Chapter 4

The Mycobacterial Comparison Project

4.1 Introduction

The possibility and usefulness of a central framework to incorporate data analyses and data storage facilities has been demonstrated. This was shown by example, using general bioinformatics analyses such as general sequence alignments, phylogenetic analyses and large scale genome alignment. What shall now be attempted is further building on this framework and applying it in the comparison of Mycobacterial genomes. The 'Mycobacterial Comparison Project' (MCP) has been implemented upon the built framework and is potentially useful for gene-by-gene comparison of the mycobacteria (though the analyses techniques employed here may be extended to any organism). Mycobacteria research in South Africa is an especially relevant area of research as tuberculosis is one of the leading infectious diseases and also responsible for millions of deaths globally (Fu & Fu-Lui, 2007). In South Africa, widespread emergence of multi and extreme drug resistant strains of mycobacteria further exacerbates the problem. Research into the biology of these micro-organisms is therefore very relevant and necessary if there is to be any hope of combating the deadly disease. In light of this, the Mycobacterial Comparison project was initiated in order to elucidate some of the biology of these organisms by means of comparative genomics. Due to the availability of many sequenced Mycobacterium genomes, a comparative genomics study using our FunGIMS-based framework as a backbone became possible. One aim of the whole project was to collect and process data from several key mycobacterial species and structure this data so that meaningful biological conclusions could be drawn. Details regarding the implementation and scientific aims will be discussed, first however, a brief overview of tuberculosis will be supplied, followed by an introduction to the Mycobacterial genome and the current state of comparative Mycobacterial genomics research.

4.2 Tuberculosis

Mycobacterium tuberculosis (*M. tb*) is the etiologic agent of tuberculosis which accounts for more deaths each year than any other disease caused by a single pathogen. Ninety-five percent of these cases are found in developing countries due to inadequacies in the healthcare resources and patient follow-ups (Nouvel *et al.*, 2006). Tuberculosis is predominantly spread by aerosol transmission whereby droplet nuclei often containing several bacilli (particle size $<$ 5 microns) gain access to alveoli of the lungs. It is here that the bacilli are engulfed by alveolar macrophages, which is a cell line equipped with multiple microbiocidal mechanisms including phagolysosome fusion. In order for the bacteria to establish infection, it must first survive this phagolysosome fusion and make its way to the lymphatics or bloodstream (McDonough *et al.*, 1993). It is this pathogen's ability to withstand the macrophages defense mechanisms and survive within the phagosomal compartment of the macrophage that makes it so deadly. Another reason contributing to *M. tb*'s high rate of successful disease establishment is its ability to develop resistance to drugs. It is this drug resistance that especially exacerbates disease control and management (Fu *et al.*, 2007) and poses a significant threat to the global control of the disease.

The relationship of mycobacterial genetics on drug resistant phenotypes and downstream clinical effects has now become a topic of great public interest and much work in this area has recently been conducted yielding very significant findings. One such finding was made by Pym *et al.* (2002) where a study was done on the KatG gene's 315 serine to threonine (S315T) mutation in *M. tb*. Clinically significant isoniazid (INH) resistance is most often linked to the S315T KatG phenotype. KatG in native form codes for a catalase-peroxidase enzyme which converts INH into its bioactive form. KatG is also known to be a virulence factor accounting for heavy attenuation of strains lacking KatG in a variety of animal models. In reality, it was observed that INH-resistant strains are transmitted even in the context of MDR phenotypes likely to be associated with other 'fitness-reducing' mutations. This phenomenon goes against the standard that resistance-conferring mutations significantly reduce bacterial fitness. This hypothesis was tested by constructing a panel of isogenic strains of *M. tb* with different katG alleles and characterizing them in an animal model of tuberculosis.

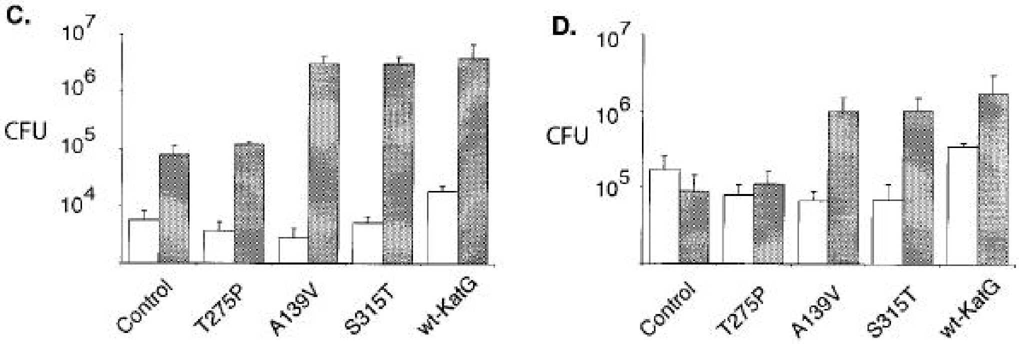


Figure 4.1: Experimental results where growth was monitored in BALB/c mice of strain INH34 complemented with the various katG allele plasmids. White bars correspond to the CFU after day one of infection and gray bars, after 40 days of infection in the lung (C) and spleen (D). Each time point represents the mean result of three or four mice and error bars correspond to standard deviation (Pym *et al.*, 2002).

One interesting point to note in this figure is that KatG-proficient strains were more prolific in their growth in the spleen and lung and the S315T recombinants grew much more vigorously than their respective negative controls.
These results conclusively demonstrated that antibiotic-resistance-conferring mutations do not necessarily carry a high fitness cost or reduction in virulence. The same concept was also ratified in a similar study showing, by knockout studies, that *M. tb*'s Nramp orthologues (mntH) are not vital determinants of virulence (Domenech *et al.*, 2002). This is a potentially dangerous situation as about a third of the world's population are allegedly latently infected with *M. tb* and a large proportion of these individuals carry MDR-TB. Due to it being shown that resistance-conferring-mutations need not negatively affect virulence, future re-activation of latent MDR-TB is a major concern (Pym *et al.*, 2002).

4.3 The Mycobacterial genome

Before moving on to comparative mycobacterial genomics, a short description of the mycobacterial genome will now be undertaken. A general understanding of the mycobacterial genome will form a foundation on which comparative genomics can be built. Using techniques such as pulsed field gel electrophoresis, fingerprinting and hybridization, scientists earlier on were able to establish the genome structure of two mycobacteria (Figure 4.2).

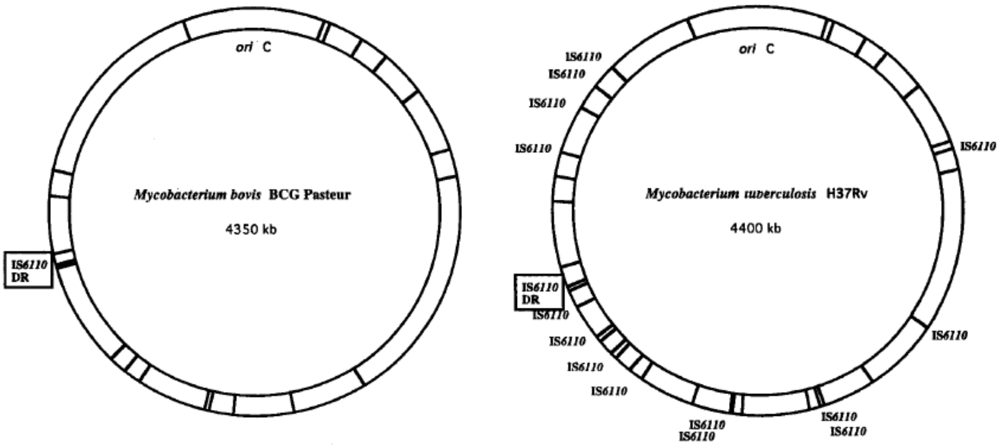


Figure 4.2: Early comparison of *M. tb* and the vaccine strain M. bovis BCG based on IS6110 sites.

Even at such an early stage, scientists were able to identify 16 copies of the insertion sequence IS6110, 8 copies of IS1081 and 26 copies of PGRS. The distribution of these insertion sequences were found to be non-random with a strong concentration of these inserts in a region of about 40% of the *M. tb* chromosome near the putative terminus of replication (terC). With the advancements of many scientific techniques, many more insights were to follow. The first *M. tb* strain to be isolated was H37Rv in 1905 (Camus *et al.*, 2002) taken from sputa from a 19 year-old male suffering from pulmonary tuberculosis (Kato-Maeda *et al.*, 2001). Even through years of passage, the bacterium has remained virulent in the animal model. In 1998, Cole *et al.* were the first to publish a completely sequenced and annotated genome of H37Rv highlighting many critical facts about the genome and opening doors for many other downstream comparative studies (see later). A few of the group's observations will be noted here.

By using large insert BACs, cosmids and random small-insert clones from whole-genome shotgun libraries, the group systematically pieced together the genome. The result was a 4,411,529 composite sequence with a 65.6% G+C (GC) content. The genome was found to be rich in repetitive DNA, particularly in insertion sequences (IS), new multi-gene families and duplicated house-keeping genes. The GC content was found to be relatively constant throughout the genome. This uniformity of the GC content is probably due to the lack of atypical base composition pathogenicity islands. There were however, regions that exhibited higher than average GC content, but these corresponded to sequences belonging to the large gene family which include the polymorphic GC-rich sequences (PGRSs).

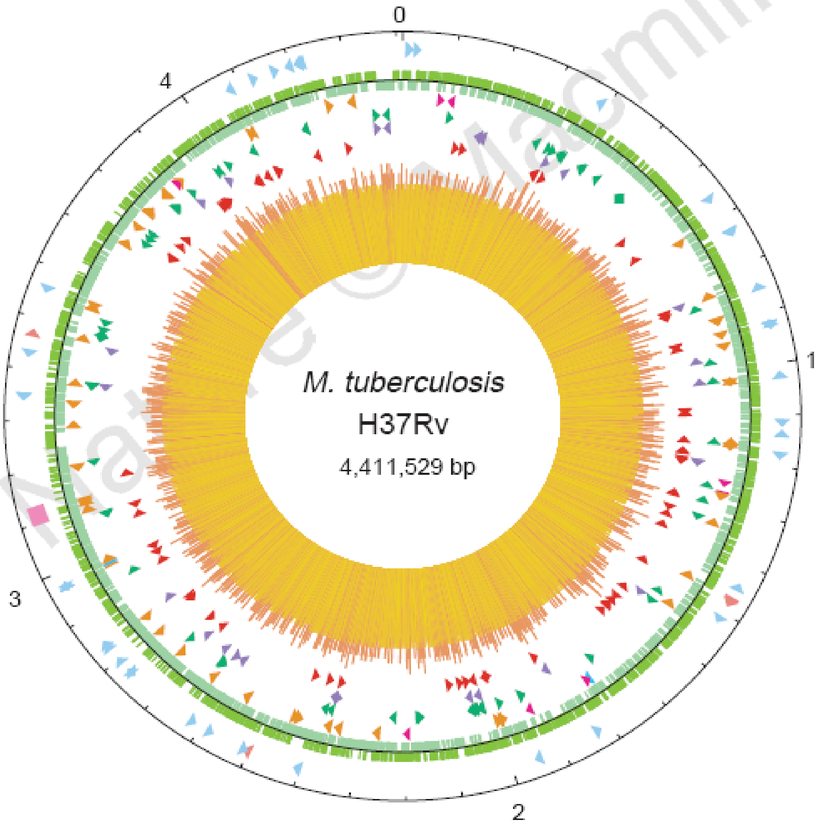


Figure 4.3: Circular map of M. tb H37Rv chromosome. The outer circle shows the scale in megabases. 0 represents the origin of replication. Moving inwards, the next ring denotes the position of the stable RNA genes (tRNA-blue; other are in pink); the second ring shows the coding sequence by strand (clockwise-darkgreen; anti-clockwise-lightgreen); the third ring denotes repetitive DNA (IS elements-orange; 13E12 REP family-dark pink; prophage-blue). The fourth ring denotes positions of the PPE family members (green). The fifth ring shows the PE family members (purple, excluding PGRS) and the sixth ring shows the positions of the PGRS sequences (dark red). The histogram (centre) represents GC content ($<$ 65% GC-yellow; $>$ 65%-red) (Cole et al., 1998).

In terms of RNA, 50 genes encoding functional RNA molecules were detected. The following were the three molecular species produced by the unique ribosomal RNA operon, the 10Sa RNA (involved in degradation of proteins encoded by abnormal mRNA), the RNA component of RNase P and 45 transfer RNA. The rrn operon is situated unusually as it is located about 1,500 kilobases (kb) from the putative oriC. This is significant as most eubacteria have one or more rrn operons proximal to oriC in order to exploit the gene-dosage effect obtained during replication. This placement may account for the slow-growth of *M. tb*.

In terms of insertion sequences and prophages, access to the full genomic sequence of H37Rv led to the detection of a further, pre-dominantly undescribed, 32 different IS elements in addition to those presented by Philipp *et al*, (1998). The newly discovered IS elements mainly belong to the IS3 and IS256 families, though 6 constitute a brand new group. Most insertion sequences found in this genome appear to have been inserted into non-coding or intergenic regions and often near tRNA genes. Many have also clustered thus suggesting the possibility of actual insertional hot-spots which may prevent gene activation. At least 2 prophages were detected in the genome explaining *M. tb*'s persistent low-level lysis in culture. Prophages phiRv1 and phiRv2, both approximately 10kb in length and some of their gene products show marked similarity to those encoded by certain bacteriophages from Streptomyces and saprophytic mycobacteria.

In terms of protein coding genes, 3,924 ORFs were identified thus accounting for about 91% of the genomes potential coding capacity. Consistent with the high GC content, GTG initiation codons (35%) are used more frequently than in *Bacillus subtilus* (9%) and *E. coli* (14%), although ATG (61%) is the most common translational start. The even distribution in gene polarity seen in *M. tb* (as opposed to fast growing bacteria such as *E. coli*) may account for the slow-growth and the infrequency of replication cycles. Recently, many more fully sequenced Mycobacteria genomes have entered the public domain such as *M. tb* strains H37Rv, CDC1551, H37Ra and F11 as well as *M. bovis* AF2122/97 and *M. bovis* BCG str. Pasteur 1173P2. These are available through via the NCBI server and other public databases (Vishnoi *et al*, 2008). A list of fully sequenced genomes may be found at the Genome News Network website (17).

4.4 Comparative genomics of Mycobacteria

Thanks to the availability of *M. tb* and other mycobacterial whole genome sequences, comparative genomic studies are now flourishing helping us gain ground in the elucidation of tuberculosis pathogenesis amongst other things. An overview of Mycobacterial comparative genomics (ranging from the last decade to the present) will be treated here thus giving a good idea of the progression and current state of the field.

Comparative studies of Mycobacteria is by no means a new field. Scientists from very early on realized the potential of these comparisons and resorted to techniques such as PFGE, genetic fingerprinting, hybridization and restriction maps to compare mycobacteria (Philipp *et al.*, 1998; Philipp *et al.*, 1996; Cole *et al.*, 1998). These techniques were extremely effective and made many great discoveries possible. One of the first questions that scientists tried to understand was the mechanism responsible for the attenuation of *Mycobacterium bovis* in becoming the reliable BCG vaccine strain. In 2000 Brosch *et al.* set out to characterize the vaccine strain of M. bovis BCG Pasteur 1173P2. Due to the inability of DNA hybridization techniques to detect insertion sequences and translocations, a complimentary method based on PFGE of HinDIII restriction fragments from selected *M. tb* H37Rv and *M. bovis* BCG Pasteur (vaccine strain) BACs was used. Using this technique, two major rearrangements were identified in BCG. These were shown to correspond to two tandem duplications, DU1 (29 668 bp) and DU2 (36 161 bp). DU1 was the result of a single duplication event but DU2 is supposed to have arisen from a 100 kb genomic segment which subsequently incurred a 64 kb internal deletion. BCG strains that still contain DU1 and DU2 were found to be diploid for at least 58 genes and contain two oriC copies (Figure 4.4). Some evidence at the time of the study suggested that DU2 was still undergoing expansion as two copies were detected in a few sub-populations of BCG Pasteur cells. Although the exact impact of the duplications on the pathogenicity of BCG is unclear, mere knowledge of these regions would aid in the quality control of BCG vaccines (Brosch *et al.*, 2000).

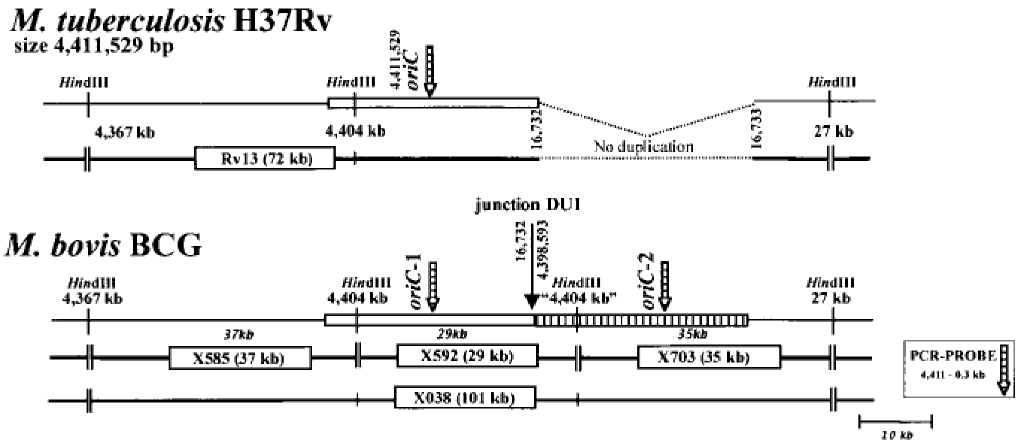


Figure 4.4: Overview of the genomic organization in the corresponding regions proximal to the origin of replication in BCG Pasteur and *M. tb* H37Rv, revealed by BAC mapping, PCR and hybridization experiments (Brosch *et al.*, 2000).

Comparison of mycobacterial genomes is also important for purposes of genotyping. Already covered in chapter one was the use of SNPs in order to categorize or group mycobacteria with their clonal relatives (Gutacker *et al.*, 2002). Fleischmann *et al.* (2002) in the same year also produced similar results. After sequencing the complete genome of the *M. tb* clinical strain, the group was able to perform comparisons with *M. tb* lab strain, H37Rv and showed that the two strains exhibited several differences on the SNP and larger-sequence level such as genomic re-arrangements. These changes were more apparent at specific loci, thus suggesting mutational hotspots and selective pressure. A more detailed look at these differences could elucidate their role in pathogenesis and host immunity (Fleischman *et al.*, 2002).

Drug target discovery is undeniably, a crucial aspect of medicine and biological sciences with far reaching implications. Comparative genomics has shown to be a powerful tool in this field as well (Marmiesse *et al.*, 2003). One study presented by Cole (2002) clearly illustrated this point. Gene duplication events are an unavoidable part of mycobacterial evolution leading to functional redundancies in biochemical pathways. It is sometimes difficult to predict with certainty, which of the duplicated genes impact on which function. In *M. tb*, there were five proteins that showed strong similarity (based on database searches) to various lipoamide dehydrogenase components of the crucial pyruvate dehydrogenase complex. The proteins in question were Rv0462, Rv0794c, Rv2855c, Rv2713 and Rv3303c. During early analyses of the *M. tb* genome it was shown that proteins Rv3303c and Rv0794c (termed lpdA and lpdB respectively) showed the highest similarity to lipoamide dehydrogenase. Subsequent biochemical studies of the gene products however, showed that it was Rv0462 that actually encoded authentic lipoamide dehydrogenase. Comparative genomics would have helped come to this conclusion much faster as only one of the five *M. tb* proteins had a functional orthologue with *M. leprae* and that was Rv0462. The remaining four genes were present in pseudogene form in M. leprae (Cole *et al.*, 2002).

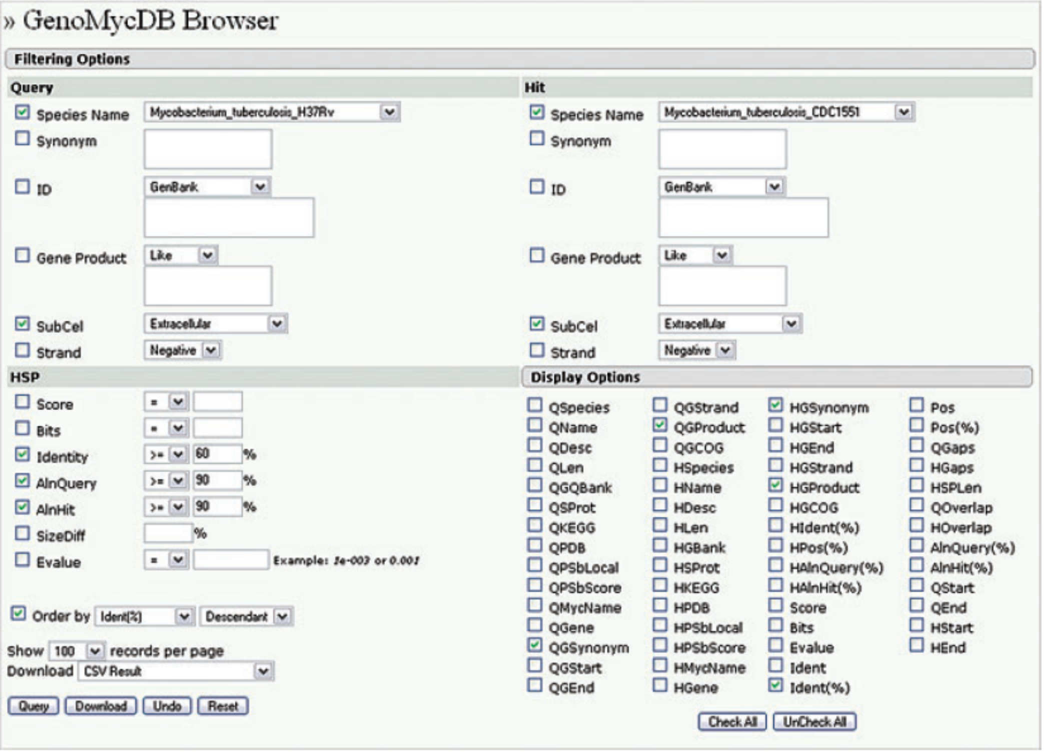
Scientists are often interested in tracing the lineage of bacteria (and organisms in general) for various reasons such as observing the effects of micro-evolution, tracing pathogenesis (Kato-Maeda *et al.*, 2001), examining what micro-evolutionary effects have on some protein families downstream in terms of drug resistance (Gagneux *et al.*, 2006) and also to determine their origins (Tsolaki *et al.*, 2005). Using comparative genomics, Brosch *et al.*, (2002) shattered the long held belief that *M. tb* evolved from *M. bovis*. By looking at a specific set of 20 variable regions among a 100 mycobacterial strains as well as specific deletion events (TbD1) the groups evidence pointed to the fact that *M. tb* existed before *M. bovis* and may have been a human pathogen far longer than previously believed. Using similar analyses Pym *et al.*, (2002) also managed to show mechanisms for *Mycobacterium bovis* BCG and *Mycobacterium microti* attenuation into vaccine strains. Comparative analyses (amongst other techniques such as gene knock-out) revealed that a specific deletion event of RD1 has a major role to play in the efficacy of the above-mentioned vaccine strains. In addition, further comparative analyses carried out by Garnier *et al.* (2003) revealed that *M. bovis* contains no unique genes when compared to other members of the *M. tb* complex. This implies that it is the differential gene expression that dictates host tropism and virulence of the various strains.

Comparative mycobacterial genomics has indeed proven to be a useful scientific tool over the years and with the availability of more sequence data, the scientific world is poised for greater insights into this pathogen. To proceed, we aim to exploit new mycobacterial sequence data by the implementation of a mycobacterial comparison project. This is dealt with next.

4.5 The Mycobacterial Comparison Project in context

Owing to the widespread effect of the disease and growing international concern, many other projects involving the comparisons of genomes of Mycobacteria have also been undertaken. Some of these projects include MycDB (Bergh & Cole, 1994), GenoMycDB (Cantanho *et al.*, 2006) and MycoperonDB (Rangan *et al.*, 2006) to name a few. MycDB is one of the earliest examples highlighting the need for an integrated platform to combine sequence data for mycobacteria. The MycDB project was essentially a database management system which combined mycobacterial specific data such as antigen lists, reagent details from CDC/WHO antibody bank, a few thousand gene sequences, reference data and physical maps (Bergh & Cole, 1994). The system was a stand-alone database management software which basically allows users to view data as well as show relationships between the various data types in the database by complex query processes. Although useful at the time, it is now obsolete, as access to completely sequenced and finished genomes with substantial annotation and post genomic analyses information is widely available.

GenoMycDB (Cantanho *et al.*, 2006) is another more contemporary, mature and comprehensive web-based database example of a mycobacterial comparative database. At the core structure of this database lies the pairwise sequence alignments (and parameters) of all predicted proteins for six mycobacterial strains including *M. tb* (strains H37Rv and CDC1551), M. bovis AF2122/97, *M. avium subsp. paratuberculosis* K10, *M. leprae* TN, and *M. smegmatis* MC2 155. For each protein, the database also stores information such as sub-cellular localization, assigned cluster of orthologous groups (COGs), corresponding gene features and more. Users are also allowed to carry out complex queries in order to produce tables containing groups of potential homologous proteins. Furthermore, queries may be restricted by localization, DNA strand of corresponding gene and/or protein annotation.


» GenoMycDB Browser

Filtering Options

Query
☑ Species Name: Mycobacterium_tuberculosis_H37Rv
□ Synonym
□ ID: GenBank
□ Gene Product: Like
☑ SubCel: Extracellular
□ Strand: Negative

Hit
☑ Species Name: Mycobacterium_tuberculosis_CDC1551
□ Synonym
□ ID: GenBank
□ Gene Product: Like
☑ SubCel: Extracellular
□ Strand: Negative

HSP
□ Score =
□ Bits =
☑ Identity >= 60 %
☑ AlnQuery >= 90 %
☑ AlnHit >= 90 %
□ SizeDiff %
□ Evalue = Example: 1e-003 or 0.001

☑ Order by Ident(%) Descendant
Show 100 records per page
Download CSV Result

Display Options
□ QSpecies
□ QName
□ QDesc
□ QLen
□ QGQBank
□ QSProt
□ QKEGG
□ QPDB
□ QPSbLocal
□ QPSbScore
□ QMycName
□ QGene
☑ QGSynonym
□ QGStart
□ QGEnd
□ QGStrand
☑ QGProduct
□ QGCOG
□ HSpecies
□ HName
□ HDesc
□ HLen
□ HGBank
□ HSProt
□ HKEGG
□ HPDB
□ HPSbLocal
□ HPSbScore
□ HMycName
□ HGene
☑ HGSynonym
□ HGStart
□ HGEnd
□ HGStrand
☑ HGProduct
□ HGCOG
□ HIdent(%)
□ HPos(%)
□ HAlnQuery(%)
□ HAlnHit(%)
□ Score
□ Bits
□ Evalue
□ Ident
☑ Ident(%)
□ Pos
□ Pos(%)
□ QGaps
□ HGaps
□ HSPLen
□ QOverlap
□ HOverlap
□ AlnQuery(%)
□ AlnHit(%)
□ QStart
□ QEnd
□ HStart
□ HEnd

Figure 4.5: Overview of the GenoMycDB user interface. Note the available options for searching and displaying (Catanho *et al.*, 2006).

GenoMycDB is extremely useful allowing users to functionally classify their mycobacterial proteins of interest as well as elucidate the genome structure of the contained mycobacteria. The mycobacterial comparison project does not contain the same information as GenoMycDB but rather seeks to create a richer source of post-analyses mycobacterial comparison data not only on the protein, but on the DNA level as well.

Yet another example is the MycoperonDB (Rangan *et al.*, 2006). This is a database of computationally predicted transcriptional units and operons from five different strains of mycobacteria. The database also combines literature information for experimentally validated mycobacterial operons with the aim of validating computationally predicted operons. All the above mentioned data is contained with a relational database with a user-friendly web-interface freely available at http://www.cdfd.org.in/mycoperondb/index.html.

The mycobacterial comparison project seeks to create a new type of comparative environment where on-the-fly comparisons can be done and meaningful conclusions can be drawn. This will be achieved by integrating a few related datatypes all of which are crucial in understanding the inter-relatedness of the various mycobacteria. The mycobacterial species chosen for analyses will now be discussed as well as how the various datatypes were generated.

4.6 Data pre-processing

4.6.1 Mycobacterial strain selection

There are quite a large number of mycobacterial strains that have been sequenced as well as fully assembled and annotated. These sequences are found in Genbank. However, initially only a few strains of mycobacteria were chosen so as to be representative of Mycobacteria as a genus. Both virulent and avirulent strains, spanning a moderate range of hosts, were chosen with which to conduct analyses. The strains chosen were *M. tb* H37Rv, *Mycobacterium tuberculosis* H37Ra, *Mycobacterium tuberculosis* F11, *Mycobacterium tuberculosis* CDC1551, *Mycobacterium ulcerans* Agy99, *Mycobacterium bovis subsp. bovis*, *Mycobacterium bovis* BCG, *Mycobacterium leprae* TN and *Mycobacterium avium subsp. paratuberculosis.*

4.6.2 Annotation Data

Annotation data for all mycobacterial strains were extracted from their full Genbank records. The files were parsed using Biopython and the annotations were subsequently stored in their respective tables in the database. Python scripts were used to extract the annotations as well as to populate the database (See figure below).

4.6.3 Gene-by-gene mutation data

After the gene data (i.e. annotations) for each species was committed to the database, the actual gene sequences for each gene was then searched with BLAST against all genes for the next species. In this way, the gene sequences for all the mycobacterial species were 'aligned' by BLAST. This analysis was a crude way of establishing whether a gene of one species was present in the next. If the gene was present in another species, CAI values were calculated for each gene and more importantly, mutations between the sequences were recorded at each of the three codon positions. This data was stored in the 'gene_mutations' table.

4.6.4 SNP Data

For each of the nine species, SNP data was generated in the following way. Using M. tb F11 as the reference strain, each of the other strains of mycobacteria were aligned to the reference strain using the 'nucmer' tool (part of the MUMmer suite (Kurtz *et al.*, 2004)). Thereafter, 'show-snps' (also part of the MUMmer suite) was used to extract single-nucleotide polymorphisms (SNPs) from the alignment data produced by nucmer. SNP data was stored in species specific tables.

4.6.5 Gene island data

Using an in-house algorithm, developed by co-workers (see Seqword Genome Browser chapter), whole genome sequences of each of the mycobacterial strains were scanned for the presence of genomic islands which imply horizontally transferred elements. The genomic island co-ordinates for each of the strains were stored in the 'gi' (gene islands) table.

4.7 Database requirements

The database schema needed to be fast, stable and cater for the various data-types mentioned above. Separate tables were created for each of the data-types and foreign keys were subsequently used to associate the tables (Figure 4.6). There is a high degree of inter-connectivity between the sub-schema and within the sub-schema thus creating a highly flexible and highly semantic relationship of all the data. This sub-schema designed for the mycobacterial project had to also be compatible with the main Fungims project schema and this was successfully accomplished.

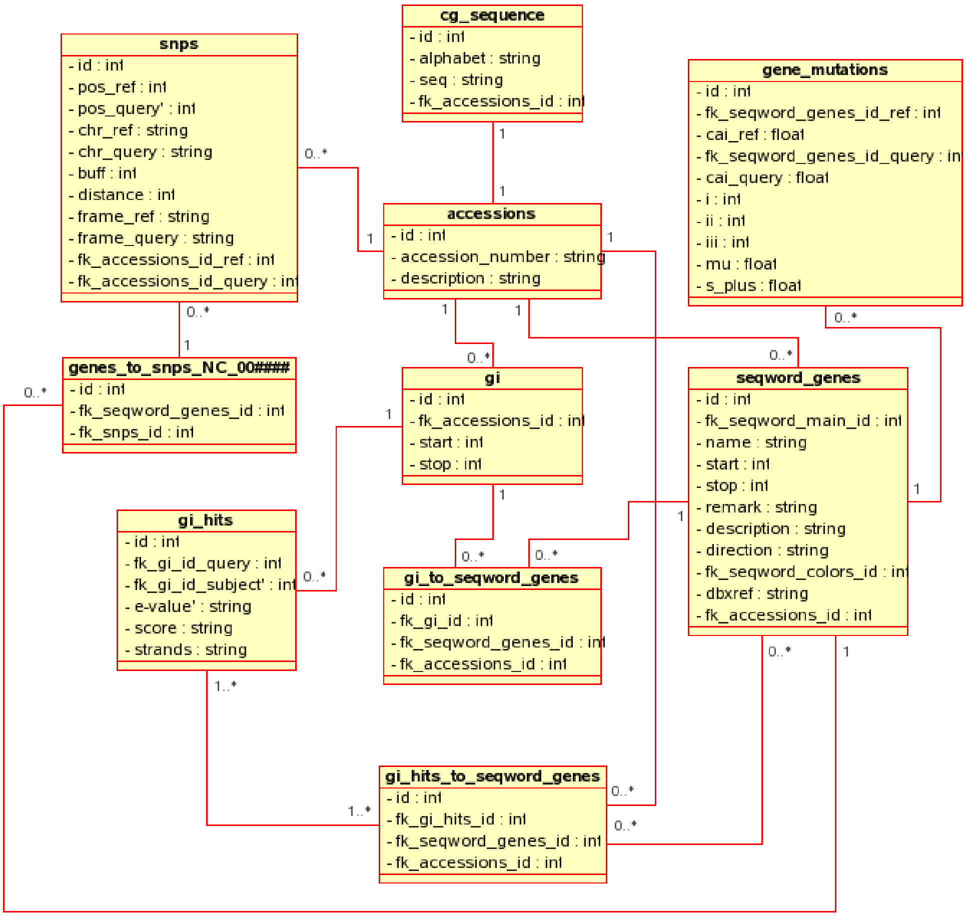


Figure 4.6: Schema of the mycobacterial comparison project database.

4.8 Graphical User Interface requirements

The design of the graphical-user interface (GUI) had to first and foremost, relay the significance of the inter-relationships of the various data-types in a manner that had the most impact. This was accomplished by the use of contrasting colors and an effective layout. Also, due to the many data-types and information that had to be displayed, a simple layout pattern had to be used to display all the data on the page while avoiding clutter and over usage of screen real estate. This was accomplished by simple tables and side-by-side display of the data. Furthermore, user-friendliness was a top priority and no more than two or three buttons had to be clicked for users to dynamically generate data. Also, the screen layout was extremely simple and un-ambiguous leaving users with little to be confused about when using mycobacterial comparison project.

4.9 Workflow summary

The mycobacterial comparison project essentially allows for the integration and visualization of several key data-types significant in mycobacteria. In addition, simple tree drawing functionality is also available. The basic underlying functions include annotation retrieval, database filtering of overlapping data, sequence alignment of genes and lastly, neighbor-joining tree generation. The following table depicts the general flow contained within the mycobacterial comparison project (Table 4.1)

Table 4.1: Table showing general order of events and options available to users when in the mycobacterial comparison project.

#	Web page/User choice	Tasks performed by server
1	Mycobacterial Comparison Project introductory Page	- Retrieval and display of mycobacterial genome metadata including genome length; accession numbers and number of gene islands found
2	Choice of strain to explore	- Retrieval and display of annotation information for chosen strain
3	Choice of gene of interest	- Retrieval of gene information of user specified gene - Retrieval of homologues (gene-by-gene data) - Retrieval of SNPs corresponding to specified gene - Autogeneration of protein alignment of homologues - Formatting and display of all above data
5	Creation of phylogenetic trees	- Redisplay of all homologues on new web page
6	Choice of DNA or Protein alignment based neighbour-joining tree	- Retrieval of all sequences (sequence conversion if necessary) - Remote-procedure call for clustalw alignment - Collection of result files and error checking - Creation of distance matrix (phylip) - Neighbour-joining tree generation (phylip) - Conversion of tree file to other formats - Display of tree and homologues

4.10 A comparative genomics investigation of key genomic loci in mycobacterial genomes and their role in virulence

In this section an investigation built off the findings of the SWGB chapter is carried out. The aim is to further identify significant genetic differences between specific mycobacterial strains such as *M. tb* H37Rv, *M. tb* H37Ra and *M. tb* F11. Starting off with the genomic islands identified by the SWGB, this system was used to analyze these specific areas in more detail and show how the genes contained within these loci are not conserved as would be expected for these essential genes. Genes that have important functions are often found conserved within a group of bacteria. For example, DNA replication genes are core to bacterial genomes and are thus expected to be conserved. An investigation into these DNA polymerase genes was also undertaken for a few mycobacterial species in order to inspect their level of conservation among the mycobacteria. Other genes of interest, identified by their atypical nucleotide usage (in chapter 3) were the PE-PGRS genes. A profile of these genes in particular was built up for the mycobacteria in order to guage their relationship with the organism's virulence status. Profiles for this gene family was created by, selecting specific genes (one at a time) and then retrieving the homologue data generated by the MCP. The data for each gene was then tabularized (see below).

In the previous chapter, the SWGB identifed genomic islands within the various mycobacterial strains and using its comparative view, several key genomic islands were identified. As an example, attention will be turned to the gross differences identified between *M. tb* H37Rv and *M. avium ssp Paratuberculosis*. With the aid of gene diagram dot-plots (RV, GRV and D) in chapter 3, several genomic islands were identified in *M. avium ssp Paratuberculosis* that were clearly absent in *M. tb* H37Rv. A closer look within these genomic islands revealed several genes and gene families. Similarly, using different dot-plot criteria, genomic islands unique to *M. tb* H37Rv were revealed.

Several genomic islands within *M. avium ssp Paratuberculosis* (K10) were identified. The table beneath shows the coordinates and annotations for the identified genomic islands.

Table 4.2: Coordinates and annotation of the genes islands in the genome of *M. avium* K10.

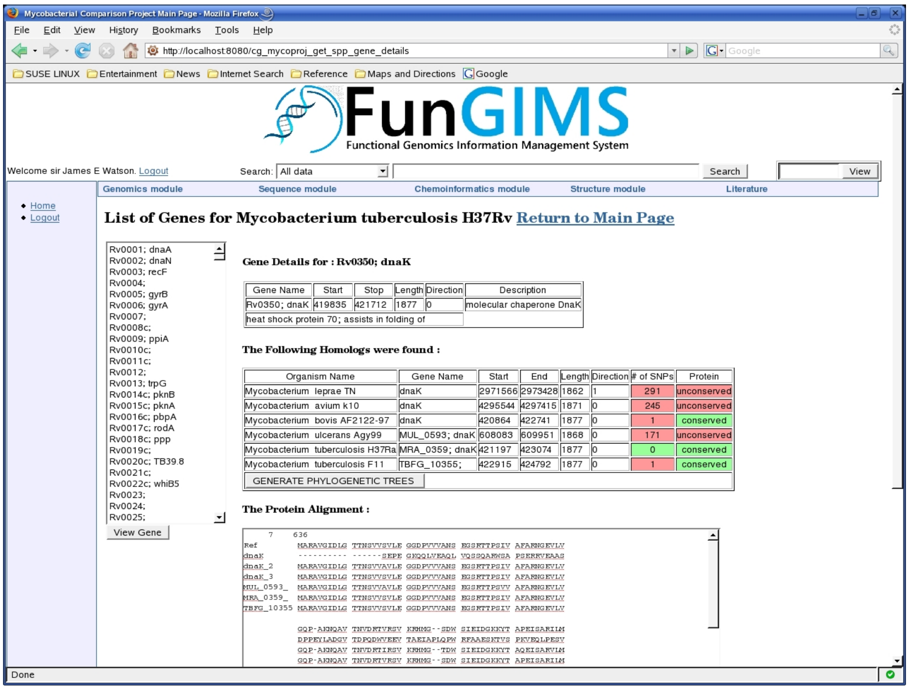
Left	Right	n1_4mer:RV	n1_4mer:GRV	n0_4mer:D	Annotations
78000	86000	11.6606	18.0853	13.7828	dnaB; mmpL4_1 and 4 genes for hypothetical proteins
870000	892000	8.6089	22.8127	22.0832	nramp and 20 hypothetical proteins
1290000	1304000	8.7803	20.6197	21.1895	lipL and a hypothetical gene

Using the mycobacterial comparison project, a closer look is taken into each of these genomic island sub-elements to see how they relate to other mycobacterial strains. dnaB was examined first. Although dnaB may not necessarily be of 'horizontal' origin it was investigated to see whether this very important gene contained homologues and to what extent the homologues differed among the various strains.

```
C:/Documents and Settings/user/My Documents/phd_Feb09/phd_thesis_draft3/phd_images_3/H37R
```

Figure 4.7: dnaB gene details for *M. tb* H37Rv and its homologues.

The gene dnaB was found present in all the database species present. The MCP system essentially extracts the annotation and its related sequence information for the gene of interest from the species of interest and nucleotide BLASTs it against the annotations and sequences for the other strains. Based on this analyses steps, the dnaB gene of *M. tb* H37Rv has been identified in (i.e aligned to) all other species in the database. However, on closer examination, it was found that the dnaB *M. avium* K10 homologue is the most different when compared to the dnaB of *M. tb* H37Rv, containing over 640 single nucleotide differences which ultimately lead to a downstream protein alteration. When aligning, the MCP also takes into account the nucleotide mismatches within the alignment and counts them. The MCP also performs on the fly translation of the genes to inspect whether the mismatches were synonymous (coded in green) or non-synonymous (coded in red) relative to the reference species protein. Although at the gene level dnaB is relatively unchanged among *M. tb* CDC1551, *M. bovis* and *M. tb* F11, the protein sequence was still affected. Note that this protein remains fully conserved between the *M. tb* strains H37Ra and H37Rv. *M. Leprae* exhibited 291 nucleotide differences relative to *M. tb* H37Rv which lead to an altered protein. This many differences in the nucleotide sequence inevitably means great differences in the amino-acid sequence and subsequent protein. The possibility that the dnaB of *M. Leprae* functions slightly different to that of *M. tb* H37Rv is great.


Mycobacterial Comparison Project Main Page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/cg_mycoproj_get_spp_gene_details
Google
SUSE LINUX Entertainment News Internet Search Reference Maps and Directions Google
FunGIMS
Functional Genomics Information Management System
Welcome sir James E Watson. Logout
Search: All data
Search
View
Genomics module
Sequence module
Chemoinformatics module
Structure module
Literature
Home
Logout
View Gene
GENERATE PHYLOGENETIC TREES
Done

List of Genes for Mycobacterium tuberculosis H37Rv Return to Main Page

Rv0001; dnaA
Rv0002; dnaN
Rv0003; recF
Rv0004;
Rv0005; gyrB
Rv0006; gyrA
Rv0007;
Rv0008c;
Rv0009; ppiA
Rv0010c;
Rv0011c;
Rv0012;
Rv0013; trpG
Rv0014c; pknB
Rv0015c; pknA
Rv0016c; pbpA
Rv0017c; rodA
Rv0018c; ppp
Rv0019c;
Rv0020c; TB39.8
Rv0021c;
Rv0022c; whiB5
Rv0023;
Rv0024;
Rv0025;

Gene Details for : Rv0350; dnaK

Gene Name	Start	Stop	Length	Direction	Description
Rv0350; dnaK	419835	421712	1877	0	molecular chaperone DnaK
heat shock protein 70; assists in folding of					

The Following Homologs were found :

Organism Name	Gene Name	Start	End	Length	Direction	# of SNPs	Protein
Mycobacterium leprae TN	dnaK	2971566	2973428	1862	1	291	unconserved
Mycobacterium avium k10	dnaK	4295544	4297415	1871	0	245	unconserved
Mycobacterium bovis AF2122-97	dnaK	420864	422741	1877	0	1	conserved
Mycobacterium ulcerans Agy99	MUL_0593; dnaK	608083	609951	1868	0	171	unconserved
Mycobacterium tuberculosis H37Ra	MRA_0359; dnaK	421197	423074	1877	0	0	conserved
Mycobacterium tuberculosis F11	TBFG_10355;	422915	424792	1877	0	1	conserved

The Protein Alignment :

```
    7    636
Ref        MARAVGIDLG TTNSVVSVLE GGDPVVVANS EGSRTTPSIV AFARNGEVLV
dnaK       ---------- ------SEPE GKQQLVEAQL VQSSQARWSA PSERRVEAAS
dnaK_2     MARAVGIDLG TTNSVVAVLE GGDPVVVANS EGSRTTPSIV AFARNGEVLV
dnaK_3     MARAVGIDLG TTNSVVSVLE GGDPVVVANS EGSRTTPSIV AFARNGEVLV
MUL_0593_  MARAVGIDLG TTNSVVAVLE GGDPVVVANS EGSRTTPSVV AFARNGEVLV
MRA_0359_  MARAVGIDLG TTNSVVSVLE GGDPVVVANS EGSRTTPSIV AFARNGEVLV
TBFG_10355 MARAVGIDLG TTNSVVSVLE GGDPVVVANS EGSRTTPSIV AFARNGEVLV

           GQP-AKNQAV TNVDRTVRSV KRHMG--SDW SIEIDGKKYT APEISARILM
           DPPEYLADGV TDPQDWVEEV TAEIAPLQPW RFAAESKTVS PKVEQLPESV
           GQP-AKNQAV TNVDRTIRSV KRHMG--TDW SIEIDGKKYT AQEISARVLM
           GQP-AKNQAV TNVDRTVRSV KRHMG--SDW SIEIDGKKYT APEISARILM
```

Figure 4.8: dnaK gene details for *M. tb* H37Rv and its homologues.

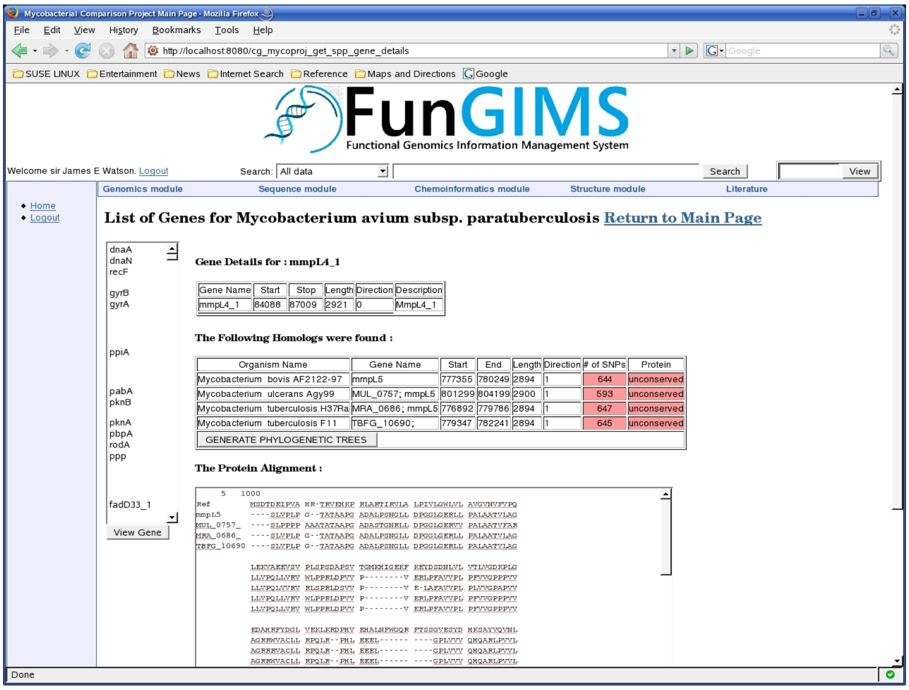
Several other dna genes were looked at using the MCP. Figure 4.8 above shows again how different *M. avium's* dnaK homologue is relative to *M. tb* H37Rv. The table below highlights the dna genes of *M. tb* H37Rv and whether there are homologues in *M. avium* K10.

Table 4.3: Summary of absence/presence of dna genes of *M. tb* H37Rv in *M. avium* K10.

Gene in *M. tb* H37Rv	Present in *M. avium* K10
dnaB	Yes
dnaA	No
dnaN	No
dnaK	Yes
dnaJ1	Yes
dnaE1	Yes
dnaJ2	Yes
dnaE2	Yes
dnaQ	Yes
dnaZX	Yes

The table above clearly shows that most dna genes of H37Rv are also present in *M. avium* K10. This may imply that the mechanisms for dna replication between these species are highly similar, however, the MCP also revealed that none of these dna genes are actually conserved in *M. avium*. In all cases, where there is an *M. avium* homologue, there are over 100 nucleotide differences between the *M. tb* H37Rv copy and its *M. avium* counterpart. This could further imply that although all the protein components may be shared among species, the mechanisms by which they are used may differ.

In terms of the other genes identified within *M. avium*'s gene-island, mmpL4_1 was examined.


Mycobacterial Comparison Project Main Page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/cg_mycoproj_get_spp_gene_details
Google
SUSE LINUX Entertainment News Internet Search Reference Maps and Directions Google
FunGIMS
Functional Genomics Information Management System
Welcome sir James E Watson. Logout
Search: All data
Search
View
Genomics module
Sequence module
Chemoinformatics module
Structure module
Literature
Home
Logout
dnaA
dnaN
recF
gyrB
gyrA
ppiA
pabA
pknB
pknA
pbpA
rodA
ppp
fadD33_1
View Gene
GENERATE PHYLOGENETIC TREES
Done

List of Genes for Mycobacterium avium subsp. paratuberculosis Return to Main Page

Gene Details for : mmpL4_1

Gene Name	Start	Stop	Length	Direction	Description
mmpL4_1	84088	87009	2921	0	MmpL4_1

The Following Homologs were found :

Organism Name	Gene Name	Start	End	Length	Direction	# of SNPs	Protein
Mycobacterium bovis AF2122-97	mmpL5	777355	780249	2894	1	644	unconserved
Mycobacterium ulcerans Agy99	MUL_0757; mmpL5	801299	804199	2900	1	593	unconserved
Mycobacterium tuberculosis H37Ra	MRA_0686; mmpL5	776892	779786	2894	1	647	unconserved
Mycobacterium tuberculosis F11	TBFG_10690;	779347	782241	2894	1	645	unconserved

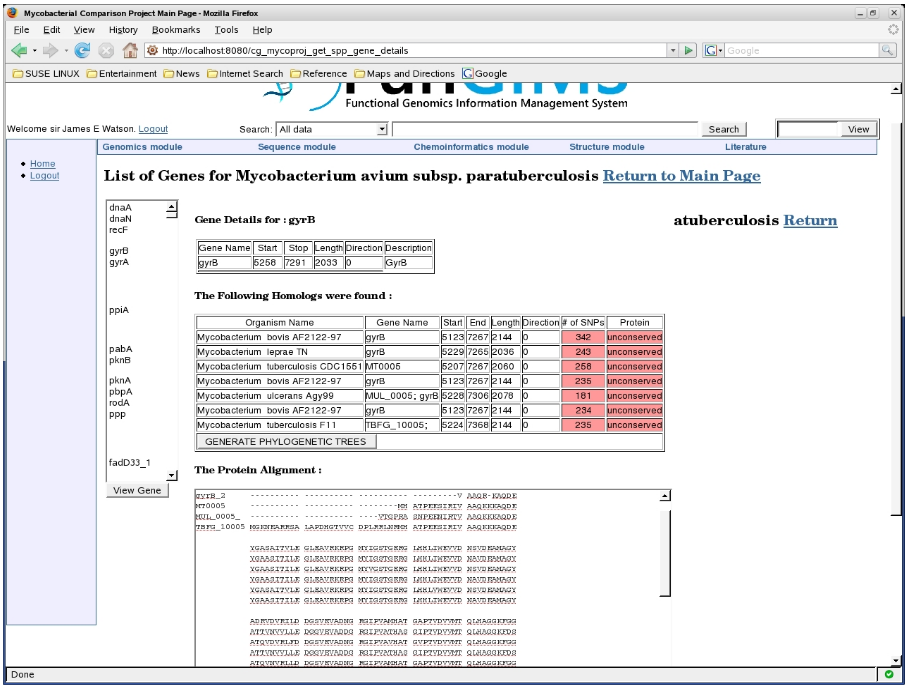
The Protein Alignment :

```
     5   1000
Ref          MSDTDEIPVA HR-TRVEHKP RLARTIRVLA LPIVLGWLVL AVGVNVFVPQ
mmpL5        ----SLVPLP G--TATAAPG ADALPSNGLL DPGGLGERLL PALAATVLAG
MUL_0757_    ----SLPPPP AAATATAAPG ADASTGNRLL DPGGLGERVV PALAATVFAR
MRA_0686_    ----SLVPLP G--TATAAPG ADALPSNGLL DPGGLGERLL PALAATVLAG
TBFG_10690   ----SLVPLP G--TATAAPG ADALPSNGLL DPGGLGERLL PALAATVLAG

             LEKVAEEVSV PLSPSDAPSV TGMKHIGEKF KEYDSDNLVL VTLVGDKPLG
             LLVPQLLVRV WLPPRLDPVV P--------V ERLPFAVVPL PFVVGPPPVV
             LLVPQLVVRV RLSPRLDSVV P--------V E-LAFAVVPL PLVVGPAPVV
             LLVPQLLVRV WLPPRLDPVV P--------V ERLPFAVVPL PFVVGPPPVV
             LLVPQLLVRV WLPPRLDPVV P--------V ERLPFAVVPL PFVVGPPPVV

             EDAHRFYDGL VEKLKRDPHV EHALNFWGQR FTSSGVESYD HKSAYVQVNL
             AGRRWVACLL RPQLR--PHL EEEL------ ----GPLVVV QHQARLPVVL
             AGRRRVACLL RPQLR--PHL EEEL------ ----GPLVVV QHQARLPVVL
             AGRRWVACLL RPQLR--PHL EEEL------ ----GPLVVV QHQARLPVVL
```

Figure 4.9: mmpL4_1 gene details for *M. avium ssp paratuberculosis* K10 and its respective homologues.

Supportive of the SWGB gene-plot findings, it is seen here that this gene, is totally absent from *M. tb* H37Rv. Furthermore, the gene does not seem to exhibit a high degree of conservation among the various mycobacterial strains as shown by the high number of SNPs in the homologues. Whether this gene is functional or not in the species tested is not known, however, if they are functional, the mechanisms are likely to be quite variable.

The gyr gene is another gene of interest commonly used to differenciate between different mycobacterial strains (Chimara *et al.*, 2004). It was decided that the MCP system would be used to check if this gene could be used to differenciate between these two strains.



List of Genes for Mycobacterium avium subsp. paratuberculosis Return to Main Page

Gene Details for : gyrB

atuberculosis Return

Gene Name	Start	Stop	Length	Direction	Description
gyrB	5258	7291	2033	0	GyrB

The Following Homologs were found :

Organism Name	Gene Name	Start	End	Length	Direction	# of SNPs	Protein
Mycobacterium bovis AF2122-97	gyrB	5123	7267	2144	0	342	unconserved
Mycobacterium leprae TN	gyrB	5229	7265	2036	0	243	unconserved
Mycobacterium tuberculosis CDC1551	MT0005	5207	7267	2060	0	258	unconserved
Mycobacterium bovis AF2122-97	gyrB	5123	7267	2144	0	235	unconserved
Mycobacterium ulcerans Agy99	MUL_0005; gyrB	5228	7306	2078	0	181	unconserved
Mycobacterium bovis AF2122-97	gyrB	5123	7267	2144	0	234	unconserved
Mycobacterium tuberculosis F11	TBFG_10005;	5224	7368	2144	0	235	unconserved

GENERATE PHYLOGENETIC TREES

The Protein Alignment :

```
gyrB_2      ---------- ---------- ---------- ---------V AAQR-KAQDE
MT0005      ---------- ---------- --------MH ATPEESIRIV AAQKKKAQDE
MUL_0005_   ---------- ---------- ----VTGPRA SNPEENIRTV AAQKKKAQDE
TBFG_10005 MGKNEARRSA LAPDHGTVVC DPLRRLNRMH ATPEESIRIV AAQKKKAQDE

            YGASAITVLE GLEAVRKRPG MYIGSTGERG LHHLIWEVVD NSVDEAMAGY
            YGAASITILE GLEAVRKRPG MYIGSTGERG LHHLIWEVVD NAVDEAMAGY
            YGAASITILE GLEAVRKRPG MYVGSTGERG LHHLIWEVVD NSVDEAMAGY
            YGAASITILE GLEAVRKRPG MYIGSTGERG LHHLIWEVVD NAVDEAMAGY
            YGASAITVLE GLEAVRKRPG MYIGSTGERG LHHLVWEVVD NSVDEAMAGY
            YGAASITILE GLEAVRKRPG MYIGSTGERG LHHLIWEVVD NAVDEAMAGY

            ADRVDVRILD DGSVEVADNG RGIPVAMHAT GAPTVDVVMT QLHAGGKFGG
            ATTVNVVLLE DGGVEVADDG RGIPVATHAS GIPTVDVVMT QLHAGGKFDS
            ATQVDVRLFD DGSVEVADNG RGIPVAVHAT GVPTVDVVMT QLHAGGKFGG
            ATTVNVVLLE DGGVEVADDG RGIPVATHAS GIPTVDVVMT QLHAGGKFDS
            ATQVNVRLLD DGSVEVADNG RGIPVAMHAT GAPTVDVVMT QLHAGGKFGG
```

Figure 4.10: gyrB gene details of *M. avium ssp paratuberculosis* K10 and its homologues.

It is interesting to note that no homologues for gyrB are detected within *M. tb* H37Ra or strain H37Rv. The gene however is known to be present in *M. tb* H37Rv, but is so vastly different on the nucleotide level to gyrB in *M. avium* that the MCP fails to detect it. Due to the way in which the MCP searches for homologues, gyrB could have gone undetected because the *M. tb* genomes (H37Rv, H37Ra etc) were not properly annotated or incompletely annotated. gyrB holds an essential function by encoding the B sub-unit of DNA gyrase and its presence is crucial to the organism. The fact that this gene exhibits such a high level of mutation among the mycobacteria is significant and not well understood. The variability for this gene though high, is controlled.

In terms of the gene islands detected for *M. tb* H37Rv, most of the features within these islands are comprised of genes belonging to the PE-PGRS/PPE family of proteins (Table 4.4)

Table 4.4: Annotations for the outlined genomic fragments for the *M. tb* H37Rv plot (Figure 3.10).

LEFT	RIGHT	n1_4mer:RV	n0_4mer:PS	GC	ANNOTATION
332000	342000	75.5	25.7	0.73	PE-PGRS and PPE family proteins
1630000	1638000	77.1	29.1	0.75	PE-PGRS and PPE family proteins
3734000	3746000	120.5	31.5	0.73	PE-PGRS and PPE family proteins
3924000	3954000	121.6	29.1	0.76	PE-PGRS and PPE FAMILY PROTEINS; acyl-CoA synthase; acyl-CoA dehydrogenase; acyl-CoA lygase FADD18; enoyl-CoA hydratase; thiamine-pyrophosphate requiring enzyme and many hypotheticals

Using the MCP, another comprehensive table was constructed containing PE-PGRS/PE/PPE comparisons (for loci 333437-3950263 of *M. tb* H37Rv) among all the database species present. The table was constructed in the following way. Starting with *M. tb* H37Rv, every annotation appearing between region 333437 and 3950263 was individually analysed using the MCP and the results stored in the table. Results that were recorded included, does the annotation appear in other species (i.e is there a homologue) or not; location of the homologue if found; is the homologue conserved or not; if not, how many SNPs appear in the homologue.

Table 4.5: Summarised table showing cross-species comparison of loci 333437-3950263 of *M. tb* H37Rv. (Note the SNP column indicates - total number of SNPs in homologue / number of SNPs per 100 bp).

H37Rv		H37Ra		F11	
Genes		NC_009525		NC_009565	
Code	Name/Position	Coord	SNP	Coord	
Rv0287c	PE_PGRS3 [333437-336310]	334799-337672	1 \ 0	336752-339472	193 \ 6
Rv1450c	PE_PGRS27 [1630638-1634627]	1631948-1636144	98 \ 2	1635000-1639223	109 \ 2
Rv3343c	PPE54 [3729364-3736935]	3740205-3746048	981 \ 14	3741691-3749289	209 \ 2
Rv3347c	PPE55 [3743711-3753184]	3752824-3762261	36 \ 0	3756176-3763594	2058 \ 24
Rv3350c	PPE56 [3755952-3767102]	3765065-3776089	126 \ 1	3768418-3779781	66 \ 0
Rv3507	PE_PGRS53 [3926569-3930714]	3935246-3939391	1 \ 0	3938922-3943133	36 \ 0
Rv3508	PE_PGRS54 [3931005-3936710]	3939682-3944319	1124 \ 21	3959787-3963047	968 \ 21
Rv3511	PE_PGRS55 [3939617-3941761]	3949108-3951261	7 \ 0	3953290-3958956	6 \ 0
Rv3512	PE_PGRS56 [3941724-3944963]	3949108-3951261	1285 \ 47	3953290-3958956	1246 \ 27
Rv3514	PE_PGRS57 [3945794-3950263]	**3939682-3944319**	**626 \ 13**	**3959787-3963047**	**727 \ 18**
H37Rv		CDC 1551		Bovis	
Genes		NC_009755		NC_002945	
Code	Name/Position	Coord	SNP	Coord	SNP
Rv0278c	PE_PGRS3 [333437-336310]	333551-336268	157 \ 5	334697-337330	360 \ 13
Rv0280	PPE3 [339364-340974]	339309-341036	30 \ 1	340366-341976	1 \ 0
Rv1450c	PE_PGRS27 [1630638-1634627]	No homologs		1632588-1634975	1287 \ 40
Rv1452c	PE_PGRS28 [1636004-1638229]	1636119-1638335	32 \ 1	1632588-1634975	112 \ 4
Rv3343c	PPE54 [3729364-3736935]	366212-375772	2018 \ 23	3686614-3690630	3708 \ 64
Rv3345c	PE_PGRS50 [3738158-3742774]	No homologs		3694689-3696308	3039 \ 97
Rv3347c	PPE55 [3743711-3753184]	424940-434767	2353 \ 24	3700427-3706717	3205 \ 40
Rv3350c	PPE56 [3755952-3767102]	424940-434767	2608 \ 24	3719324-3720628	9848 \ 158
Rv3507	PE_PGRS53 [3926569-3930714]	No homologs		3872192-3876274	215 \ 5
Rv3508	PE_PGRS54 [3931005-3936710]	No homologs		3890501-3893479	1154 \ 26
Rv3511	PE_PGRS55 [3939617-3941761]	No homologs		3883854-3889670	7 \ 0
Rv3512	PE_PGRS56 [3941724-3944963]	No homologs		3883854-3889670	1151 \ 25
Rv3514	PE_PGRS57 [3945794-3950263]	No homologs		**3890501-3893479**	**859 \ 23**
H37Rv		Ulcerans		Avium K10	
Genes		NC_008611		NC_002944	
Code	Name / Position	Coord	SNP	Coords	SNP
Rv0278c	PE_PGRS3 [333437-336310]	575111-576157	974 \ 49	4622313-4622930	889 \ 50
Rv0279c	PE_PGRS4 [336560-339073]	575111-576157	942 \ 52	No homologs	
Rv0280	PPE3 [339364-340974]	1289688-1291244	386 \ 24	3880292-3881887	300 \ 18
Rv1450c	PE_PGRS27 [1630638-1634627]	4845511-4848744	960 \ 26	No homologs	
Rv1452c	PE_PGRS28 [1636004-1638229]	4845511-4848744	562 \ 20	No homologs	
Rv3343c	PPE54 [3729364-3736935]	95273-96187	2229 \ 52	4399276-4399920	2257 \ 54
Rv3345c	PE_PGRS50 [3738158-3742774]	1221305-1223809	1057 \ 29	No homologs	
Rv3347c	PPE55 [3743711-3753184]	829924-831396	2867 \ 52	2895832-2896983	3706 \ 69
Rv3350c	PPE56 [3755952-3767102]	829924-831396	2804 \ 44	1673194-1674579	3534 \ 56
Rv3507	PE_PGRS53 [3926569-3930714]	402249-403910	1808 \ 62	No homologs	
Rv3508	PE_PGRS54 [3931005-3936710]	4456652-4457851	1764 \ 51	No homologs	
Rv3511	PE_PGRS55 [3939617-3941761]	4913637-4915358	540 \ 27	No homologs	
Rv3512	PE_PGRS56 [3941724-3944963]	4859588-4862377	935 \ 31	No homologs	
Rv3514	PE_PGRS57 [3945794-3950263]	4456652-4457851	1225 \ 43	No homologs	

The table essentially represents the gene island identified by the SWGB and all the sub-genetic elements found within this island. It is clear that this region is dominated by PE/PE-PGRS/PPE proteins. Several important findings are highlighted in the above table. Firstly the loci marked in blue for *M. tb* F11 are quite interesting as it appears to be genetically closer to *M. tb* H37Rv than strain H37Ra. This was not expected as the strain H37Ra was generated from strain H37Rv in a series of passages in petri dishes under controlled laboratory conditions. This finding will be discussed below.

It is also interesting to note that many of the genes above are not found within *M. avium* K10. Examination of the genome atlas of the *M. tb* H37Rv genome shows that loci 333437 to 3950263 are rich in repeats (direct and inverted). As already mentioned, this is characteristic is typical for this family of proteins.

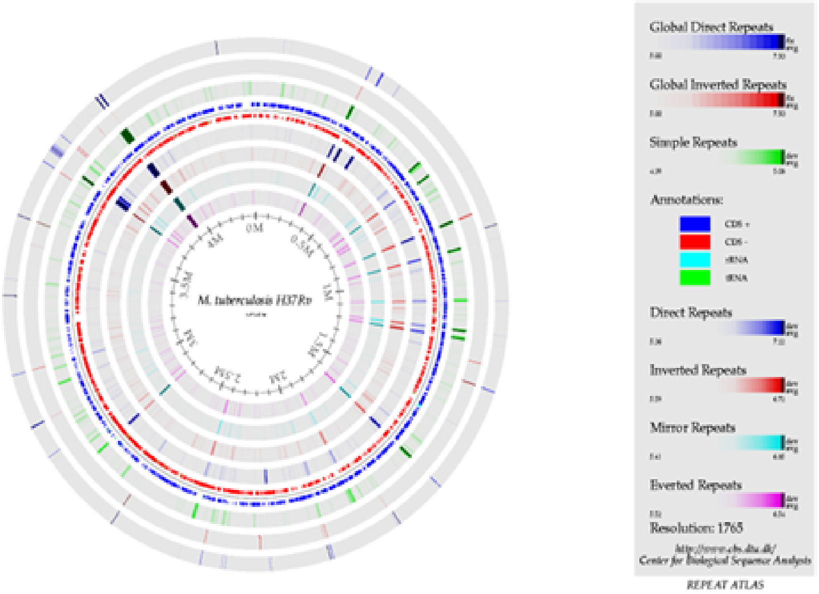


Figure 4.11: Genome atlas of *M. tb* H37Rv. Note the abundance of repeat regions especially in regions 3.9 – 4.0 MB (13).

Looking at a similar atlas for *M. avium* K10 reveals that these genes or gene area, unlike strain H37Rv, is actually core to the *M. avium* K10 genome (Figure 4.12) and does not contain an unusually high amount of repeats.

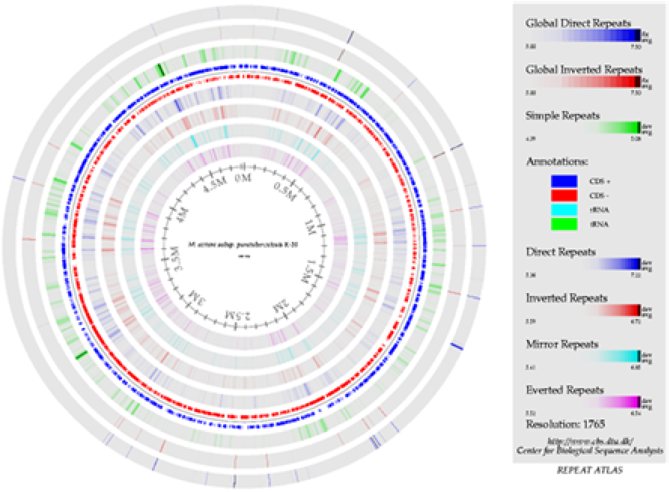
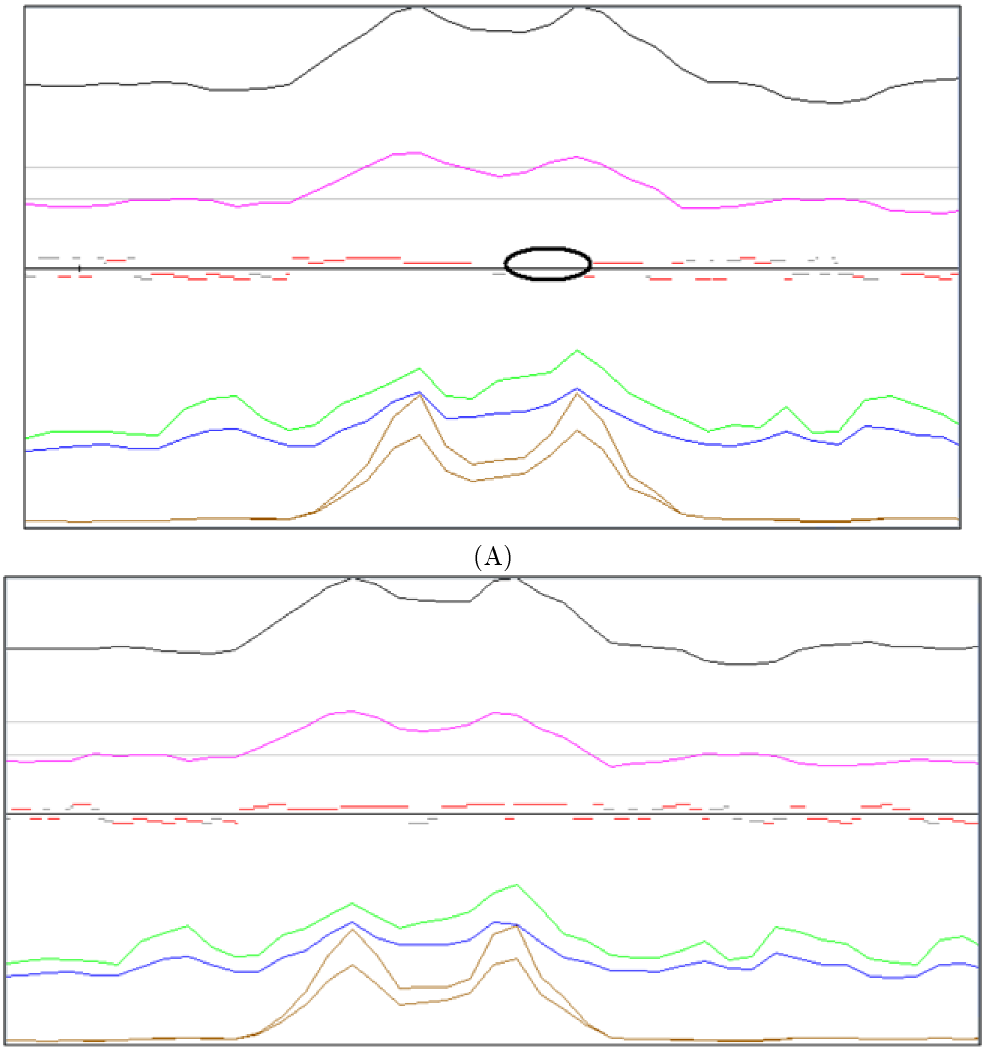


Figure 4.12: Genome atlas of *M. avium* K10 (13).

Another point of note is for that of annotation Rv3514, especially in relation to strain CDC1551. The lack of homologues is highly unusual. Going back to a SWGB view for this location in CDC1551 it can be seen that this discrepancy may have resulted from incompleted annotation information for the genome *M. tb* CDC1551 (Figure 4.13).



(A)

(B)

Figure 4.13: A Region of *M. tb* CDC1551 that appears to lack annotation information and B the corresponding region in *M. tb* H37Rv.

This illustrates that there still much vital information missing from our databases. In time however, with the increase in the amount of curated data, sequence information will definitely grow in richness.

4.11 Discussion

In this work the MCP system was used to compare rates of mutations in different genes of mycobacteria. One interesting finding was that the genes gyrA, gyrB, dnaA and dnaB, which are key players in the bacterial replication system and thus show high level of conservation among bacteria, were shown to exhibit a high degree of heterogeneity among the tested mycobacterial strains. Due to this heterogeneity, it can be suggested that the replication system among these mycobacterial strains are not conserved. There is a possibility that the genes in question may have evolved faster and these gene changes could have, in conjunction with several other events, lead up to the emergence of the slow-growing mycobacteria. However, to verify this, a genome-wide scale investigation would have to be undertaken for all genes amongst all the strains in question to acscertain with confidence if the gyr and dna genes are more susceptible to mutation relative to the other genes.

mmpL4_1 was also one of genes identified within *M. avium*'s K10 GI, and the MCP showed that this gene was entirely absent from H37Rv as suggested by the SWGB dot-plot.

The SWGB was able to identify crude regions of differences between each strain and the MCP is complementary in that it allows us to zoom into these areas of interest and allows us to see exactly how these genes differ even on the SNP level.

In terms of *M. tb* H37Rv, the GIs detected for this strain was found roughly between the loci 3.3-3.96 MB and was predominantly comprised of PE/PE-PGRS/PPE gene families. Using the MCP, a table was constructed summarizing all the genes within this GI in relation to other strains within the database collection. Several noteworthy observations will now be discussed.

M. tb strain H37Rv is the virulent precursor to strain H37Ra and thus should be the closest genetic relative to H37Ra. However, it is seen that in terms of the PE-PGRS genes in the above-mentioned loci, for several genes, *M. tb* H37Rv is actually closer to *M. tb* F11. Strain H37Ra is a laboratory strain while F11 is a clinical isolate. Therefore, the evolutionary pressure that H37Ra is under is quite different to what F11 had experienced over time. The time frames are not clear but what is clear is that these genes code for proteins that exhibit antigenic variation and thus contribute significantly to B-cell responses in TB patients (Tundup *et al.*, 2006). A global gene study needs to be unteraken to acsertion if these genes undergo greater mutation activity ('mutational hotspots') relative to the other genes within a mycobacterial genome. What is known is that strains H37Rv and F11 are both virulent whereas H37Ra is not. The contribution of these genes and the mutations they underwent, toward the virulence of *M. tb* (H37Rv and F11) and acquired avirulence of *M. tb* H37Ra is an area still not well understood.

The MCP system, showed the PE/PE-PGRS/PPE arrangement is indeed, non-random. There could be several reasons for the clustering of these gene families. van Pittius *et al.* showed that some members of the PE/PPE families are associated with the ESAT-6 (esx) genes cluster. The ESAT-6 gene clusters have been shown to be involved immunopathogenic activity. By employing techniques such as phylogenetics, DNA hybridization and comparative genomics the group showed the PE/PPE gene families expansion to be linked with the duplications occurring within the ESAT-6 gene clusters. They also propose that the duplication and distribition of the ESAT-6 gene clusters over time could explain the lineage of the slow-growing mycobacteria (van Pittius *et al.*, 2006).

The MCP was shown to be quite useful in highlighting useful features of mycobacterial genomes and seeing how various genes relate to other strains and shows the level of conservation between elements among strains. Another feature of the MCP, though not shown, is its ability to quickly construct phylogenetic trees. This is useful in allowing users to view phylogenetic distances of their genes/features of interest relative to the various mycobacterial species. The system could easily be adapted to include other mycobacterial species (as they become sequenced) as well as any other prokaryotic group with simple pre-calculations and database inclusions.

Research into tuberculosis is very much a growing concern and due to the bacteria's slow growth, sequence based research and diagnostic methods are advantageous. With the greater access researchers have to high-throughput sequencing techniques, means, these sequence analyses methods will find a niche. The ability to centralize and rapidly compare mycobacterium sequences is becoming an indispensible tool to researchers and diagnostic services alike.