





Comparative Genomic Analysis of *Mycobacteriaceae* Reveals Horizontal Gene Transfer-Mediated Evolution of the CRISPR-Cas System in the *Mycobacterium tuberculosis* Complex

 Anoop Singh,^a Mohita Gaur,^a Vishal Sharma,^a Palak Khanna,^a Ankur Bothra,^b Asani Bhaduri,^c Anupam Kumar Mondal,^b Debasis Dash,^b  Yogendra Singh,^a Richa Misra^{a,d}

^aDepartment of Zoology, University of Delhi, Delhi, India

^bCSIR-IGIB, Sukhdev Vihar, New Delhi, India

^cCluster Innovation Centre, University of Delhi, Delhi, India

^dDepartment of Zoology, Sri Venkateswara College, University of Delhi, Delhi, India

ABSTRACT Clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) genes are conserved genetic elements in many prokaryotes, including *Mycobacterium tuberculosis*, the causative agent of tuberculosis. Although knowledge of CRISPR locus variability has been utilized in *M. tuberculosis* strain genotyping, its evolutionary path in *Mycobacteriaceae* is not well understood. In this study, we have performed a comparative analysis of 141 mycobacterial genomes and identified the exclusive presence of the CRISPR-Cas type III-A system in *M. tuberculosis* complex (MTBC). Our global phylogenetic analysis of CRISPR repeats and Cas10 proteins offers evidence of horizontal gene transfer (HGT) of the CRISPR-Cas module in the last common ancestor of MTBC and *Mycobacterium canettii* from a *Streptococcus*-like environmental bacterium. Additionally, our results show that the variation of CRISPR-Cas organization in *M. tuberculosis* lineages, especially in the Beijing sublineage of lineage 2, is due to the transposition of insertion sequence IS6110. The direct repeat (DR) region of the CRISPR-Cas locus acts as a hot spot for IS6110 insertion. We show in *M. tuberculosis* H37Rv that the repeat at the 5' end of CRISPR1 of the forward strand is an atypical repeat made up partly of IS-terminal inverted repeat and partly CRISPR DR. By tracing an undetectable spacer sequence in the DR region, the two CRISPR loci could theoretically be joined to reconstruct the ancestral single CRISPR-Cas locus organization, as seen in *M. canettii*. This study retracing the evolutionary events of HGT and IS6110-driven genomic deletions helps us to better understand the strain-specific variations in *M. tuberculosis* lineages.

IMPORTANCE Comparative genomic analysis of prokaryotes has led to a better understanding of the biology of several pathogenic microorganisms. One such clinically important pathogen is *M. tuberculosis*, the leading cause of bacterial infection worldwide. Recent evidence on the functionality of the CRISPR-Cas system in *M. tuberculosis* has brought back focus on these conserved genetic elements, present in many prokaryotes. Our study advances understanding of mycobacterial CRISPR-Cas origin and its diversity among the different species. We provide phylogenetic evidence of acquisition of CRISPR-Cas type III-A in the last common ancestor shared between MTBC and *M. canettii*, by HGT-mediated events. The most likely source of HGT was an environmental *Firmicutes* bacterium. Genomic mapping of the CRISPR loci showed the IS6110 transposition-driven variations in *M. tuberculosis* strains. Thus, this study offers insights into events related to the evolution of CRISPR-Cas in *M. tuberculosis* lineages.

KEYWORDS CRISPR-Cas system, *Mycobacterium*, *Mycobacterium tuberculosis*, *Mycobacterium canettii*, horizontal gene transfer, IS6110, evolution, CRISPR-Cas system, comparative genomics, transposons

Citation Singh A, Gaur M, Sharma V, Khanna P, Bothra A, Bhaduri A, Mondal AK, Dash D, Singh Y, Misra R. 2021. Comparative genomic analysis of *Mycobacteriaceae* reveals horizontal gene transfer-mediated evolution of the CRISPR-Cas system in the *Mycobacterium tuberculosis* complex. mSystems 6:e00934-20. <https://doi.org/10.1128/mSystems.00934-20>.

Editor David W. Cleary, University of Southampton

Copyright © 2021 Singh et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Yogendra Singh, ysinghdu@gmail.com, or Richa Misra, richamisra@svc.ac.in.

Received 15 September 2020

Accepted 31 December 2020

Published 19 January 2021

The discovery of CRISPR-Cas (clustered regularly interspaced short palindromic repeats–CRISPR-associated proteins) system in *Mycobacterium tuberculosis* complex (MTBC) has been an important clinical finding for epidemiological studies (1–3). The CRISPR-Cas genomic locus has been frequently employed for strain genotyping (spoligotyping) in *M. tuberculosis*, the most dreaded infectious organism of MTBC (4). *M. tuberculosis* is the causative agent of tuberculosis (TB) and infects more than one-quarter of the world's population (5). MTBC was earlier grouped with environmental mycobacteria within a single genus, *Mycobacterium*. However, a new classification system proposed by Gupta et al. revisits the taxonomy and divides the mycobacterial species into five distinct clades based on the conserved signature indels and proteins (6). The *Mycobacterium* genus is now emended to encompass only the “Tuberculosis-Simiae” clade, which includes the group of slow-growing MTBC pathogens and nontuberculous mycobacteria (NTMs). MTBC comprises of human-adapted lineages (lineages 1 to 4 and lineage 7) of *M. tuberculosis sensu stricto*, *M. tuberculosis* variant africanum (lineages 5 and 6), and the recently discovered *M. tuberculosis* RW-TB008 (lineage 8), known for its early divergence from the rest of MTBC members (7). Besides these lineages, several animal-adapted forms, including *M. tuberculosis* variant bovis, *M. tuberculosis* variant caprae, *M. tuberculosis* variant microti, *M. tuberculosis* variant pinnipedii, *M. tuberculosis* variant origys, *M. tuberculosis* variant mungi, *M. tuberculosis* variant suricattae, *M. tuberculosis* variant dassie, and *M. tuberculosis* variant chimpanzee, are also included in MTBC (8). In addition to these classical members of MTBC, some studies occasionally include *Mycobacterium canettii* strains, also known as smooth tubercle bacilli (STBs), in MTBC based on nucleotide identity, but the present study includes only the classical members in “MTBC” (9). MTBC is said to evolve from an *M. canettii*-like ancestor that had an environmental reservoir (10). The other four novel genera are *Mycolicibacterium* gen. nov., *Mycolicibacter* gen. nov., *Mycolicibacillus* gen. nov., and *Mycobacteroides* gen. nov. corresponding to the “Fortuitum-Vaccae,” “Terra,” “Triviale,” and “Abscessus-Chelonae” clades, respectively (6).

In recent years, the CRISPR-Cas system has garnered a lot of attention in other prokaryotes such as *Streptococcus mutans*, *Pseudomonas aeruginosa*, etc., with mounting evidence on its physiological roles like gene regulation, virulence, evolutionary adaptation apart from the classical role in evasion and defense against phage predation (11, 12). Additionally, the recent evidence on the activity of *M. tuberculosis* CRISPR interference system in invader defense and potential for an active genome editing system (13, 14) has brought the focus back to the MTBC CRISPR-Cas system. The most defining feature of the CRISPR-Cas locus is the presence of a CRISPR array, comprising of short direct repeats (DR) separated by short variable DNA sequence “spacers” and flanked by *cas* genes (15). CRISPR array with no adjacent *cas* genes is known as orphan CRISPR (11, 15). Based on the effector module composition, CRISPR-Cas systems are classified into two classes with six types (types I to VI) and 33 subtypes (16). Considerable diversity of CRISPR-Cas systems exists among various prokaryotic species, possibly owing to the selective environmental and/or host pressure (17). The organization of MTBC CRISPR-Cas type III-A system is considered mostly conserved with two CRISPR loci and the *cas* gene cluster of nine genes: *cas6*, *cas10* (*csm1*), *csm2*, *csm3*, *csm4*, *csm5*, *csm6*, *cas1*, and *cas2* adjacent to the CRISPR1 locus (2, 13). However, some clinical isolates, particularly belonging to *M. tuberculosis* lineage 2 strains (Beijing sublineage), show deletion in the CRISPR-Cas locus (18). The Beijing sublineage represents one of the most virulent and drug-resistant clusters among *M. tuberculosis* isolates. This lineage also possesses a remarkably high proportion of MTBC-specific insertion sequences (IS), IS6110, which are widely used as an epidemiological marker for TB (19). IS6110 belongs to the IS3 family of IS, comprising of a 1,361-bp sequence with 28-bp terminal inverted repeats (IR) and 3-bp DR of target sequences at its extremities. The IS6110 sequence contains two partially overlapping open reading frames (ORFs), *orfA* and *orfB* encoding transposases (20). Expansion of IS is considered a key feature in the MTBC genome reduction process and is found at multiple sites in the genome, with one of the

insertion sites located in the CRISPR-Cas locus (19). However, the impact of IS6110 transposition on the evolution of the CRISPR-Cas system in *M. tuberculosis* has not been studied yet.

An earlier study reported similarities in the MTBC CRISPR-Cas type III-A system with some *M. canettii* strains but not with any NTMs, suggesting a horizontal gene transfer (HGT)-related acquisition (21); however, little is known about this evolutionary adaptation. Here, we performed a comprehensive comparative genomic analysis of 141 mycobacterial genomes, available at NCBI-RefSeq, to advance our understanding of the origin of the mycobacterial CRISPR-Cas system, its diversity, and interrelation among species with reference to the recent reclassification of *Mycobacteriaceae* genomes (6). Our results offer strong phylogenetic evidence of a HGT-mediated acquisition of CRISPR-Cas type III-A system in MTBC from an environmental *Firmicutes* as the likely source. Additionally, the analysis shows the influence of IS6110 transposition on the *M. tuberculosis* CRISPR-Cas system. Therefore, a deeper look into this genomic region gave fresh insights on the evolution of the CRISPR-Cas system in MTBC, especially in *M. tuberculosis* strains.

RESULTS AND DISCUSSION

Diversity of CRISPR-Cas systems in *Mycobacteriaceae* and potential targets of CRISPR spacers. To characterize CRISPR-Cas systems in *Mycobacteriaceae*, 141 genome sequences from NCBI-RefSeq were analyzed in the present study (see Table S1a in the supplemental material). The presence of true CRISPRs in the genome was assessed by the CRISPRCasFinder tool using default parameters (22). To discriminate spurious CRISPR-like elements from the true CRISPRs, only CRISPRs classified with evidence levels 3 and 4 were considered for further analyses. Based on the selection criteria, in total, 36 CRISPR loci/arrays containing 891 spacers in 19 genomes were selected. Among these predicted arrays, five CRISPR arrays are of evidence level 3, whereas the other 31 CRISPR arrays were assigned to level 4, projecting them as high-confidence CRISPR candidates (Table S1b).

Further, to correctly determine the presence of CRISPR-Cas systems in all 141 genomes, Cas proteins were identified using a combination of CRISPRCasFinder and HMMER 3 search against a collection of 395 Cas protein profiles obtained from a previous study (15). The results revealed the presence of true CRISPR-Cas system in 18 genomes (12 species) and an orphan CRISPR locus in one genome (*Mycobacterium avium*) (Fig. 1a). The distribution and diversity of the CRISPR-Cas systems in *Mycobacteriaceae* is shown along a phylogenetic tree generated using 16S rRNA sequences from the sequenced genomes (Fig. 1a). Our analysis revealed the presence of five monophyletic clades, in accordance with the new classification system by Gupta et al. (6) (Fig. 1a and Table S1a). We observed that CRISPR-Cas loci are predominantly present in the slow-growing monophyletic clade, Tuberculosis-Simiae. Among the 12 species that possess a true CRISPR-Cas system, eight species belong to the *Mycobacterium* genus. On the basis of the presence of signature *cas* genes, *cas3* and *cas10*, the CRISPR-Cas system was further classified, and out of the eight species of the *Mycobacterium* genus, six species were found to possess the type I system and the remaining two species, belonging to MTBC (seven members) and *M. canettii*, exclusively possess the type III-A system (Fig. 1a and Table S1a). The presence and organization of MTBC type III-A CRISPR-Cas system are as described in earlier reports (2, 13); however, the present study expands the search to 141 mycobacterial genomes, including 129 species compared to the 22 genomes, including 14 species from the earlier study (2). Among the other four genera, two species of *Mycolicibacterium* and one species each in *Mycolicibacter* and *Mycolicibacillus*, respectively, show the presence of a type I system. However, genomes of *Mycobacteroides* lack any CRISPR-Cas system (Fig. 1a). The detailed features of the CRISPR DR were predicted using the CRISPRmap program (23). The features included consensus sequences, secondary structures, conserved motifs, family, and superclass. These features are generally specific to a particular type/subtype of CRISPR-Cas system irrespective of the bacterial/archaeal species harboring them. Comparison of these conserved features among the 141 genomes revealed that all MTBC

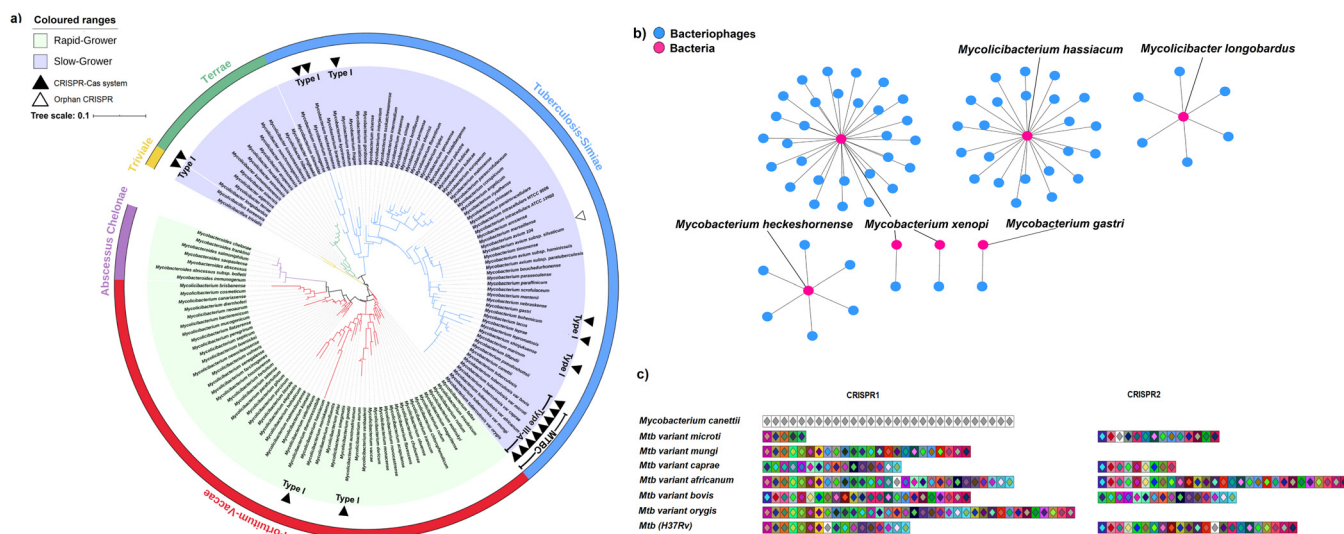


FIG 1 Analysis of CRISPR-Cas systems in *Mycobacteriaceae* reveals exclusive presence of type III-A system in MTBC and the interrelatedness of MTBC CRISPR spacers. (a) 16S rRNA gene-based phylogeny of *Mycobacteriaceae* shows the differential distribution of CRISPR-Cas systems. The different monophyletic clades/genera are represented by a color strip and colored branches. The presence of true CRISPR-Cas systems is illustrated as filled triangles, while the hollow triangle depicts an orphan CRISPR. The classification types are mentioned along with the triangles. The inner circle is color coded as light green and light purple, representing rapid-growing and slow-growing bacteria, respectively. (b) Mycobacteriophages as potential targets of CRISPR spacers. The phage-bacterium bipartite network derived from *Mycobacteriaceae* spacer sequences and their matches in the phage genome sequences showing mycobacteriophages targeted by the spacers of *Mycobacterium hassiacum*, *Mycobacter longobardus*, *Mycobacterium heckeshornense*, *Mycobacterium xenopi*, and *Mycobacterium gastri*. Pink nodes, bacteria; blue nodes, bacteriophages; edges, shared spacer-protospacer pair. (c) Comparative genomic analysis of MTBC and *M. canettii* CRISPR spacer content. The complete CRISPR loci are illustrated as two clusters of CRISPR1 and CRISPR2. Each color-coded box differentiates groups of spacer sequences. The unique spacers are depicted as gray boxes, and similar spacers are marked with the same colors across the data set. *Mtb*, *Mycobacterium tuberculosis*.

members and *M. canettii* (STB-A) share a conserved consensus CRISPR DR belonging to the same family and superclass (Table S2).

To look for potential targets of *Mycobacteriaceae* CRISPR spacers, we performed a command-line NCBI-BLASTN with 90% identity and 90% query coverage against the NCBI phage and plasmid databases, with the spacer sequences extracted from CRISPR arrays. Several putative protospacers homologous with the spacers in the phage and plasmid genomes were identified (Table S3a and S3b). We identified the targets in a few known mycobacteriophages such as Bxz1, Anaya, Iracema64, L5, etc., for spacer sequences of *Mycobacterium hassiacum*, *Mycobacter longobardus*, *Mycobacterium heckeshornense*, *Mycobacterium xenopi*, and *Mycobacterium gastri* (Fig. 1b and Table S3a). Although MTBC and *M. canettii* together possess ~53% of CRISPR arrays in the *Mycobacterium* genus, we did not find any significant similarity with any phage or plasmid database sequence. This could be due to the limited sequencing data available for uncharacterized mycobacteriophages. To overcome this issue and understand the possible origin of spacer sequences in MTBC, we looked for the conservation pattern of spacer sequences in MTBC. Most spacer sequences were conserved within or across CRISPR arrays in the MTBC members with a low proportion of unique spacers. However, we observed that *M. canettii* spacer sequences were unique and did not match with any of the MTBC members (Fig. 1c). CRISPR spacers are acquired in response to exposure to foreign invading genetic elements, which results in sequence-specific memory, protecting bacteria from future invasion (15). Therefore, lack of any shared spacer sequences may be due to the phylogenetic distance of *M. canettii* from MTBC members, which are mainly variants of *M. tuberculosis* species and so phylogenetically more related to each other (7, 21). Another possibility could be the inability of *M. tuberculosis* to incorporate new spacers, unlike *M. canettii* (2, 24).

Evidence of horizontal gene transfer of CRISPR-Cas type III-A system in MTBC.

Past findings suggest that MTBC members evolved into obligate pathogens by a bimodal evolutionary process of reductive evolution and selective genome expansion as

observed as genomic islands (GIs) in recipient bacterial genomes (29). The GIs have been considered direct evidence of the HGT of genes playing a crucial role in the evolution of bacterial genomes (30). To predict the occurrence of GIs in STB-A (*M. canettii* reference genome), we used IslandViewer 4, which integrates four methods, SIGI-HMM, IslandPath-DIMOB, IslandPick, and Islander, to most accurately analyze the GIs in the genome (31). Out of the 19 GIs predicted by IslandViewer 4, one of the GIs in STB-A was found to possess CRISPR-Cas type III-A system, supporting the hypothesis of HGT-based CRISPR-Cas acquisition. The predicted GI is around 32,729 bp in length, possessing CRISPR-Cas type III-A system along with mobility genes such as transposase and integrase, as shown in Fig. 2b. A comparative GI prediction in *M. tuberculosis* genome revealed a smaller sized (16,446-bp) GI, a probable result of genomic reduction, carrying a *cas* gene and other mobility genes (see Fig. S1 in the supplemental material).

To trace the origin of the STB-A GI, we used nucleotide BLAST but could not find a genus harboring such genomic loci. Therefore, to gain insight into the source of HGT, we analyzed the CRISPR repeats of MTBC and *M. canettii* using the CRISPRmap program. This program examines CRISPR repeat queries against the CRISPR repeat database to generate a clustering tree to determine the evolutionary relationships based on the conservation of CRISPR repeat sequence and similarities in their minimum free energy (MFE) secondary structures. The CRISPR repeat is a central regulatory element as it serves as the binding template for Cas proteins, and conservation of CRISPR DR RNA stem-loop structure is essential for interaction with effector complex, required for CRISPR biological function (23), as also recently shown in *M. tuberculosis* (13). Since MTBC members and *M. canettii* showed 100% conservation in their CRISPR DR, the consensus repeat was used as a query sequence in the program. Our results revealed clustering of the MTBC CRISPR cluster with *Streptococcus thermophilus* cluster (Fig. 3a, left panel, and Table S4b), suggesting a possible genetic exchange of CRISPR-Cas type III-A system between *Actinobacteria* and *Firmicutes*. Conservation in secondary structures of CRISPR RNA has been observed in diverse organisms that reflect conserved binding motifs and shared mechanisms of action of effector complex (23, 32). The consensus MFE structure of clustering tree members based on multiple sequence-structure alignment using LocARNA is shown in Fig. 3a, right middle panel, and a sequence logo of these aligned DR sequences, is shown in Fig. 3a, right bottom panel, respectively. The stem of the hairpin MFE structure shows the conserved compatible bases (highlighted in shades of green, blue, and yellow in Fig. 3a, right middle panel, and Fig. S2). The complete sequence-structure alignment file of these DR sequences from the cluster tree is shown in Fig. S2. Figure 3a, right top panel, shows an independent CRISPR RNA DR sequence alignment of *M. tuberculosis*, *M. canettii*, and *S. thermophilus* displaying conservation of compatible bases involved in RNA stem-loop formation that may interact with Cas endoribonucleases (13, 23).

Further, to independently validate the source, we analyzed the global collection of Cas10 protein sequences obtained from a previous study (15) (Table S4c). Cas10 was chosen for comparative analysis since it is encoded by a signature *cas* gene of CRISPR type III system which forms the major part of the effector complex and interacts with the CRISPR repeat (15). On the basis of Cas10 phylogeny, we found that the MTBC clade belonging to phylum *Actinobacteria* clustered to a clade consisting of members of *Streptococcus* spp. and *Lactobacillus ruminis* both belonging to the phylum *Firmicutes* with high bootstrap support (Fig. 3b). The evolutionary proximity of MTBC Cas10 with its corresponding homologs in *Streptococcus* spp. is in line with the observation of conserved CRISPR repeats and independently supports our finding of *Streptococcus* like *Firmicutes* bacterium to probably serve as the source for HGT-acquired CRISPR-Cas type III-A system in MTBC. A recent study has demonstrated that *M. tuberculosis* type III-A CRISPR system utilize Cas10-activated cyclic hexa-adenylate (cA6) signaling to degrade invading RNA to enhance immunity. The

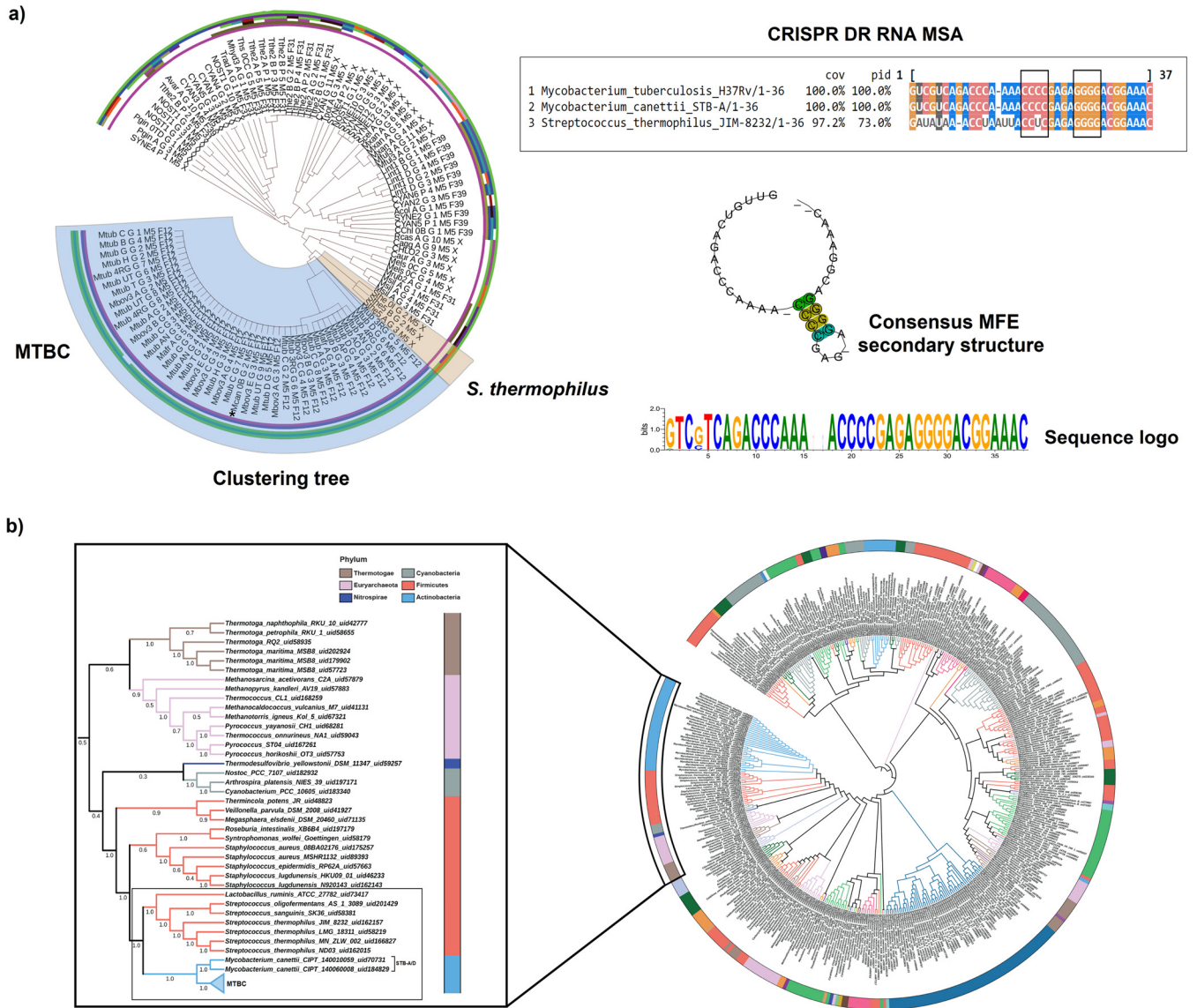


FIG 3 Phylogeny analysis based on CRISPR repeats and Cas10 predicts the acquisition source of the CRISPR-Cas system in the MTBC ancestor. (a) MTBC CRISPR repeat clusters with *S. thermophilus* cluster based on sequence and structural similarities. A hierarchical cluster tree was generated based on the multiple sequence-structure alignment of repeat sequences. The tree revealed a cluster of MTBC and *M. canettii* (highlighted in blue) along with *S. thermophilus* (highlighted in pale yellow). The right top panel shows the multiple sequence alignment of CRISPR DR RNA sequence of *M. tuberculosis*, *M. canettii*, and *S. thermophilus*. The conserved compatible bases, involved in RNA stem-loop formation, are shown inside rectangular boxes. The consensus MFE structure and sequence logo of aligned members from the cluster tree are shown in the right middle and right bottom panels, respectively. The conserved compatible bases involved in RNA stem formation are highlighted in similar color in the MFE structure. (b) Cas10 phylogeny shows evolutionary relatedness of MTBC with *Streptococcus* spp. The circular phylogenetic tree was generated from the global collection of Cas10 data obtained from the study of Makarova et al. (15). The bootstrap values are calculated from 1,000 replicates and are represented along the branches. The color strip represents different phyla. A magnified view of the area of interest is shown to the left of the circular tree highlighting the clustering of MTBC and *M. canettii* with *Streptococcus* species clade and *Lactobacillus ruminis* sharing a common ancestral node between them, indicating an HGT event.

phylum-wise comparison of characterized type III-A CRISPR systems revealed that the cA6-dependent signaling strategy is common between *Actinobacteria* members such as *M. tuberculosis* and *Firmicutes* members such as *S. thermophilus*, while other studied archaeal, *Deinococcus-Thermus*, and proteobacterial phyla utilize the cA4-modulated immunity (14). Although physical evidence of genetic exchange between *Mycobacterium* and *Streptococcus* is missing, interphylum HGT is a major evolutionary process and has been suggested to occur frequently for transfer of metabolic genes in many mesophilic bacteria (33). Thus, our results also strongly indicate that an HGT-driven acquisition of CRISPR-Cas type III-A system likely occurred in the last

common ancestor of *M. canettii* (STB-A) and MTBC, from a *Streptococcus*-like environmental bacterium.

Evolutionary role of IS6110 transposition in the diversification of CRISPR-Cas system in *M. tuberculosis* lineages. To delve deeper into the evolution of CRISPR diversification, we carried out a whole-genome core single nucleotide polymorphism (SNP)-based phylogenetic analysis of *M. tuberculosis* lineages. We obtained the reference data sets for *M. tuberculosis* lineages from four independent studies by Coll et al. (34), Phelan et al. (35), Nebenzahl-Guimaraes et al. (36), and Ngabonziza et al. (7) (Table S5a). These phylogenetic lineages were further validated using the Snippy program for SNP calling and SNP-IT for lineage classification (37). Next, CRISPR-Cas loci were predicted using CRISPRCasFinder in all lineages. The signature *cas10* gene of the CRISPR type III system was conserved in all lineages. Since Cas10-mediated cA6 generation has been shown to be critical for *M. tuberculosis* CRISPR defense (14), we also looked for the conservation of active sites in the Cas10 cyclase domain that generates the cyclic oligoadenylates. The active sites (GGDD) of Cas10 cyclase domain were conserved in all *M. tuberculosis* lineages (Fig. 4a and Fig. S3a). While the *in vitro* DNA cleavage activity of the other critical domain of Cas10, HD nuclease domain, is still challenged (14), our results showed the conservation of active site residues (HD) in all lineages (Fig. 4a and Fig. S3b). Our results confirmed the absence of *csm4* (truncated), *csm5*, *csm6*, *cas1*, and *cas2* in a strain cluster belonging to members of Beijing lineage (sublineage of lineage 2) (Fig. 4a), consistent with previous reports (18, 38). Deletions in this region suggest compromised genome defense; nonetheless, few studies have reported mutations in *cas1* and *cas2* to affect drug resistance in bacteria and the ability to accumulate DNA mutations without affecting survival (39, 40). Therefore, it has been proposed that these deletions prove advantageous and probably better adapt the Beijing strain to infect humans and spread faster, despite compromising on its phage immunity (41). A frequent cause of genomic deletions in bacteria is related to the movement of mobile genetic elements and although insertion of IS6110 in *M. tuberculosis* CRISPR-Cas locus has been observed (19), its impact on the CRISPR-Cas system is poorly understood. Apart from implication of IS6110 transposition in host adaptation (19), studying differential insertion sites in diverse strains also has potential use as molecular markers for identifying strain-specific outbreaks, as seen with the Central Asia outbreak (CAO) clade, where a specific IS6110 insertion was detected unique to this major epidemic clade of Beijing genotype (42).

Pairwise alignment of CRISPR-Cas loci was carried out for all *M. tuberculosis* lineages and *M. canettii* using BLASTN with *M. tuberculosis* H37Rv as the reference genome (Fig. 4b). While the typical organization of the CRISPR-Cas locus with a single copy of IS6110 inserted in CRISPR DR, as present in the reference genome *M. tuberculosis* H37Rv (GenBank accession number [NC_000962](https://www.ncbi.nlm.nih.gov/nuccore/NC_000962)), is common to most other strains belonging to different lineages; distinct genomic variations in locus organization were observed in some isolates (Fig. 4b). As seen in Fig. 4a and consistent with previous findings (18), deletions in the CRISPR-Cas locus in the Beijing sublineage (lineage 2) was observed (Fig. 4b), which seems to be mediated by the transposition of IS6110 in that genomic location. The evidence of active transposition is strengthened from the observation of other lineages; for example, in lineage 1, two isolates show the typical *M. tuberculosis* organization, while isolates WBB1007_LQ1975, WBB1008_SL1975, and WBB1009_SL1875 possess two copies of IS6110 in reverse orientation, with one copy inserted just outside the CRISPR locus (Fig. 4b). Presence of two IS6110 copies between three CRISPR arrays was noted in [CP002992](https://www.ncbi.nlm.nih.gov/nuccore/CP002992) (lineage 4), [CP001664](https://www.ncbi.nlm.nih.gov/nuccore/CP001664) (lineage 4), and WBB1454_IB091-1 (lineage 5). [CP003233](https://www.ncbi.nlm.nih.gov/nuccore/CP003233) (lineage 4) shows presence of only one IS6110 copy and three CRISPR arrays. Since there is no trace of another IS6110 copy between the proximal two arrays of [CP003233](https://www.ncbi.nlm.nih.gov/nuccore/CP003233) separated by a short genomic region (118 bp), we are unable to fully attribute IS6110 as the cause of this disruption. Transposition of IS6110 in the *cas* gene region of [AP012340](https://www.ncbi.nlm.nih.gov/nuccore/AP012340) (lineage 4) leaves a partial copy of IS6110 with a single ORF encoding the transposase (apart from two complete copies) and results in disruption of *csm5* and truncation of *csm6* (Fig. 4b and a). A similar

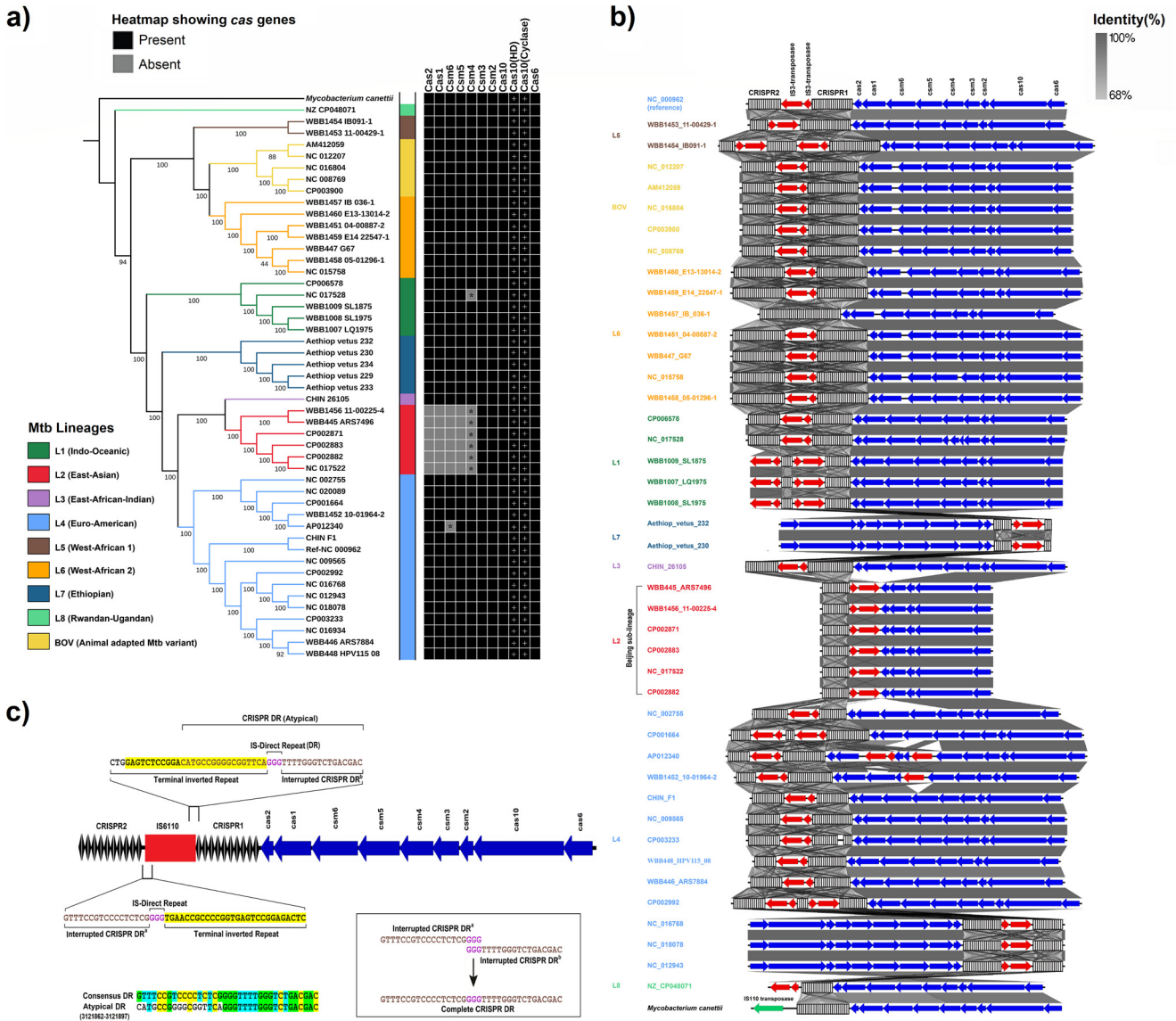


FIG 4 Comparison of genetic organization of the CRISPR-Cas region in *M. tuberculosis* lineages shows IS6110 (IS3 family transposon) element-derived interruption. (a) Phylogeny of *M. tuberculosis* lineages showing differential *cas* gene distribution. Maximum likelihood phylogeny of 43 *M. tuberculosis* strains belonging to lineages 1 to 8 and five animal-adapted strains of *M. tuberculosis* variant bovis rooted with *M. canettii* as an outgroup was inferred from core genome SNPs. The different clades are color coded according to previously defined lineages (7). The comparison shows lineage-specific absence of certain *cas* genes in lineage 2 members (gray boxes), while truncation of gene is indicated with an asterisk inside the box. The plus sign indicates the presence of active sites of Cas10 HD and cyclase domain. (b) Pattern of IS6110 transposition within the CRISPR-Cas locus in *M. tuberculosis* lineages. The CRISPR-Cas genomic clusters based upon BLASTN pairwise alignments are visualized as linear arrow comparison plots. The coding regions are represented by arrows, and CRISPR loci are shown as striated rectangular boxes. Blue arrows indicate *cas* genes, and red and green arrows represent different types of transposases, respectively. The arrow orientation represents forward/reverse positioning on the genome. The gray vertical blocks between sequences indicate regions of percent identity, shaded according to BLASTN results, and the degree of sequence identity is indicated by gray color. Sequence names are color depicted according to previously defined lineages, as in panel a. Ancestral arrangement of the CRISPR-Cas system in the genome of *M. canettii* is represented at the bottom. (c) Schematic representation of *M. tuberculosis* CRISPR-Cas locus highlighting the overlap region between CRISPR1 and IS6110. The coding regions are shown as arrows and CRISPR loci as diamonds. Blue arrows indicate *cas* genes, and the red rectangle marks the relative position of IS6110. The flanking region of IS6110 is magnified to show interruption of CRISPR repeat into two halves due to IS6110 transposition. The CRISPR1 locus repeat at position 3121862 to 3121897 is an atypical repeat with overlap between a part of IS-terminal inverted repeat (IR) and half CRISPR direct repeat (DR). IR is highlighted in yellow. The CRISPR DR sequence is in brown text. The atypical overlapping region highlighted in yellow with brown text, and the DR site of IS6110 is denoted by the text in purple. The alignment of this atypical CRISPR repeat with consensus repeat is shown at the bottom on left. Complete CRISPR DR retrieved by overlapping the interrupted CRISPR DR is shown inside the box.

organization is seen in WBB1452_10-01964-2 (lineage 4), but *csm6* is conserved. The most unique observation was seen in a lineage 6 member, WBB1457_IB_036-1, that possesses a single long CRISPR array with no nearby IS6110 insertion similar to *M. canettii*. Similarly, lineage 8 member NZ_CP048071 also has a single CRISPR array but a

nearby copy of IS6110 (Fig. 4b). These results suggest active transposition of IS6110 in *M. tuberculosis* lineages has impacted evolution of CRISPR locus genomic region.

Therefore, we analyzed the classic CRISPR locus of *M. tuberculosis* H37Rv to better understand the mechanism of transposition. We observed that the IS6110 IR overlaps with the DR of the CRISPR1 locus of the *M. tuberculosis* H37Rv reference genome (Fig. 4c). On the basis of genomic analysis, we propose that this overlapping repeat is a result of interruption of the CRISPR DR region by IS6110. The repeat at positions 3121862 to 3121897 is atypical, a combination of part of the 3' IR of IS6110 and a part of 5' DR of CRISPR1. As expected, the other half of the CRISPR DR is present at the 5' end of the IS6110 on the forward strand (Fig. 4c). The 3-bp DR duplication generated by IS6110 flanking to the point of insertion (43) is "GGG" in strain H37Rv (highlighted in purple in Fig. 4c). Next, we reconstructed the ancestral *M. tuberculosis* single array CRISPR locus by joining the two arrays seen in the present-day strain by tracing a previously undetected spacer sequence between the two loci. This spacer is not detected by commonly used identification tools such as CRISPRCasFinder and CRISPRDetect due to interruption of the repeat flanking the spacer sequence (Fig. S4). In order to validate this assumption, we performed BLASTN with the retrieved spacer sequence against all MTBC spacer sequences detected by CRISPRCasFinder. The results showed significant hit with 100% identity and coverage with spacer sequences from *M. tuberculosis* variant africanum, *M. tuberculosis* variant bovis, *M. tuberculosis* variant orygis, and *M. tuberculosis* variant caprae. This suggests yet again that the particular spacer sequence is MTBC specific and can theoretically join the two CRISPR arrays into a single continuous array locus, which was disrupted by IS6110 during the course of evolution (Fig. S5 and Table S5b). Such unique spacer sequences, which have remained undetected by common tools due to interruption by IS6110, can have potential value in strain identification.

On the basis of the results, we conclude that *M. tuberculosis* ancestor must have possessed only one long CRISPR array as seen in *M. canettii* (Fig. 4b). This array has since been interrupted by transposition of IS6110, which led to the formation of two CRISPR loci separated by IS6110, as seen in most present-day *M. tuberculosis* strains. Thus, our results show that the DR of CRISPR region acts as a hot spot for the insertion of IS6110 during transposition, as previously suggested (44), and generates a "GGG" duplication at the *M. tuberculosis* H37Rv strain locus. Such active transpositions have impacted the evolution of CRISPR locus leading to genomic variations such as gene deletions and recombination. These variations may lead to emergence of new pathogenic properties, as exemplified by the Beijing lineage, which despite genomic losses utilizes a selective advantage for infection and has emerged as a better-adapted pathogen (45).

Conclusion. Genomic comparisons of *M. tuberculosis* with related bacteria offer valuable insights into the evolutionary history and emergence of pathogenic strains. The present study, a comprehensive comparative genomic analysis of 141 mycobacterial genomes, showed the exclusive presence of the CRISPR-Cas type III-A system in MTBC. Further analysis revealed that CRISPR-Cas type III-A system was likely acquired in the last common ancestor of STB-A and MTBC by a HGT-driven acquisition. The plausible source seems to be a *Streptococcus*-like environmental bacterium. Our work reveals that although the genomic organization of CRISPR-Cas locus is conserved in *M. tuberculosis* lineages, certain specific strains show considerable deletions. These deletions, best exemplified in Beijing sublineage members, are driven by active transposition of IS6110, which utilize the DR of CRISPR region for insertion. This work delineates the evolutionary events such as HGT and IS6110-driven genomic variations in mycobacteria to better comprehend the epidemiology of *M. tuberculosis* lineages.

MATERIALS AND METHODS

Genome sequences and CRISPR-Cas classification. All available (141) genome sequences of *Mycobacteriaceae* covering five genera were downloaded from NCBI-RefSeq website on 24 August 2019. The genomic data and annotations were obtained from NCBI-FTP (<ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/>). CRISPR loci were predicted using CRISPRCasFinder (<https://crisprcas.i2bc.paris-saclay.fr/>) with default parameters (22). CRISPRCasFinder comprises of a rating system based on several features. Short candidate arrays made up of one to three spacers often do not correspond to real CRISPRs

and are therefore given the lowest evidence level, level 1. Evidence levels 2 to 4 are attributed on the basis of the combined degrees of similarity of repeats and spacers. Arrays with evidence level 1 or 2 indicate potentially false-positive results and were not considered for our analysis. Additionally, all predicted loci were manually checked, and those located in coding regions were discarded. CRISPRmap (v1.3.0-2013) (23) was used to provide conserved motifs, family, and superclass based on structural and sequence similarities.

Cas proteins were identified using CRISPRCasFinder and HMMER 3 (46) against a collection of 395 Cas protein profiles obtained from a previous study (15). *cas* genes were annotated and naming of *cas* genes, and their classification into types and subtypes was carried out as described by Makarova et al. (15). Cas proteins were also cross-verified from the respective NCBI genome annotations.

16S rRNA gene sequence-based phylogenetic tree construction. 16S rRNA gene-based comparative phylogenetic analysis was performed for all downloaded genomes. 16S rRNA sequences were obtained from the NCBI genome annotation files of the downloaded genomes (downloaded on 24 August 2019; see Table S1a in the supplemental material). To create the tree, multiple sequence alignment of the 16S rRNA gene sequences corresponding to the 141 gene copies were performed by MAFFT v7 using default parameters (47). The alignment was used to compute a maximum likelihood phylogenetic tree using the GTR+G model in RAxML-NG v1.0.1 (48), and branch support was computed with 1,000 bootstrap replicates. The tree was midpoint rooted and visualized by iTOL software (<https://itol.embl.de/>).

MTBC CRISPR spacer target identification and relatedness. All available complete genomes of phages and plasmids were downloaded from the NCBI ftp server on 7 July 2020. Redundant genomes were removed, and a database was constructed using the NCBI-BLAST+ 2.9.0 command-line tool. BLASTN was performed for all CRISPRCasFinder-identified spacers against NCBI phage and plasmid databases, with 90% identity and query coverage. Significant matches were summarized in bipartite networks with edges between spacers and their targets and visualized using the Cytoscape software (49). Edges between network nodes were assigned when a protospacer matching a spacer in a given host was identified in a phage.

CRISPRStudio (50) was used to visualize the CRISPR locus using default parameters; it compares spacer sequences present in a CRISPR array and then clusters them based on sequence similarities. To identify unique spacers at default settings, it considers spacer pairs with ≤ 2 mismatches as identical. CRISPRStudio requires gff3 file format as an input generated by CRISPRDetect. CRISPR loci common to both CRISPRCasFinder and CRISPRDetect (51) were used for visualization by CRISPRStudio.

STB phylogeny and sequence analysis. All available STB strains were downloaded from NCBI-RefSeq website on 24 August 2019. Core genome alignment was done using Roary v3.13.0 for all the genomes (52). Alignment also included genomes of *M. tuberculosis* H37RV (GenBank accession number [NC_000962](#)), *M. tuberculosis* RW-TB008 (accession number [NZ_CP048071](#)), *M. tuberculosis* variant bovis (accession number [NC_016804.1](#)), and NTMs (*M. kansasii* [accession number [NZ_CP019888.1](#)] and *M. marinum* [accession number [NZ_HG917972](#)]). A maximum likelihood phylogenetic tree was constructed using the GTR+G model in RAxML-NG v1.0.1, and branch support was computed with 1,000 bootstrap replicates, using *M. marinum* as an outgroup. The tree was visualized using iTOL.

Genomic island prediction. Genomic islands (GIs) were identified and analyzed by IslandViewer 4 (31), which integrates four different and accurate GI predictor tools: IslandPath-DIMOB, SIGI-HMM, IslandPick, and Islander. IslandViewer was used to analyze GIs in STB-A (GenBank accession number [NC_015848.1](#)) and *M. tuberculosis* genome (accession number [NC_018143.2](#)).

Phylogenetic analysis of CRISPR repeats and Cas10 families. Consensus CRISPR repeat sequence of MTBC and *M. canettii* genomes was used as a query sequence against CRISPRmap repeat database (v1.3.0-2013). CRISPRmap program constructs a hierarchical cluster tree based on multiple sequence-structure alignment of repeat sequences and the minimum free energy (MFE) structures generated by LocARNA to find relatedness (53) and provide a consensus MFE structure. A separate alignment of CRISPR DR RNA sequence of *M. tuberculosis*, *M. canettii*, and *S. thermophilus* was performed by T-COFFEE at default parameters (<https://www.ebi.ac.uk/Tools/msa/tcoffee/>). The sequence logo of CRISPR repeats of aligned family members was obtained using WebLogo (54).

To find the closest relative of *M. tuberculosis* Cas10, the global collection of Cas10 sequence data described in a study by Makarova et al. (15), was downloaded from the Batch Entrez website (<https://www.ncbi.nlm.nih.gov/sites/batchentrez>). Multiple sequence alignments were performed to align closely related sequences by MAFFT v7, and it was also used to merge these alignments. The phylogenetic tree was reconstructed using LG+G model in the IQTREE v1.6.2 (55). The same program was used for 1,000 bootstrap calculation. Phylum level classification was done manually with the help of the NCBI-Taxonomy browser (<https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi>) for tree annotation. The tree was midpoint rooted and visualized using iTOL.

Whole-genome core SNP phylogeny of *M. tuberculosis* lineages and CRISPR-Cas loci prediction. We downloaded *M. tuberculosis* lineages (lineages 1 to 8 and animal-adapted lineages) as reference data sets comprising of 48 genomes from three different studies, namely, 24 genomes from the study of Coll et al. (34), 18 genome sequences from Phelan et al. (35), five genome sequences from Nebenzahl-Guimaraes et al. (36) and one genome sequence from Ngabonziza et al. (7) (Table S5a). *M. canettii* (GenBank accession number [NC_015848.1](#)) was used as an outgroup. Core genome SNP calling was done using Snippy v4.4 (<https://github.com/tseemann/snippy>) using *M. tuberculosis* H37Rv as the reference genome. SNP-IT v1.1 (37) program was also used to predict and confirm *M. tuberculosis* lineages. Core genome SNP alignment file was used to compute a maximum likelihood phylogenetic tree using the GTR+G+ASC_LEWIS model in RAxML-NG v1.0.1 (48). Branch support was computed with 1,000

bootstrap replicates. The tree was visualized using the iTOL server. The CRISPR-Cas locus was predicted using CRISPRCasFinder. The presence of Cas10 HD and cyclase domains was predicted using Scanprosite (<https://prosite.expasy.org/scanprosite/>) and Cas10 HD and cyclase domains were aligned using MAFFT v7.

IS element annotation within the CRISPR-Cas region. CRISPR-Cas loci were predicted using the CRISPRCasFinder, as described earlier. The loci were extracted from the genomes using in-house perl script. Two sequences (NC_020089 and NC_016934) were removed due to assembly gaps in the CRISPR repeat region. IS6110 transposase was annotated using Prokka v1.14 (56) in the CRISPR-Cas locus genome segment and aligned using NCBI-BLASTN using default parameters. Easyfig v2.2.3 was used to map and compare the CRISPR-Cas loci among *M. tuberculosis* lineage genomes. *M. tuberculosis* genome (accession NC_000962) was considered the reference genome, and *M. canettii* (accession NC_015848) was considered the ancestral genome. For identification of IS6110 insertion at CRISPR loci, *M. tuberculosis* (accession number NC_000962) complete CRISPR-Cas locus sequence was extracted using in-house perl script. Overlapping region between IS6110 and CRISPR repeats was analyzed and represented with CRISPR-Cas locus visualization generated using Easyfig v2.2.3 (57). We manually extracted genome sequence present between CRISPR2 and IS6110 to identify a potential spacer sequence, as CRISPRCasFinder and CRISPRDetect were unable to detect it due to the split in the flanking CRISPR repeat. Further, BLASTN was performed with the retrieved sequence as a query against all MTBC spacer sequences using default parameters.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, TIF file, 2 MB.

FIG S2, TIF file, 2.7 MB.

FIG S3, PDF file, 2.1 MB.

FIG S4, TIF file, 0.4 MB.

FIG S5, TIF file, 2.3 MB.

TABLE S1, XLSX file, 0.04 MB.

TABLE S2, DOCX file, 0.3 MB.

TABLE S3, XLSX file, 0.02 MB.

TABLE S4, XLSX file, 0.03 MB.

TABLE S5, XLSX file, 0.01 MB.

ACKNOWLEDGMENTS

We thank the computational facility at CSIR-IGIB for providing the support for high-performance computing. We also thank Rajan for helping in data visualization. A.S. was supported by UGC senior research fellowship. M.G. and V.S. were supported by CSIR senior research fellowship.

This work was funded by the J. C. Bose research grant (SB/S2/JCB-012/2015) and research support from the University of Delhi. None of the funding organizations had a role in the study design, data analysis, or write up.

We declare that we have no conflict of interests.

R.M. and Y.S. supervised the study. A.S., A.K.M., D.D., and R.M. conceived the study design. Y.S. provided the financial support. A.S. did the data analysis with help from P.K. A.S., V.S., and P.K. generated the figures. R.M., A.S., and M.G. drafted the manuscript. A.K.M. and A. Bo. provided critical input to finalize the draft. A.Bh., D.D., and Y.S. provided input in the discussion. All authors contributed to the development of the manuscript and approved the final version.

REFERENCES

1. Driscoll JR. 2009. Spoligotyping for molecular epidemiology of the *Mycobacterium tuberculosis* complex. *Methods Mol Biol* 551:117–128. https://doi.org/10.1007/978-1-60327-999-4_10.
2. He L, Fan X, Xie J. 2012. Comparative genomic structures of *Mycobacterium* CRISPR-Cas. *J Cell Biochem* 113:2464–2473. <https://doi.org/10.1002/jcb.24121>.
3. Ishino Y, Krupovic M, Forterre P. 2018. History of CRISPR-Cas from encounter with a mysterious repeated sequence to genome editing technology. *J Bacteriol* 200:e00580-17. <https://doi.org/10.1128/JB.00580-17>.
4. Shariat N, Dudley EG. 2014. CRISPRs: molecular signatures used for pathogen subtyping. *Appl Environ Microbiol* 80:430–439. <https://doi.org/10.1128/AEM.02790-13>.
5. World Health Organization. 2020. Global tuberculosis report 2020. World Health Organization, Geneva, Switzerland. <https://www.who.int/publications-detail-redirect/9789240013131>.
6. Gupta RS, Lo B, Son J. 2018. Phylogenomics and comparative genomic studies robustly support division of the genus *Mycobacterium* into an emended genus *Mycobacterium* and four novel genera. *Front Microbiol* 9:67. <https://doi.org/10.3389/fmicb.2018.00067>.
7. Ngabonziza JCS, Loiseau C, Marceau M, Jouet A, Menardo F, Tzfadia O, Antoine R, Niyigenga EB, Mulders W, Fissette K, Diels M, Gaudin C, Duthoy S, Ssengooba W, Andre E, Kaswa MK, Habimana YM, Brites D, Affolabi D, Mazarati JB, de Jong BC, Rigouts L, Gagneux S, Meehan CJ, Supply P. 2020. A sister lineage of the *Mycobacterium tuberculosis* complex

- discovered in the African Great Lakes region. *Nat Commun* 11:2917. <https://doi.org/10.1038/s41467-020-16626-6>.
8. Riojas MA, McGough KJ, Rider-Riojas CJ, Rastogi N, Hazbon MH. 2018. Phylogenomic analysis of the species of the *Mycobacterium tuberculosis* complex demonstrates that *Mycobacterium africanum*, *Mycobacterium bovis*, *Mycobacterium caprae*, *Mycobacterium microti* and *Mycobacterium pinnipedii* are later heterotypic synonyms of *Mycobacterium tuberculosis*. *Int J Syst Evol Microbiol* 68:324–332. <https://doi.org/10.1099/ijsem.0.002507>.
 9. Brites D, Loiseau C, Menardo F, Borrell S, Boniotti MB, Warren R, Dippenaar A, Parsons SDC, Beisel C, Behr MA, Fyfe JA, Coscolla M, Gagneux S. 2018. A new phylogenetic framework for the animal-adapted *Mycobacterium tuberculosis* complex. *Front Microbiol* 9:2820. <https://doi.org/10.3389/fmicb.2018.02820>.
 10. Aboubaker Osman D, Bouzid F, Canaan S, Drancourt M. 2015. Smooth tubercle bacilli: neglected opportunistic tropical pathogens. *Front Public Health* 3:283. <https://doi.org/10.3389/fpubh.2015.00283>.
 11. Westra ER, Buckling A, Fineran PC. 2014. CRISPR-Cas systems: beyond adaptive immunity. *Nat Rev Microbiol* 12:317–326. <https://doi.org/10.1038/nrmicro3241>.
 12. Singh A, Gaur M, Misra R. 2018. Understanding the connect of quorum sensing and CRISPR-Cas system: potential role in biotechnological applications, p 231–247. In Kalia V (ed), *Quorum sensing and its biotechnological applications*. Springer, Singapore, Singapore. https://doi.org/10.1007/978-981-13-0848-2_15.
 13. Wei W, Zhang S, Fleming J, Chen Y, Li Z, Fan S, Liu Y, Wang W, Wang T, Liu Y, Ren B, Wang M, Jiao J, Chen Y, Zhou Y, Zhou Y, Gu S, Zhang X, Wan L, Chen T, Zhou L, Chen Y, Zhang XE, Li C, Zhang H, Bi L. 2019. *Mycobacterium tuberculosis* type III-A CRISPR/Cas system crRNA and its maturation have atypical features. *FASEB J* 33:1496–1509. <https://doi.org/10.1096/fj.201800557RR>.
 14. Gruschow S, Athukoralage JS, Graham S, Hoogbeem T, White MF. 2019. Cyclic oligoadenylate signalling mediates *Mycobacterium tuberculosis* CRISPR defence. *Nucleic Acids Res* 47:9259–9270. <https://doi.org/10.1093/nar/gkz676>.
 15. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJ, Charpentier E, Haft DH, Horvath P, Moineau S, Mojica FJ, Terns RM, Terns MP, White MF, Yakunin AF, Garret RA, van der Oost J, Backofen R, Koonin EV. 2015. An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* 13:722–736. <https://doi.org/10.1038/nrmicro3569>.
 16. Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, Brouns SJJ, Charpentier E, Cheng D, Haft DH, Horvath P, Moineau S, Mojica FJM, Scott D, Shah SA, Siksnys V, Terns MP, Venklovac C, White MF, Yakunin AF, Yan W, Zhang F, Garrett RA, Backofen R, van der Oost J, Barrangou R, Koonin EV. 2020. Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol* 18:67–83. <https://doi.org/10.1038/s41579-019-0299-x>.
 17. van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJ. 2009. CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem Sci* 34:401–407. <https://doi.org/10.1016/j.tibs.2009.05.002>.
 18. Freidlin PJ, Nissan I, Luria A, Goldblatt D, Schaffer L, Kaidar-Shwartz H, Chemtob D, Dveyrin Z, Head SR, Rorman E. 2017. Structure and variation of CRISPR and CRISPR-flanking regions in deleted-direct repeat region *Mycobacterium tuberculosis* complex strains. *BMC Genomics* 18:168. <https://doi.org/10.1186/s12864-017-3560-6>.
 19. Gonzalo-Asensio J, Perez I, Aguilo N, Uranga S, Pico A, Lampreaue C, Cebollada A, Otal I, Samper S, Martin C. 2018. New insights into the transposition mechanisms of IS6110 and its dynamic distribution between *Mycobacterium tuberculosis* complex lineages. *PLoS Genet* 14:e1007282. <https://doi.org/10.1371/journal.pgen.1007282>.
 20. McAdam RA, Hermans PW, van Soolingen D, Zainuddin ZF, Catty D, van Embden JD, Dale JW. 1990. Characterization of a *Mycobacterium tuberculosis* insertion sequence belonging to the IS3 family. *Mol Microbiol* 4:1607–1613. <https://doi.org/10.1111/j.1365-2958.1990.tb02073.x>.
 21. Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, Khanna V, Majlessi L, Criscuolo A, Tap J, Pawlik A, Fiette L, Orgeur M, Fabre M, Parmentier C, Frigui W, Simeone R, Boritsch EC, Debrie AS, Willery E, Walker D, Quail MA, Ma L, Bouchier C, Salvignol G, Sayes F, Cascioferro A, Seemann T, Barbe V, Loch T, Gutierrez MC, Leclerc C, Bentley SD, Stinear TP, Brisse S, Medigue C, Parkhill J, Cruveiller S, Brosch R. 2013. Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of *Mycobacterium tuberculosis*. *Nat Genet* 45:172–179. <https://doi.org/10.1038/ng.2517>.
 22. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, Rocha EPC, Vergnaud G, Gautheret D, Pourcel C. 2018. CRISPRCasFinder, an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res* 46:W246–W251. <https://doi.org/10.1093/nar/gky425>.
 23. Lange SJ, Alkhnbashi OS, Rose D, Will S, Backofen R. 2013. CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Res* 41:8034–8044. <https://doi.org/10.1093/nar/gkt606>.
 24. Refregier G, Sola C, Guyeux C. 2020. Unexpected diversity of CRISPR unveils some evolutionary patterns of repeated sequences in *Mycobacterium tuberculosis*. *BMC Genomics* 21:841. <https://doi.org/10.1186/s12864-020-07178-6>.
 25. Reva O, Korotetskiy I, Ilin A. 2015. Role of the horizontal gene exchange in evolution of pathogenic *Mycobacteria*. *BMC Evol Biol* 15:S2–8. <https://doi.org/10.1186/1471-2148-15-S1-S2>.
 26. Chiner-Oms A, Sanchez-Buso L, Corander J, Gagneux S, Harris SR, Young D, Gonzalez-Candelas F, Comas I. 2019. Genomic determinants of speciation and spread of the *Mycobacterium tuberculosis* complex. *Sci Adv* 5:eaaw3307. <https://doi.org/10.1126/sciadv.aaw3307>.
 27. Gagneux S. 2018. Ecology and evolution of *Mycobacterium tuberculosis*. *Nat Rev Microbiol* 16:202–213. <https://doi.org/10.1038/nrmicro.2018.8>.
 28. Coscolla M, Gagneux S. 2014. Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Semin Immunol* 26:431–444. <https://doi.org/10.1016/j.smim.2014.09.012>.
 29. Langille MG, Hsiao WW, Brinkman FS. 2010. Detecting genomic islands using bioinformatics approaches. *Nat Rev Microbiol* 8:373–382. <https://doi.org/10.1038/nrmicro2350>.
 30. Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol Rev* 33:376–393. <https://doi.org/10.1111/j.1574-6976.2008.00136.x>.
 31. Bertelli C, Laird MR, Williams KP, Simon Fraser University Research Computing Group, Lau BY, Hoad G, Winsor GL, Brinkman FSL. 2017. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res* 45:W30–W35. <https://doi.org/10.1093/nar/gkx343>.
 32. Plagens A, Richter H, Charpentier E, Randau L. 2015. DNA and RNA interference mechanisms by CRISPR-Cas surveillance complexes. *FEMS Microbiol Rev* 39:442–463. <https://doi.org/10.1093/femsre/fuv019>.
 33. Caro-Quintero A, Konstantinidis KT. 2015. Inter-phylum HGT has shaped the metabolism of many mesophilic and anaerobic bacteria. *ISME J* 9:958–967. <https://doi.org/10.1038/ismej.2014.193>.
 34. Coll F, McNeerney R, Guerra-Assuncao JA, Glynn JR, Perdigo J, Viveiros M, Portugal I, Pain A, Martin N, Clark TG. 2014. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat Commun* 5:4812. <https://doi.org/10.1038/ncomms5812>.
 35. Phelan J, de Sessions PF, Tientcheu L, Perdigo J, Machado D, Hasan R, Hasan Z, Bergval IL, Anthony R, McNeerney R, Antonio M, Portugal I, Viveiros M, Campino S, Hibberd ML, Clark TG. 2018. Methylation in *Mycobacterium tuberculosis* is lineage specific with associated mutations present globally. *Sci Rep* 8:160. <https://doi.org/10.1038/s41598-017-18188-y>.
 36. Nebenzahl-Guimaraes H, Yimer SA, Holm-Hansen C, de Beer J, Brosch R, van Soolingen D. 2016. Genomic characterization of *Mycobacterium tuberculosis* lineage 7 and a proposed name: 'Aethiops vetus'. *Microb Genom* 2:e000063. <https://doi.org/10.1099/mgen.0.000063>.
 37. Lipworth S, Sajou R, de Neeling A, Bradley P, van der Hoek W, Maphalala G, Bonnet M, Sanchez-Padilla E, Diel R, Niemann S, Iqbal Z, Smith G, Peto T, Crook D, Walker T, van Soolingen D. 2019. SNP-IT tool for identifying subspecies and associated lineages of *Mycobacterium tuberculosis* complex. *Emerg Infect Dis* 25:482–488. <https://doi.org/10.3201/eid2503.180894>.
 38. Wei J, Lu N, Li Z, Wu X, Jiang T, Xu L, Yang C, Guo S. 2019. The *Mycobacterium tuberculosis* CRISPR-associated Cas1 involves persistence and tolerance to anti-tubercular drugs. *Biomed Res Int* 2019:7861695. <https://doi.org/10.1155/2019/7861695>.
 39. Ribeiro SC, Gomes LL, Amaral EP, Andrade MR, Almeida FM, Rezende AL, Lanes VR, Carvalho EC, Suffys PN, Mokrousov I, Lasunskaja EB. 2014. *Mycobacterium tuberculosis* strains of the modern sublineage of the Beijing family are more likely to display increased virulence than strains of the ancient sublineage. *J Clin Microbiol* 52:2615–2624. <https://doi.org/10.1128/JCM.00498-14>.
 40. Babu M, Beloglazova N, Flick R, Graham C, Skarina T, Nocek B, Gagarinova A, Pogoutse O, Brown G, Binkowski A, Phanse S, Joachimiak A, Koonin EV, Savchenko A, Emili A, Greenblatt J, Edwards AM, Yakunin AF. 2011. A dual function of the CRISPR-Cas system in bacterial antiviral immunity and

- DNA repair. *Mol Microbiol* 79:484–502. <https://doi.org/10.1111/j.1365-2958.2010.07465.x>.
41. Ebrahimi-Rad M, Bifani P, Martin C, Kremer K, Samper S, Rauzier J, Kreiswirth B, Blazquez J, Jouan M, van Soolingen D, Gicquel B. 2003. Mutations in putative mutator genes of *Mycobacterium tuberculosis* strains of the W-Beijing family. *Emerg Infect Dis* 9:838–845. <https://doi.org/10.3201/eid0907.020803>.
 42. Shitikov E, Vyazovaya A, Malakhova M, Guliaev A, Bespyatykh J, Proshina E, Pasechnik O, Mokrousov I. 2019. Simple assay for detection of the Central Asia outbreak clade of the *Mycobacterium tuberculosis* Beijing genotype. *J Clin Microbiol* 57:e00215-19. <https://doi.org/10.1128/JCM.00215-19>.
 43. Mendiola MV, Martin C, Otaol I, Gicquel B. 1992. Analysis of the regions responsible for IS6110 RFLP in a single *Mycobacterium tuberculosis* strain. *Res Microbiol* 143:767–772. [https://doi.org/10.1016/0923-2508\(92\)90104-V](https://doi.org/10.1016/0923-2508(92)90104-V).
 44. Hermans PW, van Soolingen D, Bik EM, de Haas PE, Dale JW, van Embden JD. 1991. Insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. *Infect Immun* 59:2695–2705. <https://doi.org/10.1128/IAI.59.8.2695-2705.1991>.
 45. Hakamata M, Takihara H, Iwamoto T, Tamaru A, Hashimoto A, Tanaka T, Kaboso SA, Gebretsadik G, Ilinov A, Yokoyama A, Ozeki Y, Nishiyama A, Tateishi Y, Moro H, Kikuchi T, Okuda S, Matsumoto S. 2020. Higher genome mutation rates of Beijing lineage of *Mycobacterium tuberculosis* during human infection. *Sci Rep* 10:17997. <https://doi.org/10.1038/s41598-020-75028-2>.
 46. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39:W29–W37. <https://doi.org/10.1093/nar/gkr367>.
 47. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30:3059–3066. <https://doi.org/10.1093/nar/gkf436>.
 48. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAXML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35:4453–4455. <https://doi.org/10.1093/bioinformatics/btz305>.
 49. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–2504. <https://doi.org/10.1101/gr.1239303>.
 50. Dion MB, Labrie SJ, Shah SA, Moineau S. 2018. CRISPRStudio: a user-friendly software for rapid CRISPR array visualization. *Viruses* 10:602. <https://doi.org/10.3390/v10110602>.
 51. Biswas A, Staals RH, Morales SE, Fineran PC, Brown CM. 2016. CRISPRDetect: a flexible algorithm to define CRISPR arrays. *BMC Genomics* 17:356. <https://doi.org/10.1186/s12864-016-2627-0>.
 52. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31:3691–3693. <https://doi.org/10.1093/bioinformatics/btv421>.
 53. Will S, Joshi T, Hofacker IL, Stadler PF, Backofen R. 2012. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA* 18:900–914. <https://doi.org/10.1261/rna.029041.111>.
 54. Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* 14:1188–1190. <https://doi.org/10.1101/gr.849004>.
 55. 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. <https://doi.org/10.1093/molbev/msu300>.
 56. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>.
 57. Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: a genome comparison visualizer. *Bioinformatics* 27:1009–1010. <https://doi.org/10.1093/bioinformatics/btr039>.