filters out variants in the PE/PPE gene regions (38), in the repetitive and
316 insertion section regions identified by UVP (39) and those with lower than 30 supporting
317 reads or within 5 base pairs of an indel. These filtering options like all tool options in the
318 workflow can optionally be altered by the user.
319

320 In parallel to the SnpEff annotation and variant filtering steps, the mapped reads are
321 provided to TBProfiler (v. 2.8.4) which performs its own variant calling and lineage and
322 drug resistance prediction.
323
324 Finally the filtered, annotated variants and the TBProfiler results are fed to tb_vcf_report
325 (v. 0.1.7) which produces a report further annotated with information from the COMBAT-
326 TB eXplorer database (28) in both text and HTML formats.
327
328 The final reports provided to the user are the variant reports from tb_vcf_report, text and
329 JSON format reports from TBProfiler, variants in VCF format from SnpEff and mapping
330 statistics from samtools flagstats. These reports include both user readable and raw
331 data suitable for further downstream analysis. The metadata stored in the sample line
332 list is updated with mapping %, *M. tuberculosis* lineage and spoligotype information and
333 drug resistance information.

334 Implementation of the *M. tuberculosis* Phylogeny pipeline

335
336 The workflow in the phylogeny module starts with quality filtering of samples using fastp
337 (v. 0.19.5) (40) with default settings. The filtered reads are then aligned to the user
338 provided reference using snippy and predicted variants are filtered with tb_variant_filter
339 as described above. In addition, only single nucleotide variants (SNVs) are retained as
340 the phylogeny software used in the workflow cannot extract meaningful information from
341 indels.
342

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 visualised 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.

Funding

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 the South African Medical Research Council flagship programme MRC-RFA-UFSP-01-2013/ COMBAT-TB.

References

1. World Health Organization. 2020. Global tuberculosis report 2020. World Health

364 Organization.
365 2. Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, Farhat MR,
366 Guthrie JL, Laukens K, Miotto P, Ofori-Anyinam B, Dreyer V, Supply P, Suresh A, Utpatel
367 C, van Soolingen D, Zhou Y, Ashton PM, Brites D, Cabibbe AM, de Jong BC, de Vos M,
368 Menardo F, Gagneux S, Gao Q, Heupink TH, Liu Q, Loiseau C, Rigouts L, Rodwell TC,
369 Tagliani E, Walker TM, Warren RM, Zhao Y, Zignol M, Schito M, Gardy J, Cirillo DM,
370 Niemann S, Comas I, Van Rie A. 2019. Whole genome sequencing of Mycobacterium
371 tuberculosis: current standards and open issues. Nat Rev Microbiol 17:533–545.
372 3. Mahomed S, Naidoo K, Dookie N, Padayatchi N. 2017. Whole genome sequencing for the
373 management of drug-resistant TB in low income high TB burden settings: Challenges and
374 implications. Tuberculosis 107:137–143.
375 4. Joppich M, Zimmer R. 2019. From command-line bioinformatics to bioGUI. PeerJ 7:e8111.
376 5. Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, O'Grady J,
377 McNerney R, Hibberd ML, Viveiros M, Huggett JF, Clark TG. 2019. Integrating informatics
378 tools and portable sequencing technology for rapid detection of resistance to anti-
379 tuberculous drugs. Genome Med 11:41.
380 6. Rosenthal A, Gabrielian A, Engle E, Hurt DE, Alexandru S, Crudu V, Sergueev E,
381 Kirichenko V, Lapitskii V, Snezhko E, Kovalev V, Astrovko A, Skrahina A, Taaffe J, Harris
382 M, Long A, Wollenberg K, Akhundova I, Ismayilova S, Skrahin A, Mammadbayov E,
383 Gadirova H, Abuzarov R, Seyfaddinova M, Avaliani Z, Strambu I, Zaharia D, Muntean A,
384 Ghita E, Bogdan M, Mindru R, Spinu V, Sora A, Ene C, Vashakidze S, Shubladze N,
385 Nanava U, Tuzikov A, Tartakovsky M. 2017. The TB Portals: an Open-Access, Web-Based
386 Platform for Global Drug-Resistant-Tuberculosis Data Sharing and Analysis. J Clin
387 Microbiol 55:3267–3282.
388 7. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, Chilton J, Clements D,
389 Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H,

390 Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D. 2018. The Galaxy platform
391 for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic
392 Acids Res 46:W537–W544.
393 8. Matthews TC, Bristow FR, Griffiths EJ, Petkau A, Adam J, Dooley D, Kruczkiewicz P,
394 Curatcha J, Cabral J, Fornika D, Winsor GL, Courtot M, Bertelli C, Roudgar A, Feijao P,
395 Mabon P, Enns E, Thiessen J, Keddy A, Isaac-Renton J, Gardy JL, Tang P, Consortium
396 TI, Carrico JA, Chindelevitch L, Chauve C, Graham MR, McArthur AG, Taboada EN, Beiko
397 RG, Brinkman FS, Hsiao WW, Domselaar GV. 2018. The Integrated Rapid Infectious
398 Disease Analysis (IRIDA) Platform. bioRxiv 381830.
399 9. MariaDB Foundation. 2021. MariaDB.org.
400 10. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina
401 sequence data. Bioinformatics 30:2114–2120.
402 11. Seemann T. 2020. snippy: Rapid haploid variant calling and core genome alignment. Perl.
403 12. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SRY 2016. SNP-
404 sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genomics
405 2:e000056.
406 13. Seemann T. 2021. snp-dists: Convert a FASTA alignment to SNP distance matrix. C.
407 14. Peter van Heusden. 2021. tb_variant_filter: A tool for filtering VCF files (relative to M.
408 tuberculosis H37Rv). Python.
409 15. Thoba Lose, Peter van Heusden. 2021. tbvcfreport: Generate an interactive HTML-based
410 report from M. tuberculosis SnpEff annotated VCF(s). Python.
411 16. Jeffrey Thiessen, Dan Fornika, Peter Kruczkiewicz, Aaron Petkau, Peter van Heusden.
412 2021. irida-wf-ga2xml: Create an IRIDA workflow from a Galaxy workflow file. Clojure,
413 National Microbiology Laboratory.
414 17. Docker. Docker Web Page.
415 18. Docker. 2021. Overview of Docker Compose. Docker Doc.

416 19. Llarena A-K, Ribeiro-Gonçalves BF, Silva DN, Halkilahti J, Machado MP, Silva MSD,
417 Jaakkonen A, Isidro J, Hämäläinen C, Joenperä J, Borges V, Viera L, Gomes JP, Correia
418 C, Lunden J, Laukkanen-Ninios R, Fredriksson-Ahomaa M, Bikandi J, Millan RS, Martinez-
419 Ballesteros I, Laorden L, Mäesaar M, Grantina-Ievina L, Hilbert F, Garaizar J, Oleastro M,
420 Nevas M, Salmenlinna S, Hakkinen M, Carriço JA, Rossi M. 2018. INNUENDO: A cross-
421 sectoral platform for the integration of genomics in the surveillance of food-borne
422 pathogens. EFSA Support Publ 15:1498E.
423 20. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. 2017.
424 Nextflow enables reproducible computational workflows. Nat Biotechnol 35:316–319.
425 21. Apache Software Foundation. Apache Tomcat®.
426 22. Tania T, Sudarmono P, Kusumawati RL, Rukmana A, Pratama WA, Regmi SM,
427 Kaewprasert O, Chaiprasert A, Chongsuvivatwong V, Faksri K 2020. Whole-genome
428 sequencing analysis of multidrug-resistant Mycobacterium tuberculosis from Java,
429 Indonesia. J Med Microbiol 69:1013–1019.
430 23. Jeffrey Thiessen, Darian Hole, Peter Kruczkiewicz, Peter van Heusden, Tom Matthews.
431 2021. IRIDA Uploader: Sequence file uploader for IRIDA. Python, National Microbiology
432 Laboratory.
433 24. Simon Andrews. 2019. FastQC A Quality Control tool for High Throughput Sequence Data.
434 Babraham Bioinforma.
435 25. Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, Ernst JD, Gagneux
436 S. 2010. Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily
437 hyperconserved. Nat Genet 42:498–503.
438 26. Comas I. 2019. Genome of the inferred most recent common ancestor of the
439 Mycobacterium tuberculosis complex. Zenodo.
440 27. Goig GA, Blanco S, Garcia-Basteiro A, Comas I. 2018. Pervasive contaminations in
441 sequencing experiments are a major source of false genetic variability: a

442 meta-analysis https://doi.org/10.1101/403824.
443 28. Lose T, van Heusden P, Christoffels A. 2020. COMBAT-TB-NeoDB: fostering tuberculosis
444 research through integrative analysis using graph database technologies. Bioinforma Oxf
445 Engl 36:982–983.
446 29. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2.
447 Genome Biol 20:257.
448 30. Starks AM, Gumusboga A, Plikaytis BB, Shinnick TM, Posey JE. 2009. Mutations at embB
449 codon 306 are an important molecular indicator of ethambutol resistance in Mycobacterium
450 tuberculosis. Antimicrob Agents Chemother 53:1061–1066.
451 31. IRIDA Project. 2018. Add ability to use the results of an assembly workflow/assembled
452 genome as input to other workflows · Issue #57 · phac-nml/irida. GitHub.
453 32. Xu Y, Cancino-Muñoz I, Torres-Puente M, Villamayor LM, Borrás R, Borrás-Máñez M,
454 Bosque M, Camarena JJ, Colomer-Roig E, Colomina J, Escribano I, Esparcia-Rodríguez
455 O, Gil-Brusola A, Gimeno C, Gimeno-Gascón A, Gomila-Sard B, González-Granda D,
456 Gonzalo-Jiménez N, Guna-Serrano MR, López-Hontangas JL, Martín-González C,
457 Moreno-Muñoz R, Navarro D, Navarro M, Orta N, Pérez E, Prat J, Rodríguez JC, Ruiz-
458 García MM, Vanaclocha H, Colijn C, Comas I. 2019. High-resolution mapping of
459 tuberculosis transmission: Whole genome sequencing and phylogenetic modelling of a
460 cohort from Valencia Region, Spain. PLOS Med 16:e1002961.
461 33. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-
462 MEM. ArXiv13033997 Q-Bio.
463 34. Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing.
464 ArXiv12073907 Q-Bio.
465 35. Bush SJ, Foster D, Eyre DW, Clark EL, De Maio N, Shaw LP, Stoesser N, Peto TEA,
466 Crook DW, Walker AS. 2020. Genomic diversity affects the accuracy of bacterial single-
467 nucleotide polymorphism–calling pipelines. GigaScience 9.

468 36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin
469 R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map
470 format and SAMtools. Bioinformatics 25:2078–2079.
471 37. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM.
472 2012. A program for annotating and predicting the effects of single nucleotide
473 polymorphisms, SnpEff. Fly (Austin) 6:80–92.
474 38. Fishbein S, Wyk N van, Warren RM, Sampson SL. 2015. Phylogeny to function: PE/PPE
475 protein evolution and impact on Mycobacterium tuberculosis pathogenicity. Mol Microbiol
476 96:901–916.
477 39. Ezewudo M, Borens A, Chiner-Oms Á, Miotto P, Chindelevitch L, Starks AM, Hanna D,
478 Liwski R, Zignol M, Gilpin C, Niemann S, Kohl TA, Warren RM, Crook D, Gagneux S,
479 Hoffner S, Rodrigues C, Comas I, Engelthaler DM, Alland D, Rigouts L, Lange C, Dheda K,
480 Hasan R, McNerney R, Cirillo DM, Schito M, Rodwell TC, Posey J. 2018. Integrating
481 standardized whole genome sequence analysis with a global Mycobacterium tuberculosis
482 antibiotic resistance knowledgebase. Sci Rep 8:15382.
483 40. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor.
484 Bioinformatics 34:i884–i890.
485 41. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A,
486 Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic
487 Inference in the Genomic Era. Mol Biol Evol 37:1530–1534.
488 42. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving
489 the Ultrafast Bootstrap Approximation. Mol Biol Evol 35:518–522.
490 43. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017.
491 ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods
492 14:587–589.
493 44. Thomas Matthews. 2021. Number of pathogen samples processed by PHAC-NML using

494 IRIDA.

495

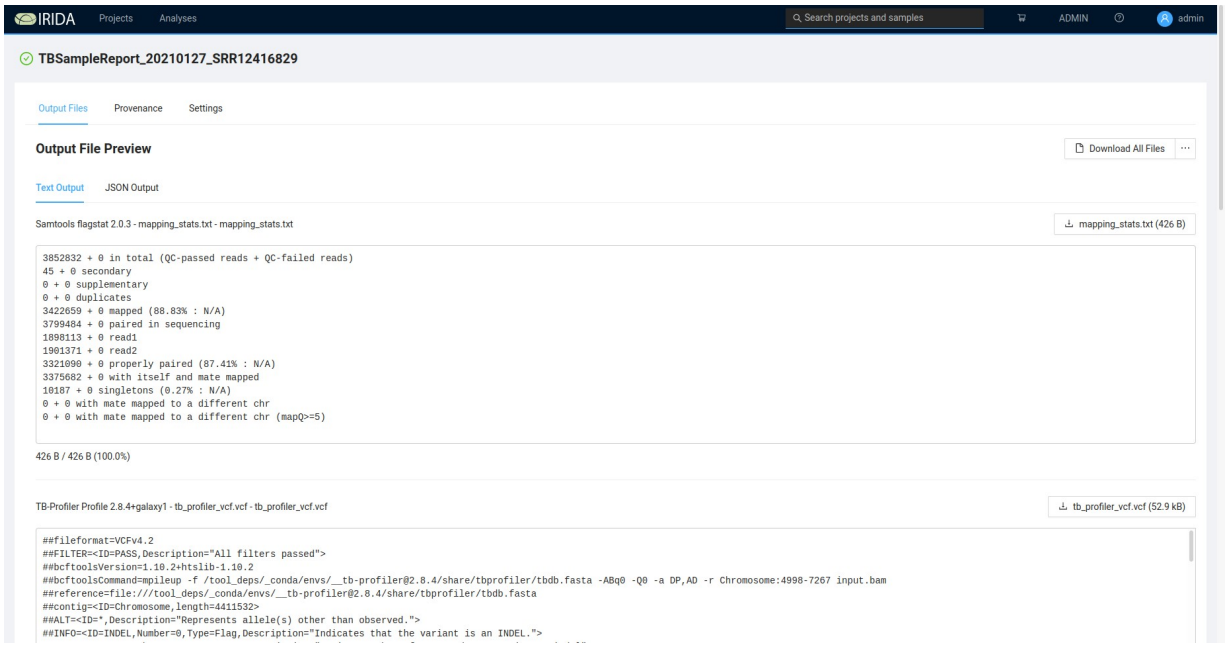
496

497

498

28

Supplemental Figures



29