Infection, Genetics and Evolution 18 (2013) 238–246

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

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

SNP typing reveals similarity in *Mycobacterium tuberculosis* genetic diversity between Portugal and Northeast Brazil

Joao S. Lopes [a,*], Isabel Marques [a], Patricia Soares [a], Hanna Nebenzahl-Guimaraes [a], Joao Costa [a], Anabela Miranda [b], Raquel Duarte [c,d], Adriana Alves [b], Rita Macedo [e], Tonya A. Duarte [f,g], Theolis Barbosa [h], Martha Oliveira [i], Joilda S. Nery [h], Neio Boechat [f], Susan M. Pereira [g], Mauricio L. Barreto [g], Jose Pereira-Leal [a], Maria Gabriela Miranda Gomes [a], Carlos Penha-Goncalves [a]

[a] *Instituto Gulbenkian de Ciencia, 2781-901 Oeiras, Portugal*
[b] *Instituto Nacional de Saude Dr. Ricardo Jorge, 4150-180 Porto, Portugal*
[c] *Centro Hospitalar de Vila Nova de Gaia/Espinho, 4434-502 Vila Nova de Gaia, Portugal*
[d] *Departamento de Epidemiologia Clinica, Medicina Preventiva e Saúde Pública, Faculdade de Medicina da Universidade do Porto, 4200-319 Porto, Portugal*
[e] *Direccao-Geral da Saude, 1049-005 Lisboa, Portugal*
[f] *Instituto de Doenças do Torax, Universidade Federal do Rio de Janeiro, 21.941-913 Rio de Janeiro, Brazil*
[g] *Instituto de Ciencias da Saude, Universidade Federal da Bahia, 40.150-510 Salvador, Brazil*
[h] *Centro de Pesquisas Goncalo Moniz, Fundacao Oswaldo Cruz, 40.296-710 Salvador, Brazil*
[i] *Centro de Pesquisa em Tuberculose, Universidade Federal do Rio de Janeiro, 21.941-913 Rio de Janeiro, Brazil*

ARTICLE INFO

Article history:
Received 14 November 2012
Received in revised form 23 April 2013
Accepted 24 April 2013
Available online 3 May 2013

Keywords:
Mycobacterium tuberculosis complex
Portugal
Brazil
SNP-typing
Spoligotyping
Phylogeny

ABSTRACT

Human tuberculosis is an infectious disease caused by bacteria from the *Mycobacterium tuberculosis* complex (MTBC). Although spoligotyping and MIRU-VNTR are standard methodologies in MTBC genetic epidemiology, recent studies suggest that Single Nucleotide Polymorphisms (SNP) are advantageous in phylogenetics and strain group/lineages identification. In this work we use a set of 79 SNPs to characterize 1987 MTBC isolates from Portugal and 141 from Northeast Brazil. All Brazilian samples were further characterized using spolygotyping. Phylogenetic analysis against a reference set revealed that about 95% of the isolates in both populations are singly attributed to bacterial lineage 4. Within this lineage, the most frequent strain groups in both Portugal and Brazil are LAM, followed by Haarlem and X. Contrary to these groups, strain group T showed a very different prevalence between Portugal (10%) and Brazil (1.5%). Spoligotype identification shows about 10% of mis-matches compared to the use of SNPs and a little more than 1% of strains unidentifiability. The mis-matches are observed in the most represented groups of our sample set (i.e., LAM and Haarlem) in almost the same proportion. Besides being more accurate in identifying strain groups/lineages, SNP-typing can also provide phylogenetic relationships between strain groups/lineages and, thus, indicate cases showing phylogenetic incongruence.

Overall, the use of SNP-typing revealed striking similarities between MTBC populations from Portugal and Brazil.

1. Introduction

Human tuberculosis (TB) is an airborne bacterial disease caused by the *Mycobacterium tuberculosis* complex (MTBC). Currently, WHO estimates that one third of the world's population is infected with this pathogen. From these, a minority progresses to disease, accounting for about 10 million new cases and 2 million deaths per year (WHO, 2011). Recent studies suggest that an increase in prevalence of immunosuppressive diseases (e.g. HIV), population ageing and changes in social patterns are leading to increasing rates of disease activation (Lönnroth et al., 2009). Furthermore, drug-resistance acquisition is also a concern, and reports of bacteria resistant to first and second-lines drugs are growing

* Corresponding author. Address: Instituto Gulbenkian de Ciencia, Apartado 14, 2781-901 Oeiras, Portugal. Tel.: +351 214407900; fax: +351 214407970.
E-mail addresses: j.sollari.lopes@gmail.com (J.S. Lopes), imarques@igc.gulbenkian.pt (I. Marques), psoares@igc.gulbenkian.pt (P. Soares), hanna.guimaraes@gmail.com (H. Nebenzahl-Guimaraes), jcosta@igc.gulbenkian.pt (J. Costa), amiranda@ibmc.up.pt (A. Miranda), raquelafduarte@gmail.com (R. Duarte), aalves@ibmc.up.pt (A. Alves), rita.macedo@dgs.pt (R. Macedo), tonya.duarte@gmail.com (T.A. Duarte), theolis@bahia.fiocruz.br (T. Barbosa), martholiveira@yahoo.com.br (M. Oliveira), joilda_nery@yahoo.com.br (J.S. Nery), n_boechat@yahoo.com (N. Boechat), susanmp@ufba.br (S.M. Pereira), mauricio@ufba.br (M.L. Barreto), jleal@igc.gulbenkian.pt (J. Pereira-Leal), ggomes@igc.gulbenkian.pt (M.G.M. Gomes), cpenha@igc.gulbenkian.pt (C. Penha-Goncalves).

1567-1348/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.meegid.2013.04.028

considerably (Gandhi et al., 2010). Thus, detailed knowledge of MTBC genetic diversity and geographical distribution is becoming of increasing importance.

MTBC genome is characterized by low substitution rates and, consequently, low DNA sequence diversity, while having marked population subdivisions (Hershberg et al., 2008). These pathogens are generally believed to be highly clonal (Achtman, 2008) with rare horizontal gene transfer (Liu et al., 2006; Namouchi et al., 2012), further decreasing the chances of diversity. Several explanations for this were put forward including the isolated life-style inside mammalian cells, the long generation time and the latent stage with little activity (Smith et al., 2003). The lack of genetic diversity in tuberculosis (TB) makes the study of short-term epidemic networks and long-term evolutionary histories particularly difficult with commonly used markers, such as spoligotype patterns (Kamerbeek et al., 1997) and MIRU-VNTR (Supply et al., 2000, 2001). However, these same traits are ideal for phylogenetic studies using vast single nucleotide polymorphism (SNP) data. In fact, the absence of horizontal gene transfer, which can derange phylogenetic trees by attracting far related branches, and the observed slow substitution rates greatly reduce the problem of convergent evolution when constructing phylogenetic trees.

Defining meaningful boundaries between groups in bacteria is complicated, yet this grouping is necessary for strain classification. Various MTBC classification schemes have been proposed in the past, but none reached a clear consensus (Gagneux and Small, 2007). Recently, Comas and co-workers (2009) defined a classification based on whole-genome data that considered the global diversity of MTBC and was phylogenetically robust (Coscolla and Gagneux, 2010). This classification consisted of six main lineages of human-adapted MTBC and one that mostly infects animals. Within the six main lineages, the authors further classified the strains with a second order grouping according to previous spolygotyping classification: lineage 1, 3, 5 and 6 were defined by single comprehensive groups called EAI, CAS, AFRI1 and AFRI2, respectively; lineage 2 was defined by a non-comprehensive group called Beijing; and lineage 4 was composed by 6 groups called Cameroon, Haarlem, LAM, T, Uganda and X.

In this work we used the two-level MTBC classification to identify samples collected from patients from Portugal and Northeast of Brazil. Previous studies on the global population structure of MTBC (Brudey et al., 2006; Gagneux et al., 2006) observed that the most frequent strain in Europe and South America is lineage 4 (comprised mostly by LAM, Haarlem, T and X). Lineage 1 is also found in both regions, although in much lower frequency. Lineage 2, on the other hand, is typically absent from South America [but see (Iwamoto et al., 2012)]. There has also been previous local-scale studies examining MTBC diversity in Portugal (David et al., 2007) and Southern regions of Brazil [Rio Grande do Sul (Borsuk et al., 2005; Scholante Silva et al., 2009), Parana (Malaghini et al., 2009) and São Paulo (Mendes et al., 2011)]. These, however, have been performed using typically less than 100 samples and genotyped only by spoligotypes. In this paper we present and extensive study using more than 2000 samples from Portugal and Northeast of Brazil genotyped using SNP-typing methods. The comparison between TB populations from these two regions can be of great importance given their possible recent shared ancestry. A demographic study on understanding major past population demographic dynamics, such as admixture, ancient population splitting or migratory trends between TB populations, is out of the scope of this work. Nevertheless, a description of the genetic diversity of the two populations may help to shed some light on the question whether one population results from a recent direct expansion of the other or if they have been evolving separately long before the large human population influx between the two continents of the last five to six centuries.

We present a novel methodology to identify MTBC samples using a reference set composed by previously studied MTBC strains, which, as far as we know, are representative of this group's global diversity. This identification is performed via SNP-based phylogenetic trees, using information on monophyletic groups and their ancestry. The construction of these phylogenetic trees allowed us to further characterize the SNPs in respect to their usefulness in identifying MTBC samples. The goal of this characterization was two-folded: to obtain information on these SNPs for future SNP-typing studies; and further exploit strain ancestry and phylogenetic incongruence in our datasets. In addition to the SNP-based classification, we also analyzed the spoligotype patterns of the Brazilian samples and compared their use in MTBC identification in terms of consistency between markers and successfulness of identification.

2. Material and methods

2.1. Sample collections and molecular typing

The dataset from Portugal consisted of 2112 MTBC samples collected between 2001 and 2011 from patients diagnosed with TB in public hospitals in four major Portuguese regions (North, Center, Lisbon and Tagus Valley and South). The dataset from Northeast Brazil consisted of 147 MTBC samples collected between 2008 and 2009 from patients diagnosed with TB in a reference hospital in Salvador, Bahia. The datasets used in this study consisted solely on sequence data and no personal data was disclosed at any point, thus, there was no need to obtain ethical approval for the analysis presented here.

Genotyping was performed from 2009 to 2012 and the SNPs used were selected from a pool of polymorphisms described until then (Dos Vultos et al., 2008; Filliol et al., 2006; Hershberg et al., 2008; Kasai and Ezaki, 2000). From this pool, 80 SNPs located outside genome regions known to be related to resistance to antibiotics were chosen for phylogenetic analyses (see Table S1 for details). SNP genotyping was performed using primer extension chemistry and mass spectrometric analysis on a Sequenom MassArray platform (Gabriel et al., 2009). The genomic sequence was amplified by multiplex polymerase chain reaction (PCR) and amplified product was treated with shrimp alkaline phosphatase and used for allele specific primer extension reaction according to the MassEXTEND protocol. The reaction mixture was then spotted onto a SpectroCHIP microarray and subjected to the MALDI-TOF mass spectrometry. The genotype calls were assigned using SpectroTYPER software from the SNP-specific peaks. Quality control of the genotyping process used *Mycobacterium tuberculosis* strains EAS054, H37Rv, Haarlem, F11, C and CDC1551, which have curated and publicly available genomes.

Microbead-based spoligotyping was performed according to Cowan et al. (2004). In brief, the direct repeat (DR) region was amplified by PCR using previously described primers (Kamerbeek et al., 1997). The amplified DNA was then incubated with a mixture containing microspheres coupled to a set of 43 oligonucleotide probes (corresponding to the spacer sequences of the DR locus). The hybridization of the PCR product to each specific spacer sequence was quantified using a solid phase fluorometer (Luminex, Austin, TX, USA). A spacer was considered to be present in the genome of a given isolate when the ratio between the average median number of relative fluorescence units (MRFU) in the isolate and the MRFU of the negative control (distilled water) exceeded 5.0.

2.2. Construction of a reference set

In order to use SNP data for identification of the collected samples we constructed a reference set by selecting 31 bacterial strains

representative of the global diversity of MTBC and whose classification was already defined. The strains were chosen in order to achieve good coverage of the MTBC global strain phylogenetic tree constructed by Hershberg et al. (2008) and by taking in consideration the availability of data in the TBDB database (http://www.tbdb.org/). Data available in TBDB consist of curated MTBC sequence alignments and whole-genome data. In order to gather the information we simply choose the loci and loci positions where our defined list of SNPs were located in. The reference set was complemented with genotyping information from the set used for quality control. Details of the reference set are summarized in Table S2.

2.3. Data curation via phylogenetic tools

Technical noise and constraints imposed by the DNA sequence genotype assays may generate genotype ambiguity (no-calls) that greatly decrease the support for a phylogenetic tree and, under some conditions, even produce biased trees. For this reason, we devised a 4-step approach to eliminate no-calls, while discarding a minimum amount of samples: step (1) discard the SNPs with more than 5% no-calls, which may indicate an inherent incapability of the technique to retrieve the genotype of the sequence position; step (2) remove samples with more than 5% no-calls, which may indicate low-quality DNA of the sample; step (3) construct a 70% consensus maximum likelihood (ML) tree and, if possible, correct no-calls for the genotype that creates the most parsimonious tree; step (4) remove samples that, after the correcting procedure still have sites with no-calls. Using this procedure, we started with a total of 1.04% no-calls (or a total of 1754 no-calls in all sites of all samples) in the Portuguese dataset and 0.85% no-calls (or a total of 100 no-calls) in the Brazilian dataset, and ended with 0.06% in both the Portuguese (103 no-calls) and the Brazilian (7 no-calls) datasets. In this process, one SNP (position 207 in locus Rv432) was discarded, since 32% and 37% of the samples in the Brazilian and the Portuguese datasets, respectively, generated an ambiguous genotype. A value of 0.06% of no-calls is not uncommon on this sensitive genotyping technique (Bradic et al., 2011) and can be due to damaged DNA samples or to DNA sequence regions which are refractory to amplification. Table 1 presents detailed results using this 4-step approach in terms of number of samples with one or more no-calls. From 2112 and 147 collected samples with ambiguous genotypes from Portugal and Brazil, respectively, we ended up with 1987 and 141 samples with complete information. The genome position of the 79 chosen SNPs and respective genotype frequencies on the selected Portuguese and Brazilian data are depicted in Fig. S1.

The ML tree used to correct the ambiguous genotypes was obtained using only the distinct genotypes of the Portuguese and the Brazilian dataset pooled together, including the ambiguous genotypes, along with the reference set previously put together. Before constructing the tree we used jModelTest 0.1.1 (Guindon and Gascuel, 2003; Posada, 2008) to determine the best fit model of nucleotide evolution using the Akaike information criterion. Following this analysis we chose the simple Hasegawa, Kishino and Yano (HKY) model with no invariable sites and homogeneous substitution rate among sites. The ML tree was obtained using RAxML 7.0.4 (Stamatakis, 2006) and the clade support was evaluated by analyzing 1000 bootstrap pseudo-replicates.

2.4. Identification of strain lineage/group using a phylogenetic analysis

In order to identify strain groups/lineages we constructed phylogenetic trees using the distinct genotypes of the Portuguese and Brazilian data pooled together (40 genotypes) along with the chosen reference set (31 genotypes). We constructed phylogenetic trees using neighbor joining (NJ), ML and Bayesian inference (BI) trees. The final identification of the strains was done using a 50% consensus BI tree by the following approach: using the reference set and assuming monophyletic strain groups, we identified the tree branches corresponding to particular groups; data samples falling in identified branches were then classified accordingly. The NJ tree was obtained using the observed number of changes to calculate distances between strains using Seaview 4 (Gouy et al., 2010), branch support was evaluated using 1000 pseudo-replicates. Before constructing the ML and BI trees we reanalyzed the dataset using jModelTest 0.1.1, again the best fit model of nucleotide evolution was determined to be the simple HKY model. For the ML tree we used again RaxML 7.0.4 and analyzed 1000 pseudo-replicates. The Bayesian analysis was performed using a Markov chain Monte Carlo method implemented in mrbayes 3.2.1 (Ronquist et al., 2012). We used two replicates of 1 million generations with four chains, samples were taken every 1000 generation and the burn-in period was set to be 0.25. Convergence was evaluated using Tracer 1.5 (http://tree.bio.ed.ac.uk/software/tracer/) and the trees were produced using FigTree 1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/).

Following the construction of the MTBC phylogenetic tree, we noticed genotypes showing phylogenetic incongruence across sites. To examine this incongruence, we constructed a recombination network using RECOMB2007 algorithm implemented in SplitsTree4 4.12.3 (Huson and Bryant, 2006) and inspected closely the phylogenetic incompatible sites of each genotype. In order to eliminate the effect of genotyping and/or clerical errors, we excluded genotypes with three or less samples from the analysis (Liu et al., 2006).

2.5. Identification of strain groups using spolygotype data

The 141 SNP-typed Brazilian samples with complete information were also characterized in respect to their spoligotype pattern. Spoligotyping has been shown not to be reliable for phylogenetic analysis, because its patterns may not reflect the evolution history of a strain and the "signature" pattern of the strain groups can be ambiguous or uninformative (Comas et al., 2009). Furthermore, spoligotyping traditionally requires a perfect match between the observed pattern and the "signature" pattern of a group. To circumvent this problem we used a novel software that relies on an extensive spoligotype database to assign new spoligotype patterns to strain groups (SPOTCLUST, Vitol et al., 2006). Strain group identification using spoligotyping was then compared with the strain groups identified with SNP-typing.

Table 1
Total number of samples (and percentage of samples with no-calls in brackets) at the various stages of the 4-step procedure proposed (see details in Material and methods).

Dataset	Initial	Step 1	Step 2	Step 3	Step 4
Brazil	147 (37.41%)	147 (8.16%)	144 (6.25%)	144 (2.08%)	141
Portugal	2112 (44.18%)	2112 (13.78%)	2077 (12.33%)	2077 (4.33%)	1987

3. Results

3.1. Identification of strain lineage/group by co-clustering

Using the SNP-based phylogenetic approach (Fig. 1), we were able to identify the lineage of the strains in the collection and most of the strain groups they belong to (Table 2). The vast majority of genotypes were classified as lineage 4 (34 genotypes); two genotypes were identified as belonging to lineage 1; and lineages 2, 6 and Animal were ascribed one distinct genotype profile each. The lineage of one SNP genotype profile (geno_5) was not identifiable against the reference set. Regarding the strain groups of lineage 4, group T is the most represented with 10 distinct genotypes, followed by group LAM (8 genotypes) and X (4 genotypes). Groups Cameroon, Uganda and Haarlem are all represented by one distinct genotype. Genotypes that were not identified directly from the tree, and that did not show phylogenetic incongruence, were all identified from their SNPs patterns (see Section 3.1.2).

3.1.1. SNP phylogenetic informativeness

The low mutation and recombination rates in MTBC allows us to map the emergence of polymorphisms as single events in a phylogenetic tree. Considering the reference set of 31 global MTBC strains, we were able to classify each SNP in terms of its usefulness to identify a particular strain group (Tables 3 and S3). We considered three types of SNP informativeness: (1) Group-specific SNP, if the polymorphism is uniquely present in all the analyzed strains of a particular strain group; (2) Intra-group SNP, if the polymorphism is only present in some of the considered strains of a particular group; (3) Supra-group SNPs, if the polymorphism is present in strains from two or more strain groups. From the 79 SNPs used we classified 36 as Group-specific, 17 as Intra-group and 22 as Supra-group. Four SNPs were unable to be classified because they were monomorphic in the observed and in the reference sets. It should be noted that the ascertainment of Group-specific SNPs was determined by the genetic diversity represented in this study. Nevertheless, we identified group-specific SNPs for 9 different groups: Animal, Beijing, Cameroon, EAI, Haarlem, LAM, T, Uganda and X.

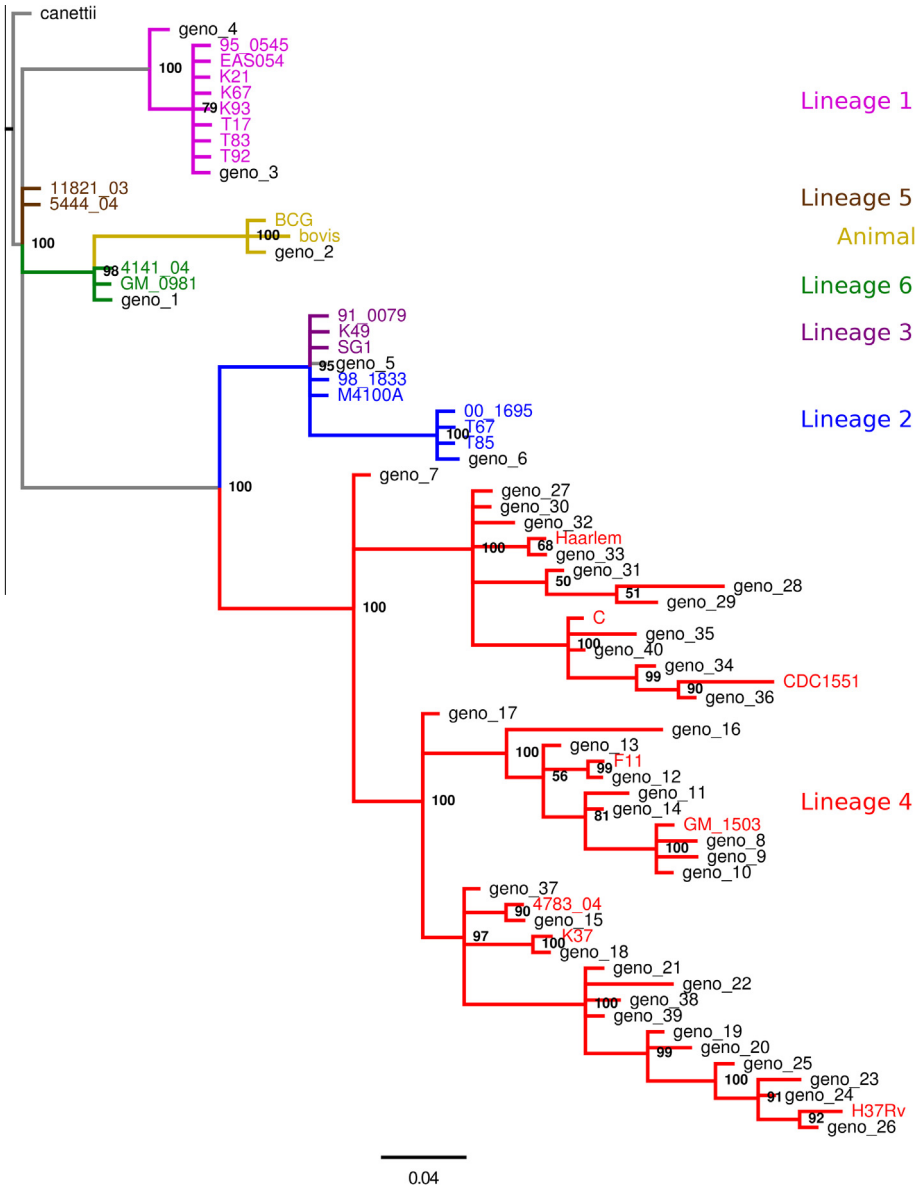

Fig. 1. Bayesian inference phylogeny based on 40 distinct genotypes identified in this study and 31 global MTBC strains previously reported (Hershberg et al., 2008) using 79 variable nucleotide positions. Six main lineages can be observed within the human MTBC as referenced in Comas et al. (2009). The 40 analysed genotypes are scattered along the tree branches, the strain group they belong to can be identified from the global strains by assuming a strictly monophyletic tree.

Table 2
Identification of the lineage and the strain group, suggested by Comas et al. (2009), to which the genotypes present in the dataset belong to.

Genotype	SNP lineage	Strain group
Geno_1	Lineage 6	AFRI1
Geno_2	Animal	
Geno_3	Lineage 1	EAI
Geno_4	Lineage 1	EAI
Geno_5	Lineage 2 or 3	
Geno_6	Lineage 2	Beijing
Geno_7	Lineage 4	
Geno_8	Lineage 4	LAM
Geno_9	Lineage 4	LAM
Geno_10	Lineage 4	LAM
Geno_11	Lineage 4	LAM
Geno_12	Lineage 4	LAM
Geno_13	Lineage 4	LAM
Geno_14	Lineage 4	LAM
Geno_15	Lineage 4	Cameroon
Geno_16	Lineage 4	LAM
Geno_17	Lineage 4	
Geno_18	Lineage 4	Uganda
Geno_19	Lineage 4	T
Geno_20	Lineage 4	T
Geno_21	Lineage 4	T
Geno_22	Lineage 4	T
Geno_23	Lineage 4	T
Geno_24	Lineage 4	T
Geno_25	Lineage 4	T
Geno_26	Lineage 4	T
Geno_27	Lineage 4	
Geno_28	Lineage 4	
Geno_29	Lineage 4	
Geno_30	Lineage 4	
Geno_31	Lineage 4	
Geno_32	Lineage 4	
Geno_33	Lineage 4	Haarlem
Geno_34	Lineage 4	X
Geno_35	Lineage 4	X
Geno_36	Lineage 4	X
Geno_37	Lineage 4	
Geno_38	Lineage 4	T
Geno_39	Lineage 4	T
Geno_40	Lineage 4	X

3.1.2. Identification of strain groups by cluster ancestry

Using a reference set as an identification framework enables the identification of strains in monophyletic groups identical to previously defined strain groups but fails the identification of genotype profiles that do not fall in groups of the reference set. Nevertheless, SNP data has the potential to further characterize these unidentified strains. SNP data provides information on the evolutionary history of the samples, and allows further analysis of phylogenetic relations between unidentified strains and defined strain groups. Nine genotypes geno_7, geno_17, geno_27, geno_28, geno_ 29, geno_30, geno_31, geno_32 and geno_37 were not identified as belonging to a particular strain group (Table 2). Four genotypes showed signs of homoplasy (i.e., geno_28, geno_29, geno_31 and geno_32) and were discarded, but the SNP pattern of the remaining five were examined by crossing against the information contained in Table S3 (see Tables S4–S8). This analysis allowed us to further classify genotypes geno_7, geno_17, geno_31 and geno_37 as ancient to particular strain groups (summarized in Table 4).

3.1.3. Examining phylogenetic incongruence

Ten identified genotypes showed phylogenetic incongruence patterns. The previously performed quality control criteria ensured that this incongruence was not due to systematic genotyping errors. Furthermore, the protocol of the SNP genotyping assures that in case of experimental errors, rather than obtaining the wrong SNP genotype, this analysis will result on a no-call (Bradic et al., 2011). Nevertheless, in order to account for other sources of error (e.g. clerical errors), we considered for analysis of phylogenetic incongruence only strain genotypes with more than 3 samples: geno_8 (117 samples), geno_16 (4 samples) and geno_20 (5 samples). These genotypes correspond to about 6% of all the samples (Table 5) and consist of 15% of all the genotypes with more than 3 samples. To examine the observed phylogenetic incongruence we analyzed the three genotypes using a recombination network approach (Fig. S2). These genotypes were previously identified as LAM (geno_8 and geno_16) and T (geno_20). Interestingly, regarding geno_8 and geno_20, we found only one site with a discrepant strain-group information (Tables S9 and S10). Contrariwise, geno_16, presents five sites characteristic of group X and four polymorphisms with a pattern unique to LAM, its estimated strain-group (Table S11). We also noted that geno_16 carries three LAM-specific SNPs (Rv3062_1212, Rv3084_0729 and Rv3088_1347) that map within a 30 kb region on the MTBC chromosome and two X-specific SNPs (Rv3221_0030 and Rv3261_0905) that map within 44 kb. Although these results are suggestive of recombination events, more robust tests using large sequence data are necessary to disentangle with certainty the origin of the observed phylogenetic discrepancies [e.g. Chi-Squared (Smith, 1999), NSS (Jakobsen and Easteal, 1996) or PHI (Bruen et al., 2006) tests].

Table 3
Characterization of the SNPs used in the study regarding presence in one or several strain groups. Dataset analyzed consisted of strains from Hershberg et al. (2008).

Identify	SNP IDs[a]
Animal and Lineage 6	Rv1375_0318, Rv3075_0588
Lineage 2 and 3	Rv1420_1301, Rv3798_1014
Lineage 2, 3 and 4	Rv3077_1002, Rv0002_0481, Rv0041_0384, Rv1696_0438, Rv3679_0471, Rv3711_0491
Animal	Rv0005_0630, Rv0005_1284, Rv1908_0087, Rv1628_0267, Rv2897_0693
Lineage 4	Rv1317_0034, Rv2979_0041, Rv3297_0229, Rv3711_0227
Cameroon, LAM, T and Uganda	Rv3088_1347, Rv3084_0729
Cameroon, T and Uganda	Rv2560_0628
Haarlem and X	Rv0189_1674, Rv0831_0645, Rv3176_0591, Rv3370_1719
Beijing	Rv0815_0153
Cameroon	Rv2949_0467
EAI	Rv0629_0870, RV1020_0256, Rv2362_0606, Rv3644_0726, Rv3644_0735
LAM	Rv0129_0309, Rv0631_1604, Rv3062_1212
Haarlem	Rv1316_0044, Rv2976_0501
Uganda	Rv0006_0238, Rv2949_0375
T[b]	Rv0006_2003, Rv0034_0165, Rv0083_1800, Rv0260_1047, Rv0956_0591, Rv1040_0243, Rv1056_0489, Rv2567_1473, Rv3077_0924, Rv3581_0075, Rv3731_0938, Rv3799_0027
X	Rv0824_0435, Rv1733_0097, Rv2330_0426, Rv3221_0030

[a] Locus ID and SNP position within the locus.
[b] Dataset contains only one strain of T group, it is not possible to distinguish between group-specific and intra-group SNPs in this group.

Table 4
Further strain group identification of genotypes from lineage 4 which do not show phylogenetic incongruence and were previously unidentified.

Genotype	Strain group
Geno_7	Ancient to Cameroon, Haarlem, LAM, T, Uganda and X groups
Geno_17	Ancient to Cameroon, LAM, T and Uganda groups
Geno_27	X group
Geno_30	Ancient to Haarlem and X groups
Geno_37	Ancient to Cameroon, T and Uganda groups

Table 5
Genotypes with more than 3 samples in the dataset in study that show phylogenetic incongruence patterns.

Genotype ID	Freq (n)	Freq (%)	SNP lineage	Strain group	Incongruent SNPs[a]
Geno_8	117	5.50	Lineage 4	LAM	1
Geno_16	4	0.19	Lineage 4	LAM	6
Geno_20	5	0.23	Lineage 4	T	1

[a] Number of SNPs which show phylogenetic incongruence.

3.2. Comparison between samples from Portugal and Brazil

In order to compare MTBC genetic diversity in datasets with different sample sizes (1987 from Portugal and 141 from Northeast Brazil) we calculated the frequency of each genotype as a percentage of the total samples in each set (Fig. 2, for absolute values see Fig. S3). The Portuguese dataset is composed by a larger number of genotypes than the Brazilian, 36 genotypes against 14, an expected result as the Portuguese dataset is more than 10-fold larger. In fact, discarding rare genotypes (i.e. frequency lower than 1% in the respective dataset), the number of distinct genotypes in the Portuguese and Brazilian datasets is 13 and 10, respectively.

3.3. Identification of strain groups using SNPs vs. spoligotypes

Fig. 3 shows a comparison between strain group identification in our collection using both types of molecular data, SNP-typing and spoligotyping. This comparison focused on the main represented strain group in the datasets (i.e. EAI, Haarlem, LAM, T and X). Following the identification of strain groups using SNP data, we classified the spoligotype identification as consistent or inconsistent with SNP identification or unassigned if the spoligotype pattern was classified as belonging to an unidentified family. SPOTCLUST provided a high assignment rate to strain groups (frequency of unassigned isolates was of about 1%) but the rate of inconsistency with group identification using SNPs was relatively high (almost 10% overall) as found by previous studies (Abadia et al., 2010).

4. Discussion

Constructing a SNP-based BI phylogenetic tree using a MTBC reference set and assuming monophyletic strain groups we were able to map the emergence of SNPs. From this mapping we established the usefulness of each SNP in the identification of a particular strain group (Table S3). This information can be used as a starting point to establish a minimum set of SNPs necessary to classify strains. For example, a set with "Group-specific" SNPs Rv0005_0630 (Animal, Bouakaze et al., 2010), Rv0006_0238 (Uganda, Bouakaze et al., 2010), Rv0006_2003 (T, Comas et al., 2009), Rv0129_0309 (LAM, Comas et al., 2009), Rv0815_0153 (Beijjing, this study), Rv1316_0044 (Haarlem, Comas et al., 2009), Rv2330_0426 (X, Comas et al., 2009), Rv2362_0606 (EAI, Abadia et al., 2010) and Rv2949_0467 (Cameroon, Comas et al., 2009) together with "Supra-group" SNPs Rv1375_0318 (Animal and AFRI1, Filliol et al., 2006) and Rv1420_1301 (lineage 2 and CAS, this study) enables us to identify almost all the MTBC strain groups. The exceptions are to distinguish between CAS and non-Beijing strains and to identify AFRI2 strains. Other combinations can be designed depending on the group or lineage to be identified. Table 3 details the SNPs which are useful to identify particular groups and lineages.

Although a minimum set of SNPs can be of utmost importance for quick identification purposes (Filliol et al., 2006), such an approach can be prone to two sources of errors: weak typing; and ambiguity. The first is due to SNP-typing errors such as producing no-calls or even, albeit rare, typing the wrong nucleotide. The second is due to the possible lack of ubiquitousness of a SNP in a strain group. The molecular identification of MTBC group strains is based on pattern signatures. Ultimately, these patterns derive from dynamic evolutionary forces that generate non-static patterns difficult to compartmentalize. Therefore, unless a strain group is defined by a particular SNP, the existence of a SNP ubiquitous to a group is not likely. Hence, one particular SNP characteristic of a strain group is likely unable to identify all the strains that belong to that group. Our need to differentiate between categories "Group-specific" and "Intra-group" of Table S3 are a good evidence of the potential lack of precision of a minimum set of SNPs. Moreover, another disadvantage of using a minimum set of SNPs is the lack of ability to define phylogenetic distances between strains.

Establishing phylogenetic distances and ancestry between strains can be particularly useful in overcoming assignment ambiguities. In Section 3.1.2 we show how this ancestry can be used to identify unassigned MTBC samples. In fact, using ancestry information we were able to identify most of the strains that were not assigned by the phylogenetic tree alone (i.e., geno_7, geno_17, geno_27, geno_31 and geno_37).

Table S3 can also be useful to further analyze samples when the SNP pattern reveals phylogenetic incongruence. In Section 3.1.3, we show the possibility to classify each SNP as in accordance with the strain-group classification or in disagreement.

The choice of the SNP set is of considerable importance when performing SNP-typing analyses. Using SNPs identified by comparing a limited number of strains or by comparing samples that do not account for global diversity may lead to phylogenetic bias (Achtman, 2008). Our choice of SNPs seems to be appropriate since the obtained phylogenetic tree (Fig. 1) overlaps other MTBC trees obtained from studies that take in account such biases (e.g., Fig. 1 of Hershberg et al., 2008). Nevertheless, because the aim of our study was to identify Portuguese and Brazil MTBC samples we gave particular importance to SNPs capable of identifying strain groups of lineage 4, the most common lineage of the Americas and Europe. We identified two effects of this conditioned SNP selection. The first one is that there seems to be a higher discriminatory power for group T. In fact, eight of its distinct genotypes encompass less than 1% of all samples, while for example the only distinct genotype of group Beijing has a frequency of more than 3%. The second consequence of our SNP selection was the lack of power to distinguish the lineage of geno_5. The lineage of this rare genotype (frequency of about 0.5%) could not be distinguished between lineage 2 or 3. In any case, our choice of SNPs proved to be successfully to not just characterize the strains collected from the regions in study, but also shed some light on their molecular evolutionary background.

In this paper we present an extensive study of the MTBC population of Portugal and Northeast Brazil. In general, our observations are in accordance with earlier studies on regional (Borsuk et al., 2005; David et al., 2007) and worldwide (Brudey et al., 2006; Gagneux et al., 2006) MTBC diversity. We observed that the most frequent strain group on both regions is lineage 4, originally called Europe and Americas lineage (Hershberg et al., 2008), with more

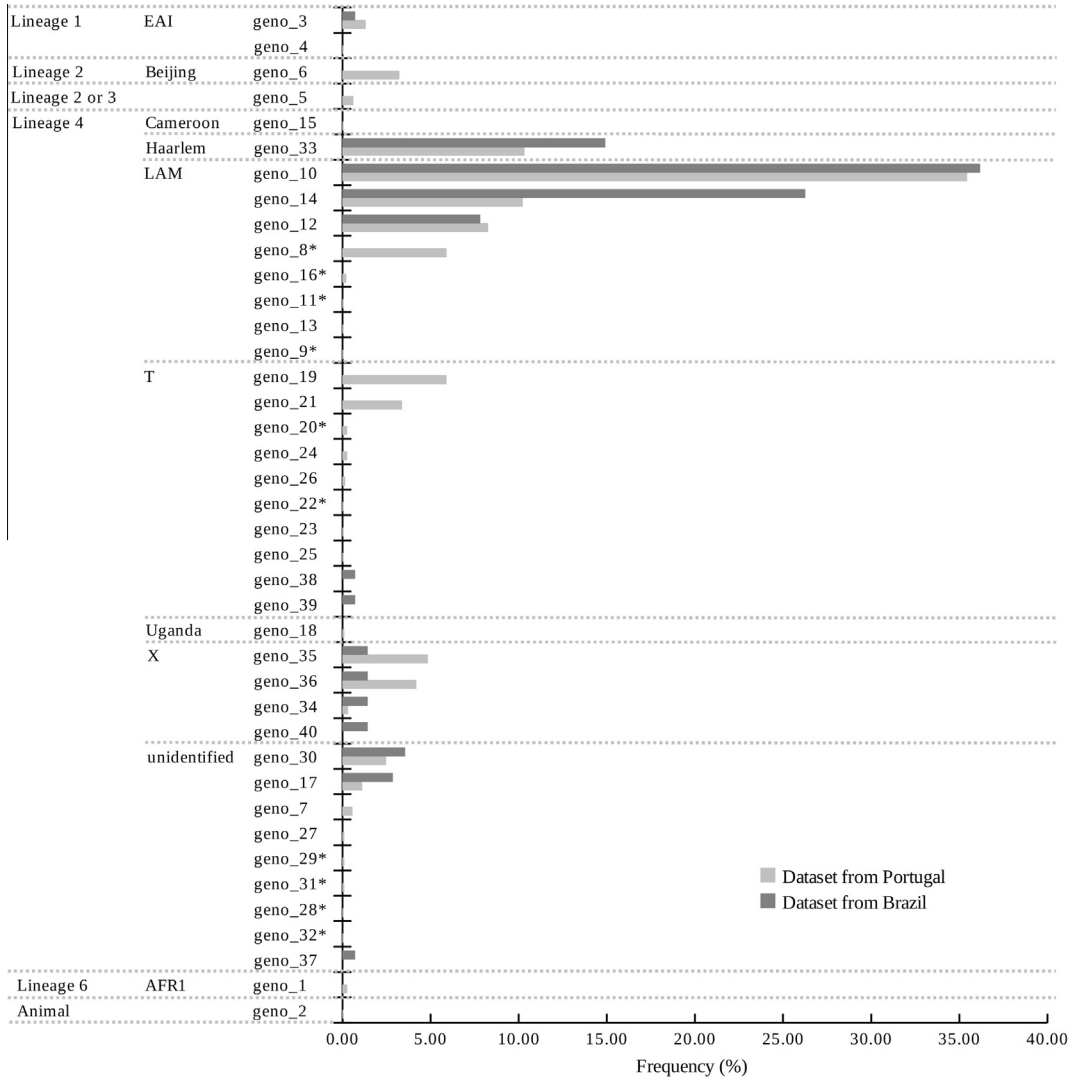

Fig. 2. Frequency distribution of the 40 genotypes when considering the Portuguese (1915 samples; light grey) and Brazilian (141 samples; dark grey) MTBC samples separately. The genotypes are grouped by strain group and SNP lineages as defined by Comas et al. (2009). The frequency of the genotypes is measured as the percentage within its belonging population. Genotypes marked with ∗ show phylogenetic incongruence in at least one SNP.

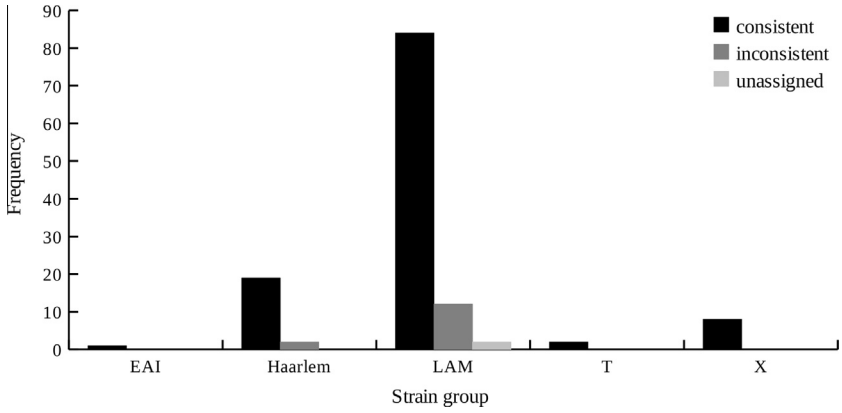

Fig. 3. Comparison between strain group identification using SNP-typing and spoligotyping on the Brazilian dataset (141 sampes). The samples were clustered in strain groups using SNP data. From these samples, 10 could not be identified. The samples were further grouped in three categories: consistent – spolygotyping identification is consistent with SNP-typing identification; inconsistent – spoligotyping identification is not consistent with SNP-typing identification; and unassigned – samples were not assign to any group of spoligotypes. The strain group identification using spolygotypes data was performed using SPOTCLUST (Vitol et al., 2006).

than 90% frequency. Lineage 1 is also present in both populations equally, but in considerably low frequencies (around 1%). These two populations, however, differ regarding lineage 2. This lineage, also called East Asia lineage (Hershberg et al., 2008), has a significant presence in Portugal (more than 3%), yet it is absent from the Northeast Brazil dataset (but see Gomes et al., 2012). This observation is not unexpected as lineage 2 is prevalent in Europe possibly due to recent migration. Another difference between Portuguese and Brazilian datasets regards lineage 6 or West Africa lineage (Hershberg et al., 2008). This lineage, also absent in the Northeast Brazil, is present in Portugal although in a very low frequency (0.25%). A possible explanation for its presence in Portugal is the existence of significant immigration from Africa to this country in the last 30 years.

Concerning lineage 4, the most frequent strain group in both Portugal and Brazil is by far LAM with more than 60%, and is followed by Haarlem (more than 10%) and then X (more than 6%). More interesting than the similarities between Portugal and Brazil, however, are their differences. For example, although the Portuguese dataset sampling was more comprehensive than the Brazilian, we have found four genotypes that only appear in Brazil, i.e. geno_37, geno_38, geno_39 and geno_40. These genotypes belong to different strain groups and fall in different parts of the phylogenetic tree (Fig. 1), so a single common origin of the four genotypes is ruled out. On the other hand, there are genotypes present in Portugal that are absent from Northeast of Brazil. Strain geno_8, for example, has a frequency of 7% in Portugal and was not found in Brazil. Similarly, genotypes geno_19 and geno_21 are absent in the Brazilian collection but account for about 7% of samples in Portugal and constitute almost all the samples of group T in this country. This yields a significant difference in the frequency of strain group T between Portugal (about 10%) and Brazil (1%). Strains geno_15 (strain group Cameroon) and geno_18 (strain group Uganda) are also observed in the Portuguese dataset, while being absent from Northeast Brazil. Their absence in the smaller Brazilian dataset is not unexpected since only 2 samples for geno_15 and 1 for geno_18 were observed in Portugal.

Despite these differences, overall, the datasets from both countries are quite similar (Fig. 2). These similarities suggest that, indeed, MTBC populations from Portugal and Brazil have been in close contact for some period of time not so long ago. This was likely due to either recent shared ancestry or massive migration events.

In this study we have also performed a comparison between strain groups identification using spoligotyping and SNP-typing. We observed that spoligotyping, in general, provided a good identification of the samples irrespective of the strain group they belong to. In fact, the rate of spoligotype identification consistent with the SNP-typing identification was about 88%, an identification rate similar to previews studies (Abadia et al., 2010). Furthermore, the methodology implemented in the software provided almost no unassigned patterns, proving itself particularly useful in this respect. Nonetheless, SNP-typing seems to be the best option when needing to identify MTBC strains because this technique can also provide the evolutionary background of the studied strains. As we demonstrated along the paper, having access to the molecular evolution history of the samples is of utmost importance to categorize them with precision, as well as, to further interpret the observations.

Acknowledgements

We thank two anonymous reviewers and editor for thoughtful comments which raise the quality of the original manuscript. The work was funded by the Portuguese Foundation for Science and Technology (FCT) and by the European Commission [grant EC-ICT-231807]. Data collection in Brazil was supported by National Counsel of Technological and Scientific Development (CNPQ) [Project Number 410498/2006-8] and Coordination of Improvement of Higher Education Personnel (CAPES) [Project Number 23038.005107/2011-83].

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2013.04.028.

References

Abadia, E., Zhang, J., Dos Vultos, T., Ritacco, V., Kremer, K., Aktas, E., Matsumoto, T., Refregier, G., Van Soolingen, D., Gicquel, B., Sola, C., 2010. Resolving lineage assignation on *Mycobacterium tuberculosis* clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method. Infect. Genet. Evol. 10, 1066–1074.

Achtman, M., 2008. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu. Rev. Microbiol. 62, 53–70.

Borsuk, S., Dellagostin, M.M., Madeira, S.D.G., Lima, C., Boffo, M., Mattos, I., Almeida da Silva, P.E., Dellagostin, O.A., 2005. Molecular characterization of *Mycobacterium tuberculosis* isolates in a region of Brazil with a high incidence of tuberculosis. Microbes Infect. 7, 1338–1344.

Bouakaze, C., Keyser, C., De Martino, S.J., Sougakoff, W., Veziris, N., Dabernat, H., Ludes, B., 2010. Identification and genotyping of *Mycobacterium tuberculosis* complex species by use of a SNaPshot Minisequencing-based assay. J. Clin. Microbiol. 48, 1758–1766.

Bradic, M., Costa, J., Chelo, I.M., 2011. Genotyping with sequenom. In: Orgogozo, V., Rockman, M.V. (Eds.), Molecular Methods for Evolutionary Genetics. Humana Press, New York, pp. 193–210.

Brudey, K., Driscoll, J.R., Rigouts, L., Prodinger, W.M., Gori, A., Al-Hajoj, S.A., Allix, C., Aristimuño, L., Arora, J., Baumanis, V., Binder, L., Cafrune, P., Cataldi, A., Cheong, S., Diel, R., Ellermeier, C., Evans, J.T., Fauville-Dufaux, M., Ferdinand, S., Garcia de Viedma, D., Garzelli, C., Gazzola, L., Gomes, H.M., Guttierez, M.C., Hawkey, P.M., Van Helden, P.D., Kadival, G.V., Kreiswirth, B.N., Kremer, K., Kubin, M., Kulkarni, S.P., Liens, B., Lillebaek, T., Ho, M.L., Martin, C., Martin, C., Mokrousov, I., Narvskaïa, O., Ngeow, Y.F., Naumann, L., Niemann, S., Parwati, I., Rahim, Z., Rasolofo-Razanamparany, V., Rasolonavalona, T., Rossetti, M.L., Rüsch-Gerdes, S., Sajduda, A., Samper, S., Shemyakin, I.G., Singh, U.B., Somoskovi, A., Skuce, R.A., Van Soolingen, D., Streicher, E.M., Suffys, P.N., Tortoli, E., Tracevska, T., Vincent, V., Victor, T.C., Warren, R.M., Yap, S.F., Zaman, K., Portaels, F., Rastogi, N., Sola, C., 2006. *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. 6, 23.

Bruen, T.C., Philippe, H., Bryant, D., 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics 172, 2665–2681.

Comas, I., Homolka, S., Niemann, S., Gagneux, S., 2009. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. PLoS ONE 4, e7815.

Coscolla, M., Gagneux, S., 2010. Does *M. tuberculosis* genomic diversity explain disease diversity? Drug Discov. Today: Dis. Mech. 7, e43–e59.

Cowan, L., Diem, L., Brake, M., Crawford, J., 2004. Transfer of a *Mycobacterium tuberculosis* genotyping method, Spoligotyping, from a reverse line-blot hybridization, membrane-based assay to the Luminex multianalyte profiling system. J. Clin. Microbiol. 42, 474–477.

David, S., Ribeiro, D.R., Antunes, A., Portugal, C., Sancho, L., De Sousa, J.G., 2007. Contribution of spoligotyping to the characterization of the population structure of *Mycobacterium tuberculosis* isolates in Portugal. Infect. Genet. Evol. 7, 609–617.

Dos Vultos, T., Mestre, O., Rauzier, J., Golec, M., Rastogi, N., Rasolofo, V., Tonjum, T., Sola, C., Matic, I., Gicquel, B., 2008. Evolution and diversity of clonal bacteria: the paradigm of *Mycobacterium tuberculosis*. PLoS ONE 3, e1538.

Filliol, I., Motiwala, A., Cavatore, M., Qi, W., 2006. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy. J. Bacteriol. 188, 759–772.

Gabriel, S., Ziaugra, L., Tabbaa, D., 2009. SNP genotyping using the Sequenom MassARRAY iPLEX platform. Curr. Protoc. Hum. Genet. 60, 2.12.11–12.12.18.

Gagneux, S., Small, P.M., 2007. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. Lancet 7, 328–337.

Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., De Jong, B.C., Narayanan, S., Nicol, M., Niemann, S., Kremer, K., Gutierrez, M.C., Hilty, M., Hopewell, P.C., Small, P.M., 2006. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc. Natl. Acad. Sci. USA 103, 2869–2873.

Gandhi, N.R., Nunn, P., Dheda, K., Schaaf, H.S., Zignol, M., Van Soolingen, D., Jensen, P., Bayona, J., 2010. Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. Lancet 375, 1830–1843.

Gomes, H.M., Elias, A.R., Cardoso Oelemann, M.A., Da Silva Pereira, M.A., Onofre, F.F., Marsico, A.G., Kritski, A.L., Filho Ldos, A., Caldas, P.C., Possuelo, L.G., Cafrune, P., Rosseti, M.L., Lucena, N., Saad, M.H.F., Cavalcanti, H.R., Leite, C.Q.F., De Brito, R.C.,

Lopes, Lima, K., Souza, M., Trindade, Rde.C., Zozio, T., Sola, C.S., Rastogi, N., Suffys, P.N. 2012. Spoligotypes of *Mycobacterium tuberculosis* complex isolates from patients residents of 11 states of Brazil. Infect. Genet. Evol. 12, 649–656.

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27, 221–224.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.

Hershberg, R., Lipatov, M., Small, P.M., Sheffer, H., Niemann, S., Homolka, S., Roach, J.C., Kremer, K., Petrov, D.A., Feldman, M.W., Gagneux, S., 2008. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. PLoS Biol. 6, e311.

Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267.

Iwamoto, T., Grandjean, L., Arikawa, K., Nakanishi, N., Caviedes, L., Coronel, J., Sheen, P., Wada, T., Taype, C.A., Shaw, M.-A., Moore, D.A., Gilman, R.H., 2012. Genetic diversity and transmission characteristics of Beijing family strains of *Mycobacterium tuberculosis* in Peru. PLoS ONE 7, e49651.

Jakobsen, I.B., Easteal, S., 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Comput. Appl. Biosci. 12, 291–295.

Kamerbeek, J., Schouls, L., Kolk, A., Van Agterveld, M., Van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., Van Embden, J., 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J. Clin. Microbiol. 35, 907–914.

Kasai, H., Ezaki, T., 2000. Differentiation of phylogenetically related slowly growing mycobacteria by their gyrB sequences. J. Clin. Microbiol. 38, 301–308.

Liu, X., Gutacker, M.M., Musser, J.M., Fu, Y.-X., 2006. Evidence for recombination in *Mycobacterium tuberculosis*. J. Bacteriol. 188, 8169–8177.

Lönnroth, K., Jaramillo, E., Williams, B.G., Dye, C., Raviglione, M., 2009. Drivers of tuberculosis epidemics: the role of risk factors and social determinants. Soc. Sci. Med. 68, 2240–2246.

Malaghini, M., Brockelt, S.R., Burger, M., Kritski, A., Thomaz-Soccol, V., 2009. Molecular characterization of *Mycobacterium tuberculosis* isolated in the State of Parana in southern Brazil. Tuberculosis 89, 101–105.

Mendes, N.H., Melo, F.A., Santos, A.C., Pandolfi, J.R., Almeida, E.A., Cardoso, R.F., Berghs, H., David, S., Johansen, F.K., Espanha, L.G., Leite, S.R., Leite, C.Q., 2011. Characterization of the genetic diversity of *Mycobacterium tuberculosis* in São Paulo city, Brazil. BMC Res. Notes 4, 269.

Namouchi, A., Didelot, X., Schöck, U., Gicquel, B., Rocha, E.P.C., 2012. After the bottleneck: Genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation, recombination, and natural selection. Genome Res. 22, 721–734.

Posada, D., 2008. JModelTest: phylogenetic model averaging. Mol. Biol. Evol. 25, 1253–1256.

Ronquist, F., Teslenko, M., Van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539–542.

Scholante Silva, A.B., Von Groll, A., Félix, C., Conceição, F.R., Spies, F.S., Scaini, C.J., Rossetti, M.L., Borsuk, S., Dellagostin, O.A., Almeida da Silva, P.E., 2009. Clonal diversity of M. tuberculosis isolated in a sea port city in Brazil. Tuberculosis 89, 443–447.

Smith, J.M., 1999. The detection and measurement of recombination from sequence data. Genetics 153, 1021–1027.

Smith, N.H., Dale, J., Inwald, J., Palmer, S., Gordon, S.V., Hewinson, R.G., Smith, J.M., 2003. The population structure of *Mycobacterium bovis* in Great Britain: clonal expansion. Proc. Natl. Acad. Sci. USA 100, 15271–15275.

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.

Supply, P., Mazars, S., Lesjean, S., Vincent, V., Gicquel, B., Locht, C., Mazars, E., 2000. Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. Mol. Microbiol. 36, 762–771.

Supply, P., Lesjean, S., Savine, E., 2001. Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. J. Clin. Microbiol. 39, 3563–3571.

Vitol, I., Driscoll, J., Kreiswirth, B., Kurepina, N., Bennett, K.P., 2006. Identifying *Mycobacterium tuberculosis* complex strain families using spoligotypes. Infect. Genet. Evol. 6, 491–504.

WHO, 2011. Global Tuberculosis Control: WHO report 2011. World Health Organization, Geneva, Switzerland; Report No: WHO/HTM/TB/2011.16.