MBE Advance Access published March 26, 2013

Evidence for Diversifying Selection in a Set of *Mycobacterium tuberculosis* Genes in Response to Antibiotic- and Nonantibiotic-Related Pressure

Nuno S. Osório,*[1,2] Fernando Rodrigues,[1,2] Sebastien Gagneux,[3,4] Jorge Pedrosa,[1,2] Marta Pinto-Carbó,[5] António G. Castro,[1,2] Douglas Young,[6,7] Iñaki Comas,[5,8] and Margarida Saraiva[1,2]

[1]Life and Health Sciences Research Institute (ICVS), School of Health Sciences, University of Minho, Braga, Portugal

[2]ICVS/3B's - PT Government Associate Laboratory, Braga/Guimarães, Portugal

[3]Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Basel, Switzerland

[4]University of Basel, Basel, Switzerland

[5]Genomics and Health Unit, Centre for Public Health Research (CSISP), Valencia, Spain

[6]MRC National Institute for Medical Research, Mill Hill, London, United Kingdom

[7]Division of Medicine and Centre for Molecular Microbiology and Infection, Imperial College London, London, United Kingdom

[8]CIBER in Epidemiology and Public Health, Madrid, Spain

***Corresponding author:** E-mail: nosorio@ecsaude.uminho.pt.

Associate editor: Howard Ochman

Abstract

Tuberculosis (TB) is a global health problem estimated to kill 1.4 million people per year. Recent advances in the genomics of the causative agents of TB, bacteria known as the *Mycobacterium tuberculosis* complex (MTBC), have allowed a better comprehension of its population structure and provided the foundation for molecular evolution analyses. These studies are crucial for a better understanding of TB, including the variation of vaccine efficacy and disease outcome, together with the emergence of drug resistance. Starting from the analysis of 73 publicly available genomes from all the main MTBC lineages, we have screened for evidences of positive selection, a set of 576 genes previously associated with drug resistance or encoding membrane proteins. As expected, because antibiotics constitute strong selective pressure, some of the codons identified correspond to the position of confirmed drug-resistance-associated substitutions in the genes *embB*, *rpoB*, and *katG*. Furthermore, we identified diversifying selection in specific codons of the genes Rv0176 and Rv1872c coding for MCE1-associated transmembrane protein and a putative L-lactate dehydrogenase, respectively. Amino acid sequence analyses showed that in Rv0176, sites undergoing diversifying selection were in a predicted antigen region that varies between "modern" lineages and "ancient" MTBC/BCG strains. In Rv1872c, some of the sites under selection are predicted to impact protein function and thus might result from metabolic adaptation. These results illustrate that diversifying selection in MTBC is happening as a consequence of both antibiotic treatment and other evolutionary pressures.

Key words: **diversifying selection, positive selection, *Mycobacterium*, tuberculosis, genetic diversity, computational molecular biology, evolution, phylogeny, drug resistance, genomics.**

Introduction

Tuberculosis (TB) is a global health concern killing approximately 1.4 million people every year (World Health Organization 2012). All members of the *Mycobacterium tuberculosis* complex (MTBC) are potentially pathogenic to humans; however, the majority of TB cases in humans are due to *M. tuberculosis* and *M. africanum* (Grange 2001). One of the most intriguing aspects of TB is the wide spectrum of outcomes observed upon infection, ranging from pathogen clearance to the establishment of latency or development of active disease (Constant and Bottomly 1997; Lin and Flynn 2010). MTBC strains are genetically monomorphic bacteria harboring relatively low genetic variability (Achtman 2008), and thus, the heterogeneous TB outcomes have been primarily attributed to host and environmental factors (Lin and Flynn 2010). However, there is an increasing body of evidence supporting that the existing variability in MTBC strains has relevance in TB pathogenesis (De Jong et al. 2008; Lari et al. 2009; Coscolla and Gagneux 2010; Rakotosamimanana et al. 2010; Portevin et al. 2011). Thus, the identification of the functionally relevant genetic variability in MTBC strains might contribute to improved diagnostic, prophylactic, and therapeutic strategies that are needed to tackle TB.

Coevolution between the host and pathogen is a powerful determinant of adaptation in interacting species (Woolhouse et al. 2002). In a process that became known as "evolutionary arms race" (Dawkins and Krebs 1979), an adaptation in the host-immune system may lead to a counter adaptation in the pathogen and vice versa. Examples of this come from several studies in viral, bacterial, and protozoan human pathogens, which have revealed that surface-exposed proteins harboring antigens tend to be under diversifying selection to evade the host-immune system (Farci 2000; Urwin et al. 2004; Jeffares et al. 2007; Kawashima et al. 2009). Most importantly, it has been shown in *M. tuberculosis* that a large subset of T-cell antigens are hyperconserved suggesting that some level of immune recognition may be beneficial for *M. tuberculosis* (Comas et al. 2010). However, the presence of a small subset of variable epitopes subjected to diversifying selection to evade the host immune system cannot be excluded and would be of relevance for TB vaccination strategies (Comas et al. 2010). Antibiotics are also well recognized to impose strong evolutionary pressures on pathogens (Maclean et al. 2010). In the case of MTBC, in which evidence for ongoing horizontal gene transfer is scarce (Krzywinska et al. 2004; Jang et al. 2008; Comas et al. 2010), the development of drug resistance appears to be mainly due to nonsynonymous single-nucleotide substitutions, insertions, and/or deletions (Sandgren et al. 2009).

Genes undergoing positive or diversifying selection can be inferred from sequence data by detecting when the ratio of nonsynonymous to synonymous substitutions (d*N*/d*S* or ω ratio) is superior to 1. Initial approaches of this method averaging ω rates over all sites in a gene had limited power. In fact, this has been found to be an overstringent criterion for detecting diversifying selection, as ω ratios averaged over all sites are rarely greater than 1 (Ward et al. 1997; Crandall et al. 1999; Bielawski and Yang 2001). The succeeding development of more powerful statistical methods has allowed new cases of diversifying selection to be identified at individual sites and lineages in various organisms (Yang and Bielawski 2000). One of these methods is implemented in the Phylogenetic Analysis by Maximum Likelihood (PAML) package (Yang 2007). It allows the application of statistical distributions to model ω variation among all codon sites in a gene and Bayes Empirical Bayes (BEB) posterior probability calculations to infer which specific sites are under significant positive selection (Anisimova et al. 2002). Therefore, positive selection can be identified even when only a small fraction of the codons in the gene is being affected (Yang et al. 2005).

Previous molecular evolution analyses in MTBC have been limited by a low number of representative genomic sequences and confined to measurements of the ω ratios averaged over whole genes (Comas et al. 2010). In this study, we have applied the statistical methods implemented in the PAML package to estimate ω ratios at individual sites in protein coding genes extracted from a set of 73 whole-genome sequences from six of the seven main lineages in MTBC. Because antibiotics and the host-immune system are strong selective constrains, we have chosen 576 genes for the analysis of diversifying selection including genes previously associated with drug resistance and genes encoding proteins that have been consistently detected in membrane fractions in two independent proteomic studies (Gu et al. 2003; de Souza et al. 2011). We found significant lines of evidence for positive selection in confirmed drug-resistance-associated genetic variants in *embB*, *rpoB*, and *katG*. In addition, we now present significant evidence for diversifying selection in specific amino acid residues of Rv0176 (MCE1-associated transmembrane protein) and Rv1872c (a putative L-lactate dehydrogenase). Our findings uncover specific amino acids in *M. tuberculosis* membrane proteins that are under diversifying selection.

Results

Phylogenetic Analysis

We have compiled 73 publicly available MTBC genomes from drug-susceptible and drug resistance strains (supplementary table S1, Supplementary Material online). The phylogeny of these strains was determined both by the presence/absence of lineage-defining large sequence polymorphisms (LSPs) (fig. 1*A*) and by a Neighbor-Joining (NJ) tree based on 52,295 genome-wide variable nucleotide positions identified across the 73 genomes (fig. 1*B*). As expected, the branching clusters obtained in the genome-wide SNPs tree (fig. 1*B*) were congruent with the ones obtained in the LSPs analysis (fig. 1*A*). The results from the phylogenetic analysis show that the genomes include strains from six of the seven main lineages of the global MTBC population structure (Gagneux and Small 2007) validating its usefulness for molecular evolution studies.

Diversifying Selection in Genes Associated with Drug Resistance

The 73 genomes used included clinical isolates with mono-, multi-, and extensive drug resistance profiles (supplementary table S1, Supplementary Material online). The antibiotics with a higher number of resistant strains included in the genome set were isoniazid (13 strains), rifampicin (6 strains), and ethambutol (6 strains). Resistance to these antibiotics is known to be associated with single-nucleotide substitutions in the genes *katG*, *rpoB*, and *embB* (Sandgren et al. 2009). Because antibiotics are well known to induce selective pressure (Maclean et al. 2010), we have used these genes to validate PAML codon-specific detection of diversifying selection in *M. tuberculosis* (Yang 2007). The likelihood ratio tests (LRTs) comparing the null models (M0, M1a, and M7) to the alternative models (M2a and M8: allow sites with $\omega > 1$) using the sequences from *katG*, *rpoB*, and *embB* extracted from the 73 genome set identified significant positive selection in all the three genes (supplementary table S2, Supplementary Material online). The BEB analysis under site models M2a and M8 detected one amino acid residue significantly (posterior probability [PP] $\geq$ 0.99) under positive selection in *katG* (codon 315) and *embB* (codon 306) and two amino acid sites significantly (PP $\geq$ 0.99) under positive selection in *rpoB* (codons 435 and 450) (table 1). Substitutions in at least one of these sites showing diversifying selection were present in the majority of the drug-resistant strains analyzed (supplementary table S3, Supplementary Material online). We have also tested 42 other genes previously associated with drug resistance (Sandgren et al. 2009), and because of the absence or low frequency of mutations in these genes in the genomes studied, no additional sites were identified (data not shown). To address the effect of sample size on

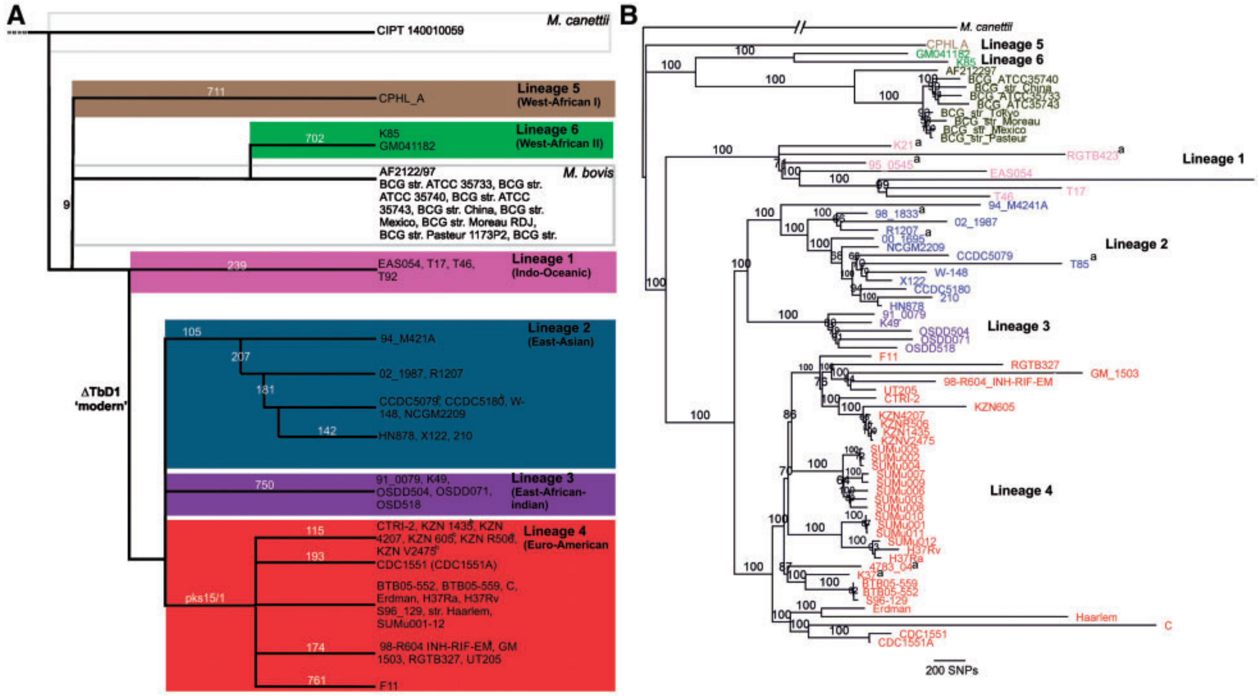

Fig. 1. Phylogeny of the 73 MTBC strains used in this study. (*A*) Phylogenetic analysis based on lineage-defining LSPs (Gagneux et al. 2006). The LSPs used are indicated by the numbers in the branches. The colored boxes indicate the main MTBC lineages that are named according to their dominance in a particular geographic area. (*B*) Distance-based NJ phylogram based on 52,295 genome-wide variable nucleotide positions across the 73 MTBC genomes. The bootstrap values shown were obtained by 100 bootstrap replications. Sequence names in each cluster are colored according to the lineage defined by LSPs analysis. Notes: [a]The sequencing information available in these strains is insufficient for LSPs classification.

the reliability and statistical power of the methodology used, we have applied PAML to a larger set of *katG* sequences (supplementary table S4, Supplementary Material online). As expected, the increase in the number of *katG* input sequences also increased the power of the method, allowing the detection of positive selection in one additional *katG* drug-resistance-associated site (*katG* codon 234, supplementary table S4, Supplementary Material online). Importantly, *katG* amino acid 315 remained significantly ($PP \geq 0.99$) under positive selection in the larger set of *katG* sequences (table 1 and supplementary table S4, Supplementary Material online). Thus, even when using a lower number of input sequences, the method proved reliable. Importantly, all sites significantly identified under diversifying selection by the conservative BEB analysis coincided with positions in which drug resistance substitutions have been previously reported (table 1). These results validate the reliability of the PAML method (Yang 2007) for diversifying selection screening in genes from this set of 73 MTBC genomes.

Diversifying Selection in Genes Encoding Membrane Proteins

Mycobacterium tuberculosis has been coevolving with its human host since at least 70,000 years ago (Comas et al., unpublished data), and membrane proteins, which are more likely to be exposed, are preferential targets for host immune-system-related selective pressure. Thus, we have focused our screen for diversifying selection on genes encoding proteins that have been consistently detected in membrane fractions in two independent proteomic studies (Gu et al. 2003; de Souza et al. 2011). From an initial list of 531 genes, we have excluded those that were hyperconserved across the 73 studied genomes and those in which paralogous sequences were found within the same genome. This resulted in a list of 238 genes (supplementary table S5, Supplementary Material online) that were screened using the PAML package for evidences of diversifying selection. The LRT statistics identified significant diversifying selection in Rv0176 ($P < 0.01$) and Rv1872c ($P < 0.001$) using both model M2a and M8 (supplementary table S2, Supplementary Material online). The BEB estimates under site models M2a and M8 detected codons under diversifying selection in Rv0176 amino acid residue 283 (M2a, $PP = 0.922$; M8, $PP = 0.957$) and 290 (M2a, $PP = 0.967$; M8, $PP = 0.984$) and in Rv1872c amino acid residues 3 (M2a, $PP = 0.991$; M8, $PP = 0.993$), 109 (M2a, $PP = 0.991$; M8, $PP = 0.993$), and 176 (M2a, $PP = 0.937$; M8, $PP = 0.938$) (table 2). We have repeated the analysis of these genes using additional sequences from a validation set of 220 MTBC genomes. The results with this extended data set fully support the previous findings with statistical significant posterior probabilities for diversifying selection in Rv0176 codons 283 (M2a, $PP = 0.962$; M8, $PP = 0.984$) and 290 (M2a, $PP = 1.000$; M8, $PP = 1.000$) and also in Rv1872c codon 3 (M2a $PP = 0.999$; M8 $PP = 1.000$), 109 (M2a $PP = 1.000$; M8

Table 1. Amino Acid Sites under Positive Selection in Rv3795 (*embB*), Rv0667 (*rpoB*), and Rv1908c (*katG*) Genes.

Gene	Description	Amino Acid Position[a]	PAML Models				Associated with Antibiotic Resistance?
			M2A		M8		
			BEB Posterior Probability of $\omega > 1$	ω Ratio ± SE	BEB Posterior Probability of $\omega > 1$	ω Ratio ± SE	
Rv3795 (*embB*)	Membrane indolylacetylinositol arabinosyltransferase, embB (3,297 nt)	306 Met	0.997**	6.193 ± 2.149	0.997**	3.990 ± 1.569	Yes, ethambutol (Ramaswamy et al. 2000)
Rv0667 (*rpoB*)	DNA-directed RNA polymerase beta chain, rpoB (3,519 nt)	435 Asp	1.000**	9.664 ± 1.094	0.999**	4.479 ± 1.378	Yes, rifampicin (Kapur et al. 1994)
		450 Ser	1.000**	9.664 ± 1.094	1.000**	4.480 ± 1.378	Yes, rifampicin (Donnabella et al. 1994)
Rv1908c (*katG*)	Catalase-peroxidase-peroxynitritase T katG (2,223 nt)	315 Ser	0.995**	7.906 ± 2.204	0.998**	7.399 ± 2.254	Yes, Isoniazid (Heym et al. 1995)

[a] Relative to H37Rv.

**Posterior probability in the BEB approach above 0.99.

Table 2. Amino Acid Sites with High Probability for Being under Positive Selection in *Mycobacterium tuberculosis* in Proteins Previously Identified in Membrane Fractions (Gu et al. 2003; de Souza et al. 2011).

Gene	Description	Amino Acid Position[a]	PAML Models			
			M2A		M8	
			BEB Posterior Probability of $\omega > 1$	ω Ratio ± SE	BEB Posterior Probability of $\omega > 1$	ω Ratio ± SE
Rv0176	MCE1-associated transmembrane protein (969 nt)	283 Pro	0.922	7.800 ± 2.805	0.957*	8.224 ± 2.479
		289 I	0.893	7.580 ± 3.014	0.939	8.085 ± 2.662
		290 Gln	0.967*	8.116 ± 2.410	0.984*	8.418 ± 2.165
Rv1629	Conserved membrane protein (1,470 nt)	660 G	0.659	1.358 ± 0.428	0.807	1.314 ± 0.428
Rv1782	Conserved membrane protein (1,521 nt)	41 R	0.851	4.094 ± 2.628	0.865	2.853 ± 1.843
Rv1872c	L-lactate dehydrogenase, lldD2 (1,245 nt)	3 V	0.991**	8.772 ± 1.847	0.993**	8.752 ± 1.927
		109 T	0.991**	8.773 ± 1.845	0.993**	8.754 ± 1.925
		253 V	0.937	8.351 ± 2.531	0.938	8.319 ± 2.619
Rv2109c	Proteasome alpha subunit, prcA (747 nt)	135 P	0.827	3.832 ± 2.691	0.848	2.651 ± 1.954
Rv3709c	Aspartokinase ask (1,266 nt)	32 Q	0.867	5.387 ± 3.110	0.899	4.829 ± 2.858

[a] Position relative to H37Rv. All codons with posterior probability of having $\omega > 1$ in the BEB approach above 0.7 are listed.

*PP $\geq$ 0.95.

**PP $\geq$ 0.99.

PP = 1.000), and 176 (M2a, PP = 1.000; M8, PP = 1.000). Because the substitution pattern found in Rv0176 codons 289, 290, and 291 is rare in *M. tuberculosis* coding regions, we have resequenced this region in additional strains from six major MTBC lineages. The results obtained (fig. 2) confirm the previously existing sequencing data and also highlight substitutions in Rv0176 codons 289, 290, and 291 that might represent intermediate states of the evolution from the sequence most prevalent in the "ancient" strains to the sequence more frequent in "modern" strains. Overall, these results constitute strong evidence for the presence of nonantibiotic-related diversifying selection in the genes Rv0176 and Rv1872c.

Ancestral State Reconstruction of Sites Showing Diversifying Selection

The detection of substitutions in different lineages constitutes strong evidence of independent events arising from convergent evolution. Thus, we compared the distribution of the nucleotide variants in the diversifying selection sites of Rv0176 and Rv1872c among the different MTBC lineages. The ancestral state reconstruction analysis showed that the substitutions in Rv0176 codons 283, 289, and 290 and in Rv1872c codons 3 and 283 are homoplastic, whereas the substitutions in Rv1872c codon 109 are only present in strains from MTBC Lineage 1. To validate the analysis in a distinct and larger population, we have used an additional set of sequences from 220 MTBC genomes from all known MTBC lineages, including four strains from Lineage 7 that has only recently been described (Blouin et al. 2012; Firdessa et al. 2013). The results confirm the presence of homoplastic substitutions in Rv0176 codons 289 and 290 and in Rv1872c codons 3 and 283 (fig. 3). With respect to the Rv0176 codon 283, no nucleotide substitutions were found in the additional 64 Lineage 4 sequences analyzed, suggesting low frequency of substitutions in this codon within this lineage. With the exception of Rv1872c codon 109, in which the nucleotide substitutions present seem to result from an isolated event in MTBC evolution, the ancestral reconstruction of the other diversifying selection sites in Rv0176 and Rv1872c is consistent with the occurrence of convergent evolution.

Amino Acid Sequence Analysis of the Sites under Diversifying Selection

To predict the functional impact of the nonsynonymous substitutions on Rv0176 and Rv1872c, we used the SIFT algorithm (Kumar et al. 2009) to compare *M. tuberculosis* Rv0176 and Rv1872c to sets of homologous sequences from other Actinobacteria. In Rv0176, the only substitution predicted to have a functional impact was I289M, with the other substitutions having SIFT scores $\geq$ 0.05 (table 3). In Rv1872c, a functional impact was predicted for T109I, A176V, and R286H (table 3). In accordance with the previous proteomic studies, further analysis of the amino acid sequence showed that both Rv0176 and Rv1872c have a high probability for membrane association (Rv0176, GRAVY score = 0.045; Rv1872c, GRAVY score = 0.058). The prediction of transmembrane domains (TMDs) suggested the existence of three TMDs in Rv0176 by using both TMHMM and PSORT tools (fig. 4*A*). As for Rv1872c, no TMDs were predicted with TMHMM, and one TMD was predicted using PSORT. The map of the diversifying selection sites and the features annotated in the major protein signature databases (Zdobnov and Apweiler 2001) highlights that Rv0176 diversifying selection hotspot is located in a large predicted extracellular domain and does not coincide with known post-translational modification regions, binding or active sites, or other functional motifs (fig. 4*A*). To investigate whether this region could be a site of recognition by the immune system, we used a semiempirical method for the prediction of antigenic regions (Kolaskar and Tongaonkar 1990) included in the EMBOSS antigenic tool (see Materials and Methods and supplementary table S6, Supplementary Material online). The nonsynonymous substitutions in this region affected the length of a predicted antigenic region (PAR) in Rv0176 (fig. 4*A*). The most common length of this PAR is 13 amino acids (286–298, score 1.079), as predicted for *M. tuberculosis* H37Rv, several other "modern" *M. tuberculosis*, *M. cannettii* CIPT 140010059, and *M. africanum* GM041182. In the other strains, the PAR varies from nine amino acids (290–298, score 1.077) in the *M. tuberculosis* RGTB423 strain, to 12 amino acids (287–298, score 1.073) in the "modern" Lineage 4 *M. tuberculosis* strains BTB05-559, BTB05-552, and SG96_129, or to 14 amino acids (285–298, score 1.079) in the Lineage 5 strain CPHL_A and *M. bovis* BCG strains. Regarding Rv1872c, with the exception of the amino acid residue 3, the sites under diversifying selection are within the Flavin mononucleotide (FMN)-dependent alpha-hydroxyl acid dehydrogenase motif and thus have the potential to influence the function of this putative lactate dehydrogenase.

Discussion

In this study, we analyzed a set of 576 genes in all the main lineages of MTBC and present evidence for the occurrence of targeted events of diversifying selection in these bacteria. As expected, because antibiotics are strong inducers of selective pressure, some of the sites we identified as undergoing diversifying selection corresponded to known drug resistance conferring mutations. These results are a validation of the use of PAML with this set of MTBC sequences. Our data are also in agreement with previous studies in other organisms, in which diversifying selection sites were associated with drug resistance (Jeffares et al. 2007; Petersen et al. 2007). In addition to antibiotics, another major selective pressure driving diversifying evolution in pathogens is the host immune system. In this context, surface-exposed protein regions are likely to be the primary targets of diversifying selection (Dawkins and Krebs 1979; Woolhouse et al. 2002). In support of this view, we found evidence for nonantibiotic-resistance-related diversifying selection in two membrane-associated proteins, denoted Rv0176 and Rv1872c.

One limitation of previous studies on the molecular evolution of MTBC was the unavailability of genomic sequences representative of the main phylogenetic lineages. By performing a phylogenetic analysis based on LSPs or single-nucleotide polymorphisms (Comas et al. 2009), we ensured a robust

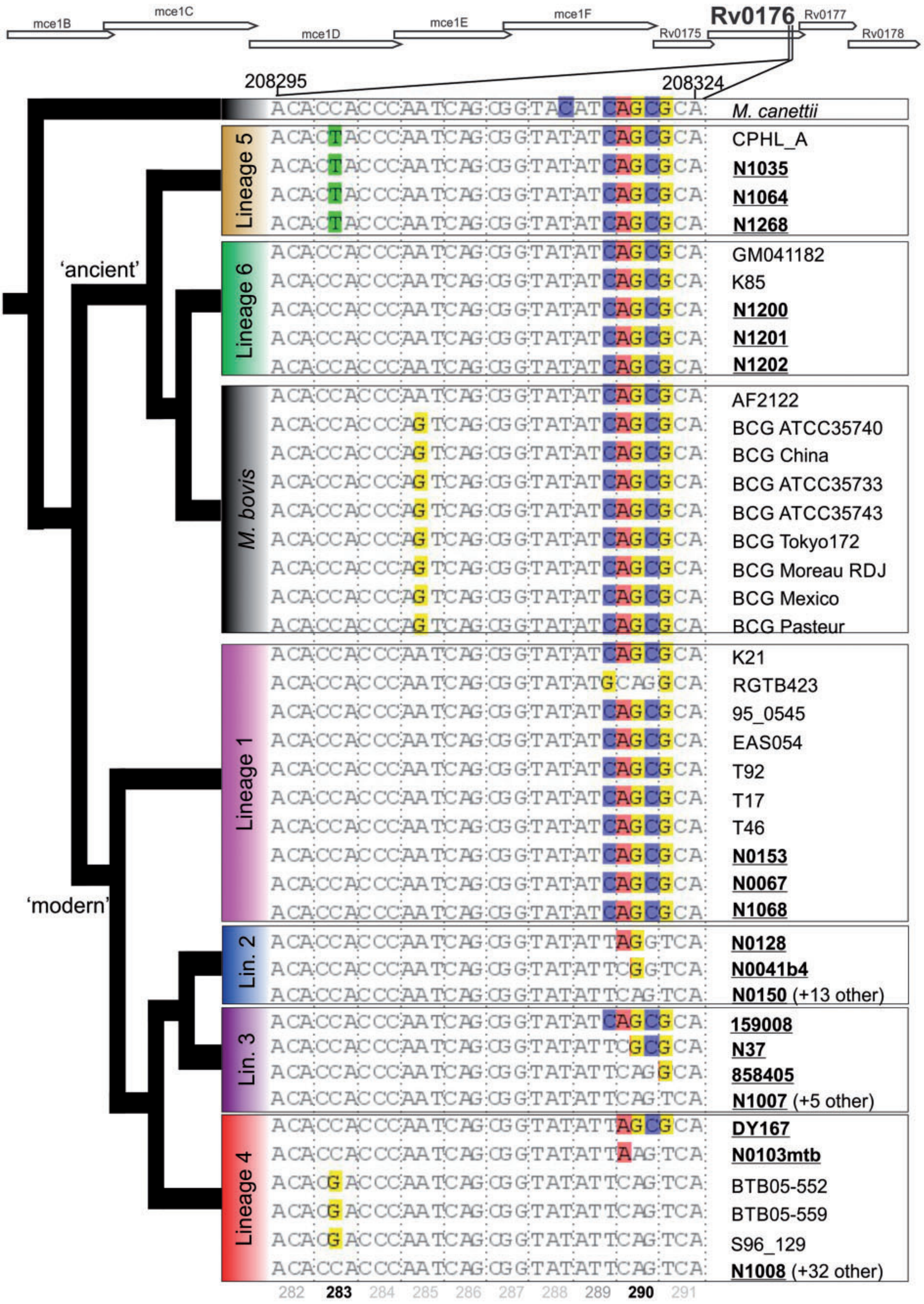
mce1B
mce1C
mce1D
mce1E
mce1F
Rv0175
Rv0176
Rv0177
Rv0178
208295
208324
'ancient'
'modern'
Lineage 5
Lineage 6
M. bovis
Lineage 1
Lin. 2
Lin. 3
Lineage 4
ACACCACCCAATCAGCGGTACATCAGCGCA M. canettii
ACACTACCCAATCAGCGGTATATCAGCGCA CPHL_A
ACACTACCCAATCAGCGGTATATCAGCGCA N1035
ACACTACCCAATCAGCGGTATATCAGCGCA N1064
ACACTACCCAATCAGCGGTATATCAGCGCA N1268
ACACCACCCAATCAGCGGTATATCAGCGCA GM041182
ACACCACCCAATCAGCGGTATATCAGCGCA K85
ACACCACCCAATCAGCGGTATATCAGCGCA N1200
ACACCACCCAATCAGCGGTATATCAGCGCA N1201
ACACCACCCAATCAGCGGTATATCAGCGCA N1202
ACACCACCCAATCAGCGGTATATCAGCGCA AF2122
ACACCACCCAGTCAGCGGTATATCAGCGCA BCG ATCC35740
ACACCACCCAGTCAGCGGTATATCAGCGCA BCG China
ACACCACCCAGTCAGCGGTATATCAGCGCA BCG ATCC35733
ACACCACCCAGTCAGCGGTATATCAGCGCA BCG ATCC35743
ACACCACCCAGTCAGCGGTATATCAGCGCA BCG Tokyo172
ACACCACCCAGTCAGCGGTATATCAGCGCA BCG Moreau RDJ
ACACCACCCAGTCAGCGGTATATCAGCGCA BCG Mexico
ACACCACCCAGTCAGCGGTATATCAGCGCA BCG Pasteur
ACACCACCCAATCAGCGGTATATCAGCGCA K21
ACACCACCCAATCAGCGGTATATGCAGGCA RGTB423
ACACCACCCAATCAGCGGTATATCAGCGCA 95_0545
ACACCACCCAATCAGCGGTATATCAGCGCA EAS054
ACACCACCCAATCAGCGGTATATCAGCGCA T92
ACACCACCCAATCAGCGGTATATCAGCGCA T17
ACACCACCCAATCAGCGGTATATCAGCGCA T46
ACACCACCCAATCAGCGGTATATCAGCGCA N0153
ACACCACCCAATCAGCGGTATATCAGCGCA N0067
ACACCACCCAATCAGCGGTATATCAGCGCA N1068
ACACCACCCAATCAGCGGTATATTAGGTCA N0128
ACACCACCCAATCAGCGGTATATTCGGTCA N0041b4
ACACCACCCAATCAGCGGTATATTCAGTCA N0150 (+13 other)
ACACCACCCAATCAGCGGTATATCAGCGCA 159008
ACACCACCCAATCAGCGGTATATTCGCGCA N37
ACACCACCCAATCAGCGGTATATTCAGGCA 858405
ACACCACCCAATCAGCGGTATATTCAGTCA N1007 (+5 other)
ACACCACCCAATCAGCGGTATATTAGCGCA DY167
ACACCACCCAATCAGCGGTATATTAAGTCA N0103mtb
ACACGACCCAATCAGCGGTATATTCAGTCA BTB05-552
ACACGACCCAATCAGCGGTATATTCAGTCA BTB05-559
ACACGACCCAATCAGCGGTATATTCAGTCA S96_129
ACACCACCCAATCAGCGGTATATTCAGTCA N1008 (+32 other)
282
283
284
285
286
287
288
289
290
291

Fig. 2. Nucleotide sequence alignment of Rv0176 diversifying selection hotspot in strains of the main MTBC lineages. The codons 283 and 290 found to be under positive selection are highlighted in bold. The alignment shows the nucleotide from 208,295 to 208,324 (coordinates relative to H37Rv reference genome) and the reading frame of the codons 282 to 291. The sequence from the underlined strains was confirmed in this study.

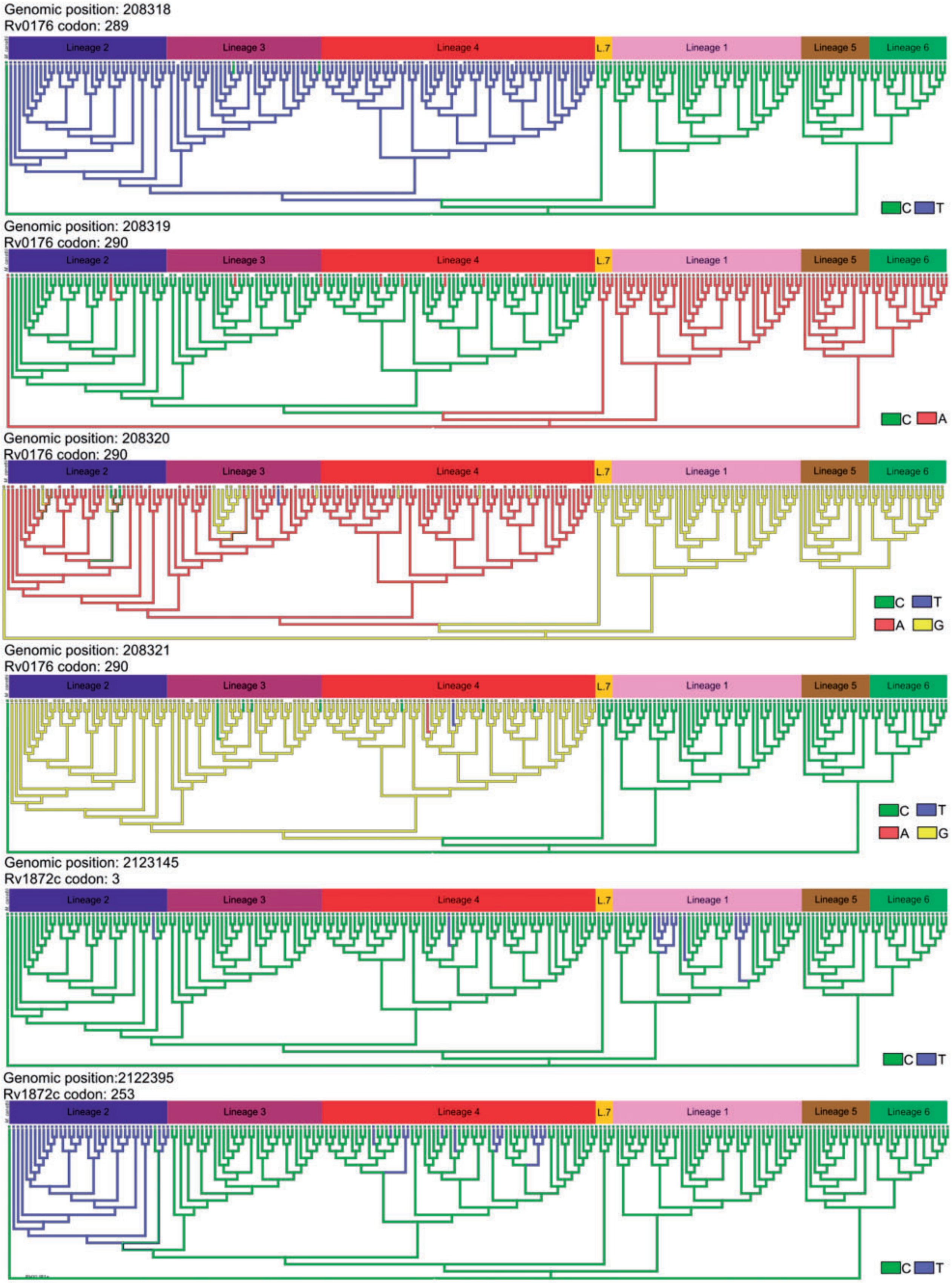

Fig. 3. Homoplastic nonsynonymous substitutions in Rv0176 and Rv1872c mapped on a phylogeny obtained using a set of 220 genomes representative of all the known MTBC lineages. For each substitution, the genomic position relative to H37Rv genome and the number of the gene codon are indicated. The results were obtained by performing parsimony ancestral reconstruction in MESQUITE (Maddison WP and Maddison DR 2001). *Mycobacterium canetti* was used as the outgroup for the ancestral reconstruction.

Table 3. Functional Impact Prediction of the Nonsynonymous Substitutions in Rv0176 and Rv1872c Using SIFT.

Gene	Substitution	Sift Score[a]	Sequences Represented at This Position[b]
Rv0176	R145L	0.05	98
	P283L	0.21	42
	N285S	0.49	42
	I289M	0.02	42
	Q290S	0.91	42
	R301H	0.16	41
	P318S	0.09	27
Rv1872c	A2S	0.26	28
	V3I	0.08	91
	A59G	0.10	119
	L96F	0.29	119
	T109I	0.01	119
	A176V	0.01	119
	D217N	0.39	119
	A237S	0.08	119
	V253M	1.00	118
	L258V	0.40	119
	R286H	0.01	119
	R291H	0.60	119
	A339V	0.06	119
	T379A	0.73	119

[a] Substitutions underlined have a SIFT score below 0.05 and are predicted to have a functional impact.

[b] The sequence database used for SIFT included 125 Actinobacteria sequences. The number in this column shows for each specific position the number of sequences included in the analysis.

analysis for these monomorphic clades. The phylogeny results obtained by both methods were congruent and indicated that the MTBC genomes under study included strains of all the main lineages, and thus representing suitable data set for molecular evolutionary studies. Additionally, as shown in our analysis of *katG*, our results indicate that the number of sequences used allowed reliable detection of diversifying selection. Indeed, diversifying selection in codon 315 from *katG* sequences extracted from whole-genome set was maintained in a larger sequence set. The inclusion of additional sequences did enhance the power of the method, as a new diversifying selection site corresponding to a known drug resistance mutation was detected. In accordance with previous studies (Yang et al. 2005), our data show that if different models are used and the posterior probabilities are estimated with a BEB approach, PAML is not prone to false positives even in small data sets. Other important factors that may mislead the methods to detect diversifying selection are recombination and horizontal gene transfer events, but these are considered rare events in MTBC (Krzywinska et al. 2004; Jang et al. 2008). In all, our study supports that the PAML method is a reliable tool for diversifying selection scans in MTBC.

Our results from amino acid sequence analysis suggest differences in the nature of the pressures underlying the site-specific diversifying selection events in Rv0176 and Rv1872c. Both genes are nonessential (Sassetti et al. 2003). However, previous studies on Rv0176 suggest it has an important role in vivo, as it is required for bacterial growth in a mouse model (Sassetti et al. 2003) and in primary murine macrophages (Rengarajan et al. 2005).

Rv0176 is an RDD family protein of unknown function with three predicted TMDs. The substitutions in amino acid residues 283, 289, and 290 are homoplastic and, in the case of codons 289 and 290, are likely to have occurred through an “ancient" to “modern" substitution trajectory at the level of each nucleotide. These loci of diversifying selection are located in a large predicted extracellular domain of Rv0176 not coinciding with any functional annotation. SIFT (Ng and Henikoff 2003) prediction of functional consequences of the substitutions in this region suggested no functional impact at the positions 283 and 290, as could be expected in a region of antigen variability. In accordance, the prediction of antigenic regions in Rv0176 using a semiempirical method (Kolaskar and Tongaonkar 1990) indicates that these substitutions affect the length of a PAR. Although coevolution of MTBC with its human host is not consistent with a classical “arms-race” model (Comas et al. 2010), our results suggest that this Rv0176 region might be a variable antigen that varies among “modern" and “ancient" MTBC/*M. bovis* BCG strains. Future research toward the identification and experimental validation of variable MTBC epitopes could be of high relevance for the design of improved vaccination strategies.

As for Rv1872c, a putative L-lactate dehydrogenase, the diversifying selection sites in codons 109 and 176 are predicted by SIFT to impact the function of Rv1872c and are located within the FMN-dependent alpha-hydroxy acid dehydrogenase motif. The analysis of the region upstream of Rv1872c start codon reveals one other homoplastic nucleotide substitution in close proximity to the substitution in codon 3. It also reveals that Rv1872c is an atypical leaderless gene without Shine-Dalgarno or TANNNT translation initiation signals. Thus, one can speculate that the diversifying selection site in codon 3 might somehow be involved in the initiation of translation. Overall, these results may uncover a possible metabolic adaptation of some MTBC strains to specific host environments such as anaerobic conditions.

In summary, we show evidence for the occurrence of targeted events of antibiotic and nonantibiotic-related diversifying selection in MTBC. The power of this analysis might increase as the number of genomes representative of wider MTBC variability increases. However, the fact that MTBC strains harbor little DNA sequence diversity and that the majority of the MTBC antigenic regions are under negative selection (Comas et al. 2010) suggests that the number of genes with sites under diversifying selection will only be a small fraction of the coding genome. This observation raises the relevance of our results. The experimental validation of the functional role of the sites herein identified might inform future strategies in the global fight against TB.

Materials and Methods

Sequence Retrieval

We have studied the phylogeny and performed diversifying selection screenings on 73 publicly available MTBC genome

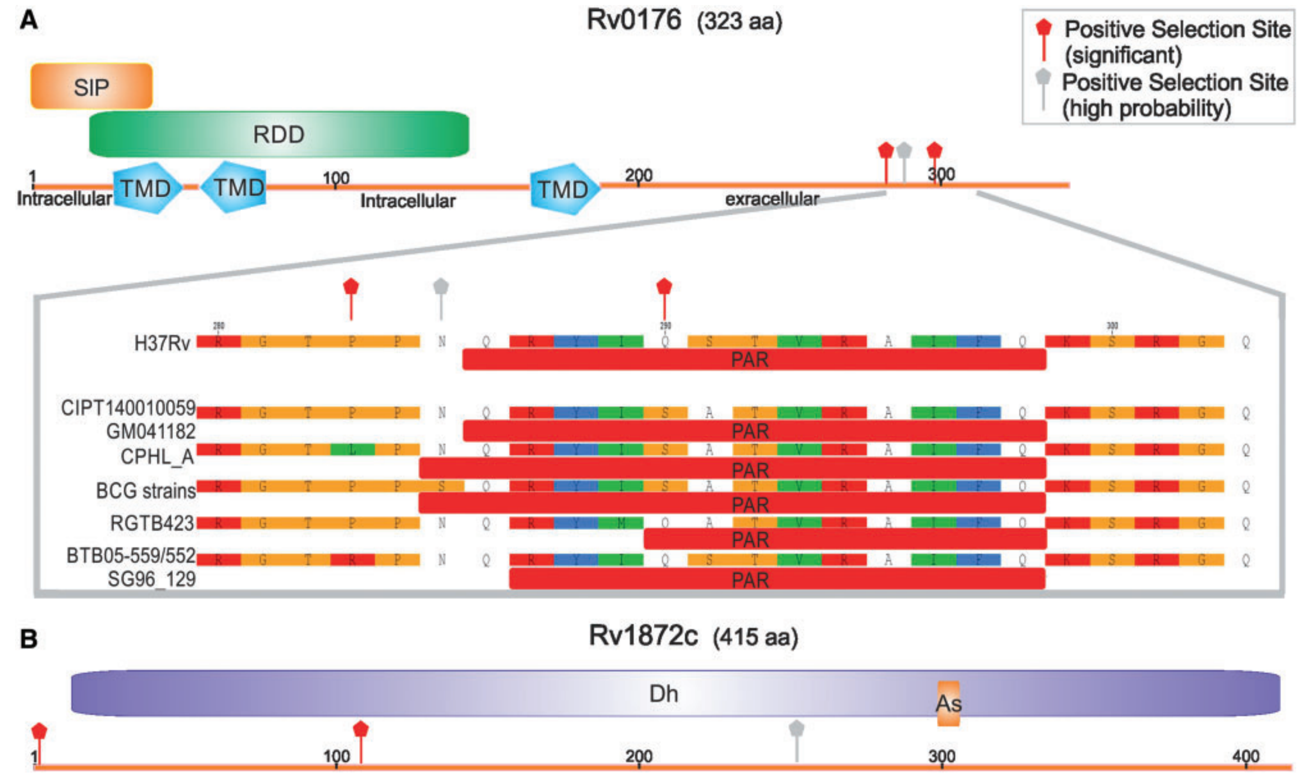
A
Rv0176 (323 aa)
Positive Selection Site (significant)
Positive Selection Site (high probability)
SIP
RDD
1
Intracellular
TMD
TMD
100
Intracellular
TMD
200
exracellular
300
280
290
300
H37Rv R G T P P N Q R Y I Q S T V R A I F Q K S R G Q
PAR
CIPT140010059 R G T P P N Q R Y I S A T V R A I F Q K S R G Q
GM041182
PAR
CPHL_A R G T L P N Q R Y I S A T V R A I F Q K S R G Q
PAR
BCG strains R G T P P S Q R Y I S A T V R A I F Q K S R G Q
PAR
RGTB423 R G T P P N Q R Y M Q A T V R A I F Q K S R G Q
PAR
BTB05-559/552 R G T R P N Q R Y I Q S T V R A I F Q K S R G Q
SG96_129
PAR
B
Rv1872c (415 aa)
Dh
As
1
100
200
300
400

Fig. 4. Schematic representation of the annotated protein sequences encoded by Rv0179 and Rv1872c. The schematic representation indicates the location of the sites under positive selection and predicted functional domains annotated in the major protein signature databases. SIP, predicted signal peptide; RDD, PFAM domain PF06271; TMD, transmembrane domains predicted using TMHMM and PSORT; PAR, antigenic region predict by EMBOSS antigenic; Dh, FMN-dependent alpha-hydroxyl acid dehydrogenase domain (PFAM PF01070); As, FMN-dependent alpha-hydroxyl acid dehydrogenase domain active site (PROFILE PS00557).

sequences retrieved from the NCBI database and the TB Diversity Sequencing Project (http://genome.tbdb.org/annotation/genome/tbdb/ReseqStrainInfo.html). It includes 62 *M. tuberculosis* clinical isolates originating from different geographic locations, 9 *M. bovis* strains (including BCG vaccine strains), and 2 *M. tuberculosis* laboratorial strains. In what regards to drug resistance profile, the *M. tuberculosis* clinical isolates used include susceptible, mono-, multi-, and extensively resistant strains as detailed in supplementary table S1, Supplementary Material online. Gene sequences were extracted by querying with the H37Rv sequence a local database of the 73 genomes using MegaBlast (Zhang et al. 2000). Hyperconserved genes (defined by a mean number of pairwise nucleotide differences—Tagima's π value—across the 73 sequences below 0.0001) and paralogous sequences (resulting in more than one complete hit within the same genome) were excluded from the diversifying selection analysis. An additional set of 220 MTBC genome sequences representative of all major MTBC lineages and geographic regions obtained in a previous study and including 44 strains from Lineage 1, 37 strains from Lineage 2, 36 strains from Lineage 3, 64 strains from Lineage 4, 16 strains from Lineage 5, 18 strains from Lineage 6, and 4 strains from Lineage 7 (Comas et al., unpublished data) was used to validate the diversifying selection analysis and to confirm the ancestral reconstruction of the nonsynonymous substitutions present in Rv0176 and Rv1872c.

DNA Sequencing

Oligonucleotide primers were designed for polymerase chain reaction (PCR) amplification and sequencing of Rv0176. DNA was amplified by PCR in a 96-well format 50-µl reaction. PCR products were purified and sequenced by Sanger sequencing method. Sequence chromatogram files were analyzed using the Staden package (Staden 1996). To identify sequence polymorphisms, the consensus sequence for each strain was compared with the corresponding gene sequence of the H37Rv reference genome using MEGA 5 software (Tamura et al. 2011).

Phylogenetic Analysis and Ancestral State Reconstruction

The phylogenetic analysis of the sequences was performed by detecting the presence/absence of previously reported lineage-defining LSPs and by the use of genome-wide variable nucleotide positions across the 73 MTBC genomes collected (Comas et al. 2009). The set of SNPs was defined relatively to the laboratory strain H37Rv and was obtained by the use of SNPfinder (Song et al. 2005). The obtained 52,295 SNPs were concatenated, and a distance-based NJ phylogram was determined with PhyML v3.0 (Guindon et al. 2010) using *M. canettii* as the root. The resulting phylogeny was then used as the guide tree for subsequent bootstrap analysis to determine the confidence of the branching.

Bayesian analysis resulted in similar tree topology and similar levels of nodal support when compared with NJ. To identify homoplastic sites, we mapped the diversifying selection sites onto the MTBC phylogeny performed using the validation set of 220 MTBC genomes and applied in MESQUITE (Maddison WP and Maddison DR 2001), the ancestral reconstruction option using parsimony.

Molecular Evolutionary Analysis

We applied the codon substitution models implemented in the CODEML program in the PAML package (Yang 2007). Only complete and unique gene sequences were analyzed, and all ambiguous sites were removed before PAML analysis. Alignments and the NJ, maximum likelihood, and Bayesian trees were used for posterior molecular evolution analyses. Similar results were obtained for all methods of tree reconstruction, reflecting PAML robustness in respect to the phylogenetic tree used. Evidence for diversifying (positive) selection was first evaluated using LRTs through the CODEML algorithm. The null models (M0, M1a, and M7: do not allow sites with $\omega > 1$) were compared with the alternative models (M2a and M8: allow sites with $\omega > 1$). The level of significance for these LRTs was calculated using a χ^2 approximation given that twice the difference of log likelihood between the models ($2\Delta\ln L$) will asymptotically have a χ^2 distribution, with a number of degrees of freedom corresponding to the difference of parameters between the nested models. The Bayesian posterior probabilities for each site were directly calculated for the models admitting selection (M2a and M8). We used the conservative BEB approach (given its higher accuracy when compared with NEB [Yang et al. 2005]) to calculate the posterior probabilities of a specific codon site and to identify those with higher probability for being under diversifying selection.

Amino Acid Sequence Analysis

Prediction of the effects of coding nonsynonymous variants on protein function in diversifying selection regions was performed using the SIFT algorithm (Kumar et al. 2009). Substituted amino acids with a SIFT score < 0.5 are predicted to have an impact on protein function. Briefly, SIFT scores the conservation of the positions where mutations are found and weighs this score by the nature of the amino acid change. These measures are then incorporated into a proxy measure of the impact of a specific substitution on protein function. As a bacterial database, we used all the available Actinobacteria sequences ($N = 125$). The "grand mean of hydropathy" (GRAVY) scores were calculated using the PROTPARAM tool (Wilkins et al. 1999) in which a score > -0.4 (mean score for the cytosolic proteins) suggests probability for membrane association; the higher the score, the greater the probability (Kyte and Doolittle 1982). The transmembrane regions were predicted using TMHMM (Krogh et al. 2001) and PSORT (Nakai and Horton 1999). Prediction of potential antigenic determinants was based on a semiempirical method for the prediction of antigenic regions (Kolaskar and Tongaonkar 1990) included in the EMBOSS antigenic tool. We have further validated this method by applying it to 30 experimentally confirmed *M. tuberculosis* epitopes and obtained an accuracy of 64% (supplementary table S6, Supplementary Material online) that is similar to the accuracy reported by the authors for non-*M. tuberculosis* epitopes (Kolaskar and Tongaonkar 1990).

Supplementary Material

Supplementary tables S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors acknowledge Margarida Correia-Neves and Catarina L. Santos for helpful comments and suggestions. N.S.O. is supported by the Portuguese Foundation for Science and Technology (FCT) SFRH/BPD/33959/2009, and M.S. is a Ciência 2007 Fellow. I.C. is supported by Marie Curie FP7 action 272086. This work was partially supported by Ministerio de Economia y Competitividad (MINECO), Spain, project BFU2011-24112.

References

Achtman M. 2008. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu Rev Microbiol.* 62:53–70.

Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and Power of Bayes prediction of amino acid sites under positive selection. *Mol Biol Evol.* 19:950–958.

Bielawski JP, Yang Z. 2001. Positive and negative selection in the DAZ gene family. *Mol Biol Evol.* 18:523–529.

Blouin Y, Hauck Y, Soler C, et al. (13 co-authors). 2012. Significance of the identification in the Horn of Africa of an exceptionally deep branching *Mycobacterium tuberculosis* clade. *PloS One* 7:e52841.

Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, Ernst JD, Gagneux S. 2010. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet.* 42:498–503.

Comas I, Homolka S, Niemann S, Gagneux S. 2009. Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PLoS One* 4:e7815.

Constant SL, Bottomly K. 1997. Induction of Th1 and Th2 CD4+ T cell responses: the alternative approaches. *Annu Rev Immunol.* 15: 297–322.

Coscolla M, Gagneux S. 2010. Does *M. tuberculosis* genomic diversity explain disease diversity? *Drug Discov Today Dis Mech.* 7:e43–e59.

Crandall KA, Kelsey CR, Imamichi H, Lane HC, Salzman NP. 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol Biol Evol.* 16: 372–382.

Dawkins R, Krebs JR. 1979. Arms races between and within species. *Proc R Soc Lond B Biol Sci.* 205:489–511.

De Jong BC, Hill PC, Aiken A, et al. (15 co-authors). 2008. Progression to active tuberculosis, but not transmission, varies by *Mycobacterium tuberculosis* lineage in The Gambia. *J Infect Dis.* 198:1037–1043.

De Souza GA, Leversen NA, Målen H, Wiker HG. 2011. Bacterial proteins with cleaved or uncleaved signal peptides of the general secretory pathway. *J Proteomics.* 75:502–510.

Donnabella V, Martiniuk F, Kinney D, Bacerdo M, Bonk S, Hanna B, Rom WN. 1994. Isolation of the gene for the beta subunit of RNA polymerase from rifampicin-resistant *Mycobacterium tuberculosis* and identification of new mutations. *Am J Respir Cell Mol Biol.* 11: 639–643.

Farci P. 2000. The outcome of acute hepatitis C predicted by the evolution of the viral quasispecies. *Science* 288:339–344.

Firdessa R, Berg S, Hailu E, et al. (23 co-authors). 2013. Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, Ethiopia. *Emerg Infect Dis.* 19:460–463.

Gagneux S, DeRiemer K, Van T, et al. (13 co-authors). 2006. Variable host–pathogen compatibility in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A.* 103:2869.

Gagneux S, Small PM. 2007. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *Lancet Infect Dis.* 7:328–337.

Grange JM. 2001. *Mycobacterium bovis* infection in human beings. *Tuberculosis* 81:71–77.

Gu S, Chen J, Dobos KM, Bradbury EM, Belisle JT, Chen X. 2003. Comprehensive proteomic profiling of the membrane constituents of a *Mycobacterium tuberculosis* strain. *Mol Cell Proteomics.* 2: 1284–1296.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 59:307–321.

Heym B, Alzari PM, Honore N, Cole ST. 1995. Missense mutations in the catalase-peroxidase gene, *katG*, are associated with isoniazid resistance in *Mycobacterium tuberculosis*. *Mol Microbiol.* 15:235–245.

Jang J, Becq J, Gicquel B, Deschavanne P, Neyrolles O. 2008. Horizontally acquired genomic islands in the tubercle bacilli. *Trends Microbiol.* 16: 303–308.

Jeffares DC, Pain A, Berry A, et al. (15 co-authors). 2007. Genome variation and evolution of the malaria parasite *Plasmodium falciparum*. *Nat Genet.* 39:120–125.

Kapur V, Li LL, Iordanescu S, Hamrick MR, Wanger A, Kreiswirth BN, Musser JM. 1994. Characterization by automated DNA sequencing of mutations in the gene (*rpoB*) encoding the RNA polymerase beta subunit in rifampin-resistant *Mycobacterium tuberculosis* strains from New York City and Texas. *J Clin Microbiol.* 32:1095.

Kawashima Y, Pfafferott K, Frater J, et al. (43 co-authors). 2009. Adaptation of HIV-1 to human leukocyte antigen class I. *Nature* 458:641–645.

Kolaskar AS, Tongaonkar PC. 1990. A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett.* 276:172–174.

Krogh A, Larsson B, Von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol.* 305:567–580.

Krzywinska E, Krzywinski J, Schorey JS. 2004. Naturally occurring horizontal gene transfer and homologous recombination in *Mycobacterium*. *Microbiology* 150:1707–1712.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding nonsynonymous variants on protein function using the SIFT algorithm. *Nat Protoc.* 4:1073–1081.

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol.* 157:105–132.

Lari N, Rindi L, Cristofani R, Rastogi N, Tortoli E, Garzelli C. 2009. Association of *Mycobacterium tuberculosis* complex isolates of BOVIS and Central Asian (CAS) genotypic lineages with extrapulmonary disease. *Clin Microbiol Infect.* 15:538–543.

Lin PL, Flynn JL. 2010. Understanding latent tuberculosis: a moving target. *J Immunol.* 185:15–22.

Maclean RC, Hall AR, Perron GG, Buckling A. 2010. The evolution of antibiotic resistance: insight into the roles of molecular mechanisms of resistance and treatment context. *Discov Med.* 10:112–118.

Maddison WP, Maddison DR. 2001. Mesquite: a modular system for evolutionary analysis. Version 2.75 http://mesquiteproject.org.

Nakai K, Horton P. 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci.* 24:34–36.

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31:3812–3814.

Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R. 2007. Genes under positive selection in *Escherichia coli*. *Genome Res.* 17: 1336–1343.

Portevin D, Gagneux S, Comas I, Young D. 2011. Human macrophage responses to clinical isolates from the *Mycobacterium tuberculosis* complex discriminate between ancient and modern lineages. *PLoS Pathog.* 7:e1001307.

Rakotosamimanana N, Raharimanga V, Andriamandimby SF, et al. (14 co-authors). 2010. Variation in gamma interferon responses to different infecting strains of *Mycobacterium tuberculosis* in acid-fast bacillus smear-positive patients and household contacts in Antananarivo, Madagascar. *Clin Vaccine Immunol.* 17: 1094–1103.

Ramaswamy SV, Amin AG, Göksel S, Stager CE, Dou S-J, El Sahly H, Moghazeh SL, Kreiswirth BN, Musser JM. 2000. Molecular genetic analysis of nucleotide polymorphisms associated with ethambutol resistance in human isolates of *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother.* 44:326–336.

Rengarajan J, Bloom BR, Rubin EJ. 2005. Genome-wide requirements for *Mycobacterium tuberculosis* adaptation and survival in macrophages. *Proc Natl Acad Sci U S A.* 102:8327–8332.

Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. 2009. Tuberculosis drug resistance mutation database. *PLoS Med.* 6:e2.

Sassetti CM, Boyd DH, Rubin EJ. 2003. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol.* 48: 77–84.

Song J, Xu Y, White S, Miller KWP, Wolinsky M. 2005. SNPsFinder—a web-based application for genome-wide discovery of single nucleotide polymorphisms in microbial genomes. *Bioinformatics* 21: 2083–2084.

Staden R. 1996. The Staden sequence analysis package. *Mol Biotechnol.* 5: 233–241.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 28:2731–2739.

Urwin R, Russell JE, Thompson EA., Holmes EC, Feavers IM, Maiden MC. 2004. Distribution of surface protein variants among hyperinvasive meningococci: implications for vaccine design. *Infect Immun.* 72: 5955–5962.

Ward TJ, Honeycutt RL, Derr JN. 1997. Nucleotide sequence evolution at the kappa-casein locus: evidence for positive selection within the family Bovidae. *Genetics* 147:1863–1872.

Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, Hochstrasser DF. 1999. Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol.* 112:531–552.

Woolhouse ME, Webster JP, Domingo E, Charlesworth B, Levin BR. 2002. Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat Genet.* 32:569–577.

World Health Organization. 2012. Global Tuberculosis Report 2012. WHO/HTM/TB/2012.6. Geneva (Switzerland): World Health Organization.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol.* 15:496–503.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol.* 22: 1107–1118.

Zdobnov EM, Apweiler R. 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17: 847–848.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *J Comput Biol.* 7:203–214.