Insights from the complete genome sequence of *Mycobacterium marinum* on the evolution of *Mycobacterium tuberculosis*

Timothy P. Stinear,1,10 Torsten Seemann,2 Paul F. Harrison,2 Grant A. Jenkin,1 John K. Davies,1 Paul D.R. Johnson,3 Zahra Abdellah,4 Claire Arrowsmith,4 Tracey Chillingworth,4 Carol Churcher,4 Kay Clarke,4 Ann Cronin,4 Paul Davis,4 Ian Goodhead,4 Nancy Holroyd,4 Kay Jagels,4 Angela Lord,4 Sharon Moule,4 Karen Mungall,4 Halina Norbertczak,4 Michael A. Quail,4 Ester Rabbinowitsch,4 Danielle Walker,4 Brian White,4 Sally Whitehead,4 Pamela L.C. Small,5 Roland Brosch,6 Lalita Ramakrishnan,7 Michael A. Fischbach,8 Julian Parkhill,4 and Stewart T. Cole9

1Department of Microbiology, Monash University, Clayton 3800, Australia; 2Victorian Bioinformatics Consortium, Monash University, Clayton 3800, Australia; 3Department of Infectious Diseases, Austin Hospital, Heidelberg 3084, Australia; 4Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom; 5Department of Microbiology, University of Tennessee, M409 Walters Life Sciences, Knoxville, Tennessee 37996-0845, USA; 6Institut Pasteur, 28 rue du Dr. Roux, 75724 Paris, France; 7Department of Microbiology, University of Washington, Seattle, Washington 98195, USA; 8Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA; 9Global Health University, Clayton 3800, Australia; 10Department of Infectious Diseases, Austin Hospital, Heidelberg 3084, Australia.

*Mycobacterium marinum*, a ubiquitous pathogen of fish and amphibia, is a near relative of *Mycobacterium tuberculosis*, the etiologic agent of tuberculosis in humans. The genome of the M strain of *M. marinum* comprises a 6,636,827-bp circular chromosome with 5424 CDS, 10 prophages, and a 23-kb mercury-resistance plasmid. Prominent features are the very large number of genes (57) encoding polyketide synthases (PKSs) and nonribosomal peptide synthases (NRPSs) and the most extensive repertoire yet reported of the mycobacteria-restricted PE and PPE proteins, and related-ESX secretion systems. Some of the NRPS genes comprise a novel family and seem to have been acquired horizontally. *M. marinum* is used widely as a model organism to study *M. tuberculosis* pathogenesis, and genome comparisons confirmed the close genetic relationship between these two species, as they share 3000 orthologs with an average amino acid identity of 85%. Comparisons with the more distantly related *Mycobacterium avium* subspecies *paratuberculosis* and *Mycobacterium smegmatis* reveal how an ancestral generalist mycobacterium evolved into *M. tuberculosis* and *M. marinum*. *M. tuberculosis* has undergone genome downsizing and extensive lateral gene transfer to become a specialized pathogen of humans and other primates without retaining an environmental niche. *M. marinum* has maintained a large genome so as to retain the capacity for environmental survival while becoming a broad host range pathogen that produces disease strikingly similar to *M. tuberculosis*. The work described herein provides a foundation for using *M. marinum* to better understand the determinants of pathogenesis of tuberculosis.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to GenBank under accession nos. CP000854 (chromosome) and CP000895 (pMM23 plasmid).]

In 1926, Joseph D. Aronson isolated a *Mycobacterium* from tubercles observed predominantly in the spleen and liver of diseased fish that had died in the Philadelphia Aquarium and named it *Mycobacterium marinum* (Aronson 1927). *M. marinum* was subsequently shown to also be a human pathogen when it was isolated again much later in a swimming pool-associated outbreak of human granulomatous skin lesions, although in this report the *Mycobacterium* was mistakenly given a new species name, *Mycobacterium balnei* (Linell and Norden 1954), a name that is no longer used.

Several phylogenetic studies have shown that *M. marinum* is most closely related to *Mycobacterium ulcerans* (97% nucleotide identity) and then *M. tuberculosis* (85% nucleotide identity) (Tønjum et al. 1998; Stinear et al. 2000; Devulder et al. 2005; Gey van Pittius et al. 2006; Yip et al. 2007). Multi-locus sequence typing (MLST) and the recently completed *M. ulcerans* genome sequence clearly show that *M. ulcerans* has recently evolved from a *M. marinum* progenitor by acquisition of a virulence plasmid and reductive evolution (Stinear et al. 2004, 2007; Yip et al. 2007). *M. ulcerans* causes the devastating human skin disease Buruli ulcer, which is quite distinct in its pathology from other mycobacterial diseases, primarily because of the production from
its virulence plasmid of mycolactone, an immunosuppressive polyketide toxin restricted to \(M. \text{ ulcerans}\) (George et al. 1999).

\(M. \text{ marinum}\) infection of humans, called fish tank or aquarium tank granuloma, typically occurs when \(M. \text{ marinum}\) is inoculated through the skin by cuts and scratches following direct contact with an infected fish or contaminated aquatic environments (Petrini 2006). The ensuing granulomatous infection generally limited to the skin and soft tissues extremities is pathologically indistinguishable from \(M. \text{ tuberculosis}\) dermal disease (Travis et al. 1985; MacGregor 1995). Like tuberculosis, fish tank granuloma can resolve spontaneously but frequently requires long-term antibiotic treatment owing to the infecting bacteria developing phenotypic antibiotic resistance in vivo (Aubry et al. 2000, 2002). \(M. \text{ marinum}\) disease can be effectively treated with the anti-tuberculosis agents rifampin, ethambutol, and the quinolones, as well as other agents such as doxycycline and clarithromycin (Aubry et al. 2000, 2002). The systemic granulomatous disease caused by \(M. \text{ marinum}\) in its poikilothermic hosts, for example, frogs and fish, also shares key histological features with human tuberculosis (Pozos and Ramakrishnan 2004; Swaim et al. 2006). For instance, zebrafish develop organized caseating granulomas that resemble active human tuberculosis (Swaim et al. 2006). The two organisms share virulence determinants, and \(M. \text{ tuberculosis}\) virulence genes complement orthologous \(M. \text{ marinum}\) genes with mutations (Cosma et al. 2003, 2006; Gao et al. 2003, 2004; Volkman et al. 2004; Bhatt et al. 2007a). Because of these findings and the relative ease of its manipulation, \(M. \text{ marinum}\) infection of its natural host species is used as a model to study tuberculosis (Saad et al. 1999; Davis et al. 2002; Guinn et al. 2004; Volkman et al. 2004; Bhatt et al. 2007a). General features of the \(M. \text{ marinum}\) genome

### Table 1. Chromosome features of \(M. \text{ marinum}\) M compared with four other mycobacteria

<table>
<thead>
<tr>
<th>Feature</th>
<th>(M. \text{ marinum}) M</th>
<th>(M. \text{ ulcerans}) Agy99</th>
<th>(M. \text{ tuberculosis}) H37Rv</th>
<th>(M. \text{ leprae}) TN</th>
<th>(M. \text{ avium}) paratuberculosis K10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chromosome size (base pairs)</strong></td>
<td>6,636,827</td>
<td>5,631,606</td>
<td>4,111,532</td>
<td>3,268,203</td>
<td>4,829,781</td>
</tr>
<tr>
<td><strong>G+C (%)</strong></td>
<td>65.73</td>
<td>65.47</td>
<td>65.61</td>
<td>57.8</td>
<td>69.3</td>
</tr>
<tr>
<td><strong>Gene density (base pairs per gene)</strong></td>
<td>1,223</td>
<td>1,353</td>
<td>1,110</td>
<td>2,036</td>
<td>1,110</td>
</tr>
<tr>
<td><strong>Average CDS length</strong></td>
<td>1,101</td>
<td>979</td>
<td>1,009</td>
<td>1,010</td>
<td>1,106</td>
</tr>
<tr>
<td><strong>Protein-coding sequences (CDS)</strong></td>
<td>5,424</td>
<td>4,160</td>
<td>3,974</td>
<td>1,605</td>
<td>4,350</td>
</tr>
<tr>
<td><strong>Conserved with assigned function</strong></td>
<td>3,987 (74%)</td>
<td>2,821 (68%)</td>
<td>3,049 (77%)</td>
<td>1,167 (73%)</td>
<td>3,223 (74%)</td>
</tr>
<tr>
<td><strong>Conserved with unknown function</strong></td>
<td>1,254 (25%)</td>
<td>1,241 (30%)</td>
<td>9,017 (22.5%)</td>
<td>3,187 (20%)</td>
<td>1,088 (25%)</td>
</tr>
<tr>
<td><strong>Unique</strong></td>
<td>183 (3%)</td>
<td>98 (2%)</td>
<td>18 (0.5%)</td>
<td>120 (7%)</td>
<td>39 (1%)</td>
</tr>
<tr>
<td><strong>Pseudogenes</strong></td>
<td>65</td>
<td>771</td>
<td>17</td>
<td>1115</td>
<td>None reported</td>
</tr>
<tr>
<td><strong>rRNA</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>tRNA</strong></td>
<td>46</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
</tr>
</tbody>
</table>

These observations suggest that \(M. \text{ marinum}\) has the characteristics of both a specialist bacterium, adapted to persist within intracellular environments, such as free-living aquatic amoeba and vertebrate hosts, as well as of a generalist, equipped to survive under changing and sometimes extracellular environmental conditions (Stamm and Brown 2004). A dichotomous lifestyle was first proposed by the seminal transmission studies of Clark and Shepard (1963) when they observed that fish, parenterally infected with \(M. \text{ marinum}\), provided a source of water-borne mycobacteria that could, in turn, infect tadpoles.

In contrast, \(M. \text{ tuberculosis}\) has become highly specialized for intracellular survival in a very restricted range of mammalian hosts, and several recent studies have shown that lateral gene transfer has been a major force in the evolution of the \(M. \text{ tuberculosis}\) complex from an environmental Mycobacterium (Kinsella et al. 2003; Gutierrez et al. 2005; Rosas-Magallanes et al. 2006; Beq et al. 2007). Here, we have determined the genome sequence of \(M. \text{ marinum}\) and then exploited its very close genetic relationship with \(M. \text{ tuberculosis}\) to provide further insight into the multiple changes that shaped \(M. \text{ tuberculosis}\) into a major human pathogen from environmental origins.

### Results

**General features of the \(M. \text{ marinum}\) genome**

Multi-locus sequence typing (MLST) among 26 diverse \(M. \text{ marinum}\) strains from around the world has revealed significant intraspecies sequence homogeneity, with <3% nucleotide variation present between all \(M. \text{ marinum}\) STs (Stinear et al. 2000; Yip et al. 2007). Nine distinct sequence types (ST) were identified, with the "M" strain representing an intermediate ST, designated ST22, and this strain was chosen for sequence analysis.

The \(M. \text{ marinum}\) genome is composed of a single circular chromosome of 6,636,827 bp with an average G+C content of 65.7%.
Complete genome sequence of *Mycobacterium marinum*

62.5%, 5424 predicted coding sequences (CDS), 65 pseudogenes, 46 tRNA genes, a single rRNA operon, and a 23-kb plasmid named pMM23 (Table 1; Fig. 1A,B). Functional information could be assigned to 74% of all CDS, 23% encode conserved hypothetical proteins, and 3% were unique (Table 1). The chromosome harbors seven distinct, low-copy-number insertion sequences, named ISMyma01-07 (Supplemental Table 1), and 10 putative prophages, named phiMmar01-10 (Supplemental Table 2). GC skew analysis demonstrated a bias in G+C content between the leading and lagging strands (Fig. 1A), indicating a probable origin of replication around *dnaA* (Lobry 1996). AT skew analysis was also used and revealed some significant sequence biases (Fig. 1C), detected to a lesser extent in *M. ulcerans* and *M. tuberculosis* and absent from other mycobacteria. These sequence deviations are caused by the presence of PE-PGRS and nonribosomal peptide synthase (NRPS) genes, with their distinctive codon usage, repeating functional domains and sequence motifs (Supplemental Fig. 1). The plasmid encodes 29 predicted CDS, including replication and partition proteins, along with a 1110-amino acid relaxase/helicase that may function in conjugation or mobilization. The plasmid also contains a functional mercury-resistance operon (M. Schue, L.G. Dover, G.S. Besra, J. Parkhill, and N.L. Brown, unpubl.), flanked by two genes encoding members of the serine-site-specific recombinase family; it is possible that these may confer independent mobility on this element. Analysis by PCR of the distribution of pMM23 among 42 diverse *M. marinum* isolates revealed that only the M strain contains this replicon (data not shown).

Figure 1. (A) Circular representation of the *M. marinum* chromosome. The scale is shown in megabases in the outer black circle. Moving inward, the next two circles show forward- and reverse-strand CDS, respectively, with colors representing the functional classification. (Red) Information pathways; (light blue) regulation; (light green) hypothetical protein; (dark green) cell wall and cell processes; (orange) conserved hypothetical protein; (cyan) IS elements; (yellow) intermediate metabolism; (gray) lipid metabolism; (purple) PE/PPE. The location of each copy of a PE or PPE gene is then shown (purple). The following two circles show forward- and reverse-strand genes that are not found in other mycobacteria (colors represent the functional classification), followed by the G+C content and finally the GC skew (G–C)/(G+C) using a 20-kb window. (B) Circular representation of the *M. marinum* mercury-resistance plasmid pMM23, following the same color scheme as the chromosome map but with a different scale in kilobase pairs as indicated. (C) AT skew analysis generated with Gene-Spaghetti, a method for visualizing DNA composition. The colors reflect AT skew (T − A)/(T + A). (Red) Extreme = 60% T-rich; (blue) extreme = 60% A-rich. The software uses a sliding Gaussian window to estimate local base usage (http://www.vicbioinformatics.com/genespaghetti.shtml). Outlier locus tags were represented by three types of CDS; (blue) NRP, (gray) PE-PGRS; (red) PPE.

Central metabolism, respiration, and gene regulation

As identified from other mycobacterial genomes, *M. marinum* has intact glycolysis and pentose phosphate pathways but no Entner-Doudoroff pathway and a bifurcated TCA cycle (Tian et al. 2005). The expanded gene families encoding the key enzymes of the beta-oxidation pathway suggest that, like *M. tuberculosis*, exogenous fatty acids are important for the growth of *M. marinum*. FadD (acyl-CoA synthase) catalyzes the formation of acyl-CoA thioesters that are then processively oxidized by FadE, EchA, and FadA (Cole et al. 1998). In *M. marinum*, 32 fadD paralogs were predicted compared with 23 in *M. tuberculosis*. There are 46 paralogous *fadE* genes compared with 34 in *M. tuberculosis*. There are also multiple copies of other key enzymes in the cycle including 35 EchA and nine FadA proteins (compared with 22 and six, respectively, in *M. tuberculosis*).

In addition to fatty acids, *M. marinum* is predicted to use a diverse catabolic repertoire including compounds like pyruvate, glucose, succinate, and acetate as sole carbon sources. This is also supported by 192 CDS predicted to be associated with carbon substrate transport. Overall, *M. marinum* has at least double the content of every class of transporter compared with *M. tuberculosis* (Table 2). There is a predominance of ATP-dependent transporters represented by 165 predicted ABC-type, 21 P-type, and eight F-type ATPase family members.

Based on gene orthology with *M. tuberculosis*, complete NADH oxidase, ubiquinol cytochrome c oxidase, and ATP synthase complexes indicate ATP generation by oxidative phosphorylation during aerobic growth. In *M. tuberculosis*, anaerobic
growth is facilitated by the nitrate and fumarate reductase systems; however, *M. marinum* lacks the corresponding genes but contains a potential alternative anaerobic respiratory pathway involving nitrite reductase (NrfD) and formate dehydrogenase (MMAR_5194, MMAR_5195), where formate oxidation is coupled to nitrite reduction. These observations are in agreement with the species-defining nitrate reductase negative phenotype of *M. marinum*. Formate dehydrogenase is a selenoprotein, and its gene is linked to a selenocysteine synthase locus (MMAR_5190). Selenoenzymes are often associated with anaerobic respiratory processes but have not previously been reported in mycobacteria. There are seven putative NarK nitrite transporters, and nitrite is a common component of aquatic ecosystems.

A versatile respiratory potential under aerobic conditions is suggested in *M. marinum* by more than 590 putative oxidoreductases, dehydrogenases, mono- and dioxygenases. The cytochrome P450 repertoire is particularly large with 47 P450 monooxygenases, some of which may play a role in specific hydroxylation of certain metabolites, while others will probably be involved in targeted oxidation of different xenobiotics encountered by the bacterium.

Versatility is also reflected in the large number of CDS (330) predicted to be involved in gene regulation, including the response to environmental change, such as those for the 17 sigma factors, 16 complete two-component regulatory systems, and 24 serine-threonine protein kinases (Wehenkel et al. 2007).

**Cell wall lipids**

Some predictions from the genome regarding the metabolism of key structural lipids of *M. marinum* have been described previously (Onwueme et al. 2005; Stinear et al. 2007). *M. marinum* has all the components of the type I and II fatty acid synthase complexes that are central to the production of cell wall mycolates. Similarly, *M. marinum* has the five type I polyketide synthases (PKS), encoded by the genes *ppsA-E*, that are required for the production of phenolphthiocerol, which can be further modified by glycosylation to form the virulence factors and immune modulators, the phenolic glycolipids (PGLs) (Daffe et al. 1992; Reed et al. 2004). Refer to Onwueme et al. (2005) for a detailed analysis of the biosynthesis of PGLs and related molecules in *M. marinum*.

Recent studies have shown that *M. marinum* also produces four distinct but highly related species of lipoooligosaccharide (LOS), the biosynthesis of which has been linked to the presence of a locus between MMAR_2307 and MMAR_2320 (Burguier et al. 2005; Ren et al. 2007). LOS production appears important for sliding motility and biofilm formation in *M. marinum*. Genome comparisons suggest that the equivalent region in *M. tuberculosis* has become nonfunctional via successive DNA deletions and has continued to decay in *Mycobacterium bovis*, where the region of difference, RD4, was identified (Brosch et al. 1998; Brosch et al. 2002).

Polypropenyl phosphate (Pol-P) forms the lipid backbone of the sugar-linked intermediates that are required by bacteria during cell wall biosynthesis. Pol-P is formed by sequential condensation of isopentenyl dipiphosphate (IPP) and dimethylallyl diprophosphate (DMAPP). In other mycobacteria, including *M. tuberculosis*, these molecules are synthesized via the non-mevalonate (DXP) pathway. However, *M. marinum* contains both an intact mevalonate (MVA) and DXP pathway for polypropenoid synthesis, the first example of such in the mycobacteria (Rohdich et al. 2004), although *M. ulcera* has a remnant, nonfunctional version of the MVA pathway. Comparative genomics suggests that the 11-kb region spanning the MVA locus was deleted from a progenitor of *M. tuberculosis* via the activity of IS1558 (Rv2177c).

**Secondary metabolism**

Little is known of the secondary metabolites produced by *M. marinum*; however, genome inspection suggests that it has the potential to produce a diverse array of compounds with 34 CDS predicted to encode PKSs. In addition to the *pps* locus (*ppsA-E*) discussed above, three other PKSs have *M. tuberculosis* orthologs of known function, namely, *pks12* (MMAR_3025) required for mycoketide production (Matsunaga et al. 2004); *pks15*/*1* (MMAR_1762), encoding a free-standing PKS module required for production of phenolphthiocerol (Constant et al. 2002); and *pks13* (MMAR_5364), encoding another free-standing PKS module required for mycolic acid synthesis (Portevin et al. 2004). *M. marinum* also has the *pks10*, 7, 8, 9 & 11 locus, encoding a PKS of unknown function, whose predicted module and domain structure are identical to that of *M. tuberculosis*, where it is essential (Sassetti et al. 2001).

Among the other predicted PKSs, two systems are worthy of comment—*MMAR_1180* and the *MMAR_3796-MMAR_3800* system. Neither of these has a strict ortholog in other sequenced mycobacterial genomes, although the second module of *MMAR_1180* and several of the *MMAR_3796-MMAR_3800* modules have strong homology with the mycolactone PKS from *M. ulcerans*. These PKs comprise modules that are more closely related to Streptomyces PKSs than other mycobacterial PKSs, which suggests that they might encode small molecule products that are novel among known mycobacteria.

*M. marinum* also has 14 CDS encoding predicted fatty acyl-AMP ligases (FAAL), six of which are shared with *M. tuberculosis* (FadD25, 26, 28, 29, 30, 32). FAALs are a family of FadD-like enzymes that have been shown in *M. tuberculosis* to adenylate long-chain fatty acids to facilitate further chain extension by PKS (Trivedi et al. 2004). They promote biochemical cross-talk between distinct FAS and PKS loci, thereby increasing metabolite diversity (Gokhale et al. 2007).

A most prominent feature of the *M. marinum* genome is its distinctive array of large CDS encoding 23 nonribosomal peptide

---

**Table 2. Principal transport proteins in *M. marinum* M compared with *M. tuberculosis* H37Rv**

<table>
<thead>
<tr>
<th>Transporter type</th>
<th>Transporter family</th>
<th><em>M. marinum</em></th>
<th><em>M. tuberculosis</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP-dependent</td>
<td>ABC</td>
<td>165</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>ATPase</td>
<td>29</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>ArsAB</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Secondary</td>
<td>MFS</td>
<td>53</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>RND</td>
<td>30</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>APC</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Nramp</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>SuIP</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Tat</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Ion channel</td>
<td>CIC</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>MIT</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>MscL</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>MscS</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

*Transporter family classification based on the scheme used in TransportDB (http://www.membranetransport.org/).
The NRPS encoded by MMAR_3268, comprising 9859 amino acids arranged in nine modules (Fig. 2A), is the largest protein in the *M. marinum* proteome, with the second being another, highly similar, NRPS, MMAR_0851, at 8760 amino acids and eight modules. These NRPSs are, in turn, closely related to two other NRPSs, the products of MMAR_3270 and MMAR_3271. With the exception of the first condensation (C) domains of MMAR_3268, MMAR_0851, and MMAR_3270, all of the other C domains are ∼88% identical at the amino acid level, and the cognate modules appear to have arisen from duplication events, in a similar manner to that seen for the mycolactone PKS in *M. alcearens* (Stine et al. 2004). From the phylogeny of the *M. marinum* C domains (Fig. 2B), it can be seen that there is a distinct clade (refer to the lower end of Fig. 2B) representing the C domains from these four NRPS. This is consistent with a model in which both a mixture of gene duplication and intragene module duplication have contributed to the evolution of these gene clusters. Supporting this view, the C domains from MMAR_3097 and MMAR_3176 also form their own clades, albeit a bit deeper, suggesting they may have evolved further.

The picture looks similar for the adenylation (A) domains, although their level of conservation is slightly lower than the C domains. This is a very intriguing situation since these NRPSs appear to have very similar A domains in each module, and as a result, the amino acid substrates of these domains cannot be predicted reliably using structure-based and homology-based models. Interestingly, the NRPS Mps, which synthesizes the tetraketidyl alcohol backbone of the mycobacterial glycopeptidolipids (GPLs) and was first described in a strain of *M. smegmatis* (Billman-Jacobe et al. 1999), comprises four modules that—like the *M. marinum* NRPS modules—seem to have arisen from some form of module duplication. Since it is known that the Mps A domains activate different amino acid substrates (one phenylalanine, one allo-threonine, and two alanine), it is likely that these differences in substrate selectivity resulted from divergent evolution following duplication. The example of Mps has two implications for the *M. marinum* NRPSs. First, it suggests that despite the high degree of amino acid sequence identity among the A domains in MMAR_3268, MMAR_0851, MMAR_3270, and MMAR_3271, these domains may have evolved divergently to accept different amino acid substrates. Since most known NRPSs are composed of modules with disparate evolutionary origins, the *M. marinum* NRPSs would be a highly unusual example of NRPS “innovation” by divergent evolution and could have important implications for efforts to custom-design NRPSs that make unnatural products. Second, it raises the possibility that the three-module NRPS MMAR_0099, which has A domains that are related to those of Mps, could synthesize an unusual nonreduced GPL similar to the serine-containing GPL isolated from *Mycobacterium xenopi* (Besra et al. 1993). In support of this conjecture, MMAR_0099 is clustered with PKS-encoding genes that could synthesize the lipid portion of a noncanonical GPL.

The secondary metabolite repertoire of *M. marinum* is not limited to fatty acids, polyketides, and nonribosomal peptides. *M. marinum*, in contrast to *M. tuberculosis*, is photochromogenic, synthesizing carotenoid pigments upon exposure to light. Mutational analysis identified several genes encoding carotenoid biosynthesis pathway enzymes that are essential for photochromogenicity (Ramakrishnan et al. 1997; Gao et al. 2003). In support of these previous findings, the remainder of this locus appears to be involved in the synthesis of a beta-carotene-like terpene that is predicted to have photochromogenic properties. MMAR_4805 is a *trans*-isoprenyl diphosphate synthase that would polymerize IPP units to form geranylgeranyl-pyrophosphate, MMAR_4807 is a phytoene synthase (CrtB ortholog) that would catalyze a head-

---

Figure 2. (A) Module and domain architecture of selected NRPSs. (B) Inferred phylogeny using the Neighbor-joining method (Saitou and Nei 1987) of selected *M. marinum* condensation domains based on aligned predicted amino acid sequences. Different colors highlight condensation domains within an operon. Replicate trees in which the associated domains clustered together on <80% of occasions in a bootstrap test (10,000 replicates) are indicated. The scale is the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated from the data set.
to-head coupling of geranylgeranyl-pyrophosphate to form phytoene. MMAR_4806 is a phytoene dehydrogenase (CrtI ortholog) that would desaturate phytoene to form the conjugated polynuclear core of the carotenoids, and MMAR_4808–4809 encode a two-component lycopene cyclase—homologous to CrtY/CrtYd from the Brevibacterium linens isorenieratene gene cluster (Krubasik and Sandmann 2000)—that would cyclize the termini to form a beta-carotene or a similar C40 terpene with photochromogenic properties.

All these observations underscore a complex and diverse secondary metabolome within M. marinum and are a point of significant difference compared with M. tuberculosis and other mycobacteria, including M. ulcerans, the latter species having lost many of these regions by deletion (Stinear et al. 2007).

ESX loci and ESAT-6 proteins

The ESX loci are large gene clusters that encode the dedicated ATP-dependent secretory apparatus required to export specific members of the 6-kDa early secreted antigenic target (ESAT-6) protein family, together with their corresponding effector proteins. In M. marinum there are 29 esx genes, compared with 23 in M. tuberculosis (Supplemental Table 3). These are within at least five ESX loci including the prototypical ESX-1 system, which is a major virulence determinant in both M. tuberculosis (Lewis et al. 2003; Brodin et al. 2004; Guinn et al. 2004) and M. gormanii (Gao et al. 2004; Volkman et al. 2004). It is noteworthy that a 3.7-kb portion of the ESX-1 region is partially duplicated (ESX-6, MMAR_0184–MMAR_0188), resulting in two copies of essA and essB that are almost identical. Although it is not clear at present whether essA,1 and essB,1 are functional, this illustrates again how members of the ESX gene family evolve by duplication and subsequent diversification (Tekaia et al. 1999).

ESX-3 (spanning nucleotides 636,000–651,200) is another system shared between M. marinum and M. tuberculosis, whereas ESX-2 may be defective in M. marinum M as indicated by partial deletion of a membrane-bound component, MMAR_5460. ESX-3 is essential for in vitro growth of M. tuberculosis (Sassetti et al. 2001), and its expression is controlled by the zinc uptake regulator Zur/FurB (Maciag et al. 2007), which is also present in M. marinum. Orthologous systems of ESX-3 have been found in all mycobacteria sequenced to date, suggesting a fundamental role for this region in the mycobacterial life cycle. The other shared ESX loci are ESX-4 and ESX-5. While ESX-4 seems to be the most ancestral ESX-system based on its presence in other actinobacteria (Gey Van Pittius et al. 2001), ESX-5 has been shown to be involved in export of the PPE41 protein in M. marinum (Abdallah et al. 2006).

In addition to ESAT-6 proteins, ESX systems also appear to secrete other effector proteins, often encoded by unlinked genes, such as EspA (Rv3616c) in M. tuberculosis. Inactivation of espA, which is paralogous to Rv3864, incapacitates the ESX-1 system, and neither ESAT-6 nor CFP-10 is secreted, thus leading to severe attenuation of virulence (Fortune et al. 2005; MacGurn et al. 2005). In M. marinum, the EspA ortholog (MMAR_4166) belongs to an 18-membered family, which also includes MMAR_5439, the ortholog of Rv3864. Surprisingly, the genes for 15 of these EspA family members occur within a 46-kb region encompassing ESX-1, with 14 of them arranged in tandem upstream of the locus. While it is clear that this is yet another example of gene duplication associated with ESX systems, it is not known whether all of these EspA paralogs are functional or if they are secreted.

In summary, it is striking that the ESX complements of M. marinum and M. tuberculosis are so similar despite their different genome sizes, while M. smegmatis, with a large genome (7.0 Mb), has only three ESX loci and is lacking ESX-1. ESX-1 is required for both intercellular bacteria spread between macrophages as well as granuloma formation—two key requirements for mycobacterial pathogenesis (Cosma et al. 2003; Guinn et al. 2004; Volkman et al. 2004; Swaim et al. 2006). Therefore, the genomic data are consistent with the acquisition of these systems being driven by the selection of an intracellular niche during evolution of these two pathogens. In contrast, M. smegmatis, the only pathogenic Mycobacterium that does not have an intracellular lifestyle within granulomas (its polyketide toxin enables the bacteria to lyse cells and dwell extracellularly), has only 13 esx genes and three intact ESX loci but not ESX-1, which has incurred two deletions (Stinear et al. 2007).

PE/PPE proteins

The PE/PPE proteins are enigmatic mycobacterial proteins, so-called because of their characteristic proline-glutamate (PE), or proline-proline-glutamate (PPE) N-terminal motifs. M. marinum has 281 CDS distributed around its genome encoding 27 PE, 148 PE-PGRS, and 106 PPE proteins (Fig. 1A). Compared with 115 in M. ulcers and 169 in M. tuberculosis, M. marinum has the most extensive repertoire yet reported of these mycobacteria-specific proteins representing 9.1% of its coding capacity, compared with 7.1% for M. tuberculosis and 3.8% for M. ulcers. It has been proposed that the PE/PPE genes coevolved with the ESX loci and then underwent a specific and significant expansion in the common progenitor of M. tuberculosis, M. marinum, and M. ulcers (Gey Van Pittius et al. 2006). Such a large collection of these genes in M. marinum may be consistent with its diverse host range as some PE-PGRS proteins from M. marinum are known to be involved in intracellular persistence (Ramakrishnan et al. 2000). Significant amino acid sequence divergence occurs, predominantly within the C-terminal region of these proteins (Garnier et al. 2003; Gey Van Pittius et al. 2006), suggesting that they are subject to constant diversifying selection, perhaps to avoid detection or destruction during intracellular life in a variety of hosts.

Lipoproteins

Like the Esp and PE/PPE proteins, lipoproteins play an important role in mycobacterial virulence by interacting with the host immune responses; for example, LpqH-dependent signaling through TOLL-like receptor 2 (Rezwan et al. 2007). Despite their different genome sizes, M. marinum and M. tuberculosis have similar numbers of CDS encoding predicted lipoproteins. M. marinum has 87 predicted lipoproteins, compared with 89 in M. tuberculosis. Furthermore, some lipoprotein sequences vary significantly between each Mycobacterium such that 14 putative lipoproteins in M. tuberculosis, encoded by lppA,B,C,G,O,Q,T,Y, lpeL,P, lpbX, lipR,F, and pstS1, have no ortholog in M. marinum. Interestingly too, M. smegmatis has only 24 CDS encoding predicted lipoproteins. Variation in the lipoprotein content may reflect the different host ranges of each Mycobacterium and perhaps indicate the importance of specific lipoproteins for M. marinum and M. tuberculosis (Rezwan et al. 2007).
Evolution of M. tuberculosis and M. marinum from a recent common ancestor

Pairwise reciprocal best-hit FASTA analysis was used to define all orthologs between each of M. marinum M, M. tuberculosis H37Rv, M. ulcerans Agy99, M. avium paratuberculosis K10, Mycobacterium leprae TN, and M. smegmatis mc²155. The results are summarized in ortholog frequency plots (Fig. 3A). These comparisons also revealed a common core set of 1072 orthologs (1755 if M. leprae is excluded). A summary of the average percent amino acid identity for these core CDS is presented in Supplemental Table 4 and was used to infer a phylogeny (Fig. 3B) that demonstrates M. marinum is most closely related to M. ulcersans and then to M. tuberculosis (3000 shared orthologs) and least related to M. smegmatis, a phylogeny in close agreement with those based on 16S rRNA gene sequences (Gey van Pittius et al. 2006).

Relationships based on shared amino acid identity contrast strikingly with the relatively high conservation of genome size and gene order between M. marinum and the 7.0-Mb genome of M. smegmatis (Fig. 3C), a conservation that probably reflects relatively tight selection constraints on genome architecture for environmental mycobacteria. Genome comparisons between M. marinum, M. tuberculosis H37Rv (4.4 Mb), and M. avium subsp. paratuberculosis K10 (4.8 Mb) show the rearrangements and genome reduction that have occurred within these more closely related species (Fig. 3C).

In a complementary approach to identifying orthologs, a BLASTCLUST analysis was applied to M. marinum, M. tuberculosis, M. ulcersans, M. avium subspecies paratuberculosis, and M. smegmatis. The results are summarized in three Venn diagrams (Fig. 4) and show CDS that are both conserved and unique between these species. In the three different three-way comparisons—taking M. tuberculosis and comparing it with either M. marinum and M. ulcersans, M. marinum and M. avium subsp. paratuberculosis, or M. marinum and M. smegmatis—there remained around 633 CDS found only in M. tuberculosis (Fig. 4). Furthermore, these CDS were the same in each three-way comparison, suggesting that M. tuberculosis possesses about 633 specific CDS that have (1) been acquired by lateral gene transfer (LGT) or (2) been subject to endogenous mutagenic processes such as duplication, recombination, or sequence amelioration since diverging from the most recent common ancestor (MRCA) of M. marinum and M. tuberculosis.

A detailed analysis of the 633 M. tuberculosis-specific CDS identified by BLASTCLUST (Fig. 4B) using the Artemis Comparison Tool in three-way whole-genome BLASTN comparisons, and Alien_Hunter to identify potential foreign DNA (Fig. 5), shows that these CDS span ~600 kb and comprise 140 distinct regions of difference between M. marinum and M. tuberculosis, of which 340 kb, spanning at least 80 regions and containing 360 CDS, appears to have been gained by LGT. Some of the “M. tuberculosis-specific” regions identified by this analysis code for known virulence factors or have been proposed to play a role in pathogenesis (Table 3; Supplemental Table 5). These include CDS that produce...
A three-way comparison with \textit{M. tuberculosis} and \textit{M. smegmatis} (Fig. 4C) revealed 785 CDS shared between \textit{M. marinum} and \textit{M. smegmatis} that are predicted to encode cell wall-associated proteins (148) or enzymes involved in intermediary metabolism such as hydroxylases, oxidases, glyoxylases, decarboxylases, and reductases (212); lipid metabolism (70); regulatory proteins (102); and conserved hypothetical proteins (207).

### Discussion

The close relationship between \textit{M. marinum} and \textit{M. tuberculosis} uncovered by this study provides deep insight into the evolution of pathogenesis in both these species. Their recent common origin is reflected in their shared and substantial common genetic core. There is a \textit{M. marinum} ortholog for 80% of all CDS in \textit{M. tuberculosis} with an average shared amino acid identity of 85%. However, somewhat paradoxically, whole-genome comparisons reveal a higher level of colinearity, synteny, and conservation of chromosome size between \textit{M. marinum} and the distant saprophyte, \textit{M. smegmatis} (76% average amino acid identity), than with \textit{M. tuberculosis}. The most parsimonious explanation for these observations is that \textit{M. marinum}, as a generalist, has preserved more of the characteristics of the hypothetical MRCA, while \textit{M. tuberculosis} has undergone a process of reductive evolution and adaptation, presumably for survival within mamalian hosts. In addition, the much lower content of IS elements in \textit{M. marinum} and \textit{M. smegmatis} has also contributed to maintaining synteny, as these elements are known to drive genome rearrangements (Parkhill et al. 2003), particularly in pathogens like \textit{M. ulcerans} that have undergone reductive evolution (Stinear et al. 2007).

Two striking features of the \textit{M. marinum} genome are the extensive range of genes encoding NRPSs and the huge repertoire of PE, PPE, and ESX systems. As seen previously in comparison of mycobacterial genomes (Garnier et al. 2003; Stinear et al. 2007), there is great variation in the size, sequence, and structure of loci encoding members of the PE, PPE, and ESX families. This variation reflects the diversity of the \textit{M. marinum} M genome not present in \textit{M. ulcerans} Agy99 (Stinear et al. 2007). A summary of these CDS and their predicted products is contained in Supplemental Table 6.

Classification by function shows that among the 55 regions, 11 appear to be involved in various metabolic processes, 16 encode predicted membrane proteins, and 10 correspond to prophages. There is a large locus, spanning MMAR\_3382--MMAR\_3414, which includes \textit{pmiA}, encoding a protein involved in synthesis of cell wall components, that inhibits phagosomal maturation (Robinson et al. 2007). Similarly, a 20-kb region between MMAR\_2955 and MMAR\_2988 (MURD73) harbors \textit{mimD}, a gene that appears to be required for survival within macrophages (Mehta et al. 2006).

A major conclusion of our study is the scale of DNA acquisition by LGT in \textit{M. tuberculosis}, and this is further evidence of adaptation. Indeed, several authors have proposed that \textit{M. tuberculosis} evolved from an environmental progenitor by LGT (Kisella et al. 2003; Rosas-Magallanes et al. 2006). Our three-way comparisons between \textit{M. marinum}, \textit{M. tuberculosis}, and \textit{M. avium} subsp. \textit{paratuberculosis} revealed at least 80 distinct regions that have probably been acquired by LGT in \textit{M. tuberculosis}. These span more than 360 CDS and account for 8% of the genome, a level twice that reported in a recent study using di-, tri-, and
Table 3. Potentially important MTB complex-specific CDS acquired by LGT

<table>
<thead>
<tr>
<th>CDS/locus description</th>
<th>Rv no. and/or gene name</th>
<th>Comments</th>
<th>AH* score</th>
<th>LGTb</th>
<th>Previous reports</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown secondary metabolite Locus of unknown function</td>
<td>Rv0031–Rv0035 (acpA, fadD34)</td>
<td>Enzymes likely to be involved in synthesis of a lipid</td>
<td>14.37</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Lipid glycosylation locus</td>
<td>Rv0057–Rv0080</td>
<td>DnaB intein, hypothetical proteins, including ATP-binding ABC transporter (Rv0073) and transcriptional regulator (Rv0067c)</td>
<td>16.52</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Phosphate transport locus</td>
<td>Rv0112–Rv0115 (gca, gnhA, gnhB, hddA)</td>
<td>Transferases predicted to be involved in lipid modification</td>
<td>17.61</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Putative virulence locus</td>
<td>Rv0933–Rv0936 (psb, psbS1, psbC1, psbA2)</td>
<td>Members of the phosphate regulon</td>
<td>12.48</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Lipid glycosylation locus</td>
<td>Rv0987–Rv0989</td>
<td>ABC transporter suggested to be involved in intracellular persistence</td>
<td>29.85</td>
<td>Yes</td>
<td>Rosas-Magallanes et al. 2006</td>
</tr>
<tr>
<td>Unknown secondary metabolite Anti-anti-sigma factor</td>
<td>Rv0057–Rv0080 (rsfA)</td>
<td>Regulates SigF</td>
<td>NP</td>
<td>Yes</td>
<td>Li et al. 2004</td>
</tr>
<tr>
<td>Lipooligosaccharide (LOS) locus</td>
<td>Rv1344–Rv1346 (fadD33, fadE14)</td>
<td>Enzymes likely involved in lipid synthesis</td>
<td>NP</td>
<td>?</td>
<td>Ren et al. 2007</td>
</tr>
<tr>
<td>Fumarate reductase operon</td>
<td>Rv1365, Rv1366 (rsfA)</td>
<td>Anaerobic respiratory chain</td>
<td>NP</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Lipoprotein locus</td>
<td>Rv1942–Rv1943 (includes gpg and locand toxin-antitoxin locus)</td>
<td>Production of an unknown metabolite; also includes an mRNA interferase-antitoxin locus; dynamic region</td>
<td>15.57</td>
<td>?</td>
<td>Barrangou et al. 2007, Caimi and Cataldi 2004</td>
</tr>
<tr>
<td>Direct repeat (DR) locus</td>
<td>Rv2100–Rv2108</td>
<td>Helicase and hypothetical proteins</td>
<td>19.97</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Lipooligosaccharide (LOS) locus</td>
<td>Rv2108–Rv2109</td>
<td>DR locus and hypothetical proteins; may play a role in resistance to phage infection</td>
<td>24.10</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Fumarate reductase operon</td>
<td>Rv2100–Rv2109</td>
<td>Hypothetical membrane proteins including the predicted lipoprotein LppY</td>
<td>NP</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>VirS locus</td>
<td>Rv2998–Rv3000 (ppp)</td>
<td>VirS regulates virulence; probably acquired by LGT as region is conserved between MM and MAP (refer orthologs MMAR_1573 and MAP3154)</td>
<td>NP</td>
<td>Yes</td>
<td>Singh et al. 2005</td>
</tr>
<tr>
<td>Locus of unknown function</td>
<td>Rv3081–Rv3089 (includes virS, lipR, adhD, and fadD12)</td>
<td>Single gene acquisition</td>
<td>NP</td>
<td>Yes</td>
<td>Altenr et al. 2007</td>
</tr>
<tr>
<td>Pilin gene</td>
<td>Rv3108–Rv3109 (ppp)</td>
<td>Required for synthesis of molybdopterin</td>
<td>16.56</td>
<td>Yes</td>
<td>Betts et al. 2002</td>
</tr>
<tr>
<td>Molybdopterin locus II</td>
<td>Rv3112–Rv3113 (includes moaA1-E1, moeB2, and cysA3)</td>
<td>Required for synthesis of molybdopterin</td>
<td>11.60</td>
<td>Yes</td>
<td>Altenr et al. 2007</td>
</tr>
</tbody>
</table>

*AH, Alien_Hunter prediction. The score reflects the compositional deviation of those regions from the genome background compositional distribution. The higher the score the more atypical the corresponding sequence. A typical score threshold suggestive of LGT is 6.157 (Vernikos and Parkhill 2006).
bLGT = “Yes” when MTBvMMvMAP comparisons were suggestive; LGT = “?” when three-way genome comparisons were inconclusive.
NP, No prediction.
tetranucleotide signatures of *M. tuberculosis* (Becq et al. 2007). Whether acquired by LGT or other means, some of these *M. tuberculosis*-specific regions contain known virulence genes or code for adaptation factors (Table 3). In the same vein, *M. marinum* has also gained genes by LGT with notable examples being some of the NRPS genes, such as MMAR\_0851 and its paralogs, that may have been acquired from aquatic *Streptomyces* spp.

These genome comparison data indicate that *M. marinum* and *M. tuberculosis* have diverged from a common generalist environmental *Mycobacterium* with *M. marinum* retaining genes required for its dual lifestyle and *M. tuberculosis* undergoing reductive evolution compatible with adaptation to its niche. It seems that, throughout the evolution of mycobacterial species, the process of genome reduction is a recurring theme in the adaptation of an environmental bacterium toward an intracellular, parasitic lifestyle. However, it is intriguing that *M. marinum* as the closest relative of *M. tuberculosis* has a genome that is 50% bigger than that of *M. tuberculosis*, while the members of the *M. avium* complex (4.8 Mb) are characterized by genome sizes that are closer to *M. tuberculosis*. This means that either significant genome downsizing has occurred independently in various phylogenetic lineages of slow-growing mycobacteria or that the contribution of LGT to the *M. marinum* gene pool must be considerable, as suggested by NRPS gene acquisition (Fig. 1B). However, the relatively large genome size of *M. ulcerans* (∼5.2 Mb after compensating for amplification of IS elements and acquisition of pMUM001, the mycolactone plasmid), a recent descendant of *M. marinum* that has adapted to particular environmental niche and entered reductive evolution, is in favor of the hypothesis where members of the *M. tuberculosis* and *M. avium* complexes downsized their genomes independently. A more comprehensive phylogenomic reconstruction of the genus *Mycobacterium* will be possible with the publication of the *M. tuberculosis* genome and the genomes of several other recently completed environmental mycobacteria such as *Mycobacterium vanbualenii*. Further insight into this process will also be provided by analyses of the genome sequences of *Mycobacterium prototuberculosis* and *Mycobacterium canetti*, which predate *M. tuberculosis* in the current evolutionary scheme (Brosch et al. 2002; Gutierrez et al. 2005).

**Methods**

*M. marinum* M strain

*M. marinum* M (ATCC BAA-535) is a human patient isolate from Moffett Hospital, University of California, San Francisco in 1992. Multi-locus sequence typing shows that this strain belongs to sequence type 22 (Yip et al. 2007).

**Genome sequencing**

Genomic DNA was sheared by sonication, size fractionated, and used to generate several shotgun libraries. Clones were sequenced from these libraries as follows: 21,773 reads from an m13mp18 library with an insert size of 1.4–2 kb; 66,388 paired reads from a pUC18 library with an insert size of 3–3.3 kb; and 55,654 paired reads from a pMAQ1b library with an insert size of 5.5–6 kb. This shotgun generated approximately 9.3-fold coverage of the genome. To scaffold these sequences, 6120 paired reads were generated from several pBACE3.6 libraries with insert sizes ranging from 15 to 48 kb, generating around 11-fold genome coverage in BAC clones. A further 27,357 directed sequences were generated during the gap-closure and finishing phase. DNA was sequenced using ABI BigDye terminator reactions and run on AB3730 capillary sequencers. Sequence assembly was performed by phrap and finished to standard criteria using GAP4 as previously described (Cole et al. 1998). PCR was used to check the distribution of the pMM23 plasmid among 41 different isolates of *M. marinum*, spanning the nine reported MLST genotypes by amplification of the plasmid genes *merB* (merB-F 5′-CTCGACGCCCTGTAGTGT-3′ and merB-R 5′-AACGCGGTCGCAATAGAG-3′; 432-bp fragment) and *paraA* (paraA-F 5′-CGTTTCGGCAGATAACGAT3′ and paraA-R 5′-GTGGTCAGCCTCCCCAAAT-3′; 421-bp fragment).

**Genome annotation**

Annotation and database construction were performed with Wasabi, an in-house, Web-interfaced MySQL database for management of genome annotation, Artemis, and GenoList. Complete databases are available (http://www.sanger.ac.uk/Projects/M_marinum/ and http://genolist.pasteur.fr/Marinolist/). CDS were predicted using a combination of Glimmer (Delcher et al. 1999), EasyGene (Anders Krogh, pers. comm.), Orpheus (Frishman et al. 1998), and GeneMark (Besemer et al. 2001); followed by manual curation as previously described (Parkhill et al. 2003); assigned the locus_tag prefix MMAR\_XXXX; and compared to public sequence databases using the BLAST suite (Altschul et al. 1990). IS were identified with reference to IS-Finder database (http://www.is.biotoul.fr/); cytochrome P450s were identified using the cytochrome P450 database (http://www.is.biotoul.fr/). CDS were identified with reference to IS-Finder database (http://www.is.biotoul.fr/); cytochrome P450s were identified using the cytochrome P450 database (http://www.is.biotoul.fr/). CDS were identified with reference to IS-Finder database (http://www.is.biotoul.fr/); cytochrome P450s were identified using the cytochrome P450 database (http://www.is.biotoul.fr/). CDS were identified with reference to IS-Finder database (http://www.is.biotoul.fr/); cytochrome P450s were identified using the cytochrome P450 database (http://www.is.biotoul.fr/).

**Whole-genome comparisons and lateral gene transfer**

Whole-genome DNA:DNA comparisons of *M. marinum* M were performed using BLASTN analysis (Expect threshold = 10) against *M. tuberculosis* H37Rv, *M. ulcerans* Agy99, *M. avium* subsp. *paratuberculosis* K10, *M. leprae* TN, and *M. smegmatis* mc²155. The results of these comparisons were visualized using the Artemis Comparison Tool (ACT) (Carver et al. 2005). In addition, to detect CDS that may have arisen from lateral gene transfer events, two new software tools were used. Gene Spaghetti is GNU public license software for visualizing base composition and codon usage in a genome. The software uses a sliding Gaussian window to estimate local base usage. Here, a window with a standard deviation of 200 was used. Further details and access to the software are available at http://www.vichioinformatics.com/genespaghetti.shtml. AlienHunter uses a combination of all nucleotide motifs of between 1 bp and 8 bp to identify regions of the genome that differ in nucleotide composition from the backbone, and then a change-point Hidden Markov Model to refine the boundaries of the anomalous DNA (Vernikos and Parkhill 2006).

**Ortholog analysis**

Orthologs were defined in two ways. Reciprocal FASTA analysis was performed on all CDS from *M. marinum*, *M. tuberculosis*, *M. ulcerans*, *M. avium* subsp. *paratuberculosis*, *M. smegmatis*, and *M. leprae*, with reciprocal top hits requiring >80% amino acid identity over 30% or more of the length of both sequences. This FASTA analysis was also used to define a core set of mycobacterial orthologs among the six species. A pairwise comparison was then undertaken of the average percentage amino acid identity from the core set of orthologs. A phylogeny of the six species was inferred from the resulting distance matrix (each distance being calculated as 100% minus the average percent identity) using the Neighbor-Joining method of tree construction (Saitou and Nei 1987) with the “Neighbor” program in the Phylip package. The robustness of the resulting tree topology was tested by bootstrap
analysis. Ten thousand bootstrap samples were constructed by selecting orthologous groups at random with replacement from the original 1072 ortholog groups identified by the FASTA analysis above. The resulting trees were compared using Phylip’s “Consense” program.

In a second complementary approach, CDS from M. marinum, M. tuberculosis, M. ulcerans, and M. avium subsp. paratuberculosis were clustered into orthologous groups using the BLASTCLUST tool from the NCBI BLAST software suite version 2.2.14, with default settings in protein mode. This enforced a maximum of 10% length variation within a cluster and a minimum bit score to length ratio of 1.75. There was high concordance obtained by both clustering techniques. All BLASTCLUST clusters were checked manually and verified using BLASTN and TBLASTX (Altschul et al. 1990).

Acknowledgments

We thank the Sanger Institute core sequencing and informatics groups for support, L. Jones and F. Tekaija for database support, Jessica L. Porter for technical assistance, Derrick Fouts for use of the unpublished M. smegmatis genome, and A. Colorni for critical review of the manuscript. We gratefully acknowledge the financial support of the Wellcome Trust, Foundation Raoul Follereau (STC), the National Health and Medical Research Council of Australia (TPS), and grant R01 AI036396 from the National Institutes of Health (LR).

References


Direct linkage of mitochondrial genome variation to risk factors for type 2 diabetes in conplastic strains

Michal Pravenec, Masaya Hyakukoku, Josef Houstek, Vaclav Zidek, Vladimir Landa, Petr Mlejnek, Ivan Miksik, Kristyna Dudová-Mothejzikova, Petr Pecina, Marek Vrbacký, Zdenek Drahota, Alena Vojtiskova, Tomas Mracek, Ludmila Kazdova, Olena Oliyarnyk, Jiaming Wang, Christopher Ho, Nathan Qi, Ken Sugimoto, and Theodore Kurtz

The authors have discovered an error in the presentation of the mitochondrial DNA haplotyping analysis described in Table 2 and wish to correct the data. The row that displays genotypes at nucleotide position 11,374 is incorrect and should be replaced by a new row that displays genotypes at nucleotide position 11,360. The corrected Table 2 is reprinted below. The authors apologize for any inconvenience this error may have caused other investigators in the field.

Table 2. Mitochondrial DNA haplotype grouping of 43 strains of rats

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1716</td>
<td>BN</td>
<td>SHR</td>
<td>SHR</td>
<td>BN</td>
</tr>
<tr>
<td>1918</td>
<td>BN</td>
<td>SHR</td>
<td>SHR</td>
<td>BN</td>
</tr>
<tr>
<td>4352b</td>
<td>BN</td>
<td>SHR</td>
<td>SHR</td>
<td>BN</td>
</tr>
<tr>
<td>5200</td>
<td>BN</td>
<td>SHR</td>
<td>SHR</td>
<td>BN</td>
</tr>
<tr>
<td>5237</td>
<td>BN</td>
<td>SHR</td>
<td>SHR</td>
<td>BN</td>
</tr>
<tr>
<td>5269</td>
<td>BN</td>
<td>SHR</td>
<td>SHR</td>
<td>BN</td>
</tr>
<tr>
<td>5326b</td>
<td>BN</td>
<td>SHR</td>
<td>SHR</td>
<td>BN</td>
</tr>
<tr>
<td>7930b</td>
<td>BN</td>
<td>SHR</td>
<td>SHR</td>
<td>BN</td>
</tr>
<tr>
<td>8021b</td>
<td>BN</td>
<td>SHR</td>
<td>SHR</td>
<td>BN</td>
</tr>
<tr>
<td>11,194</td>
<td>BN</td>
<td>SHR</td>
<td>SHR</td>
<td>BN</td>
</tr>
<tr>
<td>11,360b</td>
<td>BN</td>
<td>BN</td>
<td>SHR</td>
<td>BN</td>
</tr>
<tr>
<td>13,647b</td>
<td>BN</td>
<td>SHR</td>
<td>SHR</td>
<td>BN</td>
</tr>
<tr>
<td>13,693</td>
<td>BN</td>
<td>SHR</td>
<td>SHR</td>
<td>BN</td>
</tr>
<tr>
<td>13,974b</td>
<td>SHR</td>
<td>SHR</td>
<td>SHR</td>
<td>BN</td>
</tr>
<tr>
<td>15,135b</td>
<td>SHR</td>
<td>SHR</td>
<td>SHR</td>
<td>BN</td>
</tr>
<tr>
<td>15,400</td>
<td>SHR</td>
<td>SHR</td>
<td>SHR</td>
<td>BN</td>
</tr>
</tbody>
</table>

We identified four haplotype groups by genotype analysis of 16 mitochondrial DNA polymorphisms in 43 rat strains including the SHR and BN progenitor strains. The specific variants present at each nucleotide site in the SHR and BN strains are listed in Tables 1 and Supplemental Table 1.

*Haplotype Group 1 consisted of 20 strains: ACI/Ztm, BC/CpbU, BDE/Han, BDIV/Iiz, BDIX/Han, BN/Ztm, BUF/Han, COP/OlaHsd, DA/Han, F344/Han, NAR/SaU, NEDH/K, PVG/OlaHsdCp, R/A, RHA/Kun, SDL/Ipvc, SPRD-Cu3/Han, SR/Irjpvc, SS/Irjpvc, WF/Han. Group 2 consisted of 17 strains: AS/Ztm, AVN/Orl, BBWB/Mol, BDII/Han, BS/Ztm, LE/Han, LEW/OlaHsd, LOU/C, MNS/Gro, MW/Hsd, OM/Han, SD/Rij, SDH/Ztu, WF/Gut, WIST/Nhg, WOK.1A/K, WOK/w/K. Group 3 consisted of four strains: SHR/Ola, OKA/Wdi, GC/Kun, WKY/M. Group 4 consisted of two strains: BN/Ola, DZB/Gro. The DZB/Gro strain was derived by a cross of BN rats with AO rats.

bNucleotides causing amino acid substitutions.
FadB was inadvertently omitted from the sentence on page 731 describing the enzymes involved in beta-oxidation in *M. tuberculosis*. The sentence should have read:

“FadD (acyl-CoA synthetase) catalyses the formation of acyl-CoA thioesters that are then processively oxidized by FadE, EchA, FadB, and FadA (Cole et al. 1998).”

In addition, the page range for the Hagedorn and Soldati 2007 reference was incorrect, the corrected reference is as follows:


The authors apologize for any confusion these errors may have caused.
Insights from the complete genome sequence of *Mycobacterium marinum* on the evolution of *Mycobacterium tuberculosis*

Timothy P. Stinear, Torsten Seemann, Paul F. Harrison, et al.

*Genome Res.* 2008 18: 729-741 originally published online April 10, 2008

Access the most recent version at doi:10.1101/gr.075069.107

---

**Supplemental Material**

http://genome.cshlp.org/content/suppl/2008/04/15/gr.075069.107.DC1

**Related Content**

Errata

*Genome Res.* October, 2008 18: 1680-1681

**References**

This article cites 105 articles, 38 of which can be accessed free at:

http://genome.cshlp.org/content/18/5/729.full.html#ref-list-1

Articles cited in:

http://genome.cshlp.org/content/18/5/729.full.html#related-urls

**Open Access**

Freely available online through the *Genome Research* Open Access option.

**License**

Freely available online through the Genome Research Open Access option.

**Email Alerting Service**

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](http://genome.cshlp.org/subscriptions).

---

To subscribe to *Genome Research* go to:

http://genome.cshlp.org/subscriptions

---

Copyright © 2008, Cold Spring Harbor Laboratory Press