Patterns and Processes of *Mycobacterium bovis* Evolution Revealed by Phylogenomic Analyses

José S.L. Patané[1], Joaquim Martins Jr[1], Ana Beatriz Castelão[2], Christiane Nishibe[3], Luciana Montera[3], Fabiana Bigi[4], Martin J. Zumárraga[4], Angel A. Cataldi[4], Antônio Fonseca Junior[5], Eliana Roxo[6], Ana Luiza A.R. Osório[7], Klaudia S. Jorge[7], Tyler C. Thacker[8], Nalvo F. Almeida[3], Flabio R. Araújo[9], and João C. Setubal[1,10,*]

[1]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, SP, Brazil

[2]Universidade Federal de Mato Grosso do Sul, Campo Grande, MS, Brazil

[3]Faculdade de Computação Universidade Federal de Mato Grosso do Sul, Campo Grande, MS, Brazil

[4]Instituto Nacional de Tecnologia Agropecuária, Córdoba, Argentina

[5]Rede de Laboratórios Agropecuários do Ministério da Agricultura, Pecuária e Abastecimento, Pedro Leopoldo, MG, Brazil

[6]Instituto Biológico de São Paulo, IB-USP, São Paulo, SP, Brazil

[7]Programa em Ciência Animal Universidade Federal de Mato Grosso do Sul, Campo Grande, MS, Brazil

[8]Agricultural Research Service, United States Department of Agriculture, Ames, Iowa

[9]EMBRAPA, Campo Grande, MS, Brazil

[10]Biocomplexity Institute of Virginia Tech, Blacksburg, Virginia

*Corresponding author: E-mail: setubal@iq.usp.br.

Accepted: February 8, 2017

Abstract

Mycobacterium bovis is an important animal pathogen worldwide that parasitizes wild and domesticated vertebrate livestock as well as humans. A comparison of the five *M. bovis* complete genomes from the United Kingdom, South Korea, Brazil, and the United States revealed four novel large-scale structural variations of at least 2,000 bp. A comparative phylogenomic study including 2,483 core genes of 38 taxa from eight countries showed conflicting phylogenetic signal among sites. By minimizing this effect, we obtained a tree that better agrees with sampling locality. Results supported a relatively basal position of African strains (all isolated from *Homo sapiens*), confirming that Africa was an important region for early diversification and that humans were one of the earliest hosts. Selection analyses revealed that functional categories such as "Lipid transport and metabolism," "Cell cycle control, cell division, chromosome partitioning" and "Cell motility" were significant for the evolution of the group, besides other categories previously described, showing importance of genes associated with virulence and cholesterol metabolism in the evolution of *M. bovis*. PE/PPE genes, many of which are known to be associated with virulence, were major targets for large-scale polymorphisms, homologous recombination, and positive selection, evincing for the first time a plethora of evolutionary forces possibly contributing to differential adaptability in *M. bovis*. By assuming different priors, US strains originated and started to diversify around 150–5,210 ya. By further analyzing the largest set of US genomes to date (76 in total), obtained from 14 host species, we detected that hosts were not clustered in clades (except for a few cases), with some faster-evolving strains being detected, suggesting fast and ongoing reinfections across host species, and therefore, the possibility of new bovine tuberculosis outbreaks.

Key words: *Mycobacterium bovis*, bovine tuberculosis, phylogenomics, PE/PPE family, epidemiology.

Introduction

Bovine tuberculosis is a significant worldwide disease of cattle (Thoen, Steele, et al. 2006). *Mycobacterium bovis*, the causative agent, is also pathogenic for humans and several domestic and wild animals (Thoen, LoBue, et al. 2006). *Mycobacterium bovis* belongs to the *Mycobacterium tuberculosis* complex (MTbC), which consists traditionally of *M. tuberculosis*, the main agent of human tuberculosis, plus lineages typical of other vertebrate hosts, *M. africanum*, *M. canettii*, *M. microti*, *M. bovis*, *M. caprae*, *M. pinnipedii*, *M. mungi*, *M. orygis*, and *M. suricattae* (Fabre et al. 2004; Smith et al. 2006; Alexander et al. 2010; Van Ingen et al. 2012; Parsons et al. 2013). All of these species are genomically very similar, with more than 99.9% nucleotide identity (Thoen, LoBue, et al. 2006). It has been hypothesized that MTbC members have evolved from a common ancestor (possibly *M. canettii*) via successive DNA deletions/insertions resulting in the present *Mycobacterium* speciation and associated differences in pathogenicity (Fabre et al. 2004; Gutierrez et al. 2005).

Disease control programmes based on regular tuberculin testing, and removal of infected animals ("test-and-slaughter") have been successful in eradicating, or markedly reducing, tuberculosis (TB) from cattle in many countries, but these measures are less effective in countries with wildlife reservoirs of *M. bovis* (Cousins 2001; Miller and Sweeney 2013). The gold standard for postmortem diagnosis of bovine TB has been mycobacterial culture, although this method requires a significant amount of time (up to 90 days) (Lisle et al. 2008). This fact, together with economic globalization and the pressure from importing markets for a definitive diagnosis of tuberculosis in cattle that exhibit lesions compatible with tuberculosis and the advances in molecular biology, have provided a stimulus for improved TB molecular diagnostic techniques (Thacker et al. 2011; Araújo et al. 2014).

In order to examine the evolution of *M. bovis* infection in space and time, it is essential to study genetic differentiation of isolates and their phylogenetic relationships, together with cattle movement and wildlife ecology. A clearer understanding of *M. bovis* epidemiology may contribute to better control or to eradication of disease. High-quality epidemiological data increases the efficiency of bovine tuberculosis control programs. With new tools available in molecular epidemiology, there is an increasing possibility of finding answers to issues such as the importance of transmission between individual animals and its risk factors, and the role of wild animals as reservoirs (Hilty et al. 2005).

The use of whole genome sequencing (WGS) technology to study infectious bacterial diseases has resulted in unprecedented advances in the ability to resolve epidemiologic data at the global scale (Harris et al. 2010). WGS has provided new insights into within-host replication processes (Ford et al. 2011), and has been used to corroborate the importance of exhaustively identified transmission chains and social drivers of transmission (Gardy et al. 2011). WGS of many strains of the same species can provide the necessary statistical power required to find association between genotypes and phenotypes such as virulence, pathogenesis, host range, and antigenic variability.

In order to understand evolutionary processes of bovine tuberculosis, we undertook a large phylogenomic comparative study, addressing aspects of whole genome evolution of *M. bovis*. Our study includes 103 genomes (supplementary tables S1 and S2, Supplementary Material online, for details), isolated from 14 host species (cattle, pig, sheep, human, bobcat, deer, raccoon, opossum, coyote, jaguar, red deer/elk cross, fallow deer, elk, red deer) in eight countries (Brazil, Argentina, Uruguay, the United States, the United Kingdom, Korea, Mali, and Uganda), and obtained within a 22-year interval (1991–2013). Our goal was to obtain a thorough understanding of genomic characteristics, large-scale gains/losses/duplications, phylogenetic relationships, evolutionary timeframe, positive selection in core-coding genes, and homologous recombination rates. We also highlight important aspects that may be associated with evolution of lineages from the United States, such as patterns of reinfection among host species of *M. bovis*.

We further assess the evolutionary contribution of a specific large gene family typical of MTbC lineages, the PE/PPE family, known to be associated with evolution of these lineages in general, but whose contribution under different evolutionary forces using WGS had still not been contemplated in the particular case of *M. bovis* lineages. The PE/PPE family of genes is large, being distributed across the genome of MTbC, reaching ~170 copies per genome in *M. tuberculosis* H37Rv (Cole et al. 1998). It is known that many genes within it are associated with antigenic variation and immune evasion, being secreted by the type VII (or ESX) secretion system typical of mycobacteria and some nonmycobacterial actinomycetes, being composed of five copies of the ESX gene cluster, ESX-1 through ESX-5, the former being the most extensively studied (Abdallah et al. 2006; Bottai et al. 2012; Sayes et al. 2012; Gröschel et al. 2016). PE genes have the Pro-Glu motif at positions 8–9, with a conserved stretch of ~110 amino acids in the N-terminal domain. PPEs embrace a Pro-Pro-Glu sequence at positions 7–9 with conservation of ~180 AAs in the N-terminal domain. The C-terminal regions of both families are highly variable, also containing repetitive DNA motifs that may vary in copy number among family members (Cole et al. 1998). PE and PPE can be further subdivided by characteristics of their N-terminal regions and phylogenetic relationships (van Pittius et al. 2006; McEvoy et al. 2012). PE genes are composed of five subfamilies, the largest being the PE-PGRS (polymorphic GC-rich-repetitive sequence) embracing 65 members in H37Rv, with multiple tandem repeats of Gly-Gly-Ala or Gly-Gly-Asn as typical motifs in the C-terminal region (McEvoy et al. 2012). PPE genes are also subdivided into five subfamilies, of which the largest sets are of PPE-SVP

(for a total of 24 members in H37Rv) and the PPE-MPTR (major polymorphic tandem repeat) subfamilies (with 23 members).

Materials and Methods

Chromosomal Variations

Sibelia (Minkin et al. 2013) was used to assess chromosomal structural variations across the five *M. bovis* complete genomes (excluding BCG genomes, as these do not represent natural genomic variation) present in NCBI's RefSeq database (Pruitt et al. 2007), with comparisons based on raw fasta files to avoid annotation biases. The minimum threshold for detecting large sequence polymorphisms (LSPs) was 2,000 bp segments (this value was chosen to be high enough to uncover several interesting LSPs, but not too low to yield too many less informative polymorphisms).

Data Sets for Phylogenomics

We used two different data sets. Most of the evolutionary analyses were done with a 38-genome data set, which included a total of eight countries (supplementary table S1, Supplementary Material online, for details); unless stated otherwise, our analyses refer to this 38-genome data set. We used *M. tuberculosis* H37Rv (a canonical strain from this species) as outgroup. We also present results from a second data set (supplementary table S2, Supplementary Material online, for details), which had a narrower aim of gaining insights about *M. bovis* evolution within the United States only. This was made possible by the availability of 76 US *M. bovis* genomes from the PATRIC database (Wattam et al. 2014).

Regarding the 38-data set, we obtained all non-US genomes from GenBank (supplementary table S1, Supplementary Material online, for details). The 11 US genomes within this data set were assembled by the vcf-consensus tool within VCFtools (Danecek et al. 2011) using *M. bovis* AF2122/97 (NC_002945.3) as reference, and annotated with the Genome Reverse Compiler (Warren and Setubal 2009); these genomes can be found at <http://pintado.facom.ufms.br/mbovispub/; last accessed October 3, 2016>. There is an overlap of those 11 genomes between the 38-genome data set and the 76-genome data set. In the case of the 76-genome data set, we obtained assemblies and annotations for these 11 genomes from PATRIC (Wattam et al. 2014), as was the case for all genomes in this data set. We chose not to use our own assemblies and annotations in the analysis of the 76-genome data set in order to keep methods for this data set uniform. In order to check whether differences in assembly/annotation pipelines would introduce systematic biases in the analyses for the 38-data set, we performed *t*-tests of US versus non-US genomes in both the 38 and 76-data set (the latter set with annotations performed by the same tool, RAST, therefore with no annotation bias among genomes) considering different genomic parameters (core-coding size in base pairs, GC-content, and different codon bias measures), with significant differences ($P \leq 0.05$) between US and non-US genomes being found for all parameters in both data sets (the *t*-tests for the 38-data set specifically are further detailed in Material & Methods, and Results). Therefore, we conclude that the 38-data set was not influenced by heterogeneity in assembly/annotation.

Core- and Pan-Genome Curves

To assess whether the set of genomes used here is representative of the total genetic pool of *M. bovis*, we estimated pan- and core-genome curves (Tettelin et al. 2005). This was done for the 38-data set using the get_homologues.pl script (Contreras-Moreira and Vinuesa 2013) with the "–c" option and 20 resampling steps. At each resampling, genomes were included in an order different from the previous resampling step. Exponential regression lines were then fitted to each curve by the same script.

Inference of Orthologs

Inference of the families of homologs was done in the same way for both data sets by Get_Homologues v2.0 (Contreras-Moreira and Vinuesa 2013), using the get_homologues.pl script with "-M" flag, which activates OrthoMCL clustering (Li et al. 2003) with default values. Subsequently, strictly orthologous, unicopy core-genome genes (the *core-coding data set*) were estimated by the same software with the script compare_clusters.pl.

Multiple Alignment

Individual core-coding DNA gene alignments were obtained in Guidance v2.0 (Sela et al. 2015) anchored by codons using MAFFT v7.0 (Katoh and Standley 2013) with 100 pseudoreplicate guide trees. Concatenation of the multiple alignment was done by FasConCat (Kück and Meusemann 2010). Manual curation was then performed in Aliview (Larsson 2014), and whenever local regions were misaligned we used the software "realign selected block" option and updated the information about each gene start and end. Amount of variable and parsimony-informative sites was obtained in Mega 7.0 (Kumar et al. 2016). The fraction of nonsynonymous single nucleotide polymorphisms (nsSNPs), synonymous SNPs (sSNPs), and a reduced matrix containing only 4-fold degenerate sites were also obtained in Mega.

Genomic Characteristics of Core-Coding Data Set

Dambe v5.5.29 (Xia 2013) was used to look for any patterns of variation across the 38-genome set. For each genome, we obtained values for the effective number of codons (Nc), relative synonymous codon usage (RSCU), codon adaptation index (CAI), GC-content, and total size of core-coding alignment. RSCU (which is an individual measure of the variability

of use of each set of codons of an amino acid) per genome was obtained by summing the individual amino acid variances and then extracting the square-root of the result.

Phylogenomic Inference

Phylogenetic analysis was done by maximum likelihood (ML) within IQTree v1.3.12 (Nguyen et al. 2015), with support for each phylogeny calculated from 1,000 UFBoot pseudoreplicates (Minh et al. 2013). An important thing to consider when estimating character-based phylogenies (such as ML-based) is the amount of variation among data partitions, otherwise the resulting topology may be biased (Brown and Lemmon 2007). To test for best partitioning scheme for core-coding genes, we entered the three codon positions as data blocks (hence each block encompasses one-third of the alignment), and then estimated the best partitioning scheme among them and respective model for each partition under IQTree, using BIC as the criterion for data fit. To test for variability of rates along the genome, besides using the typical gamma distribution (+G) and a probability of invariable sites (+I) for each proposed partition, we also tested models with mixture distributions for each site, from two to five such mixtures per model (+MM), which in many cases have better fit (Venditti et al. 2008).

Further phylogenetic analyses aimed to test for (and possibly alleviate) systematic biases. As a proxy for phylogenetic reliability of each tree, the retention index (RI; Farris 1989) was used to assess the fit of the character to "sampling locality" (continents used as tip states) across the 38-genome data set, following previous conclusions showing that regionally *M. bovis* samples tend to cluster together (e.g., Hang'ombe et al. 2012; Allen et al. 2013; Hauer et al. 2015). We ascribed one among five states to each tip according to its respective continental region (South America, North America, Europe, Asia, and Africa). Subsequently, RI was calculated for each of the trees, with higher values indicating better agreement of that tree with geographic location. PAUP v4b10 (Swofford 2003) was used to calculate RI across the different topologies, using the command "describetrees".

Maximum parsimony (MP) in Mega v7.0 was used to test for topological differences due to the method employed, with support measured by 1,000 standard bootstrap pseudoreplicates (Felsenstein 1985). Besides testing for phylogenetic method, we aimed to correct for possible data artifacts that could lead to topological biases.

To test if selection, recombination, or the use of nonneutral sites were impacting the RI, reduced data sets from core-coding genes were obtained. To account for positive selection, either all genes under positive selection (see below) were removed, or were maintained but using a codon-based ML analysis to estimate a nonsynonymous to synonymous rate ratio (dN/dS) across a given topology.

In order to account for homologous recombination, genomic blocks shared among all taxa were used when considering the whole core genome (therefore also potentially spanning noncoding genes and nongenic regions) and excluding those blocks with even mild signs of recombination (see below).

To limit the analysis to an enriched set of hypothesized nearly-neutral codons (therefore maximizing true vertical signal), a reduced data set including only codons with 4-fold degenerate positions were obtained (in MEGA).

Aiming at minimizing biases due to conflicting phylogenetic signal among sites (regardless of the source of conflict), Tiger v1.0.2 (Cummins and McInerney 2011) was used, which allocates sites sharing similar relative rates within bins (we used ten bins as default). The largest bin set was used as input for a phylogenetic estimate (termed site-congruence method from here on). Models including ascertainment bias correction (+ASC) were also included for the latter phylogenomic analysis, as it lacked invariable sites (Lewis 2001).

Finally, a separate analysis independent of multiple alignment (or alignment-free) was performed by kSNP 3.0 (Gardner et al. 2015). Its algorithm works by finding core and noncore SNPs that are present across a subset, or the whole of the taxon set, therefore being very inclusive for assessment of tree topology. This SNP data set was run by ML under IQTree with the same options as described previously.

For the 76-genome data set phylogenomic analysis, we included only the variable characters, performing model-selection, tree search, and assessing confidence as explained above.

Assembly Quality

After performing selection and recombination analyses, we checked if signals could be the outcome of low quality genome assemblies by remapping paired-end reads of the genome with largest number of contigs (Uganda B2-7505, with 243 contigs) to representative genes indicated to be under these forces in genome AF2122/97. We used Trimmomatic (to filter high-quality reads, dismissing reads that were not paired-ends; Bolger et al. 2014), Bowtie 2 (mapping; Langmead and Salzberg 2012), Samtools (to order and index .bam files; Li et al. 2009), and VCFTools (to call and summarize SNVs; Danecek et al. 2011).

Detecting Homologous Recombination

The ten largest locally colinear blocks (LCBs) across the 38 genomes were obtained by the ProgressiveMauve algorithm (Darling et al. 2010). The three tests available in PHIPACK (PHI, NSS, and MaxChi) were used to detect recombination in each of these blocks (Bruen and Bruen 2005). To be conservative, an LCB was considered to be free of detectable recombination if none of three tests were significant under $P = 0.1$. Blocks with undetected recombination were concatenated and estimated by ML in IQTree as described above. No recombination

tests were performed gene-wise due to the scant variability among taxa in each gene alignment (which share >99% identity on average).

As another way to assess the effect of recombination, ClonalFrameML (Didelot and Wilson 2015) was run by concatenating all LCBs irrespective of recombination signs (the tree with best RI in the phylogenetic estimation phase was fixed). ClonalFrameML estimates relative contributions between homologous recombination and point mutations across the LCBs and for every branch of the tree.

Positive Selection

The core-coding data set was used to test for positive selection. Due to low sequence variability among *M. bovis* genomes, we had to estimate positive selection from the concatenation of core-coding genes in order to bring more power to the test. The FEL (fixed effects likelihood) method in HyPhy was able to operate on the whole concatenated core-coding data set. FEL estimates site-wise ratios of dN/dS using the entire data for inference of global alignment parameters (such as transition/transversion rates, nucleotide frequencies, etc.). The 38-genome data set used is larger than the minimum of 20 lineages recommended by the method developers (Kosakovsky Pond et al. 2009), making the test appropriate for this data set. Three different rate classes for dN and another three for dS were assumed. The best RI phylogeny was constant throughout the analysis.

Clusters of orthologous groups of proteins (COGs; Galperin et al. 2014) were inferred using the script CDD2COG.pl v0.1 (available from https://github.com/aleimba/bac-genomics-scripts/tree/master/cdd2cog; last accessed October 3, 2016). To detect which COG categories were significantly enriched in positively selected genes, the relative ratio of positively selected genes within a COG to the number of genes in the whole genome within the same COG category was calculated. A Monte Carlo resampling (bootstrap) of the ratios was performed using 1,000 pseudoreplications using an in-house R script (R Core Team 2012) with a one-tailed *P*-value.

Dating Divergences

RelTime (Mello et al. 2017) was used to estimate divergence times. It is a likelihood-based algorithm that does not incorporate explicitly rate nor tree priors, whereas providing divergence estimates comparable to the widely used Bayesian software BEAST (Drummond et al. 2012; not used here due to lack of convergence among runs having equal priors and data). We used as calibrations (within conservative minima and maxima) two different age estimates for the origin of *M. bovis*: a relatively older interval based on Comas et al. (2013), inspired by the rationale that expansion of *M. bovis* may have been associated with out-of-Africa human expansion (with a conservative upper bound of 88,000 ya obtained by the authors); and a more recent one, based on Bos et al. (2014), employing radiocarbon dating of mummies as tip calibrations (with a conservative interval of 800–7,000 ya as outcome; the same minimum was used for Comas' dating prior in our analysis). We estimated divergences using two data sets, the 4-fold degenerate data set and the site-congruence data set, both minimizing biases due to sites under reasonably different rates (due to forces such as selection). The best RI tree was fixed throughout the runs. To assess variance in date estimates, 1,000 bootstraps (generated under RaxML v8.2.4; Stamatakis 2014) for each data set were analyzed separately using the RelTime version within the MEGA-CC suite (Kumar et al. 2012). Dating outliers for each of those four run combinations (Bos et al. vs. Comas et al. calibrations; each with either 4-fold or site-congruence data sets) were discarded by calculating the 99% highest density intervals (HDI) on each case using the R library "coda" (Plummer et al. 2006). Node date intervals were calculated as [min(minima), max(maxima)] across these HDIs, therefore being conservative as it embraced altogether Comas et al. (2013) and Bos et al. (2014) calibrations.

Results

General *Mycobacterium bovis* Evolution

A total of 6,189 families of homologous genes were obtained. Of these families, 2,550 were inferred as single copy core-coding genes and subsequently multiply aligned (nucleotide sequences) by Guidance (using Mafft), 67 of them being discarded because of alignment issues and/or containing sequences that were not multiple of three (which may be indicative of at least one copy in the alignment being a pseudogene, or the effect of a wrong base call/bad assembly), resulting in 2,483 core-coding groups. The final manually curated multiple alignment of core-coding genes had 2,491,902 bp, with 6,707 variable sites (0.2%), and 1,761 parsimony-informative sites (0.07%). Regarding amino acids, 3,847 variable sites (0.46%), and 1,134 parsimony-informative sites (0.14%) were detected. Pan/core genomic curves are shown in figure 1.

RIs using either core-coding or reduced data sets were lower than the RI for the site-congruence method topology (RI = 0.87), even after changing the phylogenetic method (from ML to MP), or correcting for selection, recombination, nonneutral evolution (fig. 2; supplementary fig. S1, Supplementary Material online, for details), or by including both core and non-core SNPs (supplementary fig. S2, Supplementary Material online, for details). All US genomes clustered together in the site-congruence topology; however, one Uruguayan genome was included within the clade (fig. 2). Given its better RI, this tree was chosen for downstream analyses.

Significant differences (though small in an absolute sense) were observed in genomic characteristics such as codon usage

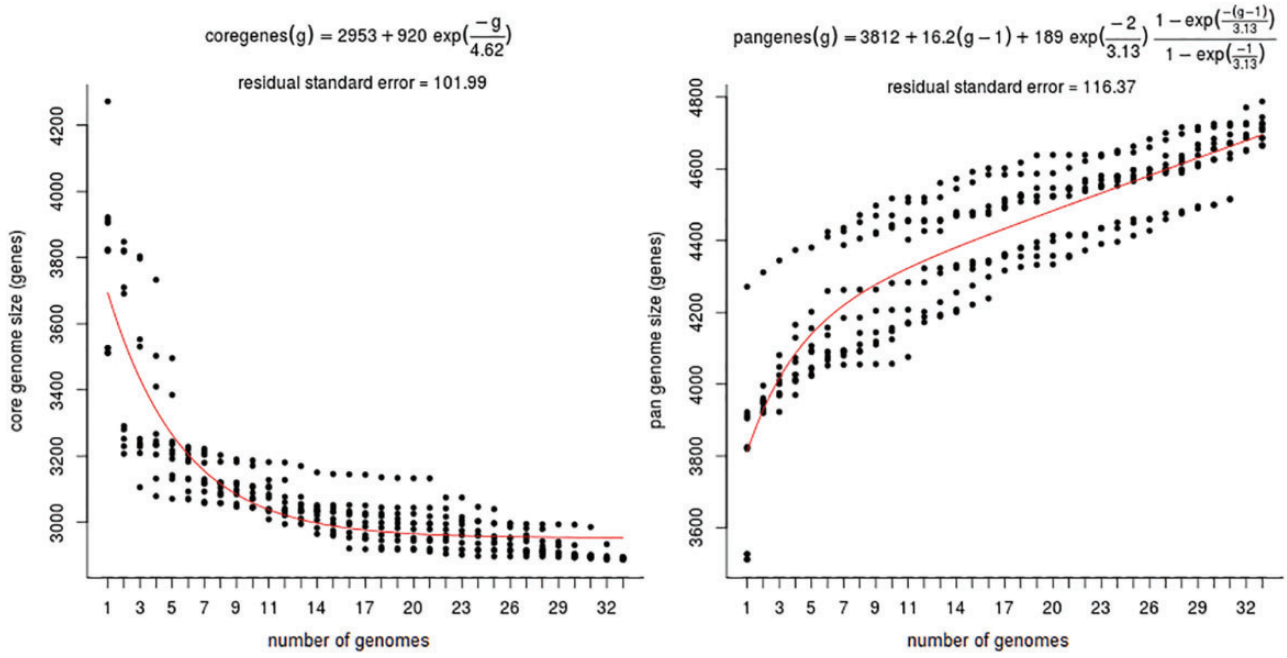

Fig. 1.—Core- (*left*) and Pan-genome (*right*) curves including only genomes of the 38-data set having <100 contigs (33 genomes). The core-genome is close to a plateau, indicating the complete core set is included in the 38-data set analysis (below 3,000 genes). The pan-genome curve indicates the pool of genes is quite high even within relatively closely related *M. bovis* genomes. Equations refer to the exponential regression for each curve.

(Nc, RSCU, and CAI), GC-content, and amount of nucleotides in the core-coding data set between US and non-US genomes (all *t*-tests with *P*-value < 1E-18) (fig. 3), agreeing with the pattern observed for the best RI tree in identifying a US lineage. Curiously, no synapomorphies were observed for the US clade, either in core-coding, or in the kSNP data sets (which includes core and noncore regions) (supplementary fig. S2, Supplementary Material online, for details).

Contrived assessment of read mapping quality (important for testing for spurious outcome in phylogenetic, recombination, and selection results) was done using the 19 PPEs genes in the core-coding, a reasonable choice as PPEs can have more than 100 copies in MTbC genomes and there are also repetitions even within a single PPE paralogue. All single nucleotide variations (SNVs) agreed 100% to SNPs observed in each of the 19 orthologues' multiple alignments (each including the 38 taxa), indicating that phylogenetic inference and genes detected under selection and recombination were not biased to any reasonable extent due to low quality assemblies.

We assessed in ClonalFrameML the amount of recombination in relation to mutations across the branches using the multiple alignment of the ten largest LCBs concatenated (550,634 bp). Rho/theta (ratio of the recombination rate to the mutation rate) and r/m (the ratio of the estimated quantity of polymorphism brought by an event of recombination in relation to an event of mutation) were: Rho/theta = 0.1, and r/m = 0.98, with an average recombination tract length of 67.7 bp. Branch individual values were in general below 1.0 for both ratios except for ten nodes which had Rho/theta ≥ 1.0 and/or r/m ≥ 1.0, seven of them including the US + Uruguay lineage (the branch subtending the clade, the two branches leading to the CO + TX clade, and the other four leading to the MI + Uruguay clade plus all branches within), one subtending the Korean strains, and the other two at more basal positions (the branch leading to the nonAfrican *M. bovis*, and the one leading to all *M. bovis* except the genome from Uganda). The branch leading to the US clade (including the Uruguayan genome) had a homologous recombination import of a PPE family protein segment of 244 bp from Korean strain 1595 (as shown by Blastn (Altschul et al. 1997)). Three different PE-PGRS (a highly polymorphic PE family member; Sampson 2011) genes had segments recombined in the UK BAA-935 genome, with best Blastn hits belonging to *M. bovis* genomes. The US genome 12_5092_S3 from Texas had an import of 244 bp of a PPE family protein gene also with best hits belonging to *M. bovis*. A recombinant PPE gene in the UK genome AN5 had a 169 bp import having best hit with *M. canettii*, a species that is probably in a sister group to MTbC (Fabre et al. 2004; Gutierrez et al. 2005). The best hit of a segment of the 23S ribosomal RNA from the Uruguayan sample (139 bp) was a distant mycobacterium with high pathogenicity (*M. abscessus*).

A total of 114 genes were inferred to be under positive selection; of these, 102 were categorized into COGs (table 1).

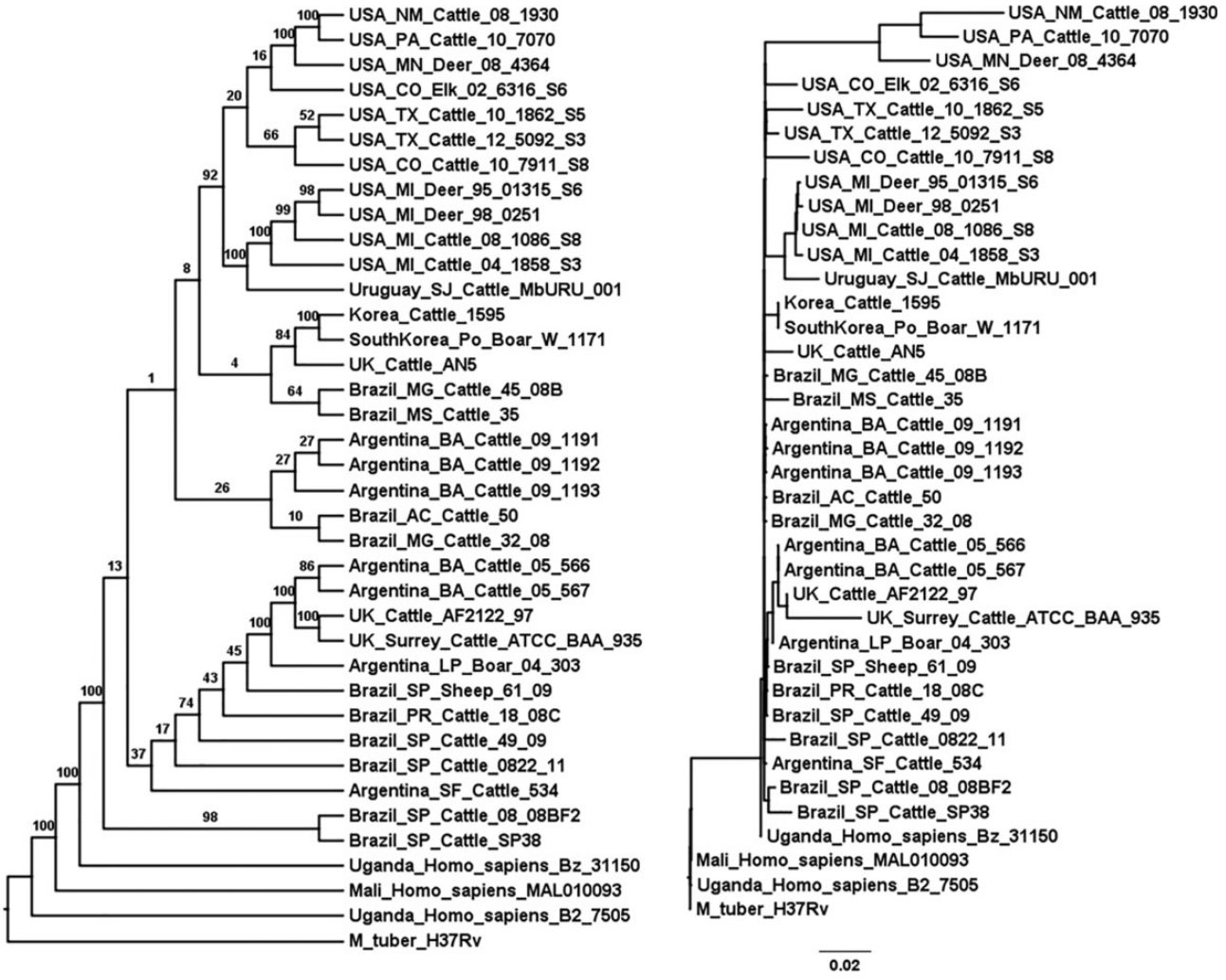

Fig. 2.—ML tree of the 38-data set matrix obtained using the site-congruence method. *Left*: branch support values (UFBoot); *Right*: same tree with actual branch lengths according to ML. RI for this tree (retention index for the character "sampling locality") = 0.81.

Monte-Carlo bootstrapping was conducted to test for COGs significantly enriched under selection. "Cell cycle control, cell division, chromosome partitioning" was the most significant category (table 1), including two copies of an *esx* component (ESX-1 secretion system protein). "Cell motility" was the second most significant; only three genes are within this category, all of them of the PE/PPE family (table 1; supplementary table S3, Supplementary Material online, for details). "Replication, recombination and repair," the third most significant category, had already been described as an enriched category within MTbC (Dos Vultos et al. 2008). Another two significant categories were found: "Lipid transport and metabolism" (14 genes), and "Posttranslational modification, protein turnover, chaperones" (the latter being borderline significant). Important genes under selection (other than PE/PPE and ESX-associated products) included cholesterol metabolism-associated products (e.g., all eight copies of acyl-CoA synthetases), CRISPR type III-associated RAMP protein Csm3, and transposition genes, among others. Genes under positive selection in each category can be found in supplementary table S3, Supplementary Material online, for details.

Large-Scale Polymorphisms

We analyzed five complete *M. bovis* genomes occurring in nature available in GenBank (therefore excluding BCG genomes): two from the United Kingdom, one from Brazil, one from the United States, and one from South Korea. We detected 12 large sequence polymorphisms (LSPs) $\geq$ 2,000 bp, including nine indels, two duplications involving two nonadjacent segments (one of the duplications was probably originally in tandem; see below), and one inversion (table 2). Only LSPs 1 to 4 are newly identified here, the remaining overlapping completely or to a great extent regions previously named

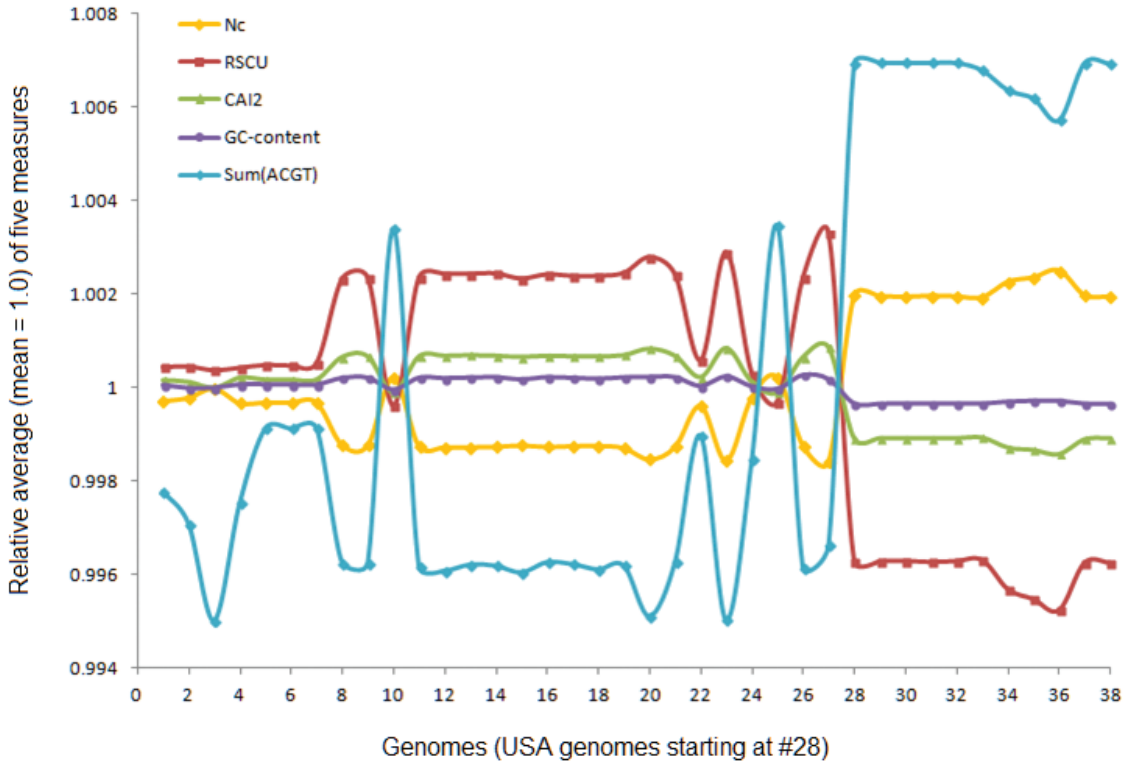

Fig. 3.—Variability of genomic characteristics across the 38 core-coding genes obtained in DAMBE. Each index was normalized to a relative average of 1.0 for better visualization. Numbers in *x* axis refer to individual genomes, with the 11 US genomes starting at position #28. Nc, effective number of codons; RSCU, relative synonymous codon usage; CAI2, corrected version of the codon adaptation index; GC-content, average genomic GC content; Sum(ACGT), size of the core-coding data set in base pairs.

as RD3, RDcap_Asia2, RD2, RD5oryx, RDbovis(c)_fadD18, and RD1 (Brosch et al. 2002; Mostowy et al. 2005). Among the three newly identified indel blocks, two of them include two copies of PE-PGRS family genes (LSP-1 and LSP-4). Further details can be found in table 2 and supplementary table S4, Supplementary Material online, for details.

Origin and Divergence of US Strains

Individual date intervals were obtained as follows ("US clade" includes the Uruguayan genome). Origin of US clade (4-fold data set): [167.27, 5209.91]; Diversification of US clade (4-fold data set): [172.73, 5209.91]; Origin of US clade (site-congruence data set): [149.7, 3492.2]; Diversification of US clade (site-congruence data set): [148.7, 3425.1].

Discussion

Phylogenomics

The phylogenetic analyses indicated that there may be substantial amounts of conflicting signal in the core-coding set of genes. The conflicting signal remains even after trying a different analytical method (maximum parsimony instead of maximum likelihood), correcting for positive selection (either by removal of genes under selection, or by maintaining them but under a global dN/dS ratio phylogeny), correcting for homologous recombination (by using only the largest collinear blocks without signs of recombination), focusing solely on 4-fold degenerate codons, or else considering the core + noncore regions. Nevertheless, the site-conguence method (an algorithm that is not based on phylogeny) was helpful in pointing at a set of sites evolving at similar rates, and these showed a better RI fit under ML, identifying a clade containing all US samples; moreover, the latter also agreed with US genomic features (codon usage bias, GC-content, and length of core-coding set) being significantly different in comparison to the other genomes. Even though the source of conflicting signal was not identified, a possible hint is the observed length of terminal branches (fig. 2; supplementary fig. S1, Supplementary Material online, for details): some of them are relatively long, suggesting that most of the mutations affected external instead of internal branches. A probable cause for the lack of synapomorphies is recent population expansion after a bottleneck, followed by selective sweeps (as suggested by Smith et al. 2006); whenever a selective sweep happens, it tends to hamper variability in the rest of the genome, potentially erasing synapomorphies that led to the evolution of

Table 1

COG Categories Enriched for Genes under Positive Selection

COG Abbr.	COG Category	All	Sel	% All	% Sel	(%Sel)/(%All)
A	RNA processing and modification	1	0	0.0004	0.0000	0.00
B	Chromatin structure and dynamics	0	0	0.0000	0.0000	—
C	Energy production and conversion	147	6	0.0644	0.0541	0.84
D	Cell cycle control, cell division, chromosome partitioning	32	6	0.0140	0.0541	3.85**
E	Amino acid transport and metabolism	165	6	0.0723	0.0541	0.75
F	Nucleotide transport and metabolism	69	4	0.0302	0.0360	1.19
G	Carbohydrate transport and metabolism	101	4	0.0443	0.0360	0.81
H	Coenzyme transport and metabolism	123	6	0.0539	0.0541	1.00
I	Lipid transport and metabolism	188	14	0.0824	0.1261	1.53**
J	Translation, ribosomal structure and biogenesis	115	5	0.0504	0.0450	0.89
K	Transcription	167	9	0.0732	0.0811	1.11
L	Replication, recombination and repair	95	10	0.0416	0.0901	2.16**
M	Cell wall/membrane/envelope biogenesis	89	4	0.0390	0.0360	0.92
N	Cell motility	22	3	0.0096	0.0270	2.80**
O	Posttranslational modification, protein turnover, chaperones	100	7	0.0438	0.0631	1.44*
P	Inorganic ion transport and metabolism	98	5	0.0429	0.0450	1.05
Q	Secondary metabolites biosynthesis, transport and catabolism	129	5	0.0565	0.0450	0.80
R	General function prediction only	334	8	0.1464	0.0721	0.49
S	Function unknown	183	5	0.0802	0.0450	0.56
T	Signal transduction mechanisms	76	2	0.0333	0.0180	0.54
U	Intracellular trafficking, secretion, and vesicular transport	21	1	0.0092	0.0090	0.98
V	Defense mechanisms	27	1	0.0118	0.0090	0.76
W	Extracellular structures	0	0	0.0000	0.0000	—
Y	Nuclear structure	0	0	0.0000	0.0000	—
Z	Cytoskeleton	0	0	0.0000	0.0000	—
	Total =	2282	111	1.00	1.00	

NOTE.—All, all genes; Sel, genes under positive selection; %, relative frequencies; "Total" represents the amount of genes for which a COG was found, and sum of relative frequencies for each COG (=1.0). The last column shows the proportion between (%Sel)/(%All). Test for COG categories significantly enriched for positive selection was done in R by bootstrapping values from this last column and then checking for significance of each category (1,000 pseudoreplicates; $\alpha = 0.05$; one-tailed critical value = 1.52, i.e., genes with Sel/All > 152% are significant).

**Significant under *P*-value = 0.05; *borderline nonsignificant.

Table 2

Polymorphic Genomic Blocks ≥2,000 bp Identified across the Five *Mycobacterium bovis* NCBI Refseqs

Block	AF212297 (UK)		BAA-935 (UK)		Strain 1595 (Korea)		SP38 (Brazil)		Strain 30 (USA)	
	Start	End	Start	End	Start	End	Start	End	Start	End
LSP-1	926157	928398	—		926052	928293	4176355	4178596	926097	928278
LSP-2	1036390	1054849	1029810	1048269	1036444	1054903	4286983	4305442	1036526	1054984
LSP-3	—		—		1413645	1417785	316471	320611	1413866	1418005
RD3	1764645	1774379	b		1769178	1778913	672665	682400	1769203	1778938
RDcap_Asia2	1988553	1997632	—		1993162	2002241	896315	905394	1993090	2002169
RD2	2199641	2210799	a		2204606	2215764	1106529	1117687	2204166	2215322
RD5oryx	2605004	2607181	—		2610050	2612227	1511708	1513885	2609452	2611629
LSP-4	2762357	2772053	2716954	2725679	2767457	2777153	1669019	1678696	c	
Rep1	—		3597147	3614668	—		—		—	
Rep2	—		3614670	3633236	—		—		—	
Rep1	—		3633243	3650816	—		—		—	
Rep2	—		3650819	3669385	—		—		—	
RDbovis(c)_fadD18	3886050	3890304	3909675	3913279	3891927	3896172	2791908	2795181	c	
RD1	4286586	4296110	—		4292815	4302330	3191235	3200759	4277324	4286846

NOTE.—UK genome BAA-935 had most of the losses (seven: LSP-1, LSP-3, RD3, RDcap_Asia2, RD2, RD5oryx and RD1), and is the one bearing the two duplications. US strain 30 had two losses (LSP-4 and RDbovis(c)_fadD18) and the inversion LSP-2, and genome AF2122/97 (UK) had one loss (LSP-3, shared with BAA-935). Gray shading: segment inversion.

[a]Only 417 bp.

clades, and possibly decreasing the correlation with geographic location.

Impact of Selection and Recombination

Homologous recombination is in general assumed to be uncommon within MTbC lineages (except for *M. canettii*) when compared with other organisms (Smith et al. 2006). Mycobacteria lack plasmids in general, and also bear cell wall peculiarities (Prozorov et al. 2014), which in theory restricts homologous recombination. Our results indicate relatively high amounts of homologous recombination instead. This agrees with a previous study focusing on 24 genomes of different species of MTbC (Namouchi et al. 2012). The r/m value of 0.98 shows that the magnitude of sites impacted by recombination are on average as much as mutation. Furthermore, it agrees with what has been found recently regarding mycobacterial DNA exchange: individuals may intercross by distributive conjugal transfer, a process in which there can be (multifragment or large-chunk) transfers and internalization by homologous recombination (Wang et al. 2003; Gray et al. 2013). Though deemed rare, when it happens many stretches of the donor DNA may get incorporated in the chromosome at once (Derbyshire and Gray 2014).

Positive selection was an important evolutionary force within *M. bovis*, impacting 114 genes according to our analyses. Previously, a study by Dos Vultos et al. (2008) had shown that many genes under the COG category "Replication, recombination and repair" were under selection. This is of high significance for the *M. bovis* lineages: being able to accumulate sustainable variability may be important if environmental changes start to select among different alleles. In our study, three other categories were shown to be enriched as well ("Cell cycle control, cell division, chromosome partitioning," "Cell motility," and "Lipid transport and metabolism"; and "Posttranslational modification, protein turnover, chaperones" being marginally nonsignificant). We highlight *esx*-associated genes (important constituents for the secretion of virulence factors; Houben et al. 2014); PPE proteins, which are themselves important virulence factors (Sampson 2011; Akhter et al. 2012; McEvoy et al. 2012); eight copies of Acyl-CoA dehydrogenases, reinforcing the importance of cholesterol-associated pathways for survival, infection, and persistence (Miner et al. 2009; Voskuil 2013); esterase/lipase lipF, shown to be important in pathogenesis (Camacho et al. 1999; Zhang et al. 2005), again reinforcing the importance of virulence-associated genes for *M. bovis* evolution.

Large-Scale Polymorphisms

Although there have been previous reports of polymorphism regarding repetitive markers, SNP polymorphisms, and structural variations within *M. bovis* lineages, this is the first report of LSPs among the five complete genomes occurring in nature. Among the nine indel regions most of the losses occur in one of the UK genomes (BAA-935), except for two that are missing solely from US strain 30 (LSP-4 and RDbovis(c)_fadD18; table 2). Furthermore, the recombination analysis suggested that there is larger-than-average level of segment imports (five of them between ~200 and 520 bp) in the BAA-935 lineage (data not shown). These results show that BAA-935 is relatively differentiated, due to both genic losses and homologous recombination. The other available complete genome from the United Kingdom, AF2122/97, is quite different in these respects (low recombination detected, and only one shared loss, LSP-3). Both types of polymorphism (recombination and indels) involve PE/PPE genes, which are known to be important for virulence (Sampson 2011; Akhter et al. 2012; McEvoy et al. 2012).

Regarding the two duplicated blocks in UK BAA-935, a parsimonious scenario for their evolution is, first, a duplication of the region encompassing the three regions (Rep1 + 63 kb intervening segment + Rep2) followed by loss of the 63 kbp segment, and then a tandem duplication of Rep1 + Rep2. Regarding the other seven polymorphic LSPs that have been described elsewhere (Brosch et al. 2002; Mostowy et al. 2005), four of them include PE/PPE genes; one of them, RD1, also contains ESX-1 secretion-associated genes.

PE/PPE Genes as Evolutionary Targets within *M. Bovis*

PE/PPE genes are known to have evolved from duplications and transpositions from MTbC ancestrals (van Pittius et al. 2006; Newton-Foot et al. 2016), whereas being intrinsically associated with *esx* clusters (ESX 1-5, the latter being the more recent cluster), even though other copies exist outside of such clusters. Such association is both in terms of genomic context (i.e., PE/PPE genes can be found within *esx* clusters) and biochemical function (*esx* proteins are largely responsible for PE/PPE secretion; Houben et al. 2014). In fact, *esx*-associated genes (including ESX clusters 1-5) form the TSS7 secretion system typical of Corynebacteriales (Houben et al. 2014).

Although it is accepted that MTbC lineages are largely clonal, it was found that these genes have large rates of recombination (Liu et al. 2006; Karboul et al. 2006, 2008; McEvoy et al. 2009; Phelan et al. 2016). This same pattern was found in our study, with PE/PPE genes being significantly associated with the evolution of different lineages. A PPE segment of 244 bp present in the US clade (including the Uruguayan genome) was probably a recombination import from a close relative of Korean strain 1595. There were also three other imports in UK BAA-935, one in the United States 12_5092_S3, and another in UK AN5, the latter with best hit with *M. canettii*, which is probably the outgroup of MTbC (Fabre et al. 2004; Gutierrez et al. 2005), suggesting genic exchanges with distant lineages may not be uncommon. Our findings once again undermine the belief that *M. bovis* lineages are strictly clonal, and that recombination may be an important force at generating variation that may be useful for

immune evasion and/or virulence. This also agrees with the observation that mycobacterial lineages can interact with one another using a particular conjugation system (distributive conjugal transfer; Wang et al. 2003; Gray et al. 2013) in which large portions of an extraneous genome can enter the cell at once, increasing the likelihood of recombination at different genomic regions.

The three genes under positive selection that were found under the "cell motility" COG category were all PE/PPE genes. This observation parallels the pattern of PE/PPE enrichment regarding selection found in closely related *M. tuberculosis* strains (Phelan et al. 2016). Finally, two of them were also found within large scale polymorphism indel regions LSP-1 and LSP-4. Overall, these findings evince that PE/PPE genes are targets of different evolutionary processes within *M. bovis*, once again highlighting the importance of the largest mycobacterial family of genes known to date.

It could be argued that, because there are so many PE/PPE copies per MTbC genome (comprising up to ~7% of the total genome's coding gene set), such genes could have been picked up by chance; in other words, the signals observed here could be simply the effect of sampling bias. This may still be a possibility concerning the identification of PE/PPE genes in the new LSP regions, but we believe this was not the case concerning selection nor recombination, as the signals were identified per site and per gene, respectively, not mattering how many copies the gene family has per genome.

This was the first time that different PE/PPE genes were detected as important targets of both selection and recombination in *M. bovis* under a phylogenomic analysis considering all core-coding genes, providing valuable evolutionary insights, especially considering that many PE/PPE proteins are highly antigenic and associated with immune evasion. These results corroborate and extend previous observations of high levels of synonymous and nonsynonymous substitutions in MTbC lineages that were based on far fewer MTbC genomes (e.g., Fleischmann et al. 2002) or analyses focused specifically on PE/PPE genes without considering the whole set of core-coding genes (e.g., McEvoy et al. 2012).

Dating and Hypotheses Regarding the Evolution of the US Lineage

Divergence time of origin and evolution of US strains [148.7, 5209.91], which included two contrasting sets of calibrations regarding origin of *M. bovis* (Comas et al. 2013, up to 88,000 ya; and Bos et al. 2014, at most 7,000 ya), suggested that this lineage evolved at most ~5,000 ya. We further acknowledge that dating estimates (ours included) in general may suffer from a bias due to change of overall evolutionary rates through time, where short time mutation rate (recent epidemics) changes to a slower longterm fixation rate (e.g., Ho et al. 2011), but we do not know of any available software that corrects for such bias. The US data set was shown to be phylogenetically informative regarding association with sampling locality in certain states, when considering at least three individuals within the same clade (fig. 4): Michigan (MI), Hawaii (HI), Minnesota (MN), Colorado (CO; three of five samples), and New York (NY; four of five samples). Particularly in the case of MI and HI, it is known that there are wildlife reservoirs of bovine tuberculosis in nature (Miller and Sweeney 2013), but a previous study did not identify monophyly in either population, nor in others (Joshi et al. 2012). This is probably because these authors used a subset of SNPs (350, with 306 being genic), whereas the present study used the whole set provided in the core-coding genes (to the exclusion of gene alignments discarded by the program Guidance) amounting to 6,707 SNPs, a set more than 20× larger, which was able to detect the actual clusterings. Therefore, our data add phylogenetic evidence to the hypothesis that MI and HI wildlife populations work as either natural reservoirs or spill-overs (such as deer, racoon, coyote, red fox and others; Miller and Sweeney 2013). MN and NY individuals, which also supposedly include wildlife populations with some *M. bovis* prevalence (Miller and Sweeney 2013), also clustered together by state. The fact that NY forms a monophyletic cluster also agrees with it being a locality prone for human transmission, because it may suggest some prevalence of the disease there. Between 2001 and 2004, 35 human cases were reported in NY (Centers for Disease Control and Prevention, 2005), but apparently these were related to contaminated cheese imported from Mexico; nevertheless, two cases of *M. bovis* were reported in white-tailed deer there (Miller and Sweeney 2013), reinforcing the hypothesis of local sources of reinfection as obtained in our phylogeny. Cattle as host was the pattern observed in terms of more basal nodes within each clade, but support values are not uniformly high and in general the host type composes paraphyletic clades, suggesting reinfections across different species (fig. 4). It is also possible to observe that there is a large variation in terms of external branches of some US strains, suggesting that some lineages may evolve faster than others (fig. 4).

The fact that most clades are paraphyletic in terms of sampling locality indicate that the spread of bovine tuberculosis is relatively fast in the United States, and some lineages evolving much faster than others suggest that different haplotypes may evolve at faster speed. This may be due to selective sweeps, which would tend to erase previous site synapomorphies hence eroding phylogenetic signal in deeper nodes within *M. bovis*, which may have been the case found here regarding the 38-set phylogeny as well (fig. 2). Overall, these results indicate the possibility of ongoing and localized outbreaks, which can be even more likely given that many of them happened in the United States in the past 20 years. In fact, the country has been considered of moderate risk concerning control of bovine tuberculosis for some time now (Clifton-Hadley and Wilesmith 1995). Such fast spreading across different states may be due to commingling of infected cattle with

Fig. 4.—ML tree of the 76- data set composed of US strains, plus outgroup, based on their core-coding set of genes. *Left:* chronogram with respective ML support. *Right:* same tree with branches in substitutions/site. *Red:* clades of geographically associated genomes (a paraphyletic clade in the case of MN due to inclusion of a sample from TX).

susceptible wildlife reservoirs, as suggested by the paraphyletic clusters in figure 4, corroborating previous assertions (Miller and Sweeney 2013).

Concluding Remarks

This was the first evolutionary study focusing on many *M. bovis* genomes and using the largest amount of genomic information available. The gains in resolution of the reconstructed tree and associated evolutionary processes are interesting compared with the use of smaller sets of SNPs, particularly in the case of more recent evolutionary scenarios, as seems to be the case in US strains. Nevertheless, the phylogeny of the 38-data set core-coding alignment was somewhat confounded by conflicting signal among different sites. It may be the case that adding more individuals from each location the synapomorphies start to show up; in any case, it would also be interesting to compare how more customary markers (such as spoligotyping or VNTR) compare in terms of phylogenetic resolution. Going one step further, it is highly likely that the concomitant use of repetitive regions and very large sets of genes will become the most effective way to understand *M. bovis* evolution and population dynamics in detail, both in shallower and deeper levels. The importance of such finer resolution comes from the fact that it may become possible to trace past outbreaks in detail, to further explore hypotheses of wildlife reservoirs, and possibly hint at previously unknown population sources of disease spreading.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was partially supported by the Brazilian Agency CAPES (grant number 3385/2013). J.S.L.P. and J.M. were supported by CAPES fellowships. J.C.S. was supported in part by a grant from CNPq (grant number 304881/2015-5). N.F.A. was supported by grants Fundect TO141/2016, TO007/2015, CNPq 305857/2013-4, 473221/2013-6, and CAPES 3377/2013.

Literature Cited

Abdallah A, et al. 2006. A specific secretion system mediates PPE41 transport in pathogenic mycobacteria. Mol Microbiol. 62:667–679.

Akhter Y, Ehebauer MT, Mukhopadhyay S, Hasnain SE. 2012. The PE/PPE multigene family codes for virulence factors and is a possible source of mycobacterial antigenic variation: perhaps more? Biochimie. 94:110–116.

Alexander KA, et al. 2010. Novel *Mycobacterium tuberculosis* complex pathogen, *M. mungi*. Emerg Infect Dis. 16:1296–1299.

Allen AR, et al. 2013. The phylogeny and population structure of *Mycobacterium bovis* in the British Isles. Infect Genet Evol. 20:8–15.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389–3402.

Araújo CP, et al. 2014. Direct detection of *Mycobacterium tuberculosis* complex in bovine and bubaline tissues through nested-PCR. Braz J Microbiol. 45:633–640.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.

Bos KI, et al. 2014. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature 514:494–497.

Bottai D, et al. 2012. Disruption of the ESX-5 system of *Mycobacterium tuberculosis* causes loss of PPE protein secretion, reduction of cell wall integrity and strong attenuation. Mol Microbiol. 83:1195–1209.

Brosch R, et al. 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. Proc Nat Acad Sci U S A. 99:3684–3689.

Brown JM, Lemmon AR. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. Syst Biol. 56:643–655.

Bruen T, Bruen T. 2005. PhiPack: PHI test and other tests of recombination. Montreal (QC): McGill University.

Camacho LR, Ensergueix D, Perez E, Gicquel B, Guilhot C. 1999. Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. Mol Microbiol. 34:257–267.

Clifton-Hadley R, Wilesmith J. 1995. An epidemiological outlook on bovine tuberculosis in the developed world. In: Proceedings of the Second International Conference on Mycobacterium bovis. Otago, Dunedin, New Zealand: University of Otago Press. p. 178–182.

Cole S, et al. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393:537–544.

Comas I, et al. 2013. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet. 45(10):1176–1182.

Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol. 79:7696–7701.

Cousins DV. 2001. *Mycobacterium bovis* infection and control in domestic livestock. Rev Sci Tech. 20:71–85.

Cummins CA, McInerney JO. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Syst Biol. 60:833–844.

Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5:e11147.

Derbyshire KM, Gray TA. 2014. Distributive conjugal transfer: new insights into horizontal gene transfer and genetic exchange in mycobacteria. Microbiol Spectr. 2:4.

Didelot X, Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol. 11:e1004041.

Dos Vultos T, et al. 2008. Evolution and diversity of clonal bacteria: the paradigm of *Mycobacterium tuberculosis*. PLoS One 3:e1538.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 29:1969–1973.

Fabre M, et al. 2004. High genetic diversity revealed by variable-number tandem repeat genotyping and analysis of hsp65 gene polymorphism in a large collection of "*Mycobacterium canettii*" strains indicates that the *M. tuberculosis* complex is a recently emerged clone of "*M. canettii*". J Clin Microbiol. 42:3248–3255.

Farris JS. 1989. The retention index and the rescaled consistency index. Cladistics 5:417–419.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791.
Fleischmann RD, et al. 2002. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J Bacteriol. 184:5479–5490.
Ford CB, et al. 2011. Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. Nat Genet. 43:482–486.
Galperin MY, Makarova KS, Wolf YI, Koonin EV. 2014. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acid Res. 43:D261–D269.
Gardner SN, Slezak T, Hall BG. 2015. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics 31:2877–2878.
Gardy JL, et al. 2011. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med. 364:730–739.
Gray TA, Krywy JA, Harold J, Palumbo MJ, Derbyshire KM. 2013. Distributive conjugal transfer in mycobacteria generates progeny with meiotic-like genome-wide mosaicism, allowing mapping of a mating identity locus. PLoS Biol. 11:e1001602.
Gröschel MI, Sayes F, Simeone R, Majlessi L, Brosch R. 2016. ESX secretion systems: mycobacterial evolution to counter host immunity. Nat Rev Microbiol. 14:677–691.
Gutierrez MC, et al. 2005. Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. PLoS Pathog. 1:e5.
Li H, et al. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079.
Hang'ombe MB, et al. 2012. *Mycobacterium bovis* infection at the interface between domestic and wild animals in Zambia. BMC Vet Res. 8:221.
Harris SR, et al. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. Science 327:469–474.
Hauer A, et al. 2015. Genetic evolution of *Mycobacterium bovis* causing tuberculosis in livestock and wildlife in France since 1978. PLoS One 10:e0117103.
Hilty M, et al. 2005. Evaluation of the discriminatory power of variable number tandem repeat (VNTR) typing of *Mycobacterium bovis* strains. Vet Microbiol. 109:217–222.
Ho SY, et al. 2011. Time-dependent rates of molecular evolution. Mol Ecol. 20:3087–3101.
Houben EN, Korotkov KV, Bitter W. 2014. Take five-Type VII secretion systems of Mycobacteria. Biochim Biophys Acta. 1843:1707–1716.
Joshi D, et al. 2012. Single nucleotide polymorphisms in *Mycobacterium bovis* genome resolve phylogenetic relationships. J Clin Microbiol. 50:3853–3861.
Karboul A, et al. 2006. Insights into the evolutionary history of tubercle bacilli as disclosed by genetic rearrangements within a PE_PGRS duplicated gene pair. BMC Evol Biol. 6:107.
Karboul A, et al. 2008. Frequent homologous recombination events in *Mycobacterium tuberculosis* PE/PPE multigene families: potential role in antigenic variability. J Bacteriol 190:7838–7846.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30:772–780.
Kosakovsky Pond SL, Poon AFY, Frost SDW. 2009. Estimating selection pressures on alignments of coding sequences. In: The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge: Cambridge University Press. p. 419–490.
Kück P, Meusemann K. 2010. FASconCAT: convenient handling of data matrices. Mol Phyl Evol. 56:1115–1118.
Kumar S, Stecher G, Peterson D, Tamura K. 2012. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. Bioinformatics 28:2685–2686.
Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol. 33:1870–1874.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9:357–359.
Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30:3276–3278.
Lewis PO. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst Biol. 50:913–925.
Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.
Lisle GW, Kawakami RP, Yates GF, Collins DM. 2008. Isolation of *Mycobacterium bovis* and other mycobacterial species from ferrets and stoats. Vet Microbiol. 132:402–407.
Liu X, Gutacker MM, Musser JM, Fu YX. 2006. Evidence for recombination in *Mycobacterium tuberculosis*. J Bacteriol. 188:8169–8177.
McEvoy CR, Van Helden PD, Warren RM, van Pittius NCG. 2009. Evidence for a rapid rate of molecular evolution at the hypervariable and immunogenic *Mycobacterium tuberculosis* PPE38 gene region. BMC Evol Biol. 9:237.
McEvoy CR, et al. 2012. Comparative analysis of *Mycobacterium tuberculosis* PE and PPE genes reveals high sequence variation and an apparent absence of selective constraints. PLoS One 7:e30593.
Mello B, Tao Q, Tamura K, Kumar S. 2017. Fast and accurate estimates of divergence times from big data. Mol Biol Evol. 34:45–50.
Miller RS, Sweeney SJ. 2013. *Mycobacterium bovis* (bovine tuberculosis) infection in North American wildlife: current status and opportunities for mitigation of risks of further infection in wildlife populations. Epidemiol Infect. 141:1357–1370.
Miner MD, Chang JC, Pandey AK, Sassetti CM, Sherman DR. 2009. Role of cholesterol in *Mycobacterium tuberculosis* infection. Indian J Exp Biol. 47:407–411.
Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol. 30:1188–1195.
Minkin I, Patel A, Kolmogorov M, Vyahhi N, Pham S. 2013. Sibelia: a scalable and comprehensive synteny block generation tool for closely related microbial genomes. In: International Workshop on Algorithms in Bioinformatics. Berlin (Germany): Springer Berlin Heidelberg. p. 215–229.
Mostowy S, et al. 2005. Revisiting the evolution of *Mycobacterium bovis*. J Bacteriol. 187:6386–6395.
Namouchi A, Didelot X, Schöck U, Gicquel B, Rocha EP. 2012. After the bottleneck: Genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation, recombination, and natural selection. Genome Res. 22:721–734.
Newton-Foot M, Warren RM, Sampson SL, van Helden PD, van Pittius NCG. 2016. The plasmid-mediated evolution of the mycobacterial ESX (Type VII) secretion systems. BMC Evol Biol. 16:62.
Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-Tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 32:268–274.
Parsons SD, Drewe JA, Gey van Pittius NC, Warren RM, van Helden PD. 2013. Novel cause of tuberculosis in meerkats, South Africa. Emerg Infect Dis. 19:2004–2007.
Phelan JE, et al. 2016. Recombination in PE/PPE genes contributes to genetic variation in *Mycobacterium tuberculosis* lineages. BMC Genomics 17(1):151.
Plummer M, Best N, Cowles K, Vines K. 2006. CODA: convergence diagnosis and output analysis for MCMC. R News 6:7–11.
Prozorov AA, Fedorova IA, Bekker OB, Danilenko VN. 2014. The virulence factors of *Mycobacterium tuberculosis*: genetic control, new conceptions. Russ J Genet. 50:775–797.
Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acid Res. 35(Suppl. 1):D61–D65.

Sampson SL. 2011. Mycobacterial PE/PPE proteins at the host-pathogen interface. Clin Dev Immunol. 2011:497203.
Sayes F, et al. 2012. Strong immunogenicity and cross-reactivity of *Mycobacterium tuberculosis* ESX-5 type VII secretion-encoded PE-PPE proteins predicts vaccine potential. Cell Host Microbe 11:352–363.
Sela I, Ashkenazy H, Katoh K, Pupko T. 2015. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res. 43(W1):W7–14.
Smith NH, Gordon SV, de La Rua-Domenech R, Clifton-Hadley RS, Hewinson G. 2006. Bottleneck and broomsticks: the molecular evolution of *Mycobacterium bovis*. Nature 4:670–681.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.
Swofford DL. 2003. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sunderland (MA): Sinauer Associates.
R Core Team. 2012. A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Tettelin H, et al. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". Proc Natl Acad Sci U S A. 102:13950–13955.
Thacker TC, Harris B, Palmer MV, Waters WR. 2011. Improved specificity for detection of *Mycobacterium bovis* in fresh tissues using IS6110 real-time PCR. BMC Vet Res. 7:50.
Thoen CO, Steele JH, Gilsdorf MJ. 2006. Mycobacterium bovis infection in animals and humans. 2nd ed. Oxford: Blackwell Publishing.
Thoen C, LoBue P, De Kantor I. 2006. The importance of *Mycobacterium bovis* as a zoonosis. Vet Microbiol. 112:339–345.
Van Ingen J, et al. 2012. Characterization of *Mycobacterium orygis* as *M. tuberculosis* complex subspecies. Emerg Inf Dis. 18:653–655.
van Pittius NCG, et al. 2006. Evolution and expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. BMC Evol Biol. 6(1):95.
Venditti C, Meade A, Pagel M. 2008. Phylogenetic mixture models can reduce node-density artifacts. Syst Biol. 57:286–293.
Voskuil MI. 2013. *Mycobacterium tuberculosis* cholesterol catabolism requires a new class of acyl coenzyme A dehydrogenase. J Bacteriol. 195:4319–4321.
Wang J, Parsons LM, Derbyshire KM. 2003. Unconventional conjugal DNA transfer in mycobacteria. Nat Genet. 34:80–84.
Warren AS, Setubal JC. 2009. The Genome Reverse Compiler: an explorative annotation tool. BMC Bioinformatics 10:35.
Wattam AR, et al. 2014. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 42 (D1):D581–D591.
Xia X. 2013. DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. Mol Biol Evol. 30:1720–1728.
Zhang M, et al. 2005. Expression and characterization of the carboxyl esterase Rv3487c from *Mycobacterium tuberculosis*. Protein Expr Purif. 42:59–66.

Associate editor: Bill Martin