Gene 518 (2013) 187–193

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

The evolutionary landscape of the *Mycobacterium tuberculosis* genome

Tai-Chun Wang [a], Feng-Chi Chen [a,b,c,*]

[a] Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, 35 Keyen Road, Zhunan, Miaoli County, 350 Taiwan
[b] Department of Biological Science and Technology, National Chiao-Tung University, Hsinchu, 300 Taiwan
[c] Department of Dentistry, China Medical University, Taichung, 404 Taiwan

ARTICLE INFO

Article history:
Accepted 27 November 2012
Available online 7 December 2012

Keywords:
Mycobacterium tuberculosis
Selection pressure
Nonsynonymous substitution
Synonymous substitution
Neutral substitution rate

ABSTRACT

Mycobacterium tuberculosis is one of the most deadly human pathogens. The major mechanism for the adaptations of *M. tuberculosis* is nucleotide substitution. Previous studies have relied on the nonsynonymous-to-synonymous substitution rate (d_N/d_S) ratio as a measurement of selective constraint based on the assumed selective neutrality of synonymous substitutions. However, this assumption has been shown to be untrue in many cases. In this study, we used the substitution rate in intergenic regions (d_i) of the *M. tuberculosis* genome as the neutral reference, and conducted a genome-wide profiling for d_i, d_S, and the rate of insertions/deletions (indel rate) as compared with the genome of *M. canettii* using a 50 kb sliding window. We demonstrate significant variations in all of the three evolutionary measurements across the *M. tuberculosis* genome, even for regions in close vicinity. Furthermore, we identified a total of 233 genes with their d_S deviating significantly from d_i within the same window. Interestingly, d_S also varies significantly in some of the windows, indicating drastic changes in mutation rate and/or selection pressure within relatively short distances in the *M. tuberculosis* genome. Importantly, our results indicate that selection on synonymous substitutions is common in the *M. tuberculosis* genome. Therefore, the d_N/d_S ratio test must be applied carefully for measuring selection pressure on *M. tuberculosis* genes.

1. Introduction

Mycobacterium tuberculosis (MTB), the causing pathogen of one of the most deadly diseases, claims millions of lives worldwide each year (Zhang et al., 2011). The MTB complex (MTBC) belongs to the slow-growing sublineage of *Mycobacteria*. Based on the geographical characteristics, MTBC can be classified into six clusters, including such species as *M. tuberculosis*, *M. bovis*, *M. africanum*, *M. microti*, *M. pinnipedii*, and *M. canettii* (Filliol et al., 2006; Gagneux et al., 2006; Gutacker et al., 2006; Schurch and van Soolingen, 2012). Members in MTBC, including *M. tuberculosis, M. bovis, M. africanum, and M. canetti*, share 99.95% of their genomic sequences and a strictly clonal population structure (Mokrousov et al., 2004; Smith et al., 2009). Compared to more ancient species (e.g. *M. marinum*), MTBC has shorter but more virulent chromosomes (Namouchi et al., 2012).

Although most bacterial species acquire new genetic materials via horizontal gene transfer (Thomas and Nielsen, 2005), it has been reported that this mechanism rarely occurs to MTBC genomes (Gutierrez et al., 2005; Veyrier et al., 2009, 2011). Therefore, nucleotide substitution is a major mechanism for the emergence of *M. tuberculosis* pathogenesis. By comparing multiple MTBC genomes, Namouchi and colleagues indicated that MTBC genomes exhibit significant regional variations in the density of single nucleotide polymorphisms (SNPs) (Namouchi et al., 2012). This observation implies that MTBC genes at different genomic positions may be evolving at very different rates. However, the authors did not distinguish between coding and noncoding regions when calculating SNP densities. Their results thus cannot reflect the variations in SNP density at selectively neutral sites.

Since the majority of the MTB genome is composed of coding sequences, genomic regions of high SNP density may harbor rapidly evolving genes. Some of these genes may be positively selected for their importance in the adaptations of MTBC to the human environments. Meanwhile, extremely conserved genes are likely to be essential for the survival and/or replication of the bacterium. These two groups of genes are good candidates of drug targets. However, an increase in evolutionary rate does not necessarily result from positive selection. An increase in mutation rate or relaxation of selective constraint can lead to the same result. An adequate reference for neutral substitution rate and a good measurement for selection pressure are thus required to infer the driver of the increased evolutionary rates in the genes of interest.

Abbreviations: MTB, *Mycobacterium tuberculosis*; MTBC, *Mycobacterium tuberculosis* complex; d_N, Nonsynonysmous substitution rate; d_S, Synonymous substation rate; d_i, Nucleotide substitution rate at intergenic regions; indel, Insertion and deletion; kb, kilobase; SNP, Single nucleotide polymorphism; NCBI, National Center for Biotechnology Information; CAI, Codon adaptation index.

[*] Corresponding author at: Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, 35 Keyen Road, Zhunan, Miaoli County, 350 Taiwan.

E-mail addresses: taichun@nhri.org.tw (T.-C. Wang), fcchen@ngri.org.tw (F.-C. Chen).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.11.033

Table 1
The genomes analyzed in this study.

	RefSeq #	Strain	Length	# Annotated genes
MTB complex	NC_015758	*Mycobacterium africanum GM041182*	4389314	3983
	NC_002945	*Mycobacterium bovis AF2122/97*	4345492	4001
	NC_008769	*Mycobacterium bovis BCG str. Pasteur 1173P2*	4374522	4033
	NC_012207	*Mycobacterium bovis BCG str. Tokyo 172*	4371711	4027
	CP001641	*Mycobacterium tuberculosis CCDC5079*	4398812	3696
	CP001642	*Mycobacterium tuberculosis CCDC5180*	4405981	3639
	NC_002755	*Mycobacterium tuberculosis CDC1551*	4403837	4293
	NC_000962	*Mycobacterium tuberculosis H37Rv*	4411532	4047
	NC_009525	*Mycobacterium tuberculosis H37Ra*	4419977	4084
	NC_009565	*Mycobacterium tuberculosis F11*	4424435	3998
	NC_012943	*Mycobacterium tuberculosis KZN 1435*	4398250	4107
Non-MTB complex	NC_015848	*Mycobacterium canettii CIPT 140010059*	4482059	3982
	NC_010612	*Mycobacterium marinum M*	6636827	5541

One commonly used test for natural selection is the ratio of nonsynonymous substitution rate (d_N) to synonymous substitution rate (d_S) (i.e., the d_N/d_S ratio) (Toll-Riera et al., 2011). In general, $d_N/d_S > 1$ indicates positive selection, and $d_N/d_S < 1$ is a sign of negative selection. However, this test is based on the assumption that synonymous substitutions are selectively neutral, which has been questioned particularly in unicellular organisms. It is known that synonymous substitutions may confer fitness effects by affecting the efficiency and/or accuracy of protein translation (Kryazhimskiy and Plotkin, 2008). An alternative neutral reference is the nucleotide substitution rate of intergenic regions (d_i) because intergenic regions are usually free from selection pressure. Therefore, theoretically, by comparing the d_S of a gene against the d_i of the neighboring intergenic region, we can infer whether the synonymous substitutions are selectively neutral or not, and determine whether we should use d_N/d_S as the measurement of selection. There are three possible scenarios in the comparison between d_S and d_i. Firstly, if d_S is approximately equal to d_i, synonymous substitutions are probably driven mainly by mutation. Alternatively, if d_S is significantly lower than d_i, synonymous substitutions are likely to be negatively selected. Finally, if d_S is significantly larger than d_i, synonymous substitutions are possibly driven by positive selection. In the latter two cases, d_N/d_i should be used instead of d_N/d_S for measuring selection pressure on the gene of interest.

Here, we examine the variations in evolutionary rates in the genomes of multiple MTB strains and the selection pressures imposed on MTB genes. We would like to address the following questions: (1) how applicable is d_N/d_S in measuring selection pressure on MTB genes; (2) which MTB genes evolve significantly more rapidly or more slowly than the genome average in terms of, separately, d_S, d_N, and d_N/d_S; and (3) what is the major driving force that leads to the variations in evolutionary rates among genes.

Our results indicate significant variations in d_i, d_S and the rate of insertions/deletions (indels) across the MTB genome, which suggests fluctuations in local mutation rate as a driving force of nucleotide substitutions. Furthermore, we found that synonymous substitutions in hundreds of MTB genes may be subject to negative or positive selection, indicating noticeable inapplicability of the d_N/d_S ratio test to the MTB genes. The molecular mechanisms and phenotypic consequences of the drastic variations in evolutionary rates in MTB genes are worth further investigations.

2. Materials and methods

2.1. Datasets

The genomic sequences of thirteen strains of *Mycobacteria* (Table 1) were downloaded from the National Center for Biotechnology Information (NCBI) at http://www.ncbi.nlm.nih.gov/. Except for *M. marinum*, all of these strains belong to the MTBC. Here, the genomes of *M. marinum* and *M. canettii* were used for comparisons with the other MTBC genomes for the calculation of evolutionary rates. The average G + C content is approximately 65% for all of the analyzed genomes.

2.2. Identification of orthologous genes

The gene annotations of the analyzed bacterial genomes were also retrieved from NCBI. The nucleotide sequences of the annotated genes were conceptually translated into peptide sequences, and input into orthoMCL (Li et al., 2003) with default parameters for identification of orthologous genes between the analyzed species/strains. OrthoMCL identified 2358 orthologous genes for the 13 analyzed *Mycobacterial* genomes. The peptide sequences of the identified orthologous genes were then aligned by using MUSCLE (Edgar, 2004) with default parameters, and then back-translated to nucleotide sequences for calculations of d_N, d_S, and the d_N/d_S ratio.

2.3. Measurements of local evolutionary rates

To analyze d_i and indel rate, we used Mauve 2.0 (Darling et al., 2004) to align the nucleotide sequences of the 13 analyzed genomes. The gaps between alignment blocks were discarded. For the comparison between d_S and d_i, we removed all of the noncoding RNAs from intergenic regions with reference to the annotations of SIPHT (sRNA identification protocol using high-throughput technologies) (Livny et al., 2008). A 50-kb non-overlapping sliding window was then used to delineate the aligned genomic regions for calculations of d_i and indel rate. Note that a window contains both genic and intergentic regions. The genic and intergenic regions were demarcated according to the NCBI annotations. d_N and d_S were calculated separately for each gene. The intergenic regions within each window were concatenated for the calculation of d_i. Therefore, for each window, we could obtain multiple d_N and d_S values (when there are multiple genes in a window), and a single d_i value. Of note, the genes that are located at the boundaries between windows were discarded. In addition, we trimmed 50 nucleotides from both ends of each alignment block to avoid potential alignment errors.

The Codeml module of PAML 4 (Yang, 2007) was used to calculate d_N and d_S. The Baseml module of PAML was applied for the calculation of d_i. We also calculated the indel rate by analyzing the MAUVE alignment files using an in-house PERL script. The indel rate was defined as the total length of insertions and deletions divided by the length of the alignable sequence.

2.4. Identification of genes with exceptional evolutionary rates

Since *M. tuberculosis* H37Rv is genetically close to *M. canettii*, in many of the cases we observe zero values of d_N, d_S, or d_N/d_S when

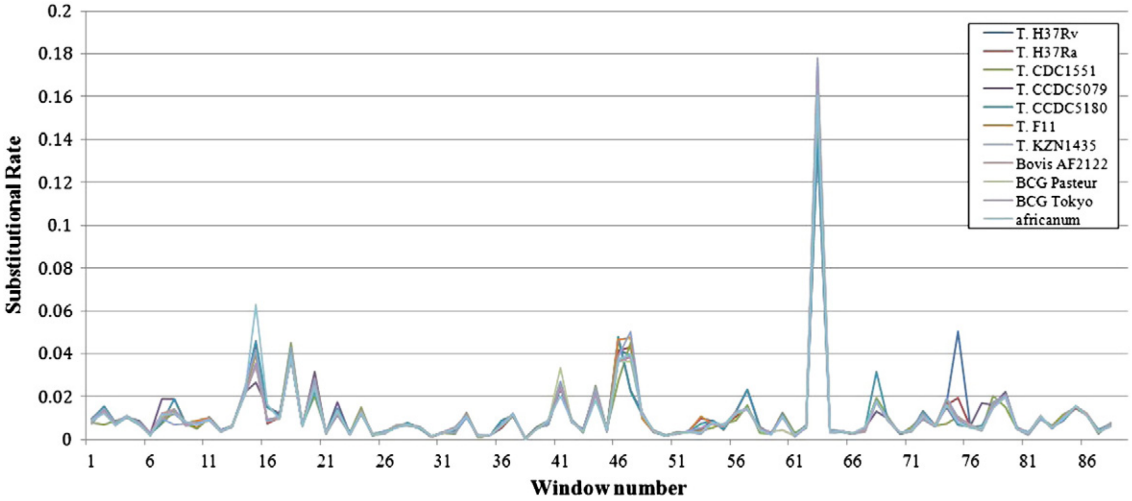

Fig. 1. The genomic landscape of d_i for eleven Mycobacterial genomes as compared with *M. canetti*.

comparing orthologous genes between the two species. For each evolutionary measurement (d_N, d_S, or d_N/d_S), we assume that the probability of a zero value is α. Then the distribution is a α: $1-\alpha$ mixture of a point probability mass at 0 and a log-normal distribution, which can be presented as:

$$F(x) = \begin{cases} \alpha, & x = 0 \\ (1-\alpha)\Phi(\log x), & x > 0 \end{cases}$$

where Φ is the function of normal distribution.

Based on this mixture probability distribution, the cut points for determining the outlier evolutionary measurements at 5% error rate should be adjusted by

$$\text{Average evolutionary rate } \pm 2^* \text{ standard deviation}^*(1-\alpha).$$

Notably, the standard deviation should be calculated based on the non-zero values. Also note that, for the analysis of d_N/d_S, the genes with $d_S = 0$ were removed. Hence, the sample size in this analysis was reduced, and α becomes the proportion of $d_N = 0$ in this reduced set.

We then defined slow- and fast-evolving genes as those that have their evolutionary rates fall outside of two standard deviations (adjusted for α) from the average. The genes thus identified were then functionally analyzed with reference to the TubercuList (http://genolist.pasteur.fr/TubercuList/) (Lew et al., 2011).

2.5. Measurement of codon usage bias

In this study, we used codon adaptation index (CAI) (Sharp and Li, 1987) as the measurement of codon usage bias for MTB genes. CAI can be calculated as follows:

$$\text{CAI} = \exp\left({}^{1}\Big/_{L}\sum_{l=1}^{L} \ln\ (w_i(l))\right),$$

where

$$w_i = \frac{f_i}{\max\left(f_j\right)} ij \in [\text{synonumous codons for amino acid}]$$

Here, f_i is the frequency of a codon i and f_j is the frequency of the synonymous codons for that amino acid.

3. Results and discussions

3.1. Significant variations in d_i and indel rate across the MTBC genomes

Genomic regions in close proximity generally have similar substitution rates. Meanwhile, recombination events may divide neighboring genomic regions into different “linkage blocks”, between which nucleotide substitution rates can differ significantly. However, if recombination events occur frequently at very small intervals (e.g., tens of base pairs), the genome would become “homogenized”, and the linkage blocks would be too small to be meaningful for large-scale studies. For the MTBC genomes, this seems to be the case according to Namouchi et al.'s recent report (Namouchi et al., 2012). By analyzing SNPs in the MTBC genomes, Namouchi and colleagues inferred that the average size of recombination DNA segments in these genomes is about 50–58 bp in length, and that recombination events occur at one-fifth the frequency of mutations, which is far more frequent than previously known (Namouchi et al., 2012). In other words, the MTBC genomes are frequently “scrambled” so that larger-scale regional variations in nucleotide substitution rates may simply reflect variations in background mutation rate and/or selection pressure. Therefore, by applying an adequate-sized widow to the genomic sequences of MTBC, we can obtain d_S, d_N, and d_i for comparison and for inferences of selection pressure on the genes in the window. To this end, we first aligned eleven MTBC genomes (Table 1) against the *M. canettii* genome by

Table 2
Spearman's correlation test for d_i and indel rate between *M. tuberculosis* H37Rv and other MTBC strains.

Strain	d_i		Indel rate	
	rho	*p*-Value	rho	*p*-Value
T. KZN1435	0.978	<2.20*E*−16	0.843	<2.20*E*−16
T. F11	0.977	<2.20*E*−16	0.907	<2.20*E*−16
T. CDC1551	0.943	<2.20*E*−16	0.691	<2.20*E*−16
T. CCDC5079	0.935	<2.20*E*−16	0.705	<2.20*E*−16
T. CCDC5180	0.959	<2.20*E*−16	0.719	<2.20*E*−16
T. H37Ra	0.994	<2.20*E*−16	0.929	<2.20*E*−16
africanum	0.973	<2.20*E*−16	0.478	3.26*E*−06
Bovis AF2122	0.973	<2.20*E*−16	0.862	<2.20*E*−16
BCG Pasteur	0.968	<2.20*E*−16	0.871	<2.20*E*−16
BCG Tokyo	0.977	<2.20*E*−16	0.842	<2.20*E*−16

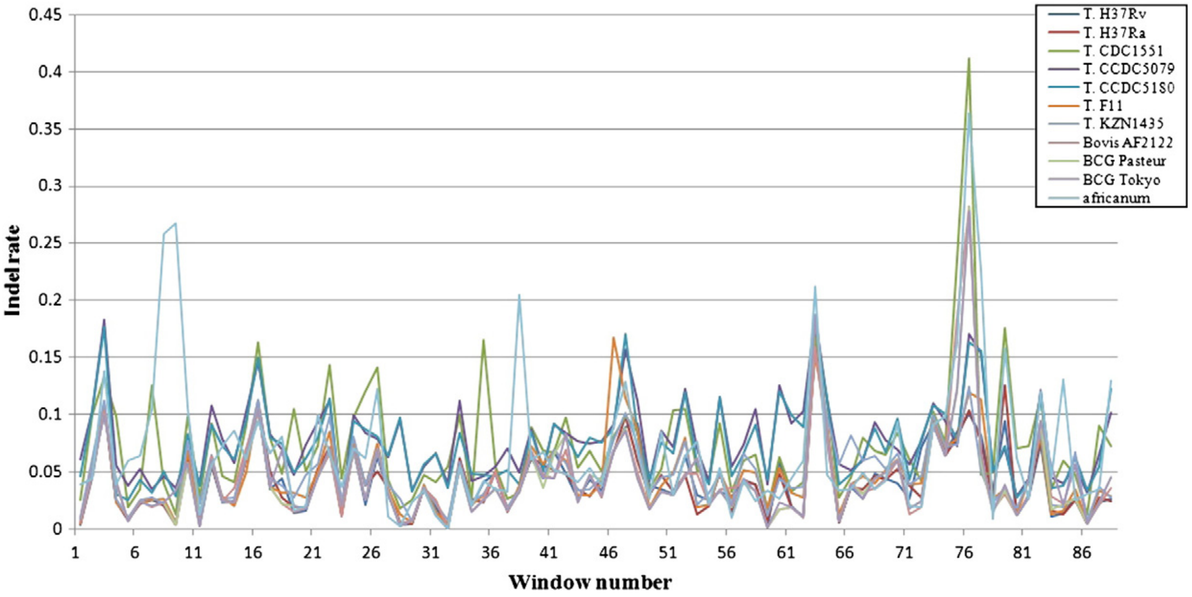

Fig. 2. The genomic landscape of indel rate for eleven Mycobacterial genomes as compared with *M. canetti*.

using MAUVE (Darling et al., 2004). We selected *M. canettii* because it is the closest outgroup species to the major pathogenic strains of *M. tuberculosis*, *M. bovis*, and *M. africanum* (the *M. marinum* genome is applied in a latter analysis). We then used a non-overlapping sliding window to examine the variations in evolutionary rates between *M. canettii* and each of the eleven MTBC genomes. Of note, a window must be large enough to include sufficient intergenic sequences for the calculation of d_i. We tested several different sizes (5 kb, 10 kb, 25 kb, 50 kb, 75 kb, and 100 kb), and determined 50 kb to be the most suitable window size for the purpose of this study. Since the MTBC genomes are approximately 4.4 Mb in length, there are a total of 88 windows for each genome.

We first examined the variations in d_i across the MTBC genomes. As shown in Fig. 1, d_i fluctuates considerably in the MTBC genomes, ranging from zero to ~0.18. Interestingly, the genomic profiles of d_i are almost identical among the MTBC strains (Table 2), which reflects the shared evolutionary histories and the small genetic distances among the analyzed genomes. Meanwhile, indel rate (defined as the total length of the identified indels divided by the alignable sequence length) also varies dramatically across the MTBC genomes (Fig. 2). Nevertheless, the profiles of indel rates are less similar among the strains as compared with d_i (Table 2). There are several potential reasons for the decreased similarity among the profiles of indel rate. One explanation is potential alignment errors, which may have led to false identification of indels. Of note, we actually trimmed 50 bp from both ends of each alignment block to reduce this possible source of error. Another possibility is genome assembly errors, which may also artificially increase indel rates (Albers et al., 2011). The third possible reason is the presence of low-complexity sequences, which might occur at different frequencies in the MTBC genomes, and lead to regional variations in indel rate. Of course, we cannot exclude the possibility that such variations in indel rate have reflected the high frequency of genomic rearrangements in the MTBC genomes, as reported previously (Namouchi et al., 2012).

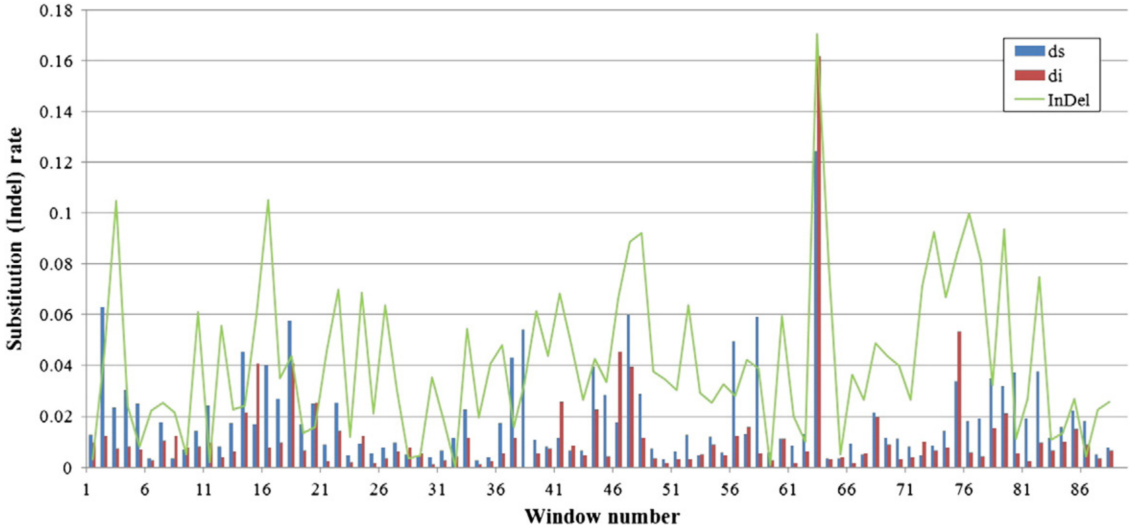

Fig. 3. The genomic landscape of d_S, d_i and indel rate of *M. tuberculosis H37Rv* as compared with *M. canetti*.

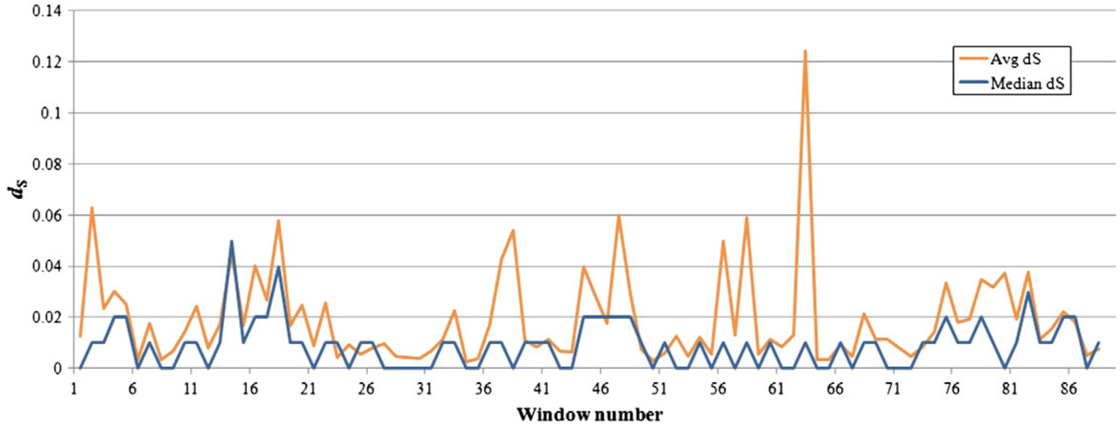

Fig. 4. Comparison between average d_S and median d_S in the case of *M. tb. H37Rv* Vs. *M. canetti*.

Next, we are interested in comparing different evolutionary measurements (d_i, d_S, and indel rate) in the same windows, and see whether they exhibit similar genomic profiles. Since the profiles of d_i are almost identical among different MTBC genomes, for simplicity, we use only the *M. tuberculosis* H37Rv genome for subsequent analyses.

3.2. Selection on synonymous substitutions in the genome of M. tuberculosis H37Rv

By comparing d_i, d_S, and indel rate in the same windows, we may be able to better understand the driving force of MTB genome evolution. Here d_i in each window is used as the neutral reference because intergenic regions should be free from selection pressure in most of the cases. Of note, annotated nocoding RNAs were removed from the intergenic sequences to avoid biased estimations of d_i in this analysis (see Materials and methods). Fig. 3 shows the genomic landscape for d_S, d_i and indel rate for the H37Rv strain as compared against the *M. canettii* genome. Apparently, the profiles of the three evolutionary measurements are similar to one another. For instance, in windows # 19, 22, 63, 68 and 82, d_S, d_i and indel rate all increase remarkably as compared with the neighboring windows. These windows may harbor recombination hotsopts, which can significantly increase both mutation rate and indel rate. Meanwhile, windows with a relatively high indel rate but low d_S and d_i are also observed (e.g. windows # 49–55). However, the pair-wise Spearman's correlations between the three evolutionary measurements are statistically insignificant (p-value > 0.05 in all three pair-wise comparisons), indicating non-mutation forces (e.g., selection) may have differentially affected different windows. Also noticeable is that d_S, d_i and indel rate may differ by more than ten folds even for two adjacent windows (e.g. windows # 63 and 64). This is actually consistent with the previous observation that the recombination DNA tracts are fairly short, so that physically close windows may have very different recombination rates (and very different d_S, d_i and indel rate). It is thus of interest to use a smaller window size for a better resolution of the variations in evolutionary rates. However, when we use a 5 kb window size for this purpose, the total length of intergenic region in each window is too short (e.g. shorter than 100 bp) for a reliable estimation for d_i.

Of note, in more than half of the windows, the average d_S is larger than $d_{i.}$ This is unexpected because synonymous substitutions are supposed to be nearly neutral or negatively selected in most of the cases. We then examined whether significant variations in d_S exist in the same windows, which can bias the average d_S upwards in the presence of exceptionally large d_S values. If this is the case, the average d_S will be larger than the median d_S, which is actually observed in 86% of the 88 windows (Fig. 4). Nevertheless, even if we replace the average d_S with the median d_S for each of the window, about 40% of all the windows still have their d_S larger than d_i (data not shown). One possible explanation for this observation is that d_i is underestimated in these windows because the intergenic regions include yet unidentified regulatory elements (although we have removed noncoding RNAs from the intergenic regions). Alternatively, the synonymous substitutions in the high-d_S genes may be subject to positive selection for enhancing translational efficiency, or for regulatory reasons related to the secondary structures of mRNAs and/or the binding of noncoding RNAs.

Next, we used the Chi-square test to examine for each gene whether d_S is significantly smaller ($d_S \ll d_i$) or larger ($d_S \gg d_i$) than d_i in the same window. We found a total of 114 and 119 genes, respectively, that have their d_S significantly smaller or larger than d_i. Interestingly, these two types of genes occasionally co-occur in the same window. This observation suggests that d_S varies to a large extent even within a relatively small distance (i.e. 50 kb).

One possible explanation for the deviation of d_S from d_i is to achieve better translational efficiency. Hence, we examined the codon usage biases (measured by the codon adaptation index, CAI) for known (non-hypothetical) genes, and compared the CAIs between $d_S \sim= d_i$ genes and $d_S \neq d_i$ genes. Fig. 5 shows that the median CAI is the highest in the genes with $d_S \ll d_i$, followed by genes with

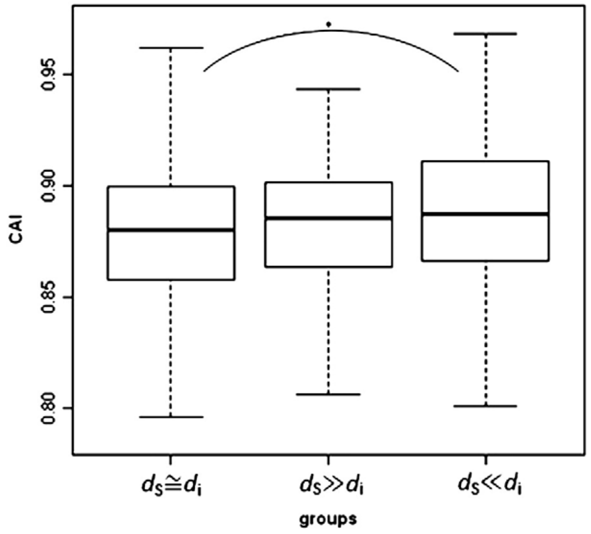

Fig. 5. The codon adaptation index (CAI) of three different groups of *M. tuberculosis* H37Rv genes ($d_S \sim= d_i$, $d_S \gg d_i$, and $d_S \ll d_i$). Note that only known genes are included in this analysis.

Table 3
The numbers (percentages) of conserved and rapidly evolving genes in *M. tuberculosis* H37Rv.

Reference species	d_N		d_S		d_N/d_S	
	Conserved	Rapidly evolving	Conserved	Rapidly evolving	Conserved	Rapidly Evolving
M. canettii	546 (40.3%)	7 (0.5%)	1 (0.1%)	55 (4.1%)	159[a] (11.7%)	6 (0.4%)
M. marinum	80 (2.9%)	51 (1.8%)	61 (2.2%)	74 (2.7%)	82 (3.0%)	43 (1.6%)

[a] Note that the number of genes in the analysis of H37Rv-*canettii* d_N/d_S is 1048 because the genes with $d_S = 0$ were excluded. In all of the other cells in this table, the number of analyzed genes is 2761 and 1356 in *M. marinum* and *M. canettii*, respectively.

$d_S \gg d_i$, and lastly by genes with $d_S \sim= d_i$. However, only the difference between $d_S << d_i$ genes and $d_S \sim= d_i$ genes is statistically significant (p-value < 0.05 by Wilcoxon Rank Sum test). Interestingly, if we conducted the same analysis while adding all of the noncoding RNA back to the intergenic regions, all of the pair-wise differences in CAI between different gene groups are statistically significant (Supplementary Fig. 1; p-value < 0.05 by Wilcoxon Rank Sum test). For the genes that have their d_S deviate significantly from d_i, the d_N/d_S ratio test may be inappropriate for measuring selection pressure. The genes with their d_N/d_S deviating significantly from d_N/d_i are listed in Supplementary Table 1.

3.3. Identification of conserved and rapidly evolving genes in M. tuberculosis H37Rv

One important issue in the development of tuberculosis therapeutics is the identification of drug target genes. As stated previously, both conserved genes and rapidly evolving genes may be good candidates in this regard. Therefore, we compared the *M. tuberculosis* H37Rv genes with their one-to-one orthologues (as identified by orthoMCL) in *M. canetti* and *M. marinum*, respectively, for the identification of conserved and rapidly evolving genes. A total of 2761 H37Rv-*canettii-marinum* orthologous gene sets were thus analyzed. We then define conserved and rapidly evolving genes as those that have their evolutionary rates fall outside of two standard deviations from the average (see Materials and methods). We conducted this analysis separately for d_S, d_N, and d_N/d_S. Since the evolutionary rates may occasionally be zero (especially for d_N), we used a mixture distribution of zero and log-normal distribution for each evolutionary measurement for this analysis (see Materials and methods). It is noteworthy that, for the analysis of d_N/d_S, we removed those genes whose d_S deviates significantly from d_i of the same window. However, we cannot evaluate the deviation of d_S from d_i in the comparison between *M. tuberculosis* H37Rv and *M. marinum* because the alignable intergenic regions are too short and fragmented for reliable estimations of d_i. Table 3 shows that 40.3%, 0.1%, and 11.7% of *M. tuberculosis* H37Rv genes have their d_N, d_S, and d_N/d_S, respectively, significantly lower than the genome average when compared against the *M. canettii* genes. Meanwhile, the corresponding numbers are 2.9%, 2.2%, and 3.0%, respectively, in the comparison against *M. marinum* genes. The most possible explanation for such differences in the percentage of conserved genes is the short divergence time between *M. tuberculosis* and *M. canettii* (Falush, 2009). This short time span is insufficient for nucleotide substitutions to accumulate, thus leading to low d_N and d_S values and lack of resolution in this analysis. The low d_N and d_S values in the *M. tuberculosis–M. canettii* comparison also render the estimations of d_N/d_S unreliable. By comparison, the small percentage (~3%) of genes conserved between *M. tuberculosis* and *M. marinum* are likely to be responsible for fundamental biological functions of *Mycobacteria*. These genes thus may serve as candidate drug targets.

Meanwhile, we observed fewer rapidly evolving genes than conserved genes in terms of d_N/d_S in both sets of pair-wise comparisons. These high-d_N/d_S genes are subject either to positive selection or relaxed negative selection. In the former case, these genes may be related to the adaptations of MTBC to the human environments. While in the latter case, the genes are biologically unimportant for MTBC, and may become pseudogenes afterwards. The gene IDs and the functional categories of the conserved and rapidly evolving genes in the H37Rv–*marinum* comparison are given in the Supplementary Tables 2 and 3.

3.4. Conclusions

In this study, we conducted a genome-wide analysis of d_S, d_N, d_N/d_S, d_i, and indel rate for MTB. We discovered significant fluctuations in all of these evolutionary rates, suggesting large variations in regional mutation rate and selection pressure across the MTB genome. We also showed that a considerable proportion of the synonymous substitutions are subject to natural selection, possibly for enhancing translational efficiency or for other regulatory reasons. The large number of genes with their d_S deviating from selective neutrality thus calls for caution in the application of the d_N/d_S ratio as a measurement for selection pressure on *M. tuberculosis* genes.

Acknowledgments

This study was supported by the National Science Council, Taiwan (under contract number NSC100-3111-B-400-001) and the intramural funding of National Health Research Institutes (PH101-SP-04) to F.-C. Chen. We thank Dr. Pei-Sheng Lin and Dr. Wen-Chang Wang for statistical advice. We also thank Dr. Horng-Yunn Dou for constructive comments and suggestions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2012.11.033.

References

Albers, C.A., Lunter, G., MacArthur, D.G., McVean, G., Ouwehand, W.H., Durbin, R., 2011. Dindel: accurate indel calls from short-read data. Genome Res. 21 (6), 961–973.

Darling, A.C., Mau, B., Blattner, F.R., Perna, N.T., 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14 (7), 1394–1403.

Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113.

Falush, D., 2009. Toward the use of genomics to study microevolutionary change in bacteria. PLoS Genet. 5 (10), e1000627.

Filliol, I., et al., 2006. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J. Bacteriol. 188 (2), 759–772.

Gagneux, S., et al., 2006. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc. Natl. Acad. Sci. U. S. A. 103 (8), 2869–2873.

Gutacker, M.M., et al., 2006. Single-nucleotide polymorphism-based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. J. Infect. Dis. 193 (1), 121–128.

Gutierrez, M.C., et al., 2005. Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. PLoS Pathog. 1 (1), e5.

Kryazhimskiy, S., Plotkin, J.B., 2008. The population genetics of dN/dS. PLoS Genet. 4 (12), e1000304.

Lew, J.M., Kapopoulou, A., Jones, L.M., Cole, S.T., 2011. TubercuList—10 years after. Tuberculosis (Edinb.) 91 (1), 1–7.

Li, L., Stoeckert Jr., C.J., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13 (9), 2178–2189.

Livny, J., Teonadi, H., Livny, M., Waldor, M.K., 2008. High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs. PLoS One 3, e3197.

Mokrousov, I., Narvskaya, O., Limeschenko, E., Vyazovaya, A., Otten, T., Vyshnevskiy, B., 2004. Analysis of the allelic diversity of the mycobacterial interspersed repetitive units in *Mycobacterium tuberculosis* strains of the Beijing family: practical implications and evolutionary considerations. J. Clin. Microbiol. 42 (6), 2438–2444.

Namouchi, A., Didelot, X., Schock, U., Gicquel, B., Rocha, E.P., 2012. After the bottleneck: genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation, recombination, and natural selection. Genome Res. 22 (4), 721–734.

Schurch, A.C., van Soolingen, D., 2012. DNA fingerprinting of *Mycobacterium tuberculosis*: from phage typing to whole-genome sequencing. Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis. 12 (4), 602–609.

Sharp, P.M., Li, W.H., 1987. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15 (3), 1281–1295.

Smith, N.H., Hewinson, R.G., Kremer, K., Brosch, R., Gordon, S.V., 2009. Myths and misconceptions: the origin and evolution of *Mycobacterium tuberculosis*. Nat. Rev. Microbiol. 7 (7), 537–544.

Thomas, C.M., Nielsen, K.M., 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat. Rev. Microbiol. 3 (9), 711–721.

Toll-Riera, M., Laurie, S., Alba, M.M., 2011. Lineage-specific variation in intensity of natural selection in mammals. Mol. Biol. Evol. 28 (1), 383–398.

Veyrier, F., Pletzer, D., Turenne, C., Behr, M.A., 2009. Phylogenetic detection of horizontal gene transfer during the step-wise genesis of *Mycobacterium tuberculosis*. BMC Evol. Biol. 9, 196.

Veyrier, F.J., Dufort, A., Behr, M.A., 2011. The rise and fall of the *Mycobacterium tuberculosis* genome. Trends Microbiol. 19 (4), 156–161.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24 (8), 1586–1591.

Zhang, Y., Zhang, H., Zhou, T., Zhong, Y., Jin, Q., 2011. Genes under positive selection in *Mycobacterium tuberculosis*. Comput. Biol. Chem. 35 (5), 319–322.