Genomewide Pattern of Synonymous Nucleotide Substitution in Two Complete Genomes of Mycobacterium tuberculosis

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi:10.3201/eid0811.020064</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:4887106">http://nrs.harvard.edu/urn-3:HUL.InstRepos:4887106</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Genomewide Pattern of Synonymous Nucleotide Substitution in Two Complete Genomes of Mycobacterium tuberculosis

Austin L. Hughes,* Robert Friedman,* and Megan Murray†

Comparison of the pattern of synonymous nucleotide substitution between two complete genomes of Mycobacterium tuberculosis at 3,298 putatively orthologous loci showed a mean percent difference per synonymous site of 0.000328 ± 0.000022. Although 80.5% of loci showed no synonymous or nonsynonymous nucleotide differences, the level of polymorphism observed at other loci was greater than suggested by previous studies of a small number of loci. This level of nucleotide difference leads to the conservative estimate that the common ancestor of these two genotypes occurred approximately 35,000 ago, which is twice as high as some recent estimates of the time of origin of this species. Our results suggest that a large number of loci should be examined for an accurate assessment of the level of nucleotide diversity in natural populations of pathogenic microorganisms.

Surveys of genetic diversity in the pathogenic bacterium Mycobacterium tuberculosis have revealed a contradictory picture. In spite of known polymorphism at the phenotypic level and abundant polymorphism associated with repetitive elements (1), surveys of single nucleotide polymorphism in protein-coding genes have shown surprisingly low levels of polymorphism in comparison with other eubacterial species (2). The apparent low level of nucleotide polymorphism has led to the hypothesis that the ancestor of this species occurred quite recently, perhaps 15,000–20,000 years ago (2,3). However, if the number of substitutions per site is low, the error of estimation of this number would be expected to be substantial unless a very large number of sites are surveyed. We addressed the question of polymorphism in M. tuberculosis by comparing protein-coding genes in two completely sequenced genotypes, H37Rv and CDC1551 (4).

Methods

We applied the BLASTP program (5) to identify, for each predicted protein sequence in the H37Rv genome (GenBank accession no. AL123456), the closest homolog in the CDC1551 genome (GenBank accession no. AE000516). Following GenBank annotations, we compared 3,972 predicted proteins in H37Rv with 4,187 predicted proteins in CDC1551. We used a strict search criterion (E = 10−50) to identify truly orthologous gene pairs. We aligned (6) the putative orthologous pairs of amino acid sequences (n=3,428), then imposed this alignment on the DNA sequences.

Visual inspection of amino acid alignments showed that certain alignments, usually near the N-terminus or C-terminus, had regions of very low sequence identity. Examination of the DNA sequences of the corresponding genes showed that these regions of low identity were typically caused by a frameshift in one of the two genomes relative to the other. Whether these frameshifts are biologically real or result from sequencing error was uncertain; therefore, we eliminated 119 such gene pairs from our data set. For the remaining gene pairs (n=3,309), we computed the proportion of synonymous substitutions per synonymous site (pS) and the proportion of nonsynonymous substitutions per site (pN) by using Nei and Gojobori’s method (7). Because values of pS and pN were very low in most cases, we did not correct for multiple hits.

Because pS values appeared to fall into two groups (see Results), we used a simple probabilistic model to separate these two sets of gene pairs. We assumed that the probability of synonymous substitution followed two separate binomial distributions, designated models A and B, with probabilities of “success” (i.e., of a synonymous difference) designated pA and pB, respectively. Using the Bayes equation, for each gene pair with a given pS value, we computed the probability that model A applies, given the observed pS: P(A|pS) = (pSA)fA / [(pSA)fA + (pSB)fB], where fA is the frequency of cases to which model A applies, fB the frequency of cases to which model B applies, pSA is the binomial probability of obtaining the observed pS, given the number of synonymous sites in the gene and a probability of a synonymous difference equal to pA; and pSB is the binomial probability of obtaining the observed pS, given the number of synonymous sites in the gene and a probability of a synonymous difference equal to pB. The probability that model B applies, given pS, is P(B|pS) = (pSB)fB / [(pSA)fA + (pSB)fB].

By comparing these two probabilities, we assigned each gene pair to one of two groups assumed to evolve according to the two models, respectively (Groups A and B). We reassigned group membership in iterative fashion, computing pA and pB from the mean pS values for each group. We started the process with fA = 0.995 and continued until group memberships were stable.

Results

Of 3,309 pairs of putatively homologous protein-coding genes in the H37Rv and CDC1551 genomes of M. tuberculosis, 2,662 (80.5%) showed no synonymous or nonsynonymous
nucleotide differences between the two genomes, and 3,010 (91.0%) showed no synonymous differences between the two genomes. However, in a small number of gene pairs, the proportion of synonymous differences per synonymous site \( p_S \) was surprisingly high. In 13 (0.4%) gene pairs, \( p_S \) was >0.01, and in 3 gene pairs \( p_S \) exceeded 4%. These extreme \( p_S \) values seen in a small number of gene pairs are much higher than generally observed between alleles at neutrally evolving loci in eukaryotes (8). Thus, the comparison of protein-coding genes between the two \( M. \) \( \text{tuberculosis} \) genomes suggested the existence of two distinct groups of gene pairs: a large group having few or no synonymous differences and a much smaller group with a substantial degree of synonymous divergence.

We used a simple probabilistic model (see Methods) to separate these two sets of gene pairs, designated Group A and Group B, respectively (Figure). The application of this method showed 11 loci with unusually high \( p_S \) values and probabilities of assignment to group A of <50% (Figure).

We assumed that Group A members are truly orthologous gene pairs that diverged at the time of the common ancestor of the \( \text{H37Rv} \) and \( \text{CDC1551} \) genomes. Group A included 3,298 pairs, with mean \( p_S \) for all genes of 0.000328 ± 0.000022 standard error. When \( p_S \) was estimated for the 3,298 genes concatenated together (a total of 934,413 synonymous sites), an estimate of \( p_S = 0.000348 ± 0.00019 \) was obtained. The range of \( p_S \) values in Group A was between zero and 0.012; a total of 288 loci in Group A had \( p_S \) values other than zero. These results show a substantial level of nucleotide diversity, approximately half the level of nucleotide diversity in humans (9).

Rates of nucleotide substitution per unit time are difficult to estimate in bacteria given the lack of calibration from the fossil record (10). To obtain an estimate of the rate of synonymous nucleotide substitution, we used published data on comparisons of \( \text{Escherichia coli} \) and \( \text{Salmonella typhimurium} \) (11,12), which are believed to have diverged approximately 100 million years ago (13,14) (Table 1). This procedure yielded estimates for the last common ancestor of \( \text{H37Rv} \) and \( \text{CDC1551} \) in the range of 34,000–38,000 years (Table 1). These estimates are approximately twice previous estimates of the age of the common ancestor of worldwide \( M. \) \( \text{tuberculosis} \) (2,3). To obtain the observed mean \( p_S \) value between \( \text{H37Rv} \) and \( \text{CDC1551} \) within 15,000–20,000 years would require a rate of synonymous substitution approximately twice that observed in \( \text{Enterobacteria} \).

Group B consisted of 11 gene pairs with mean \( p_S \) of 0.0286 ± 0.0050 (Table 2).

In \( \text{Enterobacteria} \), a negative correlation exists between observed proportions of synonymous difference and codon bias (11). In the case of \( \text{Mycobacterium} \), codon bias results mainly from the very high third position G+C content of most genes (15). In our data, however, we observed no correlation between \( p_S \) and proportion G+C at third codon positions (\( r = -0.010 \); not significant).

### Table 1. Estimates\(^a\) of the divergence time of the \( \text{H37Rv} \) and \( \text{CDC1551} \) genotypes of \( \text{Mycobacterium tuberculosis} \)

<table>
<thead>
<tr>
<th>Reference</th>
<th>No. loci</th>
<th>Synonymous substitutions/site/yr</th>
<th>Divergence time (( \text{H37Rv} ) and ( \text{CDC1551} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>67</td>
<td>4.7 ± 0.2 \times 10^{-9}</td>
<td>34,900 ± 2,300 ( ^b ) ( ^{35,500-36,400} ) ( ^{b} )</td>
</tr>
<tr>
<td>12</td>
<td>128</td>
<td>4.4 \times 10^{-9}</td>
<td>37,300 ± 2,500 ( ^b )</td>
</tr>
</tbody>
</table>

\(^a\)Based on synonymous substitutions between \( \text{Escherichia coli} \) and \( \text{Salmonella typhimurium} \), assumed to have diverged 100 million years ago (13,14).

\(^b\)Estimates are shown ± standard error, based on standard error of mean \( p_S \).

\(^c\)Range based on standard error of rate estimate.

### Discussion

A number of additional possibilities may explain the occurrence of gene pairs with higher than expected \( p_S \) values: 1) Balanced polymorphism. Selectively maintained polymorphisms are expected to be much older than neutral polymorphisms and may even predate speciation events (16). In the case of haploid organisms such as bacteria, balancing selection would take the form of frequency-dependent selection rather than overdominant selection. 2) Differential deletion. In a multi-gene family, if one member of an orthologous pair of genes were deleted in one genotype, the gene pairs would involve paralogous, not orthologous comparisons. 3) Horizontal gene transfer. A gene obtained by one of the two genotypes from another bacterial species would be expected to be more divergent than other genes in that genotype.

One indication of a balanced polymorphism is a higher rate of nonsynonymous than synonymous substitution (8). There was no strong evidence of such selection in the present case; \( p_S \) was greater than \( p_N \) at 10 of the 11 loci, and \( p_S \) exceeded \( p_N \) only slightly at one locus (Table 2). In addition, we compared \( p_S \) and \( p_N \) in sliding windows of 30 codons along the length of these genes. No regions were observed in which \( p_N \) was greater than \( p_S \) (data not shown). Thus, there was no evidence of positive selection acting on specific regions of these genes. On the other hand, differential deletion can probably explain some
Table 2. Proteins for which the nearest homologous comparison between the H37Rv and CDC1551 genotypes of Mycobacterium tuberculosis has a high pS value (Group B)

<table>
<thead>
<tr>
<th>Accession nos.</th>
<th>Protein function</th>
<th>pS</th>
<th>pN</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP_216309, NP_335079</td>
<td>unknown</td>
<td>0.0470</td>
<td>0.0000</td>
</tr>
<tr>
<td>NP_215713, NP_335504</td>
<td>unknown</td>
<td>0.0628</td>
<td>0.0043</td>
</tr>
<tr>
<td>NP_216319, NP_336310</td>
<td>PE repeat family</td>
<td>0.0115</td>
<td>0.0094</td>
</tr>
<tr>
<td>NP_215965, NP_335949</td>
<td>PE repeat family</td>
<td>0.0448</td>
<td>0.0185</td>
</tr>
<tr>
<td>NP_214910, NP_334815</td>
<td>unknown</td>
<td>0.0226</td>
<td>0.0105</td>
</tr>
<tr>
<td>NP_216104, NP_336077</td>
<td>unknown</td>
<td>0.0265</td>
<td>0.0084</td>
</tr>
<tr>
<td>NP_216281, NP_336535</td>
<td>unknown</td>
<td>0.0210</td>
<td>0.0161</td>
</tr>
<tr>
<td>NP_215835, NP_335809</td>
<td>adenylate cyclase</td>
<td>0.0229</td>
<td>0.0068</td>
</tr>
<tr>
<td>NP_216564, NP_335673</td>
<td>polyketide synthase</td>
<td>0.0093</td>
<td>0.0036</td>
</tr>
<tr>
<td>NP_217029, NP_337080</td>
<td>unknown</td>
<td>0.0148</td>
<td>0.0156</td>
</tr>
<tr>
<td>NP_216862, NP_335679</td>
<td>unknown</td>
<td>0.0313</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

**Mean ± S.E.**

0.0286 ± 0.0050  0.0085 ± 0.0019b

*Paired sample t-test of the hypothesis that pS = pN, p<0.01. The quantities pS and pN are the proportion of nucleotide difference per synonymous site and per nonsynonymous site, respectively.

In addition, the hypothesis that the single nucleotide polymorphisms (SNPs) observed between these genotypes are real received strong support from a recent study that observed a...
number of the same SNPs in clinical isolates (24). Moreover, since sequencing errors are expected to occur at random with respect to the reading frame of coding sequences, the fact that mean \( p_S \) exceeded mean \( p_N \) in both Group A and Group B was strong evidence against the hypothesis that a substantial proportion of the observed polymorphism was due to sequencing error.

Simple considerations of probability can explain why earlier studies produced relatively low estimates of this species’ age. If we assume that the per-site probability of a synonymous difference between two \( M. \) \textit{tuberculosis} genomes is equal to the mean \( p_S \) observed between H37Rv and CDC1551 (0.000328), then the probability is approximately 95% that no synonymous differences will be seen in a gene with 150 synonymous sites. The probability that no synonymous differences will be seen in 10 such loci chosen at random is approximately 60%, and the probability that no synonymous differences will be observed at 20 such loci is approximately 37%. On the other hand, the probability that no synonymous differences will be seen at 100 such loci is <1%.

These calculations emphasize the need to examine a very large number of nucleotide sites to obtain a reliable estimate of nucleotide diversity and thus of the age of the most recent common ancestor in cases where the frequency of substitution is less than one in a thousand. Even when the frequency of substitution is between one in a thousand and one in a hundred, substantial stochastic error is possible if the number of loci examined is small. Thus, any study that estimates population parameters from nucleotide sequence data needs to survey a substantial number of loci. These considerations are particularly important in the case of pathogenic microorganisms, where a number of factors (including both natural selection and horizontal gene transfer) may lead to substantial differences among loci with respect to the level of nucleotide diversity.

Comparison of two complete genomes of \( M. \) \textit{tuberculosis} showed a greater extent of sequence polymorphism than would be expected on the basis of previous studies, in turn suggesting that analysis of additional genomes will likely show further polymorphism. Polymorphism in any species of pathogen may complicate therapeutic strategies because it implies the existence of variation on which selection can act, including selection imposed by human vaccines and pharmacologic agents (20). On the other hand, known polymorphisms may prove useful to investigators in reconstructing the evolutionary relationships among clinical isolates and in providing markers for understanding the genetic basis of complex pheno-

typic traits.

This research was supported by National Institutes of Health grants GM34940 and GM66710 to A.L.H. and AI01430 and AI1046669 to M.M.

Dr. Hughes is director of the Biotechnology Institute and professor in the Department of Biological Sciences at the University of South Carolina. His research uses computational analysis of molecular sequence data to understand host-parasite co-evolution, genome evolution, and the population biology of human pathogens.

References


7. Nei M, Gojobori T. Simple methods for estimating the numbers of synon-


9. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymor-


11. Sharp PM. Determinants of DNA sequence divergence between \( Escher-\)


16. Takahata N, Nei M. Allelic genealogy under predominant and frequency-dependent selection and polymorphism of major histocompatibil-


Address for correspondence: Austin L. Hughes, Department of Biological Sciences, University of South Carolina, Coker Life Sciences Bldg., 700 Sumter St., Columbia, SC 29208, USA; fax: 803-777-4002; e-mail: austin@biol.sc.edu