International Journal of Medical Microbiology 310 (2020) 151399

Contents lists available at ScienceDirect

International Journal of Medical Microbiology

journal homepage: www.elsevier.com/locate/ijmm

Resistance Sniffer: An online tool for prediction of drug resistance patterns of *Mycobacterium tuberculosis* isolates using next generation sequencing data

Dillon Muzondiwa[a], Awelani Mutshembele[b], Rian E. Pierneef[a,c], Oleg N. Reva[a,*]

[a] Centre for Bioinformatics and Computational Biology, Dep. of Biochemistry, Genetics and Microbiology, University of Pretoria, South Africa
[b] South African Medical Research Council (SAMRC), TB Platform, Pretoria, South Africa
[c] Biotechnology Platform, Agricultural Research Council, Onderstepoort, South Africa

ARTICLE INFO

Keywords:
Mycobacterium tuberculosis
Antibiotic
Resistance
Whole-genome sequencing
Clade specific
Single nucleotide polymorphism

ABSTRACT

The effective control of multidrug resistant tuberculosis (MDR-TB) relies upon the timely diagnosis and correct treatment of all tuberculosis cases. Whole genome sequencing (WGS) has great potential as a method for the rapid diagnosis of drug resistant *Mycobacterium tuberculosis* (Mtb) isolates. This method overcomes most of the problems that are associated with current phenotypic drug susceptibility testing. However, the application of WGS in the clinical setting has been deterred by data complexities and skill requirements for implementing the technologies as well as clinical interpretation of the next generation sequencing (NGS) data. The proposed diagnostic application was drawn upon recent discoveries of patterns of Mtb clade-specific genetic polymorphisms associated with antibiotic resistance. A catalogue of genetic determinants of resistance to thirteen anti-TB drugs for each phylogenetic clade was created. A computational algorithm for the identification of states of diagnostic polymorphisms was implemented as an online software tool, Resistance Sniffer (http://resistance-sniffer.bi.up.ac.za/), and as a stand-alone software tool to predict drug resistance in Mtb isolates using complete or partial genome datasets in different file formats including raw Illumina *fastq* read files. The program was validated on sequenced Mtb isolates with data on antibiotic resistance trials available from GMTV database and from the TB Platform of South African Medical Research Council (SAMRC), Pretoria. The program proved to be suitable for probabilistic prediction of drug resistance profiles of individual strains and large sequence data sets.

1. Introduction

Tuberculosis (TB) caused by *Mycobacterium tuberculosis* (Mtb) remains the leading global infectious disease killer. Mtb is a slow-growing, Gram-positive, acid-fast bacterium. The rod shaped, intracellular aerobe has a lipid-rich cell wall which gives Mtb some of its unique characteristics, such as a high resistance to desiccation. In 2017, there were an estimated 10 million new cases of TB and an estimated 1.6 million deaths attributed to the disease (World Health Organization, 2018a). Despite the drop in global TB incidence and mortality rates in recent years, a lot of work still needs to be done if we are to attain the 2030 targets of the End TB Strategy: to reduce TB deaths by 90 % and TB incidence by 80 % (World Health Organization, 2018b). The emergence of drug resistant TB (DR-TB) remains a major challenge in the war against TB. According to the World Health Organization, over 558,000 people had developed TB that was resistant to rifampicin (RR-TB), the most potent of the first line drugs and 82 % of these were classified as multi-drug resistant TB (MDR-TB) (World Health Organization, 2018a).

MDR-TB refers to TB that has developed resistance to the two most powerful first line drugs, rifampicin and isoniazid. Extensive drug resistance TB (XDR-TB) refers to MDR-TB strains that have developed further resistance to any of the second line, injectable drugs. With 8.5 % of the MDR-TB cases classified as XDR-TB in 2017, there is a need to

Abbreviations: AMK, amikacin; CM, capreomycin; CS, cycloserin; DR-TB, drug resistant tuberculosis; DST, drug susceptibility testing; EMB, ethambutol; ETH, ethionamide; FLQ, fluoroquinolones; FN, false negative; FP, false positive; GMTV, Genome Variation Mycobacterium Tuberculosis Variation database; GWAS, genome wide association studies; INH, isoniazid; KAN, kanamycin; MDR-TB, multidrug resistant tuberculosis; Mtb, *Mycobacterium tuberculosis*; NGS, next generation sequencing; NPV, Negative Predictive Value; OFL, ofloxacin; PAS, para-amino salicyclic acid; PATRIC, Pathosystems Resource Integration Center; PPV, Positive Predictive Value; PZA, pyrazinamide; RIF, rifampicin; RR, TB -rifampicin resistant tuberculosis; SAMRC, South African Medical Research Council; SENS, sensitivity; SM, streptomycin; SPEC, specificity; TB, tuberculosis; TDR-TB, totally drug resistant tuberculosis; TN, true negative; TP, true positive; VCF, variant call format; XDR-TB, extensive drug resistant tuberculosis

* Corresponding author.
E-mail address: oleg.reva@up.ac.za (O.N. Reva).

https://doi.org/10.1016/j.ijmm.2020.151399
Received 11 July 2019; Received in revised form 13 November 2019; Accepted 29 December 2019

prioritize the development of tools for the rapid and accurate diagnosis of DR-TB isolates. Gandhi et al. (2006) reported that rapid progression to death was recorded in 98 % of XDR-TB patients during an outbreak in KwaZulu-Natal, South Africa. Other experts have also coined the term totally drug resistant TB (TDR-TB) to describe strains that are resistant to all the currently available drugs although there is no agreed definition of TDR-TB yet (Coll et al., 2015). Treatment of drug resistant TB is challenging compared to treatment of drug susceptible TB with global success rates of less than 50 % for MDR-TB. The treatment process is costly and is often associated with poor outcomes. The drugs used are also highly toxic and can lead to severe side effects, such as permanent deafness and psychiatric disorders (Yang et al., 2017). These challenges can lead to poor compliance with the treatment regime and this in turn reduces the cure rates and can even lead to the amplification of resistance (Shean et al., 2013; Coll et al., 2015). Accurate drug susceptibility testing (DST) profiles are also crucial in the improvement of treatment outcomes as they ensure that only the effective anti-TB drugs are prescribed and reduces exposure to ineffective and toxic drugs (Coll et al., 2015).

Early diagnosis and correct treatment is the key to the control of MDR-TB and incorrect treatment of TB can be catastrophic both at patient and population level. Misdiagnosis and inadequate treatment of MDR-TB or XDR-TB can lead to the positive selection of resistant subpopulations and hence the creation of novel resistant strains. This can also lead to an increase in the number of ineffective drugs against an already M/XDR-TB strain (Walker et al., 2017). Conventional phenotypic DST is still culture based and an arduous process due to the slow growth rate of Mtb. The method also requires expensive infrastructure (biosafety level 3 laboratories), which is not affordable in most low- and middle-income countries which carries the highest burden of TB. This means that DST results are only available after weeks to months which are often too late for the TB patient. For complex drugs such as ethambutol and pyrazinamide, phenotypic DST is frequently inaccurate and usually lacks reproducibility (Demers et al., 2016; World Health Organization, 2018b). Several rapid molecular assays have been developed for the diagnosis of DR-TB. The global roll out of the WHO endorsed Cepheid Xpert MTB/RIF and Ultra assays (Cepheid, Sunnyvale, CA, USA) increased the number of detected DR-TB cases (Theron et al., 2013). Despite this success, these technologies are still limited in the number of loci they can examine and the number of drugs that can be tested (Coll et al., 2015). Major diagnostic gaps still remain to be covered. It was estimated that 558,000 people developed MDR/RR-TB in 2017 of which only 160,684 cases were detected and notified. Of these, only 25 % of the patients were put on a treatment regimen that includes a second line drug (World Health Organization, 2018a). These challenges make WGS more important in the diagnostics of MDR-TB than any other infectious disease (Walker et al., 2015). Unlike in most other bacterial pathogens, resistance plasmids and horizontal gene transfer play no role in the acquisition of drug resistance in Mtb (Trauner et al., 2014). The main cause of drug resistance in Mtb is the accumulation of point mutations and indels in genes coding for drug targets or converting enzymes (Coll et al., 2015). Mechanisms of action and genetic mutations associated with drug resistance to the thirteen TB antibiotics considered in this study are listed in Table 1.

A common approach used in the identification of antibiotic resistance mutations is genome-wide association study (GWAS). A comprehensive GWAS-based analysis of mutations in 6465 clinical Mtb isolates showing various patterns of drug resistance has recently been published (Coll et al., 2018). This approach allows for the identification of novel mutations but generally ignores the evolution of drug resistance consisting of a series of mutations which may vary in different Mtb lineages. Lineage-specific development of Mtb antibiotic resistance was demonstrated in several recent publications (van Niekerk et al., 2018; Oppong et al., 2019). Multiple mutations in many genes create a complex network of epistatic interactions which play an important role in the development of the drug resistant phenotype as they are able to compensate for the fitness cost that is incurred when resistance is acquired (Borrell and Gagneux, 2011). There is a need for more sophisticated WGS-based approaches besides GWAS for the understanding and accurate detection of these processes leading to the resistant phenotype. Unlike the currently available molecular assays, which can only examine limited mutations in specific gene targets, WGS based assays can provide a near complete view of the whole resistome. This means that by using WGS assays, we have the potential to detect resistance for all available anti-TB drugs in contrast to current molecular diagnostics methods, which are limited to only a few drugs (Chen et al., 2019). WGS has the ability to detect rare mutations as well as indels that may not be detected by other molecular assays (World Health Organization, 2018b).

Several NGS platforms which can be applied in the clinical setting have been commercially developed over the years. These sequencing platforms include Illumina technologies, Ion Torrent, PacBio SMRT RSII and Sequel technologies; and Oxford Nanopore MinION. It is important to note that despite the differences associated with the functionality of these platforms, they have all been proved to be suitable for DR-TB diagnosis provided that their infrastructural and operational prerequisites are met (Phelan et al., 2016). This means that prior to setting up an NGS based service, a clinical laboratory has some factors to consider. These factors include cost, turnaround time, data quality and data output (World Health Organization, 2018b). The current study focused mostly on Illumina and Ion Torrent sequencing as they are currently the most feasible techniques for medicinal centers and public health organizations. Nevertheless, genome scale assemblies generated from other techniques are also suitable for the analysis by the online tool proposed in this paper.

The uptake of WGS based technologies in the clinical setting has been hindered by the complexity of the data generated through NGS and the general lack of bioinformatics skills among clinical microbiologists (Macedo et al., 2018). However, the decrease in the cost of NGS has resulted in a number of automated WGS based rapid DR-TB prediction tools being developed in the past few years. These tools, which include KvarQ (Steiner et al., 2014), TBProfiler (Coll et al., 2015), CASTB (Iwai et al., 2015), Mykrobe Predictor (Bradley et al., 2015) and PhyResSE (Feuerriegel et al., 2015), are freely available online and allows for processing of raw sequencing data in the form of *fastq* files. A few studies that assess the performance of these rapid tools have been done with the most extensive study published by Schleusner et al. (Schleusener et al., 2017). Although these tools are rapid and user friendly, the team cited a number of limitations associated with these platforms. Despite the improved sensitivity and specificity (when compared to DST), the number of DR loci interpreted by these tools are still not enough to fully capture the whole resistance profile of TB. None of these tools are able to interpret low frequency mutations with some of the platforms completely insensitive to indels and variants in promoter regions (Chen et al., 2019). Only two of the online platforms (PhyResSE and KvarQ) allows for batch uploads which can make the process of uploading samples one at a time cumbersome in larger settings such as national referral laboratories (Schleusener et al., 2017). Although these tools were designed to be user-friendly, some of them still require the end user to possess a certain level of bioinformatics skills, for example before using Mykrobe Predictor or KvarQ, the paired-end *fastq* files have to be merged and this might present technical challenges for end users. Another hurdle that was noted when the tools were reviewed was the lack of a standardized way of exporting and storing results. Only PhyResSE could provide comma-separated (csv) report files for all strains processed in the same session while CASTB could not offer any report on the interpreted variants. TBProfiler which yielded the best accuracy results does not offer any export and storage functionality (Schleusener et al., 2017).

In this study, a new computational tool was developed for the rapid antibiotic susceptibility profiling of Mtb isolates using WGS datasets in different file formats and on different stages of genome completion, including raw DNA reads in *fastq* format. The proposed program was

Table 1
Anti-TB antibiotics and possible mechanisms of drug resistance.

Antibiotic	Abbreviation	Mechanism of action	Mutated genes associated with drug resistance
Amikacin, Capreomycin, Kanamycin	AMK, CM, KAN	Inhibits protein synthesis through ribosomal binding.	*rrs, eis, tlyA* (Alangaden et al., 1998)
Cycloserin	CS	Cell wall biosynthesis inhibitor.	*alr, ddlA, cycA* (Chen et al., 2017; Oppong et al., 2019)
Ethambutol	EMB	The drug disrupts several pathways of actively multiplying bacilli, most importantly those that are involved in arabinogalactan biosynthesis in the cell wall. This inhibition of arabinan polymerization helps to facilitate the permeability of other drugs that are used in TB treatment.	*embB, ubiA* (Telenti et al., 1997)
Ethionamide	ETH	Structural analogue of INH (Dookie et al., 2018). The prodrug is activated by the mono-oxygenase enzyme to inhibit the binding of the enoyl-acyl carrier protein reductase and therefore inhibit cell wall synthesis.	*ethA, mshA, ndh, inhA, inhA* promoter (Hicks et al., 2019).
Isoniazid	INH	INH is a prodrug which is activated by the catalase /peroxidase enzyme encoded by the katG gene (Van Niekerk et al., 2018). The activated drug interferes with mycolic acid synthesis.	*katG, inhA, kasA* (Ramaswamy et al., 2003)
Fluoroquinolones	FLQ	Inhibit bacterial replication by blocking DNA gyrase important for the replication pathway.	*gyrA* and *gyrB* (Takiff et al., 1994)
Para-amino salicyclic acid	PAS	A para- amino benzoic acid that inhibits folate synthesis.	*thyA* (Dookie et al., 2018)
Pyrazinamide	PZA	Disrupts the proton motive force which is essential for membrane transport.	*pncA* (Demers et al., 2016)
Rifampicin	RIF	The drug is known to disrupt the elongation of mRNA by binding to the beta subunit of RNA polymerase.	*rpoB* (Koch et al., 2014)
Streptomycin	SM	Inhibits protein synthesis by irreversibly binding to the ribosomal protein S12 and 16S rRNA and therefore interfering with the binding of formyl-methionyl-tRNA to the 30S subunit of the bacterial ribosome.	*rrs* and *rpsL* (Brzostek et al., 2004), *gidB* (Okamoto et al., 2007). *whib7* (Reeves et al., 2013)

trained and validated using publicly available records on Mtb sequencing and antibiotic resistance profiling from the Genome Variation Mycobacterium Tuberculosis Variation (GMTV) database (Chernyaeva et al., 2014). Although the program Resistance Sniffer is currently available for research purposes only, the efficacy of the drug resistance prediction algorithm was demonstrated using clinical Mtb isolates available from the Tuberculosis Platform of the South African Medical Research Council (SAMRC).

2. Methods

2.1. Data sourcing and preparation

Mutation data in the form of 2501 variant call format (VCF) files was downloaded from the GMTV database (https://mtb.dobzhanskycenter.org/cgi-bin/beta/main.py#custom/world). The database consists of data from *Mycobacterium tuberculosis* isolates sourced from different regions of the Russian Federation and worldwide. The database integrates drug resistance profiles, epidemiology, TB clinical outcome, year and place of isolation as well as molecular biology data (Chernyaeva et al., 2014). The metadata includes information on drug resistance trials with respect to the following antibiotics: amikacin (AMK), capreomycin (CM), cycloserin (CS), ethambutol (EMB), ethionamide (ETH), isoniazid (INH), fluoroquinolones (FLQ), kanamycin (KAN), ofloxacin (OFL), para-amino salicyclic acid (PAS), pyrazinamide (PZA), rifampicin (RIF) and streptomycin (SM). The database also provides information on the phylogenetic clade of each sample. The quality of the microbiological, WGS and spoligotyping data is guaranteed by the institutions that provided the data to the database (Chernyaeva et al., 2014). The dataset was further split to create a training dataset of 1300 samples. The validation dataset consisted of 1201 samples whose antibiotic phenotype data was available for all the drugs included in this study. An independent testing dataset of 742 Mtb genome sequences was obtained from the SAMRC. We also obtained from the PATRIC database (Wattam et al., 2013) an additional testing dataset of 77 Mtb strains isolated in Sierra Leone (ENA accession number: PRJEB7727). Strains from this database have been described in previous studies (Homolka et al., 2008; Feuerriegel et al., 2012; Schleusener et al., 2017).

2.2. Construction of the diagnostic key

In total, 90,533 SNPs and indels leading to amino acid substitutions were selected from the GMTV database (https://mtb.dobzhanskycenter.org/cgi-bin/beta/main.py#custom/world) for the analysis of their associations with Mtb clades and antibiotic resistance patterns. Additionally, allelic states and locations of SNPs associated with antibiotic resistance were obtained from the TB Drug Resistance Mutation Database (https://tbdreamdb.ki.se/Info/Default.aspx) (Sandgren et al., 2009).

Discriminative power of SNPs used for distinguishing between Mtb clades and/or drug sensitive versus drug resistant variants in the same clade was calculated by the following equation:

$$Power_k = 1 - \frac{A \cap B}{\min(N_A, N_B)} \quad (1)$$

where A∩B is the number of strains in the clades A and B sharing the same allelic state of the locus k; N_A and N_B – sample sizes of the clades A and B, respectively. Power values were in the range from 0 to 1.

SNPs with the highest discriminative power values estimated to distinguish between clades and/or antibiotic sensitive and resistant variants were selected to create the diagnostic key as explained below. The diagnostic key comprises of 1458 selected missense SNPs. The allelic states of these diagnostic SNPs predicted for 1201 Mtb genomes including drug resistance metadata from the GMTV database are shown in Supplementary file 1 in FASTA format as a pseudo-sequence of variable amino acids. Annotation information about each SNP location in the reference genome Mtb H37Rv (NC_000962.3) is given in Supplementary file 2. These files served as the program training dataset.

2.3. Mtb lineage classification

M. tuberculosis H37Rv reference genome (NC_000962.3) was used to determine the polymorphisms. In addition to this data, discriminative single nucleotide polymorphisms (SNPs) were identified by whole genome alignment against reference genomes of *M. bovis* (NC_016804, NC_020245, NC_012207, NC_008769 and NC_002945) and *M. canettii* (NC_015848, NC_019950, NC_019965, NC_019951 and NC_019952) available from the NCBI database. Variant calling for these strains was

performed using Mauve 2.3.1 (Darling et al., 2004). Clade identification by the Resistance Sniffer program was compared to the results obtained by TBProfiler (Coll et al., 2015), CASTB (Iwai et al., 2015), Mykrobe Predictor (Bradley et al., 2015), PhyResSE (Feuerriegel et al., 2015) and KvarQ (Steiner et al., 2014).

2.4. Resistance sniffer algorithm

The program, Resistance Sniffer, was developed in Python 2.7 (also compatible with Python 2.5) and implemented as an on-line tool at http://resistance-sniffer.bi.up.ac.za/. The program is also available for download, with example input files, from http://resistance-sniffer.bi.up.ac.za/Mycobacterium_tuberculosis/help/ as a stand-alone tool. The accepted input includes complete sequences in Genbank or FASTA formats; sequences of predicted genes or proteins in FASTA format, uncompressed VCF files and raw Illumina *fastq* paired-end read files. The program maps raw sequences to the embedded reference genome sequence (*M. tuberculosis* H37Rv, NC_000962.3). The detected patterns of polymorphisms are then processed using the diagnosis key, which consists of a catalogue of clade-specific polymorphisms and genetic determinants of antibiotic resistance. The diagnosis key consists of bifurcating splits for each decision point (Fig. 1). At each intermediate node, the program calculates normalized counts of power values (Eq. 1) of diagnostic polymorphisms depending on the states of these sites in the given genome. As the program was designed to perform predictions based on partially sequenced genomes, the program does not expect to receive the states of all polymorphic sites assigned for a split and tries to make a decision based on the available sites. Optimally, the score for one bifurcating branch is expected to be 1.0, and for another branch – 0.0. If the maximal score is below 0.75, the program explores both alternative branches to avoid an erroneous decision on a top-level split. Moreover, reaching the leaf-node corresponding to an Mtb clade, the program tries a possibility that the strain may belong to a sister clade sharing similar polymorphisms. It must be emphasized that in this work we did not attempt to distinguish between phylogenetically significant traits and convergent polymorphisms. No conclusions regarding the phylogenetic relatedness between clades should be drawn from the neighboring of the clades in the diagnostic key in Fig. 1.

It should be noted that in many cases there are no clear borders between Mtb clades and intermediate strains do exist, hence in instances where the program cannot reach a confident conclusion with regards to strain affiliation, the program returns two top-scored clades.

Contrary to a general belief in the existence of DR mutations common for all Mtb strains, this approach proceeds from an assumption of parallel drug resistance evolution in Mtb clades which resulted in the creation of different clade-specific patterns of polymorphic sites associated with the antibiotic resistance phenotype (van Niekerk et al., 2018). Each clade node of the diagnostic key consists of associated sets of polymorphic sites which distinguish between antibiotic resistant and antibiotic sensitive variants for every Mtb clade. Using the same method described in the clade identification step above, the program calculates normalized counts of polymorphisms associated with the drug sensitivity (*SenCount*) and drug resistance (*ResCount*). In the next step, antibiotic resistance scores (q values) are calculated for every individual antibiotic by Eq. 2.

$$Iq = \frac{1 + \log_2\left(\frac{1 + ResCount}{1 + SenCount}\right)}{2} \quad (2)$$

In the following steps, the resistance value (R) and the standard error (Err) are calculated by Eqs. 3 and 4, respectively.

$$R = q \times 2^{(2 \times ResCount - 1)} \quad (3)$$

$$Err = \frac{2q(1-q) \times 2^{(2 \times ResCount - 1)}}{\sqrt{N-1}} \quad (4)$$

In Eq. 4, N is the number of diagnostic sites found in the given genome.

Fig. 1 details the sequence of steps taken by the program in assigning the clade to the sample. This is followed by determining whether the strain is resistant or susceptible to each of the antibiotics. It must be noted that the number of antibiotics a strain may be resistant to

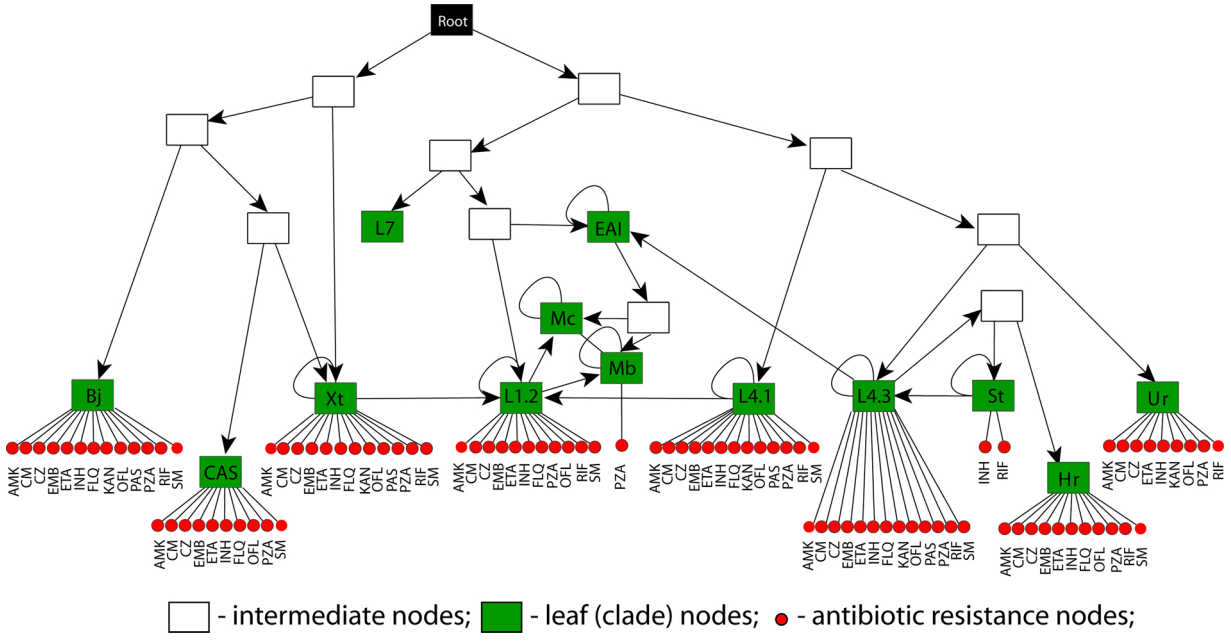

Fig. 1. The decision-making tree implemented in the Resistance Sniffer. Titled nodes denote Mtb clades: EAI – East African and Indian (Lineage 1.1); L1.2 – lineage 1.2; Bj – Beijing strains (lineage 2); CAS – Central Asian Strains (lineage 3); Xt – X-type strains; L4.1 – lineage 4.1 (H37Rv type strain clade); Ur – Ural strains (lineage 4.2); L4.3 – lineage 4.3; Hr – Haarlem strains (subtype of lineage 4.3); St – S-type (subtype of lineage 4.3); L7 – lineage 7; Mb – *M. bovis* and Mc – *M. canettii* (relate to lineages 5 and 6). Intermediate nodes represent groups of clades. Antibiotic resistance nodes are amikacin (AMK), capreomycin (CM), cycloserin (CS), ethambutol (EMB), ethionamide (ETH), isoniazid (INH), fluoroquinolones (FLQ), kanamycin (KAN), ofloxacin (OFL), para-amino salicyclic acid (PAS), pyrazinamide (PZA), rifampicin (RIF) and streptomycin (SM).

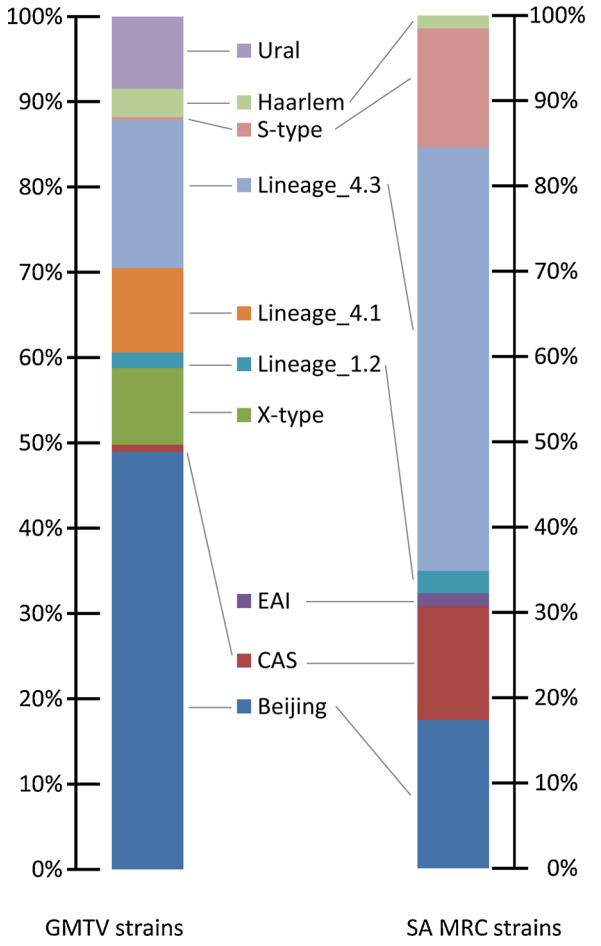

Fig. 2. Frequencies of clades assigned to Mtb strains from GMTV and SA MRC.

depends on the clade affiliation of the given strain. East African and Indian (EAI) clade, Lineage 7 and *M. canettii* are considered to be sensitive to all antibiotics by default since the antibiotic metadata from GMTV and SA MRC indicated that none of the isolates belonging to these clades was resistant to any antibiotic. *M. bovis* isolates are by default set to be resistant to PZA (Ritz et al., 2009) and sensitive to all the other antibiotics (Rousseau and Dupuis, 1990). However, this does not rule out the possibility of the future discovery of antibiotic resistance in *M. canettii* and *M. bovis* strains. The diagnostic key was implemented as an external text file *table.txt* located in the subdirectory *sources* of the local version of the program and on the server. This file may be edited to incorporate new diagnostic keys without any other changes to program. Antibiotic resistance diagnostic keys will be added to these nodes when more data becomes available.

Program validation was performed first on 1201 Mtb strains from GMTV combined with 742 new Mtb isolates from SA MRC, which were provided with antibiotic resistance/susceptibility patterns. Additionally, the program was tested on a set of 77 Mtb isolates from Sierra Leone available from the PATRIC database (Wattam et al., 2013) with data on their sensitivity/resistance to several antibiotics (project PRJEB7727). The program performance was characterized by sensitivity (SENS), specificity (SPEC), Positive Predictive Value (PPV) and Negative Predictive Value (NPV) as shown in Eqs. 5–8:

$$SENS = \frac{TP}{TP + FN} \quad (5)$$

$$SPEC = \frac{TN}{FP + TN} \quad (6)$$

$$PPV = \frac{TP}{TP + FP} \quad (7)$$

$$NPV = \frac{TN}{TN + FN} \quad (8)$$

In Eqs. 5–8, *TP* – number of true positive predictions; *FP* – false positives; *TN* – true negatives; and *FN* – false negatives.

3. Results

3.1. Phylogenetic clade classification

The accuracy of this tool depends on the correct classification of the phylogenetic lineage of the Mtb sequences. The precision of the Resistance Sniffer program for lineage classification was assessed using lineage information from the original GMTV test dataset, as well as lineage information from the Sierra Leone dataset. It is important to note that lineage information was not available for all the samples from the SAMRC, 121 samples from the GMTV as well as for 11 of the Sierra Leone samples. Supplementary file 3 presents the results of identification of clades of Mtb strains from Sierra Leone by 6 different tools, PhyResSE, KvarQ, Mykrobe Predictor, TBProfiler, CASTB and Resistance Sniffer, and compared to the original prediction provided for these strains in the PATRIC database. Generally, lineage predictions by different programs were in concordance. The main reason for discordance in classifying samples can be attributed to alternative naming of clades, different levels of resolution in classifying sub-lineages of Mtb by different tools and rather arbitrary delineation between lineages with a great abundance of intermediate strains. For example, classification ambiguities were observed for 4 strains from the Sierra Leone dataset (ERS457923, ERS457211, ERS457423, and ERS457331) which were misclassified as either CAS, or Beijing, or Lineage 1.2. For Resistance Sniffer there was ambiguous delineation between *M. bovis*, *M. africanum* and predominately African Mtb isolates of lineages 5 and 6 which may have resulted due to a small number of representatives of these groups in the initial training dataset. However, in terms of multidrug resistance development these groups are not very important (or maybe not sufficiently studied until now) except for *M. bovis* which isolates are naturally resistant to PZA (Ritz et al., 2009).

The predicted clades for the GMTV and SAMRC strains are shown in Fig. 2. The GMTV database is predominantly comprised of Mtb strains isolated in Russia, while SA MRC presents clinical isolates from South Africa. The majority of the GMTV strains belong to the highly virulent Beijing clade, European lineage 4.3, Asian lineage 4.1 (type strain lineage from India) and Ural clade specific for central Russia. European lineage 4.3 is the most prevalent clade among South African isolates. It is followed by the S-type variants of this lineage, Beijing and CAS clades. This finding indicates that the TB pandemic was introduced to South Africa by infected European settlers and to a lesser extent by Asian travelers. EAI clade is present among South African isolates but not frequent, most likely due to a relatively lower virulence. Ural, X-type and lineage 4.1 were not found among the SA MRC isolates.

3.2. Accuracy of antibiotic resistance prediction

In total, 1201 Mtb strains from GMTV and 742 strains from SA MRC were characterized by their sensitivity to one or several antibiotics resulting in 8559 data entries. This information was used to validate performance of the Resistance Sniffer program. Antibiotic resistance was predicted by *R*-values calculated using Eq. 3. This equation returns values in the range of 0–2; however, the majority of tested strains *R*-values were below 1.0, and those strains showing higher *R*-values were antibiotic resistant. The program was set to reduce *R*-values to 1.0 if they were larger than 1.0.

Assignment of strains as sensitive or resistant with respect to a given antibiotic was performed by setting a cut-off *R*-value. If the cut-off

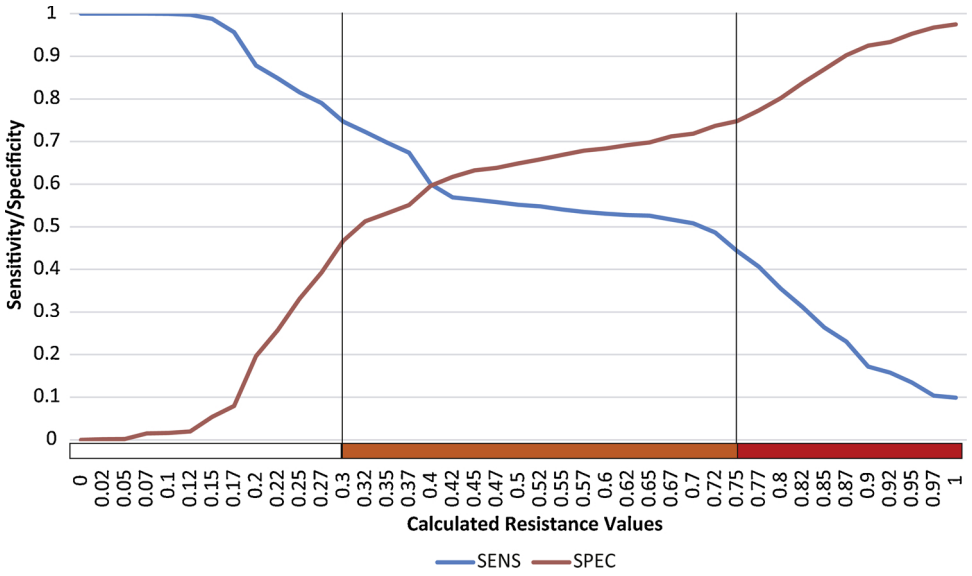

Fig. 3. Sensitivity and specificity of antibiotic resistance prediction with different R cutoff values. Vertical lines depict borders set in the program to distinguish between sensitive, potentially resistant and antibiotic resistant Mtb strains.

value is 0, all the strains will be deemed resistant and fall either into true positive (TP) or false positive (FP) categories. Sensitivity of the program will be maximal (1.0) and specificity will be minimal (0.0). Conversely, all the strains will be deemed sensitive and fall into either the true negative (TN) or false negative (FN) categories with the cut-off value of 1.0. The *R* cut-off value was gradually changed from 0 to 1.0 with a step of 0.25 and the distribution of FP, FN, TP and TN strains was evaluated using Eqs. 5–8. The calculated specificity and sensitivity of the program for different cut-off values are shown in Fig. 3.

Resistance *R*-value above 0.75 predicts the strain to be resistant against the given antibiotic with a likelihood of 55 % or higher. If the *R*-value is below 0.3, the strain is deemed sensitive to the antibiotic with a likelihood of 55 % or higher. The strains with intermediate *R*-values are either sensitive or resistant with equal likelihoods. This area of uncertainty has an important biological meaning as it depicts Mtb strains which may currently be sensitive but are rapidly approaching resistance and should be marked as potentially dangerous.

Testing of the program on 77 Mtb isolates from Sierra Leone characterized by sensitivity to EMB, INH, PZA, RIF and SM, or at least to one or several of these antibiotics (in total 285 strains per antibiotic measurements in Supplementary file 4) was performed with different cut-off *R*-values. Average values of sensitivity (0.5) and specificity (0.5) were achieved under an assumption that a strain is resistant to an antibiotic if the estimated *R*-value $\geq$ 0.25. An increase of the *R*-value cut-off led to a rapid increase in specificity and decrease in sensitivity. To improve the program performance, an additional parameter 'Sensitive' was introduced to reflect in the program output. This parameter was calculated as 1 −average of the top 6 *R*-values determined for different antibiotics. The rational of this approach was that a multidrug resistant Mtb strain may very likely show some level of resistance to other antibiotics even if no specific genetic determinants of the specific antibiotic resistance were found. The optimal program performance was achieved with the *R*-value cutoff 0.3 under an assumption that the strain is resistant to all antibiotics, if the Sensitivity coefficient is equal or lower than 0.2. With these settings, the susceptibility to antibiotics was correctly predicted in 184 cases and the resistance was correctly predicted in 19 cases. There were 41 false susceptibility predictions and 41 false resistance predictions. Calculated sensitivity and specificity were 0.32 and 0.82, respectively. As no data on the reliability of applied DST techniques was made available, it is not possible to judge whether the false negative and false positive predictions should be attributed to the experimental procedures of drug susceptibility testing or to the program algorithm. It should be noted that antibiotic resistance profiling of Mtb isolates in bacteriological laboratories is an error-prone procedure showing a relatively weak correlation with the clinical response due to the slow growth rate of this bacterium and bad standardization of the procedures (Kim, 2005). The accuracy is even worse when the data originates from different laboratories. It is expected that the sensitivity and specificity of the program cannot exceed the accuracy of the training dataset, but it seems that the antibiotic resistance prediction by Resistance Sniffer does not add significantly to the expected level of errors seen in laboratory drug resistance trials. The meaning of antibiotic resistance likelihood predicted by Resistance Sniffer will be discussed in more details below.

3.3. Program interface and output visualization

Resistance Sniffer comes with a user-friendly Web interface (Fig. 4), which allows the end user to upload files in different sequence file formats, including NGS read files. Even fragmented genomes can be used for antibiotic resistance prediction as every identification step is based on an analysis of the states of multiple diagnostic sites distributed throughout the complete genome sequence. This means that the tool can be used at different stages of whole genome completion which may be in raw reads to a SNP level.

Instead of uploading one genome at a time, users may download the stand-alone version of the program and analyze all available sequence files in a single run. However, the local version of the program cannot analyze *fastq* input files as it requires Bowtie2 to be locally installed. Depending on the format of the input file, the program run may take from several seconds to a few minutes on a regular desktop computer. The online version of the program allows optional entering of the user e-mail address to be notified of the task completion by a message with a link to the results. A detailed description on using the program is available from the download webpage at http://resistance-sniffer.bi.up.ac.za/Mycobacterium_tuberculosis/help/.

Most DR-TB databases approach the drug resistant phenotype as a binary entity which means that a strain is classified as either resistant or susceptible. However, our study suggests that the progression to the drug resistant phenotype is a stepwise process which highlights the need to develop ways to account for intermediate levels of drug resistance as well. Resistance Sniffer outputs the results as a bar plot of the probability that the strain is drug sensitive or drug resistant to the thirteen antibiotics. Fig. 5 shows several examples of graphical outputs

Fig. 4. Web user interface of Resistance Sniffer showing the accepted input file types.

of the program. The drug susceptibility pattern estimated for the strain TB0775 from GMTV is demonstrated in Fig. 5A. The strain was predicted to belong to the Beijing clade. The program prediction shows that this isolate has a high likelihood to be resistant to INH, KAN and SM, and may have an intermediate resistance towards FLQ, OFL, PAS and RIF. The experimentally detected profile of drug susceptibility available for this strain from GMTV confirms resistance to INH, SM and RIF, and sensitivity to EMB which agrees with the software prediction. This strain was not tested for other antibiotics.

Fig. 5B shows the prediction of drug resistance for a highly fragmented assembly of a clinical isolate from the SAMRC. The strain was assigned with equal likelihoods of belonging to either CAS or to X-type. It may be possible that the queried sequence is from an intermediate variant; however, the ambiguity could be attributed to the quality of sequencing. Only small fractions of diagnostic sites were found in the sequences which resulted in an increased standard error of *R*-value estimation depicted on the plot by an increased length of black vertical whiskers. Nevertheless, the program predicted a high likelihood that this strain may be resistant to ETA, KAN and OFL, and may also show an intermediate resistance to CM, EMB, FLQ, INH, RIF and SM. Because the program could not distinguish between CAS and X-type, patterns of resistance were analyzed for both these clades and the biggest *R*-values were selected. Resistance to PAS was not expected for either CAS or X-type isolates as there were no such isolates in the training dataset used for this program. This is why the program set this strain sensitive to PAS by default without any estimation. Setting of the drug sensitivity by default is depicted by short grey bar.

Only a few diagnostic sites needed for PZA resistance prediction were found in this fragmented genome and they were uninformative. For example, in PZA sensitive CAS isolates, a *Met* residue is expected on the 134th codon of the Rv0040c gene while *Ile* is expected on the same locus for PZA resistant isolates. In the given strain, *Val* was found on this locus, a finding that does not fit with the expectations. The program marked this antibiotic on the plot with a short red bar indicating an insufficiency in the information to make any decision.

In Fig. 5C, isolate ERS458164 from Sierra Leone was predicted as *M. bovis* clade, which also includes predominantly drug susceptible Western African and *M. africanum* isolates. The current version of the program was not designed to analyses the drug resistance in this clade due to lack of published data. The program displays by default that the isolate is most likely susceptible to all antibiotics except for the vaccination *M. bovis* strains reported to be PZA resistant (Ritz et al., 2009) that is indicated by highlighting the PZA resistance in the output file.

In Fig. 5D, an example is given of the analysis of historical DNA, in NGS format, obtained from human remains of an individual who died from tuberculosis in XVIII century in Hungary (Kay et al., 2015). The current analysis confirmed the affiliation of the Mtb strain with lineage 4 as reported in the original paper but with a better precision of the identification to the sub-lineage 4.3, which is common for Europe. This strain already possessed many mutations specific to future multidrug resistant Mtb variants of this lineage; however, this strain most likely was still susceptible to all antibiotics (sensitivity coef. 0.55).

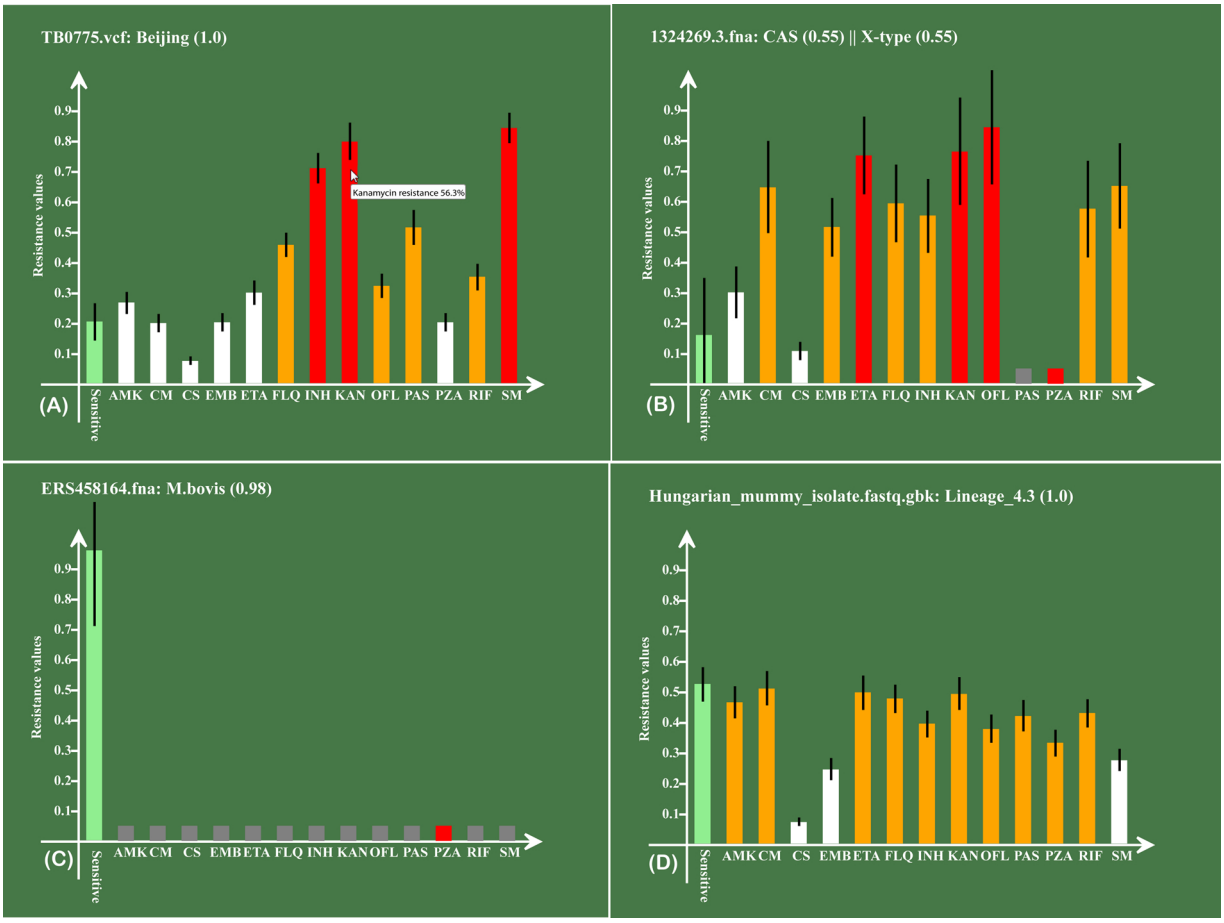

Fig. 5. Drug resistance predictions by Resistance Sniffer for the strains (A) GMTV strain TB0775; (B) SA MRC strain 1324269.3; (C) Sierra Leone isolate ERS458164; (D) DNA sequences obtained from a XVIII century Hungarian mummy who died from tuberculosis. White columns show sensitivity to antibiotics with the confidence above 55 %; red columns predict the resistance with the confidence above 55 %; and orange columns show intermediate results. The green most right column depicts the likelihood for this strain to be sensitive to all 13 antibiotics. Estimated R-values are shown along the vertical axis. Standard errors of calculation are depicted by black vertical whiskers. Antibiotic resistance nodes are amikacin (AMK), capreomycin (CM), cycloserin (CS), ethambutol (EMB), ethionamide (ETH), isoniazid (INH), fluoroquinolones (FLQ), kanamycin (KAN), ofloxacin (OFL), para-amino salicyclic acid (PAS), pyrazinamide (PZA), rifampicin (RIF) and streptomycin (SM). (For interpretation of the references to colour in the figure, the reader is referred to the web version of this article).

Both the online and stand-alone program implementations also return a text output file listing the states of all diagnostic sites in addition to the graphical output in SVG file format.

Resistance Sniffer predicts the resistance to antibiotics by analyzing patterns of diagnostic polymorphisms in genome sequences. The actual resistance of a bacterium to antibiotics may be affected by some other factors that the program does not consider; particularly by epigenetic modifications. For example, a recent study on the application of a new drug, FS-1, which causes the reversion of multidrug resistant bacteria back to the sensitive phenotype showed that the pattern of drug resistant mutations remained unchanged in the strains with reversed susceptibility to antibiotics (Ilin et al., 2017). Thus, the strains with reversed antibiotic sensitivity due to epigenetic modifications will be predicted as resistant by the pattern of diagnostic polymorphisms. Moreover, the same study showed that Mtb isolates from experimentally infected laboratory guinea pigs showed a range of antibiotic susceptibility values due to natural variations within the population even though all the animals were injected with the same MDR-TB strain. Hence the antibiotic resistance prediction is probabilistic by nature. The Resistance Sniffer program estimates a rough likelihood for an isolate to be resistant to one of thirteen antibiotics or to be sensitive to all of them. These estimations are recorded in the text output file or may be displayed on the screen when the mouse pointer is placed over the bar on the plot (see Fig. 5A).

4. Discussion

Resistance polymorphisms appear to be clade specific, which means that some mutations are more likely to be present in specific Mtb lineages (Sandgren et al., 2009; Comas et al., 2010; Spies et al., 2011; Fonseca et al., 2015). There is also evidence suggesting that strains from certain lineages are predisposed to develop into MDR-TB strains (Bifani et al., 2002; Drobniewski et al., 2005; Borrell and Gagneux, 2009). In the present study, we sought to identify clade specific patterns of polymorphisms that may be associated with the drug resistant phenotype in Mtb. Multiple mutations associated with antibiotic resistance in Mtb are known from literature and are available from public databases (e.g. https://tbdreamdb.ki.se/Info/Default.aspx) (Sandgren et al., 2009). However, for this study a decision was made to perform a *de novo* search for relevant mutations by calculating power coefficients (Eq. 1) of association between polymorphisms and the drug resistant phenotype for each antibiotic per clade. In many cases, but not always, high scored polymorphisms were consistent with known drug resistance mutations in literature. Some of these high scored polymorphisms were also located in genes associated with the antibiotic resistance and there is a need for further interrogation of these findings. Several examples are discussed below.

Mutations in the *emb*CAB operon have been known to be associated with resistance to EMB, particularly the substitutions in codons 306,

406 and 497 of *emb*B (Zhao et al., 2015). The current study confirmed that in the Beijing clade several polymorphisms were highly scored on these loci, particularly in codons 306, 354 and 406 of the *embB* gene. Strain TB0012 has mutations in codons 306 and 354 but shows a susceptible phenotype according to the GMTV database records. Strain TB0011 shows a resistant phenotype and possesses mutations in codons 354 and 406. Two other strains, TB0004 and TB0005, have mutations in all three codons although they both show a susceptible phenotype. These findings highlight the complexity of the development of antibiotic resistance in Mtb which requires accumulation of many other subordinate mutations acquired in a stepwise manner to develop sustainable drug resistance. This assumption was used as the basis of the Resistance Sniffer algorithm of estimation of the drug resistance likelihood by assessing the whole pool of genetic determinants biasedly distributed between antibiotic sensitive and resistant Mtb variants. This hypothesis was confirmed in numerous publications (Telenti et al., 1997; Safi et al., 2010; Safi et al., 2013; Fonseca et al., 2015). It was hypothesized that the *mut* gene may play a significant role in the acquisition of drug resistance in Mtb because missense mutations in these genes lead to higher mutations rates (Rad et al., 2003; Fonseca et al., 2015). Indeed, mutations in the *mutT4* gene (Rv3908) were associated with the drug resistance phenotype of Mtb strains of the Beijing clade, but no specific mutations were found in the strains of lineage 4.1 and S-type lineage, which, in contrast to Beijing strains, usually do not develop a wide spectrum of drug resistance (see Fig. 1). The Beijing clade has been associated with high levels of drug resistance and a higher propensity to develop into MDR-TB and XDR-TB. The reason why Beijing isolates are often associated with MDR-TB remains elusive. Researchers have suggested that the strain background could be more efficient in mitigating the effects of fitness cost imposed by drug resistance (Borrell and Gagneux, 2009; Fenner et al., 2012).

Another gene of interest in this study was *ogt* (Rv1316), which is known to remove methyl groups from O-6-methylguanine in DNA. Mutations in this gene were associated with SM resistance in Ural, Haarlem, lineage 4.3 and X-type clades. However, no significant associations with mutations in this gene and SM resistance were revealed in Beijing, CAS, S-type and lineage 4.1, which also have a high propensity to develop SM resistance but in a different way. Strangely enough, this study did not discover any significant correlation between mutations in genes *inhA, eis* and *tylA,* and the drug resistance phenotype despite many publications linking these genes with drug resistance. Mutations in the *inhA* regulatory regions are known to confer low level INH and ETH resistance. The *tylA* and *eis* genes are both drug targets for the second-line injectable antibiotics. This work discovered several mutations in the phenolpthiocerol synthesis polyketide gene (*ppsC)* and multifunctional mycoserosic acid synthase gene (*masA*) to be associated with rifampicin resistance in the lineage 4.3. A study by Bisson et al. (2012) showed that the expression level of *pps* can be up to 10 fold higher in *rpoB* mutant strains relative to the RIF susceptible parent strain.

Our study suggests that the progression to the drug resistant phenotype is a stepwise process involving the accumulation of multiple mutations contributing to the antibiotic resistant phenotype. Alternative hypotheses involving rare drug resistance mutations in Mtb populations were proposed by other authors. For example, in the publication by Carvalho et al. (2018) it was suggested that rare cases of resistance to D-cycloserine is caused by low frequency mutations in target genes, *cycA*, *alr* and *ddlA*, rather than fitness cost reduction mediated by other compensatory mutations. Our study confirmed the importance of CycA V67C, L322R and M343 T Alr, and T365A DdlA substitutions in the development of CS resistance; however, significant associations with multiple compensatory mutations linked to other antibiotic resistance were also observed. It may explain the insignificant fitness cost of the CS resistance mutations as they occur only in organisms already possessing the compensatory mutations. The Resistance Sniffer program may identify Mtb strains on a trajectory to developing drug resistance by accumulation of pre-requisite mutations, even if phenotypic DST results for these strains do not show any evidence of antibiotic resistance yet.

The key to the total eradication of TB globally lies in early diagnosis and correct treatment. This has been hampered by the limitations in current laboratory methodologies in performing drug susceptibility testing. Current DST procedures for Mtb are time consuming, expensive and inaccurate, especially for the second line antibiotics. Horizontal gene transfer plays no role in the development of antibiotic resistance in Mtb. This makes WGS an attractive option in the diagnosis of TB as it has the potential to determine the full antibiogram provided we have detailed knowledge of all the genetic determinants of drug resistance (Coll et al., 2015). In this study, we used a derivative of GWAS to identify clade specific patterns of polymorphisms which showed a biased distribution regarding the drug resistant phenotype. The study was designed first of all as a proof of concept of the ability to predict drug resistance or the predisposition for drug resistance acquisition by Mtb isolates. However, the designed software tool, Resistance Sniffer, showed the sensitivity and specificity of the clade and the drug resistance identification similar to that of other available tools such as TBProfiler, MyKrobe, KvarQ and PhyResSE. A recent large-scale benchmarking comparison of the available tools on 6746 Mtb isolated characterized by drug susceptibility patterns showed applicability but also some limitations of the available tools (Ngo and Teo, 2019). According to this report, the specificity and sensitivity of the programs varied from 0.6 to 0.9 depending on which antibiotic was tested with the best results achieved when they are confirmed by more than one program. All the programs showed a much better ability to predict the absence of drug resistance rather than the specific drug resistance pattern. This is also true for the Resistance Sniffer program. It was not discussed in this review to which extent the performance of the program was affected by the level of fragmentation of the genome of interest as only whole genome sequences were used in the reported study. While the estimated sensitivity and specificity of Resistance Sniffer were lower than those of the above-mentioned programs, it should be noted that the current program was developed to analyze fragmented partial genome sequences including historical sequences (see Fig. 5D) represented by different file formats. Particularly, unordered contigs of Mtb isolates from Sierra Leone in plain fasta format were used for the program evaluation The aim of the study was to estimate the propensity of a Mtb isolate to gain the antibiotic resistance rather than to delineate antibiotic resistant from antibiotic sensitive strains. The performance of the program may be improved in future studies by editing the diagnostic key table without the need to modify the program itself. A limiting factor of the current version was the size of the training dataset of Mtb strains with known drug susceptibility profiles. For certain clades, the number of available records was not sufficient to boost the association power. Although there are currently more than 20 drugs that are used in the treatment of TB, this study was limited to only thirteen antibiotics due to the unavailability of phenotypic DST data for the omitted drugs. As more data becomes available, the diagnostic key table of the program will be updated.

It is expected that in the future NGS based assays will replace phenotypic DST methods (Gröschel et al., 2018). This study has demonstrated how whole or partial genome sequence data can be used to rapidly predict drug resistance in *M. tuberculosis*. However, it should be emphasized that the current version of the program was not designed for application in clinics or for assessing antibiotic treatment regiments. The major objective of the program was to provide scientists working in public health control institutes with reliable software to estimate the distribution of drug resistant infections by using NGS datasets in different stages of genome assembly including raw *fastq* files generated by sequencers. This work also added to the current body of knowledge a valid suggestion that drug resistant phenotype is associated not with individual mutations but with clade-specific patterns of polymorphisms. Effective prediction of drug resistance should start from a proper identification of clade affiliation of Mtb isolates.

5. Conclusion

We have developed a rapid online tool that uses both complete and partial NGS datasets to predict resistance to thirteen anti-TB drugs in a lineage specific manner. Given the need for rapid and accurate diagnosis in the management of TB, WGS has the potential to bridge most of the diagnosis gaps left by current phenotypic DST methods. Although sequencing directly from sputum is currently challenging, there is a need to continue elucidating the complex evolution of drug resistance in Mtb. In future, this shall not only spur the innovation of improved diagnostic tools but will also help clinicians in designing effective treatment regimens and speed up the treatment before mutations develop. The flexibility of the proposed methodology also allows for easy updating of the diagnosis table as well as addition of new antibiotics. While many alternative tools are currently available for Mtb clade identification and antibiotic resistance prediction, the Resistance Sniffer program combines these procedures and allows analyzing of sequence datasets in multiple file formats and at different stages of genome sequence completion including files with raw DNA reads in *fastq* format generated by NGS sequencers. Other tools often require knowledge of the command line as well as the Linux operating system which often deters non-specialist end users from adopting NGS technologies; and requires also input files in sophisticated formats such as VCF. Resistance Sniffer comes with a user-friendly graphical user interface and produces easy to interpret output. The program considers the development of drug resistance in Mtb to be a multistep process which involves the sequential acquisition and accumulation of drug resistance mutations. Estimated likelihoods of resistance to thirteen anti-TB antibiotics allows for the prediction of Mtb strains possessing high levels of drug resistance as well as those strains that are likely to acquire high levels of drug resistance in future.

Funding

This work including the MSc student stipend of DM was supported by the South African National Research Foundation (NRF) [grant numbers 105996].

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijmm.2020.151399.

References

Alangaden, G.J., Kreiswirth, B.N., Aouad, A., Khetarpal, M., Igno, F.R., Moghazeh, S.L., Manavathu, E.K., Lerner, S.A., 1998. Mechanism of resistance to amikacin and kanamycin in *Mycobacterium tuberculosis*. Antimicrob. Agents Chemother. 42, 1295–1297.

Bifani, P.J., Mathema, B., Kurepina, N.E., Kreiswirth, B.N., 2002. Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains. Trends Microbiol. 10, 45–52.

Bisson, G.P., Mehaffy, C., Broeckling, C., Prenni, J., Rifat, D., Lun, D.S., Burgos, M., Weissman, D., Karakousis, P.C., Dobos, K., 2012. Upregulation of the phthiocerol dimycocerosate biosynthetic pathway by rifampin-resistant, *rpoB* mutant *Mycobacterium tuberculosis*. J. Bacteriol. 194, 6441–6452.

Borrell, S., Gagneux, S., 2009. Infectiousness, reproductive fitness and evolution of drug-resistant *Mycobacterium tuberculosis*. Int. J. Tuberc. Lung Dis. 13, 1456–1466.

Borrell, S., Gagneux, S., 2011. Strain diversity, epistasis and the evolution of drug resistance in *Mycobacterium tuberculosis*. Clin. Microbiol. Infect. 17, 815–820.

Bradley, P., Gordon, N.C., Walker, T.M., Dunn, L., Heys, S., Huang, B., Earle, S., Pankhurst, L.J., Anson, L., De Cesare, M., 2015. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. Nat. Commun. 6, 10063.

Brzostek, A., Sajduda, A., Śliwiński, T., Augustynowicz-Kopeć, E., Jaworski, A., Zwolska, Z., Dziadek, J., 2004. Molecular characterisation of streptomycin-resistant *Mycobacterium tuberculosis* strains isolated in Poland. Int. J. Tuberc. Lung Dis. 8, 1032–1035.

Carvalho, L.P., Evangelopoulos, D., Prosser, G., Rodgers, A., Dagg, B., Khatri, B., Ho, M.M., Gutierrez, M., Cortes, T., 2018. Antibiotic resistance evasion is explained by rare mutation frequency and not by lack of compensatory mechanisms. BioRxiv, 374215.

Chen, J., Zhang, S., Cui, P., Shi, W., Zhang, W., Zhang, Y., 2017. Identification of novel mutations associated with cycloserine resistance in *Mycobacterium tuberculosis*. J. Antimicrob. Chemother. 72, 3272–3276.

Chen, M.L., Doddi, A., Royer, J., Freschi, L., Schito, M., Ezewudo, M., Kohane, I.S., Beam, A., Farhat, M., 2019. Beyond multidrug resistance: leveraging rare variants with machine and statistical learning models in *Mycobacterium tuberculosis* resistance prediction. EBioMedicine 43, 356–369.

Chernyaeva, E.N., Shulgina, M.V., Rotkevich, M.S., Dobrynin, P.V., Simonov, S.A., Shitikov, E.A., Ischenko, D.S., Karpova, I.Y., Kostryukova, E.S., Ilina, E.N., Govorun, V.M., Zhuravlev, V.Y., Manicheva, O.A., Yablonsky, P.K., Isaeva, Y.D., Nosova, E.Y., Mokrousov, I.V., Vyazovaya, A.A., Narvskaya, O.V., Lapidus, A.L., O'Brien, S.J., 2014. Genome-wide *Mycobacterium tuberculosis* variation (GMTV) database: a new tool for integrating sequence variations and epidemiology. BMC Genomics 15, 308.

Coll, F., McNerney, R., Preston, M.D., Guerra-Assuncao, J.A., Warry, A., Hill-Cawthorne, G., Mallard, K., Nair, M., Miranda, A., Alves, A., Perdigao, J., Viveiros, M., Portugal, I., Hasan, Z., Hasan, R., Glynn, J.R., Martin, N., Pain, A., Clark, T.G., 2015. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med. 7, 51.

Coll, F., Phelan, J., Hill-Cawthorne, G.A., Nair, M.B., Mallard, K., Ali, S., Abdallah, A.M., Alghamdi, S., Alsomali, M., Ahmed, A.O., 2018. Genome-wide analysis of multi-and extensively drug-resistant *Mycobacterium tuberculosis*. Nat. Genet. 50, 307.

Comas, I., Chakravartti, J., Small, P.M., Galagan, J., Niemann, S., Kremer, K., Ernst, J.D., Gagneux, S., 2010. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. Nat. Genet. 42, 498.

Darling, A.C.E., Mau, B., Blattner, F.R., Perna, N.T., 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14, 1394–1403.

Demers, A.M., Venter, A., Friedrich, S.O., Rojas-Ponce, G., Mapamba, D., Jugheli, L., Sasamalo, M., Almeida, D., Dorasamy, A., Jentsch, U., Gibson, M., Everitt, D., Eisenach, K.D., Diacon, A.H., 2016. Direct susceptibility testing of *Mycobacterium tuberculosis* for pyrazinamide by use of the Bactec MGIT 960 System. J. Clin. Microbiol. 54, 1276–1281.

Dookie, N., Rambaran, S., Padayatchi, N., Mahomed, S., Naidoo, K., 2018. Evolution of drug resistance in *Mycobacterium tuberculosis*: a review on the molecular determinants of resistance and implications for personalized care. J. Antimicrob. Chemother. 73, 1138–1151.

Drobniewski, F., Balabanova, Y., Nikolayevsky, V., Ruddy, M., Kuznetzov, S., Zakharova, S., Melentyev, A., Fedorin, I., 2005. Drug-resistant tuberculosis, clinical virulence, and the dominance of the Beijing strain family in Russia. JAMA 293, 2726–2731.

Fenner, L., Egger, M., Bodmer, T., Altpeter, E., Zwahlen, M., Jaton, K., Pfyffer, G.E., Borrell, S., Dubuis, O., Bruderer, T., Siegrist, H.H., Furrer, H., Calmy, A., Fehr, J., Stalder, J.M., Ninet, B., Bottger, E.C., Gagneux, S., 2012. Effect of mutation and genetic background on drug resistance in *Mycobacterium tuberculosis*. Antimicrob. Agents Chemother. 56, 3047–3053.

Feuerriegel, S., Oberhauser, B., George, A.G., Dafae, F., Richter, E., Rüsch-Gerdes, S., Niemann, S., 2012. Sequence analysis for detection of first-line drug resistance in *Mycobacterium tuberculosis* strains from a high-incidence setting. BMC Microbiol. 12, 90.

Feuerriegel, S., Schleusener, V., Beckert, P., Kohl, T.A., Miotto, P., Cirillo, D.M., Cabibbe, A.M., Niemann, S., Fellenberg, K., 2015. PhyResSE: a web tool delineating *Mycobacterium tuberculosis* antibiotic resistance and lineage from whole-genome sequencing data. J. Clin. Microbiol. 53, 1908–1914.

Fonseca, J., Knight, G., McHugh, T., 2015. The complex evolution of antibiotic resistance in *Mycobacterium tuberculosis*. Int. J. Infect. Dis. 32, 94–100.

Gandhi, N.R., Moll, A., Sturm, A.W., Pawinski, R., Govender, T., Lalloo, U., Zeller, K., Andrews, J., Friedland, G., 2006. Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. Lancet 368, 1575–1580.

Gröschel, M.I., Walker, T.M., van der Werf, T.S., Lange, C., Niemann, S., Merker, M., 2018. Pathogen-based precision medicine for drug-resistant tuberculosis. PLoS Pathog. 14, e1007297.

Hicks, N.D., Carey, A.F., Yang, J., Zhao, Y., Fortune, S.M., 2019. Bacterial genome-wide association identifies novel factors that contribute to ethionamide and prothionamide susceptibility in *Mycobacterium tuberculosis*. mBio 10, e00616–00619.

Homolka, S., Post, E., Oberhauser, B., George, A.G., Westman, L., Dafae, F., Rüsch-Gerdes, S., Niemann, S., 2008. High genetic diversity among *Mycobacterium tuberculosis* complex strains from Sierra Leone. BMC Microbiol. 8, 103.

Ilin, A.I., Kulmanov, M.E., Korotetskiy, I.S., Islamov, R.A., Akhmetova, G.K., Lankina, M.V., Reva, O.N., 2017. Genomic insight into mechanisms of reversion of antibiotic resistance in multidrug resistant *Mycobacterium tuberculosis* induced by a nanomolecular iodine-containing complex FS-1. Front. Cell. Infect. Microbiol. 7, 151.

Iwai, H., Kato-Miyazawa, M., Kirikae, T., Miyoshi-Akiyama, T., 2015. CASTB (the comprehensive analysis server for the *Mycobacterium tuberculosis* complex): a publicly accessible web server for epidemiological analyses, drug-resistance prediction and phylogenetic comparison of clinical isolates. Tuberculosis Edinb. (Edinb) 95, 843–844.

Kay, G.L., Sergeant, M.J., Zhou, Z., Chan, J.Z., Millard, A., Quick, J., Szikossy, I., Pap, I., Spigelman, M., Loman, N.J., Achtman, M., Donoghue, H.D., Pallen, M.J., 2015. Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. Nat. Commun. 6, 6717.

Kim, S., 2005. Drug-susceptibility testing in tuberculosis: methods and reliability of results. Eur. Respir. J. 25, 564–569.

Koch, A., Mizrahi, V., Warner, D.F., 2014. The impact of drug resistance on *Mycobacterium tuberculosis* physiology: what can we learn from rifampicin? Emerg. Microbes Infect. 3, 1–11.

Macedo, R., Nunes, A., Portugal, I., Duarte, S., Vieira, L., Gomes, J.P., 2018. Dissecting whole-genome sequencing-based online tools for predicting resistance in *Mycobacterium tuberculosis*: can we use them for clinical decision guidance?

Tuberculosis Edinb. (Edinb) 110, 44–51.
Ngo, T.M., Teo, Y.Y., 2019. Genomic prediction of tuberculosis drug-resistance: benchmarking existing databases and prediction algorithms. BMC Bioinf. 20, 68.
Okamoto, S., Tamaru, A., Nakajima, C., Nishimura, K., Tanaka, Y., Tokuyama, S., Suzuki, Y., Ochi, K., 2007. Loss of a conserved 7-methylguanosine modification in 16S rRNA confers low-level streptomycin resistance in bacteria. Mol. Microbiol. 63, 1096–1106.
Oppong, Y.E., Phelan, J., Perdigão, J., Machado, D., Miranda, A., Portugal, I., Viveiros, M., Clark, T.G., Hibberd, M.L., 2019. Genome-wide analysis of *Mycobacterium tuberculosis* polymorphisms reveals lineage-specific associations with drug resistance. BMC Genomics 20, 252.
Phelan, J., O'Sullivan, D.M., Machado, D., Ramos, J., Whale, A.S., O'Grady, J., Dheda, K., Campino, S., McNerney, R., Viveiros, M., Huggett, J.F., Clark, T.G., 2016. The variability and reproducibility of whole genome sequencing technology for detecting resistance to anti-tuberculous drugs. Genome Med. 8, 132.
Rad, M.E., Bifani, P., Martin, C., Kremer, K., Samper, S., Rauzier, J., Kreiswirth, B., Blazquez, J., Jouan, M., van Soolingen, D., 2003. Mutations in putative mutator genes of *Mycobacterium tuberculosis* strains of the W-Beijing family. Emerg. Infect. Dis. 9, 838.
Ramaswamy, S.V., Reich, R., Dou, S.-J., Jasperse, L., Pan, X., Wanger, A., Quitugua, T., Graviss, E.A., 2003. Single nucleotide polymorphisms in genes associated with isoniazid resistance in *Mycobacterium tuberculosis*. Antimicrob. Agents Chemother. 47, 1241–1250.
Reeves, A.Z., Campbell, P.J., Sultana, R., Malik, S., Murray, M., Plikaytis, B.B., Shinnick, T.M., Posey, J.E., 2013. Aminoglycoside cross-resistance in *Mycobacterium tuberculosis* due to mutations in the 5′ untranslated region of *whiB7*. Antimicrob. Agents Chemother. 57, 1857–1865.
Ritz, N., Tebruegge, M., Connell, T.G., Sievers, A., Robins-Browne, R., Curtis, N., 2009. Susceptibility of *Mycobacterium bovis* BCG vaccine strains to antituberculous antibiotics. Antimicrob. Agents Chemother. 53, 316–318.
Rousseau, P., Dupuis, M., 1990. Antituberculous drug susceptibility testing of *Mycobacterium bovis* BCG strain Montreal. Can. J. Microbiol. 36, 735–737.
Safi, H., Fleischmann, R.D., Peterson, S.N., Jones, M.B., Jarrahi, B., Alland, D., 2010. Allelic exchange and mutant selection demonstrate that common clinical *embCAB* gene mutations only modestly increase resistance to ethambutol in *Mycobacterium tuberculosis*. Antimicrob. Agents Chemother. 54, 103–108.
Safi, H., Lingaraju, S., Amin, A., Kim, S., Jones, M., Holmes, M., McNeil, M., Peterson, S.N., Chatterjee, D., Fleischmann, R., 2013. Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-β-D-arabinose biosynthetic and utilization pathway genes. Nat. Genet. 45, 1190.
Sandgren, A., Strong, M., Muthukrishnan, P., Weiner, B.K., Church, G.M., Murray, M.B., 2009. Tuberculosis drug resistance mutation database. PLoS Med. 6, e1000002.
Schleusener, V., Köser, C.U., Beckert, P., Niemann, S., Feuerriegel, S., 2017. *Mycobacterium tuberculosis* resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools. Sci. Rep. 7, 46327.
Shean, K., Streicher, E., Pieterson, E., Symons, G., van Zyl Smit, R., Theron, G., Lehloenya, R., Padanilam, X., Wilcox, P., Victor, T.C., van Helden, P., Grobusch, M.P., Warren, R., Badri, M., Dheda, K., 2013. Drug-associated adverse events and their relationship with outcomes in patients receiving treatment for extensively drug-resistant tuberculosis in South Africa. PLoS One 8, e63057.
Spies, F.S., Ribeiro, A.W., Ramos, D.F., Ribeiro, M.O., Martin, A., Palomino, J.C., Rossetti, M.L.R., da Silva, P.E.A., Zaha, A., 2011. Streptomycin resistance and lineage-specific polymorphisms in *Mycobacterium tuberculosis* gidB gene. J. Clin. Microbiol. 49, 2625–2630.
Steiner, A., Stucki, D., Coscolla, M., Borrell, S., Gagneux, S., 2014. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. BMC Genomics 15, 881.
Takiff, H.E., Salazar, L., Guerrero, C., Philipp, W., Huang, W.M., Kreiswirth, B., Cole, S.T., Jacobs, W.R., Telenti, A., 1994. Cloning and nucleotide sequence of *Mycobacterium tuberculosis gyrA* and *gyrB* genes and detection of quinolone resistance mutations. Antimicrob. Agents Chemother. 38, 773–780.
Telenti, A., Philipp, W.J., Sreevatsan, S., Bernasconi, C., Stockbauer, K.E., Wieles, B., Musser, J.M., Jacobs, W.R., 1997. The *emb* operon, a gene cluster of *Mycobacterium tuberculosis* involved in resistance to ethambutol. Nat. Med. 3, 567.
Theron, G., Peter, J., Meldau, R., Khalfey, H., Gina, P., Matinyena, B., Lenders, L., Calligaro, G., Allwood, B., Symons, G., Govender, U., Setshedi, M., Dheda, K., 2013. Accuracy and impact of Xpert MTB/RIF for the diagnosis of smear-negative or sputum-scarce tuberculosis using bronchoalveolar lavage fluid. Thorax 68, 1043–1051.
Trauner, A., Borrell, S., Reither, K., Gagneux, S., 2014. Evolution of drug resistance in tuberculosis: recent progress and implications for diagnosis and therapy. Drugs 74, 1063–1072.
van Niekerk, K., Pierneef, R., Reva, O., Korostetskiy, I., Ilin, A., Akhmetova, G., 2018. Clade-specific distribution of antibiotic resistance mutations in the population of *Mycobacterium tuberculosis*. Prospects for drug resistance reversion. In: Enany, S. (Ed.), Basic Biology and Applications of Actinobacteria. IntechOpen, London, pp. 79–98.
Walker, T.M., Kohl, T.A., Omar, S.V., Hedge, J., Del Ojo Elias, C., Bradley, P., Iqbal, Z., Feuerriegel, S., Niehaus, K.E., Wilson, D.J., Clifton, D.A., Kapatai, G., Ip, C.L.C., Bowden, R., Drobniewski, F.A., Allix-Beguec, C., Gaudin, C., Parkhill, J., Diel, R., Supply, P., Crook, D.W., Smith, E.G., Walker, A.S., Ismail, N., Niemann, S., Peto, T.E.A., 2015. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. Lancet Infect. Dis. 15, 1193–1202.
Walker, T.M., Merker, M., Kohl, T.A., Crook, D.W., Niemann, S., Peto, T.E., 2017. Whole genome sequencing for M/XDR tuberculosis surveillance and for resistance testing. Clin. Microbiol. Infect. 23, 161–166.
Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., Gillespie, J.J., Gough, R., Hix, D., Kenyon, R., 2013. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 42, D581–D591.
World Health Organization, 2018a. Global Tuberculosis Report 2018. World Health Organization, Geneva, pp. 1–8.
World Health Organization, 2018b. The Use of Next-generation Sequencing Technologies for the Detection of Mutations Associated With Drug Resistance in *Mycobacterium tuberculosis* Complex: Technical Guide. World Health Organization, Geneva, pp. 1–10.
Yang, T.W., Park, H.O., Jang, H.N., Yang, J.H., Kim, S.H., Moon, S.H., Byun, J.H., Lee, C.E., Kim, J.W., Kang, D.H., 2017. Side effects associated with the treatment of multidrug-resistant tuberculosis at a tuberculosis referral hospital in South Korea: a retrospective study. Medicine 96, e7482.
Zhao, L.-l., Sun, Q., Liu, H.-c., Wu, X.-c., Xiao, T.-y., Zhao, X.-q., Li, G.-l., Jiang, Y., Zeng, C.-y., Wan, K.-l., 2015. Analysis of *embCAB* mutations associated with ethambutol resistance in multidrug-resistant *Mycobacterium tuberculosis* isolates from China. Antimicrob. Agents Chemother. 59, 2045–2050.