BMC Microbiology

Research article

High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing

Philippe Le Flèche[1,2], Michel Fabre[3], France Denoeud[2], Jean-Louis Koeck[4] and Gilles Vergnaud*[1,2]

Address: [1]Centre d'Etudes du Bouchet BP3, 91710 Vert le Petit, France, [2]GPMS, Bât. 400, Institut de Génétique et Microbiologie, Université Paris Sud, 91405 Orsay cedex, France, [3]Laboratoire de Biologie Clinique, HIA Percy, 92141 Clamart, France and [4]Département de biologie médicale, HIA Val-de-Grâce, 75230 Paris, France

E-mail: Philippe Le Flèche - lefleche@igmors.u-psud.fr; Michel Fabre - mfabre@free.fr; France Denoeud - France.Denoeud@igmors.u-psud.fr; Jean-Louis Koeck - jlkoeck@filnet.fr; Gilles Vergnaud* - Gilles.Vergnaud@igmors.u-psud.fr

*Corresponding author

Published: 27 November 2002

BMC Microbiology 2002, **2**:37

Received: 17 September 2002
Accepted: 27 November 2002

This article is available from: http://www.biomedcentral.com/1471-2180/2/37

Abstract

Background: Currently available reference methods for the molecular epidemiology of the *Mycobacterium tuberculosis* complex either lack sensitivity or are still too tedious and slow for routine application. Recently, tandem repeat typing has emerged as a potential alternative. This report contributes to the development of tandem repeat typing for *M. tuberculosis* by summarising the existing data, developing additional markers, and setting up a freely accessible, fast, and easy to use, internet-based service for strain identification.

Results: A collection of 21 VNTRs incorporating 13 previously described loci and 8 newly evaluated markers was used to genotype 90 strains from the *M. tuberculosis* complex (*M. tuberculosis* (64 strains), *M. bovis* (9 strains including 4 BCG representatives), *M. africanum* (17 strains)). Eighty-four different genotypes are defined. Clustering analysis shows that the *M. africanum* strains fall into three main groups, one of which is closer to the *M. tuberculosis* strains, and an other one is closer to the *M. bovis* strains. The resulting data has been made freely accessible over the internet [http://bacterial-genotyping.igmors.u-psud.fr/bnserver] to allow direct strain identification queries.

Conclusions: Tandem-repeat typing is a PCR-based assay which may prove to be a powerful complement to the existing epidemiological tools for the *M. tuberculosis* complex. The number of markers to type depends on the identification precision which is required, so that identification can be achieved quickly at low cost in terms of consumables, technical expertise and equipment.

Background

The precise identification of bacterial pathogens at the strain level is essential for epidemiological purposes. Consequently, constant efforts are undertaken to develop easy to use, low cost and standardized methods which can eventually be applied routinely in a clinical laboratory. Newer developments are usually genetic methods based on PCR (Polymerase Chain Reaction) to type variations directly at the DNA level. The development of polymorphic markers is now further facilitated by the availability of whole genome sequences for bacterial genomes. Recently, it has been shown that tandem repeat (usually

called minisatellites or VNTRs for Variable Number of Tandem Repeats) loci provide a source of very informative markers not only in humans where some are still in use for identification purposes (paternity analyses, forensics) but also in bacteria. Tandem repeats are easily identified from genome sequence data, the typing of tandem repeat length is relatively straight forward, and the resulting data can be easily coded and exchanged between laboratories independently of the technology used to measure PCR fragment sizes. Furthermore, the resolution of tandem repeats typing is cumulative, i.e. the inclusion of more markers in the typing assay can, when necessary, increase the identification resolution. However, the density of tandem repeats in bacterial genomes varies from species to species, and not all tandem repeats are polymorphic [1]. In addition, some tandem repeats are so unstable that they have no or little long-term epidemiological value [2]. This indicates that for each species under consideration, tandem repeats must be evaluated using representative collections of strains before they can be used. Tandem repeats for bacterial identification have already proved their utility for the typing of the highly monomorphic pathogens *Bacillus anthracis*, *Yersinia pestis*, [1] and *M. tuberculosis*. In this last case, the value of tandem repeat based identification was recognised very early [3]. The so-called DR (direct repeat) locus is a relatively large tandem repeat locus of unknown biological significance. The motif is 72 bp long, one half is highly conserved, whereas the other half (called the spacer element) is highly diverged. The spoligotyping method [4] takes advantage of these internal variations to distinguish the hundreds of different alleles at this locus, which have been reported in the *M. tuberculosis* complex among the thousands of strains typed so far [5]. Although it is quite powerful, with many advantages, spoligotyping suffers from a lack of resolution compared to the current gold-standard in *M. tuberculosis* genetic identification, IS*6110* typing [6]. IS*6110* typing is an RFLP (Restriction Fragment Length Polymorphism) method using the mobile element IS*6110* as a probe. Strains with a low-copy number of IS*6110* elements (such as most *M. bovis* strains) are poorly resolved by this method. The so-called PGRS (polymorphic GC-rich sequence) method is an other RFLP approach in which the probe used is a GC-rich tandem repeat. The polymorphisms which are scored at multiple loci simultaneously on the Southern blot are variations in the tandem repeats length (and not internal variations at a single locus as assayed by spoligotyping). The profiles generated are very informative, but in comparison with IS*6110* typing, PGRS results are more difficult to score, because the intensity of the bands are highly variable (alleles with a small tandem array yield a lower hybridisation signal) [6]. Both PGRS and IS*6110* typing are hindered by the requirement for relatively large amounts of high quality DNA which is an issue for slow-growing mycobacteria.

More recently, and owing to the release of genome sequence data, the allele-length polymorphism of tandem repeat loci has been evaluated by PCR. Essentially three complementary sets of markers have been developed [7–9]. In the first report, exact tandem repeats (ETRs) were identified by searching the existing literature as well as early versions of the *M. tuberculosis* genome sequence data [7]. The resolution provided by this first set of five loci is lower than both IS*6110* RFLP typing and spoligotyping according to a comparative study [6]. In the second report, a family of tandem repeats characterized by similar repeat units was identified by sequence similarity search in the genome sequence data. A set of 12 loci was selected (including two of the five ETR loci) and the resulting panel has a resolution close to IS*6110* typing according to [10]. In the third report tandem repeats with highly conserved (>95%) motifs longer than 50 bp identified in the *M. tuberculosis* genome sequence have been investigated. Altogether, the currently available collection of polymorphic tandem repeats for the typing of *M. tuberculosis* comprises 27 loci (taking into account duplicates) (Table 1). Fifteen have a polymorphism index above 0.5.

This collection of markers should already provide a typing resolution comparable to the current reference methods. Given that not all tandem repeats present in *M. tuberculosis* have been evaluated for polymorphism, it is likely that the typing resolution of minisatellites could further be improved. Eventually, normalisation work will have to be done in order to promote the use of tandem repeats. A number of the loci analysed are known under different names in different studies, (for instance, ETRD [7] is also known as MIRU4 in [10]; and VNTR 0580 in [11]) and the coding (number of motifs in an allele) of alleles can also be different in different studies, for reasons explained in [11]. This is due in part to the fact that the number of repeats is not necessarily an integer value (Table 1). Furthermore, because the repeats in an array are not necessarily exact repeats, there can be ambiguities in the definition of the first and last base pair of the array. Finally, in addition to length variations due to the addition or deletion of an exact number of units, microdeletions or insertions within some repeat units are sometimes observed (MIRU4 is one such instance [12]).

One purpose of the present report is to contribute to the development of Multiple Loci VNTR Analysis (MVLA) through the evaluation of new markers and the setting up of an on-line identification tool for the *M. tuberculosis* complex which can be queried very easily with the user's personal data. In the present report, we first take advantage of the availability of genome sequence from two *M. tuberculosis* strains to complement the current collection of polymorphic tandem repeat markers. We identified *in silico* tandem repeats showing a different length in the two

Table 1: Polymorphic minisatellite markers for the *M. tuberculosis* complex

Locus name	"MIRU" alias [8]	"ETR" alias [7]	"QUB" alias [9,11]	Other alias	Reference	TR location on H37Rv genome	Expected length in H37Rv (copy number)	Expected length in CDC1551 (copy number)	Expected length in M bovis AF2122 (copy number)	N° of strains	Size range observed (copy number)	N° of alleles observed	Polymorphism index
H37Rv_0024_18 bp				Mtub01	This report	24648	328 (10)	310 (9)	310 (9)	92	274–328 bp (7–10)	4	0.48
H37Rv_0079_9 bp				Mtub02	This report	79503	230 (6)	239 (7)	239 (7)	92	221–275 bp (5–11)	7	0.76
H37Rv_0154_53 bp	MIRU2				[8]	154111	508 (2)	508 (2)	508 (2)	92	455–561 bp (1–3)	3	0.09
H37Rv_0424_51 bp				Mtub04	This report	424010	269 (2.6)	371 (4.6)	269 (2.6)	28	218 – 320 (1.6–3.6)	3	0.52
H37Rv_0531_15 bp				MPTR-A	[7]	531430	328 (16)	328 (16)	328 (16)	48	(15–17)	3	0.23
H37Rv_0577_58 bp		ETR-C			[8]	577172	346 (4)	288 (3)	404 (5)	92	230–404 bp (2–5)	4	0.63
H37Rv_0580_77 bp	MIRU4	ETR-D			[7]	580546	353 (3.3)	330 (3)	483 (5)	92	253–715 bp (2–8)	7	0.35
H37Rv_0802_54 bp	MIRU40				[8]	802194	199 (1)	415 (5)	253 (2)	92	199–469 bp (1–6)	5	0.71
H37Rv_0959_53 bp	MIRU10				[8]	959868	643 (3)	750 (5)	590 (2)	92	537–1014 bp (1–10)	9	0.76
H37Rv_1121_15 bp				Mtub12	This report	1121658	215 (4)	230 (5)	215 (4)	92	200–230 bp (3–5)	3	0.19
H37Rv_1443_56 bp				Mtub16	This report	1443417	291 (1)	347 (2)	347 (2)	11	291–515 (1–5)	3	0.56
H37Rv_1451_57 bp			QUB-1451c		[9]	1451778	305 (3.8)	305 (3.8)	305 (3.8)	56	(2–4) (bovis)	2	0.12
H37Rv_1612_21 bp			QUB-23		[11]	1612529	141 (5)	162 (6)	162 (6)	20	141–203 (5–8)	3	0.18
H37Rv_1644_53 bp	MIRU16				[8]	1644026	671 (2)	724 (3)	671 (2)	92	618–777 bp (1–4)	4	0.59
H37Rv_1895_57 bp			QUB-1895		[9]	1895344	319 (4)	205 (2)	319 (4)	56	(2–4) (bovis)	3	0.35
H37Rv_1955_57 bp				Mtub21	This report	1955580	206 (2)	263 (3)	263 (3)	92	149–491 bp (1–7)	7	0.76
H37Rv_1982_78 bp			QUB-18		[11]	1982873	621 (5)	777 (7)	465 (3)	24	387–1167 (2–12)	9	0.74
H37Rv_2059_77 bp	MIRU20				[8]	2059429	591 (2)	591 (2)	591 (2)	53	(1–2)	2	0.29
H37Rv_2074_56 bp*				Mtub24	This report	2074431	805 (3.6)	693 (1.6)	693 (1.6)	44	637–749 (0.6–2.6)	3	0.52
H37Rv_2163_a_69 bp			QUB-11a	pUCD1	[11]	2163607	305 (3)	581 (7)	788 (10)	92	305–1832 bp (3–26)	15	0.88
H37Rv_2163_b_69 bp			QUB-11b	pUCD1	[11]	2163729	412 (5)	274 (3)	343 (4)	52	136–826 (1–11)	8	0.82
H37Rv_2165_75 pb		ETR-A			[7]	2165223	397 (3)	322 (2)	847 (9)	92	322–847 bp (2–9)	8	0.73
H37Rv_2347_57 bp				Mtub29	This report	2347393	350 (4)	292 (3)	293 (3)	92	236–350 bp (2–4)	3	0.55
H37Rv_2401_58 bp				Mtub30	This report	2401815	319 (2)	435 (4)	435 (4)	92	261–435 bp (1–4)	3	0.55
H37Rv_2461_57 bp		ETR-B			[7]	2461279	292 (3)	235 (2)	406 (5)	92	178–406 bp (1–5)	6	0.51
H37Rv_2531_53 bp	MIRU23				[8]	2531560	873 (6)	820 (5)	767 (4)	92	608–979 bp (1–8)	7	0.60
H37RV_2387_54 bp	MIRU24				[8]	2684427	447 (1)	447 (1)	447 (1)	53	(1–2)	2	0.24
H37Rv_2990_55 bp				Mtub31	This report	2990582	257 (2)	312 (3)	312 (3)	49	202–312 bp (1–3)	3	0.15
H37Rv_2996_51 bp	MIRU26				[8]	2996002	614 (3)	716 (5)	716 (5)	57	563–818 (2–7)	5	0.61
H37Rv_3006_53 bp	MIRU27		QUB-5		[8]	3006875	657 (3)	657 (3)	657 (3)	92	551–710 bp (1–4)	4	0.25

Table 1: Polymorphic minisatellite markers for the *M. tuberculosis* complex *(Continued)*

H37Rv_3171_54 bp				Mtub34	This report	3171465	279 (3)	225 (2)	279 (3)	11	171–225 (1–2)	2	0.3
H37Rv_3192_53 bp	MIRU31	ETR-E			[7]	3192168	651 (3)	651 (3)	651 (3)	92	545–810 bp (1–6)	6	0.67
H37Rv_3232_56 bp			QUB-3232		[9]	3232649	591 (3)	760 (6)	703 (5)	56	(4–22) (bovis)	10	0.65
H37Rv_3239_79 bp		ETR-F			[7]	3239469	476 (2.8)	476 (2.8)	421 (2.1)	48	(1–3)	3	0.49
H37Rv_3336_59 bp			QUB-3336		[9]	3336499	407 (5)	466 (6)	289 (3)	56	(3–21) (bovis)	8	0.55
*H37Rv_3663_63 bp***				Mtub38	This report	3663751	373 (2.7)	310 (1.7)	310 (1.7)	92	247–400 bp (0.7–3.1)	5	0.35
*H37Rv_3690_58 bp**				Mtub39	This report	3690947	341 (2.6)*	397 (3.6)	341 (2.6)	92	247–1349 bp (1–20)*	11	0.64
H37Rv_4052_111 bp			QUB-26		[11]	4052969	708 (5)	819 (6)	597 (4)	100	(4–14) (bovis)	5	0.41
H37Rv_4156_59 bp			QUB-4156c		[9]	4156797	224 (2)	283 (3)	165 (1)	52	106–283 (0–3)	4	0.69
H37Rv_4348_53 bp	MIRU39				[8]	4348401	646 (2)	646 (2)	646 (2)	92	593–699 bp (1–3)	3	0.31

The markers are listed according to their position in the H37Rv genome. The proposed reference name includes the size of the repeat unit. The twenty-one markers used in the present report are italicised and underlined. Alias names identified in the literature are indicated. QUB11a, QUB11b, and ETR-A (position 2163–2165) are located within the gene PPE34 [19]. The expected length assumes that the primers listed in Table 2 were used. * : the observed size (Table 3) is not the expected size. ** : the repeat unit is not easily defined, size variations do not correspond to a multiple of 63 base-pairs. Polymorphism index is calculated as $1 - \sum$ (allele frequency)2 among the 86 distinct genotypes. The values are deduced from the original report in nine cases (indicated by the absence of size range in the "size range" column). In some instances [9,11], the population of strains used is biased (*M. bovis* strains).

strains using the previously described tandem repeat database [http://minisatellites.u-psud.fr][1]. Thirteen loci with a different predicted length in the two genomes and which have not been previously investigated have been tested for polymorphism and ease of typing.

Eight among the 13 polymorphic loci were used together with 13 among the previously described markers to genotype a collection of different *M. tuberculosis* complex strains. The data produced clusters the strains as suggested by morphological observations and biochemical analyses. The resulting data can be queried from a dedicated web page [http://bacterial-genotyping.igmors.u-psud.fr/bn-server].

Table 2: Set of primers for MLVA analysis

Locus name	forward primer	reverse primer
H37Rv_0024_18 bp	GAGAAACAGGAGGGCGTTG	TATTACGACGACCGCTATGC
H37Rv_0079_9 bp	CGTGCACAGTTGGGTGTTTA	TTCGTTCAGGAACTCCAAGG
H37Rv_0154_53 bp	TGGACTTGCAGCAATGGACCAACT	TACTCGGACGCCGGCTCAAAAT
H37Rv_0424_51 bp	GTCCAGGTTGCAAGAGATGG	GGCATCCTCAACAACGGTAG
H37Rv_0531_15 bp	GGTTACCACTTCGATGCGTCTGCG	AGCCGCCGAAACCCATC
*H37Rv_0577_58 bp**	GACTTCAATGCGTTGTTGGA*	GTCTTGACCTCCACGAGTGC*
H37Rv_0580_77 bp	CAGGTCACAACGAGAGGAAGAGC	GCGGATCGGCCAGCGACTCCTC
*H37Rv_0802_54 bp**	AAGCGCAAGAGCACCAAG*	GTGGGCTTGTACTTGCGAAT*
H37Rv_0959_53 bp	GTTCTTGACCAACTGCAGTCGTCC	GCCACCTTGGTGATCAGCTACCT
H37Rv_1121_15 bp	CTCCCACACCCAGGACAC	CGGCCTACCCAACATTCC
H37Rv_1443_56 bp	GGTAATCCTGGTCGCTTGTC	ACCCAAATTGCCCTGGTC
H37Rv_1451_57 bp	GGTAGCCGTCGTCGAGAAGC	CGCCACCACCGCACTGGC
H37Rv_1612_21 bp	GCTGCACCGGTGCCCATC	CACCGGAGCCGGAACGGC
H37Rv_1644_53 bp	TCGGTGATCGGGTCCAGTCCAAGTA	CCCGTCGTGCAGCCCTGGTAC
H37Rv_1895_57 bp	GGTGCACGGCCTCGGCTCC	AAGCCCCGCCGCCAATCAA
H37Rv_1955_57 bp	AGATCCCAGTTGTCGTCGTC	CAACATCGCCTGGTTCTGTA
H37Rv_1982_78 bp*	ATCGTCAGCTGCGGAATAGT*	AATACCGGGGATATCGGTTC*
H37Rv_2059_77 bp	TCGGAGAGATGCCCTTCGAGTTAG	GGAGACCGCGACCAGGTACTTGTA
H37Rv_2074_56 bp	AAATTCAAAGAGTTTCTCGACAGTG	GATCTTGAGAACCAAGATGTCCTT
H37Rv_2163_a_69 bp	CCCATCCCGCTTAGCACATTCGTA	TTCAGGGGGGATCCGGGA
H37Rv_2163_b_69 bp	CGTAAGGGGGATGCGGGAAATAGG	CGAAGTGAATGGTGGCAT
*H37Rv_2165_75pb**	ATTTCGATCGGGATGTTGAT*	TCGGTCCCATCACCTTCTTA*
H37Rv_2347_57 bp	AACCCATGTCAGCCAGGTTA	ATGATGGCACACCGAAGAAC
H37Rv_2401_58 bp	AGTCACCTTTCCTACCACTCGTAAC	ATTAGTAGGGCACTAGCACCTCAAG
H37Rv_2461_57 bp	GCGAACACCAGGACAGCATCATG	GGCATGCCGGTGATCGAGTGG
H37Rv_2531_53 bp	CAGCGAAACGAACTGTGCTATCAC	CGTGTCCGAGCAGAAAAGGGTAT
H37RV_2387_54 bp	CGACCAAGATGTGCAGGAATACAT	GGGCGAGTTGAGCTCACAGAA
H37Rv_2990_55 bp	GTGACGTTTACCGTGCTCTATTTC	GTCGTCGGACAGTTCTAGCTTT
H37Rv_2996_51 bp	CCCGCCTTCGAAACGTCGCT	TGGACATAGGCGACCAGGCGAATA
H37Rv_3006_53 bp	TCGAAAGCCTCTGCGTGCCAGTAA	GCGATGTGAGCGTGCCACTCAA
H37Rv_3171_54 bp	GCAGATAACCCGCAGGAATA	GGAGAGGATACGTGGATTTGAG
*H37Rv_3192_53 bp**	ACTGATTGGCTTCATACGGCTTTA*	GTGCCGACGTGGTCTTGAT*
H37Rv_3232_56 bp	CAGACCCGGCGTCATCAAC	CCAAGGGCGGCATTGTGTT
H37Rv_3239_79 bp	CTCGGTGATGGTCCGGCCGGTCAC	GGAAGTGCTCGACAACGCCATGCC
H37Rv_3336_59 bp	ATCCCCGCGGTACCCATC	GCCAGCGGTGTCGACTATCC
H37Rv_3663_63 bp	GCCCAAAAAGCATGGGAACGTGCCCCT	GGTTGTCCCCGCAGTATCTC
H37Rv_3690_58 bp	AATCACGGTAACTTGGGTTGTTT	GATGCATGTTCGACCCGTAG
H37Rv_4052_111 bp	AACGCTCAGCTGTCGGAT	GGCCAGGTCCTTCCCGAT
H37Rv_4156_59 bp*	TGGTCGCTACGCATCGTGTCGGCCCGT*	TACCACCCGGGCAGTTTAC*
H37Rv_4348_53 bp	CGCATCGACAAACTGGAGCCAAAC	CGGAAACGTCTACGCCCCACACAT

* : the primers indicated are not the primers used in the princeps publication, but were designed for the present study, usually in order to reduce the size of the PCR product and consequently to improve allele size identification.

Results

Tandem repeats predicted to be of a different size in H37Rv and CDC1551

The size of tandem repeats in the two M. tuberculosis strains sequenced to date, H37Rv and CDC1551, was compared using the tandem repeat database [http://minisatellites.u-psud.fr]. Fifty-one of the tandem repeats identified in CDC1551 have repeat units longer than 9 base-pairs and a predicted overall size which differs from the H37Rv homolog estimate by at least 9 base-pairs. Seventeen have an expected product size above one kilobase. They include the DR locus and members of the family of PGRS sequences [13] and were not investigated further. Eighteen have been analyzed in previous investigations [7–9,11]. Three produced multiband patterns or inconsistent results. The results obtained for the remaining 13 loci together with the description of the 18 previously described loci are summarized in Table 1. In addition, Table 1 includes nine markers which are not polymorphic between H37Rv and CDC1551 but have already been quoted in the literature. Each locus is designated by its position (expressed in kilobases) on the H37Rv genome and by the repeat unit length as defined by the Tandem Repeat Finder software and indicated in the Tandem Repeat Database [http://minisatellites.u-psud.fr]. All thirteen newly evaluated loci are polymorphic as predicted. In two cases (Table 1) the expected product size is not the observed size. The expected size has not been observed in the collection of strains used here, which suggests that the incorrect prediction is due to an artifact along the sequencing process. Eight loci among the thirteen have polymorphism indexes above 0.50 (two are above 0.7). The vast majority of the repeats units are more than 50 bp long (Table 1) which makes them easy to assay by ordinary agarose gel electrophoresis when using the primer pairs indicated in Table 2. In one instance however (H37Rv_3663_63 bp) the PCR size products clearly do not differ by a perfect number of (63 bp) repeat units (Table 1).

Typing of strains and clustering analysis

The forty loci listed in Table 1 were used to genotype a collection of 90 strains from the M. tuberculosis complex, using the primers listed in Table 2. In our hands, some of the markers did not prove to be sufficiently robust for easy and reproducible typing in the conditions used here. On this basis, we have selected a collection of 21 markers (comprising thirteen previously described markers and eight among the new loci evaluated). The 21 markers used are italicised and underlined in Table 1 and 2. After analysis of the images using Bionumerics 3.0, and conversion of allele sizes in copy numbers of motifs in the tandem arrays, clustering analysis was done using the categorical and Ward parameters. The results of the clustering analysis are shown in Figure 1. The genotyping data from strains M. tuberculosis CDC1551 and M. bovis AF2122/97 was deduced (Table 1) from the sequence data and included in the analysis. Six major groups are defined (Figure 1). Group I contains the M. bovis strains and 5 of the M. africanum strains. Group II is composed of nine M. africanum strains. The third group includes three M. africanum strains and seven M. tuberculosis strains. Interestingly, five of these strains have been independently identified as representing the Beijing type [14] (the last two have not been tested). The last three groups comprise the vast majority of the M. tuberculosis strains. M. africanum strains which are negative for nitrate reduction (Africanum I type [15]) are among the first two groups, closer to the M. bovis strains as previously observed [16,17]. In contrast, the three M. africanum strains which are positive for nitrate reduction are in the third group, closer to M. tuberculosis strains. In order to facilitate the comparison with earlier investigations [16,17], Figure 1 displays the genotypes for the five ETR markers, extracted from the full data presented in Table 3. Group I in Figure 1 is reminiscent of group A in [17] and group A1 in [18]. Group II in Figure 1 is reminiscent of group B in [17] and group A2 in [18] which are both characterized by the 42432 ETR pattern.

The ETR panel alone discriminates 44 genotypes (instead of 84 with the panel of 21 loci; 86 genotypes when including the CDC1551 and AF2122/97 data, Figure 1) and is not sufficient to clearly separate the *M. africanum* strains from the *M. tuberculosis* strains (analysis not shown) as can be achieved using the 21 loci.

Internet-based identifications

The genotyping data presented in Table 3 can be queried directly via an internet service [http://bacterial-genotyping.igmors.u-psud.fr/bnserver/]. Figure 2 provides a brief description of the current *M. tuberculosis* query page (likely to evolve as updates are made). For each locus, allele sizes can be selected among a list of possibilities (observed sizes). Alternatively, more experienced users will go directly to a "copy-paste" page using the appropriate format. The results of the query indicate a similarity score and include links to the complete data for each strain listed. Help files are available, including a link to updated versions of Figure 1.

Testing the reproducibility of the approach

In order to test the reproducibility of the approach, ten blinded-coded control samples were typed. Figure 3 shows the typing of two markers, H37Rv_0802_54 bp (left, 54 bp unit; H37Rv allele : 1 unit, 199 bp PCR product) and H37Rv_1955_57 bp (right, 57 bp unit; H37Rv allele : 2 units, 206 bp PCR product). The number of units in each allele can be unambiguously deduced by comparison with the H37Rv control lanes and the 100 base-pairs

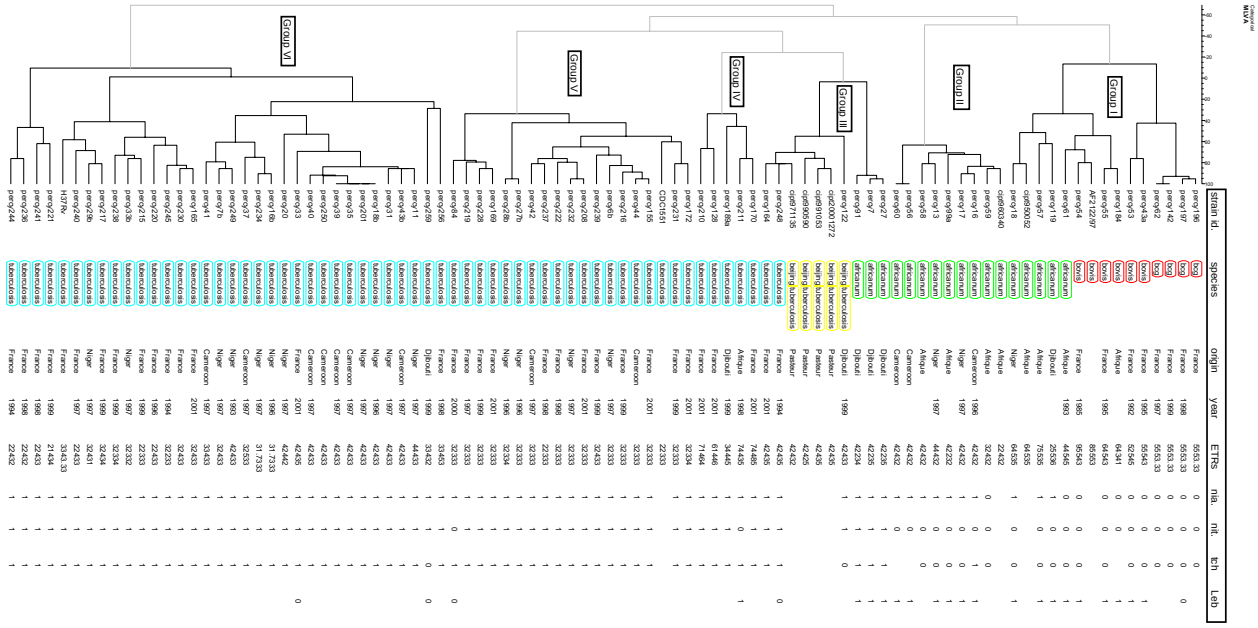

Figure 1
Dendrogram deduced from the clustering analysis of the 92 strains (including CDC1551 and AF2122/97) . The first column from the left identify the strains. The second column indicates the species (Red : *M. bovis* strains; green : *M. africanum* strains; yellow : *M. tuberculosis* strains known to be of the Beijing type and indicated "beijing tuberculosis"; blue : other *M. tuberculosis* strains). The third column indicates the geographic origin of the strain. The fourth column indicates the ETR pattern (ETR-A to ETR-E) extracted from the full data presented in Table 3. The last four columns indicate, from left to right, the result of the niacine production, nitrate reductase, TCH susceptibility and Lebek tests (0, negative ; 1, positive) when available.

ladder size marker. All ten unknown strains were correctly identified using the internet base service described above.

Discussion

The list of 40 markers given in Table 1 is close to representing the complete collection of tandem repeats of interest for MLVA typing in *M. tuberculosis*. It includes all loci with a different predicted size in H37Rv and CDC1551 and which are amenable to routine PCR typing. Nine additional loci which have been quoted in published reports are also included even if they do not fulfill this criteria. Clustering analysis (Figure 1) shows that the two strains CDC1551 and H37Rv (Figure 1) are relatively distant within the *M. tuberculosis* species. This would predict that tandem repeats of identical size in the two strains are likely to be poorly informative across the complex. However, this appears not to be absolutely true, since for instance, ETR-E (H37Rv_3192_53 bp) happens to have the same size in H37Rv, CDC1551 and even AF2122/97 (Table 1) in spite of its very high polymorphism index (0.69, Table 1). Consequently, the few additional loci, not explored here, which are of equal size in H37Rv and CDC1551, but differ with the predicted size for *M. bovis* strain AF2122/97 might also prove to be of interest.

As can be seen in Table 1, most repeat units are more than 50 bp long and allele sizes rarely exceed 1000 bp. As a result, the precision which can be achieved by ordinary agarose gel electrophoresis is sufficient to estimate the number of units in an allele. The selection of 21 markers proposed here was tested specifically in order to be easily assayed using this low-cost technological approach. Although a database system is necessary to efficiently manage a genotyping project with a high number of markers and strains, the identification of up to a few strains per day in a clinical setting for instance requires no sophisticated equipment nor costly consumables. Genotypes can be scored by visual analysis of the gel images, and a subset of the collection of available markers can be chosen for routine identification purposes. The data can then be analysed using the site described in Figure 2.

The role of tandem repeats in the *M. tuberculosis* genome is largely unknown. Twenty-one of the loci listed in Table 1 have repeat units which are a multiple of three basepairs. The majority (fifteen) falls within putative genes, often of unknown function, such as the PPE family of genes [19]. The most remarkable instance is probably PPE34 at position 2163–2165 of the genome (Rv1917c in [http://

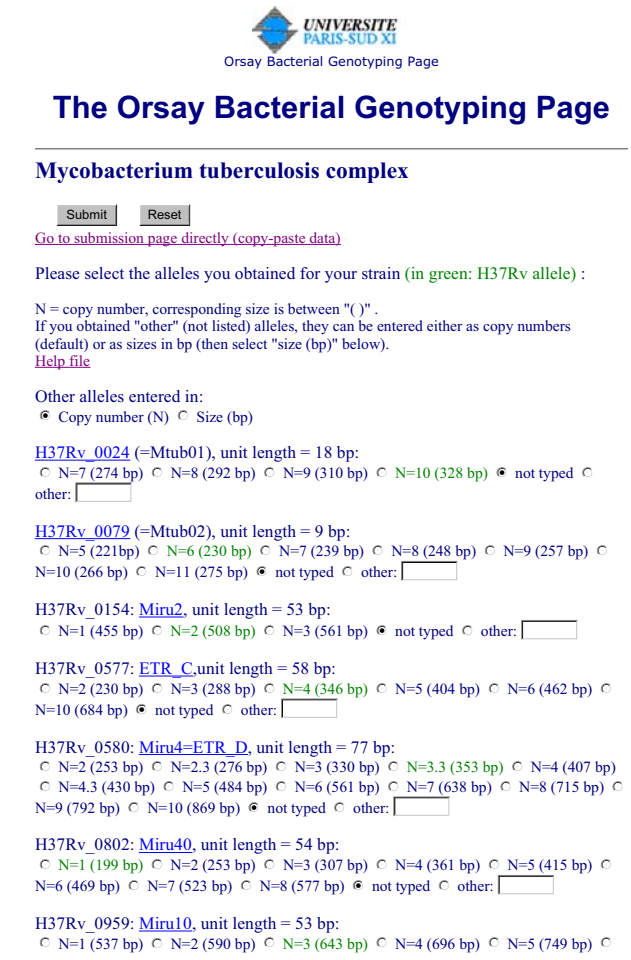

Figure 2
Internet database interrogation page . The query page can be accessed via [http://bacterial-genotyping.igmors.u-psud.fr/bnserver]. The home page (not shown) includes a link to help files (and data updates information), and links to individual species query pages. Currently, identification pages are available for *Y. pestis*, *B. anthracis* (based on the data published in [1] and some additional unpublished data) and *M. tuberculosis*. Figure 2 shows the current *M. tuberculosis* query page. For each marker, allele sizes can be selected among the list of observed sizes. Allele sizes are indicated either as number of motifs, or as fragment sizes, assuming that the primers used are the primers listed in Table 2. The allele size listed in green corresponds to the H37RV control strain allele. More experienced users can go directly to a page on which data (expressed in base-pairs or in repeat unit number) can be directly pasted using the appropriate format.

genolist.pasteur.fr/TubercuList/]) which contains three minisatellites [20] (Table 1, Qub11a, Qub11b, ETR-A).

The present study includes 17 *M. africanum* strains. All strains have been identified as such independently, based on morphological features of the colonies grown on Lowenstein-Jensen medium, and biochemical analyses. *M. africanum* has long since been recognized as showing an extensive phenotypic heterogeneity [21], suggesting that *M. africanum* could display a phenotypic continuum between *M. tuberculosis* and *M. bovis*. This was recently supported by the study of deletion events distinguishing the H37Rv *M. tuberculosis* strain and the BCG *M. bovis* strain [22] and suggesting that *M. bovis* is the most recent member of the *M. tuberculosis* complex. The analysis of deletion events in the *M. africanum* strains investigated showed that West African strains fall into two groups, clearly distinguished from the *M. tuberculosis* strains. In contrast, no deletion event distinguished East African *M. africanum* strains from *M. tuberculosis* strains. The present study includes three Africanum type II strains (positive nitrate reductase test). All three originate from East Africa (Djibouti). Although the MLVA analysis presented here does confirm that they are very close to *M. tuberculosis* strains, they are clearly distinct, at least within the collection of strains evaluated. Interestingly, they appear to be closest to the Beijing type of *M. tuberculosis* strains (Figure 1, Group III, strains percy7, percy27 and percy91).

Conclusions

In its present form, the database should be considered as preliminary. More strains must be typed in order to provide a continuous and robust coverage of the *M. tuberculosis* complex, and the clustering analysis presented in Figure 1 should be considered as provisional. If the MLVA approach is considered to be of use by the community, and given that the associated data is highly portable, then it should be relatively easy, through collaborative efforts, to significantly expand the available data. It is hoped that this data will constitute an easy-to-use high-resolution classification resource which will then help address medical and epidemiological issues regarding the *M. tuberculosis* complex.

Methods

Strains and DNA preparation

Identification of mycobacteria used conventional morphological and biochemical tests as previously described [23]. In particular, *M. tuberculosis*, *M. africanum* and *M. bovis* were distinguished according to their morphology on Lowenstein-Jensen plates. *M. tuberculosis* strains are eugonic. The dysgonic *M. africanum* strains colonies are rough and flat. The dysgonic *M. bovis* colonies are smooth, hemispheric and white. Biochemical analyses included niacin production, nitrate reduction, TCH (thiophene-2-carboxylic acid hydrazide) sensitivity tests and growth characteristics on Lebek medium. DNA for PCR analysis was prepared using a simple thermolysis procedure. Briefly, a few colonies were resuspended in 1 ml water, and in-

Table 3: Genotype data for 21 loci and 92 strains (including CDC1551 and AF2122/97)

strain id.	species	24	79	154	577	580	802	959	1121	1644	1955	2163	2165	2347	2401	2461	2531	3006	3192	3663	3690	4348
percy196	bcg	9	11	2	5	3.3	2	2	4	3	1	9	5	2	2	5	4	3	3	1.7	2	2
percy197	bcg	9	11	2	5	3.3	2	2	4	3	1	6	5	2	2	5	4	3	3	1.7	2	2
percy142	bcg	9	11	2	5	3.3	2	2	4	3	1	11	5	2	2	5	4	3	3	1.7	2	2
percy62	bcg	9	11	2	5	3.3	2	2	4	3	1	11	5	2	2	5	4	3	3	1.7	2	2
percy43a	bovis	9	10	2	5	4	2	2	4	3	1	10	5	3	4	5	4	1	3	1.7	2	2
percy53	bovis	9	9	2	5	4	2	2	4	3	1	11	5	3	4	2	4	2	5	1.7	2	2
percy184	bovis	9	8	2	3	4	2	2	4	2	3	6	6	3	4	4	4	3	1	1.7	3	2
percy55	bovis	9	8	2	5	4	2	2	4	3	3	6	6	3	4	4	4	3	3	1.7	3	2
AF2122	bovis	9	7	2	5	5	2	2	4	2	3	10	8	3	4	5	4	3	3	1.7		2
percy54	bovis	9	8	2	5	4	1	2	4	2	3	10	9	3	4	5	4	3	3	1.7	2	2
percy61	africanum	9	7	2	5	4	2	5	4	2	3	10	4	3	4	4	4	3	5	1.7	4	2
percy119	africanum	9	8	2	5	3	1	7	4	4	3	10	2	3	4	5	4	3	6	1.7	1	2
percy57	africanum	9	6	2	5	3	1	10	4	2	4	10	7	3	4	5	4	3	5	1.1	5	2
cipt950052	africanum	9	7	2	5	3	2	4	4	4	4	10	6	2	4	4	4	3	5	1.7	3	2
percy18	africanum	9	8	2	5	3	2	5	4	4	4	10	6	3	4	4	4	3	5	1.7	6	2
cipt960340	africanum	9	5	2	4	3	1	4	4	4	2	9	2	3	4	2	4	4	2	1.7	4	2
percy59	africanum	9	6	2	4	3	1	4	4	4	2	9	3	3	4	2	4	4	2	1.7	3	2
percy16	africanum	9	5	2	4	3	1	4	4	3	2	9	4	3	4	2	4	4	2	1.7	3	2
percy17	africanum	9	5	2	4	3	1	4	4	4	2	8	4	3	4	2	4	4	2	1.7	3	2
percy99a	africanum	9	6	2	2	3	1	4	4	4	2	9	4	3	2	2	4	4	2	1.7	1	2
percy13	africanum	9	5	2	4	3	1	4	4	3	2	9	4	3	4	4	4	3	2	1.7	2	2
percy58	africanum	9	5	2	4	3	1	4	4	4	2	9	4	3	2	2	4	3	2	1.7	4	2
percy56	africanum	9	5	2	4	3	1	4	3	4	2	9	4	3	4	2	2	4	2	1.7	3	2
percy60	africanum	9	5	2	4	3	1	4	3	4	2	9	4	3	4	2	2	4	2	1.7	3	2
percy27	africanum	9	9	2	2	3	3	4	3	1	4	10	4	4	2	2	5	3	5	1.7	3	3
percy7	africanum	9	9	2	2	3	3	4	3	3	4	10	4	4	2	2	5	3	5	1.7	3	3
percy91	africanum	9	9	2	2	3	3	4	3	3	4	10	4	4	2	2	5	3	4	1.7	3	3
percy122	beijing tuberculosis	9	6	2	4	3	1	3	4	3	6	11	4	4	4	2	5	1	3	0.7	3	2
cipt20001272	beijing tuberculosis	9	11	2	4	3	3	3	4	4	5	6	4	4	4	2	5	3	5	0.7	3	2
cipt991053	beijing tuberculosis	9	11	2	4	3	3	2	4	3	5	9	4	4	4	2	5	3	5	0.7	3	2
cipt990590	beijing tuberculosis	9	11	2	4	2	3	2	4	3	5	9	4	4	4	2	3	3	5	0.7	3	3
cipt971135	beijing tuberculosis	9	11	2	4	3	3	3	4	3	4	9	4	4	4	2	5	3	2	0.7	3	3
percy248	tuberculosis	9	11	2	4	3	3	3	3	3	1	14	4	4	4	2	5	3	5	0.7	3	3
percy164	tuberculosis	9	6	2	4	3	3	3	4	3	5	10	4	4	4	2	5	3	5	0.7	3	3
percy170	tuberculosis	9	8	2	4	8	2	6	3	2	5	10	7	3	2	4	5	3	5	1.7	4	3
percy211	tuberculosis	9	8	2	4	3	3	3	4	2	5	11	7	3	2	4	5	3	5	1.7	4	3
percy189a	tuberculosis	9	9	2	4	4	3	4	4	3	6	5	3	3	2	4	4	3	5	1.7	7	3
percy128	tuberculosis	9	8	2	4	4	4	3	4	3	6	6	6	3	1	1	8	3	6	1.7	4	3
percy210	tuberculosis	9	8	2	4	6	4	4	4	3	7	24	7	3	1	1	6	3	4	1.7	4	3
percy172	tuberculosis	9	10	2	3	3	3	4	5	3	3	10	3	4	4	2	5	3	4	1.7	3	2
percy231	tuberculosis	9	10	2	3	3	2	4	4	3	3	11	3	4	4	2	5	3	3	1.7	3	2
CDC1551	tuberculosis	9	7	2	3	3	5	5	5	3	3	7	2	3	4	2	5	3	3	1.7		2
percy155	tuberculosis	9	8	2	3	3	3	5	4	3	2	10	3	4	4	2	5	3	3	1.7	3	2
percy44	tuberculosis	9	8	2	3	3	3	5	4	3	2	8	3	4	4	2	5	3	3	1.7	3	2
percy216	tuberculosis	8	8	2	3	3	3	5	4	3	2	7	3	4	4	2	5	3	3	1.7	3	2
percy6b	tuberculosis	9	8	2	3	3	4	5	4	3	4	7	3	4	4	2	5	3	3	1.7	3	2
percy239	tuberculosis	9	8	2	4	3	2	5	5	3	2	24	3	4	4	2	5	3	3	1.7	3	2

Table 3: Genotype data for 21 loci and 92 strains (including CDC1551 and AF2122/97) *(Continued)*

percy208	tuberculosis	9	8	2	3	3	3	5	4	3	3	11	3	4	4	2	5	3	3	1.7	5	2
percy232	tuberculosis	9	8	2	3	3	3	5	4	3	3	11	3	4	4	2	5	3	3	1.7	3	2
percy222	tuberculosis	9	8	2	3	3	3	5	4	2	3	10	3	4	4	2	5	4	3	1.7	3	2
percy237	tuberculosis	9	8	2	3	3	3	5	4	3	3	6	2	2	4	2	5	3	3	1.7	3	2
percy42	tuberculosis	9	8	2	3	3	2	5	4	3	3	4	3	3	4	2	5	3	3	1.7	3	2
percy27b	tuberculosis	9	8	2	3	3	6	3	4	3	4	7	3	4	4	2	5	3	3	1.7	3	3
percy28b	tuberculosis	9	8	2	3	3	6	3	4	3	4	7	3	4	4	2	5	3	4	1.7	3	3
percy169	tuberculosis	9	8	2	3	3	3	5	4	3	2	5	3	2	2	2	3	3	3	1.7	3	2
percy228	tuberculosis	9	8	2	3	3	3	5	4	3	3	24	3	2	4	2	3	3	3	1.7	3	2
percy219	tuberculosis	9	9	2	3	3	3	5	4	3	2	9	3	2	4	2	3	3	3	1.7	6	2
percy84	tuberculosis	9	8	2	3	3	3	4	4	2	2	11	3	2	4	2	3	3	3	1.7	3	2
percy256	tuberculosis	9	6	1	4	5	3	7	4	4	4	12	3	4	2	3	4	3	3	1.7	1	2
percy259	tuberculosis	10	6	2	4	3	2	1	5	3	1	26	3	2	2	3	5	1	2	1.7	3	1
percy11	tuberculosis	10	6	2	4	3	3	3	4	3	3	6	4	4	2	4	5	3	3	1.7	2	2
percy43b	tuberculosis	10	6	2	4	3	1	3	4	3	3	6	4	4	2	2	5	3	3	1.7	20	2
percy31	tuberculosis	10	6	2	4	3	3	3	4	2	3	6	4	4	2	2	5	3	3	1.7	17	2
percy18b	tuberculosis	10	6	2	4	3	3	3	4	3	3	6	4	4	2	2	5	3	3	1.7	3	2
percy201	tuberculosis	10	6	2	4	3	3	3	4	3	3	6	4	4	2	2	5	3	3	1.7	3	2
percy35	tuberculosis	10	6	2	4	3	3	3	4	3	3	6	4	4	2	2	5	3	3	1.7	3	2
percy39	tuberculosis	10	6	2	4	3	3	3	4	3	3	6	4	4	2	2	5	3	3	1.7	3	2
percy250	tuberculosis	10	6	2	4	3	3	3	4	3	3	6	4	2	2	2	5	3	3	1.7	3	2
percy40	tuberculosis	10	6	2	4	3	3	3	4	3	3	6	4	4	2	2	5	3	3	1.7		2
percy33	tuberculosis	10	6	2	4	3	3	3	4	3	2	25	4	4	2	2	5	3	5	1.7	4	2
percy20	tuberculosis	10	6	3	4	4	2	3	4	3	1	4	4	4	2	2	5	3	2	1.7	3	2
percy16b	tuberculosis	10	6	2	3	3	3	3	4	3	3	24	3	4	2	1.7	5	3	3	1.7	3	2
percy234	tuberculosis	10	6	2	3	3	3	3	4	3	3	25	3	4	2	1.7	5	3	3	1.7	2	2
percy37	tuberculosis	10	6	2	5	3	3	3	4	4	3	24	3	4	2	2	5	3	3	1.7	4	2
percy249	tuberculosis	10	6		4	3	3	3	4	2	3	25	4	4	2	2	5	3	3	1.7	2	2
percy7b	tuberculosis	10	6	2	4	3	3	3	4	2	3	25	3	4	2	2	5	3	3	1.7	9	2
percy41	tuberculosis	10	6	2	4	3	4	3	4	2	3	25	3	4	2	3	5	3	3	1.7	10	2
percy165	tuberculosis	10	6	2	4	3	3	3	4	2	2	10	3	4	2	2	5	3	3	1.7	3	2
percy230	tuberculosis	10	6	2	4	3	4	2	4	2	2	6	3	4	2	2	5	3	3	1.7	3	2
percy245	tuberculosis	10	6	2	2	3	2	2	4	2	2	11	3	4	2	2	5	3	3	1.7	3	2
percy220	tuberculosis	8	5	2	4	3	2	2	4	2	2	24	2	4	2	2	5	3	3	2.7	3	2
percy215	tuberculosis	10	6	2	3	3	2	3	4	2	2	5	2	4	2	2	5	3	3	2.7	3	2
percy33b	tuberculosis	10	6	2	3	3	2	3	4	2	3	6	3	4	2	2	5	3	2	1.7	3	2
percy238	tuberculosis	10	6	2	3	3	2	3	4	1	2	6	3	4	2	2	1	3	4	3.1	3	2
percy217	tuberculosis	10	6	2	4	3	4	3	4	3	2	24	3	4	2	2	6	3	4	2.7	3	2
percy29b	tuberculosis	10	6	2	4	3	4	3	4	3	1	25	3	4	2	2	5	3	1	2.7	3	2
percy240	tuberculosis	10	6	2	4	3	4	3	4	1	2	4	2	4	2	2	5	3	3	2.7	5	2
H37Rv	tuberculosis	10	6	2	4	3.3	1	3	4	2	2	3	3	4	2	3	6	3	3	2.7	5	2
percy221	tuberculosis	10	5	2	4	3	1	4	4	2	3	8	2	4	1	1	6	1	4	1.1	2	2
percy241	tuberculosis	10	5	2	4	3	3	4	4	2	3	6	2	4	2	2	6	3	3	1.7	1	2
percy236	tuberculosis	7	5	1	4	3	6	3	4	3	3	23	2	4	1	2	6	3	2	1.7	1	2
percy244	tuberculosis	10	5	1	4	3	4	2	4	3	3	8	2	4	1	2	6	3	2	1.7	3	2

Allele sizes were converted to number of repeats according to the correspondence indicated in Table 1. In some instances, decimal values are used, reflecting the existence of alleles with intermediate size. The markers are named and listed according to their position on the genome (Table 1). The strains are listed according to their position in the clustering analysis (Figure 1). *M. tuberculosis* CDC1551 and *M. bovis* AF2122/97 are included based on the predicted allele sizes (Table 1) with the exception of locus H37Rv_3690 (disagreement between observed and expected size for H37Rv at this locus).

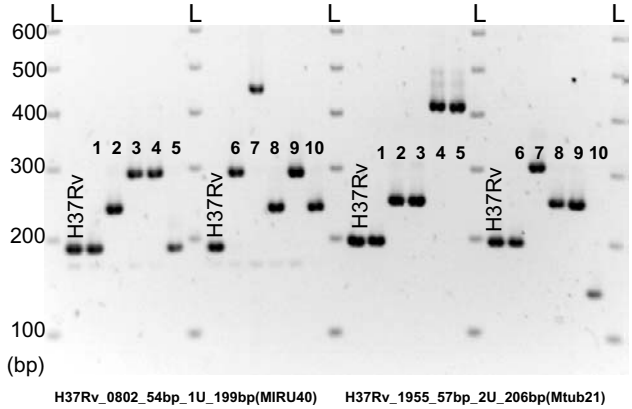

Figure 3
Set-up of the genotyping on agarose gels . The figure illustrates the usual setup for the running of pcr products on agarose gels. Twelve DNA samples (including two "H37Rv" control lanes) are typed at two loci. A 100 bp ladder size marker lane (L) flanks both sides of each group of 6 PCR products. The experiment shown is part of a reproducibility test. The ten blinded-coded samples are numbered from one to ten (percy59, percy55, percy40, percy189a, percy122, percy33, percy28b, percy33b, percy31, percy53). The number of units is easily deduced from the pattern observed, the largest alleles contain six copies of the repeat unit.

cubated at 95°C for 30 minutes. The tube was then centrifuged and the supernatant was recovered.

Identification of tandem repeats

The tandem repeats database described in [1] and accessible at [http://minisatellites.u-psud.fr] was used to identify tandem repeats with a predicted size which differs between the two strains H37Rv [24] and CDC1551 [19]. The database uses the Tandem Repeat Finder software [25] [http://tandem.biomath.mssm.edu/trf.html] to identify tandem repeats in bacterial genomes. Predicted PCR products size in *M. bovis* AF2122/97 was deduced using the *M. bovis* blast server at [http://www.sanger.ac.uk/Projects/M_bovis/blast_server.shtml].

Minisatellite PCR amplification and genotyping

PCR reactions were performed in 15 µl containing approximately 1 ng of DNA (2 µl of the thermolysate), 1× PCR buffer, 1 unit of Taq DNA polymerase, 200 µM of each dNTP, 0.3 µM of each flanking primer. The Taq DNA polymerase was obtained from Qbiogen and used as recommended by the manufacturer.

PCR reactions were run on a MJResearch PTC200 thermocycler. An initial denaturation at 94°C for five minutes was followed by 40 cycles of denaturation at 94°C for 1 minute, annealing at 62°C for one minute (except for H37Rv_0079 and H37Rv_2387 : annealing temperature 55°C), elongation at 72°C for 90 seconds, followed by a final extension step of 10 minutes at 72°C. Five microliters of the PCR products were run on standard 2% agarose gel (Qbiogen) in 0.5 × TBE buffer at a voltage of 10 V/cm (10× TBE is 890 mM Tris base, 890 mM boric acid, 20 mM EDTA, pH 8.3). Samples were manipulated and dispensed (including gel loading) with multi-channel electronic pipettes (Biohit) in order to reduce the risk of errors. Gel length of 20 cm were used. Gels were stained with ethidium bromide, visualized under UV light, and photographed.

Allele sizes were estimated using a 100 bp ladder (MBI Fermentas or Biorad) as size marker. Each 50 wells gel contained 8 regularly spaced size-marker lanes. In addition, strain H37Rv was included as a control for size assignments (one H37Rv control for each set of five DNA samples; see Figure 3). Gel images and resulting data were managed using the Bionumerics software package (version 3.0, Applied-Maths, Belgium).

Data analysis and on-line access

Band size estimates were exported from Bionumerics and converted to number of units. The resulting data was imported in Bionumerics as an opened character data set. Clustering analysis of genotyping data was performed using the Bionumerics package (categorical and Ward). The use of the categorical coefficient implies that the character states are considered as unordered. The same weight is given to a large vs. a small number of differences in the number of repeats at a locus. Among the many possibilities available for clustering analysis, the categorical and Ward combination were empirically selected for their ability to cluster the strains in almost perfect agreement with the microbiological analysis (Figure 1).

The web-page site running identifications was developed using the BNserver application (version 3.0, Applied-Maths, Belgium).

Authors' contributions

PLF has compiled and evaluated previously described markers, evaluated new markers, and genotyped the strains. FD has analyzed the H37Rv, CDC1551 and AF2122/97 sequence data to identify tandem repeats, and is the curator of the tandem repeat database [http://minisatellites.u-psud.fr] in which known data on individual markers is available. FD and GV have designed and set-up the internet strain identification service. GV conceived the study and participated in its design and coordination. MF and JLK have isolated and characterized the strains at the biochemical level, and also prepared PCR-quality DNA.

All authors contributed to the writing of the paper and approved the final manuscript.

Acknowledgements

We thank Drs V. Hervé (HIA Percy) and R. Teyssou (HIA Val de Grâce) for their support to this project. The setting up of a database for the identification of human pathogens is supported by grants from the Délégation Générale de l'Armement (DGA/DSA/SP-Num). The sequence data for *M. bovis* AF2122/97 was produced by the *M. bovis* Sequencing Group at the Sanger Institute and can be obtained from [ftp://ftp.sanger.ac.uk/pub/pathogens/mb]. We thank Dr V. Vincent, Institut Pasteur, Paris, for the provision of two *M. africanum* strains and four *M. tuberculosis* strains of the Beijing type.

References

1. Le Fleche P, Hauck Y, Onteniente L, Prieur A, Denoeud F, Ramisse V, Sylvestre P, Benson G, Ramisse F, Vergnaud G: **A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*.** *BMC Microbiol* 2001, **1:**2
2. Bayliss CD, Field D, Moxon ER: **The simple sequence contingency loci of *Haemophilus influenzae* and *Neisseria meningitidis*.** *J Clin Invest* 2001, **107:**657-666
3. Hermans PW, van Soolingen D, Bik EM, de Haas PE, Dale JW, van Embden JD: **Insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains.** *Infect Immun* 1991, **59:**2695-2705
4. van Embden JD, van Gorkom T, Kremer K, Jansen R, van Der Zeijst BA, Schouls LM: **Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria.** *J Bacteriol* 2000, **182:**2393-2401
5. Sola C, Filliol I, Gutierrez MC, Mokrousov I, Vincent V, Rastogi N: **Spoligotype database of *Mycobacterium tuberculosis*: biogeographic distribution of shared types and epidemiologic and phylogenetic perspectives.** *Emerg Infect Dis* 2001, **7:**390-396
6. Kremer K, van Soolingen D, Frothingham R, Haas WH, Hermans PW, Martin C, Palittapongarnpim P, Plikaytis BB, Riley LW, Yakrus MA, *et al*: **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* 1999, **37:**2607-2618
7. Frothingham R, Meeker-O'Connell WA: **Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats.** *Microbiology* 1998, **144:**1189-1196
8. Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C: **Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome.** *Mol Microbiol* 2000, **36:**762-771
9. Roring S, Scott A, Brittain D, Walker I, Hewinson G, Neill S, Skuce R: **Development of variable-number tandem repeat typing of *Mycobacterium bovis*: comparison of results with those obtained by using existing exact tandem repeats and spoligotyping.** *J Clin Microbiol* 2002, **40:**2126-2133
10. Mazars E, Lesjean S, Banuls AL, Gilbert M, Vincent VV, Gicquel B, Tibayrenc M, Locht C, Supply P: **High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology.** *Proc Natl Acad Sci U S A* 2001, **98:**1901-1906
11. Skuce RA, McCorry TP, McCarroll JF, Roring SM, Scott AN, Brittain D, Hughes SL, Hewinson RG, Neill SD: **Discrimination of *Mycobacterium tuberculosis* complex bacteria using novel VNTR-PCR targets.** *Microbiology* 2002, **148:**519-528
12. Supply P, Lesjean S, Savine E, Kremer K, van Soolingen D, Locht C: **Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units.** *J Clin Microbiol* 2001, **39:**3563-3571
13. Ross BC, Raios K, Jackson K, Dwyer B: **Molecular cloning of a highly repeated DNA element from *Mycobacterium tuberculosis* and its use as an epidemiological tool.** *J Clin Microbiol* 1992, **30:**942-946
14. van Soolingen D, Qian L, de Haas PE, Douglas JT, Traore H, Portaels F, Qing HZ, Enkhsaikan D, Nymadawa P, van Embden JD: **Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of east Asia.** *J Clin Microbiol* 1995, **33:**3234-3238
15. Collins CH, Yates MD, Grange JM: **Subdivision of *Mycobacterium tuberculosis* into five variants for epidemiological purposes: methods and nomenclature.** *J Hyg (Lond)* 1982, **89:**235-242
16. Haas WH, Bretzel G, Amthor B, Schilke K, Krommes G, Rusch-Gerdes S, Sticht-Groh V, Bremer HJ: **Comparison of DNA fingerprint patterns of isolates of *Mycobacterium africanum* from east and west Africa.** *J Clin Microbiol* 1997, **35:**663-666
17. Frothingham R, Strickland PL, Bretzel G, Ramaswamy S, Musser JM, Williams DL: **Phenotypic and genotypic characterization of *Mycobacterium africanum* isolates from West Africa.** *J Clin Microbiol* 1999, **37:**1921-1926
18. Viana-Niero C, Gutierrez C, Sola C, Filliol I, Boulahbal F, Vincent V, Rastogi N: **Genetic diversity of *Mycobacterium africanum* clinical isolates based on IS6110-restriction fragment length polymorphism analysis, spoligotyping, and variable number of tandem DNA repeats.** *J Clin Microbiol* 2001, **39:**57-65
19. Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, DeBoy R, Dodson R, Gwinn M, Haft D, *et al*: **Whole-Genome Comparison of *Mycobacterium tuberculosis* Clinical and Laboratory Strains.** *J Bacteriol* 2002, **184:**5479-5490
20. Sampson SL, Lukey P, Warren RM, van Helden PD, Richardson M, Everett MJ: **Expression, characterization and subcellular localization of the *Mycobacterium tuberculosis* PPE gene Rv1917c.** *Tuberculosis (Edinb)* 2001, **81:**305-317
21. David HL, Jahan MT, Jumin A, Grandry J, Lehmann EH: **Numerical taxonomy of *Mycobacterium africanum*.** *Int J Syst Bacteriol* 1978, **28:**467-472
22. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, *et al*: **A new evolutionary scenario for the *Mycobacterium tuberculosis* complex.** *Proc Natl Acad Sci U S A* 2002, **99:**3684-3689
23. Levy-Frebault VV, Portaels F: **Proposed minimal standards for the genus *Mycobacterium* and for description of new slowly growing *Mycobacterium* species.** *Int J Syst Bacteriol* 1992, **42:**315-323
24. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, *et al*: **Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence.** *Nature* 1998, **393:**537-544
25. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-580