Genomic determinants of sympatric speciation of the *Mycobacterium tuberculosis* complex across evolutionary timescales

Álvaro Chiner-Oms[1,#], Leonor Sánchez-Busó[2,#], Jukka Corander[2,3,4], Sebastien Gagneux[5,6], Simon Harris[2], Douglas Young[7], Fernando González-Candelas[1,8] Iñaki Comas[8,9]*

[1]Unidad Mixta "Infección y Salud Pública" FISABIO-CSISP/Universidad de Valencia, Instituto de Biología Integrativa de Sistemas-I2SysBio, Valencia, Spain.

[2]Pathogen Genomics, Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK.

[3]Department of Biostatistics, University of Oslo, 0317 Oslo, Norway.

[4]Helsinki Institute of Information Technology (HIIT), Department of Mathematics and Statistics, University of Helsinki, 00014 Helsinki, Finland.

[5]Swiss Tropical and Public Health Institute, Basel, Switzerland.

[6]University of Basel, Basel, Switzerland.

[7]The Francis Crick Institute, 1 Midland Road, London NW1 1AT, UK.

[8]CIBER en Epidemiología y Salud Pública, Valencia, Spain

[9]Instituto de Biomedicina de Valencia, IBV-CSIC, Valencia, Spain.

[#] Contributed equally.

*Corresponding author.

Word-count: 4,016

26

27 ## ABSTRACT

28

29 Models on how bacterial lineages differentiate increase our understanding on
30 early bacterial speciation events and about the relevant niche-specific loci
31 involved. In the light of those models we analyse the population genomics events
32 leading to the emergence of the *Mycobacterium tuberculosis* complex (MTBC)
33 from related mycobacteria species. Emergence is characterised by a combination
34 of recombination events involving multiple core pathogenesis functions and
35 purifying selection on early diverging loci. After the separation from closely related
36 mycobacteria we identify the *phoR* gene, a transcriptional regulator involved in
37 multiple aspects of MTBC virulence, as a key functional player subject to
38 pervasive positive selection. First, during the early diversification of the MTBC,
39 PhoR played a central role defining the host range of the various MTBC
40 members. Later, following adaption to the human host, PhoR mediates host-
41 pathogen interaction during human-to-human transmission. We thus show that
42 linking pathogen evolution across evolutionary and epidemiological timescales
43 lead to the identification of past and present virulence determinants of interest for
44 biomedical research.

The increasing availability of population genomics data has allowed an improved understanding of genotypic and ecological differentiation among closely related bacteria[1]. Models emanating from these data predict that sympatric speciation among bacteria occupying the same ecological niche is more likely leaving many times measurable genetic signatures in extant genomes[2,3]. However, is still underexplored how these models apply to professional pathogens, particularly those characterized by an obligate association with their host species.

Species of the *Mycobacterium tuberculosis* complex (MTBC) cause devastating morbidity and mortality in humans and animals, which also leads to important economic losses[4]. The MTBC comprises a group of bacteria with genome sequences having an average nucleotide identity of greater than 99% and sharing a single clonal ancestor. This includes the predominantly human pathogens referred to as *Mycobacterium tuberculosis* and *Mycobacterium africanum* as well as a series of pathogens isolated from other mammalian species known as *M. bovis*, *M. pinnipedi*, *M. antelope*, *M. microti*, etc. Human-adapted tuberculosis bacilli show a strong geographic association, with some lineages and even sublineages being globally distributed (e.g. lineage 4) and others geographically restricted (e.g. lineage 5, 6, 7)[5,6]. It is assumed that the causes of this variable geographical distribution are both historical (e.g. trade, conquest, globalization) and biological (e.g. interactions with different human genetic backgrounds)[6]. There is limited transmissibility of animal-adapted strains in humans and, conversely, human-adapted strains transmit poorly among animals[7]. Despite the wide range of host species infected by the different members of the MTBC, there is a maximum of ~2,500 single nucleotide polymorphisms (SNPs) separating any two MTBC genomes[8]. The most closely-related bacteria that fall outside of the MTBC include isolates initially referred to as “smooth tubercle bacilli” and now referred to as *Mycobacterium cannettii* (MCAN). MCAN strains differ from MTBC isolates by tens of thousands of SNPs[9]. MCAN strains have been isolated from the Horn of Africa, predominantly from children and often in association with extrapulmonary tuberculosis[10]. Genomic comparisons have identified gene content differences between MTBC, MCAN and other mycobacteria[9,11] as well as genetic differences in virulence-related loci[12].

Our understanding about the population genetics events mediating the divergence of the ancestor of the MTBC from an MCAN-like ancestral pool is far from complete. The availability of genome sequences from thousands of MTBC clinical strains as well as of closely relatives like MCAN enable us not only to identify molecular signatures of MTBC speciation events, but also to reveal known and new targets for biomedical research.

RESULTS

To explore this, we first analyzed the differentiation between MTBC and MCAN by searching for any hallmark of on-going recombination between and within these groups of strains. Previous reports have suggested that there might be limited but significant recombination among MTBC strains[13]. To maximize the chances of identifying potential ongoing recombination events within the MTBC, we screened a data set of 3,475 complete genome sequences of strains from global sources[14] and kept those that maximized the within-MTBC diversity (n =

1,591). These genomes are representative of the known geographic and genetic diversity of the MTBC (Supplementary Fig. 1). Among those genomes, we identified all the biallelic variant positions that were not related with known drug resistance genes and called them in all the strains. In addition, we identified potential false positive variant calls due to mapping errors. We detected 239 positions possibly involved in mapping errors due to duplicated sequences, leaving a final set of 94,780 core variant positions in the MTBC.

From the 94,780 variant positions, we identified potential homoplastic sites, i.e. polymorphic sites showing signs of convergent evolution, by parsimony mapping. A total of 2,360 homoplastic sites were identified (2.5% of all variable sites). Convergence can arise as a result of recombination, selection or neutral processes. With these variant positions, we looked at consecutive runs of homoplastic sites in the genomes. The 2,360 homoplastic positions defined 97 homoplastic runs (see Methods for definition of a homoplastic run). If the accumulation of homoplasies was due to recombination, we expect the variant positions involved in each consecutive run to share the same phylogenetic mapping. From the 97 homoplastic runs, we detected only 2 cases in which two variant positions shared phylogenetic congruence. The two regions accounted for 4 convergent variants (Supplementary Table 1) and affected strains from different MTBC lineages. Variants in positions 2195896 and 2195899 fell in the primary regulatory region of *mazE5*[15]. On the other hand, variants in positions 2,641,161 and 2,641,163 fell in the intergenic region of *glyS* and Rv2358. Although we cannot discard possible recombination events, it is more likely that the two regions have been under positive selection, a mechanism known to lead to homoplastic variants accumulation in the MTBC[16]. In summary, this large-scale variant-by-variant analysis failed to identify on-going recombination between any of the 1,591 MTBC strains analyzed.

As an additional method to identify possible on-going recombination, we evaluated linkage disequilibrium (LD) as a function of the distance between the 94,780 core variant positions. R^2 values were slightly higher at shorter distances, but high values of this parameter can arise when comparing variants with very different frequencies[17] as is revealed by the skewing of the site frequency spectra of the MTBC towards low frequency values (Fig. 1B). We also calculated D'. In this data set, as expected for a mostly clonal organism, LD measured by D' remained at its maximum value, even when focusing at distant variant positions more than 5 Kb apart, suggesting very little or no ongoing recombination (Fig. 1A).

To further validate these findings, we ran Gubbins with the same data set. Gubbins detects the accumulation of a higher than expected number of variants as a hallmark of possible recombination. We partitioned the 1,591 strain dataset into the different lineages and screened for possible tracks of recombination. This lineage-by-lineage analysis identified no potential recombination regions within lineages. Therefore, the three approaches used agreed in assigning a negligible role to recombination in the ongoing evolution of the MTBC.

Having established that recombination is not currently acting within the MTBC, we compared a representative data set of MTBC genomes (Comas 2013, n =

219) with 7 MCAN genomes to identify and quantify ongoing recombination within MCAN and between MCAN and the MTBC. From the 93,922 polymorphic sites identified, 22,718 were biallelic homoplasies (24.2%). The genomic distribution of variant positions and homoplasies was traced for both groups, showing very different landscapes (Fig. 2). A total of 22,464 (98.9%) of those homoplasies were only found among MCAN strains, representing almost half of the variability within this group (22,464/52,392 biallelic sites, 42.9%) which points to recombination as a main source of variability in MCAN. This fact is consistent with previous reports[9]. By contrast and as expected, a flat homoplastic profile was found for the 219 MTBC strains, with a low occurrence of homoplasies (488/30,056 biallelic sites, 1.6%).

To test for ongoing recombination between MCAN strains and MTBC strains, we identified runs of homoplasies involving both groups. From the total 93,922 variants when we put together MCAN and MTBC strains we found 522 variable sites (0.05% of the total) that were polymorphic in both groups, including 254 biallelic homoplasies in the MTBC group. Phylogenetic mapping showed that these homoplasies mapped to the ancestral branch leading to the MTBC clade and did not involve any extant MTBC strain. These results indicate that recombination events were common between MCAN and the ancestor of the MTBC, but were absent during subsequent diversification of the MTBC. Consistently, Gubbins identified 990 potential recombinant segments, most of which mapped on branches involving only MCAN strains (n = 907), thus corroborating that most homoplastic variants only involved MCAN strains. The remaining events mapped to the common branch of all the MTBC strains (Supplementary Fig. 2).

Sympatric and stepwise emergence of the MTBC ancestor

Our results show that recombination with closely related mycobacteria played a role in the emergence of the common ancestor of the MTBC. To gain a better insight into the distribution of genome variability in MCAN and the ancestor of the MTBC, we extracted all the variant positions that were homoplastic between the MTBC ancestor and any of the MCAN strains (7,700 positions). The MTBC ancestor genome showed a similar homoplasy profile to that of the MCAN strains (Fig. 2), suggesting that the ancestor of the MTBC speciated in sympatry with MCAN ancestral strains. Notably, both MCAN and the MTBC ancestor shared a peak around the CRISPR region, highlighting the dynamic nature of this region possibly as a result of common phage infections.

A Gubbins analysis including MCAN genomes and the most likely common ancestor of the MTBC revealed a total of 65 recombination events mapping to the branch leading to the MTBC (Supplementary Table 2). To explore whether these fragments reflected real recombination, a phylogeny was constructed with each of them. A comparison with the topology of the non-recombinant alignment (whole genome alignment subtracting the recombinant regions) using those recombinant regions with enough phylogenetic signal (Supplementary Fig. 3) revealed significant incongruence (SH test; p-value<0.05, Supplementary Fig. 4,

Supplementary Table 3). Thus, both Gubbins and phylogenetic approaches indicated that these 65 regions are likely recombinant regions.

To test whether speciation of the MTBC ancestor occurred in one single episode or in multiple episodes over time, we analyzed the relative age of divergence of the recombination fragments from the MCAN closest clade using BEAST. Given the uncertainties about the timescale of MTBC evolution, the substitution rate of the non-recombinant fragment was estimated normalizing to an arbitrary age of the time since the Most Recent Common Ancestor (tMRCA) of the MTBC of 1 (see Methods for details). We then used the inferred substitution rate to estimate the relative age of each recombinant fragment by using only variants that accumulated in the branch of the MTBC after the recombination event. Results show that the MTBC ancestor differentiated from MCAN sequentially (Fig. 3A). The estimated ages show large HPD intervals as expected from the low number of variant positions per fragment. Although the distribution of tMRCA for the fragments represents a continuum, we can still observe one peak marking “recent” events, just before the whole pathogen population became clonal, and another for “ancient” events, closer to the time of divergence from the MCAN group (Fig 3B).

If recombination played a major role in shaping the MTBC ancestral genome with regards to pathogenesis, we would expect some functions related to the interaction with the host to be affected. Indeed, we observed an enrichment in experimentally confirmed essential genes in the regions involved in the recombination, suggesting that recombination targeted important cell functions (Chi-square test; p-value < 0.01). An enrichment analysis of Gene Ontology terms for the genes contained in these regions identified functions related with growth, and most specifically with growth involved in symbiotic interactions inside a host cell as significantly overrepresented (Binomial test; adj. p-value < 0,05) (Fig 3C). Remarkably most of the genes involved have been implicated in cirulence using animal models of infection (see Discussion, Supplementary Table 4).

A sympatric model of speciation predicts that some parts of the genome will be involved in adaptation to a new niche. The hallmark would be the accumulation of variants differentiating the emerging species, at the genome-wide level or in a few loci, as a consequence of reduced recombination (i.e. incipient speciation of the region). We identified all of the variants that mapped to the MTBC ancestral branch and that had a different nucleotide in all the MCAN strains, the so called divergent variants (divSNPs). The landscape of divergent variants (n = 5,688, Fig. 4) revealed that a total of 120 genes harbored more divergent variants than expected by chance (pFDR <= 0.01).

However, bacterial genomes are highly dynamic and different processes can contribute to the genetic make-up of extant species. Consequently, not all the detected regions necessarily result from pure divergence by accumulation of substitutions. Our phylogenetic analysis identified several genes in which divSNP were introduced by horizontal gene transfer (n = 12) or by recombination to a MCAN not present in our dataset (n = 54).

A total of 53 genes in the MTBC ancestral genome were highly divergent with respect to MCAN due to substitution events (Supplementary Table 5). While the genome-wide identified divSNPs might result from genetic drift or hitchhiking events associated with selection on other loci, the accumulation in only 53 genes suggests that those regions might have played an important role during the process of niche differentiation. In agreement, these 53 genes were significantly more conserved than the rest of the genome (dN/dS = 0.154 vs genome average dN/dS = 0.279, chi-squared p-value =0.000). This result suggests that, despite the increased divergence from the MCAN strains, those 53 regions have been evolving under purifying selection. Alternatively, the accumulation of divergent variants could also represent hotspot regions for mutation. None of the genes showed a similar pattern of mutation accumulation in other MCAN (no overlap between the divSNPs probabilities distributions for these 53 genes and the rest of the genomes, t-test p-value < 0.05).

Regions under positive selection after the transition to obligate pathogen

Having established that some divSNPs accumulate in genes under purifying selection, we screened for positive selection patterns to identify additional genes important in the transition from a newly emerged pathogen to a globally established pathogen. We first revisited the evolution of antigenic proteins. Those regions are recognized by the immune system and most of them are hyperconserved within the MTBC[18,19]. Interestingly, and in agreement with previous data from MCAN genomic analyses[9], the dN/dS calculated in the branch of the ancestor showed a very similar pattern, with essential genes being more conserved than non-essential ones and T-cell epitopes being hyperconserved (Supplementary Fig. 5). Only nine divSNPs (5 synonymous and 4 non-synonymous) were found in T-cell epitopes regions, which is significantly less than expected by chance (Poisson distribution, p-value < 0.001).

Thus, antigenic regions do not show an altered pattern or intensity of selective pressure as one might expect after a speciation event. We then explored what other regions of the genome changed significantly in selective pressure by comparing the MTBC ancestor dN/dS and the actual dN/dS in extant populations. To look for robust dN/dS measures, we only took into account those genes with more than 3 synonymous variant positions and at least 1 non-synonymous variant position for each of the two sets. Due to the low number of divSNPs in individual genes, only 121 genes were evaluated. Consequently, although additional genes to those shown in the ensuing analyses may have changed the selection pattern or intensity, they cannot be evaluated properly (Supplementary Table 6). We were particularly interested in those genes with a drastic change from purifying (dN/dS < 1) to diversifying or positive selection (dN/dS >1) or vice versa.

Most of the genes evaluated did not show any sign of changing selective pressure or pattern. However, when looking at the dN/dS variation data, three genes

appeared as outliers (as defined by Tukey[20]). Genes changing to evolve under positive selection after divergence from the MTBC ancestor were Rv2464c, a DNA glycosylase involved in DNA repair, and Rv0758, also known as *phoR.* Conversely, Rv0202c was under a stronger negative selective pressure following speciation. Notably, PhoR forms part of the PhoP/PhoR virulence regulation system[21]. In the branch leading to the MTBC ancestor, this gene was as conserved at the amino acid level, as other essential genes (Chi-square test; p-value 0.4721), but when we looked within the extant MTBC diversity, the gene was significantly less conserved at the amino acid level than essential genes (Chi-square test; p-value < 0.001).

Positive selection on *phoR* linked to on-going selective pressures

Given the known central role of PhoPR in MTBC virulence, we focused our attention on the new mutations found in PhoR by expanding our MTBC dataset to 4,593 human and five animal genomes. Using this expanded data set, we observed a total of 193 nonsynonymous mutations and 31 synonymous mutations in *phoR* (Fig. 6A). The average dN/dS for this gene was well above 1 (dN/dS = 2.37), suggesting the action of positive selection. Codon-based tests of positive selection for *phoR* identified a higher dN/dS than expected by chance and at least two codons with strong evidence to be under positive selection (Supplementary Table 7). Additional evidence for the action of positive selection on this gene derives from the nonsynonymous mutations, among which we found 34 homoplastic variants, which are strong predictors of positive selection in MTBC (Supplementary Table 8). Non-synonymous mutations significantly accumulated in the sensor domain (chi-square, p-value < 0.01), further corroborating that they are involved in the fine tuning of the PhoR sensitive function to the changing environment during infection (Fig. 6C).

All the new mutations identified in our analysis were found in human clinical isolates and mapped to relatively recent branches in the MTBC phylogeny (Fig 6A). Thus, we reasoned that most mutations were associated with recent selective pressures as compared to the mutations found in the lineage 5, 6 and animal clade reported previously[22]. We tested whether novel *phoR* mutations are arising in clinical settings during infection and for their potential involvement in ongoing transmission. We used a population-based data set from Malawi[23] where more than 70% of the strains were collected during fifteen years and genome sequenced. We found 14 mutations (13 nonsynonymous and 1 synonymous) in *phoR* exclusive of the Malawi data set with *phoR* having a dN/dS of 3.93. Moreover, the mean relative age of the nonsynonymous *phoR* variants was significantly younger than that of other nonsynonymous variants in the dataset (t-test, p-value<< 0.01) and the *phoR* variants from the Malawi data set were more recent than those *phoR* mutations from the reference dataset (t-test, p-value = 0.01)(Fig. 6B). From the 13 nonsynonymous mutations in the Malawi dataset, 8 were in terminal branches and were markers of recent transmission clusters.

Moreover, *phoR* mutations in the Malawi dataset involved larger transmission clusters (permutations test, p-value < 0.001). Taken together, these data indicate that novel *phoR* mutations arise during infection and propagate in on-going human-to-human transmissions in clinical settings.

DISCUSSION

We present evidence that the MTBC ancestor transitioned to an obligate pathogenic lifestyle in sympatry from a common genetic pool including the ancestor of extant MCAN strains. Specifically, we found common patterns of genome-wide recombination between the ancestral MTBC genome and the ancestors of extant MCAN strains. The high recombination rate between MCAN strains, including the MTBC ancestor, stands in sharp contrast to the strictly clonal population structure of extant MTBC strains. By analyzing events leading to the transition from a recombinogenic to a clonal organism, we have also been able to identify genomic regions under different selective pressures. Our results suggest that mutations in *phoR* have allowed the bacteria to adapt to different mammalian hosts and they still play an important role during infection and transmission in current clinical settings.

Population genomics data has led to the development and testing of different models of how different genetic clusters of the same species can arise in sympatry[1,3,24]. In the case of *Vibrio cholera*, an appropriate combination of certain virulence-associated variants, ecological opportunity and additional virulence factors mediated the successful transition of particular clones from an environmental to a pathogenic lifestyle[25]. Other known cases such as pathogenic Salmonella[26] or Yesinia species[27] may have followed a similar scheme. The MTBC represents an extreme case of clonal emergence associated to its obligate pathogenic lifestyle. Here, we have shown that, despite the high average nucleotide identity between MCAN and the MTBC, there is complete genomic isolation between them. There is now experimental evidence in the laboratory that genetic exchange between MCAN strains occurs easily but this is not the case between MCAN and the MTBC[28]. We have shown that there is no measurable on-going recombination among the MTBC strains based on our analysis of 1,591 genomes and in agreement with other recent evidence[29,30]. Recombination in natural populations depends both on the capacity of chromosomal DNA exchange between the two groups involved and on the ecological opportunity. The mechanisms, if any, on how the MTBC bacilli lost their capacity to recombine when the ancestral genetic pool showed very similar recombination patterns to MCAN strains remains to be elucidated. Ecological opportunity may also influence on the lack of opportunities of exchange between MTBC strains. Despite the occurrence of co-infections, the bacilli occupy mainly an intracellular lifestyle, thereby reducing the opportunities for genetic exchange.

We can only speculate on how the transition from a likely environmental or opportunistic pathogen to an obligate pathogen occurred, but our analysis has identified a series of non-random evolutionary events. Notably those events involve core pathogenesis genes. We have identified highly divergent regions in the MTBC ancestor compared to MCAN. The pattern of SNP accumulation suggests that those regions were important in the transition to a closer association with the host. In addition, recombination events in the branch leading to the MTBC ancestor affected essential genes as well as genic regions known to be involved in host-pathogen interaction. The *mymA* operon (Rv3083-Rv3089) is related to the production of mycolic acids and its disruption leads to an aberrant cell-wall structure. Importantly, knock-out studies[31] have shown that this operon is essential for growth in macrophages and spleen. Furthermore, the deletion of genes in this operon leads to a higher TNF-alpha production, highlighting their role on regulating host-pathogen interaction[32]. The other major operon identified in our analysis is the *mce1* operon[33]. *mce1* knock-out mutants are hypervirulent in a mice model of infection and lose the capacity of a proper pro-inflammatory cytokine production that is needed for the establishment of the infection[34] and granuloma[33]. How all these processes are mediated by *mce1* is still not clear pointing to mce1 as a priority target for biomedical research.

Our analysis identified one gene, *phoR*, which is under positive selection in extant MTBC strains although it was under purifying selection in the MTBC ancestor. PhoR is the sensor component of the PhoPR two-component systems, which play a major role in MTBC pathogenesis[35]. Experiments with mutations identified in this study showed that the lipid composition of the membrane and the ESAT-6 secretion, a major virulence factor in the MTBC, are affected[22]. Early mutations in *phoR* were linked to adaptation of the MTBC to different animal species. Here, we show that *phoR* continues to play a role in the ongoing adaptation of MTBC strains during human-to-human transmission. We speculate that recent *phoR* mutations help to fine-tune the immunogenicity of the pathogen during infection, allowing it to manipulate the host response and increase the chances of transmission. Given that PhoPR is involved in membrane composition[36], mutations in this regulator might also be involved in the susceptibility to some antibiotics. However, antibiotic selection is an unlikely explanation for the oldest mutations in PhoPR as they likely predate antibiotic usage.

Thus, a model can be proposed in which recombination, together with acquisition of new genetic materia[11,37], generated a favorable genetic background for the MTBC ancestor to occupy or increase its association to the mammalian host. Contrarily to *Vibrio cholerae*, in which pandemic strains have emerged from the environment multiple times[25], we see this emergence only once in the MTBC, perhaps because the right combination of multiple, fortuitous genetic events and ecological conditions has only happened once. More provocative is the idea that MTBC might just be part of a spectrum of association to the host occupied by the

different MCAN-MTBC groups. The fact that the so-called Clone A MCAN strains are more common in the clinic may suggest differences in ecological niches within the MCAN group itself[38]. In agreement, previous publications[9,39] and our own analysis (Figure 2) have identified Clone A strains as the evolutionary closest MCAN group to MTBC.

In the MTBC, the stronger, obligate association with new host(s) was accompanied by new selective pressures. In accordance, we have identified genes in the MTBC genome highly diverging from MCAN and evolving under purifying selection, suggesting that they have become essential following MTBC's transition to an obligate pathogenic life-style. In the final stages of adaptation, positive selection on genes such as *phoR* and others[39,40] led to a narrowing of the host-range and later still to a further fine-tuning during the spread of the bacteria within the new host species.

ACKNOWLEDGEMENTS

We thank Alberto Marina for advice in the interpretation of the PhoR molecular structure. This work was funded by projects the European Research Council (ERC) (638553-TB-ACCELERATE) and Ministerio de Economía y Competitividad (Spanish Government) research grant SAF2016-77346-R (to IC). BFU2014-58656-R and BFU2017-89594-R from Ministerio de Economía y Competitividad (Spanish Government) and PROMETEO/2016/122 from Generalitat Valenciana (to FGC). ACO is recipient of a FPU fellowship from Ministerio de Educación y Ciencia FPU13/00913 (Spanish Government). Swiss National Science Foundation (grants 310030_166687, IZRJZ3_164171, IZLSZ3_170834 and CRSII5_177163), the European Research Council (309540-EVODRTB) and SystemsX.ch (to SG).

AUTHOR'S CONTRIBUTION

IC conceived this work. ACO, IC, LSB, SH, JC and FGC analyzed the data. ACO, IC, LSB, FGC wrote the first version of the draft. All authors critically reviewed and contributed to the final version of the manuscript.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

1. Vos, M. A species concept for bacteria based on adaptive divergence. *Trends Microbiol.* **19,** 1–7 (2011).

2. Shapiro, B. J. & Polz, M. F. Microbial speciation. *Cold Spring Harb. Perspect. Biol.* **7,** a018143 (2015).

3. Marttinen, P. & Hanage, W. P. Speciation trajectories in recombining bacterial species. *PLOS Comput. Biol.* **13,** e1005640 (2017).

4. WHO | Global tuberculosis report 2017. *WHO* (2017).

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

6. Stucki, D. *et al. Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat. Genet.* **48,** 1535–1543 (2016).

7. Davies, P. D. O. Tuberculosis in humans and animals: are we a threat to each other? *J. R. Soc. Med.* **99,** 539–40 (2006).

8. Coscolla, M. & Gagneux, S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Semin. Immunol.* **26,** 431–444 (2014).

9. Supply, P. *et al.* Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. *Nat. Genet.* **45,** 172–9 (2013).

10. Aboubaker Osman, D., Bouzid, F., Canaan, S. & Drancourt, M. Smooth tubercle bacilli: neglected opportunistic tropical pathogens. *Front. public Heal.* **3,** 283 (2015).

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

12. Brennan, P. J. Bacterial evolution: Emergence of virulence in TB. *Nat. Microbiol.* **1,** 15031 (2016).

13. Namouchi, A., Didelot, X., Schöck, U., Gicquel, B. & Rocha, E. P. C. After the bottleneck: Genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation, recombination, and natural selection. *Genome Res.* **22,** 721–34 (2012).

14. Coll, F. *et al.* A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat. Commun.* **5,** 4812 (2014).

15. Cortes, T. *et al.* Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in *Mycobacterium tuberculosis*. *Cell Rep.* **5,** 1121–31 (2013).

16. Mortimer, T. D., Weber, A. M. & Pepperell, C. S. Signatures of selection at drug resistance loci in *Mycobacterium tuberculosis*. *mSystems* **3,** e00108-17 (2018).

17. Hedrick, P. & Kumar, S. Mutation and linkage disequilibrium in human mtDNA. *Eur. J. Hum. Genet.* **9,** 969–972 (2001).

18. Comas, I. *et al.* Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat. Genet.* **42,** 498–503 (2010).

19. Coscolla, M. *et al. M. tuberculosis* T cell epitope analysis reveals paucity of antigenic variation and identifies rare variable TB antigens. *Cell Host Microbe* **18,** 538–48 (2015).

20. Tukey, J. W. (John W. *Exploratory data analysis*. (Addison-Wesley Pub. Co, 1977).

21. Gonzalo-Asensio, J. *et al.* PhoP: a missing piece in the intricate puzzle of *Mycobacterium tuberculosis* virulence. *PLoS One* **3,** e3496–e3496 (2008).

22. Gonzalo-Asensio, J. *et al.* Evolutionary history of tuberculosis shaped by conserved mutations in the PhoPR virulence regulator. *Proc. Natl. Acad. Sci.* **111,** 11491–11496 (2014).

23. Guerra-Assunção, J. A. *et al.* Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *Elife* **4,** (2015).

24. Shapiro, B. J. *et al.* Population genomics of early events in the ecological differentiation of bacteria. *Science* **336,** 48–51 (2012).

25. Shapiro, B. J., Levade, I., Kovacikova, G., Taylor, R. K. & Almagro-Moreno, S. Origins of pandemic *Vibrio cholerae* from environmental gene pools. *Nat. Microbiol.* **2,** 16240 (2016).

26. Bäumler, A. & Fang, F. C. Host specificity of bacterial pathogens. *Cold Spring Harb. Perspect. Med.* **3,** a010041 (2013).

27. McNally, A., Thomson, N. R., Reuter, S. & Wren, B. W. 'Add, stir and reduce': Yersinia spp. as model bacteria for pathogen evolution. *Nat. Rev. Microbiol.* **14,** 177–190 (2016).

28. Boritsch, E. C. *et al. pks5*-recombination-mediated surface remodelling in *Mycobacterium tuberculosis* emergence. *Nat. Microbiol.* **1,** 15019 (2016).

29. Mortimer, T. D. & Pepperell, C. S. Genomic signatures of distributive conjugal transfer among Mycobacteria. *Genome Biol. Evol.* **6,** 2489–2500 (2014).

30. Boritsch, E. C. *et al.* Key experimental evidence of chromosomal DNA transfer among selected tuberculosis-causing mycobacteria. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 9876–81 (2016).

31. Cheruvu, M., Plikaytis, B. B. & Shinnick, T. M. The acid-induced operon Rv3083–Rv3089 is required for growth of *Mycobacterium tuberculosis* in macrophages. *Tuberculosis* **87,** 12–20 (2007).

32. Olsen, A. *et al.* Targeting *Mycobacterium tuberculosis* tumor necrosis factor alpha-downregulating genes for the development of antituberculous vaccines. *MBio* **7,** e01023-15 (2016).

33. Casali, N., White, A. M. & Riley, L. W. Regulation of the *Mycobacterium tuberculosis mce1* operon. *J. Bacteriol.* **188,** 441–449 (2006).

34. Shimono, N. *et al.* Hypervirulent mutant of *Mycobacterium tuberculosis* resulting from disruption of the *mce1* operon. *Proc. Natl. Acad. Sci.* **100,** 15918–15923 (2003).

35. Broset, E., Martín, C. & Gonzalo-Asensio, J. Evolutionary landscape of the *Mycobacterium tuberculosis* complex from the viewpoint of PhoPR: implications for virulence regulation and application to vaccine development. *MBio* **6,** e01289-15 (2015).

36. Walters, S. B. *et al.* The *Mycobacterium tuberculosis* PhoPR two-component system regulates genes essential for virulence and complex lipid biosynthesis. *Mol. Microbiol.* **60,** 312–330 (2006).

37. Orgeur M, Brosch R. Evolution of virulence in the *Mycobacterium tuberculosis* complex. *Curr Opin Microbiol.* **41**, 68-75 (2018).

38. Blouin, Y. *et al.* Progenitor *Mycobacterium canettii* clone responsible for lymph node tuberculosis epidemic, Djibouti. *Emerg. Infect. Dis.* **20,** 21–28 (2014).

39. Ates, L. S. *et al.* Mutations in ppe38 block PE_PGRS secretion and increase virulence of *Mycobacterium tuberculosis*. *Nat. Microbiol.* **3,** 181–188 (2018).

40. Malone, K. M. *et al.* Comparative 'omics analyses differentiate *Mycobacterium tuberculosis* and *Mycobacterium bovis* and reveal distinct macrophage responses to infection with the human and bovine tubercle bacilli. *Microb. Genomics* **4,** (2018).

41. Walker, T. M. *et al.* Whole-genome sequencing for prediction of

Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. *Lancet. Infect. Dis.* **15,** 1193–1202 (2015).

42. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27,** 863–4 (2011).

43. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–95 (2010).

44. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–9 (2009).

45. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22,** 568–76 (2012).

46. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin).* **6,** 80–92 (2012).

47. Feuerriegel, S. *et al.* PhyResSE: a web tool delineating *Mycobacterium tuberculosis* antibiotic resistance and lineage from whole-genome sequencing data. *J. Clin. Microbiol.* **53,** 1908–14 (2015).

48. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).

49. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44,** W242-5 (2016).

50. Maddison, W. P. & Maddison, D. R. Mesquite: a modular system for evolutionary analysis. (2017).

51. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–75 (2007).

52. Team, R. C. R: a language and environment for statistical computing. (2015).

53. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* **5,** e11147 (2010).

54. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins.

Nucleic Acids Res. **43,** e15 (2015).

55. Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18,** 502–4 (2002).

56. Bos, K. I. *et al.* Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* **514,** 494–497 (2014).

57. Rambaut, A., Suchard, M., Xie, D. & Drummond, A. Tracer v1.6. (2014).

58. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. *Bioinformatics* **21,** 3448–3449 (2005).

59. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).

60. Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the q-value . 2013–2035 (2003).

61. Fedrizzi, T. *et al.* Genomic characterization of nontuberculous Mycobacteria. *Sci. Rep.* **7,** 45258 (2017).

62. Lechner, M. *et al.* Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12,** 124 (2011).

63. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539 (2011).

64. Ota, T. & Nei, M. Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. *Mol. Biol. Evol.* **11,** 613–9 (1994).

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

66. Murrell, B. *et al.* Gene-wide identification of episodic selection. *Mol. Biol. Evol.* **32,** 1365–1371 (2015).

67. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425 (1987).

68. Pennell, M. W. *et al.* geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees.

Bioinformatics **30,** 2216–2218 (2014).

69. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44,** D279–D285 (2016).

70. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* **46,** D493–D496 (2018).

620 ## FIGURE LEGENDS

621 **Fig. 1. No ongoing recombination within the MTBC** A) Linkage disequilibrium
622 as a function of genetic distance detected in a representative sample of
623 *Mycobacterium tuberculosis* complex strains (n = 1,591). B) Site frequency
624 spectrum of MTBC strains using the 94,781 variant positions.

625 **Fig. 2. Genome-wide variant profiles vary between *M. canetti*, M.**
626 **tuberculosis and the MTBC ancestor** A) Number of homoplasies (grey) as a
627 function of the total number of variants detected (orange) in the MCAN dataset,
628 in the branch leading to the MTBC most recent common ancestor and within the
629 MTBC. Black dots indicate recombination events detected in the most recent
630 common ancestor of the MTBC. B) Homoplastic variant positions mapping to the
631 branch of the most recent common ancestor of the MTBC coincide with events
632 detected by Gubbins (highlighted in yellow) and show correlated phylogenetic
633 patterns within each event.

634 **Fig. 3. Past recombination between *M. canetti* strains and the MTBC**
635 **ancestor** A) Relative age of the recombination fragments detected (black boxes).
636 The red error bars represent the 95% highest probability density (HPD). The age
637 was relativized from 0 to 1, being 0 the age of the non-recombinant fraction of the
638 genome and 1 the highest value in the confidence intervals. B) Histogram
639 distribution of the recombination fragments relative ages. C) Gene Ontology
640 terms overrepresented in the coding regions contained in the recombinant
641 fragments.

642 **Fig. 4. Divergent positions between the MTBC ancestor and *M. canetti***
643 **clade.** Average of divSNPs per 10 kb positions (green) as compared to the
644 average of homoplastic variants (gray). Blue arrows above the distribution are
645 genes that significantly accumulate more divSNPs.

646 **Fig. 5. Genes with differential selective pressures across the MTBC**
647 **speciation stages.** Genes changing selective pressure in the branch of the
648 MTBC ancestor as compared to extant MTBC strains. Red lines mark those
649 genes being outliers of the dN/dS variation distribution.

650 **Fig. 6. *phoR* is under positive selection in human affecting strains.** A)
651 Genome-based phylogeny calculated from a total of 4,598 clinical samples
652 obtained from different sources. The synonymous and nonsynonymous variants
653 found in are *phoR* mapped to the corresponding branch. Variants in internal
654 branches affect complete clades which are coloured in the phylogeny.
655 Homoplasies are marked in the outer circle of the phylogeny. B) Relative ages
656 distribution of the *phoR* variants in the reference dataset from Coll *et al.*[14] and the
657 transmission dataset[23] in comparison with the rest of the genome variants. C)
658 Schematic view of PhoR with the aminoacid changes found across the 4,598
659 samples dataset marked on it. Aminoacid changes are significantly more
660 abundant in the sensor domain (p-value < 0,01).

661

METHODS

Datasets used

Global collection 1 (1,591 isolates). This collection was obtained from Coll *et al.*[14] who gathered it from different sources. We downloaded all the available FASTQ files associated to the following ENA accession numbers: ERP000192, ERP000276, ERP000520, ERP001731, ERP000111, SRP002589, ERP002611, ERP000436, SRA065095, ERP001885, ERP001567. A total of 3,475 genomes were downloaded, aligned to the reference and their variants extracted. The transmission clusters were analyzed (transmission event defined by less than 15 SNPs) and a representative genome of each transmission cluster was kept. We ended with 1,591 strains, representative of the 7 lineages.

Transmission data set (1,646 isolates). This data set was obtained from Guerra-Assunção *et al.*[23] It includes samples taken over a 15-year period in a district of Malawi. We downloaded all the available FASTQ files associated to the following ENA accession numbers: ERP000436 and ERP001072.

Global collection 2 (219 isolates). This dataset from Comas *et al.*[5] represent a selected set of strains representing the known phylogenetic diversity of the human MTBC strains.

Global collection 3 (4,762 isolates). This dataset was obtained by joining the global collection 1 and the transmission dataset described above, as well as the isolates obtained from Walker *et al.*[41]. To the fastq files obtained previously, we added the ones obtained from the accession numbers found in the supplementary material from Walker *et al.* A total of 7,977 genomes were put together, aligned to the reference and their variants extracted. In this case, the transmission clusters (<15 SNPs) were filtered out and only one representative strain of each cluster was kept. Also, we removed potential coinfections. Coinfections were assessed by looking at lineage/sub-lineage coexisting markers[14]. Samples with evidence of more than one variant were considered as possible co-infections and removed from ensuing analyses.

***Mycobacterium canettii* dataset.** Nine *M. canettii* draft genomes were downloaded from Genbank (CIPT 140010059, NC_015848.1; CIPT 140070010, NC_019951.1; CIPT 140060008, NC_019950.1; CIPT 140070017, NC_019952.1; CIPT 140070008, NC_019965.1; CIPT 140070002, NZ_CAOL00000000.1; CIPT 140070005, NZ_CAOM00000000.1; CIPT 140070013, NZ_CAON00000000.1 and CIPT 140070007, NZ_CAOO00000000.1). However, genomes from strains CIPT 140070010 and CIPT 140070017 were discarded because they showed a larger proportion of SNPs than expected, many of them potential sequencing errors.

MTBC most likely ancestral genome. The MTBC ancestor was derived in a previous publication[18]. This ancestor is H37Rv-like in terms of genome structural variants, but H37Rv alleles were substituted by those present in the inferred common ancestor of all MTBC lineages.

FASTQ mapping and variant calling for the MTBC strains

Fastq files were trimmed to remove low quality reads using prinseq[42] and aligned to the MTBC most likely ancestral genome[18] using BWA-mem algorithm[43]. Alignments with less than 20x mean coverage per base were filtered out. The variant calling was performed using samtools[44] and VarScan[45]. Due to the low variability found in *M. tuberculosis*, to avoid mapping errors and false SNPs a variant was filtered out if: i) it was supported by less than 20 reads; ii) it was found in a frequency of less than 0.9; iii) it was found near indel areas (10 bp window); or iv) it was found in areas of high accumulation of variants (more than 3 variants in a 10 bp defined window). Variants were annotated using SnpEff[46]. Variants present in PE/PPE genes, phages or repeated sequences were also filtered-out, as they tend to accumulate SNPs due to mapping errors. High quality variant calls were combined in a non-redundant variant list and used to retrieve the most likely allele at each strain to generate a variant alignment.

Phylogenetic inference and parsimony mapping of SNPs

In the Global-1 dataset we identified 140,239 variants following the steps defined above. As we wanted to identify nucleotide variants due to recombination events, a stricter filtering was applied to remove putative recombination signal due to polymorphisms introduced by other causes. Variants related with antibiotic resistance were obtained from PhyResSe[47] and were removed from the analysis. Also, non-biallelic variants were removed from the analysis. To avoid false positives, we also removed positions in which a variant was called in at least one strain but also with a gap in at least other strain. To identify variants coming from mapping errors we generated fragments of 50 bp downstream, upstream and midstream of the variant positions in the reference genome. With these fragments, we performed a BLAST search over the reference genome to check whether they mapped to other regions. Variants identified in reads that mapped to more than one region of the reference genome (query coverage per HSP over 98% and percentage of identical matches between the query and the reference genome of 98%) were removed from the analysis.

The remaining variants (94,780) were used to infer a phylogenetic tree using RAxML[48] with the GTRCATI (GTR + optimization of substitution rates + optimization of site-specific evolutionary rates) model of evolution and represented with the iTOL software[49]. Variants were mapped to the phylogeny using the Mesquite suite[50]. Homoplastic variants were identified based on parsimony criteria. Using these homoplastic variant positions, we looked for homoplastic runs. A homoplastic run was defined if two (or more) homoplastic variants were found in the genome in correlative positions or with at least one variant between them. Variants present in the same homoplastic run were mapped on the phylogeny using Mesquite to look for coincident phylogenetic patterns.

Linkage-disequilibrium calculation

Using the filtered variant positions (94,780), we used the PLINK software[51] to calculate the linkage-disequilibrium statistics D' and R^2. To estimate these values, we took into account variants with a minimum frequency of 0.01 and we used a sliding window of 10 Kb. The results obtained were processed with R[52]. To plot the D' and R^2 pattern by variant distance, we calculated average D' and R^2 values for 50 bp sliding windows.

Multiple alignment of *M. canettii* and MTBC

Seven *M. canettii* draft genomes were aligned to each other and to the ancestor of MTBC using progressiveMauve[53]. The segmented alignment obtained in XMFA format was converted to a plain Fasta format using the MTBC ancestor as reordering reference with a custom Perl script. Positions with gaps in the reference sequence were removed from the final alignment, so the resulting aligned genomes had the same size than the reconstructed MTBC ancestor (4,411,532 Mb). The MTBC pseudogenomes reconstructed from mapping to the MTBC ancestor from the different datasets described above were concatenated to the *M. canettii* alignment obtained in the previous step for further analyses.

Recombination analyses and phylogenetic evaluation

Recombination was evaluated in the alignment containing 219 strains from MTBC and 7 *M. canettii* and in the one containing the MTBC ancestor and 7 *M. canettii*. First, repetitive regions (i.e. PPE/PGRS) were masked from both alignments and, second, recombination events were inferred using Gubbins[54], which identifies clusters of high SNP density as markers.

Gubbins identify 70 potential recombinant regions in the alignment containing the 7 *M. canetti* strains and the MTBC ancestor. Four of these regions were obviated as they fell in regions deleted in several *M. canetti* strains. One more region was removed from the analysis because it was extremely short (41 bp) and we did not obtain reliable results in the subsequent analysis.

For the remaining 65 fragments a phylogeny was calculated using RAxML[48] and applying the GTRCATI model. Also, a reference phylogeny was calculated with the same method using the complete genomes after subtracting these 65 regions. This reference phylogeny had the same topology as the one obtained from the complete genomes. To test for phylogenetic incongruence between the putative recombination fragments and the genome phylogeny we applied the Shimodaira-Hasegawa and Expected Likelihood Weight tests implemented in TREE-PUZZLE[55].

Dating analyses

To infer the age of the 65 recombinant fragments we first reasoned that most of the mutations found were contributed by recombination and not by mutation once the fragment had been integrated in the genome. Thus, before dating the

fragments we first removed all the homoplastic variants with other MCAN strain found in the fragments. The final alignments for the 65 fragments consisted of only those variants accumulated after the recombination event. We then used the non-recombinant part of the genome to infer a substitution rate assuming two different dating scenarios published for the tMRCA[5,56]. We run BEAST for each fragment pre-specifying monophyletic groups and substitution rate based on the non-recombinant genome phylogenetic reconstruction. We used an uncorrelated log-normal distribution for the substitution rate in all cases and a skyline model for population size changes. We ran several chains of up to 10E6 generations sampling every 1E3 generations to ensure independent convergence of the parameters. Convergence was assessed using Tracer[57]. For both evolutionary scenarios, the results obtained were largely congruent and proportional to the age limit imposed for the MTBC ancestor. As there is controversy about the correct MTBC ancestor age, the results were transformed to relative ages for plotting the final results.

Gene ontology enrichment analysis

Genes present in the recombinant regions between *MCAN* and the MTBC ancestor were annotated using SnpEff[46]. A Gene Set Enrichment analysis (GSE) was performed to look for enriched gene functions in these regions. The BiNGO tool[58] was used to study the enrichment in certain functional categories comparing the most abundant terms in the recombinant regions in comparison with those contained in the complete annotation. The tool uses a hypergeometric test (sampling without replacement) and the BH correction for multiple testing comparisons. The Cytoscape program[59] was used to visualize the results.

divSNP analysis

From the *M.canetti* and MTBC ancestor alignment, we extracted those positions having one variant in all the *M.canetti* strains and another variant in the MTB ancestor. The divSNP frequency by nucleotide was calculated by dividing the total number of divSNPs (5688) by the total number of bases in the alignment. Next, the expected abundance of divSNPs for each gene was calculated by multiplying the nucleotide divSNP frequency by the number of nucleotides in each gene. From the expected and the observed divSNP abundance, we used a Poisson distribution to calculate the probability of having the observed divSNPs by chance for each gene. We selected genes having a pFDR <=0,01 using the q-value Storey method[60].

Complete mycobacterial genomes for reference strains[61](Supplementary table 9) were downloaded from RefSeq and GenBank. The orthologous genes were calculated from the amino acid sequences and using the Proteinortho tool[62]. A gene was considered as orthologous if the BLAST analysis showed a minimum identity of 25%, a query coverage of 50% and a maximum e-value of 1E-05. The orthologous genes were aligned using Clustal-omega[63] and the phylogenies were constructed using RAxML and applying the PROTCATIAUTO model. The reference phylogeny was constructed using only the core genome (proteins

having orthologous in all the mycobacterial genomes downloaded) with RAxML using the same options as above. The reference and alternative phylogenies calculated with the orthologous for the divSNPs enriched genes were manually inspected to check for congruence.

dN/dS analysis

The dN/dS statistics were calculated using the R statistical language[52]. The potential synonymous and non-synonymous substitution sites for each region were calculated using the SNAP tool[64]. The dN/dS ratio for each region was calculated as follows:

$$\frac{Non-synonymousvariants/Non-synonymoussites}{Synonymousvariants/Synonymoussites}$$

The dN/dS for the MTBC ancestor was calculated using the divSNPs while the dN/dS for the MTBC were calculated using the 94,780 SNPs defined as the core variant set. To look for a robust comparison between both ratios only genes having at least 3 synonymous and 1 non-synonymous variants were taken into account. To compare the dN/dS ratios, both were normalized by the genomic dN/dS for each category (0,24 for the MTBC ancestor and 0,62 for the MTBC). The difference between the dN/dS ratio was calculated by subtracting the MTBC dN/dS to that of the MTBC ancestor. The genes that account for the largest differences in the dN/dS were identified as outliers (Q2-1.5*IQR, Q3+1.5*IQR)[20] of the differences distribution.

phoR positive selection analysis

Positive selection over *phoR* was tested using FUBAR[65] and BUSTED[66]. FUBAR was run with 5 MCMC chains of length 10,000,000. 1,000,000 states were used as burn-in and a Dirichlet prior of 0.5. BUSTED was run with default parameters.

To study the potential effect of *phoR* mutations on transmission efficacy we used the data set from Guerra-Assunção *et al.*[23]. We identified SNPs in branches leading either to leaves or to transmission clusters. Transmission clusters were categorized in large, medium or small according to the number of isolates in the cluster (large = over 75th percentile, medium=between 25th and 75th percentile, small = under 25th percentile). Each gene was scored to check for accumulation of mutations in branches leading to large transmission clusters according to the expression:

$$Score = Largeclusters * 3 + Mediumclusters$$

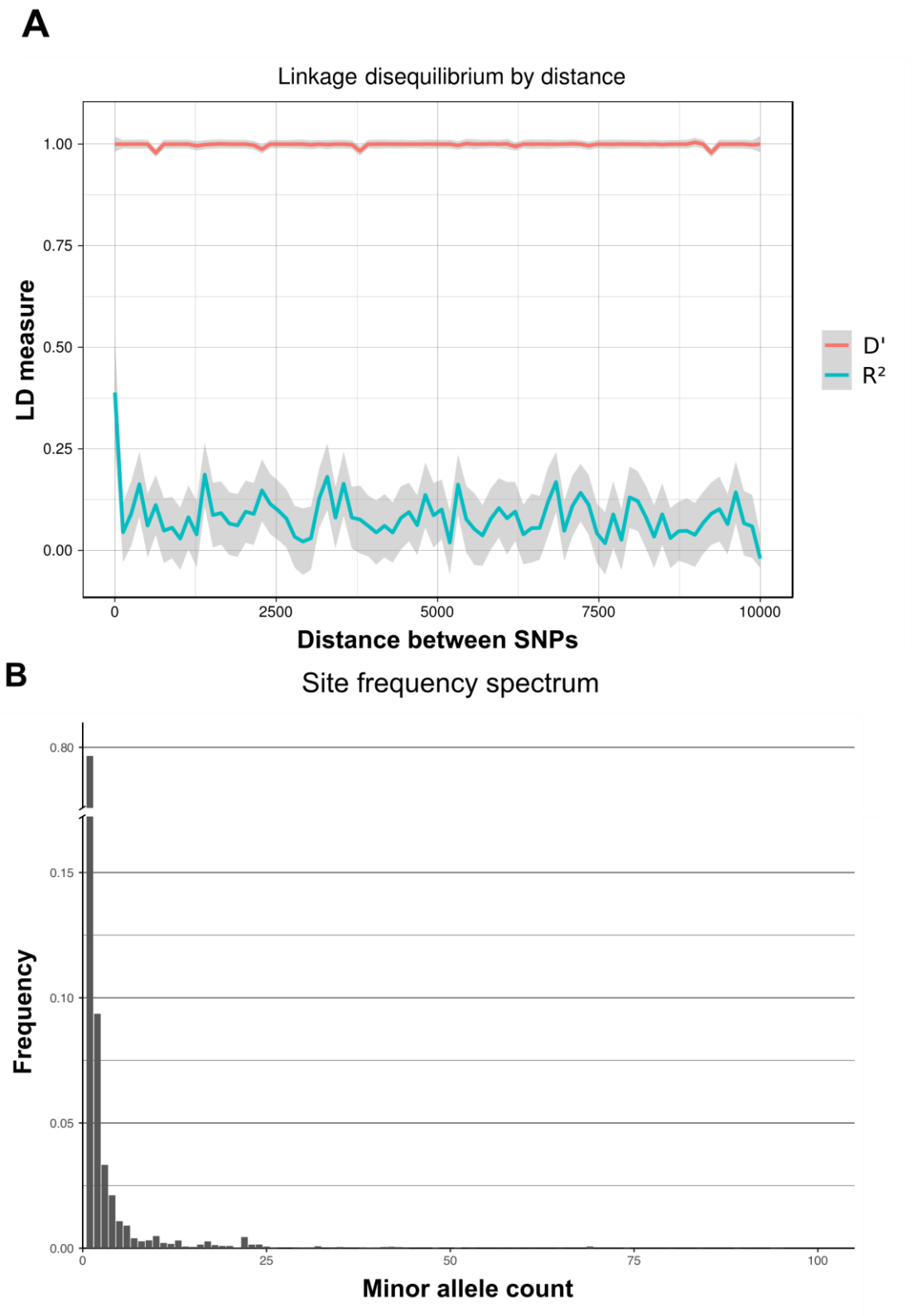
Genes with high mutation rates have a higher number of polymorphisms that could lead to a larger score by chance. To test the probability of obtaining the observed score by chance, a permutation test was carried out 10,000 times. Each of the SNPs identified was reassigned randomly to the same branches and the score was recalculated for each gene. The expected score distribution for each gene was compared to the observed score to calculate the probability. This test was performed for transmission events defined at 10 SNPs.

The relative ages for the variant positions were calculated as node-to-tip distances. In order to have a common framework, a phylogeny was constructed including all the samples from the transmission dataset and from the reference dataset. The phylogeny was constructed using the Neighbour-Joining algorithm[67]. For each variant position we first identified the node in which the variant appeared. The node-to-tip distance was calculated afterwards for each node using the geiger package[68]. Distances were normalized to obtain a relative distance. Later, all the non-synonymous variants except the *phoR* polymorphisms were used as a reference set. The nonsynonymous *phoR* variants to be compared were categorized in two groups, those exclusive to the reference dataset[14] and those derived from the transmission dataset[23].

PhoR structure representation

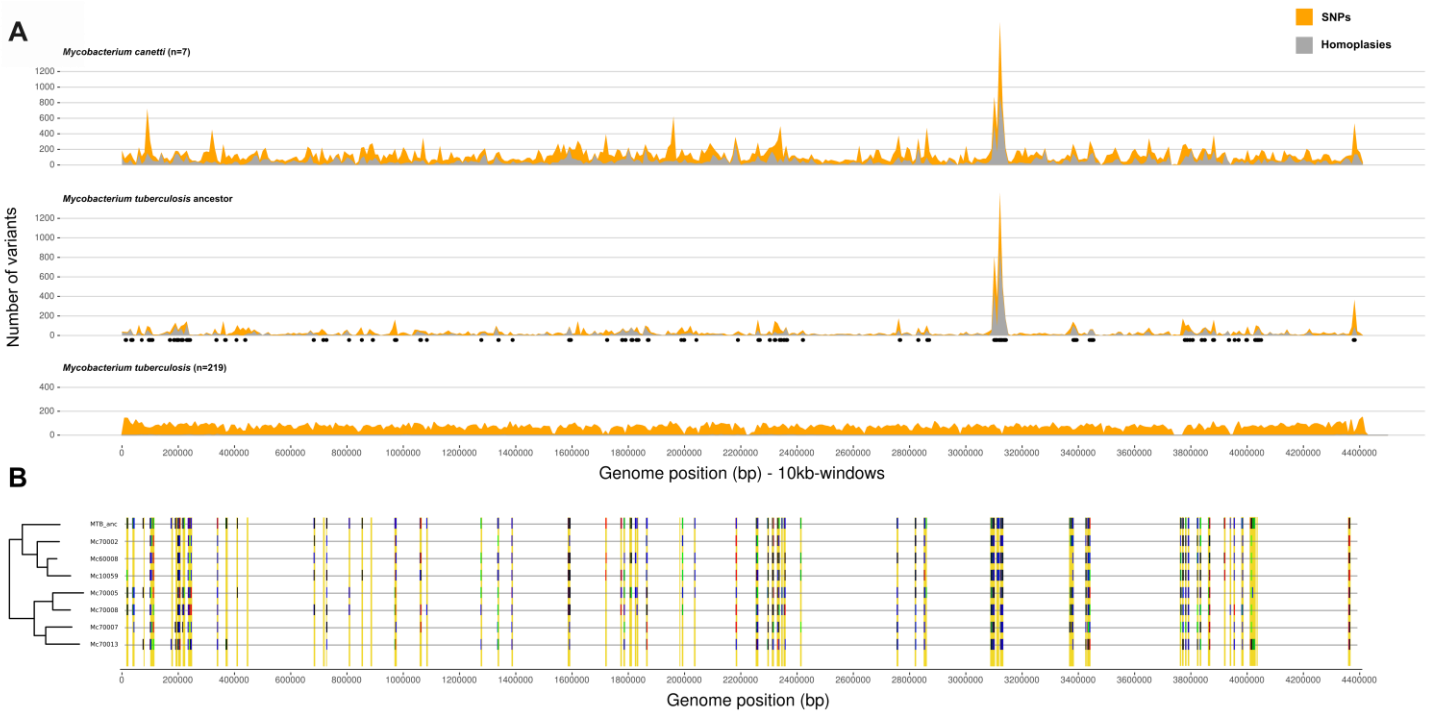
The PhoR structure was inferred by using PFAM[69] and SMART[70].

901 **FIGURES**



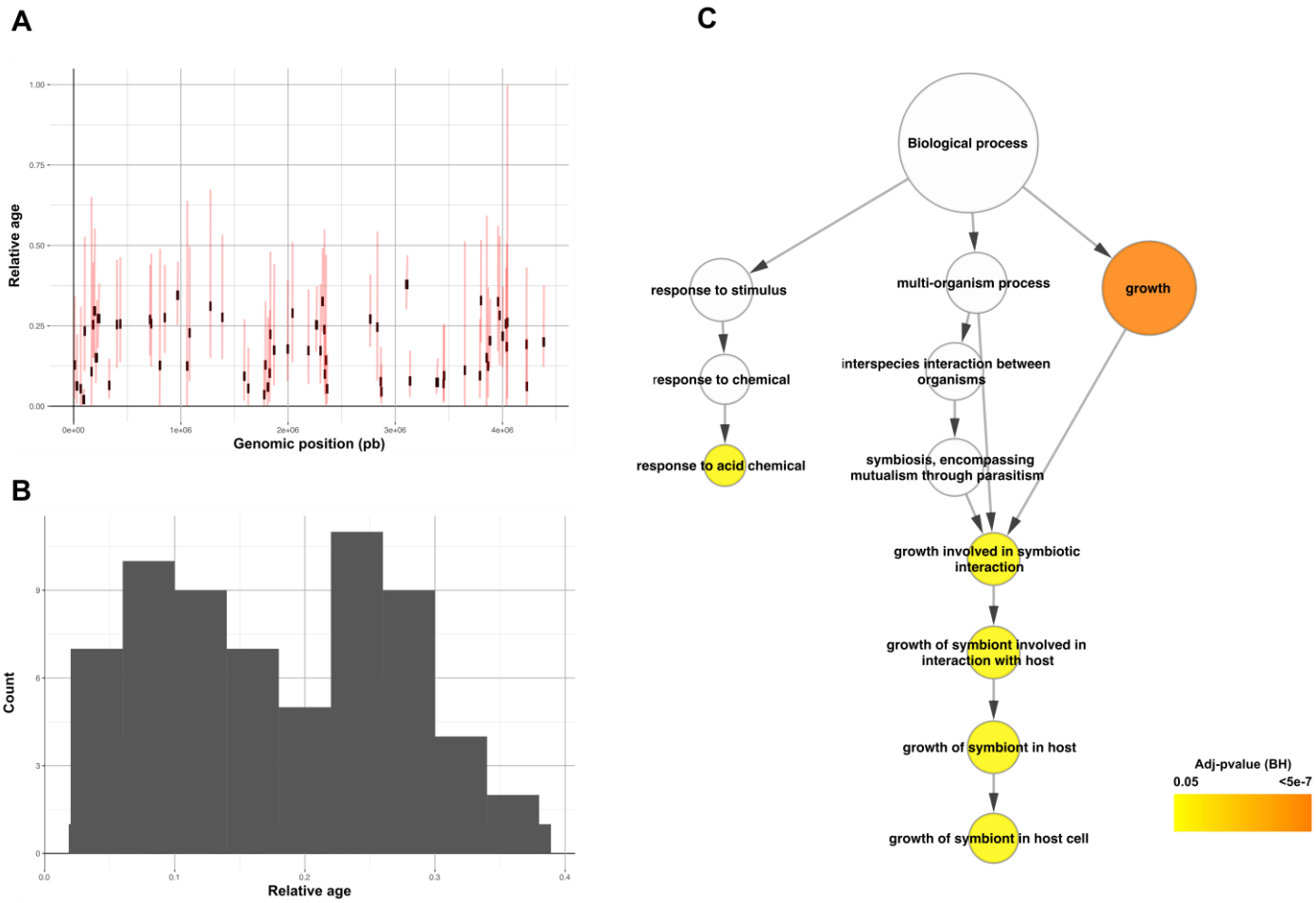
902 Figure 1

903

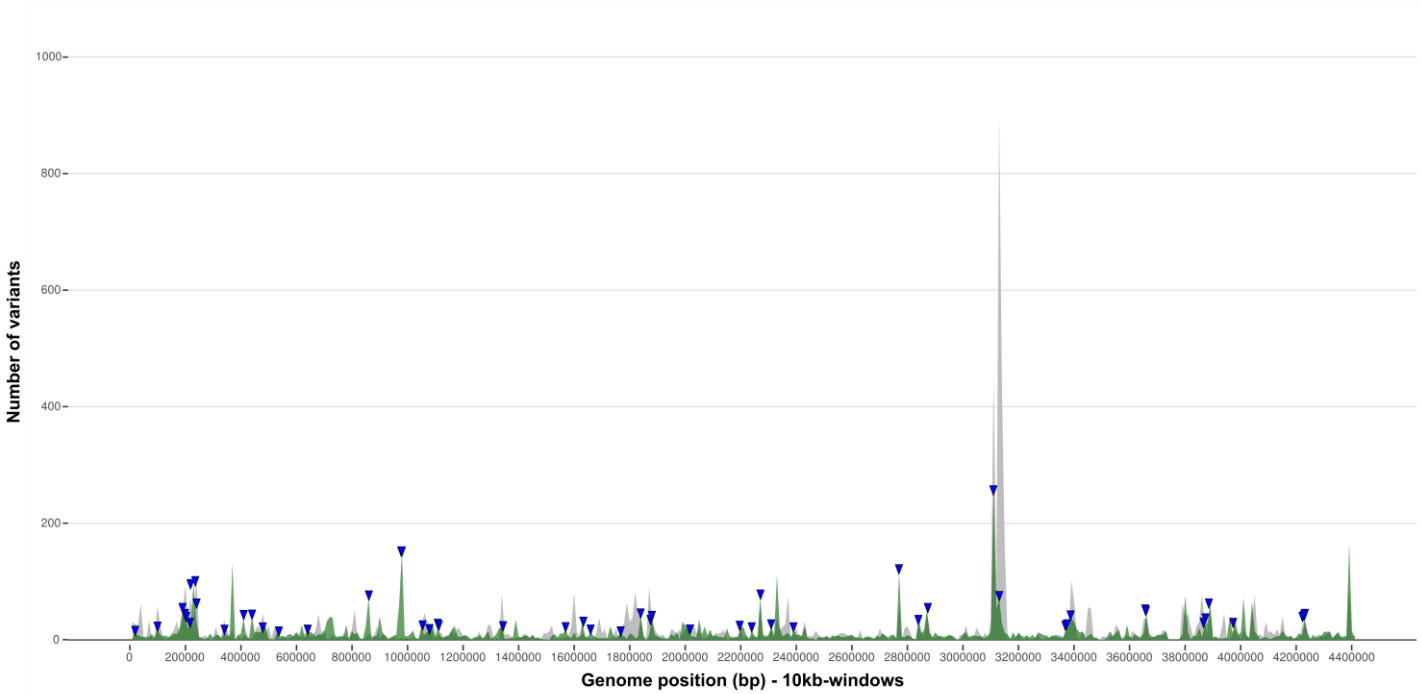


904 Figure 2

905

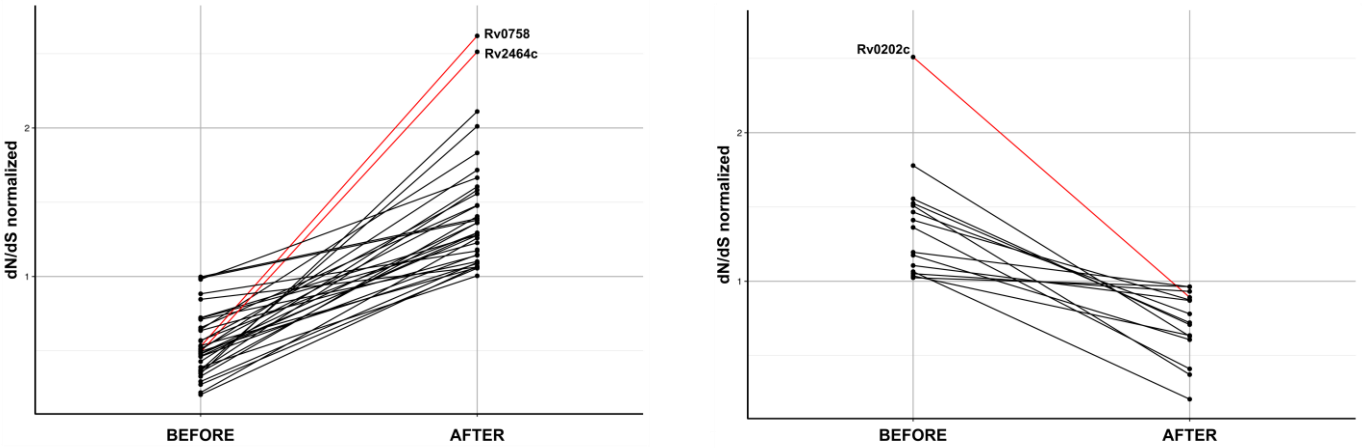


906 Figure 3



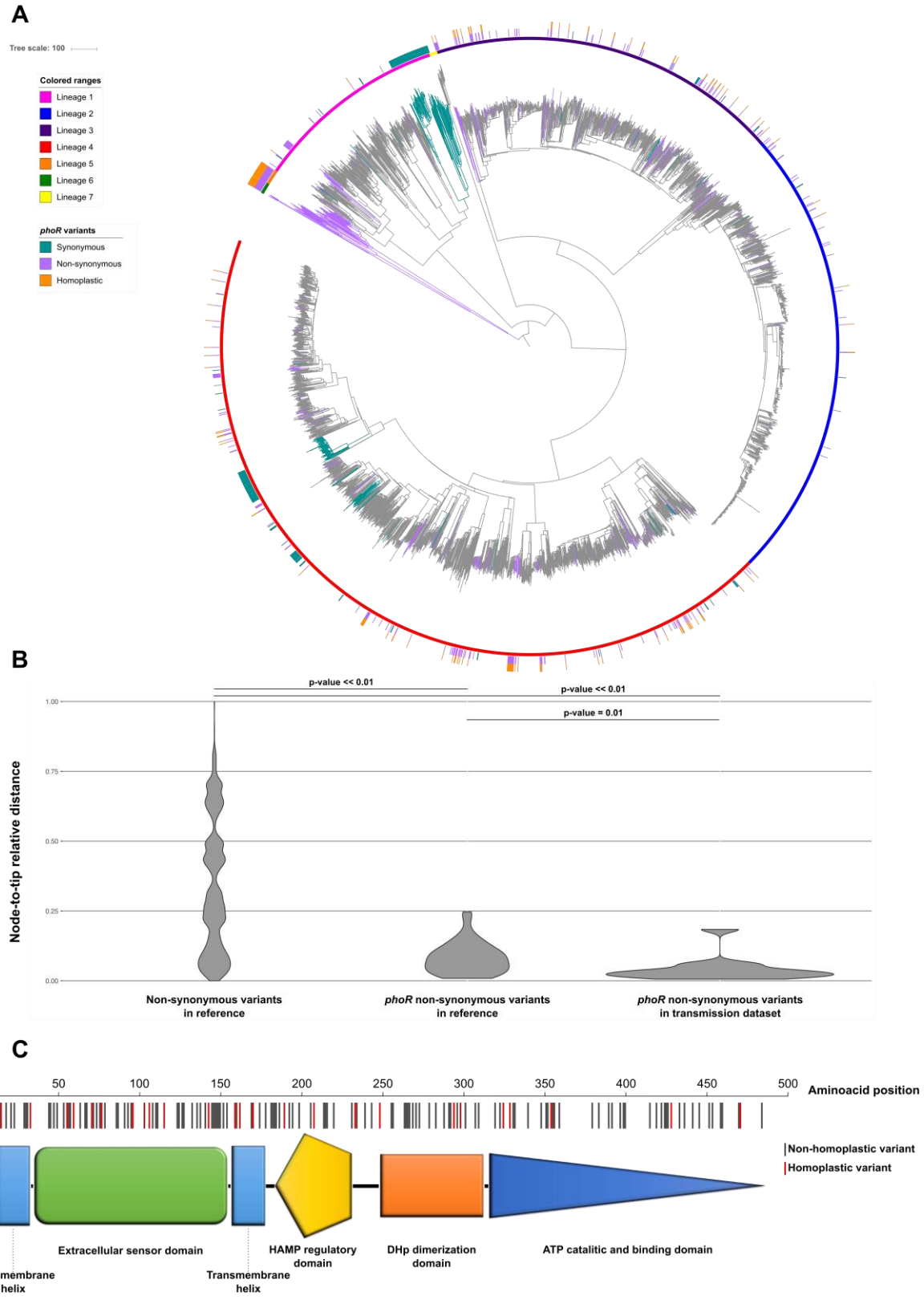
907 Figure 4

908



909 Figure 5

910



911 Figure 6