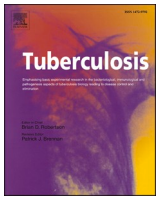
Tuberculosis 155 (2025) 102691

Contents lists available at ScienceDirect

Tuberculosis

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



Whole-genome sequencing-based surveillance system for *Mycobacterium tuberculosis* in Portugal

Miguel Pinto [a,*], Rita Macedo [b,**]

[a] *Genomics and Bioinformatics Unit, Department of Infectious Diseases, National Institute of Health, Lisbon, Portugal, Avenida Padre Cruz, 1649-016, Lisbon, Portugal*
[b] *National Reference Laboratory for Mycobacteria, Department of Infectious Diseases, National Institute of Health, Lisbon, Portugal, Avenida Padre Cruz, 1649-016, Lisbon, Portugal*

ARTICLE INFO

Keywords:
Tuberculosis
Surveillance
Whole-genome sequencing
Antimicrobial resistance
Bioinformatics
Mycobacterium tuberculosis complex

ABSTRACT

To improve TB surveillance and diagnosis, the Portuguese National Reference Laboratory (NRL) began implementing whole-genome sequencing (WGS) for all RR/MDR-TB cases in 2019. Since 2020, this approach has been expanded to indiscriminately include all received isolates. We describe the current WGS-based surveillance system in Portugal, framed in prospective and retrospective data (n = 1171), upgraded for antimicrobial resistance (AMR) prediction and epidemiological analysis. This system relies on three main steps: QC/QA and contamination assessment, with a novel data filtering step; genotyping and AMR prediction; and dynamic SNP-based approach, maximizing variable sites under analysis. While lineage 4 was the most prevalent (84.3 %) followed by lineage 2 (9.1 %), less common EU/EEA sub-lineages (e.g., lineages 3 and 6) showcased cross-border transmissions. Molecular clusters (n = 157) displayed distinct AMR profiles and diverse possible epidemiological contexts. Among the pipeline upgrades, we highlight: i) the novel filtering step that allowed the improvement of 123 out of 128 contaminated samples; ii) tolerating missing data per site more than doubled core variable site resolution; iii) automatic maximization of shared variable sites for in-depth cluster analysis, key for consolidating genetic links in epidemiological investigation. This study highlights the importance of sustained prospective genomic surveillance towards strengthening TB management and diagnosis in Portugal.

1. Introduction

Despite considerable progress made towards its control over the past few decades, tuberculosis (TB) remains a global public health challenge [1,2]. In 1993, the World Health Organization (WHO) declared TB a global health emergency [3], and, as of its latest reports, TB still accounts for an estimated 10 million new cases and 1.5 million deaths annually, underscoring the continued threat this disease poses to global health [1]. In Portugal, TB incidence remains moderate, with a reported rate of 14.9 cases per 100,000 inhabitants in 2023 [4]. While the country has made significant strides in controlling TB, the emergence of drug-resistant strains, and consequent transmission, continues to hamper its control, as also demonstrated globally. In 2021, the WHO estimated that 465,000 people worldwide were suffering from rifampicin- or multidrug-resistant TB (RR/MDR-TB), of which only 44 % were correctly diagnosed and notified [1,2]. In Portugal, the latest report from the Directorate-General of Health (DGS) and the National Institute of Health Doutor Ricardo Jorge (INSA) reported an increase of more than 100 % of the RR/MDR-TB cases in 2024 when compared with 2022 (11 vs 27 notified RR/MDR-TB cases) [4,5].

Within this challenging landscape, accurate and timely surveillance is crucial for identifying resistant strains, tracking transmission, and guiding public health interventions. In Portugal, TB surveillance involves a coordinated effort within its public health sector, between the National Reference Laboratory for Mycobacteria (NRL-TB) at INSA, responsible for laboratory-based monitoring, and the National Tuberculosis Programme (NTP) under the DGS, responsible for epidemiological surveillance and contact tracing. As the designated NRL, INSA's main role is to perform laboratory surveillance, achieved through whole-genome sequencing (WGS) on *Mycobacterium tuberculosis complex* (MTC) isolates to assess their genomic relatedness towards the identification of clusters to be flagged to the NTP-DGS as potential epidemiological links for further investigation. As the NRL does not have access to individual-level clinical, demographic, or epidemiological data, due

* Corresponding author.
** Corresponding author.
E-mail addresses: miguel.pinto@insa.min-saude.pt (M. Pinto), rita.macedo@insa.min-saude.pt (R. Macedo).

https://doi.org/10.1016/j.tube.2025.102691
Received 5 June 2025; Received in revised form 28 August 2025; Accepted 12 September 2025
Available online 12 September 2025

to data protection and governance constraints, the integration of genomic data with epidemiological context is of the responsibility of the NTP-DGS. These include standardized contact tracing procedures, structured epidemiological interviews to confirm or refute suspected transmission events and targeted interventions. The NRL-TB at the INSA has been at the forefront of TB surveillance, at the national level, receiving isolates from all TB patients with drug-resistant strains since 2004 [6–8]. The role of the NRL-TB in genotyping and resistance profiling has been vital in the ongoing effort to contain TB, particularly in identifying RR/MDR-TB clusters and tracking of genetic links between cases. In recent years, the NRL-TB has taken significant steps to enhance its surveillance capabilities through the integration of WGS [5,6,8,9]. This approach has emerged as the gold standard for TB genotyping due to its superior resolution and comprehensive ability to assess not only genetic relatedness but also antimicrobial resistance (AMR) profiles. This technology allows for more accurate tracing of infection sources, transmission routes, and the identification of previously undetected genetic variants associated with resistance. WGS also facilitates more detailed epidemiological analysis, making it a powerful tool for public health surveillance and outbreak investigation.

Since 2019, the NRL-TB has implemented WGS-based molecular surveillance for all RR/MDR-TB cases in Portugal [6–8], expanding this approach to include all MTC isolates received by the laboratory since 2020 [9]. This expansion has been accompanied by a series of technological advancements and upgrades aimed at enhancing the accuracy and utility of WGS-based surveillance. Here, we aim to present an overview of the NRL's WGS-based surveillance system for TB in Portugal, highlighting the key upgrades and innovations that have been implemented in recent years. We describe the steps involved in the WGS workflow, from quality assessment to AMR prediction and epidemiological analysis. Additionally, we present a detailed analysis of the system's impact on TB surveillance in Portugal, framed by retrospective and prospective data, and discuss the implications of these advancements for the future of TB control in Portugal. Ultimately, the integration of WGS into national TB surveillance represents a significant step in our capacity to monitor, understand, and fight TB, particularly in the face of increasing AMR.

2. Methods

2.1. Sample collection and whole-genome sequencing

All MTC strains described in the present study were received at the Portuguese NRL-TB (n = 1171), either as biological samples or MTC isolates. The dataset includes: all MDR-TB isolates collected from 2013 up to 2024, all MTC isolates received from 2020 up to 2024 for TB diagnosis confirmation, isolates analysed in the context of research projects and isolates received for potential outbreak or direct transmission investigation. For surveillance purposes, we only include in the laboratory genomic dataset one isolate per patient per TB episode (i.e., isolates from a same patient must have been collected more than six months apart). Minimum metadata for each isolate was routinely collected, namely associated patient age, gender, region of residence (when not available the Portuguese Health region is used) and date of diagnosis.

Biological samples (both respiratory and non-respiratory samples, such as sputum., BAL, urine, CSF, biopsies, etc.) were decontaminated using the N-acetyl-cisteine-NaOH BD BBL™ MycoPrep™ kit according to manufacturer's instructions (Becton Dickinson Company, East Rutherford, NJ, USA; latest version from March 2024) and used as inoculum for isolation in culture, and MTC isolates were re-cultured (usually only once) prior to downstream procedures. DNA extraction was performed using the QIAamp DNA Mini Kit (Qiagen, Düsseldorf, Germany) after heat inactivation at 95 °C for 30 min, according to the manufacturer's instructions (DNA extraction from tissues protocol). Samples were quantified using the Qubit 4 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) with the Qubit dsDNA High Sensitivity Assay Kit, following the manufacturer's instructions. All DNA samples were subjected to dual-indexed Nextera XT Illumina library preparation, cluster generation and paired-end sequencing (either 2 × 250bp or 2x150bp) on an Illumina MiSeq, NextSeq 550 or NextSeq 2000 apparatus (Illumina, SD, USA), available at the INSA.

2.2. Bioinformatics workflow

All MTC isolates described in the present study were routinely subjected to the following bioinformatics workflow implemented at the NRL-TB.

WGS data quality control and contamination assessment was performed using the INNUca v4.2.2 pipeline (described in detail at https://github.com/B-UMMI/INNUca). Briefly, reads' quality analysis (FastQC v0.11.5 - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and cleaning/improvement (using Trimmomatic v0.38 [10] and based on per sample FastQC analysis) is carried out prior to draft genome assembly (SPAdes v3.13.0 [11]) and assembly improvement (Pilon v1.23 [12]). Species taxonomic classification is performed using Kraken v2 [13], both in raw reads and final polished assemblies, using the Kraken Standard-16 DB/2023-03-13 database (available at https://benlangmead.github.io/aws-indexes/k2). Per sample comprehensive summary reports on genome statistics in every step of the pipeline are then provided at the end of the analysis. Samples are excluded when the first draft or the final polished assembly size falls below 4 Mpb or when genome estimated depth of coverage is below 15-fold. Genomes are flagged as contaminated when the first draft assembly size is higher than 4.62 Mpb, the majority of raw reads classified by Kraken do not belong to *M. tuberculosis*, or the final polished assembly contigs are mostly classified as anything other than *M. tuberculosis*. As such, whenever a contamination was detected, raw read filtering was performed by retrieving the raw reads classified by Kraken v2 as belonging to the *Mycobacterium* Genus, using the *seqtk* tool (https://github.com/lh3/seqtk), and performing a novel round of assembly with INNUca for reassessment of contamination and sequencing quality. Of note, whenever a contamination by a nontuberculous mycobacteria (NTM) is detected, the sample is excluded, due to the high genetic homology between NTM and MTC leading to potential false positive results.

In silico drug susceptibility testing and genotyping was performed directly on raw read data (or raw reads filtered with Kraken as described above) using tb-profiler v6.3.0 (*tbdb* database ID:72ef6fa) [14]. AMR data is retrieved based on the WHO mutations catalogue [15] for rifampicin, isoniazid, ethambutol, pyrazinamide, moxifloxacin, levofloxacin, bedaquiline, delamanid, pretomanid, linezolid, streptomycin, amikacin, kanamycin, capreomycin, clofazimine, ethionamide, para-aminosalicylic acid and cycloserine.

For SNP-based analysis, quality improved read data (after Trimmomatic processing) were individually mapped against the H37Rv reference genome (RefSeq accession #NC_000962.3) using Snippy v4.5.1 (https://github.com/tseemann/snippy), with SNP calling performed on variant sites with the following criteria: minimum proportion of reads differing from the reference of 70 %; minimum mapping quality of 30; minimum base quality of 20; and minimum coverage for SNP calling of 10. Full genome alignment was then extracted using Snippy's core module (*snippy-core*), with core-SNV falling in known high GC content regions, repetitive elements, and resistance-associated positions (corresponding to 8.5 % of the reference genome) excluded as previously described [6,16].

Finally, for phylogenetic and clustering analysis, Reportree v2 [17] was used over the full-alignment obtained from snippy (alongside sample metadata) using the following criteria: minimum proportion of ATCG content in informative sites of the alignment per sample of 95 %; minimum proportion of samples per site without missing data of 95 %; minimum spanning tree construction using the Grapetree MSTv2 algorithm [18]; and genetic cluster characterization performed at SNP

distance thresholds of 6, 12, 18 and 24. Exceptionally, for very large clusters, we also take advantage of the dynamic zoom-in option of Reportree to recalculate SNP distances based only on the sub-set of isolates that compose the cluster, ensuring maximum analysis resolution for these (without missing data per site and using Grapetree MST algorithm). For reporting purposes, isolates clustered at a SNP distance threshold of 12 are flagged for epidemiological investigation as potentially linked [6,16,19,20], and cluster characterization is provided regarding patient associated metadata, extended AMR profiles and genotyping data. Of note, isolates additionally clustered at a 6 or 18 SNP distance are conditionally provided as supporting information for epidemiological investigation purposes.

2.3. Data availability

All raw sequence reads were deposited in the European Nucleotide Archive (ENA) under the study accession number PRJEB29446. Of note, this BioProject is routinely populated and updated on behalf of the WGS-based surveillance activities of the NRL-TB in Portugal. All data for isolates used in the present study, including genome statistics, ENA accession numbers, antimicrobial resistance profiles, genotyping and demographic data are summarized in Supplementary Table S1.

3. Results

The implemented WGS-based laboratory surveillance workflow of the Portuguese NRL-TB is summarized in Fig. 1. Upon QC/QA analysis, MTC genomes are excluded following criteria based on mean depth of coverage, genome size and raw read/assembly contigs taxonomic classifications (see details in Methods Section). In order to maximize sequencing data and therefore reduce sequencing effort costs (i.e., reduce the number of samples that would be subjected again to culture and WGS), a raw read data filtering step, based on taxonomic classification, was introduced in the workflow to recover contaminated samples. The inclusion of this filtering step successfully removed the detected contamination for 123 out of 128 flagged samples, of which only 20 lacked enough MTC genomic data to proceed for downstream analysis and thus considered for re-sequencing efforts (Supplementary Table S2). The five remaining samples that could not be filtered were presumably mixed MTC infections. Data filtering is highly successful for cases of co-infections (see for example PT_TB1002 and PT_TB0631 in Supplementary Table S2) and residual carry-over of host material (see for example PT_TB0976 and PT_TB0987 in Supplementary Table S2). More importantly, data showed that the filtering step did not hamper the quality or breadth of coverage of the samples (mean percentage of breadth of coverage of non-filtered versus filtered samples of 98.71 % and 98.66 %, respectively) and in nine cases was able to discard the detection of false positive low frequency AMR markers that were present due to contamination (see "AMR Note" column in Supplementary Table S1). Sample exclusion was also decreased by applying an exclusion criteria based on the percentage of ATCG content over informative sites of the alignment (per sample, through Reportree), instead of

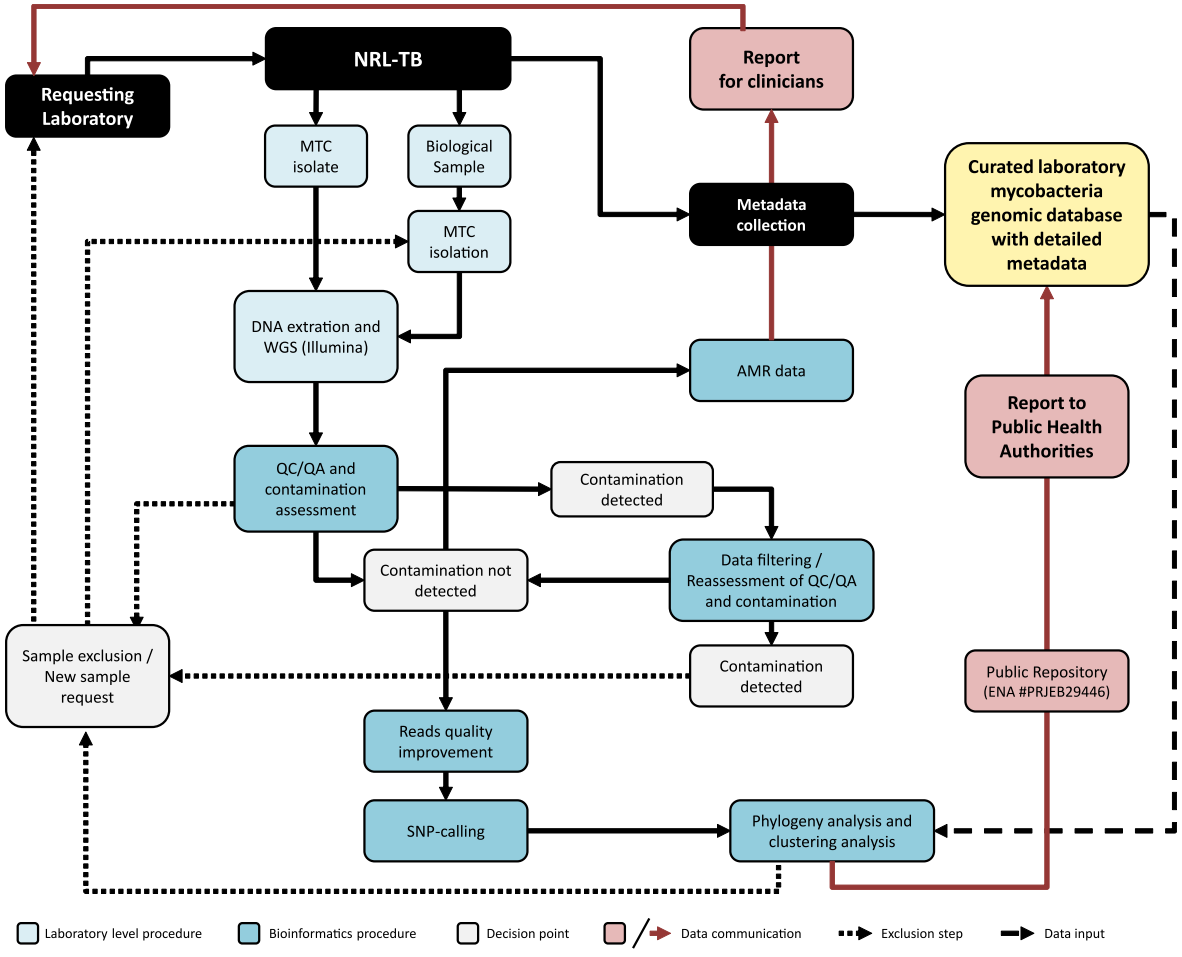


Fig. 1. | Workflow of the whole-genome sequencing-based surveillance and diagnosis support system of the Portuguese National Reference Laboratory for Mycobacteria. NRL-TB - National Reference Laboratory for Mycobacteria; MTC – *Mycobacterium tuberculosis complex*; WGS – Whole-genome sequencing; AMR – Antimicrobial resistance; QC/QA – Quality control/Quality Assessment; SNP – Single nucleotide polymorphism. ENA – European Nucleotide Archive. Note: "Reads quality improvement" refers to read trimming steps.

excluding based solely on breadth of coverage over the reference genome [e.g. below 95 %, excluding masked regions] as previously described [6]. Here, by focusing solely on informative sites, samples with overall lower breadth of coverage that affect non-informative or conserved genomic regions can still be included in the clustering analysis (this occurred for six samples within the dataset). Overall, the above-mentioned modifications in the workflow allowed the recovery of 108 samples (Supplementary Table S1), corresponding to 9.2 % of the genomic dataset, which positively impacted laboratory/sequencing efforts and costs and reduced time-to-results.

Regarding the SNP-based analysis, the implemented workflow takes advantage of Reportree to tolerate 5 % missing data per site in the full genome alignment of the dataset [17]. Data shows that using a pure core-SNP approach (i.e., only considering sites shared at 100 % between all samples) highly reduces resolution power and hampers the clustering analysis (Supplementary Fig. S1). In fact, in our dataset, of the total of 44862 informative sites that could be considered for isolate discrimination, only 23058 are included in the analysis when a pure core-SNP approach is used, compared to the 44574 sites (99.4 % of the total) that are analysed when 5 % missing data per site is applied. As such, by taking advantage of both the Grapetree MSTv2 algorithm for clustering (that ponders missing data instead of considering it as a SNP difference [18]) and the inclusion of missing data per site with Reportree, a close to maximum resolution power can be achieved during a single clustering step, providing a high-resolution overview of the genetic relatedness of all samples within the dataset. As exemplified in Supplementary Fig. S1, at the same 12 SNP distance threshold, the increase in cluster discrimination is evident when comparing the inclusion or exclusion of missing data per informative site. Additionally, as the genomic dataset increases due to routine surveillance activities, this approach reduces the need to perform re-clustering of a specific subset of isolates, i.e., performing a dynamic SNP analysis by maximizing the number of sites under analysis for a specific subset (e.g. per lineage or sub-lineage), being potentially more informative only when large clusters of isolates are present. To our knowledge, no other software apart from Reportree allows both tolerating missing data for higher resolution in the first step clustering and the ability to automatically perform re-clustering on a given subset of samples. Moreover, stable cluster nomenclatures can also be included for each identified clusters, with detailed statistical metadata reports per cluster being outputted by Reportree along with changes in cluster composition due to the addition of novel samples to the analysis during routine surveillance activities (i.e., reports on clusters increasing in size,

Table 1
| Characterization of the *Mycobacterium tuberculosis complex* genomic dataset.

	Diagnostic year													Total
	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	
Gender														
Male	3	9	74	22	22	22	26	37	15	111	128	160	186	815
Female	1	4	35	3	7	10	15	12	10	46	65	69	71	348
Unknown	0	0	0	0	0	0	0	0	0	3	0	2	3	8
Age group														
[0–5[	0	0	0	0	0	2	1	0	0	0	0	1	1	5
[5–10[	0	0	0	0	0	0	0	0	0	0	0	1	1	2
[10–15[	0	0	0	0	0	0	2	1	0	0	0	1	2	6
[15–20[	0	1	1	0	1	0	1	3	0	1	2	5	9	24
[20–25[	0	2	6	1	4	1	2	1	2	9	8	11	22	69
[25–30[	1	1	4	2	1	1	1	4	1	18	12	21	28	95
[30–35[	0	2	6	1	0	3	2	4	1	10	17	16	22	84
[35–40[	0	1	12	3	2	2	3	3	1	11	10	18	27	93
[40–45[	0	3	20	5	7	1	6	5	6	10	17	21	18	119
[45–50[	1	1	19	0	3	1	1	6	2	20	27	28	19	128
[50–55[	0	2	12	4	5	4	2	4	2	18	18	19	13	103
[55–60[	0	0	7	4	1	6	4	3	5	12	19	14	18	93
[60–65[	1	0	5	2	3	2	4	0	0	10	21	24	16	88
[65–70[	0	0	7	0	1	0	2	2	0	6	11	14	13	56
[70–75[	0	0	0	0	1	1	2	1	2	8	11	11	13	50
[75–80[	0	0	3	1	0	0	0	3	0	7	5	7	7	33
[80–85[	0	0	4	0	0	0	2	1	0	5	5	6	7	30
[85−100]	0	0	1	0	0	2	2	1	0	6	5	5	10	32
Unknown	1	0	2	2	0	6	4	7	3	9	5	8	14	61
Health Region														
LTV	0	9	97	13	14	12	17	22	15	119	156	168	205	847
North	4	4	10	10	11	19	24	23	7	15	11	20	18	176
Centre	0	0	2	0	4	0	0	2	0	6	7	6	15	42
Alentejo	0	0	0	0	0	0	0	0	0	10	11	20	10	51
Algarve	0	0	0	1	0	0	0	1	3	1	2	7	6	21
Azores-AR	0	0	0	0	0	0	0	0	0	8	6	10	5	29
Madeira-AR	0	0	0	1	0	1	0	1	0	1	0	0	1	5
Resistance Profile														
XDR-TB	0	0	0	1	0	1	0	0	0	0	0	0	0	2
Pre-XDR-TB	0	5	4	2	3	2	11	1	5	4	1	2	6	46
MDR-TB	0	8	18	16	14	8	8	6	11	8	7	16	18	138
RR-TB	0	0	0	0	0	0	0	4	0	1	2	4	3	14
HR-TB	0	0	8	0	1	6	0	4	5	10	20	14	27	95
Other	0	0	9	0	1	0	0	3	0	13	15	13	22	76
Sensitive	4	0	70	6	10	15	22	31	4	124	148	182	184	800

LTV – Lisbon and Tagus Valley; AR – Autonomous region; TB – tuberculosis; XDR– Extensive drug resistant; MDR – Multidrug resistant tuberculosis; RR - Rifampicin resistant; HR – Isoniazid resistant.

being kept the same or even merging; for details see Ref. [17]). In turn, this renders both the analytical and reporting activities of routine surveillance faster, harmonized and more reliable, due to the low computational demands of Reportree, the inclusion of cluster nomenclatures and the higher resolution power during clustering, respectively.

The above described implemented WGS-based surveillance system is here presented in light of data from the last five years (n = 869) framed with retrospective data (n = 302). Per year distribution of associated patient demographic data and respective isolate resistance profile data of the WGS dataset are presented in Table 1. The majority of the isolates were collected from male patients (69.6 %), and from patients aged between 40 and 55 years old (29.9 %). Noteworthy, from 2017 onwards we observed several cases (1.1 %) of paediatric TB (<15 years old), and 17.2 % of cases were associated with elderly patients (>65 years old). Patients' region of Health was distributed as follows: 72.3 % from the Lisbon and Tagus Valley (LTV), 15.0 % from the North, 3.6 % from the Center, 4.4 % from the Alentejo, 1.8 % from the Algarve, 2.5 % from the Azores Autonomous Region, and 0.4 % from the Madeira Autonomous Region. Regarding isolate AMR profiles, 68.3 % were fully susceptible, 17.1 % were RR/MDR-TB cases (of which 23.0 % and 1.0 % were Pre-XDR and XDR, respectively), and 8.1 % were HR-TB.

Fig. 2 shows the global phylogenetic distribution of the dataset by lineage. As expected, the most prevalent lineage is lineage 4 (84.3 %), particularly sub-lineages 4.3.4.2 (18.7 %), 4.3.4.1 (15.3 %), and 4.1.2.1 (11.3 %), followed by lineage 2 and its sub-lineages (9.1 %). In 2022 and 2024, two cases of TB caused by *M. bovis* BCG were detected. While strains from lineages 4.1 and 4.3 circulate in all regions of the country, lineages 1.1, 3, 4.2, 4.7, 4.8, and 4.9 are only detected in mainland Portugal (Supplementary Table S1). Notably, lineages 4.6 and 6 (*M. tuberculosis* spp. *africanum*) were only detected in the LTV region. While we observe RR/MDR-TB (n = 200, including XDR and Pre-XDR) isolates distributed in 21 distinct sub-lineages (Fig. 3), most cases belong to sub-lineages 4.3.4.2 (31.0 %), 2.2.1 (18.5 %) and 4.3.4.1 (14.0 %) (Supplementary Table S1). Of note, when analysing MDR-TB isolates only per definition (i.e., excluding XDR and Pre-XDR; n = 138), these are mainly from sub-lineage 2.2.1 (24.6 %), followed by 4.3.4.2 (21.0 %). The majority of Pre-XDR-TB isolates belong to lineage 4.3 (38 out of 46; 82.6 %), particularly sub-lineages 4.3.4.2 (58.7 %) and 4.3.4.1 (19.6 %), and all XDR-TB belong to sub-lineage 4.3.4.2.

For comparison purposes, the MTBseq pipeline [16] was applied (with default settings) on the complete dataset, and clusters were determined at a threshold of 12 SNP for lineage subsets (see Supplementary Table S3). Cluster congruence analysis between our workflow and MTBseq (using the *comparing_partitions.py* script [21]) showed that overall cluster congruence is high, with Adjusted Rand Index equal to 0.848 and Adjusted Wallace Coefficient equal to 0.948 (CI; 0.927–0.969). As such, although both approaches rely on different data quality, SNP calling and clustering software and parameters, cluster composition are very similar (Supplementary Table S3).

With the aim of promoting epidemiological investigations, molecular clusters to be flagged as potentially linked were thus defined in our workflow at a 12 SNP distance threshold [6,16,19,20]. As such, within this dataset, we identified 75 molecular clusters composed of more than two isolates (Fig. 4) and 82 clusters with only two isolates (Supplementary Fig. S2). Among the clusters composed of more than two isolates (Fig. 4), although some clusters with fewer cases are very recent (for example clusters 5, 25, 26, 35, 54, 55, 64, from 2023 or 2024), there are clusters that date back to 2012 (cluster 65), 2013

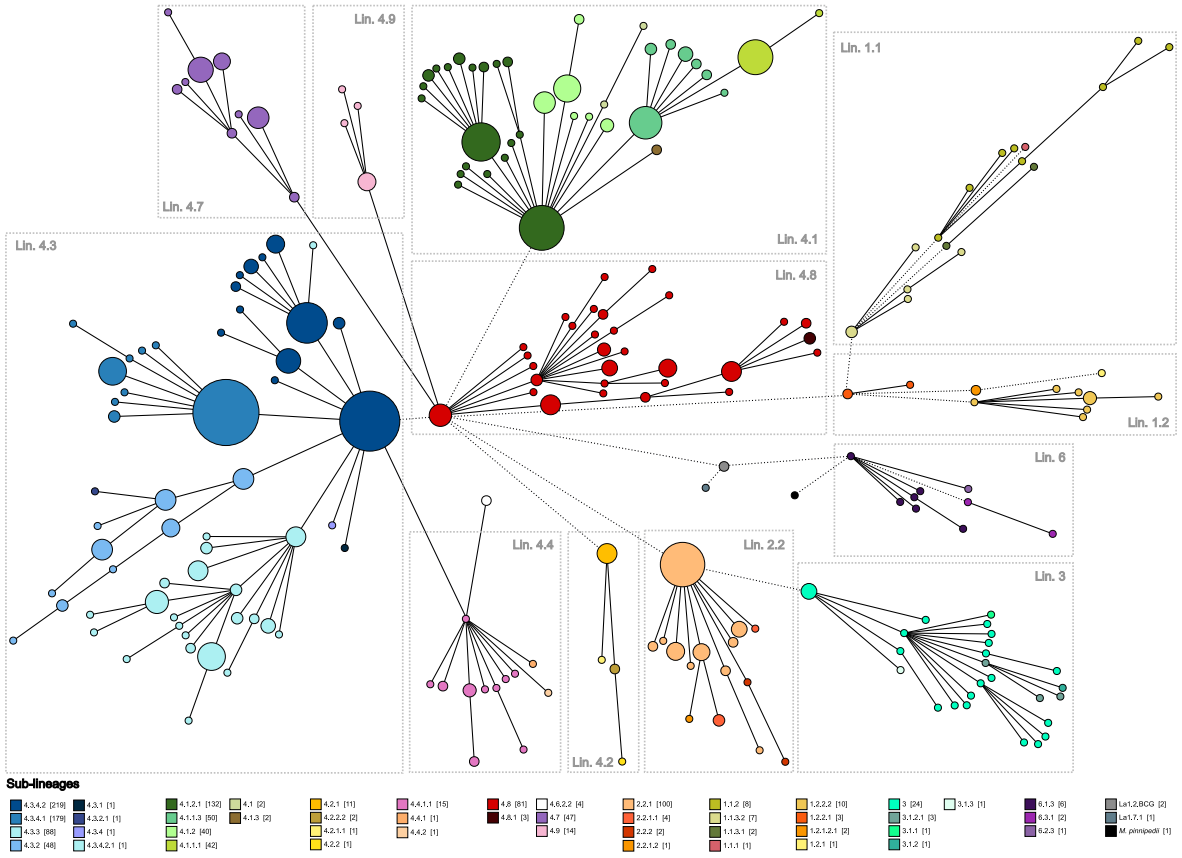


Fig. 2. | Global phylogenetic distribution by lineage of all MTC isolates enrolled in the Portuguese genomic surveillance dataset. Minimum spanning tree was constructed using Grapetree MSTv2 algorithm via Reportree [17,18]. Nodes, corresponding to different isolates, are coloured according to sub-lineage. Nodes have been collapsed at a 125 SNP distance threshold for simplified visualization purposes. Dashed lines refer to SNP distances above 500. Dashed boxes highlight main MTC lineages.

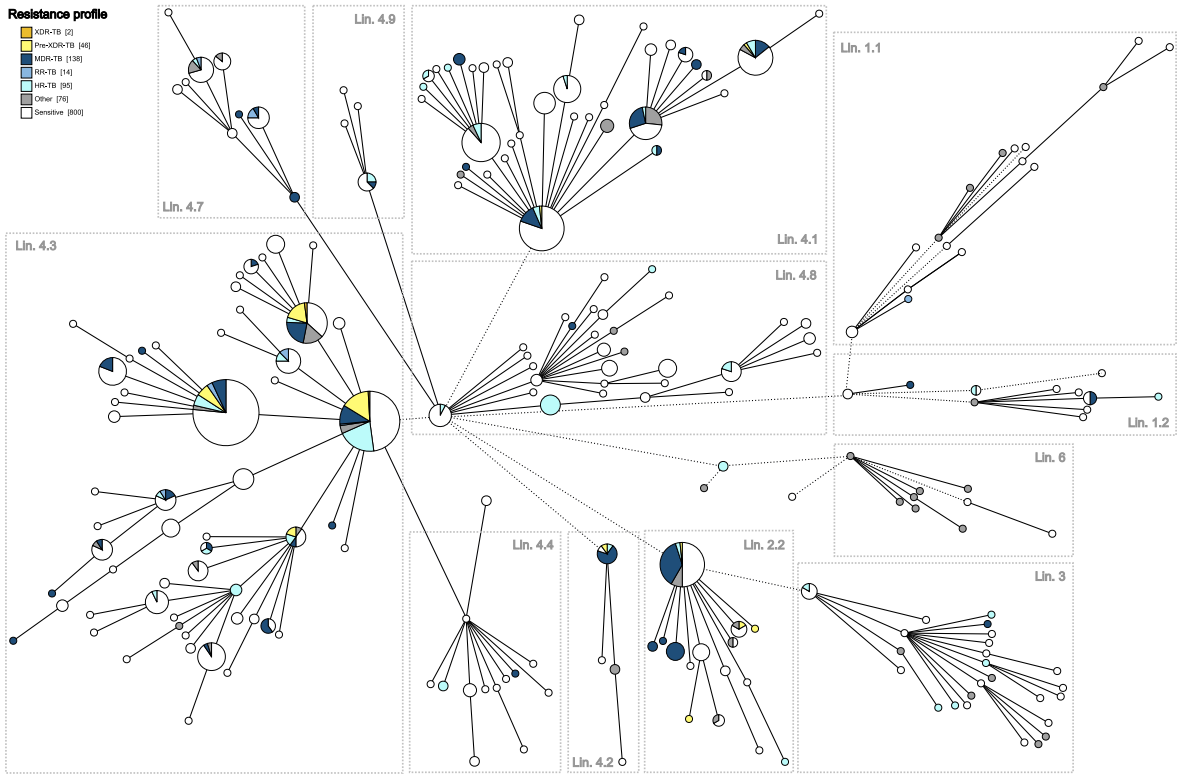


Fig. 3. | Global phylogenetic distribution by AMR profile of all MTC isolates enrolled in the Portuguese genomic surveillance dataset. Minimum spanning tree was constructed using Grapetree MSTv2 algorithm via Reportree [17,18]. Nodes, corresponding to different isolates, are coloured according to AMR profile. Nodes have been collapsed at a 125 SNP distance threshold for simplified visualization purposes. Dashed lines refer to SNP distances above 500. Dashed boxes highlight main MTC lineages. TB – tuberculosis; XDR– Extensive drug resistant; MDR – Multidrug resistant tuberculosis; RR - Rifampicin resistant; HR – Isoniazid resistant.

(clusters 13, 19, 32, 33 and 36) or 2014 (such as clusters 71 and 75, which include the highest number of cases), revealing a wide diversity of possible epidemiological contexts over the years (time span between the first and last case ranging from 0 days - cluster 11 - to 4188 days - cluster 19, Fig. 4). Most of these clusters (67 out of 75) are composed of strains with identical resistance profiles, of which 11 are RR/MDR-TB and 46 are fully susceptible. Of the total 157 identified clusters, 60 have isolates detected in 2024, suggesting that these may be continuously circulating or active transmission chains, which include: i) the cluster with the largest timespan (4188 days), with only MDR-TB strains (cluster 19); ii) the clusters with more isolates (cluster 71, n = 21; and cluster 75, n = 29), with timespans of more than 3500 days, mostly composed of susceptible strains; iii) ten clusters composed only of RR/MDR-TB strains, ranging from two to 12 isolates, with a timespan between 79 and 4188 days. Of note, eight clusters are composed by isolates with distinct AMR profiles throughout time, without evident emergence of resistance, which may render molecular epidemiology linkage alone difficult to assess with this approach, i.e., without a complete reconstruction of the potential transmission events between patients.

4. Discussion

The lineage distribution of MTC in Portugal is consistent with trends observed across Western Europe, where lineage 4 (Euro-American) is predominant [22]. In our study, 84.3 % of isolates belonged to this lineage, particularly sub-lineages 4.3.4.2, 4.3.4.1, and 4.1.2.1, reflecting established endemic transmission chains [22,23]. Countries such as Spain, Germany, and France similarly report lineage 4 as the dominant genotype, although local variation in sub-lineages may exist [22,24]. The presence of lineage 2 (East Asian, including the Beijing genotype), which represented 9.1 % of our dataset and is strongly associated with drug resistance, points to importation events, aligning with findings from the United Kingdom, the Netherlands, and Italy, where immigration from high-burden countries has increased the genomic heterogeneity of circulating strains [25]. The occasional detection of *M. bovis* BCG strains (notably in 2022 and 2024) is rare and often linked to paediatric or immunocompromised patients, a pattern similarly described in case reports from other EU/EEA countries [26,27]. This observed MTC genetic diversity, spanning multiple lineages, underscores a complex epidemiological landscape. In contrast to low-incidence countries like Sweden, where TB is largely associated with specific immigrant groups and local clustering is limited [28], Portugal seems to exhibit both sustained local transmission and evidence of repeated introductions. The detection of sub-lineages less common in the EU/EEA further emphasizes the country's unique position as a hub of TB diversity. The identification of lineage 2.2.1, particularly overrepresented by MDR- and Pre-XDR-TB cases, reflects the global spread of the Beijing strain, commonly associated with high TB burden regions in Central and East Asia [25]. Likewise, the detection of lineages 1.1 and 3, typically associated with Southeast Asia and the Indian subcontinent, aligns with also known migration patterns [29]. The exclusive presence of lineages 4.6 and 6 (*M. africanum*) in the LTV region strongly suggests importation from West Africa, where these lineages are endemic [30]. Regarding RR/MDR-TB cases, most XDR-TB cases are associated with sub-lineage 4.3.4.2, indicating persistent and potentially clonal transmission chains. Compared to other European countries, Portugal reports higher proportions of MDR- and XDR-TB, in particular in the Western European Region, where MDR-TB cases comprises <5 % of total TB cases and XDR-TB remains scarce. However, Portugal's rates are still substantially lower than high-burden Eastern European countries like Moldova, Ukraine, or Russia, where MDR-TB rates can exceed 30 %, and XDR-TB is more widespread [1,2]. Overall, these findings reinforce the role of international mobility and historical ties in shaping TB epidemiology in Portugal and underscore the need for

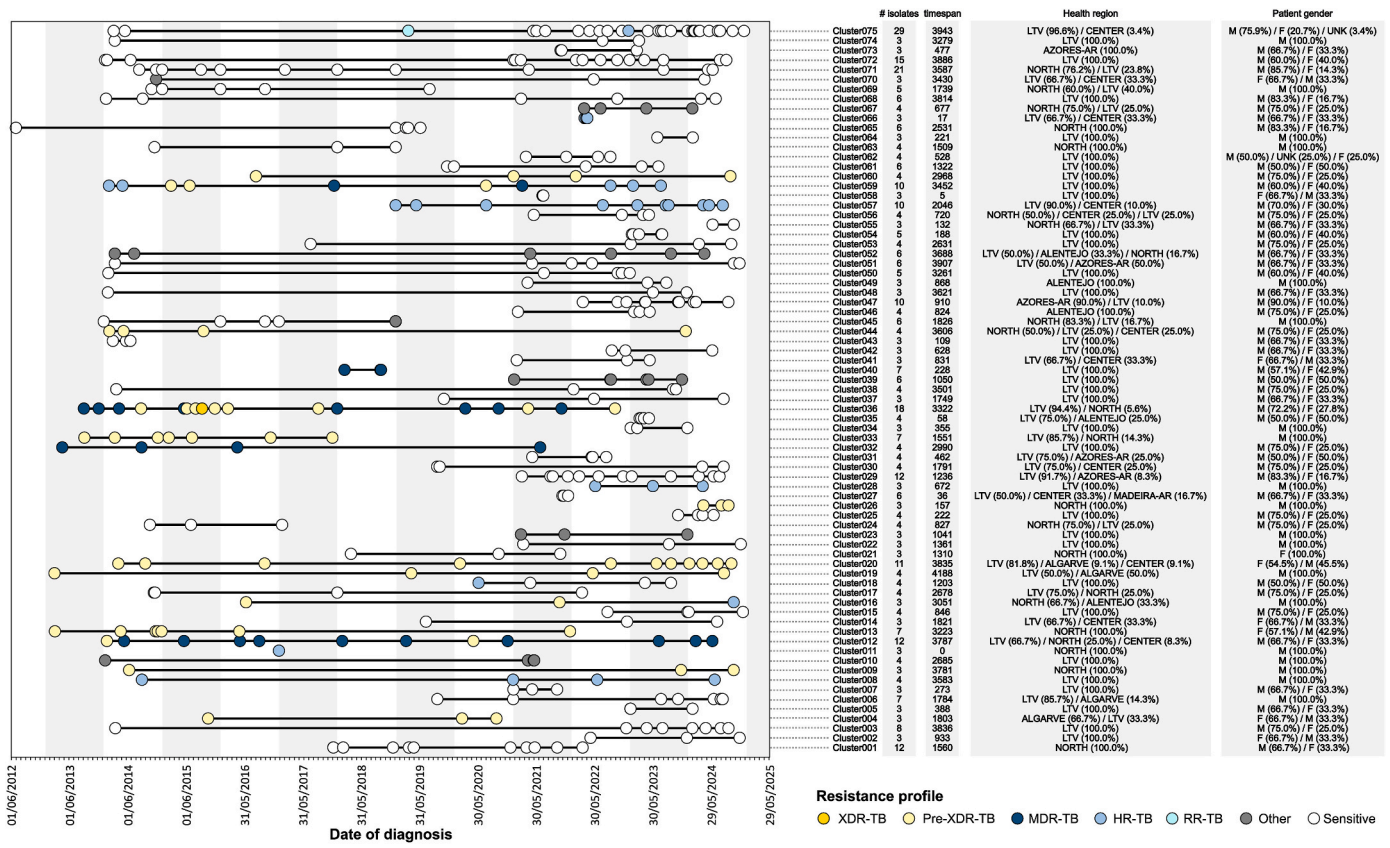


Fig. 4. | Temporal distribution of the molecular clusters with potential epidemiological linkage, composed by more than two isolates. Each node in the left panel corresponds to one MTC isolate. Right panel shows metadata associated with each cluster. LTV – Lisbon and Tagus Valley; AR – Autonomous region; TB – tuberculosis; XDR– Extensive drug resistant; MDR – Multidrug resistant tuberculosis; RR - Rifampicin resistant; HR – Isoniazid resistant.

integrated genomic and epidemiological surveillance that can track transnational transmission and resistance trends.

The present study also highlights the advantages of the upgrades implemented in our surveillance workflow (e.g., data filtering, tolerating missing data in a SNP-based approach, dynamically analysing clusters and comprehensive reporting) which have positively impacted laboratory/sequencing efforts and costs and reduced time-to-results. As many isolates are directly sent to the NRL-TB for WGS and control over "proper isolation" cannot be guaranteed *a priori*, often leading to potential contaminated samples, filtering WGS data has allowed the NRL to still timely provide quality data to public health authorities, while also reducing false positive in AMR detection. Moreover, the inclusion of Reportree software in the bioinformatics pipeline has significantly enhanced resolution power and consequently cluster detection, compared to traditional genotyping or fixed SNP approaches still used in many European settings [16]. Sequencing all MTC isolates received at the NRL (routinely performed since 2020) has extended our overview of potential epidemiologically linked molecular clusters by providing more genetic context, as shown by the number of clusters identified to date that have increased due to the inclusion of fully susceptible isolates (Fig. 3 and Supplementary Fig. S2). More importantly, the inclusion of susceptible isolates can serve as potential early warnings for the emergence of AMR isolates in our country. Additionally, the NRL-TB makes all data readily available, namely all MTC AMR data is provided to clinicians for action in treatment regimen adjustment, the ENA Bioproject is constantly updated with all new WGS raw data published routinely and identified clusters are communicated to promote epidemiological investigations (Fig. 1). Taking advantage of the Reportree cluster characterization outputs at specific thresholds [17], user-friendly reports on identified cluster data for epidemiological purposes (for NTP-DGS) and for diagnosis (for clinicians) are produced for communication, reducing the data-recipient's needed know-how to perform the laborious phylogenetic and clustering analysis with metadata integration. Besides generating flags for public health actions, the implemented WGS analysis workflow has allowed to successfully participate in External Quality Control Assessment exercises promoted by the ECDC for capacitation of the NRL-TB at the European Level. Despite all this, although the described workflow is implemented in our country, contributing towards the acceleration of the diagnosis and epidemiological surveillance of TB, significant efforts are still needed to achieve national TB surveillance coverage as MTC sample processing is decentralized which may still hamper diagnosis and Public Health actions [5,9]. In fact, we suggest that mandatory MTC isolate reporting to the NRL (i.e., systematic dispatching of all MTC isolates for WGS) would be key to achieve a more accurate nationwide overview of TB genomic epidemiology and a cost-effective AMR screening towards eventually replacing decentralized phenotypic testing. Still, an effective management and control of TB can only be achieved through the efficient data communication and coordination between the NRL-TB and the NTP-DGS. In addition, there is still no integration of TB data between the human, animal and environmental health sectors which would expand TB surveillance towards a OneHealth perspective.

In summary, this study underscores the diverse TB epidemiology in Portugal, highlighting the importance of sustained prospective genomic surveillance, improved data integration, and centralized reporting to strengthen national TB control efforts and monitor/manage drug resistant isolates.

CRediT authorship contribution statement

Miguel Pinto: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation,

Conceptualization. **Rita Macedo:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Formal analysis, Data curation, Conceptualization.

Funding

This study was co-funded by the project "Sustainable use and integration of enhanced infrastructure into routine genome-based surveillance and outbreak investigation activities in Portugal" - GENEO (Project no. 101113460) on behalf of the EU4H programme (EU4H-2022-DGA-MS-IBA-01-02).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to extend their gratitude to the clinicians and laboratories submitting TB samples and isolates to the NRL-TB in Portugal. They would also like to kindly acknowledge the Core Sequencing Facility at INSA lead by Luis Vieira, responsible for all NGS wet-lab activities, the laboratory staff of the NRL-TB, and Vítor Borges and Verónica Mixão of the Genomics and Bioinformatics Unit for the development of Reportree software.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tube.2025.102691.

References

[1] World Health Organization. Global tuberculosis report 2024. Geneva: World Health Organization; 2024. Licence: CC BY-NC-SA 3.0 IGO.
[2] European Centre for Disease Prevention and Control, WHO Regional Office for Europe. Tuberculosis surveillance and monitoring in Europe 2024 – 2022 data. Copenhagen: WHO Regional Office for Europe and Stockholm: European Centre for Disease Prevention and Control; 2024. Licence: CC BY 3.0 IGO.
[3] World Health Organization. Tuberculosis: a global emergency. World Health 1993;46(4):3–31.
[4] Directorate-General of Health. Relatório de Vigilância e Monitorização da Tuberculose em Portugal. Lisboa: Direção-Geral da Saúde; 2023. https://www.dgs.pt/em-destaque/tuberculose-em-portugal-mantem-se-estavel-pdf.aspx.
[5] National Reference Laboratory for Mycobacteria. Vigilância Laboratorial da Tuberculose em Portugal: relatório 2024. Instituto Nacional de Saúde Dr. Ricardo Jorge; 2025. http://hdl.handle.net/10400.18/10456.
[6] Macedo R, Pinto M, Borges V, Nunes A, Oliveira O, Portugal I, et al. Evaluation of a gene-by-gene approach for prospective whole-genome sequencing-based surveillance of multidrug resistant Mycobacterium tuberculosis. Tuberculosis 2019;115:81–8.
[7] Macedo R, Duarte R. Trends of multidrug-resistant tuberculosis clustering in Portugal. ERJ Open Res 2019;5(1):151–2018.
[8] 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 11 we use them for clinical decision guidance? Tuberculosis 2018;110:44–51.
[9] National Reference Laboratory for Mycobacteria. Vigilância Laboratorial da Tuberculose em Portugal: relatório 2020-2022. Instituto Nacional de Saúde Dr. Ricardo Jorge; 2023. http://hdl.handle.net/10400.18/8608.
[10] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–20.
[11] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19:455–77.
[12] Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 2014;9:e112963.
[13] Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol 2019;20(1):257.
[14] Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. Genome Med 2019;11(1):41.
[15] World Heatlh Organization. Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance. second ed. Geneva: World Health Organization; 2023. Licence: CC BY-NC-SA 3.0 IGO.
[16] Kohl TA, Utpatel C, Schleusener V, De Filippo MR, Beckert P, Cirillo DM, et al. MTBseq: a comprehensive pipeline for whole genome sequence analysis of Mycobacterium tuberculosis complex isolates. PeerJ 2018;6:e5895.
[17] Mixão V, Pinto M, Sobral D, Di Pasquale A, Gomes JP, Borges V. ReporTree: a surveillance-oriented tool to strengthen the linkage between pathogen genetic clusters and epidemiological data. Genome Med 2023;15(1):43.
[18] Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, et al. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Res 2018;28:1395–404.
[19] Kohl TA, Diel R, Harmsen D, Rothgänger J, Walter KM, Merker M, et al. Whole-genome-based Mycobacterium tuberculosis surveillance: a standardized, portable, and expandable approach. J Clin Microbiol 2014;52(7):2479–86.
[20] Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis 2013;13(2):137–46.
[21] Severiano A, Pinto FR, Ramirez M, Carriço JA. Adjusted Wallace coefficient as a measure of congruence between typing methods. J Clin Microbiol 2011;49:3997–4000.
[22] Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sub-lineages. Nat Genet 2016;48(12):1535–43.
[23] Brynildsrud OB, Pepperell CS, Suffys P, Grandjean L, Monteserin J, Debech N, et al. Global expansion of Mycobacterium tuberculosis lineage 4 shaped by colonial migration and local adaptation. Sci Adv 2018;4(10):eaat5869.
[24] Freschi L, Vargas Jr R, Husain A, Kamal SMM, Skrahina A, Tahseen S, et al. Population structure, biogeography and transmissibility of Mycobacterium tuberculosis. Nat Commun 2021;12(1):6099.
[25] Keikha M, Majidzadeh M. Beijing genotype of Mycobacterium tuberculosis is associated with extensively drug-resistant tuberculosis: a global analysis. New Microbes New Infect 2021;43:100921.
[26] Rezai MS, Khotaei G, Mamishi S, Kheirkhah M, Parvaneh N. Disseminated Bacillus Calmette-Guerin infection after BCG vaccination. J Trop Pediatr 2008;54(6):413–6.
[27] Lombardi G, Botti I, Pacciarini ML, Boniotti MB, Roncarati G, Dal Monte P. Five-year surveillance of human tuberculosis caused by Mycobacterium bovis in Bologna, Italy: an underestimated problem. Epidemiol Infect 2017;145(14):3035–9.
[28] Svensson E, Millet J, Lindqvist A, Olsson M, Ridell M, Rastogi N, et al. Impact of immigration on tuberculosis epidemiology in a low-incidence country. Clin Microbiol Infect 2011;17(6):881–7.
[29] Netikul T, Thawornwattana Y, Mahasirimongkol S, Yanai H, Maung HMW, Chongsuvivatwong V, et al. Whole-genome single nucleotide variant phylogenetic analysis of Mycobacterium tuberculosis lineage 1 in endemic regions of Asia and Africa. Sci Rep 2022;12(1):1565.
[30] Comín J, Monforte ML, Samper S. Aragonese Working Group on Molecular Epidemiology of Tuberculosis (EPIMOLA); Otal I. Analysis of Mycobacterium africanum in the last 17 years in Aragon identifies a specific location of IS6110 in Lineage 6. Sci Rep 2021;11(1):10359.