Comparative Sequence Analysis of *Mycobacterium leprae* and the New Leprosy-Causing *Mycobacterium lepromatosis*

Xiang Y. Han,¹ Kurt C. Sizer,¹ Erika J. Thompson,² Juma Kabanja,³ Jun Li,³ Peter Hu,³ Laura Gómez-Valero,⁴ and Francisco J. Silva⁵,⁶,*

Department of Laboratory Medicine,¹ DNA Analysis Core Facility,² and School of Health Sciences,³ The University of Texas M. D. Anderson Cancer Center, Houston, Texas; Institut Pasteur, Unité de Biologie des Bactéries Intracellulaires, Paris, France;⁴ Institut Cavanilles de Biodiversitat i Biologia Evolutiva, Universitat de València, València, Spain; and CIBER en Epidemiología y Salud Pública (CIBERESP), Madrid, Spain⁶

Received 12 June 2009/Accepted 20 July 2009

*Mycobacterium lepromatosis* is a newly discovered leprosy-causing organism. Preliminary phylogenetic analysis of its 16S rRNA gene and a few other gene segments revealed significant divergence from *Mycobacterium leprae*, a well-known cause of leprosy, that justifies the status of *M. lepromatosis* as a new species. In this study we analyzed the sequences of 20 genes and pseudogenes (22,814 nucleotides). Overall, the level of matching of these sequences with *M. leprae* sequences was 90.9%, which substantiated the species-level difference; the levels of matching for the 16S rRNA genes and 14 protein-encoding genes were 98.0% and 93.1%, respectively, but the level of matching for five pseudogenes was only 79.1%. Five conserved protein-encoding genes were selected to construct phylogenetic trees and to calculate the numbers of synonymous substitutions (dS values) and nonsynonymous substitutions (dN values) in the two species. Robust phylogenetic trees constructed using concatenated alignment of these genes placed *M. lepromatosis* and *M. leprae* in a tight cluster with long terminal branches, implying that the divergence occurred long ago. The dS and dN values were also much higher than those for other closest pairs of mycobacteria. The dS values were 14 to 28% of the dS values for *M. leprae* and *Mycobacterium tuberculosis*, a more divergent pair of species. These results thus indicate that *M. lepromatosis* and *M. leprae* diverged ~10 million years ago. The *M. lepromatosis* pseudogenes analyzed that were also pseudogenes in *M. leprae* showed nearly neutral evolution, and their relative ages were similar to those of *M. leprae* pseudogenes, suggesting that they were pseudogenes before divergence. Taken together, the results described above indicate that *M. lepromatosis* and *M. leprae* diverged from a common ancestor after the massive gene inactivation event described previously for *M. leprae*.

Leprosy, one of the oldest human diseases, remains a significant public health problem in many developing countries (8). *Mycobacterium leprae* was the only known cause of leprosy until recently, when a new mycobacterium, *Mycobacterium lepromatosis*, was found to be the cause of diffuse lepromatous leprosy (DLL), a unique form of leprosy endemic in Mexico and the Caribbean (17). The discovery of this new species may provide an explanation for the clinical and geographic variability of leprosy.

The initial phylogenetic analysis of *M. lepromatosis* was carried out using the sequences of the 16S rRNA gene and segments of *groEL*, *rpoB*, and other genes (total, 4.99 kb) (17). This study revealed significant sequence differences between *M. lepromatosis* and all known *Mycobacterium* species and placed *M. lepromatosis* closest to *M. leprae*. However, the sequence variation justified assigning a new species for the new organism instead of classifying it as a variant of *M. leprae*. All *M. leprae* strains collected worldwide have been found to be clonal and to differ by only single-nucleotide polymorphism or variable numbers of tandem repeats (24). Also, the genomes of two *M. leprae* strains, strain TN from India (GenBank accession numbers AL583917 to AL583926) (4) and strain Br4923 from Brazil (GenBank accession number FM211192) (N. Honore et al., unpublished data) share 99.98% identity.

Like *M. leprae*, *M. lepromatosis* has not been cultivated on artificial media. In addition, our previous study also showed other similarities between these organisms, such as degeneration of *mmaA3* into a pseudogene, the presence of unique AT-rich inserted sequences in the 16S rRNA gene, identical six-base tandem repeats in *rpoT*, similar G+C contents, and great evolutionary distance from other mycobacteria (17).

The *M. leprae* genome (3.3 Mb) is much smaller than the *Mycobacterium tuberculosis* genome (4.4 Mb) (3, 4). More intriguingly, the *M. leprae* genome has undergone reductive evolution; ~40% of the genes are inactivated (4), and ~50% of the genes of the last common ancestor of *M. leprae* and *M. tuberculosis* have been lost (13). On the other hand, the *M. leprae* genome has been far more stable than the *M. tuberculosis* genome, and the worldwide clonality of the *M. leprae* strains paralleled the global spread of *M. leprae* strains that occurred via human activity and migration during the last ~100,000 years (24). Recently, by comparing the genomes of *M. leprae* and *M. tuberculosis* and by analyzing the ages of the *M. leprae* pseudogenes, Gomez-Valero et al. (13) estimated that a massive gene inactivation event took place in the *M. leprae* genome in the last 20 million years.

The discovery of *M. lepromatosis* and its differences from *M. leprae* make it relevant for further study for diagnosis, treat-
ment, and prevention of DLL. Likewise, the many similarities between these two organisms prompted questions about their evolutionary histories and about how *M. lepromatosis* became endemic mainly in Mexico, while *M. leprae* occurs worldwide. In this study, we extended and refined our previous phylogenetic study by determining and analyzing the sequences of 20 genes and pseudogenes of *M. lepromatosis*. Our findings solidified the phylogeny of this new organism and provided new insights into the history of pseudogenes.

**MATERIALS AND METHODS**

**DNA isolation and sequencing.** DNA was extracted from the autopsy liver of a patient who died of DLL (17). Briefly, approximately 200 mg of freshly frozen infected liver tissue was homogenized and processed with N-acetyl-l-cysteine and NaOH to enrich the mycobacterium, decontaminate the preparation, and disrupt the host cells, the usual procedure for mycobacterial culture. After essentially pure mycobacterial cells were obtained, a commercial DNA extraction solution (PrepMan Ultra; Applied Biosystems, CA) was added, and the mixture was boiled to disrupt the bacterial cells. The supernatant contained the released DNA and was used for PCRs.

Many sets of PCR primers were designed to target various genes and pseudogenes of *M. lepromatosis*. Without knowledge of the specific sequences of this organism, the primers were designed using the conserved sequences of the corresponding genes of *M. leprae* and *M. tuberculosis* after alignment. For a long gene, multiple sets of forward and reverse primers were designed, with each set amplifying 500 to 600 bp, with overlaps. The PCR amplicons were sequenced directly using an ABI sequencer. All sequences were matched with GenBank data and inspected visually to ensure accuracy.

The full-length *rpoB* gene of Mycobacterium haemophilum ATCC 29548 was also sequenced for phylogenetic analysis.

**Retrieval of other sequences.** Gene and pseudogene sequences of *M. lepromatosis* were used to extract putative orthologous genes from completely sequenced mycobacterial genomes. The extraction was performed using the Microbial Genome Database for Comparative Analysis (30). The genomes used included the full-length *rgs* genes from *M. leprae* and *M. tuberculosis* after alignment. A long gene, multiple sets of forward and reverse primers were designed, with each set amplifying 500 to 600 bp, with overlaps. The PCR amplicons were sequenced directly using an ABI sequencer. All sequences were matched with GenBank data and inspected visually to ensure accuracy.

The full-length *rpoB* gene of *M. tuberculosis* was also sequenced for phylogenetic analysis.

**Sequence alignment.** The MAFFT program (18) was used to align nucleotide sequences of genes and pseudogenes. The ClustalW program implemented in MEGA 3.1 (19) was used to align the amino acid sequences of protein-encoding genes and sequences of protein-encoding genes for the corresponding nucleotide alignments. Each alignment was visually inspected to detect short misalignments or poorly aligned regions. Doubtful regions, usually at both ends, were trimmed.

To calculate the numbers of synonymous (dS) and nonsynonymous (dN) nucleotide substitutions per site in five *M. lepromatosis* pseudogenes (cita, gnd2, icd1, lld1, and mmaA3), alignment with corresponding *M. leprae* pseudogenes was performed using the following parameters: (i) the alignment length corresponded to the available length of each *M. lepromatosis* pseudogene; (ii) to maintain continuous reading frames, like those in protein-encoding genes, the nucleotides at triplet codon positions covering any indel in either organism were not counted; and (iii) nucleotide positions at which there were stop codons in either organism were removed. The final lengths of the aligned regions, with <15% reduction of the original alignments, were 918, 852, 552, 483, and 540 nucleotides for *cita*, *gnd2*, *icd1*, *lld1*, and *mmaA3*, respectively.

**Phylogenetic and relative rate tests.** To determine the relationship of *M. lepromatosis* to other mycobacteria, phylogenetic trees were constructed using a concatenated alignment of five conserved genes. These genes, *rpoB*, *l1gA*, *groEL*, *gnd1*, and *bfbA*, were chosen from 14 protein-encoding genes based on the following criteria: functional importance or conservation, the presence of orthologous genes in the 17 completely sequenced mycobacterial genomes, full length or nearly full length, and alignment accuracy. In particular, *rpoB* and *groEL* have frequently been used in phylogenetic studies (6). The alignments were also visually inspected to trim more variably aligned ends (on average, ~9% of the alignments). The final lengths of the *rpoB*, *l1gA*, *groEL*, *gnd1*, and *bfbA* genes were 3,429, 2,013, 1,575, 1,125, and 405 nucleotides, respectively, and there was total concatenation of 8,547 nucleotide sites.

A maximum likelihood phylogenetic analysis was performed with the concatenated alignment using the program PHYML (16) with 100 bootstrap replicates and the HKY model. An additional neighbor-joining phylogeny was also obtained with the program MEGA 3.1 (19) with the Tamura-Nei model and 1,000 bootstrap replicates. To include *M. haemophilum* (without genome data) in the analysis, additional phylogenies with the same conditions were constructed using the full-length *rpoB* gene of this organism. Tajima’s relative rates tests (28) were implemented in MEGA (19). Both *M. leprae* and *M. lepromatosis* branches were compared with *M. haemophilum* using *M. tuberculosis* CDC1551 as the outgroup. The analyses included the first and second codon positions (usually associated with nonsynonymous substitutions) and the third codon positions (usually resulting in synonymous substitutions due to codon wobbling).

**Estimation of dN and dS values.** The dS and dN values, along with standard errors, were estimated for gene–gene, gene–pseudogene, or pseudogene–pseudogene pairs using yrt00 from PAML (32). When a pseudogene was involved, dS and dN values were used only for comparison, and no protein-encoding function of the pseudogene was indicated.

To estimate the relative age of an *M. lepromatosis* pseudogene, the formula for the *p* parameter developed for *M. leprae* (13) was used:

\[ p = \frac{dN_{\text{an}} - f(dN_{\text{i}})}{(dS_{\text{an}} - f(dS_{\text{i}}))} \]

where *dN* and *dS* are the dN and dS values for the ancestor (i) of *M. tuberculosis* and *M. leprae* and the present pseudogene (prs) of *M. leprae* (or *M. lepromatosis*), *dN* is the dN value for the same ancestor (i) and the gene of *M. tuberculosis* and (dN{prs}) is a function that makes a small correction for the expected number of substitutions in the lineage as if the pseudogene had not formed and is estimated from the complete set of pairs of orthologous genes for *M. tuberculosis* and *M. leprae*. This formula is based on the fact that the dN or dS value for a gene–pseudogene pair (for example, an *M. tuberculosi* gene and an *M. leprae* pseudogene) is the sum for the period of evolution as a gene (for the complete lineage of *M. tuberculosis* and part of the lineage of *M. leprae*) and the period of evolution as a pseudogene (which depends on the moment of inactivation in the *M. leprae* lineage). It also takes into consideration the fact that the rates of evolution for nonsynonymous changes are gene specific and that to determine these rates in *M. leprae* (or *M. lepromatosis*), the *dN* for other metabolic genes or pseudogenes in the *M. leprae* genome was used. For the substitution values, *M. avium* subsp. *paratuberculosis* was used as an outgroup.

**Nucleotide sequence accession numbers.** The *M. lepromatosis* nucleotide sequences have been deposited in the GenBank database under the accession numbers shown in Table 1. The GenBank accession number for the nucleotide sequence of the full-length *rpoB* gene of *M. haemophilum* ATCC 29548 is ГУ254966.

**RESULTS**

**M. lepromatosis sequences and matches with *M. leprae*.** The 20 genes and pseudogenes (22,814 nucleotides) from *M. lepromatosis* and their matches with *M. leprae* genes and pseudogenes are shown in Table 1. These genes and pseudogenes were chosen from the eight initial sequences (17) with extension of the 16S rRNA, *rpoB*, *groEL*, and *groT* genes and 16 other metabolic genes or pseudogenes in the *M. leprae* and *M. tuberculosis* genomes as previously described (4). Of the 16 new genes or pseudogenes, 12 were successfully amplified and sequenced, but 4 pseudogenes (*l1gB*, *ummA*, *bfbR*, and *pfkB*) were not amplified and sequenced despite many attempts and redesign of the primers, implying that there is too much variation or these pseudogenes are not present in *M. lepromatosis*. The G+C content of the nucleotide sequences was 58.6%, similar to the 58.8% for the corresponding *M. leprae* sequences or the *M. leprae* genome (57.8%) (4).

The average levels of nucleotide identity between *M. lepromatosis* and *M. leprae* were 93.1% and 79.1% for the coding genes and pseudogenes, respectively, and the overall level of
nucleotide identity was 90.9% (Table 1). This level is well below the 95% genome-level cutoff used to distinguish species (15). In addition, the level of nucleotide identity for the rpoB gene was 94.5%, which is also less than the 98% cutoff proposed recently (1). Therefore, M. lepromatosis meets the definition of a species.

The levels of identity for the protein-encoding genes were 88.0 to 94.6% at the nucleotide level and 79.0 to 98.0% at the amino acid level (overall level of identity, 95.2%). Thirteen of the 14 genes had a 128-bp deletion in the middle. Therefore, in these pseudogenes, the lack of constraints of reading frames allowed much faster accumulation of indels.

Finally, the level of identity for the 16S rRNA gene was 98.0%, which was the highest level of identity, but there were 10 gap sites (0.65% of 1,542 bp). In all mycobacteria, 16S rRNA genes are highly conserved, and there are relatively few indels or indels are not present in closely related mycobacteria, such as M. marinum and M. avium and Mycobacterium intracellulare, and other pairs of species (data not shown). Thus, indel mutations are also a feature of 16S rRNA gene evolution in the two leprosy organisms. Other 16S rRNA gene features were analyzed previously (17).

**Phylogenetic position of M. lepromatosis.** As shown in Fig. 1A, phylogenetic trees constructed by using the maximum likelihood and neighbor-joining methods produced the same topology with bootstrap values of 100 at all nodes except one. M. leprae and M. lepromatosis were the most closely related species, yet in both trees the terminal branches leading to these species were much longer than those leading to other tight groups, such as the M. tuberculosis complex and the M. avium group. These long branches could not be explained by faster evolution or high substitution rates. On average, a gene evolving in the lineage leading to M. leprae had a slightly higher dS value than a gene in the M. tuberculosis lineage (from the last common ancestor) (0.82 and 0.72, respectively) (10), while the average dN values were 0.065 and 0.045, respectively (13). Because the dS values are similar and their contribution to
branch lengths is much more important than that of $dN$ values, we believe that a much earlier divergence time is the best explanation.

Because *M. haemophilum* was recently reported to be a close relative of *M. leprae* (22) and *M. lepromatosis* (17), phylogenetic trees based on *rpoB* (3,429 bp) were also reconstructed (Fig. 1B). The same topology was obtained despite some lower bootstrap values. It was observed that *M. haemophilum* was much closer to the leprosy organisms than to other mycobacteria. Yet the branch lengths indicated that there was faster evolution of the leprosy mycobacteria than of *M. haemophilum*. Using Tajima’s relative rate tests for this gene, the branches leading to *M. leprae* and *M. lepromatosis* were found to be significantly longer than the branch leading to *M. haemophilum* for the third codon position in both leprosy organisms ($P<0.001$ for both) and for the first and second codon positions in *M. lepromatosis* ($P = 0.042$). The main reason for this difference was the decrease in the G+C content, mainly at the third codon positions (Table 2).

A phylogeny was also obtained using the 16S rRNA gene sequences, which resulted in a different topology (Fig. 1C); the

**FIG. 1.** Maximum likelihood phylogenies of selected mycobacteria based on (A) concatenated alignment of the *rpoB*, *ligA*, *groEL*, *gnd1*, and *bfrA* genes, (B) alignment of the *rpoB* gene, and (C) alignment of the 16S rRNA gene. Bootstrap values are indicated at the nodes. At some nodes there are two values, a value obtained by the neighbor-joining method (left value) and a value obtained by the maximum likelihood method (right value). The bars indicate the numbers of nucleotide substitutions per site. The trees were rooted to separate the slow-growing mycobacterial clade (A and B) and the *M. avium* branch (C).

**TABLE 2.** G+C contents of *rpoB* gene

<table>
<thead>
<tr>
<th>Species</th>
<th>Codon positions 1 and 2</th>
<th>Codon position 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>M. lepromatosis</em></td>
<td>51.4</td>
<td>77.9</td>
<td>60.2</td>
</tr>
<tr>
<td><em>M. leprae</em></td>
<td>51.9</td>
<td>79.1</td>
<td>61.0</td>
</tr>
<tr>
<td><em>M. haemophilum</em></td>
<td>52.7</td>
<td>88.0</td>
<td>64.4</td>
</tr>
<tr>
<td><em>M. tuberculosis</em></td>
<td>53.0</td>
<td>86.9</td>
<td>64.3</td>
</tr>
</tbody>
</table>

* The first- and/or second-codon substitutions are frequently nonsynonymous, whereas the third-codon substitutions are usually synonymous due to codon wobbling.
branch between the leprosy organisms was much longer, and the M. marinum-M. ulcerans clade was closer to M. tuberculosis than to the clade formed by M. haemophilum and the leprosy organisms. This tree was less realistic than the tree obtained using the concatenated protein-encoding genes (Fig. 1A). Phylogenomic studies have shown that in a gene phylogeny, when one lineage evolves much faster than the other lineages, an incorrect topology may be obtained, in which the fast-evolving branch has a tendency to move toward the base of the tree. Thus, the true topology may be obtained only with a small subset of genes that evolve slowly (2). The 16S rRNA genes of the leprosy organisms evolved much faster than those of other mycobacteria, as shown by more indels, as noted above and previously (17).

Evolutionary distance between M. lepromatosis and M. leprae. To estimate the evolutionary distance between M. lepromatosis and M. leprae, dS and dN values were calculated using the five conserved genes described above. These values were compared with the dS and dN values for other pairs of most closely related species or subspecies, including M. avium subspecies, M. marinum and M. ulcerans, and M. tuberculosis and M. bovis. The more divergent pair, M. tuberculosis and M. leprae was also included as a gauge because these most medically important mycobacteria diverged ~66 million years ago (13). The results are shown in Fig. 2.

Since they are nearly unaffected by natural selection, the dS values are similar for many genes and thus more suitable for estimating relative divergence times. The dS values for M. tuberculosis and M. bovis were zero, consistent with the clade-level divergence of these species and a genome level of identity of 99.95% (7). Similarly, subspecies-level differences for the M. marinum-M. ulcerans and M. avium subspecies pairs were reflected by low dS values. Conversely, for M. lepromatosis and M. leprae the dS values were far higher and for the five genes ranged from 14 to 29% (average, ~20%) of the values for the much more divergent organisms M. tuberculosis and M. leprae. Similar results were obtained for the dN values, although in this case natural selection affected each gene or lineage differently, which resulted in some variability. For example, the highest dN value for ligA meant faster evolution of this gene in the M. lepromatosis and M. leprae lineages. Together, these results suggest that the divergence of the two leprosy-causing organisms was not recent and that the length of their divergence was around 20% of the age of divergence of M. tuberculosis and M. leprae or ~10 million years.

Age of M. lepromatosis pseudogenes. For all five pseudogenes (citA, gnd2, icd1, lld1, and mmaA3) in both species there were corresponding functional genes in M. tuberculosis and other mycobacteria. So when were the genes inactivated, in the last common ancestor or after divergence? This question was addressed from a few angles.

First, the dN/dS ratio for both species was estimated for each pseudogene. For an ancestral pseudogene, continued evolution after divergence in both lineages would be neutral, resulting in similar dN and dS values with a ratio of 1 or close to 1. Otherwise, selective evolution as a gene in both lineages after divergence with late inactivation would result in a much smaller dN/dS ratio. The longer the selective evolution, the smaller the dN/dS ratio would be. For example, Gomez-Valero et al. (13) showed that in the M. leprae lineage after divergence from M. tuberculosis, the average dN/dS ratio for genes was 0.079.

As shown in Table 3, the dN/dS ratios for M. lepromatosis and M. leprae were close to 1 (range, 0.680 to 0.789) for the citA, gnd2, icd1, and mmaA3 pseudogenes, indicating that there was likely neutral evolution and thus the sequences were pseudogenes prior to divergence. It is noteworthy that removal of stop codons, which was required for calculation, resulted in underestimation of the dN values, leading to smaller ratios. Mutations from codons to stop codons, which are common in pseudogenes, are essentially dN events. Yet similar dN and dS values were also indicated by the overlapping confidence intervals (mean ± 1.96 standard errors) for these values. The only exception was icd1, for which the dN value was significantly smaller than the dS value (a ratio much less than 1). The pseudogene dN/dS ratios contrasted with the dN/dS ratios for the five coding genes analyzed, for which the average dN/dS ratio was 0.074 (Fig. 2). The dN/dS ratios were also estimated for M. lepromatosis and M. tuberculosis, and as expected, they
were small (Table 3), reflecting the evolution of the sequences entirely as genes in the M. tuberculosis lineage.

Second, the relative ages of M. lepromatosis pseudogenes were estimated by using the formula used for the M. lepra pseudogenes (13). This formula required the presence of corresponding orthologous genes in M. tuberculosis and M. avium. Thus, it could not be used for mmaA3 because this pseudogene is not present in M. avium. The relative ages of the remaining four pseudogenes (Table 4) were found to be similar to the relative ages in M. leprae and also to the average age (0.13) of the 611 M. leprae pseudogenes analyzed previously (13). Thus, the similar ages of pseudogenes in the two species may indicate that they were pseudogenes in the ancestor.

Finally, indels that were present in both species and thus likely of ancestral origin were examined using sequence alignments. Several common indels were detected, except in citA, and they included a 2-nucleotide insertion in gnd2, a 5-nucleotide deletion in icd1, a 10-nucleotide deletion in lld1, and a 5-nucleotide insertion and a 8- or 9-nucleotide deletion in mmaA3. These indels were not in triplet nucleotides; as a result, they could alter the codon reading frames and possibly create stop codons. Thus, these findings in essence show the ancestral presence of these pseudogenes, rather than convergent evolution. Many unique indels in each species represented continued evolution of these pseudogenes after divergence.

Together, these analyses provided ample evidence showing that (i) all or most of the pseudogenes analyzed were inactivated prior to the divergence of M. lepromatosis and M. leprae and (ii) the massive gene inactivation described for M. leprae (4, 13) before the discovery of M. leprae (17) in fact took place in the last common ancestor of these two species.

**DISCUSSION**

The discovery of the new leprosy-causing organism M. lepromatosis and its close relationship to M. leprae provided an opportunity to discern the evolutionary features of these organisms. Using in-depth analysis of 20 genes and pseudogenes from M. lepromatosis, this study fulfilled some of our goals. First, the species-level divergence between these two organisms was confirmed to be on a larger genetic scale despite the current lack of culture and phenotypic features. Second, divergence ~10 million years ago is proposed based on the topology of robust phylogenetic trees and the levels of nucleotide substitutions in the protein-encoding genes. This time is orders of magnitude (~88 times) greater than the relatively well-established divergence time for M. tuberculosis and M. bovis (~113,000 years ago), which share 99.95% genome identity (7) and had no resolvable difference in our analyses (Fig. 1 and 2). And third, the age of the pseudogenes of M. lepromatosis, which is nearly identical to the age of M. leprae pseudogenes, their nearly neutral evolution, and the presence of codon-altering common indels suggest these pseudogenes were pseudogenes in the last common ancestor of the two species.

Bacterial taxonomy has changed along with technological advances, from morphological classification a century ago to phenotype-based chemotaxonomy to a polyphasic approach involving assessment of genetic variation, such as genomic DNA hybridization (70% cutoff to differentiate species) and divergence of the 16S rRNA gene (in general, 97% cutoff) in recent decades (21). Numerous bacterial genome sequences are now readily accessible, making genetic comparisons accurate and comprehensive for the purpose of species identification (15). Yet the phenotype of an organism, which includes complex traits involving many genetic elements, is still essential. For instance, historically as well as practically, several species with essentially identical genome sequences are retained, such as M. tuberculosis for humans and M. bovis for cows. In the case of M. lepromatosis, status as a novel species can be justified because of its remarkable genetic variation, as well as uniqueness of the disease that it causes (DLL). However, to satisfy the bacteriological code for a full description of this new species, phenotypic characterization after cultivation and/or animal passage is needed, and efforts to do this are being made.

Practically, the sequences enable us to separate the two leprosy organisms to better define the disease. For instance, we have devised a reliable PCR assay based on the species-specific 16S rRNA gene sequences to detect the presence of each species in a skin biopsy (X. Y. Han, K. C. Sizer, F. Aung, H. H. Tan, and B. Werner, unpublished data; X. Y. Han, K. C. Sizer, S. Velarde-Felix, and F. Vargas-Ocampo, unpublished data). Using this assay in a follow-up study of M. lepromatosis infections in Mexico (X. Y. Han, K. C. Sizer, S. Velarde-Felix, and...
F. Vargas-Ocampo, unpublished data), we found that far more leprosy cases were caused by M. lepromatosis than by M. leprae, suggesting that M. lepromatosis is the dominant leprosy-causing species in Mexico. We also noted that, in addition to lepromatous leprosy, M. lepromatosis also specifically caused DLL in ~30% of the infections. DLL is endemic in Mexico but rare elsewhere. These findings, as well as the lack of DLL reports or description in Europe, more specifically in Spain, suggest that DLL is native to Mexico and was not brought over by the Spaniards over the past 500 years. Native Mongoloid Mexicans initially migrated from Asia ~12,000 years ago through the Alaska Bering land bridge and settled in the Pacific coastal areas of Mexico. Some of them migrated further south and settled in the Caribbean, other central American countries, and South America. Our finding of M. lepromatosis in Brazil and Singapore Chinese (X. Y. Han, K. C. Sizer, F. Aung, H. H. Tan, and B. Werner, unpublished data) and the presence of DLL in the Caribbean (8) are also consistent with the possibility that there was Mongoloid spread of this organism. Therefore, the studies of M. lepromatosis have provided new insights into the puzzling clinical and geographic features of leprosy.

While the overall features of M. lepromatosis and M. leprae are very similar in terms of culture difficulty, G+C content, and the functional status of genes and pseudogenes, M. lepromatosis likely has some unique phenotypic traits that have yet to be discovered that may be responsible for some distinct clinical features of M. lepromatosis infections, such as DLL. DLL is chronic, and the organism is strictly intracellular (17, 20, 26). Vargas-Ocampo (31) recently studied the histopathology of 199 patients with DLL in Mexico and concluded that DLL is essentially a vascular disease caused by mycobacterial invasion, with early vaculitis to late-stage vascular occlusion, which is responsible for the clinical manifestations of the disease that include skin purpura, ulceration, and necrosis.

Humans are the only known natural host of M. leprae and have probably been infected for >100,000 years according to genomic evidence based on many M. leprae strains obtained worldwide (24). Consistent with the strictly parasitic lifestyle of this organism, the M. leprae genome is the smallest genome of all of the mycobacterial genomes sequenced to date. This genome has also undergone reductive evolution with massive inactivation of genes (4) estimated to have occurred during the last 20 million years (13). The present study further establishes that this genome decay took place in the last common ancestor of M. leprae and M. lepromatosis.

What led to the genome decay in the ancestor ~20 million years ago? We speculate that there is some relationship to adaptation from a free-living lifestyle, like that of almost all other mycobacteria (>100 species) except M. tuberculosis and M. bovis, to an increasingly parasitic lifestyle in a host (an early ancestor of hominids?)? During the last decade, sequencing and comparative analysis of many bacterial genomes have shown that changes in lifestyle from free living to host dependence or restriction to a specific host may lead to relaxation of natural selection and make many genes nonessential, initiating the process of genome downsizing. In the early stages, many genes (up to thousands of genes), usually less selectively constrained genes, are inactivated (4, 5, 13, 29). This inactivation occurs through missense or nonsense nucleotide substitutions, indels, or mobilization of insertion elements (9). There is then a tendency to lose the nonfunctional DNA (12, 23, 27). Analysis of some bacterial endosymbionts of insects, such as the aphid endosymbiont Buchnera aphidicola and the ant endosymbiont Blochmannia floridanus, has permitted workers to establish a stepwise scenario for genome reduction involving very slow and gradual degradation with scattered punctual large deletion events (11, 14). This model has been confirmed by complete sequencing of several B. aphidicola genomes (25).

In this regard, future genome sequencing of M. lepromatosis and comparative genomic analyses with M. leprae may shed light on the genome decay in these highly specialized pathogens.

Why did the ancestor evolve into two species ~10 million years ago if it had become parasitic in a primate host? The answer to this question is even more elusive. There are a few possible explanations, including subtle niche differences in the host, eventually leading to separate evolution paths; coevolution with different host populations of the same species that might be affected variably by the availability of nutrients, susceptibility, geographic restriction, and other factors; and a shift from one host species to another. The niche possibility can be tested in future clinical studies by examining tissue infected by each organism. Both organisms are known to be intracellular in macrophages; however, remarkable endothelial invasion by M. lepromatosis in DLL has been documented (17, 26, 31). In our clinical studies, we also found many cases of dual M. lepromatosis-M. leprae infection (X. Y. Han, K. C. Sizer, F. Aung, H. H. Tan, and B. Werner, unpublished data; X. Y. Han, K. C. Sizer, S. Velarde-Felix, and F. Vargas-Ocampo, unpublished data). Further studies of the coexistence of these two organisms may provide insight as well.

ACKNOWLEDGMENTS

This work was supported in part by a University Cancer Foundation grant from the University of Texas M. D. Anderson Cancer Center (MDACC) (to X.Y.H.), by the School of Health Sciences of MDACC, and by National Institutes of Health grant CA16672 for the MDACC DNA Core Facility.

We thank the DNA Core Facility staff for DNA sequencing and John Spencer for helpful discussions. X.Y.H. initiated and orchestrated this study; X.Y.H., K.C.S., E.J.T., J.K., J.L., and P.H. determined the sequences; F.J.S., L.G.V., and X.Y.H. performed data analysis; and F.J.S. and X.Y.H. wrote the paper.

REFERENCES