Pangenome Modeling for Analyzing the Evolution of *Mycobacterium tuberculosis*

ZHOU, Haokui

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy
in
Microbiology

The Chinese University of Hong Kong
February 2014

Abstract of thesis entitled:

Pangenome Modeling for Analyzing the Evolution of *Mycobacterium tuberculosis*

Submitted by ZHOU, Haokui

for the degree of Doctor of Philosophy

at The Chinese University of Hong Kong in February 2014

Abstract

Comparative analysis of multiple genomes of the same microbial species has led to the concept of *pangenome* to characterize the variations of gene content in different strains and to study their relationship to strain phenotype variations. Pangenome studies of microbial pathogens have identified strain-specific genes that may play roles in the evolution and adaptation of the pathogens. In previous studies, much attention was paid to estimate the size of the pangenome of different microbial species. But it is also important to develop bioinformatic methods for analyzing the evolution of the pangenome of a species, such as gene gain and loss or coevolution of clusters of genes, which may help to associate genotype variations with phenotype variations of a microbial species, and thus provides biological insights for further studies.

In this thesis, to analyze the pangenome consisting of complete mycobacterial genomes from public database and additional five *Mycobacterium tuberculosis* (MTB) Beijing genotype genomes sequenced by our own project, two bioinformatic approaches have been developed. The first is a local parsimony ancestral state reconstruction method, which was used to analyze genome-wide indels evolution of the MTB Beijing genotype. The key finding was that reductive evolution shaped the formation of not only different MTB species, but also different subspecies or genotypes, such as the Beijing genotype, for

which genome-wide deletions of large RDs and disruption of individual genes were identified. This finding might have implications for the virulence evolution of the Beijing genotype. The method also provides an alternative perspective to understand parsimony analysis in phylogenetics, which can be used to incorporate statistical analysis into the method.

The second approach developed is a pangenome phyletic model for analyzing the coevolution of genes in the pangenome of a microbial species. This phyletic model calculates coevolution scores of gene frequencies in a pangenome. And graph-based clustering is used to identify coevolved clusters of genes. Applying this method to the genus *Mycobacterium* helped us to identify various gene clusters, from conserved core clusters of housekeeping genes to species-specific clusters, including genes related to pathogenesis. The key finding was that different MTB species have arose from their mycobacterial ancestor mainly by loss of many environmental related genes. On the other hand, gain of genes has also occurred within the MTB genomes, especially the clusters of the PE/PPE genes. This finding implied that the MTB species have undergone reductive evolution from an environmental species to adapt to and coevolve with their specific hosts.

In conclusion, the two methods were shown to be powerful in analyzing the pangenome of the MTB species and also of the *Mycobacterium* genus, and have provided useful insights into their genome and virulence evolution for further studies, including both pathogenesis related genes and genotyping genetic markers. Future works in that direction is to introduce stochastic models of gene evolution into these two methods. Finally, this work indicated that pangenome modeling is critical and can provide a good starting point for comprehensive pangenome sequencing of mycobacteria. Therefore, a database of *Mycobacterial* genomes for integrative pangenome annotation and evolutionary analysis should be developed.

论文摘要

泛基因组的概念来源于比较分析同一微生物物种的多个基因组。泛基因组分析已经被用于研究病原微生物基因组的变化，并且揭示了与菌株进化和宿主适应相关的特异基因。目前泛基因组研究主要集中在估计不同物种的泛基因组大小。但是对泛基因组的结构进行进化分析的生物信息方法还有待开发，以便研究不同基因在不同菌株的进化，比如基因的获得或者丢失，或者共同进化的基因簇。这样的分析方法可以把基因和菌株之间的表型关联起来，为进一步的生物学实验提供线索。

为了研究结核分枝杆菌种和分枝杆菌属泛基因组进化的规律并揭示其生物学意义，本论文开发了两种泛基因组数据分析的生物信息学方法。第一个是基于局部最大简约计算的祖先状态重构算法。它被用来分析结核分枝杆菌北京型全基因组水平插入／缺失序列（indels）的进化。分析表明基因组退化不仅塑造了该物种不同亚种的形成，并且也塑造了同一亚种不同亚型的分化，比如北京型。该分析还找出了北京型全基因组水平的 RD 区域和各种被中断的基因，这些基因可能同北京型的毒力进化相关。同时，该算法提供了另一个理解简约分析的视角；该视角可以把统计分析引入到该算法中。本论文提出的第二个模型是基于泛基因组进化的基因聚类模型。通过计算泛基因组中不同基因家族的分布频率，结合基于图论的聚类算法，该模型可以找出泛基因组中共同进化的基因聚类。对分枝杆菌属的泛基因组进行聚类分析发现了不同类别的基因簇，它们与不同分枝杆菌种的表型进化相关。这些结果说明了，一方面结核分枝杆菌在进化过程中丢失大量环境相关的基因；另一方面，它可能通过水平基因转移获得一些基因，特别是 PE/PPE 基因家族。因此，结核分枝杆菌可能是通过不断的基因组收缩，从一个环境菌种进化为与宿主共进化的病原菌。

总地来说，上面的两种方法能够被有效地用于结核分枝杆菌种和分枝杆菌属的泛基因组分析。 将来的工作可以考虑进一步引进随机模型；同时需要建立分枝杆菌的泛基因组数据库，以面对大规模测序的需求。

Acknowledgements

I would like to express my sincere gratitude and appreciation to my supervisor, Prof. Guoping Zhao, for his guidance and encouragement throughout my PhD work. The many discussions between us have helped me to begin and develop my research in the right direction. And his insightful advice has also help me to combine bioinformatics and genome biology in a better way. Without his continuous support, completion of this PhD work would not have been possible.

I also want to thank the colleagues who have worked with me in the MTB project. I would like to express my appreciation to Prof. Stephen K.W. Tsui (School of Biomedical Sciences, The Chinese University of Hong Kong) for his valuable ideas and suggestions. I am also grateful to Dr. K. K. Leung (School of Biomedical Sciences, The Chinese University of Hong Kong) for his communication with me about bioinformatics and data analysis. And I want to thank Dr. Huajun Zheng (Chinese National Human Genome Center at Shanghai) for his valuable helps in the genome data and also lots of discussions about the MTB biology.

I am also grateful to the members of my department (Department of Microbiology, The Chinese University of Hong Kong). They provide such a good opportunity for me to join with them in the department seminars. This has helped me to improve my presentation skills and to broaden my microbiology knowledge so much. I am also grateful to Mrs. Corrie Leung for her helps in the university affairs during my study.

Last but not least, I want to thank my family, who have supported me all the way through with their love.

Table of Contents

List of Abbreviations

ASR	Ancestral state reconstruction
BCG	Bacilli Calmette-Guerin
CC	Coevolved cluster of genes
CMN	The *Corynebacteriaceae*, *Mycobacteriaceae* and *Norcardiaceae* group
COGs	Clusters of orthologous groups
CRISPR/Cas	Clustered regularly interspaced short palindromic repeats and CRISPR-associated (cas) genes
DGH	The distributed genome hypothesis
dN/dS	The ratio of the number of nonsynonymous substitutions per non-synonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks)
DU	Tandem duplications in the *Mycobacterium bovis* BCG genome
GBS	Group B *streptococcus*
HGT	Horizontal gene transfer
IMG	The infinitely many genes model
indel	Insertion and deletion sequence
IS	Insertion sequence
IS6110	An insertion element found exclusively within the members of the *Mycobacterium tuberculosis* complex
LCA	Lowest common ancestor
LCB	Locally collinear blocks of genome sequence
MCL	The Markov cluster algorithm
MIRU	Mycobacterial interspersed repetitive unit
MRCA	Most recent common ancestor
MTB	*Mycobacterium tuberculosis*
MTBC	The *Mycobacterium tuberculosis* complex
NBjD	Non-MTB Beijing family deletion

NOG	Non-supervised orthologous group
NTM	Nontuberculous mycobacteria
ORF	Open reading frame
PE	A multigene family encoded in *Mycobacterium tuberculosis* genomes, with motifs Pro–Glu (PE) found near the N terminus in most cases
PGRS	Polymorphic GC-rich sequence (PGRS)-containing genes
PPE	A multigene family encoded in *Mycobacterium tuberculosis* genomes, with motifs Pro–Pro–Glu (PPE) found near the N terminus in most cases
RD	Region of difference
RFLP	Restriction fragment length polymorphism
SNP	Single-nucleotide polymorphism
TB	Tuberculosis
TR	Tandem repeat
VNTR	Variable number tandem repeat

Chapter 1

Pangenome analysis of microbial species

1.1 Introduction

A vast diversity in gene content among bacterial genomes has been revealed even in strains of the same microbial species, as shown in a comparative genomics study of eight *Streptococcus agalactiae* genomes, which proposed the *pangenome* concept to quantify the variability of gene content in a microbial species (Tettelin, Masignani et al. 2005). Following this pioneer work, driven by an accelerating effort of whole genome sequencing using high throughput next generation sequencing technologies, more and more microbial species of clinical or environmental importance have been subjected to pangenome analysis (Hiller, Janto et al. 2007; Hogg, Hu et al. 2007; Kettler, Martiny et al. 2007; Rasko, Rosovitz et al. 2008; Donati, Hiller et al. 2010; Muzzi and Donati 2011). As a result, various biological findings from these species have been illustrated in a pangenome context, which characterizes species variability with different sub-species phenotypes, such as serotypes of human bacterial pathogens or ecotypes of environment species (Tettelin, Masignani et al. 2005; Johnson, Zinser et al. 2006). Most of the variations are attributed to the mobilome of the microbial world, horizontally transferred between different strains via the three well-recognized routes – transformation, transduction and conjugation (Gogarten and Townsend 2005). Biologically, these variations are associated with the phenotypes of their carriage strains in response to environmental challenges, including virulence factors for host colonization, antibiotic resistance genes and transporters for niche partition and adaptation (Dobrindt, Hochhut et al. 2004; Ambur, Davidsen et al. 2009; Fischer, Windhager et al. 2010).

In addition to the many biological findings, there is also a theoretical thread of works regarding the modeling of pangenome evolution of different microbial species (Tettelin, Riley et al. 2008; Lapierre and Gogarten 2009). The first question raised is how many genomes are required to fully describe a species. This is a pangenome size estimation problem. Intuitively, the first approach to this problem was to consider the genome sequencing process as sequential sampling and cataloging process. Because the sample space is unknown in prior, regression analysis of fitting gene-discovery curve was performed and then extrapolated to estimate different pangenome quantities, such as core-genome size, pan-genome size and discovery rates of new genes (Tettelin, Masignani et al. 2005; Tettelin, Riley et al. 2008). The most intriguing result from this approach is the inference of an 'open' pangenome for some species, which have an infinite gene repertoire, although the other species have a 'close' pangenome of finite gene content (Tettelin, Masignani et al. 2005; Tettelin, Riley et al. 2008). In contrast to this extrapolation approach and its 'open' pangenome implication, there are studies that made an explicit requirement of finite pangenome size for modeling to account for occurrence frequencies of all gene families in the pangenome. This approach has been denoted as 'supragenome' analysis, with an emphasis on the gene frequency modeling (Hogg, Hu et al. 2007; Snipen, Almoy et al. 2009). A more sophisticated probabilistic method has been also published, the 'infinitely many genes model'. In this model, dynamics of gene gain and loss is formulated along a genealogy of individual genomes (Baumdicker, Hess et al. 2012; Collins and Higgs 2012).

With the information of gene distribution within the pangenome of a species, especially those species with rich phenotype information, another important line of works is to associate gene distribution with phenotype diversity. Such works have been performed to analyze the gene gain and loss during the evolution of a species (Kettler, Martiny et al. 2007; Lefebure and Stanhope 2007). Gain or loss of gene families specific to lineages or sub-lineages has been demonstrated to be associated with their phenotypes (Kettler, Martiny et

al. 2007). For example, putative virulence factors or possible pathogenesis mechanism was revealed from comparison of pathogenic strains with their commensal counterparts from the same species (Fischer, Windhager et al. 2010). This also provides insights into the formation of the pangenome during species evolution, especially the biological roles of the dispensable genes, which are dynamic and responsible for species adaptation (Ambur, Davidsen et al. 2009).

In this chapter, I reviewed the concept of pangenome from comparative genomic studies of some bacterial species, and the mathematical models developed so far for pangenome size estimation. The assumptions and their mathematical formulas were examined for each model, and also implication and limitation of the models. Studies of gene gain and loss within the pangenome of a species with strains of different phenotypes were also included. At last I concluded this chapter with a phylogenetic perspective of pangenome analysis. In this perspective, attention will be paid to the evolutionary processes that generate the extant species pangenome. This will be more useful for both theoretical modeling of pangenome analysis and for the interpretation of pangenome structure with regard to genotype-phenotype association analysis.

1.2 The concept of pangenome

Comparative genomics analysis of multiple individual genomes of the same microbial species have shown a variability of gene content among them, from species of conserved genomes to species with a large gene repertoire (Medini, Donati et al. 2005; Read and Ussery 2006). These observations have been further extended by large-scale comparative genomics studies of clinical or environmental microbial species (Muzzi and Donati 2011). Such a variability of gene content of individual species can be explained by the interaction of two evolutionary forces, the maintenance of essential species features by vertical

inheritance of genetic materials and dynamically adaptation to environmental conditions by horizontal gene transfer. As a result, strains of the same species can have distinct sub-species phenotypes, for example, host commensal or pathogenic for some human bacterial pathogens (Chen, Hung et al. 2006).

To understand the diversity of gene content and their relation with phenotype variations in a microbial species, a pangenome concept was proposed to categorize the gene repertoire of a microbial species (Tettelin, Masignani et al. 2005). Based on the distribution of orthologous genes (or gene families) among individual genomes of the species – presence or absence in a particular genome, gene content of the whole species can be classified into core-genes, which are shared by all individuals and make up a core-genome, and dispensable-genes, which present only in a subset of individuals and form a dispensable-genome. The dispensable genes also include rare genes unique to individual genomes.

With this classification of the gene pools, it is possible to answer a species-sequencing problem: how many individual genomes are necessary to represent a whole species (Medini, Donati et al. 2005). This problem was initially tackled in a pangenome study of the human pathogen *Streptococcus agalactiae* (Tettelin, Masignani et al. 2005), which was shown to have a core-genome of about 1,800 genes, accounting for ~80% of the gene content of any single genome. The remained 20% of each genome belong to the dispensable genome, but the number of new genes contributed by adding a new genome shows no tendency to reach zero, even with hundreds of genomes sequenced. For *S. agalactiae*, this number was estimated to be 33 genes on average for each new genome, which implies an infinite pangenome of this species. Functional classification of the core and dispensable genes indicated different biological and evolutionary roles of them. The core genes consist of gene families of housekeeping functions, and also including genes involved in the cell envelope, regulatory functions, and various transporters, which in

together define the essential and basic biological features of the species. In contrast, the dispensable genes, which are dynamic and much more variable, are mainly made of groups of genes associated with genomic islands, integrated phage genomes and other transposable elements. Among them, virulence related genes, and resistance genes are commonly identified, but most of them are gene of unknown function. Therefore, collectively, the dispensable genome confers the species abilities to incorporate new functions from external sources to adapt to new environmental challenges. Furthermore, by extending the pangenome concept to the whole bacteria domain, Lapierre *et al.* described three distinct pools of gene families that comprise the bacterial pangenome, based on describing their occurrence frequencies among all the analyzed genomes (Lapierre and Gogarten 2009). The estimated average bacterial pangenome was shown to have three components of the 'extended core genes' (in total ~250 gene families, and ~8% of an average single genome), the 'assessor genes' (in total >139,000 gene families, and ~28% of an average single genome), and the 'character genes' (in total ~7,900 gene families, and ~64% of an average single genome).

1.3 Estimation of pangenome size using regression analysis

In the GBS pangenome work of Tettelin *et al*., empirical observation has been made that, along the sequential addition of new sequenced genomes, the observed number of core genes is decreasing, and the same for the observed number of new arriving genes (Tettelin, Masignani et al. 2005). This is a decaying process, and which has served as the starting point in their modeling work, that is to fit an exponential decay function to the observed data. Before regression analysis to fit an exponential function, a data simulation step has been taken to generate all possible combination of sequential inclusion of new genomes, and to observe the number of core genes and new genes.

Given *n-1* total number of genomes already included, an *n*-th additional of a new genome is simulated, and also the number of core genes shared among all the *n* genome, and number of new genes introduced by this *n*-th adding genome. Suppose the total number of available genomes to simulated is *m*, then the number of observations for such an *n*-th addition is:

$$N = \binom{m}{n-1}\binom{m+1-n}{1} = \frac{m!}{(n-1)!(m-n)!}, \quad (1)$$

This formula gives $8!/(0!\,7!) = 8$ observations, for a first inclusion of a new genome, and $8!/(1!\,6!) = 56$, for a second inclusion. All the observed values in the *n*-th inclusion are then averaged to generate data points for following regression analysis. This summarization of data points can void the heterogeneity of observations from each individual genome and capture the general tendency of data from the sample.

The exponential decay function for fitting the number of core genes is:

$$F_c(n) = \kappa_c e^{-\frac{n}{\tau_c}} + \Omega, \quad (2)$$

with κ_c, τ_c and Ω are parameters to estimated, which represent the decaying amplitude, decaying speed and the converged asymptotic value respectively. In the pangenome context, the first two parameters define the shape the exponential decay curve, and the last parameter measures the asymptotic size of the core genes when there are $n \to \infty$ genomes sampled.

In a similar way, the function for fitting the number of new genes is:

$$F_s(n) = \kappa_s e^{-\frac{n}{\tau_s}} + tg(\theta), \quad (3)$$

with the $tg(\theta)$ as a parameter measuring the estimated number of new genes.

To estimate the size of pangenome, we can add up the number of new genes added from each *n*-th addition:

$$P(n) = \sum_{i=1}^{n}\left(\kappa_s e^{-\frac{n}{\tau_s}} + tg(\theta)\right)$$

$$= n \cdot tg(\theta) + \kappa_s e^{-\frac{1}{\tau_s}} \frac{1-e^{-\frac{n}{\tau_s}}}{1-e^{-\frac{1}{\tau_s}}}. \qquad (4)$$

Equation (3) gives that, $\lim_{n\to\infty} P(n) \approx n \cdot tg(\theta)$, which is a estimation of pangenome size, and means that this size goes to infinite when $tg(\theta) > 0$.

After arguing for a reasonable conclusion of the possible 'open' pangenome estimation, when considering a vast diversity of genes in the microbial world and the observed frequent gene exchanging between species, Tettelin *et al*. proposed a new model of broad applicability (Tettelin, Riley et al. 2008). This model is based on an empirical law, called Heaps' law, arising from studies of new words discovery among with scanning more and more instance text. Heap's law, which is a form of power law, is not limited to text corpora, but also appears in a variety of phenomena with the field of complexity science. A pangenome analogy with this model is that, it is becoming increasingly harder to discover new genes with more and more genomes sequenced. In the case of pangenome, the formula is:

$$P(n) = kn^{b} \ (0 < b < 1), \qquad (5)$$

in which, *P(n)* is pangenome size and *n* the number of genomes observed. The parameter κ is a scale coefficient and β a growth parameter. From equation (5), the rate of discovering new genes can be described as:

$$F(n) = P(n) \Leftarrow kbn^{b-1} \sim n^{-a} \ (a = 1 - b) \qquad (6)$$

The values of parameter α describe different power law behaviors with respect to whether equation (5) converges or not. For $\alpha > 1$ ($\beta < 0$), the equation converges to an

asymptotic value, and presents a close pangenome, while for $\alpha \leq 1$ $(0 < \beta < 1)$, the equation doesn't have an asymptote and indicates an open pangenome. When applying to real genome data from several species, this power law function regression displayed better fitting results than regression with exponential function, particularly reflecting in the extrapolation of trends of discovering new genes. Also, using power law function has a merit of fewer free parameters to estimate, a multiplicative constant κ and a growth factor β, but together empirically describe a wide range of natural systems.

1.4 Estimation of gene frequency using a finite supragenome model

Another approach to pangenome modeling is to probabilistically model gene occurrence in the pangenome based on the observation of their frequencies, which has been adopted in the study of Hogg *et al* (Hogg, Hu et al. 2007). This approach, denoted as 'supragenome' model, assumes that the pangenome takes a finite number of gene families with different frequency of occurrence, in contrast to an open pangenome possibility from regression analysis.

The model begins with a pre-set of *K* class of gene occurrence frequency μ_k in the supragenome. These parameters are not trained from data, but pre-defined. In the Hogg *et al.*'s modeling of the *Haemophilus influenzae* pangenome, the *K* was selected to be 7, and $\vec{\mu} = \langle 0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0 \rangle$ accordingly. The class of $\mu_7 = 1.0$ represents core genes, which occur in all the genomes, and $\mu_1 = 0.01$ for rare-genes that are unique to small number of strains. For a gene of class *k* to be observed in *n* genomes within a pangenome consisting of *S* genomes, or success *n* times from *S* trials with the supragenome, the observation probability is:

$$P(x = n|\mu_k) = \binom{S}{n}\mu_k^n(1 - \mu_k)^{S-n}. \qquad (7)$$

With the *K* classes of gene occurrence frequency, the possibility of a gene that belongs to class *k* is defined with another parameter π_k. From equation (7), the probability of observing a gene in *n* genomes, considering contribution from all the *K* class, is:

$$P(x = n|\vec{\pi}, \vec{\mu}) = \sum_{k=1}^{K} \pi_k \binom{S}{n} \mu_k^n (1 - \mu_k)^{S-n}. \qquad (8)$$

By definition, the supragenome consist of all the genes of different occurrence frequency, from *0* (not observed yet) to *S* (present in all sampled genomes), designated as $\vec{c} = \langle c_0, c_1, \ldots, c_S \rangle$. And the observation a supragenome of *N* genes can be modeled as, treating occurrence of each frequency independently:

$$P(\vec{c}|N, \vec{\pi}, \vec{\mu}) = \frac{N!}{c_0! c_1! \ldots c_S!} \prod_{n=0}^{S} \left(\sum_{k=1}^{K} \pi_k \binom{S}{n} \mu_k^n (1 - \mu_k)^{S-n} \right)^{c_n}, \qquad (9)$$

which can be maximum optimized to fit the real observation of pangenome data to estimate the parameters of *N* and $\vec{\pi}$, which are the size of pangenome and portions of genes belonging to the *K* different class of occurrence ($\sum_k^K \pi_k = 1$), respectively. In Hogg *et al.*'s treatment, the maximum likelihood estimation was performed with a fixed N choosing a range (from the number of all observed genes to a quite large value) and then maximized the value of *P* with respect to $\vec{\pi}$. The log-likelihood estimation function for this is:

$$logP(\vec{c}|N, \vec{\pi}, \vec{\mu}) = C + \sum_{n=0}^{S} C_n \log\left(\sum_{k=1}^{K} \pi_k \binom{S}{n} \mu_k^n (1 - \mu_k)^{S-n} \right). \qquad (10)$$

The concepts behind this supragenome model are sound with respect its explicit modeling of occurrence frequency of gene families. There are two critical components underpinning the model, $\vec{\mu}$, which is a rule of diversity of frequency classes and assumed to be universal regardless of the specific species under studied, and $\vec{\pi}$, which is related to the nature of the studied species and the one of interest. But prediction using this model changed very much with different size of data set to train, as demonstrated in their study, from 3,078 genes with 8 trained genomes to 5,230 with 13 trained genomes. Most of the

discrepancy relies in the prediction of the number of rare genes (the frequency class of 0.01). Significantly more rare genes (or unique genes) were predicted with 13 genomes as input, which implies a similar tendency of new gene discovering as the results from regression analysis.

In response to the pre-setting of K and its associated frequency $\vec{\mu}$ in the above modeling, Snipen *et al.* tried to train these parameters from the data (Snipen, Almoy et al. 2009). First, they iterated a procedure of a 'zero-truncated log-likelihood' estimation, which omits the term of c_0 in equation (10), but with K predefined, to get estimated values for $\vec{\mu}$. Second, they tried to estimate K from the data using the Bayesian information criterion. This extension of the original model seems to select a proper numbers of frequency class and their associated frequencies, which are crucial in the estimation of pangenome size. As shown in their results, there were at least three frequency classes (or component) detected in all species, and a pre-setting of 7 classes seamed to be over selected, which could lead to an over-estimation of pangenome size. But the introduction of such many parameters ($2K$-2) in the binomial mixture model has to be justified, which define a complex model and can be over-fitting the training data.

1.5 The infinite gene models with Kingman's coalescent process

The infinite gene models (IMG) assumes that a genome consists of two parts, core-genome and dispensable genome, which are evolving with different dynamics (Baumdicker, Hess et al. 2012). The core-genome is stable and unchanged during evolution, whereas the dispensable genome evolves by gaining genes from an infinitely large external gene pool or losing genes independently. To formulate the dynamics of this changeable dispensable genome, two fundamental assumptions and the corresponding mathematical models are used. The first is a birth and death Markov process, which assumes that genes are constantly

acquired at a constant rate of $\theta/2$, or independently deleted at rate $\rho/2$. This model gives a conversed Poisson distribution of the dispensable genome size, with the expectation value of $m = \theta/\rho$.

To model a neutral evolution of the dispensable pangenome of *n* genomes, the Kingman's coalescent process is used to generate a genealogy of the *n* genomes, which gives expected time intervals for the events of two descent genomes to find their immediate common ancestor. Imposing the gene birth and death process described above on this genealogy allows for calculating a rich set of pangenome quantities. The first one, $|G^n|$, expected number of different genes in a pangenome of size *n*, is given by:

$$E(|G^n|) = C + \theta \sum_{i=0}^{n-1} \frac{1}{i+\rho}, \quad (11)$$

in which, *C* is the size of core-genome.

And number of new genes (S_n) to be discovered from an *n*-th genome is:

$$E(S_n) = \frac{\theta}{n-1+\rho}. \quad (12)$$

Denoting G_k^n as the number of genes occurred in *k* individuals of the total *n* genomes, which are the gene frequencies of the supragenome, and they are given by:

$$E(G_k^n) = \frac{\theta}{k} \cdot \frac{n \cdots (n-k+1)}{(n-1+\rho) \cdots (n-k+\rho)},$$

$$E(G_n^n) = C + \theta \cdot \frac{(n-1)!}{(n-1+\rho) \cdots \rho}. \quad (13)$$

In these equations, there are two parameters θ and ρ to be estimated from data. For a real pangenome dataset, the observed gene frequencies can be calculated, and also, a real genealogy tree can be reconstructed from phylogenetic marker genes. With these two kinds of input data, maximum likelihood can be used to estimate the optimal values for θ and ρ.

When applied this model to the case of *Prochlorococcus* pangenome, which shows a more neutral evolution that is assumed in this infinite gene model, the results are only comparable to the other two models described above. And the major difference between them is the prediction of the growth of pangenome size along with adding more and more individual genomes. This model describes a logarithmic growth, while the regress model assumes a power law growth and the binomial mixture supragenome coming with a closed pangenome.

1.6 Analysis of gene gain and loss in pangenome evolution

Besides the estimation of pangenome size and thus effort required to fully characterize the genetic diversity of a microbial species, it is also important to understand how the pangenome has evolved and adapted to exhibit the extant patterns of gene distribution and their association to phenotypes. In this regard, the dispensable genome is of more interest to analyze: genes acquired or deleted randomly would be less relevant, but those genes that show associated patterns with species phenotypes would lead to discovery of the underlying molecular mechanism and its evolutionary histories.

Usually, genotype-phenotype association analysis needs to incorporate phylogenetic relationships of the analyzed genomes, since the genomes are not independent to each other but with different evolutionary distance, as indicated by their phylogenetic tree. To analyze the evolution of the dispensable genome of a species, gene gain and loss events can be derived from inference of the presence or absence status of genes in the ancestral nodes of a reference species tree, using well-established ancestral state reconstruction methods, such as maximum parsimony. Such species tree is usually reconstructed from the core genome of the species. With this presence or absence information, the process of gene gain and loss during the evolution and divergence of the species ecotypes or lineages, will be revealed.

This kind of analysis has been taken in several pangenome studies. In a study of 12 genomes of *Prochlorococcus*, an abundant photosynthetic bacterium living in the ocean, Kettler *et al.* revealed the patterns of gene gain and loss in the evolution of its pangenome and their biological implications (Kettler, Martiny et al. 2007). Genes responding to the niche partitions of high-light and low-light adapted ecotypes have been identified, which are involved in DNA repair, protecting photosystems from oxidative damage, stress responses, and also phage-related genes that are commonly found among dispensable genes. In addition to the events of gene gain and loss affined to lineages or clads, most of the events occur in the 'leaves of the tree', meaning gene content variations among close related strains that are dynamic and instable. Another study examined gene gain and loss in a genus level, the genus *Streptococcus*, which comprises several human pathogen species (Lefebure and Stanhope 2007). With the 26 *Streptococcus* genomes compared, it was found that there are more gene gains than losses leading to species separation. And higher numbers of gene gain and loss is associated with a larger pangenome size for *Streptococcus agalactiae*, than that for *Streptococcus pyogenes*, which are with fewer numbers of gain and loss and of smaller pangenome size. This observation was attributed to the more broad and diverge niches for *S. agalactiae* to survive.

1.7 Conclusion and perspectives

Pangenome sequencing is allowing an unprecedented opportunity to understand and define microbial species, but also will challenge the analysis and interpretation of large amount of genome sequence data (Read and Ussery 2006). Attempts to tackle the problem of pangenome size estimation have been made in several studies with different approaches. These are regression analysis of fitting sample curves, binomial mixture models to estimate gene family frequency and gene birth and death model of the evolution of pangenome. In

spite of the different underlying modeling assumptions, similar trends about pangenome structures for various microbial species have been observed. And a great diversity of dispensable genes exists to boost the genetic diversity of species living in an environment with access to them. Therefore the aim of fully pangenome sequencing is likely unreachable. When extending a pangenome analysis to a higher taxonomic level, it was shown that there is a small set of very conserved core genes ('extended core'), and most of the core genes at species level move into a 'character genes' set, which are distinct and indicator of species characters. As for the dispensable genes, there is a similar situation, no matter what taxonomic scale considered. Evidently, the vast volume of genes belongs to the mobilome, associated with phage and other mobile genetic elements, and mostly short and with function unknown, a so-called 'dark matter' of the microbial world (Wu, Hugenholtz et al. 2009).

The pangenome concept accords with the distributed genome hypothesis (DGH), which states that a population of strains of a bacteria pathogen species can utilize genic characters distributed in the whole species pangenome (Ehrlich, Ahmed et al. 2010). These clonal or polyclonal strains thrive in a biofilm community of close contact for easy DNA exchange via conjugation or transformation, which may help for disease progression and bacteria persistence. Through gene reassortment, the pangenome can generate novel strains with unique phenotypes conferring the species strength against the host's immune system. Therefore, it is important to characterize the different combinations of genes from the pangenome to understand their resulting phenotypes and their coevolution with host. In this regard, it is much more relevant to analyze the evolution of the pangenome, such as ancestral genome reconstruction of the pangenome. From an ancestral reconstruction analysis, gene gains and losses along different lineages or different serotypes/ecotypes can be inferred, and thus provides information for species divergence and adaptation.

Chapter 2

Genomics and evolution of *Mycobacterium tuberculosis* and other related Mycobacteria

2.1 Introduction to the *Mycobacterium tuberculosis* complex and the genus *Mycobacterium*

The *Mycobacterium tuberculosis* complex (MTBC) is a *Mycobacterium* species of highly genetic homogeneity (Sreevatsan, Pan et al. 1997). This species comprises subspecies of causative agents of human and animal tuberculosis (TB) disease, which are *M. tuberculosis* (human tuberculosis), *M. bovis* (found in cattle), *M. microti* (found in voles, wood mice and shrews), *M. africanum* (human tuberculosis commonly found in West African countries), *M. pinnipedii* (found in marine mammals), *M. caprae* (associated with goats), *M. mungi* (found in mongooses), *M. orygis* (found in oryx) and *M. canettii* (found in human) (Brosch, Gordon et al. 2002; Alexander, Laver et al. 2010; van Ingen, Rahim et al. 2012). Additionally, this species also includes the Bacillus Calmette–Guérin (BCG) vaccine strains derived from the *M. bovis*. Despite the high similarity (>99.9%) of genome sequence, different host ranges and phenotypic characteristics are displayed in this species (Smith, Kremer et al. 2006). Although the paucity of single-nucleotide polymorphisms (SNPs), which was supposed to be resulted from a recent clonal expansion from the progenitor '*M. prototuberculosis*', a large portion of genomic diversity is associated with repeat sequences, insertion sequences and large sequence polymorphisms (RDs) (Sreevatsan, Pan et al. 1997; Brosch, Gordon et al. 2002). Together, these divergent genomic regions provide useful genetic markers for epidemiology studies, such as spacer oligonucleotide typing (spoligotyping), IS6110 RFLP fingerprinting, variable number of tandem repeats (VNTR) typing and mycobacterial

interspersed repetitive units (MIRU) typing (Djelouadji, Raoult et al. 2011). Furthermore, in combination with RD sequences, these markers also help to identify MTBC lineages associated with different hosts and human populations (Figure 2.1) (Brosch, Gordon et al. 2002; Wirth, Hildebrand et al. 2008). And it has been suggested that this evolutionary background of the MTBC lineages has implication in their transmissibility and emergence of drug resistance (Parwati, van Crevel et al. 2010). Considering the relatively recent divergence and adaptation to different hosts of the MTBC subspecies and lineages, there are emerging efforts to elucidating the underlying evolutionary processes, epically with an access to the next generation high throughput sequencing technology (Ford, Yusim et al. 2012).

The *Mycobacterium tuberculosis* complex belongs to the genus *Mycobacterium*, which also includes other human pathogens, such *M. leprae* and *M. ulcerans* (Brosch, Pym et al. 2001). In fact, the genus *Mycobacterium* is a highly diverse phylogenetic group characterized by a cell envelop rich in unusual lipids and glycolipids, comprising more than 100 identified species, spanning a broad spectrum of ecotypes, from free-living saprophytes and species adapted to soil and aquatic habitats, to obligate human and animal pathogens (Tortoli 2006). These species can be divided into slow-growing and fast-growing species according to their generation time (Figure 2.2) (Brosch, Pym et al. 2001). The slow-growing group includes some famous human pathogens, the *M. tuberculosis*, *M. ulcerans* and *M. leprae*. The nature of a long generation time (1-14 days) is an important aspect for MTB dormancy in host cells, which can be reactivated to cause active tuberculosis. In contrast to MTB, environmental mycobacteria or non-tuberculosis mycobacteria (NTM) were shown to engage in active horizontal gene transfer, and therefore with more diverge genomic architecture (Veyrier, Dufort et al. 2011).

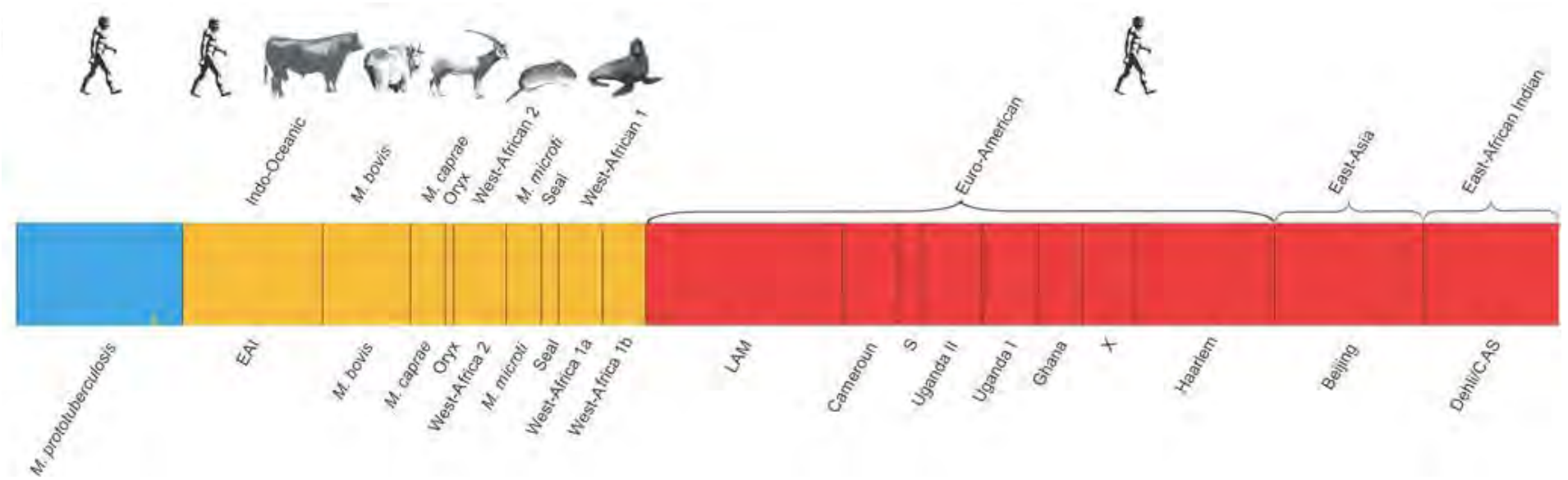

Figure 2.1: Major MTBC lineages and their associated host ranges and human populations. Colors are in accordance to different lineages or clusters, and length of the colored segments indicates their population proportions in that cluster. (Adapted from Wirth *et al.*, 2008)

Actually, the MTB species was assumed to evolve from a soil habitant and adapt to host environment by genome reduction (Gutierrez, Brisse et al. 2005; Veyrier, Dufort et al. 2011). Therefore, genomic studies of NTM can help to understand how tuberculosis mycobacteria arise from NTM and get adapted as obligate mammalian pathogens, in addition to their own biology questions. Especially, comparative genomic analysis between members of the two groups of slow-growing and fast-growing has the potential to reveal the corresponding genes and pathways, and contribute to an improved understanding of the pathogenesis of MTB (Brosch, Pym et al. 2001; Veyrier, Dufort et al. 2011).

2.2 Comparative genomics of *M. tuberculosis*

Determination of the first mycobacterium genome of *M. tuberculosis* H37Rv, has helped to elucidate various genetic aspects of the biology of MTB, and also served as an important reference genome sequence for comparative genomics studies of the species (Cole, Brosch et al. 1998; Camus, Pryor et al. 2002). It was shown that the high G+C (65.6%) genome of the *M. tuberculosis* H37Rv exhibits several features corresponding to the slow-growth nature of this species, including an even distribution in gene polarity and an *rrn* operon relatively far from the *oriC*. A large array of ~250 genes involved in lipogenesis and lipolysis were predicted, which are responsible for the synthesis of a diverse array of lipophilic molecules present in this organism, and also the fatty acid oxidation systems. The most striking feature of the coding genes was the characterization of two multi-gene families, PE and PPE genes, which accounts for ~10% of the coding capacity. These genes encode glycine-rich proteins with a conversed N-terminal plus a variable C-terminal.

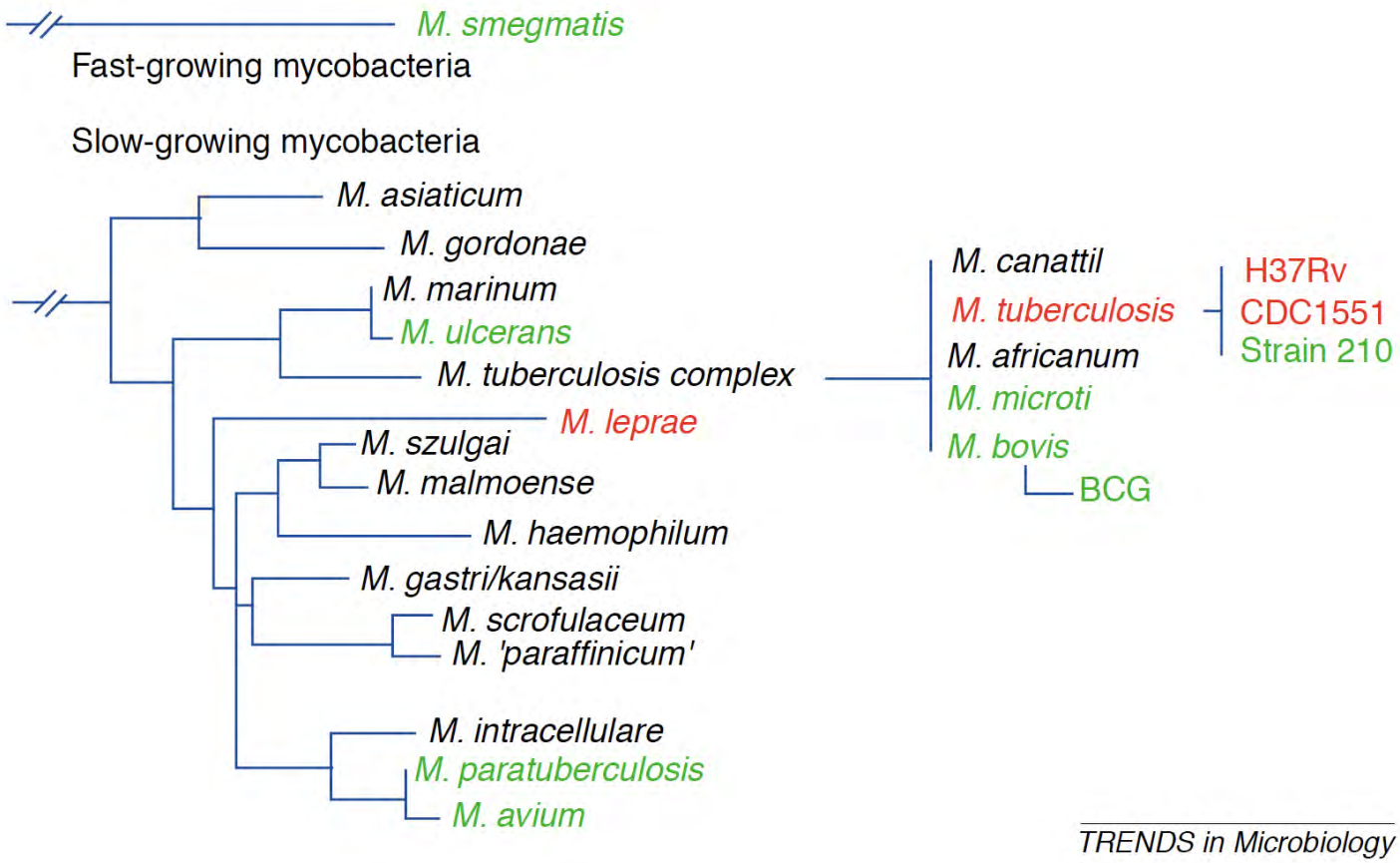

Figure 2.2: Phylogeny of selected fast-growing and slow-growing mycobacteria based on 16S rRNA sequence (Adapted from Brosch *et al.*, 2001).

The PE family is characterized by conserved N-terminals of Pro-Glu (PE) motifs, and N-terminals of Pro-Pro-Glu (PPE) motifs are characteristics of the PPE proteins. Repetitive motifs were found in the C-terminals of most PE and PPE proteins, leading to the speculation that these genes might be potentially polymorphic antigens. Furthermore, sixteen IS6110 elements, six IS1081 elements and another set of 32 newly discovered insertion sequence elements were identified, in addition to two prophages, phiRv1 and phiRv2. Although some of the IS elements were assumed to be stable, others are capable of transposition and influence gene integrity, and also resulting in gene deletions via recombination.

Although there are rare single nucleotide polymorphisms in the MTB complex, comparative genomics analyses have succeeded to uncover another source of genetic diversity. A number of indels events associated with repetitive sequences, PE and PPE genes and IS6110 elements have been identified from comparing the genomes of *M. tuberculosis* H37Rv, BCG Pasteur, *M. bovis* AF2122/97 and *M. tuberculosis* CDC1551 (Behr, Wilson et al. 1999; Gordon, Brosch et al. 1999; Fleischmann, Alland et al. 2002). These indels, or region of differences (RDs), affect genes of various functions, and vary among different strains or different members of the species (Figure 2.3, 2.4). Close examination of the distribution of these RDs among different MTB strains or subspecies provides clues to their evolution and their implications in MTB pathogenesis. Some RDs appear to be lineage or subspecies specific and thus were supposed to occur at their lowest common ancestors (Figure 2.5), while others vary among different strains of the same lineage and therefore are more recent evolutionary events (Brosch, Gordon et al. 2002). For example, the RvD2, which is absent from H37Rv, is still present in its avirulent relative H37Ra. Another conclusion that can be drawn from the RDs is that there is a reductive evolution of the MTBC members from their progenitor, since the RDs represent series of losses of genetic materials during the divergence of the species (Figure 2.5) (Brosch, Pym et

al. 2001). But there are also two cases of large sequence duplications that occurred within the BCG Pasteur genome, the DU1 and DU2, which were suggested to relate to its laboratory attenuation as a vaccine strain.

2.3 Evolutionary scenario of the MTB complex

Inspired by the results of RD distribution among the MTB described above, which were drawn primarily from comparing *M. tuberculosis* H37Rv, *M. bovis* and BCG strains, further studies have been taken to interrogate these RD more thoroughly using more MTB representative strains (Brosch, Gordon et al. 2002; Tsolaki, Hirsh et al. 2004; Alland, Lacher et al. 2007). In the study of Brosch *et al.*, a total of 20 RDs were examined among 100 MTB strains and the result led to the establishment of an evolutionary scenario of the MTB complex (Figure 2.5) (Brosch, Gordon et al. 2002). This was made possible by the observation that these RDs did not occur independently during the evolution of the different lineages within the complex, but correlated with the lineages in different evolutionary time scale. Some RDs appear to be ancient events and are useful markers for delineating earlier branching of MTBC subgroups, such as the RD9, which is deleted in the divergence of the animal MTBC groups and the *M. africanum* (Figure 2.5). In contrast, RD2 and RD14 occurred more recently and help to separate different BCG strains of Tokyo and Pasteur. Based on these RD markers and the availability of the initial framework of lineage evolution, new members of the complex can be evaluated and added to give a more comprehensive scenario. Examples are *M. mungi*, Dassie bacillus, and *M. orygis*, which are newly characterized members of the MTB complex and were added to the evolutionary framework by sharing some of the known RDs (Alexander, Laver et al. 2010; van Ingen, Rahim et al. 2012).

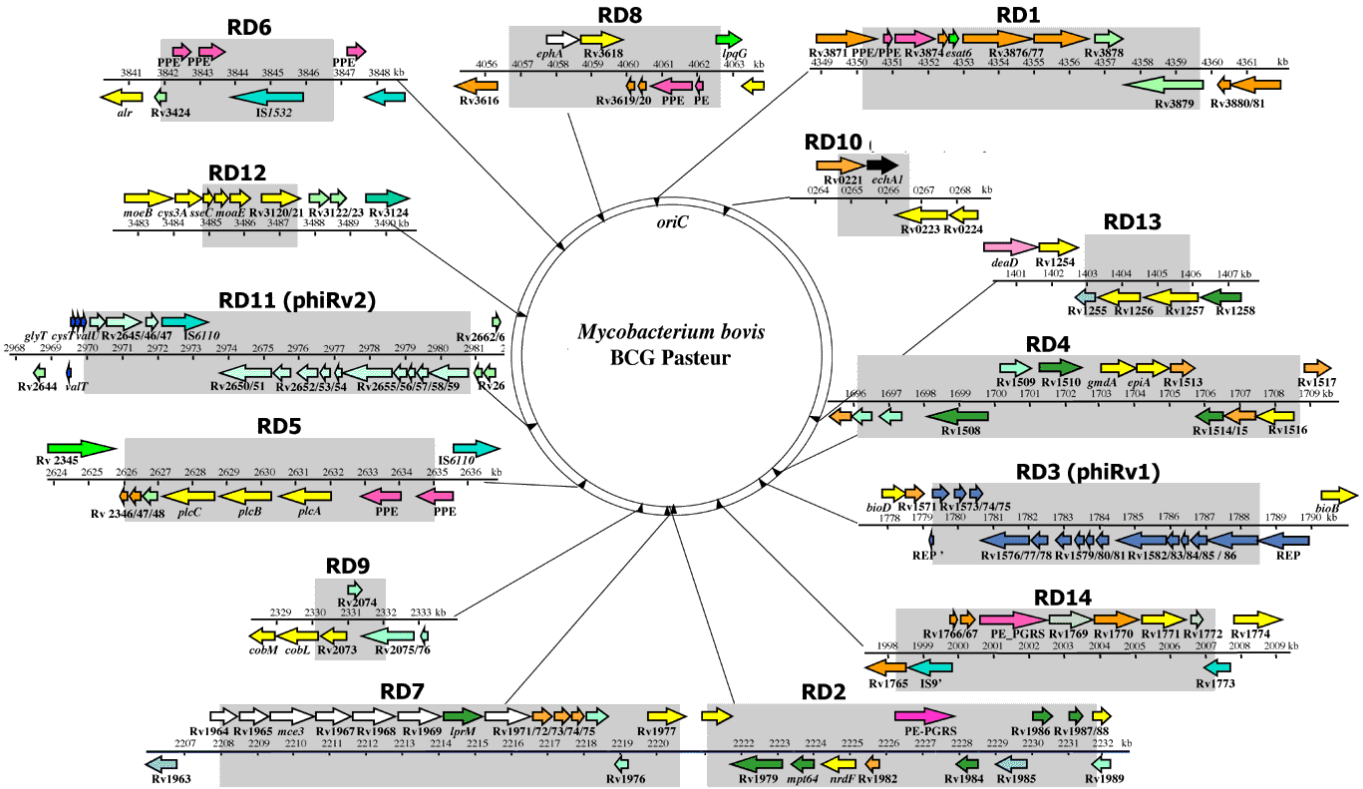

Figure 2.3: RDs absent from the genome of *M. bovis* BCG Pasteur comparing with the *M. tuberculosis* H37Rv. Two comparative genomics approaches, BAC-arrays (Brosch *et al.*, 1998, 1999, 2000, and Gordon *et al.*, 1999) and DNA microarrays (Behr *et al.*, 1999), have been used to identify these RDs. (Adapted from http://www.pasteur.fr/recherche/unites/Lgmb/Deletion.html)

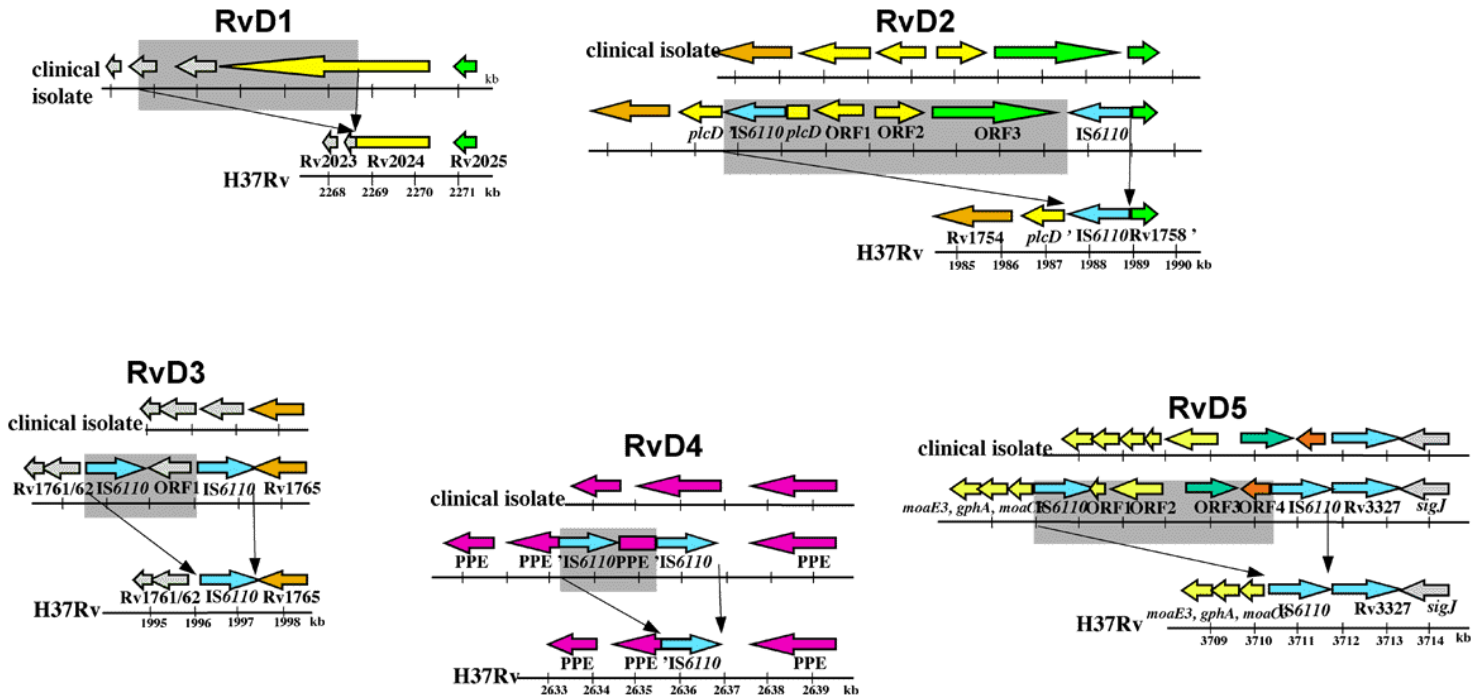

Figure 2.4: Deleted regions in the genome of *M. tuberculosis* H37Rv when comparing with other MTB strains (Brosch *et al.*, 1999). (Adapted from http://www.pasteur.fr/recherche/unites/Lgmb/Deletion.html)

This evolutionary scenario has helped to clarify a misconception about MTB evolution (Brosch, Gordon et al. 2002). It was speculated that human MTB was derived from *M. bovis* after the domestication of cows given that *M. bovis* has a broader host range (Smith, Hewinson et al. 2009). This contradicts with the situation derived from RD patterns, which indicates that, in contrary, *M. bovis* arose from the common ancestor of the complex that may more resemble the human MTB. More insights into the origin and divergence of the complex arise readily from the scenario. For instance, the deletion of the TbD1 defines a 'modern' human MTB lineage, regarding that other human MTB strains still have the RD present and are assumed to be 'ancestral'. There is also a demography aspect of the modern human MTB lineage, which is associated with most of the current worldwide epidemic of TB (Wirth, Hildebrand et al. 2008). In this regard, the *M. africanum* WA-1 and WA-2 are also ancestral, further explained by their endemic in the West Africa. From the scenario, it is evident that losses of RDs are main theme in MTB evolution, which is assumed to happen after an evolutionary bottleneck. In this point, the ancestral *M. tuberculosis* is the most similar extant subspecies to the common ancestor, or the '*M. prototuberculosis*'. And it is intriguing that the loss of RDs seems to enable a broader host range. But to what extent the *M. tuberculosis* ancestral genome resemble to the progenitor needs more information from comparative genomics studies. Most of the RDs were identified based on the reference genome of *M. tuberculosis* H37Rv, which fails to include RDs absent in the H37Rv genome but present in other MTBC subspecies and therefore this reference set of RDs are incomplete. Inclusion of the *M. canettii* as an out-group to understand how the progenitor gave rise to current MTBC lineage is also necessary in conducting comparative genomics analysis.

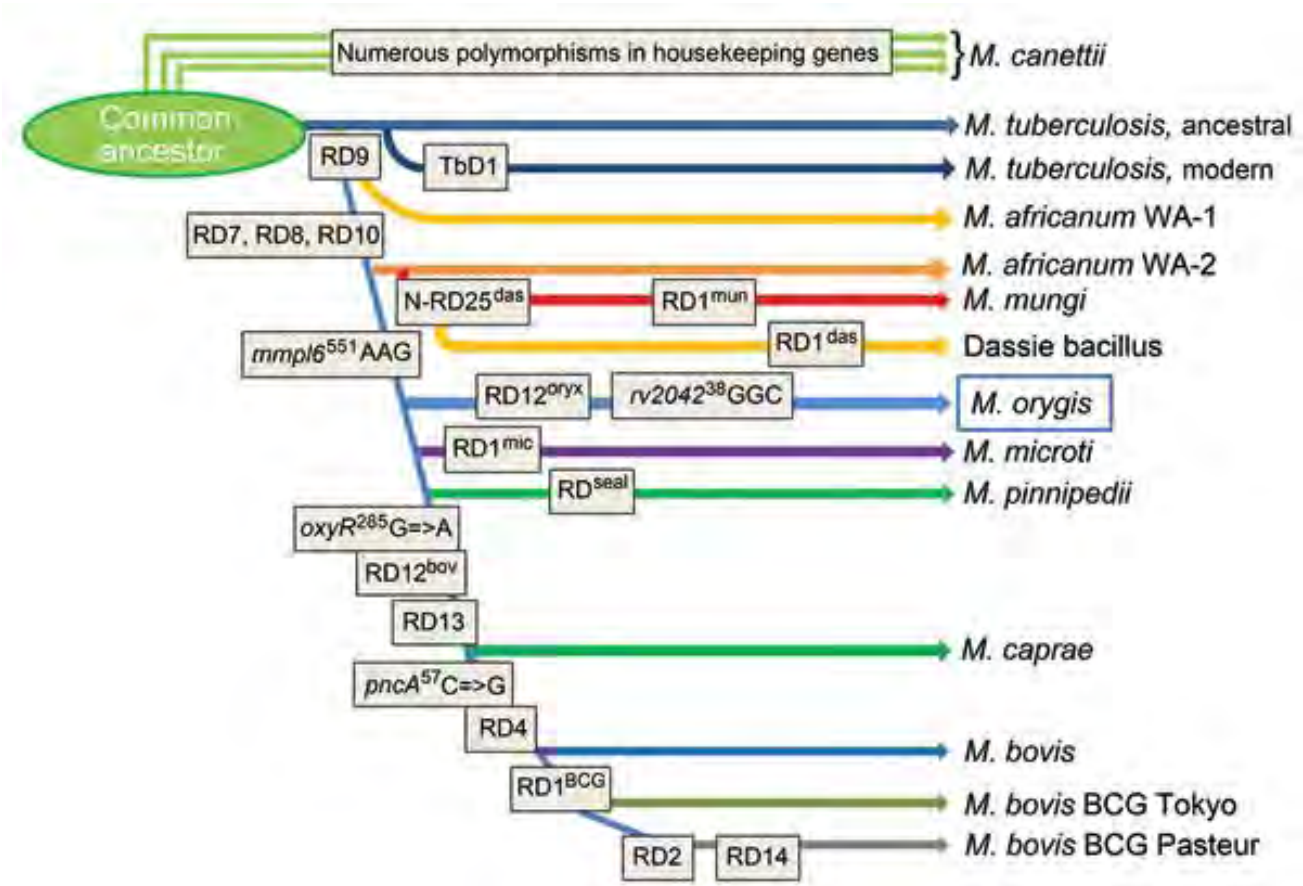

Figure 2.5: Evolutionary scenario of the MTB complex. The combination of RDs enables the delineation of different MTBC lineages and their order of divergence. Serial losses of RDs occurred from their descendant from the common ancestor. (Adapted from Brosch *et al.*, 2002; Ingen *et al.*, 2012)

2.4 MTB lineages and epidemiology studies

Much more efforts have been directed towards investigation of the population structures of the human *M. tuberculosis*, given the disease burden imposed on the world population by them (Filliol, Motiwala et al. 2006; Hershberg, Lipatov et al. 2008). In spite of the very conserved genomes among the MTB strains, genetic markers from repetitive genomic regions have been identified and applied successfully to epidemiology studies. Well-established markers include IS6110-RFLPs based on presence of different copies and genomic locations of the IS6110 elements among the MTB strains, spoligotyping based on a genomic locus of direct repeats, VNTR typing based on variable numbers of tandem repeats throughout the genome that are minisatellite-like loci, and MIRU (mycobacterial interspersed repetitive units) that is a extended set of VNTRs (currently comprise 24 loci, http://www.miru-vntrplus.org/) (Supply, Allix et al. 2006; Weniger, Krawczyk et al. 2010; Comas and Gagneux 2011). These sets of markers have traditionally provided convenient genotyping ways to catalogue and compare different strains of human epidemics, and succeeded to trace different MTB genotypes from different world regions, such as the Beijing genotypes, which are of high prevalence. From the results of epidemiological studies, it is clear that a certain degree of genetic diversity is carried in this otherwise thought-to-be extremely conversed species, and this diversity is associated with different regions of human population, and also with varying clinical phenotypes (Hershberg, Lipatov et al. 2008; Brown, Nikolayevskyy et al. 2010).

A more recent study has stepped further in illustrating the genetic diversity within the human MTB and the evolutionary force shaping them. By using a large set of DNA sequence data from 89 conserved genes spanning the whole genome, Hershberg *et al.* gave a more precise and comprehensive picture of the population structure of a global collection of

human MTB strains (Figure 2.6) (Hershberg, Lipatov et al. 2008; Brown, Nikolayevskyy et al. 2010). This approach was assumed to complement earlier epidemiology studies with more phylogenetic informative SNPs that are distributed throughout the whole genome (89 genes, 65,829bp in concatenation, ~1.5% of the ~4.4 Mbp genome of MTBC) and more suitable for phylogenetic inference and population genetics analysis. In consistence with the evolutionary scenario from RD analysis (Figure 2.5), the modern human MTB is further separated into lineages of 'India and East Africa', 'East Asia' and 'Europe and Americas'. And the ancestral group consists of 'The Philippines', 'Rim of Indian Ocean', 'Animal strains' and the two 'West Africa' lineages. Apparently, distribution of these lineages is in accordant with human migration patterns in history (Figure 2.7). Ancestral MTB group appear to have a same 'out of Africa' route as human. Further population genetic analysis of these data showed that genetic drift is the major force that shapes the diversity of these lineages. Examination of the identified 488 SNPs (from a total of 65,829 sites of the surveyed sequence, 129–145 different sites between *M. canettii* and other MTBC, a maximum of 46 sites between other MTBC) by calculating their dN/dS ratio (0.57, higher than other observed cases of strong purifying selections) supported an action of reduced purifying selection. As a possible explanation, this reduction in selection was attributed to the their functional consequences of these SNPs, which were shown to present more in conversed sites in the encoded proteins. Given the results of highly subdivided population structure and clonality and the observed bottleneck effects, it was assumed that a small effective population size of this species and therefore a random genetic drift responsible for shaping its population diversity. And this explains the observed relative relaxed purifying selection, which may happen in the clonal expansion of this species after an evolutionary bottleneck and initial adaptation to different hosts. To name a particular example, the animal lineage comprises a variety of MTBC subspecies, *M. bovi*, *M. microti*, *M. pinnipedii*, and *M.*

caprae, which are different ecotypes affecting different animal hosts, but show only comparable diversity relative to the human MTB (Figure 2.6).

Another important factor that could have shaped the population structure of the MTB complex is human migration and demographic changes. As evident from the distribution of geographic lineages of MTBC (Figure 2.6), origin and spread of MTBC appears to correlate with the history of human migration and demography, the so-called 'out of Africa' hypothesis (Figure 2.7) (Wirth, Hildebrand et al. 2008). This is plausible from the fact that two deep-branching lineages, West Africa clade I and II, were found to be more responsible for human MTB in West Africa than other areas in the world. Further, the ancient MTBC group, including another two deep-branching lineages of 'The Philippines' and 'Rim of Indian Ocean', also coincides with a land route of the early spread of human population out of Africa. Additionally, this evolutionary scenario gives hints to the origin and evolution of the 'animal strains', which might also have an 'Africa' origin. And it is tempting to speculate that the progenitor of the 'animal strains' could have arisen from domesticated animals, such as cows, and then spread into other animal species. But more recent migration and global travel of human underpinned the prevalence of the 'modern group' of MTBC, which was shown to be more virulent and more frequently reported from clinical cases worldwide, compared with its 'ancestral' counterpart (Glynn, Whiteley et al. 2002; Hanekom, Gey van Pittius et al. 2011). Such case also happened in the Africa and led to the 'out of and back to Africa' hypothesis (Hershberg, Lipatov et al. 2008). The situation would be more severe, made by the becoming frequent global travel by air route and the emergence of drug resistance strains. A pronounced example is the MTBC Beijing genotype, which was first described for isolates from the Beijing area of China and was found to be dominant in the East Asia (van Soolingen, Qian et al. 1995). Since its first description, strains of this genotype have been continually reported worldwide and were becoming prevalence in some areas. The success of this genotype was attributed to its increased

virulence and associated with drug resistance. Such case of expansion of hyper-virulent MTBC genotypes, facilitated by global transmission and drug resistance, would reshape the global population structure of the MTBC species to make it more homogeneous and has important implication for TB control (Parwati, van Crevel et al. 2010; Hanekom, Gey van Pittius et al. 2011).

2.5 Sequencing of other closely related mycobacteria

A consensus of MTBC evolution can be reached: the establishment of a progenitor through an evolutionary bottleneck followed clonal expansion accompanied by reductive evolution (Brosch, Pym et al. 2001). But how the progenitor has arisen from its non-tuberculosis mycobacterial (NTM) ancestor remains unknown. Sequencing of other closely related mycobacteria provides a potential to unravel the evolutionary process that leads to the MTBC and understand the adaptation and pathogenesis of this pathogen (Veyrier, Dufort et al. 2011). With this regard, new insights into the speciation of MTBC have been obtained from the more recent sequencing of four NTM genomes of *M. avium*, *M. marinum*, *M. kansasii* and *M. canettii* (Stinear, Seemann et al. 2008; Bannantine, Wu et al. 2012; Bentley, Comas et al. 2012). From the perspective of these genomes, horizontal gene transfer that is absent within the conservative MTBC, was recruited to explain the emergence of the MTBC from its ancestral non-tuberculosis *Mycobacterium* (Becq, Gutierrez et al. 2007; Jang, Becq et al. 2008). To synthesize this result with the evolutionary scenario established within the MTBC, a biphasic evolutionary process leading to the formation of MTBC was proposed (Figure 2.8) (Veyrier, Dufort et al. 2011).

With this biphasic paradigm, evolution of the MTBC species has undergone a ‘rise’ (gene acquisition and duplication) followed by a ‘fall’ (gene loss and deletion). Particularly, a step-wise evolutionary process can be conceived with regard to the closely related

mycobacterial species with different phenotypes on human infections (Veyrier, Pletzer et al. 2009). Different from MTBC, *M. kansasii* is only opportunistic human pathogen and cannot establish persistent infection in human population. The same is for *M. canettii*, which was described as atypical MTB and causes only sporadic human disease. They are both considered unsuccessful for their inability to cause long-term human infection and epidemic. Therefore, they may resemble the ancestral colonization of human host in the initial establishment of MTBC and carry information to reveal ancestral genomic events responsible for that. As indicated in Figure 2.8(b), the MTBC might have been becoming more adapted to the human host as an obligate pathogen after its divergent from the common ancestor of *M. canettii* and *M. tuberculosis*, involving both gene acquisition and deletion. After that, the MTBC was well established in the human population and fine-tuned by additional reductive evolution. These insights and proposed paradigms would be very promising with more closely related non-tuberculosis mycobacteria characterized and their detailed phenotype information and genome sequences available, which is the current trend in the field of mycobacterial research (Tortoli 2006).

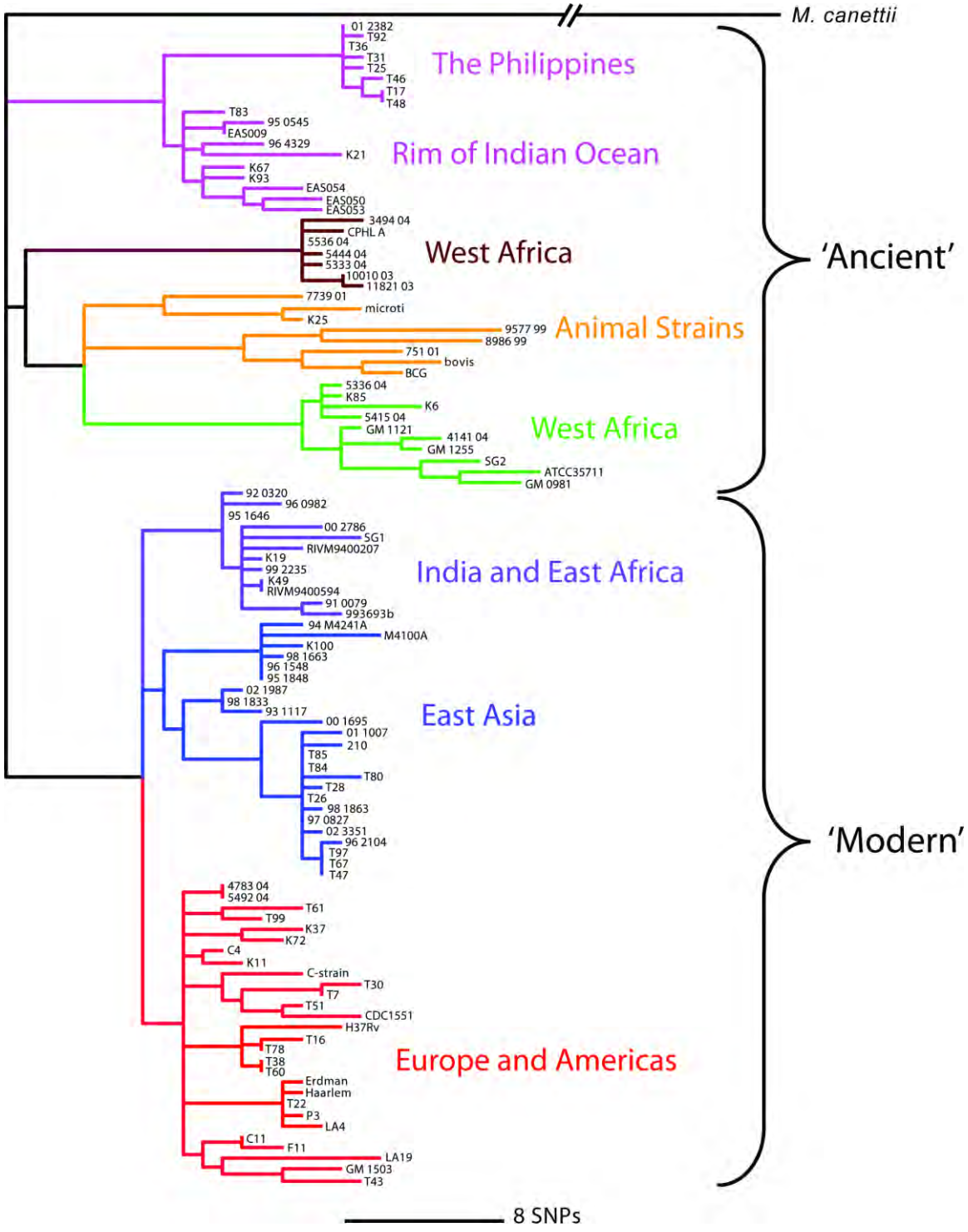

Figure 2.6: Maximum parsimony phylogeny of *M. tuberculosis* Complex using 89 concatenated gene sequences in 108 strains. Colors of the branches indicate different MTB lineages from different world areas. (Adapted from Hershberg *et al.*, 2008)

Figure 2.7: Origin and spread of the *M. tuberculosis* complex inferred from epidemiological data of 24 MIRU loci on a collection of 355 isolates. The MTBC is supposed to origin from a *M. prototuberculosis* species in Northeast Africa and then spread to the other areas in the world, following the different routes of human migrations. (Adapted from Wirth *et al.*, 2008)

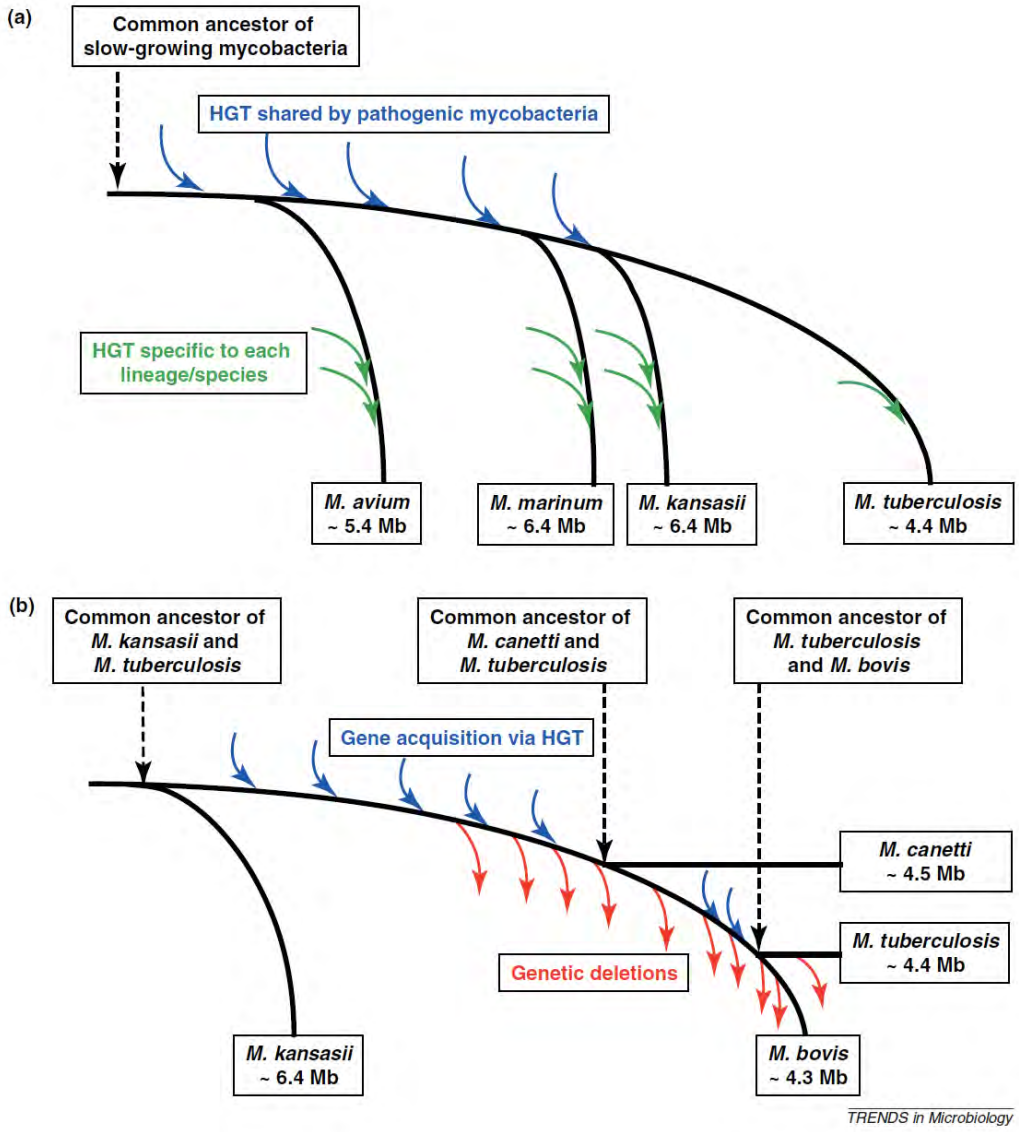

Figure 2.8: The rise and fall of *M. tuberculosis*. (a) Horizontal gene transfer (HGT) events were predominant in the speciation of different non-tuberculosis mycobacteria, and before the formation the *M. tuberculosis* species. These events introduced both core genes (blue arrows) and specific genes (green arrows) during species evolution and adaptation. (b) Step-wide evolution of the genesis of the *M. tuberculosis* can be investigated with regard to two closely related mycobacterial species, *M. kansasii* and *M. canettii*. This paradigm may help to understand the ancestral opportunistic colonization of human host by MTBC, involving both HGTs (blue arrow) and deletions (red arrow). And then the species has become obligate pathogen, tuned by genome reduction. (Adapted from Veyrier *et al.*, 2011)

2.6 Conclusion and perspectives

Since the complete sequencing of the genome of *M. tuberculosis* H37Rv, a vast amount of genetic heterogeneity has been uncovered for the species of MTBC, which was previously assumed to be very homogeneous and clonal in nature. The observed genomic heterogeneity has provided valuable genetic markers, such as RDs, IS6110 elements, the TR locus and MIRU loci, for evolutionary and epidemiology studies of the global population structure of the MTBC. Different MTBC lineages have been identified using a combination of these markers, and have been shown to be associated with different human population in different regions. Furthermore, these lineages exhibit different patterns of divergence from the MTBC progenitor and therefore provide clues to understanding the origin and spread of this host-associated pathogen and their adaptation to host. In addition to this phylogeographical pattern of the macroevolution of MTBC, strain variation has also been examined in clinical settings, especially the emerging of drug resistant strains or hypervirulent strains, which are responsible for the current transmission and epidemic of this devastating pathogen. All these results of the genomic and evolution of the MTBC have implied that genomic heterogeneity of this species is strongly associated with its phenotypic outcomes. Therefore comparative and evolutionary genomics represents a principle approach to understanding how the genetic diversity translates into relevant phenotypic variations and pathogenicity of different strains.

Given the availability of the next generation high-throughput sequencing technologies, there is a foreseeable future of accelerating efforts in large-scale sequencing of many MTBC strains at a population level. As proposed for studying the hypervirulent MTBC Beijing family, three routes are suggested to approach from different aspects, which complement to each other (Parwati, van Crevel et al. 2010). The first is to genotype epidemic strains by extensive molecular epidemiology study to bring in temporal and

special dimensions in more depth, with regard to their changes in population structure. The second is to phenotype representative epidemic strains to characterize both their interaction with human host in clinical settings and their behavior in animal models. A third route is a synergy of the genotype and phenotype information obtained from population study of epidemic strains using whole genome sequencing to reveal the evolutionary and population changes in the genomes of these epidemic strains. In this new paradigm of 'population-centric' MTBC study, analysis of microevolution and genomic dynamics will play a critical role, involving new phylogenetic methods, new genomic evolution models and large-scale genome data processing to bring the full power of whole genome sequencing at a population level.

For this purpose, related genomic and bioinformatic issues will be addressed in this thesis from the perspective of pangenome analysis. As demonstrated in earlier comparative genomics analysis, there is no obvious evidence for horizontal gene transfer in the MTBC species; instead, clonal expansion followed by genome reduction represents a main theme for its evolution. This trend of genome evolution implies that there were relatively few genomic inversion and conversion events and most of the genomic changes were associated with subspecies, lineages or strain genotypes at different evolutionary time scale. In this context, ancestral reconstruction of the evolutionary process leading to current population structure using pangenome data is feasible and promising for understanding host adaptation of this pathogen. Particularly, the term of 'palaeogenomics' was coined for this emerging theme of studying ancestral events in the evolution of MTBC, which also emphasize investigation of palaeomicrobiological samples and MTBC ancient DNAs (Djelouadji, Raoult et al. 2011). In a pangenome approach, genetic repertoire of the MTBC progenitor is assumed to distribute among the whole population, if not totally lost. This assumption seems plausible for the MTBC, because of its early clonal expansion after an evolutionary bottleneck and adaption and survival of the different lineages after divergence. Pangenome

analysis of the distributed genetic repertoire of the MTBC progenitor is also feasible and necessary given the ongoing genome sequencing of representative members of MTBC (Figure 2.9). Ancestral genome reconstruction with a pangenome comprising these strains (Figure 2.9), and more lineage representative strains in the future, is likely to reveal the key steps in their evolution and adaptation into different lineages and host-associated phenotypes.

Another component that has to take into account in investigating the evolution of MTBC is the genomes of non-tuberculosis mycobacteria. It has been hypothesized that the MTBC that causes human and animal tuberculosis arose from an environmental mycobacterial ancestor. And initial establishment of the ancestor as a human associated pathogen has involved horizontal gene transfers and genome duplications, which were suggested to be relevant to their emergence of pathogenicity. Therefore, studying MTBC from the perspective of NTM provides opportunities to unravel functional clusters of genes underlying the emergence of MTBC pathogenicity. Pangenome model of the *Mycobacterium* genus, including various species of slow-growing or fast-growing, pathogenic or environmental, will facilitate this process for identifying clusters of genes that are responsible for the formation of different species in the genus and their related phenotypes. Currently, this viewpoint is beginning to gain its momentum, as demonstrated in the increased number of newly sequenced NTM genomes in the database (Figure 2.10).

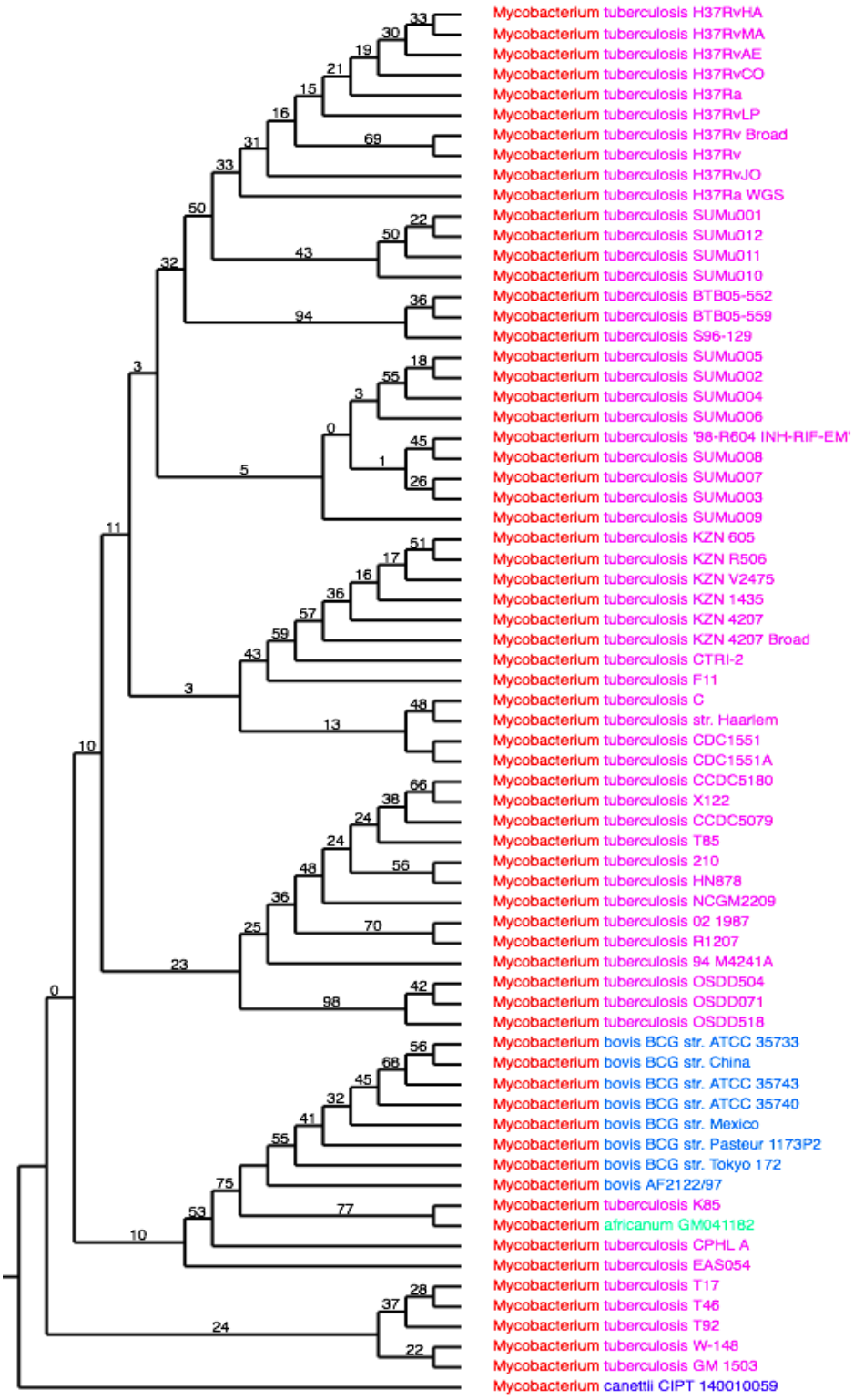

Figure 2.9: Currently available genome sequences of different MTBC strains, including both complete and draft. Data and their phylogenetic positions were obtained from the PATRIC database (http://www.patricbrc.org/portal/portal/patric/Phylogeny?cType=taxon&cId=1763) at the time of writing this thesis.

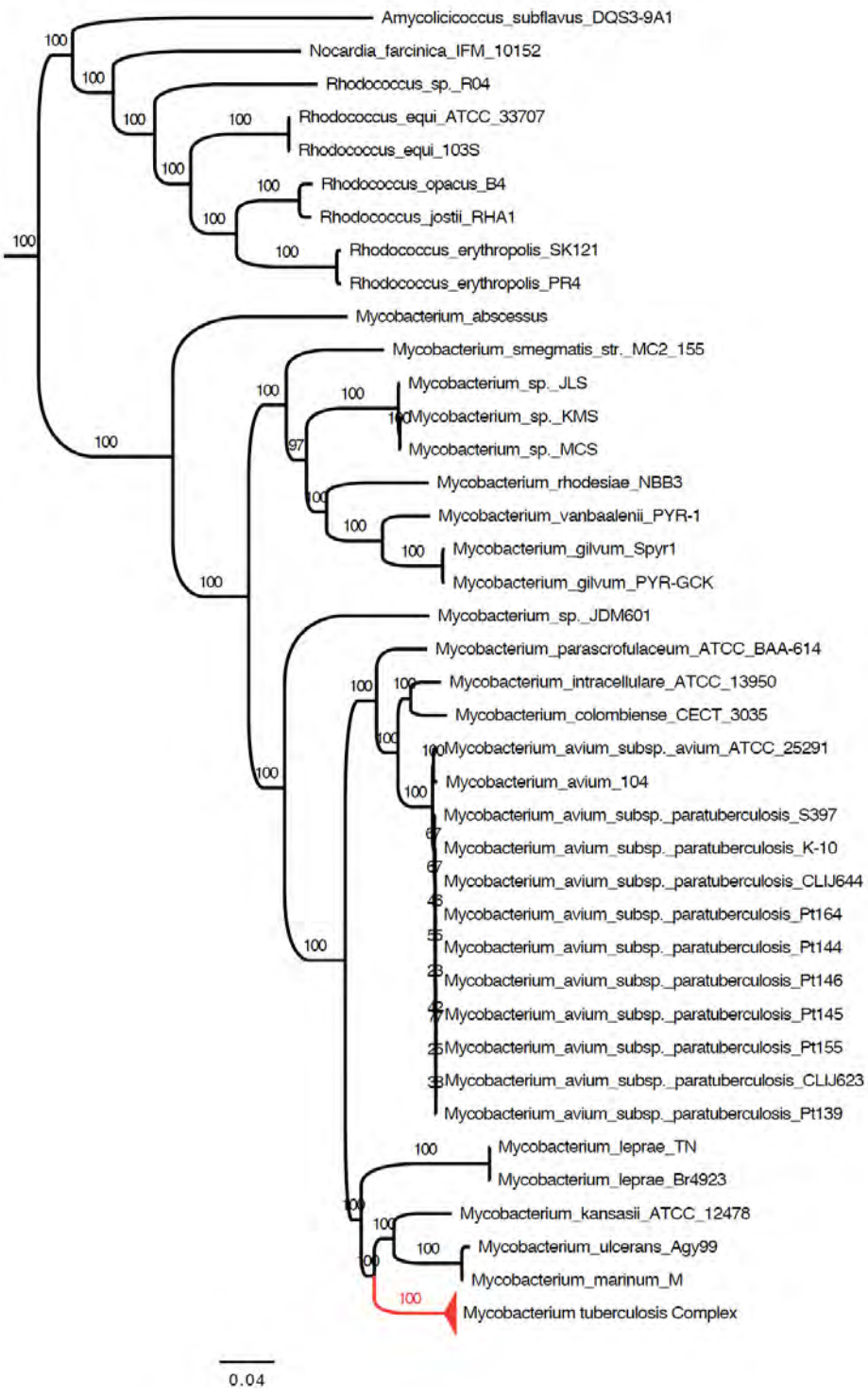

Figure 2.10: Currently available genome sequences of different non-tuberculosis mycobacterial species and strains, including both complete and draft. The branch leading to the MTBC is grouped and indicated by a red triangle to provide information on its relation to other NTM species. Data and their phylogenetic positions were obtained from the PATRIC database (http://www.patricbrc.org/portal/portal/patric/Phylogeny?cType=taxon&cId=1763) at the time of writing this thesis.

The pangenome data is continuing to accumulate at both species and genus level, but evolutionary models and bioinformatic methods suitable for analyzing such kind of data lag

behind. There are three aspects for addressing the challenge of data analysis with regard to the current consensus of the evolutionary scenario of the MTBC species and the whole genus, as reviewed above. Ancestral reconstruction and phenotype mappings in a phylogenetic framework are valuable to understand the microevolution and adaptation of the different lineages or emerging strains of MTBC. On the other hand, given the important role of HGT in the evolution of different mycobacteria species, pangenome models that are analyzing gene coevolution and adaptation of gene clusters are more suitable at the genus level. As will be explored in this thesis, graph-based clustering models that are probabilistically modeling the evolutionary process at a pangenome scale are the promising ones, because these models have been successfully applied to cluster gene families or ortholog groups of genes from large-scale whole genome data. At last, bioinformatic solutions to large-scale pangenome data processing are required to form a basis for pangenome modeling, including multiple genome alignment, genome mapping for indel and SNP calling, pangenome-based functional annotation and integration of strain genotype and phenotype information from epidemiology studies, which together make up a pangenome platform for genomics and evolution studies of the mycobacteria and MTBC.

Chapter 3

Research hypothesis and study aims and objectives

3.1 Research hypothesis I:

Ancestral genome reconstruction of the *Mycobacterium tuberculosis* species pangenome would help to identify important genetic events during its evolution, especially for the study of the MTB Beijing genotype virulence evolution.

In this thesis, motivated by the pangenome analysis approaches summarized in Chapter 1 and the knowledge of MTB genome evolution as summarized in Chapter 2, **study aims and objectives are mainly on developing an algorithm for ancestral genome reconstruction for gene gain and loss analysis, with application to the species of *Mycobacterium tuberculosis* (Chapter 4 and Chapter 5)**. The *M. tuberculosis* species or *Mycobacterium tuberculosis* complex (MTBC) is important human pathogen, and causes infections globally, and thus under extensive studies with global efforts. Therefore, the MTB provides an ideal subject for pangenome modeling, with a plethora of genome data and also lineage/sub-lineage characterizations with different virulence or associated demography information. There are also well-characterized genes or genomic regions, such as the IS6110 genes and large genomic indels, which can be served as benchmark dataset to evaluate the pangenome models and verify the prediction results. For the purpose of studying indels evolution, a local parsimony ancestral state reconstruction algorithm was developed to consider the specific evolutionary scenario involved with genome sequence

insertions and deletions (indels). After that, by mapping the ancestral indels events back to the genomes that have been sequenced, their affected genes and related functions were thoroughly examined, and also their biological implications for MTB virulence evolution. On the other hand, the algorithm itself was also discussed in depth as for its assumptions and its generalization to other evolutionary scenarios.

3.2 Research hypothesis II:

Gene clustering analysis at the whole *Mycobacterium* genus level based on the gene family frequencies of its pangenome would help to identify gene clusters related to species adaptation and pathogenesis evolution.

The genus *Mycobacterium*, including both MTB and other non-tuberculosis bacteria, appears to be suitable for pangenome modeling at the genus level and for identification of gene clusters, since this genus is made up of a wide range of host associated pathogenic species and has a dynamic pangenome with apparent horizontal gene transfer and gene loss, in contrast to the conservative pangenome of the MTB species. **In this thesis, to obtain a more comprehensive picture of the evolution of the whole *Mycobacterium* genus including the formation of different pathogenic species, another thread of study aims and objectives was to develop a gene clustering method that uses the pangenome information at the genus level (Chapter 6).** In this effort, complete genomes of different mycobacterial species were retrieved from public database to include as many species as possible, and their pangenome gene family frequencies, or the 'supragenome' model, were calculated. Based on these data, a new phyletic model was developed to quantify coevolutionary scores among different gene families. With these scores, the MCL graph clustering method was used to identify coevolved gene clusters. After that, gene clusters

involved in the pathogenesis evolution or host adaptation were examined and discussed, which demonstrated the usefulness of this pangenome gene clustering method for discovering important gene clusters underpinning species genotypes evolution, given a pangenome sequencing a the genus level. And also, these identified gene clusters supported the proposal that the MTB species evolved from an environmental ancestral species with a larger genome and then have undergone a series of steps of genome reduction to become host-associated pathogens.

Chapter 4

A local parsimony method for ancestral state reconstruction

4.1 Introduction

Ancestral state reconstruction (ASR) for internal nodes on a phylogenetic tree is useful for comparative evolutionary analysis, and for testing hypotheses of evolutionary processes (Swofford and Maddison 1992; Cunningham, Omland et al. 1998). In a reconstruction analysis, by explicitly mapping character states on a phylogenetic tree, a variety of evolutionary events can be inferred to understand the way by which the characters evolve (Coddington 1988; Donoghue 1989). For instance, gene gain and loss during the evolution of a microbial species can be inferred from comparing an ancestral node and its descendants (Mirkin, Fenner et al. 2003). In a word, ancestral state reconstruction is trying to infer the unknown states of ancestral nodes from observed states of extant organisms that are on the leaves of the referenced tree. In a bifurcating tree, there are in total *n-1* ancestral nodes and thus *n-1* states to infer, given the *n* observed states on the leaves (Omland 1999).

For a given phylogenetic tree, usually bifurcating, and the associated states on the leaves, there are three classes of methods for reconstructing ancestral states: maximum parsimony, maximum likelihood and Bayesian inference (Swofford and Maddison 1987; Yang, Kumar et al. 1995; Cunningham, Omland et al. 1998; Huelsenbeck and Bollback 2001). Among them, the maximum parsimony method is most widely used, and has been well adopted in reconstruction analysis, to obtain a reconstruction with a minimum number of state changes (Maddison and Maddison 2000; Swofford 2003; Maddison and Maddison 2006). For a two-state character, a minimum-change reconstruction is the reconstruction with fewest changes between the two states during their evolution on the tree. Algorithms

have been developed for maximum parsimony ancestral state reconstruction, including the most popular Fitch's algorithm and Sankoff's algorithms, which are dynamic programming algorithms (Felsenstein 2003). Like the dynamic programming algorithm for sequence alignment, these algorithms proceed with two phases. First, they start with the leaves and their associated initial values of states, and then proceed upward to their parent nodes, recording the number of state changes for all possible states assigned. This process is iterated until reaching the root, and the minimum number of total changes determined. In the second phase, with the determined minimum number of changes on the root, the algorithms iterate downward from root to leaves, to revolve the states for each node that are corresponding to the minimum changes (Felsenstein 2003).

Global parsimony is applied in these algorithms, that is, to obtain a minimum number of changes on the entire reference tree (Fitch 1971). And the first use of this criterion is to reconstruct the maximum parsimony tree in phylogenetic analysis. For all the candidate trees, representing all bifurcating patterns of the analyzed species, some trees have more parsimony patterns and should be chosen to represent the true evolutionary histories (Felsenstein 2003; Felsenstein 2003). But in the situation of ASR, there is a different aspect compared with phylogenetic analysis, that is, the reference phylogenetic tree is given in prior for reconstruction analysis, instead of to be inferred (Felsenstein 2003). In spite of this subtle but relevant difference, traditional ASR methods still rely on the global parsimony. Regarding that a reference tree is presumed and supposed to represent a reliable evolutionary history for the organisms to study, the global parsimony can be modified for the ASR problem to take into account the hierarchical information of the reference tree, which could be valuable for state inference.

Traditionally, hierarchical structure of a phylogenetic tree is of critical interest and helpful for evolutionary reasoning, such as cladistics analysis (Maddison, Donoghue et al.

1984; Coddington 1988). In cladistics analysis, organisms are grouped into a clade if they share a same derived character, which is inherited from their lowest common ancestor, but absent in the more distantly related organisms. In this sense, selection of an out-group is important to determine whether or not a character is derived from a common ancestor and thus an indication of a common ancestry. Out-group analysis is usually resorted to in a procedure to revolve character states, under the observed character distribution among organisms (Maddison, Donoghue et al. 1984; Nixon and Carpenter 1993). Systematical application of cladistics analysis to a group of organisms with measured character states can generate a cladogram of the organisms, which is also a phylogeny. But evolution of biological characters usually involves convergence, reversal, and transformation (Cunningham, Omland et al. 1998). For example, a derived character from the common ancestor of the clade could be lost in the descendants. There are also possibilities that different character sets conflict with a cladistic grouping of organisms. To encompass these kinds of conflicts, cladistics analysis should be performed under a flexible framework. One solution that has been developed is to combine the out-group analysis with parsimony analysis (Maddison, Donoghue et al. 1984). This solution consists of two steps, both invoking local parsimony analysis. The first step infers the state of the 'out-group node', which refers to the node connecting the out-group to the common ancestor ('in-group node') of the in-group, which is a group of organism for their cladogram to be resolved. The parsimony analysis in this step takes into account only the information of the out-group taxa, but ignoring information of the in-group taxa, therefore it is local. With the state of the 'out-group node' inferred, a second round of parsimony analysis is taken on the in-group to find a cladogram of local parsimony. As a result, different out-group patterns (e.g., out-group taxa states, out-group taxa hierarchy) can lead to different cladogram of the in-group taxa. And proper selection of the out-groups and their accurate state information can help construct reliable in-group cladogram (Maddison, Donoghue et al. 1984).

Local parsimony analysis has been used in cladistics analysis to consider the out-group and in-group relationships. Modifying and extending the traditional maximum parsimony algorithms for ancestral state reconstruction to introduce local parsimony would be promising. In this chapter, an ASR algorithm has been developed with a set of rules of local parsimony analysis. This algorithm has been applied in the ancestral reconstruction of genomic insertion/deletion events and IS6110 elements in fifteen *Mycobacterium tuberculosis* genomes. The results were shown to be consistent with previous results of the well-characterized large indels (presented in the next chapter). Starting with this simple algorithm, I went further to extend it in several ways. First, several algorithmic concepts and models were proposed to explained the local parsimony method further. And it was demonstrated that these concepts and models provide a useful framework for the extension of this algorithm. Further examination also showed that this local parsimony algorithm is a subcase the general ASRs by preferring a local minimization of state changes. Second, scoring matrices for weighting the transformation between different states were incorporated to the algorithm to allow for unequal rates of character changes. At last, this method was further generalized to consider incorporating statistical analysis, which is possible in the proposed framework and would be useful for phylogenetic data analysis.

4.2 A simple ancestral state reconstruction algorithm using local parsimony

For an in-group of species with observed character states, ancestral character state can be inferred for the lowest common ancestor (LCA, or most recent common ancestor, MRCA) of this group. The simplest case for this is that if all the species have the same state, the LCA should be inferred to have this same state by parsimony criterion. But there are situations in which not all the states are the same, and different state assignments to the LCA have to be evaluated. By using an out-group analysis, character states of the out-group

species could be used for such evaluation, as those performed in cladistics analysis. When this evaluation is performed on a bifurcating tree, which usually is the case at hand, some basic rules have to devise for an algorithmic processing. Shown in Figure 4.1 is such a set of rules, consisting of the basic scenarios of evaluation using parsimony criterion, which form the ground for the presented ancestral state reconstruction algorithm. In this algorithm, only two-state character is considered, which are designated as state '*0*' or '*1*'. These rules and scenarios, illustrated in Figure 4.1, are explained in detail as follows:

Rule 1: If the descendants have the same state ('*0*' or '*1*'), their LCA is inferred to have this same state ('*0*' or '*1*'), as shown in Figure 4.1 (A) and (B).

Rule 2: If different states are observed between the two descendants, a 'hidden' state is assigned to the LCA, which are representing uncertain about the optimal state ('*0*' or '*1*') to choose, as shown in Figure 4.1 (C). And revolving this uncertainty needs out-group information.

Rule 3: Hidden state of an ancestral node can be resolved to be the same state as of an out-group, which is a sibling of this node, descending from the same LCA, as shown in Figure 4.1 (E) and (F).

Rule 4: If the two states of the descendants are hidden, then the state of their LCA is also hidden (Figure 4.1 (D)). But they are all of the same underlying state, '*0*' or '*1*'. In other words, once the hidden state has been resolved to be '*0*' or '*1*', they are all assigned to the same explicit state of either '*0*' or '*1*' accordingly.

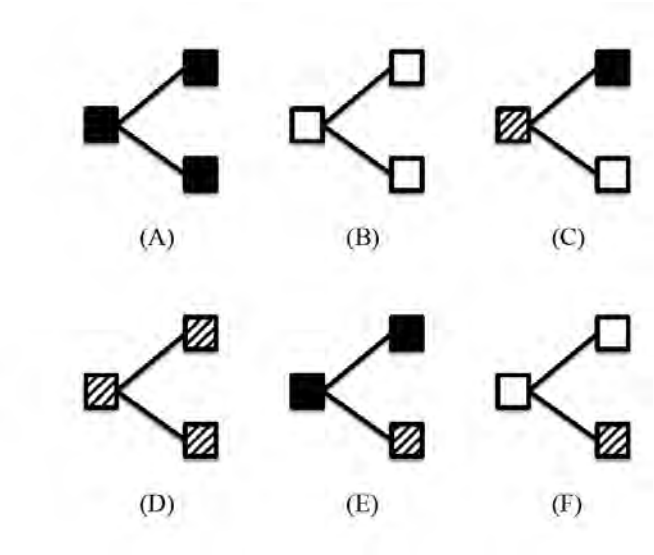

Figure 4.1: The parsimony rules and basic scenarios for the presented ancestral state reconstruction algorithm. Only two-states characters are considered in this algorithm. The states are '*0*' (in white square) and '*1*' (in black square). A 'hidden' state is depicted as striped square, representing uncertain states to be further resolved.

With the rules described above, an ancestral state reconstruction algorithm is straightforward to come out, as illustrated in Figure 4.2. This algorithm performs a post-order tree traversal, from the leaves to the root, applying the rules successively. Dealing with ***rule 1*** is simple just by assigning the state of the descendants to their LCA. The major operations involve in applying ***rule 2-4***. Briefly, there are two processes to go with, namely 'tracking' and 'triggering', which are illustrated in an example of Figure 4.2 and described as follows:

(1). ***The 'tracking' of hidden state series***: Before a hidden state to be resolved, there might be a series of hidden states, connected to each other on the tree, which are still failed to determine at that point. For instance, as illustrated in Figure 4.2 (A), depicted in blue arrows, the state of node *i* is hidden, since its two descendants (*h* and *c*) have different states (***rule 2***). The same is for the node *k*, and then the node *j* (by ***rule 4***). In the algorithm, these

hidden states are kept 'tracking' until a resolution of the state when proper information from the out-groups is available (***rule 3***).

(2). ***The 'triggering' of hidden state series***: When an out-group of determined state, such as node *l* in Figure 4.2 (B), is encountered with a hidden state (node *j*), the state of their LCA and the hidden sate can be resolved. In this case, the hidden state has been tracked in a series, then a 'triggering' is invoked to resolve the series to be of the same sate as the out-group, as shown in Figure 4.2 (B) with red arrows.

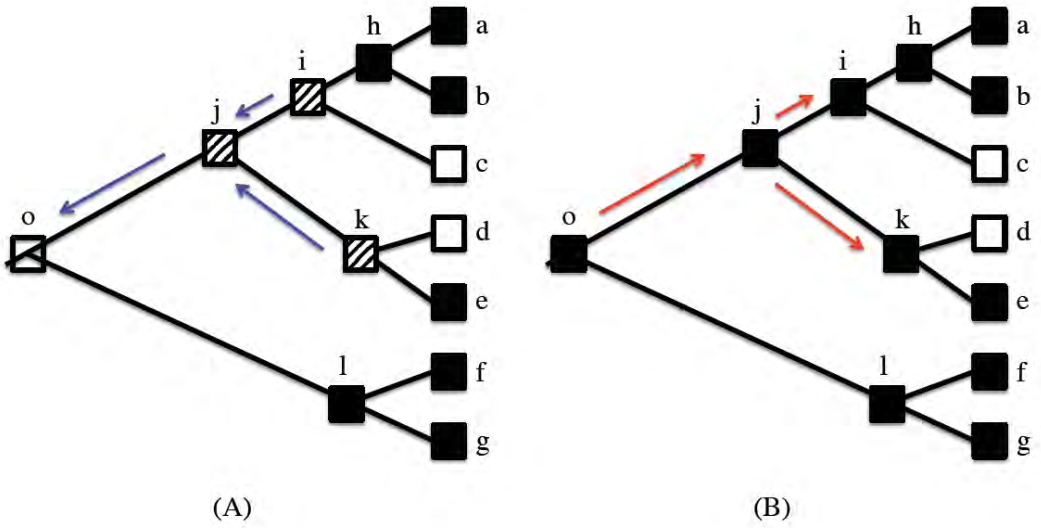

Figure 4.2: The algorithm for ancestral state reconstruction. This algorithm proceeds upward from leaves of the tree to its root, using post-order tree traversal. (A) One of the major operations in this algorithm is the tracking of a series of hidden state along the traversal, indicated with blue arrows. (B) These hidden states are then triggered to be one of the possible real states of '*0*' or '*1*', according to the state of relevant out-group during traversal, indicated with red arrows.

In some situations, the 'triggering' might be not invoked even then the traversal process is reaching the root and ended. As a result, the series of hidden states are remained undetermined, as demonstrated in Figure 4.3. In such cases, depending on the purpose of the

reconstruction analysis, different strategies can be applied. If the purpose is to reconstruct the state explicitly for the nodes on the tree, a presumed default state may be assigned to all the hidden states. For example, in a prediction of gene gain and loss during species evolution, according to some prior biological knowledge about the particular gene analyzed, it can be more possible for the gene to present in the oldest ancestors, and therefore a default '*1*' state can be used, and *vise versa*. But if the purpose is only to count the number of state changes, then the remained hidden series have no effect on the counting, and can be just left therein. For example, in Figure 4.3, there are three changes predicted, not matter what the hidden states are.

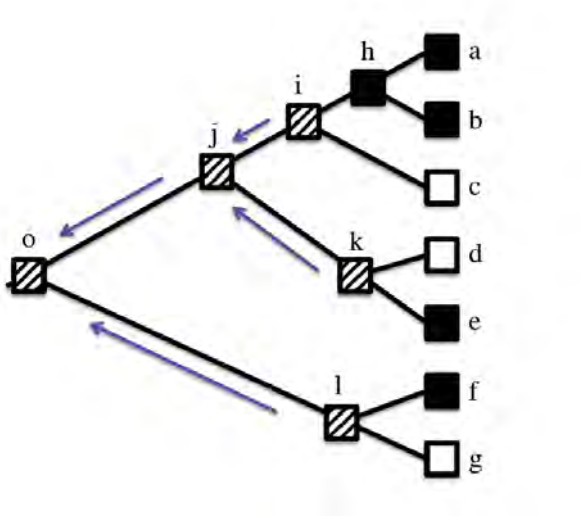

Figure 4.3: A satiation of failing to determined the hidden states. Depending on the purpose of ASR analysis, these nodes can be assigned to a default sate, or just left wherein.

4.3 Comparison with global parsimony analysis

The algorithm presented above is closely related to the traditional ASR algorithms in some aspects, but not the same. Close examination reveals that it is a strong subcase of the general ASR algorithms, with more assumptions about the biology of evolution and explicit reconstruction statements as regard to the ambiguities of possible multiple equal

reconstructions produced by the general ASR algorithms. More than that, this algorithm also provides a framework for extension to incorporate explicit statistical hypothesis testing during the reconstruction analysis, as will be describing in the following sections.

A brief proof of the local parsimony algorithm as a strong special case is given here. Traditional ASRs are instances of the more general algorithm of dynamic programming. For a problem to solve with dynamic programming, a more global optimization (or minimization) is obtained by solving small local optimizations. In ASRs, the minimum cost for a LCA (node *LCA*) is calculated as (Felsenstein 2003):

$$S_{LCA}(i) = \min_j[c_{ij} + S_l(j)] + \min_k[c_{ik} + S_r(k)].$$

For state *i* to be assigned to a node *LCA*, states of the two descendants (*j* for node *l* and *k* for node *r*) are chosen to minimize their sums of cost for state changes and the cost for the selected states in the descendants. In essence, the rules given in Figure 4.1 are implementing the equation above. ***Rule 1*** has minimum cost with no change involved by assigning LCA the same state as two descendants. The same holds also for ***rule 4***:

$$min\{S_{LCA}(a)\} = min\begin{Bmatrix} S_l(a) + S_r(a) \\ 2 + S_l(b) + S_r(b) \\ 1 + S_l(a) + S_r(b) \\ 1 + S_l(b) + S_r(a) \end{Bmatrix}$$

$= S_l(a) + S_r(a)$, given that, $S_l(a) = S_l(b), S_r(a) = S_r(b)$;

$$min\{S_{LCA}(b)\} = min\begin{Bmatrix} S_l(b) + S_r(b) \\ 2 + S_l(a) + S_r(a) \\ 1 + S_l(a) + S_r(b) \\ 1 + S_l(b) + S_r(a) \end{Bmatrix}$$

$= S_l(b) + S_r(b)$, given that, $S_l(a) = S_l(b), S_r(a) = S_r(b)$.

In ***rule 3***, because a same cost for either '*1*' or '*0*' for hidden state, choosing the state to be the same as the out-group for LCA is minimum costing. As for ***rule 2***, the hidden state represents an equal choice between '*0*' and '*1*', both of which need at least one change. With all the above together, the minimum parsimony in this algorithm has been proved.

It is well recognized that traditional ASRs may generate multiple reconstructions of equal parsimony from a same data set, and this has been considered as one of their shortcomings, because of the ambiguity introduced by the multiple choices. As shown in Figure 4.4 (A), (B) and (C), the three reconstructions are equal global parsimony, involving a same number of two changes to reconstruct ancestral states from observed data. But the presented algorithm produces only one of them, Figure 4.4 (A), with the states of node *f* and *e* assigned to '*0*' and whatever state for node *g*. Provided that a out-group to the root is available, the hidden state of the root in (A) will be resolved to '*0*' or '*1*' accordingly, without affecting the node *f* and *e*. But the situation is different for (B) and (C). A '*0*' out-group to (B) will make not parsimonious any more, and the same for a '*1*' out-group to (C).

Difference between the three alternatives, although of same parsimony, is not trivial, especially examined under some evolutionary scenarios. In terms of cladistics analysis, local clades of species and less parallel speciation are preferred in the presented algorithm than the traditional ones. In details, such as example of Figure 4.5, the local parsimony reconstruction (A) prefers small local cade of character '*0*' followed by emerging of a new species of '*1*'. In contrast, the general and global methods produce a large and global clade of '*1*' from the root, followed by two parallel new species of '*0*'. One justification for the (A) scenario could be that, evolution may prefer less parallel speciation of similar or identical phenotypes, and therefore the 'tree of life' concept gains its popular acceptance. This might be true for more complex biological traits, and also for indels in genomic

rearrangements, for which parallel insertion and deletion at the some position would be less likely.

4.4 Underlying ideas behind the local parsimony algorithm

Parsimony analysis or maximum parsimony assumes minimum changes of evolutionary events and applies this criterion to find a most parsimonious reconstruction. In the presented local parsimony algorithm in this thesis, the introduction of the 'hidden' state, augmenting with the 'tracking' and 'triggering' operations, is surrogating for this principle. Depicted in Figure 4.6 (A), the series of hidden states indicate that there is no evident support from the observed states of the leaves to invoke changes in the ancestral nodes. Although the ancestral states are unknown, needing extra information to resolve, they are the same and thus parsimonious in evolution. More general, as depicted in Figure 4.6 (B), we assume that characters have been inherited during the course of species evolution, staying similar or identical, if without evidence to reject this assumption, which is a ***null model*** (or ***null hypothesis***) of evolution, in statistics terms. This assumed trends of 'continuity' or 'inheriting' rather than more changes is considered and applied in the algorithm even more specifically to prefer local parsimonious clades, as explained in previous section. For example, as in Figure 4.6 (C), the null mode of hidden *'i-h-g-f'* states in (B) is rejected by the observation of '*c{0}-d{1}-e{0}*' on the leaves, and the ***alternative model*** (or ***alternative hypothesis***) is made by the algorithm, which is '*g{0}-f{0}*'. As a result, the null model of '*i-h-g-f*' is tested and splits into a null '*i-h*' and a alternative '*g{0}-f{0}*'. In this sense, this algorithm is named with 'local parsimony' to reflect this fact, that is, local in '*g-f*' resolution. These ideas are also in action in the traditional ASR algorithms, but implicitly.

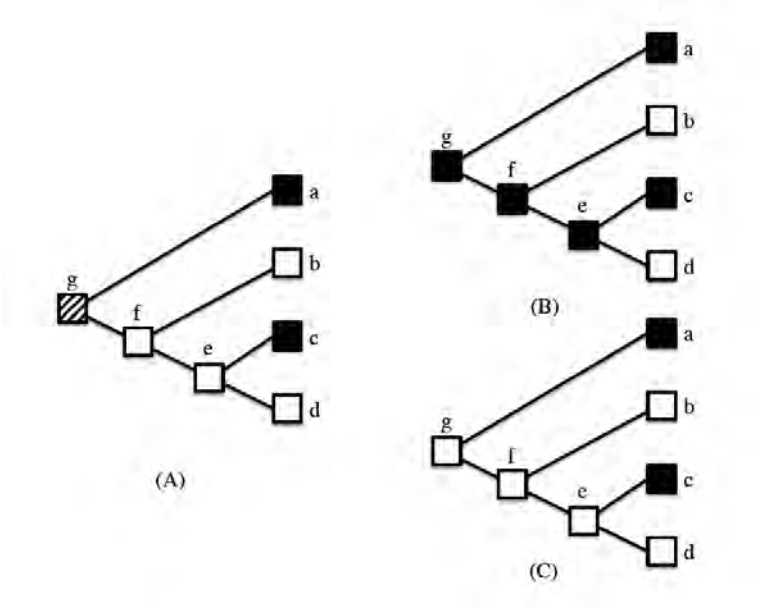

Figure 4.4: Comparison of the algorithm to the general ASR algorithms. This algorithm appears as strong case of the general ASR algorithms, and prefers a special solution (A) among the three equal parsimony reconstructions (A), (B) and (C).

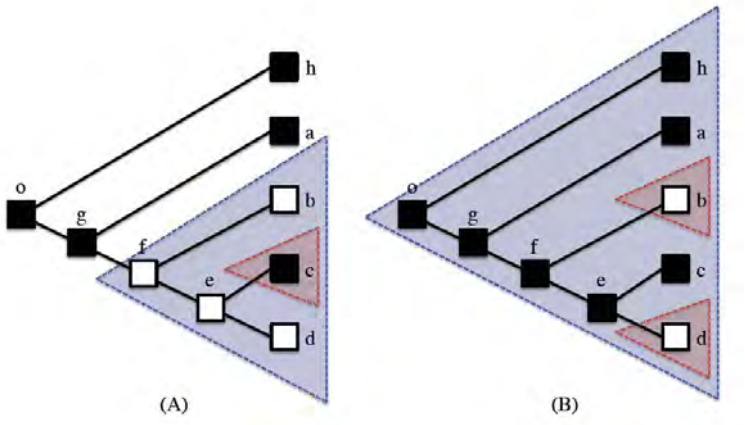

Figure 4.5: Different evolutionary meanings between the local and global parsimony reconstructions. In terms of cladistics analysis, reconstruction (A) prefers a local clade of character '*0*' (blue shaded triangle) followed by a speciation of a new sub-clade of '*1*' (red shaded triangle). Reconstruction (B) indicates a larger clade of '*1*' (blue triangle) embedded with parallel speciation of two new sub-clade of '*0*'.

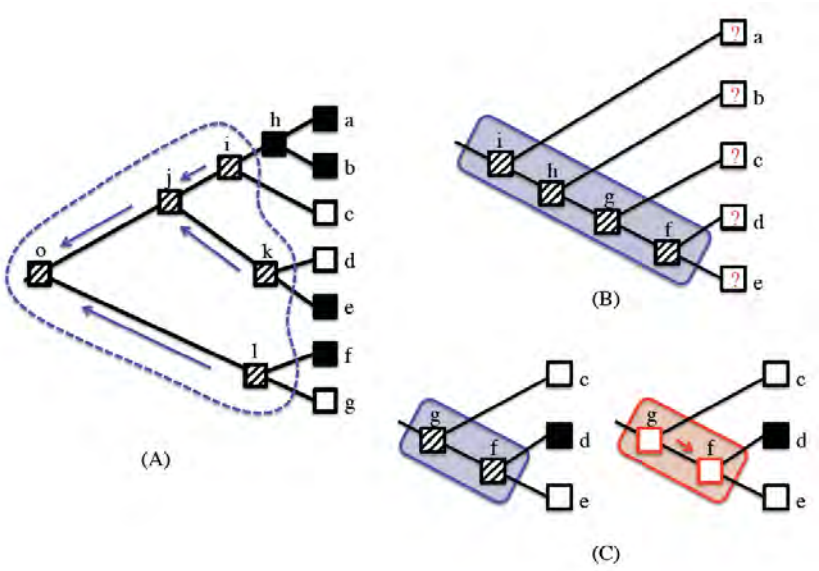

Figure 4.6: The underlying ideas behind the local parsimony algorithm. (A), the 'hidden' states and the accompany operations of 'tracking' and 'triggering' in this algorithm are surrogating for parsimony criterion of minimum changes. (B), the ***null model*** of no change is assumed, if no observation of leaves available, indicated by a blue shaded rectangle. (C), with states of leaves available, the ***null model*** is rejected (blue rectangle) and an ***alternative model*** (red rectangle) is chosen to assign the states to ancestral nodes.

It seems that the algorithm presented here brings them out explicitly for the first time, and provides an algorithmic framework to apply them. With the explicit statements of the null and alternative model, it then can go beyond to introduce multi-states characters and apply soring systems for state transformations, which is studied in the next section.

4.5 Introducing multi-state and scoring matrices

The algorithm presented above only deals with a two-state ('*0*' and '*1*') case. But in many situations, biological data involves more than two character states, or multi-states, to resolve. To encompass a multi-state ASR, scoring system has to be introduced to weight the transformation between different states, which takes the form of a scoring matrix. In this section, based on a same scoring matrix strategy, the simple two-state ASR algorithm is extended to allow for multi-state ancestral reconstruction. The extension follows the basic

principle of '***null***' and '***alternative***' models, proposed in the previous section to explain the underlying ideas behind the simple two-state algorithm.

As illustrated in Figure 4.7 (A), if there is no evidence for introducing different scores for transformation between different states, then the ***null model*** assumes that ancestral nodes are of the same state, which can be any state from the state space. In a parsimony explanation, this ***null model*** assumes no change in ancestral nodes. And any change occurs only in the descending of leaves. In this perspective, the ***null model*** acts like a maximal entropy model, because the number of changes in the leaves in this situation is the largest in all possible ASR reconstructions. Introducing scoring matrices can help to reject the ***null model***, and therefore to assign different state to ancestral nodes, which is called the ***alternative model*** in this algorithm (Figure 4.7 (B) and (C)). For a scoring matrix, it does not always reject the ***null model***, as shown in Figure 4.7 (B), which scores all transformations equally and provides no information to infer the ancestral states. Actually, many matrices do provide information for ancestral state inference, such as the case in Figure 4.7 (C), in which the '*X*-*Y*' transformation is less likely. With this unequal matrix, if used in the case of Figure 4.7 (A), the null model can be successfully rejected. In details, '*W*' or '*Y*' in the node *f* is not equal likely anymore, with regard to the '{*X, Z*}' state of the node *e*. The same is for '{*X, Z*}' in the node *e*, regarding the possible state of the node *f*. It can be seen that, either state '*Z*' or '*W*' is more likely than the other two, to be assigned to the ancestral nodes. As a result, by rejecting the null model, the ancestral nodes 'e-g-f' can be either of state '*Z*' or '*W*', with a score of 3 calculated from the scoring matrix.

To implement the ideas illustrated in Figure 4.7, the simple two-state algorithm can be extended, based on the 'tracking' and 'triggering' operations during tree traversal, as illustrated in Figure 4.8. In fact, these two operations are corresponding to the accepting or rejecting of the ***null model***. During the post-order tree traversal from leaves to the root, for

each encountered ancestral node, possible states of its two descendants are evaluated to test the ***null model***. If the ***alternative model*** is accepted, then local parsimony analysis is carried out by the triggering operation to assign states to a train of hidden states that have been tracked so far. If not, the tracking is continuing to maintain a train of hidden states, which can be any of the possible states of their leaves. For the example of Figure 4.8, leaves '*a*' and '*b*' are of different states and therefore invoke a tracking in their ancestor (node *e*) with states '{*X, Y*}'. The second step goes to the leave *c* and the ancestral node *f*. To determine the state of node *f*, an evaluation of '{*X, Y* | *Z*}' is taken to test the ***null model***. For example, if it is more likely to have a '*X-Z*' than a '*Y-Z*' transformation, or '*c(Z)*' prefers '*f(X)*', then a triggering is invoking to assign states to '*e-f*' parsimoniously, depending of the underlying scoring matrix. If not, the tracking process will continue and update the possible states from '{*X, Y*}' to '{*Z, Y, Z*}'. This process of model evaluation and algorithmic operation is repeating iteratively until reaching the root, and the result is the construction of ancestral states according to a scoring matrix.

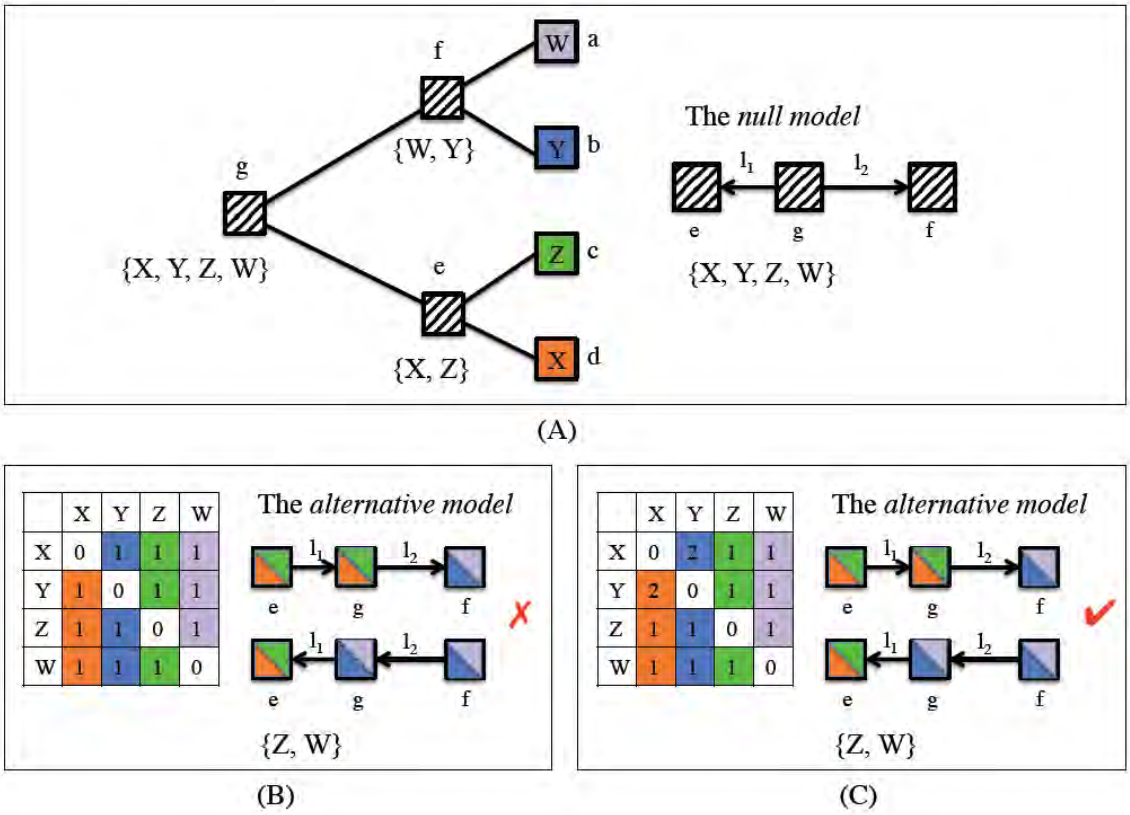

Figure 4.7: An example of multi-state ancestral reconstruction with scoring matrices. (A) The ***null model*** assumes that if there is no evidence for resolving the transformation between different states, then all the ancestral nodes have the same hidden state, which can be any of the state space (e.g., {X, Y, Z, W}). Introducing scoring matrices can assign weights to state transformation for evaluating the ***alternative model***. (B) An equal scoring matrix provides no information for rejecting the ***null model***. (C) A scoring matrix that weights more on '*X*-*Y*' transformation and rejects the ***null model***.

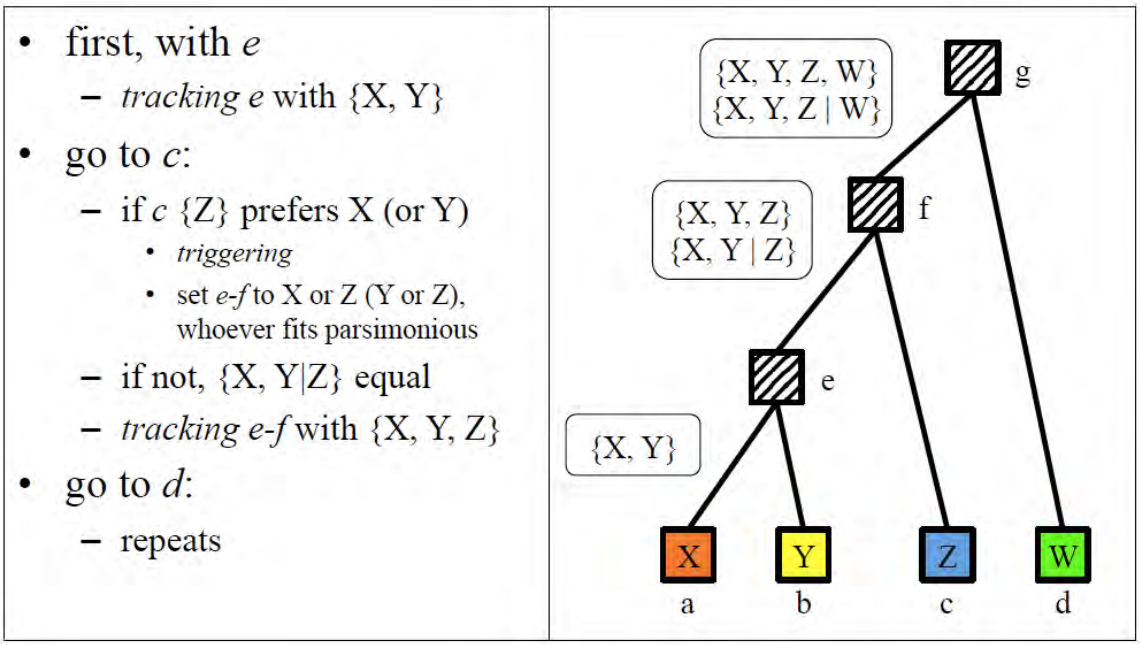

Figure 4.8: Extension of the simple two-state algorithm for multi-state ancestral reconstruction using the 'tracking' and 'triggering' operators. In this example, by a post-order tree traversal, the pair of '*X-Y*' is processed first, and their ancestor (node *e*) is 'tracked' with {*X, Y*}. Then goes to the leaf *c* and node *f*, which can be either 'tracked' or 'triggered', depending on the underlying scoring matrix. The same procedures are repeated during tree traversal until reaching the root.

4.6 Further statistical generalization of the local parsimony algorithm

Furthermore, the models and operations demonstrated above go beyond the extension of the algorithm for multi-state ancestral reconstruction. In fact, statistical techniques can be incorporated into the test of the ***null model***. In the above algorithm, comparing the score of different state transformation is used to reject the ***null model***, when there is a particular transformation that is more likely with the scoring matrix. In this manner, no obvious statistics has been used. But further generalization can be made to use model comparison statistics. As depicted in Figure 4.9, when evaluating the ***null model*** in an ancestral node, state spaces of its two descendants have to be compared. A simple treatment can be that compares the mean scores of these two spaces. For example, pairwise scores can be generated using the scoring matrix for each descendant. And then pairwise scores are calculated between the two descendants. With these score values, traditional statistics, such as *t*-test, can be used to deduce a *p*-value for between-groups difference. If this *p*-value is significant, the triggering operation will be invoked for local parsimony analysis, otherwise, tracking operation will be invoked to keep track of the hidden state and its associated state space. With this kind of statistical treatment, a single pair of transformation of small score (e.g., '*X*-*Q*') may be not enough to make a triggering operation, if the group-difference is not significant. But how to use an appropriate statistical model for the purpose of ancestral state reconstruction needs careful consideration, especially that, such statistical model should come from the underlying scoring matrix of state transformation. In this thesis, this question is not addressed. But the algorithm provides a framework for including statistics into parsimony analysis, which is promising in phylogenetic analysis of character data and will be explored in my future work.

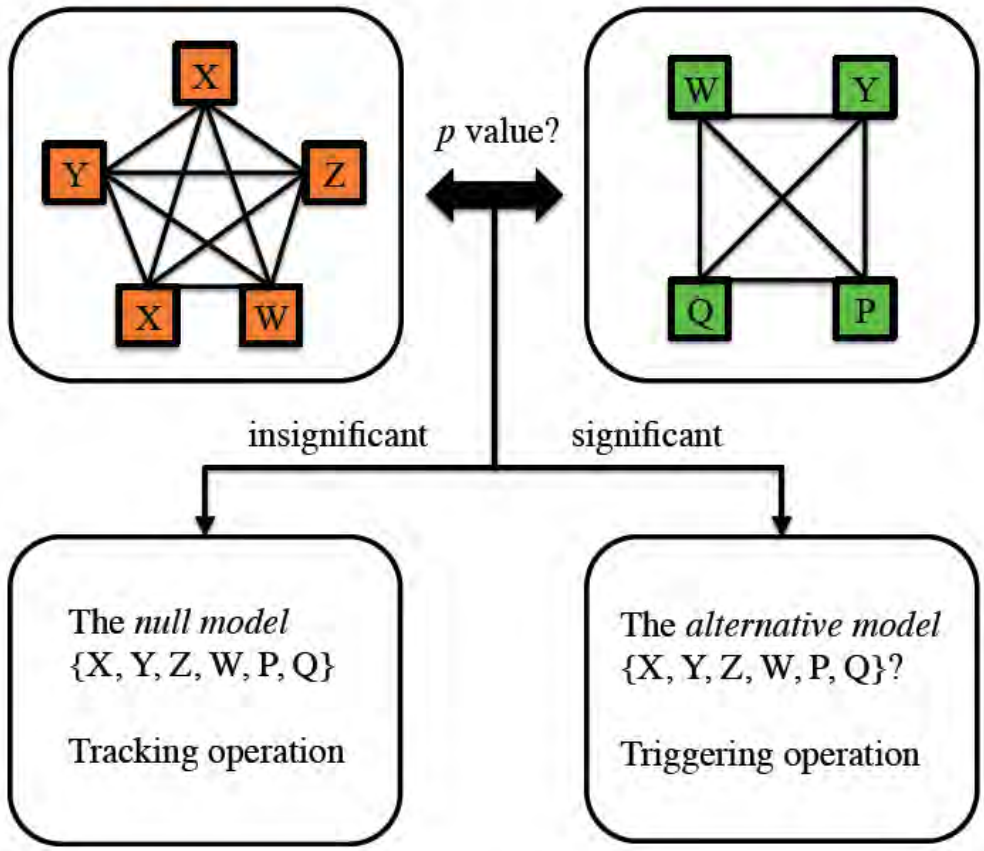

Figure 4.9: Statistical generalization of the local parsimony algorithm. A *p*-value can be deduced from group comparison statistics of the two state spaces. If this *p*-value is significant, the triggering operation will be invoked for local parsimony analysis, otherwise, tracking operation will be invoked to keep track of the hidden state and its associated state space.

4.7 Conclusion

The presented local parsimony method for ancestral state reconstruction has been used to reconstruct the ancestral genome of fifteen *Mycobacterium tuberculosis* strains, for a purpose to analyze the evolution of the *M. tuberculosis* Beijing family. Results of the ancestral reconstruction analysis, as shown in the next chapter, demonstrated that this method is useful for analyzing indel evolution with a pangenome data. Particularly, by the 'local' parsimony, evolutionary scenario of less independent parallel changes is preferred by the method. In this regard, this method is a subcase of the general parsimony ancestral reconstruction methods, which make no distinction between different evolutionary scenarios, and therefore usually produce global parsimony reconstruction with ambiguities. Especially for indel evolution in the *M. tuberculosis* genomes that have relatively few reversion and conversion, this method presents a better solution than the general ASRs.

The proposed ***null*** and ***alternative*** models and the corresponding algorithmic 'tracking' and 'triggering' operations have helped to formally illustrate the underlying ideas behind the local parsimony method. Furthermore, these concepts provide a general framework for extension of the algorithm to allow for multi-state ancestral reconstruction using scoring matrices. The framework also provides an alternative perspective to understand parsimony analysis in phylogenetics instead of the usual one of dynamic programming, which appears to be more insightful and can be used to incorporate statistical analysis into the method, as illustrated above. Although not fully explored in the current work, this statistical feature should be a good aspect of the local parsimony method, which goes beyond pure arithmetic operation of the usual ASRs and will be the direction for future works.

Chapter 5

Ancestral genome reconstruction for analyzing the evolution of the *Mycobacterium tuberculosis* Beijing family

5.1 Introduction

The *Mycobacterium tuberculosis* (MTB) Beijing family, which was first characterized as a closely related group of MTB strains with conserved IS6110 RFLP and spoligotyping patterns, which is prevalence in the Beijing region of China and nearby countries, has been shown to be emerging as a relatively recent epidemic worldwide (van Soolingen, Qian et al. 1995; Glynn, Whiteley et al. 2002). This family has several characteristics associated with its prevalence, including multidrug resistance, increased virulence, coevolved with BCG vaccination and other immunity and biochemical properties (Parwati, van Crevel et al. 2010). Regarding to the higher rate of worldwide occurrence of this family than other MTB genotypes, especially in East Asia (>75%), it has been hypothesized that this family has evolved with a selective advantage, which might hamper the global TB control (Parwati, van Crevel et al. 2010; Hanekom, Gey van Pittius et al. 2011). Therefore, understanding the evolution and formation of this closely related group, in comparison with other non-Beijing groups, may help to reveal the underlying mechanism of its selective advantage, and its implications for MTB evolution under the current trends of drug treatment and TB vaccination. One of the powerful ways to addressing this problem is genome wide evolutionary analysis, in combination with information from epidemiology studies (Hershberg, Lipatov et al. 2008; Wirth, Hildebrand et al. 2008; Ford, Yusim et al. 2012).

Previous genomics studies of the species of *Mycobacterium tuberculosis* complex (MTBC) have shown an extremely high degree of genome sequence similarity (>99.9%) between the members of this species, which was assumed to be a result of an evolutionary bottleneck before its recent dissemination in human populations (Sreevatsan, Pan et al. 1997; Brosch, Gordon et al. 2002; Fleischmann, Alland et al. 2002). But ecotypes of phylogeographical MTB lineages of different host adaptation and different virulence, including the Beijing family, have been identified and characterized with different genomic sequence polymorphisms (Tsolaki, Gagneux et al. 2005; Dou, Tseng et al. 2008; Rindi, Lari et al. 2009; Faksri, Drobniewski et al. 2011). These polymorphisms are mainly the results of indel events that occurred in different evolutionary time scale, and have been used to reconstruct an evolutionary scenario for the species. There are several classes of indels, which are evolving with different mechanisms, including insertion sequence (e.g., IS6110) mediated recombination, direct repeats, large region of genomic difference and small gene truncations. In together, these indels play an important role in MTB species evolution, which shows a tendency of genome degrading at a species level (Brosch, Pym et al. 2001).

Early comparative genomics studies, by employing differential hybridization arrays, have succeeded in establishing a set of large region of genomic differences (RD1-14 RvD1-5 and TbD1) by comparing the *M. tuberculosis* H37Rv genome to the *M. bovis* BCG Pasteur genome, or to the genomes of other strains of *M. tuberculosis* (Behr, Wilson et al. 1999; Gordon, Brosch et al. 1999; Brosch, Gordon et al. 2002). The resulting RDs are of relative large size, ranging from 2 to 12.7kb, due to the limited revolution of the comparing methods. These RDs have also been shown to be stable and ancestral, which are responding for subspecies formation (Brosch, Gordon et al. 2002). But there is evidence that MTB lineages are still undergoing small indel changes, especially those related to IS6110 activities and small gene truncations (Hanekom, van der Spuy et al. 2007; McEvoy, Falmer et al. 2007; Rindi, Lari et al. 2009). Therefore, thoroughly genome-wide analysis of the

indels would be crucial for understanding recent evolution of MTB lineages, especially for the Beijing family. Furthermore, complete genome sequencing would be more valuable to identify repeats related indels, which are usually incomplete in draft genomes but important given that about 10% of the MTB genome contains repetitive DNA (Cole, Brosch et al. 1998).

Comprehensive genome-wide indels identification is performed in this study, accompanied by complete genome sequencing of five strains of Beijing genotype. Complete genome sequences of other non-Beijing groups are also included for comparison to find out Beijing family specific indels and their affected functional genes. An ancestral genome reconstruction approach is taken to reconstruct the indel events along the evolution of the fifteen genomes to understand their association with the formation of the Beijing family. Taken together, this study reveals that there are two types of Beijing group specific indels occurred in the formation of their ancestor, including both inserted and deleted indels, which are shown to affect several genes involved in MTB pathogenesis. In particular, ancestral mapping of IS6110 elements supports the hypothesis that IS6110 activities associate with the formation of different MTB lineages, and so for the Beijing family. Two gene families of PPE24 and PPE34 are characterized with their tandem repeats of protein motifs, which appear to be maintained with relation to protein structures and associated with different MTB lineages.

5.2 Materials and Methods

5.2.1 Genome sequences and annotation

Genome sequences of five strains of Beijing family were assembled from reads generated by Roche 454 GS FLX system and the GS FLX Titanium Sequencing Kit, using the

Newbler v2.3 assembler. These strains and their associated drug resistance profiles are: BS1 and CCDC5079 (drug susceptible), CCDC5180 (MDR strain resistant to rifampicin, isoniazid, ethambutol and streptomycin), and BT1 and BT2 (resistant to all commonly anti-TB drugs, totally drug resistance or TDR). Except for BT1, gap filling was performed to obtain complete genomes for the other four strains, using PCR application and Sanger sequencing with ABI 3730xl capillary sequencers.

Open reading frames (ORFs) were predicted from the five genomes using GLIMMER, and then subjected to functional annotation by searching against the NCBI *nr* protein database, followed by manual inspection. Available complete genome sequences of MTB and their annotation retrieved from the GenBank database to be used in this study are: H37Rv (NC_000962.2), H37Ra (NC_009525.1), CDC1551 (NC_002755.2), bovis AF2122/97 (NC_002945.3), bovis BCG str. Pasteur 1173P2 (NC_008769.1), bovis BCG Tokyo 172 (NC_012207.1), F11 (NC_009565.1), KZN 4207 (NC_016768.1), KZN R506 (NZ_CM000789.1), and KZN V2475 (NZ_CM000788.1).

5.2.2 Phylogenetic analysis

A phylogenetic tree was reconstructed for the 15 MTB genomes in this study, together with the genome of *Mycobacterium canettii* CIPT 140010059 (NC_015848.1) to be used as an out-group, on the basis of a concatenation sequence of 89 phylogenetic marker genes plus 3R (DNA replication, recombination and repair) genes (Hershberg, Lipatov et al. 2008; Mestre, Luo et al. 2011), in total of 12,769 aligned columns after trimming of gaps. PHYML v2.4.5 was used to infer a maximum likelihood tree with default parameters (HKY85 model for nucleotide substitution, a BIONJ initial tree, no discrete gamma model and transition/transversion ratio of 4.480) (Guindon, Dufayard et al. 2010). The tree was rooted in the branch connecting *M. canettii* to others. This tree was used as a reference tree in the following ancestral state reconstruction analysis.

5.2.3 Whole genome alignment and indels identification

MAUVE v2.3.1 was used to produce a whole genome alignment of the 15 MTB genomes with the *progressiveMauve* method with default parameters and default LCB scoring (Darling, Mau et al. 2010). 11 locally collinear blocks (LCBs) and 24 singleton segments that were failed to align have been generated. To improve the alignment quality, the LCBs were examined manually to determine their synteny in genome location and then merge small LCBs into larger ones when synteny exists. Finally, three large LCBs were obtained, instead of one, due to a reverse of a large genomic region in the genome of KZN 4207. Furthermore, gapped regions of the whole genome alignment were identified and realigned with MAFFT v6.704b to improve gap placements, regarding a better performance in DNA sequence alignment within MAFFT (Katoh, Kuma et al. 2005) than the MUSCLE aligner used in MAUVE. Homemade scripts were developed to make this process automatically.

After obtaining the whole genome alignment, aligned columns of the 15 genomes were scanned to identify and define indels. Within each LCB, continuous columns of an identical 'gap pattern' among the 15 genomes were defined as a single indel. For an example of four species, with '0' representing a gap and '1' for a base in an aligned column, a gap pattern may be (0, 0, 1, 1) for that column. And four continuous columns of (0, 0, 1, 1), (0, 0, 1, 1), (0, 0, 0, 1) and (0, 0, 0, 1) will be defined as two independent indels, one for the first two columns and another for the latter two. The result of indels identification was then encoded into a '0-1' matrix, in which rows represent species and columns for indels states ('0' for absence in a particular species entry, and '1' for presence). This matrix was used as an input to the following ancestral state reconstruction analysis.

5.2.4 Ancestral genome reconstruction

A local parsimony algorithm has been developed to reconstruct ancestral states of the indels along the evolution of the 15 genomes on the given phylogenetic tree. This algorithm was described and discussed in detail in Chapter 4 of this thesis. In principle, it was developed to emphasize a preferred scenario of less independent parallel identical events and parsimony of local clades. In the context of the evolution of MTB species, indel events have been shown to be ancestral and responsible for discrimination between different lineages or sub-lineages, rather than independent gain and loss in different strains. Therefore, this algorithm should fit this scenario better than traditional methods. Once the presence/absence status was inferred for ancestral nodes on the tree, gain and loss of indels were derived from comparing the lowest ancestor and their descendants. Another assumption applied in this analysis is a default value of presence when there are two equal states to choose, given the observed situation of reductive evolution of MTB species.

5.2.5 Characterization of the PE/PPE gene family

Dot-plot analysis was used to inspect the pattern of self-similarity of protein sequence for each PE and PPE gene (165 genes in total) from the H37Rv genome. Tandem repeats of protein motifs were identified from the patterns, which appear as parallel lines in the dot-plot. To allow for amino acid substitutions in inexact motifs, the BLOSUM90 matrix was employed to score mismatch, with a window size of 10 amino acids and a threshold score value of 50. PE/PPE genes with tandem repeats were used to search the NCBI *nr* database for their homologous genes from other *M. tuberculosis* genomes. Protein sequences were aligned and inspected to identify the units of repeats, numbers of repeats and the boundary regions of the repeats.

5.3 Results

5.3.1 Gain and loss of indels along the evolution of MTB species

In general, there are more losses than gains of indels in the reconstruction of indel evolution in the MTB species. This trend is observed in both subspecies and lineage level, except for different strains of the same genotype. There are 293 and 125 indels lost during the separation of the subspecies of *M. bovis* and the early common ancestor of the Beijing and non-Beijing groups respectively (Figure 5.1). Another 186 and 97 indels have been lost in the split of the Beijing group and non-Beijing group from their common ancestor, but there are also a large number of indels gained. Even though, the number of losses is around twice of the number of gains. In contrast, for strains of the same genotype, the losses are not significantly more than gains, for example, within the Beijing group, H37 group and KZN group. Such case is also observed for the descending of several strains from their lowest ancestor, including H37 and F11, with only slightly more losses. The exception is the BCG group, which has significantly more gains than losses.

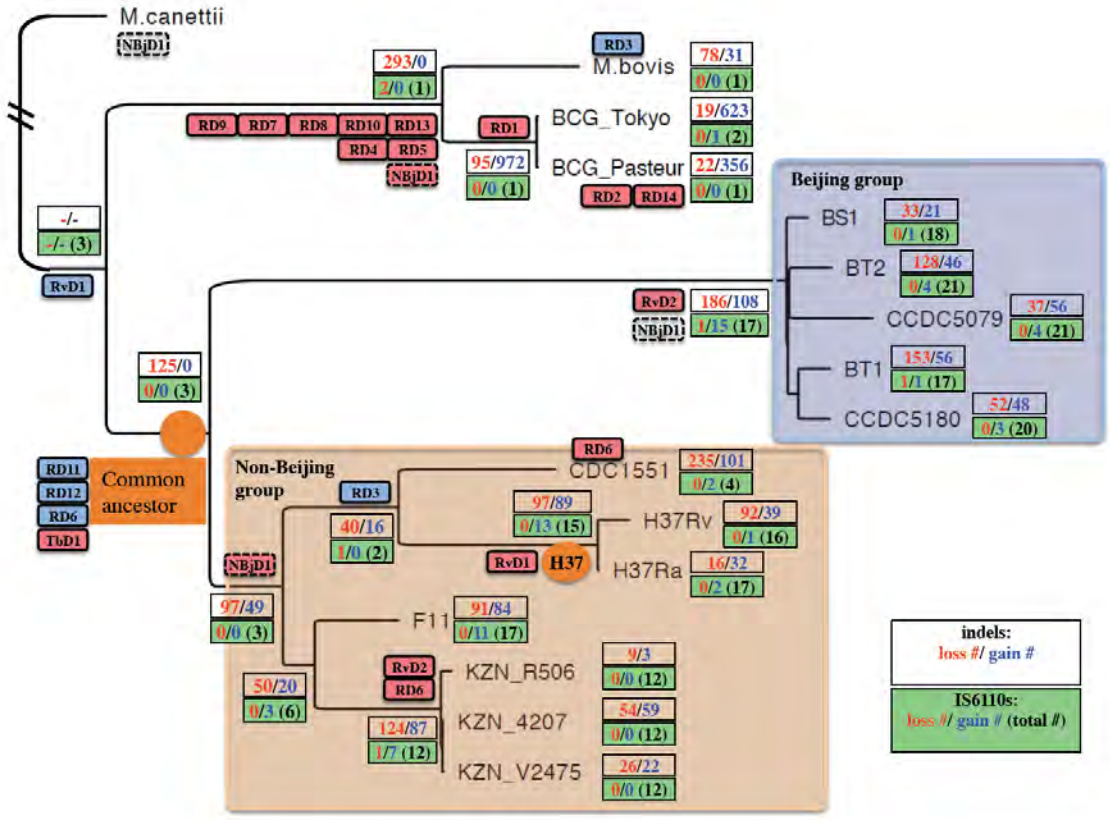

Figure 5.1: Loss and gain of indels/IS6110 along the evolution of the MTB species. Presence or absence of indels on the ancestral nodes in the tree was inferred using a local parsimony method. Gain and loss of indels was derived from comparing a lowest ancestor and its descendants. Known RDs identified in previous studies (Brosch, Gordon, et al. 2002) were also mapped on the tree. Gained RDs were depicted in blue rounded rectangles, and lost RDs in red rectangles. NBjD1, a RD specific to the Beijing group, but absent in Non-Beijing group, was shown in dash-line rectangle.

5.3.2 Ancestral indel events in the ancestor of the Beijing group

The Beijing group represents a closely related clade of strains, which diverges deeply from the progenitor of the MTB complex (the 'common ancestor'), as shown in Figure 5.1. The other genomes form the non-Beijing group in another branch from the common ancestor. In this evolutionary context, the ancestor of the Beijing group can be characterized with different kinds of ancestral indel events, which are useful for understanding the evolution of the Beijing family.

The first sets of indels that are characteristic of the Beijing group are those newly gained indels after the divergence from the early common ancestor. In total, there are 108 indels of this kind, among which 84 (~78%) are found to be located in the gene coding regions. These 84 indels are examined in detail to further clarify their influences on gene function and the genetic mechanisms associated with them. Most of the indels are results of insertions of IS element and diversity of PE/PPE genes (Figure 5.3). In details, there are 28 (~33%) indels associated with 20 IS elements, mainly IS6110. Another set of indels (26 indels, ~31%) can be attributed to repeated motifs of the PE/PPE genes. Other three genes with repeated motifs are also responsible for some of the indels (14 indels, ~17%). All these three genes are only predicted to be hypothetical, without further information on their biological functions. The remained 17 (~20%) indels consist of a large RD (4,495bp, encoding 9 genes), two large indels (106bp and 53bp) within two hypothetical genes, five in-frame indels and nine indels causing frame-shift in nine genes. Encoded genes within the 4.5k RD are listed in Table 5.1, which is a Beijing specific RD in the MTB complex, but also present in *M. canettii*. The nine disrupted genes caused by frame-shift insertion and their functions are shown in Table 5.2. As shown in Figure 5.2, some of the insertions occur frequently in a 'hotspot' region in the MTB genomes, mainly involving IS6110 and PE/PPE

genes, probably resulted from frequent intra-chromosomal recombination between the IS elements.

With comparison to the non-Beijing group, indels present in the early common ancestor can be divided into two groups. The first group comprises indels that are preserved in the Beijing ancestor but lost in the non-Beijing ancestor (97 indels). The other group consists of indels that are lost in the Beijing ancestor but still present in the non-Beijing ancestor (186 indels). Considering an observed trend of increased virulence and host adaptation by reductive evolution in different MTB species, the indels that lost in the Beijing ancestor are of particularly interest to further investigate their influences on functional genes. Among the 186 indels lost in the Beijing ancestor, 143 (~77%) of them locate inside gene coding regions. This proportion is around the same quantity of gained indels (~78%), and may reflect the relevance of indel events in the evolution of phenotypes among MTB lineages.

In the same way as the above analysis, the indels can be further divided into different groups according to their probably associated genetic mechanisms. That is, 41 indels (~29%) are related to PE/PPE gene families, 32 indels (~22%) related to IS elements, and another indel (72bp) inside the Rv3281 gene with repeated motifs. In contrast, there are many indels (34, ~24%) that occur within genes of the CRISPR/Cas system, which is not observed in gained indels in the Beijing family (Figure 5.3). The remaining 35 indels (~24%) comprise 3 large RDs (>1kbp, with 4 indels, Table 5.3), 7 small (3bp and 9bp) in-frame indels and other 24 indels causing frame-shift and gene disruption (Table 5.4).

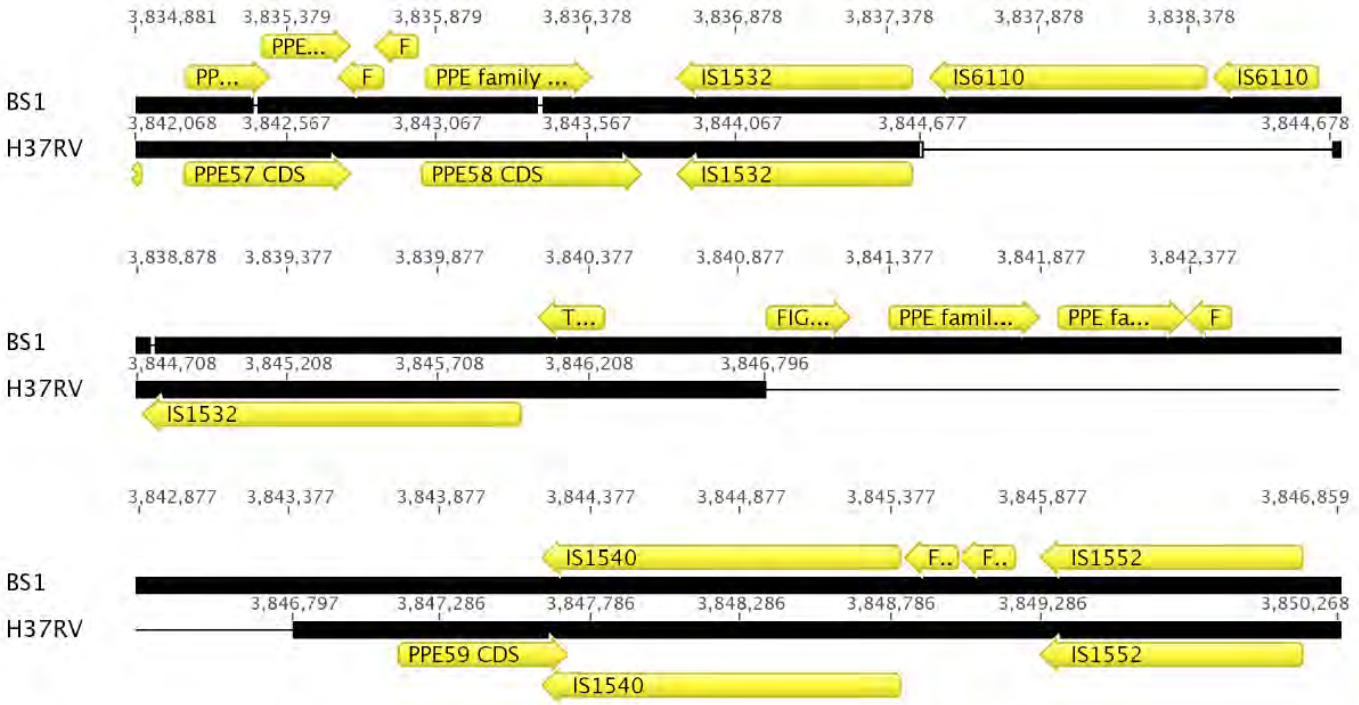

Figure 5.2: A hotspot of frequent insertion of indels in the Beijing ancestor, involving IS elements. The shown alignment is drawn from BS1 (from base 3,834,881 to 3,846,859) and H37Rv (from base 3,842,068 to 3,850, 268). Noting that, single base deletions within the PPE57 and PPE58 happened, causing truncations of these two genes.

Table 5.1: Genes encoded within the 4.5kb RD specific to the Beijing family, designated as NBjD1.

Name	Length (bp)	Direction
FIG00824736: hypothetical protein	279	reverse
probable PE family protein	129	reverse
probable PE family protein	492	reverse
FIG00827404: hypothetical protein	210	reverse
FIG00820644: hypothetical protein	750	forward
Possible secreted protein	747	forward
iron-regulated elongation factor tu Tuf-like	261	reverse
hypothetical protein	180	reverse
FIG00820216: hypothetical protein	207	forward

Table 5.2: Disrupted genes by frame-shift insertion in the Beijing ancestor. a, Disruption type describes the results of frame-shifts of the predicted CDS compared with H37Rv; b, the corresponding gene in H37Rv; c, notes of biological functions taken from the H37Rv annotations.

Indel ID	Size (bp)	Disruption type [a]	Rv No. [b]	Notes [c]
RD21	1	C terminal merged	Rv0063	FAD/FMN-containing dehydrogenases [Energy production and conversion]; Region: GlcD; COG0277
RD318	1	C terminal truncated	Rv0888	Exonuclease-Endonuclease-Phosphatase (EEP) domain superfamily; Region: EEP; cl00490
RD470	1	N terminal truncated	Rv1225c	Predicted sugar phosphatases of the HAD superfamily [Carbohydrate transport and metabolism]; Region: NagD; COG0647
RD476	1	C terminal truncated	Rv1258c	H+ Antiporter protein; Region: 2A0121; TIGR00900
RD1034	1	C terminal truncated	Rv2264c	unkown
RD1419	1	cobB and cysG merged	Rv2848c; Rv2847c	Siroheme synthase (precorrin-2 oxidase/ferrochelatase domain) [Coenzyme metabolism]; Region: CysG; COG1648; cobyrinic acid a,c-diamide synthase; Validated; Region: PRK01077
RD3699	1	fragmented	Rv3483c	LppP/LprE lipoprotein; Region: Lipoprotein_21; pfam14041
RD3989-RD3990	1 1	two genes merged	Rv3829c; Rv3829c	mycofactocin system transcriptional regulator; Region: mycofact_TetR; TIGR03968 Phytoene dehydrogenase and related proteins [Secondary metabolites biosynthesis, transport, and catabolism]; Region: COG1233

The three large RDs (RD29, RD666-691-693) are also related to IS6110 elements and PE/PPE families, and significantly, involving several genes of functions related to the pathogenicity of *M. tubercolusis* (such as, *cut1*, *plcD*, and transporter genes), although some other genes are hypothetical (Figure 5.4, 4.5 and Table 5.3). It is found that most of them are small in size (only 4 indels of length > 20bp), causing a N-terminal truncation of nine genes disrupting 9 genes into two parts. The types of disruption and the functional annotations on these genes are shown in Table 5.4.

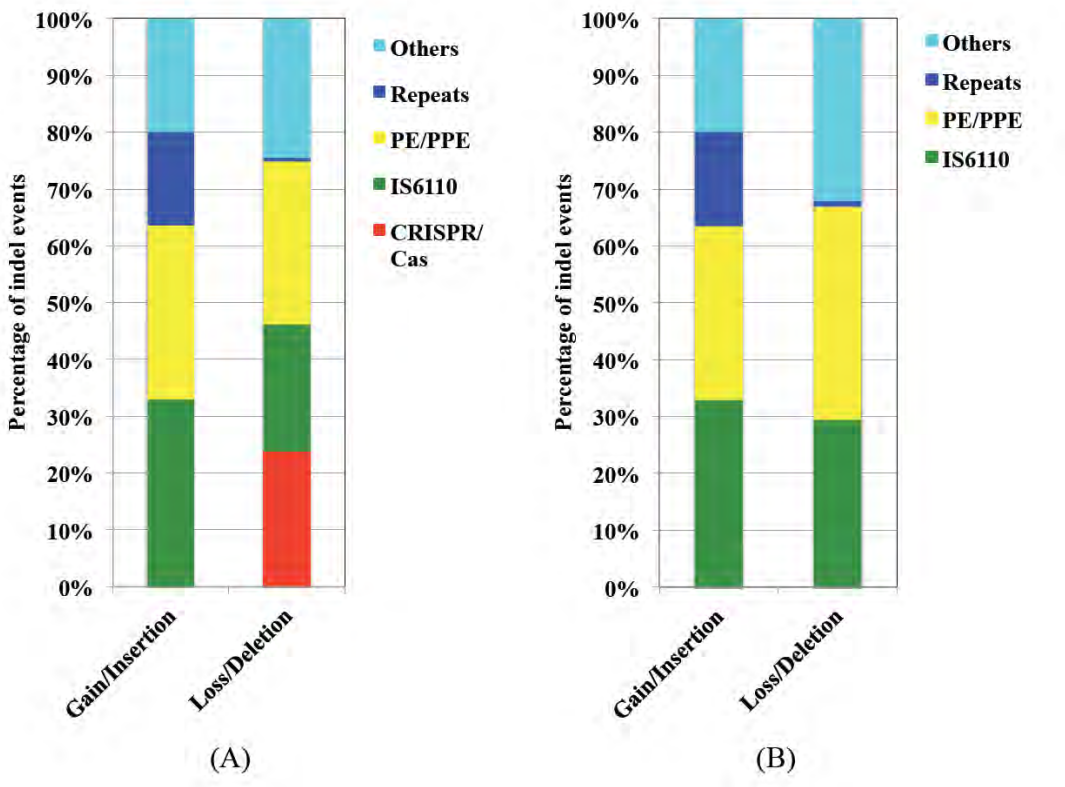

Figure 5.3: Proportion of different genetic mechanisms contributing to the gain and loss of indels in Beijing ancestor. (A). CRISPR/Cas is included, which involves indel losses in the Beijing ancestor, but not associated with indel gains; (B). CRISPR/Cas is excluded to evaluate the different contribution to gain and loss of indels by different genetic mechanisms, which are PE/PPE, IS6110, repeats and small indel polymorphisms within coding genes (others).

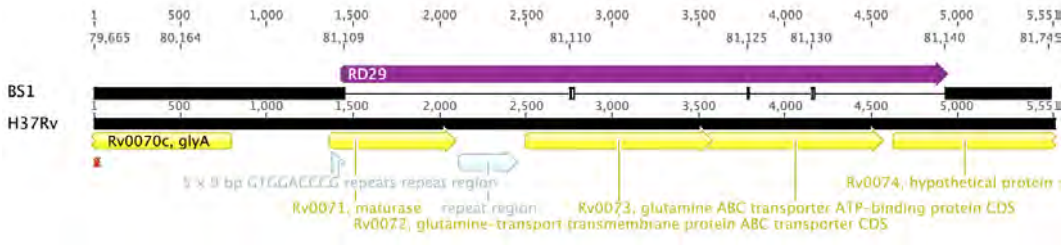

Figure 5.4: RD29, a large RD deleted in the Beijing ancestor. Shown is an alignment of BS1 (from base 79,665 to 81,745) and H37Rv (from base 78,113 to 83,663). Rv0071, Rv0072 and Rv0073 are deleted in BS1. And Rv0074 has it N terminal truncated in BS1. This RD also includes two small indels defined with frame-shifts (RD30-31).

Figure 5.5: RD666-691-693, a large RD deleted in the Beijing ancestor. Shown is an alignment of BS1 (from base 1,979,050 to 1,984,026) and H37Rv (from base 1,985,849 to 2,001,454). Rv1755c, Rv1758, Rv1759, Rv1760, Rv1761, Rv1762 and Rv1765c are deleted in BS1, in addition to the deleted IS6110 genes.

Table 5.3: Functional annotations of the detected genes in RD29 and RD666, 691-693. The annotations are taken from the H37Rv genome from the Genbank database.

Rv No.	Name	Notes
Rv0071	maturase	Retron-type reverse transcriptase [DNA replication, recombination, and repair]; Region: COG3344
Rv0072	glutamine-transport transmembrane protein ABC transporter	Thought to be involved in active transport of glutamine across the membrane (import). Responsible for the translocation of the substrate across the membrane.
Rv0073	glutamine ABC transporter ATP-binding protein	Thought to be involved in active transport of glutamine across the membrane (import). Responsible for the translocation of the substrate across the membrane.
Rv0074	hypothetical protein	Imidazolonepropionase and related amidohydrolases [secondary metabolites biosynthesis, transport, and catabolism]; Region: huti; COG1228
Rv1754c	hypothetical protein	Domain of unknown function (duf4185); Region: duf4185; pfam13810
Rv1755c	plcD	Hydrolyzes sphingomyelin in addition to phosphatidylcholine. Probable virulence factor implicated in the pathogenesis of *M. tuberculosis* at the level of intracellular survival, by the alteration of cell signaling events or by direct cytotoxicity.
Rv1758	cut1	Hydrolysis of cutin.
Rv1759c	wag22	PE-PGRS family protein
Rv1760	hypothetical protein	Wax ester synthase-like acyl-coa acyltransferase domain; Region: wes_acyltransf; pfam03007
Rv1761c	hypothetical protein	
Rv1762c	hypothetical protein	Uncharacterized conserved protein [function unknown]; Region: COG0393
Rv1765c	hypothetical protein	HNH nucleases; HNH endonuclease signature which is found in viral, prokaryotic, and eukaryotic proteins. Including members of the large group of homing endonucleases, yeast intron 1 protein, muts, as well as bacterial colicins, pyocins; Region: hnhc; CD00085

Table 5.4: Disrupted genes by frame-shift deletion in the Beijing ancestor. The start and end position of the deletions are referred to the H37Rv genome. a, Disruption type describes the results of frame-shifts of the predicted CDS compared with H37Rv; b, the corresponding gene in H37Rv; c, notes of biological functions taken from the H37Rv annotations.

Indel ID	Start	End	Size	Disruption type [a]	Name	Rv No. [b]	Notes [c]
RD17	49691	49692	2	N terminal truncated	hydrolase	Rv0045c	Alpha/beta hydrolase family; Region: Abhydrolase_6; pfam12697
RD173	485811	485811	1	N terminal truncated	pks6	Rv0405	POLYKETIDE SYNTHASE POSSIBLY INVOLVED IN LIPID SYNTHESIS.
RD354	1168010	1168010	1	N terminal truncated into a new gene	Nucleotidyl transferase of unknown function	Rv1045	Nucleotidyl transferase of unknown function (DUF1814); Region: DUF1814; pfam08843
RD440	1252054	1252054	1	Fragmented into two parts	hypothetical	Rv1128c	Domain of unknown function (DUF222); Region: DUF222; pfam02720
RD614	1756359	1756359	1	Fragmented into two parts	plsB1	Rv1551	glycerol-3-phosphate acyltransferase; Reviewed; Region: PRK11915
RD650 **RD651**	1955761 1955915	1955763 1955933	3 19	Fragmented into two parts	penicillinbinding protein	Rv1730c	Beta-lactamase class C and other penicillin binding proteins [Defense mechanisms]; Region: AmpC; COG1680
RD718	1998226	1998599	374	C terminal truncated	hypothetical	Rv1765c	HNH nucleases; HNH endonuclease signature which is found in viral, prokaryotic, and eukaryotic proteins. The alignment includes members of the

							large group of homing endonucleases, yeast intron 1 protein, MutS, as well as bacterial colicins, pyocins, and...; Region: HNHc; cd00085
RD727	2009291	2009291	1	N terminal truncated	hypothetical	Rv1775	
RD761	2153726	2153739	14	N terminal truncated	hypothetical	Rv1907c	Domain of unknown function (DUF4262); Region: DUF4262; pfam14081
RD901	2241032	2241032	1	Fragmented into two parts	ctpF	Rv1997	Cation transport ATPase [Inorganic ion transport and metabolism]; Region: MgtA; COG0474
RD910	2273734	2273734	1	Fragmented into two parts	histidine kinase response regulator; signal transduction	Rv2027c	Signal transduction histidine kinase [Signal transduction mechanisms]
RD1004	2406843	2406843	1	C terminal 2AAs modified	hypothetical	Rv2148c	Predicted enzyme with a TIM-barrel fold [General function prediction only]; Region: COG0325
RD1033	2535432	2536141	710	N terminal truncation of both Rv2262c and Rv2263	Apolipoprotein N-acyltransferase; shortchain dehydrogenase	Rv2262c; Rv2263	Oxidoreduction; Apolipoprotein N-acyltransferase [Cell envelope biogenesis, outer membrane]; Region: Lnt; COG0815;
RD1050	2614910	2614910	1	Fragmented into two parts	mmpL9	Rv2339	Predicted drug exporters of the RND superfamily [General function prediction only]; Region: COG2409

RD1202	2729619	2729832	214	Fragmented into two parts	transmembrane protein	Rv2434c	Small-conductance mechanosensitive channel [Cell envelope biogenesis, outer membrane]; Region: MscS; COG0668; cAMP-binding proteins - catabolite gene activator and regulatory subunit of cAMP-dependent protein kinases [Signal transduction mechanisms]; Region: Crp; COG0664
RD1203	2734482	2734482	1	N terminal truncated	Putative protein-S-isoprenylcysteine methyltransferase	Rv2437	Putative protein-S-isoprenylcysteine methyltransferase [Posttranslational modification, protein turnover, chaperones]; Region: STE14; COG2020
RD1227	2851162	2851166	5	N terminal truncated	hypothetical	Rv2526	
RD1232	2867881	2867881	1	N terminal truncated	Transcription regulator of the Arc/MetJ class	Rv2545	Transcription regulator of the Arc/MetJ class [Transcription]; Region: COG5450
RD3979	4231860	4231948	89	Fragmented into two parts	hypothetical	Rv3785	dTDP-glucose 4,6-dehydratase
RD3996	4322040	4322040	1	C terminal elongated	hypothetical	Rv3847	
RD4028	4379045	4379045	1	Fragmented into two parts	type VII secretion protein	Rv3894c	type VII secretion protein EccCa; Region: T7SS_EccC_a; TIGR03924; type VII secretion protein EccCb; Region: T7SS_EccC_b; TIGR03925

5.3.3 Evolution of IS6110 in the formation of MTB lineages

There are many indel events that can be attributed to the transposition of IS elements and homologous recombination mediated by them, especially the IS6110 elements (Figure 5.3). Reconstruction of the IS6110 elements along the evolution of the 15 MTB genomes shows that there are expansions of these elements in different MTB lineages (Figure 5.1). Only three IS6110s are predicted to present in the early common ancestor. And the bovis and BCG group have also very few number of IS6110 (1 or 2). After the divergence from the early common ancestor, the Beijing ancestor acquires a large number of IS6110 (17), which is the largest among all the observed lineages. Gain of the IS6110s still occurs in different strains within the Beijing family. This result indicates that expansion of IS6110s is recent evolutionary activities in the lineage level, and is still ongoing within different MTB genotypes. This is also supported by reconstruction results of the non-Beijing group, in which early ancestral nodes are inferred to have a limited number of IS6110s (from 3 to 6). But, in the more recent divergence of different lineages in this group (Figure 5.1), this number grows rapidly (15 in H37, 17 in F11 and 12 in KZN), except for CDC1551 (only 4). Furthermore, although some IS6110 are insertions in intergenic regions, but most of them cause gene disruption (Table 5.5) and therefore have phenotype influences. Careful examination shows that the IS6110s are usually associated with PE/PPE insertions or deletions; and together, they can result in large RDs involving important pathogenicity genes (Figure 5.5, Table 5.5).

Table 5.5: Occurrence of the IS6110s in the Beijing family and their influence on functional genes. The start and end position are referred to the H37Rv genome. a, occurrence of the IS6110s in the MTB genomes, BJ stands for occurrence in all five genomes of Beijing family; b, gene locus number in the H37Rv genome; c, annotation from the H37Rv genome.

Indel ID	Start	End	Occurrence [a]	Rv. No. [b]	Notes [c]
RD2	1594	1594	BJ	-	intergenic
RD267	889021	890375	BJ, Ra, Rv	-	intergenic
RD332	1076075	1076075	BT2	Rv0963c	hypothetical; Alpha/beta hydrolase; Region: Abhydrolase_8; pfam06259
RD441	1262959	1262959	BJ	Rv1135c	PPE16
RD509	1531041	1531041	BT2	-	hypothetical
RD515	1543968	1543968	BJ	Rv1371	hypothetical; Fatty acid desaturase [Lipid metabolism]; Region: DesA; COG3239
RD600	1657015	1657015	BJ	Rv1469	ctpD; Cation-transporting atpase; possibly catalyzes the transport of a cation (possibly cadmium) with the hydrolyse of atp
RD674	1987702	1988952	BJ, Ra, Rv, CDC1551, F11	-	cut1
RD732	2040443	2040443	CCDC5180	Rv1800	PPE28
RD733	2041736	2041736	CCDC5079	-	PPE
RD882	2165902	2165902	BJ but not BT2	Rv1917c	PPE34
RD905	2263624	2263624	BJ	Rv2016	hypothetical
RD909	2270812	2270812	CCDC5079	Rv2025c	Cobaltzinccadmium resistance; cation efflux system protein
RD950	2365415	2366766	BJ, Ra, Rv	-	PPE
RD1058	2634083	2634083	BT1, KZN	Rv2352c	PPE71
RD1271	3069959	3069959	CCDC5079	Rv2756c	hsdM; implicated in methylation of dna. component of type i restriction/modification system.
RD1281	3115053	3115053	BT2	Rv2808	hypothetical
RD1492	3378553	3378553	BS1, CCDC5180	-	PPE
RD1495	3335583	3335583	-	-	

Table 5.5: Occurrence of the IS6110s in the Beijing family and their influence on functional genes (continued)

RD1495	**3379024**	**3379024**	**BJ**	**-**	**intergenic**
RD1513	3493910	3493910	BJ	-	hypothetical
RD1527	3547341	3547341	BT2, CCDC5180	-	intergenic
RD1528	3549196	3549196	BJ	-	hypothetical
RD3636	3797825	3797825	BJ	Rv3383c	idsB; involved in biosynthesis of membrane ether-linked lipids.
RD3678	3844677	3844677	BJ	-	hypothetical

5.3.4 Repeat patterns of the PPE34 and PPE24 genes

Two PPE genes, Rv1917c (PPE34, 1,459aa) and Rv1753c (PPE24, 1,053aa), are identified with typical tandem repeats, among the 165 PE/PPE encoded in the H37Rv genome. There are four blocks of tandem repeats in PPE34 and two blocks in PPE24 (Figure 5.6). These repeat units, ranging from 23aa to 26aa in protein sequence, repeat in different times and are bounded by conserved sequences in different strains. Significantly, there is no singleton of repeat observed, and there are at least two units in each block even in the most reductive form. As the most repeated unit, unit 4 in PPE34 (23aa, Figure 5.6) consists of 322 amino acids (ZP_03428743 of *M. tuberculosis* EAS054, 1,717aa). And the largest PPE34 protein is from *M. bovis* AF2122/97, of length 1,910aa.

Association between the repeat profiles and different *M. tuberculosis* lineages is evident in the context of their phylogeny (Figure 5.7). These profiles are characteristic of different MTB genomes, except for strains of the same genotype groups (KZN, H37, and BCG), without considering those PPE genes of incomplete sequence. From Figure 5.7, it can be seen that the “Europe & Americas” lineage has the most reductive repeating of the unit1 and unit2 in PPE34, but the “Animal strains” lineage repeats the most. The repeat

times even distinguish between different sub-lineages of the "Europe & Americas" group, evidencing by the unit1 and unit2 of PPE24. Overall, all the lineages can be clearly distinguished by their PPE repeat profiles, and the resolution even remains for strains of the same lineage. An unusual case is that the PPE34 gene is disrupted by an insertion of the IS6110 element, with incomplete N-terminal (CCDC5180, S1 and CCDC5079) or a loss of the C-terminal (R2). For other strains, it is unclear whether the incomplete PPE34/24 genes are due to gene interruption or unfinished sequencing.

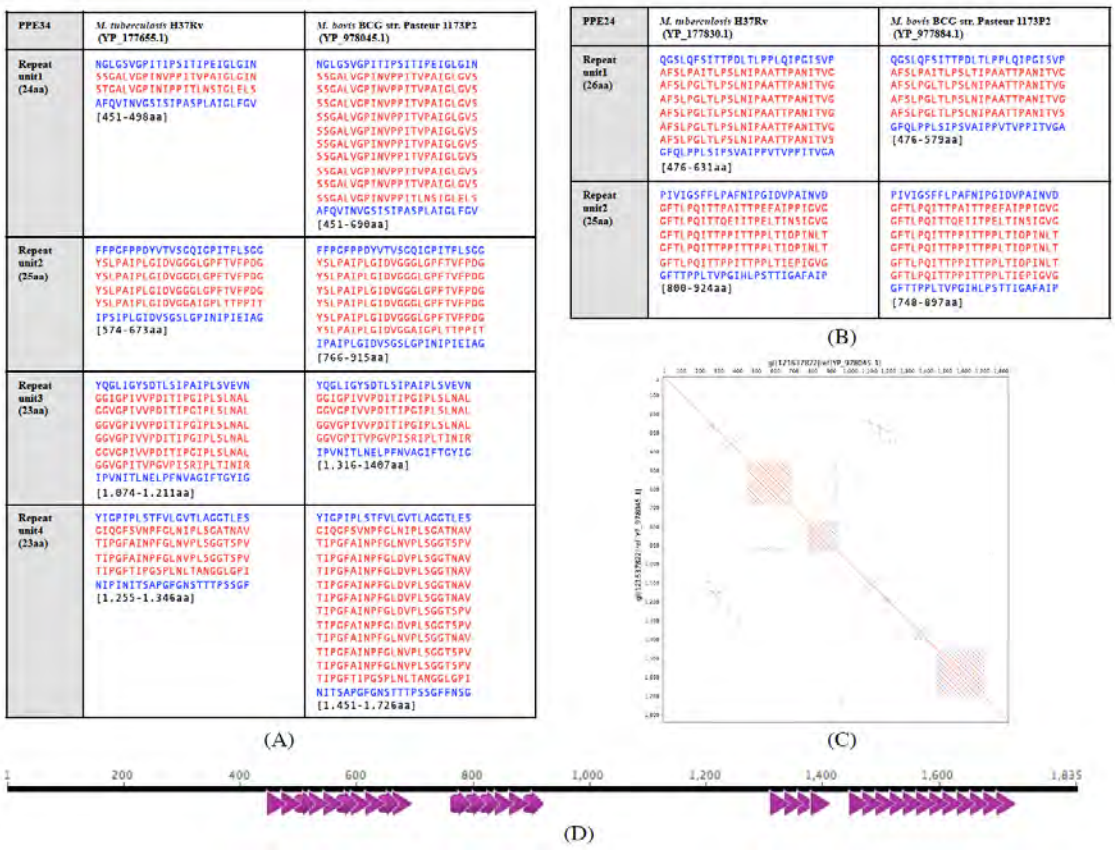

Figure 5.6: Repeat structure of the PPE24 and PPE34 genes, including repeat units and boundary sequences. PPE34 (A) and PPE24 (B) genes from H37Rv and BCG str. Pasteur 1173P2 were used to illustrate the structures. A bock of repeats comprises tandem repeat units (red) and boundary sequences (blue). Below the blocks, positions of the repeats in the particular genes were given inside square brackets. Dot-plot (C) and gene structure (D) were also presented as illustrations with the example of PPE34 gene from BCG str. Pasteur 1173P2.

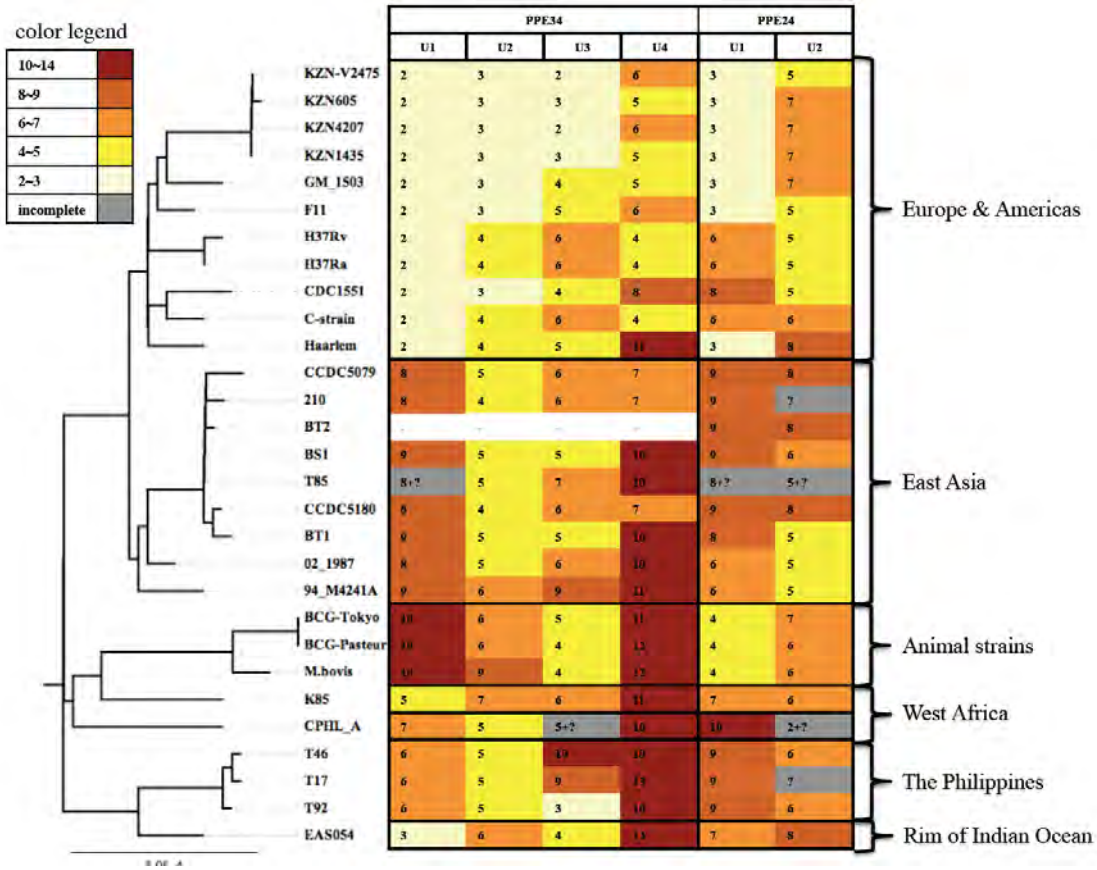

Figure 5.7: Repeat patterns of the PPE34 and PPE24 genes in the MTB genomes, and its association with MTB lineages. Phylogeny and lineages was adopted from Hershberg et al. (Hershberg, Lipatov et al. 2008). Fragmented genes of incomplete information are colored in grey, but the number of repeats was still counted when repeat units found therein. A question mark was used to indicate the absence of boundary sequence for an intact block of repeats.

5.4 Discussion

5.4.1 Local parsimony ancestral state reconstruction with genome data

A large number of indels of size ranging from 1bp to 11,345bp were identified from whole genome alignment of the 15 complete genomes of different *Mycobacterium tuberculosis* stains, comprising a Beijing group, a non-Beijing group and a *M. bovis* group. Each identified indel was mapped onto the phylogenetic tree reflecting the evolutionary history of these genomes by a local parsimony ancestral state reconstruction algorithm. This algorithm has been developed with an assumption that local parsimony prefers less independent parallel changes in the evolution of species. In contrast to evolution models for DNA mutation, in which base substitution involves reversion and conversion equally, this local parsimony model is more suitable for evolutionary events associated with species and lineage divergence. In fact, previous analysis of large RDs among a large set of clinical MTB isolates has shown that indels did not occur independently among the strains, but strongly associated with strain evolution in different lineages (Brosch, Gordon et al. 2002). To further validate the local parsimony model and its application to ancestral reconstruction of indels in MTB genomes, the known RDs were also mapped to the reference tree using this algorithm (Figure 5.1). Reconstruction of all these RDs is consistent with the proposed evolutionary scenario of MTB lineage evolution, in which RDs that occurred at the divergence of different lineages or strains were correctly reconstructed with this algorithm. But there is an ambiguous case for the Beijing specific RD (Table 5.1), the NBjD1, which also presents in the *M. canettii* genome with only a few SNPs. The algorithm infers that this RD has been lost from the lowest common ancestor of MTB, given the absence of this RD from both the non-Beijing group and the *M. bovis* group. And therefore, a regaining of this RD occurred in the Beijing group, which might be resulted from a homologous recombination with a distant MTB species, like *M. canettii*. Without evidence for such a

recombination process or other supports, this inference might not seem likely. There is also an alternative explanation for this RD, which involves two independent parallel losses in the ancestors of the non-Beijing group and the *M. bovis* group. Because there is no obvious evidence for homologous recombination within the MTB lineages (Veyrier, Dufort et al. 2011), more data is needed to rule out the alternative explanation of two losses. Such kind of data may come from strains that are more closely related to the Beijing group, such as the CAS family, or from strains that are more ancestral to examine the possibility of presence of this RD in the ancestor of MTB.

5.4.2 Reductive evolution of the MTB genomes

Reductive evolution of MTB genomes at species and lineage level is evident from the analysis of indels at a whole genome scale, as indicated by the observed more losses of indels than gains in their divergence (Figure 5.1). But this trend is not significant for evolution of different strains within the same lineage. In contrast, the BCG strains show significantly more gains of indels than losses. This contradiction may be due to the *in vitro* evolution of BCG strains as vaccines manipulated in laboratory (Mostowy, Tsolaki et al. 2003). Furthermore, reductive evolution that has been suggested for the adaptation of different MTB lineages associated with different human populations (Hershberg, Lipatov et al. 2008) seems to act at a macroevolution scale. At microevolution scale, such as recent emerging of drug resistant strains and spreading of different strains of same genotype, other evolutionary forces take over (e.g., single nucleotide polymorphisms or reversal indels for antigenic variation) (Parwati, van Crevel et al. 2010). Taken together, reconstruction of ancestral events of gain/loss of indel during the evolution and divergence of MTB species is useful for uncovering different evolutionary forces that are responsible for different evolutionary stages. Insights can be gained from the ancestral reconstruction to understand how changes in MTB genomes gave rise to different adapted phenotypes. One requirement

for this reconstruction analysis is to obtain indels from accurate whole genome alignment, which is feasible for the conserved MTB genomes. And incorporating more related strains into the analysis is also valuable to have a more precise picture of MTB species evolution.

5.4.3 Effects of gain and loss indels in the evolution of the Beijing family

From the reconstruction analysis, it is suggested that the Beijing family arose from a combination of macroevolution events of indels (186 losses and 108 gains, Figure 5.1). To further characterize 'Beijing group specific' indels that might underlie the formation of the Beijing family, these indels were compared with the non-Beijing group. This resulted in two classes of indels: newly gained indels specific to the Beijing group (108 indels) and ancestral indels loss from the Beijing group but present in the non-Beijing group (186 indels). Examination of the possible genetic mechanisms associated with these indels shows that the most dominant events are associated with PE/PPE gene and IS6110 elements, following by repeats and other indels (Figure 5.3). A group of CRISPR/Cas associated indels shows a strictly reductive change in the Beijing family, in which only losses but not gains were observed. CRISPR/Cas presents in many bacteria and archaea genomes and plays an important role in phage-bacteria interaction, conferring acquired immunity against invading nucleic acids in bacteria (Bondy-Denomy, Pawluk et al. 2012; He, Fan et al. 2012). The reductive change of this system in the Beijing group implies that its function is not fully maintained in the lineage, which is surviving in a strictly human associated environment. Reductive evolution in the ancestor of the Beijing group is also evident from two large indels that are deleted from the group: RD29 (Figure 5.4) and RD666-691-693 (Figure 5.5). There are three genes deleted (encoding maturase, and two glutamine ABC transporter proteins) and one hypothetical gene truncated in RD29. In RD666-691-693, eight genes have been totally deleted, including the *plcD*, *cut1*, a PE-PGRS gene and other hypothetical genes, together with an IS6110. Their functional implications are remained to be examined,

but these two RDs provide useful markers for identification of the Beijing genotype. Furthermore, IS6110 elements are found to involve in both of these RDs, suggesting that deletion of the RDs may be due to activities of IS6110 in the ancestor of Beijing family.

Both 'Beijing group specific' gain and loss of indels are shown to have disrupted coding genes (Table 5.2, 5.3 and 5.4). Most of the disruptions are due to ORF fame-shift caused by small indels, resulting N-terminal or C-terminal truncation or fragmentation of the encoded proteins. Pseudogenization of the disrupted genes relative to the non-Beijing group represents another force in the reductive evolution of the Beijing group, in addition to deletions of large RDs encompassing several genes that are mentioned above. This kind of inactivation of individual gene by small indels, probably resulted from replication errors, may complement the more 'systematic ways' of generating indels by repeats or IS6110 elements (Figure 5.3). And most of the pseudogenes have functional implications, in contrast to intergenic indels from structural repeats. The most significant cases include the Rv3894c gene (encoding the type VII secretion protein EccC2, a reported virulence factor in MTB (Chen, Xiong et al. 2012)), which has been fragmented into two parts in the Beijing group, the Rv0063 gene (encoding an oxidoreductase, involved in the twin-arginine translocation (Tat) pathway in MTB pathogenesis (McDonough, McCann et al. 2008)), the Rv1258c gene (encoding the Tap efflux pump, disruption of this gene in BCG strains leading to extensive change in gene expression patterns during stationary phase (Ramon-Garcia, Mick et al. 2012)), the Rv3483c gene (encoding a membrane associated protein, specific to *Mycobacterium* ESX-1 secretion system (Xu, Laine et al. 2007)), the Rv0045c gene (encoding a novel esterase, involved in ester/lipid metabolism (Zheng, Guo et al. 2011)), the Rv0405 gene (encoding the membrane bound polyketide synthase pks6, involved in MTB pathogenesis (Hisert, Kirksey et al. 2004)), the Rv2027c gene (encoding a histidine kinase response regulator, involved in the two-component signal transduction systems MprAB and DosRS-DosT (DevRS-Rv2027c) in MTB pathogenesis (Bretl, He et al.

2012)), the Rv2339 gene (encoding the *Mycobacterium* membrane protein mmpL9, involved in the transport of lipids (Domenech, Reed et al. 2005)), and the Rv2434c gene (encoding a CRP and cNMP binding protein for MTB signaling (Agarwal and Bishai 2009)). There are also many other disrupted genes that are only hypothetical and need more efforts to elucidate their molecular functions in *Mycobacterium tuberculosis*. Therefore, pseudogenization of these genes, relative to the non-Beijing group, are likely to have contributed to the formation of the Beijing genotype by fine-tuning its pathogenicity for host adaptation, and their exact roles need further experimental examination.

5.4.4 The roles of IS6110 elements in the formation of MTB lineages

It has been hypothesized that transposable activity of IS6110 elements can systematically invoke genotype diversity in MTB genomes, conferring selective advantage for species or lineage adaptation (McEvoy, Falmer et al. 2007). Reconstruction analysis of IS6110 elements from the 15 MTB genomes appears to support this hypothesis (Figure 5.1). It can be seen that expansions of IS6110 elements have occurred in the formation of different MTB lineages, especially in the Beijing group, which has a highest number of IS6110 elements (17 IS6110 in its common ancestor). The exception is the *M. bovis*-BCG group, which seems to be free from active IS6110 transpositions. In addition to those IS6110s that are ancestral and conserved in whole species or lineages, there are also relatively new IS6110s that are continuing to introduce genotype variation among strains of the same lineage (e.g., different strains in the Beijing family). Furthermore, activity of IS elements appears not to happen randomly in the whole MTB genome. For example, a hotspot of IS elements has been identified (Figure 5.2), involving PPE genes, IS1532, IS1540 and IS6110. This region distinguishes between the Beijing group (insertion of a IS6110) and the non-Beijing group (deletion of a segment of two PPE genes) (McEvoy, Warren et al. 2009; Kim, Nahid et al. 2010). Genetic variations introduced by IS6110 can translate into phenotype

variations by mediating deletion of large RDs (Figure 5.4 and 5.5) or disrupting individual genes (Table 5.5). Although exact conditions for invoking large-scale IS6110 activities in MTB are still unclear (McEvoy, Falmer et al. 2007), transpositions of IS6110 undoubtedly represent an import force for reshaping the MTB genome for adaptation, especially in a reductive manner; and this is the reason for its use as an effective molecular maker in epidemiology study. From the perspective of the Beijing family, IS6110s should have contributed to its formation and the IS6110-affected gene identified in this study may help to understand the molecular mechanisms contributed to the virulence of this family.

5.4.5 Repetitive structures of the PPE24 and PPE34 genes

Two PPE genes, PPE24 (Rv1753c) and PPE34 (Rv1917c) have been examined in detail to characterize their repeat patterns. Previously, several VNTRs (QUB-11a, QUB-11b and ETR-A from PPE34; QUB-18 from PPE24) have been designed from these two genes to be used as epidemiological makers (Skuce, McCorry et al. 2002). Further characterization of repeat motifs in their protein sequences shows that, in fact, there are four tandem repeat units in PPE34 and two tandem repeat units in PPE24 (Figure 5.6). Comparing the number of repeats of these units between different genomes shows a correlation with MTB lineages (Figure 5.7). Given the probable biological role of the PPE34 gene as involving in host immune responses, these amino acid units may have structural and functional meanings, likely to be responsible for antigenic variation and exposed partly on cell surface (Sampson, Lukey et al. 2001). In fact, these units have been maintained intactly in tandem repeats and at least two repeats remain in the most reductive form (Figure 5.7). This sequence regularity may result from structural constraints and therefore to be maintained. Considering the same regular repeats occurred in PPE24, as in PPE34, it is reasonable to suggest that PPE24 could have similar function as PPE34, probably encoding antigenic variation. From an evolutionary viewpoint, reversion and conversion of repeat numbers are obvious in the

repeat profile of the two genes, but a combination of the six units still have enough resolution to distinguish different genomes, therefore can be used for epidemiology purpose. It is noteworthy that the PPE34 gene has been disrupted in the BT2 genome by insertion of an IS6110, for which functional implication is unknown (McEvoy, Warren et al. 2009).

5.5 Conclusion

In this study, complete genome sequencing of five MTB Beijing genotype strains, together with ten other complete MTB genomes of non-Beijing genotype, allows us to reconstruct the evolution of the Beijing family. By using a local parsimony ancestral state reconstruction method and whole genome alignment, genome-wide indels are identified and reconstructed for analyzing their evolution in the MTB species. The result shows that three major genetic mechanisms have contributed to the reductive evolution for the formation of the Beijing family, including repeats-related indels, PE/PPE gene variation, and IS6110 transposition and recombination. In addition, gene pseudogenization by other small indels also plays an important role, as demonstrated by the affected genes involved in MTB pathogenesis. Further experimental examination of the functional implications of these disrupted genes may help to understand the adaptation and virulence of the Beijing family. Most significantly, reconstruction of the evolution of IS6110 elements in these fifteen MTB genomes shows expansion of copies of IS6110 elements in the formation of different MTB lineages, except for the *M. bovis* and BCG group. This observation supports the hypothesis that transposition activities of IS6110s may associate with adaption of MTB strains. And it is shown that activities of IS6110s are related to two large RDs that have been deleted in the Beijing family. Insertion of IS6110s also disrupts gene in the Beijing family, even within different strains of this family, at a microevolution scale. In conclusion, reductive evolution at a macroevolution scale has shaped the formation of the Beijing family, for which

genome-wide deletions of large RDs and disruption of individual genes are identified in this study. Although this trend of genomic reduction is not significant among strains within the family, IS6110 elements appears to continue to introduce variation between different strains at a microevolution scale.

Chapter 6

A phyletic model for pangenome clustering and its application to the *Mycobacterium* genus

6.1 Introduction

A phyletic pattern or profile of orthologous genes is their pattern of presence or absence in genomes of different organisms (Tatusov, Koonin et al. 1997; Gaasterland and Ragan 1998). Analysis of phyletic patterns of a group of different gene families can help to determine their coevolutionary relationships and infer their functional linkages (Pellegrini, Marcotte et al. 1999). For instance, gene co-occurrence in many genomes of different phylogenetic distance has strong implication in their coevolution constrained by functional coupling (Kensche, van Noort et al. 2008). The coevolutionary relationship of gene families is important in comparative genomics analysis. First, identification of a cluster of co-evolving genes can help to infer biological functions of its members when some of them have already had their functions elucidated (Cunningham, Lafond et al. 2000). Second, such cluster of genes can help to compare genomes of different organisms to understand their association with phenotype of the organisms (Li, Gerdes et al. 2004). In this sense, the most typical co-evolving clusters are genome islands, which usually carry functional related genes for species adaptation, such as virulence factors in some microbial pathogens (Dobrindt, Hochhut et al. 2004). Furthermore, analysis of phyletic patterns plays important roles in the field of phylogenomics, which uses whole genome information to infer the evolutionary relationship between different species (Delsuc, Brinkmann et al. 2005).

In a pangenome data, it has been shown that the gene repertoire of a species can be classified into different groups, including core-genes (presenting in all the strains) and dispensable genes (presenting only in a subset of strains) (Medini, Donati et al. 2005; Lapierre and Gogarten 2009). Therefore, different genes may have different phyletic patterns in a pangenome. Usually, information is carried in such patterns for identifying genes responsible for strain phenotype variation and understanding their evolution for strain divergence. Particularly, in a supragenome model, different gene families have been shown to have different frequencies in the genomes of different strains (Hogg, Hu et al. 2007). This model is intimately related to phyletic patterns of the gene families in a pangenome. But previous studies focused on estimating the pangenome size of microbial species, with little attention to analyzing gene coevolution in the supragenome model (Tettelin, Riley et al. 2008; Lapierre and Gogarten 2009; Snipen, Almoy et al. 2009; Baumdicker, Hess et al. 2012). Some efforts have also been paid to infer gene gain and loss during the evolution of the pangenomes of some species, to identify genes that are related to different ecotypes or phenotypes (Kettler, Martiny et al. 2007; Lefebure and Stanhope 2007). But these studies were based on the analyzing of individual genes, instead of identifying co-evolving gene clusters.

In this study, I use phyletic patterns of gene families to analyze gene coevolution in the pangenome of the genus *Mycobacterium*. First, an association score that characterizes the degree of coevolution of a pair of genes is proposed, based on the supragenome model. Second, the graph-based clustering method, MCL (Dongen 2000; Enright, Van Dongen et al. 2002), is used to identify co-evolving clusters of genes from all pairwise coevolutionary relationship descried in the first step. It is shown that the phyletic model can predict the protein association network in the STRING-COG database (Szklarczyk, Franceschini et al. 2011), using the pangenome data of the whole bacteria domain. Furthermore, clustering the

pangenome of the genus *Mycobacterium* identifies clusters of genes that have contributed to the evolution of the genus and also the *Mycobacterium tuberculosis* species.

6.2 Materials and methods

6.2.1 The pangenome data of the genus *Mycobacterium*

The pangenome data was obtained from the STRING database (version 9.0, Table 6.1). The information of COG mapping was used as the primary data in this study, which is the information of presence of the COGs in the genomes in the database. To analyze the pangenome of the genus *Mycobacterium*, all mycobacterial genomes were included, in which there were 13 non-tuberculosis mycobacterial genome and 8 genomes of *M. tuberculosis* complex. Also included were other 14 genomes from the out-group species. These species are closely related to the genus, together defining the CMN cluster within the phylum *Actinobacteria* (Ventura, Canchaya et al. 2007).

Table 6.1: The pangenome data of the genus *Mycobacterium* and its closely related out-group species. a, the short taxon code used in this study; b, NCBI taxonomy id for the species; c, type of COG identification used in the STRING database.

Taxon code [a]	Taxon id [b]	STRING type [c]	Official NCBI name
C.aurimucosum	169292	periphery	*Corynebacterium aurimucosum* ATCC 700975
C.diphtheriae	257309	periphery	*Corynebacterium diphtheriae*
C.efficiens	196164	core	*Corynebacterium efficiens* YS-314
C.glutamicum.B	196627	periphery	*Corynebacterium glutamicum* ATCC 13032 Bielefeld
C.glutamicum.R	340322	periphery	*Corynebacterium glutamicum* R
C.jeikeium	306537	periphery	*Corynebacterium jeikeium* K411
C.kroppenstedtii	645127	periphery	*Corynebacterium kroppenstedtii* DSM 44385
C.urealyticum	504474	periphery	*Corynebacterium urealyticum* DSM 7109

G.bronchialis	526226	core	*Gordonia bronchialis* DSM 43247
M.abscessus	561007	periphery	*Mycobacterium abscessus* ATCC 19977
M.avium.105	243243	periphery	*Mycobacterium avium* 104
M.avium.para	262316	periphery	*Mycobacterium avium paratuberculosis*
MTBC.bovis	233413	periphery	*Mycobacterium bovis*
MTBC.BCG.Pasteur	410289	periphery	*Mycobacterium bovis* BCG Pasteur 1173P2
MTBC.BCG.Tokyo	561275	periphery	*Mycobacterium bovis* BCG Tokyo 172
M.gilvum	350054	periphery	*Mycobacterium gilvum* PYR-GCK
M.JLS	164757	periphery	*Mycobacterium* JLS
M.KMS	189918	periphery	*Mycobacterium* KMS
M.leprae	272631	periphery	*Mycobacterium leprae*
M.leprae.Br	561304	periphery	*Mycobacterium leprae* Br4923
M.marinum	216594	periphery	*Mycobacterium marinum* M
M.MCS	164756	periphery	*Mycobacterium* MCS
M.smegmatis	246196	periphery	*Mycobacterium smegmatis* MC2 155
MTBC.CDC1551	83331	periphery	*Mycobacterium tuberculosis* CDC1551
MTBC.F11	336982	periphery	*Mycobacterium tuberculosis* F11
MTBC.H37Ra	419947	periphery	*Mycobacterium tuberculosis* H37Ra
MTBC.H37Rv	83332	core	*Mycobacterium tuberculosis* H37Rv
MTBC.KZN.1435	478434	periphery	*Mycobacterium tuberculosis* KZN 1435
M.ulcerans	362242	periphery	*Mycobacterium ulcerans* Agy99
M.vanbaalenii	350058	periphery	*Mycobacterium vanbaalenii* PYR-1
N.farcinica	247156	core	*Nocardia farcinica* IFM10152
N.JS614	196162	core	*Nocardioides* JS614
R.erythropolis	234621	periphery	*Rhodococcus erythropolis* PR4
R.jostii	101510	core	*Rhodococcus jostii* RHA1
R.opacus	632772	periphery	*Rhodococcus opacus* B4

6.2.2 A phyletic model based on supragenome gene frequency

Given a pangenome data, the model begins with a target gene to characterize its coevolutionary relationships with all the genes in the pangenome (Figure 6.1). First, based on the presence or absence of the target gene, all the individual genomes in the pangenome can be divided into two groups, '***pangenome 1***' (presence of the target gene) and

‘***pangenome 2***’ (absence of the target gene), as shown in Figure 6.1 (A). Second, within these two groups, supragenome models (f_1 and f_2) are calculated for each of them. In a supragenome model, occurrence frequency of the gene families in the pangenome is counted as the percentage of genomes that have the families. For example, core genes that are present in all the genomes have a frequency of 100%. Dispensable genes have frequencies less than 100%. And genes that are absent from all the genomes have a frequency of 0% (Figure 6.1 (B)).

The rationale for the above two steps comes from the observations as follow. The first is that the occurrence of a gene in an organism may contribute to its genotype. And the occurrence of a gene in all the organisms in a species strongly implies that this gene should have been associated with the evolution of the species and contributes to its phenotype. Therefore, bipartition of all the genomes in a pangenome into two sub-pangenomes based on the presence and absence of a target gene assumes that the two sub-pangenomes (‘***pangenome 1***’ and ‘***pangenome 2***’) may represent two sub-species with different phenotypes associated with the target gene. Next, for identifying genes that have coevolved with the target gene, the supragenome model provides useful information. Because the target gene is core in the ‘***pangenome 1***’ and totally absent from the ‘***pangenome 2***’, therefore genes that are strongly associated with the target gene should also be core in the ‘***pangenome 1***’ and very rare or absent in the ‘***pangenome 2***’. These two observations and assumptions provide the basis for the following steps of calculating the association score of gene coevolution in the pangenome.

To compare the ‘***supragenome 1***’ and ‘***supragenome 2***’ for characterizing true coevolutionary relationships, two more steps are taken. First, core genes in the ‘***pangenome 1***’ might be universal genes, therefore they also are core in the ‘***pangenome 2***’. To account for that, the gene frequency of the ‘*supragenome 1*’ is subtracted by that of the

'*supragenome 2*' ($f=f_1-f_2$), as shown in Figure 6.1 (C). This step produces a relative difference between the two supragenomes, which characterizes the degree of distinction of the two sub-pangenomes by the target gene. The final association score uses this relative difference to weight the '***supragenome 1***', in which the target gene is assumed to have contributed to its phenotype and co-evolved with other genes ($f=f_1*(f_1-f_2)$, as shown in Figure 6.1 (C)). After weighting, universal genes will not be significant associated with the target gene anymore. And for genes of similar frequency to the target gene, the rarer they are in the '*pangenome 2*', the more significant associated with the target gene in the '*pangenome 1*'. This association score can have a negative value, if $f_2>f_1$. This is reasonable for that some genes can replace other genes when the species got adapted to different environments. But this negative association doesn't imply the normal coevolutionary relationships of clusters of genes that are functionally related.

6.2.3 Pangenome clustering using the MCL algorithm

To cluster the genes in a pangenome into co-evolving clusters based on the association score defined above, a graph-based Markov clustering algorithm (the MCL algorithm) was used. A matrix of all phyletic vectors of all target genes in the pangenome was generated for this purpose. In this matrix, a row represents a vector of the association scores of a target gene, which includes all pairwise coevolutionary relationships between the target gene and all the genes in the '*pangenome 1*'. This matrix can then be used as input into the MCL algorithm for clustering analysis. The MCL algorithm implements random walks on a given graph using discrete time Markov process to identify clusters of its nodes, by considering different weighted edges that connects different nodes. This algorithm has been successfully applied to identify large gene families or orthologous groups of genes in comparative genomics analysis (Enright, Van Dongen et al. 2002; Li, Stoeckert et al. 2003; Chen, Mackey et al. 2006; Fischer, Brunk et al. 2011). In this algorithm, a flow on a graph is simulated by

random walks, involving two operations. The first operation is the expansion of the flow on the graph using matrix multiplication. This expansion operation tries to distribute the flow over the graph to identify potential neighbors to a node in a same cluster. But iterations of the expansion operation will eventually lead to distributing the flow over the whole graph and therefore resulting a single large cluster.

To modify this operation, the second operation of inflation is introduced to distribute more flow within clusters than between clusters. The inflation operation is implemented as entry powers followed by re-normalization of the matrix. Different entry powers (the '*I*' parameter) of this operation can have different effects on the clustering results. Usually, the inflation parameter '*I*' is set between 1.2 and 5, resulting from very coarse-grained to fine-grained clustering. Alternative iteration of these two operations is taken by the MCL algorithm for many steps until reaching the convergence of the matrix, which contains the information for defining clusters of nodes in the graph.

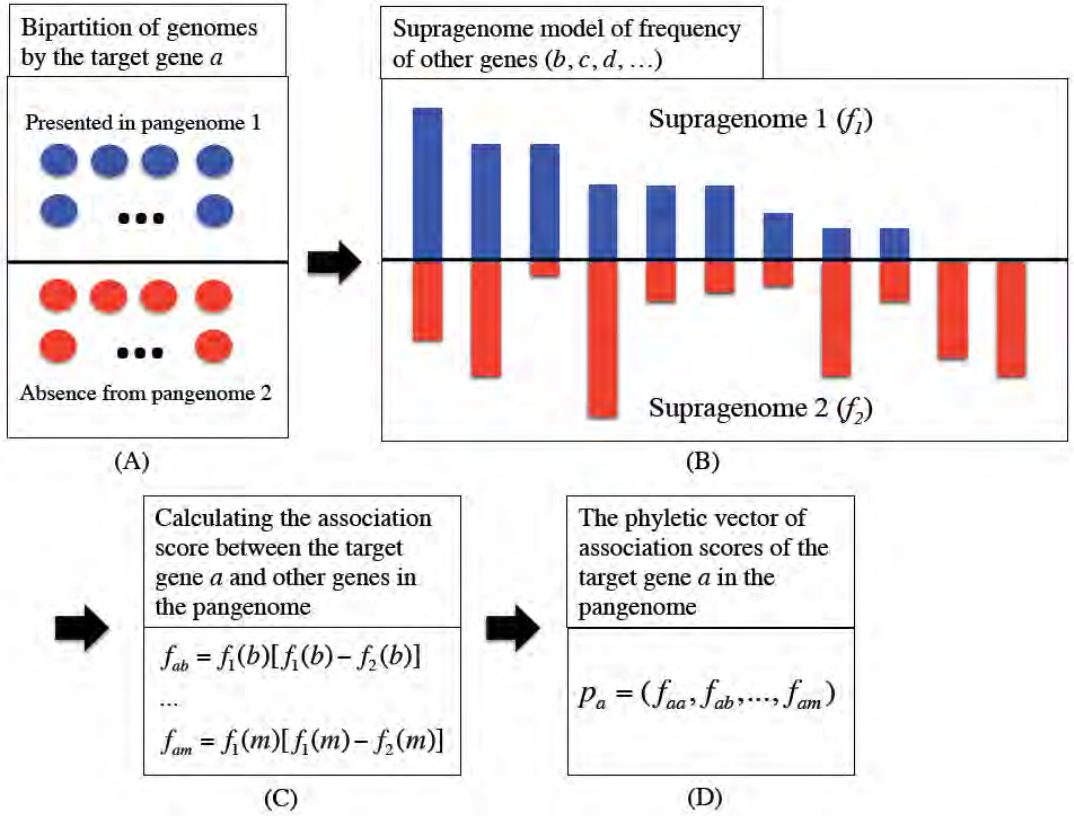

Figure 6.1: A phyletic model based on supragenome gene frequency. Two sub-pangenomes, '*pangenome 1*' and '*pangenome 2*' are defined by the presence or absence of a target gene. These two pangenomes are assumed to have different genotypes associated with the target gene. The supragenome models of the sub-pangenomes provide information on coevolutionary relationships between the target gene and other genes. After accounting for difference between the two supragenomes, a final association score is proposed to characterize the co-evolving genes for the target gene in the '*pangenome 1*'.

6.3 Results and discussion

6.3.1 Testing the phyletic model by predicting protein associations as in the STRING database

The bacteria domain as a whole can be seen as a large pangenome (Lapierre and Gogarten 2009). Therefore, the phyletic model can be used to infer gene coevolution from all the COGs in the STRING database. Several COGs were arbitrarily selected as target genes to test the phyletic model. First, COG0133 that encodes a tryptophan synthase beta chain was used as a target gene to calculate its coevolutionary relationships with all other COGs in the database. Association scores of the coevolutionary relationships were generated with the model (Figure 6.2). It can be seen from the distribution of association scores that most of them are just around 0.00 (Figure 6.2 (A)), which means that most of the COGs may have evolved independently from COG0133. But there are also genes that have scores significantly larger than 0.00, in which the largest score (of 1.00) is for the coevolution with COG0133 itself. To further understand how the scores predict protein associations, different cutoff score values were used to select different sets of best-associated COGs to be compared in the STRING database. A smaller cutoff value will result in a larger set of best-associated COGs, which may include COGs that are not so significantly associated with the target gene, in terms of their biological functions. As shown in Figure 6.3, the best associated COGs of a smaller cutoff value (0.70, 20 COGs, Figure 6.3(A)) shows a more compact association network than those of a larger cutoff value (0.60, 26 COGs, Figure 6.3(B)), as calculated from the STRING database. And a more relax cutoff value can even result in a protein association network of the best-associated COGs that have multiple components in the network, instead of a single highly connected component (Figure 6.3 (C)). Examination of the functions of the best-associated COGs as defined by a cutoff value of

0.70 shows that their biological functions are also associated (Figure 6.4). In fact, these proteins are involved in the biochemical pathway of tryptophan biosynthesis (Kagan, Sharon et al. 2008). Therefore, the cutoff value is crucial for defining the set of best-associated COGs to the target gene to include its co-evolving COGs that have significant correlated biological functions, such as partners in a biochemical pathway.

Two more cases were also tested, representing other biological scenarios of co-evolving genes. COG0056, encoding the alpha subunit of the F0F1-type ATP synthase (Deckers-Hebestreit and Altendorf 1996), was used to test the phyletic model to see if it can predict other subunits of the synthase or not. With a cutoff value of 0.80, all the six predicted best-associated COGs were shown to encode other subunits of the synthase. The third case comes from an ABC-transporter (Bowers, Pellegrini et al. 2004). Predicting the best-associated COGs with COG0555 (ABC-type sulfate transport system, permease component) successfully identified the other components of the system, including periplasmic (COG1613), permease (COG4208) and ATPase (COG1118) components. As a result, although not thoroughly tested for all the COGs in the STRING database, the phyletic model and its association score can predict the protein associations in the database. Furthermore, the cutoff value is crucial for defining highly associated proteins. For the cases that have been tested, a value of 0.70 or larger appears to be a proper choice. But this value may vary for different proteins, and correction for the variation to have a universal normalized score was not explored in this study. For this reason, the cutoff value of 0.70 was chose to simplify the following analysis of the pangenome of the genus *Mycobacterium*.

6.3.2 Clustering the pangenome of the genus *Mycobacterium*

To cluster all the genes in the pangenome of the genus *Mycobacterium* into groups of co-evolving genes, the pangenome data of the genus and its closely related species were used (see the section of materials and methods). The phyletic model was applied to this

pangenome data to generate a matrix of association scores of gene coevolution. As has been suggested in the previous section, a cutoff value of 0.70 was used to define the best-associated GOGs. For those COGs that have scores smaller than 0.70, the scores were reset to 0 in the matrix. Informally, the use of a cutoff value and the reset of the scores that are smaller than that is a kind of correction for emphasizing larger association scores that reflect coevolutionary relationships with functional implications, as demonstrated in the above section. After using the cutoff value to correct the matrix, the MCL algorithm was then used to cluster the matrix with an inflation parameter of 5 (I=5), to produce more fine-grained clusters. As a result, there were 310 clusters of co-evolving genes generated, with size (the number of COGs in the cluster) ranging from 1 (singleton) to 1072. In total, there are 13 clusters of size 3, 60 clusters of size 2 and 136 clusters of size 1. These small clusters and singleton were excluded in the following analysis, for the reason that such small clusters provide only limited information on co-evolving genes. Particularly, there is no information about gene coevolution in singleton clusters.

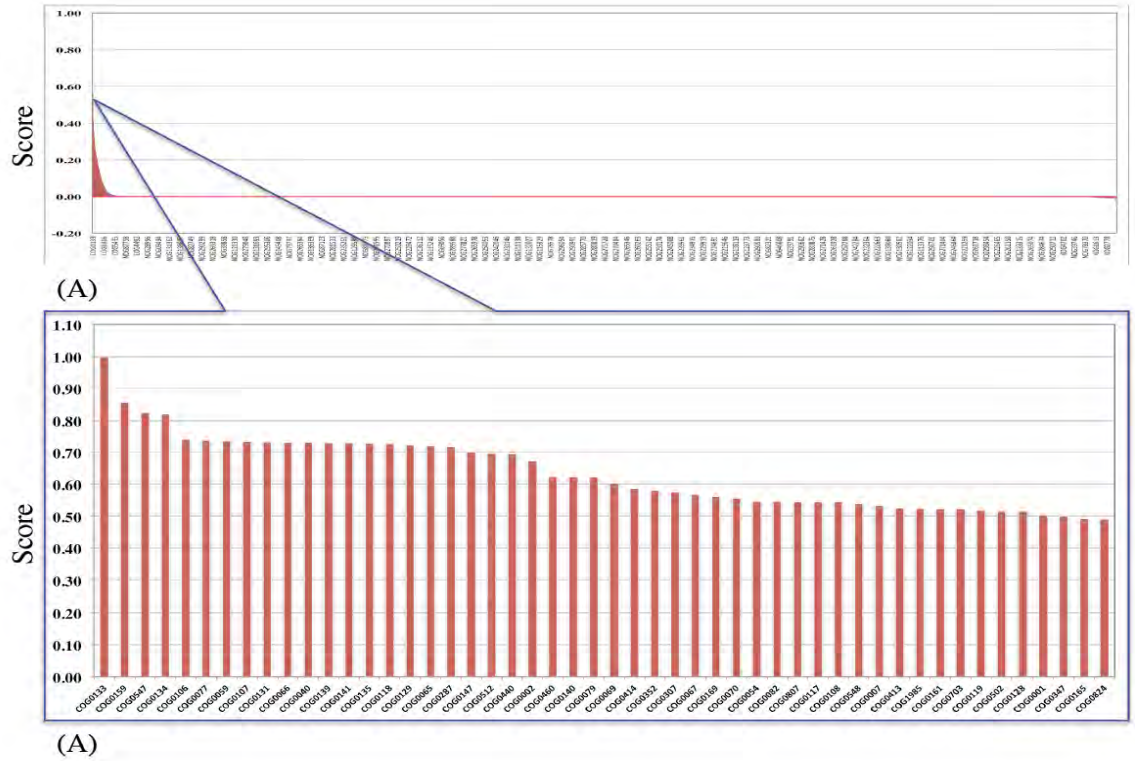

Figure 6.2: Distribution of association scores of COG0133 that characterize the coevolutionary relationships between this COG and other COGs in the STRING database, using the whole bacteria pangenome. (A) All association scores were presented in descending order. (B) An extraction of association scores that are larger than 0.5.

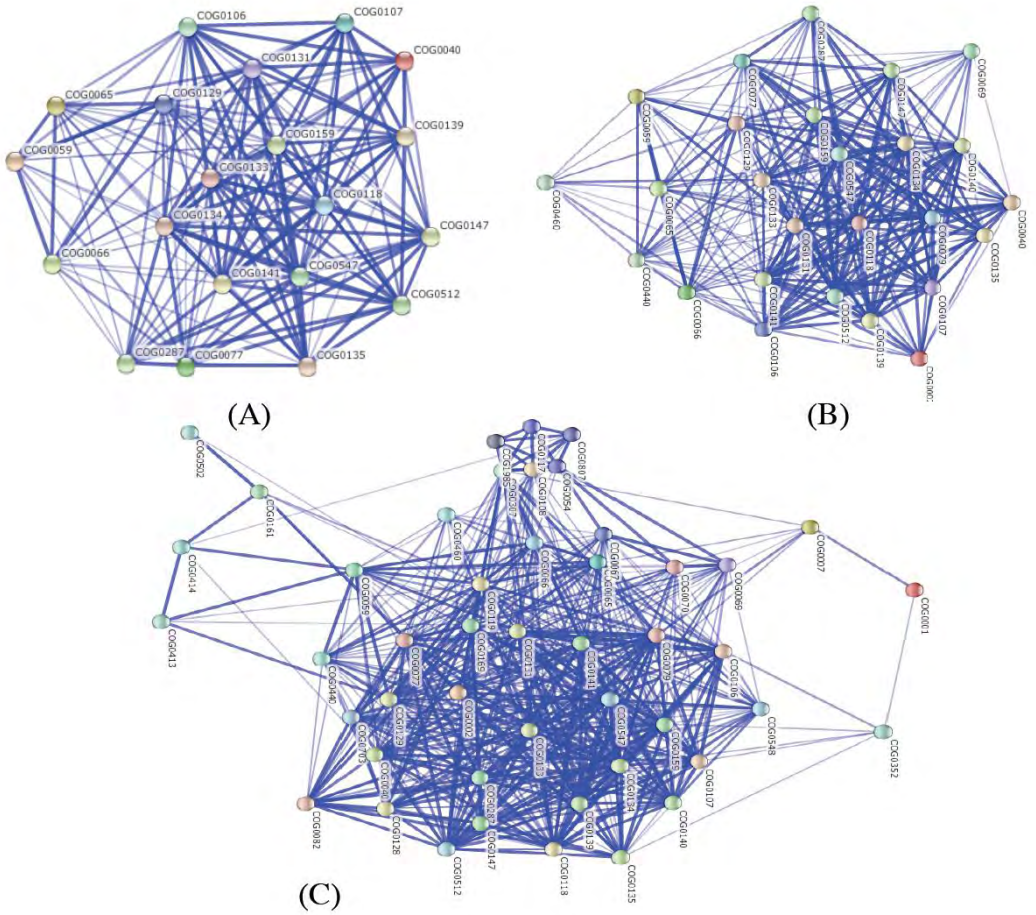

Figure 6.3: Protein associations of the best-associated COGs that were predicted to co-evolve with COG0133. (A) The best-associated COGs were selected by a cutoff score value of 0.70. And the cutoff values for (B) and (C) are 0.60 and 0.50 respectively.

COG id	Score	Function
COG0133	1.00	Tryptophan synthase beta chain
COG0159	0.86	Tryptophan synthase alpha chain
COG0547	0.82	Anthranilate phosphoribosyltransferase
COG0134	0.82	Indole-3-glycerol phosphate synthase
COG0106	0.74	Phosphoribosylformimino-5-aminoimidazole carboxamide ribonucleotide (ProFAR) isomerase
COG0077	0.74	Prephenate dehydratase
COG0059	0.74	Ketol-acid reductoisomerase
COG0107	0.73	Imidazoleglycerol-phosphate synthase
COG0131	0.73	Imidazoleglycerol-phosphate dehydratase
COG0066	0.73	3-isopropylmalate dehydratase small subunit
COG0040	0.73	ATP phosphoribosyltransferase
COG0139	0.73	Phosphoribosyl-AMP cyclohydrolase
COG0141	0.73	Histidinol dehydrogenase
COG0135	0.73	Phosphoribosylanthranilate isomerase
COG0118	0.73	Glutamine amidotransferase
COG0129	0.72	Dihydroxyacid dehydratase/phosphogluconate dehydratase
COG0065	0.72	3-isopropylmalate dehydratase large subunit
COG0287	0.72	Prephenate dehydrogenase
COG0147	0.70	Anthranilate/para-aminobenzoate synthases component I
COG0512	0.70	Anthranilate/para-aminobenzoate synthases component II

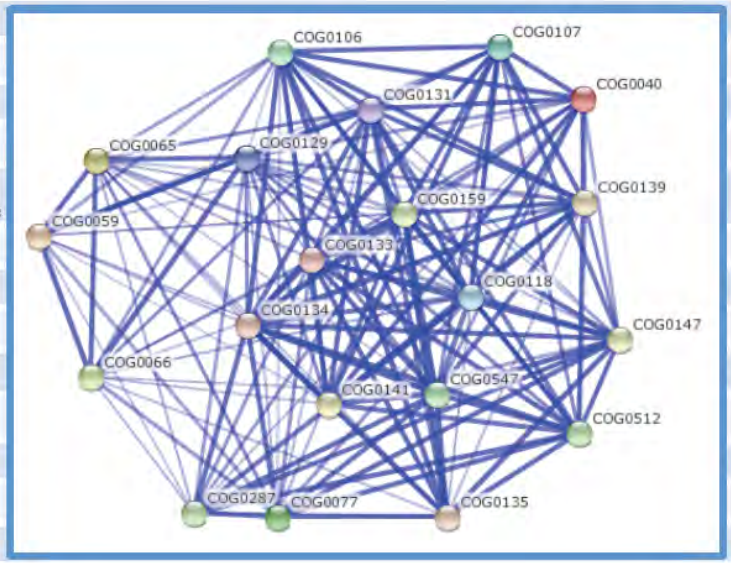

Figure 6.4: Functions of the best-associated COGs with COG0133. These COGs were selected by a cutoff value of 0.70. The top-five COGs are the most significantly related genes in terms of their function (tryptophan biosynthesis) and have been predicted by the phyletic model.

The remaining 101 clusters of size larger than 3 were further analyzed to characterize their distribution pattern among the three groups of organisms, the 'NTM' group (non-tuberculosis *Mycobacterium*), the 'MTBC' group (the *Mycobacterium tuberculosis* complex) and the 'OUTS' group (out-groups of close relatives to the genus *Mycobacterium*). These three groups were conceived to understand how gene co-evolved in the mycobacteria genus and in the MTBC species. Within each cluster, three frequencies or coverage were calculated for this purpose, corresponding to the three groups (Figure 6.5). The frequency was calculated as the fraction of the members of a cluster that are present in a given group of organisms. For example, if every COG in a cluster is present in any genome of the MTBC group, then this cluster was calculated as 100% in this group. With these three group frequencies, each cluster can be interpreted into different patterns of evolution among the groups. As shown in Figure 6.5, most of the clusters are present in the NTM group with 100%. And there are the least number of clusters that are present in the MTBC with 100%. For the purpose of this study, four types of clusters were defined based on the group frequencies. The first one is the 'core' cluster that is present in all three groups with around 100%, such cluster CC1 (Figure 6.5). The second is the 'mycobacteria-core' cluster that is present in both the NTM and MTBC group with around 100%, such as cluster CC36. The third is the cluster lost in the MTBC group (≤2%), but present universally in the other two groups. And the fourth is the cluster gained in the MTBC group (around 100%), but rarely present in the other two groups (≤2%), such as cluster CC9. These four types of clusters may represent different groups of genes that are responsible for the evolution of the mycobacteria genus and the MTBC species, and therefore have functional implication for their adaptation and speciation, which will be discussed in the following sections.

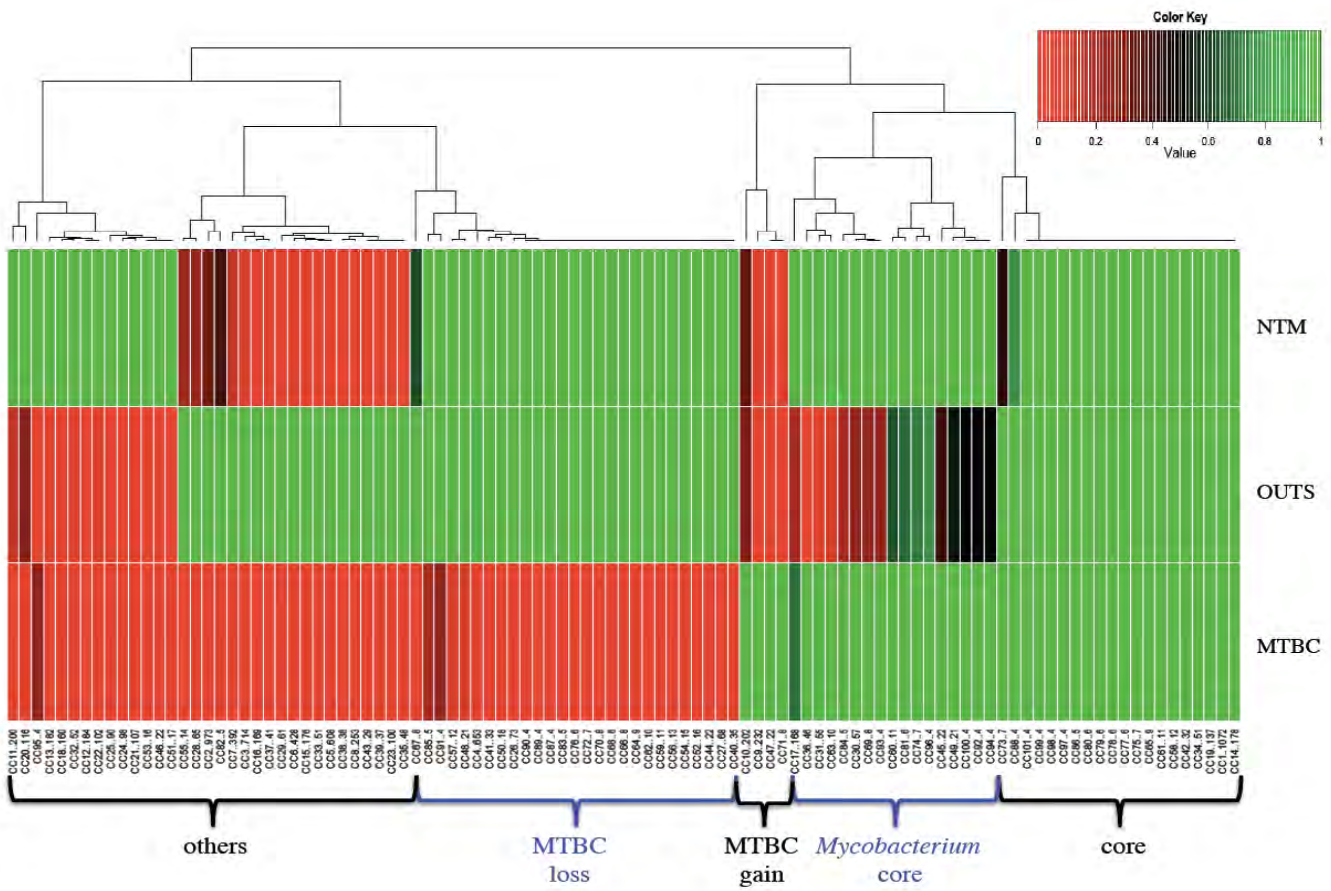

Figure 6.5: Different types of pangenome clusters that are distributed among the three groups of organisms. Four types of clusters were defined according to their frequencies in the three groups, the NTM group, the MTBC group and the out-groups. Clusters that do not belong to any of the four types were defined as 'others'.

6.3.3 Universal core clusters of essential genes

The largest cluster CC1, containing 1,072 COGs, is belonging to the 'core' clusters, which are universally present in all the organisms in the three groups. Reconstruction of protein association network in the STRING database shows that these COGs are highly connected to each other, forming a single dense component, which is surrounding by a few dispersed COGs (Figure 6.6). Further examination also shows that these COGs are of essential genes, encoding proteins involved in basic metabolic pathways, such as ribosomal proteins in the cellular process of translation and also proteins involved in DNA repair, replication and recombination. These genes are also known to be conversed in most bacterial genomes and therefore represent the universal core genome in many bacterial species (Lapierre and Gogarten 2009). In addition to those universal core genes, the cluster CC1 also includes core genes that may be specific to the three groups of organism. Examples are NOG04168 and NOG04169, which are involved in cell wall arabinan biosynthesis. Another two large clusters that are core in the three groups are cluster CC14 (of size 178) and CC19 (of size 137). These two clusters include genes that are not universal in all bacterial genomes, but essential in some specific species. Examples are COGs involved in cobalamin biosynthesis. Thirteen subunits of the NADH-ubiquinone oxidoreductase complex are also included in the cluster CC14.

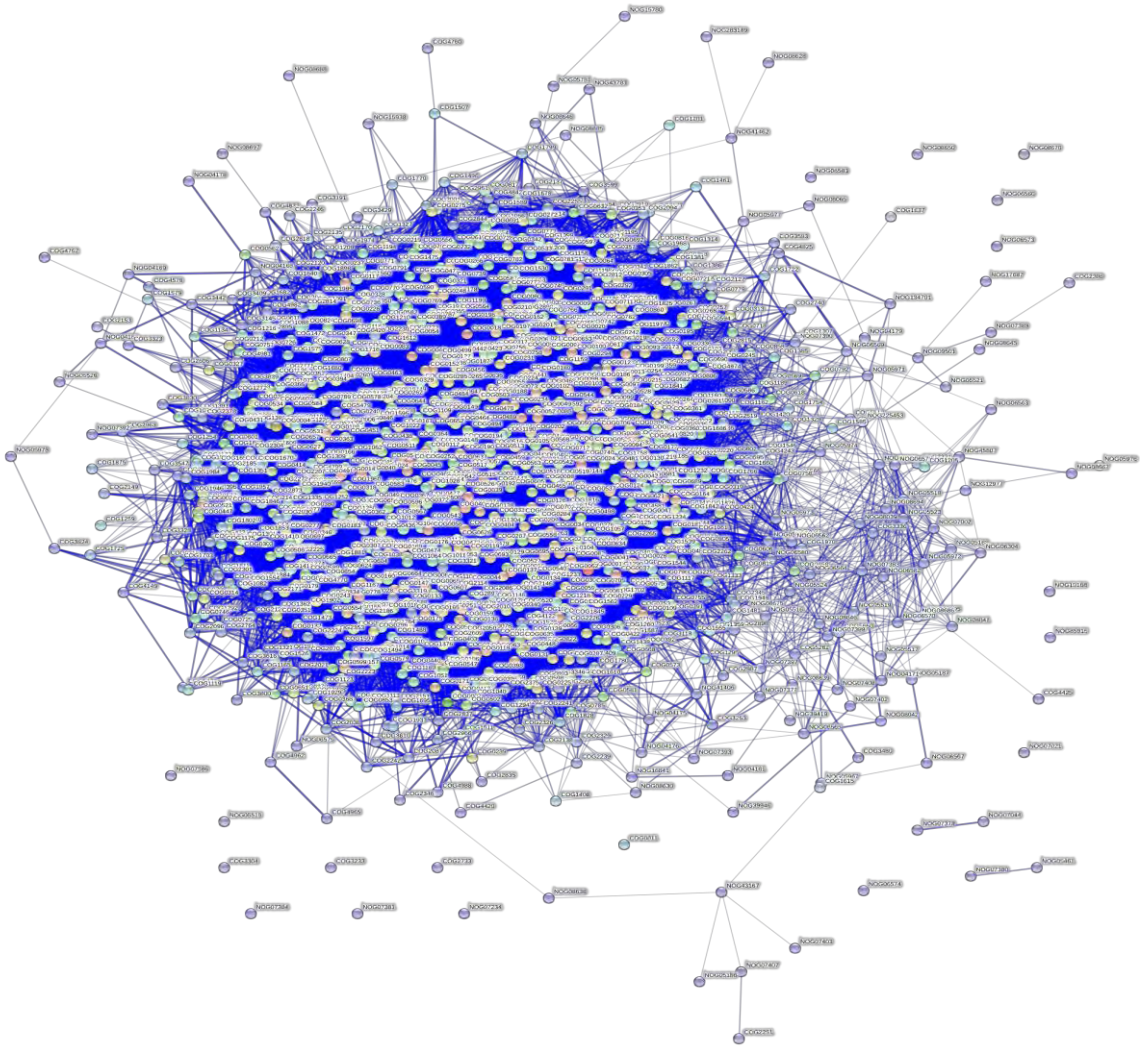

Figure 6.6: Protein association network of the cluster CC1. COGs in this cluster are mainly involved in basic metabolic pathways, and are of essential gene that are conserved in most bacterial genomes.

Results of the core clusters show that the mycobacteria genus and its related out-group species share a large common core of genes. Although a large fraction of these core genes are also universal in other bacterial genomes, there are core genes specific to this actinobacterial CMN group. Identification of the CMN group specific genes and their related clusters, such as genes involved in cell wall arabinan biosynthesis, may help to understand the evolution of this group in the *Actinobacteria* phylum, especially for identifying genes that are responsible for its characteristic phenotypes, such as cell envelope composition (Ventura, Canchaya et al. 2007). Furthermore, the success in identifying different core clusters gives a solid test to the clustering method and the pangenome phyletic model. A particularly successful example is the cluster CC1, confirmed by the STRING database as a highly connected group of COGs of essential genes.

6.3.4 Mycobacteria-core clusters

Three large clusters, CC31 (55 COGs), CC36 (46 COGs) and CC63 (10 COGs), were found to be core in both the NTM group and the MTBC group, but rarely occurred in the out-group species. The most significant is that many lipoproteins are contained in the cluster CC31, including *LppA*, *LppN*, *LppJ*, *LppP* and *LppV*, which are related to the cell envelope that is characteristic of the genus (Cole, Brosch et al. 1998). In addition to these, COGs of cell surface protein (precursor PirG, NOG11683; precursor Erp, NOG261550) and seven predicted transmembrane proteins were also found in this cluster. More examples are COGs of tuberculin related protein (NOG236487), cutinase (NOG40559) and chitinase (NOG46487). In contrast to the cluster CC31, most of the COGs in the cluster CC36 are of PPE or PE-PGRS family proteins, which are well known *Mycobacterium* specific proteins. The cluster CC63 contains an ESAT-6 like protein (NOG13301), but other members in this cluster are only hypothetical proteins.

From the above examples it can be seen that mycobacteria-core clusters mainly contain genes that are defining the characteristic phenotypes of the genus, such as cell envelope composition, cell surface protein and transmembrane proteins, and also the characteristic PE/PPE protein families. Genes of the ESAT-6 secretion system (NOG13243 and NOG13301) and exported proteins (NOG11683, NOG279689, and NOG261550) were also observed in these mycobacteria-core clusters. But many other COGs are only predicted to encode hypothetical proteins and have unknown functions. Since these hypothetical proteins belong to the core clusters of the genus, illustration the biological functions of them can further improve our understanding of the genetic basis that determine its characteristics (Xu, Laine et al. 2007). One example is the cluster CC63. Many of the hypothetical proteins in this cluster may relate to the ESAT-6 like protein (NOG13301), which is also included in the same cluster.

6.3.5 Lost and gained clusters in MTBC

There are 18 clusters (of 253 COGs) that were found to be totally lost in the MTBC group. In contrast, only 30 COGs were found gained in the MTBC group. The more prevalence of gene loss than gain is consistent with previous studies. Also, the identification of clusters that are present in the MTBC group but absent from the other two groups supports the hypothesis that gene gains or duplications occurred in the early formation of the MTBC species, which is followed by massive genome reduction (Veyrier, Dufort et al. 2011). These lost and gained clusters were further examined for their biological functions to understand what contributions they may have to the evolution of the MTBC species.

Reconstruction of protein association network of the COGs that are lost in the MTBC group indicates that these COGs fall into several groups of genes of related biological functions (Figure 6.6). Although near half of the COGs encode proteins of unknown function, a significant portion of the remaining COGs were found to encode

various ABC-type transport systems, symporters, antiporters and synthetases. The loss of these genes in the MTBC group may relate to the fact that the MTBC species are obligate human pathogens and have adapted to their host environment. Detailed examples of the loss pathways involving these genes are ABC-type branched-chain amino acid transport systems, ABC-type spermidine/putrescine transport systems, ABC-type nitrate/sulfonate/bicarbonate transport system and ABC-type Fe3+ transport system. The genes encoding these three ABC-type transport systems appear to be highly connected in the association network, implying that they may have co-evolved in the NTM and out-group species. Another set of co-evolving genes that are loss in the MTBC group includes ABC-type ribose/xylose/arabinose/galactoside transport systems, ABC-type xylose transport system, ABC-type sugar transport system and fructose-specific phosphotransferase system. This implies a limited capacity of sugar metabolisms in MTBC species (Braibant, Gilot et al. 2000; Titgemeyer, Amon et al. 2007; Niederweis 2008). Furthermore, ABC-type metal ion transport system was also found to be lost in the MTCB group. Lost symporters and antiporters include multisubunit Na+/H+ antiporter, H+/gluconate symporter, Na+/proline symporter, Na+/alanine symporter, solute-hydrogen antiporter and Ca2+/Na+ antiporter. The lost synthetases include selenocysteine synthase, selenophosphate synthase, spermidine synthase, 3-deoxy-D-arabino-heptulosonate 7-phosphate (DAHP) synthase, 2-Isopropylmalate synthase, 3-hydroxy-3-methylglutaryl CoA synthase, phosphatidylserine synthases, and isopenicillin N synthase. Because all these lost COGs are still present in the NTM and the out-group species, it can be inferred that the MTBC species have lost these genes during its step-wise evolution from its environmental ancestor. And the question of how the loss of these genes has contributed to its pathogenesis is worth further examination (Fang, Wallqvist et al. 2012).

Among the 30 COGs that were predicted to be gained in the MTBC group, 9 of them encode PE-PGRS family proteins. These PE-PGRS genes may have been duplicated in the

evolution of the MTBC species. Although their functions have not been clearly determined yet, it was suggested that the PE/PPE proteins might be of immunological importance and play roles in mycobacterial virulence (Akhter, Ehebauer et al. 2012). Therefore, gain or duplication of the gene family should have contributed to the evolution of the MTBC species as host-associated pathogens. In addition to the PE-PGRS genes, other gained genes include those encoding predicted membrane proteins, exported proteins and also transcriptional regulators. But their exact functions are still unknown. Therefore what roles they may have in the evolution of the MTBC species are remained to be elucidated.

6.4 Conclusion

A pangenome phyletic model for analyzing gene coevolution in the divergence and adaptation of microbial species was developed and tested using the STRING database. This phyletic model calculates a coevolution metric of gene frequencies in a pangenome data. With this metric, graph based clustering was performed to identify coevolutionary clusters of functional genes. Applying this method to the genus *Mycobacterium* dissected its pangenome into different clusters, from conserved and core clusters of essential biological functions to species-specific clusters of pathogenesis related genes. This result shows that the MTB species has arose from its mycobacterial ancestor mainly by loss of many environmental related genes. But gain of genes, probably by horizontal gene transfer or genome duplication, has also occurred within the MTB species, especially the clusters of PE/PPE genes.

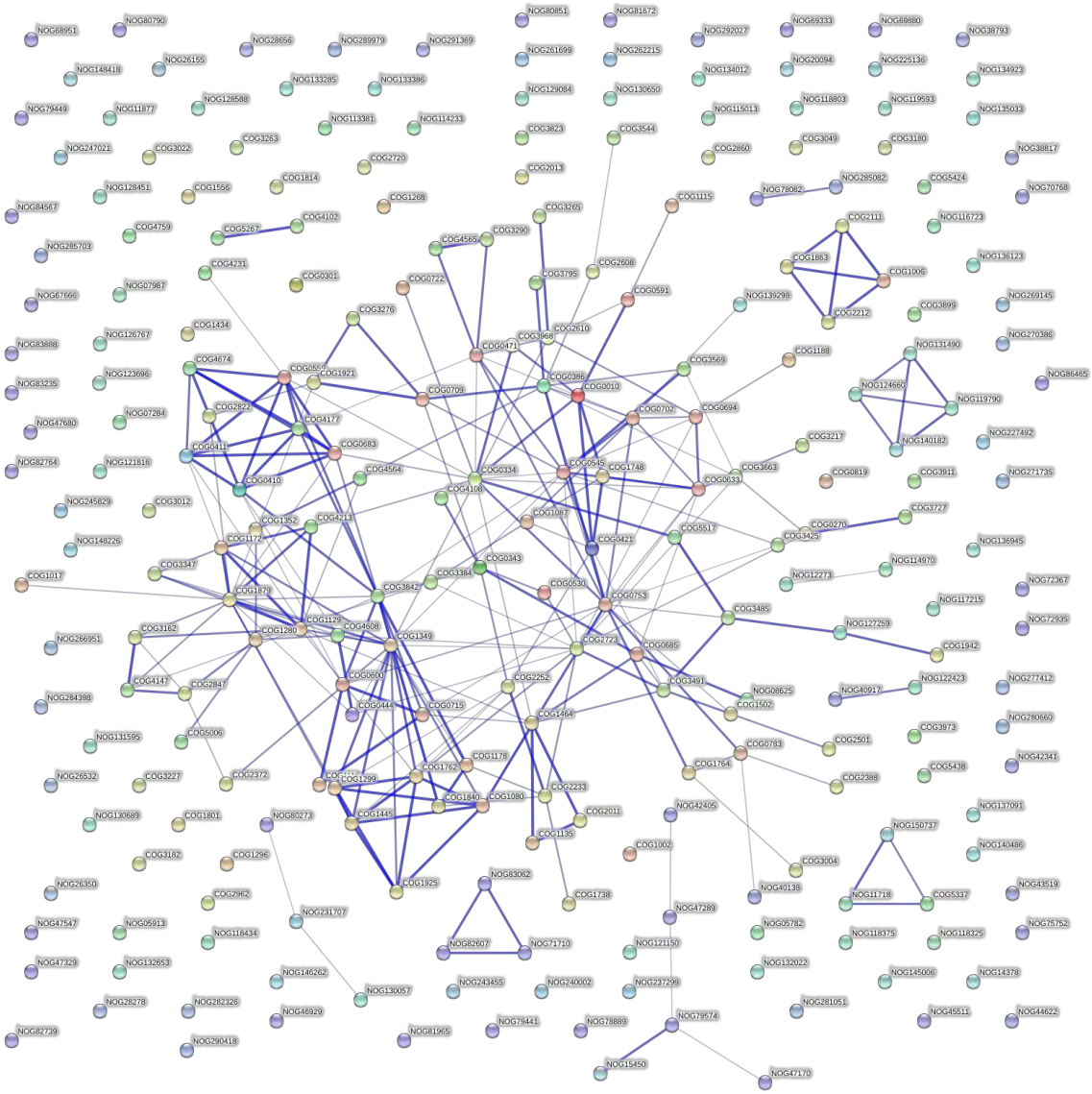

Figure 6.7: Protein association network of the COGs that are lost in the MTBC group. The lost COGs fall into several groups of genes of related biological functions, as indicated by the connections in the association network. A significant portion of the COGs of known functions was found to encode various ABC-type transport systems, symporters, antiporters, and synthetases.

Chapter 7 Concluding remarks

With the advent of high-throughput next generation sequencing technologies, pangenome sequencing will be applied to analyze more and more microbial species, including both important pathogenic bacterial species and environmental bacterial species of ecological importance (Medini, Donati et al. 2005; Bentley 2009). In addition to the analysis of pangenome variations that are associated with type strains of different phenotypes that are within the same species, genome sequencing at a population scale can even be taken to understand the short-term dynamic of pangenome evolution (Shapiro, Friedman et al. 2012). In face of these large-scale pangenome data and their associated biological information, there are emerging bioinformatic challenges (Read and Ussery 2006), from whole genome mapping and alignment, comparative genome annotation to the developments of evolutionary pangenome models and pangenome database.

In this thesis, a local parsimony ancestral genome reconstruction method and a phyletic model for pangenome clustering were developed, which were shown to be powerful in analyzing the pangenome of the *Mycobacterium tuberculosis* complex species and also of the *Mycobacterium* genus. The results showed that identification of genome-wide indels, followed by ancestral state reconstruction of indel evolution, provides a useful approach to identify ancestral genomic events that have shaped the evolution of different MTB lineages. Particularly, this approach were applied to analyze the evolution of the MTB Beijing family, leading to the identification of several large RDs that have been lost in this family and also many disrupted genes involved in MTB pathogenesis. Further experimental examination of these results may help to understand the formation of the Beijing family as a hypervirulent genotype (Parwati, van Crevel et al. 2010). Another significant result from the

ancestral genome reconstruction analysis was that ancestral expansion of IS6110 elements was found to be associated with the formation of different MTB lineages, which has been suggested to provide a systematic way for the reductive evolution and adaptation of this pathogenic species (McEvoy, Falmer et al. 2007).

It has been proposed that the MTBC species might have undergone a biphasic evolution, in which early horizontal gene transfers and duplications were followed by gene deletions and disruptions (Veyrier, Dufort et al. 2011). By clustering the pangenome of the *Mycobacterium* genus, using the pangenome phyletic model developed in this study, gene clusters that are corresponding to the two evolutionary phases were revealed. There are many clusters found to be lost in the MTBC species, most of which are involved in ABC-type transport systems for branched-chain amino acids or for sugar metabolisms. Among the lost clusters, there are also genes encoding symporters, antiporters, and various synthetases. In contrast, fewer genes were found to be gained in the MTBC species, mainly encoding PE/PPE family proteins. This result is in accordance with the hypothesis that the MTBC species might have arose from an ancestral environment mycobacterial species by loss of many environmental related genes, followed by gene gain and loss that allowed it to adapt to the host environment (Ventura, Canchaya et al. 2007; Veyrier, Dufort et al. 2011).

Furthermore, extension and formalization of the local parsimony ancestral state reconstruction method were made to allow for multi-state ancestral reconstruction using scoring matrices. The method was found to provide an alternative perspective to understand parsimony analysis in phylogenetics, which appears to be more useful and can be used to incorporate statistical analysis into the method. Testing of the null and alternative model in the tacking and triggering operations in this method explicitly considers the scoring matrix of state transformations. Therefore, development of stochastic evolutionary process based on the scoring matrix may provide better null and alternative model, instead of the simple

ones used in this study. Such kind of models can be used to derive p-values for group comparisons of different states in the clades of the reference tree, and therefore introduce statistics signification into the maximum parsimony analysis (Omland 1999; Huelsenbeck and Bollback 2001).

On the other hand, the pangenome phyletic model was developed for analyzing gene coevolution in the divergence and adaptation of microbial species, instead of analysis of individual genes in the above method. This phyletic model calculates a coevolution metric of gene frequencies in a pangenome. With this metric, graph based clustering is performed to identify coevolutionary clusters of functional genes. These two steps can be further improved in future work. First, in the comparison of gene frequencies between different pangenomes, a new coevolutionary metric should be more accurate when considering a species tree that relates the individual genome in the pangenome. Second, despite of the general purpose for the MCL clustering algorithm that was used, more accurate clustering should also be developed to considering the evolutionary tree that relates the species in the pangenome. Therefore, stochastic process on trees will be the future work in this direction.

Meanwhile, these methods are general and can be applied in pangenome analysis of other microbial species and even the whole bacteria domain. Finally, this work clearly indicated that the sample size for pangenome modeling is critical, which thus provides a starting point for future comprehensive pangenome sequencing of mycobacteria. Along with these massive sequencing efforts, a database of *Mycobacterium* for integrative pangenome annotation and evolutionary analysis should be developed.

References

Agarwal, N. and W. R. Bishai (2009). "cAMP signaling in Mycobacterium tuberculosis." Indian journal of experimental biology **47**(6): 393-400.

Akhter, Y., M. T. Ehebauer, et al. (2012). "The PE/PPE multigene family codes for virulence factors and is a possible source of mycobacterial antigenic variation: perhaps more?" Biochimie **94**(1): 110-116.

Alexander, K. A., P. N. Laver, et al. (2010). "Novel Mycobacterium tuberculosis complex pathogen, M. mungi." Emerging infectious diseases **16**(8): 1296-1299.

Alland, D., D. W. Lacher, et al. (2007). "Role of large sequence polymorphisms (LSPs) in generating genomic diversity among clinical isolates of Mycobacterium tuberculosis and the utility of LSPs in phylogenetic analysis." Journal of clinical microbiology **45**(1): 39-46.

Ambur, O. H., T. Davidsen, et al. (2009). "Genome dynamics in major bacterial pathogens." FEMS microbiology reviews **33**(3): 453-470.

Bannantine, J. P., C. W. Wu, et al. (2012). "Genome sequencing of ovine isolates of Mycobacterium avium subspecies paratuberculosis offers insights into host association." BMC genomics **13**: 89.

Baumdicker, F., W. R. Hess, et al. (2012). "The infinitely many genes model for the distributed genome of bacteria." Genome biology and evolution **4**(4): 443-456.

Becq, J., M. C. Gutierrez, et al. (2007). "Contribution of horizontally acquired genomic islands to the evolution of the tubercle bacilli." Molecular biology and evolution **24**(8): 1861-1871.

Behr, M. A., M. A. Wilson, et al. (1999). "Comparative genomics of BCG vaccines by whole-genome DNA microarray." Science **284**(5419): 1520-1523.

Bentley, S. (2009). "Sequencing the species pan-genome." Nature reviews. Microbiology **7**(4): 258-259.

Bentley, S. D., I. Comas, et al. (2012). "The genome of Mycobacterium africanum West African 2 reveals a lineage-specific locus and genome erosion common to the M. tuberculosis complex." PLoS neglected tropical diseases **6**(2): e1552.

Bondy-Denomy, J., A. Pawluk, et al. (2012). "Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system." Nature.

Bowers, P. M., M. Pellegrini, et al. (2004). "Prolinks: a database of protein functional linkages derived from coevolution." Genome biology **5**(5): R35.

Braibant, M., P. Gilot, et al. (2000). "The ATP binding cassette (ABC) transport systems of Mycobacterium tuberculosis." FEMS microbiology reviews **24**(4): 449-467.

Bretl, D. J., H. He, et al. (2012). "MprA and DosR coregulate a Mycobacterium tuberculosis virulence operon encoding Rv1813c and Rv1812c." Infection and immunity **80**(9): 3018-3033.

Brosch, R., S. V. Gordon, et al. (2002). "A new evolutionary scenario for the Mycobacterium tuberculosis complex." Proceedings of the National Academy of Sciences of the United States of America **99**(6): 3684-3689.

Brosch, R., A. S. Pym, et al. (2001). "The evolution of mycobacterial pathogenicity: clues from comparative genomics." Trends in microbiology **9**(9): 452-458.

Brown, T., V. Nikolayevskyy, et al. (2010). "Associations between Mycobacterium tuberculosis strains and phenotypes." Emerging infectious diseases **16**(2): 272-280.

Camus, J. C., M. J. Pryor, et al. (2002). "Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv." Microbiology **148**(Pt 10): 2967-2973.

Chen, F., A. J. Mackey, et al. (2006). "OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups." Nucleic acids research **34**(Database issue): D363-368.

Chen, L., Z. Xiong, et al. (2012). "VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors." Nucleic acids research **40**(Database issue): D641-645.

Chen, S. L., C. S. Hung, et al. (2006). "Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach." Proceedings of the National Academy of Sciences of the United States of America **103**(15): 5977-5982.

Coddington, J. A. (1988). "CLADISTIC TESTS OF ADAPTATIONAL HYPOTHESES." Cladistics-the International Journal of the Willi Hennig Society **4**(1): 3-22.

Cole, S. T., R. Brosch, et al. (1998). "Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence." Nature **393**(6685): 537-544.

Collins, R. E. and P. G. Higgs (2012). "Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome." Molecular biology and evolution **29**(11): 3413-3425.

Comas, I. and S. Gagneux (2011). "A role for systems epidemiology in tuberculosis research." Trends in microbiology **19**(10): 492-500.

Cunningham, C. W., K. E. Omland, et al. (1998). "Reconstructing ancestral character states: a critical reappraisal." Trends in ecology & evolution **13**(9): 361-366.

Cunningham, F. X., Jr., T. P. Lafond, et al. (2000). "Evidence of a role for LytB in the nonmevalonate pathway of isoprenoid biosynthesis." Journal of bacteriology **182**(20): 5841-5848.

Darling, A. E., B. Mau, et al. (2010). "progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement." PloS one **5**(6): e11147.

Deckers-Hebestreit, G. and K. Altendorf (1996). "The F0F1-type ATP synthases of bacteria: structure and function of the F0 complex." Annual review of microbiology **50**: 791-824.

Delsuc, F., H. Brinkmann, et al. (2005). "Phylogenomics and the reconstruction of the tree of life." Nature reviews. Genetics **6**(5): 361-375.

Djelouadji, Z., D. Raoult, et al. (2011). "Palaeogenomics of Mycobacterium tuberculosis: epidemic bursts with a degrading genome." The Lancet infectious diseases **11**(8): 641-650.

Dobrindt, U., B. Hochhut, et al. (2004). "Genomic islands in pathogenic and environmental microorganisms." Nature reviews. Microbiology **2**(5): 414-424.

Domenech, P., M. B. Reed, et al. (2005). "Contribution of the Mycobacterium tuberculosis MmpL protein family to virulence and drug resistance." Infection and immunity **73**(6): 3492-3501.

Donati, C., N. L. Hiller, et al. (2010). "Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species." Genome biology **11**(10): R107.

Dongen, S. V. (2000). Graph Clustering by Flow Simulation, University of Utrecht.

Donoghue, M. J. (1989). "PHYLOGENIES AND THE ANALYSIS OF EVOLUTIONARY SEQUENCES, WITH EXAMPLES FROM SEED PLANTS." Evolution **43**(6): 1137-1156.

Dou, H. Y., F. C. Tseng, et al. (2008). "Molecular epidemiology and evolutionary genetics of Mycobacterium tuberculosis in Taipei." BMC infectious diseases **8**: 170.

Ehrlich, G. D., A. Ahmed, et al. (2010). "The distributed genome hypothesis as a rubric for understanding evolution in situ during chronic bacterial biofilm infectious processes." FEMS immunology and medical microbiology **59**(3): 269-279.

Enright, A. J., S. Van Dongen, et al. (2002). "An efficient algorithm for large-scale detection of protein families." Nucleic acids research **30**(7): 1575-1584.

Faksri, K., F. Drobniewski, et al. (2011). "Genetic diversity of the Mycobacterium tuberculosis Beijing family based on IS6110, SNP, LSP and VNTR profiles from Thailand." Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases **11**(5): 1142-1149.

Fang, X., A. Wallqvist, et al. (2012). "Modeling phenotypic metabolic adaptations of Mycobacterium tuberculosis H37Rv under hypoxia." PLoS computational biology **8**(9): e1002688.

Felsenstein, J. (2003). Counting evolutionary changes. Inferring Phylogenies. Sunderland, Massachusetts, Sinauer Associates, Inc.**:** 11-18.

Felsenstein, J. (2003). Finding the best tree by branch and bound. Inferring Phylogenies. Sunderland, Massachusetts, Sinauer Associates, Inc.**:** 54-66.

Felsenstein, J. (2003). Finding the best tree by heuristic search. Inferring Phylogenies. Sunderland, Massachusetts, Sinauer Associates, Inc.**:** 37-53.

Filliol, I., A. S. Motiwala, et al. (2006). "Global phylogeny of Mycobacterium tuberculosis based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis

evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set." Journal of bacteriology **188**(2): 759-772.

Fischer, S., B. P. Brunk, et al. (2011). "Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups." Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.] **Chapter 6**: Unit 6 12 11-19.

Fischer, W., L. Windhager, et al. (2010). "Strain-specific genes of Helicobacter pylori: genome evolution driven by a novel type IV secretion system and genomic island transfer." Nucleic acids research **38**(18): 6089-6101.

Fitch, W. M. (1971). "TOWARD DEFINING COURSE OF EVOLUTION - MINIMUM CHANGE FOR A SPECIFIC TREE TOPOLOGY." Systematic Zoology **20**(4): 406-&.

Fleischmann, R. D., D. Alland, et al. (2002). "Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains." Journal of bacteriology **184**(19): 5479-5490.

Ford, C., K. Yusim, et al. (2012). "Mycobacterium tuberculosis - Heterogeneity revealed through whole genome sequencing." Tuberculosis **92**(3): 194-201.

Gaasterland, T. and M. A. Ragan (1998). "Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes." Microbial & comparative genomics **3**(4): 199-217.

Glynn, J. R., J. Whiteley, et al. (2002). "Worldwide occurrence of Beijing/W strains of Mycobacterium tuberculosis: a systematic review." Emerging infectious diseases **8**(8): 843-849.

Gogarten, J. P. and J. P. Townsend (2005). "Horizontal gene transfer, genome innovation and evolution." Nature reviews. Microbiology **3**(9): 679-687.

Gordon, S. V., R. Brosch, et al. (1999). "Identification of variable regions in the genomes of tubercle bacilli using bacterial artificial chromosome arrays." Molecular microbiology **32**(3): 643-655.

Guindon, S., J. F. Dufayard, et al. (2010). "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0." Systematic biology **59**(3): 307-321.

Gutierrez, M. C., S. Brisse, et al. (2005). "Ancient origin and gene mosaicism of the progenitor of Mycobacterium tuberculosis." PLoS pathogens **1**(1): e5.

Hanekom, M., N. C. Gey van Pittius, et al. (2011). "Mycobacterium tuberculosis Beijing genotype: a template for success." Tuberculosis **91**(6): 510-523.

Hanekom, M., G. D. van der Spuy, et al. (2007). "A recently evolved sublineage of the Mycobacterium tuberculosis Beijing strain family is associated with an increased ability to spread and cause disease." Journal of clinical microbiology **45**(5): 1483-1490.

He, L., X. Fan, et al. (2012). "Comparative genomic structures of Mycobacterium CRISPR-Cas." Journal of cellular biochemistry **113**(7): 2464-2473.

Hershberg, R., M. Lipatov, et al. (2008). "High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography." PLoS biology **6**(12): e311.

Hershberg, R., M. Lipatov, et al. (2008). "High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography." PLoS Biol **6**(12): e311.

Hiller, N. L., B. Janto, et al. (2007). "Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome." Journal of bacteriology **189**(22): 8186-8195.

Hisert, K. B., M. A. Kirksey, et al. (2004). "Identification of Mycobacterium tuberculosis counterimmune (cim) mutants in immunodeficient mice by differential screening." Infection and immunity **72**(9): 5315-5321.

Hogg, J. S., F. Z. Hu, et al. (2007). "Characterization and modeling of the Haemophilus influenzae core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains." Genome biology **8**(6): R103.

Huelsenbeck, J. P. and J. P. Bollback (2001). "Empirical and hierarchical Bayesian estimation of ancestral states." SYSTEMATIC BIOLOGY **50**(3): 351-366.

Jang, J., J. Becq, et al. (2008). "Horizontally acquired genomic islands in the tubercle bacilli." Trends in microbiology **16**(7): 303-308.

Johnson, Z. I., E. R. Zinser, et al. (2006). "Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients." Science **311**(5768): 1737-1740.

Kagan, J., I. Sharon, et al. (2008). "The tryptophan pathway genes of the Sargasso Sea metagenome: new operon structures and the prevalence of non-operon organization." Genome biology **9**(1): R20.

Katoh, K., K. Kuma, et al. (2005). "MAFFT version 5: improvement in accuracy of multiple sequence alignment." Nucleic acids research **33**(2): 511-518.

Kensche, P. R., V. van Noort, et al. (2008). "Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution." Journal of the Royal Society, Interface / the Royal Society **5**(19): 151-170.

Kettler, G. C., A. C. Martiny, et al. (2007). "Patterns and implications of gene gain and loss in the evolution of Prochlorococcus." PLoS genetics **3**(12): e231.

Kim, E. Y., P. Nahid, et al. (2010). "Novel hot spot of IS6110 insertion in Mycobacterium tuberculosis." Journal of clinical microbiology **48**(4): 1422-1424.

Lapierre, P. and J. P. Gogarten (2009). "Estimating the size of the bacterial pan-genome." Trends in genetics : TIG **25**(3): 107-110.

Lefebure, T. and M. J. Stanhope (2007). "Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition." Genome biology **8**(5): R71.

Li, J. B., J. M. Gerdes, et al. (2004). "Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene." Cell **117**(4): 541-552.

Li, L., C. J. Stoeckert, Jr., et al. (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome research **13**(9): 2178-2189.

Maddison, W. P., M. J. Donoghue, et al. (1984). "OUTGROUP ANALYSIS AND PARSIMONY." Systematic Zoology **33**(1): 83-103.

Maddison, W. P. and D. R. Maddison (2000). "MacClade. Analysis of Phylogeny and Character Evolution."

Maddison, W. P. and D. R. Maddison (2006). "Mesquite: a modular system for evolutionary analysis."

McDonough, J. A., J. R. McCann, et al. (2008). "Identification of functional Tat signal sequences in Mycobacterium tuberculosis proteins." Journal of bacteriology **190**(19): 6428-6438.

McEvoy, C. R., A. A. Falmer, et al. (2007). "The role of IS6110 in the evolution of Mycobacterium tuberculosis." Tuberculosis **87**(5): 393-404.

McEvoy, C. R., R. M. Warren, et al. (2009). "Multiple, independent, identical IS6110 insertions in Mycobacterium tuberculosis PPE genes." Tuberculosis **89**(6): 439-442.

Medini, D., C. Donati, et al. (2005). "The microbial pan-genome." Current opinion in genetics & development **15**(6): 589-594.

Mestre, O., T. Luo, et al. (2011). "Phylogeny of Mycobacterium tuberculosis Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair." PloS one **6**(1): e16020.

Mirkin, B. G., T. I. Fenner, et al. (2003). "Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes." BMC evolutionary biology **3**: 2.

Mostowy, S., A. G. Tsolaki, et al. (2003). "The in vitro evolution of BCG vaccines." Vaccine **21**(27-30): 4270-4274.

Muzzi, A. and C. Donati (2011). "Population genetics and evolution of the pan-genome of Streptococcus pneumoniae." International journal of medical microbiology : IJMM **301**(8): 619-622.

Niederweis, M. (2008). "Nutrient acquisition by mycobacteria." Microbiology **154**(Pt 3): 679-692.

Nixon, K. C. and J. M. Carpenter (1993). "ON OUTGROUPS." Cladistics-the International Journal of the Willi Hennig Society **9**(4): 413-426.

Omland, K. E. (1999). "The assumptions and challenges of ancestral state reconstructions." SYSTEMATIC BIOLOGY **48**(3): 604-611.

Parwati, I., R. van Crevel, et al. (2010). "Possible underlying mechanisms for successful emergence of the Mycobacterium tuberculosis Beijing genotype strains." The Lancet infectious diseases **10**(2): 103-111.

Pellegrini, M., E. M. Marcotte, et al. (1999). "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." Proceedings of the National Academy of Sciences of the United States of America **96**(8): 4285-4288.

Ramon-Garcia, S., V. Mick, et al. (2012). "Functional and genetic characterization of the tap efflux pump in Mycobacterium bovis BCG." Antimicrobial agents and chemotherapy **56**(4): 2074-2083.

Rasko, D. A., M. J. Rosovitz, et al. (2008). "The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates." Journal of bacteriology **190**(20): 6881-6893.

Read, T. D. and D. W. Ussery (2006). "Opening the pan-genomics box - Editorial overview." Current opinion in microbiology **9**(5): 496-498.

Rindi, L., N. Lari, et al. (2009). "Evolutionary pathway of the Beijing lineage of Mycobacterium tuberculosis based on genomic deletions and mutT genes polymorphisms." Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases **9**(1): 48-53.

Sampson, S. L., P. Lukey, et al. (2001). "Expression, characterization and subcellular localization of the Mycobacterium tuberculosis PPE gene Rv1917c." Tuberculosis **81**(5-6): 305-317.

Shapiro, B. J., J. Friedman, et al. (2012). "Population genomics of early events in the ecological differentiation of bacteria." Science **336**(6077): 48-51.

Skuce, R. A., T. P. McCorry, et al. (2002). "Discrimination of Mycobacterium tuberculosis complex bacteria using novel VNTR-PCR targets." Microbiology **148**(Pt 2): 519-528.

Smith, N. H., R. G. Hewinson, et al. (2009). "Myths and misconceptions: the origin and evolution of Mycobacterium tuberculosis." Nature reviews. Microbiology **7**(7): 537-544.

Smith, N. H., K. Kremer, et al. (2006). "Ecotypes of the Mycobacterium tuberculosis complex." Journal of theoretical biology **239**(2): 220-225.

Snipen, L., T. Almoy, et al. (2009). "Microbial comparative pan-genomics using binomial mixture models." Bmc Genomics **10**.

Sreevatsan, S., X. Pan, et al. (1997). "Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination." Proceedings of the National Academy of Sciences of the United States of America **94**(18): 9869-9874.

Stinear, T. P., T. Seemann, et al. (2008). "Insights from the complete genome sequence of Mycobacterium marinum on the evolution of Mycobacterium tuberculosis." Genome research **18**(5): 729-741.

Supply, P., C. Allix, et al. (2006). "Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of Mycobacterium tuberculosis." Journal of clinical microbiology **44**(12): 4498-4510.

Swofford, D. L. (2003). "PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)."

Swofford, D. L. and W. P. Maddison (1987). "RECONSTRUCTING ANCESTRAL CHARACTER STATES UNDER WAGNER PARSIMONY." MATHEMATICAL BIOSCIENCES **87**(2): 199-229.

Swofford, D. L. and W. P. Maddison (1992). Parsimony, character-state reconstructions, and evolutionary inferences. Systematics, Historical Ecology, and North American Freshwater Fishes. R. L. Mayden, Stanford University Press**:** 186–223.

Szklarczyk, D., A. Franceschini, et al. (2011). "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored." Nucleic acids research **39**(Database issue): D561-568.

Tatusov, R. L., E. V. Koonin, et al. (1997). "A genomic perspective on protein families." Science **278**(5338): 631-637.

Tettelin, H., V. Masignani, et al. (2005). "Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome"." Proceedings of the National Academy of Sciences of the United States of America **102**(39): 13950-13955.

Tettelin, H., D. Riley, et al. (2008). "Comparative genomics: the bacterial pan-genome." Current opinion in microbiology **11**(5): 472-477.

Titgemeyer, F., J. Amon, et al. (2007). "A genomic view of sugar transport in Mycobacterium smegmatis and Mycobacterium tuberculosis." Journal of bacteriology **189**(16): 5903-5915.

Tortoli, E. (2006). "The new mycobacteria: an update." FEMS immunology and medical microbiology **48**(2): 159-178.

Tsolaki, A. G., S. Gagneux, et al. (2005). "Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of Mycobacterium tuberculosis." Journal of clinical microbiology **43**(7): 3185-3191.

Tsolaki, A. G., A. E. Hirsh, et al. (2004). "Functional and evolutionary genomics of Mycobacterium tuberculosis: insights from genomic deletions in 100 strains." Proceedings of the National Academy of Sciences of the United States of America **101**(14): 4865-4870.

van Ingen, J., Z. Rahim, et al. (2012). "Characterization of Mycobacterium orygis as M. tuberculosis complex subspecies." Emerging infectious diseases **18**(4): 653-655.

van Soolingen, D., L. Qian, et al. (1995). "Predominance of a single genotype of Mycobacterium tuberculosis in countries of east Asia." Journal of clinical microbiology **33**(12): 3234-3238.

Ventura, M., C. Canchaya, et al. (2007). "Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum." Microbiology and molecular biology reviews : MMBR **71**(3): 495-548.

Veyrier, F., D. Pletzer, et al. (2009). "Phylogenetic detection of horizontal gene transfer during the step-wise genesis of Mycobacterium tuberculosis." BMC evolutionary biology **9**: 196.

Veyrier, F. J., A. Dufort, et al. (2011). "The rise and fall of the Mycobacterium tuberculosis genome." Trends in microbiology **19**(4): 156-161.

Weniger, T., J. Krawczyk, et al. (2010). "MIRU-VNTRplus: a web tool for polyphasic genotyping of Mycobacterium tuberculosis complex bacteria." Nucleic acids research **38**(Web Server issue): W326-331.

Wirth, T., F. Hildebrand, et al. (2008). "Origin, spread and demography of the Mycobacterium tuberculosis complex." PLoS pathogens **4**(9): e1000160.

Wu, D., P. Hugenholtz, et al. (2009). "A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea." Nature **462**(7276): 1056-1060.

Xu, J., O. Laine, et al. (2007). "A unique Mycobacterium ESX-1 protein co-secretes with CFP-10/ESAT-6 and is necessary for inhibiting phagosome maturation." Molecular microbiology **66**(3): 787-800.

Yang, Z., S. Kumar, et al. (1995). "A new method of inference of ancestral nucleotide and amino acid sequences." Genetics **141**(4): 1641-1650.

Zheng, X., J. Guo, et al. (2011). "Crystal structure of a novel esterase Rv0045c from Mycobacterium tuberculosis." PloS one **6**(5): e20506.