Mycobacterium tuberculosis genetic diversity in Portugal and Northeast Brazil

Joao S. Lopes[1], Isabel Marques[1], Patricia Soares[1], Hanna Nebenzahl-Guimaraes[1], Joao Costa[1], Anabela Miranda[2], Raquel Duarte[3], Adriana Alves[2], Rita Macedo[4], Antonio Fonseca-Antunes[4], Tonya A. Duarte[6, 7], Theolis Barbosa[5], Martha Oliveira[8], Joilda S. Nery[5], Neio Boechat[6], Susan M. Pereira[7], Mauricio L. Barreto[7], Jose Pereira-Leal[1], M. Gabriela M. Gomes[1], Carlos Penha-Goncalves[1]

1 - Instituto Gulbenkian de Ciencia, 2781-901 Oeiras, Portugal.

2 - Instituto Nacional de Saude Dr. Ricardo Jorge, 4150-180 Porto, Portugal.

3 - Sociedade Portuguesa de Pneumologia, 1069-130 Lisboa, Portugal.

4 - Direccao-Geral da Saude, 1049-005 Lisboa, Portugal .

5 - Centro de Pesquisas Goncalo Moniz, Fundacao Oswaldo Cruz, 40.295-001 Salvador, Brazil.

6 - Instituto de Doenças do Torax, Universidade Federal do Rio de Janeiro, 21.941-913 Rio de Janeiro, Brazil

7 - Instituto de Saude Coletiva, Universidade Federal da Bahia, 40.110-179 Salvador, Brazil.

8 - Centro de Pesquisa em Tuberculose, Universidade Federal do Rio de Janeiro, 21.941-913 Rio de Janeiro, Brazil.

E-mail addresses:

J.S. Lopes, j.sollari.lopes@gmail.com

I. Marques, imarques@igc.gulbenkian.pt

P. Soares, psoares@igc.gulbenkian.pt

H. Guimaraes, hanna.guimaraes@gmail.com

J. Costa, jcosta@igc.gulbenkian.pt

A. Miranda, amiranda@ibmc.up.pt

R. Duarte, raquelafduarte@gmail.com

A. Alves,

R. Macedo, rita.macedo@dgs.pt

A.F. Antunes, afantunes@dgs.pt

T.A. Duarte, tonya.duarte@gmail.com

T. Barbosa, theolis@bahia.fiocruz.br

M. Oliveira,

J.S. Nery, joilda_nery@yahoo.com.br

S.M. Pereira, susanmp@ufba.br

N. Boechat, n_boechat@yahoo.com

M.L. Barreto, mauricio@ufba.br

J. Pereira-Leal, jleal@igc.gulbenkian.pt

M.G.M. Gomes, ggomes@igc.gulbenkian.pt

C. Penha-Goncalves, cpenha@igc.gulbenkian.pt

Corresponding author: Joao Sollari Lopes; Instituto Gulbenkian de Ciencia, Apartado 14, 2781-901 Oeiras, Portugal; Tel: +351 214407900; Fax: +351 214407970; e-mail: j.sollari.lopes@gmail.com

Abstract:

Human tuberculosis is an infectious disease caused by bacteria from the *Mycobacterium tuberculosis* complex (MTBC). Although Spoligotyping and MIRU-VNTR are standard methodology in MTBC genetic epidemiology, recent studies suggest that Single Nucleotide Polymorphisms (SNP) are advantageous in phylogenetics and strain identification. In this work we use a set of 79 SNPs to characterize 1915 MTBC isolates from Portugal and 141 from Northeast Brazil. All Brazilian samples and a subset of 111 Portuguese isolates were further characterized using Spolygotyping. Phylogenetic analysis against a reference set revealed that about 95% of the isolates in both populations were singly attributed to bacterial lineage 4. Within this lineage, the most frequent strain groups in both Portugal and Brazil were LAM, followed by Haarlem and then X. Contrarily to these groups, strain group T showed a very different prevalence between Portugal and Brazil, with a frequency of 7% and less than 1.5%, respectively. Taking the classification by SNPs as the most accurate, we analyze the performance of Spoligotyping in strain identification. The former marker shows about 12% of miss-matches and less than 1% of unidentifiable strains. The miss-matches are observed in the most represented groups of our sample set (i.e., LAM, Haarlem and T) in almost the same proportion. Besides being more accurate in strain identification, SNPs typing can also provide phylogenetic relationships between the strain groups. Indeed, using this molecular markers we were able to observe possible signs of rare recombination events in Mtb.

Overall, the use of SNP typing revealed striking similarities between MTBC populations from Portugal and Brazil. In addition this marker suggest that, albeit rare, recombination events in MTBC are likely to occur.

Keywords:

Mycobacterium tuberculois complex; Portugal; Brazil; SNP typing; Spoligotyping; Phylogeny.

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 becoming a concern, and reports of bacteria resistant to first and second-lines drugs are growing 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 with rare horizontal gene transfer (Liu et al., 2006), further decreasing the chances of diversity. This 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. However, these same traits are ideal for phylogenetic studies using vast single nucleotide polymorphism (SNP) data. Horizontal gene transfer can derange phylogenetic trees by attracting far related branches, while fast substitution rates lead to higher chances of convergent evolution,.

Currently, the gold standard of epidemiological genotyping in TB is based on Spoligotype patterns (Kamerbeek et al., 1997) and MIRU-VNTR (Supply et al., 2000, 2001). These systems have limited use for phylogenetics and strain identification because, due to targeting polymorphisms with fast substitution rate and using limited number of markers, they may present events of convergent evolution, or homoplasy, (Comas et al., 2009). Furthermore, the use of Spoligotyping to identify genotype profiles absent from the reference database is debatable (but see Vitol et al., 2006).

Nevertheless, their usage strongly shaped the classification of MTBC strains. Defining meaningful boundaries between groups in bacteria is complicated, yet this grouping is necessary for strain classification. For MTBC, various classification schemes have been proposed in the past, but none got a clear consensus (Gagneux and Small, 2007). Recently, however, 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 consists 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 was defined by a single comprehensive group called EAI; Lineage 2 was defined by a non-comprehensive group called Beijing; Lineage 3 was defined by a comprehensive group called CAS; Lineage 4 was composed by 6 groups called Cameroon, Haarlem, LAM, T, Uganda and X; Lineage 5 was defined by a comprehensive group called AFRI2;and Lineage 6 was defined by a comprehensive group called AFRI1. This second-order grouping retains historical nomenclature, providing connection to previous studies. In this work we use this two-level MTBC classification proposed by Comas and co-workers.

The global population structure of MTBC has been analyzed using different genetic markers [e.g. large sequence polymorphisms (Gagneux et al., 2006), Spoligotypes (Brudey et al., 2006)]. Overall, populations of MTBC have been found to be highly geographically structured. In fact, previous studies shown a strong agreement between the geographical origin of a human host and the MTBC strain carried (Hirsh et al., 2004). In this work we analyzed MTBC samples collected from patients from Portugal and Northeast of Brazil. In general, the most frequent strain found in Europe and South America is Lineage 4 (comprised mostly by LAM, Haarlem, T and X). Lineage 1 (EAI) is also found in both regions, although in much lower frequency. Lineage 2, in the other hand, is typically absent from South America. 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. 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 classification of the samples was done at two levels: SNP main lineages; and, within these lineages, main historically identified strain groups. The construction of these phylogenetic trees allowed us to further characterize the used SNP 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. Analyzing phylogenetic incongruence, we devised a rudimentary process for differentiating homoplasy and recombination events based on the number of incongruencies. In addition to the SNP-based classification, we also analyzed the Spoligotype patterns of the samples and compared their use in MTBC identification in terms of consistency between markers and successfulness of identification. The datasets analyzed were from Portugal and Northeast of Brazil. Previous characterizations of MTBC populations in these regions were performed using Spoligotype data. In here we provide for the first time a study of Portuguese and Brazilian MTBC diversity using SNP typing methods. The comparison between TB populations in these two countries is of particular importance because of their possible recent shared ancestry.

2. Material and methods

2.1. Sample collections and molecular typing

The dataset from Portugal consisted of 2112 MTBC samples collected from patients diagnosed with TB

in public hospitals in four major Portuguese regions (North, Center, Lisbon and Tagus Valley and South). The dataset from Brazil consisted of 147 MTBC samples collected from patients diagnosed with TB in Salvador, Bahia.

The SNP used for genotyping were selected from a pool of previously described polymorphisms (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 analyzes (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.

Spoligotyping was carried out as previously described (Kamerbeek et al., 1997). In brief, the direct repeat (DR) region was amplified by PCR using primers derived from the DR sequence. The amplified DNA was then hybridized to a set of 43 oligonucleotides derived from the defined spacer sequences within the DR locus by reverse line blotting.

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 from public databases 32 bacterial strains that are representative of the global diversity of MTBC and whose classification is already defined. The strains to use were chosen in order to have a good coverage of the MTBC global strain phylogenetic tree constructed by Hersheberg et al. (2008) and by taking in consideration the availability of whole-genome sequence data of the strains. The primary source of data was the TBDB database (http://www.tbdb.org/). Whenever necessary, these data

was complemented using NCBI databases (http://www.ncbi.nlm.nih.gov/). Data available in TBDB is composed by sequence alignments of a previously defined list of loci. In order to gather the information we simply choose the loci where our defined list of SNPs were located in. Whenever the required loci information was not available, we used data from NCBI. The data from NCBI was collected firstly by gathering all the identified locus sequences of H37Rv, the reference strain, and then using stand-alone ncbi-blastn 2.2.25+ executable (Astchul et al., 1990) to search for H37Rv-like sequences of the required loci in NCBI databases "Nucleotide collection" and "Whole-genome shotgun contigs". The search results were carefully examined and the corresponding sequences were fetched using the Python package Biopython (Cock et al., 2009). The collected sequences were further aligned using muscle 3.8.31 (Edgar, 2004). Finally, the required SNP information for each strain was collected by taking in consideration its respective position in H37Rv in the obtained alignments. The identification details of this reference set are summarized in Table S2. The strains are classified according to the two-level classification system described previously.

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 a biased tree. For this reason, we devised a 4-steps 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 0.11% (or 173) in the Portuguese dataset and 0.06% (or 7) in the Brazilian dataset. One SNP (position 207 in locus Rv432) was discarded in step 1, since 32% and 37% of the samples in the Brazilian and the Portuguese datasets, respectively, generate an ambiguous genotype. Table 1 presents detailed results using this 4-step procedure in terms of number of samples with one or more no-calls. From 2112 and 147 collected samples from Portugal and Brazil, respectively, with ambiguous genotypes we end up with 1915 and 141 samples with complete information. After discarding one SNP we end up with a total of 79 SNPs distributed randomly through the all MTBC genome (Figure 1). The 70% consensus ML trees 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 with an 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. Strain identification using a phylogenetic analysis

For strain identification with SNP data we constructed a phylogenetic tree using the distinct genotypes of the Portuguese and Brazilian data pooled together along with the chosen reference set. We used a total of 2056 collected samples comprising 40 distinct genotypes. To these data we added 32 strains from the reference set obtaining a total of 72 distinct genotypic profiles. Using these data 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. 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 reanalyze 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 (Rambaut and Drummond, 2009), following software recommendation. All the trees were produced using FigTree 1.3.1 (Rambaut, 2009). From the 50% consensus BI tree we identified the genotypes of the collected samples by assuming monophyletic strain groups.

2.5 Detection of recombination

Following the construction of the MTBC phylogenetic tree, we noticed genotypes showing phylogenetic incongruencies across sites. These incongruencies can be due either to homoplasy or recombination events. In order to disentangle these two causes we constructed a recombination network using RECOMB2007 algorithm implemented in SplitsTree4 4.12.3 (Huson and Bryant, 2006), as well as, inspected closely the phylogenetic incompatible sites of each genotype. In order to eliminate the effect of possible genotyping errors on the construction of the recombination network, genotypes with frequency of less than three copies in the collection of samples were excluded (Liu et al., 2006).

2.6. Strain identification using spolygotype data

From the 1915 SNP typed Portuguese samples with complete information, 111 were also characterized in respect to their spoligotype pattern. As for the 141 SNP typed Brazilian samples with complete information, we determine the spoligotyping pattern of each one. These 252 samples were identified through their spoligotype profile using SPOTCLUST (Vitol et al., 2006). Strain identification using

spoligotyping was then compared with the group strains identified with SNP typing.

3. Results

3.1. Strain identification by co-clustering

Constructing a tree of the 32 samples pooled together with the 40 distinct genotype profiles of the observed data, we were able to identify the lineage of the strains in the collection and most of the strain group they belong to (Table 2). Figure 1 presents the 50% consensus BI tree of the 72 MTBC strains. Strain identification was performed by assuming monophyletic strain groups. 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. The remaining genotypes were not identified directly from the tree. However, after careful examination of the SNP data, we were able to identify all the remaining genotypes that did not show signs of homoplasy (see section 3.1.2).

Figure 2 presents the classification of the 40 distinct genotypes of the observed dataset, as well as, the frequency of each genotype. We observed that Lineage 4 is not only the most represented in respect to distinct genotypes, but also in total number of samples. In fact, almost 95% of the samples belong to this lineage. Interestingly, although this lineage is composed of more than 30 genotypes, the ten most frequent ones account for more than 90% of all samples pooled together. Regarding the sample frequency of each strain group of Lineage 4, LAM is by far the most frequent, accounting for 62.5 % of all samples. This group is followed by Haarlem and X with 11% and 9%, respectively. Group T, composed by 10 distinct genotypes, only accounts for about 6% of all samples. This discrepancy between number of distinct genotypes and total sample frequencies suggests that the selection of the

SNPs set imposed a higher discriminatory power for group T as it benefited from information of H37Rv, the most studied MTBC strain, which belongs to group T. Finally, Lineage 1 with 27 samples and Lineage 2 with 69 represent 1.3% and 3.3% of the samples, respectively

3.1.1. SNP phylogenetic informativeness

MTBC species provide a good example of clonal evolution (Smith et al., 2003). In fact, up until recently there was almost no evidence of recombination of these species (Liu et al., 2006; Namouchi et al., 2012). Because of this, in general, the emergence of a particular polymorphism in MTBC species can be mapped as a single event in a phylogenetic tree. Considering the reference set of 32 global MTBC strains, we were able to classify each SNP of our chosen set of SNPs in terms of its usefulness to identify a particular strain group (Table 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, 16 as Intra-group and 22 as Supra-group. Five SNPs were unable to be classified because they were monomorphic in the reference set. 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.

3.1.2. Strain identification by cluster ancestry

The method for strain identification we used takes advantage of monophyletic groups in phylogenetic trees. Using a reference set of strains 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_30 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 4 and S4-S7). 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 5). Genotype geno_27 presents a good example of the need for several SNP markers for the same strain group. This genotype is identified as belonging to strain group X, since it contains three of the polymorphisms that identify this group (Table 4). However, some other polymorphisms that were verified to be specific to X were not present in geno_27. This genotype is, then, likely to have branched earlier than the strains identified as X. Nevertheless, it does not share any polymorphism with strains of the group Haarlem, the phylogenetically closest strain group of X. Hence, we considered it to belong to strain group X. Possibly, geno_27 does not form a monophyletic group with the remaining X strains owing to the conservative consensus rate we chose for the BI tree.

3.1.3. Resolving phylogenetic incongruence

Convergent evolution, or homoplasy, in a phylogenetic tree arises when one polymorphism is found in different parts of the tree and shows an incongruent phylogenetic pattern. Horizontal gene transfer among far related strains can also cause phylogenetic incongruence. In general, horizontal gene transfer (i.e., recombination) is revealed by shared regions of the genome presenting similarities between unrelated strains, whereas homoplasic events are typically characterized by sharing of single polymorphisms.

Ten identified genotypes showed phylogenetic incongruence patterns (Table 3). To gain insight in the

origin of the observed phylogenetic incongruencies we used a recombination network approach (Figure 3), and selected for analysis the genotypes that were represented in more than 3 samples (i.e., geno_8, geno_15 and geno_22). These genotypes were previously identified as LAM (geno_8 and geno_15) and T (geno_22). This analysis yielded supportive evidence for homoplasic events in geno_8 and geno_22 (Tables 6-7) and suggested recombination in geno_15. Geno_8 has six polymorphisms unique to the LAM group, while one site has a polymorphism characteristic of different groups. Geno_22 is identified as belonging to strain group T by seven Group-specific sites, while one site showed discrepant information. Thus, phylogenetic incongruence in these two genotypes was attributed to one site, a scenario compatible with homoplasic events. As for geno_15, four polimorphisms reveal a pattern unique to group LAM while five sites were characteristic of group X (Table 8). This mosaic of information is typical of recombination events. Furthermore, crossing this information with the physical location of the SNPs in Figure 1 suggest that more than one recombination breakpoint is in the origin of this genotype. Although these results are suggestive, more robust tests using sequence data would be necessary to disentangle with certainty the origin of the observed phylogenetic discrepancies.

3.2. Comparison between samples from Portugal and Brazil

In this study we analyzed MTBC isolates collected from Portugal (1915) and Northeast Brazil (141). In order to compare MTBC genetic diversity in both datasets we calculated the frequency of each genotype as a percentage of the total samples in each set (Figure 4). The Portuguese dataset is composed by a larger number of genotypes than the Brazilian, 36 genotypes against 14, an expected results 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. Interestingly, four genotypes found in the Brazilian dataset were not present in the Portuguese, namely geno_37, geno_38, geno_39 and geno_40. Conversely, geno_8 had close to 7% frequency in the Portugal dataset but was not present in the Brazil

collection. Analysis of phylogenetic representation reveals that the vast majority of isolates belongs to Lineage 4, 94% in the Portuguese dataset and 99% in the Brazilian. Lineage 1 is also present in both datasets, although in very low frequencies. As for Lineage 2, we found 69 samples in the Portugal collection (accounting for more than 3% of the total samples), whereas in the Brazil sample this lineage is absent. Finally, Lineage 6 and Animal are also exclusive of the Portuguese dataset with 5 and 1 samples, respectively.

Regarding the sample frequency of each strain group of Lineage 4, LAM is still by far the most frequent, accounting for 62% and 70% of all samples in the Portuguese and Brazilian datasets, respectively. This group is followed by Haarlem with 11% in Portugal and 15% in Brazil. The next most frequent group in both datasets is X with 10% and 6% in Portugal and Brazil respectively. Group T is the one with most striking differences between datasets, in Portugal 7% of the samples are from this group, whereas in Brazil only 1% belong to it. Finally, the Portuguese dataset is also composed by 1 sample of group Cameroon and 2 of group Uganda but owing to the low frequency of these groups, their absence in the smaller Brazilian dataset is not unexpected.

3.3. Strain identification using SNPs vs. spoligotypes

To compare identification consistency between our SNP set with the Spoligotyping method we made use of 252 isolates for which Spoligotype pattern was available. 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). To circumvent these problems we used a novel software that relies on an extensive Spolygotype database to assign new Spoligotype patterns to strain groups (SPOTCLUST, Vitol et al., 2006). Figure 5 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 strain identification using SNP

data, we classified the Spoligotyping 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 is less than 1% of all isolates) but rate of inconsistency with group identification using SNPs was relatively high (approximately 12% in high frequency strain groups) as found by previous studies (Abadia et al., 2010).

4. Discussion

To analyze the genetic diversity of our MTBC collection we implemented a methodology that uses a set of previously identified MTBC strains (reference set) to further characterize the collected isolates. Assuming monophily of MTBC strain groups, we are able to identify the strain group to which a sample belongs to, provided that the sample is located within a branch of the phylogenetic tree corresponding to a particular strain group in the reference set. The required phylogenetic trees were build using molecular data from SNP typing.

Contrary to other common markers, such as Spoligotypes, VNTRs or RFLPs, SNPs are found in many sites across the MTBC genomes (in the order of tens of thousands) and can easily preserve traces of the evolutionary background of a sample (Comas et al., 2009). Furthermore, the evolutionary dynamics of MTBC species are particularly benficial for SNP-based phylogenetic analyses: on one hand, the substitution rate of MTBC is considerably low compared to other bacteria, which reduces the chance of convergent evolution in the absence of selection (Achtman, 2008); on the other hand, horizontal gene transfer has very limited expression in MTBC, reducing further the chance of phylogenetic inconguences. Several explanations for this were put forward including the isolated living style inside mammalian cells, the long generation time and the latent stage with little activity (Smith et al., 2003). After the construction of the NJ, ML and BI phylogenetic trees, we were able to map the origin of the 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 Lineage 2 strains and to identify AFRI2 strains.

Although a minimum set of SNPs can be of utmost importance for quick identification purposes (Filliol et al., 2006), such 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 not to be able 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 incongruencies. Apart from typing errors, these incongruencies can be due to either convergent evolution of genotypes or horizontal gene transfer between far related strains. Statistical tests that aim to disentangle these two causes for incompatible SNP patterns make use of sequence data [e.g. Chi-Squared (Smith, 1999), NSS (Jakobsen and Easteal, 1996) or PHI (Bruen et al., 2006) tests]. Our analyzed data, however, are composed by SNPs scattered along the genome (Table S1), making these tests unsuitable. Instead, we devised a crude method to distinguish between the two events. Crossing data from Table S3 with the genotypes showing incongruent patterns, we were able to specify the strain group information of each SNP. Two of the analyzed genotypes were characterized by one single SNP with incompatible group identification. We propose that convergent evolution or homoplasy may underlie the occurrence of these genotypes. The third analyzed genotype, in the other hand, was characterized by four SNPs characteristics of group LAM and one incompatible with it, and five characteristic of group X and four incompatible with X (Table 8). Although our results do not specify possible break points of recombination events, they suggest occurrence of horizontal gene transfer in this genotype. Sequence data from this strain would be informative to determine whether recombination event(s), known to be rare in MTBC, occurred in this genotype.

Choosing 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 (Figure 1) overlaps other MTBC trees obtained from studies that take in account such biases (e.g., Figure 1 of Hershberg et al., 2008). Nevertheless, because the aim of our study was to identify Portuguese and Brazil MTBC samples we gave particularly 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 bias on the number of distinct genotypes of group T. In fact, eight of its distinct genotypes encompasses 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 observation are in accordance with regional (Borsuk et al., 2005; David et al., 2007) and worldwide (Brudey et al., 2006; Gagneux et al., 2006) earlier studies. 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 than 90% frequency. Lineage 1 is also present in both populations equally, but in considerably low frequencies (around 1%). Within 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 and then X. More interesting than the similarities between Portugal and Brazil, however, is 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 phylognetic tree (Figure 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, the genotypes geno_20 and geno_21 are absent in the Brazil collection but account for about 7% of samples in Portugal and constitute almost all the samples of group T in Portugal. This yields a significant difference on the frequency of this strain group between Portugal and Brazil. Finally, another striking difference between both populations regards samples of Lineage 2. This lineage, previously called East

Asia lineage (Hershberg et al., 2008), have a significant presence in Portugal of more than 3%; however they are completely absent from the Northeast Brazil dataset. This observation is not unexpected as Lineage 2 is regularly prevalent in Europe possibly due to recent migration. Despite these differences, overall, the datasets from both countries are quite similar (Figure 3).

In this study we have also performed a comparison between strain identification using Spoligotyping and SNP typing. To this end, part of the Potuguese and Brazilian datasets were also identified by their spoligotype profile using SPOTCLUST (Vitol et al., 2006). The spoligotyping identification was then compared with the group strains identified with 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 86%, 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

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. Infection, genetics and evolution 10, 1066–1074.

Achtman, M., 2008. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annual review of microbiology 62, 53–70.

Astchul, S., Gish, W., Miller, W., Myers, E., 1990. Basic local alignment tool. Journal of molecular biology 215, 403–410.

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 and infection 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. Journal of clinical microbiology 48, 1758–66.

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 microbiology 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.

Cock, P.J., Antao, T., Chang, J.T., Chapman, B., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M.J.L., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423.

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 discovery today. Disease mechanisms 7, e43–e59.

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. Infection, genetics and evolution 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.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research 32, 1792–1797.

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. Journal of bacteriology 188, 759–772.

Gabriel, S., Ziaugra, L., Tabbaa, D., 2009. SNP genotyping using the Sequenom MassARRAY iPLEX platform. Current protocols in human genetics 60, 2.12.11–12.12.18.

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. Proceedings of the National Academy of Sciences of the United States of America 103, 2869–2873.

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

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. The Lancet infectious diseases 375, 1830–43.

Gouy, M., Guindon, S., Gascuel, O., 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular biology and evolution 27, 221–224.

Guindon, S., Gascuel, O., 2003. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Systematic Biology 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 biology 6, e311.

Hirsh, A.E., Tsolaki, A.G., DeRiemer, K., Feldman, M.W., Small, P.M., 2004. Stable association between strains of Mycobacterium tuberculosis and their human host populations. Proceedings of the National Academy of Sciences of the United States of America 101, 4871–4876.

Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Molecular biology and evolution 23, 254–267.

Jakobsen, I.B., Easteal, S., 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Computer applications in the biosciences 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. Journal of clinical microbiology 35, 907–914.

Kasai, H., Ezaki, T., 2000. Differentiation of phylogenetically related slowly growing mycobacteria by their gyrB sequences. Journal of clinical microbiology 38, 301–308.

Liu, X., Gutacker, M.M., Musser, J.M., Fu, Y.-X., 2006. Evidence for recombination in Mycobacterium tuberculosis. Journal of bacteriology 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. Social science & medicine 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 research 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 research 22, 721–734.

Posada, D., 2008. jModelTest: phylogenetic model averaging. Molecular biology and evolution 25, 1253–1256.

Rambaut, A., 2009. FigTree v1.3.1.

Rambaut, A., Drummond, A.J., 2009. Tracer v1.5.

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. Systematic Biology 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. Proceedings of the National Academy of Sciences of the United States of America 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., Lesjean, S., Savine, E., 2001. Automated high-throughput genotyping for study of global epidemiology of Mycobacterium tuberculosis based on mycobacterial interspersed repetitive units. Journal of clinical microbiology 39, 3563–3571.

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. Molecular microbiology 36, 762–771.

Vitol, I., Driscoll, J., Kreiswirth, B., Kurepina, N., Bennett, K.P., 2006. Identifying Mycobacterium tuberculosis complex strain families using spoligotypes. Infection, genetics and evolution 6, 491–504.

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

Tables

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 (7.80%)	1915

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	LAM
geno_16	Lineage 4	Cameroon
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

Table 3 - Genotypes in the dataset in study that show homoplasy.

Genotype ID	Freq (n)	Freq (%)	SNP Lineage	Strain group	Homoplasic SNPs[a]
geno_8	117	5.69	Lineage 4	LAM	1
geno_10	1	0.05	Lineage 4	LAM	1
geno_12	1	0.05	Lineage 4	LAM	1
geno_15	4	0.19	Lineage 4	LAM	6
geno_19	1	0.05	Lineage 4	T	4
geno_22	5	0.24	Lineage 4	T	1
geno_28	1	0.05	Lineage 4	?[b]	1
geno_29	2	0.10	Lineage 4	?[b]	4
geno_30	1	0.05	Lineage 4	?[b]	4
geno_32	2	0.10	Lineage 4	?[b]	2

[a] - number of SNPs which show evidence of homoplasy.

[b] - strain group unidentified.

Table 4 - Identification of genotype geno_27 using SNPs relevant to Haarlem and X.

SNP ID[a]	Observed bp[b]	Group-specific[c]	Intra-group[d]	Strain group[e]
Rv0189_1674	A		Haarlem and X(A)	Haarlem or X
Rv0824_0435	G	X(G)		X
Rv0831_0645	T		Haarlem and X(T)	Haarlem or X
Rv1316_0044	C	Haarlem (G)		not Haarlem
Rv1733_0097	C	X(T)		not X
Rv2330_0426	C	X (T)		not X
Rv2976_0501	G	Haarlem (A)		not Haarlem
Rv3176_0591	G		Haarlem and X (G)	Haarlem or X
Rv3221_0030	G	X (A)		not X
Rv3370_1719	T		Haarlem and X (T)	Haarlem or X

[a] - locus ID and SNP position within the locus.

[b] - base pair of the genotype considered.

[c] - unique to a family and present in all strains.

[d] - unique to more than one family.

[e] - identification of the strain group of the genotype based on the considered SNP.

Table 5 - Further strain group identification of genotypes from Lineage 4 which do not show signs of homoplasy 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_31	ancient to Haarlem and X groups
geno_37	ancient to Cameroon, T and Uganda groups

Table 6 - Inconsistent strain group identification of genotype geno_8 (identified as LAM) when considering relevant SNPs (1 of which showing evidences of homoplasy).

SNP ID[a]	Observed bp[b]	Group-specific[c]	Intra-group[d]	Supra-group[e]	Strain group[f]
Rv0129_0309	A	LAM (A)			LAM
Rv0631_1604	T	LAM (T)			LAM
Rv1411_0027	C		LAM (C)		LAM
Rv1884_0047	G		LAM (G)		LAM
Rv2959_0207	A		LAM (A)		LAM
Rv3062_1212	G	LAM (G)			LAM
Rv3084_0729	T			LAM (T)	LAM
Rv3088_1347	G			LAM (G)	LAM
Rv3176_0591[g]	G			Haarlem and X (G)	Haarlem or X

[a] - locus ID and SNP position within the locus.

[b] - base pair of the genotype considered.

[c] - unique to a family and present in all strains.

[d] - unique to a family but not present in all strains.

[e] - unique to more than one family.

[f] - identification of the strain group of the genotype based on the considered SNP.

[g] - SNP showing homoplasy.

Table 7 - Inconsistent strain group identification of genotype geno_22 (identified as T) when considering different SNPs (1 of which showing evidences of homoplasy).

SNP ID[a]	Observed bp[b]	Group-specific[c]	Supra-group[d]	Strain group[e]
Rv0006_2003	G	T(G)[g]		T
Rv0034_0165	C	T(C)[g]		T
Rv0083_1800	T	T(T)[g]		T
Rv1056_0489	T	T(T)[g]		T
Rv2560_0628	G		T (G)	T
Rv3084_0729[f]	C		T (T)	not T
Rv3088_1347	G		T (G)	T
Rv3581_0075	A	T(A)[g]		T
Rv3731_0938	G	T(G)[g]		T
Rv3799_0027	T	T(T)[g]		T

[a] - locus ID and SNP position within the locus.

[b] - base pair of the genotype considered.

[c] - unique to a family but and present in all strains.

[d] - unique to more than one family.

[e] - identification of the strain group of the genotype based on the considered SNP.

[f] - SNP showing homoplasy.

[g] - 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 8 - Inconsistent strain group identification of genotype geno_15 (identified as LAM) when considering relevant SNPs (6 of which showing evidences of homoplasy).

SNP ID[a]	Observed bp[b]	Group-specific[c]	Intra-group[d]	Supra-group[e]	Strain group[f]
Rv0129_0309[g]	G	LAM (A)			not LAM
Rv0189_1674	G			X(A)	not X
Rv0631_1604	T	LAM(T)			LAM
Rv0831_0645	A			X(T)	not X
Rv1733_0097[g]	T	X(T)			X
Rv2030_0111[g]	T		X (T)		X
Rv2330_0426[g]	T	X (T)			X
Rv3062_1212	G	LAM (G)			LAM
Rv3084_0729	T			LAM (T)	LAM
Rv3088_1347	G			LAM (G)	LAM
Rv3176_0591	A			X (G)	not X
Rv3221_0030[g]	A	X (A)			X
Rv3261_0905[g]	T		X (T)		X
Rv3370_1719	C			X (T)	not X

[a] - locus ID and SNP position within the locus.

[b] - base pair of the genotype considered.

[c] - unique to a family and present in all strains.

[d] - unique to a family but not present in all strains.

[e] - unique to more than one family.

[f] - identification of the strain group of the genotype based on the considered SNP.

[g] - SNP showing homoplasy.

Figures

Figure 1 - Location in the circular genome of MTBC of the 79 SNPs used to identify the strain groups of the collected dataset. The bars represent the proportion of SNPs with the same morphology as the reference strain H37Rv.

Figure 2 - Bayesian inference phylogeny based on 40 distinct genotypes identified in this study and 32 global MTBC strains previously reported (*, Hersheberg 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 monophiletic tree.

Figure 3 - Frequency distribution of the 40 genotypes when considering the Portuguese and Brazilian MTBC samples pooled together (2056 samples). The genotypes are grouped by strain group and SNP lineages as defined by Comas et al. (2009). The frequency of the genotypes is measured in total numbers. Genotypes marked with * show signs of homoplasy in at least one SNP.

Figure 4 - Recombination network based on the 20 distinct genotypes with more than 3 samples and 32 global MTBC strains (*) using 79 variable nucleotide positions. Genotypes geno_8 and geno_15 were identified as showing signs of recombination.

Figure 5 - 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 signs of homoplasy in at least one SNP.

Figure 6 - Comparison between strain identification using SNP typing and Spoligotyping on Portuguese (111 samples) and Brazilian (141 samples) datasets comprised of both genetic markers. The samples were clustered in strain groups using SNP data. From these samples, 14 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 identification using spolygotypes data was performed using SPOTCLUST (Vitol et al., 2006).

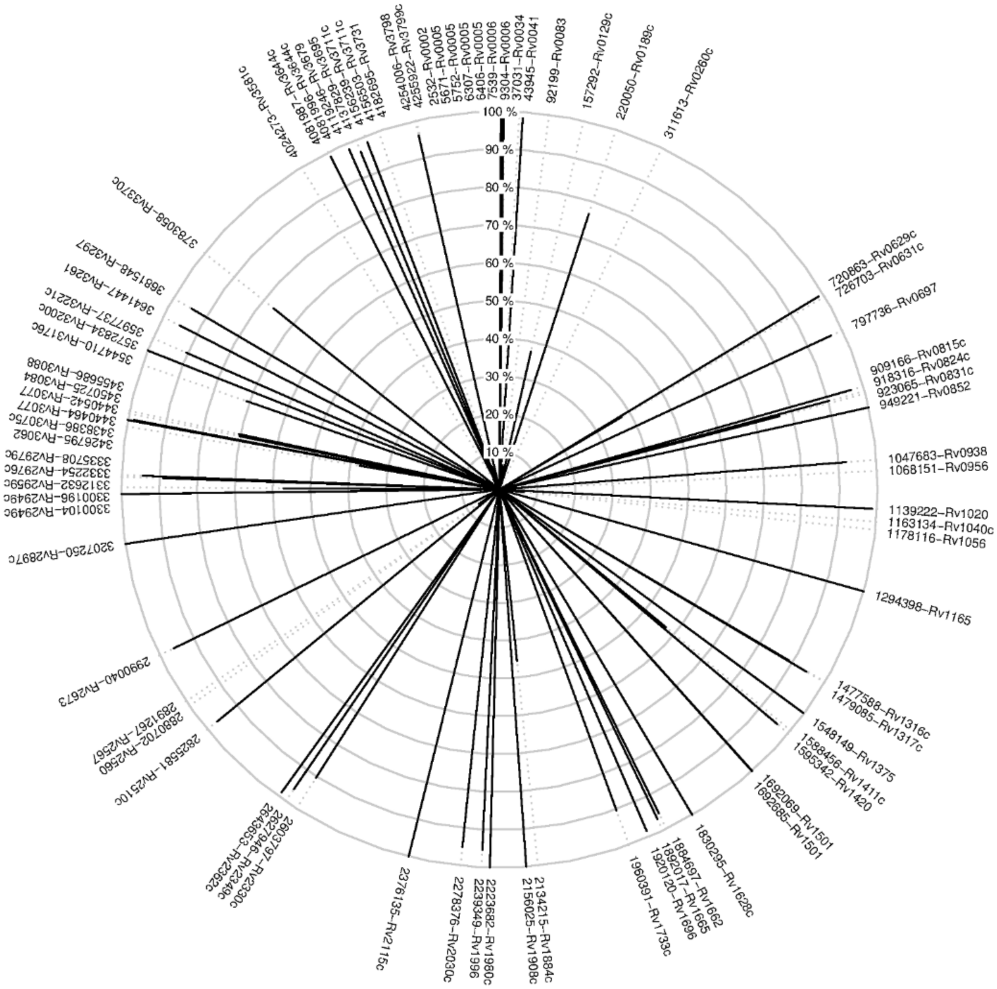
100 %
90 %
80 %
70 %
60 %
50 %
40 %
30 %
20 %
10 %
4024273–Rv3581c
4081987–Rv3644c
4081996–Rv3644c
4119246–Rv3679
4137829–Rv3695
4156239–Rv3711c
4156503–Rv3711c
4182695–Rv3731
4254006–Rv3798
4255922–Rv3799c
2532–Rv0002
5671–Rv0005
5752–Rv0005
6307–Rv0005
6406–Rv0005
7539–Rv0006
9304–Rv0006
37031–Rv0034
43945–Rv0041
92199–Rv0083
157292–Rv0129c
220050–Rv0189c
311613–Rv0260c
720863–Rv0629c
726703–Rv0631c
797736–Rv0697
909166–Rv0815c
918316–Rv0824c
923065–Rv0831c
949221–Rv0852
1047683–Rv0938
1068151–Rv0956
1139222–Rv1020
1163134–Rv1040c
1178116–Rv1056
1294398–Rv1165
1477588–Rv1316c
1479085–Rv1317c
1548149–Rv1375
1588456–Rv1411c
1595342–Rv1420
1692069–Rv1501
1692685–Rv1501
1830295–Rv1628c
1884697–Rv1662
1892017–Rv1665
1920120–Rv1696
1960391–Rv1733c
2134215–Rv1884c
2156025–Rv1908c
2223682–Rv1980c
2239349–Rv1996
2278376–Rv2030c
2376135–Rv2115c
2605797–Rv2330c
2627946–Rv2349c
2643853–Rv2362c
2825581–Rv2510c
2880702–Rv2560
2891267–Rv2567
2990040–Rv2673
3207250–Rv2897c
3300104–Rv2949c
3300196–Rv2949c
3312632–Rv2959c
3332254–Rv2976c
3335708–Rv2979c
3426795–Rv3062
3438386–Rv3075c
3440464–Rv3077
3440542–Rv3077
3450725–Rv3084
3455686–Rv3088
3544710–Rv3176c
3572834–Rv3200c
3597737–Rv3221c
3641447–Rv3261
3681548–Rv3297
3783058–Rv3370c

Fig. 1

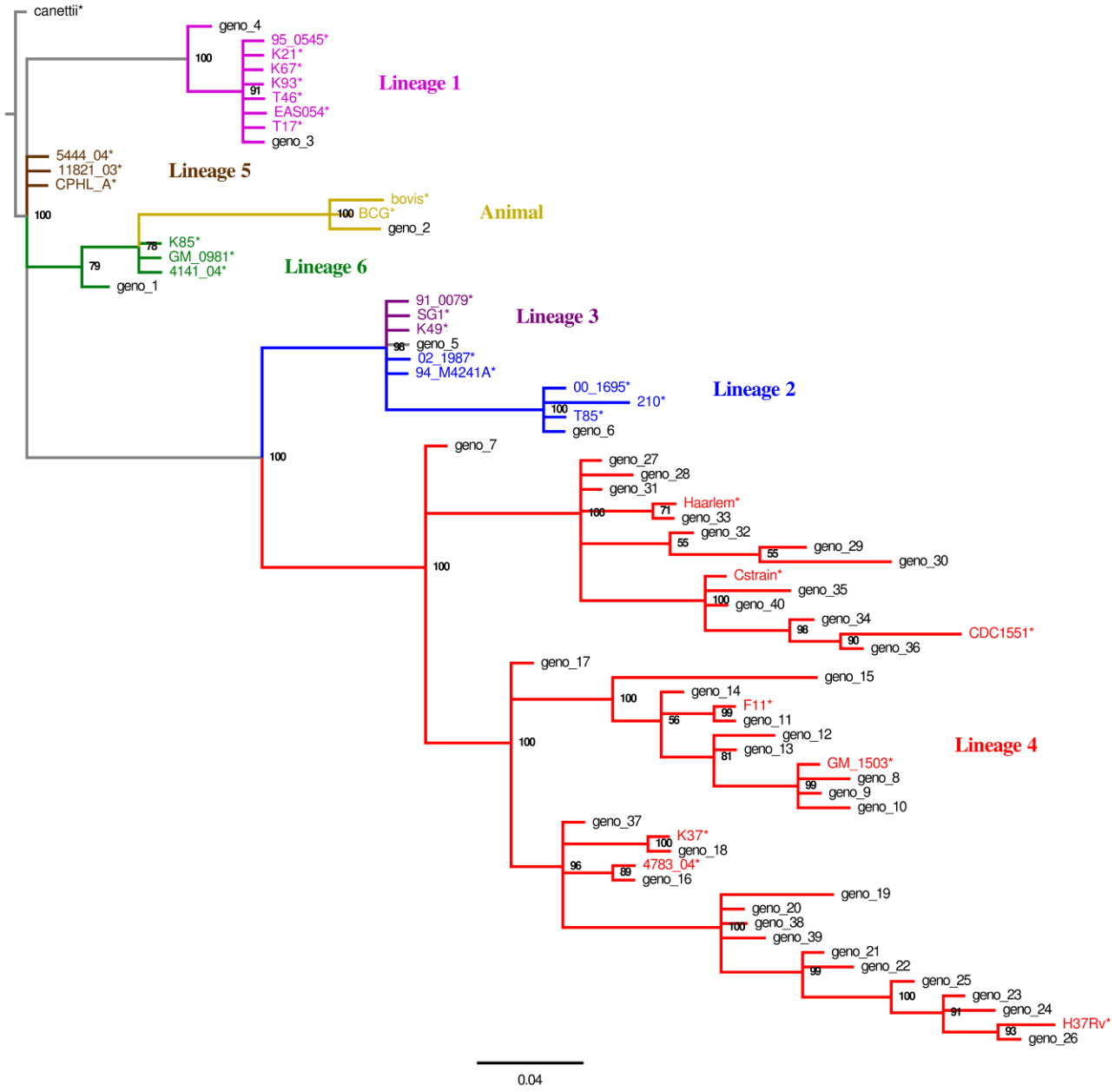
canettii*
geno_4
95_0545*
K21*
K67*
K93*
T46*
EAS054*
T17*
geno_3
Lineage 1
5444_04*
11821_03*
CPHL_A*
Lineage 5
bovis*
BCG*
geno_2
Animal
K85*
GM_0981*
4141_04*
Lineage 6
geno_1
91_0079*
SG1*
K49*
Lineage 3
geno_5
02_1987*
94_M4241A*
00_1695*
210*
Lineage 2
T85*
geno_6
geno_7
geno_27
geno_28
geno_31
Haarlem*
geno_33
geno_32
geno_29
geno_30
Cstrain*
geno_35
geno_40
geno_34
CDC1551*
geno_36
geno_17
geno_15
geno_14
F11*
geno_11
geno_12
geno_13
Lineage 4
GM_1503*
geno_8
geno_9
geno_10
geno_37
K37*
geno_18
4783_04*
geno_16
geno_19
geno_20
geno_38
geno_39
geno_21
geno_22
geno_25
geno_23
geno_24
H37Rv*
geno_26
0.04

Fig. 2

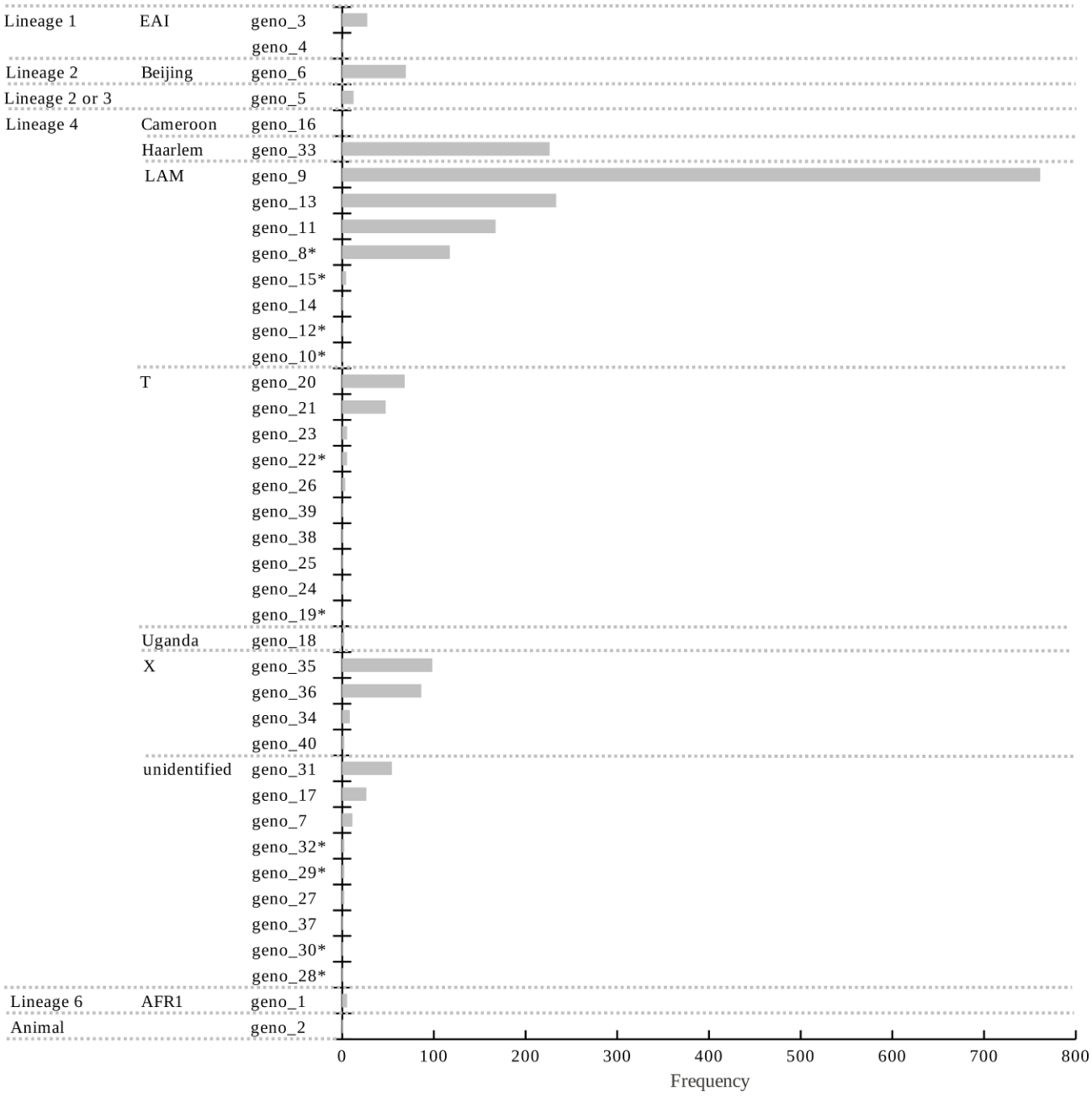
Lineage 1
EAI
geno_3
geno_4
Lineage 2
Beijing
geno_6
Lineage 2 or 3
geno_5
Lineage 4
Cameroon
geno_16
Haarlem
geno_33
LAM
geno_9
geno_13
geno_11
geno_8*
geno_15*
geno_14
geno_12*
geno_10*
T
geno_20
geno_21
geno_23
geno_22*
geno_26
geno_39
geno_38
geno_25
geno_24
geno_19*
Uganda
geno_18
X
geno_35
geno_36
geno_34
geno_40
unidentified
geno_31
geno_17
geno_7
geno_32*
geno_29*
geno_27
geno_37
geno_30*
geno_28*
Lineage 6
AFR1
geno_1
Animal
geno_2
0
100
200
300
400
500
600
700
800
Frequency

Fig. 3

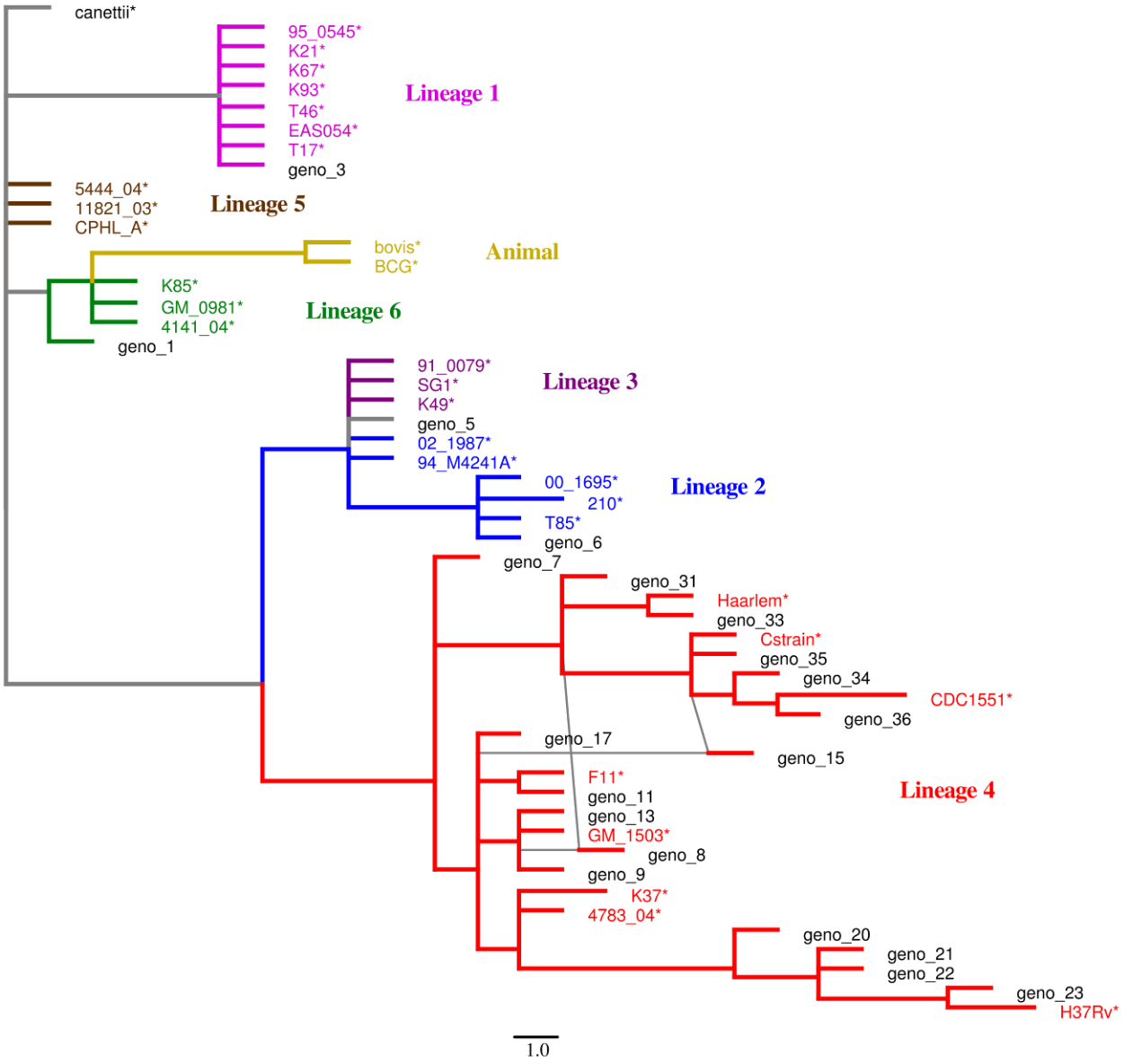
canettii*
95_0545*
K21*
K67*
K93*
T46*
EAS054*
T17*
geno_3
Lineage 1
5444_04*
11821_03*
CPHL_A*
Lineage 5
bovis*
BCG*
Animal
K85*
GM_0981*
4141_04*
geno_1
Lineage 6
91_0079*
SG1*
K49*
geno_5
Lineage 3
02_1987*
94_M4241A*
00_1695*
210*
T85*
geno_6
Lineage 2
geno_7
geno_31
Haarlem*
geno_33
Cstrain*
geno_35
geno_34
CDC1551*
geno_36
geno_17
geno_15
F11*
geno_11
geno_13
GM_1503*
geno_8
geno_9
K37*
4783_04*
Lineage 4
geno_20
geno_21
geno_22
geno_23
H37Rv*
1.0

Fig. 4

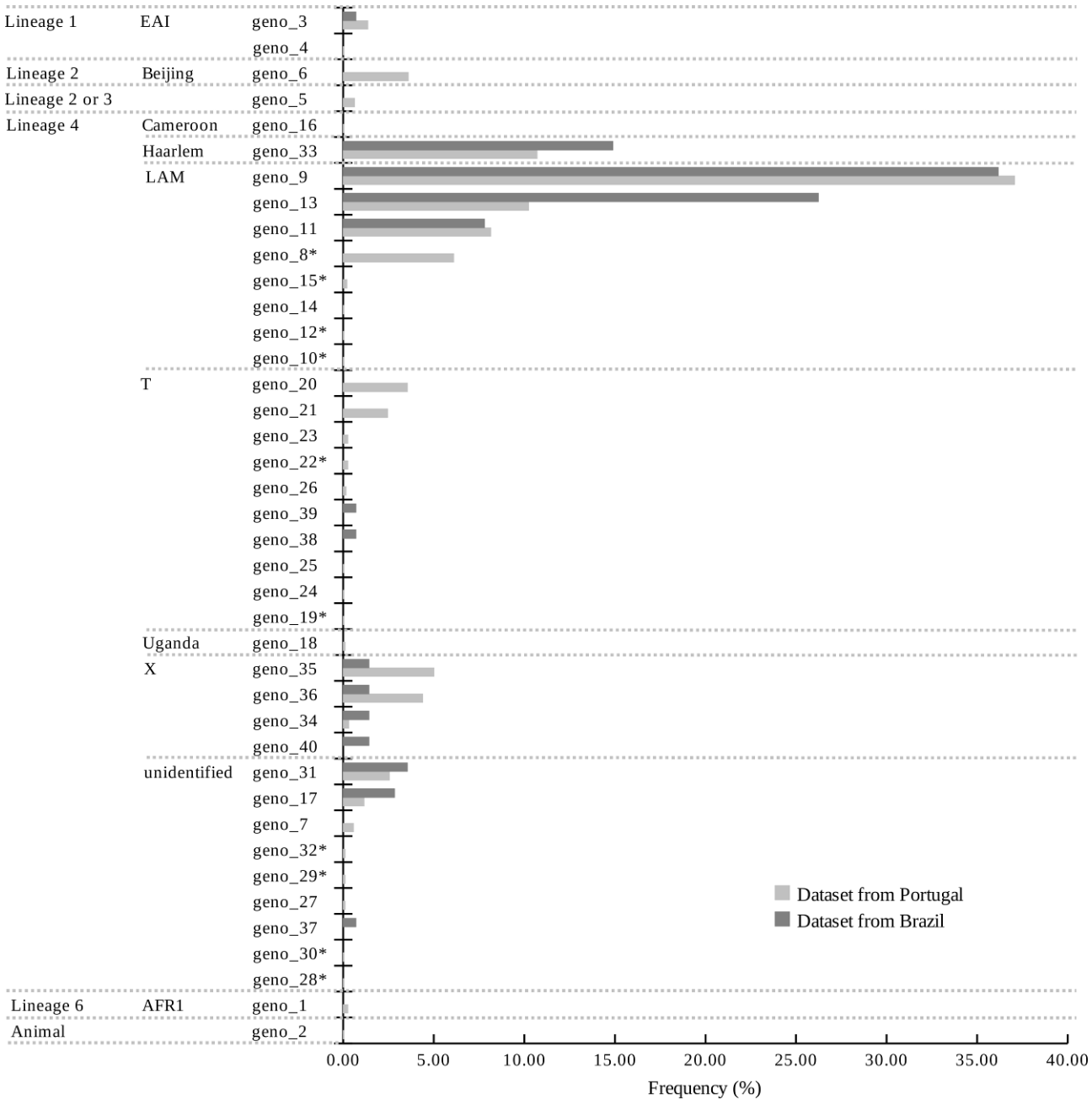
Lineage 1
EAI
geno_3
geno_4
Lineage 2
Beijing
geno_6
Lineage 2 or 3
geno_5
Lineage 4
Cameroon
geno_16
Haarlem
geno_33
LAM
geno_9
geno_13
geno_11
geno_8*
geno_15*
geno_14
geno_12*
geno_10*
T
geno_20
geno_21
geno_23
geno_22*
geno_26
geno_39
geno_38
geno_25
geno_24
geno_19*
Uganda
geno_18
X
geno_35
geno_36
geno_34
geno_40
unidentified
geno_31
geno_17
geno_7
geno_32*
geno_29*
geno_27
geno_37
geno_30*
geno_28*
Lineage 6
AFR1
geno_1
Animal
geno_2
Dataset from Portugal
Dataset from Brazil
0.00
5.00
10.00
15.00
20.00
25.00
30.00
35.00
40.00
Frequency (%)

Fig. 5

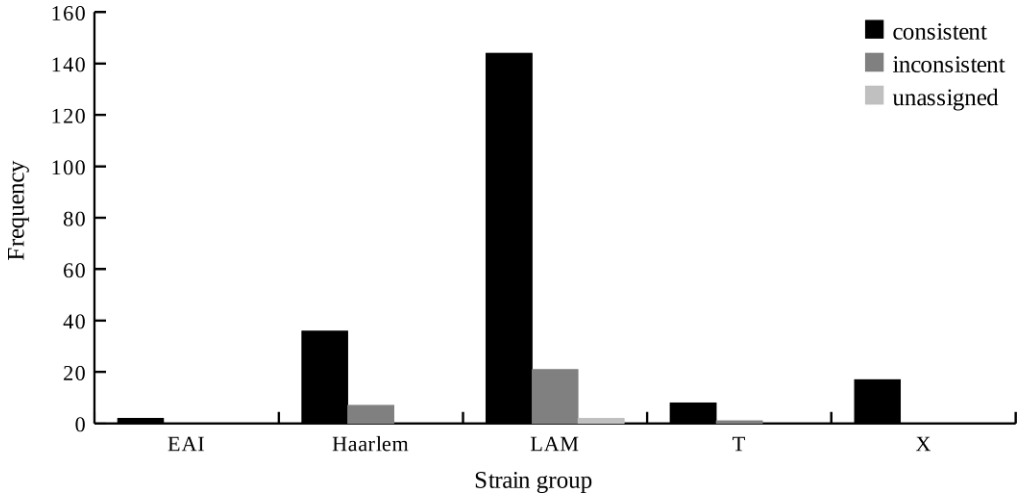
160
140
120
100
80
60
40
20
0
Frequency
consistent
inconsistent
unassigned
EAI
Haarlem
LAM
T
X
Strain group

Fig. 6