Defining mycobacteria: Shared and specific genome features for different lifestyles

Varalakshmi D. Vissa · Rama Murthy Sakamuri · Wei Li · Patrick J. Brennan

Received: 30 April 2007 / Accepted: 16 August 2008

Abstract During the last decade, the combination of rapid whole genome sequencing capabilities, application of genetic and computational tools, and establishment of model systems for the study of a range of species for a spectrum of biological questions has enhanced our cumulative knowledge of mycobacteria in terms of their growth properties and requirements. The adaption of the corynebacterial surrogate system has simplified the study of cell wall biosynthetic machinery common to actinobacteria. Comparative genomics supported by experimentation reveals that superimposed on a common core of ‘mycobacterial’ gene set, pathogenic mycobacteria are endowed with multiple copies of several protein families that encode novel secretion and transport systems such as *mce* and *esx*; immunomodulators named PE/PPE proteins, and polyketide synthases for synthesis of complex lipids. The precise timing of expression, engagement and interactions involving one or more of these redundant proteins in their host environments likely play a role in the definition and differentiation of species and their disease phenotypes. Besides these, only a few species specific ‘virulence’ factors i.e., macromolecules have been discovered. Other subtleties may also arise from modifications of shared macromolecules. In contrast, to cope with the broad and changing growth conditions,

their saprophytic relatives have larger genomes, in which the excess coding capacity is dedicated to transcriptional regulators, transporters for nutrients and toxic metabolites, biosynthesis of secondary metabolites and catabolic pathways. In this review, we present a sampling of the tools and techniques that are being implemented to tease apart aspects of physiology, phylogeny, ecology and pathology and illustrate the dominant genomic characteristics of representative species. The investigation of clinical isolates, natural disease states and discovery of new diagnostics, vaccines and drugs for existing and emerging mycobacterial diseases, particularly for multidrug resistant strains are the challenges in the coming decades.

Keywords Genomics · Evolution · Mycobacteria · Virulence · COGs

Introduction and significance

Whole genome sequencing of organisms has become a reality to the point that there is a growing interdependence between *in silico* predictions based on genomic codes and classical experimental approaches in everyday biological research for questions that span the grand description of the ‘tree of life’ to the simpler mechanism of a single enzymatic reaction. Already, vast amounts of genome and bioinformatic codes exist that beg to be harnessed. These resources are simultaneously overwhelming and unwieldy, yet necessary and desirable.

In this review we attempt to summarize the principles, methods and challenges that combine *in silico* resources, with biological tools in the understanding of genotype–phenotype relationships in mycobacteriology. We address

V. D. Vissa (✉) · R. M. Sakamuri · W. Li · P. J. Brennan
Department of Microbiology, Immunology and Pathology
Colorado State University,
Fort Collins,
CO-80523-1628,
USA

E-mail: vvissa@colostate.edu

briefly, a range of themes including ecology, epidemiology, and physiology. Furthermore, we present independent *in silico* comparative analyses of multiple sequenced genomes that summarize and delineate major genetic signatures and trends within and between mycobacterial genomes.

The objective of this review and analysis is to cite landmark studies and novel approaches to give the reader an overall appreciation of the major advances in the study and the understanding of mycobacteria that have been accelerated by availability of genome sequences. Although gaps remain, it is hoped that practical outcomes emerge from this information in due course, such as diagnostic kits, vaccine and drugs for different diseases.

Taxonomy of mycobacteria

The genus *Mycobacterium* is derived from the phylum Actinobacteria, class Actinobacteria, which includes gram-positive bacteria of high genomic G+C content. Further classification (Table 1) based on 16sRNA sequences and morphological traits places the genus within subclass Actinobacteridae, order Actinomycetales, suborder Corynebacterineae, and family Mycobacteriaceae (<http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root>) [1, 2].

Other families within suborder Corynebacterineae in this classification include Corynebacteriaceae, Dietziaceae, Gordoniaceae, Nocardiaceae and Tsukumurellaceae that share certain morphological and biochemical features with those of Mycobacteriaceae. Literature also cites relationships of mycobacteria with distant genera of other suborders such as Streptomycineae (*Streptomyces*), Propionibacterineae (*Propionibacterium*), Pseudonocardinaceae (*Amycolatopsis*), Micrococcineae (*Cellulomonas* and *Micrococcus*) (Table 1)

Availability of microbial genome sequences within Actinomycetales

With regard to microbes, the National Center for Biotechnology Information (NCBI) reports 524 complete, 320 assembled, and 462 unfinished genome sequencing projects (as of June 2007). Genome sequence information of Actinomycetales is abundant, with 146 submissions in the current database. These include complete and unfinished genomes, and plasmid sequences. Moreover the suborder Corynebacterineae is extensively represented with 82 sequence projects, of which 26 are complete genomes (Table 1).

Mycobacterium, a genus that contains species of significant pathogenic import along with a number of non-

pathogenic relatives is well represented in this genome-era, starting with the publication of the genome sequence of *M. tuberculosis*, the agent of human tuberculosis (TB) in 1998 [3]. Subsequently, sequences of *M. bovis* that causes TB in cattle and wildlife [4] and its attenuated vaccine strains *M. bovis* Pasteur [5] and BCG became known [6]. The other two sequenced human pathogens are *M. leprae* [7], and *M. ulcerans* [8] that cause skin diseases, leprosy and Buruli ulcer, respectively. *M. avium* subsp. *paratuberculosis* [9], the bacterial agent of cattle Johne's disease, is also implicated in human Crohn's disease.

The source for the taxonomy is <http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Undefined&name=Bacteria&lvl=3&srchmode=1&keep=1&unlock>. The numbers after the names of the suborder and family in columns 1 and 2 respectively, refer to the number of completed sequencing projects. When there are two numbers, the first refers to the total number of sequencing projects to date, and the second refers to the completed genome sequences. The presence of cell wall molecules trehalose monomycolate (TMM) and/or trehalose di mycolate (TDM), arabinogalactan (AG), lipomannan (LM) and/or lipoarabinomannan (LAM) and mycolic acids, is indicated by the + symbol. Truncated or shorter LAMs and mycolic acids as compared to those of mycobacteria are indicated by * and ** respectively. References for occurrence or characterization of cell wall associated molecules are included in the last column.

Genomics of mycobacteria

In an insightful essay in 2000, Weinstock [23], predicting the pace of accumulation of microbial genome sequence data, proposed a 'top-down' genomics approach in modern microbiology. He noted that while potential and limitations exist in handling large raw datasets, experimental strategies that could turn these data into new knowledge about microbial processes, even in the absence of functional gene validations, are plausible. The gains in less than a decade, have already justified 'genomics' as an independent means to study bacteria. Notable successes and future possibilities as exemplified for mycobacteria are highlighted herein.

Genetic tools for mycobacteria

The development of temperature-sensitive plasmids [24] and mycobacteriophages [25] was key to launching a series of experiments allowing site specific and random mutagenesis of mycobacterial genomes including those of *M. smegmatis*, *M. tuberculosis*, *M. bovis* BCG, *M. avium*, and *M. marinum*, fostering the functional characterization of genes involved

in nutrition, cell wall synthesis, infection, survival and persistence in various model systems. The test conditions examined growth properties in defined medium in vitro (liquid or solid bacteriological media and macrophage systems), or in vivo, in animal host models that include mouse, zebrafish, leopard frog, guinea pig, rabbit and monkey.

The systematic and impressive effort by Lamichhane et al. [26], capitalizing on genome sequence availability of *M. tuberculosis* CDC1551 is noteworthy. By using the *HimarI* phage to transpose randomly into the *M. tuberculosis* genome, picking individual surviving insertion clones, and identifying the point of insertion of the transposon in each by sequencing the flanking genomic regions, they found that up to 65% of the predicted coding sequences (CDSs) could be interrupted, i.e., these CDSs were not ‘lethal’ (or ‘non-essential’) for growth on agar plates. The remaining 35% of CDSs not represented under the experimental growth conditions were therefore interpreted to be ‘essential’ for survival. Mutant clones from this experiment serve as a valuable source of defined gene knock outs for further phenotypic characterization [25].

Array technologies accelerate screening of mycobacteria mutants

The earliest screens of mutant libraries were based on the construction of transposon vectors carrying an ‘array of signature tags’ in the insertion element [27, 28]. Pools of signature tagged mutants would then be used to infect mice. The genes required for ‘growth in vivo’ were identified by comparing hybridization patterns and intensities of ‘input’ versus ‘recovered’ transposon DNA probes on a membrane format containing an array of signature tagged DNA sequences.

A frequently cited method in this regard is that by Sasseti et al. [29, 30, 31], in which gene arrays were created to rapidly identify transposon interrupted genes by the method called TraSH (transposon site hybridization). An ‘essential gene list’ was compiled based on growth of *M. tuberculosis* H37Rv mutants on 7H10 agar containing OADC enrichment [30]. The major finding from this study was that most of the 614 essential *M. tuberculosis* genes are intact in the heavily degraded genome of *M. leprae*, while ‘non-essential’ *M. tuberculosis* genes have been deleted or mutated in *M. leprae*. Another conclusion was that one third of these essential genes are not found in other bacteria and have no assigned function, indicating that the core mycobacterial physiology requires genes beyond those in ‘minimum genomes’ of mycoplasmas [28].

Mycobacterial phylogeny

There are many ways to classify and study mycobacteria for their genotype–phenotype associations. Growth rates separate them down into rapid or slow growers, habitat defines them as environmental (free living/saprophytic) or host adapted, while disease causing properties separate the tuberculous from the non-tuberculous mycobacteria (NTM) [32]. The terms atypical mycobacteria and mycobacteria other than *M. tuberculosis* (MOTT) are also in use for NTM that cause infections. The term *M. tuberculosis* complex mainly includes *M. tuberculosis*, *M. microti*, *M. africanum* and *M. bovis*.

Recognizing that there have been many diverse criteria and schemes for bacterial taxonomy since the late 1800s, and that the taxonomy based only on 16S rDNA sequences encoding rRNAs often conflicted with existing higher level taxonomy, taxonomists called for ‘polyphasic’ systematics, which required that phylogeny be determined by DNA sequences, and that more than one class of molecules be included [33]. Now that whole genome sequences are becoming available, it should be possible to perform genome wide comparisons to demonstrate phylogenetic relatedness of species (phylogenomics) and also yield insights into factors that govern niche adaptation and virulence. However, there are no simple, single methods that can accurately compile and represent all the genomic data due to the variability in the rates of evolution of different parts of the genome, recombination, gene loss and acquisition events. Genome trees are typically built from one of five common sources of phylogenetic markers: sequence attributes such as word frequencies (alignment free techniques), shared gene content, gene order, average sequence similarity, and gene trees [34].

Literature cites multilocus sequence typing (MLST) as a means for determining species and higher level phylogeny whereby a limited set of loci, such as those for housekeeping genes are compared by PCR and sequencing techniques.

- For the genus *Mycobacterium* comprising species with highly similar or identical 16S rRNA sequences, additional genes such as *rpoB*, *gyrB*, *recA*, *hsp65*, *sodA* and ITS, have been sequenced and compared to improve discrimination [35].
- Devulder et al. [36] assembled a set of nearly 100 cultivable strains of *Mycobacterium* spp., amplified and sequenced portions of the 16rRNA, *hsp65*, *rpoB* or *sodA* genes. Comparing five phylogenetic trees, four computed from a single gene (16rRNA, *hsp65*, *rpoB* or *sodA*) and the fifth, from concatenation of all four genes, the latter (MLST tree) was more robust than that computed from single genes. Furthermore,

Table 1 Classification of members of the order Actinomycetales, availability of genome sequences and known occurrences of characteristic cell wall features

Suborder	Family	Genus and species	Number of genome(s) sequenced	Number of plasmids sequenced	Characteristic cell wall components			
					TMM/TDM	AG	LM/LAM	Mycolic acids
Actinomycineae	Actinomycetaceae							
	Catenulisporineae							
Corynebacterineae [82(26)]	Corynebacteriaceae	<i>Corynebacterium callunae</i>	-	1	+	+	+	++
	(Coryneform bacteria) [35(6)]							
		<i>Corynebacterium diphtheriae</i>	1	2	+	+	+	++
		<i>Corynebacterium efficiens</i>	1	2	+	+	+	++
		<i>Corynebacterium glutamicum</i>	3	13	+	+	+	++
		<i>Corynebacterium jeikeium</i>	1	7	+	+	+	++
		<i>Corynebacterium renale</i>	-	1	+	+	+	++
		<i>Corynebacterium striatum</i>	-	1	+	+	+	++
		<i>Corynebacterium</i> sp. L2-79-05	-	2	+	+	+	++
	Dietziaceae	<i>Dietzia</i>	1	-		+	+	++
	Gordoniaceae	<i>Gordonia westfalica</i>	1	-		+	+	++
	Mycobacteriaceae	<i>Mycobacterium avium</i> complex (MAC)	2	1	+	+	+	+
		<i>Mycobacterium celatum</i>	-	-	+	+	+	+
		<i>Mycobacterium goodii</i>	1	3	+	+	+	+
		<i>Mycobacterium leprae</i>	1	-	+	+	+	+
		<i>Mycobacterium smegmatis</i>	1	-	+	+	+	+
		<i>Mycobacterium tuberculosis</i> complex	8	-	+	+	+	+
		<i>Mycobacterium ulcerans</i>	1	1	+	+	+	+
		<i>Mycobacterium vanbaalenii</i>	1	-	+	+	+	+
		<i>Mycobacterium</i> sp. JLS	1	-	+	+	+	+
		<i>Mycobacterium</i> sp. KMS	1	-	+	+	+	+
		<i>Mycobacterium</i> sp. MCS	1	-	+	+	+	+
	Nocardiaceae [19(2)]	<i>Nocardia farcinica</i>	1	2	+	+	+	++
		<i>Rhodococcus equi</i>		2	+	+	+	++
		<i>Rhodococcus erythropolis</i>		6	+	+	+	++
		<i>Rhodococcus opacus</i>		2	+	+	+	++
		<i>Rhodococcus rhodochrous</i>		1	+	+	+	++
		<i>Rhodococcus</i> sp. B264-1		1	+	+	+	++
		<i>Rhodococcus</i> sp. RHA1	1	3	+	+	+	++

Table 1 (Continued)

Segniliparaceae						
Segniliparaceae						
Tsukamurellaceae						15, 21
Williamsiaceae						
Acidothermaceae [1]						
Frankiaceae [4]						
Geodermatophilaceae [1]						
Kineosporiaceae						
Nakamurellaceae						
Sporichthyaceae						
Glycomycetaceae						
Beutenbergiaceae						
Bogoriellaceae						
Brevibacteriaceae						
Cellulomonadaceae						
Dermabacteraceae						
Dermacoccaceae						
Dermatophilaceae						
Intrasporangiaceae						
Jonesiaceae						
Microbacteriaceae						
Micrococcaceae						16, 22
Promicromonosporaceae						
Rarobacteraceae						
Sanguibacteraceae						
Yaniaceae						
Micromonosporaceae						
NoCARDIIDACEAE						
Propionibacteriaceae						
Actinosynnemataceae						17
Pseudonocardaceae						18
Streptomycetaceae						
Streptosporangineae [1]						

The source for the taxonomy is <http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Undefined&id=2037&lvl=3&p=mapview&lin=f&keep=1&srchmode=1&unlock>. The numbers after the names of the suborder and family in columns 1 and 2, respectively refer to the number of completed sequencing projects. When there are two numbers, the first refers to the total number of sequencing projects to date and the second refers to the completed genome sequences. The presence of cell wall molecules trehalose monomycolate (TMM) and/or trehalose dimycolate (TDM), arabinogalactan (AG), lipomannan (LM) and/or lipoarabinomannan (LAM) and mycolic acids, is indicated by the + symbol. Truncated or shorter LAMs and mycolic acids as compared to those of mycobacteria are indicated by * and ** respectively. References for occurrence or characterization of cell wall associated molecules are included in the last column.

individual species except those of the *M. tuberculosis* complexes, were resolved and showed that the slowly growing species descended and separated from the rapid growers.

- Recent examples of evolutionary studies in mycobacteria include evidence for the descent of *M. bovis* from *M. tuberculosis* rather than in the reverse direction [37]; the combined downsizing of the *M. marinum* genome and the acquisition of a plasmid bearing virulence gene clusters to generate a new species, *M. ulcerans*, a human host-adapted *Mycobacterium* sp. associated in the environment with the aquatic insect *Naucoris cimicoides* [38] and also snails [39] and plants [40.] Similarly, strains within and outside the *M. avium* complex (MAC) have been studied to address evolution, strain differentiation, differential growth niches and replication rates [7, 32, 41]. Details for some such findings are presented in later sections of this review.

Common and species-specific properties of *Mycobacterium* spp.

In previous reviews, we and others have proposed and demonstrated that *M. leprae*, a paradigm genome with highest levels of known reductive evolution, serves as a ‘model minimal *Mycobacterium*’ [5, 42] because it is characterized by physical features shared by members of Mycobacteriaceae and few outside this family, the most prominent being the mycolylarabinogalactan peptidoglycan (mAGP) cell

wall complex and the presence of glycolipids/lipoglycans–PIMs (phosphatidylinositol mannosides), LM and LAM built on the phosphatidyl inositol (PI) anchor. Furthermore, contained in its smaller genome is information for pathogenicity. In accordance with this principle, it is not surprising that the ‘essential gene list’ for growth in an animal model that was experimentally discovered using the drop out saturation mutagenesis techniques, coincides with the intact genome fraction of the genome of *M. leprae*; non-essential genes of *M. tuberculosis* being those that are present in more than one functional equivalent, or else those lost by reductive evolution in *M. leprae* [43]. Nonetheless, it is obvious that beyond the small number of ‘essential genes’, the biological properties of individual species [growth habitats, replication rates, pathogenicity (tissue tropism and pathology) etc.] are not identical to that of *M. leprae*. The genetic origins of these species differences are still not clear and it is hoped that clues can be found in the genomes.

Therefore, we try to address the shared features of *Mycobacterium* as a genus and the species-specific differences. To this end, we allude to genes and gene families discussed in the literature and also present simple comparative *in silico* analyses of 10 selected genomes. We used on-line resources in the public domain, and intuitive approaches to allow us to simultaneously compare multiple sequenced and annotated genomes to recognize the shared and distinct genomic content. In Table 2 is a list of some of the web-based resources accessed during the preparation of this review:

First we selected a panel of genomes: *M. tuberculosis* H37Rv, *M. bovis* AF2122/9, *M. avium* subsp. *paratuberculosis* K10, *M. leprae* TN, *M. ulcerans* Agy99, *M. avium* 104,

Table 2 Web-based genome sequence and analysis resources used in comparing the genomes of selected *Mycobacterium* spp.

Genome	Database name	URL
<i>M. leprae</i> TN	Leproma	http://genolist.pasteur.fr/Leproma/
<i>M. tuberculosis</i> H37Rv	Tuberculist	http://genolist.pasteur.fr/TubercuList/
<i>M. bovis</i> AF2122/97	Bovilist	http://genolist.pasteur.fr/BoviList/
<i>M. bovis</i> BCG	BCGlist	http://genolist.pasteur.fr/BCGList/
<i>M. ulcerans</i> Agy99	Burulist	http://genolist.pasteur.fr/BuruList/
<i>M. marinum</i>	Marinolist	http://genolist.pasteur.fr/MarinoList/
<i>M. smegmatis</i>		http://cmr.jcvi.org/cgi-bin/CMR/GenomePage.cgi?org=gms
All other microbial genomes		http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi
Gene, protein, enzyme, metabolic pathways and whole genome searches/comparisons		http://www.jcvi.org/ http://ca.expasy.org/ http://www.ncbi.nlm.nih.gov/ http://www.genome.jp/kegg/ http://biocyc.org/ http://pages.usherbrooke.ca/gaudreau/MtbRegList/www/index.php# http://img.jgi.doe.gov

M. smegmatis MC [2]155, *Mycobacterium JLS*, *Corynebacterium glutamicum* ATCC 13032 (Bielefeld) and *Escherichia coli*. The first five species in this panel are mycobacterial host associated pathogens, while *M. avium* subsp. *avium*, an environmental *Mycobacterium* is an opportunistic pathogen in humans. *M. smegmatis* and *Mycobacterium JLS* serve as representatives of mycobacterial non-pathogen saprophytes. *C. glutamicum* is a non-pathogenic representative of the suborder shared by all of the above listed mycobacteria. *E. coli* was selected as a gram-negative, non-pathogenic distant species for genome comparisons.

We then compared the relative abundance and species distribution of genes that encode proteins belonging to conserved functional categories known as Clusters of Orthologous Group (COG) as conceived by Tatusov et al. [44].

- In an article entitled "A genomic perspective on protein families", Tatusov et al. [44] define orthologs as genes in different species that evolved from a common ancestral gene by speciation; by contrast, paralogs are genes related by duplication within a genome. In this scheme, a COG consists of individual orthologous genes or orthologous groups of paralogs from three or more phylogenetic lineages, whereby each COG can be assumed to have evolved from an individual ancestral gene through a series of speciation and duplication events. The database of COGs attempts to represent a phylogenetic classification of the proteins encoded in sequenced genomes.

We used the Integrated Microbial Genomes (IMG) data management system developed by the U.S. Department of Energy Joint Genome Institute (DOE JGI), version 2.2 at <http://img.jgi.doe.gov>. A summary of the genomes and their COGs is shown in Table 3.

We searched for the COGs that are more abundant (≥ 1) in one query species when compared to those encoded in a select panel of nine other species. In addition, we listed the genes within each of the COGs (except for the family encoding the PE-PPE genes and mobile elements/insertion se-

quence elements) and identified the genes that were unique to the query species in relation to the other nine species. These are presented in subsequent sections (Tables 4–6). A few of the COGs with skewed distribution amongst the species that we examined are shown in Fig. 3.

We are aware that approximately one third of sequenced genomes contain genes that are not assigned to any COGs. Furthermore, certain COGs belong to categories whose functions are not known. Therefore COG abundance profiling can miss a significant proportion of genes that are important in biology. Nevertheless, the results generally substantiate experimental findings.

Novel gene families in mycobacteria

PE and PPE family of proteins

Since their discovery, studies have sought to find and ascribe biological functions for these proteins that are thought to be enriched in mycobacteria [3, 45]. Owing to their overall similarity, yet sequence variability, members of these families are believed to be involved in the intracellular survival and antigenic variation [3]. The extracellular localization of several of these proteins does indicate the potential for antigen presentation. However, other biological activities have been discovered, such as a role in intracellular macrophage survival in *M. marinum* [46]. Only a few of the genes are 'essential' by the Lamichhane and Sassetti criteria [26, 30].

- The Conserved Domain database (<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>) [47, 48] describes and depicts these proteins and their domains as follows:

PE family: This family is named after a PE (Pro-Glu) motif found at the amino terminus of the domain (pfam00934) (Fig. 1). The PE family of proteins contain a conserved amino-terminal region of about 110 amino acids. The carboxyl termini of this family are variable and fall into several

Fig. 1 Pictorial representation of PE (pfam00934), PPE (pfam00823) and PE-PPE (pfam08237) conserved domains in proteins of the PE/PPE family in *M. tuberculosis*

Table 3 Comparison of genome properties and COG content of a selection of mycobacterial and non-mycobacterial species

	<i>M. tuberculosis</i>	<i>M. smegmatis</i>	<i>M. avium</i> subsp. <i>paratub</i>	<i>M. avium</i> subsp. <i>avium</i>	<i>M. leprae</i>	<i>C. glutamicum</i>	<i>E. coli</i>
DNA (total number of bases)	4,411,532	6,988,209	4,829,781	5,475,491	3,268,203	3,282,708	4,639,675
DNA coding (number of bases)	4,031,454	6,537,125	4,431,880	4,984,222	1,631,335	2,908,615	4,085,940
DNA G+C (number of bases)	2,894,588	4,710,249	3,346,971	3,777,437	1,888,915	1,767,468	2,356,477
Genes (total number)	4,060	6,925	4,413	5,290	2,749	3,148	4,590
Proteins (number of coding genes)	3,997	6,871	4,350	5,240	2,691	3,058	4,391
COG functions	Gene count	Gene count	Gene count	Gene count	Gene count	Gene count	Gene count
Amino acid transport and metabolism	204	485	228	224	129	208	367
Carbohydrate transport and metabolism	138	416	177	179	72	183	377
Cell cycle control, cell division, chromosome partitioning	44	28	29	27	21	18	36
Cell motility	68*	6	42	39	6	2	116
Cell wall/membrane/envelope biogenesis	128	170	131	123	69	108	239
Chromatin structure and dynamics	0.00	1 [^]	1	1	0.00	0.00	0.00
Coenzyme transport and metabolism	169	205	180	169	87	134	155
Defense mechanisms	38	49	35	34	12	40	49
Energy production and conversion	218	503	299	346	76	140	291
Function unknown	253	348	252	249	81	197	328
	8.00	6.25	6.85	6.10	6.16	8.40	8.17

Table 3 (Continued)

General function prediction only	440	13.91	820	14.73	515	14.00	606	14.85	418	13.07	127	9.65	282	12.03	401	9.99
Inorganic ion transport and metabolism	130	4.11	282	5.07	188	5.11	189	4.63	130	4.07	49	3.72	177	7.55	223	5.56
Intracellular trafficking, secretion and vesicular transport	26	0.82	24	0.43	25	0.68	24	0.59	24	0.75	20	1.52	27	1.15	135	3.36
Lipid transport and metabolism	258	8.15	499	8.97	396	10.76	513	12.57	311	9.72	91	6.91	71	3.03	102	2.54
Nucleotide transport and metabolism	75	2.37	103	1.85	71	1.93	72	1.76	70	2.19	55	4.18	77	3.28	97	2.42
Posttranslational modification, protein turnover, chaperones	101	3.19	136	2.44	96	2.61	103	2.52	104	3.25	62	4.71	84	3.58	138	3.44
RNA processing and modification	2	0.06	3	0.05	4	0.11	4	0.10	4	0.13	1	0.08	1	0.04	2	0.05
Replication, recombination and repair	185	5.85	193	3.47	138	3.75	216	5.29	185	5.78	65	4.94	134	5.71	215	5.36
Secondary metabolites biosynthesis, transport and catabolism	210	6.64	403	7.24	343	9.32	414	10.14	239	7.47	57	4.33	46	1.96	66	1.64
Signal transduction mechanisms	120	3.79	190	3.41	121	3.29	120	2.94	105	3.28	39	2.96	79	3.37	184	4.59
Transcription	204	6.45	525	9.43	263	7.15	289	7.08	235	7.35	74	5.62	190	8.10	308	7.68
Translation, ribosomal structure and biogenesis	153	4.84	177	3.18	145	3.94	141	3.45	154	4.82	123	9.35	147	6.27	184	4.59
	Low		Medium													

* predominantly PPE genes

^ histone deacetylase superfamily protein

classes. The largest class of PE proteins is the highly repetitive PGRS class which has a high glycine content. This PGRS domain is found to have sequences of glycine and alanine residues such as GGAGGX (where X is any amino acid), which can be repeated more than 30 times.

PPE family: This family is named after a PPE (Pro-Pro-Glu) motif near the amino terminus of the domain (pfam00823) (Fig. 1). The PPE proteins contain a conserved amino-terminal region of about 180 amino acids. The carboxyl terminus of this family is variable, and on the basis of this region are further subdivided based on their C terminal domain. PPE-SVP subgroup has a Gly-X-X-Ser-Val-Pro-X-X-Trp motif. The major polymorphic tandem repeat (MPTR) subgroup has multiple C terminal repeats of Asn-X-Gly-X-Gly-Asn-X- Gly. The third subfamily PPE-PPW consists of highly conserved Gly-Phe-X-Gly-Thr and Pro-X-X-Pro-X-X-Trp, C terminal motifs. The fourth subfamily members do not have homology at their C termini.

PE-PPE domain (pfam08237): This domain refers to the variable domain found C terminal to PE and PPE motifs. The secondary structure of this domain is predicted to be a mixture of alpha helices and beta strands.

- A large-scale gene expression study indicates that these multiple PE and PPE genes exhibit a dynamic, differential and independent mode of expression rather than by a global co-regulation mechanism. This has been borne out by comparing the expression of 128 genes in 15 major growth conditions that include a range of stress conditions such as low pH, hypoxia, high temperature, denaturants, starvation, stationary phase, peroxide, drug treatment, etc., using a microarray hybridization format [49].
- Specific expression of certain PE family genes suggests their possible role in pathogenesis or in virulence [50]. In a recent study, a similar profile of expression was observed in different host tissue by using a RT PCR approach with three PE-PGRS genes [51].
- DNA vaccine studies in mice showed that the PE domain PE-PGRS33 gene (*Rv1818c*) could induce a cellular immune response, whereas the whole

PE-PGRS domain elicited a humoral immune response, diminishing the protective immune response of the host to the PE domain in the context of the PGRS region. Likewise, in a DNA vaccine based on the *M. tuberculosis* PE-PGRS33 gene (*Rv1818c*), the PGRS domain with 21 GGAGGX repeats, inhibited the host immune response to the adjacent PE domain [52].

Mce

One or more ‘mammalian cell entry’ *mce* genes have been discovered in mycobacterial genomes. A typical *mce* gene is made of two domains, *mce* and Ttg, the former enabling cell entry, and the latter serving as a transporter as depicted and annotated below from the Conserved Domain database (Fig. 2).

‘The archetype *mce* domain in Rv0169 was isolated as being necessary for colonization of, and survival within, the macrophage. This *mce* protein family contains proteins of unknown function from other bacteria’.

‘The Ttg2C (COG1463) domain is defined as ABC-type transport system involved in resistance to organic solvents, periplasmic component [Secondary metabolites biosynthesis, transport, and catabolism]’.

Several *mce* genes are found in tandem within a *mce* operon. Moreover, multiple partial or intact *mce* operons are variably distributed across pathogenic and non-pathogenic mycobacteria. The contributions of individual *mce* genes and the operons in the natural physiology of the bacteria have been difficult to assess. Establishing different in vitro culture conditions to represent active and stationary phase, Kumar et al. [53] have shown that genes within *mce* operons 1–4 are expressed in stationary phase (such as in standing cultures), while one or more *mce* genes are expressed during active growth. Only *mce1* and *mce4* of *M. tuberculosis* have been deemed essential for survival [31]. Kumar et al. [53] proposed that the biological functions and evolution of these clusters indicate a fundamental role in transport (for nutrients, metabolites, and extrusion of toxic molecules in saprophytic organisms via the Ttg2C domain), which have then been adapted for cell entry functions by pathogenic bacteria via the *mce* domain.

Fig. 2 Pictorial representation of *mce* (pfam02470, cell entry) and Ttg (COG1463, transporter) conserved domains in *mce* proteins of *M. tuberculosis*

Esx

ESAT-6 and CFP-10 are small molecular weight secreted protein antigens, implicated as virulence factors in *M. tuberculosis*, but lacking in the attenuated *M. bovis* BCG vaccine strain. The pair of genes encoding these proteins are located within a cluster known as the *esx*-1 locus, which potentially encodes a complex of multiple proteins forming a novel transport system worthy of a separate systematic nomenclature, i.e. Type VII secretory system [54]. Similar to the *mce* locus, genome duplication events indicate that there is a scattered distribution of multiple *esx* loci in pathogenic and non-pathogenic mycobacteria and other gram-positive species. There appears to be an inter-dependence between *esx* loci for secretion of the ESAT-6-CFP-10 and other proteins, but questions concerning the actual in vivo substrates that are secreted, and the details about shared functionalities and protein–protein interactions between the proteins within and between different *esx* loci remain [55]. It is also argued that ESAT-6, CFP-10 homologs, and other proteins, may be structural components of the transport machinery, rather than the natural substrates and actual effectors of virulence in pathogenic species. The *esx* loci include members of the PE-PPE family. Gey Van Pittius et al. [56] have postulated that from this original location, extensive gene duplication resulted in non *esx* locus distribution of PE-PPE genes in pathogenic bacteria.

Defining *M. tuberculosis*

Virulence factors discovered by genetic engineering

As described earlier, genetic engineering and array technologies have aided in the search for virulence factors, i.e. proteins/pathways in processes such as attachment, infection, survival, persistence and reactivation. The *M. tuberculosis* loci that have emerged in independent experimental and *in silico* studies are:

1. Phthiocerol dimycocerosic acids (PDIMs) cluster of genes for polyketide synthesis, acylation, and lipid transport [27, 28, 57]
2. *esx* loci (7 in *M. tuberculosis*) [54, 55, 56, 58]
3. *mce* locus (4 in *M. tuberculosis*) [53, 59]
4. PE-PGRS and PPE genes (there are 66 genes in *M. tuberculosis*. One of these, *Rv3018c* was found by signature tagged mutagenesis [3, 50, 60]
5. Fatty acid metabolism (anabolism and catabolism) [61]

Of these, as described earlier, the presence of one or more copies of the *esx*, *mce* and PE-PPE genes is a feature shared by pathogenic mycobacteria. *M. avium* subsp. *avium* and

M. avium subsp. *paratuberculosis* are endowed with glycopeptidolipids (GPLs) instead of PDIMs.

- An innovative example of comparative genomics for functional validation is the discovery of the role for one of the four *mce* loci (*mce4*) in cholesterol catabolism and for survival of *M. tuberculosis* in macrophages [62]. This association was uncovered by using the soil actinomycete *Rhodococcus* sp. RHA1 as the model strain for profiling the genetics of cholesterol uptake and degradation. A large cluster (~80 genes) conserved in *M. tuberculosis* H37Rv (Rv3492c–Rv3574), *M. bovis* BCG and *M. avium* subsp. *paratuberculosis* harbored the equivalent genes found in RHA1 for cholesterol catabolism. When we look at other non-pathogenic and environmental mycobacteria and corynebacteria, this cluster of genes is also conserved in the saprophyte *M. smegmatis* and obligate parasite *M. ulcerans*, but virtually deleted in *M. leprae*, and soil organism *C. glutamicum*. Perhaps, alternate sources of energy or host enzymes overcome the absence of this pathway in *M. leprae* or may be a related to its slow rate of replication.

Multiple studies have investigated genes involved in signaling, DNA replication, DNA repair, cell division, secretion and transport of proteins and small molecules, and nutrition based on phenotypes of reference strains and their respective gene knock outs in animal models (beyond the scope of this review). Based on such approaches, a number of vaccine strains, drug targets and diagnostic reagents have been proposed, although only few of these research findings have been tested for clinically applicable products.

Genetics of natural populations of *M. tuberculosis*

Although ‘essential’ and ‘virulence’ genes are often identified through various experimental animal models using reference strains, it is of interest to verify if these findings are relevant to clinical strains and well defined study populations.

- The work of Maeda et al. [63] and Tsolaki et al. [64] attempted to study the genomes of clinical isolates (in an array hybridization format) and to relate the genotypes to transmission phenotypes. They included strains with well characterized clinical and epidemiological datasets. The Maeda et al. [63] study includes 13 representative clones from a larger collection (taken from 1744 patients studied in San Francisco during a seven year period in the 1990s that were responsible for 148 TB cases and

Table 4 COGs enriched in *Mycobacterium tuberculosis* H37Rv: shared and unique genes

COG ID	Gene function	Number of genes within the COG										
			Mfb	Mb	Map	Mav	Mul	Mlp	Ms	Mjls	Cg	Ec
COG5651	PPE-repeat proteins	66	**	61	36	35	44	6	2	3	0	0
COG3321	Polyketide synthase modules and related proteins	20	**	19	10	10	12	9	6	10	1	0
COG0277	FAD/FMN-containing dehydrogenases	12	**	11	6	6	7	4	10	6	4	5
COG0463	Glycosyltransferases involved in cell wall biogenesis	9		7	5	4	6	3	5	6	5	5
COG4842	Uncharacterized protein conserved in bacteria	8		7	5	6	5	3	6	5	2	0
COG0455	ATPases involved in chromosome partitioning	6		5	4	4	0	2	4	4	0	0
COG3511	Phospholipase C	4		1	0	0	1	0	0	0	0	0
COG0399	Predicted pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biogenesis	4		3	0	0	0	0	0	0	1	2
COG2224	Isocitrate lyase	3		2	2	2	2	1	2	2	1	1
COG0314	Molybdopterin converting factor, large subunit	3		2	1	1	1	0	1	1	1	1
COG3293	Transposase and inactivated derivatives	3		2	0	0	0	0	0	0	0	0
COG1770	Protease II	2		1	1	1	1	1	1	1	1	1
COG0820	Predicted Fe-S-cluster redox enzyme	2		1	1	0	1	0	1	1	1	1
COG1085	Galactose-1-phosphate uridylyltransferase	2		1	1	1	1	0	1	0	1	1
COG1461	Predicted kinase related to dihydroxyacetone kinase	2		1	1	1	0	0	1	1	1	0
COG0810	Periplasmic protein TonB, links inner and outer membranes	2		1	0	0	0	0	0	0	0	1
COG3740	Phage head maturation protease	2		1	0	0	0	0	0	0	0	0
COG3747	Phage terminase, small subunit	2		1	0	0	0	0	0	0	0	0
COG4653	Predicted phage phi-C31 gp36 major capsid-like protein	2		1	0	0	0	0	0	0	0	0
COG1089	GDP-D-mannose dehydratase	2		0	1	0	0	0	0	0	0	1
COG1948	ERCC4-type nuclease	1		0	0	0	0	0	0	0	0	0
COG5343	Uncharacterized protein conserved in bacteria	1		0	0	0	0	0	0	0	0	0

^a = clinical strain with a deletion in this gene was reported; Genes in bold = unique genes in *Mycobacterium tuberculosis* H37Rv; Italicized genes are either pseudogenes or split non-functional genes

The environmental species Ms and Mjls are shaded

Table 5 COGs enriched in *Mycobacterium avium* subsp. *paratuberculosis* K-10: shared and unique genes

COG ID	Gene function	Number of genes within the COG										
		Map	Mtb	Mb	Mav	Mul	Mlp	Ms	Mjls	Cg	Ec	
COG2409	Predicted drug exporters of the RND superfamily	18	14	16	15	7	5	17	12	4	0	
COG1835	Predicted acyltransferases	11	5	5	10	6	4	10	8	4	0	
COG1020	Non-ribosomal peptide synthetase modules and related proteins	10	4	4	8	6	1	9	4	1	1	
COG2227	2-polypropenyl-3-methyl-5-hydroxy-6-methoxy-1,4-benzoquinol methylase	10	7	5	8	7	1	8	5	1	2	
COG1216	Predicted glycosyltransferases	6	4	4	4	4	3	5	3	4	0	
COG266	Formamidopyrimidine-DNA glycosylase	5	4	4	4	4	1	4	4	3	2	
COG753	Catalase	5	0	0	4	2	0	3	1	1	1	
COG3707	Response regulator with putative antiterminator output domain	5	1	1	4	2	1	3	3	0	0	
COG523	Putative GTPases (G3E family)	4	1	1	1	1	0	3	2	1	2	
COG155	Sulfite reductase, beta subunit (hemoprotein)	3	2	2	2	2	0	2	2	2	1	
COG1121	ABC-type Mn/Zn transport systems, ATPase component	3	0	0	2	1	1	2	2	2	1	
COG1240	Mg-chelatase subunit ChlD	3	1	1	2	2	1	1	1	1	0	
COG707	UDP-N-acetylglucosamine:LPS N-acetylglucosamine transferase	2	1	1	1	1	1	1	1	1	1	
COG1066	Predicted ATP-dependent serine protease	2	1	1	1	0	0	1	1	1	1	
COG1074	ATP-dependent exoDNase (exonuclease V) beta subunit (contains helicase and exonuclease domains)	2	1	1	1	1	0	1	1	0	1	
COG2201	Chemotaxis response regulator containing a CheY-like receiver domain and a methyltransferase domain	2	0	0	1	0	0	0	0	0	1	
COG2216	High-affinity K+ transport system, ATPase chain B	2	1	1	1	0	0	1	1	0	1	
COG2365	Protein tyrosine/serine phosphatase	2	1	1	1	1	0	1	1	0	0	
COG3256	Nitric oxide reductase large subunit	2	0	0	1	0	0	0	1	0	0	
COG3953	SLT domain proteins	2	1	1	1	1	0	0	0	0	0	
COG4717	Uncharacterized conserved protein	2	1	1	1	1	1	1	1	0	0	
COG784	FOG: CheY-like receiver	1	0	0	0	0	0	0	0	0	0	
COG2221	Dissimilatory sulfite reductase (desulfovibrin), alpha and beta subunits	1	0	0	0	0	0	0	0	0	0	
COG2324	Predicted membrane protein	1	0	0	0	0	0	0	0	0	0	
COG3319	Thioesterase domains of type I polyketide synthases or non-ribosomal peptide synthetases	1	0	0	0	0	0	0	0	0	0	
COG4693	Oxidoreductase (NAD-binding), involved in siderophore biosynthesis	1	0	0	0	0	0	0	0	0	0	

Genes in bold = unique genes in *M. avium* subsp. *paratuberculosis*
Mtb: *M. tuberculosis* H37Rv; Map: *M. avium* subsp. *paratuberculosis*; Mav: *M. avium* subsp. *avium*; Mul: *M. ulcerans*; Mlp: *M. leprae*; Ms: *M. smegmatis*; Mjls: *M. JLS*; Cg: *C. glutamicum*; Ec: *E. coli*
The environmental species Ms and Mjls are shaded

Table 6 COGs enriched in *Mycobacterium ulcerans* Agy9: shared and unique genes

COG ID	Gene function	Number of genes within the COG											
		Mul	Mtb	Mb	Map	Mav	Mlp	Ms	Mjls	Cg	Ec		
COG3328	Transposase and inactivated derivatives	73	Too many to list	9	9	7	66	0	3	3	0	0	
COG2114	Adenylate cyclase, family 3 (some proteins contain HAMP domain)	14	MUL_0687, MUL_1472, MUL_2284 , MUL_3594, MUL_4940	13	13	10	9	2	7	10	1	0	
COG1695	Predicted transcriptional regulators	7	MUL_2388	3	3	4	4	2	5	6	3	1	
COG3239	Fatty acid desaturase	6	MUL_2564, MUL_4931	3	3	4	2	0	5	5	0	0	
COG156	7-keto-8-aminopelargolate synthetase and related enzymes	5	MUL_0241, MUL_1966, MUL_2843 , MUL_4045	2	2	1	1	1	1	1	0	2	
COG1192	ATPases involved in chromosome partitioning	4		3	3	3	3	2	3	3	2	1	
COG1773	Rubredoxin	4	MUL_2747, MUL_2748	2	2	0	2	0	2	1	0	0	
COG1231	Monoamine oxidase	3	MUL_1281, MUL_2489	1	1	1	1	0	1	1	1	0	
COG2138	Uncharacterized conserved protein	3		2	2	2	2	0	2	2	1	0	
COG2308	Uncharacterized conserved protein	3	MUL_4094	2	2	2	1	2	2	2	0	0	
COG3320	Putative dehydrogenase domain of multifunctional non-ribosomal peptide synthetases and related enzymes	3	MUL_4346	1	1	2	2	0	2	0	0	0	
COG302	GTP cyclohydrolase I	2	MUL_2233	1	1	1	1	1	1	1	1	1	
COG324	tRNA delta(2)-isopentenylpyrophosphate transferase	2	MUL_3469	1	1	1	1	1	1	1	1	1	
COG393	Uncharacterized conserved protein	2	MUL_4365	1	1	0	0	1	0	0	0	1	
COG450	Peroxiredoxin	2	MUL_2912	1	1	1	1	1	1	0	0	1	
COG509	Glycine cleavage system H protein (lipoate-binding)	2	MUL_4903	1	1	1	1	1	1	1	0	1	
COG1219	ATP-dependent protease Clp, ATPase subunit	2	MUL_2703	1	1	1	1	1	1	1	1	1	

Table 6 (Continued)

[illegible]

Genes in bold = unique genes in *M. ulcerans*

Mtb: *M. tuberculosis* H37Kv; Map: *M. avium* subsp. *paratub*; Mav: *M. avium* subsp. *avium*; Mul: *M. ulcerans*; Mlp: *M. leprae*; Msm: *M. smegmatis*; Mjls: *M. JLS*; Cg: *C. glutamicum*; Ec: *E. coli*

Fig. 3 Examples of COGs abundance differences for 10 species. In each panel, the COG number is shown along with X axis and the description on top.

implicated in 358 others). The term clone refers to a group of isolates likely to be genetically linked, i.e. derived from a common progenitor. One of these clones was involved in a cluster of 41 patients. The technique identified large sequence polymorphisms (LSPs) with a detection limit of 350 bp deletion, but was not sensitive for single nucleotide polymorphisms (SNPs) and insertions. Deletions of an average of 0.3% of the genome, accounting for ~17 open reading frames, were found. As the extent of deletions increased, the probability of pulmonary cavitation, an indicator of clinical pathogenicity decreased.

Interestingly, several of the *M. tuberculosis* unique genes we listed in Table 4 were found to be deleted from some

clinical strains (such as dehydrogenases, phospholipases, phage proteins). One or more genes within a large cluster (Rv1500-Rv1527c) that contains glycosyltransferases are deleted in some of these clones. The biological role of this cluster has not been described thus far, but may correspond in part to the lipooligosaccharide (LOS) biosynthetic locus in *M. marinum* [65]. A deletion of *pks5* (Rv1527c) in H37Rv strain however diminishes, but does not abrogate virulence or persistence in mice [66]. LOSs are considered to be ‘avirulence’ factors absent in most clinical strains of *M. tuberculosis*.

From the above San Francisco *M. tuberculosis* strain bank, Tsolaki et al. [64] further studied 100 strains to look for lineage specific and non-lineage specific genetic variations (50 that were involved in transmission clusters, and

50 that were unique or non-clustered). They identified ~250 regions of difference (RD) [64]. Theoretically, these 'RD' genes are non-essential for human disease, or may have altered levels of virulence and host specificity. Alternatively, by examining the RDs, it was postulated that dumping extra copies of mobile elements and lipoproteins reduces genomic and antigenic load (or immune evasion). Interesting was the finding of deletions of *katG* and *furA*, offering a possible antibiotic resistance mechanism; and that of genes within 'hypoxia' regulons conducive to the escape from latency, a means of promoting transmission. These events were considered to be 'positively selected' in phylogenetically unrelated isolates within vulnerable regions of the genome, while certain other genes were deleted in specific lineages only. Overall, the authors noted that the degree of LSPs was limited and not expected to exceed ~100 genes in *M. tuberculosis* isolates (5.5% of total genes), while in *H. pylori*, 22% genes can be deleted in a small sample set.

- Fleischmann et al. [67] took advantage of the complete genome sequences of *M. tuberculosis* strains CDC1551 and H37Rv and identified SNP and LSPs (greater than 10 bp in this study). From this panel of genomic markers, a few were selected for interrogating polymorphisms in 169 epidemiologically characterized clinical isolates. In this and a subsequent study, it was found that LSPs can occur multiple times, and as independent events, flanking IS6110 sequences being one of the factors. One of these LSP groups (LSP A) comprising four loci was judged to be suitable for phylogenetic interpretations, while other LSPs occurred in regions subject to selection (rich in PE and PPE genes). Association of genotype to phenotype was indicated for deletion in *plcD* [68]; extrapulmonary TB was two fold more likely with *plcD* mutant strains [69].
- Noting that *M. tuberculosis* strains from Beijing, China are more closely related to each other, Van Soolingen et al. first described the Beijing strains of *M. tuberculosis* [70]. These strains have similar IS6110 DNA containing restriction fragments. DNA polymorphisms within other repetitive DNA elements, such as the PGRS domains of PE-PPE genes and within the direct repeat (DR) used for a strain typing method known as spoligotyping is very limited.

The Beijing strains of *M. tuberculosis* are thought to be highly pathogenic strains since they acquire drug resistance [71]. They are thought to have emerged from China where BCG vaccination has been implemented for almost two decades and that this vaccination favored their

selection, resisting BCG-induced immunity [72]. Common genomic features of the Beijing family of *M. tuberculosis* are [73]:

- The copy number for insertion sequence IS6110 is in the range of 15–26.
- The Spoligotype (S00034) contains nine spacers from 35 to 43.
- IS6110 insertion in the origin of replication (corresponds to a 3.36 kb *PvuII* band in a Southern hybridization blot probed with *dnaA-dnaN* fragment). Two insertion sequences (IS) are observed in this region in the 'W' strain of the Beijing family.
- Mutations in *katG* codon 463 and *gyrA* codon 95 are associated with drug resistance.

Beijing strains are also found to demonstrate distinct expression patterns of proteins; several species of α -crystallin (a known *M. tuberculosis* virulence factor) are enhanced, while there is decreased expression of heat shock protein of 65 kDa and many others [72]. In addition, the Beijing family strains produce high levels of phenolic glycolipid (PGL-tb), not made by H37Rv. These altered expression of proteins and glycolipids are thought to contribute to the success of the Beijing family strains. The Beijing/W strains have three times the propensity of non-Beijing strains to be associated with extrathoracic TB [74].

- Dormans et al. [75] performed a comparative study of the phenotypes associated with nine different global major genotypes based on intratracheal mouse infection (progressive pulmonary tuberculosis model). They used a semi-quantitative scoring system to measure various parameters including histopathology of lung, bacillary load, survival and delayed type hypersensitivity. The genotypes could be broadly divided into three groups with low, moderate and high levels of virulence which correlated well with severity of histopathology and increased bacillary counts and reduced survival time. However the virulent strains also elicited the highest levels of DTH reactivity indicating a lack of correlation between DTH and protection. In general, the Beijing type and Africa strains were more virulent than the Amsterdam and Haarlem strains, while the H37Rv and Canetti strains were the least virulent in this study.

More studies that examine clinical isolates are needed for better evaluation of genotypes and gene functions in disease and immunity and to examine if there are interactions between host and bacteria when studied in different populations. Genetic markers, platforms for inexpensive and high throughput genome comparisons are thus warranted to extrapolate and validate the information that is generated

from reference strains as they affect the development and efficacy of diagnostics, vaccines and drugs.

Our COG abundance screening has identified a list of additional genes present in *M. tuberculosis* but not in other members of the test panel (Table 4). These include extra copies for a given function as for the phospholipase, phage proteins, and a few genes (Rv1514 and Rv1518) within a larger cluster of shared genes that are proximal to those involved in LOS in *C. glutamicum*. Genes within-frame mutation resulting in split genes and/or pseudogenes account for a few in this list. In summary (excluding the larger PPE, polyketide synthases and FAD/FMN dehydrogenase COGs), 26 genes were found only in *M. tuberculosis*, only one of these is deemed 'essential' [30]; deletion mutants have been detected in clinical strains for nine. Therefore, these 26 genes in themselves may not solely define the *M. tuberculosis* phenotype, but their presence or absence, individually or in combination with other genes may contribute to specific experimental and/or clinical states.

Defining and differentiating *Mycobacterium avium* subsp.

With regard to *M. avium*, collectively known as the *M. avium* complex (MAC), researchers are seeking better species definitions and nomenclature to enable rationale approaches for identifying source and modes of transmission of human and animal diseases and developing diagnostics and vaccines. Considering that MAC organisms are found to reside in both environmental and host associated niches, Turenne et al. [76] have recommended that MAC be considered as a 'microcosm' of mycobacteria with distinct genomic identities. It is expected that additional sequencing of representative genomes from different host niches will clarify some of the existing confusion in taxonomy.

Three subsets of *M. avium*, obtained from non-human sources have been defined according to DNA–DNA hybridization and phenotypic properties (growth and biochemical tests); *M. avium* subsp. *avium*, *M. avium* subsp. *paratuberculosis* and *M. avium* subsp. *silvaticum*. Although *M. intracellulare* is given a species status and is found more often in immune competent hosts as opposed to the *M. avium* subsets associated with immune-deficient patients, it is placed within MAC, and controlled with therapeutic regimens common to *M. avium* subsp. MAC now also includes *M. avium* subsp. *hominissuis* of which the reference strain no. 104 has been sequenced. The *M. avium* subsp. *hominissuis* differs from *M. avium* sp. by the presence of multiple copies of IS1245 insertion sequence, variable 16S–23S internal transcribed spacer, tolerance to a wide temperature range and the lack of IS901 sequences. *M. lepraemurium* is related to MACs by DNA–DNA hybridization, and in 16S rRNA sequence, but not by hsp65 sequence; it is not placed in MAC. Apparently *M. avium* subsp. *silvaticum* is hardly distinguishable from *M. avium* subsp. *avium* and doesn't warrant a subspecies classification.

- Turenne et al. [76] have pointed out that although *M. avium* subsp. *avium* are popularly referred to as environmental species, strains found in birds are not found in human MAC infections and environmental sources [71]. On the other hand the *M. avium* subsp. *hominissuis* is found in multiple sources and more likely represents an environmental species.

Due to these confounding observations, studies in MAC have often dealt with diagnostic issues and developing molecular diagnostic probes has been a major thrust for clinical and field applications. In addition, genomes have been queried to search for genes responsible for host and tissue specificity and the differences in growth rates and specific requirements in vitro culture.

Table 7 Structures of glycosyl modifications in phenolic glycolipids of mycobacteria

Species	Trivial name	Oligosaccharide
<i>M. bovis</i>	Mycoside B1	2-O-Me- α -L-Rhap-PDIM
<i>M. bovis</i>	Mycoside B2	α -L-Rhap-PDIM
<i>M. bovis</i>	Mycoside B3	L-Rhap- 2-O-Me- α -L-Rhap-PDIM
<i>M. marinum</i>	-	α -L-Rhap-PDIM
<i>M. ulcerans</i>	-	PDIM*
<i>M. tuberculosis</i>	PGL-Tb-1	2,3,4-tri-O-Me- α -L-Fucp-(1-3)- α -L-Rhap(1-3)-2-O- Me- α -L-Rhap-PDIM
<i>M. leprae</i>	PGL-1	3,6-di-O-Me- β -D-Glcp-(1-4)-2,3-di-O-Me- α -L- Rhap-(1-2)-3-O-Me- α -L-Rhap-PDIM
<i>M. kansasii</i>	PGL-K1	2,6-dideoxy-4-O-Me α -ara-Hexp-(1-3)-4-O-Ac- 2-O-Me - α -L-Fucp(1-3)-2-O-Me- α -L-Rhap-(1-3)-2,4- di-O-Me- α - L-Rhap-PDIM
<i>M. haemophilum</i>	-	2,3-di-O-Me- α -L-Rhap(1-2)-3-O-Me- α -L- Rhap(1-4)-2,3-di-O-Me- α -L-Rhap-PDIM

* Produces only the phenol phthiodolone dimycocerosic acid

Fig. 4 Comparative genomic cluster of glycosyltransferases (Gtf), methyltransferases (Mtf), ketoreductases and enolreductases of phenolic glycolipid biosynthesis. Genes represented in dotted boxes indicate pseudogenes.

- Availability of *M. avium* subsp. *avium* 104 (human isolate) and *M. avium* subsp. *paratuberculosis* K10 (cow strain) sequences allowed a search for LSPs that can be diagnostic for each. Fourteen LSPs, in *M. avium* subsp. *paratuberculosis* (LSP [P]s) and three in *M. avium* subsp. *avium* (LSP [A]s) were found by Semret et al. [77]. When tested against large panels of MAC isolates, three LSPs (LSP [P] 2, 4 and 15) were found to be 100% specific for *M. avium* subsp. *paratuberculosis* (i.e. absent in non-*M. avium* subsp. *paratuberculosis* isolates). However, due to variable distribution in isolates or PCR technicalities, only two are reliably present and suitable for diagnosis of *M. avium* subsp. *paratuberculosis*. Besides, these two LSPs convey biological information, LSP [P] 12 carries an *mce* operon, while LSP [P] 15 encodes genes for iron transport (mycobactins, mycobacterial siderophores, are absent in this species). LSPs did not include the redundant PE/PPE genes in contrast to the clinical variants in *M. tuberculosis* strains.

- *M. avium* subsp. *paratuberculosis* is an infectious agent of enteric disease in a broad range of hosts: cattle, goats, sheep and other wild ruminants. Evidence for its role in human Crohn's disease is still actively debated [76, 78]. *M. avium* subsp. *paratuberculosis* isolates are further classified according to RFLP patterns and other phenotypic properties as S for sheep and C for cattle, referring to host preferences. The C strains have a broader host range than S strains [41]. Subtractive DNA hybridization techniques lead to the identification of a large deletion in the S strain covering 10 genes

(MAP1734 to MAP1743c) of the *M. avium* subsp. *paratuberculosis* genome, which may account for the fastidious growth and host specificity of the S strains. This deletion includes the *mmpL5* gene, which belongs to a family implicated in transport of complex lipids. Several other SNPs suitable for distinguishing the S and C strains were found by this representational difference analysis (RDA) technique.

- Macrophage infection is a common feature of mycobacterial pathogens. Danelishvili et al. [79] screened a transposon mutant library of *M. avium* subsp. *avium* (strain 109 isolated from an AIDS patient), for defects in in vitro macrophage invasion. A locus absent in *M. tuberculosis* and *M. avium* subsp. *paratuberculosis* was identified. This locus, of lower G+C content postulated to have been acquired by horizontal gene transfer, is responsible for growth in environmental niches such as the protozoa *Acanthamoeba*, a property lacking in *M. tuberculosis* and *M. avium* subsp. *paratuberculosis*. As for the *mce* and *esx* loci, the macrophage invasion locus appears to encode transport proteins that are secreted into the host, in this case to promote actin polymerization for entry into the macrophage/amoeba.

By our COG abundance comparison of *M. avium* subsp. *paratuberculosis* versus the others in the panel, as expected, *M. avium* subsp. *paratuberculosis* genome is generally more similar to *M. avium* subsp. *avium* (Table 5). Overall, these two species contain COGs as seen in 'environmental' and non-pathogens such as *M. smegmatis* and *Mycobacterium*

Table 8 Genetics of mycobacterial cell wall and associated macromolecules, and shared features in *C. glutamicum*

Name	<i>M. tuberculosis</i>	<i>M. leprae</i>	<i>M. bovis</i>	<i>M. avium</i>	<i>M. avium</i> subsp. <i>paratub</i>	<i>M. smegmatis</i>	<i>M. marinum</i>	<i>M. ulcerans</i>	<i>C. glutamicum</i>	Function	References
Polyprenyl phosphate synthesis											
<i>dxs1</i>	Rv2682c	ML1038	Mb2701	MAV_3577	MAP2803c	MSMEG_2776	MMAR_2032	MUL_3319	NCg11827	1-deoxy-D-xylulose 5-phosphate synthase	90
<i>dxs2</i>	Rv3379c		Mb3413c				MMAR_0276			Probable 1-deoxy-D-xylulose 5-phosphate synthase	
<i>dxr</i>	Rv2870c	ML1583	Mb2895c	MAV_3727	MAP2940c	MSMEG_2578	MMAR_1836	MUL_2085	NCg11940	1-deoxy-D-xylulose 5-phosphate reductoisomerase	91
<i>ispD</i>	Rv3582c	ML0321	Mb3613c	MAV_0571	MAP0476	MSMEG_6076	MMAR_5082	MUL_4158	NCg12570	4-diphosphocytidyl-2C-methyl-D-erythritol synthase	92
<i>ispE</i>		ML0242	Mb1038	MAV_1149	MAP0976	MSMEG_5436	MMAR_4477	MUL_4649	NCg10874	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase	
<i>ispF</i>	Rv3581c	ML0322	Mb3612c	MAV_0572	MAP0477	MSMEG_6075	MMAR_5081	MUL_4157	NCg12569	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	93
<i>ispG</i> or <i>gcpE</i>	Rv2868c	ML1581	Mb2893c	MAV_3725	MAP2938c	MSMEG_2580	MMAR_1838	MUL_2087	NCg11938	1-hydroxy-2-methyl-2-(e)-butenyl 4-diphosphate synthase	
<i>ispH</i> or <i>htrB1</i>	Rv3382c		Mb3414c							4-hydroxy-3-methylbut-2-enyl diphosphate reductase	
<i>htrB2</i>	Rv1110	ML1938	Mb1140	MAV_1230	MAP2684C	MSMEG_5224	MMAR_0277	MUL_0168	NCg10982	4-hydroxy-3-methylbut-2-enyl diphosphate reductase 2	
<i>idi</i>	Rv1745c		Mb1774c				MMAR_3218	MUL_3526	NCg12223	isopentenyl-diphosphate delta-isomerase	
<i>upps</i> [^]	Rv1086	ML2467	Mb1115	MAV_2034	MAP2703c	MSMEG_5256	MMAR_4380	MUL_0193	NCg10951	Short (C15) chain Z-isoprenyl diphosphate synthase	94
<i>upps</i> [^]	Rv2361c	ML0634	Mb2382c			MSMEG_4490	MMAR_3671	MUL_3614	NCg12203	Long (C50) chain Z-isoprenyl diphosphate synthase	94
<i>idsA1</i> [^]	Rv3398c	ML0900	Mb3431c	MAV_4802	MAP3846	MSMEG_1133	MMAR_5095	MUL_4171*	NCg12092	Geranylgeranyl pyrophosphate synthetase	95
<i>idsA2</i> [^]	Rv2173		Mb2195	MAV_2321	MAP1911	MSMEG_4240	MMAR_2098	MUL_3516		"	
<i>idsB</i> [^]	Rv3383c		Mb3415c	MAV_3884	MAP3069		MMAR_2564	MUL_3197		polyprenyl diphosphate synthase	
<i>murA</i>	Rv1315	ML1150	Mb1348	MAV_1531	MAP2447c	MSMEG_4932	MMAR_4083	MUL_3950	NCg12470	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	96
<i>murB</i>	Rv0482	ML2447	Mb0492	MAV_4668	MAP3975	MSMEG_0928	MMAR_0808	MUL_4552	NCg10386	UDP-N-acetylenolpyruvoylglucosamine reductase	
<i>murC</i>	Rv2152c	ML0915	Mb2176c	MAV_2337	MAP1896c	MSMEG_4226	MMAR_3192	MUL_3500	NCg12077	UDP-N-acetylmuramate-alanine ligase	97
<i>murD</i>	Rv2155c	ML0912	Mb2179c	MAV_2334	MAP1899c	MSMEG_4229	MMAR_3195	MUL_3503	NCg12080	UDP-N-acetylmuramoylalanine-D-glutamate ligase	98
<i>murE</i>	Rv2158c	ML0909	Mb2182c	MAV_2331	MAP1902c	MSMEG_4232	MMAR_3198	MUL_3506	NCg12083	UDP-N-acetylmuramoylalanine-D-glutamate-2,6-diaminopimelate ligase	98
<i>murF</i>	Rv2157c	ML0910	Mb2181c	MAV_2332	MAP1901c	MSMEG_4231	MMAR_3197	MUL_3505	NCg12082	UDP-N-acetylmuramoylalanine-D-glutamate-2,6-diaminopimelate-D-alanyl-D-alanyl ligase	98
<i>murX</i>	Rv2156c	ML0911	Mb2180c	MAV_2333	MAP1900c	MSMEG_4230	MMAR_3196	MUL_3504	NCg12081	phospho-N-acetylmuramoyl-pentapeptidetransferase	98

<i>murG</i>	Rv2153	ML0915	Mb2177c	MAV_2336	MAP1897c	MSMEG_4227	MMAR_3193	MUL_3501	NCg12078	UDP-N-acetylglucosamine-N-acetylmutamyl- (pentapeptide) pyrophosphoryl-undecaprenol-N- acetylglucosamine transferase	98
<i>ponA1</i>	Rv0050	ML2688	Mb0051	MAV_0071	MAP0064	MSMEG_6900	MMAR_0069	MUL_0068	NCg12884	bifunctional penicillin-binding protein (PBP) 1A/1B	99
<i>ponA2</i>	Rv3682	ML2308	Mb3707	MAV_0446	MAP0392c	MSMEG_6201	MMAR_5171	MUL_4257	NCg10274	bifunctional membrane-associated penicillin- binding protein (PBP) 1A/1B	
Linker unit and Arabinogalactan synthesis											
dTDP-rhamnose synthesis											
<i>rmlA</i>	Rv0334	ML2503	Mb0341	MAV_4228	MAP3828	MSMEG_0384/ MEMEG_5983	MMAR_0606	MUL_0568	NCg10325	alpha-D-glucose-1-phosphate thymidyl- transferase	100
<i>rmlB</i>	Rv3464	ML1964	Mb3493	MAV_4406	MAP4225c	MSMEG_1512	MMAR_1082	MUL_0840	NCg10327	dTDP-glucose-4,6-dehydratase	100
<i>rmlC</i>	Rv3465	ML1965	Mb3494	MAV_4407	MAP4224c	MSMEG_1510/ 5977	MMAR_1081	MUL_0839	NCg10326	dTDP-4-dehydrothamnose 3,5-epimerase	100
<i>rmlD</i>	Rv3266c	ML0751	Mb3294c	MAV_4231	MAP3380c	MSMEG_1825	MMAR_1275	MUL_2612	NCg10326	dTDP-6-deoxy-L-lyxo-4-hexulose reductase	100
UDP-gactatofuranose synthesis											
<i>galE</i>	Rv3634c	ML0204	Mb3658c	MAV_0524	MAP0430	MSMEG_6142	MMAR_5133	MUL_4210	NCg10317	UDP-glucose 4-epimerase	101
<i>glf</i>	Rv3809c	ML0092	Mb3839c	MAV_0208	MAP0211	MSMEG_6404	MMAR_5373	MUL_4993	NCg12788	UDP-galactopyranose mutase	102
Lipid linked linker unit synthesis and arabinogalactan polymerization											
DPA synthase [^]	Rv3790	ML0109	Mb3819	MAV_0232	MAP0235c	MSMEG_6382	MMAR_5352	MUL_4969	NCg10187	DPA synthase	103
DPA synthase [^]	Rv3791	ML0108	Mb3820	MAV_0231	MAP0234c	MSMEG_6385	MMAR_5353	MUL_4970	NCg10186	DPA synthase	103
<i>rfe or wecA</i>	Rv1302	ML1137	Mb1334	MAV_1519	MAP2459	MSMEG_4947	MMAR_4095	MUL_3962	NCg11156	undecaprenyl-phosphate alpha-N- acetylgluco samyltransferase	104
<i>wblI</i>	Rv3265c	ML0752	Mb3293c	MAV_4230	MAP3379c	MSMEG_1826	MMAR_1276	MUL_2611	NCg10709	dTDP-rha:A-D-GlcNAc-diphosphoryl polyprenol A-3- L-rhamnosyl transferase	105
<i>glfT</i>	Rv3808c	ML0093	Mb3838c	MAV_0209	MAP0212	MSMEG_6403	MMAR_5372	MUL_4992	NCg12783	bifunctional UDP-galactofuranosyl transferase	106
<i>glfT[^]</i>	Rv3782	ML0113	Mb3811	MAV_0237	MAP0240c	MSMEG_6367	MMAR_5337	MUL_0091	NCg10195	"	107
<i>afIA[^]</i>	Rv3792	ML0107	Mb3821	MAV_0230	MAP0233c	MSMEG_6386	MMAR_5354	MUL_4971	NCg10185	Arabinosyltransferase: priming enzyme on galactan core	108
<i>embA</i>	Rv3794	ML0105	Mb3823	MAV_0229	MAP0229c	MSMEG_6389	MMAR_5356	MUL_4973	NCg10184	Arabinosyltransferase	108
<i>embB</i>	Rv3795	ML0104	Mb3824		MAP0228c	MSMEG_6388	MMAR_5357	MUL_4974		Arabinosyltransferase	108
<i>afIB[^]</i>	Rv3805c	ML0096	Mb3835c	MAV_0212	MAP0215	MSMEG_6400	MMAR_5369	MUL_4989	NCg12780	Arabinosyltransferase: Terminal β capping enzyme	109
Mycolic acid synthesis, condensation and deposition											
α-branch synthesis											
<i>fas</i>	Rv2524c	ML1191	Mb2553c	MAV_1650	MAP2332c	MSMEG_4757	MMAR_3962	MUL_3818	NCg12409	Fatty Acid Synthase	110
Meromycolic acid synthesis											
<i>accD6</i>	Rv2247	ML1657	Mb2271	MAV_2190	MAP2000	MSMEG_4329	MMAR_3340	MUL_1302	NCg10677	Acetyl/Propionyl CoA Carboxylase	111
<i>acpM</i>	Rv2244	ML1657	Mb2268	MAV_2193	MAP1997	MSMEG_4326	MMAR_3337	MUL_1305	NCg12174	meromycolate extension acyl carrier protein	112

Table 8 (Continued)

<i>fadD</i>	Rv2243	ML1653	Mb2267	MAV_2194	MAP1996	MSMEG_4325	MMAR_3336	MUL_1306	Malonyl CoA:AcylM acyltransferase	113
<i>fadH</i>	Rv0533c		Mb0547c	MAV_4612	MAP4028c	MSMEG_3953	MMAR_0879	MUL_0632	3-oxoacyl-[acyl-carrier-protein] synthase III	114
<i>kasA</i>	Rv2245	ML1655	Mb1519	MAV_2192	MAP1998	MSMEG_4327	MMAR_3338	MUL_1304	3-oxoacyl-[acyl-carrier-protein] synthase I	112, 115
<i>kasB</i>	Rv2246	ML1656	Mb2270	MAV_2191	MAP1999	MSMEG_4328	MMAR_3339	MUL_1303	3-oxoacyl-[acyl-carrier-protein] synthase 2	115
<i>fabG1</i>	Rv1483	ML1807	Mb1519	MAV_3295	MAP1209	MSMEG_3150	MMAR_2289	MUL_1491	3-oxoacyl-[acyl-carrier-protein] reductase	116
<i>inhA</i>	Rv1484	ML1806	Mb1520	MAV_3294	MAP1210	MSMEG_3151	MMAR_2290	MUL_1492	enoyl-[acyl-carrier-protein] reductase	117
Meromycolic acid modification										
<i>emaA1</i>	Rv3392c	ML0404	Mb3424c	MAV_0130	MAP0135	MSMEG_1351			cyclopropane mycolic acid synthase	118
<i>emaA2</i>	Rv0503c	ML2426	Mb0515c	MAV_4647	MAP3995c	MSMEG_1205	MMAR_0831	MUL_4575	cyclopropane-fatty-acyl-phospholipid synthase 2	119
<i>mmaA1</i>	Rv0645c	ML1900	Mb0664c	MAV_4516	MAP4117c		MMAR_0980	MUL_0732	methoxy mycolic acid synthase 1	120
<i>mmaA2</i>	Rv0644c	ML1901*	Mb0663c	MAV_4541	MAP4095c		MMAR_2920	MUL_0731	methoxy mycolic acid synthase 2	120, 121
<i>mmaA3</i>	Rv0643c	ML1902*	Mb0662c				MMAR_0978	MUL_0730	methoxy mycolic acid synthase 3	120
<i>mmaA4</i>	Rv0642c	ML1903	Mb0661c	MAV_4517	MAP4116c		MMAR_0977	MUL_0729	methoxy mycolic acid synthase 4	120
<i>umaA1</i>	Rv0469	ML2460*	Mb0478	MAV_4680	MAP3963	MSMEG_0913	MMAR_0794	MUL_4538	mycolic acid synthase	122
<i>umaA2</i> or <i>pcaA</i>	Rv0470c	ML2459	Mb0479c	MAV_4679	MAP3964c	MSMEG_3538	MMAR_0796	MUL_4539	mycolic acid synthase	123
<i>desA1</i>	Rv0824c	ML2185	Mb0847c	MAV_0772	MAP0658c	MSMEG_5773	MMAR_4856	MUL_0445	acyl-[acyl-carrier-protein] desaturase	124
<i>desA2</i>	Rv1094	ML1952	Mb1124	MAV_1216	MAP2698c	MSMEG_5248	MMAR_4374	MUL_0187	acyl-[acyl-carrier-protein] desaturase	124
<i>desA3</i>	Rv3229c	ML0789*	Mb3258c	MAV_4192	MAP3343c	MSMEG_1886	MMAR_1315	MUL_2565	linoleoyl-CoA desaturase	
<i>echA10^h</i>	Rv1142c		Mb1174c	MAV_1583	MAP2397	MSMEG_5185	MMAR_4309	MUL_0985	enoyl-CoA hydratase	
<i>echA11^h</i>	Rv4441c		Mb1173c	MAV_1283	MAP2639		MMAR_4302	MUL_3888	enoyl-CoA hydratase	
Mycolic acid condensation										
<i>accD4^h</i>	Rv3799c	ML0102	Mb3829c	MAV_0220	MAP0221	MSMEG_6391	MMAR_5363/4000	MUL_4982/3864	propionyl-CoA carboxylase beta chain 4	125
<i>accD5^h</i>	Rv3280	ML0731	Mb3308	MAV_4250	MAP3399	MSMEG_1813	MMAR_1256	MUL_2632	propionyl-CoA carboxylase beta chain 5	126
<i>fadD32^h</i>	Rv3801c	ML0100	Mb3831c	MAV_0217	MAP0219	MSMEG_6393	MMAR_5365	MUL_4984	fatty-acyl AMP ligase	125
<i>pks13^h</i>	Rv3800c	ML0101	Mb3803c	MAV_0218	MAP0220	MSMEG_6392	MMAR_5364	MUL_4983	Condensation enzyme of alkyl and hydroxy chains in mycolic acids	127
Deposition of mycolic acids										
<i>fbpA</i>	Rv3804c	ML0097	Mb3834c	MAV_0214	MAP0216	MSMEG_6398	MMAR_5368	MUL_4987	secreted antigen 85-A FbpA	128
<i>fbpB</i>	Rv1886c	ML2028	Mb1918c	MAV_2816	MAP1609c	MSMEG_2078	MMAR_2777	MUL_2970	secreted antigen 85-B FbpB	128
<i>fbpC</i>	Rv0129c	ML2655	Mb0134c	MAV_5183	MAP3531c	MSMEG_3580	MMAR_0328	MUL_4793	secreted antigen 85-C FbpC	128
Phthiocerol dimycocerosic acid (PDIM), phenol phthiocerol dimycocerosic acid and glycosylated PDIM synthesis										
Mycocerosic acid synthesis										
<i>mas</i>	Rv2940c	ML0139	Mb2965	MAV_1321		MSMEG_4727	MMAR_1767	MUL_2010	multifunctional mycocerosic acid synthase	129
<i>fadD28</i>	Rv2941	ML0138	Mb2966		MAP3752		MMAR_1765	MUL_2008	fatty-acyl AMP ligase FadD28	130
<i>mmpL7</i>	Rv2942	ML0137	Mb2967				MMAR_1764	MUL_2007	Translocation of Phthiocerol DiMycocerosate (PDIM) in the cell wall	130

 Springer

Table 8 (Continued)

<i>pks10</i> ^Δ	Rv1660	ML1415	Mb1688	MAV_3110	MAP1369	MSMEG_0808	MMAR_4313	MUL_1652*	shown to be useful in the biosynthesis of phthiocerol	140
<i>pks12</i> ^Δ	Rv2048c	ML1437	Mb2074c	MAV_2450	MAP1796c	MSMEG_0408	MMAR_3025	MUL_2266	shown to be useful in the biosynthesis of DIM	141
<i>pks8</i> ^Δ	Rv1662		Mb1690				MMAR_2472	MUL_1654*	shown to be useful in the biosynthesis of phthiocerol	142
<i>pks17</i> ^Δ	Rv1663		Mb1691						shown to be useful in the biosynthesis of DIM	142
Glycosylation of PDIM										
<i>rpf1</i>	Rv2962c	ML0125	Mb2986c				MMAR_1755	MUL_1998*	catalyzes the transfer of rhamnose on to the phenol of PHBA	82
<i>rpf2</i>	Rv2958c	ML0128	Mb2982c						Adds second rhamnose	82
<i>fufC</i>	Rv2357		Mb2981						fucosyltransferase (third sugar)	82
<i>mif1</i>	Rv2952	ML0130	Mb2976				MMAR_3170	MUL_2377	Transfers methyl group on the lipid moiety onto phthiotriol and glycosylated phenolphthiotriol to form to PDIm and PGL	83
<i>mif2</i>	Rv2959c	ML0127	Mb2983c						catalyzes O-methylation of the hydroxyl group on the carbon 2 of the rhamnose linked to phenol group of PGL	83
PIMs, LM and LAM synthesis										
<i>ppm1</i>	Rv2051c	ML1440 /1441	Mb2077c	MAV_2446/ MAV_2447	MAP1800c	MSMEG_3860	MMAR_5093/3029/ 3028	MUL_2269/4169	Polyprenol-monophosphomannose synthase	143
<i>pgxA</i>	Rv2612c	ML0454	Mb2644c	MAV_3488	MAP2714	MSMEG_2933	MMAR_2090	MUL_3254	PI synthase (CDP-diacylglyceride--inositol phosphatidyltransferase	144
	Rv2611c	ML0452	Mb2643c	MAV_3487	MAP2713c	MSMEG_2934	MMAR_2091	MUL_3253	acylation of the 6 position of mannose residue linked to 2 position of myoinositol in PIM1 and PIM2	145
<i>pimA</i>	Rv2610c	ML0453	Mb2642c	MAV_4386	MAP2712c	MSMEG_2935	MMAR_2092	MUL_3252	mannosyltransferase	146
<i>pimB</i>	Rv0557	ML2272	Mb0572	MAV_4586	MAP4054	MSMEG_1113	MMAR_0903	MUL_0656	mannosyltransferase	147, 148(†)
<i>pimC</i>			Mb1785c				MMAR_2629	MUL_3104	mannosyltransferase	149
<i>pimE</i> ^Δ	Rv1159	ML1504	Mb1100	MAV_1298	MAP2624c	MSMEG_5149	MMAR_4292	MUL_1006	mannosyltransferase	150
	Rv2181c ^Δ	ML0893	Mb2203	MAV_2312	MAP1919	MSMEG_4250	MMAR_3225	MUL_3536	mannosyltransferase	151
<i>embC</i>	Rv3793	ML0106	Mb3822	MAV_0225	MAP0232c	MSMEG_6387	MMAR_5355	MUL_4972	polymerizes arabinose into the arabinan of LM	152
Glycopeptidolipids synthesis ^Δ										
<i>rml2</i>						MSMEG_0387			Rhamnose 2-O-methyltransferase	153, 154, 155
<i>rml4</i>				MAV_3266		MSMEG_0388			Rhamnose 4-O-methyltransferase	153, 154, 156
<i>gtf1</i> or <i>dialf</i>				MAV_3265		MSMEG_0389			D-allo-threonine 6-deoxytolosyltransferase	156
<i>atf</i>				MAV_3274	MAP1229	MSMEG_0390			6-deoxytalose 3,4-O-acetyltransferase	157
<i>rml3</i>				MAV_3260		MSMEG_0391			Rhamnose 3-O-methyltransferase	153, 154, 156
<i>gtf</i>				MAV_3258	MAP3762c	MSMEG_0392			L-alaninol rhamnosyltransferase	157

<i>finI</i>		MAP3760c	MSMEG_0393		Fatty acid O-methyltransferase	153, 154, 156
<i>mpsI</i>			MSMEG_0395		Non-ribosomal protein synthase. Synthesis of the dipeptide	158
<i>mps2</i>			MSMEG_0396		Non-ribosomal protein synthase. Synthesis of the amino acid alcohol	158
<i>gap</i>			MSMEG_0397		Integral membrane protein. Required for GPL export	159
<i>pks</i>		MAV_3243 MAV_3262 MAV_3261	MSMEG_0402		Fatty acid synthesis and activation	160
<i>rtfA</i>					rhannosyltransferase	161, 162
<i>mtfC</i>					methyltransferase	153, 154, 156
<i>pks</i>			MSMEG_0398		Fatty acid synthesis and activation	
Sulfolipid synthesis[^]						
<i>pks2</i>	Rv3825c	MAP3764c			Polyketide synthase	163
<i>MmpI8</i>	Rv3823c		MSMEG_4741		Lipid transporter	164
<i>papA1</i>	Rv3824c				acyltransferase	165
<i>papA2</i>	Rv3820c	MAP1694	MSMEG_4728		acyltransferase	165
<i>Strf0</i>	Rv0295c	MAP2118	MSMEG_0630		Sulfotransferase	166
Trehalose synthesis[^]						
<i>otsA</i>	Rv3490	MAV_0666	MSMEG_5892	MUL_4052	alpha, alpha-trehalose-phosphate synthase [UDP-forming]	167, 168, 169
<i>treY</i> or <i>gIgY</i>	Rv1563c	MAV_3211		MUL1554	maltooligosyltrehalose synthase	167, 168, 169
<i>treS</i>	Rv0126	MAV_5186	MSMEG_6515	MUL4797	Trehalose synthase	167, 168, 169
<i>otsB1</i>	Rv2006		MSMEG_3954	MUL1852*	trehalose-6-phosphate phosphatase	167, 168, 169
<i>otsB2</i>	Rv3372	ML0414c	MSMEG_6043	MUL0921	Possibl	

Gene absent or not found

Bold letters indicate the genes are either characterized by recombinant (over-expression) or by mutant analysis

* Pseudogene

†Alternative pathway

[^] not cited in our previous reviews (42, 88)

Table 9 *M. leprae* genes not found in *M. tuberculosis*

<i>M. leprae</i>	<i>M. smegmatis</i>	<i>M. avium</i> subsp. <i>paratub</i>	<i>M. avium</i> subsp. <i>avium</i>	<i>M. ulcerans</i>	<i>C. glutamicum</i>	<i>E. coli</i>	Function
ML0142	MSMEG_6138			MUL_4357	cg0409		metallopeptidase
ML0333	MSMEG_6054	MAP0486c	MAV_0580	MUL_4152	cg1141	b0713	LamB/YcsF
ML0336	MSMEG_6046	MAP3775c	MAV_0582	MUL_4148	cg2912		cation ABC transporter, ATP-binding protein, putative
ML0397	MSMEG_4171				cg1412	b3750	ribose transport system permease protein RbsC
ML0398	MSMEG_3095				cg1413	b3751	D-ribose-binding periplasmic protein
ML0405	MSMEG_0723			MUL_5045			hypothetical protein MSMEG_0723
ML0458	MSMEG_6730	MAP2720c	MAV_3494				putative oxidoreductase YdbC
ML0578	MSMEG_3097	MAP1169	MAV_3336	MUL_1836	cg1787	b3956	phosphoenolpyruvate carboxylase
ML0814	MSMEG_1934	MAP3309c	MAV_4156	MUL_2530	cg0882		ATP-binding protein
ML0840	MSMEG_4536	MAP2122	MAV_2053				hypothetical protein MSMEG_4536
ML0841	MSMEG_4537	MAP2121c	MAV_2054				major membrane protein I
ML0842	MSMEG_4538	MAP2120c	MAV_2055			b1680	cysteine desulphurase, SufS
ML0845	MSMEG_4474	MAP2101	MAV_2078	MUL_1233			acyl-CoA oxidase
ML0956	MSMEG_5203	MAP2663c	MAV_1260	MUL_0140	cg1010		DoxX subfamily protein, putative
ML1267	MSMEG_3719	MAP1301	MAV_3179				sodium/calcium exchanger protein
ML1305	MSMEG_6196			MUL_4508	cg1257	b2663	gaba permease
ML1389	MSMEG_3546						hypothetical protein MSMEG_3546
ML1423	MSMEG_5743			MUL_1736			pataatin
ML1795	MSMEG_5611	MAP3268	MAV_4106	MUL_2232			spore protein

ML1796	MSMEG_0542	MAP3269	MAV_4107	MUL_3601	antar domain protein
ML1992	MSMEG_0181	MAP3729			b0368 alpha-ketoglutarate-dependent taurine dioxygenase
ML2013	MSMEG_6487				2-hydroxy-3-carboxy-6-oxo-7-methylocta-2, 4-dienoate decarboxylase
ML2045	MSMEG_3576	MAP1587c	MAV_2842	MUL_3001	cg1012 alpha-amylase 3
ML2088	MSMEG_6312	MAP0344c	MAV_0358		cytochrome P450 107B1
ML2091	MSMEG_5577	MAP0876c	MAV_1051	MUL_4435	fructokinase
ML2158	MSMEG_4778	MAP0798	MAV_0989		putative thiolase
ML2242	MSMEG_6359	MAP0583	MAV_0678	MUL_1503	trypsin domain protein
ML2313	MSMEG_6227	MAP0354c	MAV_0407	MUL_4279	transcriptional regulator, PadR family protein
ML2341	MSMEG_0228		MAV_4804	MUL_3594	adenylate and Guanylate cyclase catalytic domain protein
ML2357	MSMEG_6767				mycocerosic acid synthase
ML2359	MSMEG_4514	MAP2176c	MAV_2010	MUL_3636	Thioesterase domain protein
ML2426	MSMEG_1205				cyclopropane-fatty-acyl-phospholipid synthase 1
ML2459	MSMEG_3538	MAP0135	MAV_0130		b1661 cyclopropane-fatty-acyl-phospholipid synthase 1
ML2497	MSMEG_6556	MAP0930	MAV_1113		putative transcriptional regulator
ML2498	MSMEG_6558				putative enoyl-CoA hydratase
ML2654	MSMEG_6343	MAP3534c	MAV_5179	MUL_4792	hypothetical protein MSMEG_6343
ML2667	MSMEG_1780	MAP2066	MAV_2120		hypothetical protein MSMEG_1780

Shaded rows represent genes that are in clusters in the *M. leprae* TN genome

JLS. We found several COGs not shared with *M. tuberculosis* and *M. bovis* (COGs annotated for catalases, ABC type Mn/Zn transport, chemotaxis, and nitric oxide reductase). Two COGs (predicted as non-ribosomal peptide synthetases and acyltransferases) are considerably larger in *M. avium* subsp. *paratuberculosis* compared to *M. tuberculosis* and *M. bovis*. There are five *M. avium* subsp. *paratuberculosis* COGs absent from all other mycobacteria, *E. coli* and *C. glutamicum* (COGs 784, 2221, 2324, 3319 and 4693), and we found an extra gene for 10 other COGs.

Defining *Mycobacterium ulcerans* Agy99

After TB and leprosy, Buruli ulcer, well known in parts of Australia and Papua New Guinea is emerging as a serious disease in Africa. Therefore efforts to study the causative mycobacteria, *M. ulcerans*, have lead to deciphering its genome sequence, confirming a phylogenetic descent from that of *M. marinum*, a pathogen in ectotherms such as frogs and fish. *M. ulcerans* shows features of gene reduction, restricted host range and niche, and dependence on host for growth reminiscent of *M. leprae* [5], *M. avium* subsp. *paratuberculosis* [7], and other recently evolved bacteria (*Yersinia pestis* [80], *Burkholderia mallei* [81]). COG comparison indicate that the large number of recognized transposases contribute to genome rearrangements and loss. Similar to the analysis performed for *M. tuberculosis* and *M. avium* subsp. *paratuberculosis*, we include a list of COGs enriched in *M. ulcerans* (see Table 6).

A key genomic feature is the acquisition of a plasmid encoding mycolactone, an immunosuppressive cytotoxin macrolide [6]. Also, the glycosylation machinery for generating phenolic glycolipids is lost.

The composition, glycosyl linkages and methyl modifications of phenolic glycolipids are species specific and also antigenic (Table 7) and details can be found in a review by Onwueme et al. [57] and references therein.

The genetic locus involved in the attachment and methylation of the glycosyl residues at the phenol moiety of PGL-Tb of *M. tuberculosis* has been verified [82,83]. The comparison of this locus in the sequenced strain of *M. ulcerans* and *M. marinum* and *M. leprae* is shown in Fig. 4 verifying that the gene for the first glycosylation step is defective, while the genes for the other two glycosyltransferases and methyltransferases are absent. In *M. marinum* however, consistent with the published structure of phenolic glycolipid [6], there is only one gene for the first glycosylation reaction and none for methylation. Also, with regard to the diol lipid backbone, while *M. tuberculosis*, *M. marinum* and *M. leprae* have a ketoreductase to convert a

phthidiodolone to phthiocerol, this gene is a pseudogene in *M. ulcerans*.

The native PGL-I of *M. leprae* or its synthetic glycoconjugate antigen have been used extensively in serological investigations to aid as a tool to detect leprosy infection [84]. Thus far, there has been only one publication regarding cross reactivity of PGL-1 antibodies reactivity to *M. ulcerans* [85]. This and a prior study with references to glycosylated PGL versions in *M. ulcerans* [86] are not compatible with the current genome information. Therefore, strain variants may reconcile these issues of the presence of phenolic glycolipids in *M. ulcerans* [86, 87].

Defining *Mycobacterium leprae* TN

Since the *M. leprae* TN genome was placed in public domain in 2001 [5], and the first annotated version was accessible through the Leproma website, we and others have commented on the genome content of *M. leprae* [5, 42, 88, 89].

It was anticipated that the genome knowledge will solve challenging questions of in vitro growth, and identify virulence factors and explain pathogenesis including nerve damage [5]. The severe gene loss that has left a small repertoire of ~1600 genes explains intracellularism, but there doesn't appear to be any significant new knowledge thus far from the "*M. leprae*" unique genes that can account for its pathology and tissue specificity. In order to gain insight into the peculiar growth properties and adaptations of *M. leprae*, it may be of interest to pay attention to genes not necessarily shared with *M. tuberculosis*, but also those that are present in other species as listed in Table 9. The origin and distribution pattern of these genes is interesting, and tests of functionality can be pursued in one of the tractable species such as *C. glutamicum* and *M. smegmatis*.

The genome has provided some clues for modification of the growth conditions in vitro, however: applying and testing these in practice remains a daunting proposition (http://igs-server.cnrs-mrs.fr/axenic-cgi/generate_table?Mycobacterium+no+off+off), particularly due to the long doubling time (~2 weeks). The doubling time of *M. ulcerans* in vitro, was reduced by the addition of algal extracts in the growth media; a phenotype such as this would be a boost for the study of *M. leprae* in the laboratory.

In this review, we corroborate previous hypotheses that the 'mycobacterial cell wall core' biosynthetic machinery is intact per *in silico* evidence and furthermore we update gene lists for biosynthesis of known cell wall and associated macromolecule biosynthesis and their occurrence in mycobacteria including *M. leprae* (Table 8). Within Table

Fig. 5 Comparison of the number of genes present for each COG function between *M. tuberculosis* (blue bars) and *M. smegmatis* (magenta bars)

8, there are numerous examples of how the elucidation of gene function (as applied in other mycobacterial and related species) has been possible by a process of candidate gene selection *via* careful homology and domain searches followed by experimental "wreck and check" methodologies.

- The search for diagnostic reagents from genome based approaches has been pursued with *M. leprae* specificity as an important criterion [170, 171, 172]. The work of Duthie et al. [173], focused on the search for potential serologically reactive protein antigens prior to testing for the rigorous requirement for leprosy specificity when tested in various endemic populations. Such approaches lead to the identification of new antigens (ML0405, ML 2331) from which novel fusion proteins were designed. While sequence similarity with counterparts in other species is restricted to *M. tuberculosis* and *M. bovis* for ML0405, it extends to *M. avium*, *M. smegmatis* and *M. marinum* for ML2331. In this regard, other candidate gene lists have been put forth, including genes with occurrences in more than one sequenced mycobacteria [92].

- Regarding the evolution and origin of *M. leprae*, Gomez-Valero et al. [43] speculate that *M. leprae* is more closely related to *M. tuberculosis* than to *M. avium* (the analysis was based on *M. avium* subsp. *paratuberculosis*). They propose that a series of gene by gene inactivation events rather than loss of 'blocks of genes' lead to pseudogenes followed by a gradual loss of nucleotides in *M. leprae* and that these processes started after the *M. avium*–*M. tuberculosis* branch split. They note that the majority of the original sequence (~89%) persists in pseudogenes in the extant genome. By identifying orthologs and gene order they reconstruct the genome of the last common ancestor of *M. tuberculosis* and *M. leprae*.

Defining *Mycobacterium smegmatis* mc² 155

M. smegmatis is one among numerous environmental fast growing, avirulent mycobacteria, several of which are being sequenced due to their importance in industrial applications and bioremediation potential. The *M. smegmatis* mc²155

strain has been a model organism exploited extensively for mycobacteria research. Its sequence is larger than *M. tuberculosis* with nearly twice the coding potential. There are many COGs of higher and lower abundance in this species compared to the pathogenic species as depicted in Table 3 and Fig. 3. The relative distribution of the genes assigned to COGs in *M. smegmatis* and *M. tuberculosis* is highlighted in Fig. 5. COGs involved in transport and metabolism of amino acids, carbohydrates, inorganic ions, lipids and secondary metabolites are larger in *M. smegmatis* compared to *M. tuberculosis*. There are additional genes attributed to energy production, and transcription, and those without any specific functional prediction. Together these expanded COGs account for the additional 2402 genes in *M. smegmatis*, compared to *M. tuberculosis*. On the other hand, despite, the larger genome size, the number of genes for certain pathways and functionalities are not enriched or redundant. This type of comparison indicates that there are pathways that can be maintained by a basic minimum number of genes in most mycobacteria.

We found that *M. smegmatis*, *M. avium* subsp. *avium* and *paratuberculosis* share a single gene assigned to the COG category ‘chromatin structure and dynamics’ and an additional gene for RNA processing and modification (Table 3).

- Recently, a comprehensive genome based study by Titgemeyer et al. predicts and validates genes involved in sugar transport in *M. smegmatis* and *M. tuberculosis* [174]. The distinct excess of carbohydrate uptake systems in *M. smegmatis* (28) over that in *M. tuberculosis* (5), reflect saprophytic *versus* host dependent pathogenic lifestyles.

Conclusions and perspectives

Delving into the genomes of mycobacterial and related species has furthered our knowledge of genes associated with common and unique growth requirements, habitats, and cell wall molecules, all applicable; important pathogenic and model microbes which have been applied towards targeted approaches for controlling mycobacterial diseases via vaccines and antimicrobials. In addition, the DNA sequences has allowed for selection of appropriate probes for diagnosis, strain typing, and reconstruction of evolutionary schemes. During the preparation of this article, a comprehensive review of actinobacteria from a genomics perspective has been published [175].

The basis of pathogenicity of mycobacteria is thought to depend completely or in part on members of expanded

gene families such as *esx*, *PE-PPE*, *pks*, *mce* etc. The COG abundance profiles comparisons demonstrate these genes and others that are common or enriched in the three pathogenic species relative to the non-pathogenic species (*E. coli*, *M. avium* subsp. *avium*, *C. glutamicum*). Non-pathogenic species also have orthologs for one or more of these genes, suggesting functions common to metabolism or biosynthesis of macromolecules. However, we found that the majority of the *M. tuberculosis* restricted genes are deemed ‘non-essential’ in experimental models. It is therefore clear that redundant genes (arising from gene duplication events) preclude the precise functional assignment of individual genes, particularly within the large families. Therefore, differential expression and complex genetic interactions are likely to influence pathogenicity and fitness of individual mycobacterial species within changing host milieus.

Further studies of natural populations, particularly of clinical isolates in conjunction with epidemiology are important for a comprehensive understanding of mycobacteria and the nuances of host-bacteria interactions in their native environments and in disease. Though much emphasis is still currently placed on individual open reading frames, the future of genomics, supported by other ‘omics’ may allow for such comprehensive studies in the coming years. In parallel, it is envisioned that bioinformatics will keep pace with the large amount of data (raw genome sequence and metadata) to allow informed gene function predictions requiring minimal laboratory testing and become accessible to the average microbiologist lacking formal training in bioinformatics/computational skills.

Acknowledgements The authors acknowledge the support from grants AI-063457 and NO1-AI-25469 from the National Institute of Allergy and Infectious Diseases, National Institutes of Health.

References

1. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA and Rapp BA (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 28:10–14
2. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA and Wheeler DL (2000) GenBank. *Nucleic Acids Res* 28:15–18
3. Cole ST et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393:537–544
4. Garnier T et al. (2003) The complete genome sequence of *Mycobacterium bovis*. *Proc Natl Acad Sci* 100:7877–7882
5. Brosch R, Gordon SV, Garnier T, Eiglmeier K, Frigui W, Valenti P, Dos Santos S, Duthoy S, Lacroix C, Garcia-Pelayo

- C, Inwald JK, Golby P, Garcia JN, Hewinson RG, Behr MA, Quail MA, Churcher C, Barrell BG, Parkhill J and Cole ST (2007). Genome plasticity of BCG and impact on vaccine efficacy. *Proc Natl Acad Sci* 104:5596–5601
6. http://www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj&ncmd=Retrieve&dopt=Overview&list_uids=18059 or <http://genolist.pasteur.fr/BCGList/>
7. Cole ST et al. (2001) Massive gene decay in the leprosy bacillus” *Nature* 409:1007–1011
8. Stinear TP et al. (2007) Reductive evolution and niche adaptation inferred from the genome of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. *Genome Res* 17:192–200
9. Li L et al. (2005) The complete genome sequence of *Mycobacterium avium subspecies paratuberculosis*. *Proc Natl Acad Sci* 102:12344–12349
10. Nigou J, Gilleron M and Puzo G (2003) Lipoarabinomannans: from structure to biosynthesis. *Biochimie* 85: 153–166
11. Sutcliffe C (2000) Characterisation of a lipomannan lipoglycan from the mycolic acid containing actinomycete *Dietzia maris*. *Antonie Van Leeuwenhoek* 78:195–201
12. Flaherty C and Sutcliffe IC (1999). Identification of a lipoarabinomannan-like lipoglycan in *Gordonia rubropertincta*. *Syst Appl Microbiol* 22:530–533
13. Ma Z, Zhang J and Kong F (2004) Facile synthesis of arabinomannose penta- and decasaccharide fragments of the lipoarabinomannan of the equine pathogen, *Rhodococcus equi*. *Carbohydr Res* 339:1761–1771
14. Flaherty C, Minnikin DE and Sutcliffe IC (1996) A chemotaxonomic study of the lipoglycans of *Rhodococcus rhodnii* N445 (NCIMB 11279). *Zentralbl Bakteriol* 285:11–19
15. Gibson KJ, Gilleron M, Constant P, Brando T, Puzo G, Besra GS and Nigou J (2004) *Tsukamurella paurometabola* lipoglycan, a new lipoarabinomannan variant with pro-inflammatory activity. *J Biol Chem* 279:22973–22982
16. Pakkiri LS and Waechter CJ (2005) Dimannosyldiacylglycerol serves as a lipid anchor precursor in the assembly of the membrane-associated lipomannan in *Micrococcus luteus*. *Glycobiology* 15:291–302
17. Gibson KJ, Gilleron M, Constant P, Sichi B, Puzo G, Besra GS and Nigou J (2005) lipomannan variant with strong TLR-2-dependent pro-inflammatory activity in *Saccharothrix aerocolonigenes*. *J Biol Chem* 280:28347–28356
18. Gibson KJ, Gilleron M, Constant P, Puzo G, Nigou J and Besra GS (2003) Identification of a novel mannose-capped lipoarabinomannan from *Amycolatopsis sulphurea*. *Biochem J* 372:821–829
19. Daffe M, McNeil M and Brennan PJ (1993) Major structural features of the cell wall arabinogalactans of *Mycobacterium*, *Rhodococcus*, and *Nocardia* spp. *Carbohydr Res* 249: 383–398
20. Sutcliffe IC. 1998 Cell envelope composition and organisation in the genus *Rhodococcus*. *Antonie Van Leeuwenhoek* 74:49–58
21. Tropis M, Lemassu A, Vincent V and Daffe M (2005) Structural elucidation of the predominant motifs of the major cell wall arabinogalactan antigens from the borderline species *Tsukamurella paurometabolum* and *Mycobacterium fallax*. *Glycobiology* 15:677–686
22. Barry CE 3rd, Lee RE, Mdluli K, Sampson AE, Schroeder BG, Slayden RA, and Yuan Y (1998) Mycolic acids: structure, biosynthesis and physiological functions. *Prog Lipid Res.* 37:143–179
23. Weinstock GM (2000) Genomics and bacterial pathogenesis. *Emerg Infect Dis* 6:496–504
24. Guilhot C, Gicquel B and Martín C (1992) Temperature-sensitive mutants of the *Mycobacterium* plasmid pAL5000. *FEMS Microbiol Lett* 77:181–186
25. Bardarov S, Kriakov J, Carriere C, Yu S, Vaamonde C, McAdam RA, Bloom BR, Hatfull GF and Jacobs WR Jr (1997) Conditionally replicating mycobacteriophages: a system for transposon delivery to *Mycobacterium tuberculosis*. *Proc Natl Acad Sci* 94:10961–10966
26. Lamichhane G, Zignol M, Blades NJ, Geiman DE, Dougherty A, Grosset J, Broman KW and Bishai WR (2003) A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to *Mycobacterium tuberculosis*. *Proc Natl Acad Sci* 100: 7213–7218
27. Camacho LR, Ensergueix D, Perez E, Gicquel B, and Guilhot C (1999) Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. *Mol Microbiol* 34:257–267
28. Cox JS, Chen B, McNeil M and Jacobs WR Jr (1999). Complex lipid determines tissue-specific replication of *Mycobacterium tuberculosis* in mice. *Nature* 402:79–83
29. Sassetti CM, Boyd DH and Rubin EJ (2001) Comprehensive identification of conditionally essential genes in mycobacteria. *Proc Natl Acad Sci* 98:12712–12717
30. Sassetti CM, Boyd DH and Rubin EJ (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol* 48:77–84
31. Sassetti CM and Rubin EJ (2003) Genetic requirements for mycobacterial survival during infection. *Proc Natl Acad Sci* 100:12989–12994
32. Heifets L. 2004 Mycobacterial infections caused by nontuberculous mycobacteria. *Semin Respir Crit Care Med* 25: 283–295
33. Stackebrandt E, Frederiksen W, Garrity GM, Grimont PA, Kämpfer P, Maiden MC, Nesme X, Rosselló-Mora R, Swings J, Trüper HG, Vauterin L, Ward AC and Whitman WB (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol* 52:1043–1047
34. Snel B, Huynen MA and Dutilh BE (2005) Genome trees and the nature of genome evolution. *Annu Rev Microbiol* 59:191–209
35. Adékambi T and Drancourt M (2004) Dissection of phylogenetic relationships among 19 rapidly growing *Mycobacterium* species by 16S rRNA, hsp65, sodA, recA and rpoB gene sequencing. *Int J Syst Evol Microbiol* 54:2095–2105
36. Devulder G, Pérouse de Montclos M and Flandrois JP (2005) A multigene approach to phylogenetic analysis using the genus *Mycobacterium* as a model. *Int J Syst Evol Microbiol* 55:293–302
37. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, Parsons LM, Pym AS, Samper S, van Soolingen D and Cole ST (2002) A new evolutionary scenario for the

- Mycobacterium tuberculosis* complex. Proc Natl Acad Sci 99:3684–3689
38. Marsollier L, Aubry J, Coutanceau E, André JP, Small PL, Milon G, Legras P, Guadagnini S, Carbonnelle B and Cole ST (2005) Colonization of the salivary glands of *Naucoris cimicoides* by *Mycobacterium ulcerans* requires host plasmacytes and a macrolide toxin, mycolactone. Cell Microbiol 7:935–943
 39. Marsollier L, Séverin T, Aubry J, Merritt RW, Saint André JP, Legras P, Manceau AL, Chauty A, Carbonnelle B and Cole ST (2004) Aquatic snails, passive hosts of *Mycobacterium ulcerans*. Appl Environ Microbiol 70:6296–6298
 40. Marsollier L, Stinear T, Aubry J, Saint André JP, Robert R, Legras P, Manceau AL, Audrain C, Bourdon S, Kouakou H and Carbonnelle B (2004) Aquatic plants stimulate the growth of and biofilm formation by *Mycobacterium ulcerans* in axenic culture and harbor these bacteria in the environment. Appl Environ Microbiol 70:1097–1103
 41. Bannantine JP, Zhang Q, Li LL and Kapur V (2003) Genomic homogeneity between *Mycobacterium avium* subsp. *avium* and *Mycobacterium avium* subsp. *paratuberculosis* belies their divergent growth rates. BMC Microbiol 3:10
 42. Vissa VD and Brennan PJ (2001) The genome of *Mycobacterium leprae*: a minimal mycobacterial gene set. Genome Biol. 2:REVIEWS1023
 43. Gómez-Valero L, Rocha EP, Latorre A and Silva FJ (2007) Reconstructing the ancestor of *Mycobacterium leprae*: the dynamics of gene loss and genome reduction. Genome Res 17:1178–1185
 44. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND and Koonin EV (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29:22–28
 45. Banu S, Honoré N, Saint-Joanis B, Philpott D, Prévost MC and Cole ST (2002) Are the PE-PGRS proteins of *Mycobacterium tuberculosis* variable surface antigens? Mol Microbiol 44:9–19
 46. Ramakrishnan L, Federspiel NA and Falkow S (2000) Granuloma-specific expression of *Mycobacterium* virulence proteins from the glycine-rich PE-PGRS family. Science 288:1436–1439
 47. Marchler-Bauer A, Bryant SH (2004) CD-Search: protein domain annotations on the fly. Nucleic Acids Res 32(Web Server issue):W327–W331
 48. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki C, Liebert CA, Liu C, Lu F, Marchler GH, Mullokandov M, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Yamashita RA, Yin JJ, Zhang D, Bryant SH (2005) CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res 33:D192–D196
 49. Voskuil MI, Schnappinger D, Rutherford R and Liu Y and Schoolnik GK (2004) Regulation of the *Mycobacterium tuberculosis* PE/PPE genes. Tuberculosis 84:256–262
 50. Brennan MJ and Delogu G (2002) The PE multigene family: a 'molecular mantra' for mycobacteria. Trends Microbiol 10: 246–249
 51. Delogu G, Sanguinetti M, Pusceddu C, Bua A, Brennan MJ, Zanetti S and Fadda G. (2006). PE_PGRS proteins are differentially expressed by *Mycobacterium tuberculosis* in host tissues. Microbes Infect 8:2061–2067
 52. Delogu G and Brennan MJ (2001) Comparative immune response to PE and PE_PGRS antigens of *Mycobacterium tuberculosis*. Infect Immun 69:5606–5611
 53. Kumar A, Chandolia A, Chaudhry U, Brahmachari V and Bose M (2005) Comparison of mammalian cell entry operations of mycobacteria: in silico analysis and expression profiling. FEMS Immunol Med Microbiol 43:185–195
 54. Abdallah AM, Gey van Pittius NC, Champion PA, Cox J, Luirink J, Vandenbroucke-Grauls CM, Appelmelk BJ and Bitter W (2007) Type VII secretion--mycobacteria show the way. Nat Rev Microbiol 5:883–891
 55. Fortune SM, Jaeger A, Sarracino DA, Chase MR, Sasseti CM, Sherman DR, Bloom BR, and Rubin EJ (2005) Mutually dependent secretion of proteins required for mycobacterial virulence. Proc Natl Acad Sci 102:10676–10681
 56. Gey van Pittius NC, Sampson SL, Lee H, Kim Y, van Helden PD and Warren RM. 2006 Evolution and expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. BMC Evol Biol 6:95
 57. Onwueme KC, Vos CJ, Zurita J, Ferreras JA and Quadri LE 2005 The dimycocerosate ester polyketide virulence factors of mycobacteria. Prog Lipid Res 44:259–302
 58. DiGiuseppe Champion PA and Cox JS (2007) Protein secretion systems in Mycobacteria. Cell Microbiol 9: 1376–1384
 59. Casali N and Riley LW (2007) A phylogenomic analysis of the *Actinomycetales* mce operons. BMC Genomics 8:60
 60. Marri PR, Bannantine JP and Golding GB (2006) Comparative genomics of metabolic pathways in *Mycobacterium* species: gene duplication, gene decay and lateral gene transfer. FEMS Microbiol Rev 30:906–925
 61. Russell DG (2003) Phagosomes, fatty acids and tuberculosis. Nat Cell Biol 5:776–778
 62. Van der Geize R, Yam K, Heuser T, Wilbrink MH, Hara H, Anderton MC, Sim E, Dijkhuizen L, Davies JE, Mohn WW and Eltis LD (2007) A gene cluster encoding cholesterol catabolism in a soil actinomycete provides insight into *Mycobacterium tuberculosis* survival in macrophages. Proc Natl Acad Sci 104:1947–1952
 63. Kato-Maeda M, Rhee JT, Gingeras TR, Salamon H, Drenkow J, Smittipat N and Small PM (2001) Comparing genomes within the species *Mycobacterium tuberculosis*. Genome Res 11:547–554
 64. Tsolaki AG, Hirsh AE, DeRiemer K, Enciso JA, Wong MZ, Hannan M, Goguet de la Salmoniere YO, Aman K, Kato-Maeda M and Small PM (2004) Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains. Proc Natl Acad Sci 101: 4865–4870
 65. Ren H, Dover LG, Islam ST, Alexander DC, Chen JM, Besra GS and Liu J (2007) Identification of the lipooligosaccharide biosynthetic gene cluster from *Mycobacterium marinum*. Mol Microbiol 63:1345–1359
 66. Rousseau C, Sirakova TD, Dubey VS, Bordat Y, Kolattukudy PE, Gicquel B and Jackson M (2003) Virulence attenuation of two Mas-like polyketide synthase mutants of *Mycobacterium tuberculosis*. Microbiology 149:1837–1847

67. Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, DeBoy R, Dodson R, Gwinn M, Haft D, Hickey E, Kolonay JF, Nelson WC, Umayam LA, Ermolaeva M, Salzberg SL, Delcher A, Utterback T, Weidman J, Khouri H, Gill J, Mikula A, Bishai W, Jacobs Jr WR Jr, Venter JC and Fraser CM (2002) Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J Bacteriol* 184:5479–5490
68. Viana-Niero C, de Haas PE, van Soolingen D and Leão SC (2004) Analysis of genetic polymorphisms affecting the four phospholipase C (plc) genes in *Mycobacterium tuberculosis* complex clinical isolates. *Microbiology* 150:967–978
69. Yang Z, Yang D, Kong Y, Zhang L, Marrs CF, Foxman B, Bates JH, Wilson F and Cave MD (2005) Clinical relevance of *Mycobacterium tuberculosis plcD* gene mutations. *Am J Respir Crit Care Med* 171:1436–1442
70. van Soolingen D, Qian L, de Haas PE, Douglas JT, Traore H, Portaels F, Qing HZ, Enkhsaikan D, Nymadawa P and van Embden JD (1995) Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of east Asia. *J Clin Microbiol* 33:3234–3238
71. European Concerted Action on New Generation Genetic Markers and Techniques for the Epidemiology and Control of Tuberculosis (2006) Beijing/W genotype *Mycobacterium tuberculosis* and drug resistance. *Emerg Infect Dis* 12: 736–743
72. Abebe F and Bjune G (2006) The emergence of Beijing family genotypes of *Mycobacterium tuberculosis* and low-level protection by bacille Calmette-Guérin (BCG) vaccines: is there a link? *Clin Exp Immunol* 145:389–397
73. Bifani PJ, Mathema B, Kurepina NE and Kreiswirth BN (2002) Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains. *Trends Microbiol* 10:45–52
74. Kong Y, Cave MD, Zhang L, Foxman B, Marrs CF, Bates JH and Yang ZH (2007) Association between *Mycobacterium tuberculosis* Beijing/W lineage strain infection and extrathoracic tuberculosis: Insights from epidemiologic and clinical characterization of the three principal genetic groups of *M. tuberculosis* clinical isolates. *J Clin Microbiol* 45:409–414
75. Dormans J, Burger M, Aguilar D, Hernandez-Pando R, Kremer K, Roholl P, Arend SM and van Soolingen D (2004) Correlation of virulence, lung pathology, bacterial load and delayed type hypersensitivity responses after infection with different *Mycobacterium tuberculosis* genotypes in a BALB/c mouse model. *Clin Exp Immunol* 137:460–468
76. Turenne CY, Wallace R Jr and Behr MA (2007) *Mycobacterium avium* in the postgenomic era. *Clin Microbiol Rev* 20: 205–229
77. Semret M, Turenne CY, de Haas P, Collins DM and Behr MA (2006) Differentiating host-associated variants of *Mycobacterium avium* by PCR for detection of large sequence polymorphisms. *J Clin Microbiol* 44:881–887
78. Motiwala AS, Li L, Kapur V and Sreevatsan S (2006) Current understanding of the genetic diversity of *Mycobacterium avium subsp. paratuberculosis*. *Microbes Infect* 8:1406–1418
79. Danelishvili L, Wu M, Stang B, Harrieff M, Cirillo SL, Cirillo JD, Bildfell R, Arbogast B and Bermudez LE (2007) Identification of *Mycobacterium avium* pathogenicity island important for macrophage and amoeba infection. *Proc Natl Acad Sci* 104:11038–11043
80. Wren BW (2003) The yersiniae—a model genus to study the rapid evolution of bacterial pathogens. *Nat Rev Microbiol* 1:55–64
81. Kim HS, Schell MA, Yu Y, Ulrich RL, Sarria SH, Nierman WC and DeShazer D (2005) Bacterial genome adaptation to niches: divergence of the potential virulence genes in three *Burkholderia* species of different survival strategies. *BMC Genomics* 6:174
82. Pérez E, Constant P, Lemassu A, Laval F, Daffé M and Guilhot C (2004) Characterization of three glycosyltransferases involved in the biosynthesis of the phenolic glycolipid antigens from the *Mycobacterium tuberculosis* complex. *J Biol Chem* 279:42574–42583
83. Pérez E, Constant P, Laval F, Lemassu A, Lanéele MA, Daffé M and Guilhot C (2004) Molecular dissection of the role of two methyltransferases in the biosynthesis of phenolglycolipids and phthiocerol dimycoserolate in the *Mycobacterium tuberculosis* complex. *J Biol Chem* 279: 42584–42592
84. Cho SN, Yanagihara DL, Hunter SW, Gelber RH and Brennan PJ (1983) Serological specificity of phenolic glycolipid I from *Mycobacterium leprae* and use in serodiagnosis of leprosy. *Infect Immun* 41:1077–1083
85. Mwanatambwe M, Yajima M, Etuaful S, Fukunishi Y, Suzuki K, Asiedu K, Yamada N and Asanao G (2002) Phenolic glycolipid-1 (PGL-1) in Buruli ulcer lesions. First demonstration by immuno-histochemistry. *Int J Lepr Other Mycobact Dis* 70:201–205
86. Daffé M, Varnerot A and Lévy-Frébault VV (1992) The phenolic mycoside of *Mycobacterium ulcerans*: structure and taxonomic implications. *J Gen Microbiol* 138:131–137
87. Käser M, Rondini S, Naegeli M, Stinear T, Portaels F, Certa U and Pluschke G (2007) Evolution of two distinct phylogenetic lineages of the emerging human pathogen *Mycobacterium ulcerans*. *BMC Evol Biol* 7:177
88. Brennan PJ and Vissa VD (2001) Genomic evidence for the retention of the essential mycobacterial cell wall in the otherwise defective *Mycobacterium leprae*. *Lepr Rev* 72: 415–428
89. Eiglmeier K, Parkhill J, Honoré N, Garnier T, Tekaiia F, Telenti A, Klatser P, James KD, Thomson NR, Wheeler PR, Churcher C, Harris D, Mungall K, Barrell BG and Cole ST (2001) The decaying genome of *Mycobacterium leprae*. *Lepr Rev* 72:387–398
90. Bailey AM, Mahapatra S, Brennan PJ and Crick DC (2002) Identification, cloning, purification, and enzymatic characterization of *Mycobacterium tuberculosis* 1-deoxy-D-xylulose 5-phosphate synthase. *Glycobiology* 12:813–820
91. Dhiman RK, Schaeffer ML, Bailey AM, Testa CA, Scherman H and Crick DC (2005) 1-Deoxy-D-xylulose 5-phosphate reductoisomerase (IspC) from *Mycobacterium tuberculosis*: towards understanding mycobacterial resistance to fosmidomycin. *J Bacteriol* 187:8395–8402
92. Eoh H, Brown AC, Buetow L, Hunter WN, Parish T, Kaur D, Brennan PJ and Crick DC (2007) Characterization of the *Mycobacterium tuberculosis* 4-diphosphocytidyl-2-C-methyl-D-erythritol synthase: potential for drug development. *J Bacteriol* 189:8922–8927
93. Buetow L, Brown AC, Parish T and Hunter WN (2007) The structure of *Mycobacteria* 2C-methyl-D-erythritol-2,4-cy-

- clodiphosphate synthase, an essential enzyme, provides a platform for drug discovery. *BMC Struct Biol* 7:68
94. Schulbach MC, Brennan PJ and Crick DC (2000) Identification of a short (C15) chain Z-isoprenyl diphosphate synthase and a homologous long (C50) chain isoprenyl diphosphate synthase in *Mycobacterium tuberculosis*. *J Biol Chem* 275:22876–22881
 95. Dhiman RK, Schulbach MC, Mahapatra S, Baulard AR, Vissa V, Brennan PJ and Crick DC (2004) Identification of a novel class of omega,E,E-farnesyl diphosphate synthase from *Mycobacterium tuberculosis*. *J Lipid Res* 45: 1140–1147
 96. De Smet KA, Kempell KE, Gallagher A, Duncan K and Young DB (1999) Alteration of a single amino acid residue reverses fosfomycin resistance of recombinant MurA from *Mycobacterium tuberculosis*. *Microbiology* 145 (Pt 11): 3177–3184
 97. Mahapatra S, Crick DC and Brennan PJ (2000) Comparison of the UDP-N-acetylmuramate:L-alanine ligase enzymes from *Mycobacterium tuberculosis* and *Mycobacterium leprae*. *J Bacteriol* 182:6827–6830
 98. Mahapatra S, Yagi T, Belisle JT, Espinosa BJ, Hill PJ, McNeil MR, Brennan PJ and Crick DC (2005) Mycobacterial lipid II is composed of a complex mixture of modified muramyl and peptide moieties linked to decaprenyl phosphate. *J Bacteriol* 187:2747–2757
 99. Bhakta S and Basu J (2002) Overexpression, purification and biochemical characterization of a class A high-molecular-mass penicillin-binding protein (PBP), PBP1* and its soluble derivative from *Mycobacterium tuberculosis*. *Biochem J* 361:635–669
 100. Ma Y, Stern RJ, Scherman MS, Vissa VD, Yan W, Jones VC, Zhang F, Franzblau SG, Lewis WH and McNeil MR (2001) Drug targeting *Mycobacterium tuberculosis* cell wall synthesis: genetics of dTDP-rhamnose synthetic enzymes and development of a microtiter plate-based screen for inhibitors of conversion of dTDP-glucose to dTDP-rhamnose. *Antimicrob Agents Chemother* 45: 1407–1416
 101. Weston A, Stern RJ, Lee RE, Nassau PM, Monsey D, Martin SL, Scherman MS, Besra GS, Duncan K and McNeil MR (1997) Biosynthetic origin of mycobacterial cell wall galactofuranosyl residues. *Tuber Lung Dis* 78: 123–131
 102. Sanders DA, Staines AG, McMahon SA, McNeil MR, Whitfield C and Naismith JH (2001) UDP-galactopyranose mutase has a novel structure and mechanism. *Nat Struct Biol* 8:858–863.
 103. Mikusová K, Huang H, Yagi T, Holsters M, Vereecke D, D’Haeze W, Scherman MS, Brennan PJ, McNeil MR and Crick DC (2005) Decaprenylphosphoryl arabinofuranose, the donor of the D-arabinofuranosyl residues of mycobacterial arabinan, is formed via a two-step epimerization of decaprenylphosphoryl ribose. *J Bacteriol* 187:8020–8025
 104. Mills JA, Motichka K, Jucker M, Wu HP, Uhlik BC, Stern RJ, Scherman MS, Vissa VD, Pan F, Kundu M, Ma YF and McNeil M (2004) Inactivation of the mycobacterial rhamnosyltransferase, which is needed for the formation of the arabinogalactan-peptidoglycan linker, leads to irreversible loss of viability. *J Biol Chem* 279:43540–43546
 105. Kremer L, Dover LG, Morehouse C, Hitchin P, Everett M, Morris HR, Dell A, Brennan PJ, McNeil MR, Flaherty C, Duncan K and Besra GS (2001) Galactan biosynthesis in *Mycobacterium tuberculosis*. Identification of a bifunctional UDP-galactofuranosyltransferase. *J Biol Chem* 276: 26430–26440
 106. Mikusova K, Belanova M, Kordulakova J, Honda K, McNeil MR, Mahapatra S, Crick DC and Brennan PJ (2006) Identification of a novel galactosyl transferase involved in biosynthesis of the mycobacterial cell wall. *J Bacteriol* 188: 6592–6598
 107. Alderwick LJ, Seidel M, Sahm H, Besra GS and Eggeling L (2006) Identification of a novel arabinofuranosyltransferase (AftA) involved in cell wall arabinan biosynthesis in *Mycobacterium tuberculosis*. *J Biol Chem* 281:15653–15661
 108. Belanger AE, Besra GS, Ford ME, Mikusová K, Belisle JT, Brennan PJ and Inamine JM (1996) The embAB genes of *Mycobacterium avium* encode an arabinosyl transferase involved in cell wall arabinan biosynthesis that is the target for the antimycobacterial drug ethambutol. *Proc Natl Acad Sci* 93:11919–11924
 109. Seidel M, Alderwick LJ, Birch HL, Sahm H, Eggeling L and Besra GS (2007) Identification of a novel arabinofuranosyltransferase AftB involved in a terminal step of cell wall arabinan biosynthesis in *Corynebacteriaceae*, such as *Corynebacterium glutamicum* and *Mycobacterium tuberculosis*. *J Biol Chem* 282:14729–14740
 110. Fernandes ND and Kolattukudy PE (1996) Cloning, sequencing and characterization of a fatty acid synthase-encoding gene from *Mycobacterium tuberculosis* var. bovis BCG. *Gene* 170:95–99
 111. Daniel J, Oh TJ, Lee CM and Kolattukudy PE (2007) AccD6, a member of the Fas II locus, is a functional carboxyltransferase subunit of the acyl-coenzyme A carboxylase in *Mycobacterium tuberculosis*. *J Bacteriol* 189:911–917
 112. Mdululi K, Slayden RA, Zhu Y, Ramaswamy S, Pan X, Mead D, Crane DD, Musser JM and Barry CE 3rd. (1998) Inhibition of a *Mycobacterium tuberculosis* beta-ketoacyl ACP synthase by isoniazid. *Science* 280: 1607–1610
 113. Kremer L, Nampoothiri KM, Lesjean S, Dover LG, Graham S, Betts J, Brennan PJ, Minnikin DE, Loch C and Besra GS (2001) Biochemical characterization of acyl carrier protein (AcpM) and malonyl-CoA:AcpM transacylase (mtFabD), two major components of *Mycobacterium tuberculosis* fatty acid synthase II. *J Biol Chem* 276:27967–27974
 114. Choi KH, Kremer L, Besra GS and Rock CO (2000) Identification and substrate specificity of beta-ketoacyl (acyl carrier protein) synthase III (mtFabH) from *Mycobacterium tuberculosis*. *J Biol Chem* 275:28201–28207
 115. Schaeffer ML, Agnihotri G, Volker C, Kallender H, Brennan PJ and Lonsdale JT (2001) Purification and biochemical characterization of the *Mycobacterium tuberculosis* beta-ketoacyl-acyl carrier protein synthases KasA and KasB. *J Biol Chem* 276:47029–47037
 116. Marrakchi H, Ducasse S, Labesse G, Montrozier H, Margeat E, Emorine L, Charpentier X, Daffé M and Quémar A (2002) MabA (FabG1), a *Mycobacterium tuberculosis* protein involved in the long-chain fatty acid elongation system FAS-II. *Microbiology* 148:951–960

117. Banerjee A, Dubnau E, Quemard A, Balasubramanian V, Um KS, Wilson T, Collins D, de Lisle G and Jacobs WR Jr (1994) inhA, a gene encoding a target for isoniazid and ethionamide in *Mycobacterium tuberculosis*. *Science* 263: 227–230
118. Yuan Y, Lee RE, Besra GS, Belisle JT and Barry C.E 3rd (1995) Identification of a gene involved in the biosynthesis of cyclopropanated mycolic acids in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci* 6630–6634
119. Glickman MS, Cahill SM and Jacobs WR Jr. 2001. The *Mycobacterium tuberculosis* cmaA2 gene encodes a mycolic acid trans-cyclopropane synthetase. *J Biol Chem* 276:2228–2233
120. Yuan Y and Barry CE 3rd (1996) A common mechanism for the biosynthesis of methoxy and cyclopropyl mycolic acids in *Mycobacterium tuberculosis*. *Proc Natl Acad Sci* 93:12828–12833
121. Glickman MS (2003) The mmaA2 gene of *Mycobacterium tuberculosis* encodes the distal cyclopropane synthase of the alpha-mycolic acid. *J Biol Chem* 278:7844–7849
122. Laval F, Haïtes R, Movahedzadeh F, Lemassu A, Wong CY, Stoker N, Billman-Jacobe H and Daffé M (2008) Investigating the function of the putative mycolic acid methyltransferase UmaA: divergence between the *Mycobacterium smegmatis* and *Mycobacterium tuberculosis* proteins. *J Biol Chem* 283:1419–1427
123. Glickman MS, Cox JS and Jacobs WR Jr (2000) A novel mycolic acid cyclopropane synthetase is required for cording, persistence, and virulence of *Mycobacterium tuberculosis*. *Mol Cell* 5:717–727
124. Dyer DH, Lyle KS, Rayment I and Fox BG (2005) X-ray structure of putative acyl-ACP desaturase DesA2 from *Mycobacterium tuberculosis* H37Rv. *Protein Sci* 14: 1508–1517
125. Portevin D, de Sousa-D'Auria C, Montrozier H, Houssin C, Stella A, Lanéeelle MA, Bardou F, Guilhot C and Daffé M (2005) The acyl-AMP ligase FadD32 and AccD4-containing acyl-CoA carboxylase are required for the synthesis of mycolic acids and essential for mycobacterial growth: identification of the carboxylation product and determination of the acyl-CoA carboxylase components. *J Biol Chem* 280:8862–8874
126. Lin TW, Melgar MM, Kurth D, Swamidass SJ, Purdon J, Tseng T, Gago G, Baldi P, Gramajo H and Tsai SC (2006) Structure-based inhibitor design of AccD5, an essential acyl-CoA carboxylase carboxyltransferase domain of *Mycobacterium tuberculosis*. *Proc Natl Acad Sci* 103: 3072–3077
127. Portevin D, De Sousa-D'Auria C, Houssin C, Grimaldi C, Chami M, Daffé M, and Guilhot C (2004) A polyketide synthase catalyzes the last condensation step of mycolic acid biosynthesis in mycobacteria and related organisms. *Proc Natl Acad Sci* 101:314–319.
128. Belisle JT, Vissa VD, Sievert T, Takayama K, Brennan PJ and Besra GS (1997) Role of the major antigen of *Mycobacterium tuberculosis* in cell wall biogenesis. *Science* 276:1420–1422
129. Azad AK, Sirakova TD, Rogers LM and Kolattukudy PE (1996) Targeted replacement of the mycocerosic acid synthase gene in *Mycobacterium bovis* BCG produces a mutant that lacks mycosides. *Proc Natl Acad Sci*. 93: 4787–4792
130. Camacho LR, Constant P, Raynaud C, Laneelle MA, Triccas JA, Gicquel B, Daffé M and Guilhot C (2001) Analysis of the phthiocerol dimycocerosate locus of *Mycobacterium tuberculosis*. Evidence that this lipid is involved in the cell wall permeability barrier. *J Biol Chem* 276:19845–19854
131. Stadthagen G, Korduláková J, Griffin R, Constant P, Bottoová I, Barilone N, Gicquel B, Daffé M and Jackson M (2005) p-Hydroxybenzoic acid synthesis in *Mycobacterium tuberculosis*. *J Biol Chem* 280:40699–40706
132. Azad AK, Sirakova TD, Fernandes ND and Kolattukudy PE (1997) Gene knockout reveals a novel gene cluster for the synthesis of a class of cell wall lipids unique to pathogenic mycobacteria. *J Biol Chem* 272:16741–16745
133. Choudhuri BS, Bhakta S, Barik R, Basu J, Kundu M and Chakrabarti P (2002) Overexpression and functional characterization of an ABC (ATP-binding cassette) transporter encoded by the genes drrA and drrB of *Mycobacterium tuberculosis*. *Biochem J* 367:279–285
134. Onwueme KC, Ferreras JA, Buglino J, Lima CD and Quadri LE (2004) Mycobacterial polyketide-associated proteins are acyltransferases: proof of principle with *Mycobacterium tuberculosis* PapA5. *Proc Natl Acad Sci* 101:4608–4613
135. Constant P, Perez E, Malaga W, Lanéeelle MA, Saurel O, Daffé M and Guilhot C. (2002) Role of the pks15/1 gene in the biosynthesis of phenolglycolipids in the *Mycobacterium tuberculosis* complex. Evidence that all strains synthesize glycosylated p-hydroxybenzoic methyl esters and that strains devoid of phenolglycolipids harbor a frameshift mutation in the pks15/1 gene. *J Biol Chem* 277: 38148–38158
136. Hotter GS, Wards BJ, Mouat P, Besra GS, Gomes J, Singh M, Bassett S, Kawakami P, Wheeler PR, de Lisle GW and Collins DM (2006) Transposon mutagenesis of Mb0100 at the ppe1-nrp locus in *Mycobacterium bovis* disrupts phthiocerol dimycocerosate (PDIM) and glycosylphenol-PDIM biosynthesis, producing an avirulent strain with vaccine properties at least equal to those of *M. bovis* BCG. *J Bacteriol* 187:2267–2277
137. Sulzenbacher G, Canaan S, Bordat Y, Neyrolles O, Stadthagen G, Roig-Zambroni V, Rauzier J, Maurin D, Laval F, Daffé M, Cambillau C, Gicquel B, Bourne Y and Jackson M (2006) LppX is a lipoprotein required for the translocation of phthiocerol dimycocerosates to the surface of *Mycobacterium tuberculosis*. *EMBO J* 25:1436–1444
138. Siméone R, Constant P, Guilhot C, Daffé M and Chalut C (2007) Identification of the missing trans-acting enoyl reductase required for phthiocerol dimycocerosate and phenolglycolipid biosynthesis in *Mycobacterium tuberculosis*. *J Bacteriol* 189:4597–4602
139. Siméone R, Constant P, Malaga W, Guilhot C, Daffé M and Chalut C (2007) Molecular dissection of the biosynthetic relationship between phthiocerol and phthiodiolone dimycocerosates and their critical role in the virulence and permeability of *Mycobacterium tuberculosis*. *FEBS J* 274: 1957–1969
140. Sirakova TD, Dubey VS, Cynamon MH and Kolattukudy PE (2003) Attenuation of *Mycobacterium tuberculosis* by

- disruption of a mas-like gene or a chalcone synthase-like gene, which causes deficiency in dimycocerosyl phthiocerol synthesis. *J Bacteriol* 185:2999–3008
141. Sirakova TD, Dubey VS, Kim HJ, Cynamon MH and Kolattukudy PE (2003) The largest open reading frame (pks12) in the *Mycobacterium tuberculosis* genomes involved in pathogenesis and dimycocerosyl phthiocerol synthesis. *Infect Immun* 71:3794–3801
 142. Dubey VS, Sirakova TD, Cynamon MH and Kolattukudy PE (2003) Biochemical function of msl5 (pks8 plus pks17) in *Mycobacterium tuberculosis* H37Rv: biosynthesis of monomethyl branched unsaturated fatty acids. *J Bacteriol* 185:4620–4625
 143. Gurucha SS, Baulard AR, Kremer L, Loch C, Moody DB, Muhlecker W, Costello CE, Crick DC, Brennan PJ and Besra GS. 2002. Ppm1, a novel polyprenol monophosphomannose synthase from *Mycobacterium tuberculosis*. *Biochem J* 365:441–450
 144. Jackson M, Crick DC and Brennan PJ (2000) Phosphatidylinositol is an essential phospholipid of mycobacteria. *J Biol Chem* 275:30092–30099
 145. Korduláková J, Gilleron M, Puzo G, Brennan PJ, Gicquel B, Mikusová K and Jackson M. 2003. Identification of the required acyltransferase step in the biosynthesis of the phosphatidylinositol mannosides of mycobacterium species. *J Biol Chem* 278:36285–36295
 146. Korduláková J, Gilleron M, Mikusova K, Puzo G, Brennan PJ, Gicquel B and Jackson M. 2002. Definition of the first mannosylation step in phosphatidylinositol mannoside synthesis. PimA is essential for growth of mycobacteria. *J Biol Chem* 277:31335–31344
 147. Schaeffer ML, Khoo KH, Besra GS, Chatterjee D, Brennan PJ, Belisle JT and Inamine JM (1999) The pimB gene of *Mycobacterium tuberculosis* encodes a mannosyltransferase involved in lipoarabinomannan biosynthesis. *J Biol Chem* 274:31625–31631
 148. Tatituri RV, Illarionov PA, Dover LG, Nigou J, Gilleron M, Hitchen P, Krumbach K, Morris HR, Spencer N, Dell A, Eggeling L and Besra GS (2007) Inactivation of *Corynebacterium glutamicum* NCgl0452 and the role of MgtA in the biosynthesis of a novel mannosylated glycolipid involved in lipomannan biosynthesis. *J Biol Chem* 282:4561–4572
 149. Kremer L, Gurucha SS, Bifani P, Hitchen PG, Baulard A, Morris HR, Dell A, Brennan PJ and Besra GS (2002) Characterization of a putative alpha-mannosyltransferase involved in phosphatidylinositol trimannoside biosynthesis in *Mycobacterium tuberculosis*. *Biochem J* 363:437–447
 150. Morita YS, Sena CB, Waller RF, Kurokawa K, Sernee MF, Nakatani F, Haites RE, Billman-Jacobe H, McConville MJ, Maeda Y and Kinoshita T (2006) PimE is a polyprenol-phosphate-mannose-dependent mannosyltransferase that transfers the fifth mannose of phosphatidylinositol mannoside in mycobacteria. *J Biol Chem* 281:25143–25155
 151. Kaur D, Berg S, Dinadayala P, Gicquel B, Chatterjee D, McNeil MR, Vissa VD, Crick DC, Jackson M and Brennan PJ (2006) Biosynthesis of mycobacterial lipoarabinomannan: role of a branching mannosyltransferase. *Proc Natl Acad Sci* 103:13664–13669
 152. Zhang N, Torrelles JB, McNeil MR, Escuyer VE, Khoo KH, Brennan PJ and Chatterjee D (2003) The Emb proteins of mycobacteria direct arabinosylation of lipoarabinomannan and arabinogalactan via an N-terminal recognition region and a C-terminal synthetic region. *Mol Microbiol* 50:69–76
 153. Jeevarajah D, Patterson JH, McConville MJ and Billman-Jacobe H (2002) Modification of glycopeptidolipids by an O-methyltransferase of *Mycobacterium smegmatis*. 148: 3079–3087
 154. Jeevarajah D, Patterson JH, Taig E, Sargeant T, McConville MJ and Billman-Jacobe H (2004) Methylation of GPLs in *Mycobacterium smegmatis* and *Mycobacterium avium*. *J Bacteriol* 186:6792–6799
 155. Patterson JH, McConville MJ, Haites RE, Coppel RL and Billman-Jacobe H (2000) Identification of a methyltransferase from *Mycobacterium smegmatis* involved in glycopeptidolipid synthesis. *J Biol Chem* 275:24900–24906
 156. Miyamoto Y, Mukai T, Nakata N, Maeda Y, Kai M, Naka T, Yano I and Makino M (2006) Identification and characterization of the genes involved in glycosylation pathways of mycobacterial glycopeptidolipid biosynthesis. *J Bacteriol* 188:86–95
 157. Recht J and Kolter R (2001) Glycopeptidolipid acetylation affects sliding motility and biofilm formation in *Mycobacterium smegmatis*. *J Bacteriol*. 183:5718–5724
 158. Billman-Jacobe H, McConville MJ, Haites RE, Kovacevic S and Coppel RL (1999) Identification of a peptide synthetase involved in the biosynthesis of glycopeptidolipids of *Mycobacterium smegmatis*. *Mol Microbiol* 33:1244–1253
 159. Sonden B, Kocincova D, Deshayes C, Euphrasie D, Rhayat L, Laval F, Frehel C, Daffe M, Etienne G and Reyrat JM (2005) Gap, a mycobacterial specific integral membrane protein, is required for glycolipid transport to the cell surface. *Mol. Microbiol* 58:426–440
 160. Trivedi OA, Arora P, Sridharan V, Tickoo R, Mohanty D and Gokhale RS (2004) Enzymatic activation and transfer of fatty acids as acyl-adenylates in mycobacteria. *Nature* 428:441–445
 161. Deshayes C, Laval F, Montrozier H, Daffe M, Etienne G and Reyrat JM (2005) A Glycosyltransferase Involved in Biosynthesis of Triglycosylated Glycopeptidolipids in *Mycobacterium smegmatis*: Impact on Surface Properties. *J. Bacteriol* 187:7283–7291
 162. Miyamoto Y, Mukai T, Nakata N, Maeda Y, Kai M, Naka T, Yano I and Makino M (2006) Identification and characterization of the genes involved in glycosylation pathways of mycobacterial glycopeptidolipid biosynthesis. *J. Bacteriol*. 188:86–95
 163. Sirakova TD, Thirumala AK, Dubey VS, Sprecher H and Kolattukudy PE (2001) The *Mycobacterium tuberculosis* pks2 gene encodes the synthase for the hepta- and octamethyl-branched fatty acids required for sulfolipid synthesis. *J Biol Chem* 276:16833–16839
 164. Converse SE, Mougous JD, Leavell MD, Leary JA, Bertozzi CR and Cox JS (2003) MmpL8 is required for sulfolipid-1 biosynthesis and *Mycobacterium tuberculosis* virulence. *Proc Natl Acad Sci* 100:6121–6126
 165. Kumar P, Schelle MW, Jain M, Lin FL, Petzold CJ, Leavell MD, Leary JA, Cox JS and Bertozzi CR (2007) PapA1 and PapA2 are acyltransferases essential for the biosynthesis of the *Mycobacterium tuberculosis* virulence factor sulfolipid-1. *Proc Natl Acad Sci* 104:11221–11226

166. Mougous JD, Petzold CJ, Senaratne RH, Lee DH, Akey DL, Lin FL, Munchel SE, Pratt MR, Riley LW, Leary JA, Berger JM and Bertozzi CR (2004) Identification, function and structure of the mycobacterial sulfotransferase that initiates sulfolipid-1 biosynthesis. *Nat Struct Mol Biol* 11: 721–729
167. Tzvetkov, M, Klopprogge C, Zelder O and Liebl W (2003) Genetic dissection of trehalose biosynthesis in *Corynebacterium glutamicum*: inactivation of trehalose production leads to impaired growth and an altered cell wall lipid composition. *Microbiology* 149:1659–1673
168. Wolf A, Kramer R and Morbach S (2003) Three pathways for trehalose metabolism in *Corynebacterium glutamicum* ATCC13032 and their significance in response to osmotic stress. *Mol Microbiol* 49:1119–1134
169. Woodruff PJ, Carlson BL, Siridechadilok B, Pratt MR, Williams SJ, and Bertozzi CR (2004) Trehalose is required for growth of *Mycobacterium smegmatis*. *J Biol Chem* 279: 28835–28843
170. Spencer JS, Dockrell HM, Kim HJ, Marques MA, Williams DL, Martins MV, Martins ML, Lima MC, Sarno EN, Pereira GM, Matos H, Fonseca LS, Sampaio EP, Ottenhoff TH, Geluk A, Cho SN, Stoker NG, Cole ST, Brennan PJ and Pessolani MC (2005) Identification of specific proteins and peptides in *Mycobacterium leprae* suitable for the selective diagnosis of leprosy. *J Immunol* 175: 7930–7938
171. Aráoz R, Honoré N, Cho S, Kim JP, Cho SN, Monot M, Demangel C, Brennan PJ and Cole ST (2006) Antigen discovery: a postgenomic approach to leprosy diagnosis. *Infect Immun* 74:175–82
172. Geluk A, Klein MR, Franken KL, van Meijgaarden KE, Wieles B, Pereira KC, Bühner-Sékula S, Klatser PR, Brennan PJ, Spencer JS, Williams DL, Pessolani MC, Sampaio EP and Ottenhoff TH (2005) Postgenomic approach to identify novel *Mycobacterium leprae* antigens with potential to improve immunodiagnosis of infection. *Infect Immun* 73:5636–5644
173. Duthie MS, Goto W, Ireton GC, Reece ST, Cardoso LP, Martelli CM, Stefani MM, Nakatani M, de Jesus RC, Netto EM, Balagon MV, Tan E, Gelber RH, Maeda Y, Makino M, Hoft D and Reed SG (2007) Use of protein antigens for early serological diagnosis of leprosy. *Clin Vaccine Immunol* 14:1400–1408
174. Titgemeyer F, Amon J, Parche S, Mahfoud M, Bail J, Schlicht M, Rehm N, Hillmann D, Stephan J, Walter B, Burkovski A and Niederweis M (2007) A genomic view of sugar transport in *Mycobacterium smegmatis* and *Mycobacterium tuberculosis*. *J Bacteriol* 189:5903–5915
175. Ventura M, Canchaya C, Tauch A, Chandra G, Fitzgerald GF, Chater KF, van Sinderen D (2007) Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiol Mol Biol Rev* 71:495–548