Development of novel genotyping method of *Mycobacterium tuberculosis* strains based on discriminatory analysis of whole genomes alignment

Pedro Ceia e Oliveira

Integrated Master's student in Biological Engineering at Instituto Superior Técnico, Universidade Técnica de Lisboa, Av. Rovisco Pais, 1, 1049-001, Lisbon, Portugal

November 2017

ABSTRACT: Molecular epidemiology has been playing a major role in understanding the key issues in the epidemiology of tuberculosis (TB), contributing to the development of control policies and helping in the prevention and fight against TB. This work aimed to develop a new genotyping method for *Mycobacterium tuberculosis* (*Mtb*) strains using genes containing polymorphic GC-rich sequences (PGRS), that have been systematically linked to the antigenic properties and virulence of *Mtb*. The nucleotide sequences of thirty-two genomes of different *Mtb* strains were aligned and screened for discriminatory regions, leading to the design of fifteen novel pairs of primers, fitting mainly to *PE_PGRS* and *PPE* genes. *In silico,* the new method even outperformed the currently most widely adopted methods. To test the new method and simultaneously assess the occurrence of a possible outbreak, 20 samples from different TB patients were analyzed. Data obtained allowed the differentiation of 16 out of 20 samples. A simultaneous analysis of a set of 5 variable number of tandem repeat (VNTR) loci and the IS*6110*-Mtb2 method, was performed in order to increase the differentiation power and to allow comparisons. Although the new genotyping method proved to be useful in strain differentiation, it still requires optimization and application in larger investigations. Additionally, the mechanisms underlying the mutations in the PGRS genes require deeper studies. Only then, this new method can prove its full potential.

KEYWORDS: Molecular epidemiology; *M. tuberculosis*; genotyping; PGRS genes; antigenic properties and virulence

INTRODUCTION

TB remains one of the most serious global health problems, infecting more than 10 million, and killing almost 2 million people each year, being positioned as the biggest cause of death in the world for an infectious disease[1]. *Mtb* is the most recognized member of the *Mtb* Complex (MTBC), and the most prevailing causative agent of TB, being estimated that one-third of the world's population are infected with this bacilli[2]. This burden is more prevailing in developing countries, especially in Africa and South Asia[1], but also represents a treat in high-income countries, mainly due to TB/HIV coinfection and the increasing emergence of multidrug-resistant (MDR) and extensively drug-resistant (XDR) *Mtb* strains[2–4]. Up to two thirds of TB patients die if they do not receive appropriate treatment[5].

Molecular methods revolutionized the epidemiology of TB, and they have been serving as a resource for understanding the key issues, revealing sources of infection, quantifying recent transmission, identifying transmission links and risk factors, discerning reinfection from relapse, tracking the geographic distribution and clonal expansion of specific strains, and determining the genetic basis behind specific phenotypic characteristics, including virulence or resistance to antimicrobial drugs, and this way contributing to the development of control politics in different parts of the world, helping in the fight against and prevention of TB[2,5].

Determining the *Mtb* genotype is critical for designing efforts to control TB due to the impact of genotype on disease outcome, vaccine efficacy and drug resistance. In fact, sublineages within the main lineages show very distinct geographic association and strain-specific genomic diversity is an important factor in pathogenesis. For these reasons it is important to identify and catalog this lineage/strain specific properties, as well as identify and develop genotyping methods capable of detecting mutations associated with these properties[6–10].

An ideal typing method should be highly discriminatory, easy to perform, fast, inexpensive, 100% reproducible and highly sensitive. In the last 20 years a diversity of typing methods have been developed, however, all methods currently available have their benefits and drawbacks, so, the choice is made accordingly to the study settings[2].

WGS offers ultimate genomic knowledge about each strain[11]. It can be used in all kind of molecular epidemiologic and evolutionary studies[12], and is an extremely useful tool for detection of mutations linked to drug resistance, strain fitness and virulence[13,14]. However, for now, it requires specialized equipment, it is cost prohibitive, especially in low-income countries, and produces huge amounts of data, that is difficult and time-consuming to process. Nevertheless, even after these problems are solved, other genotyping methods might still be useful in preliminary studies and in low-resource scenarios[11].

IS*6110*-RFLP was used for many years as a golden standard due to its high discriminatory power, and now the most used genotyping methods, Spoligotyping and MIRU-VNTR, have the advantages of being fast, simple and reproducible, as well as offering the possibility of incorporation in international databases that contain genotyping results from different countries[11]. However, Spoligotyping has many limitations in the differentiation of Beijing type strains[15] and MIRU-VNTR has a higher discriminatory power, even in Beijing strains[16], however, there is still a need for new genotyping methods that are better in reflecting the variation in the genes encoding virulence features and antigenic properties of the *Mtb* strains.

Following the guidelines proposed by Kotlowski in 2015[17], in this work we present a new, simple PCR genotyping method, targeting mainly the PGRS genes, that are have been linked to virulent and antigenic features of *Mtb*. We hypothesized that identification of specific genotypes will be possible based on the DNA sequence variation within the genes that are targeted in our new genotyping method, and the analysis of defined lengths of amplicons will allow differentiation of strains using combination of agarose gel electrophoresis.

As a secondary objective, and simultaneously validating the proposed new method, it was tested in samples provided by the Pomeranian Center of Infectious Diseases and Tuberculosis, who addressed the mission of verifying the possibility of a strain outbreak within the patients.

MATERIALS AND METHODS

M. tuberculosis whole-genomes alignment

In this work, 32 *Mtb* whole genomes were analyzed. For collection of genome sequences, GenBank database[18] was used. GenBank Accession no.: CP006578.1; CP005387.1; CP003234.1; CP003233.1; CP009427.1; CP002883.1; CP002882.1; CP001641.1; CP001642.1; CP002871.1; CP007803.1; CP005082.1; CP009426.1; CP010873.1; CP007809.1; CP012506.2; HG813240.1; CP005386.1; AE000516.2; CP002992.1; CP000611.1; AL123456.3; CP001664.1; CP004886.1; CP000717.1; CP007027.1; AP012340.1; AP014573.1; HE663067.1; CP009101.1; CP009100.1; FR878060.1.

5 belonged to strains from lineage 1; 12 strains from lineage 2; 1 strain from lineage 3; 13 strains from lineage 4; and 1 strain from lineage 6. ClustalW2[19] was used for genome-wide multiple sequence alignment, using default settings.

Genome alignment discriminatory analysis

AliView[20] was used for whole genome alignment discriminatory analysis. Regions fitting the criteria defined in this work were gathered. This methodology led to the application of fifteen novel pairs of primers.

Primer design

Primer3Plus[21] was used for primer design, using default settings. In cases where primer3Plus could not retrieve any pair of primers, they were manually designed with the following criteria: size 18-22 nucleotides and similar melting temperatures (maximum of 68°C). Sets of primers used for the new method are presented in Table 3.

In silico analysis

In silico analysis of the 32 whole genomes for the new method was performed using the designed primers, and was compared with standard 24/15/12 MIRU-VNTR and Spoligotyping methods.

In silico MIRU-VNTR analysis was performed using primers for each VNTR locus. This primer sequences for amplification of the standard MIRU-VNTR loci are available elsewhere[22].

For *in silico* Spoligotyping analysis, SpoTyping software[23,24] was used. Some strains retrieved a code only with zeros, and were manually analyzed for the presence of the 43 spacer sequences in the DR locus. Sequences for these spacers are available elsewhere[25].

In silico analysis of a custom set of 5 VNTR loci (VNTR0960, VNTR1982, VNTR2372, VNTR3663, VNTR4120) was also performed. These VNTRs loci were experimentally tested in this work, to compare and evaluate the discriminatory power of the new method. Primers used are presented in Table 1.

To evaluate a method discriminatory power, the Hunter-Gaston Discriminatory Index (HGDI) was used[26,27] and is shown in equation 1. N is the total number of strains in the typing scheme, s is the total number of different patterns, and n_j is the number of strains belonging to the jth pattern.

$$HGDI = 1 - \frac{1}{N(N-1)} \sum_{j-1}^{s} n_j(n_j - 1) \qquad (1)$$

Phylogenetic tree

For phylogenetic tree creation, MVSP software[28] was used. For all analysis, the UPGMA clustering method and percent similarity was used.

M. tuberculosis DNA Samples preparation

Samples were provided by the Pomeranian Center of Infectious Diseases and Tuberculosis, in Poland. DNA samples were collected, confirmed, cultured and prepared locally, and previously to this study. 20 sputum samples from positive TB patients, were submitted to Ziehl-Neelsen staining for mycobacterial presence confirmation. The samples were then cultured in Löwenstein–Jensen (LJ) medium, in angled tubes, for 3 to 4 weeks, growing under 37°C and aerobic conditions.

In the laboratory, the protocol described by Kotlowski et al. was followed[29]. DNA extraction from cultures was prepared by scaping two or three loops of cells from LJ slants and suspended into Eppendorf tubes with 150 µl of lysis buffer (10 mM Tris/HCl, pH 8.0; 5 mM EDTA, pH 8.0; 4 M guanidinium isothiocyanate (GITC), pH 7.5; 50 g/l Sarcosyl, 2.5 g/l SDS, 5 g/l sodium citrate and 5 g/l Triton X-100), without homogenization.

Then, samples were mixed with 300 µl chloroform and 300 µl Tris-saturated phenol (pH 6.9), and placed at -20°C for 1h. Subsequently, samples were centrifuged in tubes at 4°C for 20 min at 10 000x *g*. Water phase (upper) was transferred to fresh tubes.

2-Propanol to ¼ volume of the supernatants was added and the mixtures loaded onto silica-cellulose membranes in columns (A&A Biotechnology). Samples were allowed to filter through the membrane by gravity. The membranes were washed twice with 300 µl absolute ethanol (by gravity). DNA was eluted with 400 µl of hot (about 75°C) 1x TE buffer (Tris 10 mM (Molekula); EDTA 1 mM (Sigma); pH 8.0) (by gravity) and precipitated with two portions of absolute ethanol. The resulting pellets were suspended in 25 µl 0.5x TE buffer and stored at -20°C until further analysis.

Chromosomal DNA presence was verified by running agarose gel electrophoresis and ethidium bromide DNA staining, and later by positive PCR reactions.

New method PCR protocol

For PCR, 2 µl of genomic DNA was added to a 23 µl PCR mixture containing 12.5 µl 2x Reaction Buffer (5 mM $MgCl_2$; 100 mM Tris pH 9.0; 40 mM $(NH_4)_2SO_4$ (Blirt); 10% Dimethyl sulfoxide (DMSO) (Sigma)), 4.5 µl purified water (Polpharma), 2 µl dNTP solution (solution with 2.5mM of each deoxyribonucleotide triphosphate in purified water, Sigma), 1.5 µl DMSO, 1 µl MgCl solution (25mM, DNA Gdansk), 0.5 µl of each primer in solution (25 pmol/µl, Oligo) and 0.5 µl TaqNova DNA polymerase (1U/0.5 µl, Cat. No. RP710, Blirt).

PCR was performed under cycling conditions were as follows: denaturation at 94°C for 1min, followed by amplification for 35 cycles of 94°C for 30 seconds, 60°C (65°C for V and IX primers, and 68°C for II primers) for 1 minute, and 72°C for 1 minute, followed by a final extension at 72°C for 5 minutes.

Gel electrophoresis

Agarose gels were prepared by dilution of 2 g of agarose (Abo) in 100 ml 1x TBE Buffer (Tris 0.089M; EDTA 0.002M; boric acid (Poch) 0.089M), and adding 1 µl ethidium bromide (E-7637, Sigma).

A mixture of 10 µl DNA ladder (DNA Gdansk) and 5 µl loading buffer (25 ml glycerol (Sigma); 25 ml TE buffer; 10 g xylene cyanol ff (Sigma)) was added to the first well in each row.

10 µl of amplified DNA from each sample were mixed with 5µl loading buffer, and deposited in 10 consecutive wells per row, with negative control in last well after samples. Then, were subjected to gel electrophoresis, detected by ethidium bromide staining, and visualized under UV light.

Image optimization was performed using IrfanView64[30], and band sizes were measured using GelAnalyzer2010a[31]. The results were organized and the measured band sizes for each region were normalized against the *in silico* expected sizes.

Custom 5 VNTR loci PCR protocol

The protocol followed was the same as the new method PCR protocol, using the primers in Table 1, followed by agarose gel electrophoresis and visualization under UV light.

Table 1 - Primer pairs for the custom 5 VNTR loci used in this work, and melting temperatures.

VNTR Locus			Sets of primers	Tm (°C)
VNTR 0960	FW	5'	GTGATGCGGTAGGTGTGGAC	61.4
	RV	5'	GTCGCACCGATCACGCTAC	62.8
VNTR 1982	FW	5'	CGGTGCTCGAGTTGAAGTAG	58.7
	RV	5'	CAGATCACCACCCAGGAAAT	59.8
VNTR 2372	FW	5'	GAAATGCCGTACTCGACCTC	59.7
	RV	5'	CCTGCTTGATTGTCCACCTC	60.7
VNTR 3663	FW	5'	CAGCTGCCGCCCAAAAAGCAT	70.5
	RV	5'	CTGCCCAGCACCGCGATC	68.8
VNTR 4120	FW	5'	CTCGCCGACCAGCTCACCA	68.4
	RV	5'	TGCCCAATAGCCGGATCCC	67.4

IS*6110*-Mtb2 PCR protocol

Protocol was followed mostly as described by Kotlowski *et al*.[32]. For the PCR, 3 µl aliquots of *Pvu*II (Cat. No R0151S, NEB)-digested genomic DNA were added to a 22 µl PCR mixture containing 2x Reaction Buffer, 25 pmol of each primer, 1U of TaqNova DNA polymerase, and 0.25 mM concentrations of each deoxyribonucleotide triphosphate. Primers are available elsewhere[32].

Cycling conditions were as follows: denaturation at 94°C for 5 min, followed by amplification for 35 cycles of 94°C for 1 min, 66°C for 1 min, and 72°C for 1 min, followed by a final extension at 72°C for 10 min.

5 µl of loading buffer was mixed with 20 µl of amplified DNA for each sample, was subjected to electrophoresis through a 2% agarose gel, detected by ethidium bromide staining, and visualized under UV light.

RESULTS AND DISCUSSION

The main goals of this work were to develop and test a novel genotyping method for *Mtb,* targeting the PGRS genes, and detect the possibility of an outbreak event within the samples provided. It is important to take in consideration that all these samples were collected from patients living in the Pomerania region of Poland, and genotypic diversity can be limited. 32 whole-genomes of *Mtb* strains were aligned and screened for possible regions susceptible to insertion/deletion events (indels) within the PGRS genes, that could be detected using the proposed method. For comparison, the analysis of a custom set of 5 VNTR loci and IS*6110*-Mtb2 method were performed in the same samples.

Whole genome alignment analysis

15 regions were found to fit to the previously defined research specifications. Regions within the genes *Rv0746, Rv0747, Rv0833, Rv1068, Rv1087, Rv1091, Rv1441c, Rv1450c, Rv1452c, Rv2353-Rv2354-Rv2355, Rv3060c-Rv3061c, Rv3135, Rv3345c* and *Rv3388*, were selected for designing primer sequences flanking variable regions for the new method. Of the selected regions, 12 corresponded to *PE_PGRS* genes, 2 for *PPE* genes and 1 for an intergenic region between the *Rv3060c* and *fadE22* genes. Variable regions and correspondent genes are shown in Table 3.

Each *PE* or *PPE* protein contains a highly conserved N-terminal domain with a Proline-Glutamic acid or Proline-Proline-Glutamic acid motif, respectively. The role of the *PE/PPE* genes in antigenic variation of *Mtb* has been proposed since their discovery, however, current knowledge now suggests that these proteins are not variable antigens, however, they clearly play important roles in virulence[33–35].

***In Silico* evaluation of the new method**

The new method was compared *in silico* with 24, 15 and 12 MIRU-VNTR analysis and spoligotyping methods, using all strains, or only Beijing strains, and HGDI values were calculated for each method. As a custom set of 5 VNTR loci was experimentally analyzed, it was also compared *in silico*. The calculated discriminatory power of these methods is presented in Table 2.

Table 2 – Calculated HGDI values for *in silico* Spoligotyping, 24/15/12 MIRU-VNTR, custom 5 VNTR loci and new methods.

Method	HGDI Index	
	All strains	*Mtb* Beijing-type
24 MIRU-VNTR	0.889	1.000
15 MIRU-VNTR	0.889	1.000
12 MIRU-VNTR	0.867	1.000
Spoligotyping	0.881	0.167
Custom 5 VNTR locus	0.841	0.985
New method	**0.946**	**1.000**

The new method is expected to show a higher discriminatory power, with calculated HGDI value of 0.946 for all strains, against 0.881 for spoligotyping and 0.889 for 24 MIRU-VNTR typing.

In silico results also show that spoligotyping has a low discriminatory power for differentiating Beijing strains as expected, clustering 11 out of 12 strains from that lineage, while both, 24 MIRU-VNTR and the new method are expected to be capable to differentiate all these strains.

Added to the clustered Beijing strains, spoligotyping also had two more clusters, one with 2 strains from lineage 4, and the other 3 strains from lineage 1 and 4. The 24 MIRU-VNTR analysis only had one cluster with 11 strains, including strains from lineages 1 to 4.

The new method clustered the same strains as the 24 MIRU-VNTR analysis, however, these were subdivided in two smaller clusters, of 7 and 4 strains, with these clusters being separated only by a difference in region XI.

Comparing to the new method, the 5 custom VNTR locus analysis clustered 13 strains, that included the strains clustered in the new method. Taking this observation, it is expected that, experimentally, the new proposed method will perform better than the 5 custom VNTR analysis in differentiating the 20 samples used in this work.

Table 3 – Variable regions analyzed in this work (I to XV), correspondent Rv code, gene name and sets of primers used in the new method. Primer melting temperatures are also displayed.

Variable Regions	Rv code	Gene name	Sets of primers			Tm (ºC)
I	*Rv0746*	*PE_PGRS9*	FW	5'	TCAGGTGGCTGGTTGTTG	59.2
			RV	5'	TTGAGAAAGCCACGTCCG	61.0
II	*Rv0747*	*PE_PGRS10*	FW	5'	CGGCCTCGGCGGGATTGG	73.4
			RV	5'	GAAACTCCGGCGGCGGTGCTAT	72.1
III	*Rv0833*	*PE_PGRS13*	FW	5'	TGGAGCCTTGCTGTTTGG	61.0
			RV	5'	CGTAGTGAGGCCGAATCC	59.2
IV	*Rv1068c*	*PE_PGRS20*	FW	5'	ATACCGCTGTTGCCGTTG	60.7
			RV	5'	GTTCACGCTGGCCCTGAC	62.9
V	*Rv1087*	*PE_PGRS21*	FW	5'	TGATCGGCAACGGTGGGTT	67.0
			RV	5'	ACCGATCGTCCCCTCGAAG	64.8
VI	*Rv1087*	*PE_PGRS21*	FW	5'	AGGCTGTTGGGCACTGGT	61.7
			RV	5'	TGCGCAAGCTGTAGTAGACG	60.4
VII	*Rv1091*	*PE_PGRS22*	FW	5'	AACCGGTGGGTTGCTCTT	60.5
			RV	5'	GCCATTGGTGCTGCTGAC	61.5
VIII	*Rv1441c*	*PE_PGRS26*	FW	5'	GTCACCCGTGCTTTCCTTG	61.7
			RV	5'	CCTAACAGCGGTGCCAAC	60.3
IX	*Rv1450c*	*PE_PGRS27*	FW	5'	TTAGGGTCGCCGCCAGAA	65.1
			RV	5'	CAGCCGGTGAGGACTTGTCC	65.4
X	*Rv1452c*	*PE_PGRS28*	FW	5'	GTCGCCAAATACCGTGAGAC	60.5
			RV	5'	AAGGGCGGGGAAAACATC	62.1
XI	*Rv2353 / (Rv2354-Rv2355)*	*PPE39 / (Rv2354-Rv2355)*	FW	5'	CCGAAGCCGATGTTGTTACT	60.1
			RV	5'	GGTAGTGGTGAATTTTTCGG	59.4
XII	*Rv3060c / Rv3061c*	*Rv3060c / fadE22*	FW	5'	GCTCGGTGCTCATTTCAATA	58.9
			RV	5'	GAGGTGACCCGCAATCAG	60.2
XIII	*Rv3135*	*PPE50*	FW	5'	ACACCGAGGTCCGAATTG	59,5
			RV	5'	TGCTGGTCGAGAACTGAATG	60.0
XIV	*Rv3345c*	*PE_PGRS50*	FW	5'	CACCCTTCATGGCTGGAAT	60.9
			RV	5'	CGGTAATGGCGGAAATGG	62.2
XV	*Rv3388*	*PE_PGRS52*	FW	5'	CAACGGCGGCAACAATAC	61.1
			RV	5'	GGAGTACGTGGCGGGTTAG	60.5

Region I

The amplified region is inserted within the *PE_PGRS9* gene. The *KIT87190* strain had no hybridization site for the reverse primer, with a deletion that prolonged for more than 4000 bp, and thus absence of a PCR product is a possibility.

One indel, with 90 bp, was present in only two Beijing strains (*K* and *96075*) and the *Africanum* strain. The other, with a length of 135 bp, was lacking in 7 of the 12, here represented Beijing strains, suggesting that this indel is lineage defining, absent only in Beijing strains.

Amplicons with 110 bp are expected to lack both indels. Amplicons with 245 bp are expected to lack the 90bp indel, and amplicons carrying both are expected to have 335bp.

Using the designed primers IF and IR, all samples were subjected to PCR, using 60°C during the annealing step. Figure 1 shows the results after electrophoresis.

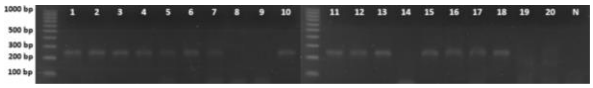

Figure 1 – PCR results after electrophoresis, using primer pair IF and IR. The annealing step was performed at 60°C. DNA ladder: 100-1000 bp. N – negative control.

The obtained results displayed weak bands, and samples 8, 9, 14, 19 and 20 had no PCR product. To obtain better results, samples 7 to 9 and 13 to 19, were submitted to a new PCR, now using 62°C during the annealing step. The results then present clear bands. Not all samples were subjected to a new PCR, including sample 20 that had no PCR product, due to limited available genomic DNA. Sample 15 had no PCR product on the second run, however, a band was visible in the first run, suggesting the occurrence of laboratory errors.

All samples (excluding sample 20), produced equal PCR products, matching the expected *in silico* 245 bp band. In Poland, only 6.5% patients were reported to be infected with Beijing strains[37], and as these indels appear to occur exclusively in Beijing strains, this result was expected.

Region II

The amplified region is inserted within the *PE_PGRS10* gene, it corresponds to the region previously described by Kotlowski, named "Variable region V"[17]. *KIT87190* and *Erdman* strains had complete deletion of the whole amplified region.

One indel contained a 42 bp repeated sequence interspersed with other 42 bp sequence, with the whole indel event having 84 bp. The other had a 26 bp repetition interspersed with a 58 bp sequence, with the indel event having 48 bp. Both indels were absent mostly in Beijing strains, only the *CDC1551* strain, from lineage 4, lacked the 48 bp indel, and *Africanum* strain lacked both.

Amplicons with 306 bp lack both indels. Amplicons with 354 bp only lack the 84 bp indel, and 390 bp products only the 48 bp indel. Products with 438 bp contain no deletions.

Using the designed primers IIF and IIR, all samples were subjected to PCR, using 62°C during the annealing step. Figure 2 shows the results after electrophoresis.

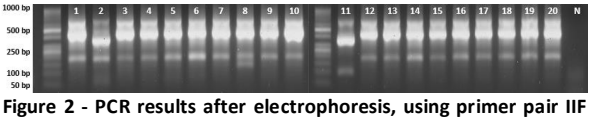

Figure 2 - PCR results after electrophoresis, using primer pair IIF and IIR. The annealing step was performed at 62°C. DNA ladder: 50-1000 bp. N – negative control.

Unspecific bands were produced probably because the annealing step was mistakenly performed at 62°C instead of 68°C, which is the optimal hybridization temperature for the set of primers used. However, the expected bands are present, showing more intensity compared to the remaining.

Two different sizes were obtained, with most samples producing the expected 438 bp band, while sample 2 and 11 produced a band with what was assumed to be the expected 354 bp product. This 354 bp product appeared to be specific of Beijing strains, according to the *in silico* analysis.

Region III

The amplified region is inserted within the *PE_PGRS13* gene. This region contains 3 indel events. *Erdman* strain had no forward primer hybridization site.

One indel consists in a 36-45 bp sequence, that was absent in *SCAID 187* and *F11* strains. An 84 bp indel was present only in *96075* Beijing strain and A*fricanum* strain, while a larger 120 bp indel was lacking only in Beijing strains, implying to be specific of this lineage. However, *CITR_2* strain, from lineage 2 had a deletion with similar size, 129 bp.

Amplicons with 352 bp are expected to lack all the indel events. Amplicons with 388 bp had absence of the 84 bp and 120 bp indels, while amplicons with 463 bp lacked the 84 bp and 36-45 bp indels. Amplicons with 508 bp lack only of the 84 bp indel, and amplicons with 592 bp contain all of them.

Using the designed primers IIIF and IIIR, all samples were subjected to PCR, using 60°C during the annealing step. Figure 3 shows the result after electrophoresis.

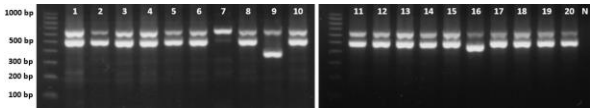

Figure 3 – PCR results after electrophoresis, using primer pair IIIF and IIIR. The annealing step was performed at 60°C. DNA ladder: 100-1000 bp. N – negative control.

The appearance of two separate bands was unexpected, so PCR was repeated with the samples 7 to 9, and 14 to 16, now using 62°C during the annealing step. In samples 7 to 9, we successfully produced one band only, however, in samples 14 to 16 the two bands were still present.

Despite the appearance of an unspecific band with approximately 600 bp, differentiation between samples was observed. Most samples had the predicted 508 bp band, while samples 9 and 16 had a 388 bp and 463 bp bands, respectively. Sample 7 had appeared to have no PCR product.

Region IV

The amplified region is inserted within the *PE_PGRS20* gene. This region exploits 4 indel events. *Erdman* and *KIT87190* strains had no forward primer hybridization site.

A bigger, 719 bp long indel was present only in *96075* and *Africanum* strains. Three more indels were present within this region. A larger, with 192 bp, was frequently present in Beijing strains, although not exclusive, only 3 other strains from lineage 4 contained this indel. Two other smaller indels, with 45 bp and 54 bp, were lacking exclusively in Beijing strains. Interestingly, the absence of the 192 bp indel never occurred together with the absence of any of the other smaller events. Likewise, whenever the 54 bp indel was lacking, the 45 bp was also absent.

Amplicons with 684 bp are expected to lack both 719 bp and 192 bp indels. Amplicons with 777 bp are absent of the large 719 bp, and both smaller 45 bp and 54 bp indels, while amplicons with 831 bp lack the 719 bp and 45 bp indels. Amplicons with 876 bp only lack the 719 bp indel, while an amplicon with 1595 bp contains all of them.

Using the designed primers IVF and IVR, all samples were subjected to PCR, using 60°C during the annealing step. Figure 4 shows the result after electrophoresis.

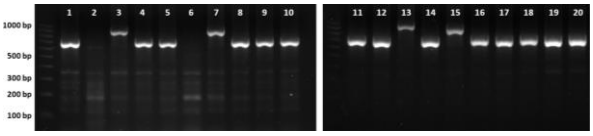

Figure 4 – PCR results after electrophoresis, using primer pair IVF and IVR. The annealing step was performed at 60°C. DNA ladder: 100-1000 bp.

Samples 2 and 6 had no PCR product, and were submitted to a new PCR, using 62°C during the annealing step. Sample 2 had again no PCR product and sample 6 produced a band with approximately 350 bp, that was defined as being unspecific. Globally, most samples produced a band with an expected 684 bp. Samples 3, 7 and 15 produced a band with 831 bp while sample 13 had 876 bp.

Region V

The amplified region is inserted within the *PE_PGRS21* gene. This region contains 3 indel events. *CCDC5180* strain had no reverse primer hybridization site.

Two of the indels have 48 bp, and one appears to be lineage defining, appearing only in Beijing strains, however, *F11* strain has a slightly different deletion site, but with the same size, and thus, cannot be distinguished using this method. A longer, 79 bp long indel is absent in *CDC1551* strain.

Amplicons with 184 bp or 215 bp, lack the 79 bp or 48 bp indels, respectively. Amplicons with 263 bp carry no deletions.

Using the designed primers VF and VR, all samples were subjected to PCR, using 65°C during the annealing step. Figure 5 shows the result after electrophoresis.

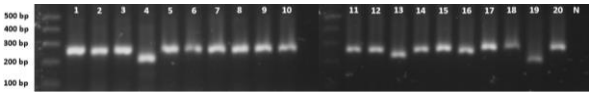

Figure 5 – PCR results after electrophoresis, using primer pair VF and VR. The annealing step was performed at 65°C. DNA ladder: 100-500 bp. N – negative control.

The differences between PCR product sizes were too small for the sensitivity of this method, and for this reason, region V was discarded from analysis.

Region VI

Like region V, this region is inserted within the *PE_PGRS21* gene. This region exploits 2 indels. *CCDC5180* and *96075* strains had no forward primer hybridization site.

One of the indels was characteristic of the *Africanum* strain, where a 149 bp indel was present. The other indel, with 45 bp, was lacking in all strains of lineage 1, and showed no specific pattern in the remaining lineages, however, it was frequently present in Beijing strains.

Amplicons with 604 bp would lack the 45 bp and 149 bp indels, while amplicons with 649 bp include the 45 bp indel. *Africanum* strain would have an amplicon with 796 bp.

Using the designed primers VIF and VIR, all samples were subjected to PCR, using 60°C during the annealing step. Figure 6 shows the result after electrophoresis.

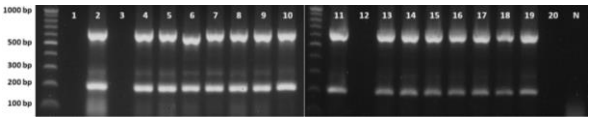

Figure 6 – PCR results after electrophoresis, using primer pair VIF and VIR. The annealing step was performed at 60°C. DNA ladder: 100-1000 bp. N – negative control.

The intended band was obtained (upper band), however, a second unspecific band was also present. Samples 1, 3 and 12 produced no PCR product, probably because of laboratory mistakes during the preparation of the PCR mixture. These samples were not repeated due to lack of available DNA for a new PCR run.

The remaining samples produced a band with 604 bp, except for sample 6, that had a smaller band, with approximately 550 bp, that was not observed *in silico*.

Region VII

The amplified region is inserted within the *PE_PGRS22* gene. This region contains 3 indel events. Forward or reverse primer hybridization sites were missing in *BS1* and *Kurono* strains, respectively. Strain *KIT87190* had a completely different sequence, in size and nucleotides, together with the presence of a repeated hybridization site for the reverse primer.

The three indels, with sizes of 75 bp, 111 bp or 654 bp, always occurred separately. However, these events appear with low frequency, with the *CDC1551* strain being the only strain lacking the large, 654 bp indel, and *CITR_2* the only lacking the 75 bp indel. In *96075* and *Africanum* strains, the 111 bp indel was absent.

Amplicons with 100 bp, 643 bp and 679 bp, lack the 654bp, 111bp or 75bp indel, respectively. Amplicons with 754bp contain all these indels. The *KIT87190* strain is expected to have a double amplicon, with 1318bp and 1857bp.

Using the designed primers VIIF and VIIR, all samples were subjected to PCR, using 60°C during the annealing step. Figure 7 shows the result after electrophoresis.

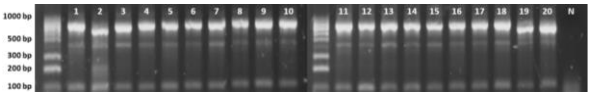

Figure 7 - PCR results after electrophoresis, using primer pair VIIF and VIIR. The annealing step was performed at 60°C. DNA ladder: 100-1000 bp. N – negative control.

Most samples produced a band with 754 bp, with samples 3, 19 and 20 producing a band that was assumed to be 679 bp.

Region VIII

The amplified region VIII is inserted within the *PE_PGRS26* gene. This region exploits 2 indel events. Reverse primer hybridization site was absent in *BT1* strain.

Two indels, with 55 bp and 438 bp were detected within this region. While the 55 bp segment was lacking in *BT2* strain, three strains from lineage 4 and *Africanum* strain, the larger 438 bp indel was absent only in *KIT87190* strain.

Amplicons lacking the 55 bp and 438 bp indels are expected to be 630 bp and 247 bp, respectively, while an amplicon with 685 bp will contain all indels.

Using the designed primers VIIIF and VIIIR, all samples were subjected to PCR, using 60°C during the annealing step. Figure 8 shows the result after electrophoresis.

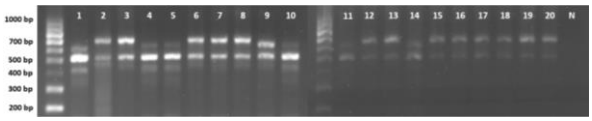

Figure 8 - PCR results after electrophoresis, using primer pair VIIIF and VIIIR. The annealing step was performed at 60°C. DNA ladder: 100-1000 bp. N – negative control.

The appearance of an unspecific band was unexpected, and a new PCR was run, now using 62°C during the annealing step. Only samples 7 to 9 and 14 to 16, were used, due to limited DNA samples. Clear, single bands were now produced.

Most samples produced a band with 685 bp, while samples 1, 4, 5, 11 and 14 had a band with 630 bp.

Region IX

The amplified region IX is inserted within the *PE_PGRS27* gene. This region exploits 2 indel events. Reverse primer hybridization site was lacking in *BT1* strain, and the sequence appeared to be completely different from the remaining strains.

An indel with 207 bp appears to be frequently absent in strains from lineage 1, however, it shows no specific lineage pattern. A larger, 456 bp long indel, was lacking in *Erdman* strain.

Amplicons lacking the 207 bp or 456 bp indels are expected to have 451 bp and 202 bp, respectively. Amplicons without deletions have 658 bp.

Using the designed primers IXF and IXR, all samples were subjected to PCR, using 65°C during the annealing step. Figure 9 shows the result after electrophoresis.

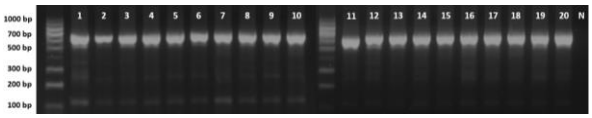

Figure 9 - PCR results after electrophoresis, using primer pair IXF and IXR. The annealing step was performed at 65°C. DNA ladder: 100-1000 bp. N – negative control.

All samples produced the expected 658bp band, except for sample 11, that produced a band with approximately 570bp, not observed *in silico*.

Region X

The amplified region X is inserted within the *PE_PGRS28* gene. This region exploits 2 indel events, that appear to occur attached together.

The first indel, is an 86 bp sequence, that is attached, or not, to a 51 bp tandem repeat, appearing with two, one, or no copies. These indels appear to be specific of Beijing strains, appearing also in the *Africanum* strain.

Amplicons with 294 bp, 243 bp and 192 bp, would contain the 86 bp indel, and two, one or no copies of the 51 bp tandem repeat, respectively. Amplicons with 102 bp would lack all these indels. The *BT1* strain had a repeated reverse primer hybridization site, and could possibly produce two amplicons, with 192 bp and 688 bp.

Using the designed primers XF and XR, all samples were subjected to PCR, using 60°C during the annealing step. Figure 10 shows the result after electrophoresis.

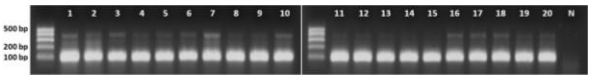

Figure 10 – PCR results after electrophoresis, using primer pair XF and XR. The annealing step was performed at 60°C. DNA ladder: 100-500 bp. N – negative control.

Like region I, this indels appeared to be present specifically in Beijing strains. All samples produced the 102 bp band.

Region XI

The amplified region is inserted within the *PPE39* gene, together with an insertion site for IS*6110*. Here we target a large, 803 bp indel event in the *PPE39* gene, together with the presence or absence of the IS*6110*. *Kurono* strain had no hybridization site for the forward primer.

Amplicons with 203 bp lack both the 803 bp indel and IS*6110*. Amplicons with 1005 bp would have no IS*6110*, and 1559 bp products would carry the IS*6110*, but not the 803 bp indel. Amplicons with 2363 bp would contain both events.

Using the designed primers XIF and XIR, all samples were subjected to PCR, using 60°C during the annealing step. Figure 11 shows the result after electrophoresis.

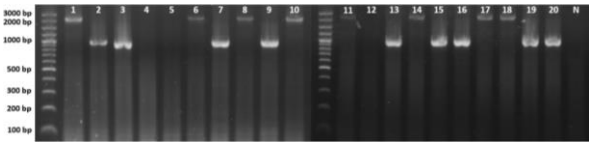

Figure 11 – PCR results after electrophoresis, using primer pair XIF and XIR. The annealing step was performed at 60°C. DNA ladder: 100-3000 bp. N – negative control.

Samples 4 and 5 had no PCR products, and were submitted to a new PCR, for verification. Sample 12 also had no PCR product, but due to lack of DNA, it was not repeated. As there were no PCR products in the second run, it was considered that would occur every time.

The sizes obtained of 1005 bp and 2363 bp, were correspondent to the presence or not, respectively, of the IS*6110* in the region, while the 803 bp indel was always present.

Region XII

The amplified region is inserted within the *Rv3060c* and *fadE22* genes. Here we target a 68bp indel that appears to occur in the intergenic region. Although this mutation does not influence the encoded protein, it may influence the expression of downstream genes.

Most strains from lineage 1 lacked this indel, however, no specific lineage pattern was observed.

Amplicons with 128 bp or 196 bp, would be absent, or not, of this indel, respectively.

Using the designed primers XIF and XIR, all samples were subjected to PCR, using 60°C during the annealing step. Figure 12 shows the result after electrophoresis.

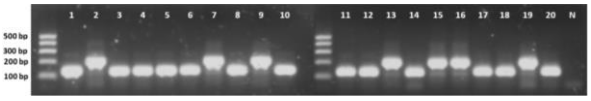

Figure 12 - PCR results after electrophoresis, using primer pair XIIF and XIIR. The annealing step was performed at 60°C. DNA ladder: 100-500 bp. N – negative control.

The expected sizes of 128 bp or 196 bp were obtained in all samples.

Region XIII

The amplified region XIII is inserted within the *PPE50* gene. This region exploits 3 indel events. *96121* strain had no hybridization site for the forward primer.

Two of the indels appeared to be lineage specific. One, with 282 bp, was lacking only in strains from lineage 4. The other indel events were present only in Beijing strains and the *Africanum* strain. Here, a 420 bp indel was present in all Beijing strains, except for the *Beijing/NITR203* strain. A larger, 947 bp sequence was inserted within the previously described indel, but only in *K* and *KIT87910* strains, while *Africanum* strain had an insertion in the same site, but only with 591 bp.

Amplicons with 165 bp and 447 bp, are expected to be absent, or not, of the 282 bp indel. Amplicons with 867 bp contain also the 420 bp indel, while amplicons with 1814 bp have the presence of the larger 947 bp indel, and *Africanum* strain would have a PCR product with 1223 bp.

Using the designed primers XIIIF and XIIIR, all samples were subjected to PCR, using 60°C during the annealing step. Figure 13 shows the result after electrophoresis.

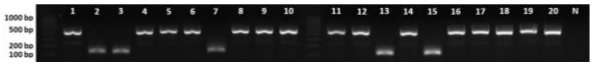

Figure 13 - PCR results after electrophoresis, using primer pair XIIIF and XIIIR. The annealing step was performed at 60°C. DNA ladder: 100-1000 bp. N – negative control.

Most samples produced a band with the expected 447 bp and 5 samples produced the 165 bp band.

Region XIV

The amplified region XIV is inserted within the *PE_PGRS50* gene, it corresponds to two regions already described by Kotlowski, named "Variable region II and III"[17]. This region exploits 2 main indel events.

One indel is present only in Beijing strains and *Africanum* strain, with 87 bp. Other indel, with 78 bp, was present in most strains, and was always present, when the other indel was present. This region also features an 83 bp deletion in the *96121* strain, a 103 bp deletion in the *Kurono* strain and a large 501 bp deletion in *K* strain.

Amplicons with 824 bp lack the 87 bp and 78 bp indels, amplicons with 902 bp only lack the 87 bp indel, and with 989 bp contain all indels, being this last product specific of Beijing and *Africanum* strains. *K*, *Kurono* and *96121* strains would have PCR products with 323 bp, 722 bp and 758 bp, respectively, with their specific deletions, and lack of the 87 bp and 78 bp indels.

Using the designed primers XIVF and XIVR, all samples were subjected to PCR, using 60°C during the annealing step. Figure 14 shows the result after electrophoresis.

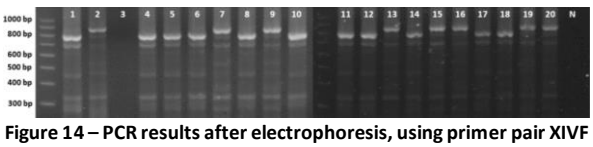

Figure 14 – PCR results after electrophoresis, using primer pair XIVF and XIVR. The annealing step was performed at 60°C. DNA ladder: 100-1000 bp. N – negative control.

Sample 3 had no PCR product, probably because to laboratory mistakes during the PCR preparation. However, due to the lack of DNA sample, it was impossible to test

again. The remaining samples produced bands with 902 bp and 824 bp.

Region XV

The amplified region XV is inserted within the *PE_PGRS52* gene and exploits 3 indels. *K* strain had no hybridization site for the forward primer, and *KIT87190* had a completely different sequence, and no matching site for the forward primer, possibly due to sequencing errors.

One indel, with 78 bp, appears to be present only in the *Africanum* strain. A large 446 bp deletion and a 186bp deletion are also observed. However, these events appear to occur in low frequency, with only *96121* and *H37Ra* strains lacking the 446 bp, and *CCDC5180* strain the 186 bp indel.

Amplicons with 655 bp and 915 bp lack the 446 bp and 186 bp indels, respectively, together with the absence of the *Africanum* specific sequence. Amplicons lacking only the 78 bp indel have 1101 bp, and the *Africanum* strain has 1179 bp.

Using the designed primers XVF and XVR, all samples were subjected to PCR, using 60°C during the annealing step. Figure 15 shows the result after electrophoresis.

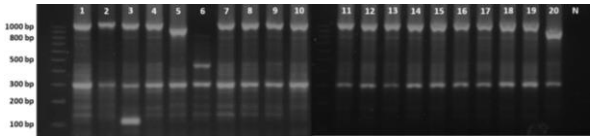

Figure 15 - PCR results after electrophoresis, using primer pair XVF and XVR. The annealing step was performed at 60°C. DNA ladder: 100-1000 bp. N – negative control.

An unspecific band with approximately 300 bp was produced in all samples. Most samples also produced the 1101 bp band, while samples 5 and 20 produced what is expected to be the 916 bp bands. Sample 6 produced a different band, with approximately 450 bp, that is possibly unspecific.

Phylogenetic tree construction and results analysis

All the measured amplicon sizes were gathered and normalized against the *in silico* results. For phylogenetic tree arrangement, only the regions where all samples produced a band, or were considered to never produce a band, were used. This excluded the regions I, V, VI and XIV. It should be noted that these regions produced results that should not be ignored, and where considered for HGDI calculation.

The created phylogenetic tree is presented in Figure 16. It indicates the formation of two clusters with 3 samples each (1, 10, 14 and 8, 17, 18), and the samples between these clusters only had differences in the analyzed region VIII. Even analyzing the excluded regions, I, V, VI and XIV, none of these samples had more differences.

The new method differentiated 16 out of 20 samples. As addressed before all these samples were collected from patients living in the same geographic region, and genotypic diversity could be limited, so these results appear to be positive. To add discriminatory power, compare and evaluate the potential of the new method, 5 different VNTR loci were analyzed.

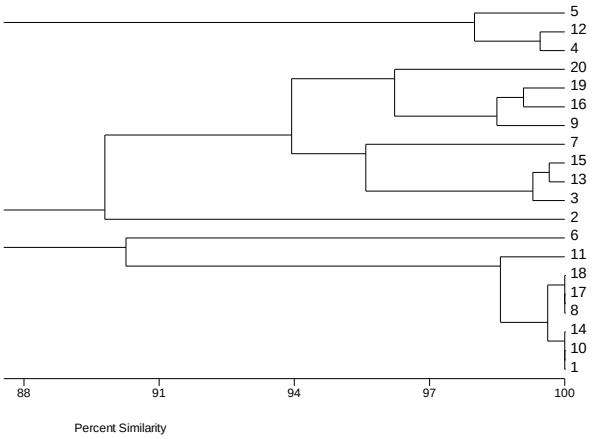

Figure 16 – Phylogenetic tree created using the new proposed PCR method.

Custom 5 VNTR loci analysis

Results for the amplification of the VNTR0960, VNTR1982, VNTR2372, VNTR3663 and VNTR4120 locus, are presented in Figure 17, from top to bottom, respectively.

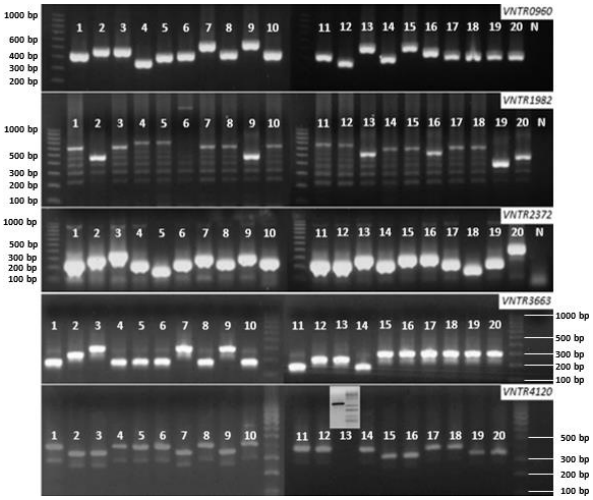

Figure 17 – Results after electrophoresis for the custom 5 VNTR locus analysis. DNA ladders: 100-1000 bp. N – negative control.

Sample 13 was absent in VNTR 4120, due to laboratorial errors, and one a second PCR run, a positive result was obtained.

All the measured amplicon sizes were gathered and normalized against the *in silico* results. Phylogenetic tree was created using these results and is presented in Figure 18.

Samples 1, 8, 10 and 14 were now clustered. Comparing to the cluster obtained in the new method, composed by samples 1, 10 and 14, these samples were again clustered, now together with the sample 8.

The other cluster obtained with the new method, comprising samples 8, 17 and 18, was now separated. However, samples 8 and 17, and samples 17 and 18, had differences of only one copy in one single locus.

Combining the two methods, one cluster was still obtained, with samples 1, 10 and 14. For further comparison and to add more differentiation power, the IS*6110*-Mtb2 PCR method was performed in all samples.

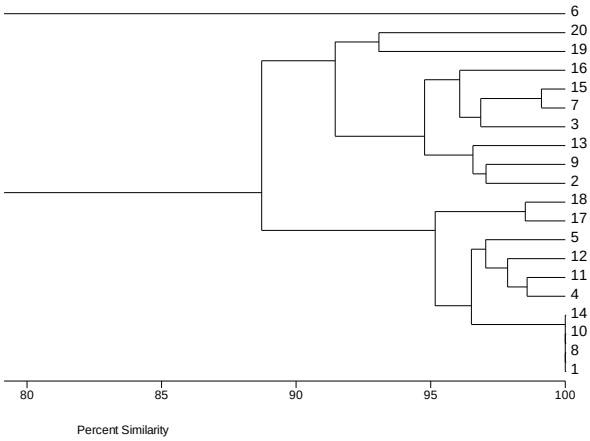

Figure 18 – Phylogenetic tree created using the custom 5 VNTR loci analysis.

IS*6110*-Mtb2 method

The method was prepared and performed as previously described, and the results are presented in Figure 19.

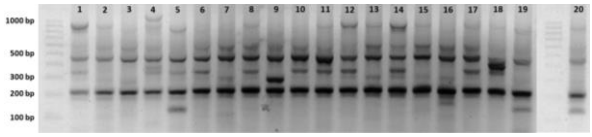

Figure 19 – Results after electrophoresis for the IS*6110*-Mtb2 PCR method, performed in all samples.

For clustering analysis and construction of a phylogenetic tree, presented in Figure 20, an array was created, using "1" for the presence of a band, and "0" for absence of that band.

This method paired samples 1 and 12, samples 2 and 15, samples 6 and 10, samples 7 and 9, and samples 19 and 20. Another cluster with samples 8, 11 and 17 was also formed.

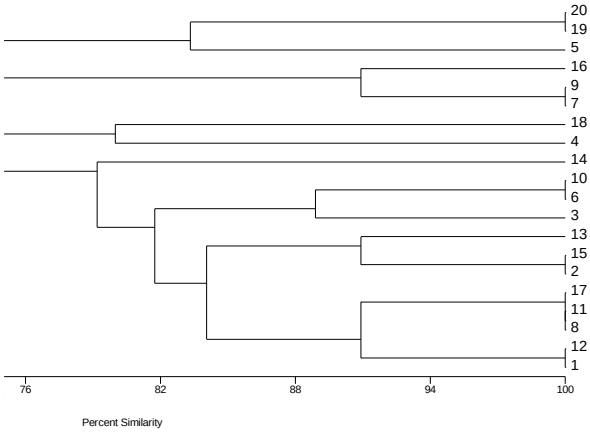

Figure 20 – Phylogenetic tree created based on the array created for the IS*6110*-Mtb2 method.

This method will be used to further differentiate samples that were clustered using the other two methods. For this effect, band patterns for the samples 1, 8, 10 and 14, and samples 8, 17 and 18, were compared. The IS*6110*-Mtb2 patterns for these samples are shown together in Figure 21.

Sample 10 appears to exhibit a different pattern, with one band clearly absent, compared to samples 1, 8 and 14. Also, sample 18 exhibits a completely different pattern compared to sample 8 and 17.

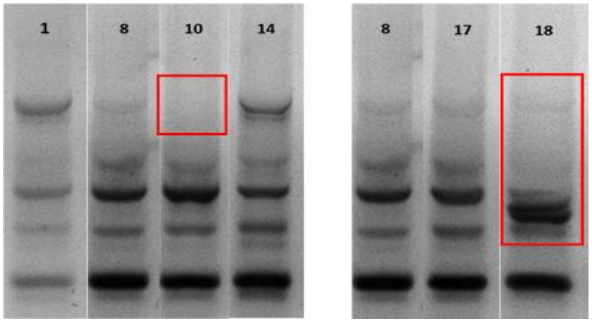

Figure 21 – IS*6110*-Mtb2 PCR results for the samples 1, 8, 10 and 14 (left), and samples 8, 17 and 18 (right).

Samples 1 and 14 displayed the same profile in all methods employed in this work, and possibly correspond to the same strain. The calculated HGDI values for each method, and combining all methods, are presented in Table 4.

Table 4 – HGDI values for the three methods experimentally employed in this study, and the combined HGDI.

Method	HGDI Index
New method	0,968
Custom 5 VNTR loci	0,968
IS*6110*-Mtb2	0,958
Combined	0.995

Experimentally, the new method calculated HGDI was equal to that of the custom 5 VNTR loci analysis, creating two smaller clusters with 3 samples each, compared to a cluster with 4 samples, respectively. But it outperformed the IS*6110*-Mtb2 method. This comparison can be extrapolated, as a previous study experimentally confirmed that the IS*6110*-Mtb2 method has similar discriminatory compared to IS*6110*-RFLP, and even higher compared to spoligotyping and 12 MIRU-VNTR[38]. As our new method outperformed the IS*6110*-Mtb2 method in this work, we can assume that it can be positioned, at least, in the same conditions.

CONCLUSIONS

The main objective of this work was to develop a new PCR genotyping method targeting the PGRS genes, that is expected to better reflect the variation in the genes encoding virulence features and antigenic properties of the *Mtb* strains.

Regarding the detection of a possible outbreak within the samples, the new method differentiated 16 out of 20 samples, creating two different clusters with 3 samples each. The analysis of 5 different VNTR loci, further separated one cluster. IS*6110*-Mtb2 PCR typing method was employed for increasing differentiation power, further separating two samples from the other cluster. Overall, the new method equaled the discriminatory power of the 5 VNTR loci analysis and outperformed the IS*6110*-Mtb2

In the end different profiles for 19 of the 20 samples were created, where only samples 1 and 14 had equal signatures in all three methods, providing a strong evidence that they might represent the same strain. Based in these results, we can address to the Pomeranian Center of Infectious Diseases and Tuberculosis, excluding the possibility of an outbreak.

In short, we believe the new proposed method proved to be useful in differentiating *Mtb* strains. Also, it is inexpensive, simple and fast to perform, and it is a PCR based method, that requires small amounts of genomic DNA. As it is based in the analysis of the size of single amplicons, it is expected to be highly reproducible, and could be incorporated in

databases, addressing information relative to the amplicon size, that is expected to be intrinsically connected to specific mutations occurring within the analyzed regions.

FUTURE WORK

Regarding the new method, three future requirements are clear. The optimization and standardization of the experimental conditions, the employment in larger investigations, integrating the variability of both the pathogen and host and the need for deeper studies regarding the mechanisms underlying the *PE/PPE* genes, to better understand the difference between polymorphism diversity and the variation in pathological and immunological properties. Only then, this new method can prove its full potential.

REFERENCES

1. WHO. Global Tuberculosis Report 2016. *Cdc* (2016).
2. Jagielski, T. *et al.* Methodological and clinical aspects of the molecular epidemiology of *Mycobacterium tuberculosis* and other mycobacteria. *Clin. Microbiol. Rev.* **29,** 239–290 (2016).
3. Niemann, S. *et al.* Impact of Genetic Diversity on the Biology of *Mycobacterium tuberculosis* Complex Strains. *Microbiol. Spectr.* 1–18 (2016).
4. García De Viedma, D. *et al.* Innovations in the molecular epidemiology of tuberculosis. *Enferm. Infecc. Microbiol. Clin.* **29,** 8–13 (2011).
5. Ribón, M. L. O. and W. Molecular Epidemiology of Tuberculosis. *RFID Technol. Secur. Vulnerabilities, Countermeas.* (2015).
6. Coll, F. *et al.* A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat. Commun.* **5,** 4812 (2014).
7. Cardoso Oelemann, M. *et al.* The forest behind the tree: Phylogenetic exploration of a dominant *Mycobacterium tuberculosis* strain lineage from a high tuberculosis burden country. *PLoS One* **6,** (2011).
8. Schleusener, V. *et al. Mycobacterium tuberculosis* resistance prediction and lineage classification from genome sequencing: comparison of automated analysis tools. *Sci. Rep.* **7,** 46327 (2017).
9. Faksri, K. *et al. In silico* region of difference (RD) analysis of *Mycobacterium tuberculosis* complex from sequence reads using RD-Analyzer. *BMC Genomics* **17,** 847 (2016).
10. Chen, H. *et al. Mycobacterium tuberculosis* Lineage Distribution in Xinjiang and Gansu Provinces, China. *Sci. Rep.* **7,** 1068 (2017).
11. Ei, P. W. *et al.* Molecular strain typing of *Mycobacterium tuberculosis*: A review of frequently used methods. *J. Korean Med. Sci.* **31,** 1673–1683 (2016).
12. Nu, U. *et al.* Whole Genome Sequencing versus Traditional Genotyping for Investigation of a *Mycobacterium tuberculosis* Outbreak : A Longitudinal Molecular Epidemiological Study. *PLoS Med.* **10,** (2013).
13. Perdigão, J. *et al.* Unraveling *Mycobacterium tuberculosis* genomic diversity and evolution in Lisbon , Portugal , a highly drug resistant setting. *BMC Genomics* **15,** (2014).
14. Coll, F. *et al.* Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.* **7,** 51 (2015).
15. Comas, I. *et al.* Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PLoS One* **4,** (2009).
16. Iwamoto, T. *et al.* Hypervariable loci that enhance the discriminatory ability of newly proposed 15-loci and 24-loci variable-number tandem repeat typing method on *Mycobacterium tuberculosis* strains predominated by the Beijing family. *FEMS Microbiol. Lett.* **270,** 67–74 (2007).
17. Kotlowski, R. A novel method of *Mycobacterium tuberculosis* complex strain differentiation using polymorphic GC-rich gene sequences. *Acta Biochim. Pol.* **62,** 317–322 (2015).
18. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41,** 36–42 (2013).
19. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).
20. Larsson, A. AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30,** 3276–3278 (2014).
21. Untergasser, A. *et al.* Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res.* **35,** 71–74 (2007).
22. Supply, P. Multilocus Variable Number Tandem Repeat Genotyping of *Mycobacterium tuberculosis* - Technical guide. 73 (2005).
23. Xia, E. *et al.* SpoTyping: fast and accurate in silico Mycobacterium spoligotyping from sequence reads. *Genome Med.* **8,** 19 (2016).
24. SpoTyping. Available at: https://github.com/xiaeryu/SpoTyping-v2.0/tree/master/SpoTyping-v2.0-commandLine. (Accessed: 1st May 2017)
25. Van der Zandem, A. *et al*. Improvement of differentiation and interpretability of spoligotyping for *Mycobacterium tuberculosis* complex isolates by introduction of new spacer oligonucleotides. *J. Clin. Microbiol.* **40,** 4628–4639 (2002).
26. Hunter, P. R. & Gaston, M. a. Numerical index of the discriminatory ability of typing systems : an application of Simpsons Index of Diversity. *J. Clin. Microbiol.* **26,** 2465–2466 (1988).
27. Mokrousov, I. Revisiting the Hunter Gaston discriminatory index: Note of caution and courses of change. *Tuberculosis* **104,** 20–23 (2017).
28. MVSP software. Available at: https://www.kovcomp.co.uk/mvsp/index.html.
29. Kotlowski, R. *et al.* One-tube cell lysis and DNA extraction procedure for PCR-based detection of *Mycobacterium ulcerans* in aquatic insects, molluscs and fish. *J. Med. Microbiol.* **53,** 927–933 (2004).
30. IrfanView. Available at: http://www.irfanview.com/. (Accessed: 1st May 2017)
31. Skosyrev, V. S. *et al.* Specialized software product for comparative analysis of multicomponent DNA fingerprints. *Russ. J. Genet.* **49,** 464–469 (2013).
32. Kotłowski, R. *et al.* PCR-Based Genotyping of *Mycobacterium tuberculosis* with New GC-Rich Repeated Sequences and IS*6110* Inverted Repeats Used as Primers. *J. Clin. Microbiol.* **42,** 372–377 (2004).
33. Fishbein, S. *et al.* Phylogeny to function: *PE/PPE* protein evolution and impact on *Mycobacterium tuberculosis* pathogenicity. *Mol. Microbiol.* **96,** 901–916 (2015).
34. Chen, X. *et al.* Structural basis of the *PE-PPE* protein interaction in *Mycobacterium tuberculosis*. *J. Biol. Chem.* (2017).
35. Brennan, M. J. The Enigmatic *PE/PPE* Multigene Family of Mycobacteria and Tuberculosis Vaccination. *Infect. Immun.* **85,** 1–8 (2017).
36. Machowski, E. E. *et al.* In vitro analysis of rates and spectra of mutations in a polymorphic region of the *Rv0746 PE_PGRS* gene of *Mycobacterium tuberculosis*. *J. Bacteriol.* **189,** 2190–2195 (2007).
37. Kozinska, M. E. A.-K. Drug Resistance and Population Structure of *Mycobacterium tuberculosis* Beijing Strains Isolated in Poland. *Polish J. Microbiol.* **64,** 399–401 (2015).
38. Sajduda, A. *et al.* Evaluation of multiple genetic markers for typing drug-resistant *Mycobacterium tuberculosis* strains from Poland. *Diagn. Microbiol. Infect. Dis.* **55,** 59–64 (2006).