# Massive gene acquisitions in *Mycobacterium indicus pranii* provide a perspective on mycobacterial evolution

**Vikram Saini[1,2], Saurabh Raghuvanshi[2], Jitendra P. Khurana[2], Niyaz Ahmed[3], Seyed E. Hasnain[4,5], Akhilesh K. Tyagi[2,6] and Anil K. Tyagi[1,*]**

[1]Department of Biochemistry, [2]Interdisciplinary Centre for Plant Genomics and Department of Plant Molecular Biology, University of Delhi South Campus, New Delhi 110021, India, [3]Pathogen Biology Laboratory, Department of Biotechnology, School of Life Sciences, University of Hyderabad, Professor C.R. Rao Road, Hyderabad, Andhra Pradesh 500046, India, [4]School of Biological Sciences, Indian Institute of Technology, New Delhi 110016, India, [5]Institute of Life Sciences, University of Hyderabad Campus, Professor C.R. Rao Road, Hyderabad 500046, Andhra Pradesh, India and [6]National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi 110067, India

## ABSTRACT

**Understanding the evolutionary and genomic mechanisms responsible for turning the soil-derived saprophytic mycobacteria into lethal intracellular pathogens is a critical step towards the development of strategies for the control of mycobacterial diseases. In this context, *Mycobacterium indicus pranii (MIP)* is of specific interest because of its unique immunological and evolutionary significance. Evolutionarily, it is the progenitor of opportunistic pathogens belonging to *M. avium* complex and is endowed with features that place it between saprophytic and pathogenic species. Herein, we have sequenced the complete *MIP* genome to understand its unique life style, basis of immunomodulation and habitat diversification in mycobacteria. As a case of massive gene acquisitions, 50.5% of *MIP* open reading frames (ORFs) are laterally acquired. We show, for the first time for *Mycobacterium*, that *MIP* genome has mosaic architecture. These gene acquisitions have led to the enrichment of selected gene families critical to *MIP* physiology. Comparative genomic analysis indicates a higher antigenic potential of *MIP* imparting it a unique ability for immunomodulation. Besides, it also suggests an important role of genomic fluidity in habitat diversification within mycobacteria and provides a unique view of evolutionary divergence and putative bottlenecks that might have eventually led to intracellular survival and pathogenic attributes in mycobacteria.**

## INTRODUCTION

*Mycobacterium indicus pranii (MIP)* is a saprophytic mycobacterial species that is known for its immunomodulatory properties (1–11). In late 70s, this bacterium, initially coded as *Mycobacterium* 'w', was selected from a panel of atypical mycobacteria for its ability to evoke cell mediated immune responses against *M. leprae* in leprosy patients (2,9). *MIP*, which shares antigens with both *M. leprae* and *M. tuberculosis*, provides protection against *M. tuberculosis* infection in mice (3,10,12,13) and accelerates sputum conversion in both type I and type II category of tuberculosis (TB) patients when used as an adjunct to chemotherapy (14,15). In HIV/TB co-infections, a single dose of *MIP* converted tuberculin −ve patients into tuberculin +ve in >95% of the cases (16). This attribute is unique to *MIP* because similar application of other saprophytic mycobacteria such as *M. vaccae* does not provide commensurate protection (17). Based on its demonstrated immunomodulatory action in various human diseases, *MIP* is the focus of several clinical trials (Table 1) and successful completion of one such trial has led to its use as an immunotherapeutic vaccine 'Immuvac' against leprosy (18). However, very little information is available about *MIP*'s molecular, biochemical, genetic and phylogenomic features.

Recently, in a molecular phylogenetic study by using candidate marker genes and FAFLP (fluorescent-amplified fragment length polymorphism techniques) fingerprinting assay, we showed that *MIP* belongs to a group of opportunistic mycobacteria and is a predecessor of *M. avium* complex (*MAC*) (19). A comprehensive analysis of cellular and biochemical features of *MIP*

*To whom correspondence should be addressed. Tel: +91 11 2411 0970; Fax: +91 11 2411 5270; Email: aniltyagi@south.du.ac.in

**Table 1.** Ongoing clinical trials of *MIP* in a diverse set of diseases

| Sr. No. | Duration | Diseases | Objective of ongoing trials | Phase | Trial No. | Intervention |
|---------|----------|----------|-----------------------------|-------|-----------|--------------|
| 1 | 2007–09 | Tuberculosis | Efficacy and safety of immunomodulator (*MIP*) as an adjunct therapy in Category I pulmonary tuberculosis along with assessment of immunological parameters | III | NCT00341328 | *MIP* alone and also along with Category I ATT drugs as per RNTCP guidelines |
| 2 | 2008–10 | Tuberculosis | To study the efficacy and safety of *MIP* in the retreatment of lung (Type 2) tuberculosis patients | III | NCT00265226 | Intra-dermal administration of *MIP* |
| 3 | 2008–11 | Tuberculous pericarditis | A pilot trial of adjunctive prednisolone and *MIP* with immunotherapy in tuberculous pericarditis | III | NCT00810849 | Prednisolone and *MIP* immunotherapy |
| 4 | 2008–11 | Superficial transitional cell carcinoma | To compare the efficacy, toxicity and time to tumor regression by treatment with *MIP* (intra-dermal) and BCG (intravesical) in patients with newly diagnosed STCC with high probability of recurrence | II | NCT00694915 | *MIP*, BCG |
| 5 | 2007–10 | Hormone refractory prostate cancer (HRPC) | To compare the overall survival, hematological toxicity, pain reduction score, response to tumor, quality of life in two arms of HRPC patients from different parts of India | II | NCT00525408 | *MIP* as an adjuvant to docetaxel |
| 6 | 2008–10 | Superficial transitional cell carcinoma | To evaluate the response rate of *MIP* treatment in patients, detecting its effect on time to tumor progression and evaluating its safety | I | NCT00694798 | *MIP* |
| 7 | 2006–10 | Stage III or Stage IV melanoma | To evaluate clinical response, immune response and safety of treating patients with advanced stage melanoma with the vaccine CADI-05 | I, II | NCT00675727 | *MIP* |
| 8 | 2008–10 | Non-small cell lung cancer | To determine efficacy of *MIP* in combination with paclitaxel plus cisplatin in advanced non-small cell lung cancer | II | NCT00680940 | *MIP*, paclitaxel and cisplatin |

along with chemotaxonomic markers such as FAME (fatty acid methyl ester) analysis and comparison with other mycobacterial species established that *MIP* is endowed with specific attributes (4). It has a growth rate (time of colony appearance ∼6–8 days) that is faster than the typical slow growers such as *M. tuberculosis* (∼3 weeks) and slower in comparison with typical fast growers, such as *M. smegmatis* (∼3 days), and thus placing *MIP* somewhere in-between the slow and fast grower mycobacterial species (4). In *Mycobacterium*, fast growers usually represent non-pathogenic organisms whereas slow growers are usually specialized pathogens. *MIP* does not cause any infection in mice, guinea pigs and monkeys, the animal models in which it has been tested (6). Biochemical analysis also showed that *MIP* shares several features that are exclusive to either slow growers or fast growers (4). Even the FAME profiling of *MIP*, a key test for appropriate taxonomic placement of microbes, and its comparison with the fatty acid complement from other mycobacterial species corroborated the placement of this saprophyte in between fast and slow growers (4). Thus, *MIP* represents an organism placed at an evolutionarily transitory position with respect to a fast grower and a slow grower or a saprophyte and a seasoned pathogen.

It is known that mycobacterial species represent one of the most dramatic examples of host tropism and habitat diversification. *Mycobacterium* has more than 125 notified species including saprophytes such as *M. smegmatis*, immunomodulators such as *M. habana*, *M. vaccae* and *MIP*, opportunist *M. avium* and strict intracellular pathogens like *M. tuberculosis* and *M. leprae*. This unmatchable competence of mycobacterial organisms and their diverse

physiological characteristics can be attributed to the genome dynamics including genome organization, gene content, coordinated gene expression and ability to interact with the host machinery. An important unanswered question in this context remains as to how the soil-living saprophytic mycobacterial species turned into one of the most notorious intracellular pathogens. Thus, understanding of the genomic basis of habitat diversification could be crucial in evolving effective control measures against mycobacterial infections. Unfortunately, despite the publication of several mycobacterial genomes (20–23), the understanding and details of advent of parasitism within mycobacterial lineages remain obscure (especially with in *MAC*) although the evolution of niche adapted parasitic forms by genomic downsizing is an accepted norm in *M. tuberculosis complex* (21,23). In fact, formal genetic studies on species differences and divergences in mycobacteria have been severely limited by the unavailability of a related organism that represents the border of optimization between saprophytic and pathogenic mycobacterial species. In prokaryotic evolution, a few species such as *Shigella flexneri* and *Yersinia pestis* have been identified, which represent an early stage of host restricted adaptation by means of genome shedding (24). *MIP* because of its unique phylogenetic placement and associated biochemical features seems to be the first case of a mycobacterium species caught in transition just before it resorted to the pathogenic adaptations. Thus, it provides a unique opportunity to understand evolutionary divergence and putative bottlenecks responsible for the advent of intracellular mode of survival and pathogenic attributes in mycobacteria.

We have sequenced complete *MIP* genome to gain an insight into its unique life style and molecular basis of immunomodulation. In addition, we have employed comparative genomics to understand the habitat diversification and bases and means of functional genetic correlates responsible for evolution of pathogenicity in ancestral mycobacterial lineages.

## MATERIALS AND METHODS

### Sequencing of *MIP* genome

The genome sequence of *MIP* was determined by employing Sanger sequencing by using a hybrid strategy of sequencing shot gun libraries (2 and 5 kb) and partial sequencing of some clones of large insert sized (>125 kb) BAC (bacterial artificial chromosome) library. Briefly, genomic DNA was isolated and whole genome shotgun libraries with average insert size of 2–3 kb and 4–5 kb were prepared by hydroshearing. Fragments of required size were gel-eluted, blunt-ended and cloned in plasmid vector pUC19. Clones were randomly picked from libraries having more than 90% insert and sequenced by Sanger's di-deoxy terminator chemistry on ABI 3700 machines. A high quality BAC library was also prepared (www.mwg-biotech.com (15 August 2012, date last accessed)) and end sequenced by employing Sanger's method to create a physical map of *MIP* genome that assisted in gap filling and resolving the ambiguities in genome assembly. Gap closing and the re-sequencing of low-quality regions were performed by sequencing the PCR products and the appropriate plasmid clones. These data were assembled by using the PHRED-PHRAP-CONSED package of software on four processor SunFire V400 series of server. Identification of open reading frames (ORFs) was carried out with the help of GLIMMER gene prediction software (25). Protein localization analysis was carried out with the help of PSORTB (26).

### Comparative proteome analysis of *MIP* with other species

Functional annotation was carried out on the basis of sequence alignment with the known mycobacterial proteins as well as the COG (clusters of orthologous groups of proteins) (27) database with the help of BLAST (28) package. Several perl scripts were developed in-house for data analysis. To understand the effect of gene variations on the habitat diversification in mycobacterial species with respect to *MIP*, we performed BLAST analysis of *MIP* proteome against the proteomes of 18 other mycobacterial species used in this study. They were assigned to specific lineages of pathogenic and saprophytic mycobacteria based on their characteristic features, habitat and available literature. The members of *M. tuberculosis* complex including *M. marinum* and *M. ulcerans* and those belonging to *M. avium* complex were categorized as pathogenic group whereas rest of them were grouped as environmental mycobacteria. The positive hits against *MIP* proteins were filtered out and remaining genes (unique with respect to species under investigation) were analysed for their function based on

COG classification and were quantified. This dataset was obtained for each species of both groups and was viewed as variation in unique gene content in each function category with respect to *MIP*.

### Analysis of rate of natural selection (Ka/Ks analysis)

To understand the role of selection on speciation in *MAC*, the orthologous group of genes between *MIP* and *M. avium* subsp. *hominissuis* (*MAH*- human strain) and *MIP* and *M. avium paratuberculosis* (*MAP*-animal strain), were identified by using InParanoid program (29). This method bypasses multiple alignments and phylogenetic tree-based conventional approaches to detect orthology and thus minimizes any bias arising due to alignment or phylogeny method in the identification of orthologs. First, all possible pair wise similarity scores that scored higher than a cutoff value (bit score $\geq 50$, overlap $\geq 70\%$, $e \leq 10^{-10}$) were detected from all-against-all BLAST comparisons and then the reciprocal genome-specific best hits were marked as orthologs. The orthologs were subsequently classified based on functional categories as per the similarity searches against COG database. The orthologs were aligned by using ClustalW (30) and each alignment was manually inspected for its correctness. Pairwise estimates of the non-synonymous (Ka) and synonymous (Ks) substitution rates were obtained by KaKs_Calculator program by using a maximum likelihood method based on the HKY85 model (31).

### Analysis of lateral gene acquisitions in *MIP*

A combination of parametric methods, comparative genomics and phylogenetic approaches was employed to predict laterally acquired genes in *MIP*. First of all, we employed the three most popular parametric approaches namely Alien Hunter (32), genomic signature analysis (33) and by analysing atypical GC content of each ORF. Alien Hunter implements an interpolated variable order motifs theory to predict compositionally deviating regions with the highest recall value. The genome sequence of *MIP* was scanned and the fine tuning of the co-ordinates of alien regions was carried out by using advance optimization algorithm available in Alien Hunter. Besides, each *MIP* ORF was analysed for its length and nucleotide composition with respect to total and positional G+C contents (G+C [T], G+C [1], G+C [2] and G+C [3]). The genes were considered as extraneous on the basis of G+C content, if their total G+C (T) content deviated by >1.5 $\varsigma$ from the mean value of their genome or if deviations of G+C [1] and G+C [3] were of the same sign and at least one of them was >1.5 $\varsigma$ (34). The genes shorter than 300 bp and the genes coding for ribosomal genes were excluded from this analysis to avoid any extraneous results. We further augmented our analysis of *MIP* genome by using genomic signature based method previously used for mycobacteria (33,35). The genes that were scored by more than one method in these analyses were considered as laterally acquired. The genes confirmed by both genomic signature and GC content based methods were referred as recently acquired and these signatures were used to ascertain their likely source of acquisition.

Further, we used the power of comparative genomics by analysing *MIP* genes for their presence/absence across available mycobacterial species $\leq e^{-10}$). *MIP* regions having a non-uniform gene distribution across various mycobacteria, which are not scored by Alien Hunter, were annotated as RRD (regions of restricted distribution of genes). RRD has been defined as the region in *MIP* genome, which harbors the genes that are absent in a minimum of 33% of species investigated in this study and is at least represented by three contiguous genes or a region of >3 kb. The genes, which are absent in more than 50% of the species investigated were then referred as laterally acquired in RRDs and elsewhere in *MIP* genome. All the genes identified as possible lateral acquisitions in *MIP* were probed against COG database to analyse the functional role of gene acquisitions. Besides, laterally acquired genes were analysed by BLASTP algorithm against ACLAME (36), a database dedicated for the classification of mobile genetic elements (MGEs).

### Other *in silico* analysis and stress experiments

CRISPR analysis was performed by using CRISPRFinder (37). Annotation of transporter genes was carried out by TransAAP (38). Pathogenic islands were inferred from PAIDB (39), the pathogenic island database. *MIP* was analysed by using Virulence factor database (VFDB) to ascertain the status of genes associated with virulence (40). *In silico* prediction of antigenicity was carried out with VAXIJEN (41). PFAM (http://pfam.sanger.ac.uk (15 August 2012, date last accessed).) was used to analyse and draw protein domains in a scaled manner. Motif scan tool at MyHits web server (http://myhits.isb-sib.ch (15 August 2012, date last accessed)) was used for further analysis of proteins and motifs (42). Phylogenetic analysis was performed by using maximum likelihood method available in Phylogeny Fr. Server (43). Influence of nutritional stress on *MIP* was evaluated on the basis of viable cell count at different time points (44).

### Statistical analyses

Variations in gene distribution across different lineages were analysed by two-way ANOVA followed by Bonferroni posttests. $P < 0.05$ was considered as statistically significant. For studying natural selection, Fisher's exact test (built in KaKs_Calculator program) for the small sample was applied to justify the validity of Ka and Ks calculated in this study. Only the ortholog pairs with $P < 0.05$ were considered for further analysis to infer the rate of natural selection. Paired *t* test was performed to ascertain the significance in the rate of selection between different organisms ($P < 0.05$).

### Total number of mycobacterial species analysed in this study ( = 18)

The genome sequence along with annotation for the following organisms were downloaded from NCBI genome databanks and used in this study: *M. marinum, M. ulcerans, M. tuberculosis* H37Rv, *M. tuberculosis* H37Ra, *M. tuberculosis* CDC1551, *M. tuberculosis* F11, *M. bovis, M. leprae, M. bovis* BCG, *M. avium* supsp. *paratuberculosis, M. avium* 104, *M. smegmatis, M. gilvum, M. abscessus, M. vanbaalenii, M.* sps. *JLS, M.* sps. *KMS* and *M.* sps. *MCS.*

## RESULTS AND DISCUSSION

### Genome sequencing and general features of *MIP* genome

Sequencing of *MIP* (DSM 45 239$^{\text{T}}$) genome was carried out by whole genome shotgun (WGS) approach. A total of 109 792 paired end reads, comprising of more than 10× coverage of *MIP* genome, were generated from randomly picked shotgun clones from both ~2 and ~5 kb shotgun libraries followed by gap filling and sequence improvement. Sequence assembly with PHRAP resulted in the assembly of 93 592 shotgun sequences leading to a single circular *MIP* chromosome of 5 589 007 bp (Figure 1). This was subsequently validated by a BAC end sequence based physical map of *MIP* genome. Mycobacterial genomes range from 3.5 to 7 Mb and *MIP* with a size of ~5.6 Mb represents a moderate genome size, which is larger than all known organisms of *MAC*. The genome contains 5270 predicted ORFs (at a density of ~1 gene/kb), a single rRNA operon and 45 tRNA genes; these ORFs account for ~91% of the genome (Table 2). The mean G + C content of *MIP* genome is 68%. However, the cumulative nucleotide skew analysis revealed several regions with a G + C content clearly divergent from this mean value, which cover considerable area in *MIP* genome and constitute potential sites to investigate for laterally acquired genes (Figure 1). The putative '*ori*' in *MIP* genome was identified by a relatively AT rich region with characteristic DnaA boxes and a typical gene order of '*rpnP-dnaA-dnaN*'. The 'ATG' was found to be the most frequent start codon (56.5%) followed by 'GTG' (37.5%) and 'TTG' (5.9%). Like *M. tuberculosis*, *MIP* has an even distribution of ORFs on both strands with respect to the direction of replication (2656 on the leading strand and 2614 ORFs on lagging strand) (4). PSORTB analysis indicated that 55.5% of *MIP* proteins are cytoplasmic in nature, 13.5% are localized in the cytoplasmic membrane and only 3.5% are extra-cellular in nature (26). However, the precise localization of 27.5% of the proteins could not be ascertained.

BLAST-based comparative analysis of *MIP* ORFs (at a cut off value of ≥70% amino acid identity) revealed their maximum similarity with *MAC* organisms, which are evolutionarily close to *MIP* (Supplementary Figure S1). This is followed by *M. marinum* with which *MIP* shares over 51% of its coding sequences (CDS) (Supplementary Table S1). This observation is consistent with the status of *MIP* as the progenitor of *MAC* and supports the idea of a shared aquatic past between saprophytic and pathogenic mycobacteria (19,45). With *M. tuberculosis*, *MIP* shares only ~40% of its proteins. However, the number of *MIP* ORFs (~68%) shared by closely related *MAC* species strikingly differs in comparison with other related mycobacteria, which usually share over 90% of coding sequences even at identity >95% (22). This divergence could be a critical component for the elicitation of a
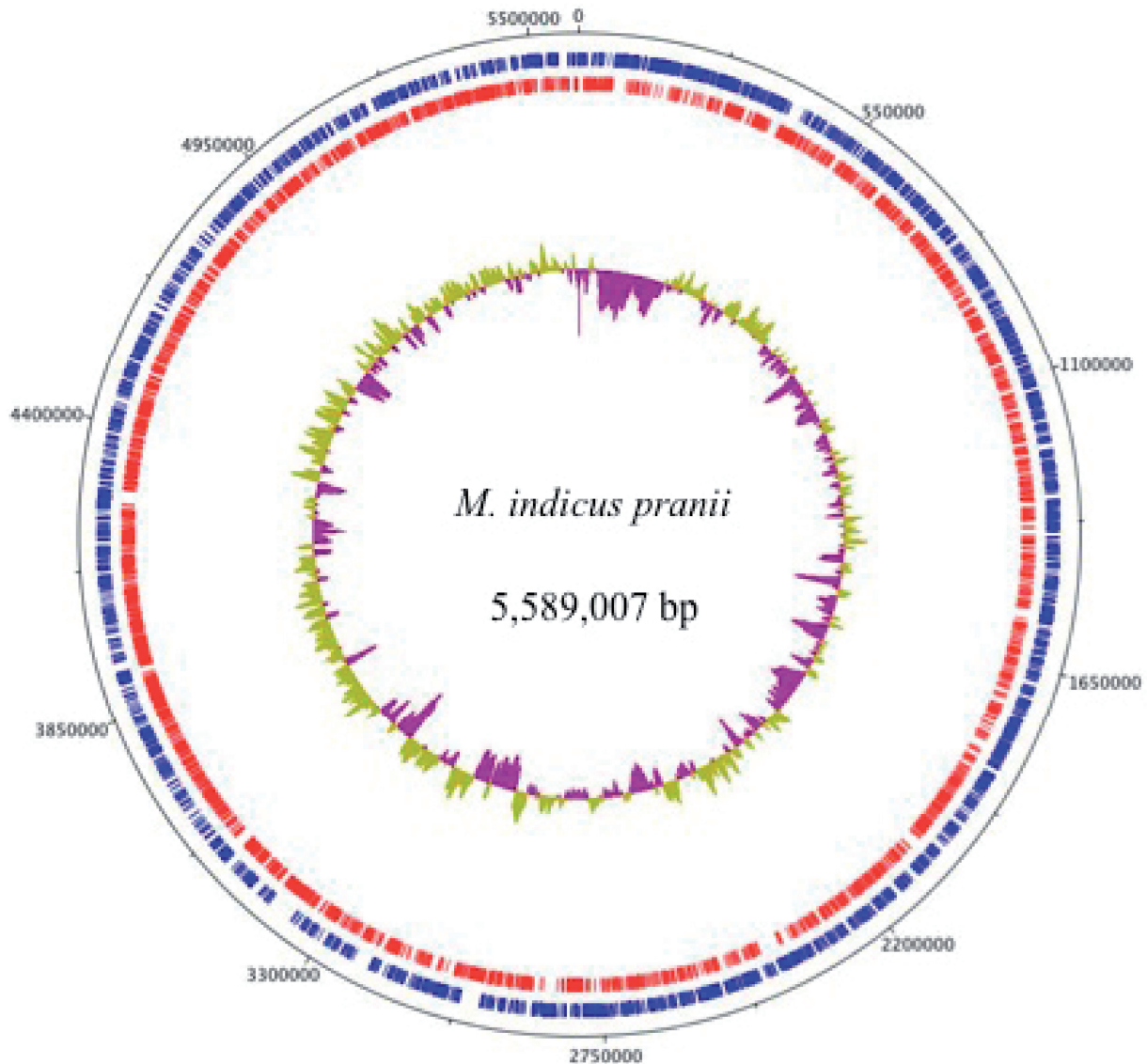
**Figure 1.** Circular representation of *MIP* genome. Whole genome sequencing of *MIP* revealed that it harbors a single circular chromosome of 5 589 007 bp. The accuracy of genome data assembly is ensured by a BAC end sequence based physical map of *MIP* genome. The size of *MIP* genome is much larger than the genome of any member of *M. avium* complex and thus is in agreement with the progenitor status of *MIP* (19). The red and blue tracks represent ORFs predicted in the sense and anti-sense orientation in relation to the *ori* (origin of replication). The inner most track represents the GC skew wherein sharp peaks of violet and yellow represent regions of AT and GC richness, respectively, and constitute potential targets for lateral gene analysis .

robust yet unique immune response upon vaccination with *MIP*.

**Functional classification of *MIP* proteins**

To facilitate functional studies, *MIP* proteins were subjected to BLAST analysis against the COG database, which serve as a platform for functional annotation of newly sequenced genomes and for studies on genome evolution (27). On the basis of similarity with COG proteins, it was possible to assign functions to ~80% of *MIP* proteins but ~20% of the proteins still remain un-annotated. More significantly, ~7.5% of proteins are unique to *MIP* and show no significant homology with other proteins present in mycobacterial proteomes.

Several of these candidate orthologs are present in gene clusters, which are absent from most of the other mycobacteria, and thus indicating the modular nature of gene acquisitions or deletions in mycobacteria. Our analysis shows that 41.5% of *MIP* proteins belong to 'Metabolism' category, 11.5% to 'ISP' (information storage and processing), and 9.5% to 'CPS' (cellular processes and signaling) whereas 16.7% are 'poorly' categorized proteins (Figure 2). Within 'Metabolism' category, the genes pertaining to lipid transport and metabolism (I) were over-represented (22.5%) closely followed by secondary metabolites biosynthesis, transport and catabolism (Q) (21.4%). In the 'ISP' category, majority of the proteins were related to transcription (K) (48.5%) followed by replication, recombination and repair

(L) (26%) and translational, ribosomal structure and biogenesis (J) (24.5%). In case of 'CPS', major representation comes from cell wall/membrane/envelope biogenesis (M) (27.6%) followed by posttranslational modifications (O) and signal transduction mechanisms (T) at 23 and 21.4%, respectively (Figure 2).

**Table 2.** General genomic features of *MIP*

| Category | Feature | Value |
|---|---|---|
| General characteristics | Size (bp) | 5 589 007 |
| | GC content (%) | 68 |
| | Coding density (%) | 91 |
| | Average ORF size (bp) | 960 |
| | Predicted ORFs | 5270 |
| | tRNA | 45 |
| | Ribosomal RNA operon | 1 |
| | IS/transposons | 38 |
| Proteome functional analysis (based on COG) | CDS with predicted function (%) | 80 |
| | Unannotated (%) | 20 |
| | No significant homology with reported mycobacterial proteomes (%) | 7.5 |
| Protein localization (by using PsortB) | Cytoplasmic (%) | 55.6 |
| | Cytoplasmic membrane (%) | 13.5 |
| | Extracellular (%) | 3.4 |
| | Unknown (%) | 27.5 |

## Comparative proteome analysis of *MIP* with other species reveals the role of genomic fluidity in habitat diversification in *Mycobacterium*

COG-based comparative analysis of gene distribution across mycobacterial proteomes highlights the presence of distinct genome fluidity. 'ISP' and 'Metabolism' proteins vary considerably with the maximum flexibility being observed in replication, recombination and repair (L), lipid transport and metabolism (I) and secondary metabolites biosynthesis and transport (Q), respectively (Figure 3). The minimum variations are observed in 'CPS' with nearly all sub-categories exhibiting a consistent representation. In 'ISP', the distribution of genes across all mycobacterial proteomes is almost consistent for translation, ribosomal structure and biogenesis (J) and chromatin structure and RNA processing (B), while a clear genomic fluidity is exhibited by the genes belonging to replication, recombination and repair (L). This category is least represented in *MIP* (3%) and maximally in *M. ulcerans* (10%). Similarly, the genes belonging to category K (transcription) are least represented in CDC1551 (5%) and maximally in *M. smegmatis* (9.4%), which is consistent with its saprophytic habitat.

In 'Metabolism', while the genes related to nucleotide transport and co-enzyme transport show a consistent
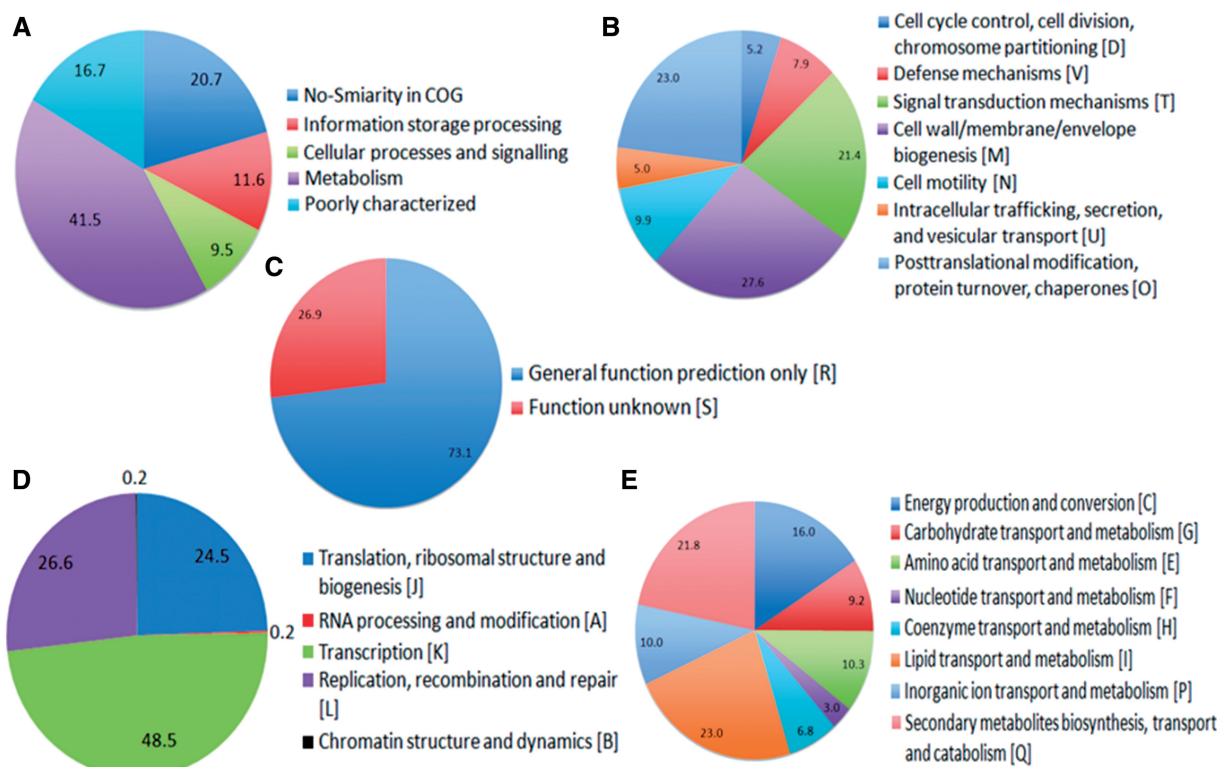


**Figure 2.** Functional classification of *MIP* proteins. (**A**) Representation of *MIP* proteome based on the similarity of its proteins with COG database (27). (**B**) represents distribution in cell processing and signaling category (CPS), (**C**) denotes distribution of poorly characterized proteins in *MIP* while (**D**) and (**E**) stands for information storage and processing (ISP) and 'metabolism' related genes, respectively. It is evident that ~42% of total *MIP* genes are involved in basic metabolic functions and ~21% do not have any homology in COG database. Within 'metabolism' category, the genes involved in lipid transport and metabolism (I) are over-represented (22.5%) closely followed by secondary metabolites biosynthesis, transport and catabolism (Q) (21.4%). In the 'ISP' category, majority of the proteins are related to transcription (K) (48.5%) followed by replication, recombination and repair (L) (26%).
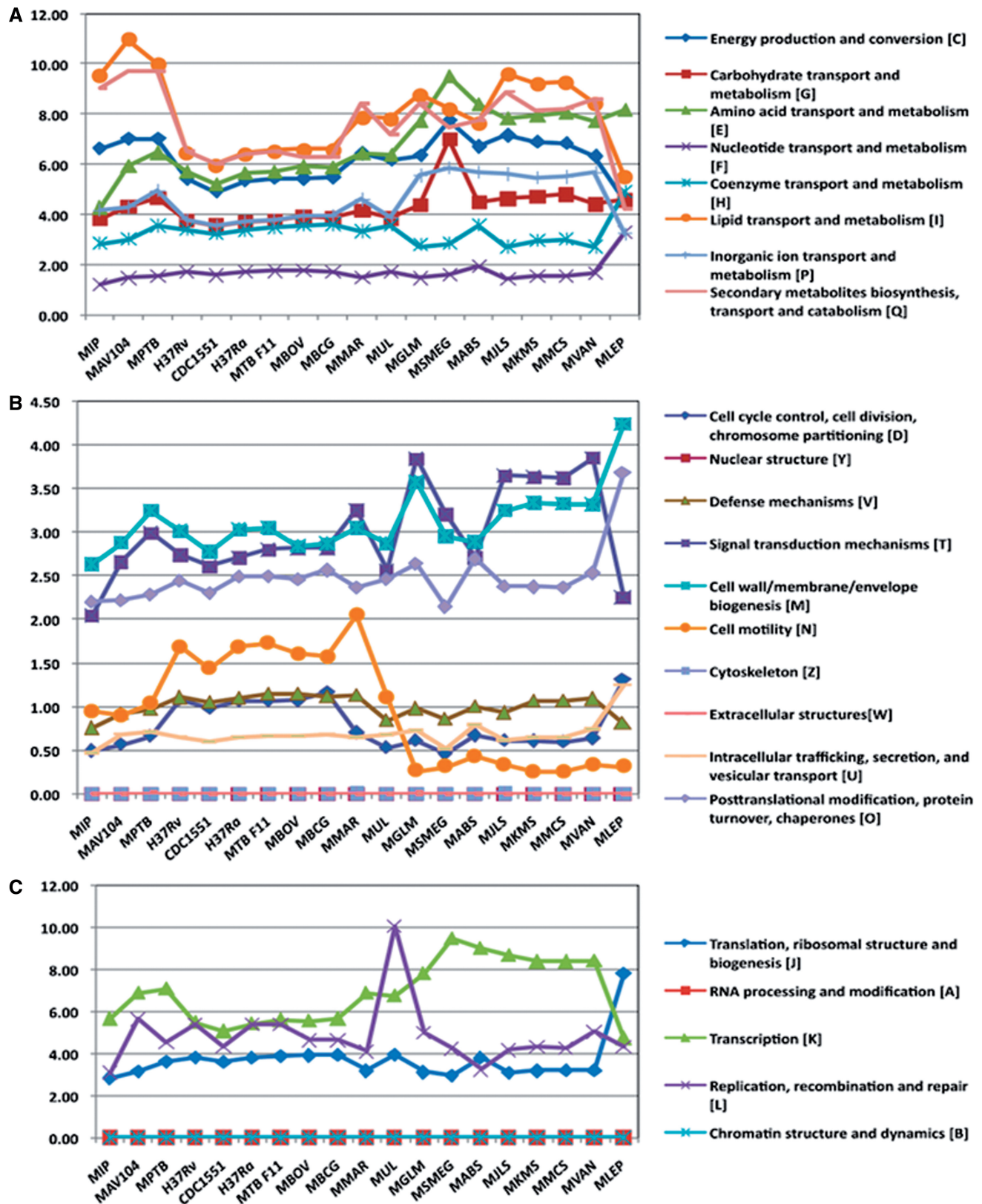
**Figure 3.** Comparative analysis of distribution of different mycobacterial proteomes under various COG functional categories. Different mycobacterial proteomes were downloaded from NCBI and subjected to COG-based BLAST analysis. The contribution of each functional category was calculated to observe the pattern of relative gene distribution across different mycobacterial species and plotted on this graph. (**A**) Distribution across 'Metabolism' category and various sub categories, (**B**) cell processing and signaling (CPS) and (**C**) information storage and processing (ISP). '*X*' and '*Y*' axis represent mycobacterial species and the number of mycobacterial proteins (in percentage), respectively. Our comparative analysis clearly highlights the presence of distinct genome fluidity in mycobacterial species across different functional groupings of genes. This genomic fluidity within different functional groups of proteins may contribute to the habitat diversification observed in mycobacterial species.

distribution, the genes belonging to secondary metabolite biosynthesis and transport (Q), amino acid transport (E) and lipid transport and metabolism (I) show major quantitative variations. 'I' has the maximum representation in *MAC* like *MAH* ($\sim$11%), *MAP* (10%) followed by *MIP* (9.5%) while 'E' and 'Q' are best represented in *M. smegmatis* (9.5%) and *MAC* organisms (9–10%), respectively (Figure 3). In case of carbohydrate transport and metabolism (G), all mycobacterial species have almost an equal representation except *M. smgematis*, which harbors almost twice (7%) the percentage of genes dedicated for this function in other mycobacterial species. In most COG categories, *M. leprae* seems to have a distinctly biased distribution of proteins probably indicative of the extensive gene-loss that the organism has undergone during evolution (21). Of all the mycobacteria, *MIP* has the least representation in L (3%) and E (amino acid transport) (4.3%) categories of genes.

Although the distribution of genes is a species-specific attribute, variations in gene distribution across different lineages could provide an idea about the role of genomic fluidity in shaping the behavior of mycobacteria as saprophytes or host-adapted pathogens. Hence, to get a comprehensive picture of habitat transformation, mycobacterial species were classified in two groups according to their known attributes: pathogenic (PGN) comprising of *M. tuberculosis* complex (including *M. marinum and M. ulcerans*) and *M. avium* complex and saprophytic or environmental (ENV) mycobacteria comprising of *M. smegmatis, M. vanabaalenii, M. gilvum* and others. *MIP* was placed in between saprophytic and pathogenic mycobacterial species because of its unique intermediate position and these two groups were investigated for effect of gene variations in different COG classes with *MIP* as a common background (4). A two-way ANOVA analysis was performed to ascertain the statistical significance of analysis.

While the transition from ENV-*MIP* was associated with a significant reduction restricted to a few COG classes, i.e. K (transcription), T (signal transduction), E (amino acid transport), P (inorganic ion transport and metabolism), R (general function) and S (unknown function) [$P < 0.001$, 0.01, 0.001, 0.01, 0.001 and 0.001, respectively], ENV–PGN transitions involved extensive gene variations (Figure 4). In addition to the gene reduction observed in the earlier mentioned classes, reduction was also noticed in genes related to energy metabolism (C, G, I and Q [$P < 0.05$, 0.05, 0.001 and 0.05, respectively]) and a significant increase in L (replication, recombination and repair, $P < 0.01$) and N (cell motility and secretion, $P < 0.01$) related genes in ENV-PGN transition. Noticeably, the habitat change from *MIP* to PGN lineages was primarily due to the loss of genes involved in I (lipid transport and metabolism, $P < 0.001$) and Q (secondary metabolite biosynthesis and transport, $P < 0.001$) and gain of genes in L, E and S [$P < 0.001$, 0.001 and 0.05, respectively] (Figure 4). This observation augurs well for a reduced habitat diversity of pathogenic mycobacteria and indicated toward the role of genomic fluidity within selected gene functions towards habitat
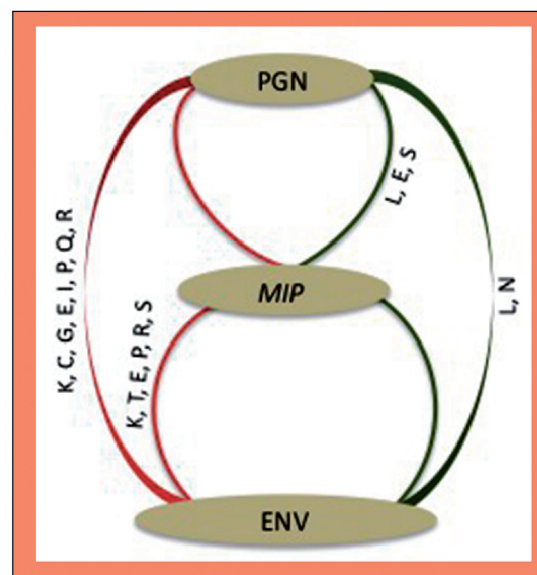


**Figure 4.** Quantitative analysis of gene variations involved in habitat transformation in mycobacteria. This cartoon depicts variations across major functional gene groupings as mycobacterial species adapted to a pathogenic lifestyle from free-living environmental mycobacteria. Red lines denote loss of genes while the green ones denote the gene gain with a change of habitat. Although the transition from *ENV-MIP* was associated with a significant reduction restricted to a few COG classes i.e. K (transcription), T (signal transduction), E (amino acid transport), P (inorganic ion transport and metabolism), R (general function) and S (unknown function), ENV–PGN transitions involved extensive gene variations and are consistent with the intermediate evolutionary position of *MIP*. Significantly, only two major gene categories reported gain of genes associated with the advent of pathogenicity: L (DNA replication, recombination and repair) and E (amino acid transport and metabolism). But transition from purely saprophytic lineage to pathogenic habitat is associated with genes categorized into 'L' only, which also contains transposon elements. Indeed, we observe that saprophytic mycobacterium like *MIP* has only 38 transposons-like elements compared with 302 found in similar-sized pathogenic mycobacterial species *M. ulcerans*. A two-way ANOVA analysis was used to ascertain statistical significance.

specification. An increase in the representation of 'L' with the advent of pathogenicity offers an interesting paradigm, which warrants further studies in the model organisms.

## Role of natural selection in speciation in *Mycobacterium*

Measurement of the rate of non-synonymous (leading to change in amino acid) and synonymous (silent) nucleotide substitutions in protein-coding DNA sequences is the most referred criterion for detecting natural selection in molecular evolutionary analysis (46). Significantly higher non-synonymous nucleotide substitutions (Ka) over the synonymous (Ks) ones are interpreted as an evidence of positive natural selection. Hence, to understand the contribution of selection in speciation, we have used closely related and phylogenetically independent species of *M. avium* complex of which *MIP* is a predecessor (19). Orthologs were identified using Inparanoid tool (29) and dataset of $\sim$2600 gene pairs representing >80% of the orthologs shared among different species of *MAC* was obtained to perform comparative analysis of
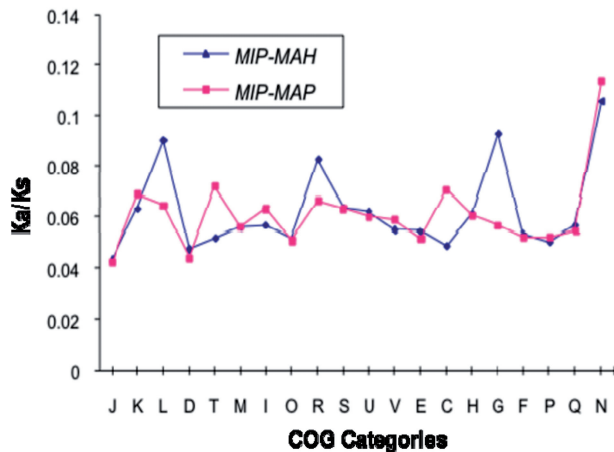
**Figure 5.** Role of natural selection in speciation in *MAC*. Analysis of average rate of natural selection (Ka/Ks) among *MIP–MAP* and *MIP–MAH* lineages revealed the presence of a similar purifying selection (46). This implies that both mycobacterial lineages have undergone an independent evolution into their respective host adapted forms from *MIP*. However, a significant skew in selection rate (Ka/Ks) is observed in genes categorized into energy production and conversion (C), and thus establish the role of metabolism-related genes in the evolution of host tropism. Also, a strong positive selection (>50 times of average selection) was observed on ComEC gene that encodes a competence protein required for DNA uptake and natural transformation (47). Such a strong selection on this gene indicates that ComEC has played an important role in modulating the efficiency of DNA uptake during mycobacterial evolution. In fact, in the case of *MAP*, this gene is found to be pseudogenized, which usually results from an excessive positive selection.

the rate of selection for human-adapted (*MIP–MAH*) and animal-adapted (*MIP–MAP*) niches from a saprophytic *MIP*. The evaluation of rate of selection (Ka/Ks) revealed strong purifying selection (~0.06) acting on both human and animal adapted lineages. However, further resolution of analysis based on protein function revealed a significant difference in the rate of selection only for the genes involved in energy production and conversion (C) ($P < 0.03$, unpaired *t* test) (Figure 5). Also, very few genes, mostly distributed in metabolic pathways, were found to have undergone strong positive selection (Supplementary Table S2) suggesting their relevance in undergoing niche-specific adaptations in *Mycobacterium*. A very strong positive selection (>50 times of average rate) was observed in ComEC (MIP2580) (47), the competence protein required for exogenous DNA uptake during natural transformation, which can critically influence the ability to acquire foreign DNA in microbial species. Incidentally, we found this gene to be pseudogenized in *MAP*, which usually results from an excessive positive selection. In a recent study (48) based on SNP analysis, it was argued that recombination may influence the rate of selection in extremely closely related species of *M. tuberculosis* complex (average nucleotide identity >98% across different species). Even though *MIP* is likely to have minimal homologous recombination events because of sequence heterogeneity with *MAP* and *MAH*, the likelihood of recombination and lateral gene transfer influencing the rate of selection cannot be completely discounted.

## Identification of laterally transferred genes reveals massive gene acquisitions and mosaic architecture of *MIP* genome

Identification of laterally acquired genes is an important paradigm, which is cardinal to gain a deeper insight into microbial evolution (49). Hence, after analysing the role of natural selection in speciation, we were keen to analyse the contribution of lateral gene acquisitions in *MIP*. The precise and accurate prediction of lateral gene transfer (LGT) events in an organism is challenging. First, detection of LGT may be influenced not only by source, size and quantity of lateral transfer but also by the genetic features associated with the recipient or host genome (50). Besides, LGT takes place by a variety of means and different tools may be required for better detection of LGT based on specific mechanisms of gene transfer (51). It is also known that different surrogate methods detect lateral acquisitions of different antiquities (52). Hence, all LGT are not amenable to detection by a single parametric method and the application of a combination of different methods is recommended to improve sensitivity of detection in different possible situations (50). However, while the simple addition of predictions from individual methods may increase false-positive rates, the consideration of strictly overlapping predictions as the inclusion criteria for LGT predictions is counterproductive because of the limited overlap of genes observed between different approaches (52,53). Nonetheless, it has been argued that even if the errors inherent to these individual methods are added, the overall benefit is worthy (50). Hence, to predict laterally acquired genes in *MIP*, we used three different parametric methods based on anomalous GC content of each ORF, genomic signature analysis and Alien Hunter predictions to score for likely LGT candidates. The genes were scored as laterally acquired only if they were predicted more than once. This would not only provide sensitivity of detection but also reduce the number of false-positive predictions associated with individual methods. We further augmented our analysis by using information on phylogenetic approaches and phyletic distribution of *MIP* genes in other mycobacterial species as an additional stand alone criterion to score LGT genes (54).

Analysis of atypical GC content of ORFs (34) identified ~28.5% (1503/5270) of *MIP* genes as putative candidates for LGT. A similar analysis with *M. tuberculosis (MTB)*, *M. avium paratuberculosis* (*MAP*) and *M. avium* subsp. *hominissuis* (*MAH*) could only identify 4.3, 6.5 and 11.3% genes, respectively (55). Genomic signature approach could identify ~33% of *MIP* genes as candidate LGT's as compared to *MTB*, *MAP* and *MAH* wherein this approach yielded only 6%, 6.3 and 9.3% genes, respectively. Alien Hunter (32) predicted 85 probable laterally acquired regions (AL) comprising of 1298 (24.63%) ORFs (Supplementary Table S3). The regions around *'ori'* (15 kb on both sides, upstream as well as downstream) and one harboring ribosomal genes were excluded from the evaluation to remove any possible bias. By using similar criteria with Alien Hunter, however, we could identify putative laterally acquired genes in *MTB* (21.5%), *MAP* (15.35%) and *MAH*

(24.2%). More than 42% of *MIP* genes predicted by Alien Hunter are also shared by genomic signature analysis. A similar analysis with *MTB*, *MAP* and *MAH* showed an overlap of 17% (148/865), 31% (215/678) and 33.7% (421/1247), respectively, between Alien Hunter–predicted genes and genomic signature–based predictions. After applying our 'majority' based inclusion criteria, ∼6.2% of the genes emerged as laterally acquired in *MTB*, while *MAP* and *MAH* have 8.3 and 10.2% genes as LGT, respectively. By using this approach, ∼34% of *MIP* genes emerged as laterally acquired, which is significantly higher than in other mycobacterial species analysed in this study. A comparative analysis of *MIP* ORFs based on their restricted distribution within the other mycobacterial genomes identified additional 939 ORFs as plausible lateral acquisitions. This included 362 ORFs harbored by 93 defined RRDs (regions of restricted distribution of genes) in *MIP* (Supplementary Table S4) and 261 ORFs present in alien regions. The incongruence observed in the phylogenetic analysis of some of these genes substantiated their laterally acquired nature. Overall, 50.5% (2664/5270) of *MIP* ORFs appear to be laterally acquired highlighting thereby the scale of evolutionary novelties undergone by this microbe (Figure 6). This study represents the first report of such massive gene acquisitions in mycobacteria and suggests mosaic architecture of *MIP* genome.

Analysis of laterally acquired genes by using COG functional classification revealed the maximum gain in lipid transport and metabolism category (I) followed by transcription (K)-related genes, which are usually under-represented among laterally acquired genes in prokaryotes (56). This was followed by the genes affiliated to secondary metabolites biosynthesis, transport and catabolism (Q) and energy production and conversion (C); these four categories together constitute ∼35% of the total lateral acquisitions. In addition, the LGT predictions based on atypical GC content of ORFs (34) and further validated by genomic signatures 1478 (28.05%) (Figure 7A) appear to retain their native genomic imprints, which are yet to be masked by natural selection. This points toward their relatively recent acquisition and hence, their likely source could be ascertained (33). Analysis based on genomic signatures revealed that majority of these recently acquired genes (∼85%) are most likely derived from actinobacterial species (Figure 7B) like *Streptomyces* (∼25%), *Amycolatopsis* (∼15%), *Rhodococcus* (7.5%) and *Frankia* (6.5%). These gene acquisitions might have been mediated by physical proximity and close interactions among different actinobacteria.

### Mobile elements based gene acquisition in *MIP* are dominated by plasmid-mediated lateral gene transfers

LGT events are usually mediated by mobile genetic elements like phages, transposons and plasmids. BLAST analysis of laterally acquired genes against ACLAME (36), a database of mobile elements comprising all known phage genomes, plasmids and transposons, indicated mobile elements as likely source to 27.4% of these putative laterally acquired ORFs ($<e^{-20}$). Majority of these genes exhibit similarity with plasmids and

extremely small fraction with phages (2%) and IS elements (∼1.2%). The relative paucity of phage and IS elements mediated gene acquisitions and abundance of plasmid-acquired genes in case of *MIP* is surprising. In comparison with other mycobacterial species of similar sized genomes such as *M. ulcerans* (chromosome size ∼5.6 Mb), which has 302 IS elements/transposons (23), *MIP* has merely 38 genes harboring sequences consistent with IS signatures. This is consistent with our earlier analysis based on genomic fluidity across different mycobacterial lineages where variation in the number of transposable elements, which are classified in category L, was found to be associated with habitat diversification. It is tempting to envisage that *MIP* may harbor specific genomic determinants that either provide immunity from phages and transposons, or else predispose *MIP* toward plasmid-based gene acquisitions. *MIP* has a relatively higher number of CRISPR (Clustered Regularly Interspaced Short Palindromic Repeat) elements compared with other mycobacterial genomes (7 as compared with 1–2 in other mycobacterial genomes) (37). These CRISPR molecules not only provide immunity against invasion by phages and viruses (57) but also limit the mobility of IS elements in genome and help in their excision from genome (58). *MIP* also lacks RD1 region, the loss of which facilitates efficient conjugation with plasmids and other chromosomes to promote rapid acquisitions of genes (59). In addition, we found that *MIP* is particularly enriched in transporters of septal DNA translocator family (6 as against 1–2 present in other mycobacteria) (Table 3) that are known to bring out rapid acquisition of genes by mediating cell to cell DNA transfer during plasmid conjugation (60). The abundance of these genomic determinants coupled with the absence of RD1 locus may contribute to the propensity of *MIP* towards plasmid-mediated gene acquisitions.

### Effect of lateral gene acquisitions on different gene families in *MIP*

Two distinct observations have emerged from this analysis: (i) a large number of lateral gene acquisitions in *MIP* have been mediated through mobile elements with only a small contribution through phages and (ii) gene distribution among laterally acquired regions in *MIP* follows a skewed pattern with respect to function as indicated by over-representation of genes belonging to certain categories.

To know the influence of LGT events on distribution of genes across *MIP* gene families, we performed a comprehensive analysis that showed that CYP450 is the largest gene family in *MIP* with 66 members and a gene density of ∼12/Mb. This is remarkably high in comparison with other mycobacterial species such as *M. tuberculosis* (4.5/Mb), *M. smegmatis* (5.6/Mb) and *M. marinum* (7.1/Mb). The analysis of Cytochrome P450 database (http://drnelson.uthsc.edu/CytochromeP450.html (15 August 2012, date last accessed)) revealed that *MIP* harbors the highest number of genes from CYP450 family among prokaryotes sequenced so far and ∼46% (30/66) of these genes are laterally acquired.
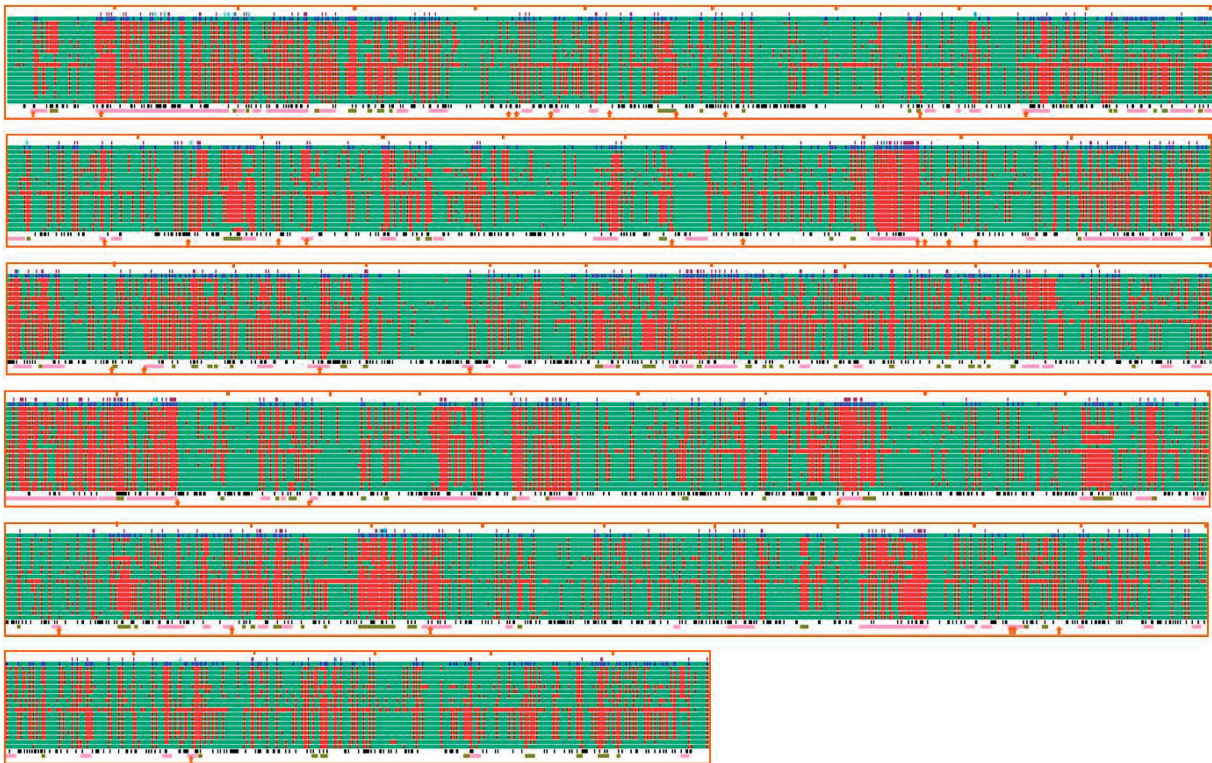
**Figure 6.** Depiction of lateral gene acquisitions in *MIP*. Each column depicts one *MIP* gene and each row depicts one mycobacterial genome (total 18 genomes comprising of *M. tuberculosis* complex, *M. avium* complex and saprophytic mycobacterial species–see Materials and Methods and Supplementary Table S1 for species list); green and red denote presence and absence, respectively, of the *MIP* gene in other genomes. Pink denotes the regions predicted by Alien Hunter (32). Dark yellow represents RRDs, while black columns denote recently acquired genes identified by using atypical gene content and gene signatures (34, 35). Orange arrow denotes position of tRNA molecules, blue denotes genes with homologs in genomes other than mycobacteria, while brown denotes absence in the COG database. A very good overlap is observed between genes identified by using different methods. Most of the alien regions and RRD's overlap with red, substantiating the effectiveness and accuracy of our approach. It is noteworthy that over 50% of *MIP* genome has emerged as laterally acquired, the highest reported so far for any of the mycobacterial species. The figure is scaled to approximation with each figure row denoting 1 Mb of genome and every tick mark denoting 100 kb along the lane.
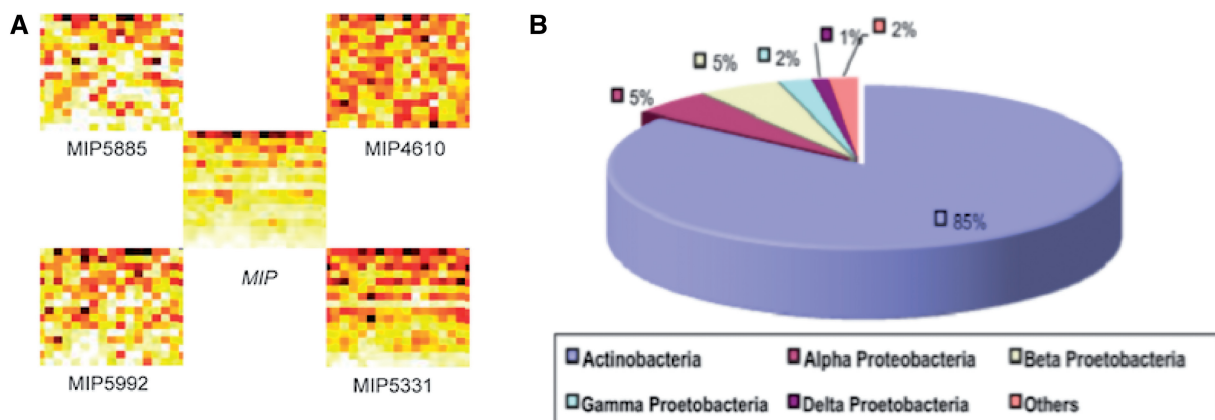


**Figure 7.** Identification of recent lateral gene acquisitions in *MIP* and their analysis. (**A**) Individual gene signatures of recently acquired *MIP* genes along with the whole genome signature of *MIP* establish the alien nature of respective genes (33). These gene signatures are based on the frequency of distribution of tetranucleotide pattern across the whole genome and individual genes of *MIP*, which are color coded to generate a visual impression. (**B**) Distribution of recently acquired genes with respect to their most likely source of acquisition. Based on the genomic signatures, our analysis revealed that majority of these recently acquired genes (~85%) are possibly derived from actinobacterial species.

Approximately 27% of these genes are recent acquisitions, suggesting a recent expansion of CYP450 family. Approximately 11% of CYP450 genes were identified as unique by International CYP450 nomenclature commission and have been classified into three new families and two new sub-families of CYP450 (Table 4). The context-based analysis based on gene neighborhood suggested the role of these genes in the utilization of unusual carbon sources, a key to adaptability and survival of *MIP* at its most likely habitat at soil–water interface (19).

**Table 3.** Comparative transporter analysis of *MIP* with other mycobacterial species

| | | | MSMEG | MIP | MAP | MTB |
|---|---|---|---|---|---|---|
| | Genome size (Mb) | | 7 | 5.58 | 4.83 | 4.4 |
| | Total transporter proteins | | 423 | 222 | 170 | 148 |
| | No. of transporters/ Mb genome | | 60.43 | 39.79 | 35.2 | 33.64 |
| | **Major membrane transporter families and their number in respective mycobacterial genomes** | | | | | |
| 1 | The ATP-binding cassette (ABC) super family | ABC | 109 | 69 | 65 | 44 |
| 2 | The Type II (General) secretory pathway (IISP) family | IISP | 8 | 9 | 0 | 0 |
| 3 | The P-type ATPase (P-ATPase) super family | PATPase | 6 | 10 | 5 | 12 |
| 4 | The septal DNA translocator (S-DNA-T) family | S-DNA-T | 1 | 6 | 0 | 0 |
| 5 | The ammonia transporter channel (Amt) family | Amt | 3 | 3 | 2 | 1 |
| 6 | The small conductance mechanosensitive ion channel (MscS) family | MscS | 4 | 3 | 3 | 2 |
| 7 | The amino acid-polyamine-organocation (APC) family | APC | 28 | 7 | 6 | 9 |
| 8 | The cation diffusion facilitator (CDF) family | CDF | 1 | 5 | 2 | 1 |
| 9 | The monovalent cation (K+ or Na+): proton antiporter-3 (CPA3) family | CPA3 | 6 | 8 | 0 | 0 |
| 10 | The drug/metabolite transporter (DMT) superfamily | DMT | 17 | 3 | 4 | 1 |
| 11 | The major facilitator superfamily (MFS) | MFS | 112 | 29 | 32 | 28 |
| 12 | The multidrug/oligosaccharidyl-lipid/polysaccharide (MOP) flippase superfamily | MOP | 4 | 5 | 2 | 2 |
| 13 | The metal ion (Mn2+-iron) transporter (Nramp) family | Nramp | 2 | 6 | 4 | 1 |
| 14 | The resistance-nodulation-cell division (RND) superfamily | RND | 18 | 19 | 19 | 15 |
| 15 | The Mg$^{2+}$ Transporter-E (MgtE) family | MgtE | 2 | 4 | 2 | 2 |
| 16 | The putative 4-toluene sulfonate uptake permease (TSUP) family | TSUP | 7 | 1 | 0 | 0 |
| 17 | Others | OTH | 95 | 35 | 24 | 30 |

MSMEG, *M. smegmatis*; *MAP, M. avium* susbsp. *paratuberculosis*; *MTB, M. tuberculosis.*

**Table 4.** List of CYP450 ORFs unique to *MIP*

| No. | ORF ID | Length (amino acid) | CYP450 nomenclature[a] | Description |
|---|---|---|---|---|
| 1 | MIP0162 | 399 | CYP1016A1 | New family |
| 2 | MIP0241 | 377 | CYP1018A1 | New family |
| 3 | MIP0261 | 420 | CYP1017B1 | New sub-family |
| 4 | MIP0272 | 409 | CYP1019A1 | New family |
| 5 | MIP0281 | 404 | CYP1018B1 | New sub-family |
| 6 | MIP0283 | 397 | CYP1018A2 | Second member of CYP1018A1 |
| 7 | MIP0295 | 405 | CYP1019A2 | Second member of CYP1019A1 |

[a]Nomenclature as per International Committee of CYP450 nomenclature. 'A1' refers to the first member of a new CYP450 family, whereas 'B1' refers to the first member of a new CYP450 sub-family.

*MIP* has 66 genes of PE–PPE family (16 PE and 50 PPE), which encompass complete repertoire of PPE genes present in various *MAC* species. These PPE genes appear to be selectively inherited by various species of *MAC* as evident from their distribution in *MAP* (36 genes), *MAH* (38 genes), *M. avium* subsp. *avium* (36 genes) and *M. intracellulare* (38 genes), respectively (61). This observation substantiates that evolutionarily *MIP* is a predecessor of *MAC* and endorses our earlier findings based on the rate of natural selection that speciation and habitat diversification has taken place independently from *MIP*. Comparative analysis of PE–PPE genes in *MIP* (with Nr database at NCBI) highlighted that five genes of PPE–SVP sub-family are unique to *MIP*. In addition, several of its PE genes are evolutionarily closer to those belonging to *M. tuberculosis* complex, which is in agreement with the shared evolutionary history of *MIP* and predecessors of *M. tuberculosis* complex (19).

The presence of PPE gene family is unique to mycobacteria among prokaryotes; however, its origin remains unknown. A large number of these genes are laterally acquired in *MIP* prompting us to speculate that PE–PPE genes might have been introduced into mycobacteria through mobile elements. Majority of PE–PPE gene clusters in *MIP* harbor genes related to mobile function activity such as phages, tRNA or 13e12 repeats in their vicinity. Besides, several of these PE–PPE genes exhibited the presence of Ig-like motifs often present in the proteins of tailed double stranded DNA bacteriophage particles. However, the most clinching evidence about the origin of these PE–PPE genes in mycobacteria emerged from the presence of a PPE protein containing intact prophage of ~40 kb that we observed in *MAH* genome during ACLAME analysis (36), a direct evidence for a phage-mediated acquisition of PE–PPE family members.

### Hemerythrins: a versatile gene family laterally acquired in *MIP*

Most surprising finding of *MIP* gene analysis, however, is the unusual presence of ORFs belonging to hemerythrin proteins (Hr) family, the oxygen-carrying non-heme diiron binding proteins, which are usually present in lower invertebrates and annelids. *MIP* has 10 ORFs having significant similarity to hemerythrin (Hr) genes (Figure 8) in comparison with one or two copies in most prokaryotes (62). The prevalence of 'Hr' proteins strongly suggests the preference of *MIP* to inhabit water-columns or sediments, where they reside predominantly at oxic–anoxic interface (OAI) and in the anoxic regions of the marine habitat or both as reported in the case of magnetotactic bacteria, which are also endowed with the abundance of hemerythrins (62).
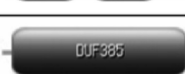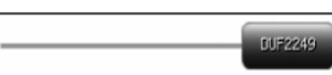
**Figure 8.** Distribution of hemerythrins in *MIP*. Domain mapping and BLAST searches indicated the presence of 10 ORFs belonging to hemerythrin genes in *MIP*. Hemerythrin proteins (Hr) are oxygen-carrying non-heme diiron binding proteins, which are usually present in lower invertebrates and annelids and they usually have only to 1 or 2 copies in most prokaryotes (62). The abundance of these genes strongly suggests the preference of *MIP* to inhabit water-columns or sediments, where they reside predominantly at oxic–anoxic interface and in the anoxic regions of the habitat or both as reported in the case of magnetotactic bacteria, which are also endowed with the abundance of hemerythrins (62,63). The ability of hemerythrins to reversibly bind to oxygen at higher oxygen concentrations and release it in anoxic conditions could also provide an explanation to intriguing behavior of *MIP*, which, not withstanding its aerobic life-style has been shown to grow at 0% oxygen, to reach a plateau in 3 days and die thereafter (64). MIP2918 and MIP2747 did not harbor a definite 'Hr' domain and are identified by BLAST searches against NCBI 'Nr' database.

Domain mapping and comparative genomics with available mycobacterial genomes indicated the presence of two Hr domains in several *MIP* ORFs such as MIP2750, MIP2759, MIP5034 and MIP6380, a trait usually restricted only to proteobacteria (62). We could also identify putative homologs of 'hr' genes in mycobacteria such as *M. marinum* and *M. ulcerans* (1 each), *M. tuberculosis* (3), *M. gilvum* and *M. smegmatis* (4 each), *M. avium* complex (5), other environmental mycobacteria like *M.* sps. *KMS*, *M.* sps. *JLS* and *M. vanabaalenii* (three each) and none in the case of *M. leprae*. Considering the variations in the number of 'hr' genes in various species of mycobacteria and their significant sequence heterogeneity, it appears that acquisition of hemerythrins could have been a selective and independent event facilitating mycobacterial evolution. Incidentally, 90% of these ORFs in *MIP* are laterally acquired with over one-half of them being recent acquisitions. It should be noted that the

efficiency of hemerythrins as oxygen storage proteins is directly dependent on oxygen concentration in its surrounding environment (63). The selective enrichment of *MIP* with Hr proteins and the ability of hemerythrins to reversibly bind to oxygen at higher oxygen concentrations and release it in anoxic conditions could provide an explanation to intriguing behavior of *MIP*, which, not withstanding its aerobic life-style, can manage to grow at 0% oxygen, reach a plateau in 3 days and die thereafter (64).

**Membrane transporters in *MIP***

Transport systems play a critical role in the life-endowing processes such as metabolism, metal homeostasis and secondary metabolite production, affecting thereby the physiology and lifestyle of the organism. In *MIP*, a total of 222 genes were annotated as membrane transporters comprising ∼4.2% of the total gene content and a transporter density of 39.73/Mb (Table 3). This is an apparent reflection on the unique evolutionary position of *MIP* as its transporter density is significantly lower than that of saprophytic *M. smegmatis* (60.43/Mb) and higher than *M. tuberculosis* (33.64/Mb), *MAP* (35.2/Mb) and *M. leprae* (17.2/Mb). It is likely that *M. smegmatis* and *MIP*, being saprophytic in nature need extensive transport machinery to support their life style, whereas intracellular organisms owing to their relatively stable environment have a reduced transporter requirement. Comparative analysis revealed a selective abundance of transporters belonging to septal DNA Translocator family (S-DNA-T) with distinct homology with FtsK/SpoIIIE proteins, which could primarily be responsible for high propensity of *MIP* toward plasmid conjugation and gene acquisitions (60). Another unique seven gene cluster of CPA3 (Proton Antiporter-3) family present in RRD79 (Supplementary Table S4), the largest region with restricted distribution, is responsible for species defining ability of *MIP* to grow on 5% NaCl (4). The phylogenetic analysis established the proximity of these genes to *Catenulispora acidiphila* (Figure 9) that can grow in high salt concentration (65). Along with $Mn^{2+}$ transporters, *MIP* also possesses an abundance of catalases (4) and superoxide dismutases (5), some of them being laterally acquired, which mitigate oxidative stress and reflect not only upon the primitive origin of *MIP* but also equip it for intracellular adaptations. *MIP* can withstand the carbon starvation as evident from our experiments on nutritional stress. After 5 days of growth in PBS without any media or nutritional supplement, *MIP* exhibited no significant reduction in log CFU, reflecting upon its potential to undergo longer period of starvation. Thus, *MIP* appears to have fine-tuned its specific transport abilities by lateral gene acquisitions to gain physiological attributes required for its unique habitat.

**Genome-enabled ecophysiological and metabolic attributes of *MIP* and influence of LGT events**

The presence of a very high number of alternate sigma factors (24 in comparison with 13 in *M. tuberculosis* and *MAP*) endows *MIP* with a complex transcriptional
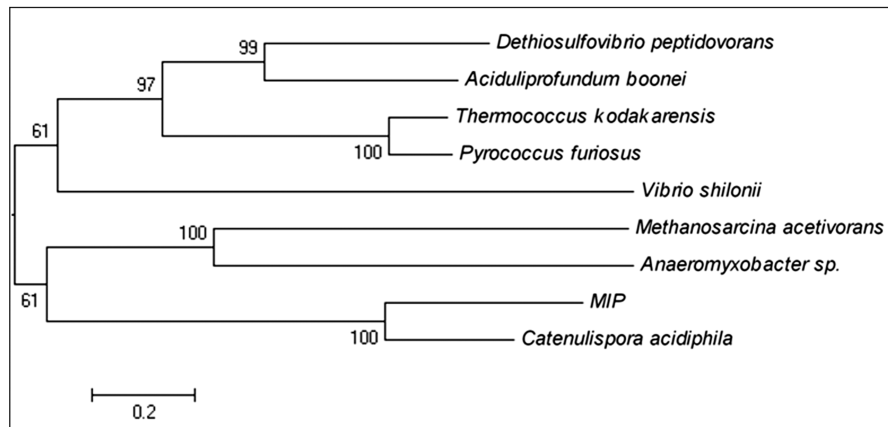
**Figure 9.** Phylogenetic analysis of CPA3 family cluster. This cluster is unique to *MIP* among mycobacterial species and each ORF of this complex encodes different subunits of a unique $Na^+/H^+$ antiporter. All genes have been laterally acquired as a unit, hence, a representative single gene is used to perform phylogenetic analysis by using maximum likelihood method available in Phylogeny Fr. Server (43). The numbers along the branches denote bootstrap values. The phylogenetic analysis established the proximity of these genes to *Catenulispora acidiphila* that can grow under salt conditions (3% NaCl w/v) (65).

flexibility necessary for it to respond to its unique life style (66). The interface life style (soil/water) of *MIP,* as substantiated by its genetic features, prompted us to look for the bio-degradative capabilities of *MIP*. In addition to the abundance of CYP450 genes, *MIP* has also laterally acquired homologs of 3-octaprenyl-4-hydroxybenzoate carboxy-lyase that are involved in anaerobic metabolism of phenol during degradation of plant substrates (67). Besides*,* as enlisted in Supplementary Table S5, it also possesses complete cyanide and thiocyanate biodegradation machinery including the complete enzyme complex (MIP3820-22) of thiocyanate hydrolase (alpha, beta and gamma subunits). This complex degrades thiocyanate and produces $CO_2$ and $NH_3$ that can be used by *MIP* as a nitrogen source (68). Notably, thiocyanate gene cluster is absent from all the pathogenic mycobacteria analyzed in this study (including *M. abscessus* and the opportunists of *MAC*).

Although the ability of *MIP* to degrade different compounds and utilize diverse sources of carbon is perspicuous, the presence of an intact hydrogenases enzyme complex (Table 5) provides evidence for its chemolithotrophic nature even though further research is required to establish the functionality of this complex. The loss of hydrogenases concurs well with the advent of pathogenicity in mycobacteria across different lineages, an observation corroborated by previous studies on mycobacterial hydrogenases (69). Loss of accessory protein coding genes which are required for the maturation and assembly of the hydrogenase complex as well as integration of different metal ions renders this complex non-functional in immediate descendents of *MIP* (i.e. *MAC* species).

In addition to these unique metabolic characteristics of *MIP*, fundamental differences were observed in the organization of lipid metabolic machinery, which is cardinal to the physiology and behavior of mycobacterial species (70). Although, the genetic machinery required for synthesis and modification of mycolic acids is present in *MIP* (70), a major reshuffle is observed in methoxy mycolic

acid synthase gene operon. Sequence analysis further suggested the absence of *papA5*, a gene encoding polyketide-associated protein (Pap) required for the synthesis of virulence associated phthiocerol dimycocerosate (PDIM) (71). These traits are in agreement with the observation that members of *MAC* do not synthesize PDIM's. Further analysis demonstrated the presence of a glycophospholipid (GPL) biosynthesis locus, which is a hallmark of antigenic diversity in *MAC* and appears to be laterally acquired in *MIP*. This gene cluster in *MIP* harbors an ORF (MIP4595) sharing significant similarity with '*gsc*' gene of *MAP*. This gene constitutes a pathogenic island in pathogenic mycobacteria including *M. tuberculosis* (72). However, a comparative analysis of this locus with other *MAC* sequevars revealed the interruption of this locus by a six-gene cluster exclusive to *MIP* with four of these six genes being transposable elements. Thus, this GPL locus acts as a hotspot for transposon integration and is likely to play an important role in *MIP*'s unique biological attributes by influencing GPL biosynthesis.

### Non-pathogenic attributes of *MIP* and immunome analysis

*MIP* as discussed before is non-infectious in mouse, guinea pig and monkey models (2,6,17,73). However, investigation of *MIP* against PAIDB (74), the pathogenic islands database identified the presence of three regions in *MIP* with genomic attributes similar to PAGI islands of *Pseudomonas aeruginosa*. These included a gene cluster (MIP227–MIP247) similar to PAGI 1 pathogenic island of *P. aeruginosa* isolated from a patient with a urinary tract infection (39); a PAG3 like genomic island (MIP272–MIP283) and another region homologous to PAI (MIP333–MIP343), a region similar to *tcd* (toxin complex D) island of *Photorhabdus luminescens* (75). A comparative analysis of *MIP* proteins with virulence factors database (VFDB), a comprehensive compilation of all known virulence factors, also revealed the presence of most of the genes in *MIP* that are reportedly associated with virulence in other mycobacterial species (40).

**Table 5.** List of *MIP* ORFs encoding hydorgenase gene cluster

| ORF ID | Name of the species giving best hit with ACLAME on BLAST analysis | E value | Annotation |
|---|---|---|---|
| MIP5324 | *Ralstonia eutropha* H16 | 5.00E−62 | Nickel-transport integral membrane protein nict |
| MIP5325 | NA | | Hydrogenase expression synthesis |
| MIP5326 | *Anabaena variabilis* | 1.00E−55 | Hydrogenase nickel incorporation protein (hypB) |
| MIP5328 | *Ralstonia eutropha* H16 | 1.00E−137 | NADH ubiquinone 20 kDa subunit (cytochrome C like Ni–Fe hydrogenase small subunit) |
| MIP5330 | NA | | Hypothetical protein |
| MIP5331 | *Ralstonia eutropha* H16 | 0 | Hydrogen:quinone oxidoreductase (cytochrome C like Ni–Fe hydrogenase larger subunit) |
| MIP5333 | *Ralstonia eutropha* H16 | 2.00E−11 | Nitrogen-fixing NifU domain protein |
| MIP5335 | *Ralstonia eutropha* H16 | 3.00E−29 | Hypothetical protein |
| MIP5336 | *Ralstonia eutropha* H16 | 1.00E−39 | Hypothetical protein |
| MIP5337 | *Ralstonia eutropha* H16 | 7.00E−42 | Hypothetical protein |
| MIP5338 | NA | | Hypothetical protein |
| MIP5340 | *Ralstonia eutropha* H16 | 5.00E−16 | Ni–Fe hydrogenase maturation factor HyaD |
| MIP5341 | *Ralstonia eutropha* H16 | 1.00E−07 | Hydrogenase assembly chaperone HypC/hupf |
| MIP5344 | *Bradyrhizobium* sp. | 1.00E−119 | Carbamoyl transferase hypF |
| MIP5345 | *Ralstonia eutropha* H16 | 6.00E−06 | Phosphoheptose isomerase |
| MIP5346 | *Ralstonia eutropha* H16 | 4.00E−22 | Sugar phosphoheptose isomerase |
| MIP5347 | *Ralstonia eutropha* H16 | 2.00E−82 | Hydrogenase expression /maturation/formation protein HypE |
| MIP5349 | *Ralstonia eutropha* H16 | 1.00E−14 | Hydrogenase maturation protein HypC |
| MIP5351 | *Ralstonia eutropha* H16 | 1.00E−137 | Hydrogenase-forming protein HypD |

Pathogenesis is a multi-factorial phenomenon that requires pathogen to attach, infect, sustain, proliferate and eventually disseminate itself inside the host. Hence, the loss of a component responsible for any of these functions is likely to result in the attenuation of virulence or pathogenicity. Thus, despite having PE-PPE genes and *mce1* operon, which enable mycobacteria to invade the host cell, *MIP* lacks both *mce2* and *mce3* operons, which are essential for causing macrophage infections by *M. tuberculosis* and *M. avium* (76–78). The *mce3* as well as *mce2* mutants of *M. tuberculosis* are attenuated in mice although the latter shows no growth defect in macrophages. The *mce2* mutant of *M. tuberculosis* elicits an altered immune response and exhibits no lung pathology along with enhanced survival in mice (76–78). Likewise, *MIP* lacks phospholipase (*plc*) ABCD genes, which are responsible for acquiring host fatty acids for their use as a potential carbon source during persistent infections both in tuberculous and non-tuberculous mycobacterial infections (79).

Another factor crucial for mycobacterial pathogenicity is associated with the presence of latency-related genes that confer on mycobacteria the ability to survive and grow in microaerophilic environment for prolonged period of time. The d*evS*/*devR* two-component system, essential for maintenance of dormant state in low oxygen conditions, is conspicuous by its absence in *MIP* (79). In addition to RD1 locus and toxin–antitoxin system, *in silico* studies further identified *MIP* as a natural mutant of anthranilate phosphoribosyl transferase gene *trpD*, which is involved in tryptophan biosynthesis (81,82). The absence of these critical determinants may severely compromise *MIP*'s ability to survive inside the host as the infection with *MIP* has been found to be self-limiting and clears off within 6–7 weeks (17). The limited survival of *MIP* in low oxygen

inside macrophages despite the absence of d*evS*/*devR* two-component system can be attributed to the prevalence of 'Hr' proteins. *In silico* analysis revealed a much higher fraction of putative antigenic proteins in *MIP* in comparison with BCG (Figure 10), and a majority among them being contributed by lateral acquisitions emphasizing the importance of LGT events in augmenting its immune potential. Besides, the significant sequence heterogeneity observed between *MIP* and *M. tuberculosis* proteins (as mentioned earlier) would render *MIP* proteins acquiescent to generate novel T-cell epitopes resulting in an enhanced immune response. Our analysis revealed that of the 36 proteins shared by *MIP* and *M. leprae*, which were absent in *M. bovis* BCG, 29 were highly immunogenic in nature (Table 6). The most prominent putative antigenic proteins were MIP0340 and MIP5962, both belonging to Hsp20 family and share a close similarity with the 18 kDa small heat shock protein of *M. leprae* (83). This protein bears several T-cell epitopes and generates CD4$^+$ T-cell mediated immune response, a hallmark of protection against tuberculosis. Similarly, MIP7697 is a homolog of *M. leprae* protein MLep2649 that encodes a protein with excellent T-cell stimulating properties, which responds to more than 60% of tuberculosis patients (84). The presence of such immunodominant and productive antigens in *MIP* may potentiate the expression of an antigenic profile better than BCG against *M. tuberculosis* infection.

In summary, different analyses performed in this study establish that *MIP* represents an organism at a unique phylogenetic point as the immediate predecessor of opportunistic mycobacterial species of *MAC*. It is also evident that natural selection in *MAC* has acted in a preferential manner on specific categories of genes leading to reduced habitat diversity of pathogenic bacteria, and thus facilitating host tropism. The genome of *MIP* is
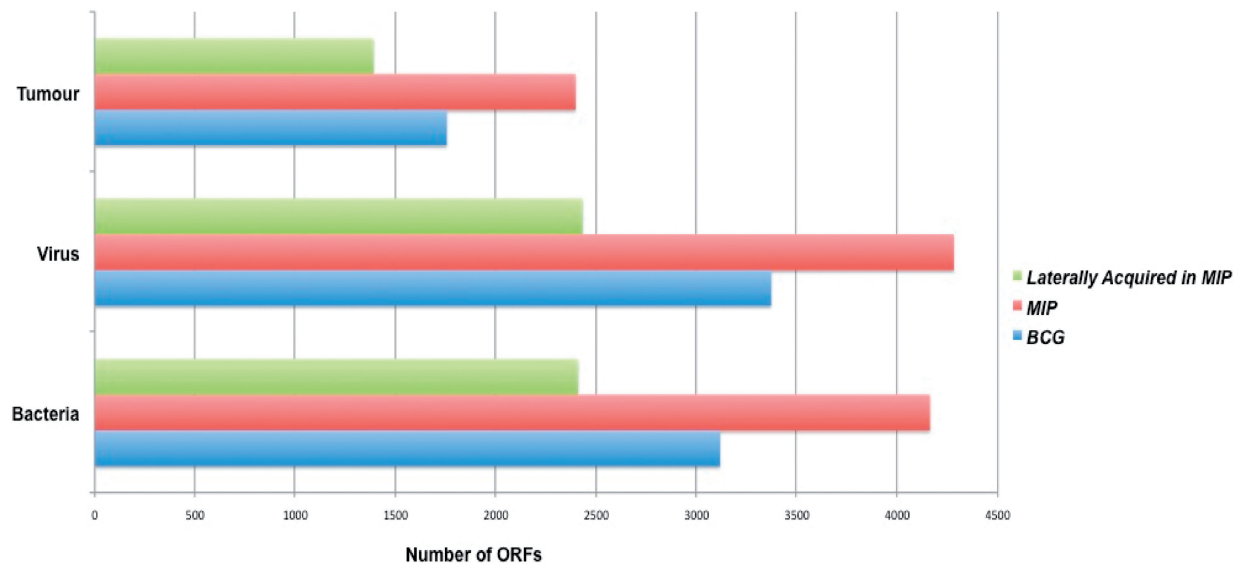
**Figure 10.** Comparative analysis of immunomes of *MIP* and BCG and contribution of LGT events. *In silico* immunome analysis of *MIP* and its comparison with BCG revealed the presence of a greater number of antigenic proteins in *MIP* (41). This may subscribe to the unique potential of *MIP* for immunomodulation against various types of infections. Noteworthily, a significant proportion of these immunogenic proteins appear to be laterally acquired in *MIP*.

**Table 6.** List of *MIP* ORFs shared between *MIP* and *M. leprae* and absent from BCG

| ORF ID | Annotation | Antigenicity index[a] |
|--------|-----------|-----------|
| MIP0340 | 18 kDa protein | 0.527 |
| MIP0593 | Hypothetical protein | 0.4996 |
| MIP0745 | Transmembrane protein | 0.661 |
| MIP0923 | LamB/YcsF family protein | 0.4527 |
| MIP0926 | $Mn^{2+}/Zn^{2+}$ transport system, permease complex | 0.5725 |
| MIP1062 | Trypsin domain-containing protein | 0.8364 |
| MIP1751 | Conserved hypothetical membrane protein | 0.537 |
| MIP2148 | TrkA-N domain-containing protein | 0.4554 |
| MIP2994 | ATPase, RecF-like protein | 0.4161 |
| MIP3248 | Hypothetical protein | 0.7363 |
| MIP4059 | Amidohydrolase family protein | 0.547 |
| MIP4149 | $Mn^{2+}/Fe^{2+}$ transporter | 0.4353 |
| MIP4787 | Phosphoenolpyruvate carboxylase | 0.4691 |
| MIP5052 | Hypothetical protein | 0.4151 |
| MIP5353 | Type 1 phosphodiesterase nucleotide pyrophosphatase | 0.596 |
| MIP5848 | Conserved hypothetical membrane protein | 0.5346 |
| MIP5962 | Molecular chaperone (small heat shock protein Hsp20) | 0.5851 |
| MIP5964 | Antar domain-containing protein | 0.649 |
| MIP6536 | Cadmium, cobalt and zinc/$H^+$/$K^+$ antiporter | 0.5209 |
| MIP6553 | $Mn^{2+}/Fe^{2+}$ transporter | 0.4015 |
| MIP6569 | $Mn^{2+}/Fe^{2+}$ transporter | 0.4332 |
| MIP6577 | Cation efflux system | 0.5447 |
| MIP6758 | Conserved hypothetical membrane protein | 0.546 |
| MIP6769 | Divalent cation-transport integral membrane protein | 0.4636 |
| MIP6809 | Cell entry related family protein | 0.4378 |
| MIP7017 | Pas pac sensor protein | 0.5252 |
| MIP7100 | Cobalt–zinc–cadmium resistance protein | 0.4206 |
| MIP7581 | Putative metallophosphor esterase protein | 0.4255 |
| MIP7697 | Probable deacetylase | 0.41 |

[a]As predicted by *in silico* analysis of *MIP* proteins by VAXIJEN software at default parameters (41).

~5.6 Mb in size and is shaped by a large number of lateral gene acquisitions thus revealing, for the first time, mosaic architecture of a mycobacterial genome. Thus, this study offers a paradigm shift in our understanding of evolutionary divergence, habitat diversification and advent of pathogenic attributes in mycobacteria. A scenario for mycobacterial evolution is envisaged wherein the earliest evolving soil derived mycobacterial species like *MIP* underwent massive gene acquisitions to attain a unique soil–water interface habitat before adapting to an aquatic and parasitic lifestyle. These lateral acquisition events were selective and possibly facilitated by the presence of specific genetic factors (i.e. ComEC) that induce competence to acquire large chunks of DNA to confer immediate survival advantage to the recipient organism. The genes, such as members of 'Hr' family, acquired to assist mycobacteria survive in fluctuating oxygen levels, would have been instrumental in the initial advent of pathogenicity in the aquatic opportunistic mycobacterial species. Subsequently, mycobacterial species tuned their genetic repertoires to respective host adapted forms with a high degree of genomic fluidity aided by selective lateral gene acquisitions and gene loss by deletion or pesudogenization (19). Importantly, a significant increase in transposon elements in the pathogenic mycobacteria as compared with *MIP*, for the first time, suggests their possible role toward mycobacterial virulence and would be interesting to explore.

In addition, comparative genomic analysis revealed a higher antigenic potential of *MIP* subscribing to its unique ability for immunomodulation against various types of infections and presents a template to develop reverse genetics based approaches to design better strategies against mycobacterial infections.

## ACCESSION NUMBERS

*MIP* genome has been submitted to the genome depository at NCBI (accession no. CP002275).

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–5 and Supplementary Figure 1.

## REFERENCES

1. Sharma,P., Mukherjee,R., Talwar,G.P., Sarathchandra,K.G., Walia,R., Parida,S.K., Pandey,R.M., Rani,R., Kar,H., Mukherjee,A. *et al.* (2005) Immunoprophylactic effects of the anti-leprosy *Mw* vaccine in household contacts of leprosy patients: clinical field trials with a follow up of 8-10 years. *Lepr Rev.*, **76**, 127–143.

2. Talwar,G.P., Zaheer,S.A., Mukherjee,R., Walia,R., Misra,R.S., Sharma,A.K., Kar,H.K., Mukherjee,A., Parida,S.K., Suresh,N.R. *et al.* (1990) Immunotherapeutic effects of a vaccine based on a saprophytic cultivable Mycobacterium, *Mycobacterium w* in multibacillary leprosy patients. *Vaccine*, **8**, 121–129.

3. Guleria,I., Mukherjee,R. and Kaufmann,S.H. (1993) In vivo depletion of CD4 and CD8 T lymphocytes impairs *Mycobacterium w* vaccine-induced protection against *M. tubercolosis* in mice. *Med. Microbiol. Immunol.*, **182**, 129–135.

4. Saini,V., Raghuvanshi,S., Talwar,G.P., Ahmed,N., Khurana,J.P., Hasnain,S.E., Tyagi,A.K. and Tyagi,A.K. (2009) Polyphasic taxonomic analysis establishes *Mycobacterium indicus pranii* as a distinct species. *PLoS One*, **4**, e6263.

5. Rakshit,S., Ponnusamy,M., Papanna,S., Saha,B., Ahmed,A. and Nandi,D. (2011) Immunotherapeutic efficacy of *Mycobacterium indicus pranii* in eliciting anti-tumor T cell responses: critical roles of IFN gamma. *Int. J. Cancer*, **130**, 865–875.

6. Talwar,G.P. (1999) An immunotherapeutic vaccine for multibacillary leprosy. *Int. Rev. Immunol.*, **18**, 229–249.

7. Ahmad,F., Mani,J., Kumar,P., Haridas,S., Upadhyay,P. and Bhaskar,S. (2011) Activation of anti-tumor immune response and reduction of regulatory T cells with *Mycobacterium indicus pranii* (*MIP*) therapy in tumor bearing mice. *PLoS One*, **6**, e25424.

8. Katoch,K., Singh,P., Adhikari,T., Benara,S.K., Singh,H.B., Chauhan,D.S., Sharma,V.D., Lavania,M., Sachan,A.S. and Katoch,V.M. (2008) Potential of *Mw* as a prophylactic vaccine against pulmonary tuberculosis. *Vaccine*, **26**, 1228–1234.

9. Zaheer,S.A., Mukherjee,R., Ramkumar,B., Misra,R.S., Sharma,A.K., Kar,H.K., Kaur,H., Nair,S., Mukherjee,A. and Talwar,G.P. (1993) Combined multidrug and *Mycobacterium w* vaccine therapy in patients with multibacillary leprosy. *J. Infect. Dis.*, **167**, 401–410.

10. Singh,I.G., Mukherjee,R. and Talwar,G.P. (1991) Resistance to intravenous inoculation of *Mycobacterium tuberculosis* H37Rv in mice of different inbred strains following immunization with a leprosy vaccine based on *Mycobacterium w*. *Vaccine*, **9**, 10–14.

11. Zaheer,S.A., Beena,K.R., Kar,H.K., Sharma,A.K., Misra,R.S., Mukherjee,A., Mukherjee,R., Kaur,H., Pandey,R.M., Walia,R. *et al.* (1995) Addition of immunotherapy with *Mycobacterium w* vaccine to multi-drug therapy benefits multibacillary leprosy patients. *Vaccine*, **13**, 1102–1110.

12. Singh,I.G., Mukherjee,R., Talwar,G.P. and Kaufmann,S.H. (1992) In vitro characterization of T cells from *Mycobacterium w*-vaccinated mice. *Infect Immun.*, **60**, 257–263.

13. Gupta,A., Geetha,N., Mani,J., Upadhyay,P., Katoch,V.M., Natrajan,M., Gupta,U.D. and Bhaskar,S. (2009) Immunogenicity and protective efficacy of "*Mycobacterium w*" against *Mycobacterium tuberculosis* in mice immunized with live versus heat-killed *M. w* by the aerosol or parenteral route. *Infect. Immun.*, **77**, 223–231.

14. Patel,N., Deshpande,M.M. and Shah,M. (2002) Effect of an immunomodulator containing *Mycobacterium w* on sputum conversion in pulmonary tuberculosis. *J. Indian Med. Assoc.*, **100**, 191–193.

15. Nyasulu,P.S. (2011) The role of adjunctive *Mycobacterium w* immunotherapy for tuberculosis. *J. Exp. Clin. Med.*, **2**, 124–129.

16. Kharkar,R. (2002) Immune recovery in HIV with *Mycobacterium w*. *J. Indian Med. Assoc.*, **100**, 578–579.

17. Marsh,B.J., Von Reyn,C.F., Arbeit,R.D. and Morin,P. (1997) Immunization of HIV-infected adults with a three-dose series of inactivated *Mycobacterium vaccae*. *Am. J. Med. Sci.*, **313**, 377–383.

18. Nath,I. (1998) A vaccine for leprosy. *Nat. Med.*, **4**, 548–550.

19. Ahmed,N., Saini,V., Raghuvanshi,S., Khurana,J.P., Tyagi,A.K. and Hasnain,S.E. (2007) Molecular analysis of a leprosy immunotherapeutic bacillus provides insights into Mycobacterium evolution. *PLoS One*, **2**, e968.

20. Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eiglmeier,K., Gas,S., Barry,C.E. *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537–544.

21. Cole,S.T., Eiglmeier,K., Parkhill,J., James,K.D., Thomson,N.R., Wheeler,P.R., Honoré,N., Garnier,T., Churcher,C., Harris,D.

*et al.* (2001) Massive gene decay in the leprosy *bacillus*. *Nature*, **409**, 1007–1011.

22. Garnier,T., Eiglmeier,K., Camus,J.C., Medina,N., Mansoor,H., Pryor,M., Duthoy,S., Grondin,S., Lacroix,C., Monsempe,C. *et al.* (2003) The complete genome sequence of *Mycobacterium bovis*. *Proc. Natl. Acad. Sci. USA*, **100**, 7877–7882.

23. Stinear,T.P., Seemann,T., Pidot,S., Frigui,W., Reysset,G., Garnier,T., Meurice,G., Simon,D., Bouchier,C., Ma,L. *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.

24. Jin,Q., Yuan,Z., Xu,J., Wang,Y., Shen,Y., Lu,W., Wang,J., Liu,H., Yang,J., Yang,F. *et al.* (2002) Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res.*, **30**, 4432–4441.

25. Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.

26. Gardy,J.L., Laird,M.R., Chen,F., Rey,S., Walsh,C.J., Ester,M. and Brinkman,F.S. (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*, **21**, 617–623.

27. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

28. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

29. Remm,M., Storm,C.E.V. and Sonnhammer,E.L.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.

30. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

31. Zhang,Z., Li,J., Zhao,X.Q., Wang,J., Wong,G.K.S. and Yu,J. (2006) KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics*, **4**, 259–263.

32. Vernikos,G.S. and Parkhill,J. (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics*, **22**, 2196–2203.

33. Fertil,B., Massin,M., Lespinats,S., Devic,C., Dumee,P. and Giron,A. (2005) GENSTYLE: exploration and analysis of DNA sequences with genomic signature. *Nucleic Acids Res*, **33**, W512–W515.

34. Kuroda,M., Ohta,T., Uchiyama,I., Baba,T., Yuzawa,H., Kobayashi,I., Cui,L., Oguchi,A., Aoki,K., Nagai,Y. *et al.* (2001) Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. *Lancet*, **357**, 1225–1240.

35. Dufraigne,C., Fertil,B., Lespinats,S., Giron,A. and Deschavanne,P. (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res.*, **33**, e6.

36. Leplae,R., Hebrant,A., Wodak,S.J. and Toussaint,A. (2004) ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic Acids Res.*, **32**, D45–D49.

37. Grissa,I., Vergnaud,G. and Pourcel,C. (2008) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.*, **7**, W52–W57.

38. Ren,Q., Kang,K.H. and Paulsen,I.T. (2004) TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Res.*, **32**, D284–D288.

39. Liang,X., Pham,X.Q.T., Olson,M.V. and Lory,S. (2001) Identification of a genomic island present in the majority of pathogenic isolates of *Pseudomonas aeruginosa*. *J. Bacteriol.*, **183**, 843–853.

40. Chen,L., Yang,J., Yu,J., Yao,Z., Sun,L., Shen,Y. and Jin,Q. (2005) VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.*, **33**, D325–D328.

41. Doytchinova,I.A. and Flower,D.R. (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics*, **8**, 4.

42. Pagni,M., Ioannidis,V., Cerutti,L., Zahn-Zabal,M., Jongeneel,C.V. and Falquet,L. (2004) MyHits: a new interactive resource for protein annotation and domain identification. *Nucleic Acids Res.*, **32**, W332.

43. Dereeper,A., Guignon,V., Blanc,G., Audic,S., Buffet,S., Chevenet,F., Dufayard,J.F., Guindon,S., Lefort,V., Lescot,M. *et al.* (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.*, **36**, W465–W469.

44. Betts,J.C., Lukey,P.T., Robb,L.C., McAdam,R.A. and Duncan,K. (2002) Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. *Mol. Microbiol.*, **43**, 717–731.

45. Djelouadji,Z., Raoult,D. and Drancourt,M. (2011) Palaeogenomics of *Mycobacterium* tuberculosis: epidemic bursts with a degrading genome. *Lancet Infect. Dis.*, **11**, 641–650.

46. Page,R.D.M. and Holmes,E.C. (1998) *Molecular Evolution: A Phylogenetic Approach*. Wiley-Blackwell, New York, pp. 11–280.

47. Draskovic,I. and Dubnau,D. (2005) Biogenesis of a putative channel protein, ComEC, required for DNA uptake: membrane topology, oligomerization and formation of disulphide bonds. *Mol. Microbiol.*, **55**, 881–896.

48. Namouchi,A., Didelot,X., Schock,U., Gicquel,B. and Rocha,E.P. (2012) After the bottleneck: genome-wide diversification of the *Mycobacterium tuberculosis* complex by mutation, recombination, and natural selection. *Genome Res.*, **4**, 721–734.

49. Ochman,H., Lawrence,J.G. and Groisman,E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.

50. Becq,J., Churlaud,C. and Deschavanne,P. (2010) A benchmark of parametric methods for horizontal transfers detection. *PloS One.*, **5**, e9989.

51. Zaneveld,J.R., Nemergut,D.R. and Knight,R. (2008) Are all horizontal gene transfers created equal? Prospects for mechanism-based studies of HGT patterns. *Microbiology*, **154**, 1–15.

52. Ragan,M.A., Harlow,T.J. and Beiko,R.G. (2006) Do different surrogate methods detect lateral genetic transfer events of different relative ages? *Trends Microbiol.*, **14**, 4–8.

53. Shi,S.Y., Cai,X.H. and Ding,D.F. (2005) Identification and categorization of horizontally transferred genes in prokaryotic genomes. *Acta Biochim Biophys Sin (Shanghai)*, **37**, 561–566.

54. Azad,R.K. and Lawrence,J.G. Towards more robust methods of alien gene detection. *Nucleic Acids Res.*, **39**, e56.

55. Garcia-Vallve,S., Guzman,E., Montero,M.A. and Romeu,A. (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res.*, **31**, 187–189.

56. Beiko,R.G., Harlow,T.J. and Ragan,M.A. (2005) Highways of gene sharing in prokaryotes. *Proc. Natl Acad. Sci. USA*, **102**, 14332–14337.

57. Andersson,A.F. and Banfield,J.F. (2008) Virus population dynamics and acquired virus resistance in natural microbial communities. *Science*, **320**, 1047–1050.

58. Fang,Z., Morrison,N., Watt,B., Doig,C. and Forbes,K.J. (1998) IS6110 transposition and evolutionary scenario of the direct repeat locus in a group of closely related *Mycobacterium tuberculosis* strains. *J. Bacteriol.*, **180**, 2102–2109.

59. Flint,J.L., Kowalski,J.C., Karnati,P.K. and Derbyshire,K.M. (2004) The RD1 virulence locus of *Mycobacterium tuberculosis* regulates DNA transfer in *Mycobacterium smegmatis*. *Proc. Natl Acad. Sci. USA*, **101**, 12598–12603.

60. Parsons,J.A., Bannam,T.L., Devenish,R.J. and Rood,J.I. (2007) TcpA, an FtsK/SpoIIIE homolog, is essential for transfer of the conjugative plasmid pCW3 in *Clostridium perfringens*. *J Bacteriol.*, **189**, 7782–7790.

61. Mackenzie,N., Alexander,D.C., Turenne,C.Y., Behr,M.A. and De Buck,J.M. (2009) Genomic comparison of PE and PPE genes in the *Mycobacterium avium* complex. *J. Clin. Microbiol.*, **47**, 1002–1011.

62. French,C.E., Bell,J.M.L. and Ward,F.B. (2008) Diversity and distribution of hemerythrin-like proteins in prokaryotes. *FEMS Microbiol. Lett.*, **279**, 131–145.
63. Terwilliger,N.B. (1998) Functional adaptations of oxygen-transport proteins. *J. Exp. Biol.*, **201**, 1085–1098.
64. Mukhopadhyay,A., Panda,A.K. and Pandey,A.K. (1998) Leprosy vaccine: influence of dissolved oxygen levels on growth of a candidate strain (*Mycobacterium w*), and storage stability of the vaccine. *Vaccine*, **16**, 1344–1348.
65. Busti,E., Cavaletti,L., Monciardini,P., Schumann,P., Rohde,M., Sosio,M. and Donadio,S. (2006) *Catenulispora acidiphila* gen. nov., sp. nov., a novel, mycelium-forming actinomycete, and proposal of *Catenulisporaceae* fam. nov. *Int. J. Syst. Evol. Microbiol.*, **56**, 1741–1746.
66. Sachdeva,P., Misra,R., Tyagi,A.K. and Singh,Y. (2010) The sigma factors of *Mycobacterium tuberculosis*: regulation of the regulators. *FEBS J.*, **277**, 605–626.
67. Breinig,S., Schiltz,E. and Fuchs,G. (2000) Genes involved in anaerobic metabolism of phenol in the bacterium *Thauera aromatica*. *J. Bacteriol.*, **182**, 5849–5863.
68. Ebbs,S. (2004) Biological degradation of cyanide compounds. *Curr. Opin. Biotechnol.*, **15**, 231–236.
69. Park,S.S. and DeCicco,B.T. (1976) Hydrogenase and ribulose diphosphate carboxylase during autotrophic, heterotrophic, and mixotrophic growth of scotochromogenic mycobacteria. *J Bacteriol.*, **127**, 731–738.
70. Takayama,K., Wang,C. and Besra,G.S. (2005) Pathway to synthesis and processing of mycolic acids in *Mycobacterium tuberculosis*. *Clin. Microbiol. Rev.*, **18**, 81–101.
71. Onwueme,K.C., Ferreras,J.A., Buglino,J., Lima,C.D. and Quadri,L.E.N. (2004) Mycobacterial polyketide-associated proteins are acyltransferases: proof of principle with *Mycobacterium tuberculosis* PapA5. *Proc. Natl Acad. Sci. USA*, **101**, 4608–4613.
72. Tizard,M., Bull,T., Millar,D., Doran,T., Martin,H., Sumar,N., Ford,J., Hermon-Taylor,J. *et al.* (1998) A low G+C content genetic island in *Mycobacterium avium* subsp. *Paratuberculosis* and *M. avium* subsp. *Silvaticum* with homologous genes in Mycobacterium tuberculosis. *Microbiology*, **144**, 3413–3423.
73. Talwar,G.P., Ahmed,N. and Saini,V. (2008) The use of the name *Mycobacterium w* for the leprosy immunotherapeutic bacillus creates confusion with *M. tuberculosis-W* (Beijing strain): a suggestion. *Infect. Genet. Evol.*, **8**, 100–101.
74. Yoon,S.H., Park,Y.K., Lee,S., Choi,D., Oh,T.K., Hur,C.G. and Kim,J.F. (2007) Towards pathogenomics: a web-based resource for pathogenicity islands. *Nucleic Acids Res.*, **35**, D395–D400.
75. Waterfield,N.R. and Daborn,P.J. (2002) Genomic islands in *Photorhabdus*. *Trends Microbiol.*, **10**, 541–545.
76. Ahmad,S., El-Shazly,S., Mustafa,A.S. and Al-Attiyah,R. (2004) Mammalian cell-entry proteins encoded by the mce3 operon of *Mycobacterium tuberculosis* are expressed during natural infection in humans. *Scand. J. Immunol.*, **60**, 382–391.
77. Marjanovic,O., Miyata,T., Goodridge,A., Kendall,L.V. and Riley,L.W. (2009) Mce2 operon mutant strain of *Mycobacterium tuberculosis* is attenuated in C57BL/6 mice. *Tuberculosis*, **90**, 50–56.
78. Aguilar,L.D., Infante,E., Bianco,M.V., Cataldi,A., Bigi,F. and Pando,R.H. (2006) Immunogenicity and protection induced by *Mycobacterium tuberculosis* mce-2 and mce-3 mutants in a Balb/c mouse model of progressive pulmonary tuberculosis. *Vaccine*, **24**, 2333–2342.
79. Gomez,A., Mve-Obiang,A., Vray,B., Rudnicka,W., Shamputa,I.C., Portaels,F., Meyers,W.M., Fonteyne,P.A. and Realini,L. (2001) Detection of phospholipase C in nontuberculous mycobacteria and its possible role in hemolytic activity. *J. Clin. Microbiol.*, **39**, 1396–1401.
80. Converse,P.J., Karakousis,P.C., Klinkenberg,L.G., Kesavan,A.K., Ly,L.H., Allen,S.S., Grosset,J.H., Jain,S.K., Lamichhane,G. and Manabe,Y.C. (2009) Role of the dosR-dosS two-component regulatory system in *Mycobacterium tuberculosis* virulence in three animal models. *Infect. Immun.*, **77**, 1230–1237.
81. Smith,D.A., Parish,T., Stoker,N.G. and Bancroft,G.J. (2001) Characterization of auxotrophic mutants of *Mycobacterium tuberculosis* and their potential as vaccine candidates. *Infect. Immun.*, **69**, 1142–1150.
82. Smith,I. (2003) *Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence. *Clin. Microbiol. Rev.*, **16**, 463–496.
83. Nerland,A.H., Mustafa,A.S., Sweetser,D., Godal,T. and Young,R.A. (1988) A protein antigen of *Mycobacterium leprae* is related to a family of small heat shock proteins. *J. Bacteriol.*, **170**, 5919–5921.
84. Geluk,A., Klein,M.R., Franken,K., van Meijgaarden,K.E., Wieles,B., Pereira,K.C., Bührer-Sékula,S., Klatser,P.R., Brennan,P.J., Spencer,J.S. *et al.* (2005) Postgenomic approach to identify novel *Mycobacterium leprae* antigens with potential to improve immunodiagnosis of infection. *Infect. Immun.*, **73**, 5636–5644.