Simple sequence repeats in mycobacterial genomes

VATTIPALLY B SREENU, PANKAJ KUMAR, JAVAREGOWDA NAGARAJU† and HAMPAPATHALU A NAGARAJARAM*

Laboratory of Computational Biology, †Laboratory of Molecular Genetics, Centre for DNA Fingerprinting and Diagnostics, ECIL Road, Nacharam, Hyderabad 500 076, India

**Corresponding author (Fax, 91-40-2715 5610 Email, han@cdfd.org.in)*

Simple sequence repeats (SSRs) or microsatellites are the repetitive nucleotide sequences of motifs of length 1–6 bp. They are scattered throughout the genomes of all the known organisms ranging from viruses to eukaryotes. Microsatellites undergo mutations in the form of insertions and deletions (INDELS) of their repeat units with some bias towards insertions that lead to microsatellite tract expansion. Although prokaryotic genomes derive some plasticity due to microsatellite mutations they have in-built mechanisms to arrest undue expansions of microsatellites and one such mechanism is constituted by post-replicative DNA repair enzymes MutL, MutH and MutS. The mycobacterial genomes lack these enzymes and as a null hypothesis one could expect these genomes to harbour many long tracts. It is therefore interesting to analyse the mycobacterial genomes for distribution and abundance of microsatellites tracts and to look for potentially polymorphic microsatellites. Available mycobacterial genomes, *Mycobacterium avium*, *M. leprae, M. bovis* and the two strains of *M. tuberculosis* (CDC1551 and H37Rv) were analysed for frequencies and abundance of SSRs. Our analysis revealed that the SSRs are distributed throughout the mycobacterial genomes at an average of 220–230 SSR tracts per kb. All the mycobacterial genomes contain few regions that are conspicuously denser or poorer in microsatellites compared to their expected genome averages. The genomes distinctly show scarcity of long microsatellites despite the absence of a post-replicative DNA repair system. Such severe scarcity of long microsatellites could arise as a result of strong selection pressures operating against long and unstable sequences although influence of GC-content and role of point mutations in arresting microsatellite expansions can not be ruled out. Nonetheless, the long tracts occasionally found in coding as well as non-coding regions may account for limited genome plasticity in these genomes.

[Sreenu V B, Kumar P, Nagaraju J and Nagarajaram H A 2007 Simple sequence repeats in mycobacterial genomes; *J. Biosci*. **32** 3–15

1. Introduction

Simple sequence repeats (SSRs) or microsatellites are the repetitive sequence motifs of 1–6 bp (Schlotterer 2000). They are scattered throughout the genomes of all the known organisms ranging from viruses to eukaryotes (Heller *et al* 1982; Ellegren 2004). The origin, evolution and ubiquitous occurrence of these repeats still pose a riddle to researchers. One of the unique properties of the SSRs is their high degree of polymorphism by virtue of variability in their repeat number at most loci. The mutations in the form of insertions and deletions (INDELS) of their repeat units are typically in the range from 10^{-6} to 10^{-2} per generation which is higher than base substitution rates (Schlotterer 2000).

Microsatellites are more than mere repetitive sequences. Their roles have been attributed to many biological functions. For instance, the genes responsible for virulence of many pathogenic bacteria have been shown to contain these repetitive elements (Moxon *et al* 1994). Regions with high occurrence of microsatellites are called as the

Keywords. Comparative genomics; genome analysis; microsatellite; mycocacteria; polymorphalism; sequence repeats

Abbreviations used: INDELS, Insertions and deletions; ML, *M.leprae*; MA, *M. avium*; MB. *M. bovis*; MTC, *M. tuberculosis*; PPM, potential polymorphic microsatellite; SSRs, simple sequence repeates.

Supplementary Data pertaining to this article is available on the *Journal of Biosciences* Website at *http://www.ias.ac.in/jbiosci/jan2007/pp3-15-suppl.pdf*

Table 1. Mycobacterial genomes that are considered for microsatellite analysis

Genome	Genome size (bp)	GC content (%)	Coding density (%)	Reference
M. avium paratuberculosis (MA)	4829781	69	91	*Li et al 2005*
M. bovis (MB)	4345492	66	90	Garnier *et al* 2003
M. leprae (ML)	3268203	58	49	Cole *et al* 2001
M. tuberculosis CDC1551 (MTC)	4403836	66	90	Fleischmann *et al* 2002
M. tuberculosis H37Rv (MTH)	4411529	66	90	Cole *et al* 1998

contingency loci (Moxon *et al* 1994). Variations in repeat numbers in microsatellites in the coding regions bring about drastic changes to their gene products, as a consequence of premature termination due to frameshift (Moxon *et al* 1994; van Belkum *et al* 1998; Sreenu *et al* 2003, 2006). Such changes in the coding regions have been shown to cause phase variations in pathogenic bacteria, which impart greater defensive capability to the pathogens to escape hostile host environment (Murphy *et al* 1989; Hood *et al* 1996; van Belkum *et al* 1998). Microsatellites also act as gene regulators where loss or gain of repeats in the promoter region can regulate transcriptional activity (van Ham *et al* 1993). In this way microsatellites inculcate a considerable genome plasticity to adjust to different physical and physiological host environments. For example, in *Escherichia coli* high mutation rates have been observed in microsatellites when they are cultured under stress conditions (Jackson *et al* 1998). Furthermore, close association of many repeats in *E. coli* ORFs related to physiological adaptations, DNA repair and recombination, is indicative of the probable function of these repeats to overcome stressful conditions (Rocha *et al* 2002). Particularly, stress response genes are reported to contain a large number of repeat tracts (Rocha *et al* 2002).

The basic mechanism behind expansion or contraction of microsatellites which happens due to INDELS of their repeat units, is thought to be strand slippage during DNA replication (Levinson and Gutman 1987; Schlotterer and Tautz 1992). Such replication error is usually corrected in the cell by the post-replicative mismatch repair system, constituted by the genes *mutS*, *mutL* and *mutH* (Levinson and Gutman 1987). Though these genes are conserved across the genre, organisms like the mycobacterial species are devoid of them (Springer *et al* 2004).

Mycobacterial genomes, the focus of the present study, have been studied for repetitive sequences for many years (Hermans *et al* 1991; van Soolingen *et al* 1993; Kamebeek *et al* 1997; Cole *et al* 1998). However, most of them focused on insertion elements, and other mycobacterial repeats like MIRU that are in use as genetic markers. Although a few published reports on microsatellite loci have shown their involvement directly or indirectly in the pathogenicity in some pathogens (Moxon *et al* 1994; van Belkum *et al* 1998), to the best of our knowledge, there is no published report on their involvement in mycobacterium.

Currently, complete genome sequences for the five mycobacteria namely, *M. avium* (MA) (Li *et al* 2005), *M. leprae* (ML) (Cole *et al* 2001), *M. bovis* (MB) (Garnier *et al* 2003) and two strains of *M. tuberculosis* [CDC1551 (MTC) (Fleischmann *et al* 2002) and H37Rv (MTH) (Cole *et al* 1998)] are available in the public domain (see table I). *M. avium* is a common bacterium in surface water and soils and is the causative agent of the Crohn's disease in humans (Cosma *et al* 2003). *M. leprae* causes leprosy in human. *M. bovis* is the causative agent of tuberculosis in many animals including human. *M. tuberculosis* is the major cause of tuberculosis in humans. All these organisms are GC-rich genomes. Approximate coding density of these genomes is about 90%, except in *M. leprae*, where it is 49% (Cole *et al* 2001). As mentioned earlier these genomes lack the post-replication DNA repair enzymes and therefore it can be surmised, as a null hypothesis, that the mutations in microsatellites occur as unregulated events and that the genomes are enriched with long microsatellites. To test this null hypothesis we analysed the five mycobacterial genomes for frequencies and distributions of microsatellites and the results are reported in this communication.

2. Materials and methods

The microsatellite data pertaining to the five mycobacterial genomes MA, MB, ML, MTC and MTH (table 1) comprising of sequence of the repeat motif, repeat size, repeat number and location with respect to coding and non-coding regions available in MICdb2.0 database (*http://210.212.212.7/MIC/index.html*) (Sreenu *et al* 2003) were used for the present analysis. The observed number of each class of microsatellites (mono, di, etc.) in a genome was compared to the number that could be expected by chance in a randomized genome of the same length and composition. For this purpose each genome sequence was randomized thousand times and from them average number of microsatellite tracts was calculated as the expected number. We used SHUFFLESEQ program of the EMBOSS software suite (*http://hgmp.mrc.ac.uk/*

Software/EMBOSS/) for randomization of the genomes and SSRF (Sreenu *et al* 2003) to identify microsatellites in them. The number of microsatellites observed per 10 kb (referred to as tract-density) was calculated for every non-overlapping window of the size 10 kb throughout the genomes. Statistical significance of the observed number of microsatellites as compared to the expected number was carried out with students *t*-test (programs were written in C by taking the functions from "Numerical recipes in C" (Press *et al* 1992)).

3. Results

3.1 *Microsatellite density profile*

Our analysis revealed that every genome, except ML, harbours as many as one million microsatellites. ML genome harbours 25% less as compared to the other genomes however its net genomic occupancy of the microsatellites expressed as the ratio of number of bases in microsatellites and in the whole genome is 0.65 which is not different from the genomic occupancies found in the other genomes (MA=0.72, MB=0.69, MTC=0.69 and MTH=0.69).

The abundance and distribution of microsatellites throughout the five genomes are represented as tract density profiles in figure I. A tract density profile represents the plot of tract densities (the number of microsatellites/10 kb) calculated for every successive non-overlapping segments of DNA in a given genome. The mean of the observed tract densities in the five genomes vary in the range of 2200–2300 tracts/10 kb. In four (MA, MTC, MTH and MB) out of the five genomes the mean of the observed tract densities (shown as horizontal thick line in figure 1) is higher than the mean of the expected tract densities (shown as horizontal thin line in figure 1) indicating characteristic abundance of microsatellites in these genomes. ML, on the other hand, shows scarcity of tracts as the observed tract density is lower than the expected tract density. All the genomes comprise of some regions conspicuously either rich or poor in microsatellites as revealed by their tract density values, which are three standard deviations higher or lower than the mean of the expected tract densities. The MTC, MTH, MB genomes as well as MA genomes harbour more of microsatellite-rich regions (in the range of 30–40) than microsatellite-poor regions; ML, however, harbours only microsatellite-poor regions which are seven in total. An examination of the microsatellite-rich regions and microsattellite–poor regions in the five genomes revealed that a large majority of ORFs in the microsatellite-poor region of MA encode hypothetical proteins. In MTC, MTH and MB genomes the microsatellite-rich regions encode proteins belonging to PE and PPE family of proteins in addition to some hypothetical and heat shock proteins (data given as supplementary information).

The profiles of MA and ML stand distinctly different from the other profiles. The profiles of MTC, MTH and MB genomes look very similar among themselves indicating similar disposition of these genomes for harbouring microsatellites as well as homology of microsatellite evolution. In our earlier studies (Sreenu *et al* 2006) it was observed that these genomes conserve microsatellite regions. Some of the homologous microsatellites showed differences in their lengths due to INDELS of repeat units, revealing inter-species and intra-species microsatellite polymorphism (Sreenu *et al* 2006). These polymorphic microsatellites act as resource elements for rendering certain amount of plasticity to the genomes.

3.2 *Microsatellite repeat numbers, motif distribution and abundance*

The number of microsatellites of different repeat sizes (mono to hexa) found in each genome, are given in table 2 and the abundance of microsatellites in terms of the sequence motifs is given in table 3.

In all the five genomes, microsatellites with small repeat sizes are more abundant than those with large size repeats. The mono tracts are the most abundant type with numbers varying in the range of 160–170 tracts per kbp. The hexa tracts are the least abundant type with mere one repeat or less per kbp. Although the genomic distribution of microsatellites is grossly uniform, occurrence of the tri and hexa repeats is slightly more biased towards coding regions, whereas mono, di, tetra and penta repeats are slightly more biased towards non-coding regions (refer to column 9 in table 2).

The mononucleotide tract lengths in MTH and MTC genomes seem to be highly restricted compared to the other genomes where the repeat numbers never exceed 9 whereas in the other three genomes the repeat numbers go up to 27 at some loci. Except for this difference all the genomes commonly show an abundance of the short tracts with repeat iteration of two which occur significantly more than expected compared to the tracts with repeat iterations greater than two which are under-represented. The ratio of the observed to the expected (O/E ratio) numbers of mononucleotide tracts decreases very steeply as the number of repeats (more than six times) in a tract increases. It may also be noted that while the enriched short tracts of two repeats occur both in coding as well as non-coding regions, longer tracts occur in the non-coding regions. About three-forth of the mono tracts are G/C–a reflection of GC richness of the genomes, and these are under-represented. Although the A/T tracts are less abundant they occur more than expected which is a probable indication of a trend in mycobacterial genome evolution characterized by accumulation of A/T tracts.

The repeat numbers of dinucleotide microsatellites are restricted to a maximum of five or six in MA, MB, MTC

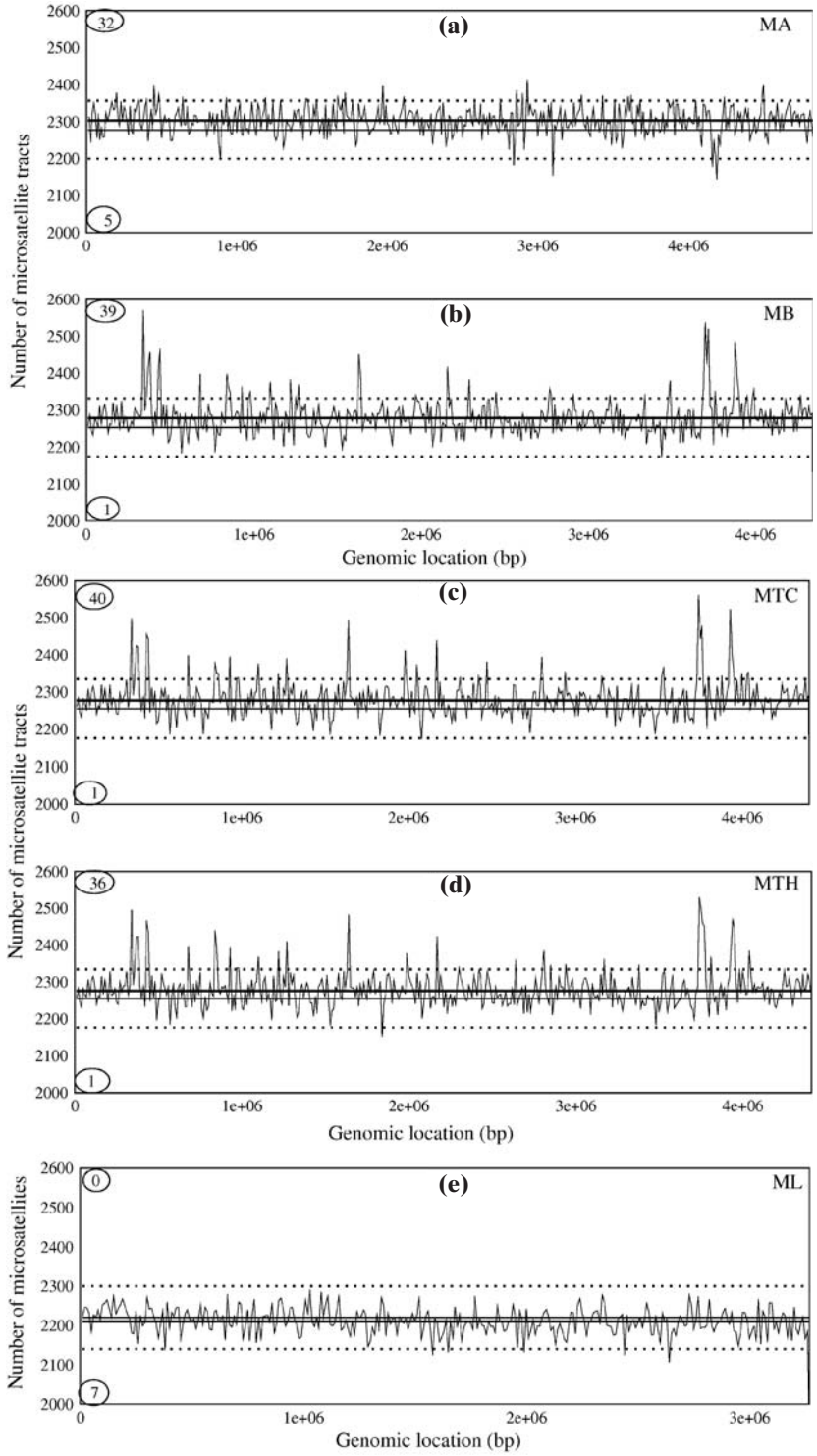

Figure 1. Tract density profiles of microsatellites in the five genomes: **(a)** *M. avium* (MA), **(b)** *M. bovis* (MB), **(c)** *M. tuberculosis* CDC1551 (MTC), **(d)** *M. tuberculosis* H37Rv (MTH) and **(e)** *M. leprae* (ML). Tract density refers to the number of microsatellite tracts found in a DNA segment of length 10 kb and the profile is a plot of tract density values of successive, non-overlapping DNA segments in a genome. The horizontal thick and thin lines respectively represent the means of observed and expected tract densities. The broken lines have been drawn at 3 standard deviation levels above and below the mean of expected tract densities and using these as reference the microsatellite rich and poor regions have been determined. The expected tract densities for a genome were calculated as the averages of the tract densities found in 1000 randomized sequences of that genome.

Table 2. The observed as well as the expected number (given within parentheses) of microsatellites of different motif sizes and repeat numbers found in the five mycobacterial genomes.

	2	3	4	5	6	7	>7	PIC	RPK
				Mycobacterium avium					
Mono	623349 + (560563)	147016 - (183508)	31201 - (61587)	6264 – (20999)	662 - (7227)	37 – (2486)	9 – (1315)	91	167
Di	152127 – (154904)	10281 - (13284)	1043 – (1319)	69 – (144)	17 (17)	2 (2)	1 (1)	91	34
Tri	106819 + (63723)	6592 + (1815)	420 + (62)	15 + (3)	1 (1)	0 (1)	0 (0)	93	24
Tetra	15871 – (19242)	119 – (173)	7 + (2)	1 (1)	0 (0)	0 (0)	0 (0)	90	3
Penta	6102 – (6262)	18 – (19)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	89	1
Hexa	4438 + (1747)	24 + (2)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	94	1
				Mycobacterium bovis					
Mono	579496 + (521214)	130389 – (159875)	27674 – (49658)	5733 – (15722)	813 – (5053)	127 – (1634)	13 – (791)	90	171
Di	129978 – (138646)	7580 – (10661)	599 – (918)	43 – (86)	1 – (8)	2 (1)	1 (1)	89	32
Tri	84224 + (53249)	4066 + (1259)	251 + (35)	28 + (2)	1 (1)	1 (1)	0 (0)	92	20
Tetra	13165 – (15269)	68 – (107)	1 – (2)	1 (1)	0 (0)	0 (0)	0 (0)	90	3
Penta	4347 – (4657)	9 – (10)	2 (1)	0 (0)	0 (0)	0 (0)	0 (0)	88	1
Hexa	2727 + (1250)	18 + (1)	3 (1)	0 (0)	0 (0)	0 (0)	0 (0)	93	1
				Mycobacterium tuberculosis CDC 1551					
Mono	587333 + (528307)	132130 – (161988)	28023 – (50264)	5786 – (15907)	826 – (5113)	138 – (1652)	7 – (798)	90	171
Di	131690 – (140503)	7691 – (10802)	607 – (929)	45 – (87)	13 (9)	2 (2)	1 (1)	90	32
Tri	85166 + (53926)	4117 + (1274)	244 + (35)	37 + (2)	1 (1)	1 (1)	0 (0)	91	20
Tetra	13294 – (15450)	70 – (107)	1 – (2)	1 (1)	1 (1)	0 (0)	0 (0)	90	3
Penta	4424 – (4716)	10 (10)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	89	1
Hexa	2750 + (1265)	19 + (1)	2 (1)	0 (0)	0 (0)	0 (0)	0 (0)	91	1
				Mycobacterium tuberculosis H37rv					
Mono	588352 + (529262)	132525 – (162271)	28137 – (50383)	5774 – (15944)	824 – (5125)	137 – (1655)	7 – (801)	90	171
Di	131866 – (140789)	7710 – (10818)	605 – (929)	45 – (87)	13 (9)	1 (2)	1 (1)	90	32
Tri	85287 + (54023)	4103 + (1275)	243 + (35)	32 + (2)	1 (1)	1 (0)	0 (0)	92	20
Tetra	13312 – (15491)	68 – (108)	1 – (2)	1 (1)	0 (0)	0 (0)	0 (0)	90	3
Penta	4457 – (4725)	9 – (10)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	88	1
Hexa	2748 + (1269)	18 + (1)	2 (0)	0 (0)	0 (0)	0 (0)	0 (0)	94	1
				Mycobacterium tuberculosis leprae					
Mono	437169 + (409791)	93349 - (113880	20933 – (30765)	4848 – (8340)	850 – (2287)	151 – (632)	38 – (246)	50	171
Di	94326 – (103781)	4825 – (6831)	260 – (463)	29 – (32)	4 + (3)	2 + (1)	11 (1)	50	30
Tri	50458 + (35921)	1649 + (631)	39 + (12)	6 + (1)	1 (1)	0 (0)	2 (0)	52	16
Tetra	9421 – (9487)	32 – (44)	1 (1)	1 (1)	0 (0)	0 (0)	0 (0)	48	3
Penta	3030 + (2626)	8 + (3)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	48	1
Hexa	1323 + (662)	4 + (1)	0 (0)	0 (0)	0 (0)	1 (0)	0 (0)	53	0.5

The "+" and "–"signs indicate the statistical significance of the differences between the observed and expected numbers and accordingly the tracts with "+" are overrepresented and those with "–"are under-represented.
PIC, percentage of repeat in coding region. RPK, number of repeat per kb of genome.

Table 3. The observed and expected numbers (shown within parentheses) of sequence motifs observed as microsatellite tracts in the five mycobacterial genomes.

	M. avium	*M. leprae*	*M. bovis*	*M. tb* CDC1551	*M. tb* H37Rv
Mononucleotide motifs					
A	93081 + (92781)	108332 + (107038)	105058 + (101116)	106686 + (102615)	106821 + (102825)
T	91709 -(91919)	109980 + (108367)	105645 + (101242)	107067 + (102695)	107129 + (102814)
G	311170 – (325894)	171119 – (176827)	264439 – (274983)	267766 – (278485)	268608 – (279174)
C	312578 – (327091)	167907 – (173710)	269103 – (276605)	272721 – (280234)	273198 – (280627)
Di nucleotide motifs					
AT	1637 – (4386)	7707 - (9856)	4164 - (6052)	4229 - (6151)	4227 - (6159)
GC	101590 + (86716)	32346 + (31124)	71466 + (64721)	72362 + (65530)	72409 + (65673)
AG	11919 – (19686)	10928 – (17601)	11617 – (19795)	11789 – (20072)	11853 – (20144)
AC	18248 – (19791)	18661 + (17218)	19532 – (19951)	19806 – (20239)	19814 – (20273)
TG	18216 – (19498)	19152 + (17846)	19842 (19829)	20108 (20086)	20141 (20125)
TC	11910 – (19591)	10663 – (17462)	11580 – (19970)	11739 – (20251)	11782 – (20257)
Tri nucleotide motifs					
TAA	60 – (159)	532 – (666)	100 – (273)	101 – (278)	103 – (278)
TAT	64 – (157)	563 – (675)	116 – (273)	116 – (278)	115 – (278)
TAG	541 – (738)	867 – (1227)	561 – (926)	579 – (939)	577 – (943)
TAC	582 – (741)	843 – (1200)	531 – (934)	531 – (949)	547 – (950)
GAA	1599 + (745)	1294 + (1209)	1394 + (925)	1416 + (939)	1417 + (944)
GAT	2452 + (738)	2244 + (1228)	2276 + (929)	2317 + (940)	2324 + (943)
GAG	2113 – (3442)	1083 – (2230)	1449 – (3129)	1472 – (3167)	1471 – (3179)
GAC	9822 + (3460)	3881 + (2179)	7253 + (3154)	7320 + (3197)	7284 + (3208)
GTT	1508 + (730)	2440 + (1246)	2040 + (928)	2067 + (941)	2060 + (943)
GTG	7462 + (3406)	4450 + (2263)	6322 + (3133)	6397 + (3171)	6397 + (3177)
GTC	9701 + (3426)	4079 + (2212)	7406 + (3158)	7504 + (3200)	7484 + (3199)
CAA	1571 + (750)	2325 + (1182)	1863 + (934)	1885 + (948)	1896 + (951)
CAT	2583 + (741)	2108 + (1199)	2263 + (934)	2299 + (948)	2296 + (950)
CAG	7757 + (3456)	4144 + (2179)	5424 + (3153)	5484 + (3194)	5491 + (3204)
CAC	7393 + (3475)	4157 + (2131)	6383 + (3182)	6458 + (3223)	6459 + (3224)
CTT	1615 + (734)	1265 +(1219)	1431 + (934)	1443 + (950)	1449 + (948)
CTG	7786 + (3425)	4005 + (2212)	5510 + (3155)	5616 + (3197)	5588 + (3201)
CTC	2079 – (3444)	1030 – (2159)	1624 – (3186)	1648 – (3222)	1647 – (3223)
CGG	23186 + (15875)	5532 + (4017)	16984 + (10607)	17199 + (10726)	17311 + (10751)
CGC	23973 + (15950)	5313 + (3921)	17639 + (10687)	17712 + (10819)	17751 + (10829)

The motif-wise numbers shown are the sum of the numbers of all sequentially permuted motifs. For example, the observed numbers for "AT" repeat is the sum of the observed numbers of AT as well as TA tracts. The differences which are statistically significant are indicated by "+" (over-representation) and "–" (under-representation).

and MTH. In ML the repeat number at some loci goes up to 18. As also observed in mono the O/E ratio falls as the number of repeats increase in all the genomes. Tracts with repeat numbers less than six occur without any striking bias towards coding or non-coding regions. The long tracts found in ML are confined to the non-coding regions. In all the genomes, GC motifs occur more frequently and the number is over represented. The least frequently found AT/TA motifs are underrepresented. For GC/CG, GA/AG, CA/AC and GT/TG repeat tracts, ML genome shows a distinct over-representation compared to the other genomes where they are under-represented.

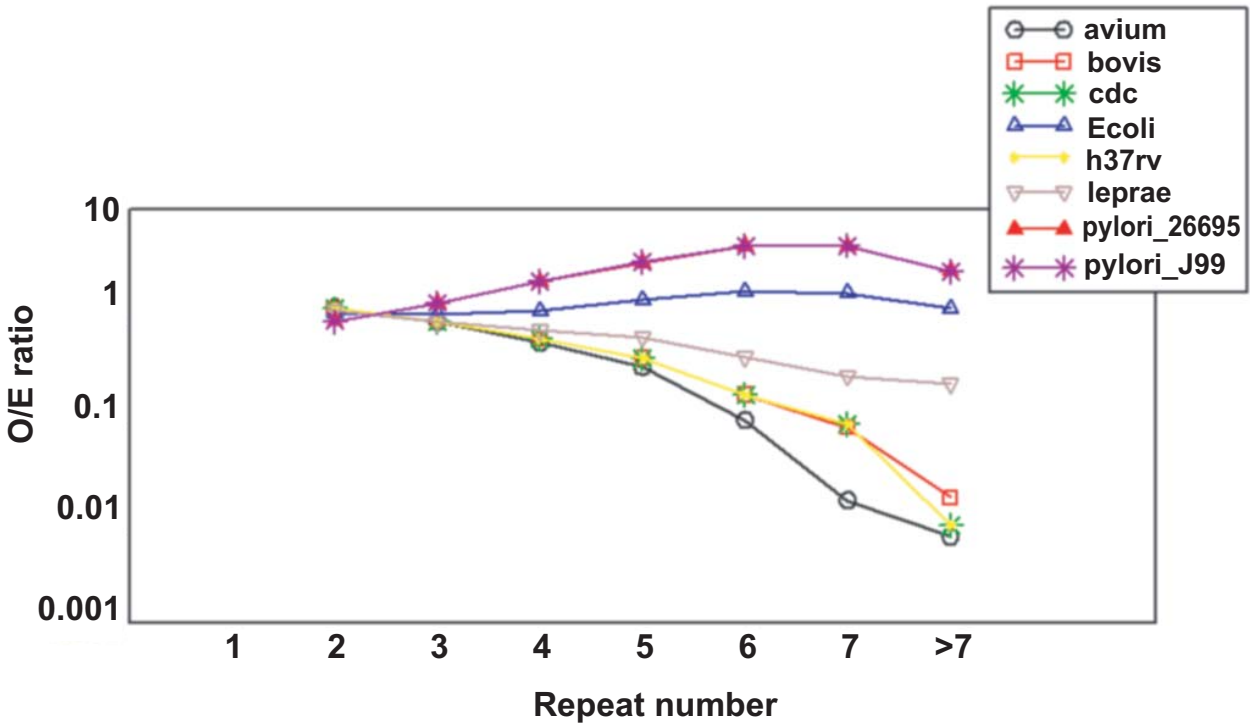

Figure 2. Graph showing ratios of observed and expected (after 1000 times randomizations) numbers of microsatellites from *E. coli* K12, *H. pylori* (J99 and 26695), *M. avium*, *M. leprae*, *M. bovis* and *M. tuberculosis* (CDC1551 and H37Rv).

All the five genomes show uniform over-representation of tri nucleotide repeat tracts, which hardly ever exceed five iterations of repeats per locus. In fact, over-representation increases (i.e. O/E ratio) as the microsatellite repeat number increases, indicating strong selection for accumulation of long trinucleotide tracts. It is interesting to note that trinucleotide microsatellites with repeat number more than 4 are in the coding regions in MA, MB and MTH except the ML where these repeats occur in non-coding regions. Of the 20 possible trinucleotide repeat motifs, 14 of them are over represented in all the genomes.

Of the higher-order microsatellites (with repeat unit size between 4 and 6), only hexanucleotide tracts of all repeat numbers are consistently over represented in all the mycobacterial genomes. It is also interesting to note that ML shows a distinct over-representation of pentanucleotide tracts. In mycobacterial genomes, in general, there is a universal over-representation of GCs-rich motifs compared to AT-rich motifs (data available as supplementary material).

From the table 2 it is also clear that mycobacterial genomes generally show scarcity of long microsatellites tracts. In fact the severity of scarcity of tracts increases proportionally with the increase in the number of repeats. We also calculated observed/expected (O/E) ratio of the microsatellites with different repeat numbers in the genomes of *E. coli* K12 with the complete presence of post-replicative DNA repair enzymes and *H. pylori* (both the strains: 26695 and J99) with partial presence of such a repair system marked by the absence of enzymes mutL and mutH (figure 2) (Tomb *et al* 1997).

Comparatively the O/E ratio of microsatellites (shown in figure 2) with higher repeat numbers are lower in *E. coli* compared to the *H. pylori.* Hence this feature correlates with regulatory role of repair systems to control long repeat tracts. However, it is surprising to note that mycobacterial genomes are represented by the lowest O/E ratios of the long repeats. This observation is quite contrary to what one might believe that complete absence of mismatch repair system would make a genome to get enriched with long repeat tracts. The absence of long microsatellites in mycobacterial genomes could arise due to one or more of the following:

(i) Strong selection pressures operating against long and unstable tracts in the mycobacterial genomes.

(ii) The absence of mis-match repair system also promoting accumulation of point mutations which in turn arrest the expansion of microsatellites.

(iii) The possible influence of the rich GC content on microsatellite expansions. However, this has been contested in the literature (Glenn *et al* 1996; Balloux *et al* 1998; Schlotterer 2000).

Table 4. The list of potential polymorphic microsatellites (PPMs) occuring in the five mycobacterial genomes.

(a) noncoding regions			
Motif	Repeat number	Start pos	Upstream ORF within 200 bp
M. leprae			
T	8	143422	-
C	8	229073	-
G	22	229625	-
C	**20**	**312039**	-
T	8	337466	-
G	10	347280	-
G	10	442993	-
T	8	514193	-
T	8	634418	-
G	8	663135	-
G	8	667968	-
G	8	741133	-
G	8	755942	-
G	9	976857	-
G	8	1197267	-
G	11	1309544	-
A	9	1414666	-
C	8	1778021	-
C	16	1987156	-
G	8	1987172	-
T	8	2486597	-
A	8	2562329	-
C	9	2658192	-
A	8	2946873	-
T	8	3215279	-
AT	14	308814	-
AT	15	948935	-
TA	**18**	**984591**	-
AC	9	1452573	-
CA	8	1531184	-
TA	10	1744091	-
AC	8	2211035	-
AT	**17**	**2597735**	-
AT	10	2844970	-
TA	11	2951820	-
AT	8	3221616	putative TetR-family transcriptional regulator (GI:15828443)
GTA	**9**	**2583814**	-
GAA	**21**	**2785433**	-

Table 4. (*Continued*)

M. bovis			
C	8	856443	Hyothetical protein (GI:31791947)
G	27	1619414	-
G	8	4036749	-
M. tuberculosis CDC1551			
C	8	191560	-
C	8	856394	HIT family protein (GI:15840174)
M. tuberculosis H37Rv			
C	9	854251	hypothetical protein Rv0759c (GI:15607899)

(b) coding regions

Motif	Repeat number	Start pos	Gene name
M. avium			
C	8	169357	hypothetical protein (41406264)
C	19	1793090	hypothetical protein (41407736)
C	8	2119844	hypothetical protein (41408016)
G	8	2467982	hypothetical protein (41408318)
C	10	2719084	hypothetical protein (41408519)
C	8	2820932	hypothetical protein (41408610)
C	8	3300015	hypothetical protein (41409061)
G	8	3880098	GuaA (41409587)
G	8	4209441	hypothetical protein (41409863)
M. leprae			
C	8	22194	putative penicillin-binding protein (15826881)
C	8	67158	putative membrane protein (15826903)
C	8	151870	possible membrane protein (15826944)
T	8	170900	hypothetical protein (15826958)
A	8	225690	putative phosphoribosylaminoimidazolecarboxamide formyltransferase / IMP cyclohydrolase (15826983)
G	8	299803	putative membrane protein (15827025)
A	8	593037	putative protein-export membrane protein (15827165)
T	8	795734	hypothetical protein (15827271)
A	8	893789	putative dTDP-rhamnose modification protein (15827318)
G	12	1116443	conserved hypothetical protein (15827450)
G	8	1145474	Cell division protein (15827463)
T	8	1511004	putative antiporter (15827650)
G	8	3093597	putative cell invasion protein (15828393)
M. bovis			
G	8	17896	serine/threonine protein kinase (31791192)
C	8	693132	MCE-family protein MCE2DA [FIRST PART] (31791774)
G	9	977363	PPE family protein (31792066)
G	8	1168426	Hypothetical protein (31792236)
C	8	1543144	Glucolipid sulfotransferase [first part] (31792567)
G	11	1744180	frdB and frdC (31792738)

Table 4. (*Continued*)

C	11	2320081	Transmembrane protein (31793264)
C	15	2771732	PE-PGRS family protein [first part] (31793670)
C	8	3321471	Lipoprotein LPPZ (31794183)
C	8	4076531	GLPKA (31794867)
M. tuberculosis CDC1551			
G	8	17897	serine/threonine protein kinase (15839389)
T	8	976902	PPE family protein (15840292)
G	9	976910	PPE family protein (15840292)
C	8	2340527	hypothetical protein (15841574)
C	8	3359231	lipoprotein, putative (15842564)
M. tuberculosis H37Rv			
G	8	17897	pknA (15607157)
T	8	976888	PPE (15608018)
G	9	976896	PPE (15608018)
C	8	1992322	PE_PGRS(wag22) (15608897)
C	8	2338193	hypothetical protein Rv2081c (15609218)
C	8	3364853	LppZ (15610143)

Repeats which are in bold have been tested and reported for their repeat variation (Groathouse *et al* 2004).

3.3 *Potential polymorphic microsatellites*

Mutations in microsatellites are believed to be dependent on their tract lengths and that the long tracts more than seven repeats are more prone to slippage than shorter tracts (Brinkmann *et al* 1998) and hence any such tract can be called as a potential polymorphic microsatellite (PPM). As mentioned earlier there is a general scarcity of long tracts in the mycrobacterial genomes. It is still worthwhile to examine the location of sparingly available long tracts in coding and non-coding regions. While polymorphism in microsatellites in coding regions can bring out an in-frame or out-of-frame mutations, in non-coding regions it may affect regulatory signals of the coding regions situated upstream of the coding regions (Gur-Arie *et al* 2000; Sreenu *et al* 2003). A screening of the microsatellites revealed 80-90% of the PPMs in MA, MB, MTC and MTH are in the coding regions whereas 74% of the PPMs in ML are found in its non-coding regions.

3.3a *PPMs in coding regions:* Among the PPMs found in the coding regions, 13 are present in ML, 9 in MA, 10 in MB while the MTC and MTH respectively harbour 5 and 6 PPMs (table 4). Interestingly, all the PPMs are the mononucleotide tracts and therefore insertion or deletion of mono repeat units (unless there is a simultaneous insertion or deletion of three mononucleotides repeats or their integral multiples) does lead to shifts in the reading frame causing either premature terminations or new translated sequences. In all the genomes PPMs are distributed in the ORFs encoding non-house keeping genes such as membrane proteins, virulence factors, PPE proteins, as well as hypothetical proteins.

3.3b *PPMs in non-coding regions:* Among the genomes, MA is completely devoid of PPMs in its non-coding regions. MB, MTC and MTH harbour some PPMs while ML is relatively richer in PPMs with 38 tracts (table 4). Most of the PPMs (38 out of 44) found in these genomes comprise of the mononucleotide tracts. Of the PPMs a large majority of PPMs are situated more than 200 bp away from the upstream coding regions thereby hinting a probable functional irrelevance of their polymorphism on the regulatory elements of the downstream coding regions. For example, in ML one of the PPMs is a dinucleotide tract $(AT)_8$ and is located 79 bp away from an ORF annotated as tetR (tetracycline resistance) family transcriptional regulator. This gene encodes for a repressor protein that regulates the expression of tetA protein which is a membrane-associated protein involved in export of tetracycline out of the bacterial cell (Hinrichs *et al* 1994; Kisker *et al* 1995). Five out of the thirty-eight PPMs from ML have already been tested and reported to be polymorphic in clinical isolates (table 4) (Groathouse *et al* 2004). All these variable microsatellites are more than 200 bp away from the coding regions and they are used as molecular markers for strain typing (Groathouse *et al* 2004). In MTC, one of the two PPMs viz. $(C)_8$ is 29 bp away from HIT (histidine triad) family protein. Function of the proteins in this family is unknown, however, they are conserved in various prokaryotes as well as in eukaryotes

(Seraphin 1992). The PPMs in MTH and MB seems to be the equivalents of the PPM in MTC but the downstream coding regions have been annotated as hypothetical proteins.

4. Discussion

As could be anticipated, the distributions of microsatellites in the closely related MB, MTC and MTH genomes are similar to each other. ML and MA show distinct distribution profiles. Although SSRs are distributed throughout the mycobacterial genomes there are some regions that are markedly either rich or poor in them. MB, MTC, MA and MTH have more number of microsatellite rich regions than the poor regions. Many stress response genes, transcription regulators and virulence factors are embedded in the repeat rich regions. Genes that are unique to mycobacteria, such as PE and PPE are also present in repeat rich regions. Hence, it appears that the repeat rich regions act as reservoirs of genes, which are capable of bringing about certain variability in virulence, antigenicity and host adaptation. (The complete list of ORFs which are falling in the microsatellite rich or poor region are given in the supplementary material.) In stressful conditions, increased microsatellite mutations could generate gene variants in different populations, which confer stress response to tolerate and survive in hostile environments. A higher number of repeat enriched regions in MB, MA, MTC and MTH as compared to ML, indicates an intrinsic plasticity of these genomes perhaps to deal with hostile environments.

By-and-large microsatellite motif distributions are similar to those found in the other prokaryotes. Under-representation of mononucleotide, di, tetra and penta repeats is commonly observed in many prokaryotic genomes (Field and Wills 1998) so also the over-representation of the trinucleotide and hexanucleotide repeats. Under-representation of di, tetra and penta motifs in the genomes, which fall mostly in the coding regions can be attributed to selection pressures to avoid chances of frameshift mutations brought out by these microsatellites in the coding regions. In ML, where nearly half the genome is non-coding, less selection pressure against frameshift mutations can be expected. Indeed our analysis shows that the tetra and penta repeats are excessively represented.

Among the microsatellites, the mono, di and tri with iterations of two are in excess, and such abundances have also been reported in some of the prokaryote genomes (Field and Wills 1998). Codon usage has been attributed to such excess of short repeat sequences (Field and Wills 1998).

In all the mycobacterial genomes, base composition of the genomes appears to be influencing the abundance as well as enrichment of microsatellites. Most of the di-hexa tracts are G+C rich and they are excess in number. The mono tracts show different characteristics in which the most frequent G/C tracts are under-represented while the less abundant A/T tracts are over-represented. This indicates a trend in the evolution of mononucleotide tracts where A/T tracts are getting accumulated. Among the dinucleotide tracts TA repeats' under-representation is also observed in many prokaryotic genomes and this repeat has been considered as "universally under-represented" (Burge *et al* 1992; Karlin *et al* 1997). TA depletion in genomes has been attributed to avoidance of inappropriate binding of the regulatory elements as the repeat TA is part of regulatory sequences (Burge *et al* 1992).

Mutations in SSRs especially small motifs (mono and dinucleotide) with high repeat number are more prone to mutations than long motifs with low repeat number (Shinde *et al* 2003). In the studied genomes, all the long repeats (PPMs) located in coding regions are mononucleotide repeats only, hinting that these coding regions may act as contingency loci. Most of the contingency genes in pathogenic bacteria code for membrane proteins and membrane associated proteins (Moxon *et al* 1994) favouring "antigenic variation", thus conferring a particular selective advantage to escape the host immune system.

Distribution of microsatellites in a genome is considered to be an equilibrium between expansion due to addition of repeat units and point mutations which break long microsatellites into smaller tracts (Kruglyak *et al* 1998). The length polymorphism of a repeat tract is primarily controlled by the selection forces which act on it (Nauta and Weissing 1996). Hence, microsatellite distribution and frequency of a genome reflects the underlying mutational processes, selection constrains as well as DNA repair mechanisms. Influence of GC-content of a genome on the microsatellite mutation rates has also been discussed in the literature apparently with no consensus opinion (Glenn *et al* 1996; Balloux *et al* 1998; Schlotterer 2000). In the case of mycobacterial genomes, there is a strong control on the tract lengths compared to known bacterial genome such as *E. coli.* In fact the per Kbp distribution of repeats in mycobacterial genomes is in the range of 220–230 which is equal to that observed in *E. coli* and *H. pylori* (data taken from MICdb, Sreenu *et al* 2003), indicating a possible tight regulation of microsatellite evolution (birth, mutation and death). The absence of post-replicative repair system, in principle, should act in favour of microsatellite expansions. However, absence of such a repair system can also promote accumulation of point mutations which in turn arrest the growth of microsatellites.

Acknowledgements

This work was supported by the core grants of CDFD and, VBS and PK greatly acknowledge support from the Council of Scientific and Industrial Research (CSIR), New Delhi. Last, but not the least, the authors thank the two anonymous referees for their helpful critical comments.

References

Balloux F, Ecoffey E, Fumagalli L, Goudet J, Wyttenbach A and Hausser J 1998 Microsatellite conservation, polymorphism, and GC content in shrews of the genus Sorex (Insectivora, Mammalia); *Mol. Biol. Evol.* **15** 473–475

Brinkmann B, Klintschar M, Neuhuber F, Huhne J and Rolf B 1998 Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat; *Am. J. Hum. Genet.* **62** 1408–1415

Burge C, Campbell A M and Karlin S 1992 Over- and under-representation of short oligonucleotides in DNA sequences; *Proc. Natl. Acad. Sci. USA* **89** 1358–1362

Cole S T, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon S V, Eiglmeier K *et al* 1998 Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence; *Nature* (*London*) **393** 537–544

Cole S T, Eiglmeier K, Parkhill J, James K D, Thomson N R, Wheeler P R, Honore N, Garnier T *et al* 2001 Massive gene decay in the leprosy bacillus; *Nature* (*London*) **409** 1007–1011

Cosma C L, Sherman D R and Ramakrishnan L 2003 The secret lives of the pathogenic mycobacteria; *Annu. Rev. Microbiol.* **57** 641–676

Ellegren H 2004 Microsatellites: simple sequences with complex evolution; *Nature Rev. Genet.* **5** 435–445

Field D and Wills C 1998 Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces; *Proc. Natl. Acad. Sci. USA* **95** 1647–1652

Fleischmann *et al* 2002 Whole genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains; *J. Bacteriol.* **184** 5479–5490

Garnier T, Eiglmeier K, Camus J C, Medina N, Mansoor H, Pryor M, Duthoy S, Grondin S *et al* 2003 The complete genome sequence of *Mycobacterium bovis*; *Proc. Natl. Acad. Sci. USA* **100** 7877–7882.

Glenn T C, Stephan W, Dessauer H C and Braun M J 1996 Allelic diversity in alligator microsatellite loci is negatively correlated with GC content of flanking sequences and evolutionary conservation of PCR amplifiability; *Mol. Biol. Evol.* **13** 1151–1154

Groathouse N A, Rivoire B, Kim H, Lee H, Cho S N, Brennan P J and Vissa V D 2004 Multiple polymorphic loci for molecular typing of strains of *Mycobacterium leprae*; *J. Clin. Microbiol.* **42** 1666–1672

Gur-Arie R, Cohen C J, Eitan Y, Shelef L, Hallerman E M and Kashi Y 2000 Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism; *Genome Res.* **10** 62–71

Heller M, van Santen V and Kieff E 1982 Simple repeat sequence in Epstein-Barr virus DNA is transcribed in latent and productive infections; *J. Virol*. **44** 311–320

Hermans P W M, van Soolingen D, Bik M, de Haas P E W, Dale J W and van Embden J D A 1991 The insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *M. tuberculosis* complex strains; *Infect. Immun.* **59** 2695–2705

Hinrichs W, Kisker C, Duvel M, Muller A, Tovar K, Hillen W and Saenger W 1994 Structure of the Tet repressor-tetracycline complex and regulation of antibiotic resistance; *Science* **264** 418–420

Hood D W, Deadman M E, Jennings M P, Bisercic M, Fleischmann R D, Venter J C and Moxon E R 1996 DNA repeats identify novel virulence genes in *Haemophilus influenzae*; *Proc. Natl. Acad. Sci. USA* **93** 11121–11125

Jackson A L, Chen R and Loeb L A 1998 Induction of microsatellite instability by oxidative DNA damage; *Proc. Natl. Acad. Sci. USA* **95** 12468–12473

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

Karlin S, Mrazek J and Campbell A M 1997 Compositional biases of bacterial genomes and evolutionary implications; *J. Bacteriol.* **179** 3899–3913

Kisker C, Hinrichs W, Tovar K, Hillen W and Saenger W 1995 The complex formed between Tet repressor and tetracycline-Mg^{2+} reveals mechanism of antibiotic resistance; *J. Mol. Biol.* **247** 260–280

Kruglyak S, Durrett R T, Schug M D and Aquadro C F 1998 Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations; *Proc. Natl. Acad. Sci. USA* **95** 10774–10778

Levinson G and Gutman G A 1987 Slipped-strand mispairing: a major mechanism for DNA sequence evolution; *Mol. Biol. Evol.* **4** 203–221

Li L, Bannantine J P, Zhang Q, Amonsin A, May B J, Alt D, Banerji N, Kanjilal S and Kapur V 2005 The complete genome sequence of *Mycobacterium avium* subspecies paratuberculosis; *Proc. Natl. Acad. Sci. USA* **102** 12344–12349

Moxon E R, Rainey P B, Nowak M A and Lenski R E 1994 Adaptive evolution of highly mutable loci in pathogenic bacteria; *Curr. Biol.* **4** 24–33

Murphy G L, Connell T D, Barritt D S, Koomey M and Cannon J G 1989 Phase variation of gonococcal protein II: regulation of gene expression by slipped-strand mispairing of a repetitive DNA sequence; *Cell* **56** 539–547

Nauta M J and Weissing F J 1996 Constraints on allele size at microsatellite loci: implications for genetic differentiation; *Genetics* **143** 1021–1032

Press W H, Flannery B P, Teukolsky S A and Vetterling W T 1992 *Numerical recipes in C* (Cambridge: Cambridge University Press)

Rocha E P, Matic I and Taddei F 2002 Over-representation of repeats in stress response genes: a strategy to increase versatility under stressful conditions?; *Nucleic Acids Res.* **30** 1886–1894

Schlotterer C 2000 Evolutionary dynamics of microsatellite DNA; *Chromosoma* **109** 365–371

Schlotterer C and Tautz D 1992 Slippage synthesis of simple sequence DNA; *Nucleic Acids Res.* **20** 211–215

Seraphin B 1992 The HIT protein family: a new family of proteins present in prokaryotes, yeast and mammals; *DNA Seq*. **3** 177–179

Shinde D, Lai Y, Sun F and Arnheim N 2003 *Taq* DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood

anlysis: (CA/GT)n and (A/T)n microsatellites; *Nucleic Acids Res.* **31** 974–980

Springer B, Sander P, Sedlacek L, Hardt W, Mizrahi V, Schär P and Böttger E C 2004 Lack of mismatch correction facilitates genome evolution in mycobacteria; *Mol. Microbiol.* **53** 1601–1609

Sreenu V B, Alevoor V, Nagaraju J and Nagarajaram H A 2003 MICdb: database of prokaryotic microsatellites; *Nucleic Acids Res.* **31** 106–168

Sreenu V B, Kumar P, Nagaraju J and Nagarajaram H A 2006 Microsatellite polymorphism across the *M. tuberculosis* and *M. bovis* genomes: implications on genome evolution and plasticity; *BMC Genomics* **7** 78

Tomb J F, White O, Kerlavage A R, Clayton R A, Sutton G G, Fleischmann R D, Ketchum K A, Klenk H P, Gill S *et al* 1997 The complete genome sequence of the gastric pathogen *Helicobacter pylori*; *Nature* (*London*) **388** 539–547

van Belkum A, Scherer S, van Alphen L and Verbrugh H 1998 Short-sequence DNA repeats in prokaryotic genomes; *Microbiol. Mol. Biol. Rev.* **62** 275–293

van Ham S M, van Alphen L, Mooi F R and van Putten J P 1993 Phase variation of *H. influenzae fimbriae*: transcriptional control of two divergent genes through a variable combined promoter region; *Cell* **73** 1187–1196

van Soolingen D, de Haas P E W, Hermans P W M, Groenen P M A and van Embden J D A 1993 Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of *Mycobacterium tuberculosis*; *J. Clin. Microbiol.* **31** 1987–1995

*e*Publication: 18 December 2006

Simple sequence repeats in mycobacterial genomes

VATTIPALLY B SREENU, PANKAJ KUMAR, JAVAREGOWDA NAGARAJU and HAMPAPATHALU A NAGARAJARAM

J. Biosci. **32(1), January 2007, 3–15**, **© Indian Academy of Sciences**

Supplementary Material 1

The list of microsatellite rich as well as poor regions in the five mycobacterial genomes.

	Local GC%	Repeat rich(+)/ Repeat poor(-)	Total ORFs	Number of hypothetical proteins	Non hypothetical proteins
			M. avium		
19-20	0.72	+	11	8	SodA, GlpQ1, prephenatedehydratase
34-35	0.70	+	8	7	DNApolymeraseIIIsubunits
39-40	0.71	+	7	5	SelD, PfpI
44-45	0.73	+	9	7	TrbB, CspA_1
47-48	0.70	+	8	4	EphA, FtsH, GTPcyclohydrolaseI, FolP
63-64	0.71	+	8	5	phosphoribosylamine--glycineligase, GgtA, adenylosuccinatelyase
92-93	0.69	+	7	4	succinyl-CoAsynthetasesubunitbeta, succinyl-CoAsynthetasealphasubunit, acetyl-CoAacetyltransferase
118-119	0.73	+	8	4	LipI, primosomeassemblyproteinPriA, methionyl-tRNAformyltransferase, Fmu
128-129	0.71	+	5	2	MutA, methylmalonyl-CoAmutase, arginine/ornithine transportsystemATPase
145-146	0.69	+	8	3	lysyl-tRNAsynthetase, translationinitiationfactorIF-3, 50SribosomalproteinL35, 50SribosomalproteinL20, TsnR
146-147	0.72	+	7	0	phenylalanyl-tRNAsynthetasebetasubunit, N-acetyl-gamma-glutamyl-phosphatereductase, bifunctionalornithineacetyltransferase/N-acetylglutamatesynthase, acetylglutamatekinase, acetylornithineaminotransferase, ornithinecarbamoyltransferase, argininerepressor
152-153	0.70	+	10	7	tyrosinerecombinase, cytidylatekinase, GTP-bindingproteinEngA
166-167	0.69	+	7	6	PE_5
171-172	0.69	+	9	5	6-phosphogluconatedehydrogenase, Ndh, shortchaindehydrogenase, ModA
176-177	0.68	+	11	6	LppE, shortchaindehydrogenase, chorismatemutase, FbpB, AdhA_2
196-197	0.71	+	0	0	
210-211	0.72	+	7	4	UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelateligase, PbpB, S-adenosyl-methyltransferase
257-258	0.69	+	5	4	glycerol-3-phosphateacyltransferase
274-275	0.66	+	6	4	AlkA, Ogt
285-286	0.69	+	6	3	alpha-ketoglutaratedecarboxylase, LpqZ, malatedehydrogenase
289-290	0.72	+	7	4	acyl-CoAsynthase, succinyl-diaminopimelatedesuccinylase, PE_6
292-293	0.70	+	7	5	acyl-CoAsynthase, sulfateadenylyltransferase
328-329	0.69	+	11	7	ribosomereleasingfactor, uridylatekinase, amidase, elongationfactorTs
342-343	0.69	+	7	5	isopentenylpyrophosphateisomerase, Phr
359-360	0.70	+	8	7	CatB

361-362	0.72	+	7	5	pyruvatedecarboxylase, PknJ
367-368	0.73	+	7	5	LipV, molybdopterinbiosynthesisproteinMoeB
389-390	0.72	+	6	4	CtpI, AdhD
404-405	0.71	+	5	3	MmpL3, tRNA(guanine-N(7)-)-methyltransferase
423-424	0.72	+	9	8	AtsG
448-449	0.73	+	10	7	glutamate-1-semialdehydeaminotransferase, CcsA, CcsB
449-450	0.74	+	9	7	1,4-dihydroxy-2-naphthoateoctaprenyltransferase,, 5'-methylthioadenosinephosphorylase
88-89	0.61	-	10	10	
283-284	0.67	-	8	6	DeaD, LprE
309-310	0.64	-	12	12	
415-416	0.63	-	8	8	
418-419	0.62	-	7	4	IS1110transposase, MmpS1, MmpL4_5
			M. leprae		
37-38	0.54	-	5	3	thiaminebiosynthesisproteinThiC, phosphomethylpyrimidinekinase
157-158	0.57	-	4	2	proteasome[beta]-typesubunit2, proteasome[alpha]-typesubunit1
164-165	0.56	-	1	1	
198-199	0.55	-	1	1	
243-244	0.56	-	6	3	probableLysR-familytranscriptionalregulator, alkylhydroperoxidereductase, L-lactatedehydrogenase
263-264	0.56	-	3	0	putativeaminopeptidase2, phosphoribosylformylglycinamidinesynthasesubunitI, phosphoribosylformylglycinamidinesynthase
326-327	0.52	-	8	2	putativecelldivisionprotein, putativecelldivisionprotein, glucose-inhibiteddivisionproteinB, putativeinnermembraneprotein, ribonucleaseP, 50SribosomalproteinL34
			M. bovis		
33-34	0.75	+	4	2	PE-PGRSFAMILYPROTEIN[FIRST, PE-PGRSFAMILYPROTEIN
35-36	0.67	+	10	3	PEFAMILYPROTEIN, PPEFAMILYPROTEIN, LOWMOLECULARWEIGHTPROTEIN, PROBABLECONSERVEDTRANSMEMBRANEPROTEIN, PROBABLEPROTEASEPRECURSOR, PROBABLECONSERVEDTRANSMEMBRANEPROTEIN, PROBABLETRANS-ACONITATEMETHYLTRANSFERASETAM
36-37	0.66	+	8	4	PROBABLESULFATASE, PE-PGRSFAMILYPROTEIN, PROBABLETRANSCRIPTIONALREGULATORYPROTEIN, PROBABLEDEHYDROGENASE/REDUCTASE
37-38	0.62	+	5	1	PPEFAMILYPROTEIN, PUTATIVEOXIDOREDUCTASE, PROBABLECONSERVEDINTEGRALMEMBRANE, POSSIBLECONSERVEDEXPORTEDPROTEIN

42-43	0.63	+	4	0	molecularchaperoneDnaK, PROBABLEGRPEPROTEIN(HSP-70, PROBABLECHAPERONEPROTEINDNAJ1, PROBABLEHEATSHOCKPROTEIN
43-44	0.64	+	5	3	adenylosuccinatesynthetase, PROBABLECONSERVEDINTEGRALMEMBRANE
67-68	0.71	+	7	5	PROBABLETRANSCRIPTIONALREGULATORYPROTEIN, PE-PGRSFAMILYPROTEIN
83-84	0.70	+	10	5	POSSIBLETRANSCRIPTIONALREGULATORYPROTEIN, PUTATIVETRANSPOSASE(FRAGMENT), PE-PGRSFAMILYPROTEIN, POSSIBLETRANSCRIPTIONALREGULATORYPROTEIN, PE-PGRSFAMILYPROTEIN
84-85	0.70	+	8	4	PE-PGRSFAMILYPROTEIN, PROBABLE3-HYDROXYISOBUTYRATEDEHYDROGENASEMMSB, PROBABLEACYL-COADEHYDROGENASEFADE9, PROBABLEMETHYLMALONATE-SEMIALDEHYDEDEHYDROGENASEMMSA
85-86	0.63	+	10	4	PPEFAMILYPROTEIN, PUTATIVETRANSPOSASE(FRAGMENT), POSSIBLETWOCOMPONENTSYSTEM, POSSIBLETWOCOMPONENTSYSTEM, POSSIBLEZINC-CONTAININGALCOHOLDEHYDROGENASE, POSSIBLEFERREDOXIN
92-93	0.70	+	8	3	PROBABLETRANSCRIPTIONALREGULATORYPROTEIN, POSSIBLEDEAMINASE, PUTATIVETRANSPOSASE(FRAGMENT), PE-PGRSFAMILYPROTEIN, PE-PGRSFAMILYPROTEIN
97-98	0.66	+	8	2	PROBABLEACYL-COADEHYDROGENASEFADE10, POSSIBLECONSERVEDEXPORTEDPROTEIN, POSSIBLECONSERVEDTRANSMEMBRANEPROTEIN, PPEFAMILYPROTEIN, POSSIBLECONSERVEDTRANSMEMBRANEPROTEIN, POSSIBLETRANSCRIPTIONALREGULATORYPROTEIN
109-110	0.69	+	7	1	PE-PGRSFAMILYPROTEIN, PE-PGRSFAMILYPROTEIN, 50SribosomalproteinL32, PE-PGRSFAMILYPROTEIN, MYCOBACTERIALPERSISTENCEREGULATORMRPA, PROBABLETWOCOMPONENTSENSOR
121-122	0.71	+	8	1	POSSIBLEHEMOLYSIN-LIKEPROTEIN, SHORT(C15)CHAINZ-ISOPRENYL, PE-PGRSFAMILYPROTEIN, PEFAMILYPROTEIN, PEFAMILYPROTEIN, PROBABLECELLULASECELA2A(ENDO-1,4-BETA-GLUCANASE), PROBABLECELLULASECELA2B(ENDO-1,4-BETA-GLUCANASE)
126-127	0.66	+	9	3	5-methyltetrahydropteroyltriglutamate--homocysteinemethyltransferase, PPEFAMILYPROTEIN, POSSIBLEACETYL-COAACETYLTRANSFERASE(ACETOACETYL-COA, POSSIBLEENOYL-COAHYDRATASE, PUTATIVEOXIDOREDUCTASE, PROBABLEINTEGRALMEMBRANEPROTEIN

162-163	0.70	+	5	0	6-phosphogluconolactonase, PUTATIVEOXPPCYCLEPROTEIN, glucose-6-phosphate1-dehydrogenase, transaldolase, transketolase
163-164	0.70	+	6	2	PE-PGRSFAMILYPROTEIN, POSSIBLETRANSCRIPTIONALACTIVATORPROTEIN, PROBABLEQUINONEREDUCTASEQOR, PROBABLEUNIDENTIFIEDANTIBIOTIC-TRANSPORTINTEGRAL
196-197	0.64	+	5	3	POSSIBLEINTEGRALMEMBRANEPROTEIN, acyl-CoAsynthase
215-216	0.61	+	3	1	isocitratelyase, PPEFAMILYPROTEIN
228-229	0.68	+	0	0	
244-245	0.65	+	9	2	Probableconservedtransmembraneprotein, PROBABLECONSERVEDMEMBRANEPROTEIN, Possible conservedintegralmembrane, PROBABLETRANSMEMBRANECYTOCHROMEC, Probableasparagine synthetaseAsnB, ProbablecarbohydratekinaseCbhK, POSSIBLECONSERVEDMEMBRANEPROTEIN
276-277	0.71	+	4	0	enoyl-CoAhydratase, PE-PGRSFAMILYPROTEIN[FIRST, PROBABLETRANSCRIPTIONALREGULATORYPROTEIN, HYPOTHETICALALANINERICHPROTEIN
277-278	0.67	+	9	4	PE-PGRSFAMILYPROTEIN[FIRST, dihydrolipoamideacetyltransferase, PROBABLEPYRUVATEDEHYDROGENASEE1, PROBABLEPYRUVATEDEHYDROGENASEE1, PROBABLECITRATE(PRO-3S)-LYASE(BETA
290-291	0.66	+	8	1	pyridoxinebiosynthesisprotein, pyridoxamine5'-phosphateoxidase, PPEFAMILYPROTEIN, PROBABLECONSERVEDMEMBRANEPROTEIN, ALPHA-MANNOSYLTRANSFERASEPIMA, lipidAbiosynthesislauroyl, PROBABLEPISYNTHASEPGSA1
312-313	0.65	+	7	1	dihydrolipoamidedehydrogenase, POSSIBLENICKEL-TRANSPORTINTEGRALMEMBRANE, shortchaindehydrogenase, PROBABLEALDEHYDEDEHYDROGENASEALDC, POSSIBLEAMIDOTRANSFERASE, PROBABLEGLUTAMINESYNTHETASEGLNA4
333-334	0.66	+	10	1	PROBABLELIPOPROTEINLPQA, PUTATIVEESAT-6LIKEPROTEIN, PPEFAMILYPROTEIN, PEfamilyprotein, PUTATIVESECRETEDESAT-6LIKE, PEFAMILYPROTEIN, PPEFAMILYPROTEIN, PEFAMILYPROTEIN, PROBABLETRANSPOSASE
348-349	0.63	+	6	0	NADHdehydrogenasesubunitN, PPEFAMILYPROTEIN, PPEFAMILYPROTEIN, POSSIBLETRANSCRIPTIONALREGULATORYPROTEIN, POSSIBLEDIOXYGENASE, POSSIBLEINTEGRALMEMBRANEPROTEIN
368-369	0.64	+	5	1	isocitratedehydrogenase, O-acetylhomoserinesulfhydrylase, homoserineO-acetyltransferase, POSSIBLEMETHYLTRANSFERASE(METHYLASE)

369-370	0.72	+	2	1	PE-PGRSFAMILYPROTEIN[FIRST
370-371	0.64	+	3	0	PPEFAMILYPROTEIN[FIRST, PROBABLETRANSPOSASE, PROBABLETRANSPOSASE
371-372	0.64	+	0	0	
375-376	0.68	+	8	2	POSSIBLETRANSPOSASE, POSSIBLETRANSPOSASE, PE-PGRSFAMILYPROTEIN, POSSIBLEDEHYDROGENASE, PROBABLECONSERVEDLIPOPROTEINLPQD, shortchaindehydrogenase
386-387	0.64	+	9	0	MCE-FAMILYPROTEINMCE4D, MCE-FAMILYPROTEINMCE4C, MCE-FAMILYPROTEINMCE4B, MCE-FAMILYPROTEINMCE4A, CONSERVEDHYPOTHETICALINTEGRALMEMBRANE, CONSERVEDHYPOTHETICALINTEGRALMEMBRANE, 3-ketoacyl-(acyl-carrier-protein)reductase, PROBABLEFERREDOXINFDXD, PROBABLEACYL-COADEHYDROGENASEFADE26
387-388	0.76	+	2	0	acyl-CoAsynthase, PE-PGRSFAMILYPROTEIN
388-389	0.75	+	3	2	PE-PGRSFAMILYPROTEIN
389-390	0.70	+	7	2	PE-PGRSFAMILYPROTEIN, acyl-CoAsynthase, enoyl-CoAhydratase, PROBABLECYTOCHROMEP450MONOOXYGENASE, PROBABLECYTOCHROMEP450MONOOXYGENASE
394-395	0.68	+	8	0	PROBABLETRANSCRIPTIONALREGULATORYPROTEIN, PPEFAMILYPROTEIN, shortchaindehydrogenase, PROBABLEACYL-COADEHYDROGENASEFADE30, acyl-CoAsynthase, PROBABLEACYL-COADEHYDROGENASEFADE31, PROBABLEACYL-COADEHYDROGENASEFADE32, PROBABLEACYL-COADEHYDROGENASEFADE33
398-399	0.66	+	7	1	PROBABLEATP-DEPENDENTCLPPROTEASE, PROBABLELSR2PROTEINPRECURSOR, lysyl-tRNAsynthetase, aspartate1-decarboxylaseprecursor, pantoate--beta-alanineligase, CONSERVEDHYPOTHETICALALANINEAND
426-427	0.64	+	9	2	POSSIBLEMEMBRANEPROTEIN, POSSIBLEHISTONE-LIKEPROTEINHNS, ribonucleaseactivityregulatorprotein, MONOOXYGENASEETHA, TRANSCRIPTIONALREGULATORYREPRESSORPROTEIN, POSSIBLEMEMBRANEPROTEIN, PUTATIVENADH-DEPENDENTGLUTAMATESYNTHASE

343-344	0.62	-	10	2	PROBABLENADPH:ADRENODOXINOXIDOREDUCTASEFPRA, POSSIBLEALKYLDIHYDROXYACETONEPHOSPHATESYNTHASEAGPS, PROBABLEMOLYBDENUMCOFACTORBIOSYNTHESIS, PROBABLEPTERIN-4-ALPHA-CARBINOLAMINEDEHYDRATASEMOAB1, molybdenumcofactorbiosynthesisprotein, PROBABLEMOLYBDENUMCOFACTORBIOSYNTHESIS, POSSIBLEPHOSPHATASE, PROBABLETRANSPOSASE
			M. tuberculosis CDC1551		
33-34	0.72	+	11	9	PE_PGRS (15839660), PPE (15839665)
35-36	0.67	+	9	5	PEfamilyprotein, secretedantigen,putative, subtilasefamilyprotein, trans-aconitatemethyltransferase
36-37	0.66	+	6	2	PE_PGRSfamilyprotein, DNA-bindingprotein, CopGfamily, transcriptionalregulator,TetRfamily, oxidoreductase,short-chaindehydrogenase/reductasefamily
37-38	0.62	+			
42-43	0.63	+	5	0	heat shock protein (grpE) (15839737), heat shock protein (dnaJ) (15839738), transcriptional regulator HspR(15839739), PPE (15839740), PPE (15839741)
43-44	0.64	+	7	5	adenylosuccinate synthetase (15839743), divalent cation transporter (15839748)
67-68	0.71	+	5	2	nitroreductase,cobalaminbiosynthesisprotein, PAP2superfamilyprotein, baiEprotein
83-84	0.69	+	15	12	transcriptionalregulator,MarRfamily, IS1557transposase, PE_PGRSfamilyprotein
84-85	0.70	+	9	5	PE_PGRSfamilyprotein, 3-hydroxyisobutyratedehydrogenase, acyl-CoAdehydrogenase, methylmalonicacidsemialdehydedehydrogenase
85-86	0.63	+	10	3	PPEfamilyprotein, DNA-bindingresponseregulator, sensorhistidinekinase, HITfamilyprotein, steroid isomerase,putative, zinc-bindingdehydrogenase, ferredoxin-relatedprotein
92-93	0.71	+	7	3	transcriptionalregulator,ArsRfamily, truncatedIS1605transposase, PE_PGRSfamilyprotein, PE_PGRSfamilyprotein
96-97	0.69	+	10	5	molybdenumcofactorbiosynthesisprotein, molybdopterinbiosynthesisMogprotein, molybdopterincofactorbiosynthesisprotein, molybdenumcofactorbiosynthesisprotein, cold-shockdomainfamilyprotein
97-98	0.66	+	9	7	acyl-CoAdehydrogenase,putative, PPEfamilyprotein
109-110	0.69	+	5	0	PE_PGRSfamilyprotein, 50SribosomalproteinL32, PE_PGRSfamilyprotein, DNA-bindingresponseregulator, sensorhistidinekinase
121-122	0.69	+	9	3	undecaprenyldiphosphatesynthase, PEfamilyprotein, PEfamilyprotein, cellulase-relatedprotein, PE_PGRSfamilyprotein, pantothenatekinase

126-127	0.66	+	12	7	PPEfamilyprotein, ketoacyl-CoAthiolase-relatedprotein, enoyl-CoAhydratase/ isomerasefamilyprotein, enoyl-CoAhydratase, enoyl-CoAhydratase
163-164	0.73	+	6	2	PE_PGRS (15840909), cytochrome c oxidase folding protein (15840911), PE_PGRS (15840912), quinone oxidoreductase (15840914)
197-198	0.63	+	6	4	PPEfamilyprotein, phospholipaseC
198-199	0.67	+	5	1	molybdopterinoxidoreductase, membraneprotein,MmpLfamily, IS6110,transposase, serineesterase,cutinasefamily
204-205	0.66	+	7	2	PPEfamilyprotein, PE_PGRSfamilyprotein, PEfamilyprotein, PPEfamilyprotein, PPEfamilyprotein
216-217	0.60	+	5	1	PPE (15841389), PPE (15841389) acyltransferase family protein, lipoprotein (15841392)
230-231	0.68	+	1	1	
241-242	0.71	+	6	0	N-acetylglucosaminyltransferase, celldivisionproteinFtsW, UDP-N-acetylmuramoyl-L-alanyl-D-glutamatesynthetase, phospho-N-acetylmuramoyl-pentapeptide-transferase, UDP-N-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate--D-alanyl-D-alanylligase, UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelateligase
246-247	0.66	+	9	4	cytochromecoxidase,subunit, asparaginesynthetase,putative, carbohydratekinase,PfkBfamily, HesB/YadR/YfhFfamilyprotein, nicotinate-nucleotide--dimethylbenzimidazolephosphoribosyltransferase
279-280	0.73	+	3	1	PE_PGRSfamilyprotein, transcriptionalregulator,LuxRfamily
293-294	0.66	+	8	0	pyridoxamine5'-phosphateoxidase, PPEfamilyprotein, MutT/nudixfamilyprotein, glycosyltransferase, lipidAbiosynthesislauroyl, CDP-diacylglycerol--glycerol-3-phosphate3-phosphatidyltransferase,putative, HITfamilyprotein, threonyl-tRNAsynthetase
315-316	0.68	+	7	1	cobyrinicacida,c-diamidesynthase, cob(I)yrinicacida,c-diamideadenosyltransferase, magnesiumchelatase,putative, ElaAfamilyprotein, malate:quinoneoxidoreductase, PE_PGRSfamilyprotein
351-352	0.66	+	8	0	ATPsynthasesubunitE, NADHdehydrogenaseI,F, NADHdehydrogenasegammasubunit, NADHdehydrogenasesubunitH, NADHdehydrogenasesubunitI, NADHdehydrogenasesubunitJ, NADHdehydrogenasekappasubunit, NADHdehydrogenasesubunitL
352-353	0.63	+	6	1	NADHdehydrogenasesubunitN, PPEfamilyprotein, PPEfamilyprotein, transcriptionalregulator,TetRfamily, Rieske2Fe-2Sfamilyprotein
373-374	0.71	+	3	1	IS1608', transposase (15842945), IS1561', transposase (15842946)
374-375	0.64	+	1	1	

375-376	0.64	+	0	0	
379-380	0.69	+	9	2	IS1560transposase, IS1560transposase, PE_PGRSfamilyprotein, MaoCfamilyprotein, lipoprotein,putative, shortchaindehydrogenase, cyclopropane-fatty-acyl-phospholipidsynthase1
387-388	0.66	+	17	15	integrase,putative, bacteriophageprotein
392-393	0.78	+	2	0	PE_PGRS (15843119), PE_PGRS (15843120)
393-394	0.74	+	2	1	fatty-acid-CoAligase-relatedprotein
394-395	0.71	+	7	3	acyl-CoAsynthase, enoyl-CoAhydratase, P450heme-thiolateprotein, N5,N10-methylenetetrahydromethanopterinreductase-relatedprotein
399-400	0.68	+	7	0	PPEfamilyprotein, shortchaindehydrogenase, acyl-CoAdehydrogenase,putative, acyl-CoAsynthase, acyl-CoAdehydrogenase,putative, acyl-CoAdehydrogenase,putative, acyl-CoAdehydrogenase,putative
402-403	0.70	+	9	5	carbonicanhydrase, A/G-specificadenineglycosylase ,putative, PE_PGRSfamilyprotein, hydrolase,alpha/betahydrolasefold
403-404	0.65	+	8	1	ATP-dependentClpprotease,ATP-binding, lsr2protein, lysyl-tRNAsynthetase, transcriptionalactivator,putative,Baf, aspartate1-decarboxylaseprecursor, pantoate--beta-alanineligase, chalcone/stilbenesynthasefamilyprotein
207-208	0.65	-	6	3	glycinedehydrogenase, haloalkanedehalogenase, PEG1/MESTprotein
			M. tuberculosis H37Rv		
33-34	0.72	+	7	3	acyl-CoA synthase (fadD27) (15607416), PE_PGRS (15607419), PE_PGRS (15607420), PPE (15607421)
35-36	0.67	+	9	3	ESAT-6LIKEPROTEINESXG, LOWMOLECULARWEIGHTPROTEIN, PROBABLECONSERVEDTRANSMEMBRANEPROTEIN, PROBABLEMEMBRANE-ANCHOREDMYCOSINMYCP3, PROBABLE CONSERVEDTRANSMEMBRANEPROTEIN, PROBABLETRANS-ACONITATEMETHYLTRANSFERASETAM
36-37	0.66	+	7	4	PE-PGRSFAMILYPROTEIN, PROBABLETRANSCRIPTIONALREGULATORYPROTEIN, PROBABLEDEHYDROGENASE/REDUCTASE
37-38	0.62	+	6	2	PPEFAMILYPROTEIN, PUTATIVEOXIDOREDUCTASE, PROBABLECONSERVEDINTEGRALMEMBRANE, POSSIBLECONSERVEDEXPORTEDPROTEIN
42-43	0.63	+	5	0	heat shock protein (grpE) (15607492), heat shock protein (dnaJ) (15607493), heat shock protein (hspR) (15607494), PPE (15607495), PPE (15607496)
43-44	0.64	+	7	5	adenylosuccinate synthase (purA) (15607498), magnesium ion transporter (mgtE) (15607503)
67-68	0.70	+	7	5	PE-PGRSFAMILYPROTEIN, PROBABLECONSERVEDLIPOPROTEINLPQN
83-84	0.71	+	10	7	PE_PGRS (15607882), PE_PGRS (15607886), PE_PGRS (15607887)

84-85	0.67	+	8	4	PROBABLE3-HYDROXYISOBUTYRATEDE HYDROGENASEMMSB, PROBABLEACYL-CoADEHYDROGENASEFADE9, PROBABLEMETHYLMALONATE-SEMIA LDEHYDEDEHYDROGENASEMMSA, PE-PGRSFAMILYPROTEIN
92-93	0.71	+	7	2	PROBABLETRANSCRIPTIONALREGULA TORYPROTEIN, POSSIBLEDEAMINASE, POSSIBLETRANSPOSASE(FRAGME NT), PE-PGRSFAMILYPROTEIN, PE-PGRSFAMILYPROTEIN
96-97	0.69	+	10	2	molybdenumcofactorbiosynthesisprotein, PROB ABLEMOLYBDOPTERINBIOSYNTHESISMO G, PROBABLEMOLYBDENUMCOFACTORB IOSYNTHESIS, POSSIBLERESUSCITATION-PROMOTINGFACTORRPFA, PROBABLEM OLYBDENUMCOFACTORBIOSYNTHESIS, molybdenumcofactorbiosynthesisprotein, POSS IBLECONSERVEDINTEGRALMEMBRANE, PROBABLECOLDSHOCK-LIKEPROTEIN
97-98	0.66	+	8	2	PROBABLEACYL-CoADEHYDROGENASEFADE10, POSSIBLEC ONSERVEDEXPORTEDPROTEIN, POSSIBLE CONSERVEDTRANSMEMBRANEPROTEIN, PPEFAMILYPROTEIN, POSSIBLECONSERVEDTR ANSMEMBRANEPROTEIN, POSSIBLETRANSCRI PTIONALREGULATORYPROTEIN
109-110	0.69	+	7	1	PE-PGRSFAMILYPROTEIN, PE-PGRSFAMILYPROTEIN, 50SribosomalproteinL32, PE-PGRSFAMILYPROTEIN, MYCOBACTERIALP ERSISTENCEREGULATORMRPA, PROBABLETW OCOMPONENTSENSOR
121-122	0.71	+	8	1	SHORT(C15)CHAINZ-ISOPRENYL, PE-PGRSFAMILYPROTEIN, PEFAMILYPROTEIN, PEFAMILYPROTEIN, PROBABLECELLULASE CELA2A(ENDO-1,4-BETA-GLUCANASE), PRO BABLECELLULASECELA2B(ENDO-1,4-BETA-GLUCANASE), PE-PGRSFAMILYPROTEIN
126-127	0.66	+	10	3	PPEFAMILYPROTEIN, POSSIBLEACETYL-CoAACETYLTRANSFERASE(ACETOACETY L-CoA, POSSIBLEENOYL-CoAHYDRATASE, POSSIBLEOXIDOREDUCTASE, PROBABLEINTE GRALMEMBRANEPROTEIN, enoyl-CoAhydratase, enoyl-CoAhydratase
163-164	0.73	+	5	1	PE_PGRS (15608588), cytochrome c oxidase assembly factor (ctaB) (15608589), PE_PGRS (15608590), quinone oxidoreductase (qor) (15608592)
198-199	0.63	+	7	2	PPEFAMILYPROTEIN, PROBABLEPHOSPHOLIPASEC4, PUTATIVETRANSPOSASE, PUTATIVETRANSPOSASE, PROBABLECUTINASECUT1
216-217	0.61	+	3	0	PROBABLEISOCITRATELYASEaceAa, isocitratelyase, PPEFAMILYPROTEIN
229-230	0.67	+	2	1	ProbablelipoproteinlppI

263-264	0.63	+	6	0	PROBABLEMEMBRANE-ASSOCIATEDPHOSPHOLIPASEC, PPEFAMILYPROTEIN, PPEFAMILYPROTEIN, PROBABLETRANSPOSASE, PROBABLETRANSPOSASE, PPEFAMILYPROTEIN
280-281	0.70	+	7	4	HYPOTHETICALALANINERICHPROTEIN, PE-PGRSFAMILYPROTEIN, dihydrolipoamideacetyltransferase
284-285	0.69	+	1	0	PROBABLEFATTYACIDSYNTHASE
293-294	0.66	+	11	4	PROBABLEACYL-CoATHIOESTERASEII, pyridoxinebiosynthesisprotein, pyridoxamine5'-phosphateoxidase, PPEFAMILYPROTEIN, PROBABLECONSERVEDMEMBRANEPROTEIN, ALPHA-MANNOSYLTRANSFERASEPIMA, lipidAbiosynthesislauroyl
316-317	0.66	+	7	2	malate:quinoneoxidoreductase, PE-PGRSFAMILYPROTEIN, dihydrolipoamidedehydrogenase, POSSIBLENICKEL-TRANSPORTINTEGRALMEMBRANE, shortchaindehydrogenase
320-321	0.68	+	9	4	PPEFAMILYPROTEIN, POSSIBLEOXIDOREDUCTASE, tyrosinerecombinase, POSSIBLEMYCOBACTINUTILIZATIONPROTEIN, formatedehydrogenaseaccessoryprotein
337-338	0.67	+	10	2	aspartyl/glutamyl-tRNAamidotransferasesubunitC, DNAligase, PROBABLELIPOPROTEINLPQA, ESAT-6LIKEPROTEINESXQ, PPEFAMILYPROTEIN, PEFAMILYPROTEIN, SECRETEDESAT-6LIKEPROTEIN, ESAT-6LIKEPROTEINESXS
373-374	0.67	+	2	0	PE_PGRS (15610480), PE_PGRS (15610481)
374-375	0.67	+	2	1	PPE (15610483)
375-376	0.65	+	3	2	PPE (15610486)
376-377	0.65	+	6	5	methylenetetrahydrofolate dehydrogenase (folD) (15610492)
380-381	0.69	+	8	0	POSSIBLETRANSPOSASE, POSSIBLETRANSPOSASE, PE-PGRSFAMILYPROTEIN, POSSIBLEDEHYDROGENASE, PROBABLECONSERVEDLIPOPROTEINLPQD, shortchaindehydrogenase, CYCLOPROPANE-FATTY-ACYL-PHOSPHOLIPIDSYNTHASE1CMAA1, PROBABLENUCLEOSIDEHYDROLASEIUNH
392-393	0.70	+	6	0	CONSERVEDHYPOTHETICALINTEGRALMEMBRANE, 3-ketoacyl-(acyl-carrier-protein)reductase, PROBABLEFERREDOXINFDXD, PROBABLEACYL-CoADEHYDROGENASEFADE26, PROBABLEACYL-CoADEHYDROGENASEFADE27, acyl-CoAsynthase
393-394	0.75	+	4	1	PE_PGRS (15610644), probable acetohydroxyacid synthase I large subunit (ilvX) (15610645), PE_PGRS(15610647)

394-395	0.79	+	3	0	PE_PGRS (15610648), acyl-CoA synthase (fadD18) (15610649), PE_PGRS (15610650)
397-398	0.66	+	7	0	PPEFAMILYPROTEIN, 4-hydroxy-2-ketovaleratealdolase, acetaldehydedehydrogenase, PROBABLEHYDRATASE, 3-ketosteroid-delta-1-dehydrogenase, PROBABLEDEHYDROGENASE, PPEFAMILYPROTEIN
403-404	0.69	+	7	2	PROBABLEADENINEGLYCOSYLASEMUTY, PE-PGRSFAMILYPROTEIN, POSSIBLEHYDROLASE, PROBABLECONSERVEDLIPOPROTEINLPQF, PE-PGRSFAMILYPROTEIN

Supplementary Material 2

The distribution of microsatellite motifs of sizes di-hexa according to their GC content.

GC% of the repeat motif	Mycobacterium avium	Mycobacterium bovis	Mycobacterium tuberculosis H37Rv	Mycobacterium tuberculosis CDC 1551	Mycobacterium leprae
			Di		
0.00	1637	4164	4227	4229	7707
0.50	60293	62571	63590	63442	59404
1.00	101590	71466	72409	72362	32346
			Tri		
0.00	124	216	218	217	1095
0.17	0	0	0	1	0
0.33	12451	12359	12566	12537	13386
0.67	54113	41371	41822	41900	26829
1.00	47159	34623	35062	34911	10845
			Tetra		
0.00	12	21	21	21	142
0.25	267	495	493	495	1128
0.50	1805	2516	2533	2539	3062
0.75	5526	5721	5810	5794	4002
1.00	8387	4481	4524	4516	1119
			Penta		
0.00	1	0	0	0	42
0.20	16	22	23	24	36
0.40	186	274	295	281	548
0.60	840	1081	1137	1117	1139
0.80	2157	1809	1828	1832	902
1.00	2920	1170	1183	1180	271
			Hexa		
0.00	0	0	0	0	13
0.17	1	5	5	5	62
0.33	43	53	52	52	151
0.50	453	413	414	416	339
0.67	1295	862	877	873	427
0.83	1941	1051	1045	1058	288
1.00	729	364	375	367	48