JOURNAL OF BACTERIOLOGY, Dec. 2006, p. 8169–8177 Vol. 188, No. 23
0021-9193/06/$08.00+0 doi:10.1128/JB.01062-06

Evidence for Recombination in *Mycobacterium tuberculosis*▽†

Xiaoming Liu,[1] Michaela M. Gutacker,[2]‡ James M. Musser,[2,3] and Yun-Xin Fu[1]*

Human Genetics Center, University of Texas at Houston, Houston, Texas 77225[1]; Laboratory of Human Bacterial Pathogenesis, Rocky Mountain Laboratories, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 903 South 4th Street, Hamilton, Montana 59840[2]; and Center for Human Bacterial Pathogenesis Research, Department of Pathology, Baylor College of Medicine, Houston, Texas 77030[3]

Received 19 July 2006/Accepted 14 September 2006

Due to its mostly isolated living environment, *Mycobacterium tuberculosis* is generally believed to be highly clonal, and thus recombination between different strains must be rare and is not critical for the survival of the species. To investigate the roles recombination could have possibly played in the evolution of *M. tuberculosis*, an analysis was conducted on previously determined genotypes of 36 synonymous single nucleotide polymorphisms (SNPs) in 3,320 *M. tuberculosis* isolates. The results confirmed the predominant clonal structure of the *M. tuberculosis* population. However, recombination between different strains was also suggested. To further resolve the issue, 175 intergenic SNPs and 234 synonymous SNPs were genotyped in 37 selected representative strains. A clear mosaic polymorphic pattern ahead of the MT0105 locus encoding a PPE (Pro-Pro-Glu) protein was obtained, which is most likely a result of recombination hot spot. Given that PPE proteins are thought to be critical in host-pathogen interactions, we hypothesize that recombination has been influential in the history of *M. tuberculosis* and possibly a major contributor to the diversity observed ahead of the MT0105 locus.

Analysis of the genetic structure and evolution of populations of pathogenic microbes is essential for understanding the mechanisms responsible for the ability to escape host immune responses, drug resistance, and the reemergence of infectious diseases (6, 27, 30, 45, 48). Exchanging genetic material can be an important factor in a bacterium's success. It increases genetic variability and the ability of the bacteria to rapidly adapt to environmental changes and, thus, to survive. Much evidence has accrued to indicate that naturally occurring recombination between strains is frequent in many bacterial species (9, 46, 48). Throughout the present paper, recombination refers to the event in which one bacterium obtains a segment of DNA from an individual of the same species through at least two chromosome crossovers (in the circular bacterial genome). However, there are also some pathogenic bacteria, such as *Yersinia pestis*, *Salmonella enterica* serovar Typhi, and *Mycobacterium tuberculosis*, that are thought to have evolved clonally with little impact from recombination.

M. tuberculosis, the etiological agent of human tuberculosis (TB), is the most widespread infectious bacterial pathogen in human beings. According to the World Health Organization fact sheet on tuberculosis, TB bacilli newly infect someone every second, and one-third of the world's population is currently infected with the TB bacillus (55). Because of its isolated living space, lack of migration between hosts, long generation time, and latent stage, *M. tuberculosis* along with other members of the *M. tuberculosis* complex remains the paradigm of clonal evolution (47). Thus, if recombination occurs, it happens between identical or nearly identical individuals and hence leaves little detectable trace. Indeed, there is virtually no evidence of recombination of *M. tuberculosis* species reported to date, and many population studies confirmed the predominant clonal evolution of *M. tuberculosis* (15, 16, 49, 51).

However, there is a logical possibility of recombination between different strains of *M. tuberculosis*. Patients simultaneously infected by two different strains have been reported (2, 7, 36, 44, 52, 56). A chromosomally coded conjugation system has recently been identified in *Mycobacterium smegmatis* (35, 53), which opened up the possibility of lateral DNA transfer via conjugation with other mycobateria. Recently, evidence of recombination has been found among strains that are considered to be the progenitors of the *M. tuberculosis* complex (17). Recent studies (18, 37) have shown that mycobacteriophage genomes are highly mosaic and contain many unexpected genes, possibly originating from the hosts, so that it is also possible that horizontal DNA transfer between strains (even between species) without coinfection can happen via transduction.

In general, recombination is revealed through examining polymorphic loci. The fact that no evidence of recombination of *M. tuberculosis* strains has been reported may be due to the fact that there is, indeed, no recombination occurring between different strains. Alternatively, it may be that recombination has not been recognized because insufficient polymorphic markers have been typed or only small strain sample sizes have been studied (49, 51). Recently, the whole-genome sequences of two *M. tuberculosis* strains, H37Rv and CDC1551, have been released (5, 14). More than 1,000 single nucleotide polymorphisms (SNPs) along with other polymorphisms were identified after the comparison of the two genomes (14). Considering the important role recombination may play in escaping host immune responses and developing multidrug resistance, a large-

* Corresponding author. Mailing address: Human Genetics Center, University of Texas at Houston, P.O. Box 20186, Houston, TX 77225. Phone: (713) 500-9813. Fax: (713) 500-0900. E-mail: Yunxin.Fu@uth.tmc.edu.

‡ Present address: Istituto Cantonale di Microbiologia, Via Mirasole 22A, 6500 Bellinzona, Switzerland.

† Supplemental material for this article may be found at http://jb.asm.org/.

▽ Published ahead of print on 22 September 2006.

scale population genetics study of *M. tuberculosis* is necessary and now possible.

Gutacker et al. (15) described the variation of 5,069 *M. tuberculosis* strains by analyzing 36 synonymous SNPs (sSNPs) and 48 genetically representative *M. tuberculosis* strains by studying 227 nonsynonymous SNPs and 121 intergenic SNPs (iSNPs). These authors found a complicated polymorphic pattern that may have been shaped by recombination during evolution. In the present paper, we report our extended analysis on both a large-scale population genetics study of 3,320 *M. tuberculosis* isolates with 36 sSNPs and an extensive polymorphic pattern study of 37 *M. tuberculosis* isolates with 409 SNPs. We show a clear evidence of recombination in *M. tuberculosis* strains that was found via genetic and statistical analyses.

MATERIALS AND METHODS

Mycobacterial isolates. The 3,320 *M. tuberculosis* isolates used for the present analysis are part of the 5,069 isolates studied by Gutacker et al. (15). All isolates were collected from population-based tuberculosis surveillances of TB patients. The distribution of patients is as follows: New York, N.Y., 1,025 isolates from the year 2000; New Jersey, 756 isolates from 2000 to 2001; Houston, TX, 682 isolates from 1995 to 2000; and Finland, 857 isolates from 2000 to 2001.

SNP data. sSNPs were first identified by aligning genome sequences of four strains (H37Rv, CDC1551, 210, and *Mycobacterium bovis* strain AF2122/97) and then verified by independent sequencing of the relevant genome segments in strains TN587, CDC1551, H37Rv, and *M. bovis* strain TN10130. iSNPs were discovered by independent sequencing of a group of representative strains. The processes of identification, selection, and verification of SNPs were described in detail by Gutacker et al. (15, 16). A total of 409 SNPs were studied in this research. All 234 sSNPs and 64 iSNPs were studied previously (15, 16), and 111 new iSNPs were discovered and genotyped in this research.

Minimum number of recombination events. The lower bounds of the minimum number of recombination events in the history of the samples were estimated using Myers and Griffiths' composite bound algorithm (31). Two methods were chosen to estimate local lower bounds for the composite bound. Considering only SNPs in the local interval, if there were no more than 15 distinctive haplotypes in the sample, Bafna and Bansal's R_I (1) was used. Otherwise, Hudson and Kaplan's $R_{\min}$ (20) was used for local lower bounds. The algorithms used here were all under the infinite-site model, under which no multiple mutations on the same locus were allowed. The estimated lower bound shows at least how many crossovers should happen in the sample history assuming no multiple hits.

DNA maximum parsimony tree. The DNA maximum parsimony tree method has been widely used for reconstructing phylogenetic trees from DNA sequences (12, 13). In principle, this method searches for the tree topology with the minimum number of mutations to explain the polymorphism in the DNA sequences of a sample, assuming no recombination. The program DNAPARS from the package PHYLIP, version 3.5c (10), was used to estimate the minimum number of mutations needed for a sample.

Neighbor-joining tree. The neighbor-joining tree method is a widely used distance-based phylogeny reconstruction method (40, 50). It is fast, robust, and suitable for conducting bootstrap tests (32). The software MEGA3 (23) was used to build neighbor-joining trees with *p* distance and pairwise deletion for missing data. When fewer than 10 strains were analyzed, a regular bootstrap test was used with 1,000 repeats. Otherwise, Efron's two-level bootstrap method (8) was used to produce a more accurate bootstrap confidence level because of the downward trend of the regular bootstrap values in large-scale phylogenies (11, 19, 33, 42, 57, 58). In the two-level bootstrap method, 2,000 repeats for the first-round bootstrap, seven edge points, and 400 second-round bootstrap repeats for each edge point were used.

Recombination detection algorithms. Many algorithms for detecting recombination have been developed over the last two decades (38, 39, 54). However, the reliability of these algorithms still needs extensive testing. Some preliminary comparisons suggested that most of the algorithms suffer from high false-positive rates (38). Therefore, several recombination detection algorithms were used for the preliminary detection of possible recombination regions for further analysis. The software RDP 2.0 (25) was used to conduct such detection. It incorporated the RDP algorithm (25), GENECONV (43), maximum chi-square (26), maximum match chi-square (39), reticulate (21), and bootscanning (41), among other methods. The first three algorithms were used for detection and tests of significance because of their relatively high consensus power and/or low false-positive rates (38). Additionally, other algorithms (maximum match chi-square, reticulate, and bootscanning) were used for validation.

Recombination Rate Estimation. The program package Ldhat2.0 (29) was used to conduct McVean's composite-likelihood estimation of $\gamma = 8N_e c\bar{t}$ with the gene-conversion model, where c and $\bar{t}$ are the per base initiation rate and the average tract length of gene conversion, respectively, and N_e is the effective population size; γ can be regarded as the population rate of recombination between two distantly linked loci.

As a prerequisite of McVean's method for allowing multiple mutations, the population mutation rate per site, $\theta = 4N_e\mu$, needs to be provided. An approximate finite-sites version of Watterson's estimator (29)

$$\hat{\theta}_w^* \left(\sum_{k=1}^{n-1} \frac{1}{k} \right)^{-1} \ln\left(\frac{L}{L-S} \right)$$

was used for estimating θ on the synonymous site based on a previous study of some structural genes (49), where n is the sample size, L is the total number of synonymous sites, and S is the total number of synonymous mutations. Most of the estimated per-site θs were near 0, while the highest one was approximately 0.01 (*gyrA*). Given that all these estimations of θ are very small, a reasonable infinite-site model can be assumed. Therefore, per site $\theta = 0.001$ and $\bar{t} = 100$ bp was used for analysis. The likelihood search grid ranged from 0 to 3 with a step (precision) of 0.1.

RESULTS

Data set I. Thirty-six sSNPs were previously genotyped in 3,320 *M. tuberculosis* isolates sampled from patients from four geographic populations (Finland, Houston, New Jersey, and New York City; see Material and Methods) (15). These sSNPs were selected as the most informative for categorizing different strains according to a previous phylogenetic analysis (16). Selection and genotyping procedures for the 36 sSNPs were described in detail by Gutacker et al. (15). In total, 257 distinctive haplotypes based on these 36 sSNPs were identified (Table 1; see Table S1 in the supplemental material). In the sample from New Jersey, the sSNP of Rv1175c has a rare third allele, T. And in the sample from New York City, the sSNP of Rv0307c_1A has a rare third allele, G. Each of them is present in only a single isolate in the sample. To simplify analysis, we treat them as missing data.

Minimum number of recombination events in data set I. Myers and Griffiths' composite bound algorithm (31) was used to estimate the minimum number of recombination events in data set I under the infinite-site model (Table 1). The lower bounds of the minimum number of recombination events ranged from 35 to 44 for the four populations. If samples are pooled together, at least 65 recombination events are needed in the history, assuming no multiple mutations.

Minimum number of mutations in data set I. Backward mutation alone can also produce a similar polymorphic pattern as recombination events. The DNA maximum parsimony tree method (12, 13) was used to estimate the minimum number of mutations needed to produce the same pattern in the sample (Table 1). The estimated minimum number of mutations ranged from 84 to 136 for the four populations. Surprisingly, 299 mutations were needed if the four populations were treated as a whole. These large numbers of mutations on only 36 segregating sites suggest that backward mutation alone is insufficient to account for all the polymorphic patterns found in the sample, particularly considering the fact that only two of the polymorphic sites have more than two alleles. We shall return to this issue later.

TABLE 1. Minimum number of recombinations or mutations in data set I

Population	No. of isolates	No. of distinctive strains[a]	No. of noninformative SNPs[b]	Minimum no. of recombinations[c]	Minimum no. of mutations[d]
New York, N.Y.	1,025	69	1	35	84
New Jersey	756	64	0	37	93
Houston, Tex.	682	111	1	44	136
Finland	857	103	7	37	121
Total	3,320	257	0	65	299
Cleaned[e]		68	3	28	75

[a] A distinctive strain has a distinctive genotype of the 36 SNPs.
[b] A noninformative SNP locus is monomorphic or its rarer allele only has one carrier in that sample.
[c] Estimated with Myers and Griffiths' algorithm assuming no multiple mutations.
[d] Estimated with DNA parsimony tree method assuming no recombination.
[e] The total data set excluding isolates with a genotype frequency of less than three copies.

Elimination of the effects of possible genotyping error in data set I. Considering the effects of possible genotyping errors on the estimation of the minimum number of mutations, isolates with a genotype frequency of less than three copies in the sample were excluded. The composite bound algorithm and the DNA parsimony tree method were repeated on this "cleaned" data set (Table 1). The estimated minimum of 75 mutations or 39 backward mutations was still too large for the 36 informative segregating sites. This result confirmed that it was most likely that some recombination events and backward mutations were responsible for producing the observed pattern of polymorphism in this data set.

Estimation of recombination rate with data set I. McVean's composite-likelihood estimator was used to estimate the population recombination rate, γ, using the program package Ldhat2.0 (29). Because of computational intensiveness, the analysis of the whole sample is inapplicable. Thus, 10 subsamples of 100 isolates each were randomly selected from each sample population (Finland, Houston, New Jersey, and New York City) for analysis. The same analysis procedures were conducted with a pooled sample of all four populations. To relieve the computational intensity of the program, only isolates with no missing data were sampled.

The estimations of γ were 0.97 ± 0.11, 0.93 ± 0.09, 0.40 ±

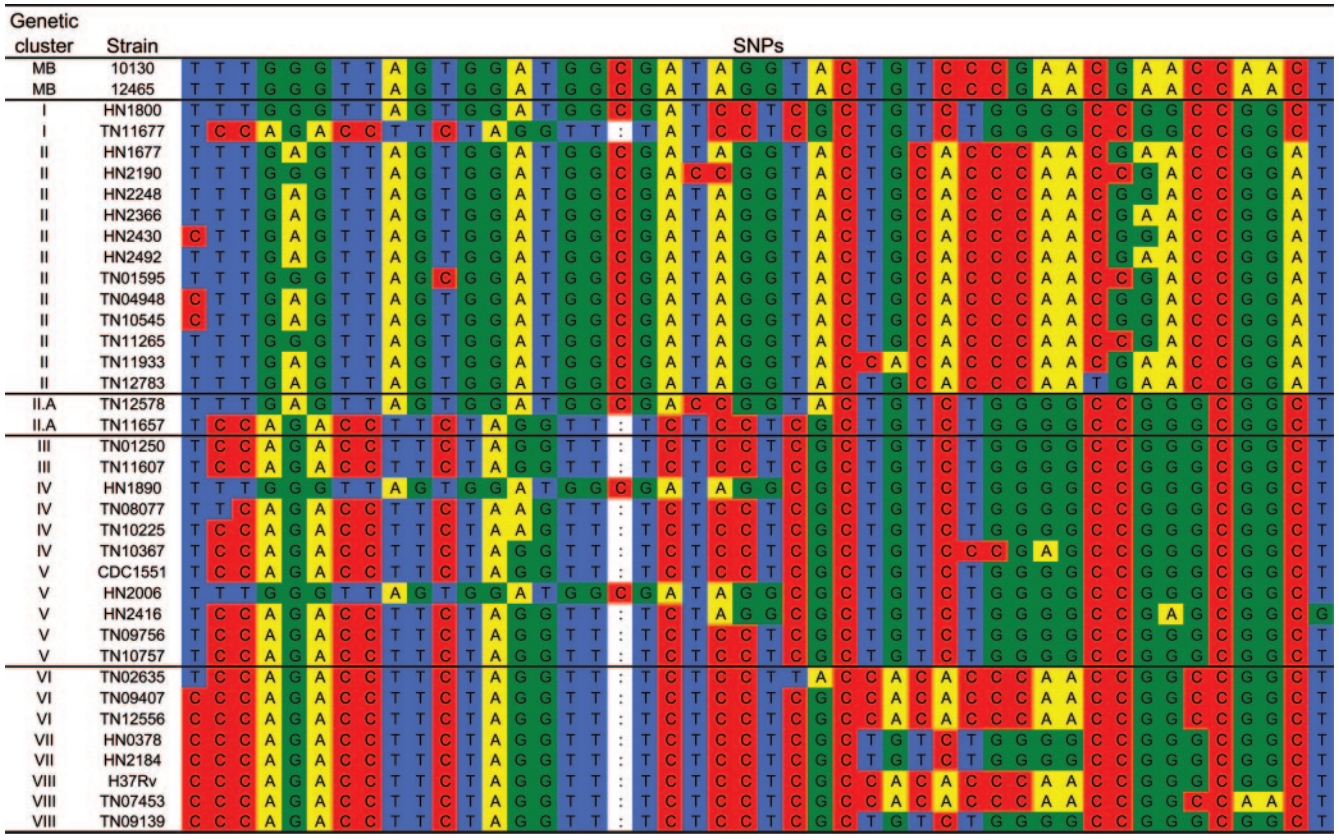

FIG. 1. Mosaic pattern of iSNP in IRMT0105. Deletions are indicated by colons.

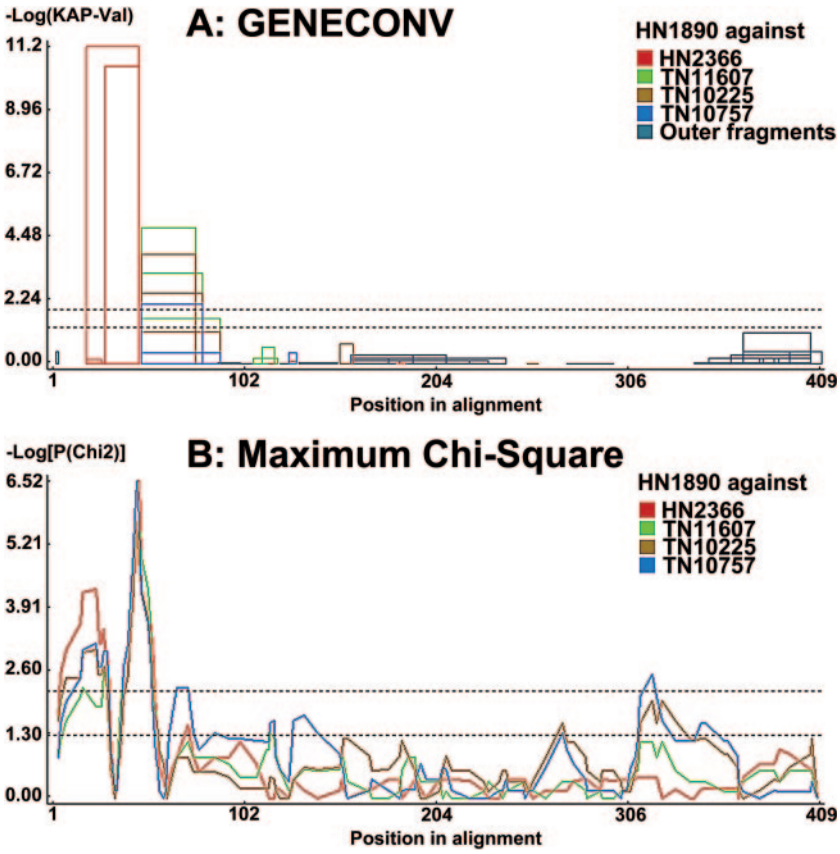

FIG. 2. Recombination detection results of GENECONV and maximum chi-square for strain HN1890 (cluster IV) versus HN2366 (II), TN11607 (III), TN10225 (IV), and TN10757 (V). Dashed lines correspond to local and global significance cuts of the test statistics used (see reference 25 for details). Both the highest red bar in plot A and the interval between the two highest peaks in plot B correspond to the first block of the IRMT0105 iSNPs of HN1890: TTTGGGTTAGTGGATGGCGATAGG. Both algorithms were run with RDP software. GENECONV used the default setting except that each indel site was treated as a polymorphism. Maximum chi-square used the default setting except that gaps were included.

0.06, 0.59 ± 0.06, and 0.83 ± 0.10 (mean ± standard error) for the Finland, Houston, New Jersey, and New York City populations and pooled samples, respectively.

These results suggest that when considering the *M. tuberculosis* genome as a whole, effective gene conversion (or recombination between distant linked loci) is relatively rare, but the results do not rule out frequent small-scale gene conversion or recombination hot spots.

Data set II. Gutacker et al. (15) analyzed 578 polymorphic SNPs in 48 representative strains. For our analysis we excluded all nonsynonymous SNPs from the above data set of 578 SNPs because they may follow a different evolutionary model than sSNPs and iSNPs. Among the 48 strains, 37 *M. tuberculosis* strains, and 2 *M. bovis* strains were used for extended iSNP genotyping. These selected *M. tuberculosis* strains still represent all eight previously defined phylogenetic clusters (I to VIII) (15, 16).

For each of these 37 isolates, a total of 409 SNPs were genotyped, including 175 iSNPs and 234 sSNPs (see Table S2 in the supplemental material). It is assumed that iSNPs and sSNPs have similar mutation rates. No SNP with more than two alleles was found in this data set. There were 265 transitions, 128 transversions, and 6 insertion/deletions. Ten SNPs turned out to be monomorphic in this data set.

An intergenic region (IR) between locus MT0103 (position 103700 to 104563 of CDC1551 genome; GenBank locus NC_002755) and the PPE (where P is and E is Glu) gene locus MT0105 (position 105314 to 106705 of CDC1551 genome; GenBank locus NC_002755) included 46 iSNPs. Within this region, there is a pseudogene called MT0104 (position 104424 to 105091 of CDC1551 genome; GenBank locus NC_002755). We called this region IRMT0105 in this paper according to its closest gene locus. Because this locus is highly polymorphic, it has the highest density of SNPs in data set II. The positions of these 46 iSNPs are 139, 194, 200, 202, 209, 212, 218, 221, 227, 228, 230, 233, 235, 236, 237, 239, 240, 240 to 241(deletion), 244, 251, 256, 257, 260, 265, 283, 285, 324, 342, 346, 363, 367, 368, 369, 370, 371, 389, 427, 434, 438, 448, 472, 477, 487, 490, 566, and 667. All positions refer to the beginning (position 104564 of GenBank locus NC_002755) of the IR on the genome of strain CDC1551. Two of the SNPs at positions 324 and 477 are monomorphic in this data set. All the remaining 363 SNPs will be referred to as background SNPs in this paper.

Detection of recombination in IRMT0105 in data set II. IRMT0105 is located ahead of a PPE gene. The PE and PPE gene families are one of the major discoveries of the *M. tuberculosis* genome projects (4, 5). These two gene families account for about 10% of all coding sequences of the *M. tuberculosis* genome. Members of the PE and PPE gene families share a conserved N-terminal domain with a Pro-Glu (PE) or a Pro-Pro-Glu (PPE) motif. Although the exact functions of these proteins are not known, evidence has accumulated that they are involved in antigenic variation and escape from B-cell and T-cell responses and growth in macrophages (3, 4, 24, 34). Thus, PE and PPE proteins may contribute to the success of the organism in escaping host immune surveillance. A genome comparison between the strain H37Rv and strain CDC1551 showed high substitution and insertion/deletion frequency within genes of these families (14), consistent with the hypotheses of a role in virulence or escape from host immune responses. Therefore, a high degree of polymorphism of these genes may be crucial for the survival of the *M. tuberculosis* population, and recombination may be an important mechanism for inflating the polymorphism present in these proteins.

The algorithms RDP (25), GENECONV (43), and maximum chi-square (26) were used to detect possible regions of recombination. If at least two of the three algorithms had significant results of recombination, it was regarded as a strong recombination signal. There was no strong signal detected in background SNPs; on the other hand, many strong signals of recombination were picked up in the IRMT0105 iSNPs.

These strong signals were confirmed to correspond to the incongruent blocks in IRMT0105 of the isolates HN1800, TN11677, TN12578, HN1890, HN2006, TN02635, TN09407, TN12556, TN07453, and H37Rv (Fig. 1). Obviously, for IRMT0105, these 37 strains can be approximately divided into two major groups. Group 1 includes the *M. bovis* and the *M. tuberculosis* clusters I and II (referred to as the principal genetic group 1 in reference 16). Group 2 includes the *M. tuberculosis* clusters III to VIII (referred to as principal genetic groups 2 and 3 in reference 16). Cluster II.A is more like a bridge between the two major groups. The previously mentioned strains each had an incongruent block of SNPs that was obviously not similar to the corresponding segment of the members of its own group but was almost identical to the corresponding segment of members of the other group. Figure 2 shows an example of the correlation between the significant signals of the detection algorithms and the SNP block of HN1890. The iSNPs in IRMT0105 of HN1890 can obviously be divided into two parts: the first part includes 24 SNPs (TTTG GGTTAGTGGTTGGCGATAGG), which are identical to the corresponding part of most of the strains of group 1; the second part includes 22 SNPs (CGCTGTCTGGGGCCGGG CGGCT), which are identical to the corresponding part of most strains of group 2.

These incongruent blocks within IRMT0105 can also easily be shown and confirmed by a change in the phylogenetic relationship. Figure 3 is an example of strain HN1890. Similar changes were also observed for other strains listed above. Both the incongruent blocks and the changes in the phylogenetic relationship suggest that the strains mentioned above obtained a segment of DNA from another strain via recombination in their history.

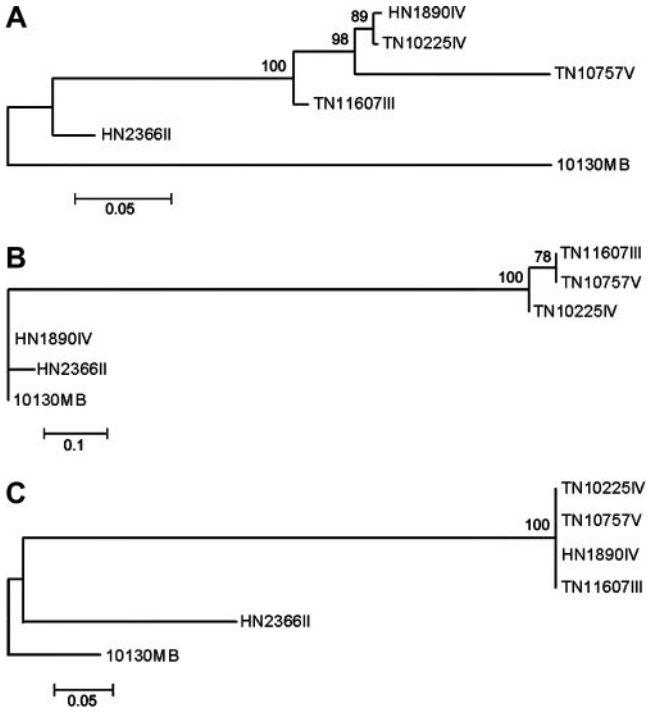

FIG. 3. Phylogenic trees of strains HN1890 (cluster IV), *M. bovis* TN10130 (10130MB), HN2366 (II), TN11607 (III), TN10225 (IV) and TN10757 (V) with *M. bovis* TN10130 as the root. (A) Neighbor-joining tree based on 363 background SNPs. (B) Neighbor-joining tree based on the first 24 iSNPs of IRMT0105. Bootstrap value 35 refers to the cluster of HN1890 and HN2366. (C) Neighbor-joining tree based on the last 22 iSNPs of IRMT0105. Numbers shown on the trees are bootstrap numbers. Assuming no recombination in background SNPs, tree A was supposed to show the true phylogenic relationship of the strains. Tree C was built with the second half of IRMT0105 iSNPs. The relationship shown is similar to that of tree A although with lower differentiating power. However, in tree B, HN1890 jumped from a close cluster with TN1075, TN11607, and TN10225 to the close cluster with HN2366 near the root. This dramatic change of phylogenic relationship suggested a recombination event.

Minimum of eight recombination events in IRMT0105 in data set II. Since no significant recombination signal was detected in the background SNPs, it is reasonable to start with the phylogenetic tree based on the background SNPs, which is assumed to represent the true phylogenetic relationship of the strains. Based on this tree, at least eight recombination events can be identified and located on the tree branches to explain the incongruent blocks in the data set (Fig. 4).

The DNA maximum parsimony tree was used to estimate the minimum number of mutations in the data set before and after any given recombination event (Table 2). The minimum number of mutations decreased from 572 (assuming no recombination ever happened in history) to 471 after the possible status of the eight recombination events was restored. In other words, eight recombination events compensated for about 100 backward mutations. For the 399 segregating sites, 471 mutations or 72 backward mutations are much more acceptable than 572 mutations or 173 backward mutations.

Possible recombination in the background SNPs of data set II. Using the DNA maximum parsimony tree method, it was estimated that about 34 backward mutations were in the back-

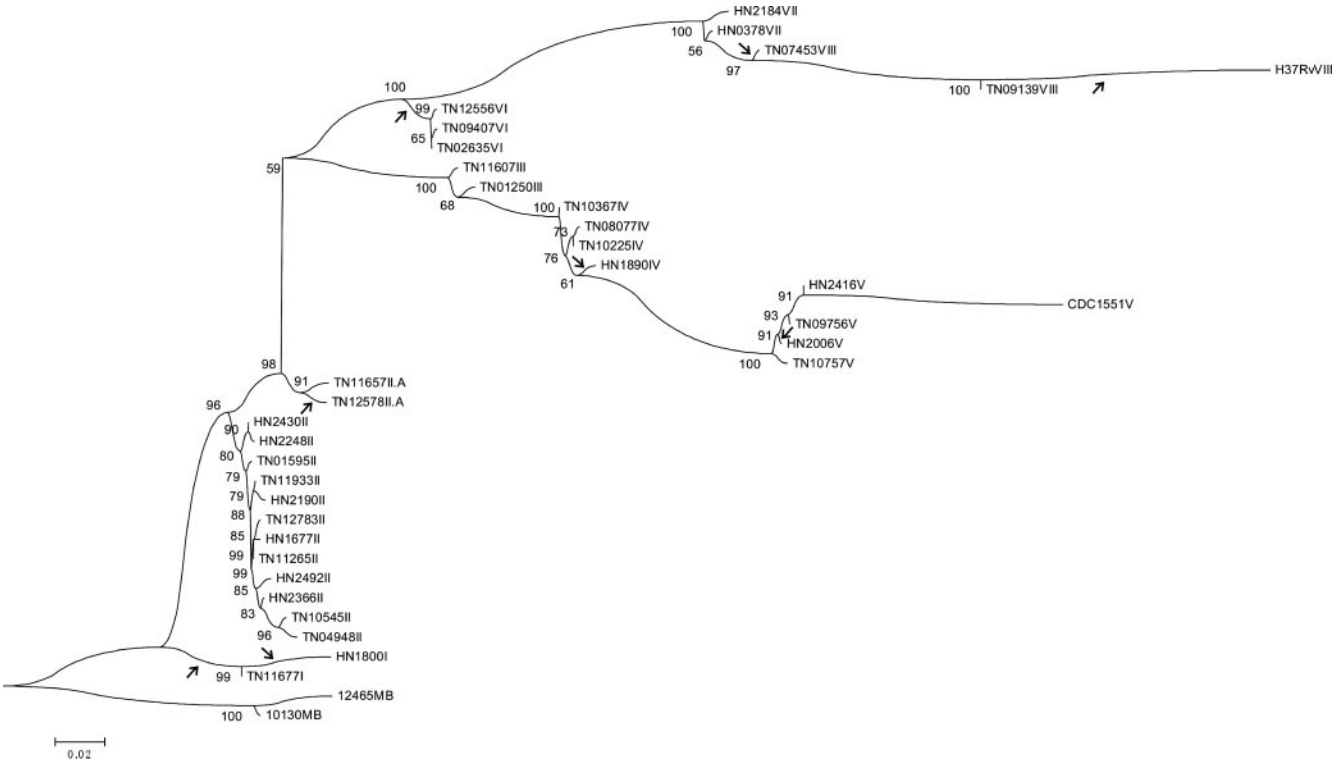

FIG. 4. Neighbor-joining tree based on background SNPs. Bootstrap numbers are shown on the tree. Arrows show the position of one possible scenario of eight possible recombination events.

ground SNPs. Given the fact that the ratio of transition to transversion is 2.07 (265 transitions with 128 transversions), the upper bound of the probability of observing no SNP with more than two alleles is about 1.5e−6. (This value of 1.5e−6 is obtained as follows. For a site experiencing double mutations, if the first mutation is a transition, the second mutation must also be a transition to make it a "backward" mutation, which has a conditional probability of 2.07/3.07. If the first mutation is a transversion, the "backward" mutation must also be a transversion and the mutated nucleotide must mutate back to the original nucleotide. Assuming the probabilities of one nucleotide transverse to the two destination nucleotides are equal, such "backward" mutation has a conditional probability of 0.5/3.07. Assuming all "backward" mutations have the same large conditional probability of 2.07/3.07, the probability of 34 independent backward mutations is calculated as follows: $[2.07/3.07]^{34} = 1.5e-6$. A simple calculation can also show that to make the probability of no triple allele a reasonable 0.01, the ratio of transition versus transversion should be at least 6.7). If the 25 transitions and 17 transversions in IRMT0105 are excluded, the ratio of transition to transversion is 2.16. Accordingly the probability is at most 2.4e−6. These low probabilities again suggest that backward mutation alone is not a probable explanation of the observation. Either recombination or transition hot spots may also contribute.

TABLE 2. Number of mutations with the eight possible recombination events shown in Fig. 4

No. of recombinations	Strain(s) with recombination	Replaced fragment[a]	No. of mutations needed
0			572
1	H37Rv	SNPs 28-36 from TN10757	563
2	TN07453	SNPs 28-36 from TN10757	554
3	Group VI[b]	SNPs 28-36 from TN10757	545
4	HN2006	SNPs 1-46 from TN10757	528
5	HN1890	SNPs 1-24 from TN10225	508
6	TN12578	SNPs 1-26 from TN10225	487
7	HN1800, TN11677	SNPs 21-46 from HN2366	486
8	TN11677	SNPs 1-20 from HN2366	471

[a] Incongruent SNP blocks were replaced by corresponding congruent SNPs from other strains to restore the status before recombination.
[b] TN02635, TN09407, and TN12556.

DISCUSSION

Recombination may occur at some *M. tuberculosis* loci. The estimated population recombination rate γ of ≈0.83 is very small. This result supported the mostly clonal growth of the *M. tuberculosis* population. On the other hand, there may exist recombination hot spots in the genome such as the IRMT0105 region. Although as a whole our findings do not contradict the assumption that the *M. tuberculosis* population is mostly clonal, they suggest that recombination events might be more common than previously thought, particularly considering that many recombinations will not leave any trace or will be quickly purged away by selection.

Gutacker et al. (15) reported three intergenic regions showing complicated iSNP patterns and suggesting that recombination may help to form such pattern. However, in their study

each intergenic region only has five iSNPs. Although the complicated pattern could be explained by recombination, the possibility that backward mutations alone shaped that pattern cannot be easily excluded. Our analysis of data set I showed that mutation alone can hardly explain the polymorphic pattern of the sSNPs. However, it is only an indirect support of the existence of recombination hot spots. The mosaic pattern found in IRMT0105 iSNPs of our data set II seems to be direct evidence of the existence of such a recombination hot spot. Because gene locus MT0105 belongs to a member of PPE gene family, if the hot spot hypothesis is true, it may help to explain the possible function of this gene family in escaping host immune surveillance. Interestingly, among the three intergenic regions reported by Gutacker et al. (15), one is close to gene locus Rv0980c, which is a PE-PGRS family protein gene. We think their finding is consistent with our hypothesis that there may be recombination hot spots within or close to PE and PPE gene family members.

The phylogenetic analysis of background SNPs and the IRMT0105 iSNPs also suggests that the detected incongruent SNP blocks are not relics of ancient recombinations from the progenitors of *M. tuberculosis* but recent events following the whole evolution of *M. tuberculosis* populations. As shown in Fig. 4, *M. tuberculosis* strains form clear genetic clusters (15–17), and all of them form a clade branching out of its ancestors. According to coalescence theory, this means that, based on their history, strains of *M. tuberculosis* found their most recent common ancestor before coalescing with any strains of their ancestors. Thus, polymorphism in the ancestral population of ancestors of *M. tuberculosis* is irrelevant to the polymorphism among strains of *M. tuberculosis*. Furthermore, as shown in Fig. 4, all eight recombination events used to explain the mosaic pattern of the IRMT0105 iSNPs are located at branches within the clade and clusters. Some recombinations shaped some of the isolates from genetic clusters VIII, V, and IV but not the others in the same cluster. These results suggest that these recombinations occurred after *M. tuberculosis* strains diverged from their ancestors and throughout the whole history of their further diverging into different genetic clusters.

Our hypothesis for the incongruent block of the IRMT0105 iSNPs is that it may be due to ancient recombination events between different *M. tuberculosis* strains (and/or ancient horizontal DNA transfer events via transduction). However, there is a possibility that the incongruent block is the product of ancient gene conversions of different segments of the same genome (43). Two other significant homologs of IRMT0105 were found in the complete genome sequences of *M. tuberculosis* strains CDC1551 and H37Rv with the BLAST program (28). However, none of them is in the near upstream or downstream regions. Our preliminary analysis showed a mosaic-like pattern in one of the homologs found in the CDC1551 genome, which leaves open the possibility of the hypothesis of gene conversion. However, based on the available data no definite conclusion can be made. The main purpose of this paper, though, is to provide working hypotheses based on our observations instead of giving definite explanations. Understanding the cause and molecular mechanism behind the SNP diversity demands extensive studies in the future.

Rationale for using background SNPs. Part of our analysis relies on having a correct phylogeny of the sample, and we chose to pool together part of the iSNPs and sSNPs as background SNPs to obtain such a phylogeny. The rationale for the pooling is that certain iSNP and sSNP are supposed to be selectively neutral or nearly so. This should be generally true for most cases. However, genome comparisons showed that the ratio of synonymous substitution versus nonsynonymous substitution in *M. tuberculosis* is about 1.6, which is quite low compared to other bacteria like *Escherichia coli* and *Salmonella enterica* (14). This observation suggested an increased selection pressure on synonymous substitutions or a decreased selection pressure on nonsynonymous and intergenic substitutions. Considering that *M. tuberculosis* may have emerged as a human pathogen as late as 10,000 to 15,000 years ago, evolutionarily recent global dissemination supports the second explanation; that is, selection did not have sufficient time to purge nonsynonymous and intergenic substitutions (16, 22, 49). So, assuming a similar selection pressure also on certain iSNPs, which we used as background SNPs together with sSNPs, is still reasonable in this case.

Our results suggest that there could be some recombination in the background SNPs although recombination detection algorithms may not be able to detect such events. The hidden recombinations may make the phylogenetic relationship of strains less clear and the recombination signal less significant, especially for phylogenies based on recombination detection algorithms. In other words, the hidden recombinations can only increase the false-negative rate but will not increase the false-positive rate. They only make the detection more conservative but are not able to invalidate detected significant recombination. In this case, hidden recombination in background SNPs should not have large effects, given the fact that high bootstrap values suggested that the phylogenetic relationship is quite reliable (Fig. 4).

Low SNP density may explain the lack of recombination detection in background SNPs. To examine the effect of the sampled SNP density on the strength of the signal, we resampled the IRMT0105 iSNPs and analyzed the partial data. Four sampling schemes were used corresponding to randomly removing, respectively, 10, 20, 30, and 40 SNPs from the original data. Each scheme was repeated 10 times. The RDP, GENECONV, and maximum chi-square algorithms were used to detect recombination in each resulting data set. The algorithms could still pick up some strong signals when the SNP number was reduced to 26 (removing 20 SNPs). When the SNP number was reduced to 16, the algorithms failed to detect any signal from one sampled data set. When the SNP number was reduced to 6, the algorithms could not detect any signal from any of the sampled data sets. These results suggest that the high density of polymorphic markers is crucial in detecting recombination, which must be considered in the experimental design of any future population genetics study of *M. tuberculosis*.

Alternative possibilities to explain the incongruent blocks. A minimum of eight recombinations was needed to explain all the incongruent blocks in data set II. Figure 4 only shows one of the possible scenarios. There are other possibilities. For example, two recombinations were needed to explain the incongruent blocks of the two cluster I strains, HN1800 and TN11677. One recombination may occur on the branch leading to the most recent common ancestor of these two strains (in-

ternal branch), and the other one may occur on the branch leading to one of the two strains (either HN1800 or TN11677) after their divergence (external branch). Or both of the recombinations may occur on the external branches, that is, one on the branch leading to HN1800 and the other on the branch leading to TN11677.

As previously mentioned, 37 strains can be divided into two groups according to the polymorphic pattern of IRMT0105. This is consistent with the phylogeny based on the background SNPs: one group includes *M. bovis* and clusters I and II and the other one includes clusters III to VIII. This puzzling polymorphic pattern change is very likely to be the result of an ancient recombination event instead of more than 30 mutations. However, because the donor of this recombination is unknown, we did not include it in the detected recombinations. Additionally, because this donor was absent in the phylogenetic tree, the advantage of compensating multiple mutations with the eight recombinations was actually underestimated (in this case, the mutation number compensated by recombination on the most recent common ancestor of HN1800 and TN11677 in Table 2 was underestimated). Gutierrez et al. (17) showed that *M. tuberculosis* has evolved from a broader progenitor species of smooth tubercle bacilli, which may experience recombination in history. Given that the smooth strains may hold the key of the puzzling polymorphic pattern change, we think it will be very interesting to examine the polymorphic pattern of IRMT0105 of these smooth strains in the future.

In conclusion, the incongruent SNP blocks discovered in this study indicate strongly that multiple recombinations have occurred in the intergenic region ahead of the MT0105 locus in the *M. tuberculosis* genome. Hypothetically, these recombination events could be due to either gene conversions from the same genome or recombination (or horizontal DNA transfer) between different strains. Since the MT0105 locus encodes a PPE protein, which may be critical in host-pathogen interactions, we further hypothesize that recombination at hot spots near PE or PPE gene families has been an important mechanism for *M. tuberculosis* to escape immune surveillance. Further genetic study of other genes encoding PPE proteins will help us to understand the potential important roles of these proteins.

ACKNOWLEDGMENTS

This study was supported in part by NIH grant R01 GM6077 (Y.F.).

We thank two anonymous reviewers for their thoughtful comments and suggestions. We thank Sara Barton for assistance for manuscript preparation.

REFERENCES

1. **Bafna, V., and V. Bansal.** 2006. Inference about recombination from haplotype data: lower bounds and recombination hot spots. J. Comput. Biol. **13:**501–521.
2. **Braden, C. R., G. P. Morlock, C. L. Woodley, K. R. Johnson, A. C. Colombel, M. D. Cave, Z. Yang, S. E. Valway, I. M. Onorato, and J. T. Crawford.** 2001. Simultaneous infection with multiple strains of *Mycobacterium tuberculosis*. Clin. Infect. Dis. **33:**e42–e47.
3. **Choudhary, R. K., S. Mukhopadhyay, P. Chakhaiyar, N. Sharma, K. J. R. Murthy, V. M. Katoch, and S. E. Hasnain.** 2003. PPE antigen Rv2430c of *Mycobacterium tuberculosis* induces a strong B-cell response. Infect. Immun. **71:**6338–6343.
4. **Cole, S. T.** 2002. Comparative and functional genomics of the *Mycobacterium tuberculosis* complex. Microbiology **148:**2919–2928.
5. **Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M. A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead, and B. G. Barrell.** 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature **393:**537–544.
6. **Cooper, B. S.** 2001. Pathogen population dynamics: the age of the strain. Trends Microbiol. **9:**199–200.
7. **Das, S., S. Narayanan, L. Hari, N. S. Mohan, S. Somasundaram, N. Selvakumar, and P. R. Narayanan.** 2004. Simultaneous infection with multiple strains of *Mycobacterium tuberculosis* identified by restriction fragment length polymorphism analysis. Int. J. Tuberc. Lung Dis. **8:**267–270.
8. **Efron, B., E. Halloran, and S. Holmes.** 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. USA **93:**13429–13434.
9. **Feil, E. J., E. C. Holmes, D. E. Bessen, M. S. Chan, N. P. Day, M. C. Enright, R. Goldstein, D. W. Hood, A. Kalia, C. E. Moore, J. Zhou, and B. G. Spratt.** 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.
10. **Felsenstein, J.** 1989. PHYLIP: phylogeny inference package (version 3.2). Cladistics **5:**164–166.
11. **Felsenstein, J., and H. Kishino.** 1993. Is there something wrong with the bootstrap on phylogenies—a reply. Syst. Biol. **42:**193–200.
12. **Fitch, W. M.** 1977. Problem of discovering most parsimonious tree. Am. Nat. **111:**223–257.
13. **Fitch, W. M.** 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20:**406–416.
14. **Fleischmann, R. D., D. Alland, J. A. Eisen, L. Carpenter, O. White, J. Peterson, R. DeBoy, R. Dodson, M. Gwinn, D. Haft, E. Hickey, J. F. Kolonay, W. C. Nelson, L. A. Umayam, M. Ermolaeva, S. L. Salzberg, A. Delcher, T. Utterback, J. Weidman, H. Khouri, J. Gill, A. Mikula, W. Bishai, W. R. Jacobs, Jr., J. C. Venter, and C. M. Fraser.** 2002. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J. Bacteriol. **184:**5479–5490.
15. **Gutacker, M. M., B. Mathema, H. Soini, E. Shashkina, B. N. Kreiswirth, E. A. Graviss, and J. M. Musser.** 2006. Single-nucleotide polymorphism-based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. J. Infect. Dis. **193:**121–128.
16. **Gutacker, M. M., J. C. Smoot, C. A. L. Migliaccio, S. M. Ricklefs, S. Hua, D. V. Cousins, E. A. Graviss, E. Shashkina, B. N. Kreiswirth, and J. M. Musser.** 2002. Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. Genetics **162:**1533–1543.
17. **Gutierrez, M. C., S. Brisse, R. Brosch, M. Fabre, B. Omais, M. Marmiesse, P. Supply, and V. Vincent.** 2005. Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. PLOS Pathog. **1:**e5.
18. **Hatfull, G. F., M. L. Pedulla, D. Jacobs-Sera, P. M. Cichon, A. Foley, M. E. Ford, R. M. Gonda, J. M. Houtz, A. J. Hryckowian, V. A. Kelchner, S. Namburi, K. V. Pajcini, M. G. Popovich, D. T. Schleicher, B. Z. Simanek, A. L. Smith, G. M. Zdanowicz, V. Kumar, C. L. Peebles, W. R. Jacobs, Jr., J. G. Lawrence, and R. W. Hendrix.** 2006. Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. PLOS Genet. **2:**e92.
19. **Hillis, D. M., and J. J. Bull.** 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. **42:**182–192.
20. **Hudson, R. R., and N. L. Kaplan.** 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:**147–164.
21. **Jakobsen, I. B., and S. Easteal.** 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Comput. Appl. Biosci. **12:**291–295.
22. **Kapur, V., T. S. Whittam, and J. M. Musser.** 1994. Is *Mycobacterium tuberculosis* 15,000 years old? J. Infect. Dis. **170:**1348–1349.
23. **Kumar, S., K. Tamura, and M. Nei.** 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief. Bioinform. **5:**150–163.
24. **Li, Y., E. Miltner, M. Wu, M. Petrofsky, and L. E. Bermudez.** 2005. A *Mycobacterium avium* PPE gene is associated with the ability of the bacterium to grow in macrophages and virulence in mice. Cell. Microbiol. **7:**539–548.
25. **Martin, D., and E. Rybicki.** 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics **16:**562–563.
26. **Maynard Smith, J.** 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. **34:**126–129.
27. **McDonald, B. A., and C. Linde.** 2002. Pathogen population genetics, evolutionary potential, and durable resistance. Annu. Rev. Phytopathol. **40:**349–379.
28. **McGinnis, S., and T. L. Madden.** 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. **32:**W20–W25.

29. **McVean, G., P. Awadalla, and P. Fearnhead.** 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160:**1231–1241.
30. **Musser, J. M.** 1996. Molecular population genetic analysis of emerged bacterial pathogens: selected insights. Emerg. Infect. Dis. **2:**1–17.
31. **Myers, S. R., and R. C. Griffiths.** 2003. Bounds on the minimum number of recombination events in a sample history. Genetics **163:**375–394.
32. **Nei, M., and S. Kumar.** 2000. Molecular evolution and phylogenetics. Oxford University Press, New York, N.Y.
33. **Newton, M. A.** 1996. Bootstrapping phylogenies: large deviations and dispersion effects. Biometrika **83:**315–328.
34. **Okkels, L. M., I. Brock, F. Follmann, E. M. Agger, S. M. Arend, T. H. M. Ottenhoff, F. Oftung, I. Rosenkrands, and P. Andersen.** 2003. PPE protein (Rv3873) from DNA segment RD1 of *Mycobacterium tuberculosis*: strong recognition of both specific T-cell epitopes and epitopes conserved within the PPE family. Infect. Immun. **71:**6116–6123.
35. **Parsons, L. M., C. S. Jankowski, and K. M. Derbyshire.** 1998. Conjugal transfer of chromosomal DNA in *Mycobacterium smegmatis*. Mol. Microbiol. **28:**571–582.
36. **Pavlic, M., F. Allerberger, M. P. Dierich, and W. M. Prodinger.** 1999. Simultaneous infection with two drug-susceptible *Mycobacterium tuberculosis* strains in an immunocompetent host. J. Clin. Microbiol. **37:**4156–4157.
37. **Pedulla, M. L., M. E. Ford, J. M. Houtz, T. Karthikeyan, C. Wadsworth, J. A. Lewis, D. Jacobs-Sera, J. Falbo, J. Gross, N. R. Pannunzio, W. Brucker, V. Kumar, J. Kandasamy, L. Keenan, S. Bardarov, J. Kriakov, J. G. Lawrence, W. R. Jacobs, Jr., R. W. Hendrix, and G. F. Hatfull.** 2003. Origins of highly mosaic mycobacteriophage genomes. Cell **113:**171–182.
38. **Posada, D.** 2002. Evaluation of methods for detecting recombination from DNA sequences: empirical data. Mol. Biol. Evol. **19:**708–717.
39. **Posada, D., and K. A. Crandall.** 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc. Natl. Acad. Sci. USA **98:**13757–13762.
40. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406–425.
41. **Salminen, M. O., J. K. Carr, D. S. Burke, and F. E. McCutchan.** 1995. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Res. Hum. Retrovir. **11:**1423–1425.
42. **Sanderson, M. J., and M. F. Wojciechowski.** 2000. Improved bootstrap confidence limits in large-scale phylogenies, with an example from Neo-Astragalus (leguminosae). Syst. Biol. **49:**671–685.
43. **Sawyer, S.** 1989. Statistical tests for detecting gene conversion. Mol. Biol. Evol. **6:**526–538.
44. **Shafer, R. W., S. P. Singh, C. Larkin, and P. M. Small.** 1995. Exogenous reinfection with multidrug-resistant *Mycobacterium tuberculosis* in an immunocompetent patient. Tuber. Lung Dis. **76:**575–577.
45. **Smith, J. M., E. J. Feil, and N. H. Smith.** 2000. Population structure and evolutionary dynamics of pathogenic bacteria. Bioessays **22:**1115–1122.
46. **Smith, J. M., N. H. Smith, M. O'Rourke, and B. G. Spratt.** 1993. How clonal are bacteria? Proc. Natl. Acad. Sci. USA **90:**4384–4388.
47. **Smith, N. H., J. Dale, J. Inwald, S. Palmer, S. V. Gordon, R. G. Hewinson, and J. M. Smith.** 2003. The population structure of *Mycobacterium bovis* in Great Britain: clonal expansion. Proc. Natl. Acad. Sci. USA **100:**15271–15275.
48. **Spratt, B. G., and M. C. Maiden.** 1999. Bacterial population genetics, evolution and epidemiology. Philos. Trans. R. Soc. Lond. B **354:**701–710.
49. **Sreevatsan, S., X. Pan, K. E. Stockbauer, N. D. Connell, B. N. Kreiswirth, T. S. Whittam, and J. M. Musser.** 1997. Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. Proc. Natl. Acad. Sci. USA **94:**9869–9874.
50. **Studier, J. A., and K. J. Keppler.** 1988. A note on the neighbor-joining algorithm of Saitou and Nei. Mol. Biol. Evol. **5:**729–731.
51. **Supply, P., R. M. Warren, A.-L. Banuls, S. Lesjean, G. D. V. D. Spuy, L.-A. Lewis, M. Tibayrenc, P. D. V. Helden, and C. Locht.** 2003. Linkage disequilibrium between minisatellite loci supports clonal evolution of *Mycobacterium tuberculosis* in a high tuberculosis incidence area. Mol. Microbiol. **47:**529–538.
52. **Theisen, A., C. Reichel, S. R. Gerdes, W. H. Haas, J. K. Rockstroh, U. Spengler, and T. Sauerbruch.** 1995. Mixed-strain infection with a drug-sensitive and multidrug-resistant strain of *Mycobacterium tuberculosis*. Lancet **345:**1512.
53. **Wang, J., L. M. Parsons, and K. M. Derbyshire.** 2003. Unconventional conjugal DNA transfer in mycobacteria. Nat. Genet. **34:**80–84.
54. **Wiuf, C., T. Christensen, and J. Hein.** 2001. A simulation study of the reliability of recombination detection methods. Mol. Biol. Evol. **18:**1929–1939.
55. **World Health Organization.** 2004. WHO fact sheet on tuberculosis. World Health Organization. [Online.] http://www.who.int/mediacentre/factsheets/fs104/en/index.html.
56. **Yeh, R. W., P. C. Hopewell, and C. L. Daley.** 1999. Simultaneous infection with two strains of *Mycobacterium tuberculosis* identified by restriction fragment length polymorphism analysis. Int. J. Tuberc. Lung Dis. **3:**537–539.
57. **Zharkikh, A., and W. H. Li.** 1995. Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique. Mol. Phylogenet. Evol. **4:**44–63.
58. **Zharkikh, A., and W. H. Li.** 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. Mol. Biol. Evol. **9:**1119–1147.