SCIENTIFIC REPORTS

Corrected: Author Correction

OPEN

Extensive genomic diversity among *Mycobacterium marinum* strains revealed by whole genome sequencing

Received: 26 March 2018
Accepted: 25 July 2018
Published online: 13 August 2018

Sarbashis Das[1], B. M. Fredrik Pettersson[1], Phani Rama Krishna Behra[1], Amrita Mallick[2], Martin Cheramie[2], Malavika Ramesh[1], Lisa Shirreff[2], Tanner DuCote[2], Santanu Dasgupta[1], Don G. Ennis[2] & Leif. A. Kirsebom[1]

Mycobacterium marinum **is the causative agent for the tuberculosis-like disease mycobacteriosis in fish and skin lesions in humans. Ubiquitous in its geographical distribution, *M. marinum* is known to occupy diverse fish as hosts. However, information about its genomic diversity is limited. Here, we provide the genome sequences for 15 *M. marinum* strains isolated from infected humans and fish. Comparative genomic analysis of these and four available genomes of the *M. marinum* strains M, E11, MB2 and Europe reveal high genomic diversity among the strains, leading to the conclusion that *M. marinum* should be divided into two different clusters, the "M"- and the "Aronson"-type. We suggest that these two clusters should be considered to represent two *M. marinum* subspecies. Our data also show that the *M. marinum* pan-genome for both groups is open and expanding and we provide data showing high number of mutational hotspots in *M. marinum* relative to other mycobacteria such as *Mycobacterium tuberculosis*. This high genomic diversity might be related to the ability of *M. marinum* to occupy different ecological niches.**

The genus *Mycobacterium* comprises more than 177 species including, pathogenic, opportunistic pathogens and non-pathogenic environmental species. Mycobacteria are free-living, acid fast, robust organisms, which can sustain themselves, and even thrive, in widely diverse environments ranging from soil and tap water to animals and humans. Of interest to this study, *Mycobacterium marinum* (*Mma*) was first described in 1926 by Joseph D. Aronson[1]. The bacterium was isolated from infected fish suffering from mycobacteriosis exhibiting similarities with tuberculosis in humans. Later, it was shown that *Mma* could also infect humans and cause skin lesions at body extremities[2]. Phylogenetic analysis based on the 16S rRNA gene sequences suggested that *Mycobacterium tuberculosis* (*Mtb*) and *Mycobacterium ulcerans* are its closest neighbours[3] and *Mma* is a member of the *M. ulcerans* clade. Together with *M. ulcerans* it constitutes the *M. ulcerans*-*M. marinum* complex, MuMC[1,4,5].

The complete genome of the *M. marinum* M strain was released in 2008. The genome size is 6.5 Mb, which is 2.1 Mb longer than the *Mtb* genome. Comparison studies showed that *Mma* and *Mtb* have more than 85% genome similarity and share major virulence factors[6]. The *M. ulcerans* genome is roughly one Mb smaller than the *Mma* M strain genome[7] but their genomes are highly similar[8]. They share a common ancestor and available genetic data suggest that *M. ulcerans* diverged from *M. marinum*[9]. The MuMC and the evolution of *M. ulcerans* and its capacity to produce the toxin mycolactone have been studied using comparative genomic approaches[7,8]. However, only a few *Mma* genomes are available and therefore knowledge of the genomic diversity as well as the global *Mma* gene repository is rather limited. Hence, to understand the evolutionary relationship and genomic diversity of *Mma* we decided to determine the genome sequences of strains isolated from different sources. In our analysis we also included four published *Mma* genomes of the Europe (acc no. ANPL00000000), MB2 (acc no. ANPM01000000), E11 (acc no. HG917972.2)[10], and M (GCA_000018345)[6] strains.

Here we provide genomic data for 15 different *Mma* strains including type strains and isolates from infected fishes and humans from different geographical regions. Comparative analysis of 19 genomes suggests two distinct

[1]Department of Cell and Molecular Biology, Box 596, Biomedical Centre, SE-751 24, Uppsala, Sweden. [2]Department of Biology, University of Louisiana, Lafayette, Louisiana, USA. Correspondence and requests for materials should be addressed to L.A.K. (email: Leif.kirsebom@icm.uu.se)

Mma types (or lineages) that share a common ancestor. This raises the question that the lineage to which the *Mma* M strain belongs and the other lineage, referred to as the "Aronson-lineage", should be considered as two separate subspecies. Consistent with this notion, during the course of evolution, "Aronson-lineage" members acquired an additional ribosomal operon likely through duplication. We have identified the presence of plasmid sequences, various IS elements and their distribution, as well as sequences of phage origin and their translocation in the different strains. Additionally, we characterized the *Mma* pan-genome, the phylogenetic relationship and identified mutational hot spots. Altogether, these data provide insight into the evolutionary mechanisms of mycobacterial strain diversification.

Results

Whole genome sequencing was performed for 15 *Mma* strains (Table 1). Of these, the type strains *Mma* CCUG20998 (hereafter referred to as CCUG) and *Mma* 1218R (referred to as 1218R) were sequenced using Pacific Biosciences (PacBio) technology and the remaining 13 strains with Illumina sequencing technology. In addition, in our comparative analysis we included four published *Mma* genome sequences referred to as M, E11, Europe and MB2[10]. The genomes of the M and E11 strains are complete while the other two are draft genomes.

Overview of genomic features. *De novo* assembly of the long reads (average length 10 kb) derived from PacBio platform generated complete genomes (CCUG and 1218R) comprised of single scaffolds while assembly of the Illumina sequencing reads (average length 100 bp) for the other *Mma* strains resulted in near complete genomes split into multiple scaffolds. Sequencing read statistics, average read depths and assembly qualities are shown in Fig. S1. The average GC-content is 65%, in keeping with the 65.7% determined for the M strain[6] while the genome length ranged from 5.7 Mb (DL240490) to 6.6 Mb (M) representing 5343 to 5573 coding sequences (CDS; Fig. 1a and Table S1). The number of predicted tRNA genes varied between 46 and 53 with 1218R having the highest number. Interestingly, the complete genomes of three strains (1218R, CCUG and E11) carry six ribosomal genes, corresponding to two ribosomal RNA operons (rRNA; 16S rRNA, 23S rRNA and 5S rRNA), while the M strain carries only one rRNA operon (Fig. 1b). For the draft genomes, we predicted the presence of two 5S rRNA genes for the strains closely related to CCUG and 1218R (see below) while only one was predicted for the others (see discussion). No tRNA genes were predicted within any of the rRNA operons. Depending on strain, we also predicted the presence of 45 to 74 non-coding (nc) RNA genes (Table S1).

Whole genome alignment for the 19 genomes suggested that the genomes are well conserved without major genomic rearrangements (Fig. 1c). However, two regions of genomic variations were apparent: the "1.5–2 Mb" and "3.5–4.5 Mb" regions, in all the strains. A likely reason for this variation is the presence of different phage sequences in these regions. For example, the M strain carries a fragment in the 1.4 Mb region, which is conserved in E11, MSS4, and KST but absent in the other strains. Similarly, CCUG, NCTC2275, DSM44344 and DSM43519 have two conserved prophages located at 4.3 Mb and 4.7 Mb. The DSM43518 strain was predicted to have three prophages. Two of these (located at 4.5 Mb and 5.5 Mb) were also detected in two other strains, the "4.5 Mb-phage" in Davis-1 and the "5.5 Mb-phage" in VIMS-9 (Fig. 1c). Moreover, two inversions were detected in the Europe genome and one of these covers nearly 5.5 Mb. Genome alignment, however, suggests that this inversion is likely the result of scaffolding of the contigs.

With the exception of DE4381 (also called "1218S") and DE4576 (referred to as "Huestis"), the 15 new *Mma* genomes were predicted to have plasmid sequences (Table S2; see Methods). But, no plasmid was present in the MB2 and Europe strains, while CCUG and 1218R harboured complete circular plasmids encompassing 127 kb and 130 kb, respectively. Of the draft genomes, DSM43519 was predicted to have the largest plasmid of 181 kb encoding 161 genes. The average GC-content for the plasmid scaffolds is 63.9%, which is lower than for the chromosomal sequences (see above) and the plasmid present in the M strain (67.9%). Plasmid alignment revealed that plasmids/plasmid sequences could be grouped into four types: (I) CCUG, NCTC2275, DSM43519, VIMS-9, DSM44344, Davis-1 and 1218R carry the same plasmid, which is similar to the plasmid present in E11, (II) present in MSS2 and MSS4, (III) BB170200 and DL240490 have similar plasmid fragments and the sequences show high similarity compared to the pMUM003 plasmid previously reported to be present in *Mma* DL240490[10], and (IV) pMM23 is present only in the M strain (Fig. 1d)[6]. Interestingly, the type (I) plasmid carries genes encoding for two secretion systems, type IV and type VII, where the type VII genes show high homology to the ESX-5 category. The significance of the presence of these genes remains to be studied but it is noteworthy that ESX-5 has been suggested to be associated with slow growing pathogenic mycobacteria and to have an impact on virulence[11]. These findings are in keeping with previously published data[10,12–14].

Average nucleotide identity revealed two distinct *M. marinum* strain clusters. The average nucleotide identity (ANI) value, which is useful for discriminating between species and strains[15], was calculated pairwise for the homologous regions in the 19 *Mma* genomes and the two phylogenetically closest neighbours *M. ulcerans* and *Mycobacterium liflandii*. Although the ANI values for any pairs are higher than 97% (Fig. 2a), hierarchical clustering on the basis of the ANI values resulted in two clusters: cluster I including the M, Huestis, MB2, DL240490, BB170200 and MSS2 strains while cluster II encompasses the remaining strains including E11, 1218R and CCUG (Fig. 2a and b). Strains belonging to cluster II show significant similarity (>98.5% ANI score), while for cluster I strains the ANI scores range from 97 to 99%. We interpret this difference to reflect higher genomic diversity among cluster I members. Including *M. ulcerans* and *M. liflandii* revealed high ANI scores comparing either of these two species with several of the strains in cluster I, *e.g.*, ≈99% comparing *M. liflandii* and BB170200 or DL240490. Hence, it appears that *M. ulcerans* and *M. liflandii* are evolutionarily closer to cluster I strains than to cluster II members. In summary, *Mma* strains can be grouped into two clusters, albeit that all the ANI scores are high and above species threshold (97%)[15].

Strains	Host	Isolated from	Accession no	Comments	Reference
CCUG20998	Salt water fish	Philadelphia, USA	CP024190	Derivative of the *Mma* Aronson isolate. Strain collection passage variant.	Strain collection, Gothenborg, Sweden
BB170200	Silver perch (*Bidyanus bidyanus*)	Israel, freshwater, cultured	PEDI00000000		ref.[63]
DL240490	European sea bass (*Dicentrarchus labrax*)	Israel, marine (RS), cultured	PEDJ00000000		ref.[63]
NCTC2275	Salt water fish		PEDD00000000	Derivative of the *Mma* Aronson isolate. Strain collection passage variant.	Dr B. Herrmann, Uppsala Academic Hospital, Sweden
MSS4	Human skin lesions infection during outbreak in an aquaculture facility	Mississippi, USA	PEDG00000000	Clinical isolate.	ref.[44]
MSS2	Hybrid striped bass outbreak in an aquaculture facility	Mississippi, USA	PEDH00000000	Clinical isolate.	ref.[44]
Davis-1	Farmed striped bass	Davis California, USA	PEDF00000000	Clinical isolate.	ref.[44]
VIMS9	Striped bass in the wild	Virginia, USA	PECY000000	Clinical isolate.	ref.[44]
KST-214	Hybrid striped bass, an outbreak at Kent Sea Tech aquaculture facility	Central Valley of California, USA	PEDE00000000	Clinical isolate.	ref.[44]
DSM44344	Salt water fish	Philadelphia, USA	PEDC00000000	Derivative of the *Mma* Aronson isolate. Strain collection passage variant.	DSMZ strain collection
DSM43518	Salt water fish	Philadelphia, USA	PEDA00000000	Derivative of the *Mma* Aronson isolate. Strain collection passage variant.	DSMZ strain collection
DSM43519	Salt water fish	Philadelphia, USA	PEDB00000000	Derivative of the *Mma* Aronson isolate. Strain collection passage variant.	DSMZ strain collection
1218R	Salt water fish	Philadelphia, USA	CP025779	Derivative of *Mma* Aronson, TMC 1218, rough colony morphology variant (1218R).	Trudeau Mycobacterial Collection (TMC), ref.[64]
DE4381	Salt water fish	Philadelphia, USA	PECZ00000000	Derivatives of TMC 1218 smooth colony morphology variant (1218S).	P. L. Small, L. P. Barker and D. G. Ennis
DE4576/ Huestis	Zebrafish (*Danio rerio*), outbreak of the ZIRC National zebrafish facility	Oregon, USA	PEDK00000000	Clinical isolate ("Heustis").	K. Guillemin and D. G. Ennis
M	Human patient	San Francisco, USA	GCA_000018345.1		ref.[6]
E11	European sea bass (*Dicentrarchus labrax*)	Israel	GCA_000723425.2		refs[10,65]
MB2	Fish	Thailand	GCA_000419335.1		ref.[66]
Europe	Fish	Europe	GCA_000419315.1		ref.[66]

Table 1. *Mma* strains source, apparent derivation that was employed in this study. List of the *Mma* strains and corresponding sources of isolation.

Cluster I and II pan-genome and core-genome sizes are different. The pan-genome includes all genes identified in all members of a species while the core-genome represents the set of genes present in all species members[16]. We used power law regression analysis:

$$y = A_{pan}x^{B_{pan}} + C_{pan} \quad (1)$$

to model the distribution of the pan-genome where y = pan-genome size, x = number of genomes, A_{pan}, B_{pan} and C_{pan} are fitting parameters. Similarly, the core genome was modelled using

$$y = A_{core}e^{B_{core}x} + C_{core} \quad (2)$$

where A_{core}, B_{core} and C_{core} are fitting parameters and x = number of genome. The data are shown in Fig. 3a.

For the pan-genome, B_{pan} should be close to 1 if it is closed, which means that sequencing of additional genomes would not add any new gene to the gene repository. However, the B_{pan} value is 0.47 suggesting that the *Mma* pan-genome is open and evolving. Moreover, based on the 19 *Mma* genomes the pan-genome size is 8725, while the core-genome comprises 4300 genes. From Fig. 3b, we also estimated that approximately 100 "new" genes, not present in any of the current strains, would be identified upon sequencing an additional *Mma* genome.

Calculation of the pan- and core-genomes for cluster-I and cluster-II members revealed that cluster-I strains share 85% of their genes while any six cluster-II members share 89% (for a reliable comparison we analysed any six cluster-II strains since only six strains belong to cluster-I). Moreover, pan- and core-genome curves for cluster-I are slightly more separated than the corresponding curves for genomes belonging to cluster-II, suggesting modestly higher genomic variation among cluster-I strains compared to cluster-II members (Fig. 3c; see also above).

Variation of IS elements in *M. marinum*. Insertion (IS) elements are important factors responsible for genomic variations and dynamics[17]. The M strain carries multiple copies of eight different IS element types referred to as 'ISMyma1–7,11'[6], color-coded as shown in Fig. 4a. We first identified the IS elements present in the four complete genomes using ISsaga (IS-semi automated annotation[18]). Subsequently, we predicted the copy number and distribution of these IS elements in the draft genomes using raw reads (see Methods); the number of IS elements of each type is indicated by lengths and coloured patches (Fig. 4a). The MSS4 strain carried the highest number of predicted IS elements (n = 44) encompassing six "IS-types" (excluding ISMyma4 and ISMyma11). Of these, 11 copies were classified as ISMyma2 and 18 as ISMyma7. Moreover, the ISMyma4 type was only

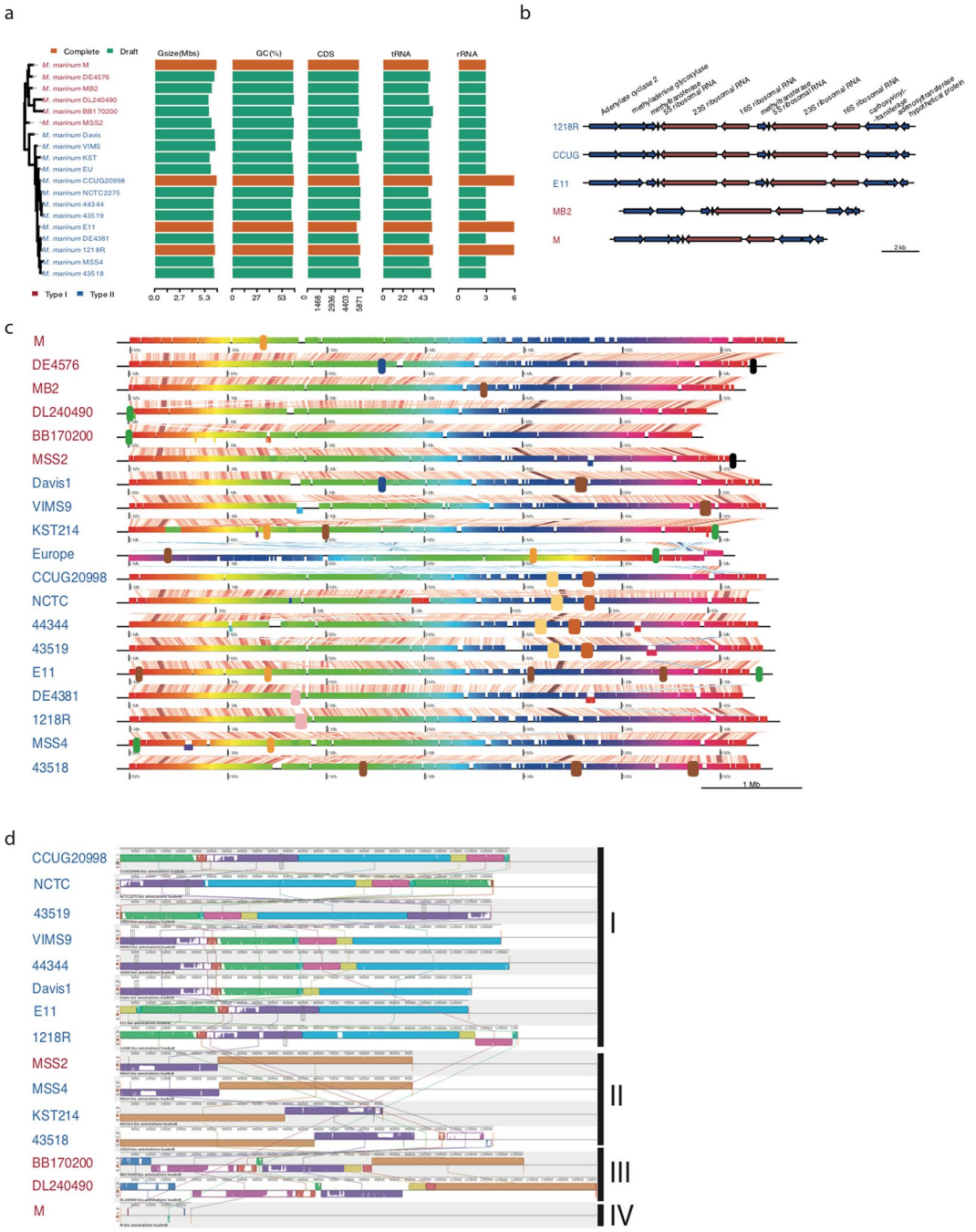

Figure 1. Overview of genomic features and genome alignments in *Mma* strains. (**a**) Bar plot showing genome size, GC-content (%), number of CDS, number of tRNA and number of rRNA in different strains along with the phylogenetic relationship. In the phylogenetic tree the strain names are in two colours representing the cluster-I (red) and cluster-II (blue) members. Complete and draft genomes are coloured by orange and green, respectively. (**b**) Synteny for the rRNA genes, *rrnA* and *rrnB*, present in cluster-I and cluster-II strains. Arrows represent genes and strand information. Right and left arrows indicate positive and negative strands. Red arrows refer to the rRNA operons and blue arrows mark flanking genes. (**c**) Whole-genome alignment of the 19 *Mma* strains where each of the coloured horizontal blocks represents one genome and the vertical bars represent homologous regions. Diagonal lines represent genomic rearrangements, whereas white gaps represent insertions/deletions. Presence of phage sequences are marked as large blocks in blue, green, yellow, and black. The same colour (except the black blocks) indicates that the phage fragments are the same while black blocks mark non-conserved phage sequences. (**d**) Alignment of the plasmid scaffolds in different *Mma* strain. Homologous regions in the plasmids are indicated by same coloured blocks connected with vertically lines. Partially filled regions and white regions in the blocks represent less similar sequence or unique regions respectively. All the plasmids are classified into four classes as indicated on the right side, see also the main text.

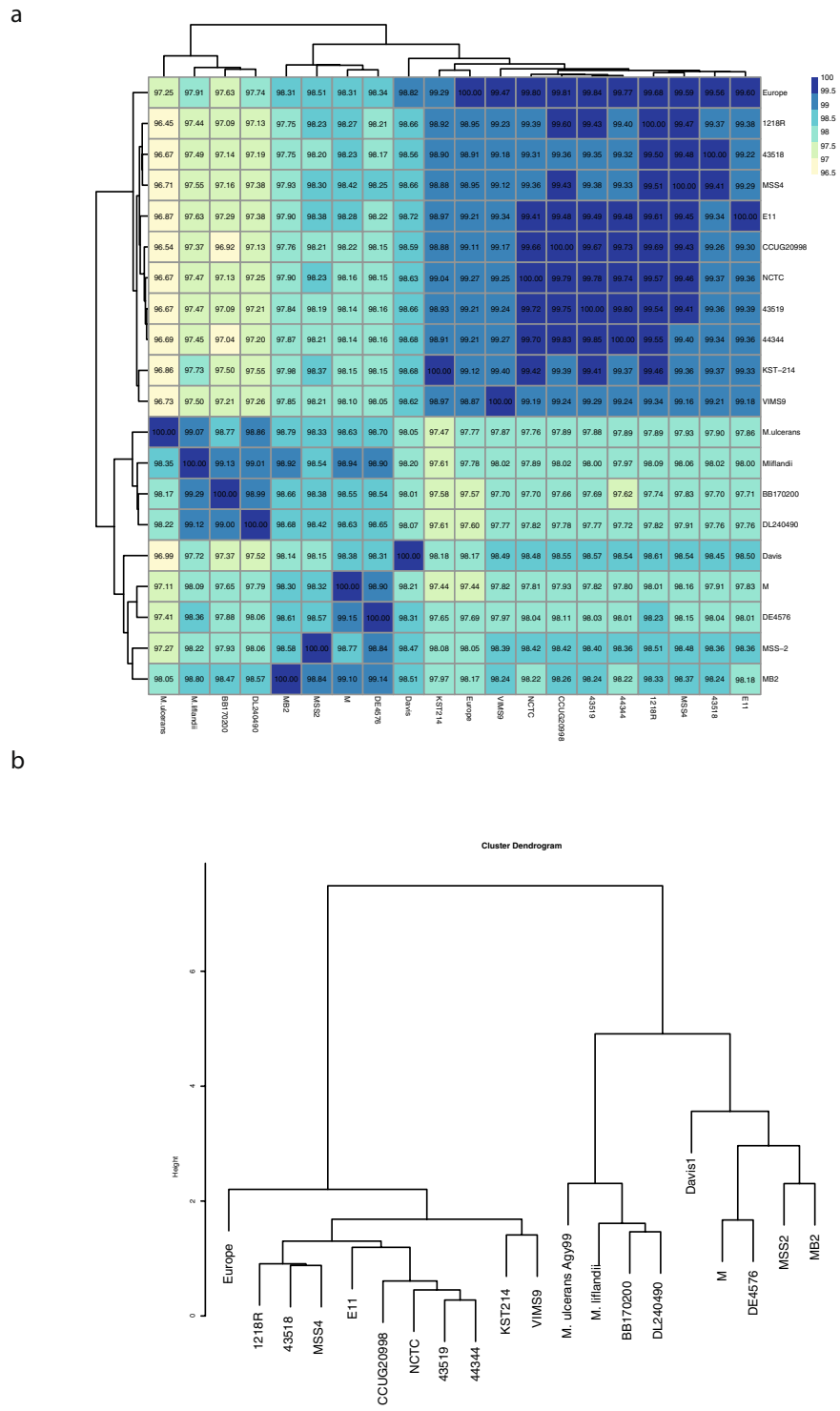

Figure 2. Clustering of *Mma* strains based on Average Nucleotide Identity (ANI). (**a**) Heat map showing ANI values for all versus all *Mma* strains including *M. ulcerans* and *M. liflandii*. (**b**) ANI values were clustered using unsupervised hierarchical clustering and plotted as dendogram.

detected in the M, Huestis and Davis1 strains while only two strains (EU and E11) carry ISMyma11. Our result also suggested that ISMyma6 and ISMyma7 are the dominant types in the cluster II strains. Remarkably, comparing 1218R with DE4381 (also designated 1218S, which was isolated as a smooth colony variant of 1218R; Small PLC, personal communication), revealed that DE4381 has a higher number of as well as different IS elements than 1218R. Interestingly, both 1218R and 1218S passage strains were likely derived from a common ancestor (TMC1218, Table 1), these results in turn, suggest that a number of IS elements from TMC1218 were lost during passage of 1218R when compared to 1218S.

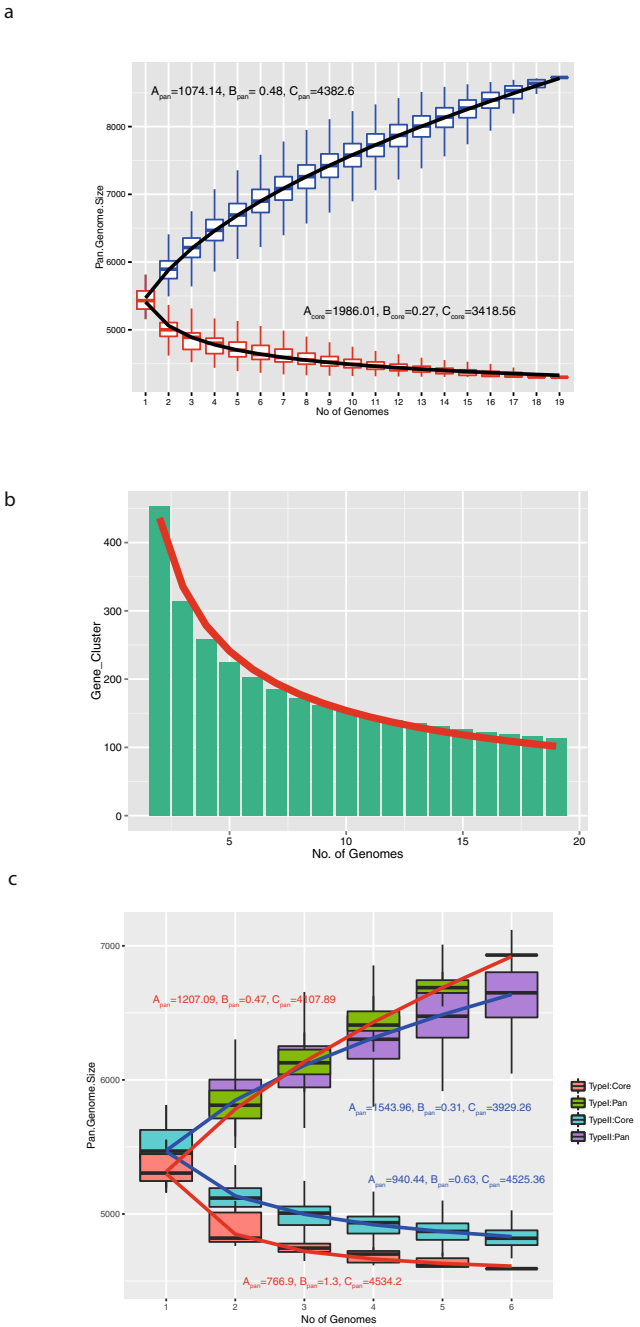

Figure 3. Pan-genome and core-genome of *Mma*. (**a**) Boxplots showing pan-genome (blue) and core-genome (red) for progressively increasing number of genomes. The black line is a fitted-line model by a regression formula (see text for details). (**b**) The number of new genes identified with increasing number of genomes. The red line is a fitted-line model generated by regression analysis (see text for details). (**c**) Similar plot as in (**a**) showing the results for cluster-I and cluster-II genomes separately.

We also compared the genome-wide distribution of the predicted IS elements in the complete M, CCUG and 1218R genomes. The mapping of predicted IS elements along with whole genome alignment of the three-complete genomes revealed that many divergence regions in the genomes are adjacent to the predicted IS elements (Fig. S2).

Comparative analysis of the gene content: core and auxiliary genes. Overall, the gene content across the different strains is highly conserved with respect to gene synteny and percentage identity. However, several gene clusters present in the M strain are absent in most of the other strains. The numbers of unique genes in the M strain is 277. Of these, 145 (55%) were annotated as hypothetical proteins. These genes are organized in clusters/regions encompassing 9 to 51 genes, and the majority of these are localized in the 3.7–4.9 Mb region (Fig. 1c and Table S2). Within these regions in the M strain genome, we identified 14 transposase and 19

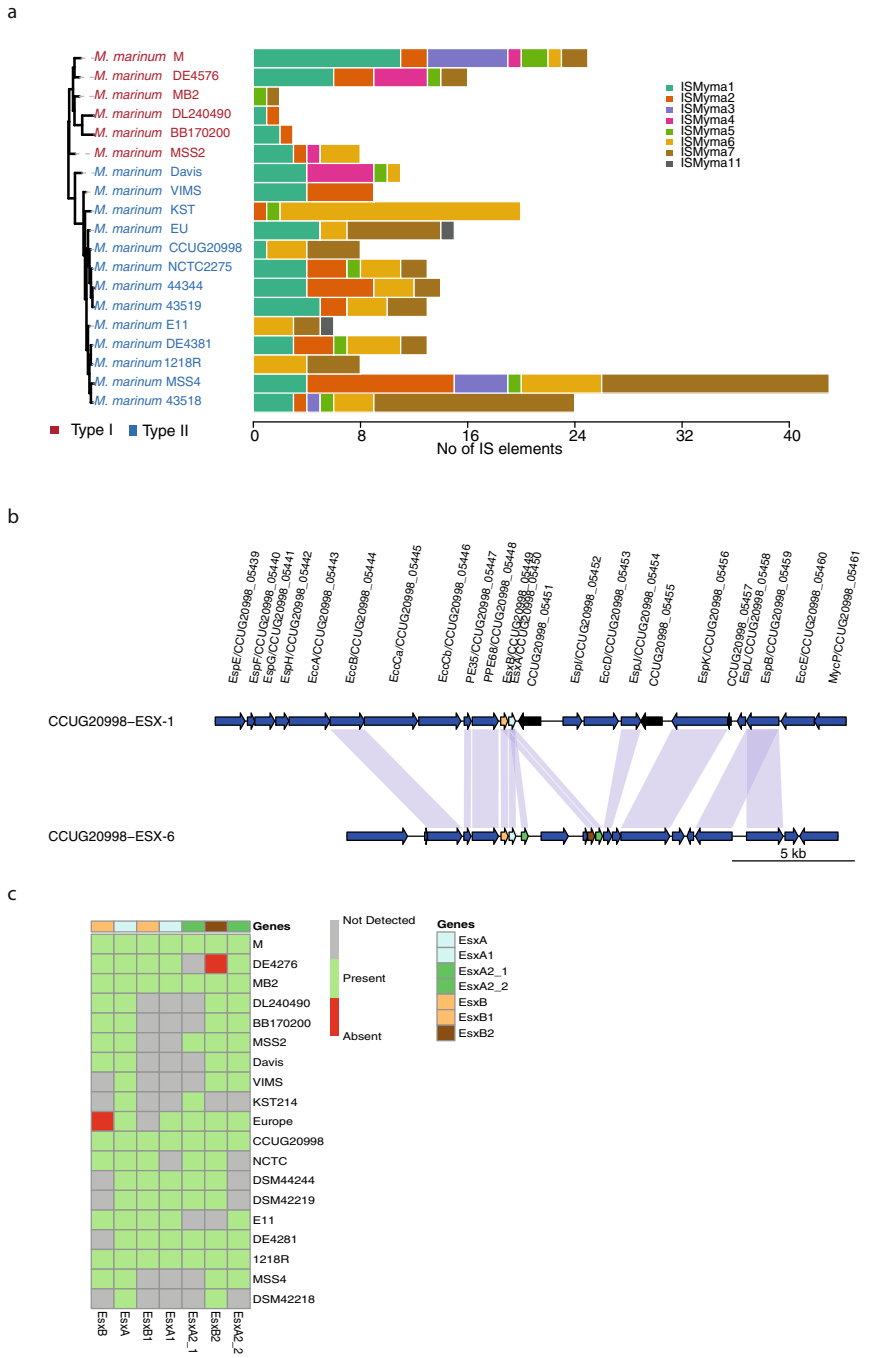

Figure 4. Genomic variations in *Mma* strains. (**a**) Distribution of IS elements in the *Mma* strains: Bar plot showing copy numbers and distribution of the eight types of IS elements in each of the strains along with their phylogenetic relationship. The phylogenetic tree is the same as shown in Fig. 1a. (**b**) Gene synteny for ESX-1 and the partially duplicated ESX-6 gene clusters. Arrows represent genes and direction of the arrow indicates strand information while vertical connections indicate orthologous genes. Genes are drawn to scale. Color code: blue (ESX-1 related genes) and black (hypothetical protein) arrow mean upstream/downstream of *esxB* and *esxA* gene loci. The regular *esxB* and *esxA* genes are colored as yellow and light green respectively and the connecting light blue shaded vertical line between the arrows indicates homologous genes. (**c**) Heat map showing presence and absence of *esxA* and *esxB* orthologous and paralogous genes in different *Mma* strains.

integrase genes, which raises the possibility that genes within these regions are "readily" mobilized. The unique regions in the M genome also overlap with different phage sequences that add to the diversification of these regions. Furthermore, the presence of prophage sequences raises the possibility that expression of genes within this region(s) is regulated in response to genome rearrangements of prophages, referred to as active lysogeny[19].

Within the unique region near 3.7 Mb, the σ^{2997} gene (MMAR_2997; σ^{2997}) was predicted to be present only in the M and NCTC2275 strains [Fig. 1 and S3; σ^{2997} is expressed and functional[20] (and unpublished)]. Since the

σ^{2997} gene is missing in the other *Mma* genomes, it is plausible that it was present in the ancestor but was lost in the majority of the *Mma* strains during evolution.

Many mycobacteria have two potassium (K^+)-uptake systems, the *trk*- and *kdp*-system[21] and all 19 *Mma* strains were predicted to have genes encoding the *trk*-system (Table S3). Within one of the unique gene clusters in the M genome we identified the *kdp*-system genes, *kdpABCDEF*. This gene cluster is absent in all the other *Mma* genomes as well as in *M. ulcerans*. This shows that the *kdp*-system is dispensible for growth *in vivo*. It has also been reported that inactivation of the *kdp*-system in *Mtb* increases its virulence[22]. Whether this also applies to *Mma* warrants further studies.

Core genes constitute the backbone of the genomes, while non-core genes have an impact on the phenotypic variation among different strains. The non-core genes were extracted and plotted in Fig. S3b. Of these, 142 genes were predicted to be present in all *Mma* genomes with the exception of BB170200 and DL240490. The predicted number of unique non-core genes varies from 26 (CCUG) to 345 (VIMS) (Fig. S3c and Supplementary Table S3). More than 50% of the unique genes in the different genomes were annotated as hypothetical proteins. For many genes, we detected variation in copy numbers comparing the different strains. For example, for genes encoding the dimodular nonribosomal peptide synthase, the tyrosine recombinase (XerD), a few ESX proteins, integrases and transposases. We cannot exclude the possibility that some of the genes identified as unique for the 14 draft genomes may be false positives due to the absence of reads in the corresponding loci.

Duplication of *esxA* and *esxB*. The type VII secretion system was discovered in mycobacteria and ESX-1 genes are major virulence factors for both *Mtb* and *Mma*[6,23–27]. As reported for the M strain, all *Mma* strains carry a partial duplication of ESX-1 (the homolog to the prototypical ESX-1 in *Mtb*) gene cluster, resulting in more than one copy of several genes including *esxA*/ESAT6 and *esxB*/CFP10 (Figs 4b,c and S4a–d). Interestingly, for the 1218R variant DE4381 (see above), genes positioned upstream of *esxB* in the ESX-1 region have been lost (Fig. S4a) but homologues for some of these genes are present in the duplicated ESX-1 region (referred to as ESX-6; Fig. S4b)[6]. It is noteworthy that the ESX-1 *esxB* is missing in the Europe and DSM43518 strains, while truncated *esxB* variants are present in MSS2, Davis and Huestis, resulting in shorter protein sequences (Fig. S5a). The *esxB* homolog, *esxB*1, in the ESX-6 region in the Europe strain is also missing, while it is present in Huestis but not in MSS2 and Davis1, which might be due to that they are draft genomes. Hence, for Huestis the loss of *esxB* could be functionally complemented by *esxB*1 since *esxB* and *esxB*1 are sequentially identical (Fig. S5a,b; see also below), which was discussed in a recent report[28]. Moreover, *esxA* was predicted to be present in all the *Mma* strains (Figs 4b and S4a). For cluster-II members the *esxA* sequence is highly conserved with no sequence variation, while for strains belonging to cluster-I it varies at several positions (Fig. S5a). It therefore appears that while the ESX-1 *esxB* gene is dispensable this is not the case for *esxA*.

Comparing the different *Mma* strains the ESX-6 region appears to be more variable than ESX-1 (Figs 4b and S4a,b). In addition to the predicted homologs of *esxB* and *esxA*, *esxB*1 and *esxA*1, we identified the presence of one *esxB* paralog, *esxB*2, and two *esxA* paralogs, *esxA*2 and *esxA*3. The *esxB*2, *esxA*2 and *esxA*3 were predicted to be present in all *Mma* strains with few exceptions, Huestis, KST214 and E11 in the case of *esxB*2 (Fig. 4c). EsxB2 is highly conserved and show 54% sequence identity compared to EsxB and EsxB1 (Fig. S5). Comparing EsxA and EsxA1 revealed roughly 90% sequence identity, and interestingly, E11 EsxA1 is identical to EsxA1 present in the M strain (Fig. S5b). This is in contrast to EsxA (see above) and might possibly be due to gene transfer and homologous recombination. On the other hand, the sequences of the EsxA paralogs, EsxA2 and EsxA3, are almost identical across the different strains. However, variation for NCTC and MSS4, where only one paralog was predicted, might be due to the genomes being draft genomes. As in the case of EsxB2, comparing EsxA2 and EsxA3 with EsxA and EsxA1 revealed lower sequence identities (40–45%) than EsxA and EsxA1 (≈90% sequence identity; Fig. S5). Notably, genes within the *Mycobacterium smegmatis* ESX-1 region that influence mating identity have been implicated as having a role for mycobacterial conjugation[29]. However, these genes are not present in the *Mma* or *Mtb* H37Rv ESX-1 regions (Figs 4b and S4a).

Together, these analyses suggest that the duplication of ESX-1 is present in all *Mma* strains and was probably present in the *Mma* ancestor. Furthermore, it appears that the *esxB* and *esxA* genes of the ESX-1 region underwent a second duplication event during evolution to yield an additional ESX region ESX-6 (Figs 4b and S4a; see discussion)[6,30].

Identification of SNVs and mutational hotspots in *M. marinum* strains. Single nucleotide variations (SNVs) were predicted for all the genomes with the M strain as reference using the program MUMmer[31]. For cluster-I members, the number of SNVs ranged between 45000 and 56000, while for cluster-II members it is significantly higher, between 70000 and 89000 (Fig. 5a). This is consistent with the proposal that the *Mma* strains can be divided into two clusters (see above).

Next, we identified the mutational hotspots in the *Mma* genomes Das *et al.*[32]. Mutational hotspots are genomic regions where the SNV frequencies are much higher relative to the background. One hundred seventy-six mutational hotspots were identified in the *Mma* genomes, which corresponds to a frequency of 26.5/Mb (Fig. 5b,c). A similar analysis of 20 *Mtb* isolates suggested only 45 mutational hotspots corresponding to 10/Mb (Fig. 5c)[32]. We therefore determined the hotspot frequencies for three other mycobacteria, *M. avium* subsp. *paratuberculosis* (*MAP*), *M. bovis* (*Mbo*) and *M. phlei* (*Mph*), for which genomic data for several strains are available (see Methods). This analysis revealed that *Mma* carries a higher number of hotspots also compared to these mycobacteria (Figs 5c and S6a–c). Moreover, analysing the two *Mma* clusters separately indicated that the number of mutational hotspots in cluster-I strains is 180, while cluster-II strains have 253 hotspots. However, the average number of SNVs per genomic regions is higher in cluster-I (≈18) than in cluster-II (≈8) consistent with higher divergence among cluster-I members compared to cluster-II members (Fig. S7a,b).

Considering all 19 *Mma* strains, 621 genes map in the hotspot regions. Of these, 300 were annotated as hypothetical genes. The remaining 321 genes were classified into different subsystem categories (Fig. 5d), and >20% were predicted to belong to the category "Fatty Acids, Lipids and Isoprenoids". Since *Mma* strains occupy widely different ecological niches this would be consistent with an evolutionary pressure on genes involved in building the outer boundaries.

Phylogenetic analysis. To understand the phylogenetic relationship between the *Mma* strains, we generated phylogenetic trees based on the 16S rRNA genes using the neighbour-join method with 1000 cycles of bootstrapping. This tree displays three main branches. However, it could not discriminate between closely related strains (Fig. 6a). In addition, the 16S rDNA tree is not very robust since many branches have low or zero bootstrap values.

We previously reported the use of core genes to generate robust phylogenetic trees for other mycobacteria[33,34]. Hence, we used the 4300 core genes (see above) present in all 19 *Mma* strains and generated the tree shown in Fig. 6b. This tree is supported by high bootstrap values and separates the different strains into two branches/clusters, which is similar to the clustering obtained based on ANI values (cluster-I and -II; cf. Figs 2b and 6b).

M. ulcerans and *M. liflandii* are *Mma*'s closest neighbours. Therefore, we were interested in understanding their positions relative to the *Mma* strains. Considering a tree based on 2297 core genes present in all 19 *Mma* strains, *M. ulcerans* and *M. liflandii* cluster together with cluster-I members with MB2, DL240490 and BB170200 as their closest neighbours (Fig. 6c). This grouping is in accordance with previous studies that position *M. ulcerans* and *M. liflandii* close to the M strain[4,8,35].

Finally, we generated a tree based on the SNVs identified in the *Mma* strains, *M. ulcerans* and *M. liflandii* (see above and Methods). The resulting tree corroborated the core gene-based phylogeny and clustering of ANI values with few exceptions at the leaf levels (Fig. 6d). In agreement with the 2297 core gene tree (Fig. 6c) the SNV-based tree suggests that strains belonging to cluster-I are more closely related to *M. ulcerans* and *M. liflandii* than cluster-II members (Fig. 6d).

In conclusion, the "M strain lineage" (or M-lineage; cluster-I) members cluster together with *M. ulcerans* and *M. liflandii* and separated from cluster-II (referred to as the "Aronson-lineage") members. This suggests that members of these two lineages should be classified as two separate *M. marinum* subspecies. Notably, strains of the "Aronson-lineage" (Table 1, and marked with an * in the figures) that originate from the originally isolated *Mma* strain (Aronson)[1] have diverged, presumably as a result of handling in different laboratories (see discussion).

Discussion

To trace the evolutionary history and relationships among organisms, the 16S rRNA gene has served as an important biomarker. Specifically, it has been useful in providing a reliable phylogenetic tree for bacteria. However, now we have access to large number of bacterial genomes and whole genome comparison can be used to reveal more detailed and better-resolved evolutionary relationships among bacterial species and strains considered phylogenetically unique on the basis of 16S rRNA gene comparison. Consequently, this has expanded the repertoire of genes that can be used to study the evolutionary relationships among bacteria and that have had an impact on their evolution, diversity and use as biotechnological vehicles and documenting the microbial biosphere. Phylogeny based on multiple genes and genomic information have generated more robust trees and in many cases also provided evidence that known species should be considered as separate species or subspecies of known bacteria[33,36,37].

We present the genome sequences for 15 *Mma* isolates including the complete genomes of two type strains CCUG20998 and 1218R, both derivatives of the original *Mma* strain isolated by Aronson[1]. Our comparative genomic studies, ANI analysis and phylogenetic trees based on core genes and SNVs, covering 19 *Mma* genomes suggested that the *Mma* strains cluster in two distinct branches, cluster-I and -II. Cluster-I encompasses six strains including the M strain, while the remaining 13 strains constitute cluster-II. In cluster-II we find derivatives of the originally isolated *Mma* strain, *e.g.*, the CCUG20998 and 1218R strains. Including *M. ulcerans* and *M. liflandii* revealed that the "M-lineage" (cluster-I) members are their closest neighbours while "Aronson-lineage" (cluster-II) strains are more distantly related. On the basis of these findings, we propose that these two branches should be considered as two separate *Mma* subspecies. We suggest that the "Aronson-lineage" should be named *M. marinum* subsp. *marinum* and the "M-lineage" *M. marinum* subsp. *moffett* (*moffett* since it was first isolated at the Moffett Hospital, Universtity of California, San Francisco)[6]. To distinguish the current "Aronson-" and "M-" strains and for strain identification we further suggest including the name of the strain, *e.g.*, *M. marinum* subsp. *marinum* strain 1218R. Moreover, since available data indicate that *M. ulcerans* evolved from *Mma*[7,8], our findings indicate that its nearest ancestor belonged to the "M-lineage". In this context, we note that the plasmid fragments present in BB170200 and DL240490, referred to as pMUM003[38], are highly similar compared to the pMUM001 plasmid present in *M. ulcerans* Agy99[7], while plasmids (or plasmid fragments) detected in cluster-II members are different. In addition, plasmid type (I) is only present in strains belonging to the "Aronson-lineage". Given that *Mma* subsp. *moffett* strains are closely related to *M. ulcerans* and *M. liflandii* with, *e.g.*, ANI values 98.63 and 98.94 compared to the *Mma* "M-strain", raises the question whether *M. ulcerans* and *M. liflandii* should be considered as subspecies of *M. marinum*.

Interestingly, MB2, Europe and Huestis, which all lack plasmids, were isolated as wild outbreak strains in fish; in particular, Huestis has been shown to be a highly virulent outbreak strain in medaka and zebrafish models (Ennis and Shirreff unpublished), which might suggest that plasmids did not play a critical role for virulence for these strains. Moreover, the two *Mma* strains BB170200 and DL240490 were reported to be hyper-virulent due to the production of the mycolactone F toxins *i.e.*, presence of the pMUM003 plasmids[6]. Both these mycolactone F producing strains conferred only moderate virulence when compared to the 1218R strain in a controlled infection medaka model[6]. In summary, it therefore appears that there is no clear correlation between the presence of plasmid carried in the *Mma* strains studied in this report and virulence in animals; however, 1218R carries a Type I

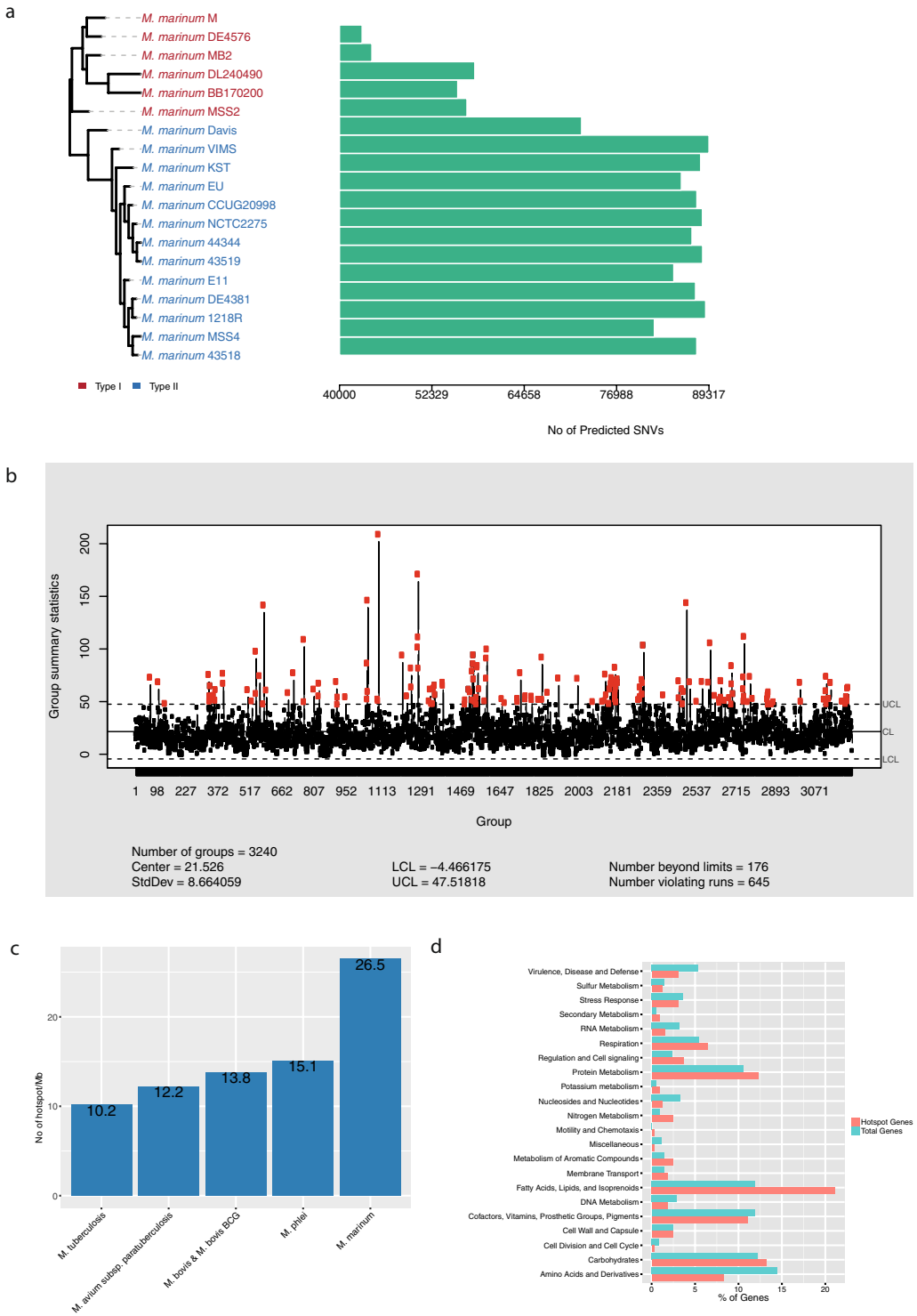

Figure 5. Analysis of mutational hotspots in *Mma*. (**a**) Bar plot showing the predicted number of SNVs in the different strains compared to the M strain along with their phylogenetic relationship. The phylogenetic tree is the same as in Fig. 1a. (**b**) Shewhart control chart showing the average SNVs frequencies in all the strains. Red and black dots indicate out of control (hotspots) and in-control SNV frequencies, respectively. (**c**) SNV frequencies per one Mb in different mycobacteria as indicated. (**d**) Functional classification of genes located in the predicted hotspot regions.

plasmid, while strain DE4381 (a related "smooth" passage variant also called 1218S; see below) has apparently lost this plasmid and may be a product of plasmid segregation. Hence, more detailed genetic and molecular analyses would be required to better document the role that specific plasmids may play in virulence.

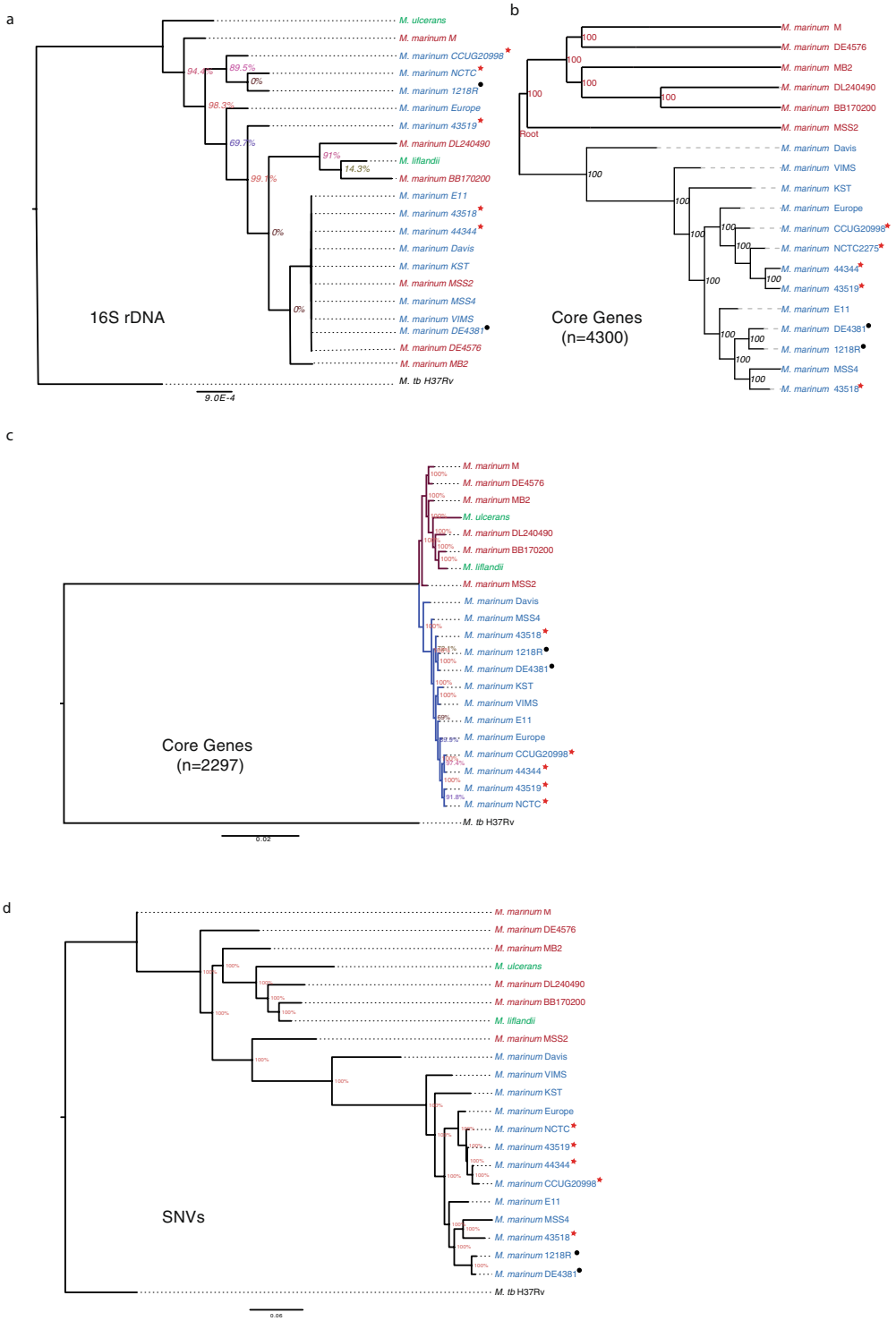

Figure 6. Phylogenetic trees for *Mma*. Phylogenetic trees were based on: (**a**) 16S rDNA, (**b**) core genes (n = 4300) present in all *Mma* strains, (**c**) core genes (n = 2297) in all *Mma* strains, *M. ulcerans* and *M. liflandii* and (**d**) predicted SNVs compared to the M strain. The percentage values in the nodes represent bootstrap values generated by 1000 cycles. Lab and passage variants derived from the TMC1218 strain (Table 1) are marked with red stars and black circles as indicated.

The average number of SNVs per region was found to be higher (18.4 vs 7.9; Fig. 5c) in members of the "M-lineage" compared to "Aronson-lineage" members. Our data also showed that the "Aronson-lineage" strains CCUG and 1218R (both complete genomes) have two rRNA operons, while the M strain has one (see below). Based on that, we predict the presence of two 5S rRNA genes in all the cluster-II draft genomes (one for cluster-I draft genomes) and we assume that all strains in the "Aronson-lineage" carry two rRNA operons.

As for other bacteria, the *Mma* pan-genome is open ($B_{pan} = 0.47$; Fig. 3a; see also *e.g.*, refs[39–42]). Its size amounts to 8725 genes. Of these, roughly 50% constitute the core genome. However, comparing the pan- and core-genomes of the "M-" and "Aronson-" lineages revealed that the pan-genome for the "M-lineage" is larger, while its core-genome is smaller (Fig. 3c). It should be noted that the pan- and core-genome sizes vary for individual species within the *Streptococcus* genus[43]. Taken together, these findings suggest that the "M-lineage" members show higher diversity compared to members belonging to the "Aronson-lineage" and thus, support the notion that the two lineages should be considered as separate subspecies.

It has previously been reported that mutational hotspots in *Mtb*, expressed as SNVs per genome size (Mb), cluster in certain genomic regions[32]. We provide data that the number of mutational hotspots is significantly higher for the 19 *Mma* strains compared to other mycobacteria (*Mtb*, *MAP*, *Mbo* and *Mph*; Fig. 4c). A major fraction of these hotspots were mapped to genes involved in fatty acid, lipid and isoprenoid metabolism. Many of these genes play important roles in building and altering the outer boundaries in response to environmental changes, consistent with that *Mma* inhabits different ecological niches. In this context, we note that of the two strains MSS2 and MSS4, belonging to different lineages, MMS2 was isolated from an infected fish cultivated from a striped bass aquaculture facility and MMS4 from a patient working at the same fish farm[44]. This suggests that the two strains occupy the same ecological niche. One expectation would be that the same strain causing disease in the fish would also be the one infecting the human. However, our genomic analysis showed that this was not the case. Rather, this might be related to different strains having a selective advantage for infecting different hosts. Alternatively, this finding could be random and insignificant.

The sole rRNA operon (*rrnA*) in the M strain is located downstream of *murA* and upstream of the *apt* gene (coding for O_6-alkylguanine DNA alkyltransferase). In general, rapidly growing mycobacteria harbour two rRNA operons, *rrnA* and *rrnB*, and these are located downstream of the *murA* and *tyrS* genes, respectively[45]. However, *rrnA* and *rrnB* in the cluster-II members are located next to each other separated by a duplicated copy of the *apt* gene (Fig. 1b). This suggests that the presence of two *rrn* genes in cluster-II members likely is the result of a duplication event. Since the closest neighbours of *Mma*, *i.e. Mtb*, *Mycobacterium kansasii* and *Mycobacterium gastri*, are equipped with one rRNA operon it is likely that the duplication occurred after the "M-" and "Aronson-" lineages diverged. However, we cannot exclude that their ancestor had two rRNA operons and that one was lost after the two lineages diverged. In this context, we note that the M and CCUG strains grow with similar rates in 7H9 media at 30 °C (generation times in hours 8.1 ± 0.8 and 8.9 ± 0.1, respectively) consistent with the number of rRNA operons not affecting the growth rate (see *e.g.*, refs[46,47]).

IS elements have a key role in generating diversity among bacteria. As such, IS elements can be used as biomarkers for strain identification, epidemiological tracking and predicting spread of antibiotic resistance[48–50]. We identified the presence of eight known IS elements (ISMyma1–7, 11) in all 19 *Mma* strains. Their distribution and copy number varied among the different strains with MSS4 having the highest number, 44 IS elements (Fig. 4). Hence, these data open up the possibility for the development of strain specific *Mma* probes for use in clinical settings that relate to, *e.g.*, fisheries and aquariums. For example, using probes to determine whether an infection is caused by MSS2 or MSS4 (see above). In this context, we note that more than four ISMyma3 copies are present in the M and MSS4 strains, which were both originally isolated from human patients.

Of specific interest is the ESX-1 region and in this context the variants, 1218S and 1218R, which are passage variants derived from the TMC1218 lineage (Table 1). We have studied the properties of the DE4381 ("1218S") smooth colony variant, and our data showed that it only conferred a modest (approx. four-fold) reduction of virulence when compared to the rough colony variant 1218R using the established Japanese medaka infection model system (Fig. S8)[51]. The DE4381 strain carries a deletion encompassing ten genes within the ESX-1 interval, including the loss of *esxB* and nine other well-defined ESX-1 genes (Fig. S4a). The loss of one of these genes, *eccA*1 (which corresponds to Rv3868 in *Mtb*H37Rv), has been reported to confer large pleiotropic effects on virulence[23]. In keeping with this, infection of zebrafish with a *Mma* Δ*eccA*1 mutant strain influenced virulence[23,25] and it displays a substantial reduction (>1000-fold) in spread and colonization upon infection of Japanese medaka (Mallick and Ennis, unpublished). We are therefore endeavouring to better characterize these unexpected subtle effects on virulence that was conferred by this ten-gene deletion and this may suggest that the DE4381 (1218S) strain carries suppressor mutations for virulence. We expect that by comparing other smaller rearrangements and mutational changes between the 1218R and DE4381 (1218S) strains, and differences in transcriptional profiles, will gain insights into the compensational mutations that presumably result in this partial suppression of virulence associated with the DE4381 strain. In this context we also note that absence of ESX-1 genes have been discussed to be associated with changes in colony morphologies such as exhibition of smooth or rough colony morphology[52].

Our study provides insight into the diversity of *Mma* strains in terms of genomic variations. Several factors contribute to this diversity, such as the presence of phage and IS elements as well as plasmids. The diversity is also expressed in terms of higher numbers of mutational hot spots compared to other mycobacteria. Together this emphasizes that the *M. ulcerans-M. marinum* complex, MuMC, constitute a group of bacteria useful to identify factors and study their importance for bacterial evolution[9].

Methods

Strain information and cultivation. Information about the *Mma* strains is compiled in Table 1. DSM44344, DSM43518 and DSM43519 were purchased from DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH). MSS2 was isolated from Hybrid striped bass during an outbreak in Mississippi, USA in an aquaculture facility, and MSS4 from a lesion of a human patient who worked in this facility. VIMS9 was isolated from wild striped bass in Virginia, USA and Davis from farmed striped bass in Davis, California, USA. The different strains were cultivated as described elsewhere[33,34].

DNA sequencing and assembly. Complete genomes of the *Mma* type strains, CCUG20998 and 1218R were sequenced using PacBio technology. The remaining 13 strains were sequenced on HiSeq. 2000 (Illumina platforms) at the SNP@SEQ Technology Platform, Uppsala University. Genomic DNA was isolated and prepared for sequencing as described elsewhere[33,34].

PacBio sequencing reads with an average length more than 10,747 bp and read depth of around 100x were assembled using the SMRT-analysis HGAP3 assembly pipeline[53], polished using Quiver (Pacific Biosciences, Menlo Park, CA, USA) and generated single scaffolds for CCUG20998 and 1218R.

For the Illumina sequencing reads, a total of 12 million short reads was generated for each strain with an average read length of 100 nucleotides (Table 1). Filtering of the short reads was done to remove low quality reads and ambiguous bases. *De novo* assembly of the short reads was done using the A5 assembly pipeline (version 1.05)[54]. Final genomes consist of contigs of more than 200 bases. Scaffolds were re-ordered using MAUVE with the CCUG20998 genome as reference.

Genome Annotations. All the genomes, including the available M, Europe, E11 and MB2 genomes, were annotated using the Prokka pipeline[55], which uses Prodigal[56] to predict coding sequences (CDS). tRNA and rRNA genes were predicted by Aragorn[57] and RNAmmer[58]. Annotated genes were functionally classified using the RAST Subsystem[59].

Identification of plasmid fragments and foreign DNA. Plasmid fragments were identified by pairwise alignments of the scaffolds with sequences from the NCBI plasmid database (ftp://ftp.ncbi.nlm.nih.gov/genomes/Plasmids/).

Prophage sequences were identified using concatenated ordered scaffolds for all the genomes in the PHAST server[60].

Prediction of IS elements. IS elements were predicted for the complete genomes using the ISsaga server[18]. All the predicted IS elements were used as reference to identify IS elements in the draft genomes using ISmapper[61]. ISmapper uses raw reads and reference IS elements and predicts possible positions of the reference IS elements in the genome enquired.

Identification of orthologous genes. Homologous coding sequences were identified using an all-versus-all BLAST search of the protein sequences from all 19 strains. Orthologous genes were predicted using PanOCT (v3.23)[62] from the BLAST output. PanOCT follows two criteria to consider a gene as orthologous, sequence homology and gene synteny. All necessary PanOCT input files were generated using in-house shell scripts, and PanOCT was executed to detect the genomic differences based on coding regions.

Identification of SNVs and mutational hotspots. Whole genome alignments were performed in a pairwise manner using MUMmer[31]. SNVs were identified using the "show-snps" program of the MUMmer package. Single nucleotide insertions/deletions were filtered out, and only SNVs were used for further analysis. Mutational hotspots were identified using Shewhart Control Chart, as described by Das *et al.*[32]. Briefly, the genome of the M strain was divided into non-overlapping windows of 2000 bases and the average number of SNVs in each of the windows was determined. The average SNV values were subsequently used in Shewhart Control Chart for the prediction of hotspots. Mutational hotspots were identified for *Mma* (n = 19), *Mbo* (n = 28), *Mph* (n = 5), and *MAP* (n = 23). Mutational hotspots for cluster-I and cluster-II were identified separately using the genome sequences of the M and CCUG strains as references for cluster-I and cluster-II, respectively, and follow the procedure as described above.

Ethics Statement. All methods were carried out in accordance with relevant guidelines and regulations.

All animal experimental protocols were approved by both the Institutional Biosafety Committee and the Institutional Animal Care and Use Committee at the University of Louisiana (Animal Assurance Identification number A3029-01 and IACUC approval No. 2016 8717-029; see also figure legend Fig. S8).

Data deposition. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the project id PRJNA414948 and PRJ414525.

References

1. Aronson, J. D. Spontaneous Tuberculosis in Salt Water Fish. *J Infect Dis* **39**, 315–320 (1926).
2. Sette, C. S. *et al*. Mycobacterium marinum infection: a case report. *J Venom Anim Toxins Incl Trop Dis* **21**, 1 (2015).
3. Stinear, T. P. *et al*. Comparative genetic analysis of Mycobacterium ulcerans and Mycobacterium marinum reveals evidence of recent divergence. *J Bacteriol* **182**, 6322–6330 (2000).
4. Qi, W., Käser, M., Röltgen, K., Yeboah-Manu, D. & Pluschke, G. Genomic Diversity and Evolution of Mycobacterium ulcerans Revealed by Next-Generation Sequencing. *Plos Pathog* **5**, e1000580 (2009).
5. Pidot, S. J., Asiedu, K., Käser, M., Fyfe, J. A. M. & Stinear, T. P. Mycobacterium ulcerans and Other Mycolactone-Producing Mycobacteria Should Be Considered a Single Species. *Plos Negl Trop Dis* **4**, e663 (2010).
6. Stinear, T. P. *et al*. Insights from the complete genome sequence of Mycobacterium marinum on the evolution of Mycobacterium tuberculosis. *Genome Res* **18**, 729–741 (2008).
7. Käser, M. *et al*. Evolution of two distinct phylogenetic lineages of the emerging human pathogen Mycobacterium ulcerans. *BMC Evol Biol* **7**, 1 (2007).
8. Doig, K. D. *et al*. On the origin of Mycobacterium ulcerans, the causative agent of Buruli ulcer. *BMC Genomics* **13**, 1–1 (2012).
9. Röltgen, K., Stinear, T. P. & Pluschke, G. The genome, evolution and diversity of Mycobacterium ulcerans. *Infect Genet Evol* **12**, 522–529 (2012).
10. Ummels, R. *et al*. Identification of a novel conjugative plasmid in mycobacteria that requires both type IV and type VII secretion. *mBio* **5**, e01744–14 (2014).

11. Gröschel, M. I., Sayes, F., Simeone, R., Majlessi, L. & Brosch, R. ESX secretion systems: mycobacterial evolution to counter host immunity. *Nat Rev Micro* **14**, 677–691 (2016).
12. Abdallah, A. M. *et al*. The ESX-5 secretion system of Mycobacterium marinum modulates the macrophage response. *J Immunol* **181**, 7166–7175 (2008).
13. Abdallah, A. M. *et al*. Mycobacterial secretion systems ESX-1 and ESX-5 play distinct roles in host cell death and inflammasome activation. *J Immunol* **187**, 4744–4753 (2011).
14. Weerdenburg, E. M. *et al*. ESX-5-deficient Mycobacterium marinum is hypervirulent in adult zebrafish. *Cellular Microbiology* **14**, 728–739 (2012).
15. Richter, M. & Rosselló-Móra, R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci USA* **106**, 19126–19131 (2009).
16. Lapierre, P. & Gogarten, J. P. Estimating the size of the bacterial pan-genome. *Trends Genet.* **25**, 107–110 (2009).
17. Siguier, P., Gourbeyre, E. & Chandler, M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol Rev* **38**, 865–891 (2014).
18. Varani, A. M., Siguier, P., Gourbeyre, E., Charneau, V. & Chandler, M. ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biol* **12**, R30 (2011).
19. Feiner, R. *et al*. A new perspective on lysogeny: prophages as active regulatory switches of bacteria. *Nat Rev Micro* **13**, 641–650 (2015).
20. Pettersson, B. M. F. *et al*. Comparative Sigma Factor-mRNA Levels in Mycobacterium marinum under Stress Conditions and during Host Infection. *Plos One* **10**, e0139823 (2015).
21. Cholo, M. C., van Rensburg, E. J. & Anderson, R. Potassium uptake systems of Mycobacterium tuberculosis: genomic and protein organisation and potential roles in microbial pathogenesis and chemotherapy. *South Afr J Epidemiol Infect* **23**, 13–16 (2008).
22. Parish, T. *et al*. Deletion of two-component regulatory systems increases the virulence of Mycobacterium tuberculosis. *Infect Immun* **71**, 1134–1140 (2003).
23. Gao, L.-Y. *et al*. A mycobacterial virulence gene cluster extending RD1 is required for cytolysis, bacterial spreading and ESAT-6 secretion. *Mol Microbiol* **53**, 1677–1693 (2004).
24. Guinn, K. M. *et al*. Individual RD1-region genes are required for export of ESAT-6/CFP-10 and for virulence of Mycobacterium tuberculosis. *Mol Microbiol* **51**, 359–370 (2004).
25. Joshi, S. A. *et al*. EccA1, a component of the Mycobacterium marinum ESX-1 protein virulence factor secretion pathway, regulates mycolic acid lipid synthesis. *Chem Biol* **19**, 372–380 (2012).
26. Houben, E. N. G., Korotkov, K. V. & Bitter, W. Take five — Type VII secretion systems of Mycobacteria. *Biochim Biophys Acta* **1843**, 1707–1716 (2014).
27. Unnikrishnan, M., Constantinidou, C., Palmer, T. & Pallen, M. J. The Enigmatic Esx Proteins: Looking Beyond Mycobacteria. *Trends Microbiol* **25**, 192–204 (2017).
28. Bosserman, R. E., Thompson, C. R., Nicholson, K. R. & Champion, P. A. Esx paralogs are functionally equivalent to ESX-1 proteins but are dispensable for virulence in M. marinum. *J Bacteriol* **200**, e00726–17 (2018).
29. Derbyshire, K. M. & Gray, T. A. Distributive Conjugal Transfer: New Insights into Horizontal Gene Transfer and Genetic Exchange in Mycobacteria. *Microbiol Spectr* **2**, 1–19 (2014).
30. Tobias, N. J. *et al*. Complete genome sequence of the frog pathogen Mycobacterium ulcerans ecovar Liflandii. *J Bacteriol* **195**, 556–564 (2013).
31. Delcher, A. L. *et al*. Alignment of whole genomes. *Nucl Acids Res* **27**, 2369–2376 (1999).
32. Das, S. *et al*. Identification of hot and cold spots in genome of Mycobacterium tuberculosis using Shewhart Control Charts. *Sci Rep* **2**, 297 (2012).
33. Das, S. *et al*. Characterization of Three Mycobacterium spp. with Potential Use in Bioremediation by Genome Sequencing and Comparative Genomics. *Genome Biol Evol* **7**, 1871–1886 (2015).
34. Das, S. *et al*. The Mycobacterium phlei Genome: Expectations and Surprises. *Genome Biol Evol* **8**, 975–985 (2016).
35. Wang, J. *et al*. Insights on the emergence of Mycobacterium tuberculosis from the analysis of Mycobacterium kansasii. *Genome Biol Evol* **7**, 856–870 (2015).
36. Devulder, G., Pérouse de Montclos, M. & Flandrois, J. P. A multigene approach to phylogenetic analysis using the genus Mycobacterium as a model. *Int J Syst Evol Microbiol* **55**, 293–302 (2005).
37. Loman, N. J. & Pallen, M. J. Twenty years of bacterial genome sequencing. *Nat Rev Micro* **13**, 1–9 (2015).
38. Pidot, S. J. *et al*. Deciphering the genetic basis for polyketide variation among mycobacteria producing mycolactones. *BMC Genomics* **9**, 462 (2008).
39. Lukjancenko, O., Wassenaar, T. M. & Ussery, D. W. Comparison of 61 Sequenced *Escherichia coli* Genomes. *Microb Ecol* **60**, 708–720 (2010).
40. Jacobsen, A., Hendriksen, R. S., Aaresturp, F. M., Ussery, D. W. & Friis, C. The Salmonella enterica Pan-genome. *Microb Ecol* **62**, 487–504 (2011).
41. Gordienko, E. N., Kazanov, M. D. & Gelfand, M. S. Evolution of Pan-Genomes of Escherichia coli, Shigella spp., and Salmonella enterica. *J Bacteriol* **195**, 2786–2792 (2013).
42. Vernikos, G., Medini, D., Riley, D. R. & Tettelin, H. Ten years of pan-genome analyses. *Curr Opin Microbiol* **23**, 148–154 (2015).
43. Lefébure, T. & Stanhope, M. J. Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. *Genome Biol* **8**, R71 (2007).
44. Ostland, V. E. *et al*. Biochemical, molecular, and virulence characteristics of select Mycobacterium marinum isolates in hybrid striped bass Morone chrysops x M. saxatilis and zebrafish Danio rerio. *Dis Aquat Org* **79**, 107–118 (2008).
45. Menendez, M. C. *et al*. Characterization of an rRNA operon (rrnB) of Mycobacterium fortuitum and other mycobacterial species: implications for the classification of mycobacteria. *J Bacteriol* **184**, 1078–1088 (2002).
46. Gonzalez-y-Merchand, J. A., Colston, M. J. & Cox, R. A. Roles of Multiple Promoters in Transcription of Ribosomal DNA: Effects of Growth Conditions on Precursor rRNA Synthesis in Mycobacteria. *J Bacteriol* **180**, 5756–5761 (1998).
47. Gonzalez-y-Merchand, J. A., Colston, M. J. & Cox, R. A. Effects of growth conditions on expression of mycobacterial murA and tyrS genes and contributions of their transcripts to precursor rRNA synthesis. *J Bacteriol* **181**, 4617–4627 (1999).
48. Eisenach, K. D. Use of an insertion sequence for laboratory diagnosis and epidemiologic studies of tuberculosis. *Ann Emerg Med* **24**, 450–453 (1994).
49. Gunisha, P., Madhavan, H. N., Jayanthi, U. & Therese, K. L. Polymerase chain reaction using IS6110 primer to detect Mycobacterium tuberculosis in clinical samples. *Indian J Pathol Microbiol* **44**, 97–102 (2001).
50. Warren, R. M., van Helden, P. D. & Gey van Pittius, N. C. Insertion element IS6110-based restriction fragment length polymorphism genotyping of Mycobacterium tuberculosis. *Methods Mol Biol (Clifton, N.J.)* **465**, 353–370 (2009).
51. Broussard, G. W. & Ennis, D. G. Mycobacterium marinum produces long-term chronic infections in medaka: A new animal model for studying human tuberculosis. *Comp Biochem Physiol C: Toxicol Pharmacol* **145**, 45–54 (2007).
52. Bosserman, R. E. & Champion, P. A. Esx systems and the mycobacterial cell envelope: What's the connection. *J Bacteriol* **199**, e00131–17 (2017).

53. Chin, C.-S. *et al*. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Meth* **10**, 563–569 (2013).
54. Tritt, A., Eisen, J. A., Facciotti, M. T. & Darling, A. E. An integrated pipeline for de novo assembly of microbial genomes. *Plos One* **7**, e42304 (2012).
55. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)* **30**, 2068–2069 (2014).
56. Hyatt, D. *et al*. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
57. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucl Acids Res* **32**, 11–16 (2004).
58. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucl Acids Res* **35**, 3100–3108 (2007).
59. Aziz, R. K. *et al*. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).
60. Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. PHAST: a fast phage search tool. *Nucleic Acids Research* **39**, W347–52 (2011).
61. Hawkey, J. *et al*. ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. *BMC Genomics* **16**, 667 (2015).
62. Fouts, D. E., Brinkac, L., Beck, E., Inman, J. & Sutton, G. PanOCT: annotated clustering of orthologs using conserved gene neighborhood for pan-genome analysis of bacterial strains and closely related species. *Nucl Acids Res* **40**, e172–e172 (2012).
63. Ucko, M. & Colorni, A. Mycobacterium marinum infections in fish and humans in Israel. *J Clin Microbiol* **43**, 892–895 (2005).
64. Hall-Stoodley, L., Brun, O. S., Polshyna, G. & Barker, L. P. Mycobacterium marinum biofilm formation reveals cording morphology. *FEMS Microbiol Lett* **257**, 43–49 (2006).
65. Weerdenburg, E. M. *et al*. Genome-wide transposon mutagenesis indicates that Mycobacterium marinum customizes its virulence mechanisms for survival and replication in different hosts. *Infect Immun* **83**, 1778–1788 (2015).
66. Kurokawa, S. *et al*. Bacterial classification of fish-pathogenic Mycobacterium species by multigene phylogenetic analyses and MALDI biotyper identification system. *Mar Biotech* **15**, 340–348 (2012).

Acknowledgements

We thank our colleagues for discussions. Sequencing was performed by the SNP@SEQ Technology Platform in Uppsala, which is part of the Science for Life Laboratory at Uppsala University and supported as a national infrastructure by the Swedish Research Council. The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project b2011072. This work was funded by the Swedish Research Council (M and N/T), the Swedish Research Council for Environment, Agricultural Sciences, and Spatial Planning (FORMAS), and Uppsala RNA Research Center (Swedish Research Council Linneus support).

Author Contributions

L.A.K. and D.G.E. conceived the study. S. Das and P.R.K.B. performed the bioinformatics analysis, and M.R. the growth rate experiments. S. Das, B.M.F.P., D.G.E. and L.A.K. analyzed and interpreted the data. A.M., M.C., L.S. and T.D. maintained, cultivated and prepared DNA from different *Mma* strains. B.M.F.P. and D.G.E. generated culture extracts and DNA isolation. S. Das, S. Dasgupta, M.C., D.G.E. and L.A.K. wrote the manuscript. All authors read and approved the final version of the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-30152-y.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.