FACULDADE · DE · CIÊNCIAS UNIVERSIDADE · DE · LISBOA
DEPARTAMENTO DE BIOLOGIA VEGETAL

# GENETIC DIVERSITY IN *MYCOBACTERIUM TUBERCULOSIS*

## The Beijing/W family

Olga Maria Elviro Mestre

**Mestrado em Biologia Molecular e Genética**

**2007**

FACULDADE · DE · CIÊNCIAS UNIVERSIDADE · DE · LISBOA
DEPARTAMENTO DE BIOLOGIA VEGETAL

# GENETIC DIVERSITY IN *MYCOBACTERIUM TUBERCULOSIS*

## The Beijing/W family

Olga Maria Elviro Mestre

Este trabalho foi realizado na Unidade de Genética Micobacteriana no Instituto Pasteur em Paris, sob a orientação da Profª. Brigitte Gicquel e do Profº. Mário Santos da Faculdade de Ciências da Universidade de Lisboa.

**Mestrado em Biologia Molecular e Genética**

**2007**

# INDEX

# INDEX OF FIGURES

# INDEX OF TABLES

## ACKNOWLEDGEMENTS

I would like to thank to my supervisors, Pr. Brigitte Gicquel Head of the 'Unité de Génétique Mycobacterienne' at the Institut Pasteur for accepting me as her Master student in her laboratory and to give me the opportunity to do this interesting work and Pr. Mário Santos from Faculty of Sciences at the University of Lisbon for accepting me as his Master student and for his support during this period.

I would like to thank the entire laboratory 'Unité De Génétique Mycobacterienne' at the Institut Pasteur, especially to Tiago Dos Vultos and Jean Rauzier for their crucial support and help in the development of this work.

I thank also Pablo Bifani, Kristin Kremer and Rasolofo Voahangy for providing the DNA of the isolates studied in the present work and Alan Murray for his important English corrections and his helpful comments on this manuscript.

I especially thank to all my family and friends for their support during this period.

## SUMMARY

The Beijing/W family of *Mycobacterium tuberculosis* represents a global threat in tuberculosis control. It has been frequently associated with drug resistance and its wide distribution suggests that these strains may have selective advantages over others.

On the present study a collection of Beijing/W strains isolated in different countries was analyzed under the genetic markers: Large Sequence Polymorphisms (LSPs) and Single Nucleotide Polymorphisms (SNPs).

Through a Multiplex-PCR isolates were tested for the presence or absence of the LSPs, described by a previous study for the Beijing/W family (RD105, RD181, RD150, RD142). The analysis of these regions proved to be very useful for defining and broadly subdivide the Beijing/W family in four groups. Furthermore, it provided a support for selection of representative isolates for the further study of other markers like SNPs in recently identified polymorphic DNA repair, recombination and replication (3R) genes.

Selected 3R genes were sequenced for representative Beijing/W isolates to search for variations and comparative analysis with *M.tuberculosis* H37Rv identified SNPs for all genes. Almost half corresponded to not previously described SNPs and a large proportion seemed specific to Beijing/W strains. SNPs were much more informative than LSPs distinguishing twenty-four groups for which a hypothetic phylogenetic network was constructed. Phylogenetic relations originated by SNPs were found to be congruent with the ones originated by LSPs and also with previous studies on *mutT2*, *muT4* and *ogt* putative DNA repair genes. Analysis according to geographic origin of isolates suggested that the collection used might be representative of the Beijing/W family. Certain genes were found to be more polymorphic than others with the accumulation of several non-synonymous SNPs (nsSNPs) that may potentially affect the function of the protein. Altered function in proteins involved in such important mechanisms in genome dynamics and stability might be associated with mutator phenotypes – strains having high mutation rates – in Beijing/W strains, as previously suggested. This might explain their higher adaptability and frequent association with resistance to antibiotics.

Overall, this study provided useful information that, although may need to be validated using a larger set of isolates, it could be the start point for several future studies. It may contribute to clarify the role of the Beijing/W family in the tuberculosis pandemic.

**Keywords:** *Mycobacterium tuberculosis*, The Beijing/W family, Large Sequence Polymorphisms (LSPs), Single nucleotide polymorphisms (SNPs), DNA repair recombination and replication (3R) genes.

# RESUMO

A tuberculose também chamada de tísica pulmonar ou "doença do peito" é uma das doenças infecciosas documentadas desde mais longa data. Apesar dos esforços feitos a nível global para combater a doença, a tuberculose permanece ainda um dos principais problemas de saúde pública a nível mundial. Em cada 15 segundos alguém morre de tuberculose o que significa aproximadamente 1.6 milhões de mortes todos os anos desta doença curável. Actualmente existem dois principais problemas em relação ao controlo da tuberculose: co-infecção com o vírus HIV e resistência da bactéria responsável pela doença aos antibióticos utilizados no seu combate.

O complexo *Mycobacterium tuberculosis* agrupa os diferentes agentes responsáveis pela tuberculose – *M.tuberculosis*, *M.bovis*, *M.africanum*, *M.microtti*, *"M.canetti"*, *M.caprae* e *M.pinipedii*. Estes apresentam diferentes hospedeiros, e potenciais zoonóticos com reservatórios de infecção em diferentes espécies animais. Nos humanos a tuberculose é principalmente causada pela espécie *Mycobacterium tuberculosis* (MTB).

Vários marcadores genéticos têm sido caracterizados e utilizados de forma a discriminar isolados do complexo MTB assim como para diferenciar e caracterizar estirpes específicas de MTB. Um dos métodos mais amplamente aplicados baseia-se na distribuição ao longo do cromossoma da sequência de inserção IS*6110* especifica do complexo MTB. Porém, a análise de padrões IS*6110*-RFLP (*Restriction Length Polymorphims*) tem algumas limitações e desvantagens e isto levou ao desenvolvimento de outras técnicas sendo o *Spacer oligonucleotide typing (spoligotyping*) e o *Variable Number Tandem Repeats* (VNTR) *typing* dos métodos mais frequentemente utilizados. Mais recentemente, análises genómicas comparativas identificaram *Single Nucleotide Polymorphisms* (SNPs) e *Large Sequence Polymorphisms* (LSPs) e vários estudos têm sido feitos com base na análise destes marcadores. Os LSPs correspondem a delecções genómicas enquanto que os SNPs são variações ao nível de uma única base no X. Informação obtida a partir destes marcadores é robusta, portátil e comparável, útil para estudos de variabilidade populacional e tem sido especialmente utilizada para inferir relações filogenéticas.

Em geral, apenas 5-10% das pessoas infectadas com a bactéria desenvolvem a doença em algum momento da sua vida. Uma das principais questões que tem ocupado os investigadores é então o que é que determina o desenvolvimento da doença nestes casos. Na verdade, parece existir uma complexa associação entre factores ambientais, factores relacionados com o hospedeiro e factores relacionados com a bactéria. Relativamente a estes últimos, dados epidemiológicos sugeriram que diferenças em termos de transmissibilidade e virulência entre diferentes estirpes de MTB podem estar relacionadas com o genoma dos organismos. Desta forma, a variabilidade genética de MTB pode

influenciar o desenvolvimento da doença. Na verdade, um grupo de estirpes geneticamente relacionadas, a família Beijing/W, tem demonstrado características particulares quando comparando com outras famílias de MTB, responsáveis pela transmissão da tuberculose.

A família Beijing foi descrita pela primeira vez em 1995. Uma grande proporção de estirpes de MTB foi encontrada na área de Beijing, na China, com semelhantes padrões de IS*6110*-RFLP e de *spoligotyping*. Mais tarde a estirpe multi-resistente encontrada em Nova York, estirpe "W", foi reconhecida como membro da família Beijing, tornando-se assim família Beijing/W. Membros desta família partilham várias características moleculares, e estirpes Beijing/W são actualmente principalmente definidas pelo seu padrão de *spoligotyping*. A sua vasta distribuição reflecte o seu sucesso e levou à hipótese de que estirpes Beijing/W podem apresentar vantagens selectivas em relação a estirpes de outras famílias de MTB. Estudos demonstraram ainda que estirpes desta família induzem uma patologia mais severa e tal sugeriu a existência de uma virulência aumentada nestas estirpes. Isto, aliado a frequente associação a resistência a antibióticos está a preocupar os investigadores e torna a família Beijing/W uma séria ameaça no que diz respeito ao controlo da tuberculose.

Através de hibridação total do genoma, estudos comparativos demonstraram a existência de LSPs associados à família Beijing/W. Estes correspondem a regiões genómicas presentes em MTB H37Rv mas variavelmente deletadas em estirpes Beijing/W. Designadas por *Regions of Difference* (RDs), quatro destas regiões, RD105, RD181, RD150 e RD142 permitem definir e ainda subdividir esta família em quatro diferentes grupos ou sublinhagens. Para além disto, foi ainda sugerido por outro estudo que podem existir diferenças entre estas sublinhagens na sua capacidade de transmissão e de causar doença.

Outro tipo de polimorfismos, SNPs, têm sido identificados em genes de reparação, recombinação e replicação (3R) de ADN em estirpes Beijing/W. Isto pode estar associado a maior adaptabilidade destas bactérias uma vez que a reparação, replicação e recombinação do DNA são importantes mecanismos na dinâmica e estabilidade do genoma. Variações nestes genes podem conduzir a um aumento nas taxas de mutação (estirpes mutadoras). Este estado mutador pode constituir uma vantagem selectiva uma vez que significa uma adaptação mais rápida em determinadas condições.

No presente estudo uma colecção de estirpes Beijing/W foi analisada usando os marcadores genéticos LSPs e SNPs.

Este trabalho incluiu 104 isolados clínicos de *M.tuberculosis* sendo que, de acordo com o *spoligotyping,* 101 pertenciam à família Beijing/W e 3 isolados eram "não-Beijing/W" usados como controlo. A origem geográfica dos isolados Beijing/W era diversa, com um grupo proveniente de Madagáscar e um grupo de vários países como: Holanda, Estados Unidos,

Coreia, África do Sul, China, Malásia, Mongólia e Tailândia. Ainda assim alguns isolados eram de origem desconhecida.

Numa parte inicial deste estudo a colecção de isolados Beijing/W foi analisada em termos de LSPs. Os respectivos isolados foram então testados quanto a ausência ou presença das regiões RD105, RD181, RD150 e RD142 previamente descritas para a família Beijing/W, como anteriormente referido. Para tal, foi realizado um *Multiplex-Polimerase Chain Reaction* (Multiplex-PCR), com 3 primers. De acordo com o tamanho dos produtos obtidos para cada isolado e em comparação com os produtos obtidos para *M. tuberculosis* H37Rv (usada como controlo positivo) foi possível determinar se cada uma destas regiões se encontrava presente ou ausente. Diferentes padrões de LSPs foram encontrados e a análise destas regiões demonstrou ser bastante útil não só para definir estirpes desta família como também para a sua divisão em 4 sublinhagens de acordo com o previamente descrito por Tsolaki *et al.* (2005). Para além disto esta análise permitiu a selecção de isolados para o seu posterior estudo usando outros marcadores, como SNPs em genes 3R*,* recentemente descritos como polimórficos em MTB.

Numa fase seguinte, um grupo de genes 3R foi seleccionado com base no seu carácter polimórfico para estirpes Beijing/W, descrito em estudos anteriores. Estes foram sequenciados de forma a procurar variações em isolados representativos de cada uma das quatro sublinhagens previamente definidas pelos LSPs. No total esta análise originou aproximadamente 1,6 Mbps de informação sob a forma de sequências e análise comparativa das mesmas, com as de *M.tuberculosis* H37Rv, permitiu identificar variações (SNPs) em todos os 22 genes analisados. Foram identificados 48 SNPs em 58 isolados Beijing/W e 1 isolado "não-beijing" sendo que aproximadamente metade destes corresponderam na verdade a SNPs nunca antes descritos para estirpes Beijing/W. Uma grande proporção (85%) foram sugeridos como sendo SNPs específicos de estirpes Beijing/W uma vez que se encontravam ausentes na estirpe "não-Beijing/W" usada como controlo. A análise dos SNPs demonstrou que estes marcadores foram mais informativos que os LSPs permitindo distinguir 24 grupos a partir dos quais uma hipotética árvore filogenética foi construída. Contudo, as relações filogenéticas originadas demonstraram ser congruentes com as originadas a partir da análise dos LSPs. Foi demonstrado ainda existir congruência entre os resultados deste estudo e os descritos em estudos anteriores para os putativos genes de reparação de ADN *mutT2*, *mutT4* e *ogt*. A análise dos resultados de acordo com a origem geográfica dos isolados e sugeriu que a colecção de isolados Beijing/W usada neste estudo possa ser representativa da família. Desta forma os SNPs descritos poderão ser ferramentas úteis para futuros estudos envolvendo estirpes da família Beijing/W.

Alguns dos genes analisados apresentaram-se mais polimórficos que outros com a acumulação de vários SNPs não-sinónimos. Estes por sua vez podem potencialmente afectar a função da proteína já que levam a uma mudança de aminoácido na sequência proteica. Como anteriormente referido, os genes analisados codificam para proteínas envolvidas em mecanismos importantes para a dinâmica e estabilidade do genoma. Desta forma, alterações na função destas proteínas pode estar associada ao aparecimento de estados mutadores – elevadas taxas de mutação – em estirpes Beijing/W como estudos anteriores sugeriram. Isto poderia explicar uma maior adaptabilidade e frequente associação com resistência a antibióticos em estirpes desta família.

Em conclusão, este estudo forneceu informação importante, que apesar de necessitar ser validada usando um maior número de isolados, poderá ser o ponto de partida para diversos estudos futuros. Como exemplo, poderão ser realizados estudos de forma a analisar uma possível relação entre determinados padrões de SNPs ou LSPs e virulências aumentadas assim como patogéneses mais severas, com base no estudo feito em ratos por Lopez *et al.* (2003) para diferentes famílias de MTB. Poderão ainda ser feitos estudos de associação entre alguns dos SNPs descritos em determinados genes e resistências a antibióticos assim como a fenótipos mutadores. Estes, entre outros estudos poderão contribuir para o aprofundamento do conhecimento acerca do papel desta família de sucesso na pandemia da tuberculose.

**Palavras-chave**: *Mycobacterium tuberculosis*, Família Beijing/W, *Large Sequence Polymorphisms* (LSPs), *Single Nucleotide Polymorphisms* (SNPs), genes de reparação, recombinação e replicação (3R) de ADN.

# INTRODUCTION

## Tuberculosis

- Brief History

Tuberculosis (TB) has claimed its victims throughout much of known human history. It is an ancient disease, with skeletal abnormalities of tuberculosis being found in Egyptian mummies from several thousand years ago. TB reached epidemic proportions in Europe and North America during the 18$^{th}$ and 19$^{th}$ centuries and it became the leading cause of death in most European countries (1, 2). In 1882, Robert Koch described the tubercle bacillus and in 1890 announced a compound (tuberculin) that inhibited the growth of tubercle bacilli in guinea pigs. Results of treatment with tuberculin were disappointing but the discovery of this substance was valuable for the diagnosis of TB. Nowadays partially purified derivative (PPD) of tuberculin is still used for the Tuberculin Skin Test which is used to determine if prior exposure to the tubercle bacilli has occurred. In 1921, Albert Calmette and Camille – Guérin developed the live attenuated vaccine Bacille Calmette – Guérin (BCG), which is still the only widely available TB vaccine. Campaigns of mass BCG vaccination were promoted as a mean of controlling the disease by the World Health organization (WHO) in the late 1940's. In 1943, streptomycin was reported as the first antibiotic and first bactericidal agent against the tuberculosis bacilli. In 1952 and 1965 isoniazid and rifampicin were discovered also as anti-TB drugs, respectively (for a review (1)).

- A global health emergency

With tools available to control disease, TB-control programs were established between 1950 and 1965 although these initiatives have had limited success in developing countries. The cost of mass case finding and specialized case management was far beyond the resources of those countries (3). In industrialized countries there was a pronounced reduction of infection and death rates. TB programs became a lower priority because the disease was considered close to elimination. In the early 1980s, cases of TB began to rise in industrialised countries because of factors such as increases in prison and homeless populations, intravenous drug use and especially Human Immunodeficiency Virus (HIV)/Acquired Immunodeficiency Syndrome (HIV/AIDS) epidemics. Indeed, the HIV/AIDS epidemic has had a devastating effect on TB control worldwhile. By 1993 increasing TB could no longer be ignored so the WHO declared the disease a global emergency (2). In 2000, the Word Health assembly endorsed the establishment of a Global Partnership to Stop TB with several targets for the next 50 years.

Global efforts have been made to combat the disease but TB remains a major public-health problem. Every 15 seconds someone dies from tuberculosis (TB), meaning 1.6 million deaths are caused each year from this curable disease. The slow decline in incidence rates per capita is offset by population growth. Consequently, the number of new cases arising each year is still increasing globally and in the WHO regions of Africa, the Eastern Mediterranean and South-East Asia (2, 5, 6). Nowadays in addition to a lack of resources for case finding and treatment there are two major problems regarding control of tuberculosis: coinfection with HIV and resistance of the bacillus to the currently used regimen of antibiotics.

## The tubercule bacillus

The *Mycobacterium* genus bacteria are non-motile and non-sporulated rods. They are grouped in the suprageneric rank of actinomycetes that have high content (61-71%) of guanine plus cytosine (G+C) in the genomic desoxiribonucleic acid (DNA), and high lipid content in the wall. Several mycolic acids in envelope structure distinguish the mycobacteria from other closely related genera. The waxy coat confers the distinctive characteristics of the genus: acid fatness, extreme hydrophobicity, resistance to injury, including that of many antibiotics, and distinctive immunological properties.

| | |
|---|---|
| Kingdom | Bacteria |
| Phylum | Actinobacteria |
| Class | Actinobacteria |
| Subclass | Actinobacteridae |
| Order | Actinomycetales |
| Suborder | Corynebacterineae |
| Family | Mycobacteriaceae |
| Genus | *Mycobacterium* |
| | unique genus |
| Species | *M. tuberculosis* |
| | *M. bovis* |
| | *M. africanum* |
| | *M. microti* |
| | *"M. canettii"* |
| | *M. caprae* |
| | *M. pinnipedii* |

**Figure I** – Lineage of the agents of tuberculosis. Adapted from the textbook 'Tuberculosis 2007 – From basic Science to Clinical care' (6).

Only a few species of mycobacteria have became successful pathogens of higher vertebrates, preferentially inhabiting the intracellular environment of mononuclear phagocytes. Members of the *Mycobacterium tuberculosis* complex are among the host-dependent mycobacteria.

The *M.tuberculosis* complex, generally referred to as the tubercle bacillus, comprises the various etiologic agents of tuberculosis – *M.tuberculosis*, *M.bovis*, *M.africanum*, *M.microti*, "*M.canettii*", *M.caprae* and *M.pinnipedii*. They exhibit host preference, and zoonotic potential

and with reservoirs of infection existing in various animal species. Tuberculosis in Humans is mainly caused by *M.tuberculosis* (MTB). Based on several genetic markers, such as the ones described in the next section, a number of MTB genotype families have already been identified (6).

- Molecular genetics of the *Mycobacterium tuberculosis* complex

Several polymorphic or hypervariable genetic markers have been characterized and used to discriminate MTB complex isolates and even sub-speciate clinical isolates of MTB.

One of the most widely applied molecular typing methods applied in MTB complex isolates has been the IS*6110* Restriction Fragment Length Polymorphism (RFLP) typing. IS*6110* is an insertion sequence specific for strains belonging to the MTB complex (7). Individual strains vary in the number of copies and in the sites into which IS*6110* is inserted. A method using a restriction enzyme has been developed and IS*6110* Southern blots of the digested DNA separed by electrphoresis on agarose gels provides a RFLP pattern (8). Unfortunately, this method has some limitations and disadvantages such as poor discrimination between isolates with low copy numbers (<5) of IS6110 and its even absence from certain strains. In addition, it is a technically demanding and expensive technique (9). This has led to the development of other methods including spacer oligonucleotide typing (Spoligotyping) (10), variable number of tandem repeat (VNTR) typing (11), polymorphic GC-rich repetitive sequence (PGRS) typing (12) and mycobacterial interspersed repetitive units (MIRU) typing (13). Spoligotyping and VNTR typing are most frequently used. These techniques have been especially important and useful genotyping methods used in molecular epidemiological studies and also in molecular evolution studies (6).

More recently, systematic comparative genomic analysis have identified Single Nucleotide Polymorphisms (SNPs) and Large Sequence Polymorphisms (LSPs). This kind of polymorphisms provide robust, portable, and comparable data for studying population variation and have been especially used to infer phylogenetic relations (14, 15, 16, 17, 18, 19). LSPs, correspond to genomic deletions that have been suggested as an important mechanism for genetic variation in MTB. Affected sequences may have functional relevance in pathogenesis and immunity since deleted sequences can include putative open reading frames as well as intergenic regions and housekeeping genes (20). LSPs may be informative markers for discrimination of strains and have been, therefore, proposed for genotyping and also for constructing phylogenies (21). Also SNPs contribute to strain variability by affecting expression of genes, for example drug-resistance characteristics, which are mainly due to SNPs in MTB (22). Therefore SNPs have, as well, provided researchers with markers to differentiate strains and to study the phylogenetic relatedness of MTB strains (20).

- Pathogenesis

The bacillus is estimated to infect one-third of the world's population (4) and is transmitted primarily via the respiratory route. Infection occurs by inhaling particles, expelled from infectious patients, containing tubercle bacilli. Most microorganisms are trapped in the mucosa of the upper respiratory tract, trachea and bronchi, and are eliminated by mucocilliary defense mechanisms. However, droplet nuclei small enough to avoid these mechanisms penetrate into the lower respiratory tract, especially inside the alveoli. Macrophages are the main targets for infection by MTB and, at the same time, the first line of defense against the bacilli. Alveolar macrophages phagocytose the MTB bacillus but if the macrophage is not capable of arresting bacterial growth, development of TB will occur a few months after infection. Disease during this initial phase is called primary TB (23, 24, 6).

In most of the cases individuals are able to mount a robust immune response, culminating in the formation of a cellular mass, the granuloma ('tubercle') that apparently contains MTB bacilli. A kind of host-parasite balance is achieved on which the bacteria appears dormant. This latent, asymptomatic state can persist for long periods of time. A rupture in this host-parasite balance results in post-primary tuberculosis (6, 23, 24).

- Progression to disease

In general, only 5-10% of people who are infected with TB bacilli will develop clinical disease at some time during their lifetime (3, 4). One of the principal questions that has occupied researchers is what determines development of TB. It appears to be a complex interplay of host, bacterial and environmental factors that influences the outcome of the disease.

In general, diseases and conditions that weaken immunity such as for example malnutrition, alcohol, air pollution, stress, diabetes mellitus and silicosis play a role in determining the development of TB (3, 6, 25). None of these risk factors is, however, as important as HIV infection, which is one of the top-ranking risk factors on the World Health Organization (WHO) list (4).

Host genetic factors clearly influence the course of the disease. Several genes have been identified has being associated with susceptibility (26, 27) although this is a complex subject confused by conflicting results obtained from studies on different populations. The risk of progressing from infection to symptomatic disease is polygenic, determined by the balance between the additive and even synergistic effects of susceptibility genes, and protective genes (26).

The host immune response against infection is one of the most investigated factors and progression to disease has been related to a switch to expression of Th2 cytokines (28). One study about immunological and pathological comparative analysis between experimental

latent tuberculosis and progressive pulmonary tuberculosis in mice revealed that high numbers of bacteria were related with progressive disease in contrast with infection with very low doses of bacilli which produced a latent infection (29). Besides bacterial load as suggested by this study, pathogenicity/virulence of the bacilli is also a factor that can influence the development of tuberculosis (30). In fact, epidemiological data suggested that differences in transmissibility and virulence among MTB strains might be related to the genetic background of the organisms (31, 32). Therefore, genetic variability of MTB may have a role in the outcome of the disease. The Beijing/W genotype, a group of genetically related MTB strains has exhibited particular properties when comparing to the others MTB families responsible for transmission of TB.

## The Beijing/W family

The Beijing family was first described in 1995. A large proportion of MTB strains in the Beijing area of China had highly similar multi-banded IS*6110* RFLP patterns and identical spoligotyping patterns (33). In the second half of the 1990's the multidrug resistant MTB strain identified in New York, "W" strain, was recognized as a member of the Beijing genotype family (34, 35) which became the Beijing/W family.

Members of this family share various distinct molecular characteristics (36):

- **a)** Principal genetic group 1 of Sreevatsan - Sreevatsan *et al.* (1997) (36) have suggested a subdivision of MTB strains into three principal genetic groups (G-1, G-2, G-3) based on single-nucleotide polymorphisms (SNPs) in two gene codons, *katG*463 and *gyrA*95 (36).
- **b)** Similar IS*6110*-RFLP patterns (copy number 15-26),
- **c)** IS*6110* insertion in the origin of replication (A1 insertion)
- **d)** IS*6110* insertion in the NTF region
- **e)** Spoligotype pattern

Beijing/W lineage strains are nowadays mainly defined by the spoligotype pattern. The Direct-Repeat (DR) locus is a single region of MTB chromosome that contains multiple copies of a 36-bp direct repeat, separated by spacer DNA with various sequences. MTB strains have the same overall arrangement of spacers but differ in terms of presence or absence of specific spacers. Spoligotyping involves Polymerase Chain Reaction (PCR) amplification of the DR locus, followed by hybridization of the labeled products to a membrane that contains covalently bound oligonucleotides corresponding to each of 43 variable-spacer sequences present in the laboratory strain *M.tuberculosis* H37Rv, and the attenuated *M.bovis* BCG strain, used as a vaccine for tuberculosis. Beijing/W strains

spoligotype patterns correspond to hybridization of at least three of the nine spacers, 35 to 43 with an absence of hybridization to spacers 1 to 34 (37).

The global distribution reflects their success and this have led to the hypothesis that Beijing/W strains may possess selective advantages over other MTB strains (35, 38).

Lopez *et al.* (2003) (39) have demonstrated that this genotype induces a more severe pathology related with high bacillary loads and the host immune responses driven towards a non-protective pattern of cytokine production. Relation between this genotype and increased virulence was suggested also by other studies (32, 40, 41) and may be related to the success of this specific group of strains. In the same study BCG was less protective against this genotype and for some authors it has been considered even as a risk factor for TB in populations infected with Beijing/W strains (42). This allied with its frequent association with drug resistance (43, 38) is alarming researchers and makes the Beijing/W family a serious threat to TB control.

By comparative whole-genome hybridization of Beijing/W strains Tsolaki *et al.* (2005) (44) have shown that there are LSPs associated with this family. These correspond to genomic regions variably deleted in Beijing/W strains in relation to *M.tuberculosis* H37Rv and have been previously designated regions of difference (RDs). RD105, RD181, RD150, RD142, were found to, not only define but also subdivide this family into four different subgroups or lineages (figure II). In another study, it was also suggested that the Beijing/W family can be divided into different lineages and there might even exist differences between the lineages in their ability to spread and cause disease (45). Therefore, individual lineages within this family may have evolved specific genetic characteristics that may even reflect different pathogenic characteristics.

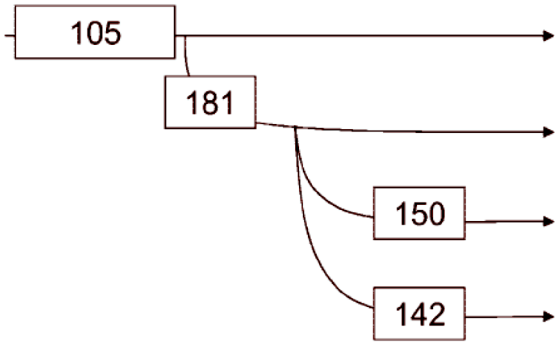


**Figure II** – Phylogeny of Beijing/W lineage based on regions of difference (RDs) found for this family. Adapted from Tsolaki *et al.* (2005) (44).

Polymorphisms have been identified in DNA repair, recombination and replication (3R) genes in Beijing/W strains (46, 22). This can be related to the higher adaptability of these bacteria because DNA repair, recombination and replication are major mechanisms in genome dynamics and stability (47). Variations on these genes can even led to higher

mutation rates (mutator strains), and high mutation rate may be, in certain conditions, a selective advantage once it means faster adaptation (48). These genes have been found to be highly polymorphic MTB strains so their analysis can provide useful information for genetic variability studies.

## OBJECTIVES

In the present study a collection of Beijing/W strains was analyzed using the genetic markers LSPs and SNPs. Analysis of the LSPs previously described by Tsolaki *et al.* (2005) (44) for the Beijing/W family (RD105, RD181, RD150, RD142) was performed along with search for SNPs in selected 3R genes based on previous studies (22). Information on the frequency and diversity of polymorphisms can provide useful information about the variability and the phylogenetic relations that may exist within this successful family. In addition, these markers may also provide useful information for future studies that may contribute to better understand the role of the Beijing/W strains in the worldwide epidemic of tuberculosis.

## MATERIALS AND METHODS

### Study isolates

This study involved a total of 104 clinical isolates of *M.tuberculosis*. On the basis of spoligotyping 101 isolates were Beijing/W – hybridization of at least three of the nine spacers 35 to 43 with an absence of hybridization to spacers 1 to 34 (37), and 3 were non-Beijing/W used as controls. The DNA of each isolate was provided from different sources (except for *M.tuberculosis* H37Rv available in the laboratory): 21 isolates from Pasteur Institute in Madagascar isolated in Madagascar (designated - MAD), 82 isolates from both the RIVM Institute (Netherlands) (designated - R), and kindly provided by Pablo Bifani (designated - PB) from different countries of origin – USA, The Netherlands, Korea, South Africa, China, Malaysia, Mongolia, Thailand. Some of the isolates were from unknown origin.

### LSPs analysis

Tsolaki *et al.* (2005) have described the Beijing/W family LSPs, designated regions of Difference (RDs), RD105, RD181, RD150, RD142, that correspond to genomic regions present in *M.tuberculosi*s H37Rv but variably deleted in Beijing/W strains (44).

Strains were tested for the presence or absence of these RDs through a Multiplex PCR with 3 primers (P1, P2 and P3). The same method was meanwhile described by Jing Chen *et al.*

(2007) as a Deletion-Targeted Multiplex PCR (DTM-PCR) for detection of the RD105 deletion (49). We have applied this method to detect not only RD105 but all the referred RD regions. One pair of primers (P1 - forward and P2 - reverse) was designed flanking each one of the RD regions. A third primer (P3 – internal reverse) is an internal primer hybridizing inside the RD region. *M.tuberculosis* H37Rv was used as a positive control for the presence of these regions and water was used as a negative control. If the region is present, amplification with the internal primer should occur (P1+P3). When the region is absent an amplification product is produced from the flanking primers (P1+P2). This PCR product will be a different size when compared to *M.tuberculosis* H37Rv (in which the region is present). In table IV are described the expected sizes (base paires (bp)) of products resulting from P1+P2 or P1+P3 amplification for each RD region tested. Primers were tested and prediction of amplification products size was calculated by software that simulates PCR (Amplify v.3.1.4, Bill Engels, University of Winsconsin).

Multiplex-PCR reactions were performed using different sets of primers for each RD, listed on table I in a total volume of 25 µl of general reaction mixture for *TaKaRa LA taq* with GC buffer (TaKaRa, Shiga, Japan) with 0.4 uM of each of the 3 primers (P1, P2 and P3) and approximately 20ng of DNA template (or water as a negative control) (table II).

**Table I** – List of primers used in Multiplex-PCR for Large Sequence Polymorphisms (LSPs) analysis. P1 – forward primer, P2 – reverse primer, P3 – internal reverse primer.

| Region of difference (RD) | Primer sequence (5' – 3') | Melting Temperature (Tm) |
|---|---|---|
| **RD105** | P1(105) - GGAGTCGTTGAGGGTGTTCATCAGCTCAGT | 67 °C |
| | P2(105) - GCGCCAAGGCCGCATAGTCACGG | 70 °C |
| | P3(105) - GGTTGCCCACTGGTCGATATGGTGGACTT | 69 °C |
| **RD181** | P1(181) - CGCAACGGCCGCGGTGAACTCT | 69 °C |
| | P2(181) - CGGGCGGCTGCGGGAACCTT | 70 °C |
| | P3(181) - CTGGCCTGGTTCGGCTTGGTCCC | 68 °C |
| **RD150** | P1(150) - TGTGGCGTGGCTCGGCAAATAG | 64 °C |
| | P2(150) - CGGGACGGCAAACGGGTGAT | 66 °C |
| | P3(150) - CGGCGTCATCGCGTATCTGA | 64 °C |
| **RD142** | P1(142) - CGTCCGCGACGACGAACAA | 62 °C |
| | P2(142 - TCACTTTCCATTTCCAGCGGCAACT | 64 °C |
| | P3(142) - CGTCGATGGCAACACACCGAAA | 62 °C |

**Table II** – General PCR reaction mixture used in Multiplex – PCR for Large Sequence Polymorphis (LSPs) analysis.

| PCR Reaction mixture | |
|---|---|
| **2x GC Buffer** | 12,5 µl (1x) |
| **dNTPs** | 4 µl (400 µM) |
| **Primers (P1, P2, P3)** | 1 µl each (0,4 µM each) |
| **Sterilized distilled water** | 2,85 µl |
| **Template** | 2,5 µl (~ 20 ng) |
| **TaKaRa LA *Taq*** | 0,15 µl (0,75 U) |
| **Total** | 25 µl |

The thermocycling parameters varied depending on the RD target as follows:

**RD105**, 1 cycle at 94°C for 1 min, followed by 30 cycles of 94°C for 30 sec, 63°C for 30 sec, and 72°C for 3 min, with completion with a final cycle at 72°C for 10 min.

**RD181**, 1 cycle at 94°C for 1 min, followed by 30 cycles of 94°C for 30 sec, 63°C for 30 sec, and 72°C for 1min 50 sec, with completion with a final cycle at 72°C for 10 min.

**RD150**, 1 cycle at 94°C for 1 min, followed by 30 cycles of 94°C for 30 sec, 59°C for 30 sec, and 72°C for 3 min, with completion with a final cycle at 72°C for 10 min.

**RD142**, 1 cycle at 94°C for 1 min, followed by 30 cycles of 94°C for 30 sec, 57°C for 30 sec, and 72°C for 3 min, with completion with a final cycle at 72°C for 10 min.

Multiplex - PCR products were analysed following electrophoresis on a 0.8% agarose gel using 1X TBE buffer. The presence or absence of deletions was determined by the size of the product (table III) and by comparison with *M.tuberculosis* H37Rv respective products, using the molecular weight marker 'Smart Ladder' (Eurogentec s.a., Belgium) (figure A1 – Annex).

**Table III** – Expected products from Multiplex-PCR amplification of regions of difference (RDs). P1+P2, flanking primers, or with P1+P3, hibridization inside the RD in case is present.

| Region of difference (RD) | Expected product size (base pairs) P1+P2 | P1+P3 |
|---|---|---|
| **RD105** | 787 | 1495 |
| **RD181** | 1002 | 639 |
| **RD150** | 780 | 1385 |
| **RD142** | 791 | 1376 |

## Sequencing of target genes and detection of SNPs

- **Selection of genes for sequencing**

  Variations were searched for in 3R genes because they have previously been described as polymorphic in *M.tuberculosis* strains (22). In addition, variations have also been reported for Beijing/W strains in some of these genes (46, 22). Using this information a group of 26 3R genes was selected. These genes were sequenced for 66 representative isolates of the Beijing/W family selected based on the analysis of the RDs, along with two Non-Beijing/W strains.

- **Sequencing**

  - Amplification of genes for sequencing

  Each target gene was first amplified, for each strain, by a PCR using the primers listed on table A1 (Annex). General reaction mixture for TaKaRa LA *Taq* with GC Buffer was prepared in a total volume of 25µl according to the manufacter's instructions, using approximately 20ng of DNA template (or water as a negative control). Table IV describes

in detail the reaction mixture. General thermocycling parameters were as follows: 1 cycle at 94°C for 1 min, followed by 30 cycles of 94°C for 30 sec, Annealing temperature (Ta °C) for 30 sec, and 72°C for Extension time (ET) min, with completion with a final cycle at 72°C for 10 min, holding at 4°C until use. Annealing temperature (Ta) varied according to primers melting temperature (Tm) (Ta = Tm – 5 ºC). Extension time (ET) depended on gene length and processing time of DNA polymerase (~1Kb/2min). Table A1 (Anex) lists information about genes and primers used.

An aliquot of each reaction product was loaded on a 0.8% agarose gel for electrophoresis in 1X TBE buffer. Correct amplification of the proper target was confirmed using the molecular weight marker 'Smart Ladder' (Eurogentec s.a., Belgium). In addition, this marker was useful to approximately quantify the PCR samples. Nonspecific products as primer-dimer artifacts and secondary PCR products, which can affect the quality of sequencing data, could be also visualized on the agarose gel.

**Note:** All reactions were performed in 96-well plates to maximize the number of reactions that could be done in each experiment.

**Table IV** - General PCR reaction mixture used for amplification of genes to sequence.

| PCR Reaction mixture | |
|---|---|
| **2x GC Buffer** | 12,5 µl (1x) |
| **dNTPs** | 4 µl (400 µM) |
| **Primers (P1, P2)** | 1 µl each (0,4 µM each) |
| **Sterilized distilled water** | 2,85 µl |
| **Template** | 2,5 µl (~ 20 ng) |
| **TaKaRa LA Taq** | 0,15 µl (0,75 U) |
| **Total** | 25 µl |

- Purification of PCR products

Amplification products have to be purified before sequencing. The purification method aim is to remove unincorporated dNTPs and primers during the PCR reaction that can interfere with the sequencing reaction. Excess PCR primers compete with the sequencing primer for binding sites and reagents in the sequencing reaction and results in noisy data. Excess of dNTPs can affect the balance of the sequencing reaction, resulting in decreased termination in shorter extension fragments. The purification method adopted was Shrimp Alkaline Phosphatase (SAP) and Exonuclease I (Exo I) Treatment based on the method described by the 'Automated DNA Sequencing – Chemistry Guide' – Applied Biosystems (50). Werle *et al.* (1994) (51) had first described this method, Exo I degrades excess of primers while SAP is responsible for de-phosphorylation of the excess dNTPs. After 1 hour of incubation at 37°C and 15 min at 72°C to inactivate enzymes, purified products were prepared for using in sequencing reaction. Table V describes in detail mixture for purification of the same volume of PCR products. PCR product may need to be diluted before sequencing, depending on the

quantification of the sample determined by comparison of the PCR product with the molecular weight marker on the agarose gel.

**Table V** – Reaction mixture for purification of same volume of PCR amplified products.

| PCR purification Reaction Mixture | |
|---|---|
| **Exol** | 0,5 µl (0,5 U) |
| **Sap** | 0,5 µl (0,5 U) |
| **Sterilized distilled water** | 4 µl |
| **Total** | 5 µl |

- Sequencing reaction

Purified PCR reaction products were used to perform DNA sequencing by the dideoxy chain-termination method (52) using the Big Dye Terminator v3.1 cycle sequencing Kit (Perkin Elmer Applied Biosystems, Courtaboeuf, France). Cycle sequencing is a simple method in which successive rounds of denaturation, annealing and extension in a thermal cycler result in linear amplification of extension products. There are many advantages of cycle sequencing and it provides the most reproducible results for sequencing PCR templates. The Kit provides the reagents for performing fluorescence-based cycle sequencing reactions: dye terminators, deoxynucleotides triphosphates (ddNTPs), DNA polymerase, magnesium chloride, and buffer. With dye terminator labeling, each of the four ddNTPs is tagged with a different fluorescent dye. Each of the growing chains is simultaneously terminated and labeled with the dye that corresponds to that base (Figure III). The DNA sequencers detect fluorescence from four different dyes that will correspond to each one of the four bases A, C, G, T .

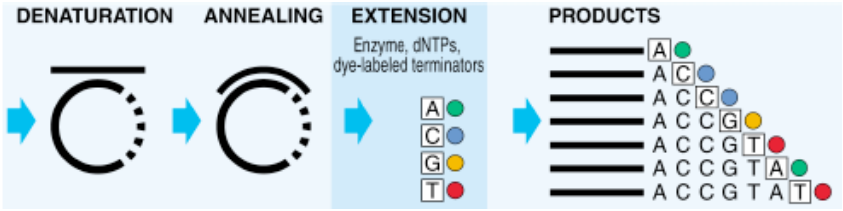


**Figure III** – One cycle of dye terminator cycle sequencing. Adapted from 'Automated DNA Sequencing – Chemistry Guide' Applied Biosystems (2000) (50).

Table VI describes reaction mixture for sequencing in a 20 µl final volume - 10 µl of PCR purified template, previously diluted (if needed), was added to 10 µl of sequencing reaction. Primers used to sequencing are listed on Table A1. Thermocycling parameters used were according to the manufacter's instructions on a GeneAmp PCR System 9600: 25 cycles of 96°C 10 sec, 50°C 5 sec and 60°C 4 min, final step holding at 4°C. Sequencing products correspond to fluorescently labeled DNA fragments.

**Table VI** – Sequencing Reaction mixture for amplified PCR genes.

| Sequencing Reaction mixture | |
|---|---|
| **Terminator Ready reaction mixture** | 1,5 µl |
| **BigDye Sequencing Buffer** | 4 µl (1x) |
| **Primer** | 0,3 µl (3,2 pmol) |
| **DMSO** | 2 µl (5% (v/v)) |
| **Template** | 10 µl |
| **Sterilized distilled water** | 1,2 µl |
| **Total** | 20 µl |

- Purification of extension products

Purification of products resulting from sequencing reaction is extremely important in order to remove unincorporated dye terminators during sequencing reaction that can obscure data. Extension products from sequencing reaction were purified by ethanol precipitation. Ethanol was added to the 20µl of sequencing reaction products at a final concentration of 60% +/-3%. The mixture was kept for at least 15 min at room temperature, followed by centrifugation at maximum speed (4000 rpm) for 15 min to precipitate the extension products. Supernatants were discarded by inversion of plates on a paper towel. Inverted plates in a towel were centrifuged at 1000rpm for 1 minute.

After wells were well dried, pellets were re-suspended with the appropriate buffer for capillary electrophoresis (water or formamide).

- Capillary electrophoresis

Re-suspended DNA fragments fluorescently labeled were run in an ABI PRISM 3100 Genetic Analyzer for automated capillary electrophoresis. Briefly, the sequencing reaction sample 96-well plates were placed in an appropriate autosampler tray. The autosampler successively brings each sample into contact with the cathode electrode on one end of a capillary filled with separation polymer. An anode electrode at the other end of the capillary is immersed in buffer. The sample enters the capillary as current flows from the cathode to the anode. When the DNA fragments reach a detector window in the capillary, a laser excites the fluorescent dye labels. Emitted fluorescence from the dyes is collected once per second by a camera at particular wavelengths and stored as digital signals on a computer for processing. The Sequencing Analysis software interprets the result, calling the bases from fluorescence intensity at each data point.

- Alignment of sequences and detection of polymorphisms

Sequence data files of the target genes obtained were analyzed with the software Genalys (53) in order to identify variations. Hence the sequences of target genes of each isolate were compared with the reference sequence for the same genes of the laboratory strain *M.tuberculosis* H37Rv, obtained from Institut Pasteur at http://genolist.pasteur.fr

(54). Multiple sequences can be introduced on this software for simultaneous alignments and detection of SNPs and other polymorphisms such as indels. Although the software automatically corrects and aligns the sequences, confirmation and manual correction is still needed to minimise errors.

Identified SNPs were correctly noted down by indicating position and codon on the gene. For each gene, the DNA sequence was translated in frame, and each nucleotide polymorphism was classified as synonymous SNPs (sSNPs), in cases where the mutation didn't lead to an amino acid change or non-synonymous SNPs (nsSNPs), if the mutation lead to an amino acid change (example in Figure A2 - Annex).

### SNP Data analysis and Phylogenetic tree construction

The SNPs were concatenated resulting in one character string (nucleotide sequence) for each clinical isolate analyzed, a FASTA file was created to run in the software. DNA Sequence Polymorphism (DnaSP) software was used to display general information of the polymorphisms data file.

This Data file was then used to build a phylogenetic tree by the Network software based on the Media-Joining (MJ) method. Using parsimony – selects the trees that require the fewest evolutionary changes - MJ network begins with the minimum spanning trees all combined within a single network. A minimum spanning tree for a set of sequence types connects all given types without creating any cycles or inferring additional nodes, such that the total length is minimal. Median vectors (mv) might be created and can be interpreted as possibly extant unsampled sequences or extinct ancestral sequences. This software assumes that there is no recombination between genomes (55).

## RESULTS

### Detection and analysis of LSPs regions

A set of Beijing/W strains isolated in several different countries were tested for the presence or absence of specific Large Sequence Polymorphisms (LSPs) previous described by Tsolaki *et al.* (2005) (44). These, designated Regions of Difference (RDs), RD105, RD181, RD150, RD142, correspond to genomic regions present in *M.tuberculosi*s H37Rv but variably deleted in Beijing/W strains, as already refered (44). Through a Multiplex-PCR with 3 primers it was possible to determine whether each one of those regions was present or absent in each of the 101 Beijing/W isolates of this study. Amplification should occur with an internal primer in case the region is present. In contrast, if the region is absent the product expected should be the one resulting from amplification with the flanking primers. Analysis of product sizes (see

table III material and methods for expected product sizes for each RD region) and comparison with the positive control (*M.tuberculosis* H37Rv), for which all regions are present, enabled the presence or absence of the RD region to be determined in each of the isolates as shown below.

- **RD 105 (2488 bp)**

Except for 2 strains for which no amplification could be obtained, all strains gave a 787 bp size product for region RD105. This corresponds to the expected size of amplification products with the flanking primers (i.e. this region was deleted) (table III). This is a different product from the one obtained for the positive control, for which the region is present, and consequently amplification with the primer hybridizing inside the region results in a 1495 bp product (figure IV).

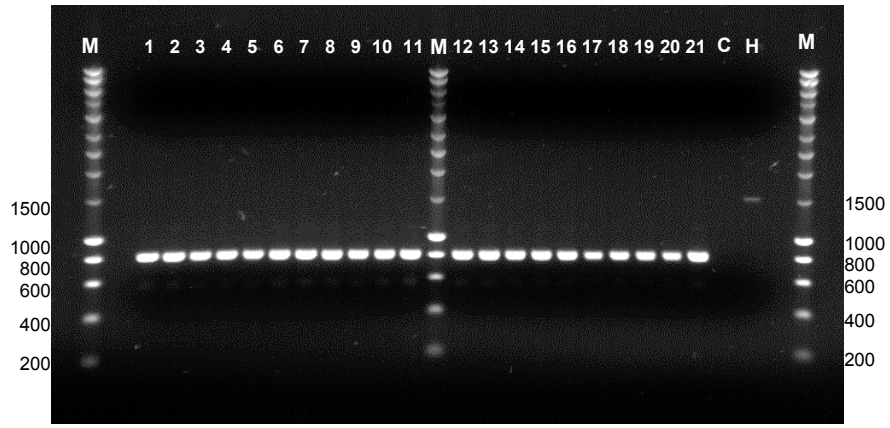


**Figure IV** – RD105 Multiplex-PCR analysis.
All the 101 Beijing/W isolates generated fragments of 787 bp in size as exemplified in the figure: Lanes: 1-21. This corresponds to amplified products in case the RD105 region is deleted. The positive control *M.tuberculosis* H37Rv (H), for which region is not deleted, generated a 1495 bp fragment.
Lanes 1-21: Beijing/W isolates. C: water as negative control. M: weight molecular marker, respective band sizes are shown on both sides of the figure.

According to these results we can conclude that all strains, except for the referred 2 strains carry deletion of RD105 region.

- **RD 181 (1713 bp)**

For region of difference RD181, most of the isolates gave a 1002 bp amplification product, which corresponds to a deletion. However 3 isolates had a 639 bp size product, as in the case of the positive control (figure V). For these strains it means that the region is not deleted and therefore amplification resulted from the internal primer. Conditions favored the production of only one amplification product as a result of the Multiplex PCR. However an additional product of approximately 1713 bp can be seen in lanes 11, 15, 20 and H (figure V). This represents amplification by the flanking primers across the RD region in addition to amplification also with the internal primer.

For one isolate represented in figure V (lane 16) results were not consistent, even after several repetitions. No conclusions could be achieved with this result so this isolate was excluded from the analysis.

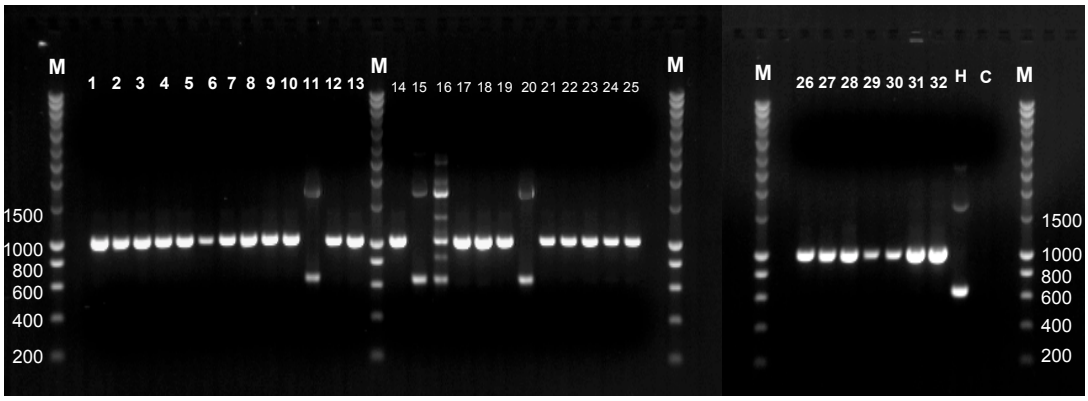


**Figure V** – RD181 Multiplex PCR analysis.
Most of the isolates generated a fragment of 1002 bp size. Few isolates (exemplified in the figure lanes 12, 15 and 21) originated a 639 bp size fragment the same as for the positive control *M. tuberculosis* H37Rv (H). This means that those isolates, in opposite to the others, don't have deletion of RD181 region.
Lanes: 1-32 Beijing/W isolates. C: water as a negative control. M: weight molecular marker, respective band sizes are shown on both sides of the figure.

- **RD 150 (2488 bp) and RD142 (2852 pb)**

According to Tsolaki *et al.* (2005) (44), Beijing/W strains with deletion of RD181 may also have deletion of one of two other regions of difference: RD150 or RD142 (see figure II Introduction).

Six of the strains with deletion of RD181 had also deletion of RD150. Amplification products resulting from Multiplex PCR for these strains had the expected size (785 bp) signifying deletion of this region (see figure VI lane 3 for an example). The remaining isolates gave a 1385 bp product (as with the positive control H37Rv), the expected product when the region is present (figure VI).

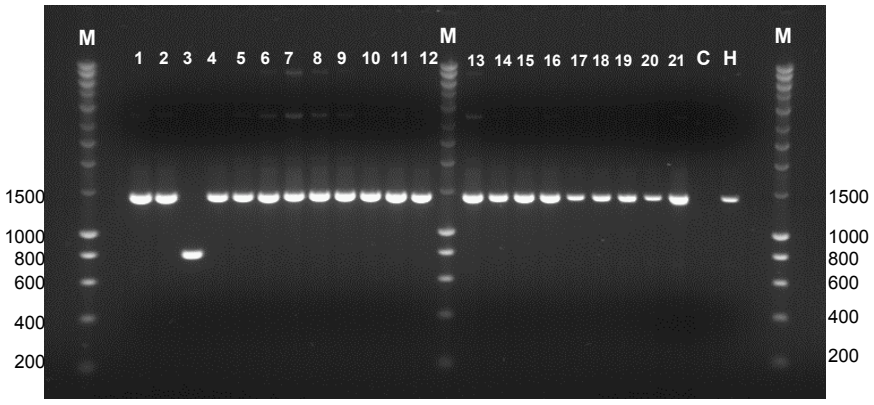


**Figure VI** – RD150 Multiplex-PCR analysis.
Almost all isolates presented a 1385 bp size product, as the positive control, represented in the figure as H (*M.tuberculosis* H37Rv). Still for some strains, multiplex PCR resulted in a product of different size, 785 bp as exemplified in this figure (lane 3). An example of some of those isolates is shown on this figure.
Lanes: 1-21 Beijing/W isolates. C: water as a negative control. M: weight molecular marker, respective band sizes are shown on both sides of the figure.

Five other isolates carried deletion of the RD142 region along with deletion of the RD181 region. For these isolates analysis of the products resulting from Multiplex PCR allowed identity of a 790bp product expected to result from flanking primer binding (table IV).On the other hand, isolates without deletion of this region, as for the H37Rv, gave a product of 1376 bp (figure VII).

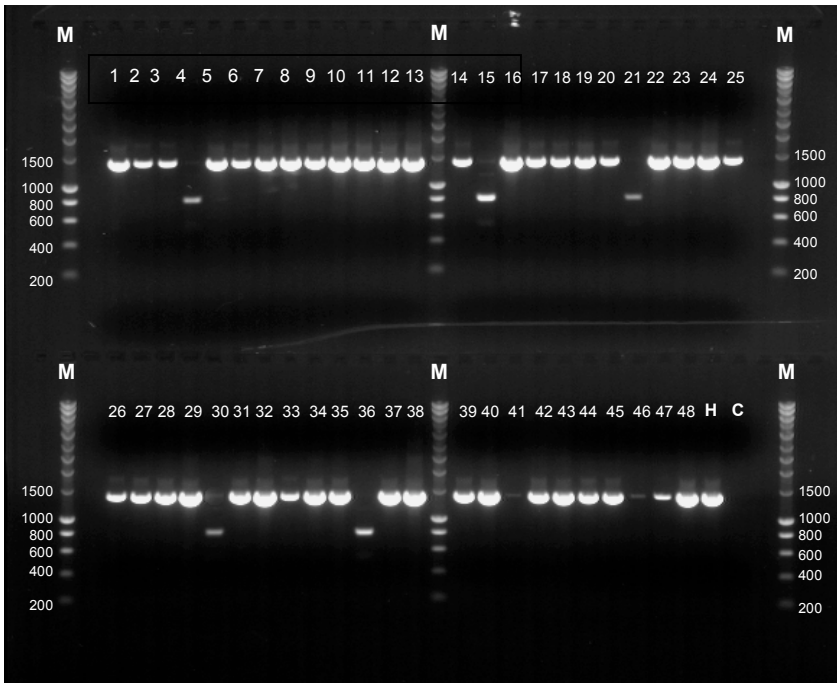


**Figure VII** – RD142 Multiplex-PCR analysis.
Most of the isolates presented a 1376 bp size product, as the positive control represented in the figure as H (*M.tuberculosis* H37Rv). Still for some strains, multiplex PCR resulted in a product of different size, 791 bp, which corresponds to amplification product expected in case of deletion of RD142 region (exemplified in the figure lane 4, 15, 21, 30, 36).
Lanes: 1-48 Beijing/W isolates. C: water as a negative control M: weight molecular marker, respective band sizes are shown on both sides of the figure.

In total, 11 of the isolates carrying a deletion of RD181 carried also a deletion of RD150 or RD141 regions. For the remaining isolates these regions were present according to the products obtained in the Multiplex PCR. As described, these products had the same size of products obtained for the positive control. (see examples in figure V for RD150 and figure VI for RD142).

Six isolates were excluded from final analysis of RDs due to difficulties in amplification or inconsistent amplification products. This may have happened due to the quality of the DNA used.

Four different LSPs patterns were found: a) deletion of RD105 only, b) deletion of RD105 and RD181 c) deletion of RD105, RD181 and RD150 and finally d) deletion of RD105, RD181 and RD142. Each pattern corresponds to a different sublineage of the phylogeny for Beijing/W family suggested using these markers (44). Figure VIII places each one of the isolates on these sublineages according to the referred patterns. The analysis of these LSPs

provided information for selection of representative isolates for the SNPs analysis on 3R genes.

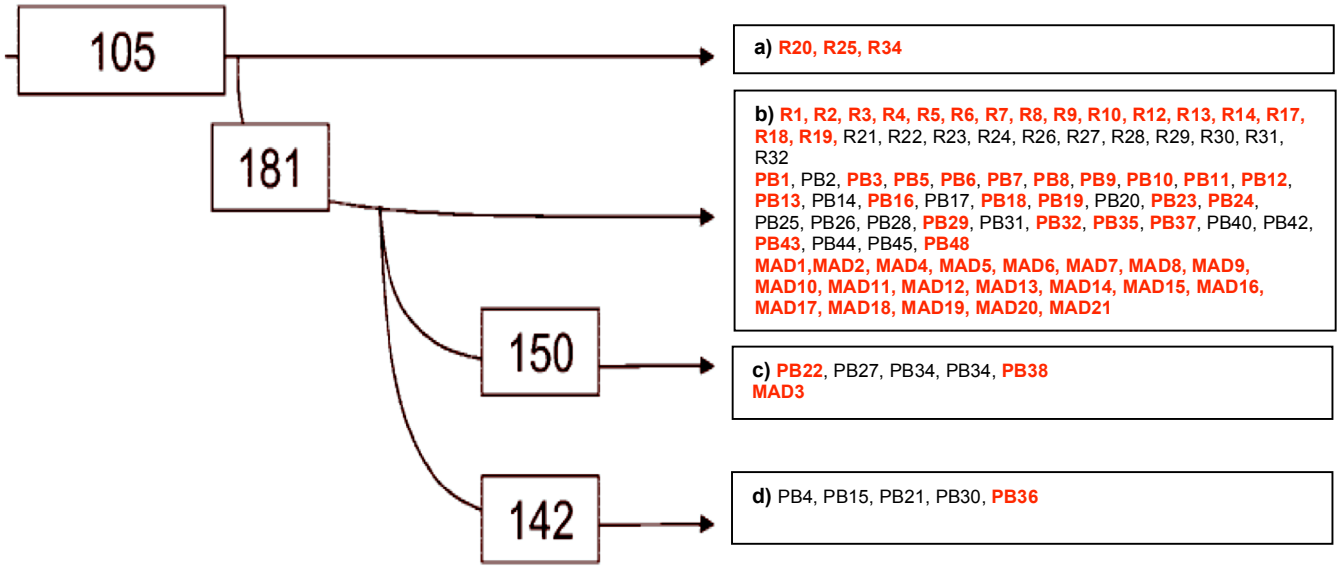


**Figure VIII** – Phylogeny of Beijing/W lineage based on regions of difference (RDs) for the Beijing/W isolates of this study.
According to the deletion patterns found, isolates were placed in one of 4 different sublineages previously defined byTsolaki *et al* (2005) (44)*:* a) RD105 deletion; b) RD105 and RD181 deletion; c) RD105, RD181 and RD 150 deletion; d) RD105, RD181 and RD142 deletion.
In red are the isolates that were selected for the SNP analysis.
R – Name attributed for DNA corresponding to isolates from RIVM, PB – name attributed for DNA corresponding to isolates from P. Bifani, MAD – name attributed to DNA from isolates from Institut Pasteur of Madagascar.

## SNPs analysis

Selected isolates representing each of the groups/sublineages of the Beijing/W family defined by the LSPs (figure VII) were then further analyzed in terms of SNPs. Twenty six genes involved in DNA repair, recombination and replication were selected for this analysis based on their polymorphic character previously found in Beijing/W strains (22). Due to problems with amplification and quantity of DNA for several isolates, the complete data set included 22 of the 26 selected genes (Table A2 - Annex). These 22 genes were sequenced initially for 66 Beijing/W isolates selected from the four different sublineages defined by the deletion patterns (see figure VIII) plus one non-Beijing/W isolate used as control.

Each one of the target genes was first amplified by PCR before the sequencing reaction. Figure IX shows the analysis of PCR products of genes, for some isolates, to confirm the correct amplification and to assure absence of products that could interfere with the sequencing reaction as described on the Materials and Methods.

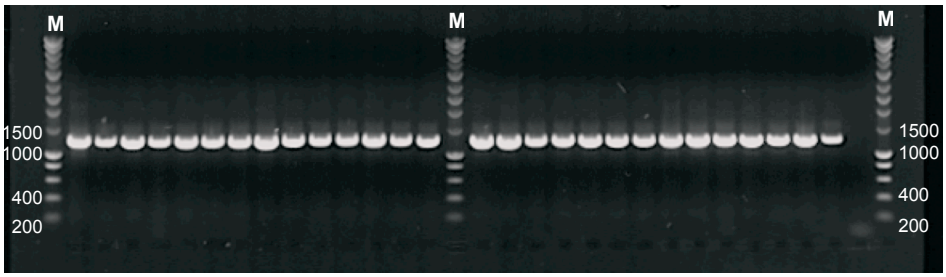


**Figure IX** - Amplification PCR products of gene *recF.*
Each target gene was amplified by PCR reaction before sequencing. Obtained PCR products were analyzed in an agarose gel as exemplified in this figure for *recF*. All lanes except last one are Beijing/W isolates. Last lane is the negative control (water). M: weight molecular marker, respective band sizes are shown on both sides of the figure.

The DNA ladder allows approximation of the quantity of PCR product for use in the sequencing reaction (Figure A1 - Annex for information about the DNA ladder).

Although genes were completely sequenced, it was not possible to obtain a good sequence for certain parts of certain genes. These isolates, including one of the non-Beijing/W isolates, were eliminated from the study. In total this work provided roughly 1,6 Mbps of sequence data.

Comparison of the sequences with the reference strain *M.tuberculosis* H37Rv enabled variations to be identified in all the 22 analyzed genes for the final set of 58 analyzed Beijing/W isolates and one non-Beijing/W. Fourty-eight polymorphisms were identified (Table A2 - Annex) and forty-four percent (21/48) of these polymorphisms corresponded to new variations not previously described for the Beijing/W strains (22, 46).

Based on the inferred proteins, 16 corresponded to Synonymous SNPs (sSNPs), which means no change on the amino acid, and 32 were Non-synonymous SNPs (nsSNPs), which do cause a change in the amino acid (Table VIII). Therefore, non synonymous base substitutions were much more frequent than synonymous substitutions. Figure A2 (Annex) is an example of a polymorphism detected by compared analysis of the sequence of the gene *dnaZX* for some isolates.

Not all these variations were exclusively found in Beijing/W strains. As can be seen in Table A2 (Annex) for example for gene *radA* the mutation in position 456, was present in all analyzed isolates including the non-Beijing/W (PBJ7). The same is true for gene *recF* position 734, *dnaQ* position 631, *ruvB* position 843, *ligD* position 1030 and *ligC* position 938. Forty-one of the 48 polymorphisms, which correspond to 85%, seemed to be specific for Beijing as they were absent from the non-Beijing/W.

Some genes were found to be more prone to accumulate different variations than others for this set of isolates. For example for *recG,* only one variation was identified on position 853 in one strain from Madagascar (Table A2 - Annex). For all the others isolates there was no variation detected on this gene. In contrast, *dnaQ*, *uvrC*, *ligD* were the most polymorphic genes. *UvrC* was also, along with *recX*, *nth,* and *recD,* the ones with the highest number of

different non-synonymous substitutions (Table VIII). According to the recent study of SNPs on 3R genes for a global collection of MTB strains *recX* and *nth* were among the less polymorphic genes (22).

| Gene | sSNPs | nsSNPs | Total |
|---|---|---|---|
| ***radA*** | 1 | 2 | 3 |
| ***recF*** | 1 | 1 | 2 |
| ***recX*** | 0 | 3 | 3 |
| ***rv2979*** | 0 | 1 | 1 |
| ***dnaQ*** | 2 | 2 | 4 |
| ***recR*** | 1 | 1 | 2 |
| ***recG*** | 0 | 1 | 1 |
| ***uvrC*** | 1 | 3 | 4 |
| ***ruvB*** | 1 | 0 | 1 |
| ***ligB*** | 0 | 2 | 2 |
| ***ligD*** | 2 | 2 | 4 |
| ***recD*** | 0 | 3 | 3 |
| ***tagA*** | 1 | 1 | 2 |
| ***uvrD1*** | 0 | 1 | 1 |
| ***dnaZX*** | 2 | 0 | 2 |
| ***nei*** | 1 | 0 | 1 |
| ***nth*** | 0 | 3 | 3 |
| ***alkA(1/4/2)*** | 0 | 2 | 2 |
| ***ligC*** | 1 | 1 | 2 |
| ***mutT2*** | 0 | 1 | 1 |
| ***ogt*** | 1 | 1 | 2 |
| ***mutT4*** | 1 | 1 | 2 |
| **Total** | **16** | **32** | **48** |

**Table VII** – Distribution of synonymous SNPs (sSNPs) and non-synonymous SNPs (nsSNPs) found in this study.

In certain genes, specific SNPs were consistently found in all Beijing/W isolates. These are called invariable sites. It was the case, for example, for the *radA* variation found in position 456 and variation found for *recF* in position 734 and for *rv2979* variation found in position 41, and others (Table A2 - Annex). By contrast, 30 of the 48 SNPs corresponded to variable sites. Thirteen of these being informative sites (SNPs) (table A3 - Annex) with the remaining 17 being non-informative sites. Informative sites are those found in more than one isolate, the term non-informative site is used when a variation is found in only one isolate.

Some variations were found to be present only in the group of isolates from Madagascar (Table A2 – Annex). The most informative were on gene *dnaQ* position 483, for *ligB* position 230, for *tagA* position 385, *uvrD1* position 1384 and *mutT4* position 297. These could be polymorphisms related with isolates from this specific geographic area.

The following hypothetical phylogenetic tree for the Beijing/W family was constructed based on SNPs described by this study, with the media-joining method using the Network software as described in the Materials and Methods. It represents the differently represented 24 Beijing/W groups or haplotypes distincted by SNPs (figure X). SNPs that allowed

discrimination of between those haplotypes are also represented in the figure by the name of the gene and codon affected.

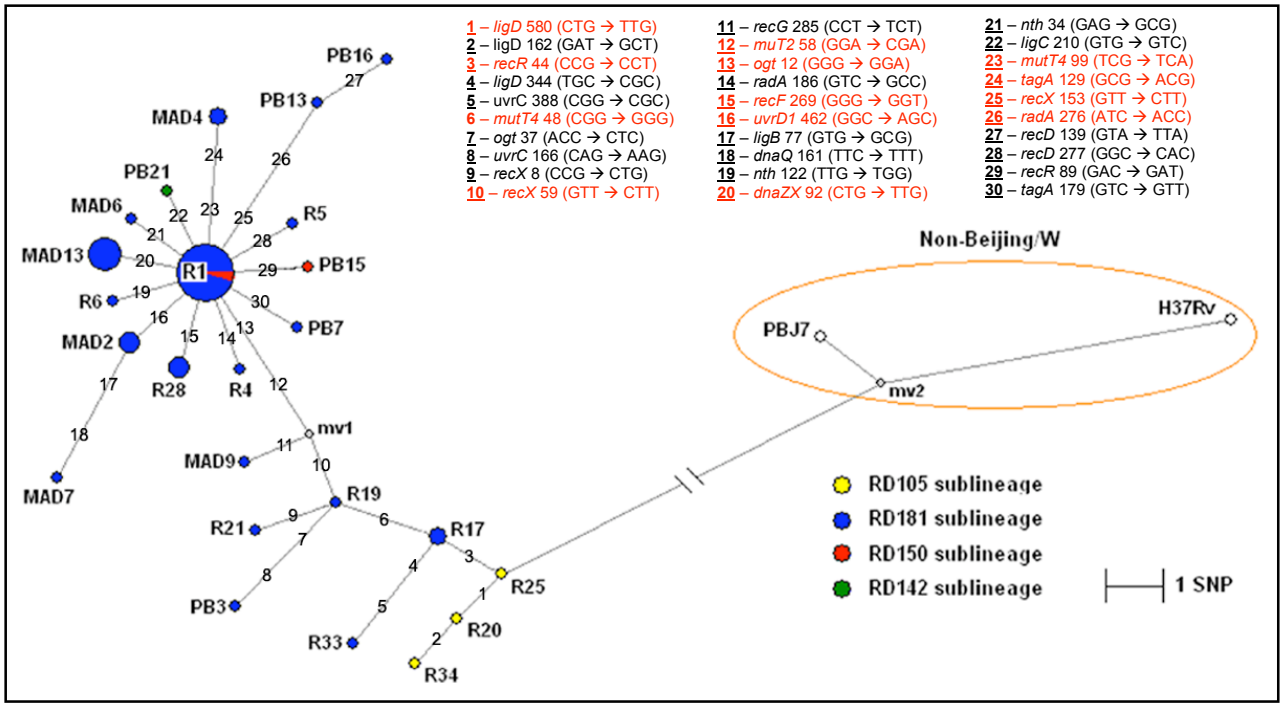


**Figure X** – Hypothetical phylogenetic network constructed by using the Median-Joining algorithm with the final set SNPs characterized in the 58 Beijing/W isolates plus one non-Beijing/W isolate, according to regions of difference sublineages.
*M.tuberculosis* H37Rv, the reference strain, was also included along with PBJ7 as it is non-Beijing/W. Isolates are classified with a color code according to their RD sublineage (Figure VIII). The numbers in each branch correspond to SNPs that allowed discrimination of haplotypes. SNPs in red correspond to the most informative ones described in this study (table A3 Annex).
Nodes sizes are proportional to the number of isolates beloging to the same haplotype:.**R17 node (2)** – R17, PB26; **R28 node (3)** – R28, PB6, PB9; **MAD4 node (2)** – MAD4, MAD5; **MAD2 node (3)** – MAD2, MAD8, MAD12; **MAD13 node (7)** – MAD13, MAD14, MAD15, MAD16, MAD17, MAD19, MAD20; **R1 node (23)** – R1, R2, R3, R7, R18, R23, R24, R31, PB1, PB5, PB8, PB10, PB12, PB14, PB17, PB19, PB20, PB27 PB23, MAD3, MAD10, MAD11, MAD21.
R – Name attributed for DNA corresponding to isolates from RIVM, PB – name attributed for DNA corresponding to isolates from P. Bifani, MAD – name attributed to DNA from isolates from Institut Pasteur of Madagascar. mv – median vector

According to the tree it is possible to distinguish a more ancient sublineage of Beijing/W strains, characterized by isolates R25, R20, R34, from which all the others might have diverged. This sublineage shares fewer differences with the non-Beijing/W strains. These were the isolates associated with the most ancient sublineage (RD105) also according to the RDs analysis (figure VIII). Isolates from sublineage RD181 were the most easily discriminated using the SNPs, (this was the most represented RD sublineage). Some of these isolates were more associated with the most ancient sublineage and others are more associated with the isolates of the more recent sublineages of the RDs (RD150 and RD142). Suggesting that it was possible to distinguish intermediate types, inside the RD181 sublineage, that may have emerged during the evolution of this family by divergence from the most ancient sublineage.

In the same phylogenetic network a color code was attributed according to geographic origin of isolates (figure XI).

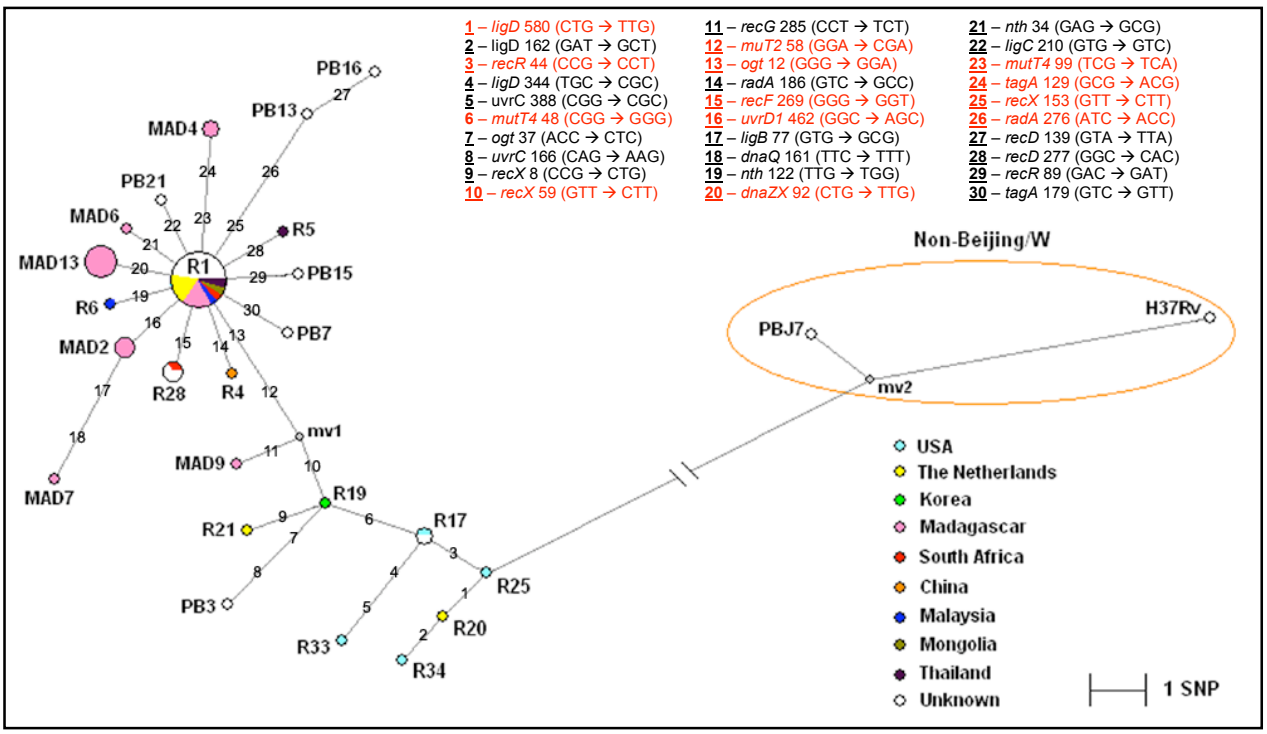


**Figure XI** – Hypothetical phylogenetic network constructed by using the Median-Joining algorithm with the final set SNPs characterized in the 58 Beijing/W isolates plus one non-Beijing/W isolate, according to geographic origin.
*M.tuberculosis* H37Rv, the reference strain, was also included along with PBJ7 as it is non-Beijing/W. Isolates are classified with a color code according to their geographic origin. The numbers in each branch correspond to SNPs that allowed discrimination of haplotypes. SNPs in red correspond to the most informative ones described in this study (table A3 Annex).
Nodes sizes are proportional to the number of isolates beloging to the same haplotype:.**R17 node (2)** – R17, PB26; **R28 node (3)** – R28, PB6, PB9; **MAD4 node (2)** – MAD4, MAD5; **MAD2 node (3)** – MAD2, MAD8, MAD12; **MAD13 node (7)** – MAD13, MAD14, MAD15, MAD16, MAD17, MAD19, MAD20; **R1 node (23)** – R1, R2, R3, R7, R18, R23, R24, R31, PB1, PB5, PB8, PB10, PB12, PB14, PB17, PB19, PB20, PB27 PB23, MAD3, MAD10, MAD11, MAD21.
R – Name attributed for DNA corresponding to isolates from RIVM, PB – name attributed for DNA corresponding to isolates from P. Bifani, MAD – name attributed to DNA from isolates from Institut Pasteur of Madagascar. mv – median vector

Isolates from different origins were found associated, belonging to same haplotype. It means that the same SNP pattern was found in isolates from different geographical areas (see in the figure XI for example haplotype R1 which includes isolates from The Netherlands, Madagascar, Korea, South Africa, etc.).

To compare results of this study with the ones described by Rad *et al.* (2003) (46), a color code was attributed to each node according to SNPs found in *mutT2, mutT4* and *ogt* DNA repair genes (figure XII).

Four of the five SNP patterns described by Rad *et al.* (2003) (46) for these genes were found for the present collection of Beijing/W strains. The pattern *mutT2* (WT) *mutT4* (WT) *ogt* (37) was not found in this study, thus confirming that it might be due to a reversion of the *mutT4* variation (46).

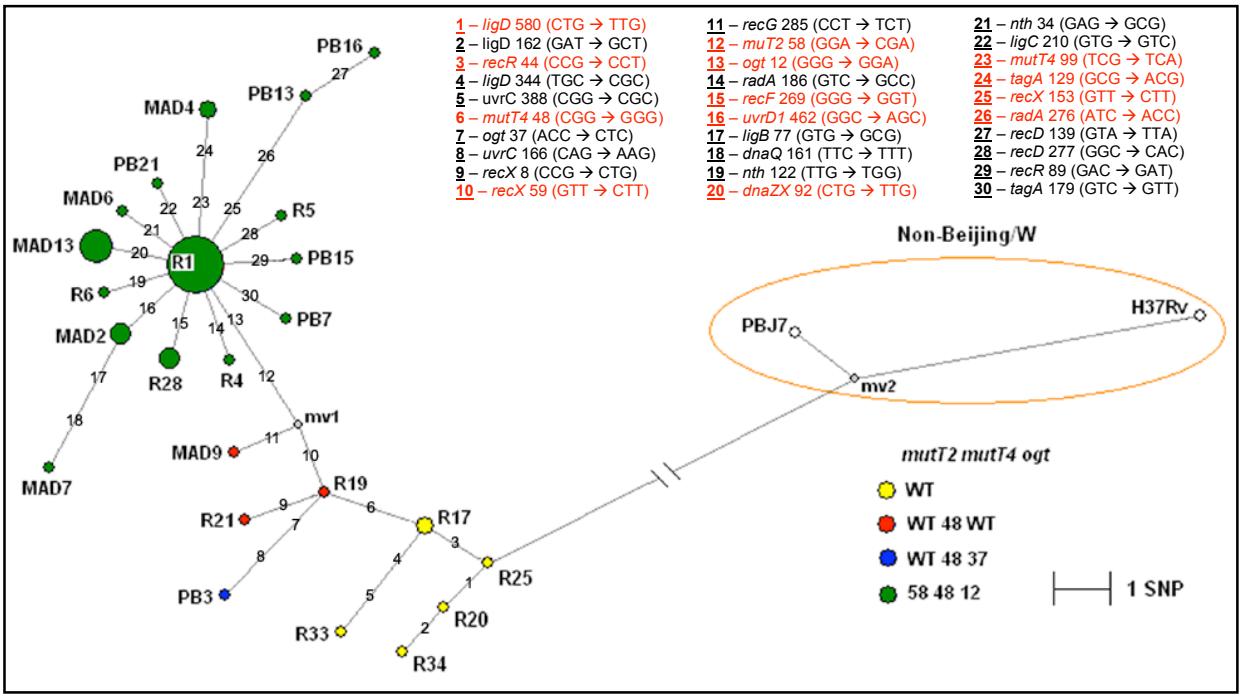


**Figure XII** – Hypothetical phylogenetic network constructed by using the Median-Joining algorithm with the final set SNPs characterized in the 58 Beijing/W isolates plus one non-Beijing/W isolate, according to variations in *mutT2 mutT4* and *ogt* genes.
*M.tuberculosis* H37Rv, the reference strain, was also included along with PBJ7 as it is non-Beijing/W. Isolates are classified with a color code according to SNPs found in *muT2*, *mutT4* and *ogt* genes. According to Rad *et al.* (2003) WT means wild-type sequence for the respective gene in relation with *M.tuberculosis* H37Rv, 58 is variation found in *mutT2* codon 58 (12), 48 is variation found in *mutT4* codon 48 (6), 37 and 12 are variations found in *ogt* codon 37 (7) and 12 (13), respectively.
The numbers in each branch correspond to SNPs that allowed discrimination of haplotypes. SNPs in red correspond to the most informative ones described in this study (table A3 Annex).
Nodes sizes are proportional to the number of isolates beloging to the same haplotype:.**R17 node (2)** – R17, PB26; **R28 node (3)** – R28, PB6, PB9; MAD4 node (2) – MAD4, MAD5; **MAD2 node (3)** – MAD2, MAD8, MAD12; **MAD13 node (7)** – MAD13, MAD14, MAD15, MAD16, MAD17, MAD19, MAD20; **R1 node (23)** – R1, R2, R3, R7, R18, R23, R24, R31, PB1, PB5, PB8, PB10, PB12, PB14, PB17, PB19, PB20, PB27 PB23, MAD3, MAD10, MAD11, MAD21.
R – Name attributed for DNA corresponding to isolates from RIVM, PB – name attributed for DNA corresponding to isolates from P. Bifani, MAD – name attributed to DNA from isolates from Institut Pasteur of Madagascar. mv – median vector

Still, accumulation of variations in these genes found for the Beijing/W isolates analyzed in the present study followed the pathway suggested by these authors on their study (figure XIII). The most ancient isolates in the hypothetic phylogenetic network (figure XII) presented the wild type (WT) sequence for these genes. Isolates diverging from this ancestral ones presented accumulation of variations *mutT4* codon 48 and isolates diverging from these presented variation in *ogt* codon 37 or variations in *mutT2* codon 58 along with *ogt* variation in codon 12 (see figure XII and figure XIII).

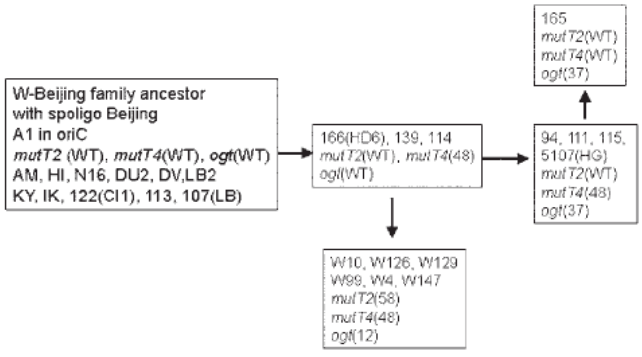


**Figure XIII** – Schematic representation of a plausible pathway to explain the accumulation of mutations in *mutT2*, *muT4* and *ogt* genes. Adapted from Rad *et al.* (2003) (46).

## DISCUSSION

Epidemiological investigations and studies have used molecular techniques to determine transmission, based on whether clinical isolates are identical or different. Insertion sequences, repetitive elements, genomic deletions and single nucleotide polymorphisms cause genetic polymorphisms. These genetic markers can be visualized by different genotyping techniques, such as spoligotyping and VNTR-MIRU typing which allows discrimination and characterization of specific MTB strains (35, 6).

The Beijing/W family is a group of genetically highly conserved strains and has been extensively studied. Recently, one the most studied markers for discrimination of these strains has been the VNTR-MIRU and different studies have shown different discriminatory power (56, 57, 58). However, other markers have been used to analyze genetic polymorphisms on this family that have provided useful information, which is the case for the LSPs described by Tsolaki *et al.* (2005) (44). In a first part of this study LSPs analysis provided a broad discrimination among isolates in different sublineages. This allowed the selection of representative Beijing/W isolates to analyze polymorphisms on recently described polymorphic DNA repair, replication and recombination (3R) genes (22).

Tsolaki *et al.* (2005) (44) has described LSPs specific for the Beijing/W family. As previously mentioned some of these, RD 105, RD142, RD150 and RD181, were used to define and subdivide this family of strains. These markers were analyzed in the set of Beijing/W isolates in the study described here. The **RD105** region is a specific marker that phylogenetically defines the Beijing/W family and can be useful for identification of Beijing/W strains (44, 45, 59). Accordingly, analysis of the RD105 identified all of the analyzed isolates as Beijing/W strains since they all carried deletion of this region. The results confirm that this is a robust and unambiguous method to identify strains of this family as proved by Jing Chen *et al.* (2007) (49) due to its congruence with spolygotyping. It can be especially useful to analyze large number of strains in a high-throughput manner since it is simpler and quicker than spoligotyping. In the present study, the same method was also adopted to target all the other RDs regions referred above. Based on this, screens can be used to facilitate further epidemiological studies to elucidate the prevalence of strains from this family, and its different sublineages in the worldwide epidemic of tuberculosis by this simple PCR-based analysis.

Analysis of the remaining regions **RD181, RD150, RD142** allowed each isolate to be attributed with a specific pattern of deletions. Four different patterns were found which allowed differentiation of strains into the four sublineages. The results were consistent with

the work of Tsolaki *et al.* (2005) (44) on the Beijing/W family when investigated by LSP analysis (figure VIII).

Hanekom *et al.* (2007) (45) has proposed that the sublineage characterized by the deletion of RD150 is associated with increased ability to spread when analyzing isolates from South Africa. Figure VII shows that not all the lineages were equally represented in the different groups of the Beijing/W family defined by these markers. The most represented sublineage (81%) was the one characterized by deletion of RD181 and not the one with deletion RD150. To what extend this is in fact an overrepresented sublineage of the Beijing/W family (possibly due to a different ability to spread) is unclear. The distribution of the RD181 deletion could be further investigated using larger number of isolates from different geographic locations. It would also be interesting to study the differences in terms of virulence between these different sublineages defined by RDs 105, 181, 150 and 142. It remains unknown if there is in fact differences in pathogenesis among sublineages of the Beijing/W family (45).

The analysis of these LSPs proved to be very useful for defining and subdividing the Beijing/W family. Besides, it provided a support for the selection of representative isolates for the further study of other markers like SNPs in recently identified polymorphic 3R genes (22).

A recent study has identified a group of highly polymorphic 3R genes in a global set of *M.tuberculosis* complex strains. The study has shown that evolution of strains of the MTB complex can be studied using 3R gene family polymorphisms and potential useful SNPs were identified as a new powerful tool for clinicians in surveillance of certain MTB haplotypes (22). This study included few Beijing/W strains and in the present study SNPs were analyzed on these genes in a larger set of strains of the Beijing/W family provided by different laboratories. Based on this study selected 3R genes were sequenced for Beijing/W strains to search for variations, providing approximately 1,6 Mbps of nucleotide data. Sequencing errors might occur and to minimize this problem putative changes should be rechecked by re-sequencing and sequence data files should be re-analyzed. Although this is time consuming and for most of the isolates used DNA quantity didn't allow re-sequencing in this study. But there are reasons to be confident that sequencing errors will not generally compromise the studies, especially for the variations found in several isolates. Besides, Gutacker *et al.* (2002) (18) in their comparison of the genomes of *M.tuberculosis* checked around 300 synonymous changes and reported that 91% of these changes were accurate. Qualitatively similar results were obtained for the analysis of close genome sequences of *Bacillus anthracis* (60) and *M.tuberculosis* (21). For other genomes sequences accuracy was even found to be 100% (61).

Comparative analysis of sequences identified variations for all the 22 analyzed genes. This was expected, as these are highly polymorphic genes for which variations have been already described in Beijing/W strains (46, 22). Therefore it was expected to find at least the previous

described polymorphisms however almost half of the total SNPs found in this study corresponded to previously undescribed polymorphisms. Only one non-Beijing/W isolate was used for comparison of sequences with the Beijing/W isolates. This could be a limitation of the work specially when assuming that certain SNPs might be specific for the Beijing/W family. However the new polymorphisms described on this study that were suggested to be specific for the Beijing/W family were also not previously found on the global set of non-Beijing/W strains in the study of Dos Vultos *et al.* (Unpublished) (22).

SNPs proved to be more informative in terms of genetic variability than the LSPs previously analyzed in this study. It was possible to distinguish 24 Beijing/W different haplotypes within the four sublineages defined by the LSPs. A hypothetic phylogenetic network was built with those haplotypes (figure X). Phylogenetic relations originated by SNPs were found to be congruent with the ones originated by LSPs. Isolates of the different LSPs sublineages were found associated in the same way when analyzed by the SNPs. The most ancient strains defined by SNPs belonged to the most ancient sublineage of LSPs (RD105). Isolates that were found to subsequently diverge, according to SNPs, belonged to sublineages that diverged in the same order from the ancient sublineages according to LSPs.

In the analysis according to geographic origin, although some of them were from unknown origin, SNP patterns were not found specific to certain geographic regions. The same SNP patterns were found in isolates from different geographic regions (figure XI). Therefore this suggests that the present collection of Beijing/W isolates might be representative of this family. The described SNPs, although may need to be validated, may be useful tools for future studies of this family of MTB.

SNPs results were found to be congruent also with the previous study done by Rad *et al.* (2003) (46) for the DNA repair genes *muT2*, *muT4* and *ogt* (figure XII and XIII).

Some of the analyzed genes were found to be more polymorphic than others (Table VIII). The most powerful way of assessing the effects of polymorphisms in coding regions is by focusing on the fraction that alter the encoded amino acid sequence, (non-synonymous SNPs). The *uvrC* along with *dnaQ* and *ligD*, were the most polymorphic genes having accumulated 4 different SNPs but *uvrC* was shown to accumulate a higher number of nsSNPs (Table VIII). This gene codes for a component involved in the Nucleotide excision repair (NER) pathway, in which genomic damage is repaired by means of incisions in regions flanking the damaged DNA, leading to excision of the oligonucleotide. UvrC is responsible for the catalysis of incisions 5' and 3' to the DNA damage (62). For the global collection of strains (22) it was suggested that this accumulation of nsSNPs may be due to redundancy of function with another gene involved in the same pathway, *uvrB*. Since this gene was not analyzed in the present study it's not possible to suggest the same for the Beijing/W strains analyzed here. Although on the study of Dos Vultos *et al.* (Unpublished) (22), variations of

*uvrC* were associated with no variations in the *uvrB* gene in the Beijing/W strains included in their study (22).

One interesting result was that *recX* and *nth* were among the genes with the highest degree of non-synonymous substitutions (table VIII) although in the previous study with the global collection of MTB strains they were among the less polymorphic genes, with only 1 non-synonymous substitution found. With a more representative set of strains of the Beijing/W family in the present study these genes showed a higher degree of variations, especially of the non-synonymous type.

The r*ecX* gene codes for a regulatory protein. RecX is necessary to overcome deleterious effects of overexpression of RecA protein, implying that RecX is a negative modulator of RecA expression or function (63). The RecA protein in MTB is involved in regulation of nucleotide excision repair (NER), in genetic recombination, and in induction of the SOS response. This indicates that the absence of RecX leads to an excessive amounts of RecA that might cause lethal defects in these processes (64). The detected nsSNPs are not lethal, suggesting that the function of the protein may not be essential or was not abolished but still it may be altered. If it is the case, these cells may have a defect in the control of the RecA levels, and subsequently may have defects in the processes the protein is involved in (even if subtle).

Nth belongs to a family of DNA glycosylases. Together with Nei, probably forms part of the repair system dealing with 8-oxo-7,8-dihydroguanine (8oxoG). This lesion has strong promutagenic properties (65) and during replication, 8oxoG frequently mispairs with the nucleotide A leading to G:C to T:A transversions (66). Such base changes in the DNA are normally repaired by the base excision repair (BER) pathway (67). BER is initiated by DNA glycosylases, which cleave the N-glycosylic bond and remove the damaged base. This is followed by strand cleavage of the sugar-phosphate backbone, either by the AP (meaning apurinic and apyrimidinic)-lyase activity inherent to many DNA glycosylases or by an AP endonuclease. The repair process is completed by the successive actions of phosphonucleotide kinase or a 3' – or 5' – deoxyribophosphodiesterase, a DNA polymerase and a DNA ligase (67, 68). Nth nsSNPs may lead to an altered function of this enzyme. Although a Nei homolog serves as backup for Nth activity in Knockout mice (69) and under some conditions, such as the repair of X-ray or hydrogen peroxide-induced lethal lesions, Nei and Nth can substitute each other (70). Although a SNP was also found in Nei on the Beijing strains, this was a sSNP. It suggests that Nei might be compensating for the altered function of Nth.

Non-synonymous SNPs were almost twice the number of sSNPs (roughly 2:1). This is so far unusual for bacteria but it was previously reported for MTB (19). Although all nsSNPs may potentially affect the function of the protein it's not correct to assume that all this

detected non-synonymous substitutions will have an effect on the function of the protein. It depends on the effect on the stability of the native protein structure and the folding rate and the protein interactions with other molecules. It could also depend on for example if the polymorphism resides in ligand-binding and catalytic sites or it could influence the the level of transcription, translation and post-translational modification, A recent study has tested the functional significance of nsSNPs in Human DNA repair genes and it was predicted that only a small part of that nsSNPs were functional, potentially damaging nsSNPs (71).

Nevertheless, even if the nsSNP is predicted to have an effect on the function of the protein, it doesn't mean that the pathway on which that protein is involved will be affected, since different genes may code for proteins with redundant functions as in the case for the *uvrC* gene and the *uvrB* gene, and also for the *nth* and *nei* genes. Analysis of the functional effects of these polymorphisms might be a crucial aspect to better understand the real meaning of such variations.

However sSNPs may be also relevant despite being synonymous. Although synonymous codons encode the same amino acids, they are not used randomly and some are used more frequently than others. Such codon usage biases occur in most species from all kingdoms of life. It means that sSNPs might also have an effect on the protein or on the mRNA that could be altered in his stability or translation efficacy.

To what extent the observed variations are associated with phenotypic characteristics is not known but mutator phenotypes - strains having high mutation rates - in general results from alterations in genes coding for DNA repair enzymes and for proteins that assure accuracy of DNA replication. When adaptation is limited by mutation availability, natural selection may favor increased mutation rates by increasing allelic variation of the genetic systems that control fidelity of DNA repair and replication. Therefore, mutator genotypes might allow adaptation to constantly changing environmental conditions, which is the case for *M.tuberculosis*. However, a continuous mutator state leads to fitness loss due to continuous generation of deleterious mutations. Reduction in the mutation rate, for example by acquisition of suppressor/compensatory mutations, occurs before the load of deleterious mutations becomes too high (48, 72). Polymorphisms found in 3R genes in previous studies have suggested that in fact they may be related to a transient mutator phenotype of the Beijing/W strains (46). The study described here also found variations but for a higher number of different genes involved in DNA repair, recombination and replication in Beijing/W strains. A transient mutator state might have in fact existed in these successful strains. If different patterns of SNPs in DNA repair and replication genes among strains reflect different degrees of mutator states might indicate the existence of differences in adaptation/pathogenesis among strains of this family remains to be studied.

Mutator strains may also generate mutations that confer antibiotic-resistance at a higher rate than non-mutator strains (48). When resistance is associated with the acquisition of several mutations, which is the case for MTB (16) the advantage of being a mutator increases significantly (73). Beijing-W genotype strains have been often associated with drug resistance (38) and although the fitness of non-Beijing/W drug resistant strains is slightly reduced, this may not happen for drug resistant Beijing strains (74). One study tried to evaluate contribution of polymorphism on *muT* genes (putative DNA repair genes) with resistance to antibiotics. Although it was not detected any statistically significant association with missense mutations on this genes and increased prevalence of resistance, couldn't rule out the hypothesis that mutator phenotypes might increase the rate of drug resistance (74). Further studies remain to be done about this subject, namely for the genes for which polymorphisms were found in Beijing/W strains in this study.

Overall, this study provided useful information that, although might need to be validated using a larger set of isolates to see how representative they are for the Beijing/W family can be the start point for several future studies. Prediction of functional effect of the observed variations might be done. In addition the effect on the phenotype of those altered proteins might also be studied. Knockout mutants can be constructed to analyze the effect on the phenotype. This will also contribute to better understand the role of these proteins in DNA repair, recombination and replication mechanisms in MTB. It could be studied the association of some of those functional relevant mutations with drug resistance, like it was done by Lari *et al.* (2006) (75) for *mutT* genes in Beijing/W isolates. Some studies about virulence and pathogenesis could be made to analyse if there is any relation with certain SNPs or LSPs patterns. Similarly with the work of Lopez *et al.* (2003) (39) BALB/c mice could be used to examine the course of infection in terms of survival, lung bacillary load, pathology and immune responses produced by Beijing/W strains with certain genotype characteristics. The most informative SNPs, could be useful genetic markers for the Beijing/W family. These SNPs could be applied to further work on phylogenetic and population genetic investigations, molecular epidemiology, studies of drug resistance, and research on host-pathogen interactions. To date, SNPs have not been routinely employed but with the ease and availability of high throughput technology, this technique may provide a useful and relatively less laborious way to type isolates for the referred kind of studies in this successful family of MTB strains. This will contribute to better understand the role of the Beijing/W strains in the worldwide epidemic of tuberculosis.

## BIBLIOGRAPHY

1. Daniel TM. The history of tuberculosis. *Respir Med* 2006; 100: 1862-70.
2. Harries AD and Dye C. Tuberculosis. *Ann Trop Med Parasitol* 2006;100:415-31
3. Raviglione MC, Pio A. Evolution of WHO policies for tuberculosis control, 1948- 2001. *Lancet* 2002, 359 : 775-780.
4. World Health Organization (WHO) (2007) - http://www.who.int/en/
5. World Health Organization (WHO). Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country.
6. Palomino, Leão, Ritacco. Tuberculosis 2007 – From basic science to patient care. *TuberculosisTextbook.com*. 2007
7. Hermans PW, *et al.* Insertion element IS*986* from *Mycobacterium tuberculosis*: a useful tool fro diagnosis and epidemiology of tuberculosis. *J Clin Microbiol* 1990; 28:2051-2058.
8. van Embden JD *et al.* Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol* 1993; 31:406-409
9. Van Sooligen D. Molecular epidemiology of tuberculosis and other mycobacterial infections: main methodologies and achievements. *J Intern Med* 2001; 249:1-26
10. Kamerbeek J, *et al.* Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J Clin Microbiol* 1997; 35: 907-14.
11. Frothingham R, *et al.* Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. *Microbiology* 1998; 144: 1189-96.
12. Chaves, F *et al.* Usefulness of the secondary probe pTBN12 in DNA fingerprinting of *Mycobacterium tuberculosis*. *J Clin Microbiol* 1996 ; 34 : 1118-1123.
13. Supply P, *et al.* Variable human minisatellite-like regions in *Mycobacterium tuberculosis* genome. *Mol Microbiol* 2000; 36: 762-771.
14. Filliol I, *et al.* 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 set. *J Bacter* 2006; 188:759-772.
15. Gutacker MM, *et al.* Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics* 2002 ; 162**:**1533–1543.
16. Gutacker MM, *et al.* Single-nucleotide polymorphism based population genetic analysis of *Mycobacterium tuberculosis* strains from 4 geographic sites. *J Infect Dis* 2006 ; 193**:**121–128.
17. Baker L, *et al.* Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. *Emerg Infect Dis* 2004; 10**:**1568–1577.
18. Fleishman RD, *et al.* Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J Bacteriol* 2002 ; 184**:**5479–5490.
19. Dos Vultos *et al.* Unpublished

20. Mathema B, *et al.* Molecular epidemiology of tuberculosis: current insights. *Clin Microbiol* 2006; 19:658-685.
21. Alland D. *et al.* Role of Large Sequence Polymorphisms (LSPs) in Generating Genomic Diversity among Clinical Isolates of *Mycobacterium tuberculosis* and the Utility of LSPs in Phylogenetic Analysis. *J Clin Microbiol* 2006; 45:39-46.
22. Ramaswany S and Musser JM. Molecular genetics basis of antimicrobial agent resistance in *Mycobacterium tuberculosis*: 1998 update. *Tuber Lung Dis* 1998; 79:3-29.
23. Meya DB and McAdam KP. The TB pandemic: an old problem seeking new solutions. *J Intern Med* 2007; 261:309-29.
24. Flynn JL, *et al.* Tuberculosis: latency and reactivation. *Infect Immun* 2001; 69(7):4195-201.
25. Kramnik I, *et al.* Genetic control of resistance to experimental infection with virulent *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A* 2000; 97: 560-5.
26. Fernando SL, *et al.* Genetic susceptibility to mycobacterial disease in humans. Immunol *Cell Biol* 2006; 84: 125-37.
27. Hill AV. Aspects of genetic susceptibility to human infectious diseases. *Annu Rev Genet* 2006; 40: 469-86.
28. Rook GAW, *et al.* A new look at the role of IL-4 in tuberculosis: implications for vaccine design. *Trends Immunol* 2004; 25: 483-8.
29. Arriaga AK, *et al.* Immunological and pathological comparative analysis between experimental latent tuberculosis infection and progressive pulmonary tuberculosis. *Clin Exp Immunol* 2002; 128: 229-37.
30. American Thoracic Society. Diagnostic standards and classification of tuberculosis in adults and children. *Am J Respir Dis Crit Care Med* 2000; 161: 1371-95.
31. Valway *et al.* An outbreak involving extensive transmission of a virulent strain of *Mycobacterium tuberculosis*. *N Engl J Med* 1998; 338**:** 633-9.
32. Caminero *et al.* Epidemiologic evidence for the spread of a *Mycobacterium tuberculosis* strain of the “Beijing” genotype on Gran Canaria Island. *Am J Respir Crit Care Med* 2001; 164: 1165-70.
*33.* Dick van Sooligen *et al.* Predominance of a single genotype of *Mycobacterium tuberculosis* in countries of East Asia. *J Clin Microbiol* 1995. 33**:** 3234-8.
34. Kurepina *et al.* Characterization of the phylogenetic distribution and chromosomal insertion sites of five IS*6110* elements in *Mycobacterium tuberculosis*: non-random integration in the dnaA-dnaN region. *Tuber Lung Dis* 1998; 79: 31-42.
35. Bifani P, *et al.* Global dissemination of the *Mycobacterium tuberculosis* W-Beijing family strains. *Trends Microbiol* 2002 ; 10: 45-52.
36. Sreevatsan S, *et al.* Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci U S A* 1997; 94: 9869-74.
37. Kremer K, *et al.* Definition of the Beijing/W lineage of *Mycobacterium tuberculosis* on the basis of genetic markers. *J Clin Microbiol* 2004; 42:4040-9.

38. Glynn JR, *et al.* Worldwide occurrence of Beijing/W strains of *Mycobacterium tuberculosis*: a systematic review. *Emerg Infect Dis* 2002; 8:843-9.
39. Lopez *et al.* A marked difference in pathogenesis and immune response induced by different *Mycobacterium tuberculosis* genotypes. *Clin Exp Immunol* 2003; 133: 30-7.
40. Dorman SE, *et al.* Correlation of virulence, lung pathology, bacterial load and delayed type hypersensitive responses after infection with different *Mycobacterium tuberculosis* genotypes in a BALB/c mouse model. *Clin Exp Immunol* 2004; 137:460-468.
41. Tsenova *et al.* Virulence of selected *Mycobacterium tuberculosis* clinical isolates in the rabbit model of meningitis is dependent on phenolic glycolipid produced by the bacilli. *J Infect Dis* 2005; 192:98-106.
42. Abebe F and Bjune G. The emergence of Beijing family genotypes of *Mycobacterium tuberculosis* and low-level protection by bacille Calmette-Guerin (BCG) vaccines: is there a link? *Clin Exp Immunol* 2006; 145: 389-97.
43. European Concerted Action on New Generation Genetic Markers and Techniques for the Epidemiology and Control of Tuberculosis. Beijing/W genotype *Mycobacterium tuberculosis* and drug resistance. *Emerg Infect Dis* 2006; 12:736-43.
44. Tsolaki AG, *et al.* Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of *Mycobacterium tuberculosis*. *J Clin Microbiol* 2005; 43: 3185-91.
45. Hanekon M, *et al.* Evidence that the spread of *Mycobacterium tuberculosis* strains with the Beijing genotype is human population dependent. *J Clin Microbiol* 2007; 45:2263-6.
46. Rad ME, *et al.* Mutations in putative mutator genes of *Mycobacterium tuberculosis* strains of the W-Beijing family. *Emerg Infect Dis* 2003; 9:838-45.
47. Tonjum T and Seeberg E. Microbial fitness and genome dynamics.*Trends Microbiol* 2001; 9:356-8.
48. Denamur E and Matic I. Evolution of mutation rates in bacteria. *Mol Microbiol* 2006; 60:820-7.
49. Chen J, *et al.* Deletion-targeted multiplex PCR (DTM-PCR) for identification of Beijing/W genotypes of *Mycobacterium tuberculosis*. *Tuberculosis* (Edinb). 2007; 87:446-9.
50. Automated DNA Sequencing – Chemistry Guide'. *Applied Biosystems* 2000.
51. Werle E, *et al.* Convenient single-step, one tube purification of PCR products for direct sequencing. *Nucleic Acids Res* 1994; 22:4354-4355.
52. Sanger F, *et al.* DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A.* 1977; 74:5463-7.
53. http://software.cng.fr/docs/genalys.html
54. Tuberculist – *Mycobacterium tuberculosis* strain H37Rv GenoList browser (2007): http://genolist.pasteur.fr.
55. Bandelt HJ, *et al.* Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol.* 1999; 16:37-48.
56. Yokoyama E, *et al.* Improved differentiation of *Mycobacterium tuberculosis* strains, including many Beijing genotype strains, using a new combination of variable number of tandem repeats loci. *Infect Genet Evol* 2007; 7:499-508.

57. Iwamoto T, *et al.* Hypervariable loci that enhance the discriminatory ability of newly proposed 15-loci and 24-loci variable-number tandem repeat typing method on *Mycobacterium tuberculosis* strains predominated by the Beijing family. *FEMS Microbiol Lett* 2007; 270:67-74.
58. Surikova OV, *et al.* Efficient differentiation of *Mycobacterium tuberculosis* strains of the W-Beijing family from Russia using highly polymorphic VNTR loci. *Eur J Epidemiol* 2005; 2011:963-74.
59. Kong Y, t al. Population-based study of deletions in five different genomic regions of *Mycobacterium tuberculosis* and possible clinical relevance of the deletions. *J Clin Microbiol* 2006; 44:3940-6.
60. Read TD, *et al.* Comparative genome sequencing for discovery of novel polymorphisms in Bacillus anthracis. *Science* 2002; 296:2028-33.
61. Rocha EP, *et al.* Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol* 2006; 239:226-35.
62. Tye BK, *et al.* Uracil incorporation: a source of pulse-labeled DNA fragments in the replication of the Escherichia coli chromosome. *Proc Natl Acad Sci U S A* 1978; 75:233-7.
63. Papavinasasundaram KG, *et al.* Construction and complementation of a recA deletion mutant of Mycobacterium smegmatis reveals that the intein in *Mycobacterium tuberculosis* recA does not affect RecA function. *Mol Microbiol*. 1998; 30:525-34.
64. Venkatesh R, *et al.* RecX protein abrogates ATP hydrolysis and strand exchange promoted by RecA: insights into negative regulation of homologous recombination. *Proc Natl Acad Sci U S A* 2002; 99:12091-6.
65. Demple B and Harrison L. Repair of oxidative damage to DNA: enzymology and biology. *Annu Rev Biochem* 1994; 63:915-48.
66. Cheng KC, *et al.* 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G----T and A----C substitutions. *J Biol Chem* 1992; 267:166-72.
67. Seeberg E, *et al.* The base excision repair pathway. *Trends Biochem Sci* 1995; 20:391-7.
68. Slupphaug G, *et al.* The interacting pathways for prevention and repair of oxidative DNA damage. *Mutat Res* 2003; 531:231-51.
69. Takao M, *et al.* A back-up glycosylase in Nth1 knock-out mice is a functional Nei (endonuclease VIII) homologue. *J Biol Chem* 2002; 277:42205-13.
70. Jiang D, *et al.* Escherichia coli endonuclease VIII: cloning, sequencing, and overexpression of the nei structural gene and characterization of nei and nei nth mutants. *J Bacteriol* 1997; 179:3773-82
71. Nakken S, *et al.* Computational prediction of the effects of non-synonymous single nucleotide polymorphisms in human DNA repair genes. *Neuroscience* 2007; 145:1273-9.
72. Taddei F, *et al.* Role of mutator alleles in adaptive evolution. *Nature* 1997; 387:700-2.
73. Tenaillon O, *et al.* Mutators, population size, adaptive landscape and the adaptation of asexual populations of bacteria. *Genetics* 1999; 152:485-93.
74. Toungoussova OS, *et al.* Impact of drug resistance on fitness of *Mycobacterium tuberculosis* strains of the W-Beijing genotype. *FEMS Immunol Med Microbiol* 2004; 42:281-90.
75. Lari N, *et al.* Mutations in mutT genes of *Mycobacterium tuberculosis* isolates of Beijing genotype. *J Med Microbiol* 2006; 55:599-603.

# ANNEXES

**Table A1** – Selected genes and respective primers for sequencing to search for variations.
PCR amplification reaction was performed using forward (f) and reverse (r) primers. Annealing temperature for PCR reaction was calculated based on the melting temperature (Tm) on table for primers used for amplification of each gene (see materials and methods). Extension time depended on gene length (bp - base pairs). Primers named with a number correspond to intermediate primers used in sequencing reaction. Last column describes putative gene function.

| | Primer sequence (5' – 3') | Tm °C | Length (bp) | Putative function |
|---|---|---|---|---|
| ***ligD*** | f-GTCACGGCGAAATTCCACGCGATATTTGA | 70 | 2280 | Possible ATP-dependent ligase |
| | r-CCCGACCAGATCCAGCAACGACACGTC | | | |
| | 2-TCACCAGCGGCAGCAAGGGATTGCAT | | | |
| | 3-GATACACACCGAGGACCACCCGCTGGAATA | | | |
| ***ligB*** | f-CCACATAGCCCCCAGGCGGTATTGGTA | 67 | 1524 | Possible ATP-dependent ligase |
| | r-CGCTTGGTCGACGAGCGTGAATCTG | | | |
| | 2-GGCACTCTACCGGGCAAAGGGTCTCAG | | | |
| ***ligC*** | f-ACCCCAGCTTCGGGAAATACATCCTGT | 66 | 1077 | Possible ATP-dependent ligase |
| | r-TCGCCACACAGACGACAAGTCCCAA | | | |
| ***recC*** | f-TGTCGTTCGGTGATCTCGCGTCTGTTTATG | 69 | 3294 | Probable exonuclease V |
| | r-AGACCGGCCAGCGCGAACGTCTTAC | | | |
| | 2-CATGTCCGCCACGACAAGACCATC | | | |
| | 3-CTGACCGTCTGCACGATGGTCCCGAT | | | |
| | 4-AGGGGTTCTTCCGGGCGCTGGACTACA | | | |
| ***recG*** | f-CATGTGCACGACCACCATCCAGGCAC | 68 | 2214 | Possible ATP-dependent DNA helicase |
| | r-CGATGATCCCAGCGTCTGATACGCGA | | | |
| | 3-CAGCACAAAAGTGCAGAGCTGGGACATCTT | | | |
| | 4-GATGACGGCAGGGCAGAAGAAGCAAGTTC | | | |
| ***uvrD1*** | f-CCCGCAAAAACTTGGCGGGAAAAGTG | 69 | 2316 | Possible ATP-dependent DNA helicase II |
| | r-GGACTTAGCGTCGGCAATTACACCGGTTGA | | | |
| | 2-CAACCTGAAGAACGAGTTGATCGACCC | | | |
| | 3-CGAGGGTAGCGAGATCACCTACAACGAT | | | |
| ***uvrC*** | f-CAATGCACCCGACCAACAGTGGGATAGC | 69 | 1941 | Possiblr excinuclease ABC |
| | r-CCGGACAGCCCGGTTACCAAGACGA | | | |
| | 2-TACATCGACAAATGTTCCGCGCCGTGT | | | |
| | 3-CGGTGCACCGAAACGCAGAAGATGC | | | |
| ***polA*** | f-AGCCCGGGCGTAAAACTGAAACGTGTTG | 69 | 2715 | Probable DNA polymerase I |
| | r-CGACGGGTACACGCTGGACAAACTCGGT | | | |
| | 2-GTCAGCGAACTTACGCGCTTCACAC | | | |
| | 3-CGAAGGCGCTTACCTCGATACCGCGACG | | | |
| | 4-TTGTTCGACAAGACCGGGCATCCGTT | | | |
| ***ruvB*** | f-GATACGGTGCTGGCCGCCAACCAT | 68 | 1035 | Probable holliday junction DNA helicase |
| | r-GGGGTCATTGCCAACGGCTCCTTTG | | | |
| ***recD*** | f-GGTGTGTTCACCTGGAACCCGCCCA | 70 | 1728 | Probable exonuclease V |
| | r-GTCGCCGTGCTGTTCGTGTATGCGATGT | | | |
| | 2-TCTCGCAAGGTGTTACGGTGTTGACTGG | | | |
| ***mfd*** | f-CAATGTTGACTAACCTCGGCGCCCTAGAAT | 67 | 3705 | Probable transcription repair coupling factor |
| | r-ACCGGCATTTCCTCGGTGTATTGCACCT | | | |
| | 2-ATTGGCTCAACGTCCACCACGGATGA | | | |
| | 3-GTGAATTCCTGGAAGCCTCGTGGTCGGT | | | |
| | 4-CGTCAGGATGGTCGACATCTCGCGAATC | | | |
| | 5-CAGACCGGGGTGCGCTGGAAGGA | | | |
| | 6-GGGATACATTCGCTTCAGCCGCACCT | | | |
| ***nudC*** | f-AGGCCAGCGACCGGCTGCTCTATATT | 66 | 942 | Probable NADH pyrophosphatase |
| | r-ACAGAACTGTTCCCACGGTGAAGTTCGC | | | |
| ***dnaQ*** | f-CGGGTGGTTACCACCCGGGCAGTTTAC | 67 | 990 | Probable DNA polymerase III |
| | r-TCTCGCAAGGTGTTACGGTGTTGACTGG | | | |
| ***recX*** | f-CCGACGTGGCTGACGAGATCGAGAAGAA | 68 | 525 | Regulatory protein |
| | r-CCGCCATCAAGTCGAGGTAAATTCGTTCA | | | |
| ***radA*** | f-TAATGGTGCCGATCTCGGCCGGATT | 68 | 1443 | DNA repair protein |
| | r-GTTGCTGCATAGCGGACATCGAGGGAGAA | | | |
| | 2-GAGATCTACCTCGCCGCACAGTCCGA | | | |
| ***recF*** | f-GGAGCGAGTGTCTTTCGGGTTTACGACTGC | 69 | 1158 | DNA replication and repair protein |
| | r-CGCCCTCGACCGGCGTCTTGTCC | | | |
| ***tagA*** | f-TGAGCTCGAGGCGCTACGCTCTCAGC | 68 | 615 | Probable DNA-3-methyladenine glycosylase I |
| | r-CCCCGCCATTGGATTTCCAGCCATA | | | |
| ***nei*** | f-TCTGGTCGAGCGGGCCGACGGCAT | 74 | 768 | Probable endonuclease VIII |
| | r-GGTGGCAGGCAATATCTGCCCAAGGCGG | | | |
| ***rv2979*** | f-GTTCGAAGGTCCACAGGGCCAGAACG | 67 | 585 | Probable resolvase |
| | r-TCCAGTTGTATGCCTTGCGACGAGCA | | | |
| ***nth*** | f-ATGACACAAGGAGAGTAAACATGGC | 55 | 738 | Probable endonuclease III |
| | r-AATAGTCATGCAGTTGGGCAACCA | | | |
| ***mutT2*** | f-CTGCCAGCCGTTGAGGTCGT | 59 | 426 | Probable 8-oxo-dGTPase |
| | r-CGGGCATGCAAACCCAAGTTA | | | |
| ***mutT4*** | f-TCGAAGGTGGGCAAATCGTG | 62 | 747 | Probable nudix hydrolase |
| | r-TGGGGTTCGCTGGAAGTGG | | | |
| ***ogt*** | f-CAGCGCTCGCTGGCGCC | 58 | 498 | 6-O-methylguanine-DNA methyltransferase |
| | r-GACTCAGCCGCTCGCGA | | | |
| ***alkA*** | f-AGCCGCGTAGGTAACCT | 54 | 1491 | Probable ada regulatory protein alkA |
| | r-GCTGACGATGCCGTTGCC | | | |
| | 2-CGCATGCAGACCGCCCG | | | |
| ***recR*** | f-AAGATGGCGCAGGAACGGCTGGGT | 68 | 612 | Probable recombination protein |
| | r-GAGATCAACATTTTGCAGGCAAGGTGCG | | | |
| ***dnaZX*** | f-CGCCGAAATCACGCCGAACGTTCA | 68 | 1737 | DNA polymerase III |
| | r-CGAACGAAACAACCTGCAGCTACATCACG | | | |
| | 2-AACACCTGATCTTCATATTCGCCACCA | | | |
| | 3-CTGCTGCTGGAAGTGGTTTGCG | | | |