Infection, Genetics and Evolution 10 (2010) 1066–1074

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Resolving lineage assignation on *Mycobacterium tuberculosis* clinical isolates classified by spoligotyping with a new high-throughput 3R SNPs based method

Edgar Abadia [a], Jian Zhang [a], Tiago dos Vultos [b], Viviana Ritacco [c], Kristin Kremer [d], Elif Aktas [e], Tomoshige Matsumoto [f], Guislaine Refregier [a], Dick van Soolingen [d], Brigitte Gicquel [b], Christophe Sola [a,b,*]

[a] *Institute of Genetics and Microbiology, UMR8621, IGEPE Team, Universud, CNRS Université Paris-Sud 11, France*
[b] *Unité de Génétique Mycobactérienne, Institut Pasteur, Paris, France*
[c] *Instituto Nacional de Enfermedades Infecciosas ANLIS Carlos Malbrán, Vélez Sarsfield 563, 1281 Buenos Aires, Argentina*
[d] *National Institute for Public Health and the Environment, Bilthoven, The Netherlands*
[e] *Department of Clinical Microbiology, Zonguldak Karaelmas University Hospital, Zonguldak, Turkey*
[f] *Osaka Prefectural Hospital Organization Osaka Prefectural Medical Center for Respiratory and Allergic Diseases, Japan*

ARTICLE INFO

Article history:
Received 19 May 2010
Received in revised form 1 July 2010
Accepted 2 July 2010
Available online 17 July 2010

Keywords:
Mycobacterium tuberculosis complex
CRISPR
Spoligotyping
Population studies
Phylogeny
Evolution
SNPs
Multiplex
High-throughput

ABSTRACT

We developed a new multiplexed-PCR assay to accurately classify *Mycobacterium tuberculosis* complex (MTC) isolates at the sublineage level by single nucleotide polymorphisms (SNPs). This method relies on 7 SNPs located in different genes of the MTC strains (*recC*, *rec0*, *recR*, *ligB*, *ligC*, *alkA*, and *mgtC*). Most of these genes are involved in replication, repair and recombination (3R) functions of *M. tuberculosis* strains, four of the mutations are synonymous, and thus neutral. Genes were chosen as a first empirical approach to assess the congruence between spoligotyping-based phylogeographical classification and SNP typing.

This scheme efficiently classifies most of MTC phylogeographical groups: (1) confirming and identifying new sublineage-specific SNPs, (2) unraveling phylogenetical relationships between spoligotyping-defined MTC sublineages, (3) appropriately assigning sublineages to some spoligotypes and reassigning sublineages to other mis-labeled spoligotype signatures. This study opens the way to a more meaningful taxonomic, evolutionary and epidemiological classification. It also allows evaluation of spoligotype-signature significance towards a more comprehensive understanding of the evolutionary mechanisms of the clustered regularly interspaced short palindromic repeat (CRISPR) locus in MTC.

1. Introduction

Despite the availability of effective antituberculosis chemotherapy for over 50 years (Styblo and Bumgarner, 1991), TB remains a major global health problem since the World Health Organization declared tuberculosis (TB) a global emergency in 1994 (Nakajima, 1993). The spread of multi-drug resistant tuberculosis (MDR-TB) and more recently of extremely drug resistant tuberculosis (XDR-TB) (Ralph et al., 2009), makes the implementation of public health measures, and molecular epidemiological investigations using rapid and high-throughput molecular methods an important point to follow TB transmission.

Current genotyping techniques used to study the epidemiology of *Mycobacterium tuberculosis* complex (MTC) clinical isolates are based on repetitive genetic elements: Clustered Repetitive Interspersed Short Palindromic Repeat (CRISPR) loci through the spoligotyping technique (Groenen et al., 1993; Kamerbeek et al., 1997; van Embden et al., 2000; Sorek et al., 2008) and Mycobacterial Interspersed Repetitive Units-Variable Number of Tandem repeats (MIRU-VNTR) (Frothingham and Meeker-O'Connell, 1998; Supply et al., 2000, 2001, 2006; Le Fleche et al., 2002; Skuce et al., 2002). Indeed, these markers have proven to be highly useful for epidemiological, population structural and evolutionary studies to distinguish between MTC clinical isolates (Abadia et al., 2009; Allix-Beguec et al., 2008a; Baranov et al., 2009; Brudey et al., 2006; Helal et al., 2009; Rohani et al., 2009; Stavrum et al., 2009) and have been used as the alternative to the classical IS*6110*-RFLP method (van Embden et al., 1993). These methods have in addition received recent technological improvements enabling fast and large-scale analyses to be performed (Cowan et al., 2004; Mazars et al., 2001). The classical spoligotyping procedure (Kamerbeek et al., 1997) relies on a reverse-line blot hybridization, a procedure that takes one full day of work to produce 43 profiles without interpretation; with the automatization now, 96 profiles can be

* Corresponding author at: Institute of Genetics and Microbiology, UMR8621, IGEPE Team, Universud, CNRS Université Paris-Sud 11, Campus d'Orsay, F-91405 Orsay Cedex, France; Tel.: +33 01 69 15 46 48; fax: +33 01 69 15 66 78.
E-mail address: christophe.sola@u-psud.fr (C. Sola).

1567-1348/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2010.07.006

obtained in half a day (Cowan et al., 2004; Zhang et al., 2010). This new technology relies on microbeads of different spectrum signatures to which capture probes are coupled depending on the targets (Cowan et al., 2004).

For phylogenetical evolutionary studies there is a concern using the fast evolving loci due to the presence of convergent evolutionary events within CRISPRs (Warren et al., 2002) as well as within MIRU-VNTR loci (Hanekom et al., 2008). To define phylogenetic associations unambiguously, genetic markers need to be unique and, ideally, irreversible (Comas et al., 2009).

In *M. tuberculosis*, these markers are large sequence polymorphisms (LSPs) (Hirsh et al., 2004; Mostowy et al., 2002) and single nucleotide polymorphisms (SNPs) (Alland et al., 2003; Dos Vultos et al., 2008; Filliol et al., 2006; Gutacker et al., 2006; Hershberg et al., 2008). LSPs are powerful markers in MTC because horizontal DNA transfer is extremely rare (Supply et al., 2003), but genetic distances based on genomic deletions are difficult to interpret in phylogenies (Gagneux et al., 2006). SNPs are less mutable than other forms of polymorphisms, making them unlikely to converge (Schork et al., 2000) so they are most appropriate markers for phylogenetic studies. In MTC strains structural genes exhibit rare polymorphism (Achtman, 2008; Kapur et al., 1994; Musser et al., 2000; Sreevatsan et al., 1997). However, recently Dos Vultos et al. (2008) have found higher polymorphisms in several genes involved in replication, recombination and repair functions (3R genes). For the first time, we provide a starting point of a new SNP typing of *M. tuberculosis* complex clinical isolates based mainly on 3R genes. We implemented this schedule on a high-throughput platform using a direct hybridization assay (Dunbar, 2006), that was used before for spoligotyping analyses (Cowan et al., 2004; Zhang et al., 2010).

We think that this could be the first step to provide a reliable high-throughput 3R SNP-based method for population structural studies and for further phylogenetical studies on *M. tuberculosis* complex clinical isolates.

2. Materials and methods

2.1. Chemicals, buffers and microbeads

All main materials and reagents required for microbead-based flow cytometry techniques were the same as described before (Zhang et al., 2010).

2.2. Oligonucleotides

To design capture probes and primers we used a demo of PrimerPlex (http://www.premierbiosoft.com/primerplex/index.html) and Primer3Plus (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi). All capture oligonucleotides (Eurogentec, Liège, Belgium) were manufactured and coupled as described before (Zhang et al., 2010). In this study we are targeting 7 genes that are polymorphic within *M. tuberculosis* sublineages, 6 of them from the 3R system described in Dos Vultos et al. (2008) and one (*mgtC*) reported by Alix et al. (2006). The primers and capture probes used here are listed in Tables 1 and 2, respectively.

2.3. Sequences analysis

Gene sequences were downloaded from http://genolist.pasteur.fr/TubercuList/. The DNAsp package (Rozas et al., 2003), Multalin (Corpet, 1988) and BioEdit sequence alignment editor (Hall, 1999) were used to locate SNPs on gene sequences and to align the gene and probe sequences in the design and/or verification of pre-designed capture probes.

Table 1
Genes, lineage or sublineage association, genomic position of targeted SNPs (up) and primers designed for the multiplex-PCR (down).

Gene[a]	Lineage or sublineage association	SNP targeted[b]
alkA	Bovine	807
recO	EAI	606
ligB	LAM	1212
recR	T2-related	94
mgtC	Haarlem	545
recC	X	1491
ligC	TUR-T3-Osaka	809

Primers

Gene	Sequence (5′–3′)	bp
alkA-F	CACGCTACGGCTCCCATG	18
alkA-R[c]	CCTTCGTCGATACCTGTTGGG	21
recO-F	TGTTGGACGCCTATCTGCTG	20
recO-R[c]	CCGTTCCAGATGCCATTGC	19
ligB-F	GGCTGGCTGAAGGTCAAG	18
ligB-R[c]	CATGGCGTCGGTCATTCC	18
recR-F	GGACCTGATTGACGAACTCG	20
recR-R[c]	GCCTGGATGTCTTTGGGTTC	20
mgtC-F	TCGTCGCTGTCCATCTCC	18
mgtC-R[c]	CACCAACCGCTCTAGCTTG	19
recC-F	CGCCGAACTGCTCACCATC	19
recC-R[c]	GCCACGCTTGGGAATCCTC	19
ligC-F	CGCGTCGGTCGGCGTGAT	18
ligC-R[c]	CGGGTCGCAGCGGCACGA	18

[a] Genes from the 3R system (Dos Vultos et al., 2008), except *mgtC* (Alix et al., 2006).
[b] nt position related to the gene.
[c] Reverse primers are biotin labeled.

2.4. SNPs typing PCR protocol

For direct hybridization of multiplexed-PCR assays, product length is recommended to be between 150 and 300 bp. We amplified segments around 200 bp of *alkA* (291 bp), *recO* (298 bp), *ligB* (174 bp), *recR* (255 bp), *mgtC* (272 bp) and *recC* (272 bp) to analyse the correlation between SNPs and major MTC lineages. To increase signals/cut-off ratio, PCR-multiplex was firstly divided into 3 sets, set1: *ligB*, *recR*, *mgtC* and *recC*; set2: *alkA* and *recO* and set3: *ligB* and *ligC*; however, running a single 7-Plex PCR protocol now provides similar results (see Supplemental Table 1). PCR assays were performed in 25 μL volumes of the following mixture: PCR buffer (100 mM Tris–HCl, pH 8.3, 500 mM KCl), 0.2 mM each dNTP, a final quantity of primers in a solution of 25 pmol of each one, 1.0 U Taq and 2 μL of DNA. The following PCR program was run: 5 min at 95 °C, followed by a touch-down PCR of 30 s at 95 °C,

Table 2
Sequence of the capture probes used to target the SNPs in this study.

Probes			
Probe name	Sequence 5′–3′	bp	Targeted site (nt)
alkA-wt	CGCGACCTGATGACGGC	17	807 ct**g** (leu) to
alkA-mut	CGCGACCTAATGACGGC	17	807 ct**a** (leu)
recO-wt	GTACGACGGCGATTGGGA	18	606 gg**c** (gly) to
recO-mut	GTACGACGGTGATTGGGA	18	606 gg**t** (gly)
ligB-wt	GGCAAGCTCTCCAATATTCACC	22	1212 tc**c** (ser) to
ligB-mut	GGCAAGCTCTCGAATATTCACC	22	1212 tc**g** (ser)
RecR-wt	CTTCCACCTGTTGTCGGTAGA	21	94 **t**tg (leu) to
RecR-mut	CTTCCACCTGCTGTCGGTAGA	21	94 **c**tg (leu)
mgtC-wt	GGGGTATACGCACGGGGC	18	545 c**g**c (arg) to
mgtC-mut	GGGGTATACACACGGGGC	18	545 c**a**c (his)
RecC-wt	GTGGCGGTTCGGACTCGA	18	1491 tt**c** (phe) to
RecC-mut	GTGGCGGTTGGGACTCGA	18	1491 tt**g** (leu)
ligC-wt	GACCACCCATGGAACTGGGCC	21	809 t**g**g (trp) to
ligC-mut	GACCACCCATTGAACTGGGCC	21	809 t**t**g (leu)

All capture probes have at 5′ a C-12 terminal linker. We are targeting both alleles, the wild type and the mutant for each site.

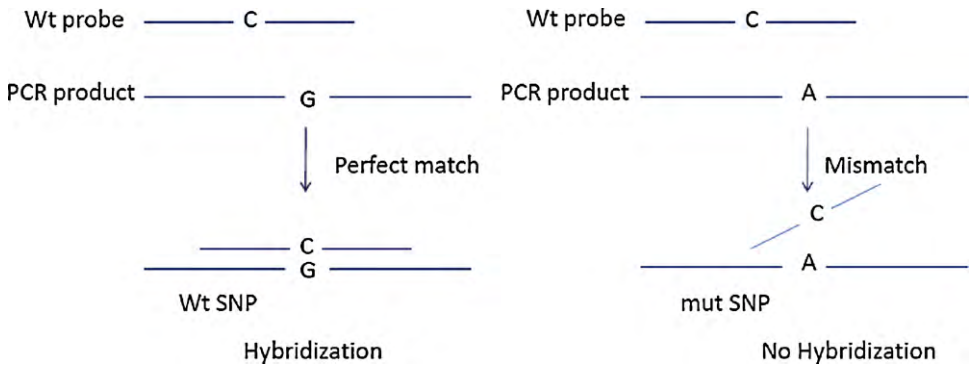

Fig. 1. Allele-specific hybridization, image modified from (Kwok, 2001).

30 s at 65 °C (−1 °C per step), 1 min at 72 °C 10 times and 30 s at 95 °C, 30 s at 56 °C, 1 min at 72 °C for 29 more cycles with a final extension step of 10 min at 72 °C. The PCR amplifications were checked during first tests by agarose gel electrophoresis.

2.5. Hybridization

The procedure detecting the two possible alleles relies on an allele-specific hybridization (Kwok, 2001). Fig. 1 shows how the SNPs were discriminated based on this approach. If the PCR product harbors the wild type (wt) SNP it will match to the wt probe that is complementary to the sequence. This hybridized product is very stable because there is a 100% complementarity. If the PCR product has the mutated (mut) SNP allele, the hybridization is not thermodynamically stable due to the nucleotide mismatch and it will be lost. For this kind of procedure the mismatch base should be placed in the middle of the probe. The hybridization procedure on microbeads was done as described previously (Zhang et al., 2010) except for the second part of the hybridization that was performed at 52 °C for 20 min.

2.6. Mycobacterium isolates, origin, DNA extraction

327 DNA samples (a summary of the lineages being targeted can be found in Supplemental Table 2) were genotyped by our novel SNP typing scheme using a high-throughput microbead-based method, as reported previously (Cowan et al., 2004; Zhang et al., 2010). The clinical isolates for this study were chosen to cover the main lineages of *M. tuberculosis* complex clinical isolates described before by spoligotyping (Filliol et al., 2002; Brudey et al., 2006). They were selected from an international quality control study on membrane-based spoligotyping vs. high-throughput-based spoligotyping (Abadia et al., unpublished results) as well as from the 2004–2008 collection of the TB National Reference Laboratory in the Netherlands, from Zonguldak hospital, Turkey and from a Medical Center in Japan. More specifically, 59 samples were from Buenos Aires – Argentina (Servicio de Micobacterias, Instituto Nacional de Enfermedades Infecciosas, ANLIS "Carlos G. Malbran", Buenos Aires, Argentina), 109 samples came from Bilthoven – The Netherlands (National Institute for Public Health and the Environment – RIVM), 120 DNA samples from the Faculty of Medicine, Department of Microbiology and Clinical Microbiology, Zonguldak Karaelmas University (Zonguldak – Turkey), 40 DNA samples from Japan (Osaka Prefectural Medical Center for Respiratory and Allergic Diseases). The DNA from these samples was extracted either by a simple thermolyzate or by the classical cetyltrimethylammonium bromide (CTAB) procedure.

2.7. Data analysis

Standard MTC lineage assignation (*i.e.* according to spoligotype pattern) was reported either according to spolDB4 database (Brudey et al., 2006) or using the SPOTCLUST algorithm (Vitol et al., 2006) online access http://cgi2.cs.rpi.edu/~bennek/Run.html).

To infer lineages according to SNPs, the presence of allele putative-lineages associated (Dos Vultos et al., 2008) was reported. Hybridization signals for the mutated or wild type SNP alleles, collected as RFI (Relative Fluorescence Intensity) were transformed in a binary code (presence/absence of each allele) using a signal/noise cut-off value of 2.

3. Results

The aim of this study was: (1) to test some associations between spoligotype-defined lineages and a set of potentially lineage-specific SNPs mainly located on 3R genes (Alix et al., 2006; Dos Vultos et al., 2008) on a representative set of samples; (2) to develop a new multiplexed high-throughput assay for this purpose; (3) to assign a lineage to those clinical isolates for which spoligotype signatures were uninformative, the so-called "U" clinical isolates in SpolDB4 (Brudey et al., 2006); and finally (4) to confirm previous lineage associations or correct possible mis-assignations. We thus aimed to further contribute to the study of the MTC molecular evolution.

We searched on the 3R gene database (Dos Vultos et al., 2008) for SNPs that were polymorphic between main MTC lineages and we found that SNPs on: *alkA* (807 ctg [leu] to cta [leu]), *recO* (606 ggc [gly] to ggt [gly]), *ligB* (1212 tcc [ser] to tcg [ser]), *recC* (1491 ttc [phe] to ttg [leu]), *recR* (94 ttg [leu] ctg [leu]), *ligC* (809 tgg [trp] to ttg [leu]) apparently were associated with *M. bovis*, East African (EAI), the Latin-American Mediterranean (LAM), the X, the T2 and the LAM7_TUR/T3_Osaka *M. tuberculosis* lineages, respectively. We added to this set of genes the SNP on *mgtC* 545 cgc [arg]) to cac [his] because it was described previously as being associated with the Haarlem lineage (Alix et al., 2006) and recently had been tested by Chuang et al. (2008). The mgtC protein is a common virulence factor to several intracellular pathogens (Buchmeier et al., 2000).

Our scheme thus targets 7 lineages: 5 large lineages covering a total of 31 sublineages, and two specific sublineages. Other SNPs belonging to other panel of genes could also be added in a near future to our assay. We have developed a single multiplexing reaction including this set of 7 genes to test the previous finding of 3R SNP MTC lineage association in a representative set of samples. Central-Asian (CAS) specific and East-Asian (Beijing) specific SNPs are not reported at this stage of work but are also in progress.

To validate our new SNP typing method, we used 327 DNA samples of *M. tuberculosis* clinical isolates, all having been previously spoligotyped on microbeads using the high-throughput technique (Zhang et al., 2010). These samples were representative of a highly diverse MTC population within the total worldwide genetic diversity of MTC: they covered unclassified strains, as well as 32 sublineages over the 62 that have been named in SpolDB4, representing all main lineages infecting humans including some that were not targeted by our SNP scheme (Beijing, CAS, EAI, LAM, bovis, Haarlem, X, S and T). Only *Mycobacterium africanum* and *Mycobacterium canettii* and Manu lineages were not represented. Out of the 32 sublineages that were represented in our data set, 21 were targeted by our scheme.

Among the seven mutations tested within our scheme we never detected two mutations simultaneously in any of the 327 typed strains, leading either to SNP-specific assignment or to SNP-nonspecific assignment. In addition, none of the isolates assigned to lineages Beijing, CAS, S and T4 did show any of the mutation present in our SNP scheme as expected. This highlights the specificity of the SNP we chose, and confirms the absence of convergence events in this collection. For sixteen sublineages (76%, covering 73 different isolates) targeted by our scheme and present in our data set, complete congruence between spoligotyping-based classification and SNP-based assignation was observed, indicating a high global fit between the two classifications: 6/6 *M. bovis*, 55 EAI (EAI1_SOM, EAI2_MANILLA, EAI5, EAI3), 28/38 H and 14/14 from sublineages H1 and H2, 23/37 LAM and 23/23 from sublineages LAM2, LAM3, LAM5, LAM9, 2/2 X (X1, X3), 23/23 T3-Osaka. For five sublineages that were targeted by the scheme, part or all of the assignations were different between the two classifications: 4/18 clinical isolates from H3, 6/6 from H4, 28/34 from T2, 7/7 from LAM3-S convergent, 40/40 LAM7_TUR (Table 3 for an overview; spoligotype patterns and SNP genotypes in Supplemental Table 3). One T5 sublineage (T5_RUS1) and 6 T1-Tuscany, and one T1 however carried the LAM-specific SNP, 2 other T1 isolates carried the H-specific mutation, and 2 other T1 isolates the T3-Osaka specific SNP, confirming that T-clade is poorly defined.

More specifically, not all of the clinical isolates labeled as Haarlem lineage (H1 to H4) according to their spoligotype signatures in SpolDB4 ($n = 38$) harbored the SNP Haarlem associated on *mgtC*$^{545\ (cgc\text{-}cac)}$ (Alix et al., 2006). All H1 spoligo-signatures ($n = 11$) and all H2 spoligo-signatures ($n = 3$) had this specific mutation whereas 14/18 of H3 spoligo-signature and none of the H4 ($n = 6$) had. Among the four H3 clinical isolates that did not harbor the *mgtC* mutation, two belonged to SIT316, one to

Table 3
SNP lineage assignation on a subset of analyzed strains and comparison with SpolDB4 linage classification to show in some cases the lineage confirmation and in others a lineage re-assignation.

		Lineage[a]		
SIT	Spoligotype	SpolDB4	SNP	No. of strains[e]
35	■■■■■■■■■■■□■■■■■■■■■■□■□□□■■■■■■	H4	-	2
262	■■■■■■■□■■■■■■■■■■■■■□□□■□□□■■■■■■	H4	-	3
777	■■■■■■■■■■■■■■■■■■■■■■■■■□□□■□□□■■■■■■	H4	-	1
760	■■■■■■■□□□□□■■■■■□□□□□□□□□□□□■□□□■■■□□□	H3	-	1
134	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□■■□□■■	H3	X	1
316	■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□■■■□■■	H3	T2	2
47[b]	■■■■■■■■■■■■■■■■■■■■■■■■□□□□□□■□□□■■■■■■	H1	H	1
50[b]	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□■■■■■■	H3	H	9
41	■■■■■■■■■■■■■■■■□□□□■□□■■■■■□□□■■■■■■	LAM7_TUR	TUR- T3- Osa	6
367	■■■■■■■■■■■■□■■■■■□□□■□□■■■■■□□□■■■■■■	LAM7_TUR	TUR- T3- Osa	1
4	□□□□□□□□□□□□□□□□□□□□□□■■■■■■■□□□■■■■■■	LAM3- S/conv	-	7
42[b]	■■■■■■■■■■■■■■■■■■■■□□□■■■■■■■□□□■■■■■■	LAM9	LAM	10
33[b]	■■■■■■■■■□□■■■■■■■■□□□■■■■■■■□□□■■■■■■	LAM3	LAM	4
93[b]	■■■■■■■■■■■■■■■■■■■□■■■■■■■□□□■■■■■■	LAM5	LAM	1
17[b]	■■□■■■■■■■■■□■■■■■■■□□□■■■■■■■□□□■■■■■■	LAM2	LAM	2
52	■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□■■■□■■	T2	T2[c]	31[c]
254	■■■■■■■■■■■■■■■□□□□□□□□□■■■■■■■□□□■■■■■■	T5_RUS1	LAM	1
736	■■■■■■■■■■■■■■■■■■■□■■■■■■■■■□□□■■■□■■	T2	-	1
627	■■■■□□□■■■■□■■■■■■■■■■■■■■■■■□□□■■■■■■	T3-OSA	TUR- T3- Osa	40
78	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□■■■□■■	T1-T2	TUR- T3- Osa	2
53**	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□■■■■■■	T1	TUR- T3- Osa[d]	25[d]
159	■■■■■■■■■■■■■□□□□□□□□■■■■■■■□□□■■■■■■	T1 (Tuscany v.)	LAM	7
370	■■■■■■■■■■■■■■■■■■■■□□■■■■■■■□□□■■□■■	T1	LAM	1
53	■■■■■■■■■■■■■■■■■■■■■■■■■■■■■□□□■■■■■■	T1	H	1
120	■■■■■■■■■■■■■■■■■■■□■■■■■■■■■□□□■■■■■■	T1	H	1

[a]Family assignation according to spoligo-signatures and according to SNP identification.
[b]Reference strains for each lineage.
[c]Just 6/31 had the mutation on *recR*$^{94\ (ttg\text{-}ctg)}$ T2-related.
[d]2/25 strains SIT 53 had the *ligC*^{809} SNP.
[e]Number of strains related to the frequency of the spoligotype pattern.
To read and interpret the spoligotype, spacers are placed in order from 1 to 43 (left to right).
"-": Means no detectable SNP.

SIT760 and the other to SIT134. The SIT134 "H3" harbored the *recC*$^{1491\ (ttc\text{-}ttg)}$ X lineage associated SNP and the two SIT316 "H3" clinical isolates harbored the mutation on *recR*$^{94\ (ttg\text{-}ctg)}$ hypothesized to be linked to the T2 lineage isolates from Central African Republic clinical isolates (Dos Vultos et al., 2008) (see also Table 3 and Supplemental Table 3).

Within LAM-labeled clinical isolates (n = 72), we found that 47 did not harbor the expected mutation on *ligB*$^{1212\ (tcc\text{-}tcg)}$, forty of these were LAM7_TUR and the seven others were LAM3-S (Table 3). All others LAM sublineages tested so far (LAM9, LAM5, LAM3, and LAM2) did harbor the expected *ligB* mutation.

Within the T lineage, there are a lot of sublineages defined by spoligotype signatures (Brudey et al., 2006), among which one is T2. *recR*$^{94\ (ttg\text{-}ctg)}$ SNP was found associated with Central African Republic isolates, all of them carrying a T2 signature (absence of spacer 40) (Dos Vultos et al., 2008). Here, 83 DNA samples belonged to the T lineage (sublineages: either T1, T1-Tuscany variant, T1–T2 undefined, T2, T4_CEU1, T5 and T5_RUS1), out of 32 T2 isolates, six only harbored the mutation on *recR*$^{94\ (ttg\text{-}ctg)}$ (Table 3). These results show that the *recR*^{94} mutation cannot be considered as being linked to the absence of spacer 40, which defines the T2 sublineage.

Table 4
Lineage identification of spoligotype that lack of it, the so-called "U" spoligotypes in SpolDB4 (Brudey et al., 2006), either because they were not reported in the database, or because this identification was not clear. Lineage assignment by SNP and by SPOTCLUST (Vitol et al., 2006) is shown.

			Family			
Pos[a]	SIT	Spoligotype	SpolDB4	SNP	No. of strains[c]	SPOTCLUST[b]
1	NR		U	EAI	2	EAI5 (0.95)
2	NR		U	EAI	1	EAI5 (1.00)
3	NR		U	H	1	Family 34 (1.00)
4	NR		U	H	3	Family 35 (1.00)
5	NR		U	H	1	H1 (1.00)
6	NR		U	H	2	H3 (0.77)
7	NR		U	H	1	H3 (0.96)
8	1274		U	H	1	H1 (1.00)
9	NR		U	H	1	H3 (0.77)
10	NR		U	H	1	H1 (1.00)
11	NR		U	H	1	H1 (1.00)
12	NR		U	H	2	H3 (0.77)
13	NR		U	H	1	H1 (1.00)
14	NR		U	LAM	2	T1 (0.80)
15	NR		U	LAM	1	LAM3 (0.93)
16	NR		U	LAM	4	LAM9 (1.00)
17	NR		U	LAM	1	LAM8 (0.94)
18	NR		U	LAM	1	LAM3 (1.00)
19	NR		U	LAM	1	LAM8 (0.99)
20	NR		U	LAM	1	LAM3 (1.00)
21	NR		U	LAM	1	LAM9 (0.96)
22	NR		U	LAM	1	LAM9 (1.00)
23	NR		U	LAM	1	T4 (1.00)
24	NR		U	LAM	2	LAM (1.00)
25	105		U (LAM3?)	LAM	2	LAM7 (1.00)
26	NR		U	LAM	1	LAM8 (1.00)
27	NR		U	LAM	1	LAM9 (0.97)
28	NR		U	LAM	1	T4 (1.00)
29	1531		U	X	1	X1 (0.65)
30	NR		U	X	2	X3 (1.00)
31	NR		U	X	1	H3 (0.77)
32	2125[d]		U	TUR- T3- Osa	1	LAM9 (1.00)
33	NR		U	TUR- T3- Osa	1	LAM9 (0.99)
34	NR		U	TUR- T3- Osa	1	T1 (1.00)
35	NR		U	TUR- T3- Osa	1	T1 (1.00)

[a]This number is provided just to locate the pattern in this list.
[b]SPOTCLUST Identification according the spoligotype pattern and the algorithm probability of the result of being true in parenthesis.
[c]Number of strains related to the frequency of the spoligotype pattern.
[d]SIT number from Aktas et al. (2008).
To read and interpret the spoligotype, spacers are placed in order from 1 to 43 (left to right).

An interesting secondary result found in this study is the confirmation of the phylogenetic link between the clinical isolates from the T3_Osaka sublineage and the ones from the LAM7_TUR. These two groups of strains share the same MIRU-VNTR[1] 12 loci signature (MIRU-VNTR International Type, MIT310) and harbor divergent spoligotypes (SIT41 and SIT627). Here, using a large set of DNA samples from Zonguldak – Turkey (LAM7_TUR) and from Japan (T3_Osaka and other variants), our results show that they shared the same SNP on *ligC*$^{809\ (tgg\ [trp]\ to\ ttg\ [leu])}$. Moreover, we observed that 2/2 SIT78 (T1–T2) and 2/25 SIT53 (T1) also had this SNP.

Our study also included 84 spoligotypes for which the lineage information was unknown, the so-called "U" isolates (for Unknown), *i.e.* isolates for which the spoligotype signature either did not allow assignation to a lineage ($n = 7$) (Brudey et al., 2006) or had not been reported in the international spoligotype database, SpolDB4 ($n = 77$, referred here as "NR"). With our SNP-based scheme, 47 clinical isolates (56%, with same frequency for strict "U" and "NR") could now be assigned at a lineage level (Table 4). The identification of these clinical isolates is distributed as follows: 21 isolates were assigned to the LAM-lineage, 15 to Haarlem, 4 to X and 3 to the EAI lineage. We also tested the congruence between our SNP classification and SPOTCLUST, a Probabilistic Multivariate Bernoulli Mixture Model that uses Naïve Bayes assumptions, as reported previously (Vitol et al., 2006). This algorithm automatically provides a probability for a spoligotype pattern to be part of a lineage using previous assignations reported in SpolDB4 (Brudey et al., 2006). Interestingly, we found a good correlation (72%) between SPOTCLUST assignation and our SNP identification (Table 4). However, we also observed discrepancies (positions 3, 4, 14, 23, 28 and 31 in Table 4). Indeed, SPOTCLUST classified these clinical isolates as: Lineage 34, Lineage 35, T1, T4, T4 and H3, respectively with probabilities of higher than 0.77 whereas our results classify these samples as H, H, LAM, LAM, LAM and X, respectively. Note that TUR_T3_Osaka strains were identified by SPOTCLUST as LAM9 or T1 with probabilities close to 1.00 (Table 4).

4. Discussion

We aimed to confirm the spoligotype-based classification of MTC using a restricted SNP-based scheme designed to be performed in a high-throughput way. Most commonly used classification of MTC relies on spoligotype patterns. Lineages and sublineages have been defined based on recurrent so-called spoligo-signatures (absence and/or presence of specific spacers) that have proven to be phylogeographically meaningful (Brudey et al., 2006). The existence of these lineages was confirmed using independent techniques such as LSP (Gagneux et al., 2006), MIRU-VNTR typing (Allix-Beguec et al., 2008b; Wirth et al., 2008)), or via sequencing projects (Comas et al., 2009).

To provide phylogenetically reliable information, markers should not be prone to converge. SNPs most of the time are unlikely to convergence and also are less prone to distortion by selective pressures (Schork et al., 2000). Notwithstanding drug resistance genes, there is just one evidence of convergence evolution due to a nonsense variation in ada/alkA gene observed in *M. tuberculosis* and in *M. bovis* that could confer a selective advantage (Nouvel et al., 2007), a SNP that is not included in our data set.

In *M. tuberculosis*, and maybe other clonal organisms, the genes involved in replication, repair and recombination (3R) seemed to play a key role in adaptation (Dos Vultos et al., 2008) so they might be prone to experience selection. However, in our study, four of the targeted SNPs are synonymous (sSNPs) so they are unlikely to be selected for. In addition, the three other non-synonymous mutations chosen are conservative so the replaced amino acid is biochemically similar to the pre-existing one. Altogether, the SNPs we provide are likely to be neutral. Regarding classification, we confirmed that EAI, BOVIS and X lineages are monophyletic, *i.e.* that the signature that defines them is phylogenetically relevant: EAI lineage is defined by the absence of spacers 29–32, absence of 34 and the *recO*$^{606\ (ggc\text{-}ggt)}$ mutation; the BOVIS lineage is both defined by the absence of spacers 39–43 and the *alkA* $^{807\ (ctg\text{-}cta)}$ mutation; the X lineage is defined by the absence of spacer 18, 33–36 group of spacers and by the *recC*$^{1491\ (ttc\text{-}ttg)}$ mutation. The results we provide come in addition to previous results obtained (Dos Vultos et al., 2008) so that we can say that there is strong support for these correlations.

This study in contrast provides some evidence of mis-labeling in LAM, T and Haarlem lineages due to the used marker itself. Regarding Haarlem lineage, H4 sublineage did not carry the expected *mgtC*$^{545\ (cgc\text{-}cac)}$ mutation. Kovalev et al. (2005) showed that according to their MIRU-VNTR signatures (especially MIRU10 with 7–10 copies) this lineage was unlikely to be phylogenetically related to Haarlem so it was renamed as URAL. Here, we confirm that the loss of spacer 29–31 in URAL is a genetic event that is independent from the loss of spacer 26–31 in H1 and H2, or of spacer 31 in H3. For the time being we have not identified a SNP that correlates with this lineage. Interestingly, all the strains that carried this signature in our scheme came from Turkey that is geographically close to Russia.

Two SITs (134 and 316) that belong to the H3 lineage (according to their spoligotype signatures) lacked the *mgtC*^{545} SNP Haarlem-specific mutation but instead harbored the *recC*^{1491}-X and *recR*^{94}-T2 lineage-specific mutations, respectively that may indicate that the deletion of spacer 31 could have occurred several times independently so that it could be used cautiously as a phylogenetic marker. We suggest that a "Haarlem" lineage assignment for strains classified as H3 could therefore be confirmed by the SNP method described here.

The X lineage was first defined by the absence of spacer 18 (Sebban et al., 2002; Filliol et al., 2002) in addition to the absence of spacers 33–36. Here, we found a strain that was not labeled as X according to its spoligotype, but in fact it was according to its SNP pattern. This suggests that the deletion of spacer 18 occurred after the *recC*^{1491} SNP event. Hence, we suggest to broaden X lineage to all strains carrying this SNP.

Within the T2 sublineage (defined by the absence of spacer 40, in addition to the absence of 33–36 group of spacers) many clinical isolates sharing SIT52 or SIT736 spoligotype did not harbor the SNP on *recR*^{94}. In addition a relatively low number of *recR*^{94} mutations were found in the whole set of strains, suggesting that this mutation occurred too recently. For the time being, we are not able to well describe the T lineage and resolve all the relationships between modern strains carrying the 33–36 spacer deletion, because there is no single SNP available to do that by now and due to the big genetic diversity inside this main lineage, it remains "ill-defined" or poorly defined (Brudey et al., 2006). Also, Vitol et al. (2006) were unable to create models that discriminate well among the members of the T lineage based on their spoligotype signatures.

Another contribution to the T sublineage markers is the confirmation (with strong support, $n = 62$) of the existence of a TUR_T3_Osaka lineage based on the *ligC*^{809} SNP. Spoligotypes that harbors this SNP are very different between them (Table 3 and Supplementary Table 1). Millet et al. (2007) reported before a shared MIRU12-VNTR pattern called MIT310 (215125113322) between strains from T3_Osaka lineage (Takashima and Iwamoto, 2006) and the previously reported LAM7_TUR lineage (Zozio et al., 2005; Millet et al., 2007). All the LAM7_TUR strains tested so far lacked the mutation on the LAM-associated *ligB*^{1212} SNP, consequently they are

[1] Mycobacterial Interspersed Repetitive Unit-Variable Number of Tandem Repeat.

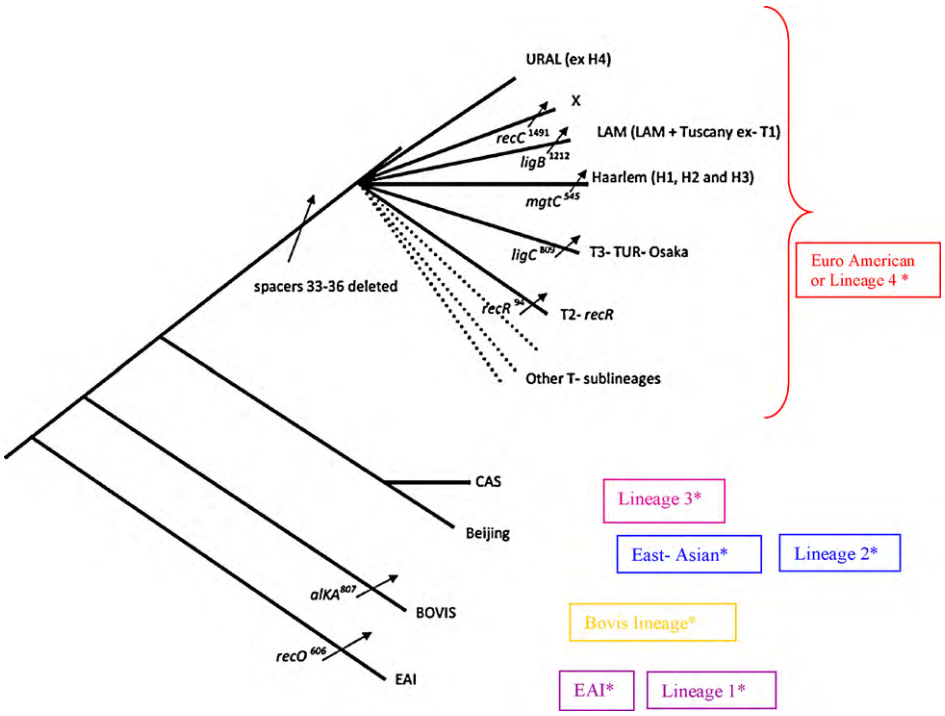

Fig. 2. Evolutionary proposal to identify *M. tuberculosis* families based on the SNP used in this study. Small arrows show the genes in which the SNP is present to allow the identification. The big arrow shows the deletion of spacers 33–36 that is common to all the modern strains. It is shown the correspondence between major lineages defined by Spoligotyping, our SNP proposal and inside boxes the one related to LSP plus a recently work done in 89 sequenced genes. *The LSP classification is related to the work of (Gagneux et al., 2006) and recently to an SNP work of (Comas et al., 2009).

not a sublineage of LAM so we give further support to renaming them TUR (Table 3). We also observed that strains carrying SIT78 T1–T2, SIT53 T1 (2/25), or SIT553 T2–T3 and 4 NR strains also harbored the TUR-T3-Osaka-*ligC*^{809} mutation. This suggests that this mutation could be useful digging inside a scheme targeting T_lineage genetic diversity.

Another result is that one SIT53 T1 strain, which is both wt for *recR*^{94} and for *ligC*^{809}, harbors the Haarlem-specific mutation for *mgtC*^{545} (Table 3). This finding could corroborate the hypothesis made before by others (Duchene et al., 2004) about the relation between T and Haarlem strains. In this case, one possibility is that strains harboring the SIT53 T1 spoligo-signature may have evolved at a genome level to Haarlem in spite of retaining the CRISPR region of the progenitor. This could be also the case for SIT120 T1 for which we found that it also harbors the mutation for Haarlem on *mgtC*^{545}, the only difference at a spoligo level between SIT53 and 120 is one spacer missing, spacer 20 (Table 3).

With our SNP-signature panel we also provide evidence favoring the renaming of several spoligotypes previously assigned to T-clade based on spoligo rules, these are: SIT159 T1 (Tuscany variant), SIT254 (T5_RUS1) and SIT370 (T1) that harbored the *ligB*^{1212} SNP-LAM-associated mutation so we assigned the right lineage to these strains. The LAM assignation rule is based on the simultaneous absence of spacers 21–24 and 33–36 that is not accomplished by SIT159, SIT 254 T5_RUS1or SIT370. Taken together, this finding could be an indication that spacers 21 and 22 are not so informative to define the LAM lineage as is the absence of spacer 23.

One of the most important features of our new SNP proposal relies on the identification of those spoligotypes that lack the lineage identification. Emerging clones continuously will appear as a result of genetic variation of pre-existing ones (Brudey et al., 2006). The SNP assignment is unambiguous, providing a precise tool that targets several sites (7 in our actual proposal) with a high precision performing the lineage assignation. We could successfully assign a lineage to 47 out of 84 U strains. One interesting lineage assignment in the group of "U" strains was the SIT 105 that had been responsible for an MDR-TB outbreak in Spain. This SIT was thought to be "LAM3?" in SpolDB4 with doubts and now we could confirm it because it was found to harbor the *ligB*^{1212}-LAM-specific mutation.

Even though Spoligotyping remains a first line tool to delineate the molecular ecology of the circulating strains, we have shown that sometimes the spoligotype signature will not always reflect the real lineage of the MTC strains. With this work, we have solved some intraspecies taxonomic issues and we provide an accurate lineage assignation using a minimal set of SNPs almost as informative as large sequencing projects similar to those of Comas et al. (2009). However, we still need to include other SNPs for other lineages, because there are still spoligotype patterns that we were not able to identify with this proposal. This is also true to reach the level of spoligotyping sublineage discrimination. Nevertheless we could correctly identify strains belonging to the major MTC lineages or families such as EAI, BOVIS, X, LAM, Haarlem as well as some T sublineages, such as T2-*recR*^{94} and TUR_T3_OSAKA. We do not have one specific SNP for the T or S strain lineage for the time being. In Fig. 2 we show an evolutionary scenario built with the set of SNPs used in this study and a hypothesis of when the deletion of spacers 33–36 may have occurred.

Some SNP-based studies have already been performed in *M. tuberculosis* to date (Alland et al., 2003; Baker et al., 2004; Filliol et al., 2006; Gutacker et al., 2006; Gutacker et al., 2002; Sreevatsan et al., 1997). However, the composition and number of SNP cluster groups (SCGs) within *M. tuberculosis* have remained unclear since it also used some inadequate genes and SNPs (i.e. involved in drug resistance) (Baker et al., 2004) whereas others were selected from a non-representative set of available genomes (Filliol et al., 2006; Gutacker et al., 2002, 2006). These previous studies had not taken full advantage of the power of SNP-based methods. Here we show that SNP typing could serve as a "gold standard" for DNA typing. We think that in the near future the inclusion of more SNP information will become an important parameter to efficiently

classify a given MTC clone, either for molecular epidemiological or for evolutionary purposes. High-throughput multiplexing among various techniques and platforms will be one of the best ways to achieve this with reasonable economic costs (Dunbar, 2006; Bergval et al., 2008).

Acknowledgments

This study was supported by the "Excellency Chair in microbiology of the Université Paris-Sud 11" granted to CS in September 2007, through a 12-month post-doctoral fellowship grant given to EA. EA is a senior mycobacteriology scientist of the IVIC (Instituto Venezolano de Investigaciones Científicas, Caracas, Venezuela). JZ is a PhD fellow of the Université Paris-Sud 11 through the "Ecole Doctorale Gènes Génomes Cellules". We are grateful to François Topin, Luminex BV, The Netherlands, and to Luminex Corporation, Austin, Texas for their support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.meegid.2010.07.006.

References

Abadia, E., Sequera, M., Ortega, D., Mendez, M.V., Escalona, A., Da Mata, O., Izarra, E., Rojas, Y., Jaspe, R., Motiwala, A.S., Alland, D., de Waard, J., Takiff, H.E., 2009. *Mycobacterium tuberculosis* ecology in Venezuela: epidemiologic correlates of common spoligotypes and a large clonal cluster defined by MIRU-VNTR-24. BMC Infect. Dis. 9, 122.

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

Aktas, E., Zozio, T., Comert, F.B., Kulah, C., Aydin, O., Rastogi, N., Sola, C., 2008. A first insight into the genetic diversity and population structure of *Mycobacterium tuberculosis* in Zonguldak, Turkey. Clin. Microbiol. Infect. 55–59.

Alix, E., Godreuil, S., Blanc-Potard, A.B., 2006. Identification of a Haarlem genotype-specific single nucleotide polymorphism in the mgtC virulence gene of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 44, 2093–2098.

Alland, D., Whittam, T.S., Murray, M.B., Cave, M.D., Hazbon, M.H., Dix, K., Kokoris, M., Duesterhoeft, A., Eisen, J.A., Fraser, C.M., Fleischmann, R.D., 2003. Modeling bacterial evolution with comparative-genome-based marker systems: application to *Mycobacterium tuberculosis* evolution and pathogenesis. J. Bacteriol. 185, 3392–3399.

Allix-Beguec, C., Fauville-Dufaux, M., Supply, P., 2008a. Three-year population-based evaluation of standardized mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 46, 1398–1406.

Allix-Beguec, C., Harmsen, D., Weniger, T., Supply, P., Niemann, S., 2008b. 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, 2692–2699.

Baker, L., Brown, T., Maiden, M.C., Drobniewski, F., 2004. Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. Emerg. Infect. Dis 10, 1568–1577.

Baranov, A.A., Mariandyshev, A.O., Mannsaker, T., Dahle, U.R., Bjune, G.A., 2009. Molecular epidemiology and drug resistance of widespread genotypes of *Mycobacterium tuberculosis* in northwestern Russia. Int. J. Tuberc. Lung Dis. 13, 1288–1293.

Bergval, I.L., Vijzelaar, R.N., Dalla Costa, E.R., Schuitema, A.R., Oskam, L., Kritski, A.L., Klatser, P.R., Anthony, R.M., 2008. Development of multiplex assay for rapid characterization of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 46, 689–699.

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

Buchmeier, N., Blanc-Potard, A., Ehrt, S., Piddington, D., Riley, L., Groisman, E.A., 2000. A parallel intraphagosomal survival strategy shared by *Mycobacterium tuberculosis* and *Salmonella enterica*. Mol. Microbiol. 35, 1375–1382.

Chuang, P.C., Liu, H., Sola, C., Chen, Y.M., Jou, R., 2008. Spoligotypes of *Mycobacterium tuberculosis* isolates of a high tuberculosis burden aboriginal township in Taiwan. Infect. Genet. Evol. 8, 553–557.

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

Corpet, F., 1988. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res. 16, 10881–10890.

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

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

Duchene, V., Ferdinand, S., Filliol, I., Guégan, J.F., Rastogi, N., Sola, C., 2004. Phylogenetic reconstruction of the *Mycobacterium tuberculosis* complex within four settings of the Caribbean region: tree comparative analysis and first appraisal on their phylogeography. Infect. Gen. Evol. 4, 5–14.

Dunbar, S.A., 2006. Applications of Luminex xMAP technology for rapid, high-throughput multiplexed nucleic acid detection. Clin. Chim. Acta 363, 71–82.

Filliol, I., Driscoll, J.R., Van Soolingen, D., Kreiswirth, B.N., Kremer, K., Valetudie, G., Anh, D.D., Barlow, R., Banerjee, D., Bifani, P.J., Brudey, K., Cataldi, A., Cooksey, R.C., Cousins, D.V., Dale, J.W., Dellagostin, O.A., Drobniewski, F., Engelmann, G., Ferdinand, S., Gascoyne-Binzi, D., Gordon, M., Gutierrez, M.C., Haas, W.H., Heersma, H., Kallenius, G., Kassa-Kelembho, E., Koivula, T., Ly, H.M., Makristathis, A., Mammina, C., Martin, G., Mostrom, P., Mokrousov, I., Narbonne, V., Narvskaya, O., Nastasi, A., Niobe-Eyangoh, S.N., Pape, J.W., Rasolofo-Razanamparany, V., Ridell, M., Rossetti, M.L., Stauffer, F., Suffys, P.N., Takiff, H., Texier-Maugein, J., Vincent, V., De Waard, J.H., Sola, C., Rastogi, N., 2002. Global distribution of *Mycobacterium tuberculosis* spoligotypes. Emerg. Infect. Dis. 8, 1347–1349.

Filliol, I., Motiwala, A.S., Cavatore, M., Qi, W., Hernando Hazbon, M., Bobadilla Del Valle, M., Fyfe, J., Garcia-Garcia, L., Rastogi, N., Sola, C., Zozio, T., Guerrero, M.I., Leon, C.I., Crabtree, J., Angiuoli, S., Eisenach, K.D., Durmaz, R., Joloba, M.L., Rendon, A., Sifuentes-Osornio, J., Ponce de Leon, A., Cave, M.D., Fleischmann, R., Whittam, T.S., Alland, D., 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.

Frothingham, R., Meeker-O'Connell, W.A., 1998. Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. Microbiology 144 (Pt 5), 1189–1196.

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

Groenen, P.M., Bunschoten, A.E., van Soolingen, D., van Embden, J.D., 1993. Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. Mol. Microbiol. 10, 1057–1065.

Gutacker, M.M., Mathema, B., Soini, H., Shashkina, E., Kreiswirth, B.N., Graviss, E.A., Musser, J.M., 2006. Single-nucleotide polymorphism-based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. J. Infect. Dis. 193, 121–128.

Gutacker, M.M., Smoot, J.C., Migliaccio, C.A., Ricklefs, S.M., Hua, S., Cousins, D.V., Graviss, E.A., Shashkina, E., Kreiswirth, B.N., Musser, J.M., 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.

Hall, T., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for indows 95/98/NT. Nucleic Acids Symp. Ser. 41, 95–98.

Hanekom, M., van der Spuy, G.D., Gey van Pittius, N.C., McEvoy, C.R., Hoek, K.G., Ndabambi, S.L., Jordaan, A.M., Victor, T.C., van Helden, P.D., Warren, R.M., 2008. Discordance between mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing and IS6110 restriction fragment length polymorphism genotyping for analysis of *Mycobacterium tuberculosis* Beijing strains in a setting of high incidence of tuberculosis. J. Clin. Microbiol. 46, 3338–3345.

Helal, Z.H., Ashour, M.S., Eissa, S.A., Abd-Elatef, G., Zozio, T., Babapoor, S., Rastogi, N., Khan, M.I., 2009. Unexpectedly high proportion of ancestral Manu genotype *Mycobacterium tuberculosis* strains cultured from tuberculosis patients in Egypt. J. Clin. Microbiol. 47, 2794–2801.

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

Hirsh, A.E., Tsolaki, A.G., DeRiemer, K., Feldman, M.W., Small, P.M., 2004. Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. Proc. Natl. Acad. Sci. U.S.A. 101, 4871–4876.

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

Kapur, V., Whittam, T.S., Musser, J.M., 1994. Is *Mycobacterium tuberculosis* 15,000 years old? J. Infect. Dis. 170, 1348–1349.

Kovalev, S.Y., Kamaev, E.Y., Kravchenko, M.A., Kurepina, N.E., Skorniakov, S.N., 2005. Genetic analysis of *Mycobacterium tuberculosis* strains isolated in Ural region, Russian federation, by MIRU-VNTR genotyping. Int. J. Tuberc. Lung Dis. 9, 746–752.

Kwok, P.Y., 2001. Methods for genotyping single nucleotide polymorphisms. Annu. Rev. Genomics Hum. Genet. 2, 235–258.

Le Fleche, P., Fabre, M., Denoeud, F., Koeck, J.L., Vergnaud, G., 2002. High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing. BMC Microbiol. 2, 37.

Mazars, E., Lesjean, S., Banuls, A.L., Gilbert, M., Vincent, V., Gicquel, B., Tibayrenc, M., Locht, C., Supply, P., 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.

Millet, J., Miyagi-Shiohira, C., Yamane, N., Sola, C., Rastogi, N., 2007. Assessment of mycobacterial interspersed repetitive unit-QUB markers to further discriminate the Beijing genotype in a population-based study of the genetic diversity of *Mycobacterium tuberculosis* clinical isolates from Okinawa, Ryukyu Islands, Japan. J. Clin. Microbiol. 45, 3606–3615.

Mostowy, S., Cousins, D., Brinkman, J., Aranaz, A., Behr, M.A., 2002. Genomic deletions suggest a phylogeny for the *Mycobacterium tuberculosis* complex. J. Infect. Dis. 186, 74–80.

Musser, J.M., Amin, A., Ramaswamy, S., 2000. Negligible genetic diversity of *Mycobacterium tuberculosis* host immune system protein targets: evidence of limited selective pressure. Genetics 155, 7–16.

Nakajima, H., 1993. Tuberculosis: a global emergency. World Health Forum 14, 438.

Nouvel, L.X., Dos Vultos, T., Kassa-Kelembho, E., Rauzier, J., Gicquel, B., 2007. A nonsense mutation in the putative anti-mutator gene ada/alkA of *Mycobacterium tuberculosis* and *M. bovis* isolates suggests convergent evolution. BMC Microbiol. 7, 39.

Ralph, A.P., Anstey, N.M., Kelly, P.M., 2009. Tuberculosis into the 2010s: is the glass half full? Clin. Infect. Dis. 49, 574–583.

Rohani, M., Farnia, P., Nasab, M.N., Moniri, R., Torfeh, M., Amiri, M.M., 2009. Beijing genotype and other predominant *Mycobacterium tuberculosis* spoligotypes observed in Mashhad city, Iran. Indian J. Med. Microbiol. 27, 306–310.

Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X., Rozas, R., 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19, 2496–2497.

Schork, N.J., Fallin, D., Lanchbury, J.S., 2000. Single nucleotide polymorphisms and the future of genetic epidemiology. Clin. Genet. 58, 250–264.

Sebban, M., Mokrousov, I., Rastogi, N., Sola, C., 2002. A data-mining approach to spacer oligonucleotide typing of *Mycobacterium tuberculosis*. Bioinformatics 18, 235–243.

Skuce, R.A., McCorry, T.P., McCarroll, J.F., Roring, S.M., Scott, A.N., Brittain, D., Hughes, S.L., Hewinson, R.G., Neill, S.D., 2002. Discrimination of *Mycobacterium tuberculosis* complex bacteria using novel VNTR-PCR targets. Microbiology 148, 519–528.

Sorek, R., Kunin, V., Hugenholtz, P., 2008. CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat. Rev. Microbiol. 6, 181–186.

Sreevatsan, S., Pan, X., Stockbauer, K.E., Connell, N.D., Kreiswirth, B.N., Whittam, T.S., Musser, J.M., 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.

Stavrum, R., Mphahlele, M., Ovreas, K., Muthivhi, T., Fourie, P.B., Weyer, K., Grewal, H.M., 2009. High diversity of *Mycobacterium tuberculosis* genotypes in South Africa and preponderance of mixed infections among ST53 isolates. J. Clin. Microbiol. 47, 1848–1856.

Styblo, K., Bumgarner, J., 1991. Tuberculosis can be controlled with existing technologies: evidence. Tuberculosis Surveillance Research Unit of the IUATLD Progress Report 2, 60–72.

Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rusch-Gerdes, S., Willery, E., Savine, E., de Haas, P., van Deutekom, H., Roring, S., Bifani, P., Kurepina, N., Kreiswirth, B., Sola, C., Rastogi, N., Vatin, V., Gutierrez, M.C., Fauville, M., Niemann, S., Skuce, R., Kremer, K., Locht, C., van Soolingen, D., 2006. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 44, 4498–4510.

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

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

Supply, P., Warren, R.M., Banuls, A.L., Lesjean, S., Van Der Spuy, G.D., Lewis, L.A., Tibayrenc, M., Van Helden, P.D., Locht, C., 2003. Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area. Mol. Microbiol. 47, 529–538.

Takashima, T., Iwamoto, T., 2006. New era in molecular epidemiology of tuberculosis in Japan. Kekkaku 81, 693–707.

van Embden, J.D., Cave, M.D., Crawford, J.T., Dale, J.W., Eisenach, K.D., Gicquel, B., Hermans, P., Martin, C., McAdam, R., Shinnick, T.M., et al., 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J. Clin. Microbiol. 31, 406–409.

van Embden, J.D.A., van Gorkom, T., Kremer, K., Jansen, R., van der Zeijst, B.A.M., Schouls, L.M., 2000. Genetic variation and evolutionary origin of the Direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. J. Bacteriol. 182, 2393–2401.

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

Warren, R.M., Streicher, E.M., Sampson, S.L., Van Der Spuy, G.D., Richardson, M., Nguyen, D., Behr, M.A., Victor, T.C., Van Helden, P.D., 2002. Microevolution of the direct repeat region of *Mycobacterium tuberculosis*: implications for interpretation of spoligotyping data. J. Clin. Microbiol. 40, 4457–4465.

Wirth, T., Hildebrand, F., Allix-Beguec, C., Wolbeling, F., Kubica, T., Kremer, K., van Soolingen, D., Rusch-Gerdes, S., Locht, C., Brisse, S., Meyer, A., Supply, P., Niemann, S., 2008. Origin, spread and demography of the *Mycobacterium tuberculosis* complex. PLoS Pathog. 4, e1000160.

Zhang, J., Abadia, E., Refregier, G., Tafaj, S., Boschiroli, M.L., Guillard, B., Andremont, A., Ruimy, R., Sola, C., 2010. *Mycobacterium tuberculosis* complex CRISPR genotyping: improving efficiency, throughput and discriminative power of 'spoligotyping' with new spacers and a microbead-based hybridization assay. J. Med. Microbiol. 59, 285–294.

Zozio, T., Allix, C., Günal, S., Saribas, S., Alp, A., Durmaz, R., Fauville-Dufaux, M., Rastogi, N., Sola, C., 2005. The previously identified "LAM7" genotype typifies a Turkish-phylogeographically specific clonal complex of *Mycobacterium tuberculosis*, 26th Annual Congress of the European Society for Mycobacteriology June 26th–29th 2005. Abstract Book, O-8 *presenting author, Istanbul, Turkey, 26–29 June 2005, p. 31.