OPEN ACCESS Freely available online

PLoS one

Genotyping of Genetically Monomorphic Bacteria: DNA Sequencing in *Mycobacterium tuberculosis* Highlights the Limitations of Current Methodologies

Iñaki Comas[1], Susanne Homolka[2], Stefan Niemann[2☯], Sebastien Gagneux[1☯*]

1 Division of Mycobacterial Research, Medical Research Council, National Institute for Medical Research, London, United Kingdom, **2** Molecular Mycobacteriology, Research Center Borstel, Borstel, Germany

Abstract

Because genetically monomorphic bacterial pathogens harbour little DNA sequence diversity, most current genotyping techniques used to study the epidemiology of these organisms are based on mobile or repetitive genetic elements. Molecular markers commonly used in these bacteria include Clustered Regulatory Short Palindromic Repeats (CRISPR) and Variable Number Tandem Repeats (VNTR). These methods are also increasingly being applied to phylogenetic and population genetic studies. Using the *Mycobacterium tuberculosis* complex (MTBC) as a model, we evaluated the phylogenetic accuracy of CRISPR- and VNTR-based genotyping, which in MTBC are known as spoligotyping and Mycobacterial Interspersed Repetitive Units (MIRU)-VNTR-typing, respectively. We used as a gold standard the complete DNA sequences of 89 coding genes from a global strain collection. Our results showed that phylogenetic trees derived from these multilocus sequence data were highly congruent and statistically robust, irrespective of the phylogenetic methods used. By contrast, corresponding phylogenies inferred from spoligotyping or 15-loci-MIRU-VNTR were incongruent with respect to the sequence-based trees. Although 24-loci-MIRU-VNTR performed better, it was still unable to detect all strain lineages. The DNA sequence data showed virtually no homoplasy, but the opposite was true for spoligotyping and MIRU-VNTR, which was consistent with high rates of convergent evolution and the low statistical support obtained for phylogenetic groupings defined by these markers. Our results also revealed that the discriminatory power of the standard 24 MIRU-VNTR loci varied by strain lineage. Taken together, our findings suggest strain lineages in MTBC should be defined based on phylogenetically robust markers such as single nucleotide polymorphisms or large sequence polymorphisms, and that for epidemiological purposes, MIRU-VNTR loci should be used in a lineage-dependent manner. Our findings have implications for strain typing in other genetically monomorphic bacteria.

Citation: 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(11): e7815. doi:10.1371/journal.pone.0007815

Editor: Anastasia P. Litvintseva, Duke University Medical Center, United States of America

Received July 31, 2009; **Accepted** October 15, 2009; **Published** November 12, 2009

Funding: This work was supported by the Medical Research Council, UK, and USA National Institutes of Health grants HHSN266200700022C and AI034238. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: gagneux@nimr.mrc.ac.uk

☯ These authors contributed equally to this work.

Introduction

Some of the most important bacterial pathogens of humans exhibit strikingly low DNA sequence diversity. On average, these organisms harbour one nucleotide difference every 2–28 k base pairs and are thus referred to as genetically monomorphic [1]. Some prominent examples include *Yersinia pestis* (the etiologic agent of plague) [2], *Salmonella enterica* serovar Typhi (typhoid fever) [3], *Bacillus anthracis* (anthrax) [4], as well as the three most important pathogenic mycobacteria, *Mycobacterium leprae* (leprosy) [5], *Mycobacterium ulcerans* (buruli ulcer) [6], and *Mycobacterium tuberculosis* complex (MTBC) [7]. MTBC includes several sub-species that cause tuberculosis in humans and in various other mammals.

Understanding the diversity of bacterial pathogens is important, both for epidemiological and biological reasons. However, because of the low DNA sequence variation in monomorphic bacteria, studying the genetic diversity of these microbes is challenging. Standard sequence-based methods like multilocus sequence typing (MLST) are not applicable because of low phylogenetic resolution [8,9]. Alternative non-sequence-based tools, such as Pulsfield Gel Electrophoresis (PFGE) and Restriction Fragment Length Phylomorphism (RFLP) have been used for fine typing of monomorphic bacteria. However, these gel-based techniques have many drawbacks and are difficult to reproduce within and between laboratories [1]. More recently, PCR-based genotyping methods have been developed. Two of the most popular techniques are based on Clustered Regulatory Short Palindromic Repeats (CRISPR) and Variable Number Tandem Repeats (VNTR), respectively (Figure 1) [10,11]. CRISPRs are regions of the bacterial genome characterized by series of direct repeats interspersed by short unique regions called spacers. CRISPRs have been shown to encode a specialized defence mechanisms against bacteriophages, and changes in the number of spacers have been associated with phage-susceptibility [12]. VNTR-typing on the other hand, compares the strain-specific numbers of repeats of short DNA sequences at different positions of the bacterial genome [11].

CRISPR- and VNTR-based genotyping has been established for many genetically monomorphic bacterial pathogens, including

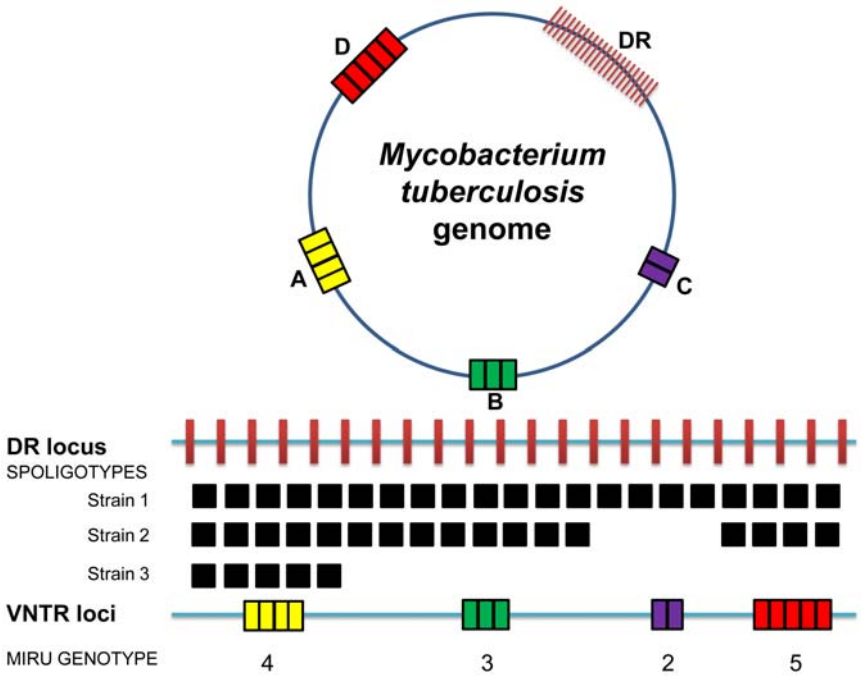

Figure 1. Schematic illustrating the principles of the CRISPR- and VNTR-based genotyping in MTBC. These genotyping methods are known as 'spoligotyping' and 'MIRU-VNTR-typing', respectively. Spoligotyping is based on the detection of 43 unique spacers located between direct repeats at a specific locus of the MTBC genome known as the direct repeat (DR) locus. Spoligotyping patterns are commonly represented by black and white squares indicating presence or absence of particular spacers, respectively. The deletion of some of these 43 spacers allows to differentiate between strains. MIRU-VNTR analysis relies on the identification of different number of repeats at several loci scattered around the bacterial genome (marked by A, B, C, and D in the figure). The number of repeats at each locus is combined to generate a unique numerical code used to establish phylogenetic and epidemiological links between strains.
doi:10.1371/journal.pone.0007815.g001

Y. pestis [13,14,15], *B. anthracis* [16], *Salmonella eneterica* serovar Typhi [17], *Francisella tularansis* [18], *Escherichia coli* O157 [19], and *M. leprae* [20]. In MTBC, the corresponding CRISPR- and VNTR-based methodologies are known as spoligotyping and MIRU-VNTR, respectively (Figure 1) [21,22]. These two genotyping techniques were originally developed for molecular epidemiological applications, and are routinely used to trace ongoing chains of tuberculosis transmission [23], to differentiate cases of disease relapse from re-infections [24], and to detect laboratory cross-contamination [25]. Over the years, databases have been populated with spoligotyping and MIRU-VNTR-typing results from thousands of patient isolates. For example, the spoligotyping database SpolDB4 contains data from close to 40,000 MTBC isolates from more than 120 countries [26], and MIRU-VNTR*plus* has been put up as a new online reference database for standard genotyping of MTBC [27].

In addition to routine molecular epidemiological applications, spoligotyping and MIRU-VNTR are increasingly also being applied to study evolutionary questions. Two complementary sets of MIRU-VNTR loci have been developed for MTBC [28]; 15-loci-MIRU-VNTR, which includes 15 loci, originally found to be the most discriminatory, and 24-loci-MIRU-VNTR that includes the same 15 loci plus an additional nine, which provide additional phylogenetic information. While 15-loci-MIRU-VNTR is mainly being applied for routine molecular epidemiology, spoligotyping and 24-loci-MIRU-VNTR have been proposed for phylogenetic and population genetic analyses of MTBC [26,29,30].

We recently performed a multilocus sequence analysis (MLSA) of 108 MTBC strains in which we generated the complete coding sequences of 89 genes, corresponding to ~70 k base pairs per strain [31]. We used these DNA sequences to generate a highly robust phylogeny of MTBC [31,32]. Here we used this MLSA-based phylogeny to evaluate the phylogenetic accuracy of spoligotyping and MIRU-VNTR in MTBC. In addition, we used this MLSA dataset to investigate the discriminatory power of the 24 standard MIRU-VNTR loci in the different strain lineages of MTBC.

Results

DNA Sequencing Defines a New Gold-Standard for the Phylogenetic Classification of MTBC

We previously showed that MLSA of 108 global strains of MTBC resulted in a single most parsimonious phylogenetic tree with negligible homoplasy (Figure 2) [31]. This phylogeny was also highly congruent with our previous analyses based on large sequence polymorphisms (LSPs) [33,34], and earlier DNA sequencing work [8,35]. To further probe the robustness of our MLSA-based phylogeny, we re-analyzed our DNA sequence data by the Neighbour-joining, Maximum likelihood, and Bayesian inference methods. All three analyses yielded identical tree topologies, which were highly congruent with our previous findings based on Maximum parsimony (Figure 2, Figure 3A, Figure S1). Furthermore, high statistical support was obtained for all main clades and for each method, despite the fact that some lineages were defined by only a small number of single nucleotide polymorphisms (SNPs). Because MTBC is strictly clonal [36,37], and our LSP and MLSA analyses were highly congruent, we conclude that our MLSA-based phylogeny is robust and appropriate for classification of MTBC strains into discrete strain lineages.

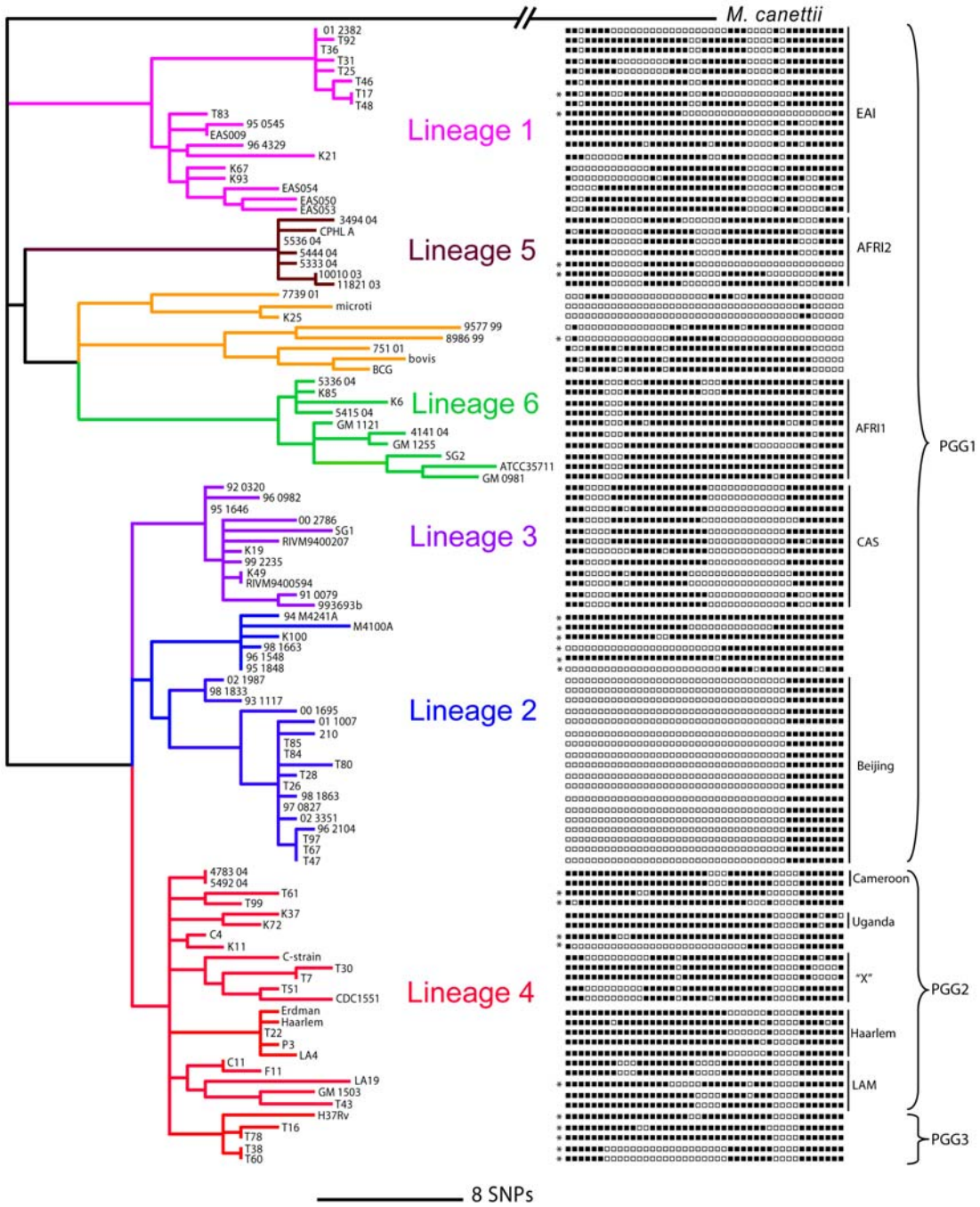

Figure 2. Maximum parsimony phylogeny based on concatenates of 89 gene sequences from 108 MTBC strains from global sources as previously reported [31]. Six main lineages can be observed within the human MTBC (numbered 1 to 6 and indicated in different colours). As shown previously, these lineages are highly congruent to the ones defined based on genomic deletions or large sequence polymorphisms (LSPs) [31,33,34]. Corresponding spoligotyping data for each strain are shown on the right, where black squares indicate the presence of a particular spacer and a white square the absence of a particular spacer (see Figure 1 for details on the methodology). Because the various typing techniques have classified MTBC strains into several lineages and strain families using differing nomenclatures, some of the traditional names are also shown. Some of the traditional groupings defined by spoligotyping correlate with SNP-based lineages (see also Table S1). For example, EAI (East-African-Indian) corresponds to the pink lineage, AFR1 and AFR2 correspond to the green and brown lineage, respectively (these strains are also known as *M. africanum*), and CAS (Central-Asian) corresponds to the purple lineage. However, other strain groupings defined by spoligotyping should be regarded as sub-lineages within the main lineages. For example, the 'Beijing' strain family is part of the blue lineage, and the five spoligotyping groups 'Cameroon', 'Uganda' 'X', 'Haarlem', and 'LAM (Latin-American-Mediterranean)' are sub-lineages within the main red lineage. This highlights another limitation of spoligotyping, which is that phylogenetic relationships between strain groupings cannot be defined. In addition, asterisks indicate spoligotyping patterns that cannot be classified at all using standard 'signature patterns' [26]. PGG1, PGG2, and PGG3 indicate Principal Genetic Group 1, 2, and 3, respectively. The PGG nomenclature is based on two SNPs originally described by Sreevatsan at al. [7]. Comparison to the MLSA data shows these groups are not phylogenetically equivalent as most of the MTBC diversity groups within PGG1, and PGG3 includes only a small subset of strains.

doi:10.1371/journal.pone.0007815.g002

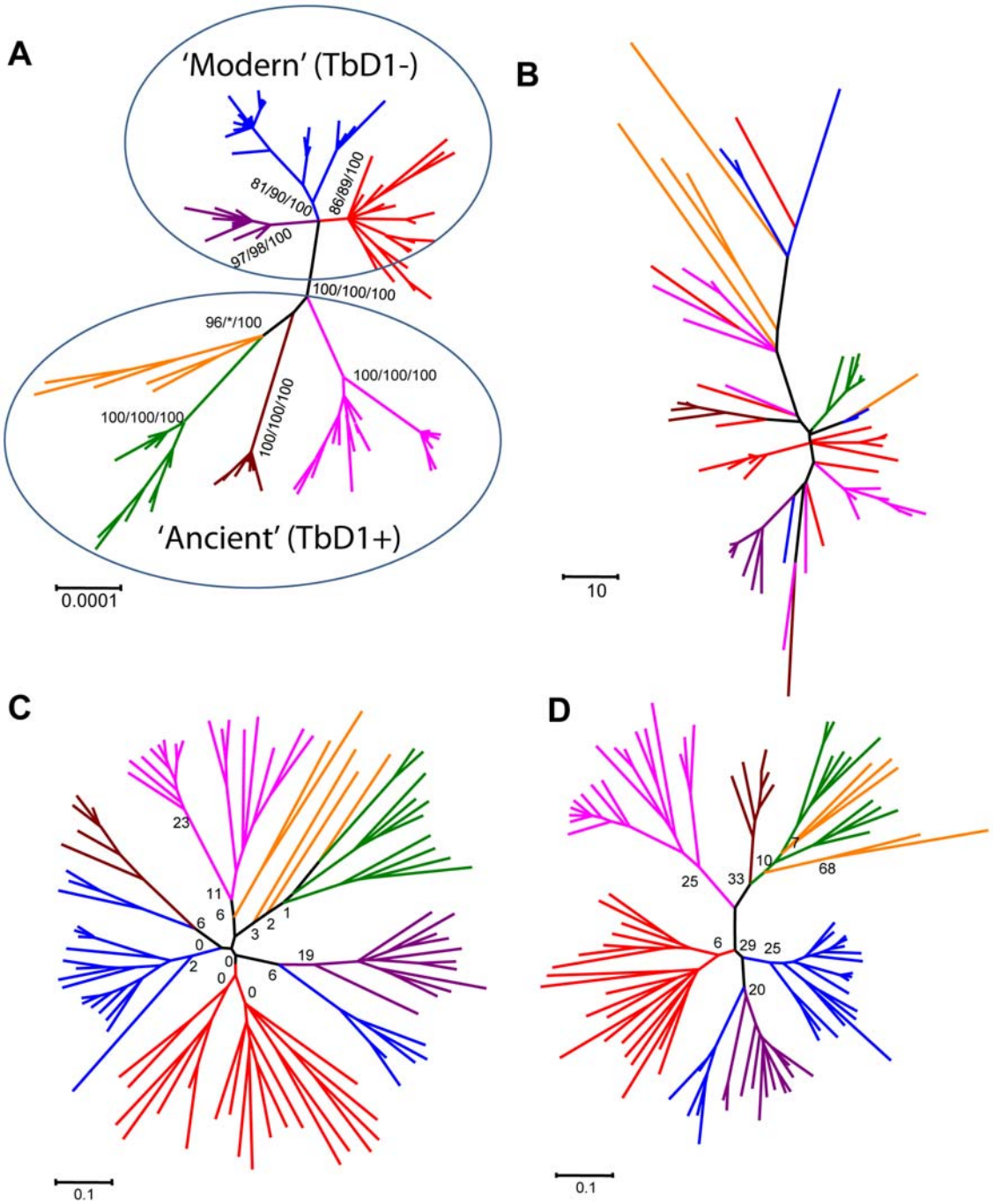

Figure 3. Comparison of unrooted phylogenies of MTBC based on 97 global strains using various molecular markers. Colours indicate the main MTBC lineages as defined by MLSA and LSPs [31]. (**A**) Neigbour-joining (NJ) phylogeny based on 339 variable nucleotide positions in 89 genes using number of SNPs as distance. The same topology was obtained using NJ, Maximum likelihood (ML), and Bayesian inference (BI). Numbers indicate bootstrap support after 1,000 pseudoreplicates for NJ and ML, and *a posteriori* probabilities for BI, respectively (Figure S1). MTBC can be divided in two main clades, one evolutionary 'modern' (also known as 'TbD1-negative'), which includes the blue, purple, and red strain lineages, and one evolutionary 'ancient' (TbD-positive), which includes the remaining strain lineages. (**B**) NJ phylogeny based on spoligotyping data and Jaccard distances. No bootstrap values could be calculated using these markers. (**C**) NJ phylogeny based on 15-loci-MIRU-VNTR data and Nei distances. Numbers indicate bootstrap support after 1,000 pseudoreplicates. (**D**) NJ phylogeny based on 24-loci-MIRU-VNTR data and Nei distances. Numbers indicate bootstrap support after 1,000 pseudoreplicates.
doi:10.1371/journal.pone.0007815.g003

To develop a new SNP-based classification system for MTBC, we used our MLSA-based phylogeny to extracted all lineage-defining SNPs (Table S1). Many of these SNPs are redundant and can be used interchangeably to identify the same phylogenetic groupings. Because the technical requirements of various SNP-typing technologies may vary [38,39], we believe being able to choose among more than one lineage-specific SNP will facilitate the design of SNP-based assays using either of these platforms.

Furthermore, the SNPs proposed here are more appropriate for typing of MTBC compared to most of the ones reported previously, because we used *de novo* DNA sequence data from 108 global strains to discover these SNPs [31]. By contrast, SNP collections published previously were identified by comparing only three or four MTBC genome sequences [40,41,42], and thus suffer from phylogenetic discovery bias [1,32,43,44,45].

CRISPR- and VNTR-Based Genotyping Results in Unreliable Phylogenetic Inference

CRISPR- and VNTR-based genotyping techniques have been used for phylogenetic and population genetic studies of genetically monomorphic bacteria [1]. Here we evaluated the performance of some of these methods using MTBC as an example. We used both qualitative and quantitative methods to determine the phylogenetic congruence of CRISPR-based spoligotyping and MIRU-VNTR-typing using our MLSA-based phylogeny as a gold standard.

We first generated the corresponding spoligotyping and 24-loci-MIRU-VNTR-typing data from strains included in our MLSA study following internationally standardized protocols [21,28]. Our final dataset comprised 97 strains with complete MLSA, spoligotyping, 15-loci-MIRU-VNTR, and 24-loci-MIRU-VNTR data (Table S2). We then used a qualitative approach to see whether the main MTBC lineages inferred by MLSA (indicated in different colours in Figure 2 and Figure 3A) were reproduced across the different genotyping datasets. To test this, we mapped the seven main MTBC lineages onto the tree topologies generated from the spoligotyping, 15-loci-MIRU-VNTR, or 24-loci-MIRU-VNTR data (Figures 3B, 3C, and 3D, respectively). Our results showed that spoligotyping was unable to retrieve five out of the seven main strain lineages as monophyletic groupings (Figure 3B). Even the clear separation between the evolutionary 'Ancient' (TbD1+) and 'Modern' (TbD1-) clades was missed [31,32]. It has been argued that even though spoligotyping data may not be ideal for formal phylogenetic analyses, particular "signature" patterns can still be informative for population genetic analyses [26]. For example, the "Beijing" lineage of MTBC has a characteristic loss of 34 spacers (Figure 2), which is caused by a deletion of a genomic region known as RD207 [46]. In other words, this spoligotyping pattern reflects a large sequence polymorphism that is phylogenetically informative (Table S1). Comparison of our MLSA dataset to the corresponding spoligotyping data shows that indeed, many strains can be grouped using such "signature" patterns. However, others cannot be classified properly because their spoligotyping patterns are either ambiguous or uninformative (Figure 2) [47].

The phylogenetic accuracy of MIRU-VNTR-typing was better than spoligotyping overall, but depended on the number of loci included in the analysis. As observed previously [28], while 15-loci-MIRU-VNTR was prone to phylogenetic misclassification (Figure 3C), 24-loci-MIRU-VNTR was more informative with most strain lineages appearing as monophyletic groups (Figure 3D). However, a closer look revealed several qualitative incongruencies with respect to the MLSA gold standard. At the main lineage level, the green, orange, and blue strains did not appear as monophyletic groupings in the 24-loci-MIRU-VNTR phylogeny (Figure 3D). Furthermore, additional incongruence became evident at the sub-lineage level. To show this, we performed an analysis restricted to strains from the red lineage. We mapped onto the 24-loci-MIRU-VNTR phylogeny all 27 phylogenetically informative SNPs found in red strains based on our MLSA dataset (Figure 2 and Figure 3). We found that 13 of these (48%) were incompatible with the 24-loci-MIRU-VNTR topology (Figure 4 and Figure S2).

A more quantitative way of evaluating the phylogenetic performance of different genotyping methodologies is by comparing the statistical support for each clade. As discussed above, our MLSA-based phylogenies exhibited high statistical support for all strain lineages, irrespective of the phylogenetic method used (Figure 3A, Figure S1). By contrast, for both MIRU-VNTR methodologies bootstrap values were low and thus multiple alternative topologies equally likely (Figure 3C and D). Phylogenetic congruence testing is another quantitative way of comparing phylogenetic topologies. It provides a statistical framework to evaluate how well the MLSA data fits the non-sequence-based phylogenies by calculating a likelihood value associated with each of the methods. Our analysis revealed that the phylogenies derived from spoligotyping and MIRU-VNTR were significantly incompatible with the MLSA data (Table 1). This result is particularly important given that among the various tests available, the Shimodaira-Haegawa test we used here tends to be the most conservative [48].

We suspected the reason for the low bootstrap support in the non-sequence-based phylogenies, and the statistically significant incongruence between these phylogenies and the MLSA data was

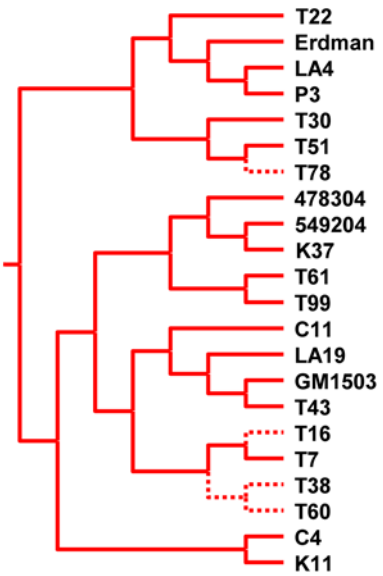

Figure 4. One example of homoplasy in the MIRU-VNTR-based phylogeny for the red strain lineage. The SNP C→G is shared by the strains T60, T38, T16, and T78 (dashed branches). These strains form a monophyletic group in the MLSA phylogeny (Figure 2). By contrast, the MIRU-VNTR-based topology splits these strains into three artificial groups, implying the same C→G change occurred three times independently.
doi:10.1371/journal.pone.0007815.g004

Table 1. Phylogenetic congruence test.

Topology	logL	difference	SH (p-value)
Spoligotyping	−95297.8	2826.06	<0.01
15-loci-MIRU1-VNTR	−93459.9	988.21	<0.01
24-loci-MIRU-VNTR	−93158.4	686.65	<0.01
MLSA (SNPs)	−92471.7	0	n.s.

For each topology the likelihood associated to the MLSA alignment and the difference between this value with the highest likelihood is shown (fourth column).
doi:10.1371/journal.pone.0007815.t001

because of homoplasy. Both spoligotyping and MIRU-VNTR are based on a limited number of loci, and the markers used evolve rapidly with a tendency to converge [40]. To test this hypothesis, we calculated the homoplasy index for each marker (Figure 5). As expected based on our qualitative analysis (Figure 3B), the highest homoplasy was found in spoligotyping patterns. Moreover, both the 15-loci-MIRU-VNTR and 24-loci-MIRU-VNTR data sets retained high levels of homoplasy, whereas in the MLSA data, homoplasy was virtually absent (Figure 5). To further explore instances of convergent evolution in MIRU-VNTR, we mapped all VNTR alleles for each MIRU-VNTR locus onto our MLSA phylogeny. We found that 23 out of 24 (96%) loci showed evidence of convergent evolution (Figure S3).

Taken together, our qualitative and quantitative analysis of CRISPR- and VNTR-based genotyping methods revealed that both types of markers are characterized by significant amounts of homoplasy. Hence using these tools to define deep phylogenetic groupings in MTBC or other bacteria, can be misleading [40,47]. By contrast, DNA sequencing allows to identify true phylogenetic relationships, and to discover SNPs that can be used as powerful genotyping markers (Table S1) [1].

Discriminatory Power of MIRU-VNTR Markers Vary by Strain Lineage

Even though SNPs are ideal for defining deep phylogenetic groupings, these markers offer insufficient discriminatory power for routine molecular epidemiological investigation in genetically monomorphic bacterial pathogens [1]. CRISPR-, VNTR-, and other related genotyping methods will thus likely remain important genotyping tools for epidemiological purposes. However, because the relative discriminatory power of particular VNTR- loci has been shown to vary depending on the specific strain background [49], we decided to use our MLSA dataset to study this phenomenon in more detail.

The discriminatory power of a given genotyping technique can be assessed using the Hunter Gaston Index (HGI) [50]. A high HGI indicates a given molecular marker or methodology is able to correctly classify closely related strains. To test whether the discriminatory power of the standard 24 MIRU-VNTR loci differed by MTBC lineage, we calculated the HGI for each locus separately for each of the main MTBC lineages (Figure 2, Figure 3A). We found that for most strain lineages, the majority of the MIRU-VNTR loci exhibited limited discriminatory power (Figure 6, Table S3). Moreover, the MIRU-VNTR loci that exhibited the highest HGI in one strain lineage were not necessarily the ones with the highest discriminatory power in other strain lineages. The fact MIRU-VNTR loci show the highest discriminatory power for the red lineage suggests that red strains were overrepresented during the original development of this genotyping technique [28]. Some strain lineages in our MLSA dataset were represented by fewer strains than other lineages, which could have influenced our analysis. To test this possibility, we analyzed the intra-lineage nucleotide diversity and compared it to the number of discriminatory loci in each lineage. We found no significant correlation between these two factors (Spearman's rho 0.62, p-value 0.14). Furthermore, in three out of four strain lineages harbouring equal or greater amounts of nucleotide diversity compared to the red lineage, the number of discriminatory loci was lower than in the red lineage (Figure 7).

In sum, our results demonstrate that VNTR loci can exhibit different discriminatory power in different bacterial strain lineages. These findings caution that selection of molecular markers for epidemiological typing should be based on large and globally representative strain collections. Moreover, our findings indicate that to maximize discriminatory power and minimize genotyping costs, only those VNTR markers should be used that offer the highest discriminatory power within a particular strain lineage (Table S3).

Discussion

Discrimination between strains of pathogenic bacteria is crucial. From an epidemiological perspective, molecular investigation contributes to the control of infectious diseases, both locally and globally. In addition, molecular typing improves our understanding of the basic biology of bacterial pathogens, including differences in virulence and transmissibility, or the variable effectiveness of vaccines and drugs. Unfortunately, the properties of molecular markers required to address both local and global levels of bacterial diversity are unlikely to be met by a single marker [51]. This problem is particularly acute in genetically monomorphic bacteria [1]. Because standard sequence-based genotyping such as MLST are not applicable in these bacteria, non-sequence-based tools, including CRISPR- and VNTR-based

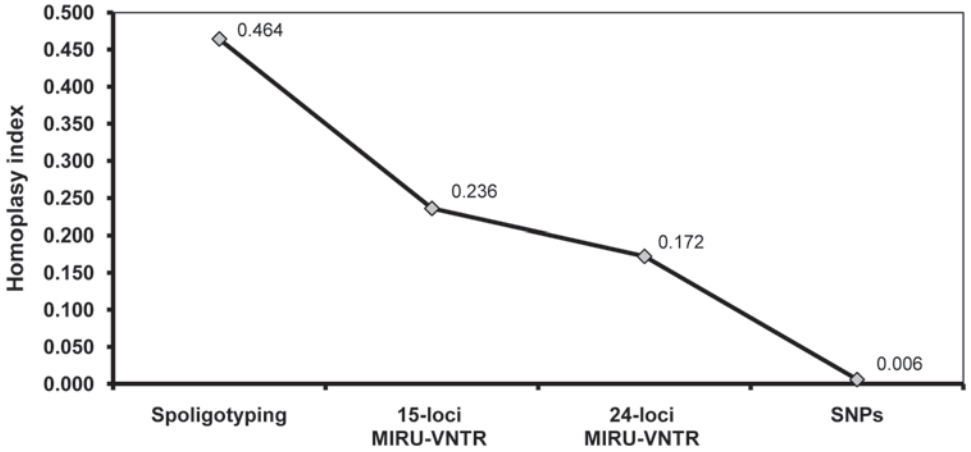

Figure 5. Comparison of the homoplasy index (HI) across the different genotyping methods. HI was calculated based on the number of observed changes at each character compared to the expected number of changes assuming absence of homoplasy. Figure S3 shows several examples of homoplasy for individual MIRU-VNTR loci where the same number of repeats appear in unrelated branches of the tree.
doi:10.1371/journal.pone.0007815.g005

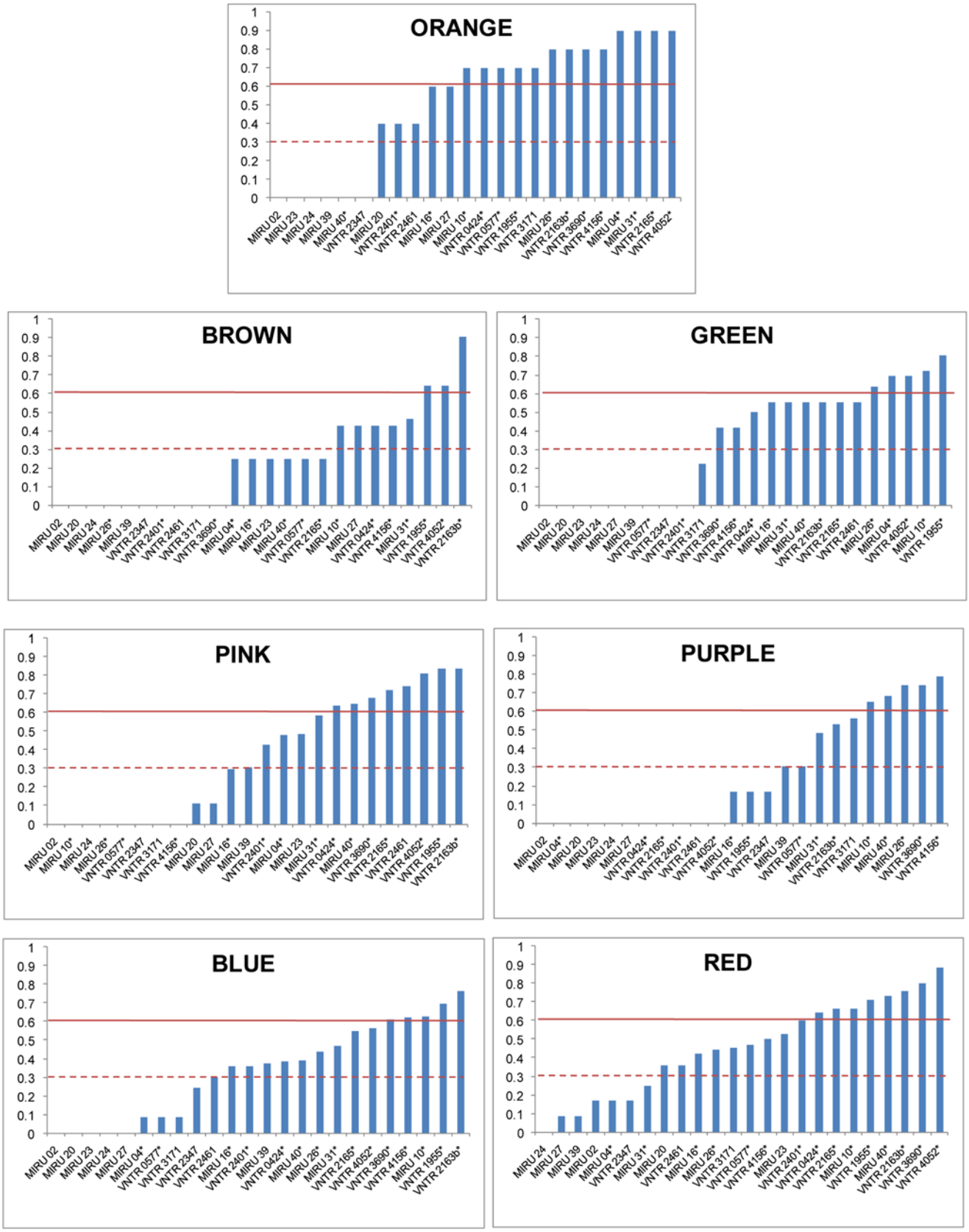

Figure 6. Measure of discriminatory power (HGI) of individual MIRU-VNTR loci by MLSA-defined MTBC strain lineage. Red lines indicate HGI thresholds for highly discriminatory loci (HGI≥0.6, continuous), and intermediate discriminatory loci (HGI≥0.3, dashed), as previously defined [28]. Asterisks indicate MIRU-VNTR loci that have been proposed for standard molecular epidemiological typing of MTBC [28]. See also Table S3.
doi:10.1371/journal.pone.0007815.g006

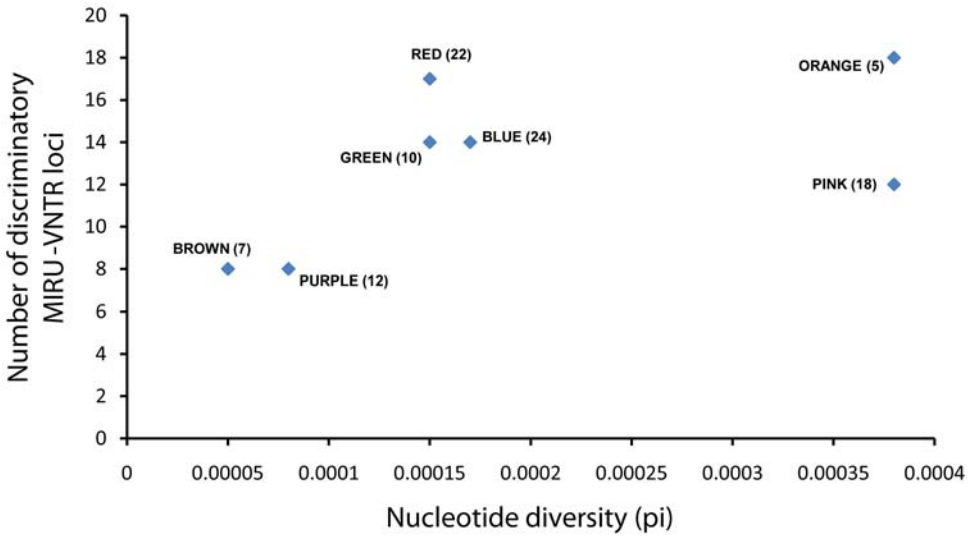

Figure 7. Number of discriminatory MIRU-VNTR loci (HGI≥0.3) as a function of intra-lineage nucleotide diversity (Pi). The number next to the lineage designation indicates the number of strains analyzed for each MTBC lineage.
doi:10.1371/journal.pone.0007815.g007

techniques, have become the gold standard for routine genotyping of these species. These tools have been applied very successfully to address a variety of epidemiological questions [25].

However, the results presented here argue against the use of these methods for evolutionary studies. The high propensity for convergent evolution and the resulting homoplasies are a significant drawback for defining deep phylogenetic relationships. Although the phylogenetic performance of VNTR-based typing was superior to that of the CRISPR-based method, phylogenies inferred using these markers show little statistical support. Furthermore, both of these typing methods are limited because little information exists with respect to the mode of molecular evolution of the respective molecular markers. This limitation is particularly important for spoligotyping and other CRISPR- based methods where it is virtually impossible to know whether the loss of multiple adjacent sequence spacers is due to a single or multiple evolutionary events. A single-step model of evolution has recently been proposed for VNTR loci in MTBC [30], but more studies are needed to confirm this model. For DNA sequence data on the other hand, multiple models of molecular evolution have been developed based on empirical data, and a robust statistical framework exists to evaluate the validity of these models for inferring phylogenetic relationships [52].

While deep phylogenetic information might be of little relevance for molecular epidemiology, unequivocal classification of bacterial strains is essential for many other applications. For example, elucidating the evolutionary history and global spread of bacterial pathogens requires robust strain assignment [1]. Furthermore, being able to define strains unambiguously is essential if phenotypic associations are to be unveiled. The mere fact that genetically monomorphic bacteria harbour little DNA sequence variation does not necessarily mean all strains of a given species behave the same [53]. In fact in MTBC, there is mounting evidence that strain diversity plays a role in the outcome of TB infection and disease [54,55,56,57]. To detect putative clinical or experimental phenotypes, assignment of individual bacterial strains to specific clades or strain lineages has to rely on phylogenetically well-defined groupings. If bacterial strains are misclassified because of inappropriate genotyping methods, the statistical power to detect true associations will be reduced.

DNA sequencing costs have been decreasing exponentially [58], and full genome sequencing of bacteria has the potential to replace standard bacterial genotyping in the near future [59]. This prospect is particularly relevant for genetically monomorphic pathogens [60]. By interrogating the whole genome, sufficient sequence diversity will be detected to differentiate between individual strains. Furthermore, because of the comprehensive nature of full genome data, they can be used for both fine typing in an epidemiological context and large-scale evolutionary analyses. Several recent reports in *S. typhi* [3], *Brucella* spp. [61], and *Francisella tularensis* [62], have highlighted the potential of comparative whole genome sequencing for elucidating the global population structure of genetically monomorphic bacterial pathogens. However, even though next-generation DNA sequencing is becoming more readily available, such large-scale projects are likely to remain limited to specialized sequencing centers for some time. Until high-throughput genome-sequencing of bacteria becomes more affordable, generating genotyping data for local epidemiology and broader applications in monomorphic microbes will remain challenging. One way to address this challenge is to combine robust lineage-specific markers with highly discriminatory molecular epidemiological typing. Our results demonstrate that CRISPR- and VNTR-based markers can be used for initial exploratory screening of strains. However, because of the inherent phylogenetic limitations of these tools, final strain assignment to specific strain lineages should be based on more robust markers such as SNPs or LSPs.

The data presented here for MTBC suggest an approach, in which the main strain lineages are first identified by SNP-typing. Many SNP-typing technologies have been developed over the years, some of which are more affordable than others [38,39]. Because lineage-specific SNPs are mutually exclusive in MTBC (Table S1), not all need to be typed in every strain, which can reduce costs. Once the main strain lineages are known, but further molecular epidemiological discrimination is necessary, lineage-specific sets of most discriminatory VNTR markers can be used to separate individual strains within each lineage (Table S3). Such an approach would generate accurate data for epidemiological and evolutionary applications, as well as for classification of strains during clinical or experimental association studies. Similar combined SNP/VNTR-typing schemes could be developed for other genetically monomorphic bacterial pathogens.

Materials and Methods

Bacterial Strains and Molecular Typing

The bacterial strains included in this study are representative the global diversity of MTBC as shown previously [31]. MLSA data including 89 genes or 70 kbp per strain was generated by direct DNA sequencing of PCR products as described [31]. Spoligotyping and 15-loci-MIRU-VNTR and 24-loci-MIRU-VNTR genotyping was performed according to internationally standardized protocols [21,28]. A total of 97 strains which had all genotyping data available was used for the further analyses.

Data Analysis

To determine the spoligotyping pattern of each strain, each of the 43 spacers was treated as a binary character indicating presence (1) or absence (0). Distances between isolates were calculated using the Jaccard index as implemented in Bionumerics 5.1. A neighbour-joining (NJ) tree was obtained using these distances. For the MIRU-VNTR analysis, Populations 1.3 was used to generate a distance matrix and a phylogenetic tree using the Nei distance [63]. From the distance matrix a NJ tree was obtained and support for each clade was evaluated by generating 1,000 bootstrap pseudoreplicates. For the sequence based phylogenetic inference a concatenate alignment of the 89 genes sequenced by Hershberg et al. [31] was obtained after removing those strains with no MIRU-VNTR or spoligotype information. From the resulting 65,829 base pair alignment we extracted the variable positions (n = 339, Supplementary Table S2) and used them for the phylogenetic analysis. The MLSA tree was obtained by the NJ method, maximum likelihood and Bayesian inference. The NJ analysis was implemented in MEGA 4 [64] using the observed number of changes and 1,000 pseudo-replicated. More complex models were implemented for the maximum-likelihood and Bayesian analyses. We used Modeltest 3.7 [65] to determine the best fit model of nucleotide evolution following the Akaike information criterion [66]. After Modeltest analysis, the TVM model was applied for both analyses. The maximum-likelihood estimation was implemented in PHYML 3.0 [67] without substitution rate heterogeneity correction or invariant estimation as recommended by Modeltest. Clade support was evaluated by analyzing 1,000 bootstrap pseudo-replicates. The Bayesian analysis was run with MrBayes 3.1.2 [68]. This program approximates the posterior probabilities of the phylogenetic tree using a Markov Chain Monte Carlo (MCMC) method. Four chains in two replicates were run during 2 million generations, convergence was evaluated using Tracer 1.4 and accepted when the effective sample sizes of all parameters combining both runs reached 100 as recommended. The final topology and Bayesian *a posteriori* support values for clades were obtained from the consensus tree after discarding the first 10% generations as burn-in.

Based on the high congruence of our LSP and MLSA analyses [31,33,34,35], we assumed that the MLSA topology and lineage classification reflects the true evolutionary history of MTBC. Therefore we tested the specific hypothesis of whether the topologies obtained from the three non-MLSA topologies (spoligotypes, 15-loci-MIRU-VNTR, 24-loci-MIRU-VNTR) were congruent with the MLSA data. The Shimodaira-Hasegawa maximum likelihood test [69] of competing phylogenetic hypothesis was used with 1,000 RELL pseudo-replicates as implemented in Tree-Puzzle [70] to test whether the difference of likelihoods between the best tree and the competing hypotheses were significantly different from zero (alpha at 0.005 after correcting for multiple trees comparisons). The homoplasy index was calculated by fitting the data from each marker to the corresponding topology using PAUP 4.0 b [71]. We used Mesquite 2.6 to map characters across phylogenies. The discriminatory power of each MIRU-VNTR locus was evaluated using the Hunter-Gaston discriminatory index (HGI) [50].

Supporting Information

Figure S1 Multilocus sequence analysis phylogeny of 97 MTBC strains.
Found at: doi:10.1371/journal.pone.0007815.s001 (0.03 MB DOC)

Figure S2 Incongruence of the 24-loci-MIRU-VNTR phylogeny of the strains belonging to the red lineage.
Found at: doi:10.1371/journal.pone.0007815.s002 (0.68 MB PDF)

Figure S3 Homoplasy in the MIRU-VNTR loci.
Found at: doi:10.1371/journal.pone.0007815.s003 (1.05 MB PDF)

Table S1 Phylogenetically informative SNPs for genoptyping of MTBC.
Found at: doi:10.1371/journal.pone.0007815.s004 (0.05 MB XLS)

Table S2 24-loci-MIRU-VNTR, spoligotyping, and SNP data from the 97 MTBC strains included in this study.
Found at: doi:10.1371/journal.pone.0007815.s005 (0.58 MB XLS)

Table S3 Discriminatory MIRU-VNTR loci by MTBC lineage.
Found at: doi:10.1371/journal.pone.0007815.s006 (0.04 MB XLS)

Acknowledgments

We thank Sonia Borrell, Philip Supply, and Mark Achtman for valuable comments on the manuscript.

Author Contributions

Conceived and designed the experiments: SN SG. Performed the experiments: SH SN SG. Analyzed the data: IC SN SG. Contributed reagents/materials/analysis tools: SN SG. Wrote the paper: IC SN SG.

References

1. Achtman M (2008) Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu Rev Microbiol 62: 53–70.
2. Achtman M, Zurth K, Morelli G, Torrea G, Guiyoule A, et al. (1999) *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. Proc Natl Acad Sci U S A 96: 14043–14048.
3. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, et al. (2008) High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. Nat Genet 40: 987–993.
4. Van Ert MN, Easterday WR, Huynh LY, Okinaka RT, Hugh-Jones ME, et al. (2007) Global genetic population structure of *Bacillus anthracis*. PLoS ONE 2: e461.
5. Monot M, Honore N, Garnier T, Araoz R, Coppee JY, et al. (2005) On the origin of leprosy. Science 308: 1040–1042.
6. Demangel C, Stinear TP, Cole ST (2009) Buruli ulcer: reductive evolution enhances pathogenicity of *Mycobacterium ulcerans*. Nat Rev Microbiol 7: 50–60.
7. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, et al. (1997) Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc Natl Acad Sci U S A 94: 9869–9874.
8. Baker L, Brown T, Maiden MC, Drobniewski F (2004) Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. Emerg Infect Dis 10: 1568–1577.

9. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, et al. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A 95: 3140–3145.
10. Grissa I, Vergnaud G, Pourcel C (2008) CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 36: W145–148.
11. Lindstedt BA (2005) Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria. Electrophoresis 26: 2567–2582.
12. Andersson AF, Banfield JF (2008) Virus population dynamics and acquired virus resistance in natural microbial communities. Science 320: 1047–1050.
13. Cui Y, Li Y, Gorge O, Platonov ME, Yan Y, et al. (2008) Insight into microevolution of *Yersinia pestis* by clustered regularly interspaced short palindromic repeats. PLoS ONE 3: e2652.
14. Klevytska AM, Price LB, Schupp JM, Worsham PL, Wong J, et al. (2001) Identification and characterization of variable-number tandem repeats in the *Yersinia pestis* genome. J Clin Microbiol 39: 3179–3185.
15. Achtman M, Morelli G, Zhu P, Wirth T, Diehl I, et al. (2004) Microevolution and history of the plague bacillus, *Yersinia pestis*. Proc Natl Acad Sci U S A 101: 17837–17842.
16. Keim P, Price LB, Klevytska AM, Smith KL, Schupp JM, et al. (2000) Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. J Bacteriol 182: 2928–2936.
17. Ramisse V, Houssu P, Hernandez E, Denoeud F, Hilaire V, et al. (2004) Variable number of tandem repeats in *Salmonella enterica* subsp. enterica for typing purposes. J Clin Microbiol 42: 5722–5730.
18. Johansson A, Farlow J, Larsson P, Dukerich M, Chambers E, et al. (2004) Worldwide genetic relationships among *Francisella tularensis* isolates determined by multiple-locus variable-number tandem repeat analysis. J Bacteriol 186: 5808–5818.
19. Lindstedt BA, Vardund T, Kapperud G (2004) Multiple-locus variable-number tandem-repeats analysis of *Escherichia coli* O157 using PCR multiplexing and multi-colored capillary electrophoresis. J Microbiol Methods 58: 213–222.
20. Truman R, Fontes AB, De Miranda AB, Suffys P, Gillis T (2004) Genotypic variation and stability of four variable-number tandem repeats and their suitability for discriminating strains of *Mycobacterium leprae*. J Clin Microbiol 42: 2558–2565.
21. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, et al. (1997) Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J Clin Microbiol 35: 907–914.
22. Mazars E, Lesjean S, Banuls AL, Gilbert M, Vincent V, et al. (2001) High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. Proc Natl Acad Sci U S A 98: 1901–1906.
23. van Deutekom H, Supply P, de Haas PE, Willery E, Hoijng SP, et al. (2005) Molecular typing of *Mycobacterium tuberculosis* by mycobacterial interspersed repetitive unit-variable-number tandem repeat analysis, a more accurate method for identifying epidemiological links between patients with tuberculosis. J Clin Microbiol 43: 4473–4479.
24. Cox HS, Sibilia K, Feuerriegel S, Kalon S, Polonsky J, et al. (2008) Emergence of extensive drug resistance during treatment for multidrug-resistant tuberculosis. N Engl J Med 359: 2398–2400.
25. Mathema B, Kurepina NE, Bifani PJ, Kreiswirth BN (2006) Molecular epidemiology of tuberculosis: current insights. Clin Microbiol Rev 19: 658–685.
26. Brudey K, Driscoll JR, Rigouts L, Prodinger WM, Gori A, et al. (2006) *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol 6: 23.
27. Allix-Beguec C, Harmsen D, Weniger T, Supply P, Niemann S (2008) Evaluation and strategy for use of MIRU-VNTRplus, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. J Clin Microbiol 46: 2692–9.
28. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rusch-Gerdes S, et al. (2006) Proposal for standardization of optimized Mycobacterial Interspersed Repetitive Unit-Variable Number Tandem Repeat typing of *Mycobacterium tuberculosis*. J Clin Microbiol 45: 691–7.
29. Sola C, Filliol I, Legrand E, Lesjean S, Locht C, et al. (2003) Genotyping of the *Mycobacterium tuberculosis* complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics. Infect Genet Evol 3: 125–133.
30. Wirth T, Hildebrand F, Allix-Beguec C, Wolbeling F, Kubica T, et al. (2008) Origin, spread and demography of the *Mycobacterium tuberculosis* complex. PLoS Pathog 4: e1000160.
31. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, et al. (2008) High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. PLoS Biol 6: e311.
32. Smith NH, Hewinson RG, Kremer K, Brosch R, Gordon SV (2009) Myths and misconceptions: the origin and evolution of *Mycobacterium tuberculosis*. Nat Rev Microbiol 7: 537–44.
33. Gagneux S, Deriemer K, Van T, Kato-Maeda M, de Jong BC, et al. (2006) Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A 103: 2869–2873.
34. Comas I, Gagneux S (2009) The past and future of tuberculosis research. PLoS Pathog 5: e1000600.
35. Gagneux S, Small PM (2007) Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. Lancet Infect Dis 7: 328–337.
36. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM (2004) Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. Proc Natl Acad Sci U S A 101: 4871–4876.
37. Supply P, Warren RM, Banuls AL, Lesjean S, Van Der Spuy GD, et al. (2003) Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area. Mol Microbiol 47: 529–538.
38. Black WCt, Vontas JG (2007) Affordable assays for genotyping single nucleotide polymorphisms in insects. Insect Mol Biol 16: 377–387.
39. Kim S, Misra A (2007) SNP genotyping: technologies and biomedical applications. Annu Rev Biomed Eng 9: 289–320.
40. Filliol I, Motiwala AS, Cavatore M, Qi W, Hernando Hazbon M, et al. (2006) Global Phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J Bacteriol 188: 759–772.
41. Gutacker MM, Mathema B, Soini H, Shashkina E, Kreiswirth BN, et al. (2006) Single-nucleotide polymorphism-based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. J Infect Dis 193: 121–128.
42. Gutacker MM, Smoot JC, Migliaccio CA, Ricklefs SM, Hua S, et al. (2002) Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. Genetics 162: 1533–1543.
43. Pearson T, Busch JD, Ravel J, Read TD, Rhoton SD, et al. (2004) Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. Proc Natl Acad Sci U S A 101: 13536–13541.
44. Alland D, Whittam TS, Murray MB, Cave MD, Hazbon MH, et al. (2003) Modeling bacterial evolution with comparative-genome-based marker systems: application to *Mycobacterium tuberculosis* evolution and pathogenesis. J Bacteriol 185: 3392–3399.
45. Pearson T, Okinaka RT, Foster JT, Keim P (2009) Phylogenetic understanding of clonal populations in an era of whole genome sequencing. Infect Genet Evol 9: 1010–1019.
46. Tsolaki AG, Gagneux S, Pym AS, Goguet de la Salmoniere YO, Kreiswirth BN, et al. (2005) Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of *Mycobacterium tuberculosis*. J Clin Microbiol 43: 3185–3191.
47. Flores L, Van T, Narayanan S, Deriemer K, Kato-Maeda M, et al. (2007) Large sequence polymorphisms classify *Mycobacterium tuberculosis* with ancestral spoligotyping patterns. J Clin Microbiol 45: 3393–5.
48. Goldman N, Anderson JP, Rodrigo AG (2000) Likelihood-based tests of topologies in phylogenetics. Syst Biol 49: 652–670.
49. Murase Y, Mitarai S, Sugawara I, Kato S, Maeda S (2008) Promising loci of variable numbers of tandem repeats for typing Beijing family *Mycobacterium tuberculosis*. J Med Microbiol 57: 873–880.
50. Hunter PR, Gaston MA (1988) Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J Clin Microbiol 26: 2465–2466.
51. Feil EJ (2004) Small change: keeping pace with microevolution. Nat Rev Microbiol 2: 483–495.
52. Felsenstein J (2004) Inferring phylogenies. Sunderland, Massachusetts: Sinauer Associates, Inc.
53. Nicol MP, Wilkinson RJ (2008) The clinical consequences of strain diversity in *Mycobacterium tuberculosis*. Trans R Soc Trop Med Hyg 102: 955–65.
54. Caws M, Thwaites G, Dunstan S, Hawn TR, Thi Ngoc Lan N, et al. (2008) The influence of host and bacterial genotype on the development of disseminated disease with *Mycobacterium tuberculosis*. PLoS Pathog 4: e1000034.
55. de Jong BC, Hill PC, Aiken A, Awine T, Antonio M, et al. (2008) Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in The Gambia. J Infect Dis 198: 1037–43.
56. de Jong BC, Hill PC, Brookes RH, Gagneux S, Jeffries DJ, et al. (2006) *Mycobacterium africanum* elicits an attenuated T cell response to Early Secreted Antigenic Target, 6 kDa, in patients with tuberculosis and their household contacts. J Infect Dis 193: 1279–1286.
57. Thwaites G, Caws M, Chau TT, D'Sa A, Lan NT, et al. (2008) The relationship between *Mycobacterium tuberculosis* genotype and the clinical phenotype of pulmonary and meningeal tuberculosis. J Clin Microbiol 46: 1363–8.
58. Service RF (2006) Gene sequencing. The race for the $1000 genome. Science 311: 1544–1546.
59. Medini D, Serruto D, Parkhill J, Relman DA, Donati C, et al. (2008) Microbiology in the post-genomic era. Nat Rev Microbiol 6: 419–430.
60. Niemann S, Koser CU, Gagneux S, Plinke C, Homolka S, et al. (2009) Genomic diversity among drug sensitive and multidrug resistant isolates of *Mycobacterium tuberculosis* with identical DNA fingerprints. PLoS One 4: e7407.
61. Foster JT, Beckstrom-Sternberg SM, Pearson T, Beckstrom-Sternberg JS, Chain PS, et al. (2009) Whole-genome-based phylogeny and divergence of the genus *Brucella*. J Bacteriol 191: 2864–2870.
62. Vogler AJ, Birdsell D, Price LB, Bowers JR, Beckstrom-Sternberg SM, et al. (2009) Phylogeography of *Francisella tularensis*: global expansion of a highly fit clone. J Bacteriol 191: 2474–2484.
63. Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. J Mol Evol 19: 153–170.

64. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. Mol Biol Evol.
65. Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14: 817–818.
66. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr AC-19: 716–723.
67. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.
68. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.
69. Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16: 1114–1116.
70. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502–504.
71. Swofford DL (2002) PAUP*. Phylogenetic Aanalysis Using Parsimony (*and Other Methods). Version 4. Suderland, Massachusetts: Sinauer Associates.