Tuberculosis 115 (2019) 81–88

Contents lists available at ScienceDirect

Tuberculosis

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

Epidemiology

Evaluation of a gene-by-gene approach for prospective whole-genome sequencing-based surveillance of multidrug resistant *Mycobacterium tuberculosis*

Rita Macedo[a,1], Miguel Pinto[b,1], Vítor Borges[b,1], Alexandra Nunes[b], Olena Oliveira[c,d,e], Isabel Portugal[f], Raquel Duarte[e,g,h], João Paulo Gomes[b,*]

[a] National Reference Laboratory for Mycobacteria, Department of Infectious Diseases, National Institute of Health, 1649-016 Lisbon, Portugal
[b] Bioinformatics Unit, Department of Infectious Diseases, National Institute of Health, 1649-016 Lisbon, Portugal
[c] Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, 4710-057 Braga, Portugal
[d] ICVS/3B, PT Government Associate Laboratory, 4710-057, Braga/4805-017, Guimarães, Portugal
[e] EPI Unit, Institute of Public Health, University of Porto, 4050-600 Porto, Portugal
[f] iMed.ULisboa–Research Institute for Medicines, University of Lisbon, Lisbon, Portugal
[g] Clinical Epidemiology, Predictive Medicine and Public Health Department, Faculty of Medicine, University of Porto, 4200-319 Porto, Portugal
[h] Pulmonology Department, Centro Hospitalar de Vila Nova de Gaia/Espinho EPE, 4400-129 Vila Nova de Gaia, Portugal

ARTICLE INFO

Keywords:
Mycobacterium tuberculosis
Whole-genome sequencing
Surveillance
Gene-by-gene approach
Multidrug-resistance

ABSTRACT

Whole-genome sequencing (WGS) offers unprecedented resolution for tracking *Mycobacterium tuberculosis* transmission and antibiotic-resistance spread. Still, the establishment of standardized WGS-based pipelines and the definition of epidemiological clusters based on genetic relatedness are under discussion. We aimed to implement a dynamic gene-by-gene approach, fully relying on freely available software, for prospective WGS-based tuberculosis surveillance, demonstrating its application for detecting transmission chains by retrospectively analysing all M/XDR strains isolated in 2013–2017 in Portugal. We observed a good correlation between genetic relatedness and epidemiological links, with strongly epilinked clusters displaying mean pairwise allele differences (AD) always below 0.3% (ratio of mean AD over the total number of shared loci between same-cluster strains). This data parallels the genetic distances acquired by the core-SNV analysis, while providing higher resolution and epidemiological concordance than MIRU-VNTR genotyping. The dynamic analysis of strain subsets (i.e., increasing the number of shared loci within each sub-set) also strengthens the confidence in detecting epilinked clusters. This gene-by-gene strategy also offers several practical benefits (e.g., reliance on freely-available software, scalability and low computational requirements) that further consolidated its suitability for a timely and robust prospective WGS-based laboratory surveillance of M/XDR-TB cases.

1. Introduction

In 2015, the World Health Organization (WHO) proposed to fight the global drug resistant tuberculosis (TB) crisis as one of its five high priority actions [1]. In 2017, of the 10.4 million that fell ill with TB worldwide, 1.7 million died from the disease, with 460 000 cases caused by multidrug-resistant strains (MDR-TB; i.e., resistant to, at least, rifampicin and isoniazid) of which 8.5% were also extensively drug-resistant (XDR-TB; i.e., MDR-TB with additional resistance to any fluoroquinolone and one or more of amikacin, kanamycin or capreomycin) [2]. Globally, TB incidence is falling at a rate of about 2% *per* year but in order to reach the 2020 milestones of the "End TB Strategy" [2] this rate needs to accelerate to a 4–5% annual decline. In Portugal, TB incidence has been steadily decreasing in the last years, with an average of about 5% *per* year [3]. From 2013 to 2017, about 10.000 new TB cases were reported to the Portuguese General Health Directorate (GHD), and the proportion of patients with MDR-TB remained stationary at 1% of total cases [3,4]. M/XDR-TB are more difficult and

* Corresponding author. Bioinformatics Unit, Department of Infectious Diseases, National Institute of Health, Avenida Padre Cruz, 1649-016 Lisbon, Portugal.
E-mail addresses: rita.macedo@insa.min-saude.pt (R. Macedo), miguel.pinto@insa.min-saude.pt (M. Pinto), vitor.borges@insa.min-saude.pt (V. Borges), alexandra.nunes@insa.min-saude.pt (A. Nunes), radomoliveira@gmail.com (O. Oliveira), iportugal@ff.ulisboa.pt (I. Portugal), raquelafduarte@gmail.com (R. Duarte), j.paulo.gomes@insa.min-saude.pt (J.P. Gomes).
[1] These authors contributed equally to this manuscript.

https://doi.org/10.1016/j.tube.2019.02.006
Received 8 November 2018; Received in revised form 15 January 2019; Accepted 22 February 2019

expensive to treat and have poorer survival rates when compared to drug sensitive TB, so monitoring and controlling these multi-drug resistant cases is key for successful TB control programs and for achieving the targets of the "End-TB Strategy" [2].

Targeted interventions to stop transmission, especially from M/XDR-TB cases, requires in depth epidemiological knowledge that can only be provided by a combination of effective genotyping with classical epidemiological enquiries. Concerning *M. tuberculosis* complex (MTC) strains, there are three methods that are commonly used for typing purposes: IS*6110* Restriction Fragment Length Polymorphism (RFLP) [5], spoligotyping (interspaced palindromic repeats) [6], and Mycobacterial Interspersed Repetitive Units-Variable Number of Tandem Repeats (MIRU-VNTR) [7]. These methods have been widely applied to answer a variety of questions, from the investigation of cross-contamination in the laboratory to the investigation of outbreaks or population-based studies [8–13]. Still, although traditional typing techniques provide standardized and easily computable typing results, these have limited discriminatory power. In this context, whole-genome sequencing (WGS) has emerged as a very powerful tool for the surveillance, outbreak investigation and drug resistance monitoring of human infectious pathogens including MTC. As the ongoing technological developments are rapidly decreasing costs, WGS has the potential to become the ultimate tool for diagnostics and pathogen typing [8,14–16], providing comprehensive genetic information, virtually including all possible genomic targets, as well as additional valuable information on drug resistance, virulence and genome evolution [8,17–19]. Additionally, efforts are being directed to overcome the time-consuming *M. tuberculosis* growth prior to WGS, as recent developments show that acquisition of WGS data directly from clinical samples may be possible [20–22].

The recent recommendations for TB molecular surveillance by the European Center for Disease Control and Prevention (ECDC) propose to introduce WGS-based surveillance methodologies in the National Reference Laboratories (NRL) of the EU/EEA regions [23]. While the time between sample collection and WGS data analysis is increasingly being shortened, there is still a lack of standardized guidelines for the use of WGS for molecular epidemiological analysis, which is crucial for Public Health Authorities to act. In particular, two practical issues remain in order to sustain alerts for possible ongoing transmission and to maintain data sharing, either locally or globally: i) the definition of the more suitable approach to apply in a WGS-based surveillance era, and ii) the establishment of cut-offs of genetic relatedness that display the highest congruence with epidemiological data for cluster definition. While the latter will become more consistent as more WGS-based studies are being performed (for which consistent epidemiological data is essential), there is still a discussion regarding the methodology to use, namely gene-by-gene and single nucleotide polymorphism (SNP) based approaches [14,15,24,25]. In addition, laboratories already performing WGS-based surveillance essentially rely on in-house command-line-based pipelines or commercial platforms (e.g., Bionumerics from Applied Maths or Ridom SeqSphere + from Ridom Bioinformatics), which may not be accessible to all laboratories, thus potentially delaying the implementation of a harmonized and globally accepted strategy.

In Portugal, since 2014, there are specific centres for the diagnosis, consultancy, monitoring, and treatment of the M/XDR-TB cases. As such, these centres constitute a major driving force for linking the epidemiological survey performed within the community by the Public Health Authorities [26] and the systematic molecular genotyping performed by the NRL at the National Institute of Health (NIH). Since the Portuguese NRL receives all the strains isolated from all the MDR-TB patients from Portugal (mandatory since 2007) [27], it is of utmost importance to set up a centralized and robust molecular typing system that potentiate the establishment of correlations between genetic and epidemiological information towards the detection/monitoring of the resistance profiles and transmission chains [28,29]. In light of the transition to a WGS-based laboratory surveillance in Portugal, the present study aimed at comparing two WGS-based strategies between them (as well as with the conventional genotyping method) for the detection of potential transmission chains among the M/XDR-TB cases isolated during the last 5 years (2013–2017). This work led to the implementation of a dynamic gene-by-gene approach for routine WGS-based TB surveillance, while contributing for the understanding on how cut-offs of genetic relatedness can be applied to strengthen epidemiological investigation and public health actions.

2. Material and methods

2.1. Sample dataset characterization and MIRU-VNTR genotyping

All MDR-TB strains isolated in Portugal are mandatorily sent to the TB NRL from the Portuguese NIH for drug susceptibility testing (DST) and genotyping [27,30]. Whenever a TB case is notified, public health services intervene in the community in order to identify other potentially associated TB cases or contacts at risk. An epidemiological enquiry is deployed and the information is stored in the Public Health Departments, at regional level. In addition, whenever a cluster is identified by molecular-based laboratory typing techniques, public health authorities are asked to confirm if there is an epidemiological link between the cases. This workflow was applied during the course of the present study with novel WGS-derived links being investigated retrospectively.

From the NIH collection of ~10 000 TB strains isolated at country level from 2013 to 2017, a total of 96 M/XDR TB strains were identified and notified to the Portuguese GHD. From these, 83 (86.5%) strains were available for molecular and genomic analysis (the remaining 13 were diagnosed based on molecular methods with no culture isolation). For 48 strains, DST, 24-loci MIRU-VNTR genotyping and WGS had been previously performed on behalf of a recent study focused on *in silico* prediction of antibiotic resistance [19]. For the remaining isolates, DST was performed for first- (isoniazid-INH, rifampicin-RMP, etambutol-EMB, streptomycine-STR and pyrazinamide-PZA) and second-line (amikacin-AMK, kanamycin-KAN, capreomycin-CAP, ofloxacin-OFX, moxifloxacin-MOX, ethionamide-ETH, linezolid-LNZ, cicloserine-CICLO and para-aminosalicylic acid-PAS) anti-TB drugs, using MGIT960 system (Becton Dickinson), according to manufacturer's instructions, or the proportion method using solid media (CICLO and PAS).

For the traditional MIRU-VNTR genotyping and WGS, total DNA was extracted from solid cultures using commercial extraction kits (QIAmp, Qiagen) after an initial step of cell inactivation where samples were subjected to 95 °C for 1 h followed by enzyme digestion for 3 h with proteinase K. The 24-loci MIRU-VNTR experimental genotyping was performed, as previously described [19], by standardized protocols using MIRU-VNTR typing kit (Genoscreen) according to manufacturer's instructions. Dendrograms were constructed using the online free software MIRU-VNTR plus [31,32]. For WGS, high-quality DNA samples (quantified using Qubit, ThermoFisher) were subjected to dual-indexed NexteraXT Illumina library preparation using the KAPA HiFi Hot Start Ready Mix PCR Kit (KAPA Biosystems) in the indexing step to improve amplification of the GC-rich genome regions, as previously described [19]. Libraries were subsequently subjected to cluster generation and paired-end sequencing (2 × 250bp) on an Illumina MiSeq (Illumina), available at the Portuguese NIH. All data for isolates used in the present study, including drug susceptibility, genotyping and demographic data, are summarized in Supplementary Table S1.

2.2. Genome de novo assembly

All 83 genomes were *de novo* assembled using the INNUca v3.1 pipeline (https://github.com/B-UMMI/INNUca), which consists of integrated modules for reads QA/QC, *de novo* assembly and post-assembly optimization steps. Briefly, after reads' quality analysis (FastQC v0.11.5 - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and

cleaning (Trimmomatic v0.36) [33], genomes are assembled with SPAdes v3.11 [34] and subsequently improved using Pilon v1.18 [35]. Based on a comparative analysis of the impact of performing "post-assembly polishment" of the assemblies using Pilon (data not shown), this step was skipped when Pilon introduced changes in more than 1/3 of the contigs. The affected genomes (n = 8) were also among the ones presenting high genome fragmentation (i.e., more than 200 contigs). Draft genome sizes, mean depth of coverage, number of contigs and sequencing read size are described in Supplementary Table S1. *In silico* DST were performed using TB-profiler v0.3.0 [36] as previously described [19]. MIRU-VNTR profiles were also predicted *in silico* with a recently developed tool, MIRU-profiler (first release), using default parameters [37].

2.3. Gene-by-gene analysis

For gene-by-gene analysis, two panel of loci were retrieved (27th of July 2018) from the RIDOM Nomenclature Server (https://www.cgmlst.org/ncs), developed by Kohl and colleagues [14,15]. These panels include 2891 genes from the *M. tuberculosis* complex "cgMLSTschema" and additional 755 "accessory" loci. In the present study, we inferred allelic diversity against either only the "short schema" (cgMLST 2891 loci) or the "extended schema" (cgMLST plus accessory loci; 3646 loci). For simplification purposes, the terms "short" and "extended" will be used throughout the manuscript to describe both schemas.

Allele calling was performed for the 83 genomes against both schemas, using chewBBACA v2.0.11 [38] with default parameters and a training file generated by Prodigal v2.6.3 from the H37Rv reference genome (RefSeq Accession NC_000962.3). After allele calling using a Blast Score Ratio of 0.6, exact and inferred matches were used to construct an allelic profile matrix, where other allelic classifications (see https://github.com/B-UMMI/chewBBACA/wiki) were assumed as "missing" loci.

Genomes with less than 90% (2602 loci) called in the short schema were removed for subsequent phylogenetic inferences. This occurred only for 3 out of the 83 genomes under analysis (Supplementary Table S1). Minimum spanning trees (MST) were constructed taking advantage of the goeBURST algorithm [39] implemented in the PHYLOViZ online web-based tool [40], based on 100% shared loci between all strains (i.e., shared-genome MLST). This approach was applied to both schemas and every cluster under evaluation. To take advantage of accessory genome loci, likely increasing the resolution power (which may be key for discriminating cases during outbreak investigation), we took advantage of PHYLOViZ online 2.0 Beta version (http://online2.phyloviz.net/), which allows maximizing the shared genome in a dynamic manner, i.e., for each sub-set of strains under comparison, the maximum number of shared loci between them is automatically used for tree construction. In this regard, all allelic distance thresholds used during cluster investigation were expressed as percentages of allele differences (AD) over the total number of shared loci under comparison. Initial potential clusters to explore were defined by applying an allelic distance cut-off of 0.56% to the initial tree enrolling 80 strains (i.e., 12 AD out of 2202 loci shared in the "short schema"; or 15 AD out of 2657 shared loci in the "extended schema"). The use of this conservative threshold relies on literature data that pointed out 12 AD as an accurate value for cluster investigation in gene-by-gene based surveillance era for *M. tuberculosis* [14,15,35]. Finally, in order to evaluate the clustering agreement between MIRU-VNTR and the gene-by-gene approach, the Wallace's coefficient, with 95% confidence intervals, was calculated using the web-based framework "Comparing partitions" (http://www.comparingpartitions.info/) [41].

2.4. Core-single nucleotide variant (SNV)-based analysis

A core-SNV-based analysis was also performed as it is one of the current methodologies used for MTB laboratory surveillance. For this, we applied Snippy v3.1 in the set of all 83 *M. tuberculosis* isolates. (https://github.com/tseemann/snippy). Briefly, quality improved read data (after Trimmomatic processing) of all isolates were individually mapped against the H37Rv reference genome (NC_000962.3), and SNP calling was performed on variant sites with the following criteria: minimum proportion of reads differing from the reference of 90%, a minimum mapping quality of 20 and a minimum coverage for SNP calling of 10. Core-SNPs were extracted using Snippy's core module (*snippy-core*) ensuring that all genomes reached at least 95% of aligned bases with the reference. To achieve this, a minimum coverage of 6 was set for five out of the 83 genomes and one genome was excluded from the analysis (TB82). In summary, this analysis enrols 82 strains (TB30 and TB35 had been excluded in gene-by-gene approach). Core-SNV falling within known *M. tuberculosis* genomic regions with high GC content or repetitive elements (Supplementary Table S2), as well as known SNVs in resistance-associated positions, were excluded (compiled by Kohl and colleagues, available at https://github.com/ngs-fzb/MTBseq_source/tree/master/var/res), as their inclusion would likely bias the phylogeny. Following the same rationale applied for gene-by-gene approach, we applied a conservative cut-off of a maximum of 12 differences [14,15] to the global core-SNV-based MST for comparison purposes.

2.5. Data availability

All raw sequence reads used in the present study were deposited in the European Nucleotide Archive under the study accession number PRJEB29446 (detailed in Supplementary Table S1), which will be the Bioproject including all MTB strains analyzed for prospective WGS-based surveillance in Portugal. As such, reads that were previously released (BioProject SRP131205) have also been deposited in PRJEB29446.

3. Results

3.1. MIRU-VNTR analysis

Our first approach consisted in analyzing the M/XDR-TB strains by MIRU-VNTR genotyping, the conventional method used for the molecular epidemiology of MTC strains. Overall, we found 41 distinct MIRU-VNTR profiles, constituting 10 different genomic clusters involving 2 to 15 strains each (Supplementary Table S1), thus yielding a clustering rate of 63.9% (53/83). Detailed analysis of these clusters will be integrated in the subsequent WGS-based analysis (see next sections). Although the *in silico* determination of VNTR profiles is challenging, a recently developed software (MIRU-profiler) was tested on our assembly data. The performance was unsatisfactory, with only a mean of 15 out of 24 VNTR profiles were correctly assigned for each strain, and was dependent on the assembly statistics (results detailed in Supplementary Fig. S1).

3.2. Gene-by-gene analysis

To investigate phylogenetic relationships between strains and their potential epidemiological link ("epilink"), we applied a gene-by-gene approach by taking advantage of publicly available schemas [using both a short (2891 loci) and an extended schema (3646 loci) [14,15]] and a freely available software for allele calling (chewBBACA) [38]. As both schemas contain "accessory" genes, i.e., genes not present in *M. tuberculosis* spp *tuberculosis*, only genomes yielding more than 90% of called loci (80 out of 83 strains) in the short schema were subjected to gene-by-gene analysis (Supplementary Table S1). In a first step, a MST was generated for each schema based on allelic diversity found among the shared loci: 2202 for the short (Supplementary Fig. S2A) and 2657 for the extended schema (Fig. 1A). As a conservative approach to

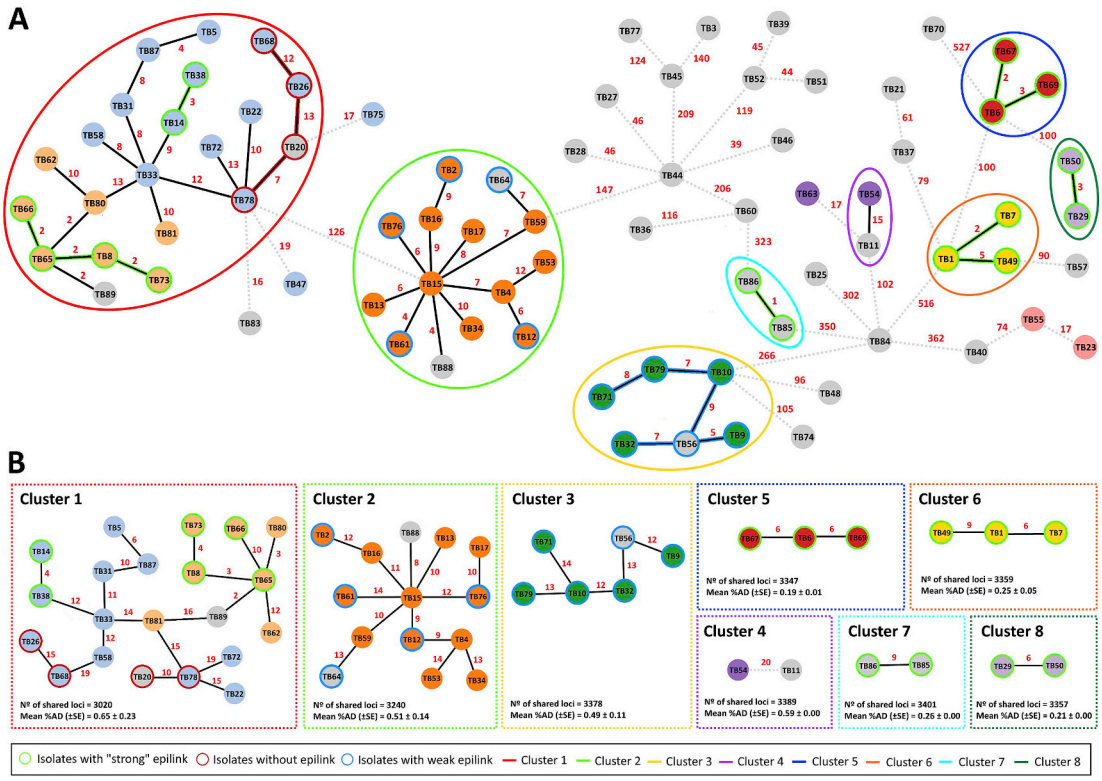

Fig. 1. Phylogeny of 80 M/XDR-TB strains based on a dynamic gene-by-gene approach using an extended schema (3646 loci). **A** – Initial Minimum spanning tree (MST) constructed based on allelic diversity found among the 2657 genes shared by 100% of the strains. Potential clusters defined for fine-tune analysis are highlighted by colored circles. **B** – Sub-MST reconstruction based on the maximum number of shared loci between strains within a potential cluster. Each circle (node) contains the strain's designation and represents a unique allelic profile. Nodes are colored according to different MIRU-VNTR profiles, where gray nodes refer to singletons. The numbers in red on the connecting lines represent the allele differences (AD) between strains. Straight and dotted lines reflect nodes linked with the ADs below and above the conservative threshold applied for initial cluster investigation (0.56%), respectively. MST were constructed using the goeBURST algorithm implemented in the PHYLOViZ Online platform, and are based on allelic profiles relying on distinct number of shared loci (indicated near eachsub-tree). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

generate potential clusters to be subjected to the dynamic MST analysis, the same cut-off of 0.56% was applied for both short and extended schemas (i.e., strains with ≤12 and ≤ 15 AD, respectively, were kept interconnected), resulting in seven (Supplementary Fig. S2A) and eight (Fig. 1A) clusters, respectively. The additional cluster in the extended schema (Cluster 4) was borderline in the short schema (the two enrolled strains, TB54 and TB11, displayed 14 AD; Supplementary Fig. S2). The content of the remaining clusters overlapped between both approaches. Of note, at the applied cut-off, the probability of two strains having the same MIRU-VNTR profile also belonging to the same cluster is about 85% (CI 69%–100%) (using either the short or the extended schema), while the probability of two strains that belong to the same cluster (using either the short or extended) also share the same MIRU-VNTR profile is around 50% (CI 35%–64%) (calculations using the Adjusted Wallace coefficient and 95% confidence intervals).

Within each cluster, the mean percentage of pairwise AD (i.e., the ratio of mean AD over the total number of shared loci) fell within the same magnitude for the "initial MST" (containing all 80 isolates) regardless of the schema used (Fig. 2A). A similar analysis after reconstruction of a sub-MST for each cluster, maximizing the number of shared loci between the sub-set of strains, revealed a general increase in mean percentage of AD, with this trend being more marked for the extended schema (Fig. 2A), indicating that accessory genes contributed to diversity. As the extended schema enrolls a much higher number of shared loci (Fig. 2B), this dynamic approach increases sensitivity, thus allowing a better evaluation of the close-relatedness of strains and, consequently, making cluster definition and exclusion of outliers more robust. For instance, the sub-MST of the additional cluster 4 in the extended schema consolidated the weak genetic link between the two enrolled strains (Fig. 1B). On the other hand, for cluster 5, despite the addition of almost 700 shared loci to the analysis, the absolute AD between strains increased by up to 4 AD, thus strengthening the potential link between the three strains. In general, regardless the schema that was used, the diversity within each cluster showed two patterns: 4 clusters (clusters 1 to 4) with mean %AD of 0.4–0.7 and 4 clusters (clusters 5 to 8) with mean %AD of 0.0–0.3 (Fig. 2A).

After correlating phylogenetic data with epidemiological data provided by the Public Health Authorities for each potential cluster, strong or weak epilinks could be established among strains within the analyzed clusters. In particular, while all strains within clusters 5 to 8 (with mean %AD below 0.3) had a strong epilink (the authorities confirmed these "strong" epi links as family members or very close contacts) (Fig. 2C), a weak epidemiological link was reported for some strains of cluster 2 and all strains of cluster 3 (representing strains from patients of the same geographic region and/or same chest clinic attendance), which could explain the higher mean pairwise %AD observed for the latter. In cluster 1, the largest cluster analyzed, while the overall mean %AD was high, very strong epilinks were reported for two clusters of patients (Fig. 2C; Fig. 1B). Concordantly, the mean %AD among epilinked strains was below 0.3%, while confirmed non-epilinked strains had %AD within ~0.3–0.5 (Fig. 2C). If we applied this empirical observed mean %AD (≤0.3%; equivalent to ≤9 ADs) as a linkage cut-off to the sub-MST of cluster 1, three sub-clusters would be obtained (Fig. 1B), including the two already confirmed and a novel cluster. In fact, while TB14 and TB38 remained linked, this fine-tune analysis raises the hypothesis that the other confirmed epilinked cluster containing strains TB66, TB65, TB73 and TB8, may also include TB80 and TB89. Also, TB5 and TB87 may also be epilinked (novel cluster).

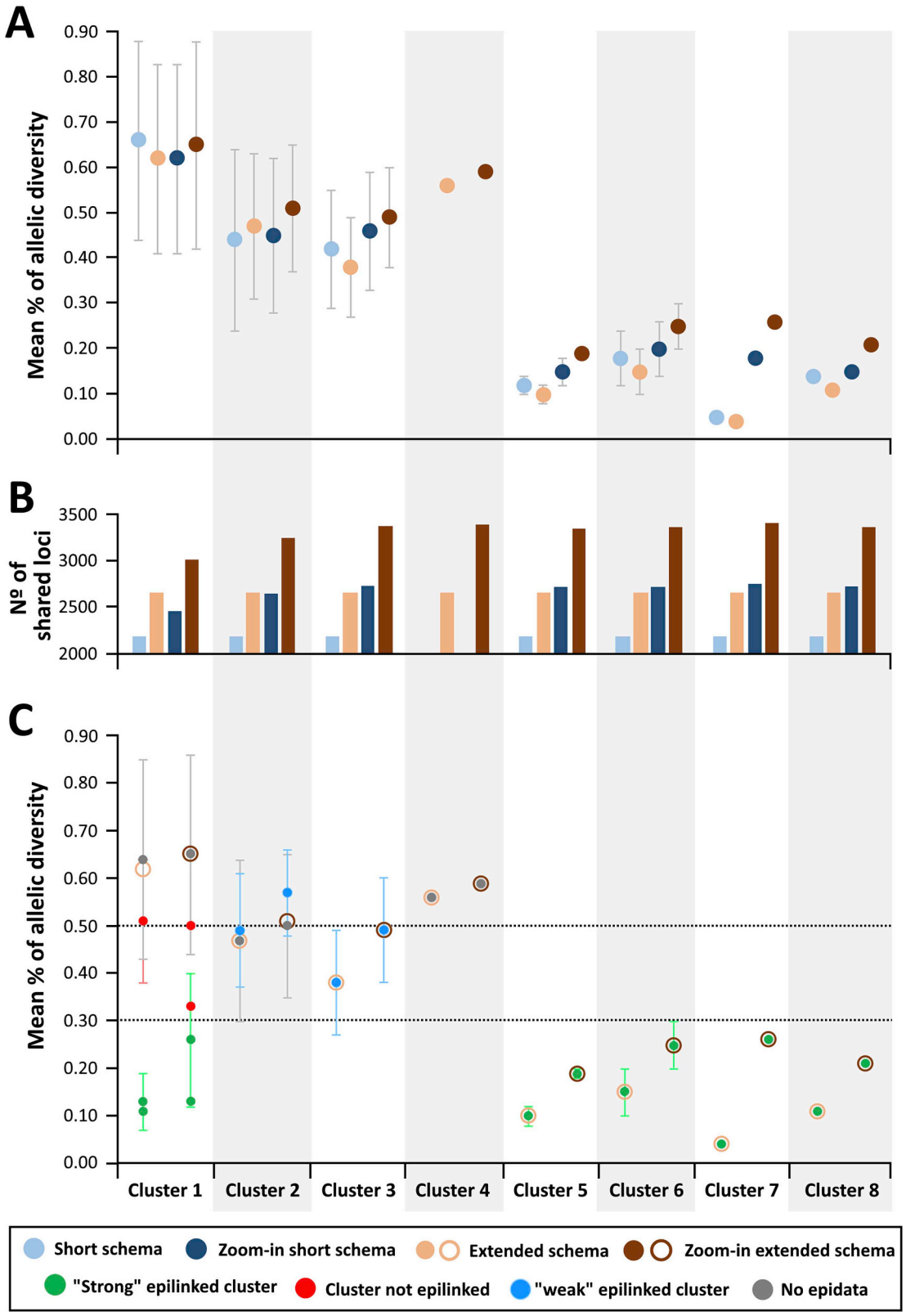

Fig. 2. Allelic diversity with potential and confirmed clusters. **A** – Comparison of the mean percentage of allelic diversity (with standard deviation) observed within each potential cluster using the short and extended schemas, either with or without sub-minimum spanning tree (MST) reconstruction. **B** – Number of shared loci analyzed for each cluster detected by each approach. **C** – Comparison of the mean percentage of allelic diversity (with standard deviation) observed within each potential cluster, after its fragmentation into sub-sets according to the available epidemiological data. Calculations are presented when using the extended schema, with or without dynamic sub-MST reconstruction. Epilinks are highlighted by different colors. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

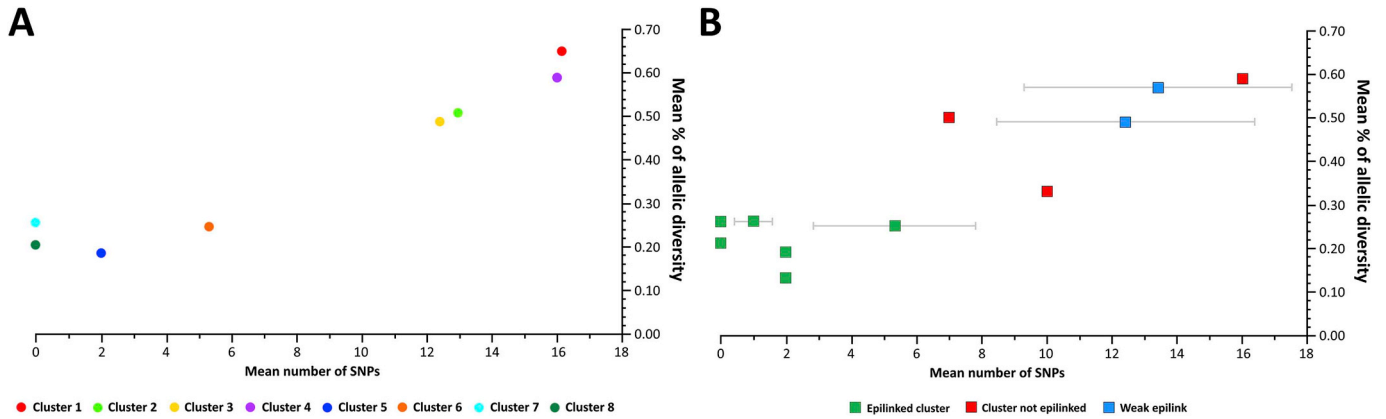

Fig. 3. Genetic diversity within clusters evaluated by the extended gene-by-gene and the core-SNV approach. Mean pairwise percentage of allelic diversity versus mean pairwise SNP distance within (**A**) each potential cluster defined in the extended gene-by-gene approach using a threshold of 0.56% (Fig. 1) and (**B**) their fragmentation according to the available epidemiological data.

Unfortunately, both potential epilinks (which arose from this dynamic analysis) could not be retrospectively traced. Regarding cluster 2 and 3, although only weak links or no data were retrieved, we cannot discard some possible links among the clustered strains, as observed when applying this cut-off. Of note, adjusted Wallace coefficient analysis at 0.3% threshold showed that the probability of two strains having the same MIRU-VNTR profile also belonging to the same cluster decreased to about 28% (using either the short or the extended schema), while same-cluster strains have about 60% probability of belonging to the same MIRU-VNTR profile. Nevertheless, a proper inference of the level of congruence will require further large-scale studies with more diverse datasets.

3.3. Core-SNP-based analysis

A core-SNV-based approach matrix was used to infer phylogenetic relationships between isolates, after filtering repetitive loci and resistance-associated SNVs. The resulting MST, enrolling 82 strains, is presented in Supplementary Fig. S3. As expected, we observed a strong correlation between the mean %AD and the mean number of SNPs (Fig. 3) within each of the eight clusters investigated by the gene-by-gene approach (extended schema). Most importantly, when looking at the 10 clusters with epidemiological support (including eight with confirmed epilink and two with confirmed non-epilink), we observed two threshold patterns, where all clusters with strong epilinks have less than 6 SNPs/0.3 %AD, whereas the weak or confirmed non-epilinks have pairwise distances ranging from 6 to 16 SNPs/0.3–0.7 %AD, sustaining the parallelism with the gene-by-gene approach.

4. Discussion

The current comparative study led to the implementation of a WGS-based laboratory workflow for TB surveillance in the Portuguese NRL, in the frame of the demanding supra-national short-term transition from traditional genotyping methods (MIRU-VNTR). A collection of 83 M/XDR-TB strains, isolated between 2013 and 2017, was characterized by traditional genotyping as well as by distinct methods for WGS data analysis (i.e., a gene-by-gene approach using both a short schema and an extended schema, and a core-SNV strategy) in order to set up a WGS-based TB surveillance focused on unveiling possible ongoing transmission chains to be flagged for action and further analysis by the health authorities. This study is the first to be performed in Portugal with these specific aims and involves the M/XDR strains isolated in the last 5 years. In general, the same clusters were essentially detected regardless the WGS-based approaches, both in strain number and content (Fig. 1; Supplementary Figs. S2 and S3). As expected, strains with the same MIRU-VNTR profile revealed high genetic relatedness at the genome level, even though we could unveil intricate genetic links between strains with distinct MIRU-VNTR patterns. Additionally, our attempt to extract MIRU-VNTR profiles *in silico* (using MIRU-profiler; [37]) yielded unsatisfactory results (Supplementary Fig. S1). Although we believe that efforts in this field are needed in order to maintain backward compatibility with traditional genotyping data, our results consolidate the expectation that repetitive genomic regions are still challenging to be extracted from short-read WGS data.

Although similar results were obtained using a gene-by-gene analysis with both short and extended schemas, we observed that the increase in the number of shared loci analyzed may have an important role in the sensitivity (Fig. 2), which may be key for outbreak investigation. In fact, the introduction of accessory genes in the schema not only allows performing a dynamic analysis of epidemiological clusters, but also increases the confidence in results, as the link between strains is strengthened when AD decrease or remain very similar after generating sub-MST. This concept of dynamically reconstructing phylogenies for sub-sets of strains, maximizing the number of shared loci, as well as the bioinformatics tools needed for its operationalization, have been set-up on behalf of a recent internationally collaborative project (INNUENDO; https://sites.google.com/site/theinnuendoproject/) that precisely aimed at developing analytical frameworks for the use of WGS in routine surveillance and epidemiologic investigations [42]. Noteworthy, by providing an in-depth perspective on the accessory genomes content, this strategy also opens new possibilities to enrich our knowledge on MTC biology and population genetics, by identifying for instance, lineage specific genetic markers to use for rapid cluster identification [21,43,44] or by linking specific phenotypes with gene presence/absence profiles [45].

In the present study, a good correlation was observed between the magnitude of strain genetic relatedness (defined both by gene-by-gene and core-SNV-based strategies) and the strength of the epidemiological link between patients, as reported by the Public Health Authorities. After conservatively defining potential clusters using the implemented dynamic gene-by-gene approach, we could subsequently observe that the mean pairwise distance among strongly epilinked strains was always below 0.3%AD, while confirmed non-epilinked and weakly epilinked strains had > 0.3%AD. This trend strongly correlated with data acquired through the core-SNV strategy, with strains sharing strong epilinks displaying mean pairwise distances of less than 6 SNPs and strains with weak or confirmed non-epilinks with up to 16 SNPs (Fig. 3). Of note, we found very close related strains within this threshold for which epidemiological data was not available, raising the hypothesis that some retrospective links may have been missed. These results are in agreement with previous studies, where genetic distance

among strains from recent transmission chains reveal a maximum number of 12 SNPs/AD [14,15,22,35,46–48], and also corroborate the expectation that applying lower thresholds will increase the concordance with epidemiological data. Several studies have already applied even more stringent thresholds [22,25]. Still, we believe that cut-offs should not be applied in a static manner, but instead should be dynamically fit to the sub-set of strains/lineages under investigation, which fits the evolutionary and epidemiological dynamics of *M. tuberculosis* that relies on a low mutation rate coupled with a long-term transmissibility. In fact, this can be operationally achieved with the gene-by-gene strategy described in this study as the maximum number of shared loci is automatically used for each potential cluster, with allelic distance thresholds being expressed as percentages of allele differences (AD) over the total number of shared loci under comparison. Moreover, this dynamic analysis should always be contingent on the epidemiological data, when available. In this context, this study illustrates the need to reinforce the conventional epidemiological tracing (e.g., by improving clinical inquiries and strengthening the communication workflow between stakeholders) in order to potentiate the benefits of an enhanced WGS-based laboratory surveillance focused on detecting/confirming more robust contact tracing investigation.

While the advantages of WGS-based typing analysis are undeniable for routine TB laboratory surveillance, there are still several drawbacks that need to be overcome before a multi-country harmonized WGS-based surveillance [49], such as: i) the lack of standardized pipelines; ii) the need of empirical-derived thresholds for cluster definition and, iii) the existence of an internationally agreed nomenclature to be used in order to facilitate data sharing, especially for cross-border outbreaks [50,51]. Also, there is still an ongoing debate on which approach will be the method of choice for TB, i.e., gene-by-gene-based, SNP-based or a combination of both [14,15,22,35,46–48], which will always be a trade-off between usability and reliability. In contrast with pay-per-use platforms (such as RIDOM SeqSphere) currently used by several reference laboratories, our approach fully relies on freely available software which, despite requiring more bioinformatics training, may constitute a valuable alternative for most laboratories, especially those with less resources. Although the gene-by-gene strategy applied here may offer several benefits (such as amenability to standardization, scalability, and less constrains to implement a harmonized nomenclature) [42,52,53], it is expected that the “ultimate” evidence about which method provides better usability and concordance with epidata will certainly emerge from studies at population level enrolling extensive datasets.

In summary, this study stands as the “starting point” for the application of a freely-available standardized routine WGS-based laboratory surveillance of M/XDR-TB cases. Together with the centralization of the diagnosis and molecular typing in the NIH and a strong articulation with the Public Health Authorities, the described framework constitutes the driving force towards a faster and robust prospective surveillance of M/XDR-TB cases, from antibiotic resistance prediction to transmission chain detection. As such, future studies based on this approach will be crucial to consolidate the benefits of this technological transition for Public Health.

Acknowledgements

We would like to thank to Prof. João A. Carriço and Bruno Ribeiro-Gonçalves for making the PHYLOViZ online 2.0 Beta version available for the analysis performed in the present study.

O. Oliveira is supported by the Project NORTE-08-5369-FSE-000041, financed by the Operational Program NORTE 2020 and co-financed by the European Social Fund through a doctoral grant (grant number UMINHO/BD/47/2016).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tube.2019.02.006.

References

[1] World Health Organization. Global tuberculosis report 2015. 2015http://www.who.int/iris/handle/10665/191102, Accessed date: 1 November 2018.
[2] World Health Organization. Global tuberculosis report 2018. 2018https://www.who.int/tb/publications/global_report/en/, Accessed date: 1 November 2018.
[3] Portuguese General Health Directorate. Tuberculose em Portugal: desafios e estratégias. 2018https://www.dgs.pt/paginas-de-sistema/saude-de-a-a-z/tuberculose1/relatorios.aspx, Accessed date: 1 November 2018.
[4] Portuguese General Health Directorate. Portugal: Infeção VIH, SIDA e Tuberculose em números. 2017https://www.dgs.pt/paginas-de-sistema/saude-de-a-a-z/tuberculose1/relatorios.aspx, Accessed date: 1 November 2018.
[5] van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, Hermans P, Martin C, McAdam R, Shinnick TM. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol 1993;31(2):406–9.
[6] Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol 1997;35(4):907–14.
[7] Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, Savine E, de Haas P, van Deutekom H, Roring S, Bifani P, Kurepina N, Kreiswirth B, Sola C, Rastogi N, Vatin V, Gutierrez MC, Fauville M, Niemann S, Skuce R, Kremer K, Locht C, van Soolingen D. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number Tandem repeat typing of *Mycobacterium tuberculosis*. J Clin Microbiol 2006;44(12):4498–510.
[8] Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S, Supply P, Kalinowski J, Niemann S. Whole Genome Sequencing versus Traditional Genotyping for Investigation of a *Mycobacterium tuberculosis* Outbreak: a Longitudinal Molecular Epidemiological Study. PLoS Med 2013;10(2):e1001387.
[9] Mazars E, Lesjean S, Banuls AL, Gilbert M, Vincent V, Gicquel B, Tibayrenc M, Locht C, Supply P. High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. P Natl Acad Sci USA 2001;98(4):1901–6.
[10] Oelemann MC, Diel R, Vatin V, Haas W, Rüsch-Gerdes S, Locht C, Niemann S, Supply P. Assessment of an optimized mycobacterial interspersed repetitive- unit-variable-number tandem-repeat typing system combined with spoligotyping for population-based molecular epidemiology studies of tuberculosis. J Clin Microbiol 2007;45(3):691–7.
[11] van Deutekom H, Supply P, de Haas PE, Willery E, Hoijng SP, Locht C, Coutinho RA, van Soolingen D. Molecular typing of *Mycobacterium tuberculosis* by mycobacterial interspersed repetitive unit-variable-number Tandem repeat analysis, a more accurate method for identifying epidemiological links between patients with tuberculosis. J Clin Microbiol 2005;43(9):4473–9.
[12] Rasoahanitralisoa R, Rakotosamimanana N, Stucki D, Sola C, Gagneux S, Rasolofo Razanamparany V. Evaluation of spoligotyping, SNPs and customised MIRU-VNTR combination for genotyping *Mycobacterium tuberculosis* clinical isolates in Madagascar. PLoS One 2017;12(10):e0186088.
[13] Jonsson J, Hoffner S, Berggren I, Bruchfeld J, Ghebremichael S, Pennhag A, Groenheit R. Comparison between RFLP and MIRU-VNTR Genotyping of *Mycobacterium tuberculosis* Strains Isolated in Stockholm 2009 to 2011. PLoS One 2014;9(4):e95159.
[14] Kohl TA, Diel R, Harmsen D, Rothgänger J, Walter KM, Merker M, Weniger T, Niemann S. Whole-genome-based *Mycobacterium tuberculosis* surveillance: a standardized, portable, and expandable approach. J Clin Microbiol 2014;52:2479–86.
[15] Kohl TA, Harmsen D, Rothgänger J, Walker T, Diel R, Niemann S. Harmonized genome wide typing of tubercle Bacilli using a web-based gene-by-gene nomenclature system. EBioMedicine 2018;34:131–8.
[16] Bravo LT, Tuohy MJ, Ang C, Destura RV, Mendoza M, Procop GW, Gordon SM, Hall GS, Shrestha NK. Pyrosequencing for rapid detection of *Mycobacterium tuberculosis* resistance to rifampin, isoniazid, and fluoroquinolones. J Clin Microbiol 2009;47(12):3985–90.
[17] Boehme CC, Nabeta P, Hillemann D, Nicol MP, Shenai S, Krapp F, Allen J, Tahirli R, Blakemore R, Rustomjee R, Milovic A, Jones M, O'Brien SM, Persing DH, Ruesch-Gerdes S, Gotuzzo E, Rodrigues C, Alland D, Perkins MD. Rapid molecular detection of tuberculosis and rifampin resistance. N Engl J Med 2010;363(11):1005–15.
[18] Satta G, Atzeni A, McHugh TD. *Mycobacterium tuberculosis* and whole genome sequencing: a practical guide and online tools available for the clinical microbiologist. Clin Microbiol Infect 2017;23(2):69–72.
[19] Macedo R, Nunes A, Portugal I, Duarte S, Vieira L, Gomes JP. Dissecting whole-genome sequencing-based online tools for predicting resistance in Mycobacterium tuberculosis: can we use them for clinical decision guidance? Tuberculosis 2018;110:44–51.
[20] Perdigão J, Silva H, Machado D, Macedo R, Maltez F, Silva C, Jordao L, Couto I, Mallard K, Coll F, Hill-Cawthorne GA, McNerney R, Pain A, Clark TG, Viveiros M, Portugal I. Unraveling Mycobacterium tuberculosis genomic diversity and evolution in Lisbon, Portugal, a highly drug resistant setting. BMC Genomics 2014;15:991.
[21] Brown-Elliott BA, Simmer PJ, Trovato A, Hyle EP, Droz S, Buckwalter SP, Borroni E,

Branda JA, Iana E, Mariottini A, Nelson J, Matteelli A, Toney NC, Scarparo C, de Man TJB, Vasireddy R, Gandhi RT, Wengenack NL, Cirillo DM, Wallace RJ, Tortoli E. *Mycobacterium decipiens sp. nov.*, a new species closely related to the *Mycobacterium tuberculosis* complex. Nt J Syst Evol Micr 2018;10.

[22] Cabibbe AM, Walker TM, Niemann S, Cirillo DM. Whole genome sequencing of *Mycobacterium tuberculosis*. Eur Respir J 2018;52(5):1801163.

[23] Tagliani E, Cirillo DM, Ködmön C, van der Werf MJ, Consortium EUSeqMyTB. EUSeqMyTB to set standards and build capacity for whole genome sequencing for tuberculosis in the EU. Lancet Infect Dis 2018;18(4):377.

[24] Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TE. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis 2013;13(2):137–46.

[25] Cabibbe AM, Trovato A, De Filippo MR, Ghodousi A, Rindi L, Garzelli C, Baretti S, Allodi G, Mannino R, Rossolini GM, Bartoloni A, Tortoli E, Cirillo DM. Countrywide implementation of whole genome sequencing: an opportunity to improve tuberculosis management, surveillance and contact tracing in low incidence countries. Eur Respir J 2018;51(6):1800387.

[26] Rito T, Matos C, Carvalho C, Machado H, Rodrigues G, Oliveira O, Ferreira E, Gonçalves J, Maio L, Morais C, Ramos H, Guimarães JT, Santos CL, Duarte R, Correia-Neves M. BMC Infect Dis 2018;18(1):53.

[27] Portuguese General Health Directorate. Circular Normativa Nº 1/DT de 11/01/2007. 2007https://www.dgs.pt/directrizes-da-dgs/normas-e-circulares-normativas/circular-normativa-n-1dt-de-11012007.aspx, Accessed date: 1 November 2018.

[28] Portugal I, Barreiro L, Vultos T, Macedo R, Furtado C, Fonseca-Antunes A, Brum L. Molecular epidemiology of *Mycobacterium tuberculosis* in Lisbon. Rev Port Pneumol 2008;14(2):239–59.

[29] Macedo R, Antunes AF, Villar M, Portugal I. Multidrug and extensively drug-resistant tuberculosis in Lisbon and Vale do Tejo, Portugal, from 2008 to 2010. Int J Mycobacteriol 2012;1(3):131–6.

[30] Portuguese General Health Directorate. Circular Normativa Nº 12/DSCS/PNT de 17/07/2008. 2008https://www.dgs.pt/...da...e.../circular-normativa-n-12dscspnt-de-17072008-pdf.aspx, Accessed date: 1 November 2018.

[31] Weniger T, Krawczyk J, Supply P, Niemann S, Harmsen D. MIRU-VNTR *plus*: a web tool for polyphasic genotyping of *Mycobacterium tuberculosis* complex bacteria. Nucleic Acids Res 2010;38:W326–31.

[32] Allix-Béguec C, Harmsen D, Weniger T, Supply P, Niemann S. Evaluation and user-strategy of MIRU-VNTR*plus*, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. J Clin Microbiol 2008;46(8):2692–9.

[33] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30(15):2114–20.

[34] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19(5):455–77.

[35] Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. Pilon: an Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS One 2014;9(11):e112963.

[36] Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G, Mallard K, Nair M, Miranda A, Alves A, Perdigão J, Viveiros M, Portugal I, Hasan Z, Hasan R, Glynn JR, Martin N, Pain A, Clark TG. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med 2015;7(1):51.

[37] Rajwani R, Shehzad S, Siu GKH. MIRU-profiler: a rapid tool for determination of 24-loci MIRU-VNTR profiles from assembled genomes of *Mycobacterium tuberculosis*. PeerJ 2018;6:e5090.

[38] Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, Ramirez M, Carriço JA. chewBBACA: a complete suite for gene-by-gene schema creation and strain identification. Microb Genom 2018;4(3):e000166.

[39] Francisco AP, Bugalho M, Ramirez M, Carriço JA. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinf 2009;10:152.

[40] Ribeiro-Gonçalves B, Francisco AP, Vaz C, Ramirez M, Carriço JA. PHYLOViZ Online: web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees. Nucleic Acids Res 2016;44(W1):W246–51.

[41] Carriço JA, Silva-Costa C, Melo-Cristino J, Pinto FR, Lencastre H, Almeida JS, Ramirez M. Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant *Streptococcus pyogenes*. J Clin Microbiol 2006;44(7):2524–32.

[42] Llarena A-K, Ribeiro-Gonçalves BF, Nuno Silva D, Halkilahti J, Machado MP, Da Silva MS, Jaakkonen A, Isidro J, Hämäläinen C, Joenperä J, Borges V, Viera L, Gomes JP, Correia C, Lunden J, Laukkanen-Ninios R, Fredriksson-Ahomaa M, Bikandi J, San Millan R, Martinez-Ballesteros I, Laorden L, Mäesaar M, Grantiņa-Ieviņa L, Hilbert F, Garaizar J, Oleastro M, Nevas M, Salmenlinna S, Hakkinen M, Carriço JA, Rossi M. INNUENDO: a cross-sectoral platform for the integration of genomics in the surveillance of food-borne pathogens. EFSA supporting publication; 2018. p. 142. EN-1498.

[43] Vasconcellos SE, Huard RC, Niemann S, Kremer K, Santos AR, Suffys PN, Ho JL. Distinct genotypic profiles of the two major clades of *Mycobacterium africanum*. BMC Infect Dis 2010;10:80.

[44] Mikheecheva NE, Zaychikova MV, Melerzanov AV, Danilenko VN. A nonsynonymous SNP catalog of *Mycobacterium tuberculosis* virulence genes and its use for detecting new potentially virulent sublineages. Genom Biol Evol 2017;9(4):887–9.

[45] Wilson BA, Garud NR, Feder AF, Assaf ZJ, Pennings PS. The population genetics of drug resistance evolution in natural populations of viral, bacterial and eukaryotic pathogens. Mol Ecol 2016;25(1):42–66.

[46] Schürch AC, Kremer K, Hendriks AC, Freyee B, McEvoy CR, van Crevel R, Boeree MJ, van Helden P, Warren RM, Siezen RJ, van Soolingen D. SNP/RD Typing of *Mycobacterium tuberculosis* Beijing strains reveals local and worldwide disseminated clonal complexes. PLoS One 2011;6(12):e28365.

[47] Sengstake S, Bablishvili N, Schuitema A, Bzekalava N, Abadia E, de Beer J, Tadumadze N, Akhalaia M, Tuin K, Tukvadze N, Aspindzelashvili R, Bachiyska E, Panaiotov S, Sola C, van Soolingen D, Klatser P, Anthony R, Bergval I. Optimizing multiplex SNP-based data analysis for genotyping of *Mycobacterium tuberculosis* isolates. BMC Genomics 2014;15(1):572.

[48] Chen YY, Chang JR, Huang WF, Kuo SC, Su IJ, Sun JR, Chiueh TS, Huang TS, Chen YS, Dou HY. Genetic Diversity of the *Mycobacterium tuberculosis* Beijing Family Based on SNP and VNTR Typing Profiles in Asian Countries. PLoS One 2012;7(7):e39792.

[49] Merker M, Kohl TA, Niemann S, Supply P. The evolution of strain typing in the *Mycobacterium tuberculosis* complex. Adv Exp Med Biol 2017;1019:43–78.

[50] Satta G, Lipman M, Smith GP, Arnold C, Kon OM, McHugh TD. *Mycobacterium tuberculosis* and whole-genome sequencing: how close are we to unleash its full potential? Clin Microbiol Infect 2017;24(6):604–9.

[51] Pankhurst LJ, Del Ojo Elias C, Votintseva AA, Walker TM, Cole K, Davies J, Fermont JM, Gascoyne-Binzi DM, Kohl TA, Kong C, Lemaitre N, Niemann S, Paul J, Rogers TR, Roycroft E, Smith EG, Supply P, Tang P, Wilcox MH, Wordsworth S, Wyllie D, Xu L, Crook DW. COMPASS-TB Study Group. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. Lancet Resp Med 2016;4(1):49–58.

[52] Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches. Clin Microbiol Infect 2018;24(4):350–4.

[53] Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, Gilpin B, Smith AM, Man Kam K, Perez E, Trees E, Kubota K, Takkinen J, Nielsen EM, Carleton H. FWD-NEXT Expert Panel. PulseNet International: vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveill 2017;22(23):30544.

[54] Merker M, Kohl TA, Roetzer A, Truebe L, Richter E, Rüsch-Gerdes S, Fattorini L, Oggioni MR, Cox H, Varaine F, Niemann S. Whole genome sequencing reveals complex evolution patterns of multidrug-resistant *Mycobacterium tuberculosis* Beijing strains in patients. PLoS One 2013;8(12):e82551.