Title: Evolutionary thrift: mycobacteria repurpose plasmid diversity during adaptation of type VII secretion systems

Authors:
Tatum D. Mortimer[1,2], Alexandra M. Weber[1], Caitlin S. Pepperell[1*]

[1] Department of Medicine, Division of Infectious Diseases and Department of Medical Microbiology and Immunology, University of Wisconsin-Madison
[2] Microbiology Doctoral Training Program, University of Wisconsin-Madison
* Author for Correspondence: Caitlin Pepperell, Departments of Medicine and Medical Microbiology and Immunology, University of Wisconsin-Madison, Madison, WI, cspepper@medicine.wisc.edu

Abstract
Mycobacteria have a distinct type of secretion system, termed type VII (T7SS), which is encoded by paralogous chromosomal loci (ESX) and associated with pathogenesis, conjugation, metal homeostasis and other functions. Gene duplication is an important mechanism by which novel gene functions can evolve. There are, however, potential conflicts between adaptive forces that stabilize duplicated genes and those that enable the evolution of new functions. Our objective was to delineate the adaptive forces underlying functional diversification of T7SS using genomic data from mycobacteria and related Actinobacteria. Plasmid-borne ESX were described recently, and we found evidence that the initial duplication and divergence of ESX systems occurred on plasmids and was driven by selection for advantageous mutations. We speculate that differentiation of plasmid ESX was driven by development of plasmid incompatibility systems. Plasmid ESX systems appear to have been repurposed following their migration to the chromosome, and there is evidence of positive selection driving further differentiation of the chromosomal ESX. We hypothesize that ESX loci were initially stabilized on the chromosome by mediating their own transfer. These results emphasize the diverse adaptive paths underlying the evolution of novelty, which in this case involved plasmid duplications, selection for advantageous mutations in both the core and mobile genomes, migration of the loci between plasmids and chromosomes, and lateral transfer among chromosomes. We discuss further implications of these results for the choice of model organism to study ESX functions in *Mycobacterium tuberculosis*.

Key words: type VII secretion system, ESX, mycobacteria, gene duplication, plasmid, selection

Introduction
Gene duplications are an important mechanism by which novel gene functions evolve (Zhang 2003). Duplications have been shown to occur frequently during experimental evolution of bacterial populations, and can be adaptive, for example in producing antibiotic resistance (Sandegren and Andersson 2009). Most duplications are transient, due to their intrinsic instability and associated fitness costs, as well as general mutational biases toward deletion (Sandegren and Andersson 2009; Adler et al. 2014). These observations have led researchers to investigate the selective forces allowing

duplicate genes to persist and to diverge from the parent gene (Bergthorsson et al. 2007; Bershtein and Tawfik 2008; Näsvall et al. 2012). 'Ohno's dilemma' refers to the potential conflict between selection that enables persistence of a gene duplication from its inception and that which enables evolution of novel functions (Ohno 1970). Several solutions have been proposed (Bergthorsson et al. 2007; Hittinger and Carroll 2007; Elde et al. 2012) to the problem of how to maintain a duplication long enough for it to acquire adaptive mutations conferring a new function.

Bacterial species within the genus *Mycobacterium* have a distinct secretion system, termed the type VII secretion system (T7SS), which is encoded by six paralogous chromosomal loci referred to as ESX (ESX-1, -2, -3, -4, -5, and –4-bis/-4_{EVOL}). The ESX loci share a core consisting of 6 genes (*eccB*, *eccC, eccD, mycP*, *esxA*, *esxB*); the loci typically encode an additional 4 genes (a PE, PPE (Bottai and Brosch 2009), *eccA*, and *eccE*) as well as a variable complement of locus specific gene content (Figure S1).

Some ESX loci have been characterized, and there is evidence from these studies of functional divergence among T7SS. ESX-1 is associated with several aspects of virulence in *Mycobacterium tuberculosis*, including growth in macrophages (Stanley et al. 2003; McLaughlin et al. 2007), cytosolic translocation (Houben et al. 2012), and antigen presentation (Sreejit et al. 2014). In *Mycobacterium smegmatis*, a non-pathogenic, environmental mycobacterium, ESX-1 is involved in distributive conjugal transfer, a mechanism of lateral gene transfer (Flint et al. 2004; Coros et al. 2008; Gray et al. 2013). ESX-3 is essential for *M. tuberculosis* growth *in vitro* (Sassetti et al. 2003) and is involved in iron acquisition in mycobacteria (Serafini et al. 2009; Siegrist et al. 2009; Serafini et al. 2013). ESX-3 is also thought to contribute to *M. tuberculosis* virulence independent of its role in metal homeostasis (Mehra et al. 2013; Tufariello et al. 2016). ESX-5 has been shown to secrete PE/PPE proteins in *Mycobacterium marinum* (Abdallah et al. 2009) and *M. tuberculosis* (Bottai et al. 2012). The emergence of ESX-5 coincides with the expansion of PE/PPEs in mycobacteria (Pittius et al. 2006). Both ESX-1 and ESX-5 additionally play roles in membrane integrity (Garces et al. 2010; Ates et al. 2015). The functions of ESX-2 and ESX-4 in mycobacteria are unknown.

The goal of the present study was to delineate the adaptive processes underlying divergence of mycobacterial T7SS, and to define groups of T7SS that are likely to be functionally related. In our analyses of genomic data from 33 mycobacterial species and related Actinobacteria, we found evidence of positive selection driving differentiation of T7SS loci. Loci within groups that diverged from each other as a result of positive selection are likely to be functionally related. We speculate that the development of plasmid incompatibility systems drove the initial diversification of ESX loci prior to their migration onto the chromosome, and that the loci were initially stabilized on the chromosome by mediating their own lateral transfer.

Materials and Methods

Data set

All genomes selected were complete and were obtained from the National Center for Biotechnology Information (NCBI) database. Accession numbers for these genomes can be found in Table S1. Members of the *M. tuberculosis* complex (MTBC) without finished genomes (*M. caprae, M. pinnipedii*, *M. orygis*) were assembled by the reference guided assembly pipeline available at https://github.com/tracysmith/RGAPepPipe using *M. tuberculosis* H37Rv as the reference. Briefly, reads were trimmed for quality and adapters using Trim Galore! v 0.4.0 (Kreuger 2013); trimmed reads were mapped to the reference genome using BWA-MEM v 0.7.12 (Li 2013); Picard-tools v 1.138 (https://broadinstitute.github.io/picard/) marked duplicates and added read group information; and variants were called using GATK v 3.4.46 (DePristo et al. 2011).

Ortholog detection

Genomes were annotated using Prokka v 1.11 (Seemann 2014). We used OrthoMCL v 2.0.9 (Li et al. 2003) to cluster proteins from these genomes into orthologous groups. Genes known to be located in the ESX loci of *M. tuberculosis* H37Rv were obtained from (Bitter et al. 2009). Orthologous groups containing any of the genes in ESX loci of *M. tuberculosis* were identified. ESX loci were identified as at least three orthologs of genes present in *M. tuberculosis* ESX loci in close proximity to one another in the genome. Identification of ESX loci were confirmed by phylogenetic analysis of conserved genes as described below.

ESX loci and core genome alignment

Protein sequences from paralogs and orthologs of genes present in the majority of ESX loci in mycobacteria (*eccA*, *eccB*, *eccC*, *eccD*, *eccE*, *mycP*) were aligned with MAFFT v 7.245 (Katoh and Standley 2014), low quality alignment columns were identified and removed using GUIDANCE v 2.01 (Sela et al. 2015), and trimmed alignments were concatenated to produce an alignment of ESX loci. We additionally identified orthologous groups present in every genome only one time as the core genome. Alignments of core proteins produced with MAFFT were concatenated for phylogenetic analysis. Scripts used to automate OrthoMCL analysis and alignment can be found here: https://github.com/tatumdmortimer/core-genome-alignment.

Plasmid assembly and annotation

Since there are few finished, mycobacterial plasmids available that contain ESX loci, we screened publically available sequence data for evidence of plasmid-borne ESX. Sequence reads identified as *Mycobacterium*, excluding those belonging to the MTBC or *Mycobacterium leprae*, which are not known to harbor plasmids, were downloaded and assembled using plasmidSPAdes v 3.5.0 (Antipov et al. 2016). Resulting plasmid contigs were annotated using Prokka v 1.11 (Seemann 2014). Plasmids with at least one annotated ESX gene were chosen for further quality control processing, including checking for at least 3 ESX genes, checking that all ESX genes were on the same component when multiple components were assembled, and ruling out chromosomal ESX loci misidentified as plasmid-borne. In total, we downloaded and assembled reads from 1300 *Mycobacterium* strains, resulting in 732 strains with assembled plasmids. We sampled at least one strain from 67% of named *Mycobacterium* species with sequence

data available in NCBI, and 50% of *Mycobacterium* strains without a species designation. The majority of nontuberculous mycobacteria reads available in NCBI are *M. abscessus* (*n* = 1990), and we assembled 20% of these strains. 248 strains contained a plasmid with at least one ESX gene, and 16 plasmids passed all quality control checks (Table S2). Final identification and alignment of ESX loci in these assembled plasmids as well as publically available plasmid sequences (Table S3) was performed as described above for the chromosomal loci. While *M. ulcerans* plasmids were not included in the downstream analyses because they did not contain a complete ESX locus, we did create a core gene alignment (*n* = 21) and phylogeny in a small sample of the total *M. ulcerans* plasmids assembled.

Phylogenetic analysis

We performed all phylogenetic analyses using RAxML v. 8.2.3 (Stamatakis 2014). The best protein model was determined automatically using the –m PROTGAMMAAUTO option. The best-scoring maximum likelihood tree was calculated from 20 trees, and bootstrap values were calculated using the autoMR bootstrap convergence criteria. We used Dendroscope v 3 (Huson and Scornavacca 2012) and ggtree (Yu and Lam) for tree visualization and editing. Phylogenetic networks were created using Splitstree 4 (Huson and Bryant 2006), and we used the PHI test (Bruen et al. 2006) to assess the presence of recombination in the alignments. In order to address the congruence of core plasmid genes, we performed Bayesian phylogenetic analysis using MrBayes v 3.2.5 (Ronquist and Huelsenbeck 2003) and visualized tree clusters using Treescape (Kendall and Colijn 2016). MrBayes analysis was run for 1,000,000 generations for each gene, and trees were sampled every 500 generations. We discarded the first 25% of trees, randomly sampled 200 trees from each gene, and performed pairwise calculations of the Kendell Colijn metric and multidimensional scaling in Treescape.

Selection analysis

We used the aBSREL method implemented in HyPhy (Smith et al. 2015) to test for episodic directional selection in a tree of mycobacterial ESX loci. Nucleotide sequences from ESX genes were aligned with MAFFT, trimmed with Guidance, and concatenated for input into the HyPhy analysis. Additionally, a nucleotide alignment was created using translatorX (Abascal et al. 2010), which back-translates an amino acid alignment to preserve the reading frame of codons, and trimmed with Gblocks v 0.91b (Castresana 2000). Both alignments were used for maximum likelihood phylogenetic inference with RAxML and HyPhy analysis.

Data availability

Unless stated otherwise above, all scripts and data, including text files for supplementary tables, used in these analyses are available at https://github.com/tatumdmortimer/t7ss.

Results

Figure 1 shows a core genome phylogeny of 56 species of Actinobacteria along with a presence/absence matrix of associated T7SS. Our analyses are consistent with an initial emergence of the FtsK/WXG100 gene cluster (as proposed in (Pallen 2002)),

followed by ESX-4-bis and ESX-4, with subsequent duplications giving rise to ESX-3, ESX-1, ESX-2, and ESX-5. Interestingly, the loci have been lost on several occasions. For example, ESX-2 in the common ancestor of *M. marinum*, *M. liflandii*, and *M. ulcerans*, and ESX-1 from *M. sinense*, *M. avium* and related species, as well as from *M. ulcerans*, have all been lost.

Figure 2 shows a network of the ESX loci (see also Figure S2). The network has a pronounced star-like configuration, consistent with rapid diversification of these loci. This pattern is particularly evident when the plasmid loci are considered separately (Figure 3). The most basal lineages on the combined network are all plasmid-associated, which suggests that the common ancestor of ESX loci was plasmid associated.

We identified several new, plasmid-borne ESX lineages and found that each of the six chromosomal ESX is paired with one or more plasmid lineages that root basal to it. This suggests that duplication of the ESX loci occurred on plasmids, and the extant chromosomal loci all result from transfers from plasmid to chromosome.

The two most basal mycobacterial species, *M. abscessus* and *M. chelonae*, have a chromosomal ESX-3 locus, but not an ESX-1 locus. ESX-1 is, however, basal to ESX-3 on the ESX phylogeny, on a branch with low bootstrap values (Figure S3). We speculate that this conflict and phylogenetic uncertainty is due to the plasmid-borne ancestor of ESX-1 having emerged earlier than ESX-3, but ESX-3 being first to migrate to the chromosome.

A phylogeny of ESX-4 and related loci is shown in Figure 4. ESX-N, found on the chromosome of *Nocardia brasiliensis* and *N. cyriacigeorgica*, pairs with a plasmid-associated ESX locus, and is basal to ESX-4 and related ESX from a range of actinobacterial species. ESX-4 and ESX-4-bis appear to be fixed among Nocardia species and are stably associated with flanking gene content, suggesting vertical inheritance in the genus. ESX-N, by contrast, is variably present among sampled Nocardia species, and we found it to be associated with variable flanking gene content (Figure S4). We also found ESX-N to be associated with T4SS genes and other gene content otherwise specific to plasmids. We hypothesize that the ESX-N loci were horizontally transferred from an unsampled (or extinct) plasmid.

The chromosomal ESX-4 phylogeny is not concordant with that of the core genome (e.g., the placement of corynebacteria), which suggests that the locus was laterally transferred during divergence of the Actinobacteria. The patchy distribution of ESX-4-bis among mycobacterial species, as well as branching patterns among these loci, suggest ESX-4-bis has been laterally acquired on a few occasions in the genus. The ESX-4-bis locus in *M. goodii* includes *espI*, which is not found in other chromosomal ESX-4 loci but is part of the plasmid core genome (discussed further below). This suggests the locus was transferred relatively recently from a plasmid.

Although broad groupings seen on the core genome (e.g., separation of slow-growing from rapid-growing mycobacteria) are reflected in the phylogeny of the combined ESX loci (Figure S3), the branching within these groups does not always reflect the patterns of the core genome. Branching patterns within these groups were sensitive to the sampling scheme and alignment, whereas internal branching patterns were stably supported across multiple analyses. This pattern could be due to a lack of fine scale phylogenetic signal in the gene content shared among ESX loci or lateral transfer of the loci. When we created an alignment and phylogeny of only ESX-5, which contains information from two additional genes, we found that phylogenetic uncertainty remained (Figure S5). This suggests that T7SS were laterally transferred among mycobacterial species during their divergence, which could contribute to both phylogenetic uncertainty (e.g., as a result of fluctuating selection pressures) and conflicts with the core genome phylogeny.

There are few reticulations in the ESX networks (Figures 2 and 3), suggesting that within-locus recombination has not played a major role in adaptation of these loci. The PHI test for recombination (Bruen et al. 2006) was not significant (p=1.0) for an alignment of chromosomal and plasmid-associated loci, nor for the plasmid-associated loci considered separately. The PHI test was, however, significant ($p=1.2 \times 10^{-5}$) for the ESX-5 alignment, suggesting that within-locus recombination has occurred among more closely related loci.

We tested for episodic directional (positive) selection in the ESX phylogeny using HyPhy (Figure 5). Branches under selection in this model mark periods during which there is evidence of advantageous mutations driving divergence from an ancestral state. We found evidence of positive selection at each ESX duplication event, as well as during each migration event from plasmid to chromosome. These results were replicated across multiple analyses, including different sampling schemes and alignment trimming methods (Figures S6 and S7).

Summarizing the results outlined above, the ESX gene family expansion likely occurred on plasmids, and this diversification appears to have been driven by selection for advantageous mutations. A simple explanation of this pattern would be that the plasmids diverged in response to divergence of their host mycobacterial species. In this case, we would expect to observe congruence between the plasmid ESX phylogeny and the host genome phylogeny. However, in this sample of plasmids harboring ESX, the phylogenetic signals are clearly at odds with that of the host genomes (Figure 3): for example, *M. kansasii* pMK12478 pairs with *M. yongonense* pMyong1, rather than *M. marinum* pRAW. There are also multiple divergent plasmid ESX lineages associated with the same host species (e.g., *M. abscessus*) or the same host cell (Figure 3).

Another possible explanation of the plasmid ESX radiation is that it was driven by adaptation to accompanying gene content on the plasmid. To investigate this possibility, we analyzed gene content across related groups of plasmids (Table S4). Gene content on the plasmids was highly variable, and little to no gene content was uniquely shared among plasmids with similar ESX. This indicates that divergence of plasmid borne ESX

is unlikely to have been driven by interactions between ESX and gene content mobilized on plasmids.

Bacteria can protect themselves from foreign DNA, including plasmids, using CRISPR-Cas nucleases (Barrangou et al. 2007; Garneau et al. 2010). It's possible that plasmid ESX diverged in response to CRISPR found among mycobacterial host genomes. CRISPR-Cas systems have been identified previously in *M. tuberculosis, M. bovis,* and *M. avium* (He et al. 2012). We searched the annotations of 33 mycobacterial species for which finished genome sequence data were available, and only identified CRISPR loci in *M. canettii*, *M. kansasii*, *M. avium,* and the *M. tuberculosis* complex (MTBC). This indicates that plasmid borne ESX divergence is unlikely to have been shaped by adaptation to host CRISPR, at least as they are currently recognized.

As a final possibility, we investigated adaptation of plasmid conjugation systems as a driving force for divergence of plasmid ESX. Both T7SS and T4SS were found to be essential for plasmid conjugation in a recently discovered plasmid in *Mycobacterium marinum* (Ummels et al. 2014). With one interesting exception discussed below, we found T7SS to be invariably accompanied by T4SS in our plasmid sample, suggesting that their functions are interdependent across diverse mycobacterial plasmids.

Several plasmids found in *M. ulcerans* encoded an ESX 2P-like locus that was not invariably accompanied by a T4SS. Two other features distinguished these plasmids from those found in other species of mycobacteria. First, there were numerous transposable and other mobile elements on the plasmids, and second, the ESX locus showed evidence of progressive degradation with multiple, independent examples of loss of one or more genes within the locus (Figure 7).

Excluding *M. ulcerans*-associated plasmids, we found the core genes of mycobacterial ESX-encoding plasmids to consist of the T7SS genes (by definition), as well as T4SS genes (*virB4*, *tcpC*) and *espI*, which was in some cases located within the ESX locus and in others was located separately. Individual gene phylogenies within the core plasmid genome were congruent (Figure 6), consistent with their having a shared evolutionary history. This congruence also suggests that the paralogous ESX systems trace to whole plasmid duplications, as opposed to duplications of ESX loci on individual plasmids.

Positive selection was evident on the ESX phylogeny following migration of the loci to the chromosome. This suggests that novel functions evolved for ESX following their incorporation into the chromosome. There is also evidence of positive selection along the branches separating various species of mycobacteria. This suggests that individual ESX systems may have functions that are specific to species or groups of species. Another possibility is that the advantageous mutations driving divergence of chromosomal ESX loci did not confer novel functions, but were advantageous as a result of interactions with loci elsewhere on the genome. Distinct functions have been identified for different ESX loci (i.e. ESX-1, -2, -3, -4, -5, -4-bis) and for the same loci in

different species (e.g., ESX-1 in *M. tuberculosis* and *M. smegmatis*), indicating that at least in some cases the advantageous mutations conferred novel functions.

Discussion

Adaptation on plasmids

Much of the prior research on gene duplication has focused on chromosomal duplications, either of the entire chromosome or one of its segments (Lynch and Conery 2000; Zhang 2003). The recent discovery of plasmid-borne ESX (Ummels et al. 2014; Dumas et al. 2016; Newton-Foot et al. 2016) opens the possibility of a more complex evolutionary path underlying the paralogous chromosomal ESX systems currently extant among mycobacteria. In addition to the previously described plasmid-borne lineages that root basal to ESX-1, -2, -3, and -5, we have identified a plasmid lineage that roots basal to ESX-4 and clarified the relationships among this ancestral group of loci (Figure 4). The finding that all extant chromosomal ESX are associated with basally rooting plasmid-borne lineages provides strong support for the hypothesis that the most recent common ancestor of these loci was plasmid-borne. Our proposed model for the adaptation of the canonical ESX is shown in Figure 8.

Duplication, divergence, migration, deletion

The evolutionary history of mycobacterial ESX is evidently quite complex, with duplication and divergence occurring on plasmids, several migrations from plasmid to chromosome, lateral transfer among chromosomes (with or without a plasmid intermediary) as well as vertical inheritance, divergence on the chromosome and occasional loss of the loci from the chromosome. We saw evidence of ancient plasmid to chromosome migrations (e.g., of ESX-4 and -3 to the MRCA of mycobacteria) as well as more recent events (i.e. migration of ESX-N to *Nocardia* and ESX-4-bis to *M. goodii*).

This complex history provides an interesting new paradigm for the evolution of novelty following gene duplication. Our analyses suggest that ESX duplication and divergence occurred on plasmids, and that this divergence was driven by positive selection. Recent work in *Yersinia pestis* identified a positively selected phenotype associated with increased plasmid copy number (Wang et al. 2016). Positive selection for increased gene dosage may have similarly enabled the initial plasmid duplications underlying diverse T7SS. Such selection could operate at the level of the host cell, as in *Y. pestis*, or the plasmid, if, for example, it resulted in more efficient transfer of one or more plasmid copies.

We found that the T4SS and T7SS evolved in concert on the plasmids, along with *espI*. Diversification of these loci did not appear to have been driven by adaptation to different host species, CRISPR-Cas systems, or the gene content delivered by the plasmids. A possible alternative selection pressure is that imposed by development of plasmid incompatibility systems: i.e. the conjugation machinery differentiated in order to prevent conjugation between cells harboring incompatible plasmids. A possible model for such adaptation is that it occurred by differentiation of signaling molecules secreted by T7SS (e.g., WXG100), which delineated incompatible groups of plasmids, with consequent co-adaptation of secretion machinery and regulatory molecules. ESX secreted proteins are

specifically exported by the membrane associated proteins encoded on the same ESX locus, and loss of function of these genes abolishes secretion despite the presence of paralogous genes encoded in other loci on the chromosome (Guinn et al. 2004; Abdallah et al. 2007). EspI has been shown to regulate ESX-1 in *M. tuberculosis* (Zhang et al. 2014); given its apparent essentiality in ESX encoding plasmids, we speculate it could play a similar role regulating plasmid borne ESX. Gene content on the plasmids was highly variable, suggesting that there is frequent recombination among them. Our finding that T7SS, T4SS and *espI* behave like a single locus, with little evidence of intra-locus recombination, provides further evidence that differentiation of these systems is maintained by selection, such as would be imposed by a plasmid incompatibility regime.

ESX-encoding plasmids in *M. ulcerans* provide an interesting example of apparent relaxation of selection to maintain conjugation machinery, with progressive degradation of the locus evident in extant plasmids. The *M. ulcerans* ESX plasmids also encoded the gene for mycolactone, which is essential for causing the ulcerative disease associated with *M. ulcerans* infection (George et al. 1999). Selection for plasmid-delivered gene content can stabilize non-transmissible plasmids (San Millan et al. 2014). We speculate that selection on *M. ulcerans* to maintain mycolactone-encoding plasmids relaxes selection on the plasmid to maintain its own conjugative machinery.

We found evidence of directional selection – i.e. acquisition of specific advantageous mutations - in ESX following their migration to the chromosome. These advantageous mutations are the mechanism by which novel functions for T7SS would have been acquired. As noted above, it's also possible that the new ESX duplicated the function of existing loci and that the mutations occurred as a result of co-adaptation with other loci on the genome. Increased gene dosage is thought to be an important mechanism by which gene duplications are selected (Bergthorsson et al. 2007; Bershtein and Tawfik 2008; Andersson and Hughes 2009; Sandegren and Andersson 2009). In the case of ESX, however, the duplicate loci had already diverged from existing chromosomal loci at the time of their migration to the chromosome. While it's possible that further divergence of the migrant loci enabled functional convergence with existing loci, this scenario seems quite complex and distinct functions have already been identified for some loci.

Both plasmid-borne and chromosomal ESX have been shown to mediate conjugation (Flint et al. 2004; Coros et al. 2008; Gray et al. 2013; Ummels et al. 2014). We found evidence suggesting that the chromosomal loci have been laterally transferred among bacterial species. Since ESX can mediate its own lateral transfer, it raises an interesting potential solution to Ohno's dilemma. Ohno's dilemma is the problem of how duplicate genes survive in the genome long enough to acquire mutations conferring a novel function, given the deleterious impacts of duplication (Bergthorsson et al. 2007). We speculate that the migrant ESX loci acted initially as selfish genetic elements mediating their own transfer among chromosomes. We found *espI* and T4SS genes retained in association with more recent migrations, suggesting that all of the plasmid conjugation machinery was transferred initially, with subsequent remodeling of the locus. The

laterally spreading chromosomal ESX loci would provide a large genetic target for adaptive mutations conferring a new function. Fixation of these mutations would have been hastened by their lateral spread if the loci retained the capacity to mediate LGT. Whether they spread laterally or not, benefits provided by novel mutations would favor retention of ESX and resolve potential conflicts between the locus and its host genome.

Our analyses of directional selection on chromosomal ESX delineate groups of loci that are likely to have differentiated from each other as a result of the acquisition of new functions. Functions have been identified for a small number of ESX loci, and these results can aid further research in this area. For example, *M. tuberculosis* ESX-3 is closely related to *M. marinum* ESX-3, without evidence of directional selection in the branches separating them (Figure 5). This suggests that *M. marinum* is likely to be a useful model for the study of ESX-3 functions in *M. tuberculosis*. The same is true of ESX-4, whereas *M. kansasii* may be a good model for *M. tuberculosis* ESX-2. ESX-1, which is an important virulence locus (Pym et al. 2002: 1), appears to perform functions that are unique to *M. tuberculosis*, as does ESX-5. Experimental results from ESX-5 mutants in *M. tuberculosis* and *M. marinum* suggest that this locus performs different functions in these closely related species (Shah and Briken 2016).

The paralogous ESX loci are the product of a complex evolutionary history during which mycobacteria capitalized on diversity found among plasmid loci and repurposed the loci to perform diverse functions. This is an interesting paradigm for the generation of novelty via gene duplication, and such complex dynamics between mobile and core genomes may be important for other bacteria as well. Positive selection has played an important role in diversification of these loci, and we propose two potential solutions to the problem of how the duplicate loci were maintained long enough to acquire novel, adaptive mutations. Selection for increased plasmid gene dosage may have fostered the plasmid duplications, whereas we propose that an initial (or stable) LGT function may have favored retention of chromosomal loci following their migrations from plasmids. Delineation of this evolutionary history aids our understanding of the generation of evolutionary novelty and we propose ways in which these results can guide the choice of model organism and functional studies of these loci in *M. tuberculosis*.

Acknowledgements

We thank Andrew Kitchen (University of Iowa) for his input on the manuscript. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program [grant number DGE-1256259] and the National Institute of Health National Research Service Award [grant number T32 GM07215] to TDM. CSP is supported by National Institutes of Health [grant number R01AI113287].

References

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 38:W7–W13.

Abdallah AM, Gey van Pittius NC, DiGiuseppe Champion PA, Cox J, Luirink J, Vandenbroucke-Grauls CMJE, Appelmelk BJ, Bitter W. 2007. Type VII secretion — mycobacteria show the way. Nat. Rev. Microbiol. 5:883–891.

Abdallah AM, Verboom T, Weerdenburg EM, Gey van Pittius NC, Mahasha PW, Jiménez C, Parra M, Cadieux N, Brennan MJ, Appelmelk BJ, et al. 2009. PPE and PE_PGRS proteins of Mycobacterium marinum are transported via the type VII secretion system ESX-5. Mol. Microbiol. 73:329–340.

Adler M, Anjum M, Berg OG, Andersson DI, Sandegren L. 2014. High Fitness Costs and Instability of Gene Duplications Reduce Rates of Evolution of New Genes by Duplication-Divergence Mechanisms. Mol. Biol. Evol. 31:1526–1535.

Andersson DI, Hughes D. 2009. Gene Amplification and Adaptive Evolution in Bacteria. Annu. Rev. Genet. 43:167–195.

Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, Pevzner P. 2016. plasmidSPAdes: Assembling Plasmids from Whole Genome Sequencing Data. bioRxiv:48942.

Ates LS, Ummels R, Commandeur S, van der Weerd R, Sparrius M, Weerdenburg E, Alber M, Kalscheuer R, Piersma SR, Abdallah AM, et al. 2015. Essential Role of the ESX-5 Secretion System in Outer Membrane Permeability of Pathogenic Mycobacteria. PLoS Genet 11:e1005190.

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science 315:1709–1712.

Bergthorsson U, Andersson DI, Roth JR. 2007. Ohno's dilemma: Evolution of new genes under continuous selection. Proc. Natl. Acad. Sci. 104:17004–17009.

Bershtein S, Tawfik DS. 2008. Ohno's Model Revisited: Measuring the Frequency of Potentially Adaptive Mutations under Various Mutational Drifts. Mol. Biol. Evol. 25:2311–2318.

Bitter W, Houben ENG, Bottai D, Brodin P, Brown EJ, Cox JS, Derbyshire K, Fortune SM, Gao L-Y, Liu J, et al. 2009. Systematic Genetic Nomenclature for Type VII Secretion Systems. PLoS Pathog 5:e1000507.

Bottai D, Brosch R. 2009. Mycobacterial PE, PPE and ESX clusters: novel insights into the secretion of these most unusual protein families. Mol. Microbiol. 73:325–328.

Bottai D, Di Luca M, Majlessi L, Frigui W, Simeone R, Sayes F, Bitter W, Brennan MJ, Leclerc C, Batoni G, et al. 2012. Disruption of the ESX-5 system of Mycobacterium tuberculosis causes loss of PPE protein secretion, reduction of cell wall integrity and strong attenuation. Mol. Microbiol. 83:1195–1209.

Bruen TC, Philippe H, Bryant D. 2006. A Simple and Robust Statistical Test for Detecting the Presence of Recombination. Genetics 172:2665–2681.

Castresana J. 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Mol. Biol. Evol. 17:540–552.

Coros A, Callahan B, Battaglioli E, Derbyshire KM. 2008. The specialized secretory apparatus ESX-1 is essential for DNA transfer in Mycobacterium smegmatis. Mol. Microbiol. 69:794–808.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–498.

Dumas E, Boritsch EC, Vandenbogaert M, Vega RCR de la, Thiberge J-M, Caro V, Gaillard J-L, Heym B, Girard-Misguich F, Brosch R, et al. 2016. Mycobacterial pan-genome analysis suggests important role of plasmids in the radiation of type VII secretion systems. Genome Biol. Evol.:evw001.

Elde NC, Child SJ, Eickbush MT, Kitzman JO, Rogers KS, Shendure J, Geballe AP, Malik HS. 2012. Poxviruses Deploy Genomic Accordions to Adapt Rapidly against Host Antiviral Defenses. Cell 150:831–841.

Flint JL, Kowalski JC, Karnati PK, Derbyshire KM. 2004. The RD1 virulence locus of Mycobacterium tuberculosis regulates DNA transfer in Mycobacterium smegmatis. Proc. Natl. Acad. Sci. U. S. A. 101:12598–12603.

Garces A, Atmakuri K, Chase MR, Woodworth JS, Krastins B, Rothchild AC, Ramsdell TL, Lopez MF, Behar SM, Sarracino DA, et al. 2010. EspA Acts as a Critical Mediator of ESX1-Dependent Virulence in Mycobacterium tuberculosis by Affecting Bacterial Cell Wall Integrity. PLoS Pathog 6:e1000957.

Garneau JE, Dupuis M-È, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadán AH, Moineau S. 2010. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature 468:67–71.

George KM, Chatterjee D, Gunawardana G, Welty D, Hayman J, Lee R, Small PLC. 1999. Mycolactone: A Polyketide Toxin from Mycobacterium ulcerans Required for Virulence. Science 283:854–857.

Gray TA, Krywy JA, Harold J, Palumbo MJ, Derbyshire KM. 2013. Distributive Conjugal Transfer in Mycobacteria Generates Progeny with Meiotic-Like Genome-Wide Mosaicism, Allowing Mapping of a Mating Identity Locus. PLOS Biol. 11:e1001602.

Guinn KM, Hickey MJ, Mathur SK, Zakel KL, Grotzke JE, Lewinsohn DM, Smith S, Sherman DR. 2004. Individual RD1-region genes are required for export of

ESAT-6/CFP-10 and for virulence of Mycobacterium tuberculosis. Mol. Microbiol. 51:359–370.

He L, Fan X, Xie J. 2012. Comparative genomic structures of Mycobacterium CRISPR-Cas. J. Cell. Biochem. 113:2464–2473.

Hittinger CT, Carroll SB. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. Nature 449:677–681.

Houben ENG, Bestebroer J, Ummels R, Wilson L, Piersma SR, Jiménez CR, Ottenhoff THM, Luirink J, Bitter W. 2012. Composition of the type VII secretion system membrane complex. Mol. Microbiol. 86:472–484.

Huson DH, Bryant D. 2006. Application of Phylogenetic Networks in Evolutionary Studies. Mol. Biol. Evol. 23:254–267.

Huson DH, Scornavacca C. 2012. Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. Syst. Biol.:sys062.

Katoh K, Standley DM. 2014. MAFFT: iterative refinement and additional methods. Methods Mol. Biol. Clifton NJ 1079:131–146.

Kendall M, Colijn C. 2016. Mapping phylogenetic trees to reveal distinct patterns of evolution. Mol. Biol. Evol.:msw124.

Kreuger F. 2013. TrimGalore! Available from: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio [Internet]. Available from: http://arxiv.org/abs/1303.3997

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 13:2178–2189.

Lynch M, Conery JS. 2000. The Evolutionary Fate and Consequences of Duplicate Genes. Science 290:1151–1155.

McLaughlin B, Chon JS, MacGurn JA, Carlsson F, Cheng TL, Cox JS, Brown EJ. 2007. A mycobacterium ESX-1-secreted virulence factor with unique requirements for export. PLoS Pathog. 3:e105.

Mehra A, Zahra A, Thompson V, Sirisaengtaksin N, Wells A, Porto M, Köster S, Penberthy K, Kubota Y, Dricot A, et al. 2013. Mycobacterium tuberculosis Type VII Secreted Effector EsxH Targets Host ESCRT to Impair Trafficking. PLoS Pathog 9:e1003734.

Näsvall J, Sun L, Roth JR, Andersson DI. 2012. Real-Time Evolution of New Genes by Innovation, Amplification, and Divergence. Science 338:384–387.

Newton-Foot M, Warren RM, Sampson SL, van Helden PD, Gey van Pittius NC. 2016. The plasmid-mediated evolution of the mycobacterial ESX (Type VII) secretion systems. BMC Evol. Biol. 16:62.

Ohno S. 1970. Evolution by Gene Duplication. Springer-Verlag

Pallen MJ. 2002. The ESAT-6/WXG100 superfamily–and a new Gram-positive secretion system? Trends Microbiol. 10:209–212.

Pittius NCG van, Sampson SL, Lee H, Kim Y, Helden PD van, Warren RM. 2006. Evolution and expansion of the Mycobacterium tuberculosis PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. BMC Evol. Biol. 6:95.

Pym AS, Brodin P, Brosch R, Huerre M, Cole ST. 2002. Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines Mycobacterium bovis BCG and Mycobacterium microti. Mol. Microbiol. 46:709–717.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

San Millan A, Pena-Miller R, Toll-Riera M, Halbert ZV, McLean AR, Cooper BS, MacLean RC. 2014. Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. Nat. Commun. 5:5208.

Sandegren L, Andersson DI. 2009. Bacterial gene amplification: implications for the evolution of antibiotic resistance. Nat. Rev. Microbiol. 7:578–588.

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

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics:btu153.

Sela I, Ashkenazy H, Katoh K, Pupko T. 2015. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res. 43:W7–W14.

Serafini A, Boldrin F, Palù G, Manganelli R. 2009. Characterization of a Mycobacterium tuberculosis ESX-3 Conditional Mutant: Essentiality and Rescue by Iron and Zinc. J. Bacteriol. 191:6340–6344.

Serafini A, Pisu D, Palù G, Rodriguez GM, Manganelli R. 2013. The ESX-3 Secretion System Is Necessary for Iron and Zinc Homeostasis in Mycobacterium tuberculosis. PLoS ONE 8:e78351.

Shah S, Briken V. 2016. Modular Organization of the ESX-5 Secretion System in Mycobacterium tuberculosis. Front. Cell. Infect. Microbiol.:49.

Siegrist MS, Unnikrishnan M, McConnell MJ, Borowsky M, Cheng T-Y, Siddiqi N, Fortune SM, Moody DB, Rubin EJ. 2009. Mycobacterial Esx-3 is required for mycobactin-mediated iron acquisition. Proc. Natl. Acad. Sci. 106:18792–18797.

Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Pond SLK. 2015. Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. Mol. Biol. Evol. 32:1342–1353.

Sreejit G, Ahmed A, Parveen N, Jha V, Valluri VL, Ghosh S, Mukhopadhyay S. 2014. The ESAT-6 Protein of Mycobacterium tuberculosis Interacts with Beta-2-Microglobulin (β2M) Affecting Antigen Presentation Function of Macrophage. PLoS Pathog 10:e1004446.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Stanley SA, Raghavan S, Hwang WW, Cox JS. 2003. Acute infection and macrophage subversion by Mycobacterium tuberculosis require a specialized secretion system. Proc. Natl. Acad. Sci. 100:13001–13006.

Tufariello JM, Chapman JR, Kerantzas CA, Wong K-W, Vilchèze C, Jones CM, Cole LE, Tinaztepe E, Thompson V, Fenyö D, et al. 2016. Separable roles for Mycobacterium tuberculosis ESX-3 effectors in iron acquisition and virulence. Proc. Natl. Acad. Sci.:201523321.

Ummels R, Abdallah AM, Kuiper V, Aâjoud A, Sparrius M, Naeem R, Spaink HP, Soolingen D van, Pain A, Bitter W. 2014. Identification of a Novel Conjugative Plasmid in Mycobacteria That Requires Both Type IV and Type VII Secretion. mBio 5:e01744-14.

Wang H, Avican K, Fahlgren A, Erttmann SF, Nuss AM, Dersch P, Fallman M, Edgren T, Wolf-Watz H. 2016. Increased plasmid copy number is essential for Yersinia T3SS function and virulence. Science:aaf7501.

Yu G, Lam TT-Y. ggtree: a phylogenetic tree viewer for different types of tree annotations.

Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol. Evol. 18:292–298.

Zhang M, Chen JM, Sala C, Rybniker J, Dhar N, Cole ST. 2014. EspI regulates the ESX-1 secretion system in response to ATP levels in Mycobacterium tuberculosis. Mol. Microbiol.:n/a-n/a.

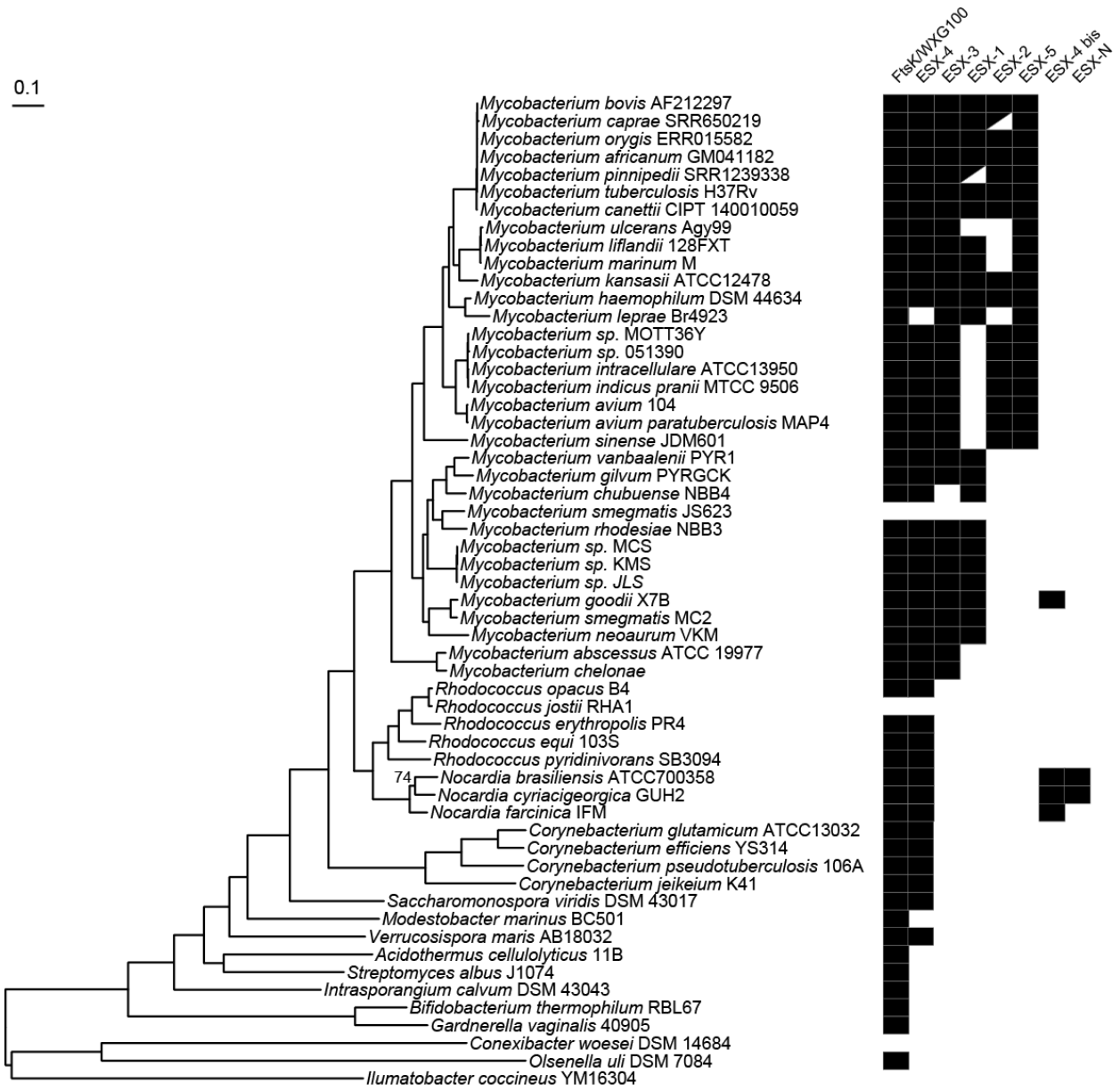

Fig. 1. Maximum likelihood phylogeny of Actinobacteria with presence of type VII secretion system loci. RAxML was used for phylogenetic inference of the Actinobacteria core genome alignment (concatenated amino acid alignments of genes (*n* = 171) present in all genomes without duplications). The phylogeny is midpoint rooted, and branches without labels have a boostrap value of 100. Presence of ESX loci is indicated with black boxes. Some MTBC species have characteristic deletions located in ESX loci. Loci with deletions are represented by black triangles. *M. caprae* has a deletion in ESX-2 spanning PE/PPE, *esxC*, *espG2*, *Rv3888c*, *eccD2*, and *mycP2. M. pinnipedii* has a deletion in ESX-1 spanning PE/PPE and a portion of *eccC1b*.

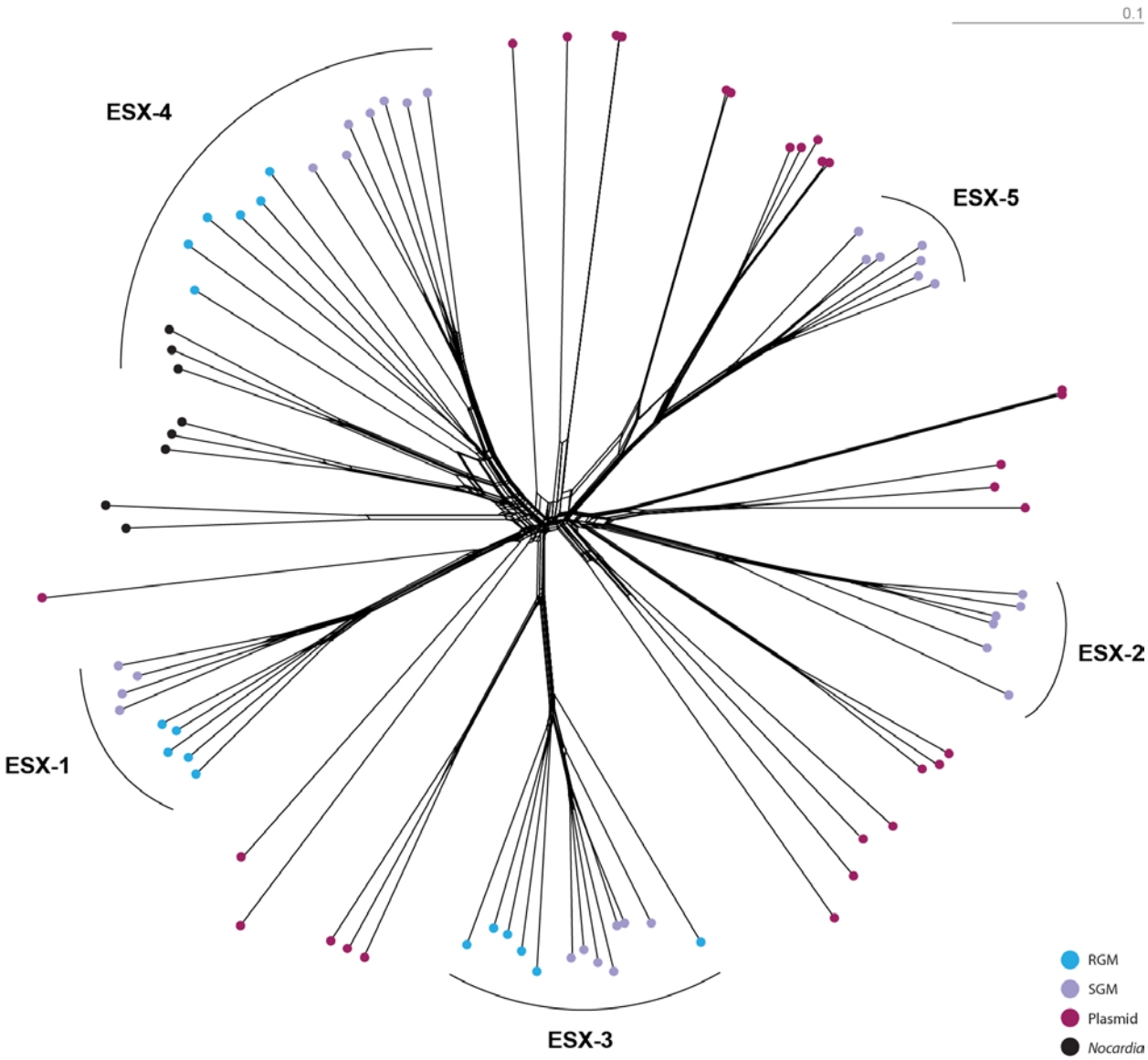

Fig. 2. Network of ESX loci in mycobacteria, *Nocardia*, and mycobacterial plasmids. Light blue dots correspond to ESX loci from rapid growing mycobacterial chromosomes, light purple dots correspond to ESX loci from slow growing mycobacterial chromosomes, magenta dots correspond to ESX loci from mycobacterial plasmids, and black dots correspond to ESX loci from *Nocardia* chromosomes. All chromosomal ESX duplications (ESX-1-5) have basal plasmid loci, suggesting that the ancestral ESX locus was plasmid-borne. The network was created in SplitsTree4 from a concatenated alignment of *eccA, eccB, eccC, eccD, eccE,* and *mycP* .The PHI test was insignificant for this alignment.

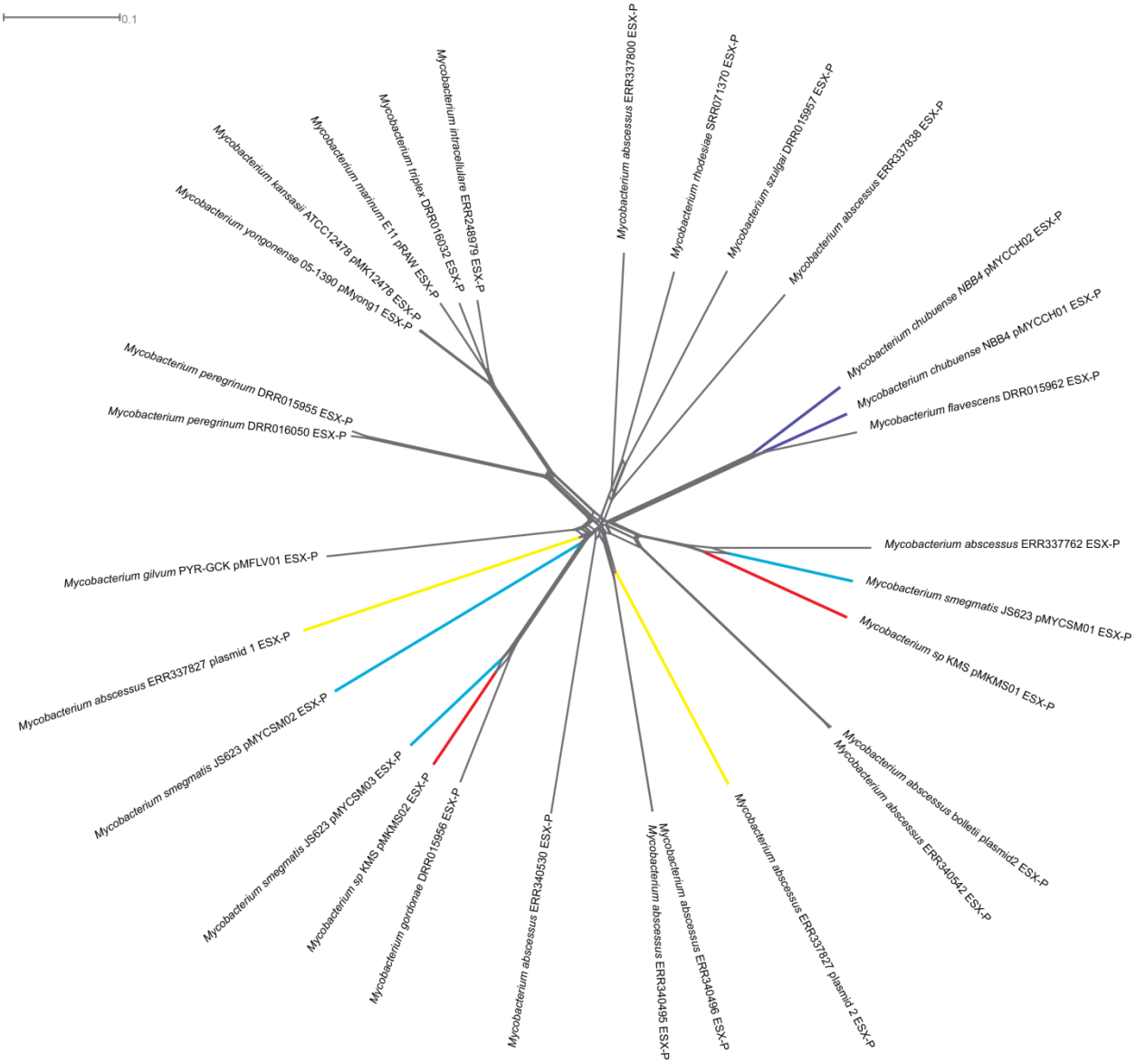

Fig. 3. Network of plasmid-borne ESX loci. Colored branches correspond to ESX loci found on plasmids within the same cell. The phylogenetic relationships of the ESX loci do not follow the core genome phylogeny, and plasmids with divergent ESX loci can be found within the same host species or even the same cell. The network was created in SplitsTree4 from a concatenated alignment of *eccA, eccB, eccC, eccD, eccE,* and *mycP.* We did not find evidence of recombination in this alignment with the PHI test.

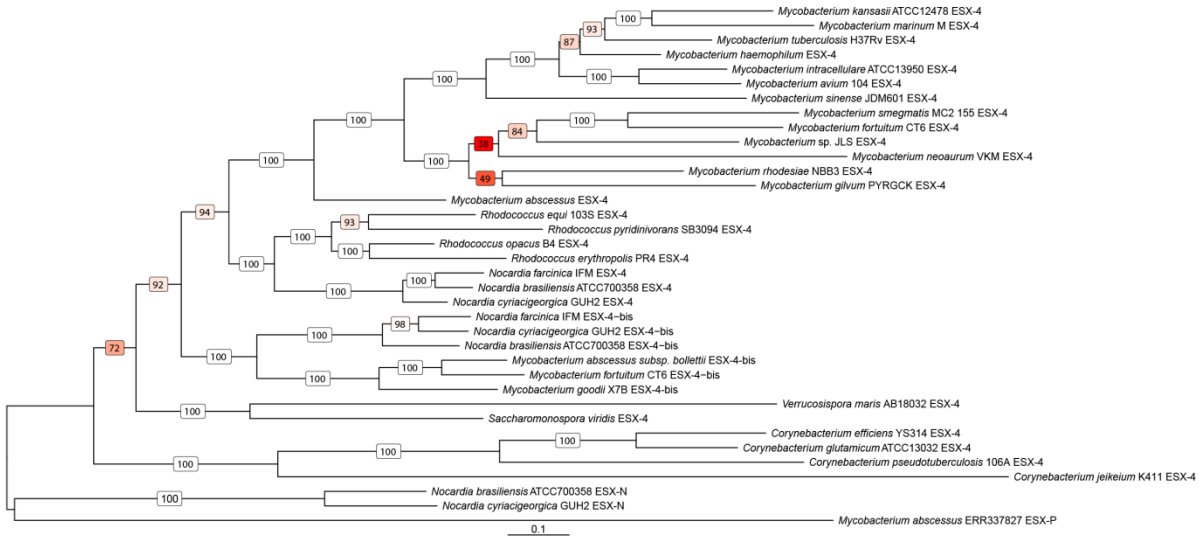

Fig. 4. Maximum likelihood phylogeny of ESX-4 in Actinobacteria. The phylogeny is rooted using ESX-N and a basal plasmid-borne ESX locus. Bootstrap values are colored based on support (white = 100, red = lowest support).The location of *Corynebacterium* ESX-4 and *M. goodii* ESX-4-bis are in conflict with the core genome phylogeny.

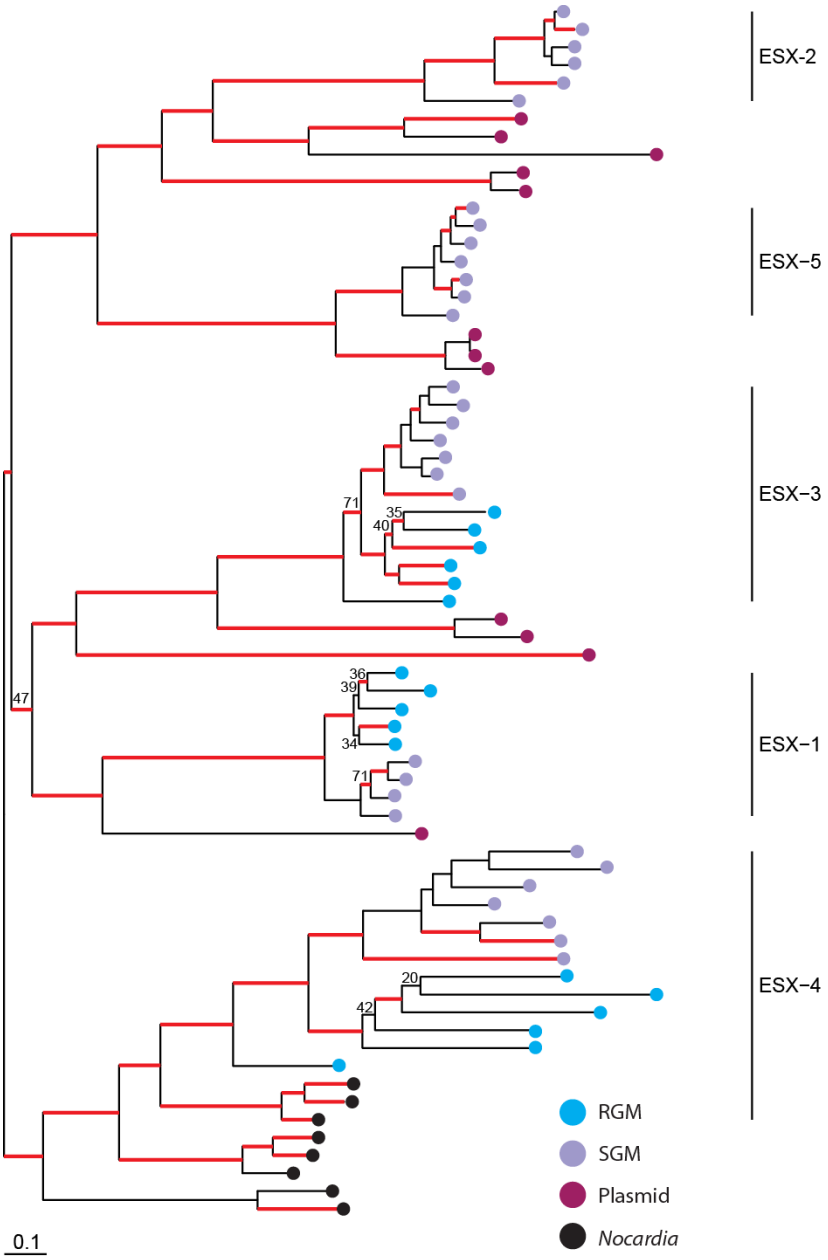

Fig. 5. Episodic directional selection during the evolution of ESX loci. Light blue dots correspond to ESX loci from rapid growing mycobacterial chromosomes, light purple dots correspond to ESX loci from slow growing mycobacterial chromosomes, magenta dots correspond to ESX loci from mycobacterial plasmids, and black dots correspond to ESX loci from *Nocardia* chromosomes. All chromosomal ESX duplications (ESX-1-5) have basal plasmid loci, suggesting that the ancestral ESX locus was likely plasmid-borne. The maximum likelihood phylogeny was created with RAxML from a concatenated alignment of *eccA, eccB, eccC, eccD, eccE,* and *mycP*. Branches without labels have a bootstrap value greater than 75. We used the aBSREL test implemented in HyPhy to identify branches with significant evidence of episodic directional selection; these branches are highlighted in red.

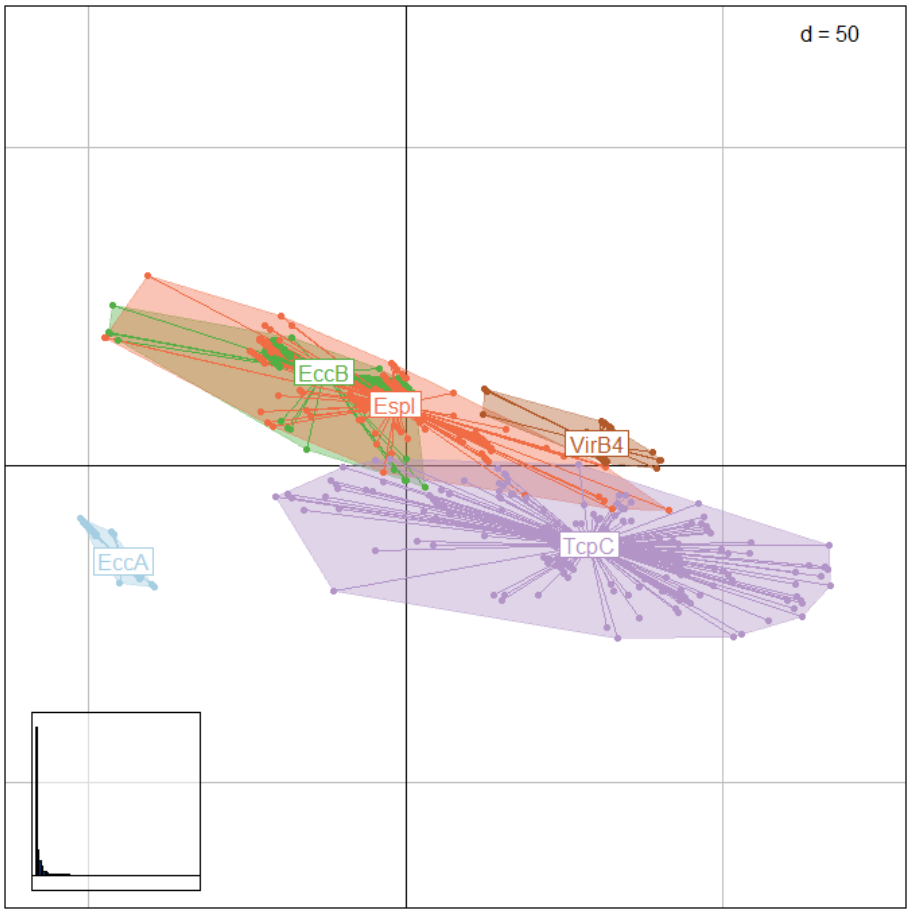

Fig. 6. Congruence of tree topologies from T7SS, T4SS, and EspI. Bayesian phylogenetic analysis was performed in MrBayes using amino acid alignments of EccA, EccB, VirB4, TcpC, and EspI. We used TreeScape to calculate the Kendell Colijn metric between pairs of trees and perform multidimensional scaling (MDS). Clusters of trees are visualized as a scatterplot of the first two principal components from the MDS. The inset bar chart is a scree plot showing the eigenvalues for the principal components.

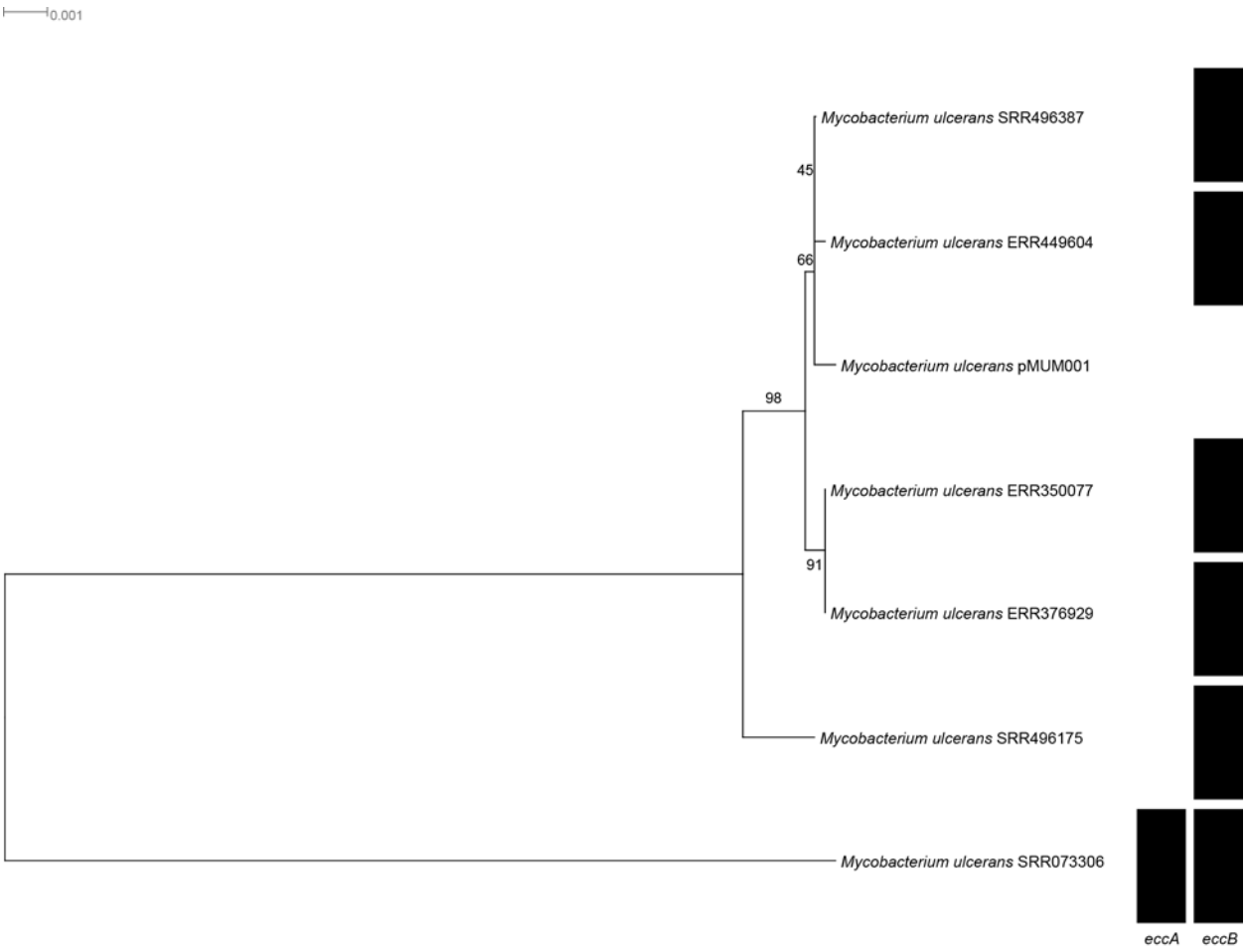

Fig. 7. Core gene phylogeny of *Mycobacterium ulcerans* plasmids and presence of T7SS genes. RAxML was used for phylogenetic inference of a core gene alignment (concatenated amino acid alignments of genes (n = 21) present in all *Mycobacterium ulcerans* plasmids without duplications). The phylogeny is midpoint rooted. Presence of ESX genes is indicated with black boxes.

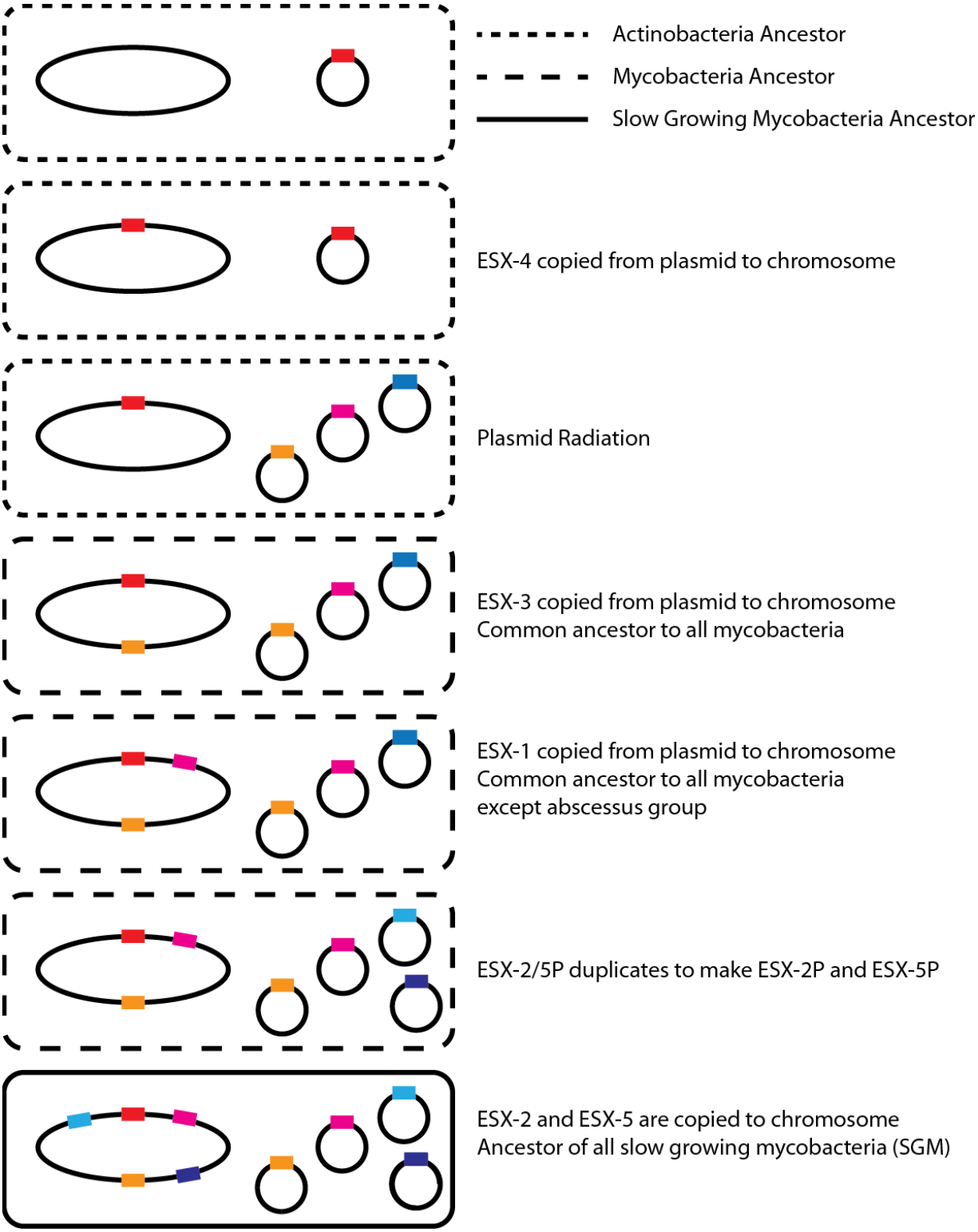

Fig. 8. Model for ESX duplication and migration to the chromosome. ESX loci are colored as follows: Ancestral/ESX-4: red, ESX-3: orange, ESX-1: pink, ESX-2: light blue, ESX-5: dark blue.