Gene xxx (2012) xxx–xxx

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 xxxx

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

Please cite this article as: Wang, T.-C., Chen, F.-C., The evolutionary landscape of the *Mycobacterium tuberculosis* genome, Gene (2012), http://dx.doi.org/10.1016/j.gene.2012.11.033