Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/5/6/eaaw3307/DC1

Supplementary Materials for

Genomic determinants of speciation and spread of the *Mycobacterium tuberculosis* complex

Á. Chiner-Oms, L. Sánchez-Busó, J. Corander, S. Gagneux, S. R. Harris, D. Young, F. González-Candelas, I. Comas*

*Corresponding author. Email: icomas@ibv.csic.es

Published 12 June 2019, *Sci. Adv.* **5**, eaaw3307 (2019)
DOI: 10.1126/sciadv.aaw3307

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/5/6/eaaw3307/DC1)

SUPPLEMENTARY TEXT

Definition of species based on genomic nucleotide identity

Currently, it is accepted that the average nucleotide identity (ANI) between two genomes can be used to define how likely the two genomes are part of the same bacterial species. Estimates have led to a threshold of 95% to define species (*61*). This threshold has been shown to correlate well with the standard 70% threshold of DNA/DNA-hybridization (DDH) used in microbial taxonomy during the last decades. Calculations of the average nucleotide identity between the different members of the MTBC showed two distinct patterns. On the one hand, the main causative agents of tuberculosis in humans (*M. tuberculosis* and *M. africanum*) and animals (*M. bovis*, *M. pinnipedi*, *M. antelope*, *M. microti*) have ANI values larger than 99% (our own data). Thus, at the genome level, the members of the MTBC form a single genomic species that is extremely monomorphic. The second pattern arises when MCAN is compared with the rest of the complex. In this case, an average identity of 98% (range 97.71%-99.30%) is observed and thus both groups are on the limit of what is generally accepted as the same species at the whole genome level (*7*).

Thus, while maintaining species names maybe clinically and ecologically relevant, it is clear that the different members of the MTBC can be considered the same genomic species. The close genomic similarity of the MCAN strains to the rest of the MTBC suggests that both groups have diverged very recently.

Recombinant fragments analysis

Gubbins identified 70 recombination events between the MTBC ancestor and MCAN strains. 5 of these fragments were filtered out due to a high accumulation of gaps (see Methods). A phylogeny was obtained for each of the 65 remaining fragments. A likelihood mapping analysis was performed for each fragment. Fragments 13 and 14 had not enough phylogenetic signal to resolve a reliable phylogeny. The variants found in these regions were present only in the MTBC ancestor branch, so recombination with other organism not present in our dataset is likely to have occurred. The topologies of the trees of the remaining fragments were compared with that of the tree derived from the non-recombinant alignment (whole genome alignment subtracting the recombinant regions). These analyses revealed significant incongruence for all the 63 fragments compared to the non-recombinant phylogeny (SH test; p-value < 0.05). To identify whether different regions share a common phylogenetic story we tested each fragment for congruence against all other trees and the non-recombinant phylogeny (fig. S2). The analysis identified consecutive fragments with similar topologies implying not only that the event involved similar donor/strains but also that they likely are part of a larger, unique event. This is the case for the genes in fragments 40, 41, 42. The genes involved are almost consecutive and only separated by PE/PPE genes that are not analyzed in this study. The fact that they share a common phylogenetic story indicates that they belong to a unique recombination event involving almost 28 Kb. A similar pattern can be observed for fragments 9-12 in which the fragments are not only consecutive in the genome but they also share a common phylogenetic story. In addition, events falling apart in the genome maybe also share a common phylogenetic story. For example, regions 6, 63, 62 are more than 3.8 Mb apart on the genome but they share the same phylogenetic topology. The genes involved are part of the same regulon, KtsR (*62*), suggesting that selection may have played a role in fixing those independent recombination events.

divSNP enriched genes analysis

Bacterial genomes are highly dynamic and different processes can lead to the genetic make-up of extant species. Consequently, not all the detected regions with a high number of divSNPs necessarily result from pure divergence by accumulation of substitutions. To ascertain the evolutionary origin of the 120 genes containing a high number of divergent variants, we checked whether the abnormal accumulation was due to: (i) horizontal gene transfer with other mycobacteria; (ii) recombination with MCAN strains that were not present in our dataset; or (iii) other evolutionary processes such as mutation combined with natural selection and/or genetic drift, thus genes that genuinely have accumulated divergence during the speciation process.

To check for horizontal gene transfer events, we downloaded from Refseq and GenBank a set of 155 complete genomes from distinct mycobacterial species. We looked for orthologues of the 120 genes accumulating divSNPs between the MTBC ancestor and the rest of the mycobacterial species. For each gene, we reconstructed a ML phylogenetic tree each of these phylogenies was compared to a ML reference built from the concatenated core mycobacterial gene set. Phylogenies for 53 of these genes placed MTBC within the MCAN clade, which is compatible with the accumulation of variants by mutation. Sixty-seven of the phylogenies were not topologically congruent with the reference tree. For all these genes, a BLAST search was performed against the NCBI database. In 54 cases the BLAST search give a better hit with *M. canettii* and in one case no hits were returned. The most plausible explanation for this alternative topology is that recombination with other MCAN strains not included in our data set had occurred. On the other hand, in 12 occurrences the BLAST search showed a best hit with other mycobacteria, more specifically with *M. chimaera*, *M. kansasii*, *M. sp. 3/86Rv*, and *M. shinjukuense*. Interestingly, the correlative genes from Rv2798c to Rv2803 followed this pattern, giving a better hit with *M. shinjukuense* than with MCAN. The *mazF9* and *mazE9* genes are in this region and were previously reported as a genomic island related with virulence and pathogenesis (*63*). Finally, one gene, Rv2804c, returned no results for the BLAST search.

Thus, a total of 53 genes in the MTBC ancestral genome were highly divergent with respect to MCAN due to substitution events. Most of the genome-wide identified divSNPs might result from genetic drift or hitchhiking events associated with selection on other loci. However, the accumulation in only 53 genes suggests that those regions might have played an important role during the process of niche differentiation.

SUPPLEMENTARY FIGURES

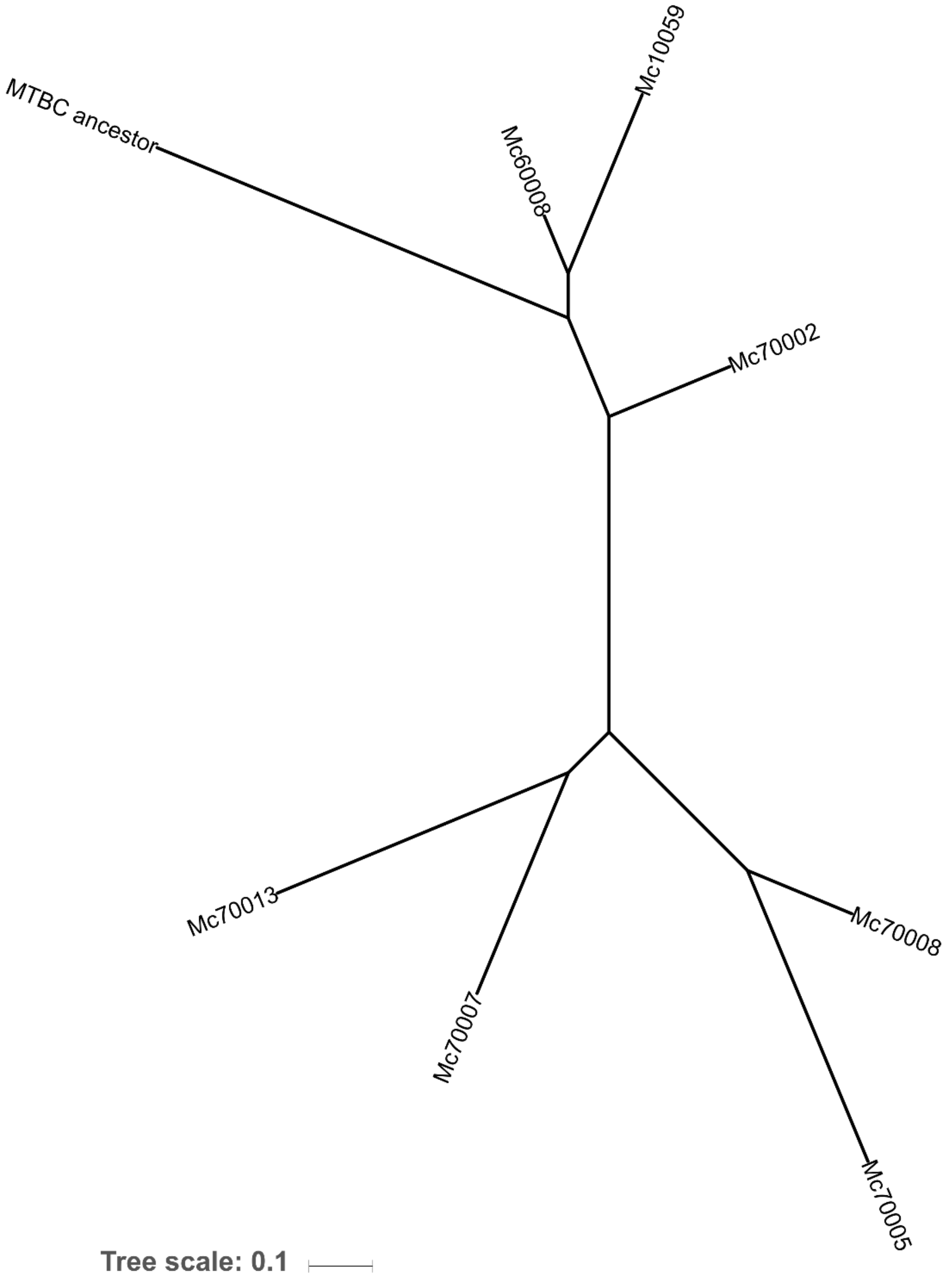

Fig. S1. Maximum likelihood phylogeny of the MCAN group, including the most likely inferred ancestor of MTBC.

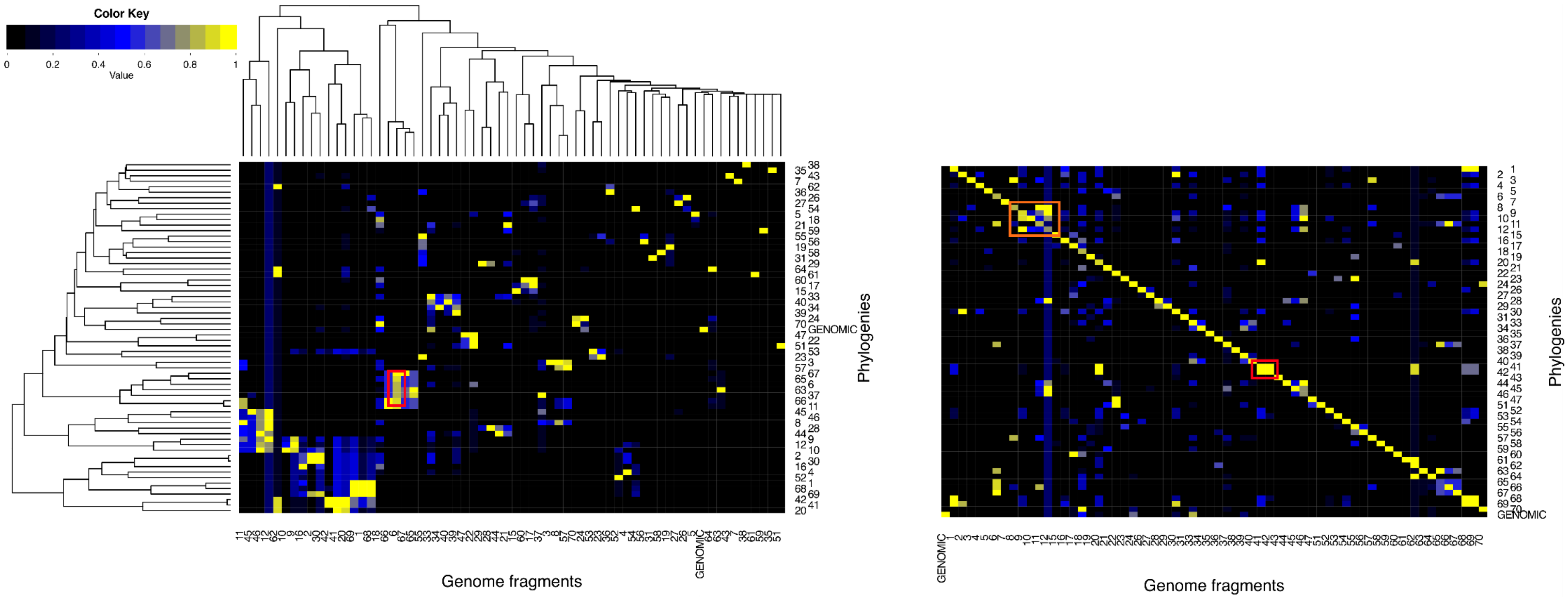

Fig. S2. Phylogenetic incongruence test. Each fragment alignment was compared against all recombinant fragments phylogenies and the non-recombinant genomic phylogenetic topology. Dark blue indicates strong incongruence and yellow no evidence to reject the topology. In the left side we show a double clustering of fragments and phylogenies in which each row corresponds to a phylogeny and each column to a fragment. In the right plot fragments are organized following their position in the genome. Fragments 13 and 14 were not included in the analysis as they did not have enough phylogenetic signal to reconstruct a reliable phylogeny. Fragments 40, 41, 42 are marked with a red square in the right panel. They share a common phylogenetic story and are correlatives suggesting that they belong to a unique recombination event. A similar pattern can be observed for fragments 9-12 (orange box, right panel). Regions 6, 63, 62 also share the same phylogenetic topology although not being consecutive (left panel, red square). The genes involved are part of the KtsR regulon.

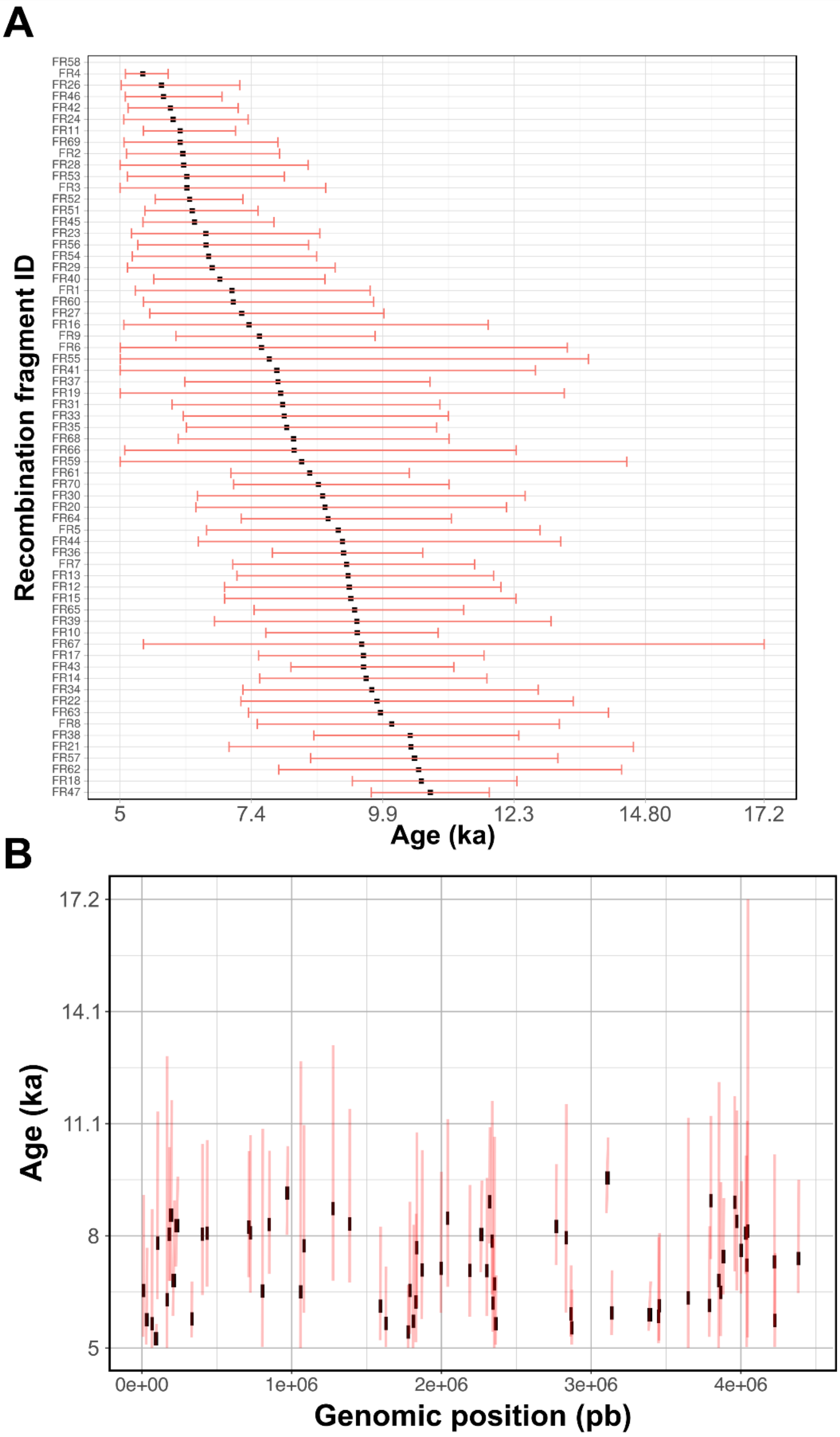

Fig. S3. Recombination fragments ages derived from BEAST. A) Ages of the recombinant fragments (x-axis), sorted by age. B) Ages of the recombination fragments (y-axis) sorted by its genomic position (x-axis). In both panels, the red error bars represent the 95% highest probability density (HPD).