Insights into horizontal acquisition patterns of dormancy and reactivation regulon genes in mycobacterial species using a partitioning-based framework

VARUN MEHRA, TARINI SHANKAR GHOSH and SHARMILA S MANDE*

Bio-sciences R&D Division, TCS Research, Tata Consultancy Services Ltd., Pune 411 013, India

**Corresponding author (Email, sharmila.mande@tcs.com)*

Horizontal Gene Transfer (HGT) events, initially thought to be rare in *Mycobacterium tuberculosis*, have recently been shown to be involved in the acquisition of virulence operons in *M. tuberculosis*. We have developed a new partitioning framework based HGT prediction algorithm, called Grid3M, and applied the same for the prediction of HGTs in Mycobacteria. Validation and testing using simulated and real microbial genomes indicated better performance of Grid3M as compared with other widely used HGT prediction methods. Specific analysis of the genes belonging to dormancy/reactivation regulons across 14 mycobacterial genomes indicated that horizontal acquisition is specifically restricted to important accessory proteins. The results also revealed Burkholderia species to be a probable source of HGT genes belonging to these regulons. The current study provides a basis for similar analyses investigating the functional/evolutionary aspects of HGT genes in other pathogens. A database of Grid3M predicted HGTs in completely sequenced genomes is available at *https://metagenomics.atc.tcs.com/Grid3M/*.

[Mehra V, Ghosh TS and Mande SS 2016 Insights into horizontal acquisition patterns of dormancy and reactivation regulon genes in mycobacterial species using a partitioning-based framework. *J. Biosci.* **41** 475–485]

1. Introduction

Microbes are one of the most diverse organisms on our planet which inhabit even the uninhabitable environments. Constant evolution of microbial genomes, through events like mutations and genomic rearrangements (such as insertions, deletions or recombination events), allow them to gradually adapt to extreme ecological niches. On the other hand, their remarkable capability to respond quickly to sudden environmental pressures (for example, upon antibiotic treatment) has been mainly attributed to their ability to acquire genomic regions from other distantly related microbes or phages or plasmids (Doolittle 1999; Ochman *et al.* 2000; Dutta and Pan 2002). Occurrence of this phenomena, referred to as Horizontal (or Lateral) Gene Transfer (HGT), was inferred from comparative analyses of the genomes of different organisms (Ochman *et al.* 2000; Dutta and Pan 2002). HGT events allow large regions of foreign DNA from 'donor' genome(s) to be inserted into the native 'recipient' genome(s) (Doolittle 1999; Ochman *et al.* 2000; Dutta and Pan 2002). Such inserted regions are also referred to as Genomic Islands, or GIs. Some of the important functions encoded in these GIs include antibiotic resistance, virulence properties, and specialized metabolic functions. Such functional islands enhance the survival chances of the recipient organisms in diverse environments (Ochman *et al.* 2000; Shrivastava *et al.* 2010).

Most of the pathogenicity islands in pathogenic bacterial genomes are reported to be acquired through HGT phenomena. This phenomenon was initially thought to be rare in mycobacteria (Rosas-Magallanes *et al.* 2006; Becq *et al.* 2007). However, some of the experimental as well as com-

Keywords. Dormancy and reactivation; horizontal gene transfer; *Mycobacterium tuberculosis*; partitioning framework

Supplementary materials pertaining to this article are available on the Journal of Biosciences Website.

DOI: 10.1007/s12038-016-9622-0

Published online: 22 July 2016

putational studies have reported instances of HGT in *M. tuberculosis* (Kinsella *et al.* 2003; Rosas-Magallanes *et al.* 2006; Becq *et al.* 2007), a pathogen causing tuberculosis in humans. For example, virulence operon, Rv0986-88, encoding ATP-binding cassette (ABC) transporter, has been reported to be horizontally acquired in ancestor of *M. tuberculosis* from γ-proteobacteria (Rosas-Magallanes *et al.* 2006). Similarly some of the other virulence islands, reported to have been acquired through HGT events in Tubercle bacilli, have also been suggested to be responsible for the evolution of Mtb-complex genomes (Becq *et al.* 2007). In addition, phylogenetic-based approach has indicated the horizontal acquisition of fatty acid synthesis genes from α-proteobacteria, followed by their adaptive evolution in *M. tuberculosis* (Kinsella *et al.* 2003). Although these studies indicate probable HGTs of certain genes in *M. tuberculosis*, identification of other HGT regions as well as their probable source (donor organism) may aid in furthering our current understanding of the pathogenicity of this organism.

Different computational methods have been developed for identifying HGT regions in microbial genomes. For example, various studies have utilized oligonucleotide composition based parametric measures (e.g. G+C content, codon usage, higher/variable order oligonucleotide frequencies) to identify compositionally distinct regions within a given microbial genome (Lawrence and Ochman 1998; Ochman *et al.* 2000; Tsirigos and Rigoutsos 2005; Vernikos and Parkhill 2006; Rajan *et al*. 2007; Shrivastava *et al.* 2010; Becq *et al.* 2011). These strategies are based on the observation that the acquired GIs generally have an oligonucleotide usage pattern which is distinct from that of the native recipient genome (Ochman *et al.* 2000; Dutta and Pan 2002). In spite of the availability of these methods, it has been observed that some of them under-perform in certain genome specific scenarios (Vernikos and Parkhill 2006; Becq *et al.* 2011). For example, some of the parametric measures fail to efficiently distinguish between the foreign and the native regions of genome when the foreign regions originate from a closely related genome (Becq *et al.* 2011). In addition, none of the methods mentioned above have the ability to identify the probable origins (donor organism) of the predicted HGT regions for a given genome.

In this study, we have evaluated the HGT regions of *M. tuberculosis* using a novel HGT prediction algorithm, called Grid3M. Grid3M has been developed as part of this paper to address the above-mentioned limitations. Apart from utilizing the tetranucleotide frequencies of genomic fragments, the algorithm specifically takes into account the compositional heterogeneity of any genome by considering all sequenced microbial genomes together in a single framework. We have evaluated and compared Grid3M algorithm on various simulated as well as real microbial genomes. Apart from the Grid3M algorithm, results of our analysis using Grid3M on *M. tuberculosis* as well as other mycobacterial genomes are also presented in this paper.

2. Results

2.1 *Testing of Grid3M algorithm*

Figure 1 summarizes the results of Grid3M algorithm in detecting the well annotated cytotoxicity-associated pathogenicity island (cagPAI) (Tegtmeyer *et al.* 2011) of genomes of six *Helicobacter pylori* strains as HGT and their comparisons with two well-known HGT prediction methods, namely, Alien Hunter (Vernikos and Parkhill 2006) and Wn (Chi) (Tsirigos and Rigoutsos 2005). Grid3M algorithm was able to predict the entire cagPAI in all the six strains of *H. pylori*. On the other hand, while Alien Hunter could predict the entire cagPAI in five out of the six strains, Wn (Chi) was unable to predict the complete cagPAI in any of the six strains. Apart from predicting the entire cagPAI as HGT, performance of Grid3M in identifying the HGT regions in various simulated genomes was also observed to be better as compared to the above two algorithms (details summarized in supplementary text 1). These results indicate the suitability of Grid3M algorithm in identifying HGT regions of any length in any given genome.

2.2 *Predicted HGTs in M. tuberculosis H37Rv genome*

Figure 2 highlights the major category of genes, predicted as HGTs by Grid3M, in *M. tuberculosis* H37Rv genome. A total of 691 genes were predicted as HGTs. These included 44 genes belonging to PE, PPE and PE-PGRS family. The HGT genes also included 18 ribosomal proteins, 8 belonging to 30S ribosomal proteins and 10 belonging to 50S ribosomal proteins. Eight CRISPR-associated proteins, known to be involved in inhibiting phage-mediated HGT transfers in bacterial genomes, were also detected as HGTs. Interestingly, Rv0986-88 virulence operon, encoding an ABC transporter, which was earlier suggested to be acquired from γ-proteobacteria (Rosas-Magallanes *et al.* 2006), was also identified as HGT by Grid3M. Detection of all the above genes which are known to be involved in HGT events indicates the efficacy of Grid3M approach in identifying HGT regions.

Apart from identifying the known HGT genes in *M. tuberculosis*, Grid3M also identified a few interesting mycobacterial genes as HGTs. Such HGT genes included five ABC transporters (Rv1273c, Rv1667c, Rv1819c, Rv1877 and Rv2333c) which have recently been re-annotated to be involved in drug resistance. For example, one of these proteins, namely Rv1667c, has been annotated

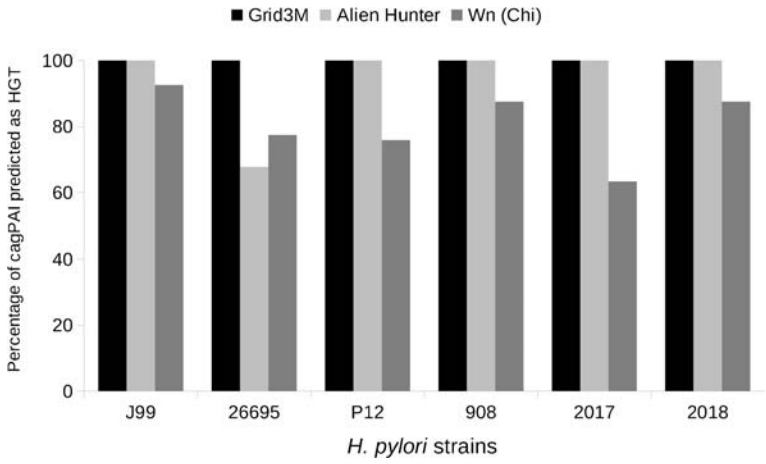

Figure 1. Percentage of cagPAI region predicted as HGT in six strains of *H. pylori* by Grid3M, Alien Hunter and Wn (Chi).

to encode a macrolide resistance gene. In addition, Grid3M also identified 22 genes encoding proteins of fatty acid (fad) metabolism, six genes encoding inner membrane transporter (mmpL) proteins, seven genes encoding polyketide synthesis (pks) proteins and eight genes encoding lipoprotein synthesis (lpp or lpq) proteins. Eight genes encoding proteins from the fatty acid biosynthesis proteins in *M. tuberculosis* was reported earlier to be horizontally acquired from α-proteobacteria (Kinsella *et al.* 2003). Five of these eight genes matched with the Grid3M-predicted HGT genes. Thus, the results of the predicted HGTs indicate that most of the genes related to fatty acid, polyketide or lipoprotein metabolism in *M. tuberculosis* are possibly being horizontally acquired from other bacteria.

Additionally, 19 genes corresponding to various known components of type-VII secretion system (T7SS) in *M. tuberculosis* were also detected as HGT genes. While three out of these 19 HGT genes were observed to be associated with ESX-5 transport machinery of T7SS, the remaining 16 HGT genes corresponded to the single GI composed of the region ranging from Rv3869-84 in *M. tuberculosis* H37Rv genome. The T7SS island formed by these 16 HGT genes contained genes encoding for ESX-1 transport machinery (including the effector proteins ESAT-6 and CFP-10). The probable HGT events of genes belonging to ESX-1 and ESX-5 regions, identified by Grid3M, is interesting since the genes belonging to these regions are known to enhance virulence properties of *M. tuberculosis* (Abdallah *et al.* 2007; Das *et al.* 2011).

A number of genes belonging to the reactivation regulon of *M. tuberculosis* were also predicted by Grid3M to have been acquired through HGT. *M. tuberculosis* has the

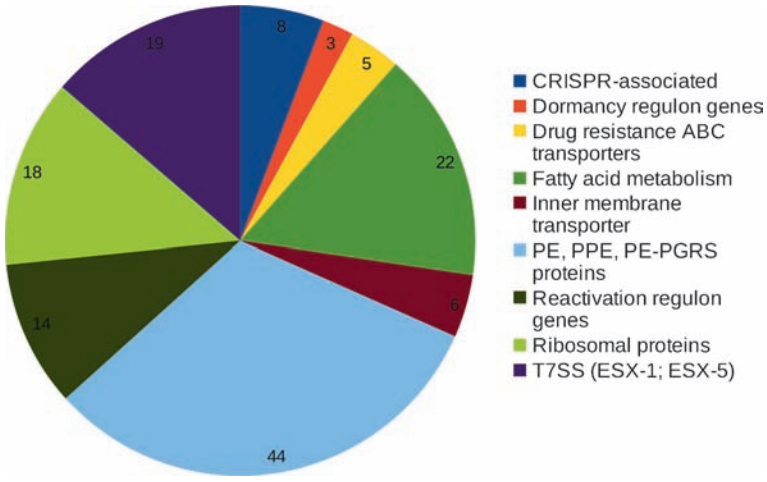

Figure 2. Pie-chart showing various categories of genes predicted as HGT by Grid3M algorithm in *M. tuberculosis* H37Rv.

remarkable ability to survive in host cells for months to decades without causing potent infection and observed to reactivate only in 2–10% of the individuals (Sherrid *et al.* 2010; Hegde *et al.* 2012). The genes belonging to the reactivation regulon are known to bring the pathogen out of the state of dormancy and cause virulence in host cells. Grid3M predicted three dormancy regulon genes and 14 reactivation regulon genes as HGT in *M. tuberculosis* H37Rv. Given the key roles of the dormancy/reactivation genes in the virulence of *M. tuberculosis*, we subsequently focused on identifying HGTs in these regulon genes across all species belonging to the *Mycobacterium* genus.

2.3 *Horizontally acquired genes in dormancy and reactivation regulons in mycobacteria*

Orthologs of 50 dormancy (Comstock 1982) and 103 reactivation regulon genes (Manabe and Bishai 2000) of *M. tuberculosis* H37Rv in 14 mycobacterial species were first identified and subsequently evaluated for their HGT nature using Grid3M (details summarized in supplementary text 2). Previous studies have indicated that mycobacterial lineage was initially composed of rapidly growing species, followed by evolution of the slow-growing ones and eventually the species belonging to the Mtb-complex (van Pittius *et al.* 2006). Based on this, the 14 mycobacterial species were divided into two major groups, namely, species belonging to Mtb-complex and those belonging to non-Mtb-complex. While all Mtb-complex species belong only to slow-growing mycobacteria (SGM), the non-Mtb-complex species were further divided into slow-growing (SGM) and rapidly-growing mycobacteria (RGM), referred to as non-Mtb-SGM and non-Mtb-RGM species, respectively. Thus, the current study classified the 14 mycobacterial genomes into three major categories, namely, Mtb-SGM, non-Mtb-SGM and non-Mtb-RGM (figures 3 and 4). In these figures, the trees depicting the phylogenetic relationship between 14 mycobacterial species were constructed using 16S rRNA gene sequences corresponding to their respective genomes (details in supplementary text 2). Results of the identified HGTs for dormancy and reactivation regulon genes are explained below.

Dormancy regulon genes: Orthologs of only two out of the 50 dormancy regulon genes were predicted as HGTs in seven (out of 14) mycobacterial species (figure 3). One of these genes corresponded to Rv0570 (*nrdZ*), a gene encoding a ribonucleotide reductase which is known to be involved in DNA replication. The orthologs of this gene were detected as HGTs in all three categories of mycobacterial species (Mtb-SGM, non-Mtb-SGM and non-Mtb-RGM). The second predicted HGT gene corresponded to Rv3130 c (*tgs1*), a triacylglycerol synthase gene which is known to be involved in storage of lipids as energy source during latency period of *M. tuberculosis* (Galagan *et al.* 2013). However, with the exception of *M. smegmatis* (belonging to non-Mtb-RGM), this gene was predicted as HGT only in slow-growing mycobacterial species (Mtb-SGM and non-Mtb-SGM). The detection of this gene as HGT is important since a recent study has reported this gene (*tgs1*) to play a key role in the accumulation of triacylglycerides (TAGs) during hypoxia-induced latency (Galagan *et al.* 2013). These TAGs are utilized by the *Mycobacterium* during re-aeration (or reactivation) period for virulence in the host cells (Manabe and Bishai 2000; Galagan *et al.* 2013). The above two predicted HGTs in dormancy regulon indicate that important accessory proteins involved in DNA replication (*nrdZ*) and lipid storage (*tgs1*) might have been acquired in mycobacteria from other bacteria to efficiently survive during the dormancy and cause virulence.

Furthermore, seven dormancy regulon genes were commonly detected as HGT in three mycobacterial genomes, namely, *M. marinum* M, *M. ulcerans* Agy99 and *M. smegmatis* MC2 155 (figure 3). While *M. smegmatis* belonged to the category of non-Mtb-RGM species, *M. marinum* and *M. ulcerans* belonged to the category of non-Mtb-SGM species. Interestingly, five of these seven orthologs (Rv1996, Rv2005c, Rv2028c, Rv2623, Rv2624 and Rv3134c) belong to the family of universal stress proteins (USPs). For example, it has been reported that product of Rv2623 (*TB31-7*) binds to ATP and helps in chronic persistence of infection by *M. tuberculosis* (Drumm *et al.* 2009). Additionally, Rv3134 c encoding a hypothetical protein (HP), has been suggested to play an important role in adaptation to hypoxia in *M. tuberculosis* by participating in phosphorelay of two-component system of *dosS*/*dosR* (*devS*/*devR*) forming a Rv3134 c-devR-devS operon (Bagchi *et al.* 2005). These results suggest that the above three mycobacterial species might have acquired these important USPs in order to adapt a latency mechanism similar to that utilized by *M. tuberculosis* species.

Reactivation regulon genes: Among the 103 reactivation regulon genes, a set of seven genes (Rv0384c, Rv1013, Rv1218c, Rv1472, Rv3515c, Rv3530c and Rv3545c) were predicted as HGT in 12 out of 14 mycobacterial species with representations from all three categories (Mtb-SGM, non-Mtb-SGM and non-Mtb-RGM species) (figure 4). One of these genes, namely, Rv0384c (*clpB*) encodes an ATPase subunit of an intracellular ATP-dependent protease and has been reported to be essential for *in vivo* survival and pathogenicity in *M. tuberculosis* (Ribeiro-Guimarães and Pessolani 2007). Three of these genes, namely, Rv1013 (*pks16*), Rv1472 (*echA12*) and Rv3515 c (*fadD19*) are known to be involved in polyketide and fatty acid metabolisms in mycobacteria. Interestingly, one of these seven HGT predicted gene (Rv1218) has also been reported to encode an ABC transporter which probably provides resistance against

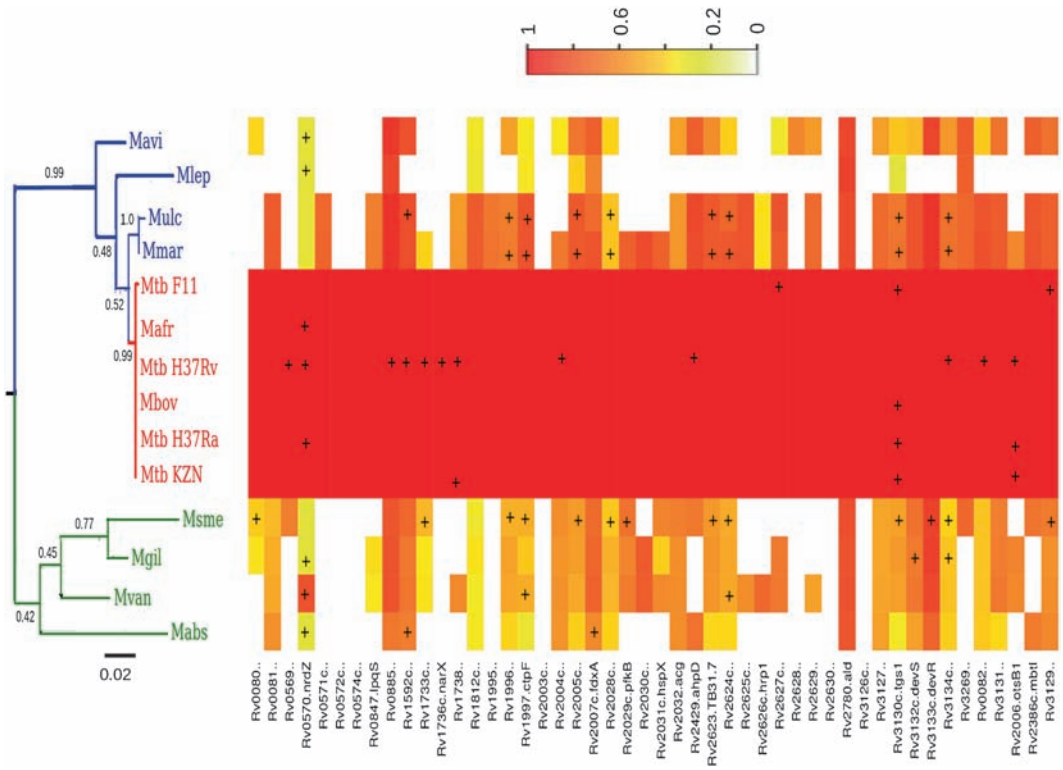

Figure 3. Heatmap summarizing the orthologs of 50 dormancy regulon genes and their HGT detection patterns in 14 mycobacterial genomes. 16S rRNA tree: Blue, Red and Green represent the non-Mtb-SGM, Mtb-SGM and non-Mtb-RGM, respectively. HGT nature: Symbol (+) represents those genes which are detected as HGT using Grid3M. Vertical colour scale: Percentage identity of a gene with its corresponding ortholog in the reference genome of *M. tuberculosis* H37Rv. While white represents that the ortholog of the gene is absent in that mycobacterial genome, deep red indicates 100% identity of the gene with its ortholog. Abbreviations used are: Mavi, *M. avium subsp. paratuberculosis* K-10, Mafr, *M. africanum* GM0411 82, Mbov, *M. bovis* BCG str. Pasteur 1173 P2, Mlep, *M. leprae* TN, Mmar, *M. marinum* M, Mtb F11, *M. tuberculosis* F11, Mtb H37Ra, *M. tuberculosis* H37Ra, Mtb H37Rv, *M. tuberculosis* H37Rv, Mtb KZN, *M. tuberculosis* KZN 1435, Mulc, *M. ulcerans* Agy99, Mabs, *M. abscessus* ATCC 1997 7, Mgil, *M. gilvum* PYR-GCK, Msme, *M. smegmatis* str. MC2 155, and Mvan, *M. vanbaalenii* PYR-1.

tetronasin antibiotic by an unknown export mechanism (Braibant *et al.* 2000). The detection of the above seven reactivation genes as HGT in all three categories of mycobacteria suggests that they might have been horizontally acquired in the mycobacterial lineage before any speciation event.

A set of seven additional reactivation genes (Rv0011c, Rv0350, Rv0351, Rv0352, Rv1168c, Rv1809 and Rv3429) were predicted as HGT in the slow-growing mycobacterial species, i.e. in Mtb-SGM and non-Mtb-SGM (figure 4). Three of these four HGT genes, namely, Rv1168c, Rv1809 and Rv3429 are PPE family proteins which are known to be specifically present in SGMs. Since the remaining three genes, namely, Rv0350, Rv0351 and Rv0352 are adjacently located on the *M. tuberculosis* H37Rv chromosome, they might have been acquired horizontally as part of a single GI. These three genes belong to the family of heat shock proteins and chaperons. During reactivation, these chaperons are suggested to play important roles in repair or replacement of proteins that are damaged by hypoxia-induced latency (Manabe and Bishai 2000). The important role played by these proteins in reactivation might have been a probable reason for their horizontal acquisition in slow-growing mycobacteria (SGM) from other bacteria to efficiently synthesize or repair genes required for revival and virulence.

Interestingly, another four HGT predicted genes (Rv0327c, Rv0753, Rv0791c and Rv3260c) were observed to be present in species belonging to non-Mtb complex, i.e. non-Mtb-SGM and non-Mtb-RGM (figure 4). While Rv0327c (*cyp135A1*) encodes for cytochrome P450-like monooxygenase, Rv0753c (*mmsA*) is known to play role in valine and pyrimidine metabolism. The third gene, Rv0791c

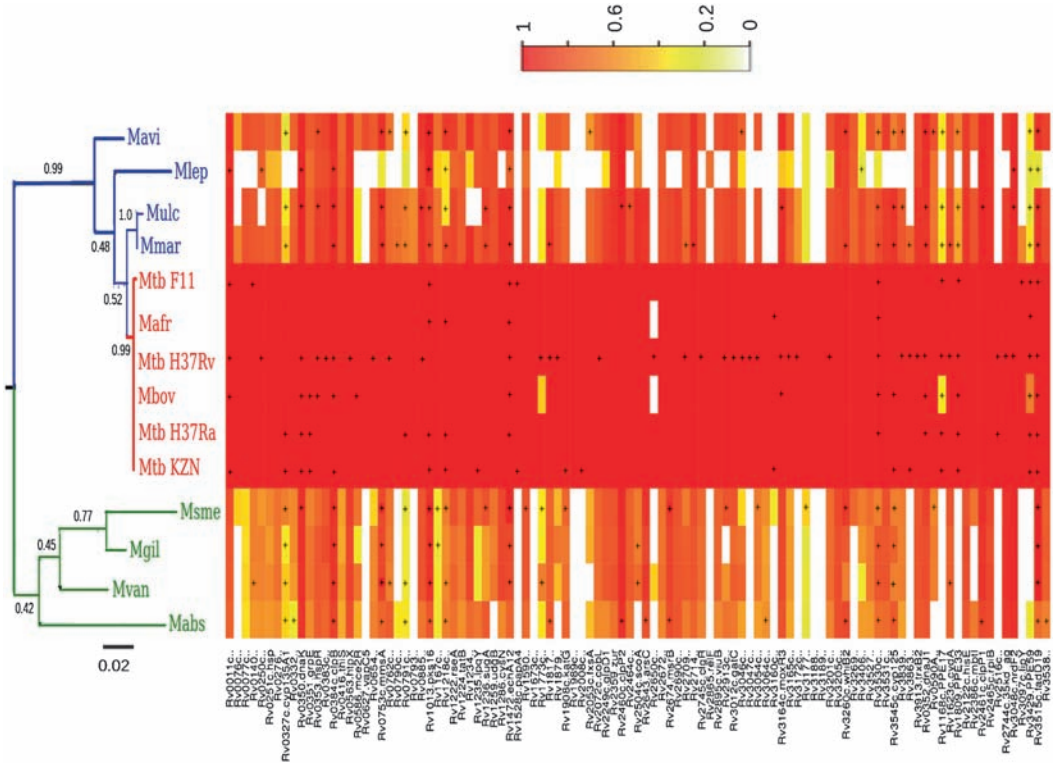

Figure 4. Heatmap summarizing the orthologs of 103 reactivation regulon genes and their HGT detection patterns in 14 mycobacterial genomes. 16S rRNA tree: Blue, Red and Green represent the non-Mtb-SGM, Mtb-SGM and non-Mtb-RGM, respectively. HGT nature: Symbol (+) represents those genes which are detected as HGT using Grid3M. Vertical colour scale: Percentage identity of a gene with its corresponding ortholog in the reference genome of *M. tuberculosis* H37Rv. While white represents that the ortholog of the gene is absent in that mycobacterial genome, deep red indicates 100% identity of the gene with its ortholog. Abbreviations used are: Mavi, *M. avium subsp. paratuberculosis* K-10, Mafr, *M. africanum* GM0411 82, Mbov, *M. bovis* BCG str. Pasteur 1173 P2, Mlep, *M. leprae* TN, Mmar, *M. marinum* M, Mtb F11, *M. tuberculosis* F11, Mtb H37Ra, *M. tuberculosis* H37Ra, Mtb H37Rv, *M. tuberculosis* H37Rv, Mtb KZN, *M. tuberculosis* KZN 1435, Mulc, *M. ulcerans* Agy99, Mabs, *M. abscessus* ATCC 1997 7, Mgil, *M. gilvum* PYR-GCK, Msme, *M. smegmatis* str. MC2 155, and Mvan, *M. vanbaalenii* PYR-1.

is a conserved HP whereas the fourth gene, *whiB2* (Rv3260c) is involved in cell division. Since these four genes are mainly required during reactivation for actively dividing cells, our results suggest that these genes might have been horizontally acquired in non-Mtb-complex species for quicker revival from latency.

A set of 40 genes were predicted as HGT only in *M. tuberculosis* H37Rv (figure 4). Interestingly, two of these genes, namely, Rv2745c (*clgR*) and Rv3164c (*moxR3*), are known to be major regulators of reactivation regulon (Manabe and Bishai 2000). For example, the transcription factor (ClgR) has been regarded as an important regulator which helps the *M. tuberculosis* to resume DNA replication during reactivation (Manabe and Bishai 2000). Thus, the above results indicate that some of the key regulators of the reactivation process in *M. tuberculosis* might have been acquired through HGT event.

2.4 *Probable source(s) of horizontally acquired genes in dormancy and reactivation regulons in mycobacteria*

Probable source of HGT genes, obtained using the Grid3M framework (details in section 4), is given below.

HGT genes in Dormancy regulon: Species belonging to *Burkholderia* and *Mycobacterium* were detected as probable sources for the HGT genes belonging to the dormancy regulon (supplementary figure 3). Species belonging to *Burkholderia* were predicted to be the probable source of the *nrdZ* gene which was detected as HGT in all three categories of mycobacteria (Mtb-SGM, non-Mtb-SGM and non-Mtb-RGM) (column A in supplementary figure 3). This result suggests that this gene probably was horizontally acquired from species belonging to *Burkholderia* before any speciation event in the mycobacterial lineage.

Species belonging to both *Burkholderia* and *Mycobacterium* were predicted as probable sources of horizontally acquired dormancy regulon genes in SGMs (Mtb-SGM and non-Mtb-SGM) as well as non-Mtb-complex (non-Mtb-SGM and non-Mtb-RGM) genomes (columns B and C respectively in supplementary figure 3). For example, in case of slow-growing mycobacteria (SGM), majority of the burkholderial species, along with two mycobacterial species (*M. africanum* and *M. abscessus*) were predicted as probable sources of horizontally acquired *tgs1* gene (column B in supplementary figure 3). On the other hand, in non-Mtb-complex genomes, species belonging to either *Burkholderia* or/and *Mycobacterium* were predicted as probable sources for seven HGT genes in dormancy regulon (column C in supplementary figure 3). The above results suggest that either intra-mycobacterial transfers or/and transfers from species belonging to *Burkholderia*, might play important roles in HGT events occurring either after the speciation event of slow-growing mycobacteria or after the delineation of species belonging to the Mtb-complex.

A few species belonging to *Burkholderia* were predicted to be the probable sources for the set of 13 dormancy genes detected as HGT only in *M. tuberculosis* H37Rv (column D in supplementary figure 3). In summary, all the above results suggest that most of the horizontally acquired dormancy regulon genes were probably obtained from species belonging to *Burkholderia* before the speciation within mycobacterial lineage. On the other hand, the dormancy genes horizontally acquired after the speciation were possibly transferred either from *Burkholderia* or other mycobacterial species.

HGT genes in Reactivation regulon: Unlike in the case of dormancy regulon, species belonging to six genera, namely, *Bordetella*, *Burkholderia*, *Methylobacterium*, *Mycobacterium*, *Pseudomonas* and *Streptomyces* were predicted as the probable sources of genes identified as HGT belonging to reactivation regulon (supplementary figure 4). The results indicate that the seven genes predicted as HGTs in all three categories of mycobacteria (Mtb-SGM, non-Mtb-SGM and non-Mtb-RGM) were equally likely to be acquired from species belonging to any of the six genera (column A in supplementary figure 4). However, all the species belonging to *Streptomyces*, a few species from *Pseudomonas* as well as *Methylobacterium* and most of the RGMs were predicted with higher likelihood as probable sources (column A in supplementary figure 4). This suggests that these seven HGT genes in reactivation regulon were probably not acquired from one single source, but may have been acquired from different sources. The presence of most of the species from RGMs might also indicate that probably some of the core genes required for coming out of dormancy was already present in the soil dwelling ancestral species of RGMs. These core reactivation genes might have been vertically transmitted from some ancestral RGM to other mycobacteria during the course of evolution. Species belonging to *Bordetella*, *Burkholderia* and *Pseudomonas* along with *M. leprae*, *M. africanum* and *M. abscessus* were predicted as the probable sources for seven horizontally acquired reactivation regulon genes in slow-growing mycobacteria (column B in supplementary figure 4). On the other hand, species belonging to *Streptomyces* and few species from *Burkholderia* and *Pseudomonas* were predicted as probable donors for four reactivation regulon genes detected as HGT in species belonging to non-Mtb-complex (column C in supplementary figure 4). Species from five genera, namely, *Bordetella*, *Burkholderia*, *Methylobacterium, Mycobacterium* and *Pseudomonas*, were identified as probable sources for gene transfer for the set of 40 reactivation genes predicted as HGT only in *M. tuberculosis* H37Rv (column D in supplementary figure 4).

In summary, the above results suggest that unlike in the case of HGTs in dormancy regulon, a multitude of species might act as donors for the HGT genes belonging to the reactivation regulon. For example, unlike dormancy HGT genes which have probable source mainly from one genus (*Burkholderia*), reactivation HGT genes have probably been acquired from various sources to achieve the present day regulon. Interestingly, genomes belonging to all the five genera (*Burkholderia*, *Pseudomonas*, *Bordetella*, *Streptomyces* and *Methylobacterium*) have also been reported to have genes required for both dormancy (or persistence) and reactivation (Gan 2005; Lewis 2010; Wood *et al.* 2013). Apart from this, no clear trend could be inferred regarding the probable sources of HGT events occurring at different stages of evolution of mycobacterial lineage.

3. Discussion

In the current study, we have analysed the nature of HGT genes in genomes of different mycobacterial species. Our HGT prediction method, Grid3M utilizes a novel partitioning-based framework for predicting HGT regions in microbial genomes. One of the important features of the algorithm pertains to judging the heterogeneity within a genome (irrespective of its length) and utilizing the same for prediction of HGT regions. The validation and testing of Grid3M, using simulated and real microbial genomes, indicated its high detection sensitivity, irrespective of the taxonomic origin, compositional characteristics and phylogenetic closeness of the donor genomes.

Using Grid3M, a total of 691 HGT genes were detected in *M. tuberculosis* H37Rv genome. Most of these predicted regions included genes corresponding to PE, PPE, PE-PGRS and a few virulence operons like Rv0986-88. The other genes identified as HGTs included drug resistance ABC transporters, fatty acid metabolism genes, inner

membrane transporters, lipoproteins and polyketide biosynthesis genes. Notably, the entire GI encoding for ESX-1 and ESX-5 transport machineries of T7SS were also identified as HGTs. These T7SS transport machineries are known to enhance the virulence properties of *M. tuberculosis* by secreting effector proteins which promote cell-to-cell migration of *M. tuberculosis* (Abdallah *et al.* 2007; Das *et al.* 2011). Interestingly, the current study, for the first time, predicted a number of horizontally acquired genes belonging to dormancy as well as reactivation genes, known to play a key role in the virulence of *M. tuberculosis*, across several mycobacterial species. However, the predicted HGTs in these regulons consisted of only accessory proteins. In other words, none of the core proteins (Comstock 1982; Manabe and Bishai 2000; Galagan *et al.* 2013) belonging to the dormancy (*dosR*, *dosS*, *narX* and hub protein Rv0081) and reactivation (*lpqY*, *sugA*, *zur*, *clgR* and *moxR3*) regulons were predicted as HGTs in most of the mycobacteria. One of the key results indicated that although the core genes belonging to both dormancy and reactivation regulons were part of the native genome of an ancestral *Mycobacterium* species (e.g. *M. abscessus*), a few important accessory genes in these regulons were probably acquired through horizontal gene transfers in order to enhance the survival capabilities during the dormancy and reactivation stages of mycobacteria. Another important prediction made from our results pertains to the role of HGT events in the likely adaptation of the latency mechanism (similar to that observed in *M. tuberculosis*) in three species belonging to non-Mtb-complex by horizontally acquiring a few important universal stress proteins. Given the importance of dormancy and reactivation regulon genes in the overall virulence of *M. tuberculosis*, these results provided interesting insights into the important yet hitherto unknown role played by HGT events in the genomic evolution of the mycobacterial clade.

A unique feature of the partitioning-based Grid3M framework is that it allowed the identification of the probable source organisms of an HGT region (at a desired taxonomic rank). The current study, for the first time, indicated the probable HGT sources in the dormancy and reactivation gene regulons in mycobacterial species. Another key finding of this study is that the HGTs in dormancy regulon were predicted to have been either acquired from species belonging to *Burkholderia* or transferred within the mycobacterial lineage. In contrast, a multitude of species were predicted to be the source of the HGT genes in the reactivation regulon. These included species belonging to six genera, namely, *Bordetella*, *Burkholderia*, *Methylobacterium*, *Mycobacterium*, *Pseudomonas* and *Streptomyces*. Interestingly, most of the members belonging to these genera have genes required for both dormancy (or persistence) and reactivation (Gan 2005; Lewis 2010; Wood *et al.* 2013). For example, *B. pseudomallei* has been reported to show characteristics of both dormancy and reactivation, similar to those shown by *M. tuberculosis* (Sherrid *et al.* 2010). Interestingly, *B. pseudomallei* has also been reported to cause infection in the respiratory tract (called Melioidosis), that has similar symptoms as Tuberculosis (Gan 2005). Some studies have also reported the co-occurrence of species belonging to *Burkholderia*, *Mycobacterium* and *Pseudomonas* in the sputum samples of patients having respiratory infections and diseases including cystic fibrosis (Torrens *et al.* 1998; Shetty *et al.* 2010; Lobo and Noone 2014). It is also known that *Burkholderia* and *Mycobacterium* are both soil-dwelling bacteria, and thus might have been involved in the horizontal transfer of few or more such genes due to physical proximity in the similar micro-environment. Previous studies have suggested that physical proximity rather than phylogenetic closeness plays a more important role in HGT events (Syvanen 1994; Dutta and Pan 2002). Thus, based on our results as well as inferences from previous studies, it is likely that the species belonging to the above genera, particularly *Burkholderia*, might have been the closest probable donors of HGT genes belonging to dormancy and reactivation regulons in mycobacteria. Thus, the analysis of the HGT regions predicted by Grid3M provided interesting insights into the role of horizontally acquired dormancy and reactivation regulon genes in various mycobacterial species. The predicted HGTs in these regulons probably play important roles in enhancing the survival and virulence properties of mycobacterial species. Similar analyses can also be performed to understand the functional as well as the evolutionary aspects of horizontally acquired genes in other pathogenic bacteria.

4. Methods

4.1 *Principle*

HGT regions in a genome are known to be compositionally distinct from the rest (native part) of the genome. Thus, if genomic fragments from all the sequenced microbial genomes are represented as three-dimensional (3D) coordinates (based on their oligonucleotide composition), then the native and HGT regions of a given genome are expected to localize in distinct regions in this 3D space. In addition, certain regions in the space will have higher abundances of fragments from the native regions of a given genome as compared to those from other microbial genomes. On the other hand, fragments from the HGT regions, being atypical from its native genome, are likely to be localized in regions having higher abundances of fragments from some other genome(s). The current algorithm employs a series of steps to identify such regions in order to identify possible HGT regions in a genome. The detailed description of all the steps is given below.

4.2 *Steps to predict HGT regions in a given genome*

The flowchart summarizing the pre-processing steps of Grid3M is shown in supplementary figure 1. The details of all the pre-processing steps are explained in supplementary text 3. The flowchart depicting the steps for predicting HGT regions in a given microbial genome is shown in supplementary figure 2. These HGT prediction steps are briefly summarized below whereas the additional details are given in supplementary text 4.

Step 1: Selection of grids tagged to a given genome

In the first step, the algorithm identifies all the grids tagged to the genome of interest. In other words, grids containing fragment points corresponding to this genome are identified.

Step 2: Classification of grids as Minority, Majority and Mixed grids

The pattern of localization of fragment points belonging to a particular genome (in this 'Grid' framework) is illustrated in figure 5. Based on the relative abundances of the fragment points belonging to a genome, a grid can be classified under three categories. These categories are described below.

In the first category, called the 'Majority grids', the fragment points originating from a given genome are observed to be higher in these grids than those originating from all other genomes. These grids are expected to contain genomic fragments that are native to that genome. On the other hand, the compositionally distinct regions of this genome are expected to be localized to those grids wherein the fragment points (originating from this genome) are in relatively lower numbers as compared to those belonging to other microbial genomes. Such grids are referred to as 'Minority grids'. The third category, referred to as 'Mixed grids', corresponds to those grids wherein the fragment points belonging to the genome under study are neither in minority nor in majority. In order to classify each grid into one of the three categories, appropriate thresholds of abundance values (corresponding to fragment points belonging to a genome) are used. For identifying these thresholds, a quantile function value based approach was used. This approach utilizes the abundance patterns of the fragment points in all the grids tagged to a genome. A detailed description of this approach is provided in supplementary text 4.

Step 3: Identification of probable HGT fragments using a centroid-based approach

This step is performed specifically on the minority and mixed grids corresponding to the genome of interest. In this step, the centroid of each (minority/mixed) grid is obtained. Subsequently, in each of these grids (referred to as the 'home' grids), the distance of each fragment point belonging to the given genome from the corresponding centroid is computed. Fragment points lying within a certain threshold distance from the corresponding centroid are identified as

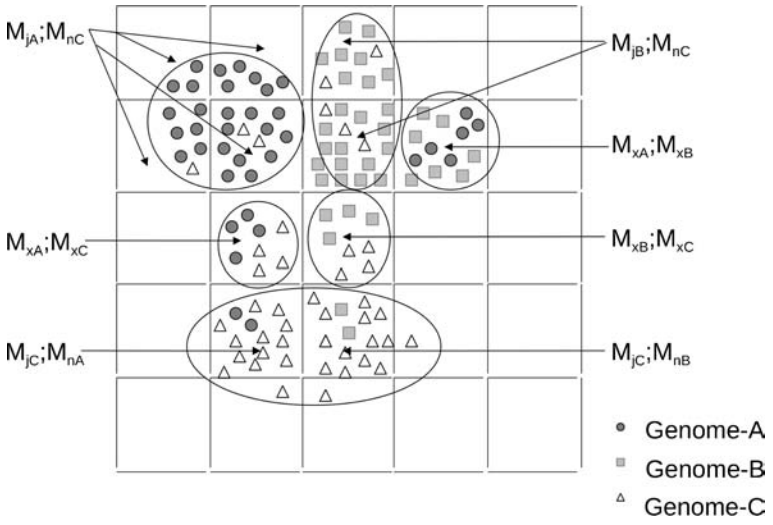

Figure 5. A representation of the Grid3M framework, indicating 'Majority', 'Minority' and 'Mixed' grids corresponding to three sample genomes (A, B, C). Mj, Mn and Mx denote the Majority, Minority and Mixed grids, respectively. For example, a grid tagged as 'MjA;MnC' denotes that the grid is a Majority grid for A and a Minority grid for C. Similarly, the tag 'MxA;MxB' denotes the grid to be Mixed grid for both 'A' and 'B'. Regions in the Grid3M framework having similar genome-level compositions of the constituting fragments are demarcated as circles.

probable HGT regions. The detailed methodology for this step is summarized in supplementary text 4.

4.3 *Identification of probable source(s) of all potential HGT genes*

The current study also probed for the probable source (or donor organism) for each of the predicted HGT gene using our framework of Grid3M. The strategy adopted for identification of probable source of each HGT gene using our Grid3M framework is summarized below.

Firstly, the percentage abundances of all the organisms localized in a particular grid (occupied by the respective HGT gene) were computed at a desired taxonomic level (like genus or species level). Further, weighted percentage abundance was computed for the desired taxonomic level based on the total number of hits obtained for that HGT gene in the given genome. Additionally, an overall percentage abundance of the desired taxonomic level was obtained for their respective occurrences in the entire Grid3M framework. Thereafter, a ratio of weighted percentage abundance to the overall percentage abundance (called as enrichment value) was computed for each desired taxonomic level present in the grids occupied by the potential HGT candidates. Further, a relative normalization was computed for the enrichment value of each taxa by dividing it by the total number of organisms in that taxa. And finally, any desired taxonomic level which was observed to be present in at least two-thirds of all the genes belonging to a respective group were regarded as potential probable source of HGT for that group of genes.

4.4 *Validation of Grid3M on simulated microbial genomes*

4.4.1 *Simulated microbial genome datasets:* Validation of Grid3M algorithm was performed using seven simulated microbial genomes corresponding to the species, *Archaeoglobus fulgidus, Bacillus subtilis, Escherichia coli, Methanococcus jannaschii, Neisseria gonorrhoeae, Ralstonia solanacearum,* and *Sinorhizobium meliloti.* These seven simulated microbial genomes were created as follows. Compositionally homogeneous genomes of the above seven species, artificially created in a previous study (Azad and Lawrence 2005), were obtained from the authors. Subsequently, in order to mimic HGT events, for each artificial genome, a corresponding simulated genome was created by inserting 'foreign' fragments (into the given genome) from each of the remaining six artificial genomes (referred to as donors). The details of various donors and recipient genomes are provided in supplementary table 1. Further explanation on creation of simulated genomes and details of the validation methodology and the corresponding results are given in detail in supplementary text 1.

4.4.2 *Testing of Grid3M using H. pylori genomes:* Grid3M algorithm was tested on six *H. pylori* strains, namely, J99 26695, P12, 908 2017 and 2018. The results obtained using the current method were also compared with those obtained using two other well-known HGT prediction methods, namely, Alien Hunter (Vernikos and Parkhill 2006) and Wn (Chi) (Tsirigos and Rigoutsos 2005). Since all HGT regions of *H. pylori* genomes are hitherto uncharacterized, the performances of the methods were compared based on their sensitivity for the detection of the well-annotated cagPAI (Tegtmeyer *et al.* 2005) of these six *H. pylori* genomes.

4.4.3 *Specific analysis on dormancy and reactivation regulon genes in mycobacteria:* The completely sequenced genomes of 14 mycobacterial species, belonging to both slow growing mycobacteria (SGM) and rapidly growing mycobacteria (RGM), were downloaded from NCBI (*ftp://ftp.ncbi.nih.gov/genomes/Bacteria/*) database (details in supplementary text 2). In the current study, a total of 153 genes including 50 dormancy upregulated genes (Comstock 1982) and 103 reactivation upregulated genes (Manabe and Bishai 2005) from *M. tuberculosis* H37Rv (considered as the 'reference genome') were analyzed for all possible occurrences of horizontal acquisition. The extensive list of all these 153 genes is provided in supplementary table 2. Orthologs of all the 153 genes were identified at the protein level, among all the 14 mycobacterial genomes, using the reciprocal Blast-based strategy (Altschul *et al.* 1990) (explained in supplementary text 2).

Acknowledgements

We thank Dr RK Azad and Dr JG Lawrence for providing artificial genome sequences for our validation studies. We also thank Mr Purnachander Gajjalla for his help in creating the simulated genomes. VM is also a Junior Research Fellow of the Department of Chemical Engineering, Indian Institute of Technology (IIT), Bombay, and would like to acknowledge Department of Chemical Engineering, IIT Bombay, for its support.

References

Abdallah AM, Gey van Pittius NC, Champion PAD, Cox J, Luirink J, Vandenbroucke-Grauls CMJE, Appelmelk BJ and Bitter W 2007 Type VII secretion–mycobacteria show the way. *Nat. Rev. Microbiol.* **5** 883–891

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ 1990 Basic local alignment search tool. *J. Mol. Biol.* **215** 403–410

Azad RK and Lawrence JG 2005 Use of artificial genomes in assessing methods for atypical gene detection. *PLoS Comput. Biol.* **1**, e56.c

Bagchi G, Chauhan S, Sharma D and Tyagi JS 2005 Transcription and autoregulation of the Rv3134 c-devR-devS operon of Mycobacterium tuberculosis. *Microbiology.* **151** 4045–4053

Becq J, Gutierrez MC, Rosas-Magallanes V, Rauzier J, Gicquel B, Neyrolles O and Deschavanne P 2007 Contribution of horizontally acquired genomic islands to the evolution of the tubercle bacilli. *Mol. Biol. Evol.* **24** 1861–1871

Becq J, Churlaud C and Deschavanne P 2011 A benchmark of parametric methods for horizontal transfers detection. *PLoS One* **5** e9989

Braibant M, Gilot P and Content J 2000 The ATP binding cassette (ABC) transport systems of Mycobacterium tuberculosis. *FEMS Microbiol. Rev.* **24** 449–467

Comstock GW 1982 Epidemiology of tuberculosis. *Am. Rev. Respir. Dis.* **125** 8–15

Das C, Ghosh TS and Mande SS 2011 Computational analysis of the ESX-1 region of Mycobacterium tuberculosis: insights into the mechanism of type VII secretion system. *PLoS One.* **6**, e2798

Doolittle WF 1999 Lateral genomics. *Trends Cell Biol.* **9** M5–8

Drumm JE, Mi K, Bilder P, Sun M, Lim J, *et al*. 2009 Mycobacterium tuberculosis universal stress protein Rv2623 regulates bacillary growth by ATP-binding: requirement for establishing chronic persistent infection. *PLoS Pathog.* **5** e1000460

Dutta C and Pan A 2002 Horizontal gene transfer and bacterial diversity. *J. Biosci.* **27** 27–33

Galagan JE, Minch K, Peterson M, Lyubetskaya A, Azizi E, *et al*. 2013 The Mycobacterium tuberculosis regulatory network and hypoxia. *Nature* **499** 178–183

Gan Y-H 2005 Interaction between Burkholderia pseudomallei and the host immune response: sleeping with the enemy? *J. Infect. Dis.* **192** 1845–1850

Hegde SR, Rajasingh H, Das C, Mande SS and Mande SC 2012 Understanding communication signals during mycobacterial latency through predicted genome-wide protein interactions and boolean modeling. *PLoS One* **7** e33893

Kinsella RJ, Fitzpatrick DA, Creevey CJ and McInerney JO 2003 Fatty acid biosynthesis in Mycobacterium tuberculosis: lateral gene transfer, adaptive evolution, and gene duplication. *Proc. Natl. Acad. Sci. USA* **100** 1032-0–1032-5

Lawrence JG and Ochman H 1998 Molecular archaeology of the Escherichia coli genome. *Proc. Natl. Acad. Sci. USA* **95** 9413–9417

Lewis K 2010 Persister cells. *Annu. Rev. Microbiol.* **64** 357–372

Lobo LJ and Noone PG 2014 Respiratory infections in patients with cystic fibrosis undergoing lung transplantation. *Lancet Respir. Med.* **2** 73–82

Manabe YC and Bishai WR 2000 Latent Mycobacterium tuberculosis-persistence, patience, and winning by waiting. *Nat. Med.* **6** 1327–1329

Ochman H, Lawrence JG and Groisman EA 2000 Lateral gene transfer and the nature of bacterial innovation. *Nature.* **405** 299–304

Rajan I, Aravamuthan S and Mande SS 2007 Identification of compositionally distinct regions in genomes using the centroid method. *Bioinformatics* **23** 2672–2677

Ribeiro-Guimarães ML and Pessolani MCV 2007 Comparative genomics of mycobacterial proteases. *Microb. Pathog.* **43** 173–178

Rosas-Magallanes V, Deschavanne P, Quintana-Murci L, Brosch R, Gicquel B and Neyrolles O 2006 Horizontal transfer of a virulence operon to the ancestor of Mycobacterium tuberculosis. *Mol. Biol. Evol.* **23** 1129–1135

Sherrid AM, Rustad TR, Cangelosi GA and Sherman DR 2010 Characterization of a Clp protease gene regulator and the reaeration response in Mycobacterium tuberculosis. *PLoS One.* **5**, e11622

Shetty AK, Boloor R, Sharma V and Bhat GHK 2010 Melioidosis and pulmonary tuberculosis co-infection in a diabetic. *Ann Thorac Med.* **5** 113–115

Shrivastava S, Reddy CV and Mande SS 2010 INDeGenIUS, a new method for high-throughput identification of specialized functional islands in completely sequenced organisms. *J. Biosci.* **35** 351–364

Syvanen M 1994 Horizontal gene transfer: evidence and possible consequences. *Annu. Rev. Genet.* **28** 237–261

Tegtmeyer N, Wessler S and Backert S 2011 Role of the cag-pathogenicity island encoded type IV secretion system in Helicobacter pylori pathogenesis. *FEBS J.* **278** 1190–1202

Torrens J, Dawkins P, Conway S and Moya E 1998 Non-tuberculous mycobacteria in cystic fibrosis. *Thorax.* **53** 182–185

Tsirigos A and Rigoutsos I 2005 A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res.* **33** 922–933

van Pittius NCG, Sampson SL, Lee H, Kim Y, van Helden PD and Warren RM 2006 Evolution and expansion of the Mycobacterium tuberculosis PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. *BMC Evol. Biol.* **6** 95

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

Wood TK, Knabel SJ and Kwan BW 2013 Bacterial persister cell formation and dormancy. *Appl. Environ. Microbiol.* **79** 7116–7121

MS received 07 February 2016; accepted 30 May 2016

Corresponding editor: Seyed E Hasnain