Molecular Microbiology (2003) **47**(2), 529–538

Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area

Philip Supply,[1]* Robin M. Warren,[2] Anne-Laure Bañuls,[3] Sarah Lesjean,[1] Gian D. van der Spuy,[2] Lee-Anne Lewis,[2] Michel Tibayrenc,[3] Paul D. van Helden[2] and Camille Locht[1]

[1]*Laboratoire des Mécanismes Moléculaires de la Pathogenèse Bactérienne, INSERM U447, Institut Pasteur de Lille, 1, rue du Prof. Calmette, F-59019 Lille Cedex, France.*

[2]*MRC Centre for Molecular and Cellular Biology, Department of Medical Biochemistry, University of Stellenbosch, PO Box 19063, Tygerberg, 7505, South Africa.*

[3]*Génétique des maladies infectieuses, Unité mixte de Recherche CNRS/IRD 9926, IRD, BP 5045, 34032 Montpellier Cedex 1, France.*

Summary

Deciphering the structure of pathogen populations is instrumental for the understanding of the epidemiology and history of infectious diseases and for their control. Although *Mycobacterium tuberculosis* is the most widespread infectious agent in humans, its actual population structure has remained hypothetical until now because: (i) its structural genes are poorly polymorphic; (ii) adequate samples and appropriate statistics for population genetic analysis have not been considered. To investigate this structure, we analysed the statistical associations (linkage disequilibrium) between 12 independent *M. tuberculosis* minisatellite-like loci by high-throughput genotyping within a model population of 209 isolates representative of the genetic diversity in an area with a very high incidence of tuberculosis. These loci contain variable number tandem repeats (VNTRs) of genetic elements named mycobacterial interspersed repetitive units (MIRUs). Highly significant linkage disequilibrium was detected among the MIRU-VNTR loci in this model. This linkage disequilibrium was also evident when the MIRU-VNTR types were compared with the IS*6110* restriction fragment length polymorphism types. These results support a predominant clonal evolution of *M. tuberculosis*.

Accepted 14 October, 2002. *For correspondence. E-mail philip.supply@pasteur-lille.fr; Tel. (+33) 3 20 87 11 54; Fax (+33) 3 20 87 11 58.

Introduction

Understanding the structure and dynamics of pathogen populations gives unique insights into crucial public health issues, such as the appearance and persistence of variants escaping immunity or the emergence of resistance to antibiotics (Musser, 1996; Spratt and Maiden, 1999; Smith *et al*., 2000; Cooper, 2001). Furthermore, the knowledge of population structures provides a powerful basis to exploit molecular epidemiological data fully and is essential for studying the genetic history of diseases. The clonal structure of *Escherichia coli* has been considered as a paradigm for the majority of bacterial populations. However, the results from multilocus-based analytical approaches have revealed that genetic exchanges are frequent in natural populations of many bacterial species. Analyses of linkage disequilibrium (non-random association of genotypes scored at different chromosomal loci) or based on comparison of gene trees have shown that, in fact, only a few of them appear to be truly clonal (e.g. Smith *et al*., 1993; Spratt and Maiden, 1999; Feil *et al*., 2001).

Mycobacterium tuberculosis is probably the most widespread human pathogen, as it is estimated to infect one-third of the human population. Tuberculosis is the leading cause of death in adults due to a single infectious agent, killing about 3 million people every year, mainly in developing countries (Dye *et al*., 1999). *M. tuberculosis* has potential opportunities for DNA exchange. Some reports have documented simultaneous infection of patients by two different strains, especially in high incidence areas (Chaves *et al*., 1999; Yeh *et al*., 1999; Braden *et al*., 2001). Furthermore, lysogenic mycobacteriophages have been identified that can transduce exogenous DNA into *M. tuberculosis* experimentally (Hatfull, 2000), and naturally occurring conjugation has been demonstrated for mycobacteria (Parsons *et al*., 1998). However, the occurrence and significance of recombinational exchanges

within natural populations of this species have remained completely elusive up to now. Contrary to a widespread idea, the restricted gene sequence diversity within this species (Sreevatsan *et al*., 1997) and empirical observation of some predominant genotypes in various epidemiological studies (van Soolingen *et al*., 1995; Bifani *et al*., 1996; 1999; 2002; Kremer *et al*., 1999; Le *et al*., 2000) provide no indication of its population structure, as they are compatible with distinct population structures with variable levels of recombination (Smith *et al*., 1993; Spratt and Maiden, 1999; Feil *et al*., 2001).

One reason for the poor knowledge of the *M. tuberculosis* population structure is the lack of fully independent polymorphic markers necessary to test potential recombinational exchanges in *M. tuberculosis*. The very restricted polymorphism of structural gene sequences within this species (Sreevatsan *et al*., 1997; Musser *et al*., 2000) has precluded linkage disequilibrium analysis by multilocus enzyme electrophoresis (MLEE) or comparison of gene trees by multilocus sequence typing (MLST). Furthermore, previous methods used to investigate the molecular epidemiology of *M. tuberculosis*, such as IS*6110* restriction fragment length polymorphism (RFLP), cannot be used to analyse linkage disequilibrium, as they do not reveal the variability of independent genetic loci. Spoligotyping, an alternative method based on polymorphisms of a single locus, cannot be assumed to be independent from IS*6110* RFLP, as this locus is a hot-spot for IS*6110* insertions, and changes within this region are often caused by IS*6110*-associated events (e.g. Hermans *et al*., 1991; Groenen *et al*., 1993; Fang *et al*., 1998; Filliol *et al*., 2000; Legrand *et al*., 2001). A second reason for the lack of knowledge of *M. tuberculosis* population structure is the difficulty of constituting adequate representative population sampling, because of the widespread but very unequal distribution of the disease among developing and developed countries.

New genetic markers that allow linkage disequilibrium analysis of mycobacteria have recently become available. They are based on variable number tandem repeats (VNTRs) in mammalian-like minisatellites present in 12 independent loci of the *M. tuberculosis* genome (Supply *et al*., 2000). These elements have been named mycobacterial interspersed repetitive units (MIRUs) (Supply *et al*., 1997). Their potential for *M. tuberculosis* genotyping and for the study of global molecular epidemiology has been demonstrated recently (Mazars *et al*., 2001). In this study, we used MIRU-VNTR high-throughput genotyping (Supply *et al*., 2001) to investigate the population structure of *M. tuberculosis* by measuring the statistical association of the 12 loci in a sample of 209 isolates representative of the genetic diversity in metropolitan Cape Town, South Africa, an area with a very high incidence of tuberculosis (Beyers *et al*., 1996).

Results

Genetic diversity in the M. tuberculosis *population*

The 209 *M. tuberculosis* isolates used in this study correspond to ≈25% of a large collection harvested during a 7 year period from patients with active tuberculosis, which represents ≈70% of the notified adult cases in the high-incidence Cape Town suburban area. The 209 strains were selected for their representativeness of the diversity and frequencies of IS*6110* RFLP patterns in the entire collection (Table S1 in *Supplementary material*). The genotypic diversity in this sample was lower than that of the isolates collected in Paris (Mazars *et al*., 2001). Ninety-seven distinct MIRU-VNTR genotypes were identified (Table S2 in *Supplementary material*), and the mean allelic diversity among the 12 loci amounted to 0.36. One of the least variable MIRU-VNTR loci in other *M. tuberculosis* collections, locus 24 (Mazars *et al*., 2001; Supply *et al*., 2001), was completely monomorphic in this study, with a single MIRU copy in all isolates (Table 1). The individual allelic diversities of the other loci ranged from 0.06 (locus 20) to 0.70 (locus 10). The overall hierarchy of the polymorphisms of the different MIRU-VNTR loci was similar to that observed in previous studies on strains from different countries (Mazars *et al*., 2001; Supply *et al*., 2001). Five (loci 10, 40, 26, 31 and 23) of the six most variable loci identified in previous studies were also the most variable in this population, suggesting that the relative degrees of genetic information carried by the different loci are globally conserved among distinct epidemiological settings and bacterial populations.

MIRU-VNTR linkage disequilibrium analysis

The degree of linkage of the 11 polymorphic MIRU-VNTR loci in the Cape Town population was measured using the genotype-wide and pairwise linkage disequilibrium values

Table 1. MIRU-VNTR allelic diversity in 209 *M. tuberculosis* isolates from metropolitan Cape Town.

Locus	No. of distinct alleles	Allelic diversity (h)[a]
2	3	0.14
4	5	0.26
10	7	0.70
16	4	0.28
20	2	0.06
23	4	0.54
24	1	0.00
26	8	0.59
27	3	0.14
31	5	0.55
39	4	0.36
40	6	0.65

a. Defined as $h = 1-\Sigma x_i^2$, where x_i is the frequency of the ith allele at the locus (Graur and Li, 2000).

Table 2. Measures of genotype-wide association between MIRU-VNTR loci.

		Statistical test		
		f	sI_A	
Sample	Sample size	p_{MC}	p_{para}	p_{MC}
All isolates	209	$<10^{-4}$	$<10^{-6}$	ND
Genotypes	97	$<10^{-4}$	0.01	0.04
Year 93	34	$<10^{-4}$	$<10^{-6}$	4×10^{-4}
Year 94	57	5×10^{-3}	10^{-3}	0.02
Year 95	34	$<10^{-4}$	$<10^{-6}$	$<10^{-4}$
Year 96	26	0.04	$<10^{-6}$	2×10^{-3}
Year 97	17	$<10^{-4}$	$<10^{-6}$	$<10^{-4}$
Year 98	29	$<10^{-4}$	$<10^{-6}$	$<10^{-4}$

P-values give the probability of observing a linkage disequilibrium by chance as high or higher than that observed in the samples, using f (Tibayrenc *et al*., 1990) and sI_A-based test (Brown *et al*., 1980; Smith *et al*., 1993; Haubold *et al*., 1998). p_{para}, significance level calculated by the parametric method; p_{MC}, significance level calculated by 10^4 Monte Carlo simulations. ND, not done because of a too large sample size. Year 92 was not analysed because of the small size of the sample.

between all 11 loci. For the genotype-wide linkage analysis, we assessed the non-random association of independent genetic loci, based on the f test (Tibayrenc *et al*., 1990) and the standardized index of association (sI_A) test (Haubold *et al*., 1998). When these tests were applied to the 209 isolates, highly significant departure from linkage equilibrium was obtained ($P < 10^{-4}$ for the f test, $P < 10^{-6}$ for the sI_A-based parametric test) (Table 2). Pairwise analyses also revealed a highly significant linkage disequilibrium between many pairs of loci (Table 3), with an average of 6.7 linked loci per polymorphic locus.

To assess the structure of this mycobacterial population further, we tested the degree of over-representation of multilocus genotypes and the absence of recombinant genotypes, two additional complementary criteria of clonality (Tibayrenc *et al*., 1990). When all loci of a particular species are considered jointly, some multilocus combinations may be over-represented, whereas others may be missing, although they would be expected with fairly high frequencies if recombination occurred. In that case, it is likely that a few, highly successful genotypes have resulted from clonal expansion. When the statistical tests to evaluate over-representation (d1 and d2) of certain genotypes and the deficiency (e) of others were applied to the Cape Town collection, the values were found to be highly significant (Table 4).

These results therefore provide converging indications of different manifestations of linkage disequilibrium, which represent evidence for clonal population structure (Tibayrenc *et al*., 1990; Tibayrenc, 1999). However, linkage disequilibrium can arise in populations within which lineages are subject to temporal and/or geographical separation (Wahlund effect) that act as physical barriers to gene flow (Souza *et al*., 1992; Smith *et al*., 1993). Such risks of artifacts resulting from geographical isolation between the strains within our population sample are minimized because, in contrast to industrialized countries, infections in the Cape Town area are mostly of regional origin on account of the reduced mobility of the patients outside the region. To assess the potential occurrence of geographical and/or temporal bias within our population model further, tests to evaluate genotype-wide linkage as well as over-representation or absence of certain multilocus genotypes were applied separately per site of the patient's residence within the suburban area (not shown) and per year of collection (Tables 2 and 4). Highly significant values were obtained in virtually every case, thus ruling out a bias resulting from geographical and/or temporal separation. Once the potential bias of the Wahlund effect has been discarded, it is important to distinguish

Table 3. Measures of pairwise association between MIRU-VNTR loci.

Locus	2	4	10	16	20	23	26	27	31	39	40
2	*	++	–	–	–	++	+	–	+	–	+
4	++	*	+	–	–	++	+	–	+	+	+
10	–	+	*	+	+	++	+	–	++	++	+
16	–	–	+	*	–	–	+	+	+	–	++
20	–	–	+	–	*	++	–	–	–	–	–
23	++	++	++	–	++	*	+	+	+	+	++
26	+	+	+	+	–	+	*	–	++	++	+
27	–	–	–	+	–	+	–	*	–	+	+
31	+	+	++	+	–	+	++	–	*	++	+
39	–	+	++	–	–	+	++	+	++	*	+
40	+	+	+	++	–	++	+	+	+	+	*

+, significant linkage disequilibrium when the 209 isolates are considered; ++, significant linkage disequilibrium when the 209 isolates or the 97 genotypes are considered; –, no significant linkage disequilibrium ($P = 0.02$).

Table 4. Statistical tests for over-representation or absence of MIRU-VNTR genotypes.

		Statistical test		
Sample	Sample size	d1	d2	e
All isolates	209	0	$<10^{-4}$	$<10^{-4}$
Genotypes	97	NA	NA	NA
Year 93	34	5×10^{-10}	4×10^{-2}	$<10^{-4}$
Year 94	57	2×10^{-8}	0.19	$<10^{-4}$
Year 95	34	6×10^{-12}	$<10^{-4}$	$<10^{-4}$
Year 96	26	0	$<10^{-4}$	$<10^{-4}$
Year 97	17	10^{-7}	$<10^{-4}$	$<10^{-4}$
Year 98	29	0	$<10^{-4}$	$<10^{-4}$

P-values give the probability of observing as many or more isolates with identical genotypes by chance (d1 based on combinatorial analysis, d2 based on 10^4 Monte Carlo iterations) or as few different genotypes as actually observed (e) (Tibayrenc *et al*., 1990).
NA, not applicable. Year 92 was not analysed because of the small size of the sample.

predominant clonal evolution from epidemic clonality, defined as occasional spread of ephemeral clonal genotypes in a basically sexual species (Smith *et al*., 1993). This can be done by treating each genotype as a single individual in the linkage disequilibrium tests. If linkage persists, this supports the hypothesis of predominant clonal evolution (Smith *et al*., 1993). When the 97 distinct MIRU-VNTR genotypes of the 209 strains were treated as individual units, the genotype-wide linkage disequilibrium was reduced (sI_A of 0.0132 compared with 0.0587 when the 209 isolates were used) but remained significant (Table 2). Significant linkage disequilibrium was also still detected between several pairs of loci (Table 3), with an average reduced to 2.2 linked loci per polymorphic locus, as well as significant over-representation or lack of some genotypes (Table 4). These results support the hypothesis that the population under survey undergoes predominant clonal evolution rather than epidemic clonality.

Ten of the 12 polymorphic MIRU loci display sequence variations between repeat units in addition to variations in repeat numbers (Supply *et al*., 2000). We assume that MIRU-VNTR loci with identical numbers of repeats also have identical sequences among different isolates. This assumption is based on previous sequence analyses indicating that the changes in these loci among different strains nearly always consist of simple additions or deletions in blocks of identical repeats. Among more than 50 alleles sequenced so far in any of the 12 minisatellites, only one case of homoplasy by fragment length (two strains containing identical numbers of repeats but with different sequences) has been observed, in locus 23 (Supply *et al*., 2000). Exclusion of locus 23 in this study did not substantially reduce the levels of significance of genotype-wide linkage disequilibrium and of over- and under-representation of multilocus genotypes (not shown), and only slightly changed the average numbers of linked loci per polymorphic locus (Table 3).

*Linkage disequilibrium between MIRU-VNTRs and IS*6110 *RFLP*

One hundred and thirty-five distinct IS*6110* RFLP profiles were present in the population analysed in this study, compared with 97 MIRU-VNTR genotypes (see above). The discrimination level of MIRU-VNTRs relative to that of IS*6110* RFLP was lower in this study than in previous ones (Mazars *et al*., 2001; Supply *et al*., 2001). This fact can be explained by the slower evolution rate of the 12 MIRU-VNTR compared with that of IS*6110* RFLP, as suggested previously (Mazars *et al*., 2001; Supply *et al*., 2001), and by the reduced bacterial diversity in the local Cape Town population compared with populations from France or from various countries studied previously. In these conditions, a significant proportion of isolates differed from one another by only a few IS*6110* RFLP bands, although having identical MIRU-VNTR types.

Nevertheless, when a matrix of genetic distances based on the IS*6110* RFLP profiles was built for the complete collection and compared with a matrix based on the MIRU-VNTR types using a non-parametric Mantel test, both matrices were found to correlate ($r = 0.499$) with a high degree of significance ($P < 10^{-4}$). This correlation between two independent sets of genetic markers (*g*-test; Tibayrenc *et al*., 1990) provides additional, particularly strong evidence for linkage disequilibrium of genotypes at different loci in the *M. tuberculosis* population. This correlation also indicates that MIRU-VNTR evolution is not dominated by phenomena of convergence or homoplasy by sequence (alleles that are identical in both the number of repeats and the sequence, although not derived from the same ancestral allele because of convergence).

Discussion

Known since antiquity, tuberculosis has disseminated globally, but it is not distributed equally throughout the world. Developing countries have by far the highest burden but are, in most cases, devoid of epidemiological surveillance systems. Even in developed countries, existing culture collections are very incomplete and most often biased towards 'outbreak' isolates. This explains why no collection representative of the global mycobacterial genetic diversity is currently available. For this study of population structure, we selected a collection of *M. tuberculosis* strains that is representative of the genetic diversity from an exceptional setting with both extremely high incidence, notification and recovery rates and high levels of ongoing transmission. We reasoned that, if frequent recombinational gene exchanges occur in *M. tuberculosis*, it would be likely to be detected in such a setting, which potentially provides ample opportunity for genetic exchanges. The use of this collection also minimizes risks of artifactual linkage disequilibrium resulting from geographical isolation between the strains within the same population (Souza *et al*., 1992) because, in contrast to western countries, infections are mostly of regional origin on account of the reduced mobility of the patients outside the region.

The highly significant linkage disequilibrium detected in this study supports a predominantly clonal evolution of *M. tuberculosis* populations in the Cape Town area during the last decade. This linkage disequilibrium was assessed by measuring the association among independent MIRU-VNTR loci and between independent sets of markers, such as MIRU-VNTRs and IS*6110* RFLP. The correlation between independent sets of markers is particularly significant, as it is never significant in species in which recombination is very frequent (Tibayrenc, 1999). Even

when genotypes instead of isolates were used as individual units, the evidence remained strong, although the genotype-wide as well as the pairwise linkage disequilibrium among MIRU-VNTR loci was reduced. This reduction typically reflects the dominance in the overall population of some very frequent genotypes. This dominance is apparent from both IS*6110* RFLP and MIRU-VNTR analyses (see Fig. 1 and Table S1), which both indicate high frequencies of some groups of closely related strains. Previous population-based studies have also indicated restricted genotype polymorphisms in regions with a high prevalence of tuberculosis, compared with regions with low prevalence, which is thought to reflect local ongoing transmission of related strains with little input from other regions (Hermans *et al*., 1995; van Soolingen *et al*., 1995; Kallenius *et al*., 1999). The restricted genotypic diversity found in this study is also consistent with the hypothesis of a relatively recent introduction of tuberculosis into South Africa (Stead, 1997), resulting in weak individualization of lineages from a restricted pool of common recent ancestors. Such a founder effect, leading to an overall low phylogenetic diversity of this sample, could explain the low bootstrap values (not shown) observed for the branching of the dendrogram in Fig. 1, although some limited horizontal gene transfer could also have played a role in generating these low bootstrap values.

However, the pattern observed remains definitely different from the population structure referred to as epidemic clonality, of which *Neisseria meningitidis* is a prototype (Smith *et al*., 1993; 2000). In these structures, dominant clones, emerging from a few ancestral genotypes, persist for a few months or a few years above a background of actively recombining unrelated isolates. These clones may then diversify progressively mainly through horizontal gene transfer (Smith *et al*., 2000). In contrast to such a structure and to panmictic structures such as those observed for *Helicobacter pylori* (see Table 5), the results shown in this study suggest that, if horizontal gene transfer occurred in this *M. tuberculosis* population, it is not sufficient rapidly to distort linkage disequilibrium between alleles from independent MIRU-VNTR loci among the different clonal lineages.

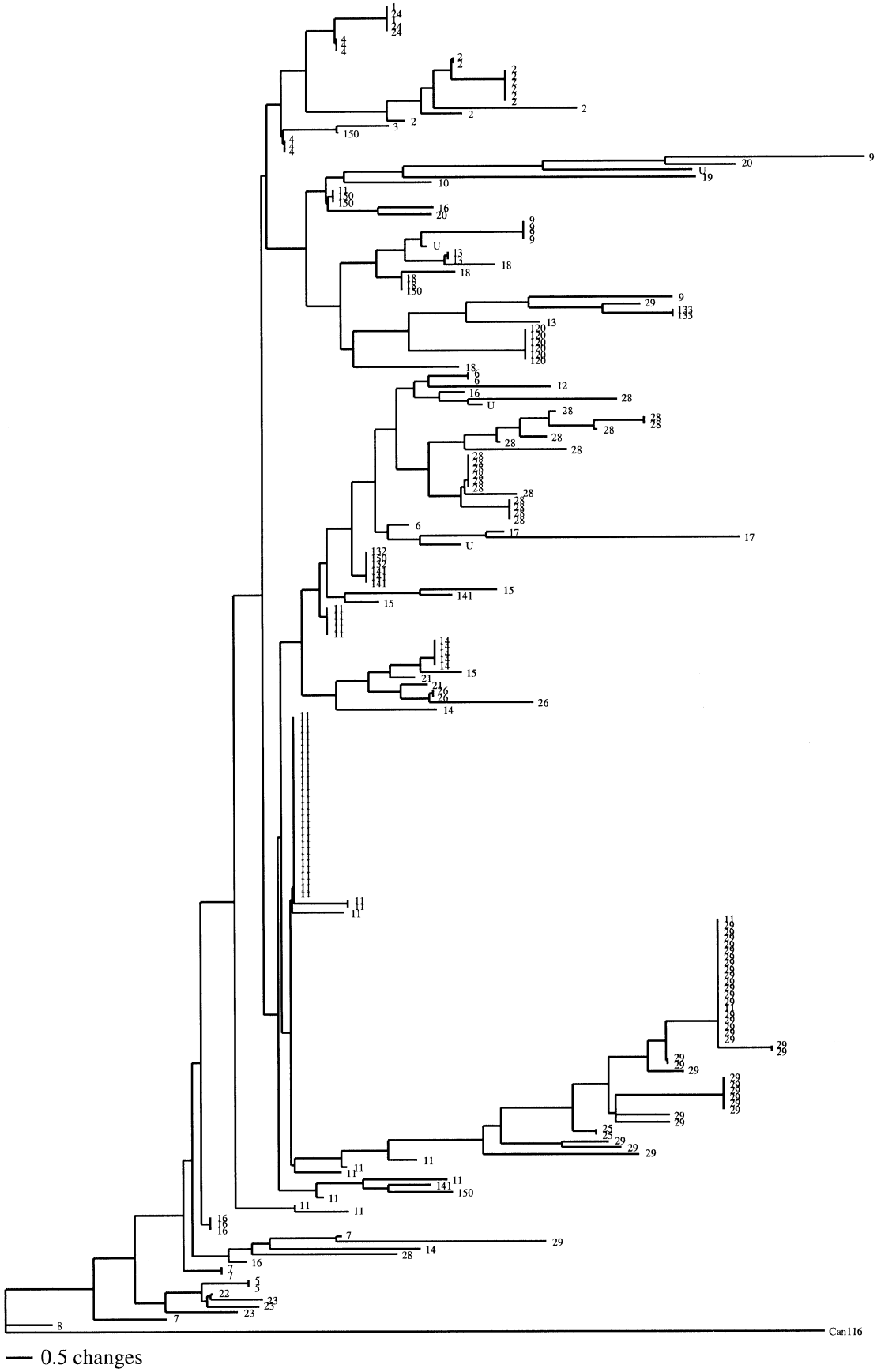
Estimates of relative ratios of recombination to point mutation, which have been obtained for other species based on analysis of single or multiple nucleotide changes in structural genes in clonal complexes defined by MLST (Feil *et al*., 2001), cannot be directly inferred from our results. Nevertheless, analysis of the MIRU-VNTR differences between isolates within given IS*6110* RFLP genotype families suggests that mutations, rather than horizontal genetic exchanges, account for most of these differences. Indeed, distinct MIRU-VNTR genotypes within these families typically differ at only one or a few MIRU-VNTR loci. In these loci, differences between the commonest allele and other alleles often consist of single unit changes. Such stepwise changes are typical mutation patterns of DNA replication-driven VNTR polymorphisms and have been observed in MIRU-VNTR locus 4 in the clonal progeny of the original *Mycobacterium bovis* BCG strain (Supply *et al*., 2000 and references therein). Consistent with the linkage disequilibrium results, these observations suggest that MIRU-VNTR variability is mainly driven by mutation, rather than by recombinational exchanges.

The strong linkages between multilocus markers detected at all levels of analysis reveal a lack of frequent genetic exchanges in *M. tuberculosis* and, thus, a predominantly clonal evolution for this species at least over several years in the highly endemic Cape Town area analysed here. This picture may be an actual reflection of *M. tuberculosis* ecology, in which different *M. tuberculosis* clones are most frequently separated in different individuals, even in a high incidence area such as Cape Town. The lack of frequent recombination inferred from this study on a relatively small spatial and temporal scale does not exclude the occurrence of rare recombination events at

Table 5. Linkage between loci in bacterial populations.

Bacteria	No. of isolates	No. of loci	Methods	I_A[a]	Linkage	References
Helicobacter pylori	74	6	MLEE	0.21	Equilibrium	Go *et al*. (1996)
Neisseria gonorrhoeae	227	9	MLEE	0.04	Equilibrium	Smith *et al*. (1993)
Rhizobium meliloti division B	23	14	MLEE	0.24	Equilibrium	Smith *et al*. (1993)
Porphyromonas gingivalis	57	4	MLST	0.206	Disequilibrium[b]	Frandsen *et al*. (2001)
Streptococcus pneumoniae	187	9	RFLP-PCR	0.43	Disequilibrium[b]	Müller-Graf *et al*. (1999)
Neisseria meningitidis	688	15	MLEE	1.96	Disequilibrium[b]	Smith *et al*. (1993)
Haemophilus influenzae	2209	17	MLEE	5.4	Disequilibrium	Smith *et al*. (1993)
Salmonella typhimurium	340	24	MLEE	1.03	Disequilibrium	Smith *et al*. (1993)
Vibrio cholerae	397	17	MLEE	1.248	Disequilibrium	Beltrán *et al*. (1999)
Campylobacter jejuni	32	7	MLST	1.536	Disequilibirum	Suerbaum *et al*. (2001)
Mycobacterium tuberculosis	209	11	VNTR	0.59	Disequilibrium	This study

a. $I_A = sI_A \times (l-1)$ where l is the number of loci (Haubold *et al*., 1998). Given for information. Absolute values do not necessarily reflect relative degrees of clonality and depend on various factors including sampling and markers used (Smith *et al*., 1993).
b. Epidemic structure, i.e. linkage disequilibrium disappeared when electrophoretic or sequence types, instead of isolates, were used.


Can116
0.5 changes

larger scales. However, such rare recombination events are unlikely to break the prevalent clonal population pattern. Moreover, the results fully corroborate our preliminary observations obtained on a limited sample of strains from other regions of the world (Mazars *et al*., 2001) and thus suggests that a prevalent clonal population structure could be generalized for the whole *M. tuberculosis* species.

The identification of such a clonal structure for a given pathogen has important implications for phylogeny, as well as for molecular epidemiology and public health aspects. In predominantly clonal species, multilocus genotypes are stable in space and time (Spratt and Maiden, 1999; Tibayrenc, 1999). This stability permits relevant exploitation of global molecular epidemiological databases (e.g. Supply *et al*., 2001) to track the past and present spread of the multiple strains of this pathogen. Furthermore, in predominantly clonal species, phylogenetic divergence is not clouded by frequent recombination. Therefore, phylogeny can be used as a framework onto which relevant pathogenicity traits of this pathogen can be mapped (Selander *et al*., 1990).

VNTR loci are powerful markers for the study of population genetics and evolutionary history of many higher eukaryotic species, including humans. So far, they have virtually not yet been used to identify bacterial population structures. Based on analyses of clonal isolates cultivated separately for decades (Supply *et al*., 2000; 2001) or originating from extended transmission chains (P. Supply, unpublished data), the MIRU-VNTR loci appear to be stable for tens of years. These bacterial minisatellites are thus globally free of strongly diversifying selective pressure, indicating their potential for population genetics analysis (Achtman, 1996; Spratt and Maiden, 1999). Recent studies have disclosed the presence of many VNTR-containing minisatellite regions in several bacterial genomes, including those of *Yersinia pestis* and *Bacillus anthracis* (Keim *et al*., 1999; Le Fleche *et al*., 2001), which could be useful in unravelling the population structure of other genetically highly homogeneous microorganisms, including these important human pathogens.

Experimental procedures

Mycobacterium tuberculosis *isolates*

The 209 *M. tuberculosis* isolates used in this study were selected from a large database containing ≈800 isolates from the MRC Centre for Molecular and Cellular Biology at the University of Stellenbosch, Cape Town, South Africa (Warren *et al*., 2000). These isolates were collected between mid-1992 and end-1998 from different patients with active tuberculosis. Most patients ($n = 198$) were resident in a 2.4 km^2 suburban zone of Cape Town, an area with a very high incidence rate (>1000/100 000 per year) and an estimated recovery of 70% of all culture-positive patients (Beyers *et al*., 1996). The remaining isolates ($n = 11$) were from patients residing in neighbouring areas of the metropolis. These isolates were selected to be representative of the diversity of IS*6110* RFLP patterns in the MRC database (Warren *et al*., 2000) (Table S1 in *Supplementary material*). In an initial screening, all the isolates were assembled into groups of similar IS*6110* RFLP patterns (IS-3′ similarity index of ≥65%, calculated using the Dice coefficient and UPGMA clustering method with GELCOMPAR 4.1 software) as described previously (Warren *et al*., 2000). Isolates not sharing this threshold value of similarity with any other isolate in the database were regarded as 'unique' strains. The numbers of isolates selected per group were weighted according to the respective frequencies of strains for these different groups (Table S1 in *Supplementary material*). Within each group, the strains were chosen randomly, regardless of their IS*6110* RFLP patterns.

MIRU-VNTR genotyping

The 209 isolates were genotyped by amplifying the 12 MIRU-VNTR loci in four different multiplex polymerase chain reactions (PCRs) and by analysing the PCR products on 96-well ABI 377 automated sequencers using the GENESCAN and GENOTYPER software packages (PE Applied Biosystem), as described previously (Supply *et al*., 2001).

Genotype diversity and linkage disequilibrium analysis

The MIRU-VNTR allelic diversity (h) at a given locus and the mean allelic diversity (H) were calculated as $h = 1 - \Sigma x_i^2$ and $H = (1/n)\Sigma h_i$, respectively, where x_i is the frequency of the ith allele at the locus, h_i the allelic diversity at locus i and n the number of loci (Graur and Li, 2000). Previously described f, g (Tibayrenc *et al*., 1990) and sI_A-based (Brown *et al*., 1980; Smith *et al*., 1993; Haubold *et al*., 1998) and d1, d2 and e tests (Tibayrenc *et al*., 1990; Tibayrenc, 1999) were used to assess genotype-wide linkage disequilibrium and over-representation and absence of MIRU-VNTR genotypes respectively. The g test (Tibayrenc *et al*., 1990) was used to evaluate linkage disequilibrium between MIRU-VNTR and IS*6110* RFLP patterns. These tests take panmixia (random genetic exchange) as a null hypothesis, which is a very classical approach in all population genetic analyses looking for departures from random genetic exchange expectations. The program LIAN 3.0 (http://soft.ice.mpg.de/lian) (Haubold and Hudson, 2000) was used to carry out the sI_A-based test. This program tests the null hypothesis by a parametric method and, for small data sets, by Monte Carlo simulation with 10^4 iterations. The g test was done by calculating the correlation

Fig. 1. Dendrogram of genetic relationships among 209 *M. tuberculosis* isolates from metropolitan Cape Town based on the 12 MIRU-VNTR loci. The numbers at the right indicate the correspondence with the IS*6110* RFLP groups indicated in Table S1. Linkage distance is indicated at the bottom. Can116, *Mycobacterium canettii* 116 strain (Supply *et al*., 2001) taken as an outgroup.

between the genetic distances inferred from MIRU-VNTR and IS*6110* RFLP patterns for any possible pair of isolates with a non-parametric Mantel test based on Monte Carlo simulation with 10^4 iterations (Mantel, 1967; Tibayrenc, 1995), using the GENETIX program (Laboratoire Génome et Populations, CNRS UPR 9060, Montpellier, France). The distance matrices for the MIRU-VNTR and IS*6110* RFLP data were calculated using Jaccard distances. Pairwise linkage disequilibrium between MIRU-VNTR loci was tested using an extension of the Fisher exact probability test on contingency tables, provided by the program ARLEQUIN (S. Schneider, D. Roessli and L. Excoffier, Genetics and Biometry Laboratory, University of Geneva, Switzerland). The tests were done using a Markov chain with 10^5 steps and 10 000 dememorization steps. Genetic relationships among isolates were represented by neighbour-joining analysis, using the PAUP 4.0 software (D. Swofford), 4.0 beta version (Sinauer Associates).

Acknowledgements

We thank Pablo Bifani for helpful discussion and critical reading of the manuscript, Frederique DeMatos for excellent technical assistance, Vincent Vatin and Philippe Boutin for providing laboratory facilities, and Bruno Pot for help in computing. The work was supported by Institut National de la Santé et de la Recherche Médicale, Institut Pasteur de Lille, Région Nord-Pas-de-Calais and a joint grant from the Ministère de l'Education Nationale, de la Recherche et de la Technologie and Ministère des Affaires Etrangères, and the South African Medical Research Council and NRF. P.S. is a Chercheur du Centre National de la Recherche Scientifique.

Supplementary material

The following material is available from http://www.blackwellpublishing.com/products/journals/suppmat/mole/mole3315/mmi3315sm.htm

Table S1. Selection of 209 *M. tuberculosis* isolates representative of the distribution of the IS*6110* RFLP patterns in metropolitan Cape Town.

Table S2. MIRU-VNTR genotypes of 209 *M. tuberculosis* isolates from metropolitan Cape Town, South Africa.

References

Achtman, M. (1996) A surfeit of YATMs? *J Clin Microbiol* **34:** 1870.

Beltran, P., Delgado, G., Navarro, A., Trujillo, F., Selander, R.K., and Cravioto, A. (1999) Population genetic analysis of *Helicobacter pylori* by multilocus enzyme electrophoresis: extensive allelic diversity and recombinational population structure. *J Clin Microbiol* **37:** 581–590.

Beyers, N., Gie, R.P., Zietsman, H.L., Kunneke, M., Hauman, J., Tatley, M., and Donald, P.R. (1996) The use of a geographical information system (GIS) to evaluate the distribution of tuberculosis in a high-incidence community. *S Afr Med J* **86:** 44.

Bifani, P.J., Plikaytis, B.B., Kapur, V., Stockbauer, K., Pan, X., Lutfey, M.L., *et al.* (1996) Origin and interstate spread of a New York City multidrug-resistant *Mycobacterium tuberculosis* clone family. *J Am Med Assoc* **275:** 452–457.

Bifani, P.J., Mathema, B., Liu, Z., Moghazeh, S.L., Shopsin, B., Tempalski, B., *et al.* (1999) Identification of a W variant outbreak of *Mycobacterium tuberculosis* via population-based molecular epidemiology. *J Am Med Assoc* **282:** 2321–2327.

Bifani, P.J., Mathema, B., Kurepina, N.E., and Kreiswirth, B.N. (2002) Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains. *Trends Microbiol* **10:** 45–52.

Braden, C.R., Morlock, G.P., Woodley, C.L., Johnson, K.R., Colombel, A.C., Cave, M.D., *et al.* (2001) Simultaneous infection with multiple strains of *Mycobacterium tuberculosis*. *Clin Infect Dis* **33:** 42–47.

Brown, A.H.D., Feldman, M.W., and Nevo, E. (1980) Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics* **96:** 523–526.

Chaves, F., Dronda, F., Alonso-Sanz, M., and Noriega, A.R. (1999) Evidence of exogenous reinfection and mixed infection with more than one strain of *Mycobacterium tuberculosis* among Spanish HIV-infected inmates. *AIDS* **13:** 615–620.

Cooper, B.S. (2001) Pathogen population dynamics: the age of the strain. *Trends Microbiol* **9:** 199–200.

Dye, C., Scheele, S., Dolin, P., Pathania, V., and Raviglione, M.C. (1999) Consensus statement: global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. *J Am Med Assoc* **282:** 677–686.

Fang, Z., Morrison, N., Watt, B., Doig, C., and Forbes, K.J. (1998) IS*6110* transposition and evolutionary scenario of the direct repeat locus in a group of closely related *Mycobacterium tuberculosis* strains. *J Bacteriol* **180:** 2102–2109.

Feil, E.J., Holmes, E.C., Bessen, D.E., Chan, M.S., Day, N.P., Enright, M.C., *et al.* (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc Natl Acad Sci USA* **98:** 182–187.

Filliol, I., Sola, C., and Rastogi, N. (2000) Detection of a previously unamplified spacer within the DR locus of *Mycobacterium tuberculosis*: epidemiological implications. *J Clin Microbiol* **38:** 1231–1234.

Frandsen, E.V., Poulsen, K., Curtis, M.A., and Kilian, M. (2001) Evidence of recombination in Porphyromonas gingivalis and random distribution of putative virulence markers. *Infect Immun* **69:** 4479–4485.

Go, M.F., Kapur, V., Graham, D.Y., and Musser, J.M. (1996) Genetic diversity and population structure of *Vibrio cholerae*. *J Bacteriol* **178:** 3934–3938.

Graur, D., and Li, W.-H. (2000) Dynamics of genes in populations. In *Fundamentals of Molecular Evolution.* Graur, D., and Li, W.-H. (eds). Sunderland, MA: Sinauer Associates, p. 58.

Groenen, P.M., Bunschoten, A.E., van Soolingen, D., and van Embden, J.D. (1993) Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*;

application for strain differentiation by a novel typing method. *Mol Microbiol* **10:** 1057–1065.

Hatfull, G.F. (2000) Molecular genetics of mycobacteriophages. In *Molecular Genetics of Mycobacteria*. Hatfull, G.F., and Jacobs, W.R., Jr (eds). Washington, DC: American Society for Microbiology Press, pp. 37–54.

Haubold, B., and Hudson, R.R. (2000) LIAN 3.0: detecting linkage disequilibrium in multilocus data linkage analysis. *Bioinformatics* **16:** 847–848.

Haubold, B., Travisano, M., Rainey, P.B., and Hudson, R.R. (1998) Detecting linkage disequilibrium in bacterial populations. *Genetics* **150:** 1341–1348.

Hermans, P.W., van Soolingen, D., Bik, E.M., de Haas, P.E., Dale, J.W., and van Embden, J.D. (1991) Insertion element IS*987* from *Mycobacterium bovis* BCG is located in a hotspot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. *Infect Immun* **59:** 2695–2705.

Hermans, P.W., Messadi, F., Guebrexabher, H., van Soolingen, D., de Haas, P.E., Heersma, H., *et al.* (1995) Analysis of the population structure of *Mycobacterium tuberculosis* in Ethiopia, Tunisia, and The Netherlands: usefulness of DNA typing for global tuberculosis epidemiology. *J Infect Dis* **171:** 1504–1513.

Kallenius, G., Koivula, T., Ghebremichael, S., Hoffner, S.E., Norberg, R., Svensson, E., *et al.* (1999) Evolution and clonal traits of *Mycobacterium tuberculosis* complex in Guinea-Bissau. *J Clin Microbiol* **37:** 3872–3878.

Keim, P., Klevytska, A.M., Price, L.B., Schupp, J.M., Zinser, G., Smith, K.L., *et al.* (1999) Molecular diversity in *Bacillus anthracis*. *J Appl Microbiol* **87:** 215–217.

Kremer, K., van Soolingen, D., Frothingham, R., Haas, W.H., Hermans, P.W., Martin, C., *et al.* (1999) Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility. *J Clin Microbiol* **37:** 2607–2618.

Le, T.K., Bach, K.H., Ho, M.L., Le, N.V., Nguyen, T.N., Chevrier, D., and Guesdon, J.L. (2000) Molecular fingerprinting of *Mycobacterium tuberculosis* strains isolated in Vietnam using IS*6110* as probe. *Tuber Lung Dis* **80:** 75–83.

Le Fleche, P., Hauck, Y., Onteniente, L., Prieur, A., Denoeud, F., Ramisse, V., *et al.* (2001) A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis* BMC. *Microbiology* **1:** 2.

Legrand, E., Filliol, I., Sola, C., and Rastogi, N. (2001) Use of spoligotyping to study the evolution of the direct repeat locus by IS*6110* transposition in *Mycobacterium tuberculosis*. *J Clin Microbiol* **39:** 1595–1599.

Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res* **27:** 209–220.

Mazars, E., Lesjean, S., Banuls, A.L., Gilbert, M., Vincent, V., Gicquel, B., *et al.* (2001) High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. *Proc Natl Acad Sci USA* **98:** 1901–1906.

Müller-Graf, C.D., Whatmore, A.M., King, S.J., Trzcinski, K., Pickerill, A.P., Doherty, N., *et al.* (1999) Population biology of *Streptococcus pneumoniae* isolated from oropharyngeal carriage and invasive disease. *Microbiology* **145:** 3283–3293.

Musser, J.M. (1996) Molecular population genetic analysis of emerged bacterial pathogens: selected insights. *Emerg Infect Dis* **2:** 1–17.

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

Parsons, L.M., Jankowski, C.S., and Derbyshire, K.M. (1998) Conjugal transfer of chromosomal DNA in *Mycobacterium smegmatis*. *Mol Microbiol* **28:** 571–582.

Selander, R.K., Beltran, P., Smith, N.H., Barker, R.M., Crichton, P.B., Old, D.C., *et al.* (1990) Genetic population structure, clonal phylogeny, and pathogenicity of *Salmonella paratyphi* B. *Infect Immun* **58:** 1891–1901.

Smith, J.M., Smith, N.H., O'Rourke, M., and Spratt, B.G. (1993) How clonal are bacteria? *Proc Natl Acad Sci USA* **90:** 4384–4388.

Smith, J.M., Feil, E.J., and Smith, N.H. (2000) Population structure and evolutionary dynamics of pathogenic bacteria. *Bioessays* **22:** 1115–1122.

van Soolingen, D., Qian, L., de Haas, P.E., Douglas, J.T., Traore, H., Portaels, F., *et al.* (1995) Predominance of a single genotype of *Mycobacterium tuberculosis*. countries of east Asia. *J Clin Microbiol* **33:** 3234–3238.

Souza, V., Nguyen, T.T., Hudson, R.R., Pinero, D., and Lenski, R.E. (1992) Hierarchical analysis of linkage disequilibrium in *Rhizobium* populations: evidence for sex? *Proc Natl Acad Sci USA* **89:** 8389–8393.

Spratt, B.G., and Maiden, M.C. (1999) Bacterial population genetics, evolution and epidemiology. *Phil Trans R Soc London B Biol Sci* **354:** 701–710.

Sreevatsan, S., Pan, X., Stockbauer, K.E., Connell, N.D., Kreiswirth, B.N., Whittam, T.S., and Musser, J.M. (1997) Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci USA* **94:** 9869–9874.

Stead, W.W. (1997) The origin and erratic global spread of tuberculosis. How the past explains the present and is the key to the future. *Clin Chest Med* **18:** 65–77.

Suerbaum, S., Lohrengel, M., Sonnevend, A., Ruberg, F., and Kist, M. (2001) Allelic diversity and recombination in *Campylobacter jejuni*. *J Bacteriol* **183:** 2553–2559.

Supply, P., Magdalena, J., Himpens, S., and Locht, C. (1997) Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Mol Microbiol* **26:** 991–1003.

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

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

Tibayrenc, M. (1995) Population genetics of parasitic protozoa and other microorganisms. In *Advances in Parasitology*, Vol. 36. Baker, J.R., Muller, R., and Rollinson, D. (eds). London: Academic Press, Inc., pp. 47–115.

Tibayrenc, M. (1999) Toward an integrated genetic epidemiology of parasitic protozoa and other pathogens. *Annu Rev Genet* **33:** 449–477.

Tibayrenc, M., Kjellberg, F., and Ayala, F.J. (1990) A clonal theory of parasitic protozoa: the population structures of *Entamoeba*, *Giardia*, *Leishmania*, *Naegleria*, *Plasmodium*, *Trichomonas*, and *Trypanosoma* and their medical and taxonomical consequences. *Proc Natl Acad Sci USA* **87:** 2414–2418.

Warren, R.M., Sampson, S.L., Richardson, M., Van Der Spuy, G.D., Lombard, C.J., Victor, T.C., and van Helden, P.D. (2000) Mapping of IS*6110* flanking regions in clinical isolates of *Mycobacterium tuberculosis* demonstrates genome plasticity. *Mol Microbiol* **37:** 1405–1416.

Yeh, R.W., Hopewell, P.C., and Daley, C.L. (1999) Simultaneous infection with two strains of *Mycobacterium tuberculosis* identified by restriction fragment length polymorphism analysis. *Int J Tuberc Lung Dis* **3:** 537–539.