bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 secretion system, termed type VII (T7SS), which is encoded by paralogous chromosomal loci (ESX) and associated with pathogenesis, conjugation, and metal homeostasis. Evolution of paralogous gene families is of interest because duplication is an important mechanism by which novel genes evolve, but there are potential conflicts between adaptive forces that stabilize duplications and those that enable evolution of new functions. Our objective was to delineate the adaptive forces underlying diversification of T7SS. 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. Plasmid conjugation has been linked to T7SS and type IV secretion systems (T4SS) in mycobacteria, and we discovered that T7SS and T4SS genes evolved in concert on the plasmids. We hypothesize that differentiation of plasmid ESX helps to prevent conjugation among cells harboring incompatible plasmids. Plasmid ESX appear to have been repurposed following migration to the chromosome, and there is evidence of positive selection driving further differentiation of 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 evolution of novelty, which in this case involved plasmid duplications, selection for advantageous mutations in the mobile and core genomes, migration of the loci between plasmids and chromosomes, and lateral transfer among chromosomes. We discuss further implications 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 & Andersson 2009). Most duplications are transient, due to their intrinsic instability and associated fitness costs, as well as general mutational biases toward deletion (Sandegren & Andersson 2009; Adler et al. 2014).

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

These observations have led researchers to investigate the selective forces allowing duplicate genes to persist and to diverge from the parent gene (Bershtein & Tawfik 2008; Bergthorsson et al. 2007; 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). Selection that stabilizes the initial duplication is likely to preserve the gene's original function, which would be at odds with selection acting to differentiate the gene copies and create new functions. Several solutions have been proposed (Bergthorsson et al. 2007; Hittinger & 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 & Brosch 2009), *eccA*, and *eccE*) as well as a variable complement of locus specific gene content (Figure 1).

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 hypothesize that incompatibility systems have played a role in diversification of plasmid ESX loci, and that ESX loci were initially stabilized on the chromosome by mediating their own lateral transfer.

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Materials and Methods

Data set

We obtained finished genomes from all available *Mycobacterium* species (*n* = 30, as of December 2015) and 23 representative Actinobacteria genomes 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. 2003a) 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 & 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 plasmid sequences available that contain ESX loci, we screened publicly 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

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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. Two hundred and forty eight 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 publicly 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 (21 genes) and phylogeny in a sample (*n* = 7) 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 & Scornavacca 2012) and ggtree (Yu et al. 2016) for tree visualization and editing. Phylogenetic networks were created using Splitstree 4 (Huson & 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 & Huelsenbeck 2003) and visualized tree clusters using Treescape (Kendall & 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 as burn in, randomly sampled 200 trees from each gene, and performed pairwise calculations of the Kendell Colijn metric and multidimensional scaling in Treescape. This analysis was performed on a subset of plasmids encoding the gene *nrdH*, as there were no genes common to all plasmids outside of T7SS and T4SS.

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. The method initially assumes that each branch in the phylogeny can be modeled with only one rate class, and rate classes are added to each branch in a step-wise manner only if there is an improvement in the likelihood of the alignment given the model. The resulting model allows rate variation across branches and sites. Additionally, this method identifies branches on the phylogeny where there is evidence for a proportion of sites to be modeled with an ω (dN/dS) greater than 1 (indicative of positive selection). We tested all branches for positive selection, and branches with a p-value < 0.05 after the Holm-Bonferroni multiple testing correction were considered to have statistically significant evidence for directional selection. Nucleotide sequences from ESX genes were aligned

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

ESX loci are shown in Figure 1. Figure 2 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 (as proposed in (Pittius et al. 2006; Dumas et al. 2016; Newton-Foot et al. 2016)). Interestingly, the loci have been lost on several occasions. For example, ESX-2 was lost in the common ancestor of *M. marinum*, *M. liflandii*, and *M. ulcerans*, and ESX-1 has been lost in *M. sinense*, *M. avium* and related species, as well as from *M. ulcerans*.

Figure 3 shows a network of the ESX loci (see also Figure S1). 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 4).

Plasmid ESX that are basal to chromosomal ESX-1, -3, -2 and -5 have been described previously (Dumas et al. 2016; Newton-Foot et al. 2016). We identified several new, plasmid-borne ESX lineages, including a plasmid lineage that is basal to ESX-4 (Figures 3 and 5B). Each of the six chromosomal ESX is paired with one or more plasmid lineages that root basal to it. The most parsimonious explanation of these observations is that the common ancestor of ESX-1 through -5 was plasmid-associated, that duplication of the ESX loci occurred on plasmids, and that extant chromosomal loci all result from transfers from plasmid to chromosome (Figure 5A). It is possible that the common ancestor of ESX 1-5 was chromosomal, as suggested by Newton-Foot and Dumas et al, but this scenario would require more migration events.

Chromosomal ESX-5, found exclusively among slow growing mycobacteria (SGM), is related to plasmid loci from both SGM and rapid growing mycobacteria (RGM), suggesting that ESX-5 like loci diversified on RGM and SGM plasmids prior to their migration to the chromosome of SGM. Although we did not identify any complete ESX-2 like loci in SGM plasmids, the incomplete ESX locus we identified in *M. ulcerans* (SGM) is most closely related to ESX-2, suggesting that the same may be true of ESX-2.

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (57%, Figure 5B). We speculate that this conflict – i.e. between the species ranges and phylogenetic positions of ESX-1 and ESX-3 – as well as the uncertainty in the phylogeny 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 6. 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 S2). We also found ESX-N in association 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 also been laterally transferred (from plasmid to chromosome and/or between chromosomes) 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 5B), 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 S3). This suggests that T7SS were laterally transferred among mycobacterial species during their divergence, which could contribute to both phylogenetic uncertainty and conflicts with the core genome phylogeny.

There are few reticulations in the ESX networks (Figures 3 and 4), 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

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 7). 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 (internal branches connecting duplicate loci), and the branches leading to chromosomal ESX loci in all cases showed evidence of positive selection. Since migration events could have occurred anywhere along the branches connecting plasmid-associated and chromosome-associated nodes, selection associated with this transition may have acted on plasmid ESX, chromosomal ESX or both. The proportion of sites under positive selection varied substantially, with the highest proportion associated with long, plasmid associated tips. These results were replicated across multiple analyses, including different sampling schemes and alignment trimming methods (Figures S4 and S5).

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 those of the host genomes (Figure 4): 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 4).

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 (plasmid gene content is discussed in more detail below). 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 is 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.

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 9).

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. This extends earlier observations of mycobacterial plasmids encoding an ESX-5 like locus (Ummels et al. 2014). Individual phylogenies of T4SS and T7SS genes were congruent (Figure 8) and distinct from other gene content on the plasmids. This observation is consistent with the conjugation loci having a shared evolutionary history while other loci on the plasmid evolved independently. Congruence among T4SS and T7SS genes also suggests that the paralogous ESX systems trace to whole plasmid duplications. An alternative explanation is that the combined T4SS/T7SS locus was duplicated on individual plasmids. However, the combined locus is ~40kbp in size, and it seems unlikely that such a large duplication would be stable on a plasmid. In addition, we did not identify any plasmids with more than one ESX locus. This modular organization, with congruence among genes involved in conjugation, has been observed in other families of plasmids (Thomas 2000; Fernández-López et al. 2006).

Positive selection was evident on the ESX phylogeny in association with 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, -3, -5) 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.

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

Duplication, divergence, migration, deletion

Much of the prior research on gene duplication has focused on chromosomal duplications, either of the entire chromosome or one of its segments (Lynch & 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 (Figure 3) provides support for the hypothesis that the most recent common ancestor of these loci was plasmid-borne. Our proposed schematic outlining major steps during adaptation of the canonical ESX is shown in Figure 5A. The underlying history is necessarily very simplified in such a schematic, and the model is likely to be modified as further data become available.

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; Figure 3) as well as more recent events (i.e. migration of ESX-N to *Nocardia* and ESX-4-bis to *M. goodii*; Figure 6). A similarly complex history has been observed previously, e.g. in IncW plasmids, where exchange of T4SS genes with the chromosome has occurred on several occasions (Fernández-López et al. 2006).

Adaptation on plasmids: *increased gene dosage, discriminatory conjugation*

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*. *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. Diversification of the conjugative 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 plasmid incompatibility systems: i.e. the conjugation machinery

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

differentiated in order to prevent conjugation between cells harboring incompatible plasmids. Surface exclusion is mechanistically distinct, but related to plasmid incompatibility and could also drive and maintain divergence of associated plasmid loci (Paulsson 2002; Garcillán-Barcia & de la Cruz 2008). Discriminatory transfer to host cells that lack incompatible plasmids is predicted to increase the fitness of the discriminatory plasmid (Paulsson 2002). 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 (Figure 8), 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. Further studies are needed to investigate this hypothesis.

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 (Figure 9). 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 hypothesize that selection on *M. ulcerans* to maintain mycolactone-encoding plasmids relaxes selection on the plasmid to maintain its own conjugative machinery.

Adaptation on the chromosome: lateral transfer and novel functionality

We found evidence of directional selection – i.e. acquisition of specific advantageous mutations - in ESX following their migration to the chromosome (Figure 7). These advantageous mutations are the mechanism by which novel functions for T7SS would have been acquired. As noted above, it is 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 & Tawfik 2008; Sandegren & Andersson 2009; Andersson & Hughes 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 is 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 instability and potentially 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

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 7, Figure S4). 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), 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 & 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*.

Acknowledgments
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].

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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. doi: 10.1093/nar/gkq291.

Abdallah AM 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. doi: 10.1111/j.1365-2958.2009.06783.x.

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. doi: 10.1093/molbev/msu111.

Andersson DI, Hughes D. 2009. Gene Amplification and Adaptive Evolution in Bacteria. Annu. Rev. Genet. 43:167–195. doi: 10.1146/annurev-genet-102108-134805.

Antipov D et al. 2016. plasmidSPAdes: Assembling Plasmids from Whole Genome Sequencing Data. bioRxiv. 48942. doi: 10.1101/048942.

Ates LS et al. 2015. Essential Role of the ESX-5 Secretion System in Outer Membrane Permeability of Pathogenic Mycobacteria. PLoS Genet. 11:e1005190. doi: 10.1371/journal.pgen.1005190.

Barrangou R et al. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 315:1709–1712. doi: 10.1126/science.1138140.

Bergthorsson U, Andersson DI, Roth JR. 2007. Ohno's dilemma: Evolution of new genes under continuous selection. Proc. Natl. Acad. Sci. 104:17004–17009. doi: 10.1073/pnas.0707158104.

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. doi: 10.1093/molbev/msn174.

Bitter W et al. 2009. Systematic Genetic Nomenclature for Type VII Secretion Systems. PLoS Pathog. 5:e1000507. doi: 10.1371/journal.ppat.1000507.

Bottai D et al. 2012. Disruption of the ESX-5 system of Mycobacterium tuberculosis causes loss of PPE protein secretion, reduction of cell wall integrity and strong attenuation. Mol. Microbiol. 83:1195–1209. doi: 10.1111/j.1365-2958.2012.08001.x.

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. doi: 10.1111/j.1365-2958.2009.06784.x.

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Bruen TC, Philippe H, Bryant D. 2006. A Simple and Robust Statistical Test for Detecting the Presence of Recombination. Genetics. 172:2665–2681. doi: 10.1534/genetics.105.048975.

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. doi: 10.1111/j.1365-2958.2008.06299.x.

DePristo MA et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–498. doi: 10.1038/ng.806.

Dumas E et al. 2016. Mycobacterial pan-genome analysis suggests important role of plasmids in the radiation of type VII secretion systems. Genome Biol. Evol. evw001. doi: 10.1093/gbe/evw001.

Elde NC et al. 2012. Poxviruses Deploy Genomic Accordions to Adapt Rapidly against Host Antiviral Defenses. Cell. 150:831–841. doi: 10.1016/j.cell.2012.05.049.

Fernández-López R et al. 2006. Dynamics of the IncW genetic backbone imply general trends in conjugative plasmid evolution. FEMS Microbiol. Rev. 30:942–966. doi: 10.1111/j.1574-6976.2006.00042.x.

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. doi: 10.1073/pnas.0404892101.

Garces A 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. doi: 10.1371/journal.ppat.1000957.

Garcillán-Barcia MP, de la Cruz F. 2008. Why is entry exclusion an essential feature of conjugative plasmids? Plasmid. 60:1–18. doi: 10.1016/j.plasmid.2008.03.002.

Garneau JE et al. 2010. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. 468:67–71. doi: 10.1038/nature09523.

George KM et al. 1999. Mycolactone: A Polyketide Toxin from Mycobacterium ulcerans Required for Virulence. Science. 283:854–857. doi: 10.1126/science.283.5403.854.

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. doi: 10.1371/journal.pbio.1001602.

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Gremme G, Steinbiss S, Kurtz S. 2013. GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. IEEE/ACM Trans. Comput. Biol. Bioinform. 10:645–656. doi: 10.1109/TCBB.2013.68.

He L, Fan X, Xie J. 2012. Comparative genomic structures of Mycobacterium CRISPR-Cas. J. Cell. Biochem. 113:2464–2473. doi: 10.1002/jcb.24121.

Hittinger CT, Carroll SB. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. Nature. 449:677–681. doi: 10.1038/nature06151.

Houben ENG et al. 2012. Composition of the type VII secretion system membrane complex. Mol. Microbiol. 86:472–484. doi: 10.1111/j.1365-2958.2012.08206.x.

Huson DH, Bryant D. 2006. Application of Phylogenetic Networks in Evolutionary Studies. Mol. Biol. Evol. 23:254–267. doi: 10.1093/molbev/msj030.

Huson DH, Scornavacca C. 2012. Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. Syst. Biol. sys062. doi: 10.1093/sysbio/sys062.

Katoh K, Standley DM. 2014. MAFFT: iterative refinement and additional methods. Methods Mol. Biol. Clifton NJ. 1079:131–146. doi: 10.1007/978-1-62703-646-7_8.

Kendall M, Colijn C. 2016. Mapping phylogenetic trees to reveal distinct patterns of evolution. Mol. Biol. Evol. msw124. doi: 10.1093/molbev/msw124.

Kreuger F. 2013. *TrimGalore!* 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. http://arxiv.org/abs/1303.3997 (Accessed September 29, 2015).

Li L, Stoeckert CJ, Roos DS. 2003a. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 13:2178–2189. doi: 10.1101/gr.1224503.

Li L, Stoeckert CJ, Roos DS. 2003b. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 13:2178–2189. doi: 10.1101/gr.1224503.

Lynch M, Conery JS. 2000. The Evolutionary Fate and Consequences of Duplicate Genes. Science. 290:1151–1155. doi: 10.1126/science.290.5494.1151.

McLaughlin B et al. 2007. A mycobacterium ESX-1-secreted virulence factor with unique requirements for export. PLoS Pathog. 3:e105. doi: 10.1371/journal.ppat.0030105.

Mehra A et al. 2013. Mycobacterium tuberculosis Type VII Secreted Effector EsxH Targets Host ESCRT to Impair Trafficking. PLoS Pathog. 9:e1003734. doi: 10.1371/journal.ppat.1003734.

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Näsvall J, Sun L, Roth JR, Andersson DI. 2012. Real-Time Evolution of New Genes by Innovation, Amplification, and Divergence. Science. 338:384–387. doi: 10.1126/science.1226521.

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. doi: 10.1186/s12862-016-0631-2.

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.

Paulsson J. 2002. Multileveled Selection on Plasmid Replication. Genetics. 161:1373–1384.

Pittius NCG van et al. 2006. Evolution and expansion of the Mycobacterium tuberculosis PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. BMC Evol. Biol. 6:95. doi: 10.1186/1471-2148-6-95.

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. doi: 10.1046/j.1365-2958.2002.03237.x.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574. doi: 10.1093/bioinformatics/btg180.

San Millan A et al. 2014. Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. Nat. Commun. 5:5208. doi: 10.1038/ncomms6208.

Sandegren L, Andersson DI. 2009. Bacterial gene amplification: implications for the evolution of antibiotic resistance. Nat. Rev. Microbiol. 7:578–588. doi: 10.1038/nrmicro2174.

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. doi: 10.1093/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. doi: 10.1093/nar/gkv318.

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. doi: 10.1128/JB.00756-09.

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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. doi: 10.1371/journal.pone.0078351.

Shah S, Briken V. 2016. Modular Organization of the ESX-5 Secretion System in Mycobacterium tuberculosis. Front. Cell. Infect. Microbiol. 49. doi: 10.3389/fcimb.2016.00049.

Siegrist MS et al. 2009. Mycobacterial Esx-3 is required for mycobactin-mediated iron acquisition. Proc. Natl. Acad. Sci. 106:18792–18797. doi: 10.1073/pnas.0900589106.

Smith MD et al. 2015. Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. Mol. Biol. Evol. 32:1342–1353. doi: 10.1093/molbev/msv022.

Sreejit G et al. 2014. The ESAT-6 Protein of Mycobacterium tuberculosis Interacts with Beta-2-Microglobulin (β2M) Affecting Antigen Presentation Function of Macrophage. PLoS Pathog. 10:e1004446. doi: 10.1371/journal.ppat.1004446.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313. doi: 10.1093/bioinformatics/btu033.

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.

Thomas CM. 2000. Paradigms of plasmid organization. Mol. Microbiol. 37:485–491. doi: 10.1046/j.1365-2958.2000.02006.x.

Tufariello JM et al. 2016. Separable roles for Mycobacterium tuberculosis ESX-3 effectors in iron acquisition and virulence. Proc. Natl. Acad. Sci. 201523321. doi: 10.1073/pnas.1523321113.

Ummels R et al. 2014. Identification of a Novel Conjugative Plasmid in Mycobacteria That Requires Both Type IV and Type VII Secretion. mBio. 5:e01744-14. doi: 10.1128/mBio.01744-14.

Wang H et al. 2016. Increased plasmid copy number is essential for Yersinia T3SS function and virulence. Science. aaf7501. doi: 10.1126/science.aaf7501.

Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2016. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. n/a-n/a. doi: 10.1111/2041-210X.12628.

Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol. Evol. 18:292–298. doi: 10.1016/S0169-5347(03)00033-8.

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Zhang M et al. 2014. EspI regulates the ESX-1 secretion system in response to ATP levels in Mycobacterium tuberculosis. Mol. Microbiol. n/a-n/a. doi: 10.1111/mmi.12718.

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure Legends

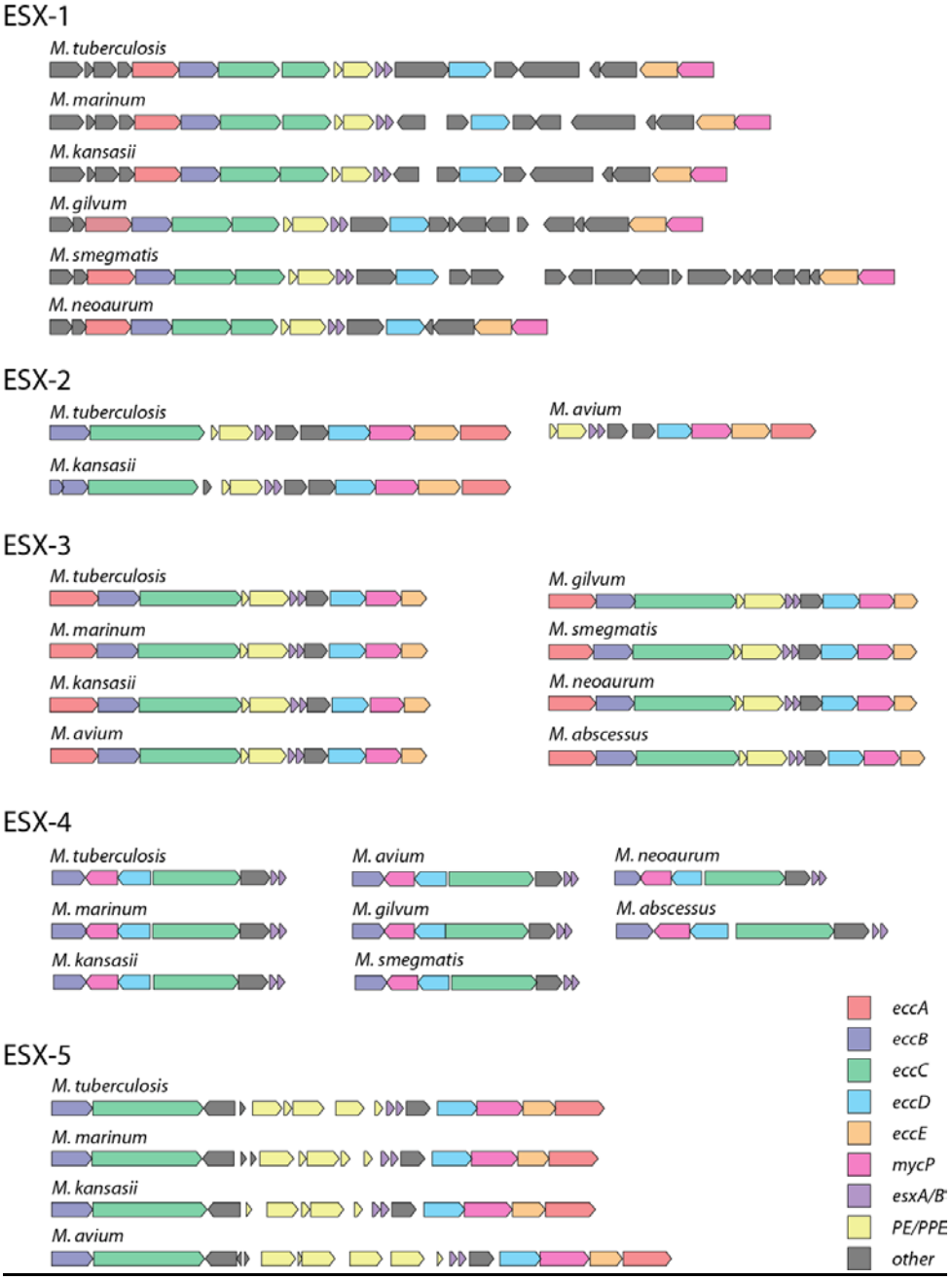

Fig. 1. Mycobacterial chromosomal ESX loci. Core gene content in the ESX loci are colored as follows: *eccA*- red, *eccB*- dark blue, *eccC*- green, *eccD*- light blue, *eccE*- orange, *mycP*- pink, *esxA/B*- purple, *PE/PPE*- yellow. Other variable genes in the loci are black. Orthologs and paralogs are based on OrthoMCL (Li et al. 2003b) output. Locus diagrams were made using GenomeTools (Gremme et al. 2013). Each locus has a distinct structure, which developed during adaptation on mycobacterial plasmids and chromosomes (see text for details).

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

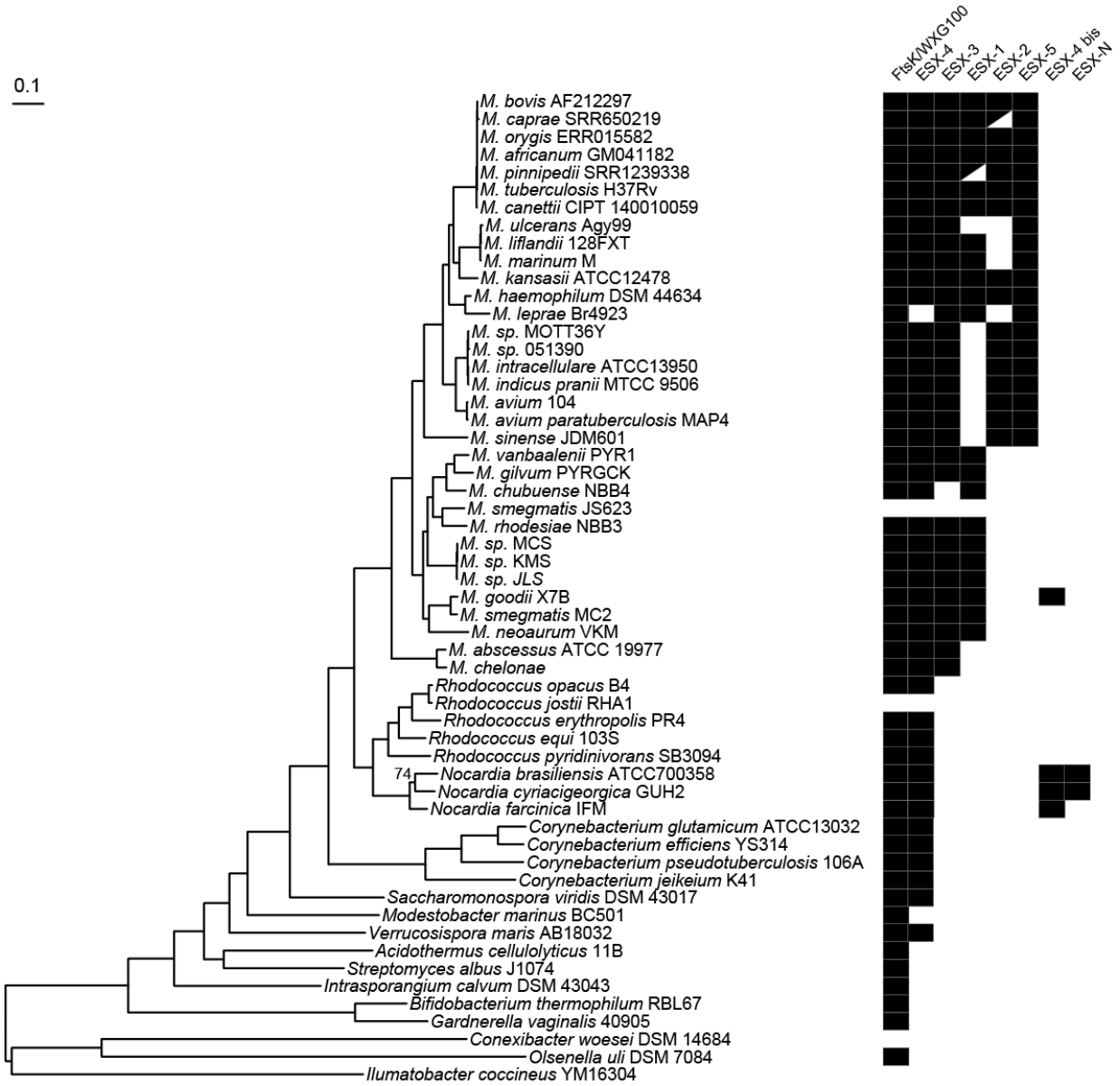

Fig. 2. Maximum likelihood phylogeny of Actinobacteria with presence/absence matrix of type VII secretion system loci. RAxML was used for phylogenetic inference of the Actinobacteria core genome alignment (concatenated amino acid alignments of genes (n = 171) present in all genomes without duplications). The phylogeny is midpoint rooted, and branches without labels have a boostrap value of 100. Presence of ESX loci is indicated with black boxes. We have abbreviated the genus *Mycobacterium* in the tip labels. Some *M. tuberculosis* complex (MTBC) species have characteristic deletions located in ESX loci. Partially deleted ESX loci 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*. Patterns of ESX presence/absence are consistent with an initial emergence of the FtsK/WXG100 gene cluster, followed by ESX-4 bis and ESX-4, with subsequent duplications giving rise to ESX-3, ESX-1, ESX-2 and ESX-5.

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

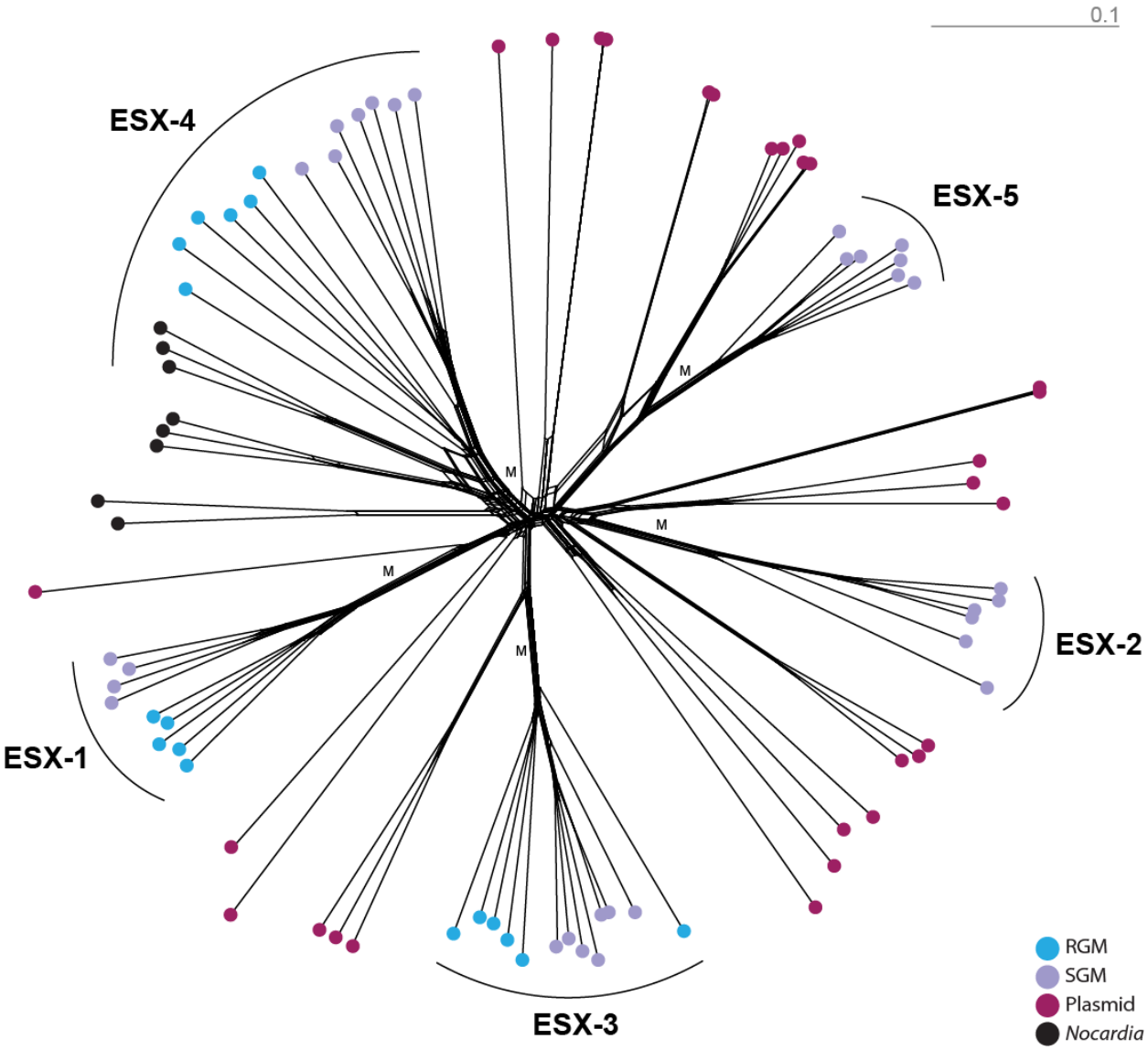

Fig. 3. Network of ESX loci in mycobacteria, *Nocardia*, and mycobacterial plasmids. The network was created in SplitsTree4 from a concatenated alignment of *eccA, eccB, eccC, eccD, eccE,* and *mycP*. Light blue dots correspond to ESX loci from rapid growing mycobacterial chromosomes (RGM), light purple dots correspond to ESX loci from slow growing mycobacterial chromosomes (SGM), magenta dots correspond to ESX loci from mycobacterial plasmids, and black dots correspond to ESX loci from *Nocardia* chromosomes. All chromosomal ESX (ESX-1-5) have basal plasmid loci, suggesting that the ancestral ESX locus was plasmid-borne (putative location of migration events to the chromosome marked 'M' on the network). The PHI test was insignificant (p = 1.0) for this alignment, indicating that there was no evidence for intra-locus recombination. A version of this figure with tip labels is available in the supplement (Figure S1).

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

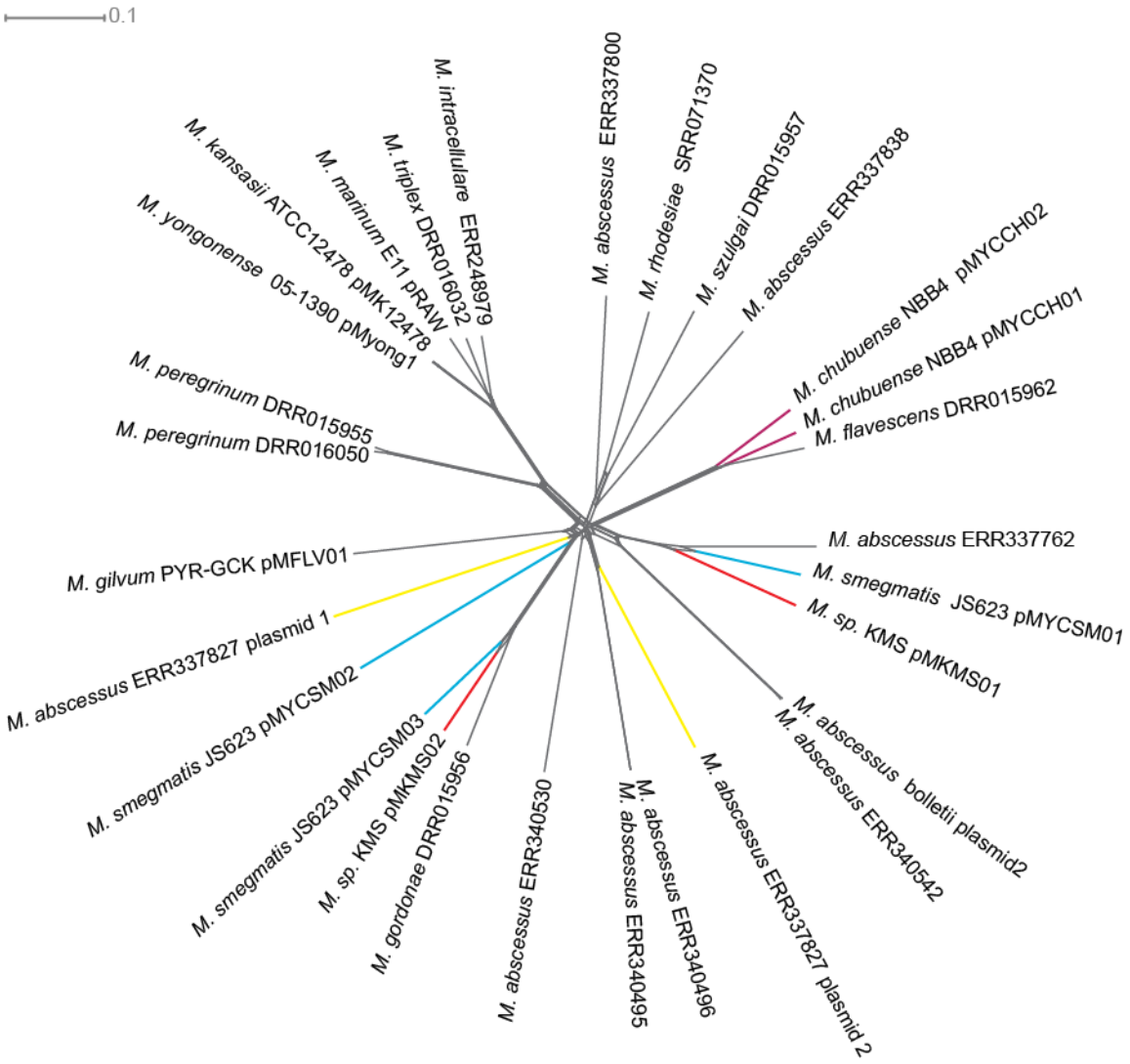

Fig. 4. Network of plasmid-borne ESX loci. The network was created in SplitsTree4 from a concatenated alignment of *eccA, eccB, eccC, eccD, eccE,* and *mycP* from ESX loci encoded on mycobacterial plasmids. The star-like appearance of the network is consistent with rapid diversification of this gene family on the plasmids. Some bacterial strains harbored multiple plasmids, and these are indicated with colored branches. The phylogenetic relationships among plasmid 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. This suggests that plasmid ESX diversification has not been shaped by adaptation to bacterial host species. We did not find evidence of intra-locus recombination in this alignment with the PHI test (p = 1.0).

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

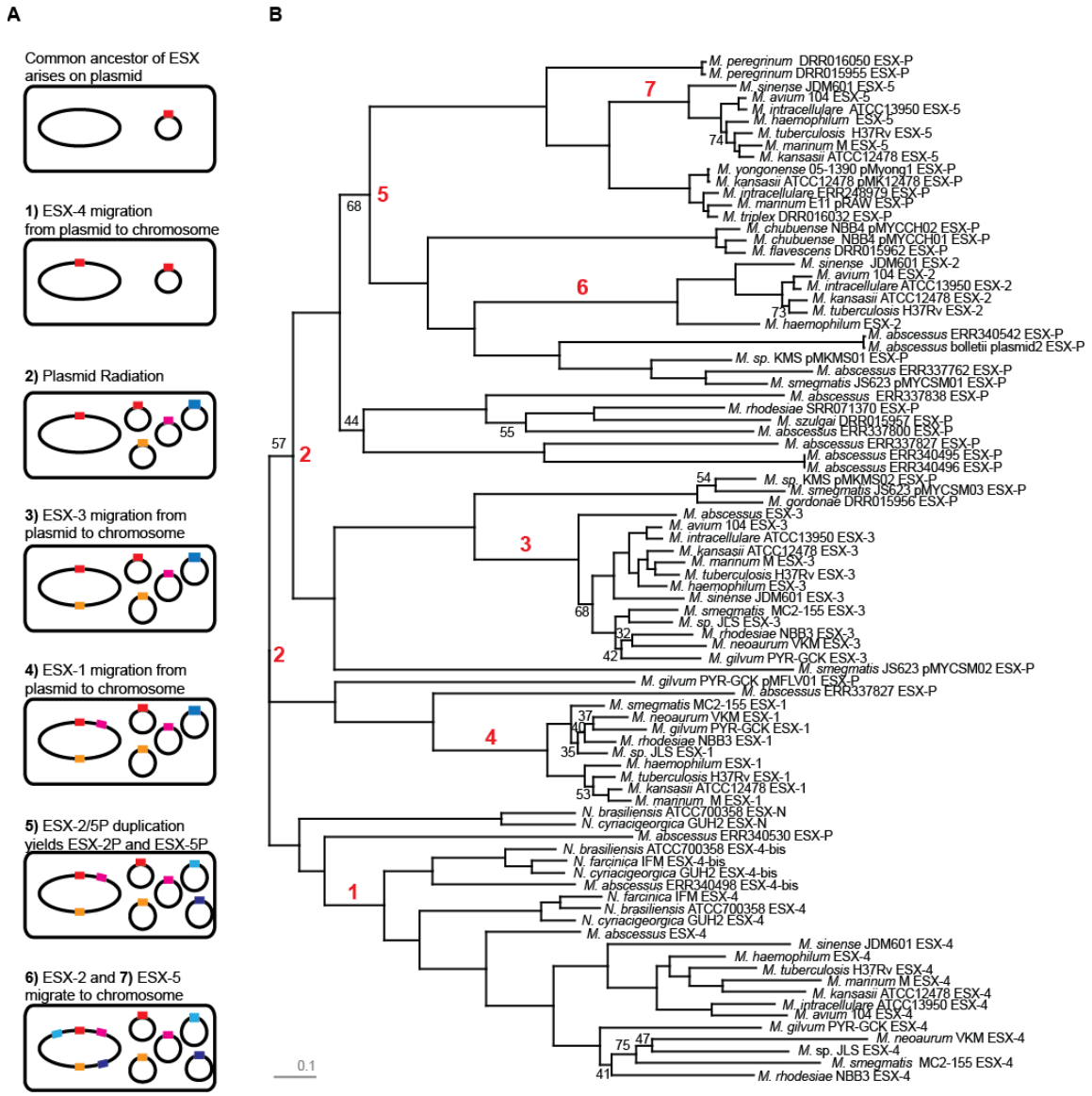

Fig. 5. ESX plasmid-mediated duplication and migration to the chromosome. A) Simplified schematic of major steps in the evolutionary history of mycobacterial ESX loci. ESX loci are colored as follows: Ancestral/ESX-4: red, ESX-3: orange, ESX-1: pink, ESX-2: light blue, ESX-5: dark blue. B) Maximum likelihood phylogeny of ESX loci (*eccA*, *eccB*, *eccC*, *eccD*, *eccE*, and *mycP*) in mycobacteria, *Nocardia*, and mycobacterial plasmids. Branches without black labels have a bootstrap value greater than 75. Red labels correspond to events presented in the schematic. ESX-N, found on the chromosomes of some *Nocardia* species, appear to have been recently transferred from a plasmid (see text). ESX-N and other plasmid-associated ESX are basal to chromosomal ESX 1-5. This suggests that their common ancestor was plasmid-borne and that extant chromosomal loci trace to migrations from plasmid to chromosome. The model outlined here is highly simplified: for example, there were likely several migrations of ESX-4 like loci to the chromosome (step 1 in the schematic) and the chromosomal loci show a mixture of vertical and horizontal inheritance (details in text).

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

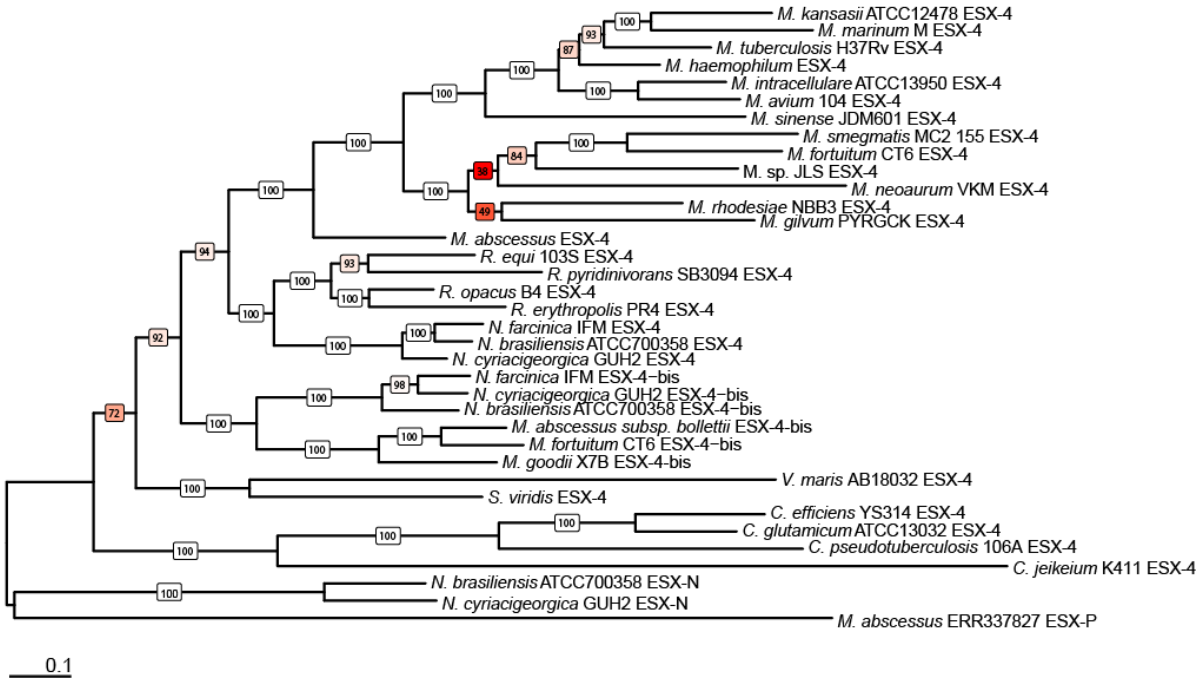

Fig. 6. Maximum likelihood phylogeny of ESX-4 in Actinobacteria. The phylogeny is rooted using ESX-N and a basal plasmid-borne ESX locus. Bootstrap values are colored based on support (white = 100, red = lowest support). The location of chromosomal *Corynebacterium* ESX-4 and *M. goodii* ESX-4-bis are in conflict with the core genome phylogeny (Figure 2). In the core genome phylogeny, *Corynebacterium* is more closely related to *Nocardia* and *Rhodococcus* than *Verrucosispora* or *Saccharamonaspora.* However, in the ESX-4 phylogeny, this relationship is reversed. Based on the relationship of *Mycobacterium* species in the core genome phylogeny, we would expect *M. goodii* ESX-4-bis and *M. fortuitum* ESX-4-bis to be more closely related to one another than either is to *M. abscessus* (the most basal *Mycobacterium* species). These conflicts suggest that chromosomal ESX-4 like loci have been laterally transferred among species. Genus names have been abbreviated, but full length names can be found in the core genome phylogeny (Figure 2).

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

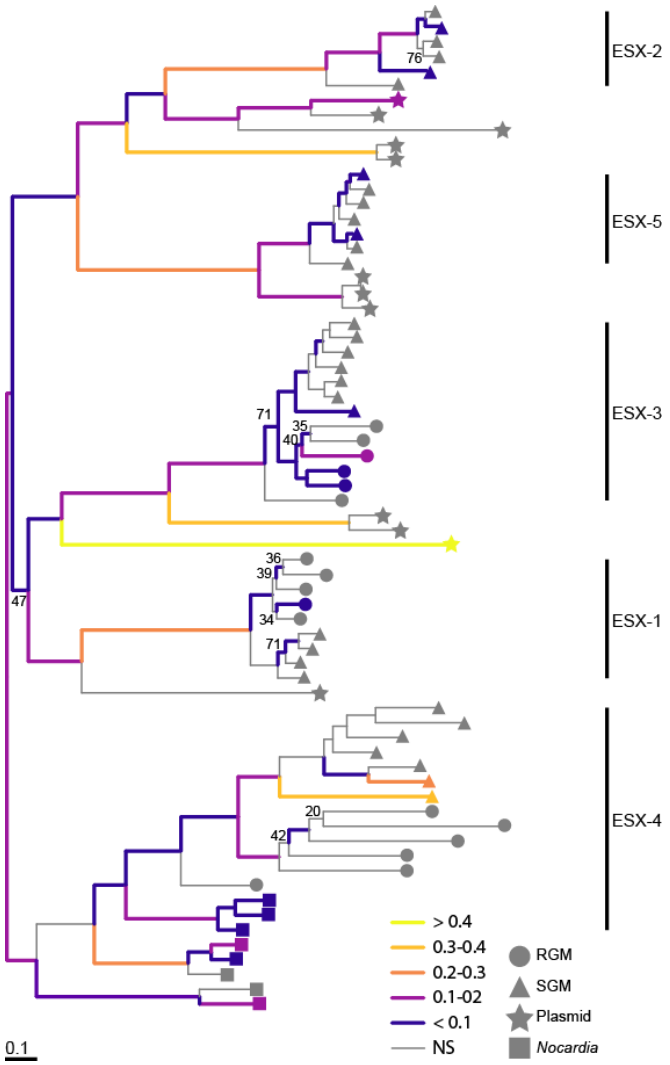

Fig. 7. Episodic directional selection during the evolution of ESX loci. Maximum likelihood phylogeny inferred using RAxML from a concatenated alignment of *eccA, eccB, eccC, eccD, eccE,* and *mycP*. In order to minimize potential effects of misalignment on inference of selection, only data from finished genomes were included in this analysis: see Figures 3 and 5B for network and phylogenetic analyses of the complete dataset. Plasmid associated taxa for ESX-4, ESX-2 and ESX-5 are not shown on this phylogeny for this reason. Branches in this phylogeny without labels have a bootstrap value greater than 75. We used the aBSREL test implemented in HyPhy to identify branches with significant evidence ($p < 0.05$) of episodic directional selection; these branches are colored based on the proportion of sites affected by positive selection ($\omega > 1$). Circles correspond to ESX loci from rapid growing mycobacterial chromosomes, triangles correspond to ESX loci from slow growing mycobacterial chromosomes, stars correspond to ESX loci from mycobacterial plasmids, and squares correspond to ESX loci from *Nocardia* chromosomes. A version of this figure is included in the supplement (fig S4) that shows the tips labeled with species names. There is evidence of directional (positive) selection at each duplication event (short internal branches), as well as on the branches leading to the extant chromosomal loci.

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

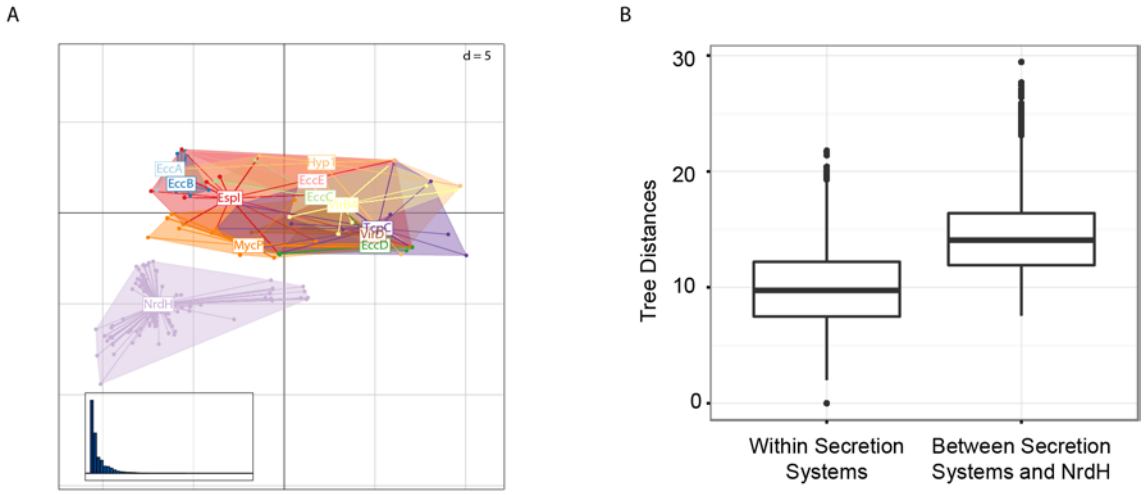

Fig. 8. Congruence of tree topologies of plasmid encoded secretion systems. Bayesian phylogenetic analysis was performed in MrBayes using amino acid alignments of EccA, EccB, EccC, EccD, EccE, MycP, VirB4, VirD, TcpC, EspI, NrdH, and a hypothetical protein (Hyp1) encoded proximal to known T4SS genes. A) 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 and third principal components from the MDS. The inset bar chart is a scree plot showing the eigenvalues for the principal components. The T4SS and T7SS gene trees overlap in the MDS, whereas topologies of NrdH gene trees are incongruent with those of T4SS and T7SS. This suggests that plasmid-encoded T4SS and T7SS have co-diverged during their evolutionary history and that they have evolved independently of other gene content on the plasmids. B) Kendell Colijn distances among secretion system gene trees and between secretion system and NrdH gene trees. The means of these distributions are significantly different ($p < 2.2 \times 10^{-16}$) according to a Mann-Whitney U test.

bioRxiv preprint doi: https://doi.org/10.1101/067207; this version posted September 30, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

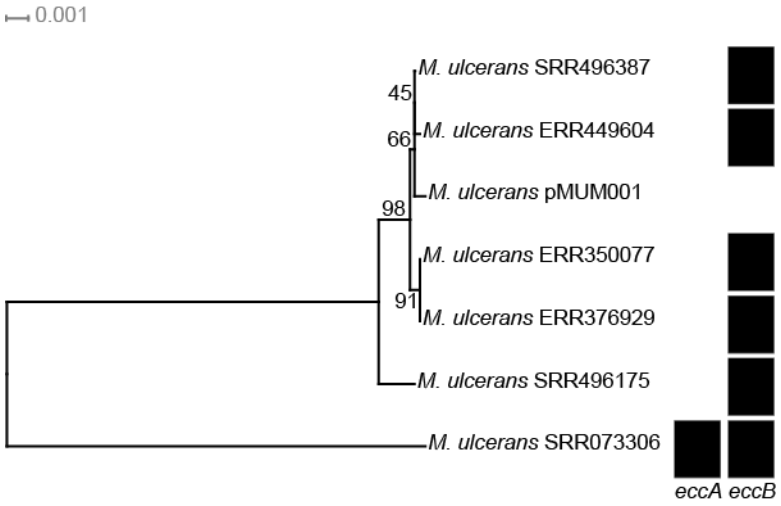

Fig. 9. Core gene phylogeny of *Mycobacterium ulcerans* plasmids and presence/absence of T7SS genes. RAxML was used for phylogenetic inference of a core gene alignment (concatenated amino acid alignments of genes (n = 21) present in all *Mycobacterium ulcerans* plasmids without duplications). The phylogeny is midpoint rooted. Presence of ESX genes is indicated with black boxes. ESX genes found on *M. ulcerans* plasmids are most closely related to ESX-2. The most basal *M. ulcerans* plasmid encodes both *eccA* and *eccB*. Most *M. ulcerans* plasmids only encode *eccB*, and some have lost all ESX genes (e.g. pMUM001). The locus shows other signs of degradation (discussed in the text). We hypothesize that selection to maintain the conjugation machinery of these plasmids has been relaxed as a result of host selection for its other gene content, likely mycolactone.