Purifying Selective Pressure Suggests the Functionality of a Vitamin B12 Biosynthesis Pathway in a Global Population of *Mycobacterium tuberculosis*

Alina Minias¹,*, Piotr Minias², Bożena Czubat¹,³ and Jarosław Dziadek¹,*

¹Laboratory of Genetics and Physiology of Mycobacterium, Institute of Medical Biology, Polish Academy of Sciences, Łódź, Poland
²Department of Biodiversity Studies and Bioeducation, Faculty of Biology and Environmental Protection University of Łódź, Łódź, Poland
³Department of Biochemistry and Cell Biology, Faculty of Biology and Agriculture, University of Rzeszów, Rzeszów, Poland

*Corresponding authors: E-mails: aminias@cbm.pan.pl; jdziadek@cbm.pan.pl.

Accepted: July 27, 2018

Abstract

*Mycobacterium tuberculosis* is one of the deadliest and most challenging pathogens to study in current microbiological research. One of the issues that remains to be resolved is the importance of cobalamin in the metabolism of *M. tuberculosis*. The functionality of a vitamin B12 biosynthesis pathway in *M. tuberculosis* is under dispute, and the ability of this pathogen to scavenge vitamin B12 from the host is unknown. Here, we quantified the ratios of nonsynonymous and synonymous nucleotide substitution rates (dN/dS) in the genes involved in vitamin B12 biosynthesis and transport and in genes encoding cobalamin-dependent enzymes in nearly four thousand strains of *M. tuberculosis*. We showed that purifying selection is the dominant force acting on cobalamin-related genes at the levels of individual codons, genes and groups of genes. We conclude that cobalamin-related genes may not be essential but are adaptive for *M. tuberculosis* in clinical settings. Furthermore, the cobalamin biosynthesis pathway is likely to be functional in this species.

Key words: *Mycobacterium tuberculosis*, cobalamin, vitamin B12, natural selection, nucleotide substitution rates.

Introduction

Tuberculosis, one of the top ten causes of death worldwide, is an infectious disease caused by the bacterium *Mycobacterium tuberculosis*. Over ten million people contracted tuberculosis and 1.5 million people died from this disease in 2015 (World Health Organization 2017). The World Health Organization (WHO) describes the fight against tuberculosis as one of the top priorities of modern medicine. Tuberculosis is very difficult to treat and requires 6–24 months of antibiotic therapy (Nahid et al. 2016). The most important issue regarding effective chemotherapy for tuberculosis is the complex life cycle of the pathogen, which involves long periods of latency. After infecting a host and progressing through the active stage of an infection, *M. tuberculosis* settles into a dormant state that can last for decades. When an infected individual becomes immunocompromised, the infection reactivates (Gorna et al. 2010). Furthermore, *M. tuberculosis* is a challenging bacterium to study. First, due to the danger to the research staff, the bacterium has to be cultivated in biosafety level three laboratories. Next, the doubling time of *M. tuberculosis* is 24h, which means that it takes from 4–6 weeks to grow a colony forming unit (Piddington et al. 2000). To efficiently control tuberculosis, it is crucial to understand the metabolic mechanisms that underlie the virulence of *M. tuberculosis*. In past years, empirical results have been greatly aided by the analysis of the whole-genome sequencing data.

A puzzling problem regarding *M. tuberculosis* is its metabolism of cobalamin. Cobalamin, also known as vitamin B12, is a molecule that is thought to influence mycobacterial metabolism through two mechanisms. First, it is a cofactor for multiple enzymes. *M. tuberculosis* encodes three vitamin B12-dependent enzymes, methionine synthase (MetH), methylmalonyl CoA mutase (MutAB), and ribonucleotide reductase (NrdZ) (Savvi et al. 2008; Gopinath, Moosa et al. 2013). Furthermore, vitamin B12 may regulate gene expression by binding to riboswitches in miRNA to prevent protein translation. *M. tuberculosis* encodes two riboswitch-regulated operons, one is related to vitamin B12 synthesis and transport,
Vitamin B12 Biosynthesis Pathway in M. tuberculosis

whereas the other encodes the vitamin B12-independent methionine synthase MetE (Warner et al. 2007; Gopinath, Moosa et al. 2013; Young et al. 2015).

Vitamin B12 is a complex molecule, the de novo production of which requires over 20 energy demanding enzymatic reactions (fig. 1). Homologues of nearly all proteins necessary for the aerobic synthesis of vitamin B12 have been identified in M. tuberculosis. Possible replacements for those proteins that are absent have been suggested (Gopinath, Moosa et al. 2013). Intriguingly, the model M. tuberculosis strain H37Rv has not been observed to synthesize cobalamin during various indirect experiments. M. tuberculosis was shown to not be able to use propionate as a sole carbon source using a vitamin B12-dependent methylmalonyl pathway without the exogenous supplementation of vitamin B12 (Savi et al. 2008). Furthermore, exogenous supplementation with vitamin B12 is required for the growth of a ΔmetE M. tuberculosis H37Rv mutant, which requires a functional cobalamin-dependent methionine synthase (MetH) (Warner et al. 2007). In turn, clinical strain of M. tuberculosis CDC1551 has been shown to possibly be able to synthesize cobalamin, because it was suspected to partially activate recombinant MetH in the absence of vitamin B12 supplementation (Guzzo et al. 2016). Finally, genes involved in cobalamin biosynthesis have been shown to be upregulated in M. tuberculosis H37Rv in a dormant state induced by growth in K⁺-deficient medium (Ignatov et al. 2015). Therefore, it is unclear whether the species of M. tuberculosis are unable to synthesize vitamin B12 or whether it is a specific phenotype of the model strain M. tuberculosis H37Rv. Finally, M. tuberculosis is suspected to scavenge vitamin B12 from the host through a BacA transporter (Gopinath, Venclovas et al. 2013). However, it is unclear whether vitamin B12 is actually available for M. tuberculosis to scavenge, especially in the closed environment of granulomas.

The aim of this study was to assess the functionality of a vitamin B12 biosynthesis pathway in a global population of M. tuberculosis by assessing the mode and strength of natural selection acting on genes involved in vitamin B12 synthesis and transport as well as on cobalamin-dependent enzymes. For this purpose, we compiled a global data set of genomic sequences for nearly four thousand M. tuberculosis genome sequences (Periwal et al. 2015; Lahlou et al. 2017). The final data set consisted of genome sequences from 3,798 M. tuberculosis strains (supplementary table S1, Supplementary Material online) and the M. tuberculosis CDC1551 reference strain (accession number: NC_002755).

Materials and Methods

Construction of a Database of M. tuberculosis Genome Sequences

We used assembled genome sequences obtained from various studies and sequenced with various technologies to obtain the maximum sample size and variability of the analyzed M. tuberculosis population. Genome sequences of M. tuberculosis were downloaded from the NCBI Genome Sequence Database. To quality filter the sequences, we eliminated all the strains for which the genomes consisted of >150 contigs. The cutoff limit for number of contigs within each strain is linked to the N50 value, which is used to evaluate the quality of the sequencing. Because the N50 length is defined as the shortest sequence length at 50% of the genome, the average contig size must be equal to or lower than the N50 value. Therefore, the lowest probable N50 value in our data set was over 28 kbp, within the range of acceptable quality for M. tuberculosis genome sequences (Periwal et al. 2015; Lahlou et al. 2017). The final data set consisted of genome sequences from 3,798 M. tuberculosis strains (supplementary table S1, Supplementary Material online) and the M. tuberculosis CDC1551 reference strain (accession number: NC_002755).

Virtual Spoligotyping

Spoligotyping has long been a “gold standard” in epidemiological studies of M. tuberculosis and is still widely used to infer genetic relatedness between strains. Virtual spoligotyping was performed with Geneious R11 (Biomatters Ltd., Auckland, New Zealand) (Kearse et al. 2012) using BLAST searches of genomic sequences for previously described oligonucleotides (Kamerbeek et al. 1997). Spacer sequences are unique in the genomes of mycobacteria and are not homologous to other M. tuberculosis DNA sequences (van Embden et al. 2000). A BLAST search approach to type M. tuberculosis has previously been implemented in SpoTyping (Xia et al. 2016). As in SpoTyping, a maximum of one mismatch in a 25 bp spacer was considered a hit. The accuracy of our approach was verified by comparing our results with previously published spoligotypes for laboratory strains included in our database (loeger et al. 2010; Mathema et al. 2012). Determination of genetic lineage and family based on spoligotypes was performed online using TB-Insight with the
TB-Lineage (Shabbeer et al. 2012) and SPOTCLUST (Vitol et al. 2006) tools, respectively. A phylogenetic tree based on unique spoligotype sequences was built with Geneious R11 using the UPGMA method (supplementary fig. S1, Supplementary Material online).

Retrieval of Cobalamin Synthesis-Related Gene Sequences

The genes related to vitamin B12 synthesis were chosen based on references in the literature (Gopinath, Moosa et al. 2013). The majority of these genes are presumably involved in aerobic vitamin B12 synthesis (fig. 1). Three genes, *nrdZ*, *metH* and *mutB*, encode vitamin B12-dependent enzymes. The *bacA* gene encodes a vitamin B12 transporter, whereas the remaining genes are predicted to play a role in vitamin B12 biosynthesis, although their exact involvement in the pathway remains to be determined. The retrieval of gene sequences was performed with Geneious R11. Genomic sequences of *M. tuberculosis* were transformed into a custom database. This database was searched using BLAST algorithm to query sequences of *M. tuberculosis* CDC1551 genes (accession number: NC_002755). An exception was the *metH* gene sequence because CDC1551 contains a truncated variant of this gene. A variant search for *metH* was performed using the *M. tuberculosis* H37Rv (accession number: NC_000962) *metH* gene as a query sequence. For each gene, we excluded all the sequences that did not cover the entire length of the query gene from further analysis. We estimated single-nucleotide polymorphism (SNP) variability within genes in comparison to our reference sequence using Geneious R11 (supplementary table S2, Supplementary Material online).

Tests of Selection

We used Geneious R11 to extract unique sequences devoid of insertions/deletions, ambiguous bases and protein truncations. Unique sequence polymorphisms were assessed as the total number of mutations, the average number of nucleotide differences, and the average nucleotide diversity using DnaSP v5 (Librado and Rozas 2009) (supplementary table S3, Supplementary Material online). Differences in the polymorphism estimates between different groups of genes were examined with a t-test in Statistica v12 (StatSoft, Tulsa, OK, USA).

![Fig. 1.—Schematic representation of the vitamin B12 synthesis pathway in *M. tuberculosis*.](image)
Genes were tested for the presence of codon-specific signatures of negative or positive selection using the DataMonkey online server for HyPhy package (Delport et al. 2010). Significant departures of codon-specific nucleotide substitution rates were tested with three different methods: single likelihood ancestor counting (SLAC), fixed effects likelihood (FEL), and fast unconstrained Bayesian approximation (FUBAR) (Kosakovsky Pond and Frost 2005; Delport et al. 2010; Murrell et al. 2013). SLAC and FEL use a combination of maximum-likelihood (ML) and counting approaches to infer nonsynonymous (dN) and synonymous (dS) substitution rates on a per-site basis for a given coding alignment and corresponding phylogeny, whereas FUBAR uses hierarchical Bayesian methods that implement a Markov chain Monte Carlo (MCMC) routine to ensure robustness against model misspecifications (Murrell et al. 2013). Under most scenarios, SLAC and FEL are conservative in terms of type I errors, and thus, they are likely to have lower actual rate of false positives than the significance level (Kosakovsky Pond and Frost 2005). In fact, simulation analyses have shown that the rate of false positives made by these methods is lower than the expected error rate within the range of 0 < \( P < 0.08 \) (Sorhannus and Kosakovsky Pond 2006); consequently, we used the \( P < 0.08 \) significance level to infer selection in all FEL and SLAC analyses. In FUBAR, the cutoff limit to infer selection was set to 0.9 posterior probability. All analyses were run with default settings, and suitable nucleotide substitution models were automatically selected for each gene with the DataMonkey web server (supplementary table S4, Supplementary Material online). Codon-specific estimates of dN/dS ratios for each gene were obtained with MEGA software (Tamura et al. 2013) All results are reported as the means ± SE unless otherwise stated.

**Results**

**Construction of the Database**

Using data deposited in the genome database of NCBI, we generated a custom genomic database consisting of 3,798 *M. tuberculosis* strains, 92.24% of which were isolated in 29 countries scattered across the globe. Information on the country of origin was unavailable for 7.42% of strains, whereas the remaining 0.34% of strains were classified as laboratory isolates (fig. 2 and supplementary table S1, Supplementary Material online).

The geographical distribution of the origin of the strains within our data set was compared with WHO data for worldwide TB prevalence in 2016 (fig. 3). In general, the geographical distribution of our data set matched the distribution of the global population of *M. tuberculosis*. However, our data set appeared to be overrepresented by African strains and underrepresented by European strains.

Most of the strains included in this study (65.3%) were isolated in three countries: Peru, Russia, and the Republic of South Africa. These strains are likely to include several clonal descendants of the common ancestor. Furthermore, according to information provided in the NCBI genome database, several strains included within this study were sequenced because they were extensively drug resistant, multidrug resistant, or caused a large outbreak in the country of origin. Therefore, the quantitative reference to different variants of *M. tuberculosis* genes presented in this study may not reflect the actual global distribution of the gene variants present within the entire population of this pathogen.

**Virtual Spoligotyping**

Spoligotyping was performed through BLAST homology searches of the genome sequences with oligonucleotide spacer sequences described for standard spoligotyping. Within our data set, we found 1,089 different spoligotypes. Spoligotyping patterns allowed determination of the genetic lineages and families of the *M. tuberculosis* strains. It needs to be emphasized that lineages determined by spoligotyping may not be completely concordant with lineages determined by the most recent method proposed by Comas and coworkers, which is based on SNP variability (Comas et al. 2013). Our data set represented all seven spoligotype-based lineages of *M. tuberculosis* (fig. 4). As expected, modern lineages constituted vast majority of our data set (Euro-American, East Asian, and East-African Indian) and were represented by 97.5% of strains (fig. 4). These results are consistent with the previously described limited distribution of ancient lineages (O’Reilly and Daborn 1995; Coscolla and Gagneux 2014; Yimer et al. 2015).

To further infer relatedness between the strains, we assessed the spoligotyping results with the international spoligotype database SpolDB3. According to the SpolDB3 database, 3,680 of the strains (96.9%) have previously described profiles, whereas the profiles of 118 strains (3.1%) are not yet listed (orphan profiles). The strains included in our study represented all large genomic clades. The predominant families in the data set are Beijing (28.9%), Latin-American-Mediterranean (LAM) (23.4%) and T clade (16.8%) (table 1).

**Retrieval of Gene Sequences and Gene Polymorphisms**

Gene sequences were retrieved from *M. tuberculosis* genome database using Geneious BLAST search for query sequences of *M. tuberculosis* CDC1551 (except for *metH* gene, which was searched against *M. tuberculosis* H37Rv). All sequence variants that did not cover the entire length of query sequences were excluded. We observed complete sequences of the analyzed genes in 95–100% (mean: 99.4% ± 0.002) of the strains and analyzed their variability (supplementary tables S2 and S3, Supplementary Material online). The high coverage of genes located in distant loci of mycobacterial genomes suggests the adequate quality of genomic sequences included in this study.
When compared with the query sequences, the variability of the sequences included in the database was low. Many gene mutations were present in only a single strain (0.03% of the analyzed population). The consensus sequences of *M. tuberculosis* genes involved in vitamin B12 metabolism are generally similar to sequences of *M. tuberculosis* CDC1551, although several modifications exist (table 2 and figs. 5 and 6).
An interesting observation was made regarding hemY (fig. 5). According to the Database of Prokaryotic Operons, hemY (Rv2850c) is in an operon with cobB (Rv2848c), cobO (Rv2849c), and a gene encoding a GCN5-related N-acetyltransferase (Rv2851c). The genes are in the reverse orientation on the M. tuberculosis chromosome. The start codon of hemY overlaps the stop codon of GCN5-related N-acetyltransferase in M. tuberculosis H37Rv, whereas an alternative hemY start codon located 57 bp upstream the initial codon in H37Rv is predicted to be used in M. tuberculosis CDC151. CDC1551 contains a two bp deletion just upstream of the alternative hemY start codon. Without the use of the alternative start codon, such a deletion would render the HemY protein nonfunctional. The variant containing the TT insertion (resembling H37Rv) is present in 82.2% of the analyzed strains. However, since M. tuberculosis CDC1551 was suspected to biosynthesize vitamin B12, we decided to use its sequence as our query.

Another gene that differed between CDC1551 and the majority of our population was metH. CDC1551 contains a truncated variant of this gene that was previously reported (Warner et al. 2007). We observed the truncated variant in 38 strains (1%) in our population. Since the metH gene of H37Rv was reported to be a functional variant of the protein (Warner et al. 2007; Guzzo et al. 2016), we decided to use it as our query sequence.

Next, we investigated the differences between M. tuberculosis CDC1551 and M. tuberculosis H37Rv. Apart from the differences already mentioned in hemY and metH, we observed several SNPs that differ between M. tuberculosis H37Rv and CDC1551 (table 2 and figs. 5 and 6). Mutations in nrdZ, Rv2067 and cobL G979C were observed to be the most prevalent variants in our population.

We also assessed the mutations in genes predicted through transposon mutagenesis to be essential for the survival of M. tuberculosis (table 3). Transposon mutagenesis studies have greatly improved our understanding of the gene requirements of M. tuberculosis during growth in broth (Sassetti et al. 2003; Griffin et al. 2011), in macrophages (Sassetti and Rubin 2003) and in animal models (Dutta et al. 2010). However, the conditions in which M. tuberculosis naturally persists and its gene requirements are expected to be somewhat different from experimental conditions. Furthermore, transposon mutants are tested over a relatively short period of time and fitness loss may be overlooked unless it is very apparent. Finally,

### Table 1

| Genomic Family | Frequency (%) of Genomic Families in Our Data Set |
|---------------|--------------------------------------------------|
| Beijing       | 28.9                                             |
| LAM           | 23.4                                             |
| T             | 16.8                                             |
| H37Rv         | 8.5                                              |
| Haarlem       | 7.3                                              |
| X             | 5.7                                              |
| S             | 2.3                                              |
| EAI           | 1.8                                              |
| M. microti    | 1.3                                              |
| 34            | 1                                                |
| 35            | 1                                                |
| 36            | 1                                                |
| CAS           | 0.7                                              |
| M. bovis      | 0.1                                              |
| 33            | 0.1                                              |
| M. afric anus | 0.1                                              |

**Note:** The description of the spoligotype is based on SpolDB3 database through TBInsight online server.

**Fig. 4.**—Distribution of the strains within our data set by lineage. Lineage distribution was determined through virtual spoligotyping and using the TBInsight online server.

**Table 2**  
Distribution of Genomic Clades in Strains of Our Data Set

| Genomic Family | Frequency (%) of Genomic Families in Our Data Set |
|---------------|--------------------------------------------------|
| Beijing       | 28.9                                             |
| LAM           | 23.4                                             |
| T             | 16.8                                             |
| H37Rv         | 8.5                                              |
| Haarlem       | 7.3                                              |
| X             | 5.7                                              |
| S             | 2.3                                              |
| EAI           | 1.8                                              |
| M. microti    | 1.3                                              |
| 34            | 1                                                |
| 35            | 1                                                |
| 36            | 1                                                |
| CAS           | 0.7                                              |
| M. bovis      | 0.1                                              |
| 33            | 0.1                                              |
| M. afric anus | 0.1                                              |
### Table 2
The Comparison of Gene Variants between *M. tuberculosis* H37Rv, *M. tuberculosis* CDC1551, and *M. tuberculosis* Population Included in Our Database

| Gene          | Codon Number | Amino Acid Variant | Variant Frequency (%) | Amino Acid Variant | Amino Acid Variant |
|---------------|--------------|--------------------|-----------------------|--------------------|--------------------|
| *nrdZ*        | 666          | SY                 | 100                   | SS                 | SY                 |
| *cobl*        | 327          | D                  | 100                   | H                  | D                  |
| *cbiX*        | 182          | V                  | 99.5                  | V                  | A                  |
| *cobl*        | 205          | P                  | 93.1                  | P                  | L                  |
| *cobD*        | 79           | C                  | 93                    | C                  | S                  |
| Rv2067 homolog| 288          | E                  | 82.1                  | K                  | E                  |
| *cobM*        | 145          | M                  | 33.9                  | I                  | I                  |
| *cobB*        | 440          | Frame shift        | 30.6                  | —                  | —                  |
| *cobN*        | 145          | H                  | 27.3                  | R                  | R                  |
| *cobN*        | 677          | K                  | 27.1                  | E                  | E                  |

**FIG. 5.**—An alignment of two alternative start codons for HemY in *M. tuberculosis* using the representative strains H37Rv (Accession Number: NC_000962) and CDC1551 (Accession Number: NC_002755).

**FIG. 6.**—The *metH* variants of *M. tuberculosis* H37Rv (Accession Number: NC_000962) and CDC1551 (Accession Number: NC_002755).
transposon mutants may not consider the existence of compensatory mutations. With the exception of cobQ2, we observed frameshift mutations and truncations in all of the essential genes of *M. tuberculosis*. The mutations are often represented by only one strain in the entire population, and ten were observed to be localized in tandem repeat regions. It cannot be excluded that at least some of these mutations are in fact sequencing errors. It is also possible that the strains carrying these mutations carry compensatory mutations in other parts of their genome. Finally, these genes may not be essential for *M. tuberculosis* survival in clinical settings.

For tests of selection, we excluded all the sequences containing insertions/deletions, ambiguous bases and truncated variants of the protein, leaving 98 to 100% (mean: 98.5% ± 0.01) of the initial gene sequences, with the exception of cobB sequences. Over 30% of the strains contained an insertion at nucleotide position 1318 of cobB (the size of the gene is 1,374 bp), resulting in a frameshift. Therefore, 69% of initial sequences were included in tests of selection (supplementary table S3, Supplementary Material online). The protein CobB is 457 aa, and the frame shift occurred at codon 440. It remains to be determined whether this mutation at such a distal part of the protein might disrupt protein structure and impede its function.

The number of alleles of each gene varied between 18 and 115 (mean 42.6 ± 4.2), but only five genes showed >60 alleles (table 4 and supplementary table S3, Supplementary Material online). Genes coding for the enzymes that require cobalamin as a cofactor had significantly higher allelic diversities than those involved in the cobalamin synthesis (76 ± 12.3 vs. 39.1 ± 4.1 alleles; t = 2.84, df = 28, P = 0.008), but this was likely due to their longer sequences (2,620 ± 482 vs. 1,109 ± 117 bp; t = 3.96, df = 28, P < 0.001). When accounting for the sequence length of genes, we did not observe significant differences in the number of polymorphic sites between genes involved in cobalamin synthesis and those encoding enzymes that require cobalamin (3.43 ± 0.14 vs. 3.05 ± 0.32 polymorphic sites per 100 bp; t = 0.86, df = 28, P = 0.40). The number of polymorphic sites for these genes ranged from 1.95 to 5.85 sites per 100 bp. The cobalamin transporter encoded by bacA had a lower number of polymorphic sites than all the other genes (1.71

### Table 3

| Gene            | Gene Size | Essentiality | Nucleotide Position | Change                      | Polymorphism Type         | Protein Effect | Variant Frequency (%) |
|-----------------|-----------|--------------|---------------------|-----------------------------|---------------------------|----------------|-----------------------|
| hemD            | 1,679     | (Griffin et al. 2011) | 148 | (G)2 → (G)3 Insertion (tandem repeat) | Frame Shift | 0.03 |
|                 |           |              | 190 | –G Deletion | Frame Shift | 0.03 |
|                 |           |              | 588 | (G)3 → (G)4 Insertion (tandem repeat) | Frame Shift | 0.03 |
|                 |           |              | 1,488 | –C Deletion | Frame Shift | 0.03 |
| cobL            | 1,173     | (Sassetti and Rubin 2003) | 376 | C → T SNP (transition) | Truncation | 0.03 |
|                 |           |              | 743 | –G Deletion | Frame Shift | 0.03 |
|                 |           |              | 1,103 | –A Deletion | Frame Shift | 0.03 |
|                 |           |              | 1,150 | C → T SNP (transition) | Truncation | 0.03 |
| Rv2228c homolog | 1,095     | (Griffin et al. 2011) | 552 | (G)4 → (G)5 Insertion (tandem repeat) | Frame Shift | 0.03 |
|                 |           |              | 859 | –G Deletion | Frame Shift | 0.03 |
| cobC            | 1,095     | (Sassetti and Rubin 2003) | 18 | (G)3 → (G)2 Deletion (tandem repeat) | Frame Shift | 0.1 |
|                 |           |              | 589 | (G)3 → (G)2 Deletion (tandem repeat) | Frame Shift | 0.03 |
|                 |           |              | 595 | (G)3 → (G)4 Insertion (tandem repeat) | Frame Shift | 0.03 |
|                 |           |              | 595 | (G)2 → (G)3 Insertion (tandem repeat) | Frame Shift | 0.03 |
|                 |           |              | 792 | (G)6 → (C)7 Insertion (tandem repeat) | Frame Shift | 0.03 |
|                 |           |              | 792 | (G)6 → (C)5 Deletion (tandem repeat) | Frame Shift | 0.3 |
| cysH            | 765       | (Griffin et al. 2011) | 106 | –A Deletion | Frame Shift | 0.03 |
|                 |           |              | 687 | (A)3 → (A)2 Deletion (tandem repeat) | Frame Shift | 0.03 |
| che1            | 846       | (Griffin et al. 2011) | 5 | –C Deletion | Frame Shift | 0.03 |
|                 |           |              | 57 | AGC → CT Deletion | Frame Shift | 0.1 |
|                 |           |              | 59 | CA → T Deletion | Frame Shift | 0.03 |
|                 |           |              | 60 | –A Deletion | Frame Shift | 0.3 |
|                 |           |              | 90 | –G Deletion | Frame Shift | 0.1 |
| cysG            | 1,218     | (Griffin et al. 2011) | 360 | –C Deletion | Frame Shift | 0.03 |
|                 |           |              | 432 | –G Deletion | Frame Shift | 0.03 |
|                 |           |              | 891 | –C Deletion | Frame Shift | 0.03 |
| cobQ2           | 696       | (Griffin et al. 2011) | not found | not found | not found | not found |

### Vitamin B12 Biosynthesis Pathway in *M. tuberculosis*

**GBE**

*Genome Biol. Evol.* 10(9):2326–2337 doi:10.1093/gbe/evy153 Advance Access publication July 30, 2018
polymorphic sites per 100 bp). The average number of nucleotide differences was $2.33 \pm 0.05$ (range: 1.88–3.20), whereas the average nucleotide diversity was $0.0023 \pm 0.0001$ (range: 0.0007–0.0044).

**Tests of Selection**

Unique sequences of genes involved in vitamin B12 synthesis and metabolism were tested for the presence of selective pressure using three codon-based methods included within the HyPhy package: SLAC, FEL and FUBAR. We observed signatures of selection acting on all genes associated with vitamin B12 (table 4 and supplementary tables S4 and S5, Supplementary Material online). In total, traces of selection (negative or positive) were observed for 14 codons using SLAC, 128 codons using FEL and 134 codons using FUBAR. Overall, we identified 190 codons under selection, with an average of $6.13 \pm 0.81$ codons per gene analyzed. Codons under selection constituted $16 \pm 0.1\%$ of identified polymorphic sites. All of the codons (100%) identified by SLAC were identified by at least one additional method. Other methods were less consistent, with 59% of codons confirmed for FEL and 54% for FUBAR (fig. 7).

We observed a signature of purifying selection acting on the large majority of genes associated with vitamin B12 in *M. tuberculosis*. With the exception of *cobO*, all genes had at least one codon under purifying selection, and the mean number of codons under purifying selection was $5.29 \pm 0.77$ per gene. In total, 162 codons were identified as being under purifying selection. The average proportion of polymorphic sites under purifying selection was $12.88 \pm 0.01\%$ per gene, with a maximum of $32\%$ of polymorphic sites being under purifying selection in *cobS*. The maximum number of codons under purifying selection ($n = 21$) was detected in *cobN*, but this was consistent with only $17\%$ of polymorphic sites present in this gene ($n = 122$).

| Gene Name | H37Rv Homolog | Total Number of Sites | No. of Polymorphic Sites Per 100 bp | dN/dS | Total No. of Codons Under Selection | No. of Codons Under Purifying Selection |
|-----------|---------------|----------------------|-------------------------------------|-------|-----------------------------------|--------------------------------------|
| cobU      | Rv0254c       | 525                  | 3.05                                | 2.22  | 4                                 | 3                                    |
| cobO      | Rv2849c       | 624                  | 2.72                                | 2.32  | 2                                 | 0                                    |
| bacA      | Rv1819c       | 1,920                | 1.72                                | 1.83  | 6                                 | 4                                    |
| bluB      | Rv0306        | 672                  | 3.13                                | 1.27  | 1                                 | 1                                    |
| pduO      | Rv1314c       | 582                  | 3.61                                | 1.20  | 3                                 | 2                                    |
| cobT      | Rv2207        | 1,086                | 3.04                                | 1.07  | 5                                 | 5                                    |
| cobI      | Rv2066        | 1,527                | 3.73                                | 1.00  | 12                                | 12                                   |
| hemY      | Rv2850c       | 1,833                | 4.31                                | 0.97  | 12                                | 9                                    |
| cobD      | Rv2236c       | 942                  | 3.29                                | 0.94  | 3                                 | 2                                    |
| cobN      | Rv2062c       | 3,585                | 3.40                                | 0.93  | 23                                | 21                                   |
| cobS      | Rv2208        | 750                  | 4.13                                | 0.91  | 12                                | 10                                   |
| metZ      | Rv0391        | 1,221                | 2.54                                | 0.90  | 8                                 | 8                                    |
| cobL      | Rv2072c       | 1,173                | 4.18                                | 0.87  | 5                                 | 5                                    |
| metH      | Rv2124c       | 3,579                | 2.93                                | 0.84  | 14                                | 13                                   |
| hemD      | Rv0511        | 1,698                | 3.47                                | 0.84  | 5                                 | 4                                    |
| nrdZ      | Rv0570        | 2,054                | 3.65                                | 0.76  | 11                                | 10                                   |
| cobK      | Rv2070c       | 735                  | 5.85                                | 0.72  | 5                                 | 3                                    |
| cobB      | Rv2848c       | 1,374                | 1.97                                | 0.71  | 7                                 | 6                                    |
| che1      | Rv2393        | 846                  | 3.90                                | 0.71  | 2                                 | 2                                    |
| cysG      | Rv2847c       | 1,218                | 2.55                                | 0.66  | 4                                 | 3                                    |
| cobC      | Rv2231c       | 1,095                | 3.56                                | 0.63  | 3                                 | 3                                    |
| cobQ1     | Rv0255c       | 1,485                | 3.43                                | 0.62  | 4                                 | 4                                    |
| cobM      | Rv2071c       | 756                  | 3.44                                | 0.61  | 7                                 | 7                                    |
| —         | Rv2067c       | 1,224                | 3.02                                | 0.61  | 7                                 | 6                                    |
| cobG      | Rv2064        | 1,092                | 2.93                                | 0.54  | 4                                 | 3                                    |
| cboX      | Rv0259c       | 744                  | 4.03                                | 0.49  | 2                                 | 2                                    |
| mutB      | Rv1493        | 2,253                | 2.57                                | 0.47  | 8                                 | 6                                    |
| cysH      | Rv2392        | 765                  | 3.14                                | 0.44  | 5                                 | 5                                    |
| cobH      | Rv2065        | 627                  | 3.83                                | 0.43  | 3                                 | 3                                    |
| —         | Rv2228c       | 1,095                | 3.74                                | 0.42  | 13                                | 13                                   |
| cobQ2     | Rv3713        | 696                  | 2.73                                | 0.35  | 3                                 | 3                                    |
Vitamin B12 Biosynthesis Pathway in *M. tuberculosis*

Fig. 7.—Overlap of the codons identified under selection by SLAC, FEL, and FUBAR. Numbers in brackets represent the number of codons identified by each method, whereas numbers without brackets represent the number of codons identified by overlapping methods.

Although purifying selection was the predominant evolutionary force acting on the vitamin B12 pathway genes, we also identified some signatures of positive selection in our data set. In total, 26 codons from 17 genes were identified as being under positive selection. The maximum number of codons under positive selection (n = 4) was observed in *hemY*, whereas only one or two codons showed a signature of positive selection in the remaining 16 genes. All these codons were identified exclusively with FUBAR, and none of them were confirmed with FEL or SLAC.

The median dN/dS ratios across all codons in all genes was 0.76 (range: 0.35–2.32). Consistent with the assumption of purifying selection, 24 genes had a dN/dS of < 1, whereas four genes had a dN/dS that was slightly higher than one (1 < dN/dS < 1.27). Three genes (*bacA*, *cobO*, and *cobU*) showed high overall dN/dS ratios (1.83–2.32), indicating that they might be predominantly under positive selection. The lowest dN/dS ratio was observed for *cobQ2* (dN/dS = 0.35), a *Rv2228c* homologue (dN/dS = 0.42), *cobH* (dN/dS = 0.43), and *cysH* (dN/dS = 0.44). Of note, three of these genes, *cobQ2*, *cysH* and a *Rv2228c* homologue, were determined to be essential for *M. tuberculosis* growth in a transposon mutant library screen (Griffin et al. 2011). The median dN/dS for genes involved in vitamin B12 synthesis was 0.72 (range: 0.35–2.32). Similarly, the median dN/dS for genes encoding proteins that require vitamin B12 as a cofactor was 0.76 (range: 0.47–0.84).

**Discussion**

Various indirect experiments have shown that the model strain *M. tuberculosis* H37Rv does not synthesize vitamin B12, suggesting possible redundancy of cobalamin biosynthesis genes in this organism. This hypothesis is supported by the presence of an encoded cobalamin transporter in the genome of *M. tuberculosis*, which suggests that instead of synthesizing the vitamin, this bacterium is able to scavenge cobalamin from the host (Domenech et al. 2009; Lawrence et al. 2018). In this study, we investigated selective pressures acting on genes involved in vitamin B12 metabolism in a global population of *M. tuberculosis*. Signatures of purifying selection were observed in the large majority of genes involved in cobalamin biosynthesis and transport as well as in those encoding enzymes that require cobalamin as a cofactor. Codon-specific nucleotide substitution rates indicated that purifying selection acted on up to or over 30% of polymorphic sites recognized within particular genes and that nearly 80% of all genes had an overall rate of nonsynonymous substitutions that was lower than the rate of synonymous substitutions (dN/dS < 1), consistent with purifying selection.

Genome-wide analyses of selection in *M. tuberculosis* indicated that purifying selection is a major evolutionary force acting on this pathogen (Namouchi et al. 2012), and the overall dN/dS ratio across the entire *M. tuberculosis* genome was estimated at ~0.6 (Liu et al. 2014). However, purifying selection in *M. tuberculosis* was suggested to be relaxed when compared with that in free-living bacteria (Hershberg et al. 2008). For example, an analysis of 63 globally extant *M. tuberculosis* genomes showed that the signature of purifying selection was several orders of magnitude weaker than recent estimates in other eukaryotic and prokaryotic organisms (Pepperell et al. 2013). However, it was suggested that this pattern could be possibly attributed to the relatively low effective population size of *M. tuberculosis* rather than to the low strength of selection against deleterious mutations associated with a pathogenic lifestyle (Pepperell et al. 2013). The hallmark of relaxed purifying selection could also be a consequence of the recent evolutionary age of *M. tuberculosis*, as purifying selection may not have had enough time to remove all the nonsynonymous mutations that are only slightly deleterious (Rocha et al. 2006; Stucki and Gagneux 2013). The strongest signature of purifying selection in *M. tuberculosis* was observed for genes that are essential for establishing infection in an animal model (dN/dS = 0.33), genes encoding ribosomal proteins (dN/dS = 0.30), and genes involved in the transport and metabolism of inorganic ions (dN/dS = 0.33) (Pepperell et al. 2013). Genes that are essential for in vitro growth showed weaker signature of purifying selection (dN/dS = 0.50), indicating a high specificity of constraints imposed on *M. tuberculosis* by the natural environment. The highest level of nonsynonymous substitutions was observed for genes related with defence mechanisms (dN/dS = 1) (Pepperell et al. 2013).
A study by Comas et al. that investigated dN/dS ratios across 22 genomes observed that dN/dS values for essential and nonessential genes ranged from 0.45 to 0.67 and 0.56 to 0.78, respectively (Comas et al. 2010). In comparison, we observed that cobalamin-related genes showed an overall dN/dS value of 0.76, suggesting that genes involved in cobalamin metabolism may not be essential for pathogen survival. This finding is supported by the detection of various nonfunctional variants of analyzed genes. However, following the assumption of purifying selection (dN/dS < 1), our results also provide support for the functionality of the vitamin B12 biosynthesis pathway in this organism and for the adaptive function of vitamin B12 for M. tuberculosis.

Although purifying selection has had a predominant role in shaping molecular evolution of genes associated with vitamin B12 in M. tuberculosis, we also identified several genes that showed signature of positive selection. Only three genes showed a strong excess of nonsynonymous versus synonymous mutations (dN/dS > 1.8). Although two of these genes (cobO and cobU) are directly involved in cobalamin biosynthesis, the third gene (bacA) encodes a cobalamin transporter that may confer resistance to bleomycin (Domenech et al. 2009). Genes that harbour mutations associated with a selective advantage to drug-resistant strains often show an excess of nonsynonymous nucleotide substitutions, and genome-wide searches for the signatures of positive selection have been performed to identify potential genes responsible for drug resistance in M. tuberculosis (Farhat et al. 2013). Other genomic regions of M. tuberculosis that are likely to be under positive selection include genes involved in cell wall biosynthesis, transcriptional regulation, and DNA repair (Farhat et al. 2013), as well as pe-pgrs genes (Copin et al. 2014) and genes encoding membrane proteins (Osório et al. 2013). Mutations in these regions have been suggested to possibly compensate for fitness costs associated with drug resistance (Farhat et al. 2013). However, an elevated dN/dS ratio may also be indicative of selective sweeps, that is, the selection for advantageous mutations under a regime of restricted migration (Pepperell et al. 2013). Although the exact molecular mechanism responsible for an elevated rate of nonsynonymous mutations in cobO and cobU genes remains unresolved, this signature of positive selection seems to provide further support for the general functionality of the vitamin B12 biosynthesis pathway in M. tuberculosis.

In conclusion, our results suggest that cobalamin-related genes are not essential but are adaptive for M. tuberculosis in clinical settings. Furthermore, a cobalamin biosynthesis pathway is likely to be functional in this species.

**Supplementary Material**

Supplementary data are available at Genome Biology and Evolution online.
Vitamin B12 Biosynthesis Pathway in M. tuberculosis

Lawrence AD, et al. 2018. Construction of fluorescent analogs to follow the uptake and distribution of cobalamin (vitamin B12) in bacteria, worms, and plants. Cell Chem Biol. doi: 10.1016/j.chembiol.2018.04.012.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25(11):1451–1452.

Liu F, et al. 2014. Comparative genomic analysis of Mycobacterium tuberculosis clinical isolates. BMC Genomics 15(1):469.

Mathema B, et al. 2012. Epidemiologic consequences of microvariation in Mycobacterium tuberculosis. J Infect Dis. 205(6):964–974.

Murrell B, et al. 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol Biol Evol. 30(5):1196–1205.

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

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.

O’Reilly LM, Daborn CJ. 1995. The epidemiology of Mycobacterium bovis infections in animals and man: a review. Tuberc Lung Dis. 76(Suppl 1):1–46.

World Health Organization. 2017. Annual TB Report 2017.

Osorio NS, et al. 2013. Evidence for diversifying selection in a set of Mycobacterium tuberculosis genes in response to antibiotic- and nonantibiotic-related pressure. Mol Biol Evol. 30(6):1326–1336.

Pepperell CS, et al. 2013. The role of selection in shaping diversity of natural M. tuberculosis populations. PLoS Pathog. 9(8):e1003543.

Perival V, et al. 2015. Comparative whole-genome analysis of clinical isolates reveals characteristic architecture of Mycobacterium tuberculosis pangenome. PLoS One 10(4):e0122979.

Piddington DL, Kashkouli A, Buchmeier NA. 2000. Growth of Mycobacterium tuberculosis in a defined medium is very restricted by acid pH and Mg2+ levels. Infect Immun. 68(8):4518–4522.

Rocha EPC, et al. 2006. Comparisons of dNdS values are time dependent for closely related bacterial genomes. J Theor Biol. 239(2):226–235.

Sassetti CM, Boyd DH, Rubin EJ. 2003. Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol. 48(1):77–84.

Sassetti CM, Rubin EJ. 2003. Genetic requirements for mycobacterial survival during infection. Proc Natl Acad Sci USA. 100(22):12989–12994.

Savvi S, et al. 2008. Functional characterization of a vitamin B12-dependent methylmalonyl pathway in Mycobacterium tuberculosis: implications for propionate metabolism during growth on fatty acids. J. Bacteriol. 190(11):3886–3895.

Shabbeer A, et al. 2012. TB-Lineage: an online tool for classification and analysis of strains of Mycobacterium tuberculosis complex. Infect Genet Evol. 12(4):789–797.

Shastry BS. 2009. SNPs: impact on gene function and phenotype. Methods Mol. Biol. 578:3–22.

Sorhannus U, Kosakovky Pond SL. 2006. Evidence for positive selection on a sexual reproduction gene in the diatom genus Thalassiosira (Bacillariophyta). J Mol Evol. 63(2):231–239.

Stucki D, Gagneux S. 2013. Single nucleotide polymorphisms in Mycobacterium tuberculosis and the need for a curated database. Tuberculosis 93(1):30–39.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 30(12):2725–2729.

Vitol I, Driscoll J, Kreswirth B, Kurepina N, Bennett KP. 2006. Identifying Mycobacterium tuberculosis complex strain families using spoligotypes. Infect. Genet. Evol. 6(6):491–504.

Warner DF, Savvi S, Mizrahi V, Dawes SS. 2007. A riboswitch regulates expression of the coenzyme B12-independent methionine synthase in Mycobacterium tuberculosis: implications for differential methionine synthase function in strains H37Rv and CDC1551. J Bacteriol. 189(9):3655–3659.

Xia E, Tea Y-Y, Ong RT-H. 2016. SpoTyping: fast and accurate in silico Mycobacterium spoligotyping from sequence reads. Genome Med. 8(1):19.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155(1):431–449.

Yirmer SA, et al. 2015. Mycobacterium tuberculosis lineage 7 strains are associated with prolonged patient delay in seeking treatment for pulmonary tuberculosis in Amhara Region, Ethiopia. J Clin Microbiol. 53(4):1301–1309.

Young DB, Comas I, de Carvalho LPS. 2015. Phylogenetic analysis of vitamin B12-related metabolism in Mycobacterium tuberculosis. Struct Biol. 2:6.

Associate editor: Purificación López-García