

SNP/RD Typing of *Mycobacterium tuberculosis* Beijing Strains Reveals Local and Worldwide Disseminated Clonal Complexes

Anita C. Schürch^{1,2}, Kristin Kremer¹, Amber C. A. Hendriks¹, Benthe Freyee¹, Christopher R. E. McEvoy^{3,4}, Reinout van Crevel⁵, Martin J. Boeree⁶, Paul van Helden³, Robin M. Warren³, Roland J. Siezen², Dick van Soelingen^{1,6*}

1 Tuberculosis Reference Laboratory, National Institute for Public Health and the Environment (RIVM), Centre for Infectious Disease Control, (CIb/LIS, pb 22), Bilthoven, The Netherlands, **2** Radboud University Nijmegen Medical Centre/NCMLS, Centre for Molecular and Biomolecular Informatics, Nijmegen, The Netherlands, **3** Department of Science and Technology, National Research Foundation Centre of Excellence in Biomedical Tuberculosis Research, Medical Research Council Centre for Molecular and Cellular Biology, Stellenbosch University, Tygerberg, Cape Town, South Africa, **4** Department of Microbiology and Immunology, University of Melbourne, Victoria, Australia, **5** Department of Medicine, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands, **6** Department of Pulmonary Diseases, Radboud University Nijmegen Medical Centre/University Lung Centre Dekkerswald, Nijmegen, The Netherlands

Abstract

The Beijing strain is one of the most successful genotypes of *Mycobacterium tuberculosis* worldwide and appears to be highly homogenous according to existing genotyping methods. To type Beijing strains reliably we developed a robust typing scheme using single nucleotide polymorphisms (SNPs) and regions of difference (RDs) derived from whole-genome sequencing data of eight Beijing strains. SNP/RD typing of 259 *M. tuberculosis* isolates originating from 45 countries worldwide discriminated 27 clonal complexes within the Beijing genotype family. A total of 16 Beijing clonal complexes contained more than one isolate of known origin, of which two clonal complexes were strongly associated with South African origin. The remaining 14 clonal complexes encompassed isolates from different countries. Even highly resolved clonal complexes comprised isolates from distinct geographical sites. Our results suggest that Beijing strains spread globally on multiple occasions and that the tuberculosis epidemic caused by the Beijing genotype is at least partially driven by modern migration patterns. The SNPs and RDs presented in this study will facilitate future molecular epidemiological and phylogenetic studies on Beijing strains.

Citation: Schürch AC, Kremer K, Hendriks ACA, Freyee B, McEvoy CRE, et al. (2011) SNP/RD Typing of *Mycobacterium tuberculosis* Beijing Strains Reveals Local and Worldwide Disseminated Clonal Complexes. PLoS ONE 6(12): e28365. doi:10.1371/journal.pone.0028365

Editor: Sebastien Gagneux, Swiss Tropical and Public Health Institute, Switzerland

Received: July 5, 2011; **Accepted:** November 7, 2011; **Published:** December 5, 2011

Copyright: © 2011 Schürch et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This study was supported by the Strategical Research (SOR) fund of the RIVM (S/3230136/01/GA) and the EU-funded TBadapt project (LSHP-CT-2007-037919). 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: dick.van.soelingen@rivm.nl

Introduction

The tubercle bacillus is one of the most important human bacterial pathogens, with an estimated 9.4 million incident cases of tuberculosis globally in 2009 [1]. Using molecular genotyping methods numerous genotypes of *Mycobacterium tuberculosis* have been identified [2,3]. The Beijing genotype is one of the most studied genotypes, and causes approximately 50% of the tuberculosis cases in Asia [4]. Beijing strains are also a major driving force behind the multidrug-resistant tuberculosis epidemic in Eastern Europe and South Africa [5,6] and “Typical Beijing” strains may be able to circumvent the BCG vaccine-induced immunity [7,8].

The *M. tuberculosis* Beijing genotype is easily identified by a highly characteristic spoligotype pattern, resulting from the deletion of the RD207 [9,10,11]. Strains of the Beijing genotype were previously grouped into two lineages; “Typical” and “Atypical” [11] according to the presence or absence of an IS6110 insertion in the NTF region [12]. Strains of the “Typical Beijing” lineage can be defined by the presence of 51 SNPs [13]

and form a monophyletic group, whereas strains that were formerly indicated as “Atypical Beijing” were shown to be genetically diverse and paraphyletic and do not form a separate lineage [13]. We address both groups of Beijing strains as “Beijing genotype”. Strains of the dominant “Typical Beijing” lineage are isolated from about 80% of Beijing-infected cases [13] however, the ratio of “Typical” versus other Beijing strains differs significantly by region [7,14,15]. The lack of genetic diversity, especially among the “Typical Beijing” strains, points to a recent, clonal expansion of this lineage [13].

Whole genome sequencing and subsequent SNP typing is a method which could be used to analyze the population structure of clonal bacterial pathogens that lack genetic diversity [16]. Due to the recent progress in sequencing technologies [17,18] and high-throughput SNP typing approaches [19], an increasing number of SNP typing systems have been developed from whole-genome sequencing data of bacterial pathogens [20,21,22,23,24,25]. *M. tuberculosis* is a highly clonal microorganism, and no recent horizontal gene transfer or recombination events between different strains have been identified so far [26,27,28]. Given the recent

ancestry of the *M. tuberculosis* complex (MTBC) and low selective pressure on particular loci, SNPs and RDs (if independent from IS6110-directed recombination) represent unique events at unidirectional time points in the genealogy of a *M. tuberculosis* strain. Synonymous SNPs and RDs are representative of ancestral states of strains and can act as molecular markers for clonal complexes. Here, we applied SNPs and newly identified RDs to characterize the population structure of the Beijing genotype strains.

Results

Development of SNP/RD typing assay for Beijing strains

To validate identified genome-wide variations and to identify informative and robust markers that could support studies on the phylogeny of the Beijing genotype of *M. tuberculosis* through a redundancy analysis, we used SNPs identified from whole genome sequence comparison of six Beijing strains originating from China, Vietnam, and South Africa, as described in detail in a recent paper [13]. The 275 SNPs selected represented 14.6% of the total 1889 SNPs identified and 21.3% (of 1294 SNPs) of the Beijing-specific or polymorphic SNPs.

Typing of 178 strains with 275 SNPs in a high-throughput mass-spectrometry typing assay resulted in a total of 48,950 SNP positions being analyzed. A small subset of 289 SNP positions (0.59%) was undetermined and 51 positions (0.1%) had an ambiguous result (both indicated as “?” in Table S1). From the SNP matrix a phylogenetic tree was inferred using maximum likelihood (Figure S1-A). The same phylogeny inferred with maximum parsimony exhibited a consistency index of 0.93, which can be caused by either homoplasious or erroneous SNPs (i.e. due to technical failures). Non-clonally distributed SNPs in the tree included a SNP in *katG* which is known to be involved in antibiotic resistance [29], six other non-synonymous SNPs (in genes *pta*, *rpsL*, *lipU*, *lppF*, *eis* and *serA1*) and a synonymous SNP (in gene *oysA3*). Information on drug resistance frequencies between Typical and other Beijing strains was previously reported in [13].

Two duplicate isolates, that were used to confirm the consistency of the SNP typing (isolates NLA000200230 and NLA009701940) clustered at the same node. Moreover, some genome-sequenced strains were also typed with the SNP typing assay. As expected, the *in silico* determined SNP type of strain CHIN- (that corresponds to strain NLA000700872) was consistent with the high-throughput mass-spectrometry-typed isolate NLA000700872, as was CHIN+ consistent with NL000700873 and SA+ with SAWC5527 [13]. Strain V+ clustered slightly different from its counterpart NLA000800162. After exclusion of the eight non-clonally distributed SNPs from the SNP matrix, the strains V+ and NLA000800162 clustered at the same node of the maximum-likelihood tree (Figure S1-B) which lead us to believe that these SNPs were at least partially the results of errors of the SNP detection assay. However, genes involved in drug resistance are known targets of strong selective pressure which could lead to independent occurrence of a polymorphism on several branches of a tree [30].

To develop an assay that types Beijing strains in a reliable and efficient way we reduced the number of the 275 SNPs initially selected to type Beijing strains to 51 in a redundancy analysis as described in Material and Methods. During the course of our study, Niemann and colleagues [31] reported the genome sequence analysis of two Beijing strains from Uzbekistan. In order to reduce branch collapse [32,33] we selected ten SNPs identified in the Beijing strains from Uzbekistan (K1 and K2) [31]. The genomes of these two strains differed by 130 SNPs and one

deletion [31] that we named RD131, according to the naming scheme of Tsolaki and colleagues. Table S2 lists the 61 SNPs in detail and Figure 1 shows the SNP matrix.

Application of SNP/RD typing assay for Beijing strains

These 61 SNPs were assayed on 259 MTBC strains with the high-throughput mass-spectrometry method (Table S3 and Figure S1-C). The clonal complexes were defined by collecting strains with the same patterns of presence and absence of the markers as indicated in Figure 1. Non-Beijing SNP types were called H37Rv, O1 (outgroup 1) and O2 (outgroup 2), where O2 contains the isolate with the complete spoligotype (all 43 direct variable repeats) and O1 consisted of the sets of MTBC- and non-Beijing strains described in Material and Methods. The inclusion of information on the absence/presence of RD131 (described below in more detail) divided one clonal complex into two distinct complexes (Beijing SNP/RD types (BST) BST18 and BST19). RD131 was, therefore, included in the SNP/RD matrix (Figure 1, number 36), allowing 27 Beijing clonal complexes to be distinguished. These clonal complexes are represented by types BST1 to BST19, and the types with the name of the respective genome-sequenced isolate (Figure 1). The presence of marker 29 and 30 indicate Typical Beijing strains. The SNP/RD assay showed an overall discriminatory power (D) of 0.9.

Comparison of SNP/RD typing and Beijing lineage designation

For 58 isolates from South Africa, SNP/RD typing results were compared to lineage designation (see Material and Methods). The South African lineage designation [15] had a better discriminatory power for isolates that clustered at SNP/RD types BST7 and BST10, which were differentiated into three and two lineages. The SNP/RD typing system achieved a higher discriminatory power for strains in lineage 6 and 7, which were split up in three SNP/RD types each (Table 1). Overall, the SNP/RD assay had a better discriminatory power compared to the lineage designation on this specific set of isolates ($D=0.78$ versus $D=0.54$).

Comparison of SNP/RD typing and RFLP typing of K1 and K2

In general the isolates represented the diversity of RFLP patterns within the Beijing genotype. Beijing strains K1 and K2 however exhibited an identical IS6110 RFLP pattern [31] and were isolated in the same geographical region. With the application of RD131 it was possible to identify 15 strains that share a more recent ancestor with K1 than K2, despite the identical RFLP pattern of K1 and K2 (for the IS6110 RFLP pattern see [31]).

Identification of RDs within Beijing sublineages

The distribution of RDs known to be polymorphic among Beijing genotype strains (RD105, RD142, RD149, RD150, RD152, RD207 and RD181) in the genome sequenced strains was previously described [13]. In this study we investigated the occurrence of newly identified RDs among Beijing genotype strains (Table 2). *In silico* analysis identified DNA fragments that were absent in one or more of the genome-sequenced strains when compared to the reference strain H37Rv which had not been previously described [9,10,34,35,36,37]. These RDs were assayed on a selection of the 259 Beijing strains for absence or presence of RDs with PCR and gel electrophoresis. The strains were selected based on their DNA availability. The strain selection and the results for each of the assayed RDs are shown in the Figures S2,

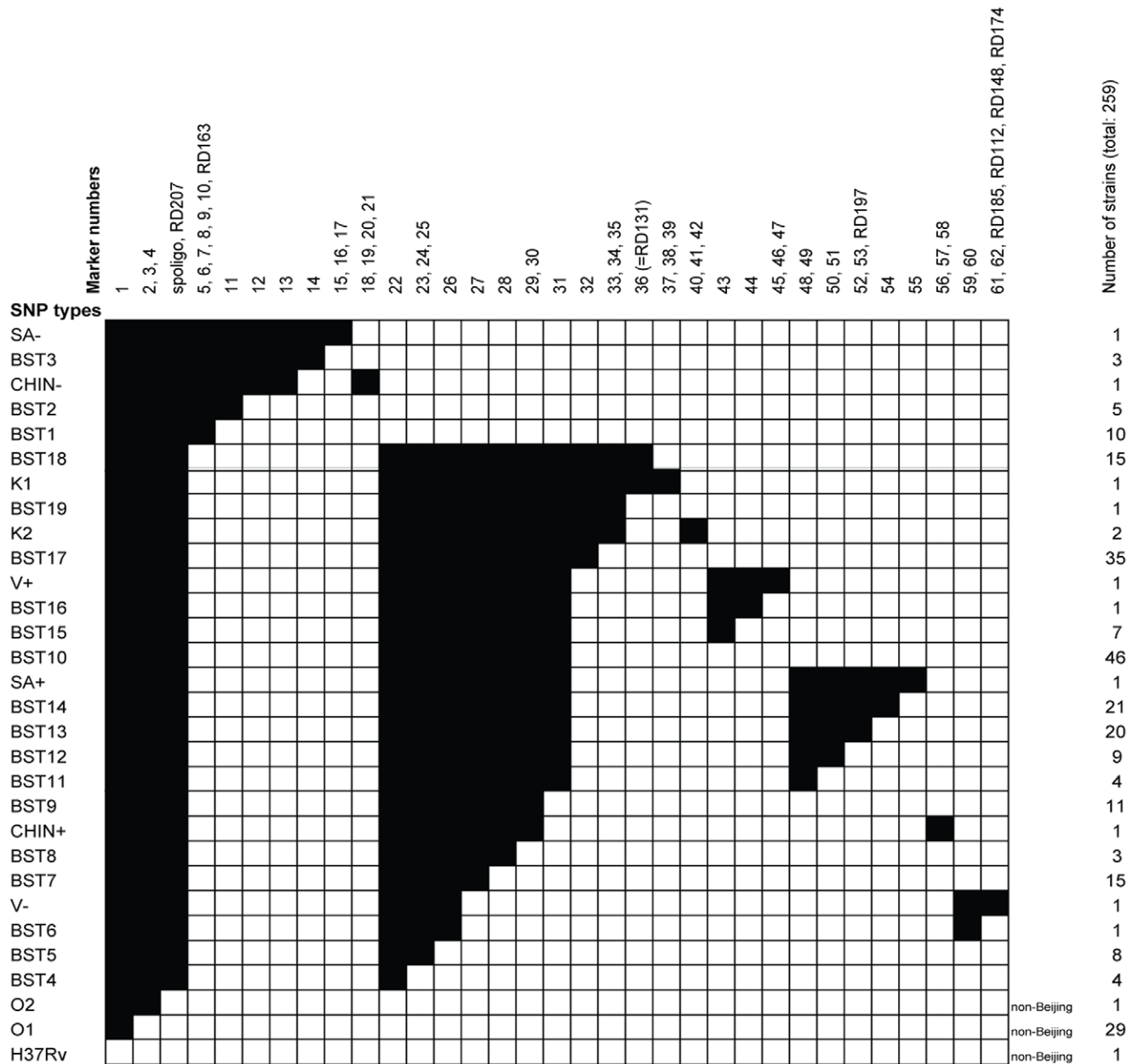


Figure 1. SNP/RD typing results matrix. Matrix of single nucleotide polymorphisms (SNPs)/region of difference (RD) typing scheme results for different clonal complexes of *M. tuberculosis* complex (outgroups H37Rv, O1, O2 and Beijing-SNP types BST1 to BST19 and SA-, CHIN-, V-, SA+, V+, CHIN+, K1 and K2). For the positions and polymorphic sites of the SNPs see Table S2, for details on RDs see Table 2. Black: SNP is present or RD is absent. SNPs and RDs that confer the same information are summarized in one column. doi:10.1371/journal.pone.0028365.g001

S3, S4, S5, S6 and S7. In summary, the patterns of presence and absence of seven RDs was concordant with the grouping on basis of SNP patterns for all strains assayed. The presence/absence pattern of one RD, RD131 (see above) differentiated one of the clonal complexes defined by SNP typing (BST18 and BST19). The chromosomal position and occurrences of all RDs are summarized in Table 2.

Phylogeography of the Beijing genotype

To assign a strain to a particular geographic area the country of origin was indicated for strains isolated in countries other than the Netherlands. For isolates from the Dutch database the patient's

country of origin which often coincides with the country of birth of the patients was used.

The typed *M. tuberculosis* strains originated from across the globe: 68 strains were isolated in Asia, 43 strains in Europe (including three strains from countries that are also on the Asian continent, Azerbaijan and Russia), 74 from Africa, (of which 65 from South Africa), 20 from North- or South America or the Caribbean and 54 with an unknown origin (Figure 2A and Table S4).

All but two clonal complexes that comprised more than one isolate were associated with diverse geographic origins (Figure 2). On the other hand, all isolates from the Archangel Oblast region

Table 1. Comparison of clustering of 58 *M. tuberculosis* isolates.

clonal complex	SA–	BST7	BST10	BST17	BST9	BST14	BST13	SA+
Lineage								
1	1							
2		2						
3		1						
4		1						
5			4					
6			3	7	1			
7						20	17	1

Compared were isolates of the Beijing genotype from South Africa, clustered by the lineage designation of Hanekom et al. (Hanekom et al., 2007) to the clonal complex designation resulting from SNP/RD typing in this study. The numbers in the table represent the number of strains in the respective clonal complex/lineage.
doi:10.1371/journal.pone.0028365.t001

that were part of this study (Figure 2B) were from clonal complexes that have SNP 32 present, showing that a wide distribution of the clonal complexes is not necessarily coupled to a wide variety locally. However, even a clonal complex that is highly resolved with 13 SNPs present, the K2 clonal complex, comprised two isolates from different origins. This clonal complex consists of isolate K2 (from Uzbekistan) and an isolate from the Archangel Oblast region. The distance between these two locations is more than 2000 km.

In contrast, two clonal complexes (BST13 and BST14) contain only samples from South Africa. These isolates belong to a group of Beijing strains that were associated with an increased transmissibility and ability to cause disease in South Africa (lineage 7, [15]). The more basal clonal complex BST12, which constitutes a clonal complex ancestral to the BST13/BST14 lineage, comprised three isolates from Indonesia and four from Hong Kong, along with two Beijing strains from an unknown origin. These isolates share a common ancestor with the South African strains from BST13 and BST14. Whether this ancestor originated in one of the three countries or a fourth unknown country remains unresolved.

Materials and Methods

A scheme of the workflow can be found as Figure S8.

Selection of SNPs

Whole genome sequences from eight Beijing genotype *M. tuberculosis* isolates representing “Typical” (+) and other (–) Beijing strains [11,31] were analyzed to identify SNPs and regions of difference (RDs). These Beijing strains originated from four different countries; China (CHIN+ and CHIN–), South Africa (SA+ and SA–), Vietnam (V+ and V–) and Uzbekistan (K1 and K2) [13,31]. H37Rv was used as the reference genome for SNP calling.

From the 1889 SNPs that were identified in a previous study [13], 275 SNPs were selected for a SNP typing assay. The selected SNPs were annotated to be in coding regions, with a preference for characterized genes as opposed to genes that are annotated as “hypothetical gene”. The selection contained only SNPs that did not have any other SNP closer than 500 bp, to avoid hypervariable regions and mutational hotspots. From this selection, 34 SNPs were located in 3R genes (genes involved in DNA repair, recombination and replication [38,39,40,41]). The SNP selection by Mestre et al. [41] based on 3R genes is the current best selection for SNP typing of Beijing strains but has only

low discriminatory power within Typical Beijing strains. Of all 275 SNPs, seven SNPs were H37Rv-specific (only present in H37Rv), 10 SNPs were Beijing-specific (with *Mycobacterium bovis* as out-group), and 258 SNPs were polymorphic among the six Beijing genomes (CHIN+, CHIN–, SA+, SA–, V+, V–). A total number of 168 SNPs were annotated as non-synonymous and 107 SNPs were synonymous.

Strain selection

One-hundred and seventy-two Beijing genotype strains were selected to cover the diversity of IS6110 RFLP patterns within the Beijing genotype [11] and comprised strains previously characterized with other markers [42,43,44]. The selected strains were representative of 45 countries on five continents: they were isolated in that country or isolated from a patient born in the respective country. H37Rv and five strains from other genotypes were included. Each strain was genotyped using the selected 275 SNPs.

One-hundred and fifty-nine of these strains were subsequently tested in a second assay of 61 informative SNPs (Table S2) and these strains were complemented with five Beijing strains from the same database that met the same inclusion criteria (representing different countries and the diversity of RFLP patterns) as described above. In addition, the 61 SNPs were assayed on 11 additional *M. tuberculosis* Beijing strains from Indonesia (in addition to six strains from Indonesia that were already present in the collection, total Indonesian strains n = 17) and 55 additional Beijing strains from South Africa (in addition to ten South African strains that were already present in the Dutch database, total South African strains n = 65). These 55 strains represented the 7 Beijing lineages described by Hanekom et al. [15] and the abundance of the different Beijing strains in South Africa. To determine the specificity of the markers for the Beijing genotype, an additional two sets of strains consisting of 13 strains of other species within the MTBC; including *Mycobacterium africanum* (n = 2), *Mycobacterium bovis* (n = 6), *M. bovis* BCG (n = 2), *Mycobacterium canettii* (n = 1) and *Mycobacterium microti* (n = 2) and 14 *M. tuberculosis* genotypes other than Beijing were included. One of these strains exhibited a complete spoligotype; such a strain is regarded to be closely related to Beijing strains as shown by large sequence polymorphisms [45]. Furthermore SNP data of two genome-sequenced strains K1 and K2 were included [31]. In total, 259 MTBC strains were assayed using 61 SNPs.

High-throughput SNP typing and clustering

Bacterial isolates were typed on the Sequenom genotyping platform (Sequenom GmbH, Hamburg, Germany) with iPLEX

Table 2. Regions of difference (RD) identified in *M. tuberculosis* Beijing strains.

Position in H37Rv	Reference	Genes (partially) deleted	RD name	Co-occurrence with marker (Figure 1)	Corresponding Figure
859243 – 859496	[31]	Rv0766c	RD131	36	S2
3120521– 3127920	[9].	Rv2814c – Rv2820c	RD207	spoligo	S3
2949906–2955132	this study	Rv2623 – Rv2627c	RD197	52, 53	S4
2626969–2633061	this study	Rv2394c – Rv2350c	RD185	61, 62	S5
358030–363748	this study	Rv0294–Rv0299	RD112	61, 62	S6
1715870–1733378	this study	Rv1522c–Rv1531	RD148	61,62	S6
2238647–2242137	this study	Rv1995–Rc1997	RD174	61, 62	S6
2128379–2129584	this study	Rv1878	RD163	5, 6, 7, 8, 9, 10	S7

doi:10.1371/journal.pone.0028365.t002

Gold biochemistry. To this end, primer extension reactions were carried out using oligonucleotide primers designed with the AssayDesigner 3.1 software. Single base extension of a primer that annealed directly adjacent to the SNP was measured on a compact MALDI-TOF mass spectrometer, following automated protocols in a 384-well format [46]. The SNP positions of the genome-sequenced strains (CHIN+, CHIN–, SA+, SA–, V+, V–, K1 and K2) were extracted from the literature [13,31].

The results of the high-throughput SNP typing for the 275 SNPs or the 61 SNPs were concatenated for each isolate and treated as alignments. A phylogenetic tree was inferred with maximum-likelihood using Phylml version 2.4.4 [47]. The trees were visualized as cladograms in Figures S1-A, S1-B and S1-C with Dendroscope [48]. Subsequently, for each internal and external branch of the phylogenetic tree that was established from 275 SNPs, SNPs were identified that represented the respective branch by their presence or absence in the isolates. To determine the consistency index, a tree was established with maximum-parsimony in MEGA 5 [49]. Non-clonally distributed SNPs in the tree were determined with mixed method discrete character parsimony carrying out Camin-Sokal parsimony as implemented in the Phylip suite (Phylip 3.69).

Redundancy analysis

SNP results of the assay with 275 SNPs were classified as reliable if no ambiguous bases were detected by the assay and could not be determined less than twice in the different strains (after isolates with more than ten empty SNP positions were excluded). Moreover we did not select the eight SNPs that were potentially the result of convergent evolution or SNP detection errors. Synonymous SNPs were chosen as these are generally assumed to be selectively neutral. We included one coding SNP that was not part of the first assay, but that was representative for a group of four SNPs that marked a short internal branch in the phylogeny of the six genome-sequenced Beijing strains [13]. This SNP causes a conservative amino acid change (valine-to-alanine substitution) in a probable acid-maltase protein.

Of the SNPs that were either specific for K1 or K2 [10] or present in both strains, but not reported in CHIN+, SA+, V+, SA–, CHIN– or V–, ten synonymous SNPs were randomly selected (three specific for strain K1, three specific for strain K2 and four SNPs specific for both K1 and K2).

The nucleotides at the 61 positions in the chromosome of each isolate were arranged in a SNP matrix and positions with the same and therefore redundant information content were merged (see Figure 1). RD131 was treated as one mutation event, and added to the concatenated 61 SNPs.

Identification and verification of RDs

A BLAST search (BLAST 2.2.19, [50]) of all genes of H37Rv with a sliding-window of 200 bp against the raw read collection of six 454/Roche-sequenced Beijing genomes (CHIN–, CHIN+, SA–, SA+, V+, V–, [13]) with read lengths of 250 bp on average was performed. Hits with query coverage of less than 0.8 were used to identify potential RDs.

To confirm the RDs identified *in silico* and to assay their presence or absence in selected Beijing isolates, PCR primers were designed adjacent to the putative deletion sites with Primer3plus (<http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi>) with standard parameters. For RDs larger than 2000 bp, two sets of primers were designed (primer sequences are indicated in Table S5). PCR reactions were performed with either PuReTaq Ready-To-Go™ PCR beads (GE Healthcare LifeScience, Little Chalfont, UK) or HotStarTaq Master Mix (Qiagen, Hilden, Germany) as mono- or multiplex reactions. The sizes of the PCR products were estimated on a 1% agarose gel to specify the presence or absence of the RD. Selected products were sequenced on an ABI 3730x sequencer (Applied Biosystems, Foster City, CA, USA) following standard protocols.

Discriminatory power

The discriminatory power (D) is the average probability that a typing system will assign a different type to two unrelated strains and was calculated according to the method described by Hunter and Gaston [51,52]. To compare the discriminatory power of the SNP/RD typing system relative to other typing systems, we compared another lineage designation (Lineage 1 to 7, established by Hanekom and colleagues [15]) of 58 isolates from South Africa to our classification in clonal complexes. These 58 isolates comprised three strains from the Dutch database from lineage 7. The lineage designation was based on IS6110 insertion sites as well as on synonymous SNPs, RDs, and SNPs in mismatch repair genes [15].

Discussion

We attempted to define a set of evolutionary stable, non-homoplasious genomic changes in the Beijing genotype identified in eight genome-sequenced strains that could serve as phylogenetic markers, and selected nucleotide changes at neutral or nearly-neutral sites. Furthermore, we included newly identified RDs in the SNP/RD typing scheme for Beijing strains, since deletions that occurred independently from IS6110-directed recombination have been successfully used to identify lineages in the MTBC

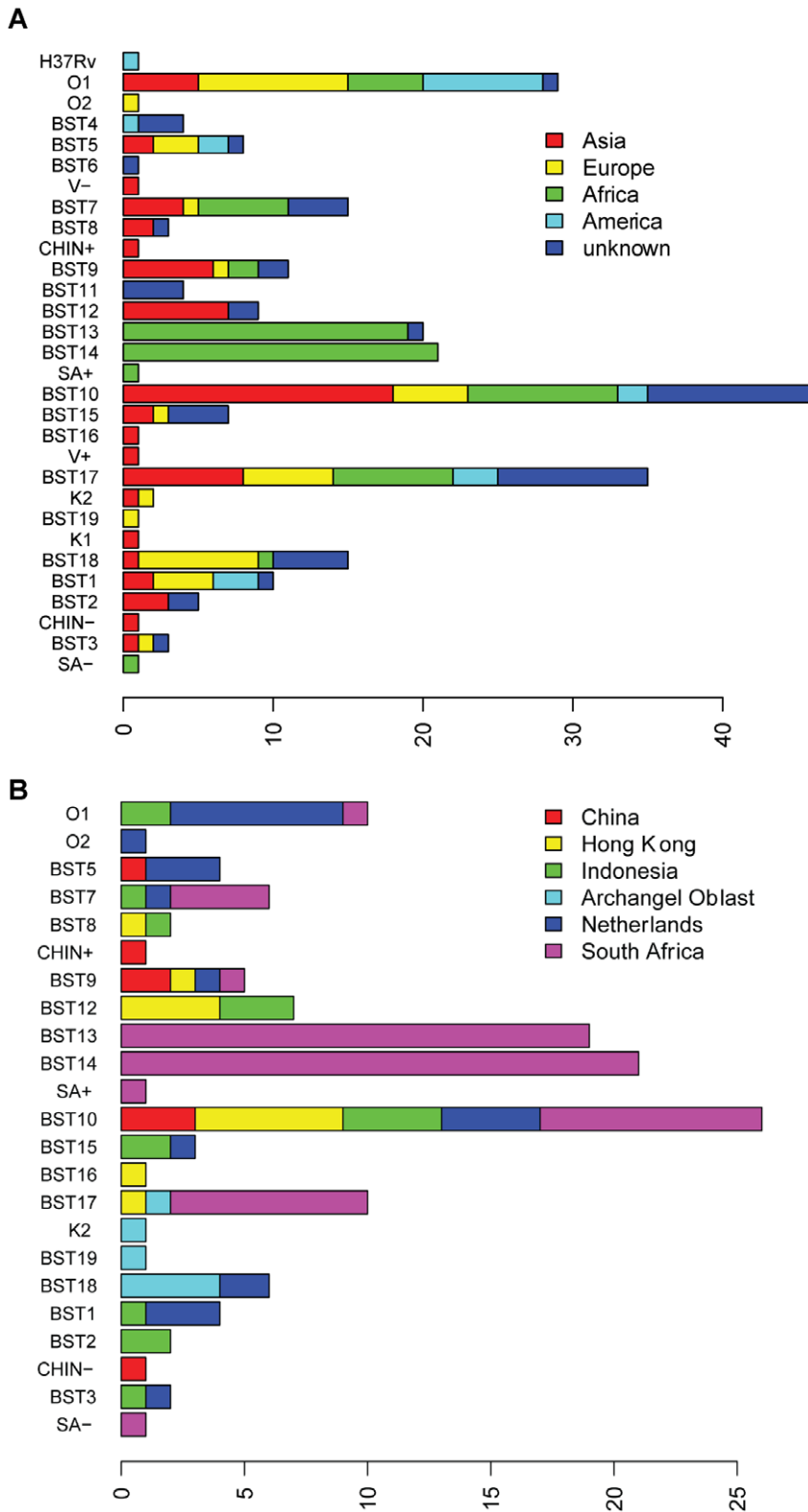


Figure 2. Origins of clonal complexes. A. Continent of origin of the *M. tuberculosis* isolates belonging to clonal complex H37Rv, O1, O2 and BST1 to BST19 and SA-, CHIN-, V-, SA+, V+, CHIN+, K1 and K2. "America" includes isolates from the South- and North American continent. B. Distribution of country of origin within clonal complexes of the *M. tuberculosis* isolates. Here, only isolates belonging to countries or regions with more than five isolates assayed in this study were considered. doi:10.1371/journal.pone.0028365.g002

[9,53,54,55,56,57,58]. With the application of the 61 SNPs and one RD to MTBC strains, 30 clonal complexes, of which 27 strictly comprised Beijing genotype strains, were distinguished.

A SNP/RD assay has many advantages over spoligotyping, IS6110 RFLP and VNTR typing. Synonymous SNPs and RDs are phylogenetically robust, and do not suffer from convergent evolution (at least if deletions were independent from repeat regions [59]) and moreover, they are relatively easy to determine by a PCR-based detection method [54]. In this study, the discriminatory power of RDs is low compared to that of the SNPs: only five clonal complexes could be distinguished by the RDs described ($D=0.52$). Moreover, RDs did not improve the resolution of the SNP-based assay, with one exception (RD131). However, the SNPs from the lineage leading to K1 and K2 have not been selected by a redundancy analysis but stochastically and thus may not provide the optimal discriminatory power that might have been achieved by a different set of SNPs. It is likely that some SNPs exist that can split up the identified clonal complexes, as does RD131. For these reasons we combined SNPs and RDs in one typing scheme.

A particularly striking result of this study is the almost complete absence of strong geographical associations of all but two clonal complexes. This in contrast to phylogeographical studies of diseases such as the plague [60], lepra [22], buruli ulcer [23], and in contrast to the strong phylogeographic association of major tuberculosis strain lineages reported in several studies [28,61,62,63]. In those studies, the Beijing/W lineage was strongly associated with patients of Asian origin. It has been suggested that this lineage has evolved so as to become adapted to the specific host background and that transmission of disease among patients with the same ethnicity is more likely [61]. In our study, patients from Asian origin form the second largest group (with 68 patient isolates) after patients from African origin (74 isolates). However, our study was not designed to include a representative sample from each region, but to include the genetic diversity among the Beijing RFLP patterns in our database combined with varying countries of origin of the isolates. Thus our study could not estimate the prevalence by region, or give information about the likelihood of transmissibility of clonal complexes. Our study does show that most Beijing clonal complexes can be isolated from patients with diverse genetic backgrounds.

The main difference between studies that find strong phylogeographic associations and our study is the scale of genetic diversity of the strains studied. Our results suggest that several Beijing ancestors spread successfully to different parts of the world and to hosts from diverse ethnicity on multiple occasions. One of the driving forces of the Beijing epidemic might be migration patterns which allow the wide distribution of closely related strains, such as air travel. In contrast, two clonal complexes (BST13 and BST14) were populated exclusively with samples from South Africa and show a strong association to a single country. It is unclear whether these strains are adapted to the South African population or if these clonal complexes did not have the chance to spread to other parts of the world yet, because of limited time or the socioeconomic circumstances of their hosts. Recently, van Helden et al. [64] suggested that Beijing strains might have been introduced to South Africa following the sea trade route from East Asia to Europe that started 400 years ago. Indeed, in the 17th and 18th centuries, Dutch colonists at the Cape of Good Hope largely imported slaves from Indonesia, Madagascar, Mozambique, and India. However, while the conjecture that Beijing strains were transported via the slave or trade route is plausible the study design and sampling do not preclude this conclusion.

The application of SNPs and RDs to non-genome-sequenced strains yields linear phylogenies with a complete absence of secondary branches (branch collapse), a phenomenon inherent to SNP-based phylogenies [32,33,65]. As expected in a branch-collapsed typing scheme, clonal complexes that are determined with a maximum number of SNPs present discriminate better than SNP types that are defined by fewer markers. These cannot distinguish between different lineages, as exemplified by comparing the lineage designation to the SNP/RD typing scheme. All samples that were typed in this study were more or less contemporary with respect to TB infections (they were isolated between 1993 and 2009). If fully sequenced, all these isolates would be separated by long branches [32,33]. In addition, the number of SNPs is not representative for the genetic distance between the clonal complexes, because we selected one to six SNPs to mark a clonal complex, regardless of evolutionary distance.

We postulate that, with the increasing number of SNPs identified from whole genome sequencing efforts, the resolution to determine the population structure will increase, and possibly more clonal complexes could show an association with country of origin. This improved discriminatory power will reach its optimum by genome sequencing of all *M. tuberculosis* isolates in the future. It is already clear however, that almost every Beijing clonal complex has disseminated to different parts of the world on multiple occasions. The SNPs and RDs presented in this study could – partially or jointly – be used in typing assays that aid in future molecular epidemiological studies of the Beijing genotype.

Supporting Information

Figure S1 Maximum-likelihood trees from SNP data. A: Maximum-likelihood tree of 275 concatenated SNPs in 178 *M. tuberculosis* complex strains. B: Maximum-likelihood tree of 267 concatenated SNPs in 178 *M. tuberculosis* complex strains. Eight SNPs that were non-clonally distributed in Figure 1A were excluded for this figure. C: Maximum-likelihood tree of 61 concatenated SNPs in 259 strains. Outgroups O1 and O2 are indicated.

(PDF)

Figure S2 Distribution of RD131 in the phylogenetic tree. Strains with background colors were assayed for the absence or presence of the RD. No background color: strain not assayed. Red: RD is present (deletion was identified). Yellow: RD is absent (no deletion has occurred). Green: product of other size than the expected product. For corresponding RDs see Table 2.

(PDF)

Figure S3 Distribution of RD207 in the phylogenetic tree.

(PDF)

Figure S4 Distribution of RD197 in the phylogenetic tree.

(PDF)

Figure S5 Distribution of RD185 in the phylogenetic tree.

(PDF)

Figure S6 Distribution of RD112, RD148 and RD174 in the phylogenetic tree.

(PDF)

Figure S7 Distribution of RD163 in the phylogenetic tree.

(PDF)

Figure S8 Scheme of workflow applied in this study.
(PDF)

Table S1 SNP matrix of 275 SNPs in 178 *M. tuberculosis* complex strains. 1: derived from 454 sequencing. 2: derived from high-throughput mass-spectrometry typing.?: no or ambiguous base determined. The Supporting Tables contain tab-delimited file that can be copied into spreadsheets.
(TXT)

Table S2 61 Single nucleotide polymorphisms and RD 131 (number 36) used in this study.
(TXT)

Table S3 SNP matrix of 61 SNPs in 259 *M. tuberculosis* complex strains.
(TXT)

Table S4 Country of origin of the isolates used in this study.
(TXT)

Table S5 Primer sequences and PCR conditions used to determine regions of difference (RDs). RD 207 was determined by spoligotyping and is therefore not present.
(TXT)

Acknowledgments

We thank Dorien van de Berg and the staff at the Tuberculosis Reference Laboratory, RIVM for technical assistance. We would also like to thank all collaborators worldwide that provided strains for typing.

Author Contributions

Conceived and designed the experiments: ACS KK DvS. Performed the experiments: ACS KK BF ACAH. Analyzed the data: ACS ACAH CREM. Contributed reagents/materials/analysis tools: RvC MJB PvH RMW. Wrote the paper: ACS KK RMW CREM RJS DvS. Reviewed the paper: ACS KK ACAH BF CREM RvC MJB PvH RMW RJS DvS.

References

- WHO (2010) Global tuberculosis control: key findings from the December 2009 WHO report. *Wkly Epidemiol Rec* 85: 69–80.
- 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.
- van Soolingen D, Kremer K, Hermans PWM (2007) Molecular epidemiology; breakthrough achievements and future prospects. Single-Nucleotide Polymorphism-Based Population Genetic Analysis of *Mycobacterium tuberculosis* Strains from 4 Geographic Sites. *Tuberculosis* 2007 1: Chapter 9.
- Parwati I, van Crevel R, van Soolingen D (2010) Possible underlying mechanisms for successful emergence of the *Mycobacterium tuberculosis* Beijing genotype strains. *Lancet Infect Dis* 10: 103–111.
- Devaux I, Kremer K, Heersma H, Van Soolingen D (2009) Clusters of multidrug-resistant *Mycobacterium tuberculosis* cases, Europe. *Emerg Infect Dis* 15: 1052–1060.
- Glynn JR, Kremer K, Borgdorff MW, Mar PR, van Soolingen D (2006) Beijing/W Genotype *Mycobacterium tuberculosis* and Drug Resistance. European Concerted Action on New Generation Genetic Markers and Techniques for the Epidemiology and Control of Tuberculosis. *Emerging Infectious Disease* 12: 736–743.
- Kremer K, van-der-Werf MJ, Au BK, Anh DD, Kam KM, et al. (2009) Vaccine-induced immunity circumvented by typical *Mycobacterium tuberculosis* Beijing strains. *Emerg Infect Dis* 15: 335–339.
- Lopez B, Aguilar D, Orozco H, Burger M, Espitia C, et al. (2003) A marked difference in pathogenesis and immune response induced by different *Mycobacterium tuberculosis* genotypes. *Clin Exp Immunol* 133: 30–37.
- 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.
- Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, et al. (2004) Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains. *Proc Natl Acad Sci U S A* 101: 4865–4870.
- Kremer K, Glynn JR, Lillebaek T, Niemann S, Kurepina NE, et al. (2004) Definition of the Beijing/W Lineage of *Mycobacterium tuberculosis* on the Basis of Genetic Markers. *J Clin Microbiol* 42: 4040–4049.
- Pliakytis BB, Marden JL, Crawford JT, Woodley CL, Butler WR, et al. (1994) Multiplex PCR assay specific for the multidrug-resistant strain W of *Mycobacterium tuberculosis*. *J Clin Microbiol* 32: 1542–1546.
- Schürch AC, Kremer K, Warren RM, Hung NV, Zhao Y, et al. (2011) Mutations in the regulatory network that underlie the recent clonal expansion of a dominant subclone of the *Mycobacterium tuberculosis* Beijing genotype. *Infect Genet Evol* 49(4): 1318–22.
- Iwamoto T, Yoshida S, Suzuki K, Wada T (2008) Population structure analysis of the *Mycobacterium tuberculosis* Beijing family indicates an association between certain sublineages and multidrug resistance. *Antimicrob Agents Chemother* 52: 3805–3809.
- Hanekom M, van der Spuy GD, Streicher E, Ndabambi SL, McEvoy CRE, et al. (2007) A Recently Evolved Sublineage of the *Mycobacterium tuberculosis* Beijing Strain Family Is Associated with an Increased Ability to Spread and Cause Disease. *J Clin Microbiol* 45: 1483–1490.
- Achtman M (2008) Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu Rev Microbiol* 62: 53–70.
- Fournier P-E, Drancourt M, Raoult D (2007) Bacterial genome sequencing and its use in infectious diseases. *Lancet Infect Dis* 7: 711–723.
- MacLean D, Jones JD, Studholme DJ (2009) Application of 'next-generation' sequencing technologies to microbial genetics. *Nat Rev Microbiol* 7: 287–296.
- Ding C, Jin S (2009) High-throughput methods for SNP genotyping. *Methods Mol Biol* 578: 245–254.
- Kuroda M, Serizawa M, Okutani A, Sekizuka T, Banno S, et al. (2010) Genome-wide single nucleotide polymorphism typing method for identification of *Bacillus anthracis* species and strain among *B. cereus* group species. *J Clin Microbiol* 48: 2821–2829.
- Gilmour MW, Graham M, Van Domselaar G, Tyler S, Kent H, et al. (2010) High-throughput genome sequencing of two *Listeria monocytogenes* clinical isolates during a large foodborne outbreak. *BMC Genomics* 11: 120.
- Monot M, Honore N, Garnier T, Zidane N, Sherafi D, et al. (2009) Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. *Nat Genet* 41: 1282–1289.
- Röltgen K, Qi W, Ruf MT, Mensah-Quainoo E, Pidot SJ, et al. (2010) Single nucleotide polymorphism typing of *Mycobacterium ulcerans* reveals focal transmission of buruli ulcer in a highly endemic region of Ghana. *PLoS Negl Trop Dis* 4: e751.
- Qi W, Käser M, Röltgen K, Yeboah-Manu D, Pluschke G (2009) Genomic diversity and evolution of *Mycobacterium ulcerans* revealed by next-generation sequencing. *PLoS Pathog* 5: e1000580.
- van Gent M, Bart MJ, van der Heide HG, Heuvelman KJ, Kallonen T, et al. (2011) SNP-Based Typing: A Useful Tool to Study *Bordetella pertussis* Populations. *PLoS ONE* 6: e20340.
- Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393: 537–544.
- Gutacker MM, Smoot JC, Migliaccio CAL, 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.
- 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.
- van Soolingen D, de Haas PE, van Doorn HR, Kuijper E, Rinder H, et al. (2000) Mutations at amino acid position 315 of the *katG* gene are associated with high-level resistance to isoniazid, other drug resistance, and successful transmission of *Mycobacterium tuberculosis* in the Netherlands. *J Infect Dis* 182: 1788–1790.
- Baker L, Brown T, Maiden MC, Drobniwski F (2004) Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. *Emerg Infect Dis* 10: 1568–1577.
- Niemann S, Koser CU, Gagneux S, Plinke C, Homolka S, et al. (2009) Genomic diversity among drug resistant and multidrug resistant isolates of *Mycobacterium tuberculosis* with identical DNA fingerprints. *PLoS One* 4: e7407.
- 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.
- 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: 537–544.
- Cubillos-Ruiz A, Morales J, Zambrano MM (2008) Analysis of the genetic variation in *Mycobacterium tuberculosis* strains by multiple genome alignments. *BMC Res Notes* 1: 110.
- Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, et al. (2002) Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J Bacteriol* 184: 5479–5490.

36. Stavrum R, Valvatne H, Bø TH, Jonassen I, Hinds J, et al. (2008) Genomic Diversity among Beijing and non-Beijing *Mycobacterium tuberculosis* Isolates from Myanmar. *PLoS ONE* 3: e1973.
37. Mokrousov I, Narvskaya O, Otten T, Vyazovaya A, Limeschenko E, et al. (2002) Phylogenetic reconstruction within *Mycobacterium tuberculosis* Beijing genotype in northwestern Russia. *Res Microbiol* 153: 629–637.
38. Dos Vultos T, Mestre O, Raugier J, Golec M, Rastogi N, et al. (2008) Evolution and Diversity of Clonal Bacteria: The Paradigm of *Mycobacterium tuberculosis*. *PLoS ONE* 3: e1538.
39. Abadia E, Zhang J, dos Vultos T, Ritacco V, Kremer K, et al. (2010) Resolving lineage assignment on *Mycobacterium tuberculosis* clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method. *Infect Genet Evol* 10: 1066–1074.
40. 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.
41. Mestre O, Luo T, Dos Vultos T, Kremer K, Murray A, et al. (2011) Phylogeny of *Mycobacterium tuberculosis* Beijing Strains Constructed from Polymorphisms in Genes Involved in DNA Replication, Recombination and Repair. *PLoS ONE* 6: e16020.
42. Kremer K, Au BK, Yip PCW, Skuce R, Supply P, et al. (2005) Use of Variable-Number Tandem-Repeat Typing To Differentiate *Mycobacterium tuberculosis* Beijing Family Isolates from Hong Kong and Comparison with IS6110 Restriction Fragment Length Polymorphism Typing and Spoligotyping. *J Clin Microbiol* 43: 314–320.
43. Kremer K, van Soolingen D, Frothingham R, Haas WH, Hermans PW, et al. (1999) Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility. *J Clin Microbiol* 37: 2607–2618.
44. Rad ME, Bifani P, Martin C, Kremer K, Samper S, et al. (2003) Mutations in Putative Mutator Genes of *Mycobacterium tuberculosis* Strains of the W-Beijing Family. *Emerging Infectious Diseases* 44: 838–845.
45. Flores L, Van T, Narayanan S, DeRiemer K, Kato-Maeda M, et al. (2007) Large sequence polymorphisms classify *Mycobacterium tuberculosis* strains with ancestral spoligotyping patterns. *J Clin Microbiol* 45: 3393–3395.
46. Jurinke C, van den Boom D, Cantor CR, Koster H (2001) Automated genotyping using the DNA MassArray technology. *Methods Mol Biol* 170: 103–116.
47. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52: 696–704.
48. Huson DH, Richter DC, Rausch C, DeZulian T, Franz M, et al. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* 8: 460.
49. Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinform* 9: 299–306.
50. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* 25: 3389–3402.
51. Hunter PR (1990) Reproducibility and indices of discriminatory power of microbial typing methods. *J Clin Microbiol* 28: 1903–1905.
52. 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.
53. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, et al. (2002) A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proceedings of the National Academy of Sciences* 99: 3684–3689.
54. Gagneux S, Small PM (2007) Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *Lancet Infect Dis* 7: 328–337.
55. Huard RC, Lazzarini LC, Butler WR, van Soolingen D, Ho JL (2003) PCR-based method to differentiate the subspecies of the *Mycobacterium tuberculosis* complex on the basis of genomic deletions. *J Clin Microbiol* 41: 1637–1650.
56. Mostowy S, Cousins D, Brinkman J, Aranaz A, Behr MA (2002) Genomic deletions suggest a phylogeny for the *Mycobacterium tuberculosis* complex. *J Infect Dis* 186: 74–80.
57. Narayanan S, Gagneux S, Hari L, Tsolaki AG, Rajasekhar S, et al. (2007) Genomic interrogation of ancestral *Mycobacterium tuberculosis* from south India. *Infect Genet Evol* 8(4): 474–483.
58. Smith NH, Kremer K, Jacqueline I, Dale J, Driscoll JR, et al. (2006) Ecotypes of the *Mycobacterium tuberculosis* complex. *Journal of theoretical biology* 239: 220–225.
59. Mathema B, Kurepina NE, Bifani PJ, Kreiswirth BN (2006) Molecular epidemiology of tuberculosis: current insights. *Clin Microbiol Rev* 19: 658–685.
60. Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, et al. (2010) *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nat Genet* 42: 1140–1143.
61. 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.
62. Reed MB, Pichler VK, McIntosh F, Mattia A, Fallow A, et al. (2009) Major *Mycobacterium tuberculosis* lineages associate with patient country of origin. *J Clin Microbiol* 47: 1119–1128.
63. Mokrousov I (2008) Genetic geography of *Mycobacterium tuberculosis* Beijing genotype: a multifacet mirror of human history? *Infect Genet Evol* 8: 777–785.
64. van Helden PD, Warren RM, Victor TC, van der Spuy G, Richardson M, et al. (2002) Strain families of *Mycobacterium tuberculosis*. *Trends Microbiol* 10: 167–168. author reply 168.
65. 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.