Chapter 5

Concluding Discussion

The development of a comparative genomics environment is a task that has been tackled by many research groups in the past and indeed has proven to be quite successful for scientific purposes. In this study however, the aim was to create an integrated comparative genomics environment that combined together not only basic comparative genomics tools but novel tools as well. In the end, the usefulness of the various tools and approaches is shown by making some interesting discoveries about mycobacterial genomes and the behaviour of key mycobacterial sub-genetic elements amongst a few representative species. This project formed part of the greater Functional Genomics Information Management System (FunGIMS). Before dealing with the specifics of the findings made, a summary of the steps taken in development from conception to design and implementation will now be discussed.

The concept of a web-based, integrated comparative genomics environment in this context meant the combining of various tools and algorithmic approaches, in a highly accessible way, in order to facilitate scientific investigation of genomic sequence data. Combining various software tools under one environment while achieving integration relied on one very important factor. There had to be a highly versatile, robust and flexible data structure (MODEL) which would be able to handle the inputs and outputs from the various tools and data types within the environment. This was the first step toward realization of the project bar technology choice.

Before embarking on detailed design for the project, decisions had to be made regarding which software would be employed to construct the system. The project policy held for development was in favor of the open-source community, thus all software used for this project was borrowed from the open source community. Turbogears, a rapid web-development framework was used to develop the core of the system and the interface. CherryPy was the server of choice where the modules were deployed. SQLAlchemy, an object relational mapper was used in conjunction with Turbogears for the development of the database schema. Likewise, all scientific tools example

BlastZ, phylogenetic packages and sequence alignment programs were all sourced from the open-source community.

The model for this system was built upon the already established Functional Genomics Experiment (FuGE) model (Pizarro *et al.*, 2005). FuGE was designed with the prospect for extension to include other biological data types under a single model (Jones *et al.*, 2007) and thus laid the platform upon which a comprehensive customized model was built.

The basic analysis tools incorporated into the system included BlastZ for genome alignment, MAFFT and ClustalW for regular sub-genomic sized sequence alignment (protein and DNA), BLAST and Phylip. Using the Turbogears web-development framework, a common web-interface allowing access to the various tools was designed. The web-interface also allowed communication with the back-end database thus giving users access to their stored data and simultaneous usage of the software tools.

Together with the comparative genomics project, the novel offering of this system was the SeqWord Genome Browser (SWGB) and the Mycobacterial Comparison Project (MCP). These were included as sub-modules of the larger system. The SWGB provided users with a novel way of visualizing genomes based on its novel algorithmic approach to sequence analyses. The MCP system on the other hand, allowed users to perform gene-by-gene analyses on several mycobacterial genes simultaneously for various mycobacterial species, this, combined with SNP and annotation data allowed users a detailed view of specific parts of the genome and allowed rapid cross-species comparison.

Tuberculosis is a very relevant area of scientific interest, both locally and abroad. Due to the devastating impact this disease has had on the human population, mycobacteria was chosen as a test organism with which to validate the performance and value of the system. Using the unique offerings of both the SWGB and the MCP, a scientific investigation was carried out to explore a few of the major genetic differences between several key strains of mycobacteria.

By using the SWGB and adjusting its various plot parameters several genomic islands (i.e regions of atypical oligonucleotide usage) were identified in *M. avium* K10 and *M. tb* H37Rv. These gene islands oligonucleotide usage patterns vary so greatly from the rest of the genome that they are deemed to be of horizontal origin or may represent mutational hotspots. Gene islands may cover large areas of a genome and thus contain may sub-genomic elements that are of interest. The SWGB further enabled zooming into these genomic islands and retrieve annotations for the sub-genomic elements contained within the genomic islands.

In the case of *M. avium* K10, the sub-genomic elements contained within its gene islands were all

found to be linked with virulence or drug metabolism. For instance MmpL was one of the genes found. It is known that MmpL-mediated lipid secretion affects both the pathogens ability to survive intracellularly as well as host-pathogen dialogue which determines the ultimate outcome of infection (Domenech *et al.*, 2005). lipL was one of the other elements identified. lipL belongs to the hormone-sensitive lipase family and is responsible for fatty acid metabolism during an adverse nutrient climate (Deb *et al.*, 2006). This activity may account for *M. avium* K10's ability to utilize stored triacylglycerols during dormancy and its subsequent reactivation. These findings may be key to the understanding of the variation in pathogenesis and host interaction exhibited by *M. avium* and *M. tb* H37Rv.

In terms of genomic islands identified for *M. tb* H37Rv. It was found that PE-PGRS and PPE family of proteins were the dominant feature comprising the islands. PGRS genes are GC rich and have been shown to be a source of antigenic variation. This set of genes plays a role in the bacterium's survival within macrophages. This again points to the differences in the survival mechanisms between these species. Furthermore, these PE-PGRS genes were shown to cluster together. This phenomenon has been observed by other researchers and it has been shown that the duplication dynamics of the PE/PPE gene families is strongly linked to the duplication of the ESAT-6 gene clusters within a genome (van Pittius *et al.*, 2006). Although we have seen (amongst our tested strains), that the PE/PPE gene families do exhibit mutation activity, we are yet to show that it occurs at a rate greater than that observed for the rest of the other gene families within the genome. In a related study however, an examination of the PPE38 genomic region was performed. PPE38 belongs to the PPE gene family of *M .tb* (and other related mycobacteria). By looking at the PPE38 region in a cohort of clinical *M. tb* isolates, it was demonstrated that this region was in fact hypervariable due to insertion events involving IS6110 elements, IS6110-associated recombination, homologous recombination and gene conversion events between its PPE38 and PPE71 constiutent homologues (McEvoy *et al.*, 2009). Rapid molecular evolution was therefore established for the region. This then lends support to the idea that PPE/PE related genomic loci within our tested strains may also be under similar 'increased' rates of mutational activity.

Also revealed by the SWGB was that the genomes of *M. ulcerans* and *M. marinum* show intermediate states of evolution (based on RV-PS plots) from the pattern of *M. avium* toward that of *M. tb* H37Rv. This was interesting in that it almost summarizes the mutational events occurring as these species evolve.

The SWGB afforded us a unique view of various mycobacterial genomes and highlighted areas of 'non-normal' oligonucleotide usage patterns. Within these gene islands, specific genes were then identified. The mycobacterial comparison project was then used to look deeper into these genes in order to gauge the roles they may play within the mycobacteria.

Continuing from the SWGB findings, the MCP system was used to make more interesting findings. Amongst the elements found within the genomic islands of *M. avium* K10, were *dna*A/B genes. These genes have important roles to play in the bacterial replication system and are thus expected to have high levels of conservation. The MCP system revealed, however, that the genes showed a high-degree of non-conservation among the tested mycobacteria. This may suggest that the replication system of mycobacteria has evolved faster. Whether the changes within this gene set is one of the contributers to the eventual appearance of slow-growing mycobacterial phenotypes remains to be proved.

The SWGB highlighted a high PE-PGRS/PPE gene content within the gene islands of *M. tb* H37Rv. Using the MCP system, a PE-PGRS profile was created for all the mycobacterial genomes within the system. The results achieved were very interesting. Mycobacterial strain H37Ra is essentially a laboratory derivative of *M. tb* H37Rv. As a result, it was expected that the genetic profile of these two strains would be closest to each other. However, the MCP system revealed that *M. tb* H37Rv was actually closer to *M. tb* F11(a virulent strain) in its PE-PGRS profile than it was to H37Ra. This further underscored the roles these genes play in virulence of mycobacteria. It is accepted that the evolutionary pressures experienced by H37Ra is quite different to that of *M. tb* F11. The time frames over which these changes took place are also not clear but what is clear from the literature is that some gene families within the PPE family do give insight into the lineage of various species within the mycobacterial genus (McEvoy *et al.*, 2009).

The ability to identify the similarities and differences between mycobacterial strains is of the utmost importance to understanding the genetic causes of diseases and the differences in disease processes and tropisms exhibited by the various mycobacterial strains. In this project, several novel approaches to identify key areas of interest between various mycobacteria were used. In identifying these areas scientists can now be more focused in their research and solidify the links between clinical phenotypes and genomic loci. TB researchers will find it very useful to be able to now pin-point and analyse specific genomic areas and immediately see how comparative loci in other species compare. This functionality is also helpful in the ability to rapidly detect drug resistant/susceptible phenotypes. In this day and age as high-throughput sequencing techniques becomes more cost effective and faster, comparative sequenced based methods will become a most valuable research and discovery tool.