 10.1111/j.1469-0691.2006.01637.x

Molecular evolution of *Mycobacterium tuberculosis*

C. Arnold

Applied and Functional Genomics, Centre for Infections, Health Protection Agency, London, UK

ABSTRACT

Tuberculosis continues to be the main cause of death from a single infectious agent in developing countries. The causative agent, *Mycobacterium tuberculosis*, is thought to have diverged from its common ancestor as recently as 15 000 years ago. Subsequently, various genetic elements have evolved over time at different rates and can be used to elucidate patterns of infection. When individual elements are studied within genetic families, very low rates of variation are observed for almost every marker. For example, when all *M. tuberculosis* genetic families are considered, the number of alleles observed at each mycobacterial interspersed repetitive unit (MIRU) locus usually drops when viewed within a single genetic family, indicating that the rate of repeat variation may be low, as each member of that family is a descendant of a single common ancestor. Also, the low level of silent nucleotide variation observed indicates that *M. tuberculosis* is, in evolutionary terms, very young. Mapping the variation of the different markers used in molecular epidemiology within a genetic framework enables the relative rates of variation of these markers to be determined and, together with a complete chronology, allows the identification of more informative panels of markers tailored to individual genetic families.

Keywords Evolution, genetic variation, molecular clock, molecular epidemiology, *Mycobacterium tuberculosis*, review

Accepted: 4 September 2006

Clin Microbiol Infect 2007; **13**: 120–128

INTRODUCTION

Although a cure for tuberculosis (TB) was developed >50 years ago, it still remains one of the world's deadliest infectious diseases. The WHO reports that TB kills 5000 people a day, and between two and three million people annually, 98% of whom live in the developing world (http://www.who.int/mediacentre/factsheets/fs104/en/). Approximately one-third of the world's population is infected with TB, and hundreds of thousands of children will become TB orphans this year. One in three patients infected with human immunodeficiency virus (HIV) or with AIDS has TB, and drug resistance is emerging at an alarming rate in some geographical areas. Effective treatment is available for the majority of cases, but the disease often goes undiagnosed. The standard test for *Mycobacterium tuberculosis* infection is culture; however, the organism can take several weeks, or even months, to grow on solid culture. The advent of rapid liquid culture methods has reduced the time required to confirm a diagnosis to just 1–2 weeks.

Currently, the fastest method for confirming *M. tuberculosis* infection is a sputum smear test, stained for acid-fast bacilli, which can be carried out in just a few hours. However, a bacterial load of 10 000 cells/mL in a sputum sample is required for detection using this method, and >50% of 'smear-negative' patients are found subsequently to be culture-positive. Different strain-typing methods have been employed to link epidemiologically-related strains in order to follow and better understand the evolution and spread of this disease, and to perform the more difficult task of differentiating epidemiologically-unrelated strains of this organism.

The molecular clocks of genetic markers exploited to study the spread of *M. tuberculosis* differ, and can be used to investigate molecular evolution over shorter or longer periods. Study of the patterns of

Corresponding author and reprint requests: C. Arnold, Applied and Functional Genomics, Centre for Infections, Health Protection Agency, 61 Colindale Avenue, London NW9 5EQ, UK
E-mail: catherine.arnold@hpa.org.uk

variation and evolution of different genetic markers gives insight into their usefulness for different applications, and analysis of the way in which this genetic variation occurs provides further information concerning their efficacy for epidemiological purposes. This review considers some of the genetic variation strategies employed by *M. tuberculosis*, and the effect of these strategies on the relative molecular clocks of some of the commonly used markers within an evolutionary framework.

ORIGIN OF *M. TUBERCULOSIS*

Mycobacteria can be divided into two groups: fast-growing and slow-growing. Sequence data from various genes, including 16S rRNA, *rpo*B and *hsp*65 genes, have been used to construct phylogenetic relationships [1,2]. One of the sequenced strains of *M. tuberculosis*, H37Rv, has been shown to contain 20 cytochrome P450-containing mono-oxygenases that catalyse mixed oxidation of hydrophobic compounds; this is an activity that is associated with free-living saprophytes in soil, which perhaps indicates that the ancestor of *M. tuberculosis* was a soil mycobacterium [3].

M. tuberculosis is part of the *M. tuberculosis* complex (MTBC), a group of closely-related slow-growing mycobacteria that includes *M. tuberculosis*, *Mycobacterium bovis*, *Mycobacterium africanum*, *Mycobacterium microti* and *Mycobacterium canettii*. Until recently, it was assumed that cattle transmitted the disease to man, as the host range of *M. bovis* was much broader than that of *M. tuberculosis*. However, genetic information has revealed that the reverse is the case [4]. Gutierrez *et al*. [5] used sequence analysis of seven genes to show that *M. tuberculosis* appears to be a composite assembly of a relatively diverse group of smooth tubercle bacilli, including *M. canettii* strains. However, the generally low level of genetic variation seen in *M. tuberculosis* indicates that the total population resulted from clonal expansion following an evolutionary bottleneck, estimated to have occurred between 15 000 and 35 000 years ago [4–7].

GENERAL PRINCIPLES OF BACTERIAL EVOLUTION

There is significant evolutionary pressure towards smaller bacterial genomes, as smaller chromosomes can replicate faster, resulting in the bacteria being able to grow faster and out-populate bacteria with larger chromosomes. If a bacterial population moves into an environment where, for example, an essential amino-acid is abundant, some members of the bacterial population, after many generations, will lose the ability to synthesise that amino-acid. Eventually, every member of the population will also lose by deletion the DNA sequence encoding the pathway necessary for producing that amino-acid. However, this process occurs extremely slowly, as the evolutionary pressure to decrease genome size needs to also protect the organism from becoming extinct by degeneration of its genome [8].

Within the general premise that bacterial species change in response to their environment, it is clear that there are different molecular mechanisms for growth-dependent mutation and adaptive or stationary-phase mutation [9,10]. When growing cells are confronted with a change in their environment, they still have metabolic capabilities for a specific compensatory response (e.g., increased levels of transcription and mutation). However, after 4–5 days with no exogenous source of energy, cells resort to non-specific increases in all mutation rates in a final effort to produce a mutant that will survive. Mutation rates will also be affected by other factors, e.g., oxidising agents, UV light, the activity of DNA repair enzymes, and variables that could be influenced by starvation conditions (e.g., nucleotide pool levels). In a study of 26 structural genes of 842 *M. tuberculosis* isolates, >95% of nucleotide substitutions caused amino-acid substitutions in genes linked to antibiotic resistance [7]. A particular mutation arises independently in response to challenge with isoniazid, the front-line drug for tuberculosis treatment, more often than is expected by random mutation. Oxygen limitation induces dramatic and specific changes in mycobacteria, including enhanced resistance to isoniazid [11]. Thus, it appears that the rate at which mutations occur is elevated under conditions in which an excess of variation is most needed, so the question arises as to whether mechanisms have evolved to regulate the mutation rate. Most molecular processes leading to spontaneous and induced mutagenesis depend on the action of particular enzymes [12]. An accumulation of DNA lesions in cells under stress may lead to the saturation of DNA repair systems, which would lead to an increased mutation rate, as

would down-regulation of DNA repair systems in static cells and the inactivation of DNA repair enzymes by mutation. Other generators of genetic variations also exist, e.g., IS elements, some of which carry outward promoters that serve no regulatory function for genes internal to the element [13].

STRATEGIES FOR GENERATION OF GENETIC VARIATION

Genetic variation analysis, or genotyping, is used to track different strains of *M. tuberculosis*, both nationally and internationally, in order to obtain information concerning the patterns of spread, infectivity and pathogenicity. The molecular clocks of the various elements used as markers are different, and can be used to look at change over shorter or longer periods of time. A marker with a fast molecular clock would be required to determine whether an infection is a re-activation of an old infection, while a marker with a slow molecular clock would be required to monitor evolution over tens of thousands of years. Evolutionary clocks estimating evolutionary distance are often based on local sequence changes within a specific gene. Within functional genes, lethal and heavily contra-selective mutations will not be maintained. Local sequence changes are known to affect different DNA regions with different efficiencies, thus limiting the precision of evolutionary clocks. Study of the patterns of variation and evolution of different genetic markers gives considerable insight into their usefulness for different applications, and analysis of the way in which this genetic variation occurs can further improve their efficacy for epidemiological purposes by improving their precision.

Overall, genetic variation generation strategies in bacteria can be divided into three categories [12]: (i) small local changes in nucleotide sequence of the genome, e.g., single nucleotide polymorphisms (SNPs); (ii) intra-genomic rearrangements of segments of genomic sequence, e.g., recombination between repeat sequences and deletions; and (iii) acquisition of DNA sequences from other organisms.

Small local changes in nucleotide sequence

This strategy can be viewed as incremental improvement of already available biological functions and, as such, development of completely new biological functions is unlikely. Silent mutations (synonymous or non-coding) in DNA are very useful for evolutionary study as they exert little or no selective pressure on the organism and thus have the ability to be maintained through many generations, enabling tracing of related organisms.

Intra-genomic rearrangements

Intra-genomic rearrangements can result in gene conversion (or non-reciprocal transfer) if related sequences in the genome undergo recombination. This process can result in novel combinations of available capacities by the fusion of different functional domains. Reassortment of expression control signals, involving different reading frames concerned with protein production, can result in fitter organisms in different environmental niches. Duplication of DNA segments as a result of rearrangement may also serve as substrates for further evolution [14]. Deletion of DNA segments can also occur, and can help, in conjunction with natural selection, in removing non-essential sequences from the genome. However, the evolutionary pressure to reduce genome size acts as an obstacle to novel protein evolution. Most mutational alterations to proteins result in a decrease in protein function, leading to cell death. For a protein to evolve a novel function, it would probably need to develop through stages in which it was useless, but could then be deleted from the genome. In eukaryotes, protein evolution occurs by gene duplication, followed by mutation to the duplicate gene. However, prokaryotes have few or no proteins that do not have a function, as such proteins would most likely be deleted.

Acquisition of DNA sequences from other organisms

Acquisition of DNA from other organisms is probably the most common way in which bacteria gain the ability to produce novel proteins. Horizontal transfer is common among bacteria, and is thought to be a widespread evolutionary process [15].

The theory that specialist organisms evolved from multifunctional bacteria with larger genomes by deletion mutations may have some validity, but it is likely that horizontal transfer

also plays a large part. However, for *M. tuberculosis*, horizontal exchange is thought to occur rarely, if at all [7,16], as evidenced by spoligotyping and deletion analysis (see below), resulting in the loss of an important route of genetic variation. Before the emergence of *M. tuberculosis*, horizontal exchange was thought to be more frequent, and the successful progenitor clone is thought to be a chimeric genome, composed of a variety of *M. canettii* strains [5].

In the following sections, specific examples of the two remaining variation strategies employed by *M. tuberculosis* will be discussed in more detail.

SMALL LOCAL CHANGES IN NUCLEOTIDE SEQUENCE IN *M. TUBERCULOSIS*

Synonymous SNPs in *M. tuberculosis* have been used for evolutionary studies, as they are less subject to selective pressure than other genetic markers [16,17]. Non-synonymous, or coding, SNPs identified by Sreevatsan *et al*. [7] have also been used to genotype *M. tuberculosis*. Sreevatsan *et al*. studied 26 structural genes of 842 isolates and suggested a broad evolutionary scenario for MTBC organisms, characterised by *katG* codon 463 and *gyrA* codon 95, in which *M. tuberculosis* could be split into three major genetic groups (MGGs), with an evolutionary bottleneck approximately 15 000 to 20 000 years ago, possibly around the time of speciation of *M. tuberculosis*. A recent SNP-based study by Filliol *et al*. [6] supported these MGGs and split them further, confirming that the species *M. tuberculosis* consists of several very distinct strain families.

Whole genome comparison of *M. tuberculosis* strains CDC1551 and H37Rv (MGG2 and MGG3, respectively) by Fleischmann *et al*. [18] revealed 1075 SNPs, with *c*. 85% of the substitutions occurring in coding regions (95% of the genome). Transitions (purine to purine, and pyrimidine to pyrimidine) were more common (61%) than transversions, and the ratio of synonymous to non-synonymous substitutions (D_s/D_n) in *M. tuberculosis* was *c*. 1.6, much lower than expected when compared with *Escherichia coli* and *Salmonella*, indicating that there is either additional selective pressure on synonymous substitutions in *M. tuberculosis*, or decreased selective pressure against non-synonymous mutations. This restricted level of silent nucleotide variation is considerably less than that observed in other pathogenic bacteria (including other mycobacteria and *Neisseria meningitidis*) that are strict host specialists. Ecological host specialisation alone does not account for the restricted genetic variation observed.

Mokrousov *et al*. [19] also indicated a C > T bias for mutations in the *rpoB* gene that conferred rifampicin resistance. Mismatch–repair genes were not found by Cole *et al*. [20] in *M. tuberculosis*, but several copies of the *mutT* gene were present, coding for a protein that removes oxidised guanines, which counteracts replication or transcription errors. Perhaps a mismatch repair system is not required in *M. tuberculosis* but, as described above, inactivation or down-regulation of some *mutT* genes would lead to the rate of mutation increasing. The two *rpoB* mutations described most frequently are both C > T transitions, which occur by spontaneous cytosine deamination to uracil; as a G + C rich organism, *M. tuberculosis* is at higher risk for cytosine deamination [19].

INTRA-GENOMIC REARRANGEMENTS

Transposable elements: IS*6110*

Transposable, or mobile, elements are common in the genomes of all plants and animals, as well as bacteria, where they are called insertion sequences (IS) [21]. These are DNA sequences that have the ability to integrate into the genome at a new site by a 'cut and paste' mechanism, whereby the element is generally cut out from one site and inserted into a new location ('hotspot') on the chromosome, resulting, potentially, in a disrupted gene. Active transposons produce a transposase enzyme encoded by a gene located between inverted–repeat termini.

IS*6110* DNA fingerprinting is the genotyping technique used most widely for *M. tuberculosis* [22]. Whole genome digestion, followed by hybridisation with an IS*6110*-based probe reveals differential patterns based on the IS*6110* copy number (up to 25 copies per genome) and location within the genome. These patterns may be very different between epidemiologically unlinked isolates, but the banding patterns of serial isolates from an individual are relatively stable over the period of the disease [23]. Warren *et al*. [24] reviewed the literature and concluded that

IS*6110* patterns were used primarily to answer epidemiological questions, and that their relationship in an evolutionary context had not been fully addressed. In addition, the two groupings of *M. tuberculosis* strains, i.e., low copy (≤ five copies of IS*6110*) or high copy (> five copies IS*6110*), did not reveal clear evolutionary patterns. It was also unclear whether low copy-number isolates represented a single evolutionary lineage, or whether they had evolved independently and demonstrated similar patterns due to transposition of IS*6110* into preferential integration sites [25].

Variable number tandem repeat (VNTR) variation

Whole genome comparison has also revealed the presence of short sequence repeats, termed mycobacterial interspersed repeat units (MIRUs [26]) and variable number tandem repeat units (VNTRs [27]), which are tandemly repeated sequences of 40–100 bp. The location and number of repeats varies in different *M. tuberculosis* strains and can be measured by a variety of PCR-based methods. The data generated are portable between laboratories [28]. Ferdinand *et al*. [29] described the use of particular loci to classify isolates of the *M. tuberculosis* complex into different genetic families, including the East African Indian (EAI), Beijing, Haarlem and X, Latin-American and Mediterranean (LAM) families. The discriminatory power of this technique depends on the number and set of loci used, and there is evidence that this may depend on the genetic family to which the isolates being examined belong [30]. The genesis of some of these repeats in *M. tuberculosis*, which have highly similar sequences, may have started with an initial 53-bp single copy repeat, which then spread to different loci throughout the genome by recombination. An alignment of repeat sequences from the published sequence of *M. tuberculosis* CDC1551 [18], split into their individual units (Fig. 1), shows that the repeats at each locus have very similar sequences, although each repeat consensus differs slightly from that of other loci. It is possible that, once a single copy of the sequence spreads throughout the genome, slight changes in sequence occurred at these different loci [30], which then started to duplicate themselves, as sequence differences in the first repeat are also duplicated in subsequent repeats in a process described by Benson *et al*. [31]. The presence of only one copy of this 53-bp repeat in *Mycobacterium leprae* (Fig. 1) and two copies in *Mycobacterium avium* supports this hypothesis.

```
                                  10        20        30        40        50
                         ....|....|....|....|....|....|....|....|....|....|....|...
MIRU10 repeat 1          ATGGCGCCGCTCCTCCTCATCGCTGCGCTCTGCATCGTCGCCGGCGGTAGTTA
MIRU 10 repeat 2         ..................................................C.
MIRU 10 repeat 3         ..................................................C.
MIRU 10 repeat 4         ..................................................C.
MIRU23 repeat 1          --.....................T..............A......CG.C.C.
MIRU 23 repeat 2         TCT....................T..............A......CG.C.C.
MIRU 23 repeat 3         TCT........G............T..............A......CG.C.C.
MIRU 23 repeat 4         TCT........G............T..............A......CG.C.C.
MIRU26 repeat 1          -AA.......................................AG..C.
MIRU 26 repeat 2         --A.......................................AG..C.
MIRU26 repeat 3          --A.......................................AG..C.
MIRU26 repeat 4          --A.......................................AG..C.
MIRU16 repeat 1          -----.................T...................CGGT.C.
MIRU16 repeat 2          CGA...................T...................CGGT.C.
ETRE repeat 1            ----------.....................................CC.ACC.
ETRE repeat 2            TCT.......................................CC.ACC.
MIRU2 repeat 1           -A.....................T....G..............CG...C.
MIRU 2 repeat 2          TA.....................T....G...........T....CG...C.
MIRU39 repeat 1          ---.....................T.................CGG.CCG
MIRU 39 repeat 2         T.........................T...............CGG..C.
MIRU24 (rev comp)        ..............AG.........................C
QUB repeat 2             C..............C.......TT..................CGG..C.CTGGC
QUB5 repeat 3            C..............C.......TT..................CGG..C.
QUB repeat 4             C..............C.......TT..................CGG..C.ATC
M. avium AE017235 nt271034 --.................G.....T....G...G...C....C...CGG..C.
M. avium AE017238 nt39493  --.............C.......A..............T...G.CGG
M. leprae MLU15181 nt22940 --......A..............TA....C.......A.......CGG
SenX3-RegX3 53-mer (MIRU4) TCA.....A..............A..............T.....CAG
```

Fig. 1. Alignment of individual repeats found at various loci in *Mycobacterium tuberculosis* CDC1551 and similar sequences from *Mycobacterium avium* and *Mycobacterium leprae*. Bases identical to the first sequence are represented by a dot and missing bases are represented by a dash (reproduced with permission from [31]).

Direct repeat (DR) analysis: spoligotyping

Spoligotyping is a widely used PCR-based reverse-hybridisation blotting technique that detects the presence or absence of 43 unique short sequences, or direct variable repeats (DVRs) in the direct repeat locus. Spacer regions separate the unique sequences with identical sequences, enabling the use of PCR targeting the spacers to amplify and test small amounts of template. Spoligotyping has been shown to be useful in the clinical laboratory, as well as for molecular epidemiology and population genetics. Being PCR-based, spoligotyping is robust, inexpensive and produces digital (and therefore portable) numerical results [32]. Spoligotyping data can also be used to elucidate evolution, based on the assumption that the DVRs in the DR region analysed can only be lost (together or in groups) following movement of the IS*6110* element, and cannot be regained as little or no recombination appears to take place between strains [33] and there appears to be no rearrangement within the DR (i.e., the DVRs are always in the same order). Genetic convergence has been demonstrated, and, although thought to be rare, this possibility should be borne in mind when using spoligotyping for evolution studies [32].

Deletion analysis

Brosch *et al*. [4] revealed the distribution of 20 variable regions in the genomes of the MTBC, and showed that the majority of these polymorphisms did not occur independently, but resulted from ancient, irreversible genetic events in common progenitor strains. Based on the presence or absence of an *M. tuberculosis*-specific deletion (TbD1), strains can be divided into so-called 'ancestral' and 'modern' strains, e.g., the Beijing and Haarlem *M. tuberculosis* clusters. The presence of TbD1 also correlates with the presence of two copies of MIRU 24 [29,34,35]. It is important to clearly define what is meant by 'ancestral', and, in this context, such strains are more similar to the common ancestor at the loci examined. Successive loss of DNA was identified in an evolutionary lineage represented by *M. africanum*, *M. microti* and *M. bovis* strains that diverged from the progenitor of today's *M. tuberculosis* strains before the deletion of TbD1 occurred.

Until recently, it was assumed that cattle originally transmitted the disease to man, as the host range of *M. bovis* was much broader than that of *M. tuberculosis*. The findings of Brosch *et al*. argue against this hypothesis as *M. canettii* and ancestral *M. tuberculosis* strains showed no loss of the specific regions identified in the study, and therefore seem to be direct descendants of tubercle bacilli that existed before the *M. africanum/M. bovis* lineage separated from the *M. tuberculosis* lineage [4].

If data from the evolutionarily informative markers described above (SNPs, spoligotype, deletion analysis and some VNTR data) are combined for a large collection of diverse strains, an evolutionary lineage can be created (Fig. 2) [30]. A data-mining study, which analysed evolutionarily informative markers, split VNTR profiles into two groups for analysis (ancestral and modern) according to whether two copies of MIRU 24 were present. The use of SNPs at *katG463* and *gyrA95* further divided strains into major MGGs 1–3, whereby MGG1 is thought to be evolutionarily older [7]. *M. tuberculosis* strains belonging to MGG1 that still have TbD1 also retain two copies of MIRU 24 and represent ancestral strains. MGG1 strains lacking TbD1 and with a single copy of MIRU 24 are considered to be more 'modern'. Once separated into ancestral and modern strains, spoligotyping data were then used to further map the genetic evolution on the timeline, based on the assumption that the DVRs in the DR region analysed can only be lost.

Following the combination and mapping of these data, it was possible to superimpose the mean numbers of repeats from VNTR/MIRU profiles on the timeline to investigate patterns of their evolution. Among 121 ancestral isolates analysed, 80 unique VNTR profiles were found, and among 369 isolates representing two of the modern lineages (Beijing and Haarlem), 180 unique VNTR profiles were found. In total, the averaged VNTR repeats were calculated from a dataset containing 478 isolates, representing ancestral and modern isolates from nine different studies. VNTR variation is thought to occur in both directions (i.e., both increases and decreases in the number of repeats), by a mechanism called slipped-strand mispairing (SSM) [36]. The number of repeats is thought to increase during replication if there is slippage in the replicated

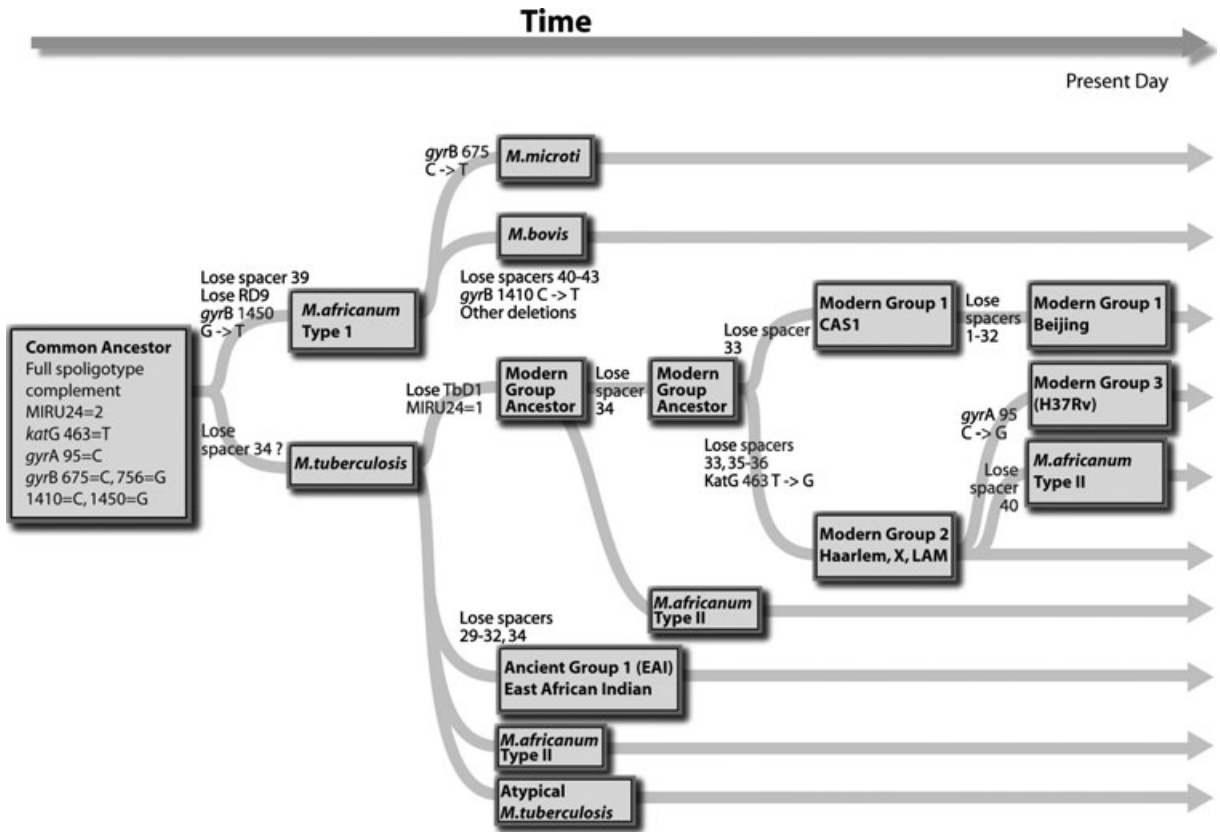

Fig. 2. Evolutionary history of selected genetic markers in *Mycobacterium tuberculosis* (reproduced with permission from [30]).

strand, and to decrease if it occurs in the template strand. However, data-mining suggests that the loss of repeats is more likely in modern strains than in ancestral strains [30].

This kind of mapping exercise demonstrates the relative molecular clocks of these widely used genetic markers in *M. tuberculosis*. Sequencing the DR region from a group of closely related isolates identified variants that enabled Fang *et al*. [37] to calculate the evolutionary rate of change over several hundred years. A similar slow evolutionary rate was observed upon examination of IS*6110* transposition, with only four transposition events occurring during this time. However, this differs from the majority of reports that identify IS*6110* transposition as being a much more frequent event than change in the DR. The discriminatory power of VNTR variation analysis is thought to approach that of IS*6110* RFLP typing, a method with a fast molecular clock [38], but only when low copy number IS*6110* isolates (often ancestral) are included in the analysis.

When VNTR pattern variation is examined among and within genetic families, clearer pictures emerge concerning the variation at different loci [30]. For example, MIRU 26 is the most variable locus in modern strains, but appears to be stable in ancestral strains where it has a low copy number. Analysing these data within a framework of genetic families indicates that repeat variation may be occurring much more slowly than previously thought.

Speculation concerning putative dates of events in the evolutionary history of *M. tuberculosis* is more difficult. As described above, Sreevatsan *et al*. [7] indicated the possibility of an evolutionary bottleneck that occurred *c*. 15 000–20 000 years ago, perhaps around the time of speciation of *M. tuberculosis*, while supposed modern strains have existed for perhaps 4000 years or more [39]. This hypothesis is supported by the presence or absence of spacers typical for MGG2 and MGG3 in the spoligotype of strains from Egyptian mummies. Mycobacterial

DNA isolated from a 17 000-year-old bison skeleton recovered from a natural trap cave revealed the presence of DVRs or spacers not present in modern *M. bovis*, which suggests that the animal was infected with a precursor around the time of speciation [40], in line with the estimate of Sreevatsan *et al*. [7].

CONCLUSIONS

Different genetic markers indicate that the species *M. tuberculosis* consists of very distinct strain families, each with a single ancestor. An analysis of genetic marker deviation from the common ancestor of each of these families will improve understanding of the individual marker evolution rates and the degree of convergence within genetic families. For accurate estimation of molecular clocks in more contemporaneous isolates, the combined power of multiple genetic markers should be harnessed. The ability to study epidemiologically-unrelated groups of isolates, that are identical using some or all of the methods described, within an evolutionary framework, will provide further clues concerning the rate of evolution of these individual markers, and will add to knowledge of the molecular mechanisms of evolution in *M. tuberculosis*. This, in turn, will have a significant impact on epidemiological studies, enabling the molecular identification of outbreaks rather than molecular confirmation.

ACKNOWLEDGEMENTS

The author would like to thank N. Thorne, A. Underwood and, especially, S. Gharbia for valuable discussions.

REFERENCES

1. Devulder G, Perouse de Montclos M, Flandrois J. A multigene approach to phylogenetic analysis using the genus *Mycobacterium* as a model. *Int J Syst Evol Microbiol* 2005; **55**: 293–302.
2. Adekambi T, Drancourt M. Dissection of phylogenetic relationships among 19 rapidly growing *Mycobacterium* species by 16S rRNA, *hsp65*, *sodA recA* and *rpoB* gene sequencing. *Int J Syst Evol Microbiol* 2004; **54**: 2095–2105.
3. Cole S. Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv. *FEBS Lett* 1999; **452**: 7–10.
4. Brosch R, Gordon S, Marmiesse M *et al.* A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc Natl Acad Sci USA* 2002; **99**: 3684–3689.
5. Gutierrez M, Brisse S, Brosch R *et al.* Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. *PLoS Pathog* 2005; **1**: 55–61.
6. Filliol I, Motiwala A, Cavatore M *et al.* Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *J Bacteriol* 2006; **188**: 759–772.
7. Sreevatsan S, Pan X, Stockbauer K *et al.* Restricted gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci USA* 1997; **94**: 9869–9874.
8. Maniloff J. The minimal cell genome: 'On being the right size'. *Proc Natl Acad Sci USA* 1996; **93**: 10004–10006.
9. Saint-Ruf C, Matic I. Environmental tuning of mutation rates. *Environ Microbiol* 2006; **8**: 193–199.
10. Zinser E, Kolter R. *Escherichia coli* evolution during stationary phase. *Res Microbiol* 2004; **155**: 326–328.
11. Wayne L, Sramek H. Metronidazole is bactericidal to dormant cells of *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* 1994; **38**: 2054–2058.
12. Arber W. Elements for a theory of molecular evolution. *Gene* 2003; **317**: 3–11.
13. Safi H, Barnes P, Lakey D *et al.* IS*6110* functions as a mobile, monocyte-activated promoter in *Mycobacterium tuberculosis*. *Mol Microbiol* 2004; **52**: 999–1012.
14. Yamanaka K, Fang L, Inouye M. The CspA family in *Escherichia coli*: multiple gene duplications for stress adaptation. *Mol Microbiol* 1998; **27**: 247–255.
15. Lorenz M, Wackernagel W. Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol Rev* 1994; **58**: 563–602.
16. Gutacker MM, Smoot J, Lux Migliaccio C *et al.* Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics* 2002; **162**: 1533–1543.
17. Baker L, Brown T, Maiden M *et al.* Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. *Emerg Infect Dis* 2004; **10**: 1568–1577.
18. Fleischmann R, Alland D, Eisen J *et al.* Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J Bacteriol* 2002; **184**: 5479–5490.
19. Mokrousov I. Multiple *rpoB* mutants of *Mycobacterium tuberculosis* and second-order selection. *Emerg Infect Dis* 2004; **10**: 1337–1338.
20. Cole ST. Comparative mycobacterial genomics. *Curr Opin Microbiol* 1998; **1**: 567–571.
21. Kazazian HH. Mobile elements. drivers of genome evolution. *Science* 2004; **303**: 1626–1632.
22. Thierry D, Cave M, Eisenach K *et al.* IS*6110*, an IS-like element of *Mycobacterium tuberculosis* complex. *Nucleic Acids Res* 1990; **18**: 188.
23. Cave M, Eisenach K, Templeton G *et al.* Stability of DNA fingerprint pattern produced with IS*6110* in strains of *Mycobacterium tuberculosis*. *J Clin Microbiol* 1994; **32**: 262–266.
24. Warren R, Richardson M, Sampson S *et al.* Molecular evolution of *Mycobacterium tuberculosis*: phylogenetic reconstruction of clonal expansion. *Tuberculosis* 2001; **81**: 291–302.
25. Fomukong N, Beggs M, el Hajj H *et al.* Differences in the prevalence of IS*6110* insertion sites in *Mycobacterium tuberculosis* strains: low and high copy number of IS*6110*. *Tuber Lung Dis* 1997; **78**: 109–116.

26. Supply P, Mazars E, Lesjean S *et al.* Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. *Mol Microbiol* 2000; **36**: 762–771.
27. Frothingham R, Meeker-O'Connell W. Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. *Microbiology* 1998; **144**: 1189–1196.
28. Mazars E, Lesjean S, Banuls A *et al.* High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. *Proc Natl Acad Sci USA* 2001; **98**: 1901–1906.
29. Ferdinand S, Valetudie G, Sola C *et al.* Data mining of *Mycobacterium tuberculosis* complex genotyping results using mycobacterial interspersed repetitive units validate the clonal structure of spoligotyping-defined families. *Res Microbiol* 2004; **155**: 647–654.
30. Arnold C, Thorne N, Underwood A *et al.* Evolution of short sequence repeats in *Mycobacterium tuberculosis*. *FEMS Microbiol Lett* 2006; **256**: 340–346.
31. Benson G, Dong L. Reconstructing the duplication history of a tandem repeat. *Proc Int Conf Intell Syst Mol Biol* 1999; 44–53.
32. Brudey K, Driscoll J, Rigouts L *et al. Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiol* 2006; **6**: 23.
33. van Embden JDA, van Gorkom T, Kremer K *et al.* Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. *J Bacteriol* 2000; **182**: 2393–2401.
34. Banu S, Gordon SV, Palmer S *et al.* Genotypic analysis of *Mycobacterium tuberculosis* in Bangladesh and prevalence of the Beijing strain. *J Clin Microbiol* 2004; **42**: 674–682.
35. Sun Y-J, Lee ASG, Ng ST *et al.* Characterisation of ancestral *Mycobacterium tuberculosis* by multiple genetic markers and proposal of genotyping strategy. *J Clin Microbiol* 2004; **42**: 5058–5064.
36. Levinson G, Gutman G. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* 1987; **4**: 203–221.
37. Fang Z, Morrison N, Watt B *et al.* IS*6110* transposition and evolutionary scenario of the direct repeat locus in a group of closely related *Mycobacterium tuberculosis* strains. *J Bacteriol* 1998; **180**: 2102–2109.
38. Kremer K, Au B, Yip P *et al.* Use of variable-number tandem-repeat typing to differentiate *Mycobacterium tuberculosis* Beijing family isolates from Hong Kong and comparison with IS*6110* restriction fragment length polymorphism typing and spoligotyping. *J Clin Microbiol* 2005; **43**: 314–320.
39. Zink A, Sola C, Reischl U *et al.* Characterisation of *Mycobacterium tuberculosis* complex DNAs from Egyptian mummies by spoligotyping. *J Clin Microbiol* 2003; **41**: 359–367.
40. Rothschild B, Martin L, Lev G *et al. Mycobacterium tuberculosis* complex DNA from an extinct bison dated 17,000 years before the present. *Clin Inf Dis* 2001; **33**: 305–311.