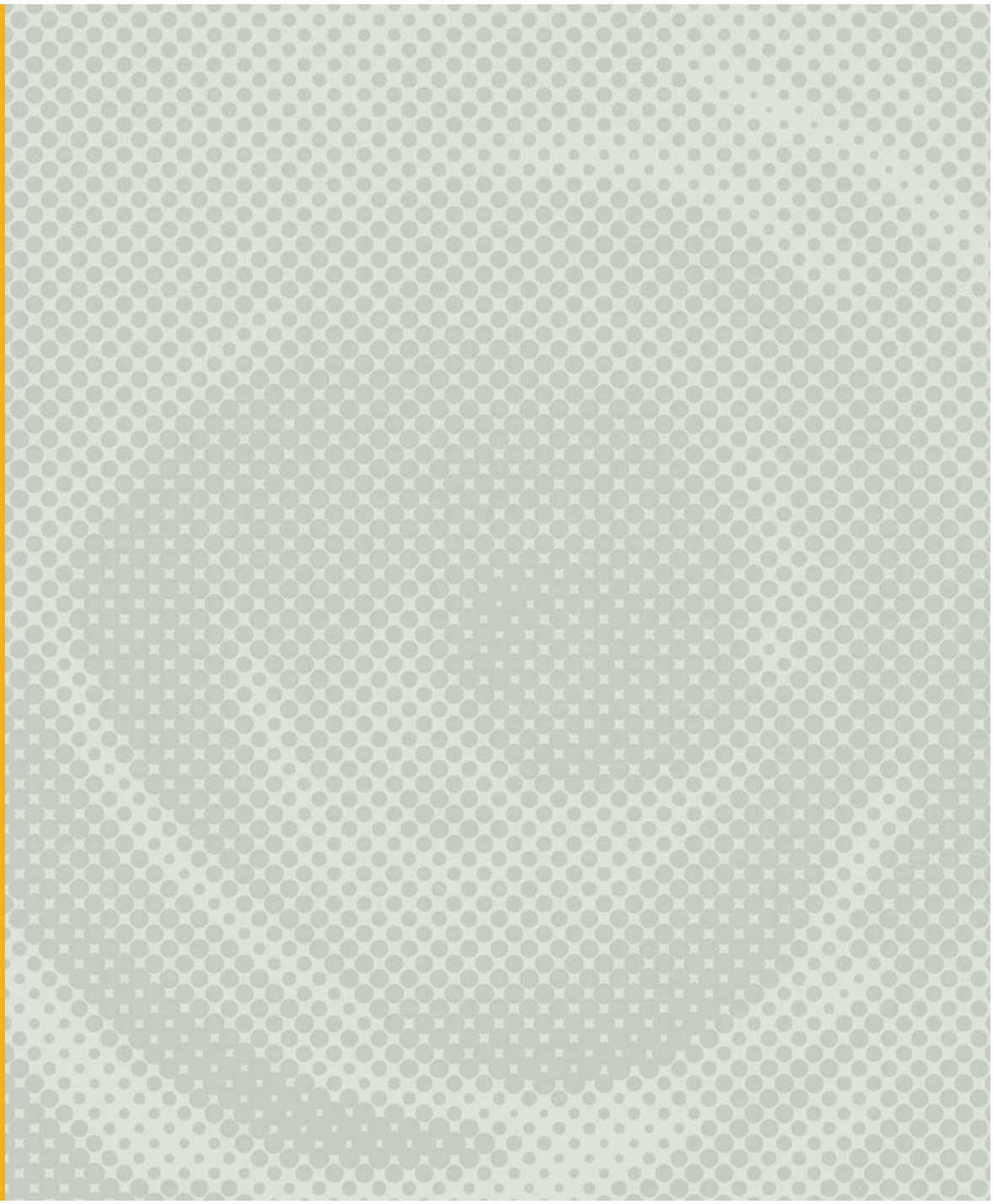
MOLECULAR EPIDEMIOLOGY OF *MYCOBACTERIUM TUBERCULOSIS* IN DAR ES SALAAM AND NORWAY

MOLEKYLÆR MYCOBACTERIUM TUBERCULOSIS EPIDEMIOLOGI I DAR ES SALAAM OG NOREG

VEGARD ELDHOLM

NORWEGIAN UNIVERSITY OF LIFE SCIENCES
DEPARTMENT OF NATURE RESOURCE MANAGEMENT
MASTER THESIS 60 CREDITS 2006



PREFACE

Mycobacterium tuberculosis is the frightening, yet intriguing etiological agent of most cases of human tuberculosis (TB). TB remains a common and deadly disease, responsible for ~2 million deaths globally each year. I am thankful for having been allowed to gain some insight into the immensity of the situation and play a minor part in the massive apparatus that is in place to fight the disease.

I would like to express my gratitude towards main supervisors Ulf R Dahle and Manfred Heun;
Ulf gave me the opportunity to conduct my MSc thesis on *M. tuberculosis* at the Norwegian Institute of Public Health (NIPH). He has been a wise and supportive supervisor who allowed me to part take substantially in the design of the study. His thorough knowledge of all aspects of TB is impressive and his help, patience and support are truly appreciated.
Manfred was the one to really open my eyes to the amazing field of genetics when I attended a course for which I was not qualified in january 2005, and he has helped me a lot since then. His expertice in genetics made it natural for me to ask him to be a co-supervisor when I was offered a master thesis at NIPH. He has pushed me a lot and spent more time and energy on me and my work than what could be expected, and for this I am extremely thankful.

I am indebted to Mecky Matee and Sayoki Mfinanga for running the labs in Tanzania and for shipping high quality TB cultures to Norway, Jørn Henrik Sønstebø for valuable help with phylogenetic analyzes, Bente Forsdahl for teaching me RFLP analysis, Arne Roseth for his help with the capilalary electrophoresis MIRU analyses, Turid Mannsåker for running things smoothly at the lab and Kari Nilsen, Ann-Christine Øvrevik, Elisabet Rønnild and Annika Reichmann for daily help and a good environment at the lab.

Finally, I would like to thank Janne for her support when the thesis demanded long hours, my parents for supporting my choices whatever they were, my sisters for seeming genuinely interested in my work and my friends for making the time spent in Ås and in Oslo a joyful one.

ABSTRACT

Tanzania has a high tuberculosis incidence, and genotyping studies of *Mycobacterium tuberculosis* in the country are necessary in order to improve our understanding of the epidemic. *M. tuberculosis* isolates from 147 sputum-positive TB patients in Dar es Salaam were spoligotyped and the spoligopatterns were checked against the global SpolDB4 database and analyzed with 'Spotclust' to assign the isolates to families, subfamilies and variants.
The TB epidemic in Dar es Salaam was found to be caused mainly by three families, namely the CAS (37%), LAM (22%) and EAI (17%) families. Import of strains was a minor problem.
MIRU typing of the isolates was initiated but not finished.
In the second part of the thesis, Norwegian TB isolates from 2003-2005 were spoligotyped in order to assess the presence of Beijing strains. The family was found to be present, but none of the Beijing strains caused outbreaks in Norway.

SAMANDRAG

Tanzania har ein høg tuberkulose insidens. Genotyping av *Mycobacterium tuberculosis* isolat er naudsynt for å forstå epidemien i landet bedre. *M. tuberculosis* isolat frå 147 sputum positive pasientar i Dar es Salaam vart spoligotypa. Spoligomønstra vart undersøkt mot den internasjonale SpolDB4 databasen og analysert med 'Spotclust' for å plassere dei i familiar og underfamiliar.
TB epidemien i Dar es Salaam viste seg å vere forårsaka av i hovudsak tre familiar; CAS (37%), LAM (22%) og EAI (17%) . Import av stammer var ikkje eit stort problem.
MIRU typing av stammene vart initiert, men ikkje fullført.
I del to av oppgåva, vart norske TB isolat frå 2003-2005 spoligotypa for å undersøke om Beijing familien er til stades i Noreg. Familien viste seg å vere representert i Noreg, men ingen av Beijing tilfella forårsaka utbrot.

1. INTRODUCTION

1.1. Tuberculosis – A Global Emergency

Tuberculosis (TB) has been a frightening disease for millennia and remains one of the most common infectious causes of adult mortality in the world [1], with 98% of all TB deaths occurring in developing countries [2]. The TB burden in the developing world has been high for centuries, while the developed world witnessed a reduction in the TB burden during the 20th century. The disease thus received less attention in high-income countries, both by the lay man and by scientists. This changed after 1984, when New York city and the United States of America as a whole, suddenly witnessed an increase in notification of TB cases [3].

At the same time Europe some countries saw a levelling-off in the rate of decline, while others saw an increase of cases [1]. Foreign born residents were identified as a major factor [4]. TB is a global challenge and one might argue that a come-back for TB in the western world was to be expected as so little progress had been made internationally.
Despite stable or declining TB rates in the rest of the world, sub-Saharan African countries contributed to a rise of the global TB incidence of 1 % in 2003 [5]. The HIV and TB epidemics have proved a dangerous combination and TB has been identified as the major cause of deaths in HIV-infected [6], and a unified approach to controlling these two diseases has been sought in recent years.

It is estimated that 1.8 million people die from TB annually [1].

1.2. Tuberculosis in Tanzania

Tanzania, a country of 37 million people, was one of the first countries to implement the WHO endorsed “directly observed treatment” (DOTS) strategy of TB control [5]. Nationwide DOTS coverage was attained in 2002. The incidence doubled between 1990 and 2004 [5], but the incidence appears to have levelled off in recent years, mainly due to the DOTS strategy and efforts to strengthen the national TB program. The rate of all forms of the disease is

estimated at 524/100,000 and the rate of new sputum smear positive disease is approximately 157/100,000 [1] with Dar es Salaam contributing about 26 % of all TB cases [7].
The World Health Organization estimates that Tanzania has the 14th highest TB burden in the world [1]. The HIV epidemic is a major contributor to the TB epidemic in sub-Saharan Africa, including Tanzania, as immunocompromised patients are easy targets for circulating *M. tuberculosis* strains and often fall victim to reactivation of latent TB.
Tanzania has 701 district laboratories diagnosing TB, three laboratories culturing *M. tuberculosis* and one National reference laboratory that perform drug susceptibility testing of *M. tuberculosis* isolates. Measures are undertaken to establish molecular genotyping methods such as spoligotyping [4], but currently no laboratory in Tanzania offers this service.

1.3. Tuberculosis in Norway

Norway has a very low TB burden, with a TB incidence of 6.6 / 100,000 per year in the period in 2001 [8] The majority of cases (71%) are first- or second-generation immigrants [9]. The age distribution among TB patients is also noteworthy, with most ethnic Norwegians being old people suffering from reactivated disease, whereas immigrants tend to be younger and newly infected [10].

The last decade, the incidence rates of TB have fallen in wealthy countries and risen in many developing countries. Programs to keep TB from rising in many low-incidence countries are challenged by international travelling and immigration from high-burdened countries [9, 11-14]. TB has been the subject of much concern in Norway and other countries in recent years. The public debate on immigration to wealthy countries has often focused on the negative effects immigration may have on public health.

One important potential impact of imported TB cases from high-incidence countries to low-incidence countries would be an increase in the rate of transmission in the recipient country. The impact of such import is difficult to measure or forecast, due to incomplete strain collections, low rates of transmission and the chronic nature of TB. However, the comparison of complete *M. tuberculosis* collections is feasible in some small countries, such as Norway, due to the manageable number of *M. tuberculosis* strains present. The fact that patients may

stay latently infected for many years before they develop symptoms of active TB, can be compensated by extending observation periods.

It has been demonstrated that the number of notified TB cases within Norway has increased significantly during the last 15 years [9]. The percentage of immigrants has increased in general and among TB patients. Each year established outbreaks recruit a few new patients, but in general, only one strain per year is able to establish an epidemic with a potential to include more than 5 patients within 5 years. This inability to establish outbreaks included the infamous Beijing lineage (current study), members of which are commonly believed to harbour increased ability for transmission. These findings indicate that the importation of *M. tuberculosis* isolates to Norway do not represent an immediate challenge to the national public health and that the TB control program in Norway are able to cure patients before they transmit their disease to others.

1.4. The morphology of *Mycobacterium tuberculosis*

Mycobacterium tuberculosis is the most common causative agent of TB. The bacillus is a rod-like obligate aerobe and is classified as Gram positive bacterium due to the lack of an outer cell membrane [15], despite the fact that it does not retain crystal violet stain. *M. tuberculosis* is a slow grower and divides every 15 to 20 hours, rendering culturing of the bacterium time consuming

Two types of molecules covalently linked to the peptidoglycan are characteristic of mycobacteria and play important roles in virulence. These are arabinogalactan and mycolic acids, important for sustaining the impermeability of the envelope [1](the anti-mycobacterial drug isoniazid works by preventing mycolic acid synthesis in susceptible strains). Outside the lipid layer there seems to be a thick layer of proteins and carbohydrates that protects the lipid layer.

The envelope is resistant to destruction by host enzymes and its impermeability denies host factors access to susceptible structures covered by it [1]. The envelope is also likely to play an important role in preventing insertion of proton-ATPase in the membranes of phagocytic vacuoles, enabling the bacilli to survive within macrophages [1].

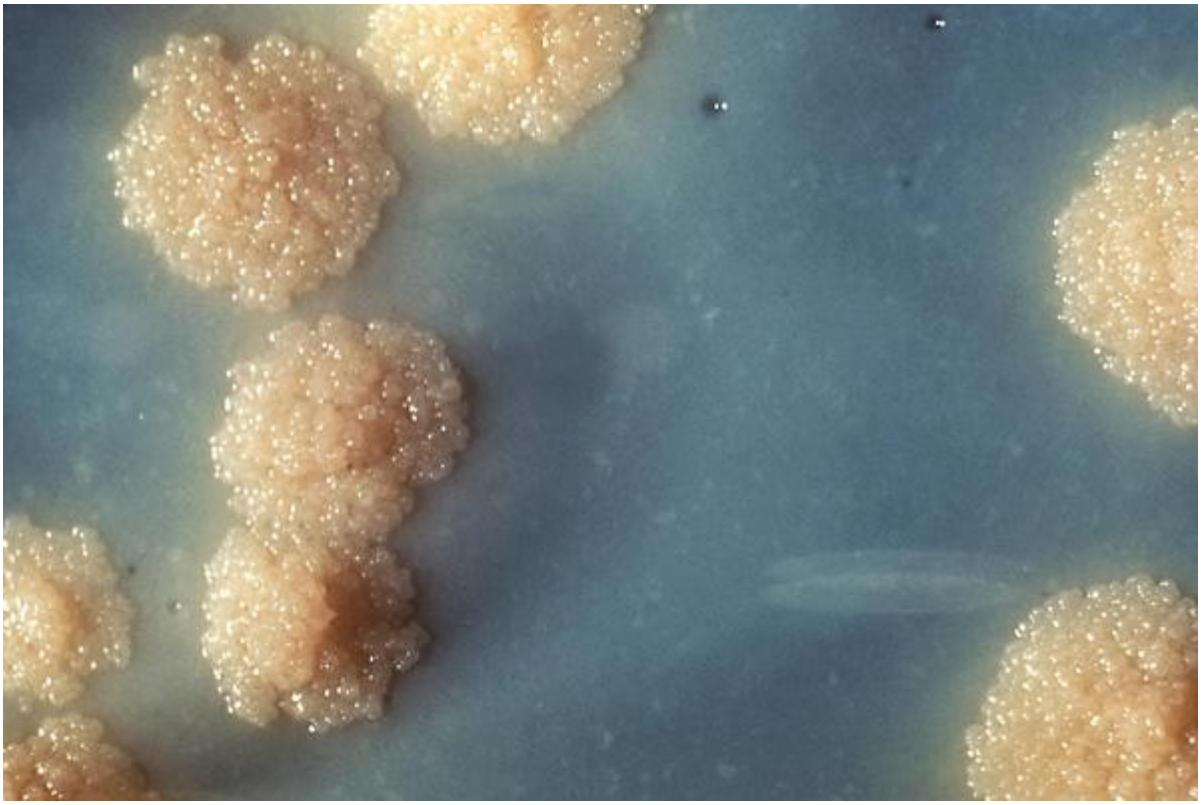


Figure 1.1. Close-up of a *M. tuberculosis culture* (Center for Disease Control, Public Domain. http://phil.cdc.gov/phil_images/20030811/8/PHIL_4428_lores.jpg).

1.5. Clinical tuberculosis

Infection with *M. tuberculosis* is most often asymptomatic in healthy persons. After infection, the life-time risk of clinical TB is only 10 % [16], that is, 90% of those infected never experience clinical disease. Yet, individuals harbouring latent *M. tuberculosis* strains are at increased risk of developing clinical TB, which may develop after decades of latent infection, and often occurs when the persons immune system is challenged due to disease and / or aging. The diverse array of possible outcomes of infection is intriguing, and subject to intense research.

Host factors such as BCG vaccination and HIV infection have been shown to be important determinants for the response of a host to infection. Before the HIV epidemic, 85 % of cases were restricted to the lungs. But in HIV-infected individuals, only 38% have pulmonary infection only, 30 % have extra-pulmonary infections, and 32 % have both. Disseminated disease is typical in immune-compromised hosts [1].

Bacterial factors such as virulence and infectivity are important but not well understood [1]. In recent years, use of microarray based approaches has begun to shed light on some of these factors though [see e.g. [17]

Latent infection is traditionally diagnosed by tuberculin skin test (Pirquet / Mantoux test). The test was developed in 1907, and is used globally despite not being very accurate. Twenty-five percent of symptomatic individuals have negative skin test and infection with nontuberculous Mycobacteria and BCG vaccination can result in false positive results [1]. New tests have been developed, such as *QuantiFERON-TB*, which detects the release of interferon-gamma in blood samples when it is incubated with mixtures of synthetic peptides representing two proteins present in *M. tuberculosis*: early secretory antigenic target--6 (ESAT-6) and culture filtrate protein-10 (CFP-10) [18]. This test is believed to be more specific than the tuberculin skin test, as these proteins are expressed by *M. tuberculosis* and pathogenic *M. bovis* strains, but are absent from most nontuberculous species.

Chest radiography is often used to diagnose TB, as abnormalities almost always are present. TB is transmitted via the respiratory route, and the presence of Mycobacteria in the sputum is often detected by sputum smear microscopy. If no Mycobacteria are detected, the sputum samples are cultured. And positive culture means that a person has active pulmonary TB.

There are also differences between major human populations when it comes to manifestation of the disease. For example, skeletal TB is more common in blacks and whites, whereas lymph node TB is more common in Asians and pacific islanders than in blacks and whites. [1]

1.6. How old is *Mycobacterium tuberculosis*?

Strains of *M. tuberculosis* exhibit extreme homogeneity. Comparisons between the fully sequenced genomes of the strains H37Rv and CDC 1551, have shown the proportion of synonymous single nucleotide polymorphisms (sSNPs) to be about 3.5×10^{-4} [19, 20], and even this low value is more than three times higher than that described before whole genomes could be compared [19]. Hughes *et al.* [20] calculated that the last common ancestor of these two strains occurred about 35 000 years ago. Before this, a popular theory held that *M. tuberculosis* was a descendant of *M. bovis* and had evolved to spread among humans following the domestication of cattle about 10 000-15 000 years ago [21].

1.7. The origin of *Mycobacterium tuberculosis*

The question of where *M. tuberculosis* originated remains open, but recent research has provided some answers. Filliol and co-workers [22] used 159 genome-wide sSNPs to generate a neighbor-joining (NJ) tree from a global collection of M. tuberculosis isolates. From this they concluded that the radiation of human tuberculosis began in India, as isolates collected here mainly belonged to the most ancestral groups close to the middle of the tree. A recent study conducted on Indian strains, came to the same results [23]. The *M. tuberculosis*-specific deletion region 1 (TbD1) has been used to characterize strains, with TbD1+ strains, thought to constitute an ancestral lineage. TbD1+ strains all belong to the East African Indian (EAI) family, and this study found 45% of the strains collected nation-wide to be TbD1+ and belong to the EAI family. According to the SpolDB3 database, the EAI family constitutes about 5.5 % of the global *M. tuberculosis* population [24]. The predominance of these ancestral strains in India supports the hypothesis that India was an early step in the later world-wide expansion of *M. tuberculosis* [23].

An elegant study by Gutierrez *et al*. [25], however, set out to disclose where and when the bacterial pool prior to the putative evolutionary bottleneck event leading to the present-day strains of *M. tuberculosis* existed. *M. canettii* is an uncommon tubercle bacillus with smooth colony phenotype, and is not a member of the so-called *M. tuberculosis* complex (MTBC).[1] The authors used repetitive and long sequence polymorphisms to investigate a group of smooth tubercle bacilli including *M. canettii* and identified eight clonal groups. Some of these groups exhibited mosaicism of certain genes, with interspersed blocks of sequences identical to *M. tuberculosis*. This finding suggests that horizontal gene transfer has taken place before the expansion of the MTBC. Given that both *M tuberculosis* and *M canettii* infects humans, the most parsimonious hypothesis is that the most recent common ancestor (MRCA) of the two could already have caused TB in humans. Given previous studies of substitution rates, the sSNP variation found in this study led to an estimated age of the MRCA of all tubercle bacilli of 2.6-2.8 million years [25]. Nearly all strains of *M. canettii* found so far have been isolated in East Africa and this species thus seems to be confined to this region of the world, which again makes East Africa a prime candidate as the cradle of this bacterium, which the authors coin *M. prototuberculosis*. It is thus possible that the tubercle bacilli may have hitch-hiked on all hominid expansions out of Africa [25]. The hypothesis of an African origin of

[1] The MTBC comprises *M. tuberculosis, M. bovis, M. microti, M. africanum, M. pinnipedii* and *M. caprae*

M. tuberculosis was further strengthened by a study that defined six major lineages based on large sequence polymorphisms (LSPs) and then found all these to be present in Africa, whereas four of them could be found in Asia [26].
The most likely scenario is probably that *M. tuberculosis* has been present both in Africa and the Indian subcontinent for a very long time, possibly originating in Africa and spreading to the rest of the world via India with India as an important focus of its subsequent worldwide spread.

1.8. Population structure

Knowledge of the dynamics of bacterial populations is important for public health issues. Escherichia coli, the most extensively studied bacterial species exhibits a clonal population structure, but research has shown that exchange of DNA happens frequently in natural bacterial populations [27].
M. tuberculosis has plenty of opportunities for DNA exchange; co-infection is not uncommon and many mycobacteriophages exist. Yet, extensive linkage disequilibrium (LD) between minisatellite loci supports a theory of clonal evolution. In an important study carried out in Cape Town, South Africa, the extensive LD remained significant when the possibility of Wahlund effect and epidemiological clonality were taken into account. Over-representation of multilocus genotypes and absence of recombinant types, further supported the notion of a clonal evolution [27].

1.9. Host-pathogen relationship

The observation, in Norway and some other countries, that immigrants from high-incidence to low-incidence countries seldom infect the native population [9, 28] and that the bacille Calmette-Guérin (BCG) seems to vary in efficacy in different parts of the world has led to speculations about co-evolution between this pathogen and different human host populations.
A study in San Francisco, an ethnically highly diverse city, aimed to investigate whether there exists defined human populations between which transmission of TB is limited, but within which transmission is more common [29]. *M. tuberculosis* strains were interrogated with microarrays to identify LSPs. LSPs are thought to behave as unique event polymorphisms

because the absence or near absence of recombination between the bacilli means that genomic deletions are irreversible. The identified LSPs were used to analyze a larger panel of isolates from immigrants, and the authors found that hosts who are infected with *M. tuberculosis* while in San Francisco, tend to contract a genotype associated with the world region in which they had been born. This association was not clear on the finer geographic scale of country of origin, though.

In a follow-up study in the same city, Gagneux *et al.* [26] investigated 875 strains isolated from diverse regions around the world for the presence of defined LSPs, revealing six main lineages. The authors then screened strains isolated in San Francisco for these main lineages. They hypothesized that lineages rare in a specific human population were less able to spread in this population, and found this to be true. The secondary case-rate of sympatric lineages were significantly higher than that of allopatric lineages [26], for example, the East Asian lineage caused fewer secondary cases of tuberculosis than the Euro-American lineage which is a sympatric lineage in San Francisco.

These findings strongly suggest that co-evolution has taken place between specific human populations and specific *M. tuberculosis* lineages.

1.10. Molecular Epidemiology

1.10.1. RFLP

A seminal paper by Daley et al. (1992) [30], describing the use of Restriction Fragment Length Polymorphism (RFLP) based on the repetitive unit IS6110 to investigate transmission of *M. tuberculosis*, opened up the field of molecular epidemiological studies of this bacterium. (At the time of writing, the paper has been cited 644 times in peer-reviewed journals). The development was spurred by a sudden increase of TB cases in New York in 1986 [31], after years of steady decline. The authors reasoned that strains with identical RFLP patterns had been recently transmitted, a notion that has been fundamental to molecular epidemiological investigation of tubercle bacilli since (Traditional contact-tracing must be undertaken to disclose the index case of an outbreak, though).

There are, alas, major drawbacks with the method; it is both time consuming, requiring weeks of culturing the slow-growing bacterium to obtain sufficient amounts of DNA, and laborious.

Thus, it came as no surprise when the PCR based method of spacer oligotyping (spoligotyping) [32] grew popular at the end of the 1990s.

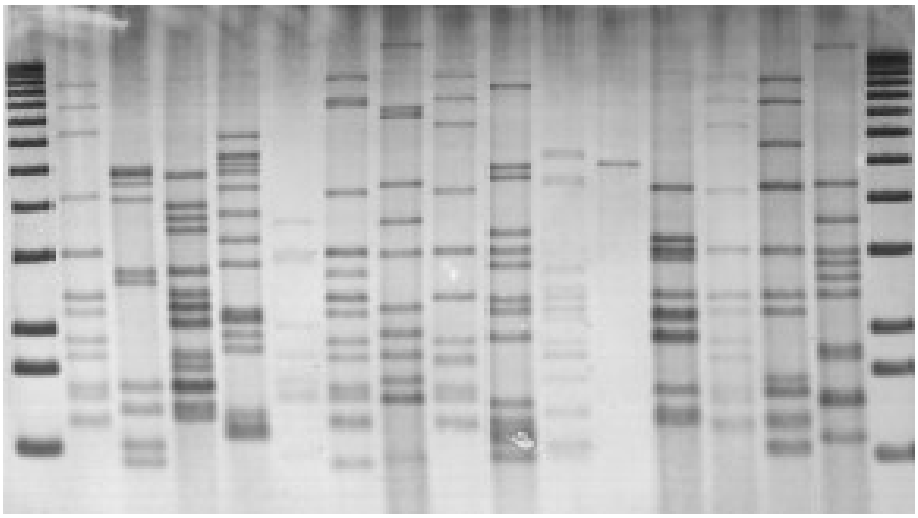


Figure 1.2. IS6110 RFLP patterns of 15 M. tuberculosis strains with DNA ladders in the left and right lane

1.10.2. Spoligotyping

Spoligotyping takes advantage of polymorphisms nested within the Direct Repeat (DR) region in the *M. tuberculosis* genome. The region is a member of the so-called Clustered regularly interspaced short palindromic repeats (CRISPRs) class of repetitive DNA found in many prokaryotic genomes [33]. In *M. tuberculosis* the CRISPRs are interspersed with nonrepetitive sequences as direct repeats to form the DR region. This region contains several 36 base pair (bp) long conserved DRs, between which variable non-repetitive spacer sequences, 34-41 bp long are interspersed. In many prokaryotic species, the spacers have been found to resemble sequences of mobile genetic elements such as bacteriophages and conjugative transposons [34] and the CRISPRs have been found to be associated with genes of unknown function. Phylogenetic analysis of some of the CRISPR associated genes show that Archaean and Eubacterial species are distributed more or less randomly on the phylogenetic tree [35], indicating that the genes have spread independently among the species. It has therefore been hypothesized that the CRISPRs and the associated genes have moved by horizontal gene transfer [35].

In *M. tuberculosis*, strains vary regarding the number of DRs present and the presence or absence of specific spacers, which can be used to differentiate them. Differentiation is

achieved by PCR amplification of the variable regions. The PCR amplification is carried out with primers complimentary to the DR sequences, resulting in amplification of all the variable spacers between different pairs of DRs. One of the primers is biotinylated, and subsequent hybridization to a membrane spotted with 43 synthetic spacer oligonucleotides reveals different hybridization patters. Spoligotyping does not require large amounts of DNA, and the method is easy to carry out in the lab. Figure X shows a picture of a spoligomembrane after hybridization.

Unique to spoligotyping results are tools like the SpolDB4 database [36] and the web-based computer algorithm *Spotclust* [37] that can be used to assign new isolates to families, subfamilies and variants (SpolDB4 only). SpolDB4 is the largest and most up to date available global database for spoligotypes. For previously not reported spoligopatterns, the *Spotclust* database is a good additional tool in that it can assign these patterns to families by using a computer algorithm based on SpolDB3 [37], an earlier version of the database. The results from local studies can thus be analyzed and compared to the global *M. tuberculosis* population. This may help us better understand the world-wide spread of common *M tuberculosis* families and subfamilies.

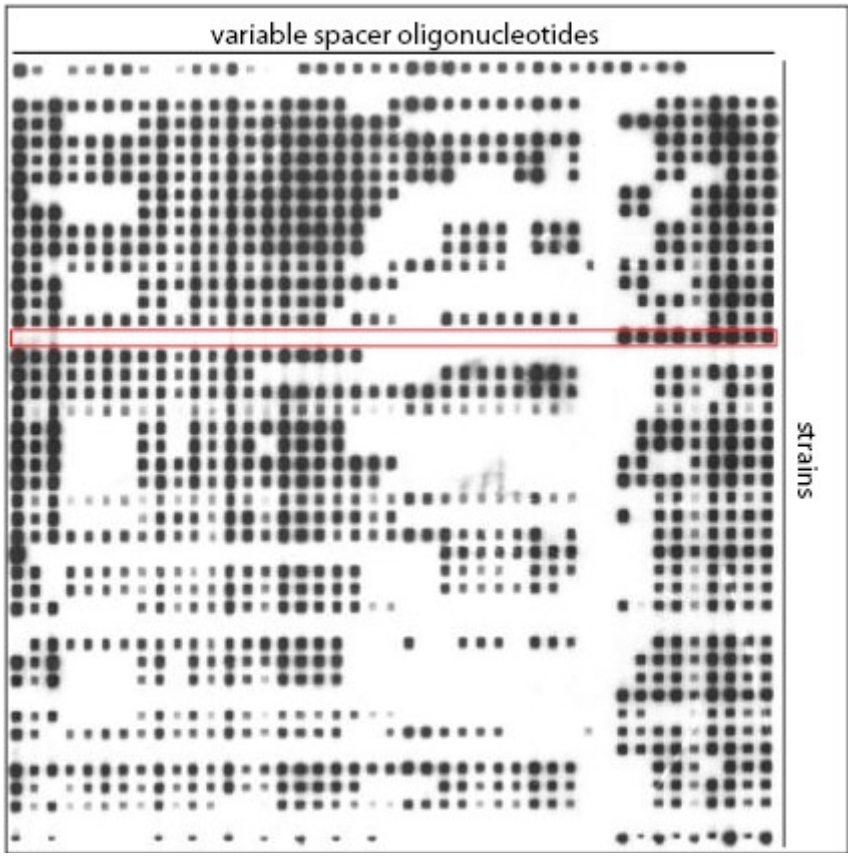


Figure 1.3. Photograph of a typical spoligomembrane. As an example, a typical pattern of the Beijing strain is highlighted (photo V. Eldholm).

1.10.3. MIRU-VNTR

Despite the advantages of spoligotyping, IS6110 based RFLP remained the gold standard in molecular *M. tuberculosis* epidemiology. This was mainly due to the higher discriminatory power of RFLP compared to spoligotyping [38].

In recent years, approaches based on analyses of Variable Number Tandem Repeats (VNTRs) / minisatellites have gained momentum. These methods have until recently had a lower discriminatory power than RFLP, but the power depends on the amount and choice of loci used. In 1997, Supply *et al*. [39] discovered a class of VNTRs with certain distinguishing properties, namely the absence of dyad symmetries common in repeat sequences and that they contain small open reading frames overlapping those of contiguous reading frames and oriented in the same direction. The authors coined the term “Mycobacterial Interspersed Repetitive Units” (MIRUs) for this class of repeats. Supply and co-workers later described 12 MIRU loci [40] which gained popularity in molecular epidemiological studies of *M.*

tuberculosis when an automated high-throughput protocol was described [41] . Other authors have identified other repeat arrays [42, 43] and named these Exact Tandem Repeats and Major Polymorphic Tandem Repeats etc. Here, we refer to all VNTRs as MIRUs, as this is becoming standard nomenclature in the literature. There is still no consensus regarding which loci to use for epidemiological studies, but the 12 MIRU loci [40] described by Supply et al. have been extensively utilized. The discriminatory power of this combination of loci is rivaling, but not quite as high as that of IS6110 RFLP [44]. By combining some of the originally described loci with repeat arrays described by other authors, van Deutekom *et al.* showed MIRU typing to yield a higher discriminatory power than IS6110 RFLP typing in one study [45]. An effort has recently been made by Philip Supply, with the help of an international panel of researchers, to establish a standard protocol for MIRU typing. This method utilizes 15 repeat array loci for epidemiological studies (Supply P, personal communication and [46]), and might well become the standard procedure for epidemiological studies of *M. tuberculosis* in the near future.

Important advantages of VNTRs include high discriminatory power, the speed of the procedure and the ability to discover co-infection. With RFLP and spoligotyping, co-infection manifests itself as more bands or dots on the membranes, but as one does not probe for specific loci one at a time, the presence of two strains in the human host cannot be detected with certainty. Another draw-back with RFLP typing is the low discriminatory power obtained when few bands are detected. Isolates exhibiting less than five bands with IS6110 RFLP has to be typed with a secondary method [10].

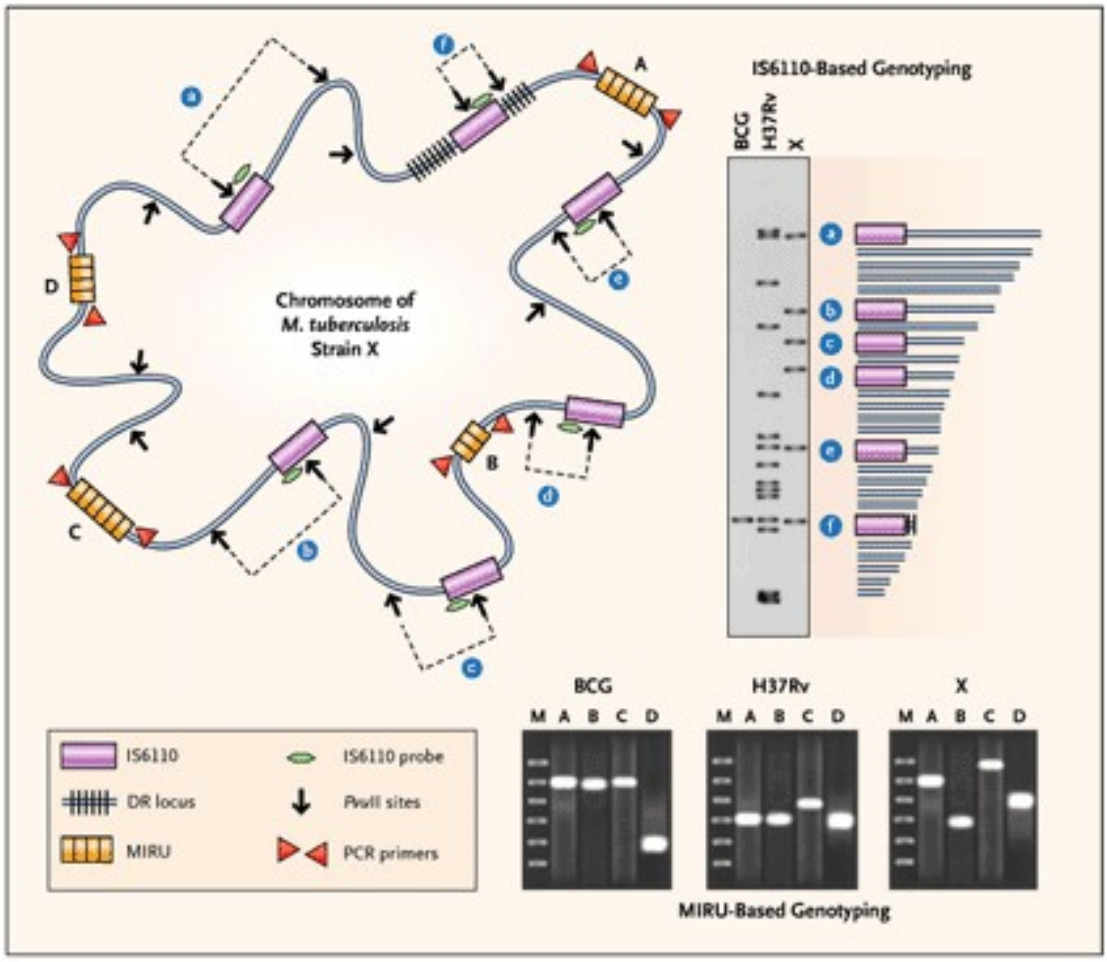


Figure 1.4. Chromosome of *Mycobacterium tuberculosis* Hypothetical Strain X and Genotyping of *M. bovis* Bacille Calmette–Guérin (BCG), the *M. tuberculosis* Laboratory Strain H37Rv, and Strain X on the Basis of IS6110 Insertion Sequences and Mycobacterial Interspersed Repetitive Units (MIRUs).

The top left-hand panel shows the chromosome of hypothetical strain X, as shown by the arrows. The top right-hand panel shows the results of IS6110-based genotyping. Mycobacterial DNA is digested with the restriction enzyme *Pvu*II. The IS6110 probe hybridizes to IS6110 DNA to the right of the *Pvu*II site in IS6110. The size of each hybridizing fragment depends on the distance from this site to the next *Pvu*II site in adjacent DNA (fragments a through f), as reflected by gel electrophoresis of the DNA fragments of BCG, H37Rv, and X. The horizontal lines to the right of the electrophoretic strip indicate the extent of the distribution of fragments in the gel, including *Pvu*II fragments that contain no IS6110. The three bottom panels show the results of MIRU-based genotyping. MIRUs contain repeat units, and MIRU analysis involves the use of polymerase-chain-reaction (PCR) amplification and gel electrophoresis to categorize the number and size of repeats in independent loci, each of which has a unique repeated sequence. shown. The specific sizes of the various MIRUs in each strain result in a distinctive fingerprint for the strain (reprint with kind permission of New England Journal of Medicine).

1.10.4. AFLP

Amplified Fragment Length Polymorphism (AFLP) method is based on the selective PCR amplification of restriction fragments from a total digest of genomic DNA [47]. The genome is first digested with restriction enzymes. The adaptors fitting the sticky ends of the fragments are added, and PCR is carried out using primers complementary to the adaptors. Then selective amplification is carried out with primers carrying one or more additional nucleotide(s), so that only a subset of the original amplification products is amplified (figure X). This is carried out in such a way that between 100 and 200 fragments between ~70 and ~500 base pairs long are produced in the end. The products are analyzed on polyacrylamide gels or by capillary electrophoresis, as these are sensitive enough to separate large numbers of bands with one basepair resolution. Fragments smaller than 70 and larger than 500 base pairs fall outside the easily analyzed size range.

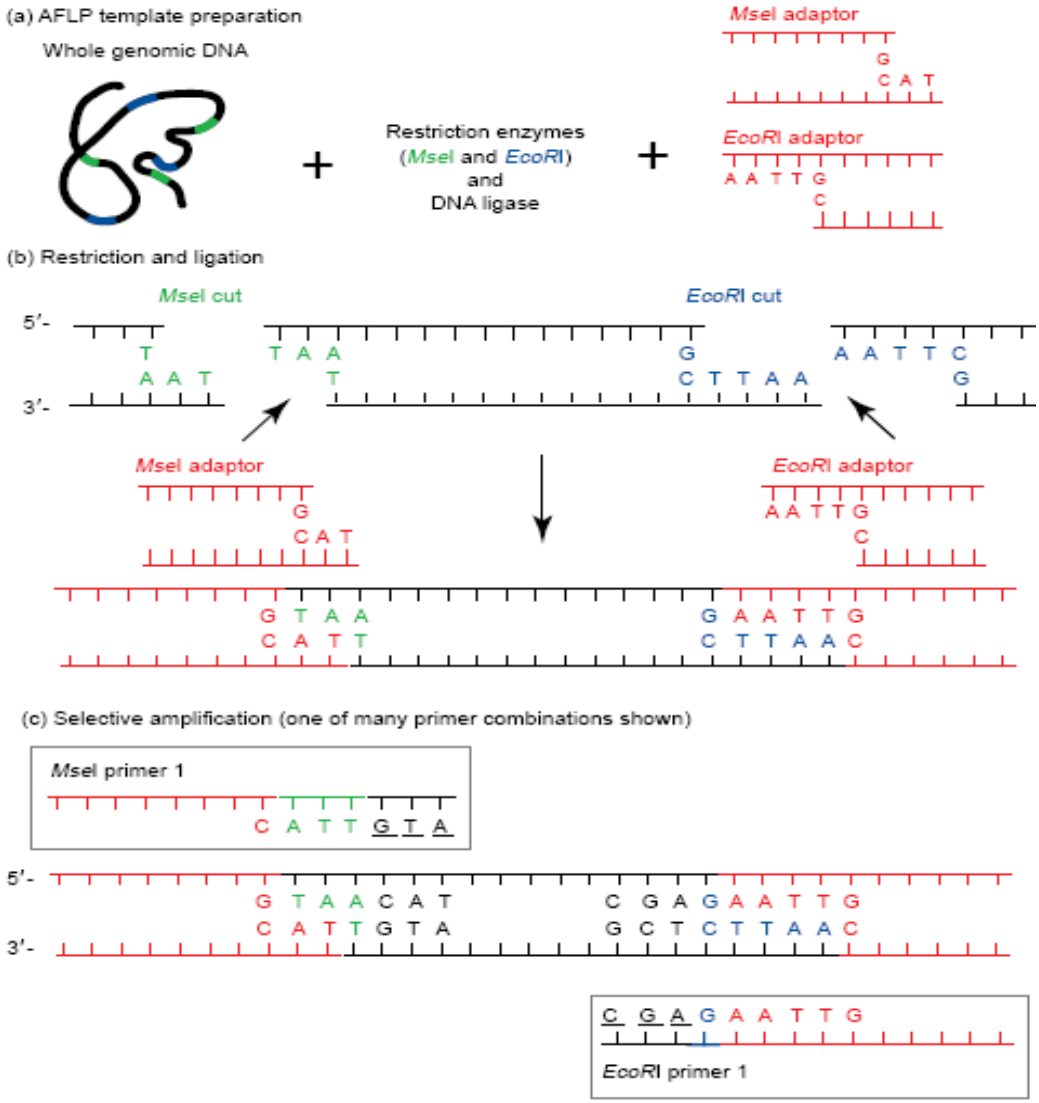


Figure 1.5. The principle of AFLP analysis. From Mueller and Wolfenbarger [48]

When the genome sequence of an organism is known, the restriction digest can be carried out with different enzymes in silico followed by a virtual electrophoresis. The enzymes to be used and the number of selective nucleotides can thus be determined before any lab procedures are carried out. The main advantage of ALFP compared to e.g. RFLP is that a lot more bands are generated, increasing the discriminatory power of the method. Another advantage is that the method investigates genome-wide differences. As a comparison, spoligotyping only covers a small region of the genome.

The major drawbacks are that the method is dominant, that is, one cannot separate alleles from the two parents (only relevant for polyploid organisms), and that the method is technically challenging (which is illustrated by low reproducibility e.g. in studies of *M. tuberculosis*).

1.11. Major TB lineages

After spoligotyping studies became popular, several families of *M. tuberculosis* with characteristic spoligopatterns were found to be widespread globally. The Beijing family was first identified in large numbers in China and neighbouring countries, and was named the Beijing family [49]. The researchers hypothesized that this might be an aggressively expanding clone radiating from the Beijing area. Simultaneously a large outbreak of the multidrug resistant W-strain occurred in New York City [50]. It was later found that the W and the Beijing strains were identical [51]. Beijing is now the most commonly used name for this family of strains, although the name W-Beijing is often used by American authors.

The sudden realization that the family was widespread and possibly more dangerous than other families, led researchers to believe they were dealing with an emerging epidemic. This notion now seems exaggerated, as the discovery that this family is extremely widespread seems to be a result of the ease with which it can be detected with spoligotyping, and thus partly a result of the growing popularity of spoligotyping. There are now many examples of outbreaks caused by drug resistant and multidrug resistant strains of the Beijing family are associated with, but many outbreaks of Beijing family strains are pan susceptible [51]. A SNP based study found the Beijing family to be relatively diverse [22], also indication that this family is less different from other families of *M. tuberculosis* than initially thought.

All members of the Beijing lineage carry a characteristic spoligopattern (figure). The homogeneity of the spoligopatterns is apparently due to the large (non-reversible) deletion in the DR locus and the consistently stable spacers 35-43. RFLP typing, however, shows that this is a more heterogeneous group of strains than spoligotyping is able to detect [52]. An extended version of spoligotyping, which probes for additional variable regions in the DR region, 103 compared to the standard 43 used, still found the Beijing lineage to be relatively homogenous, splitting two clusters into three.

The family has also been found to be widespread in Russia and the strains here typically exhibit high IS6110 RFLP diversity [51].

The recent definitions of the Central Asian (CAS), Latin-American Mediterranean (LAM), East Indian African (EAI), T and Haarlem lineages from the SpolDB3 and SpolDB4 databases [36, 37] have demonstrated that these families are widespread globally as well. Tuberculosis bacilli exhibit strong phylogeographic clustering, reflecting co-evolution with human populations and a relatively stable host relationship [26, 29], and the major lineages have been found to dominate different areas of the world, yet with considerable overlap.

The CAS lineage has been found to be dominant in the Middle-East and Central Asia (~20%) [36]. Few spoligotyping studies have been carried out in east Africa, but the current study demonstrates that the CAS lineage is highly widespread in Dar es Salaam, Tanzania, causing 37% of the TB cases (I). CAS strains have also been found to be the causative agents of an unusually large outbreak of TB in Norway [53].

The stronghold of the EAI lineage is the Middle-East and Central Asia (~25%) and Far-East Asia (~35%). In the Far-East Asia, the Beijing family is even more dominating though, causing ~45 % of all TB cases [36].

The LAM lineage is a major lineage in South and Central America and Africa as well, whereas the T and Haarlem lineages are more evenly spread globally. The distribution of the Haarlem lineage has been linked to post-Columbian European colonization [36].

1.12. The Neighbor-Joining method

To construct a phylogenetic tree using distance methods such as the NJ method, the DNA fingerprinting results must be converted into a distance matrix using a distance algorithm. In this study, the Jaccard distance [54] was used. The algorithm for calculation of Jaccard distance is as follows:

$$J' = \frac{M_{01} + M_{10}}{M_{01} + M_{10} + M_{11}},$$

where M_{01} and M_{10} are number of bands / loci differing between individuals, whereas M_{11} is the number of shared bands / loci. Note that bands / loci absent in both individuals under scrutiny are excluded. This rests on the assumption that the lack of bands is generally less reliable than bands present, and should thus be given less weight.

The NJ method is a tree-building algorithm that is based on the principle of minimum evoultion [55]

To construct a NJ tree, the NJ algorithm assumes that there is no clustering (figure X), and the total branch length is computed. The next step is to cluster two taxa at random and calculate the new total branch length. The cluster (two 'neighbours') that gives the lowest total branch length is kept. This cluster is now regarded as one taxa and further clustering is calculated. When the total branch length no longer decreases, the tree topology has been found [56]. The NJ method has proven to be reliable in most cases [55].
The NJ algorithms are quite complicated and will not be handled in detail here.

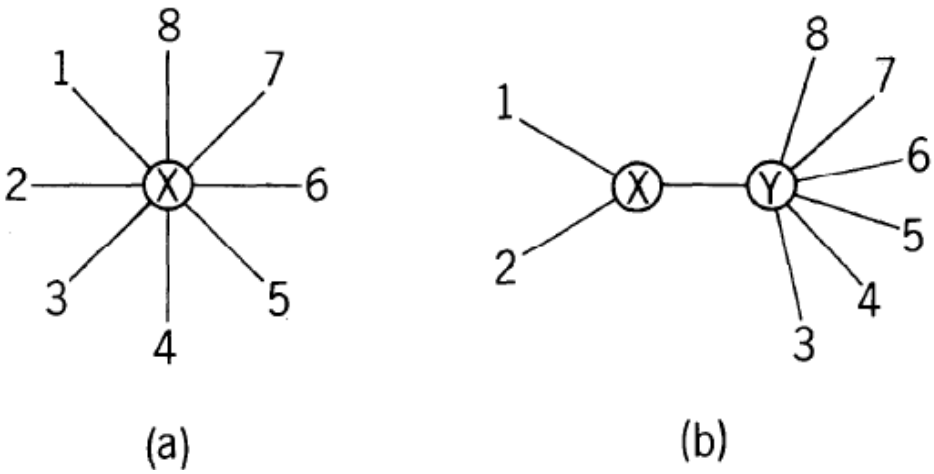


Figure 1.6. Illustration of the NJ process. No hierarchical structure is present (a) before the clustering process begins (b).

1.13. The current studies

The current study is divided into two main themes; the molecular epidemiology in Dar es Salaam (Tanzania), and in Norway. The molecular epidemiology of TB in Dar es Salaam receives by far the most attention.

Previous studies have described the molecular epidemiology of Tanzanian *M. tuberculosis* collections from the first half of the 1990s [57-59]. In order to improve our understanding of the TB epidemic in this high-incidence country, the current study included *M. tuberculosis* strains collected in Dar es Salaam during October and November 2005. Tanzania has 37 million inhabitants. Measures are undertaken to establish molecular genotyping methods such as spoligotyping [32], but currently no laboratory in Tanzania offers this service.

The first study describes the diversity of *M. tuberculosis* isolates from Dar es Salaam, Tanzania, based on spoligotyping, and identifies the families and subfamilies responsible for the current persistence and spread of TB in this high-incidence community. MIRU-VNTR typing was initiated on isolates clustered by the spoligotyping in order to split or confirm the clustering and thus reveal recent transmission events. The combination of loci described by van Deutekom *et al.* [45] were used for this analysis, because of the high discriminatory power of this combination.

Before it was decided apply for MIRU analysis of the clustered Tanzanian strains, the Amplified Fragment Length Polymorphism (AFLP) method [47] was considered. Some initial work was carried out before it was decided discontinue. Researchers have had problems with the reproducibility of this method when applied on mycobacteria [44], and it does not seem to gain popularity. It would therefore probably be difficult to compare the obtained results with other current and future studies. As MIRU typing has become the preferred typing method, it was decided to go with this method.

The second part of the thesis assesses the presence of TB caused by the Beijing lineage in Norway.

The first part of the study, concerning Dar es Salaam, resulted in a published paper covering parts of the work carried out. This paper is included in this thesis, and will be referred to in the Roman numeral (I). To avoid repetition, the ‘Materials and methods’ and the ‘Results’ of this part of the thesis will only be briefly mentioned.

2. MATERIALS AND METHODS

This section adds another level of detail to the materials and methods described in the paper (I) and covers the materials and methods that were not part of it.

2.1. Bacterial Isolates and DNA Isolation

Isolates of *M. tuberculosis* were collected from sputum smear positive TB cases in consecutive patients in Dar es Salaam during October and November 2005. Heat-killed samples were shipped to the Norwegian Institute of Public Health, and the DNA was extracted from the suspensions with chloroform-isoamyl alcohol as described previously [60]. DNA was successfully extracted from 147 isolates (I). These 147 *M. tuberculosis* isolates were spoligotyped according to Kamerbeek et al. [32].

2.2. Spoligotyping

PCR was carried out in 50 µl reaction volumes including 2 µl DNA (concentration not measured), 5 µl of 10x PCR buffer, 4 µl of each of the primers (DRa and DRb), 4 µl dNTP, 0.25 µl amplitaq polymerase and 30.75 µl off distilled autoclaved H_2O.

The PCR cycling parameters were as follows: an initial denaturing step at 96˚C for three minutes, followed by 30 cycles of the following steps: **Denaturing**: 96˚C, 1 minute, **Annealing**: 55˚C, 1 minute, **Extension**: 72˚C, 30 seconds. The cycling was terminated with 5 additional minutes extension at 72˚C.

The spoligomembrane (Isogen Lifescience) contains 43 immobilized variable spacer oligonucleotides arranged in parallel. The PCR products are loaded in a 90 degree angle, so that the products hybridize to all the probes. This is achieved by placing the membrane in a ‘miniblotter’ and loading the PCR products via hybridization channels. The membrane has room 45 samples.

The PCR products are left to hybridize with the oligonucleotides on the membrane for 60 minutes at 60˚C. Subsequently, the membrane is washed for one minute in ECL detection fluid (Amersham bioscience), wrapped in saran wrap and put in a developer box together with a photographic hyperfilm (Amersham bioscience). After 30 minutes, the films are developed, revealing the hybridization patters of the strains in question.

2.3 MIRU-VNTR typing

All the primers (n = 24) were blasted (short near-exact matches NCBI). As two strains (H37 Rv and CDC 1551) of *M. tuberculosis* are fully sequenced in addition to the genetically highly similar *M. bovis* strain BCG P3, perfect matches were returned for these three strains. The sequences flanked by the 12 respective primer pairs were aligned with Clustal X [61]. The repeat motifs were thus identified, and expected fragment sizes could be deduced (table 2.1). The lengths of sequence in which the repeat arrays are embedded vary between the different amplification products. As an example, the amplification product of VNTR 4156 contains 504 bp length of sequence including the primers in addition to the repeat arrays, which were found to contain four, three and two repeats respectively for the strains CDC 1551, H37 Rv and BCG P3.

The primers described [45] for amplification of locus 4156 had a very low melting temperature (Tm =57.3˚C and 53.3˚C). It was decided to design new primers with a melting temperature in the same range as the other 22 primers. As the sequence embedding the repeat motifs was quite long, the sequence of the expected amplification product was used to design new primers with the web-based program ‘Primer3’ [62] The primers for locus 10-12 were then analyzed with ‘Autodimer’ [63] to check for dimerization, as these primers are used together in multiplex reactions. The primer combination with new primers for VNTR 4156 was found to be no more prone to dimerization than the original set. The complete set of primers are shown in table 2.2.

Table 2.1. MIRU fragment sizes.

Locus	VNTR length	Flanking sequence (bp)	Expected number of repeats		
			H37Rv	*CDC*	*BCG*
VNTR 0580	77	121	3	2.5	2
VNTR 2996	51	285	3	5	5
VNTR 0802	54	354	1	5	2
VNTR 0960	53	484	3	5	2
VNTR 1644	53	566	2	3	2
VNTR 3192	53	493	3	3	3
VNTR 0424	51	539	2	4	2
VNTR 0577	58	184	4	3	5
VNTR 1982	78	230	6	8	4
VNTR 2401	58	251	3	4	4
VNTR 3690	58	270	2	3	2
VNTR 4156	59	273	3	4	2

All primers were tested as single PCR reactions and as multiplex reactions on a large panel of strains and the products analyzed on agarose gels (1.6% agarose, 3 hour runs at 100 V). Different annealing temperatures, number of PCR cycles and $MgCl_2$ concentrations were tested for all combinations. Based on these results, multiplex reactions A - C and D were run with different PCR programs (table 2.3).

Table 2.2. Overview of primers and multiplex reactions.

	Locus	$MgCl_2$ (mM)	Primer sequence and label	Tm˚C
Mix A	VNTR 0580	3.0	GCGCGAGAGCCCGAACTGC **FAM**	65.3
			GCGCAGCAGAAACGTCAGC	61.0
	VNTR 2996	3.0	TAGGTCTACCGTCGAAATCTGTGAC	63.0
			CATAGGCGACCAGGCGAATAG **HEX**	61.8
	VNTR 0802	3.0	GGGTTGCTGGATGACAACGTGT **NED**	62.1
			GGGTGATCTCGGCGAAATCAGATA	62.7
Mix B	VNTR 0960	3.0	GTTCTTGACCAACTGCAGTCGTCC	64.4
			GCCACCTTGGTGATCAGCTACCT **FAM**	64.2
	VNTR 1644	3.0	TCGGTGATCGGGTCCAGTCCAAGTA	66.3
			CCCGTCGTGCAGCCCTGGTAC **HEX**	67.6
	VNTR 3192	3.0	ACTGATTGGCTTCATACGGCTTTA	59.3
			GTGCCGACGTGGTCTTGAT **NED**	58.8
Mix C	VNTR 0424	3.0	CTTGGCCGGCATCAAGCGCATTATT	64.6
			GGCAGCAGAGCCCGGGATTCTTC **FAM**	67.8
	VNTR 0577	3.0	CGAGAGTGGCAGTGGCGGTTATCT **HEX**	66.1
			AATGACTTGAACGCGCAAATTGTGA	59.7
	VNTR 1982	3.0	CCGGAATCTGCAATGGCGGCAAATTAAAAG	66.8
			TGATCTGACTCTGCCGCCGCTGCAAATA **NED**	68.1
Mix D	VNTR 2401	2.0	CTTGAAGCCCCGGTCTCATCTGT **FAM**	64.2
			ACTTGAACCCCCACGCCCATTAGTA	84.6
	VNTR 3690	2.0	CGGTGGAGGCGATGAACGTCTTC **HEX**	66.0
			TAGAGCGGCACGGGGGAAAGCTTAG	67.9
	VNTR 4156	2.0	CCAGGTGTGGCTCACAAGAC	61.7
			ATCCGTGTGGTGGTCGACTT **NED**	62.7

The polymerase used was TEMPase Hot Start DNA Polymerase (Bergman). This is a chemically modified heat activated enzyme. The enzyme is inactive at room temperature, which prevents extension of misprimed primers during setup. This increases sensitivity and improves the results of multiplex PCR. For activation of the enzyme, an initial activation step of 95ºC for 15 minutes was added to the PCR cycling conditions (table 2.3).

Table 2.3. PCR conditions.

	Multiplex A - C			**Multiplex D**		
	Touchdown PCR			**Ordinary PCR**		
Step	*Time*	*˚C*	*Cycles*	*Time*	*˚C*	*Cycles*
Initial polymerase activation step	15 min	95˚C		15 min	95˚C	
Denaturing	60 sec	94˚C	2	60 sec	94˚C	35
Annealing	60 sec	59˚C		60 sec	59˚C	
Extension	90 sec	72˚C		90 sec	72˚C	
Denaturing	60 sec	94˚C	2			
Annealing	60 sec	58˚C				
Extension	90 sec	72˚C				
Denaturing	60 sec	94˚C	2			
Annealing	60 sec	56˚C				
Extension	90 sec	72˚C				
Denaturing	60 sec	94˚C	22			
Annealing	60 sec	55˚C				
Extension	90 sec	72˚C				
Final extension	15 min	72˚C		7 min	72˚C	

The PCR reactions worked sub-optimally for most of the loci, despite extensive testing of PCR parameters. The PCR products were run on agarose gels in this step of the study, for inspection of PCR products and fine-tuning of PCR parameters. The most important experimental parameters tested were the following:

-$MgCl_2$ concentrations between 1.5 and 4.5 mM

-Annealing temperatures between 55 and 59 ºC and touchdown PCR
(Stepwise decrease in annealing temperature)

-Hot-start polymerase enzymes:
AmplitaqGold (Applied Biosystems) and Tempase Hot-start (Bergman)

For detailed experimental procedures, see appendix.

The final PCR reactions were run in 96 well plates, one plate covering all the strains investigated including positive and negative controls. The 4 multiplex combinations were run

on separate plates. The PCR products were examined by capillary electrophoresis on a 3730 DNA Analyzer (Applied Biosystems). Each well on the 96 well plates contained 0.5 µl PCR product, 0.5 µl ROX size standard (Applied Biosystems) and 9 µl Hi-Di Formamide (solvent).

2.4 The Hunter-Gaston discriminatory index

The Hunter-Gaston discriminatory index (HGDI) [64] is a measure of the discriminatory power of a typing method. It can also be used to quantify the discriminatory power of individual loci used for typing. The HGDI is calculated as follows:

$$D = 1 - \frac{1}{N(N-1)} \sum_{j=1}^{S} n_j(n_j - 1)$$

Where D is the discriminatory power, N is the total number of strains in the sample, S is the number of types described and n_j is the number of strains belonging to the jth type.

The HGDI was calculated for our spoligotyping results, the MIRU typing results, for the individual MIRU loci and for the spoligtyping and MIRU typing results combined.

2.5 Phylogenetic analyses

2.5.1 Tree building

The NJ tree was constructed as described in (I)

2.5.2. '*Structure'*

The program *Structure* [65, 66] implements a Bayesian approach to identify population structure. It was decided to use a "no-admixture" model, as *M. tuberculosis* is a bacterium with no known horizontal gene transfer. That is, the model assumes that each bacterium has

its ancestry from one population, with no admixture having taken place. The model also assumes that the markers are not in linkage disequilibrium (LD) *within* populations. That means that the markers should not be too close together, but the program can handle some linkage. The variable sequences and direct repeats in the DR region of *M. tuberculosis* are quite closely linked, but it is difficult to say whether the linkage is too tight or not. 'Structure' uses the input information to estimate how many populations the data set is most likely made up of. With the no-admixture model, the prior probability of K populations is 1/K [65]. The posterior probabilities are then estimated with a Markov Chain Monte Carlo algorithm with a burn-in of 100 000 repeats and 400 000 Markov Chain Monte Carlo repeats. 65% assigned membership to a group was used as a threshold value in figure **X**.

2.6. Analysis of Beijing lineage incidence in Norway

We wanted to find out whether the Beijing lineage was present in Norway in recent years. One commonality in the RFLP patterns of Beijing strains is a relatively large number of bands (15-26) [51]. It was decided to analyze all strains with more than 10 bands from IS6110 RFLP typing in the last decade with spoligotyping. Time only allowed strains from 2003-2005 to be analyzed. It has also been proposed that Beijing isolates can be identified with RFLP by comparison to 16 reference strains and using a threshold of similarity to decide whether a strain is a member of the Beijing lineage or not, but it is uncertain how accurate this method is. It was thus decided to carry out spoligotyping on strains selected on the above criterion which is quite certain not to leave out any Beijing strains.

2.7. AFLP preparatory work

The TB genome was uploaded to the web-based software NEBcutter V2.0. Then a virtual restriction cutting was performed to see how often the various restriction enzymes cut. A combination of the EcoRI and Mse restriction enzymes resulted in about 650 fragments between 70 and 500 base pairs. The GC content of the M. tuberculosis genome is 64 %, leaving 36% A and T nucleotides. Thus the addition of an A or T nucleotide for selective amplification would produce about 650 x 0.18 = 117 fragments which is a suitable number. At this stage it was decided not go any further, and no lab work was carried out.

3. RESULTS

3.1. TB in Dar es Salaam

The main findings of the study are that three main families, namely the East African, Central Asian and East African Indian lineages, are responsible for the ongoing TB epidemic in Dar es Salaam and that import of new strains appears to be a minor problem (I)

The spoligotypes were used to construct an NJ-tree, which is shown as figure 1 in (I). Figure 3.1 shows a bar plot of the assignment to populations by *Structure*, which is not included in the paper. This assignment is also shown in figure 1 (I), but only for isolates with >65% assigned membership to a group, that is 135 of 147 isolates.

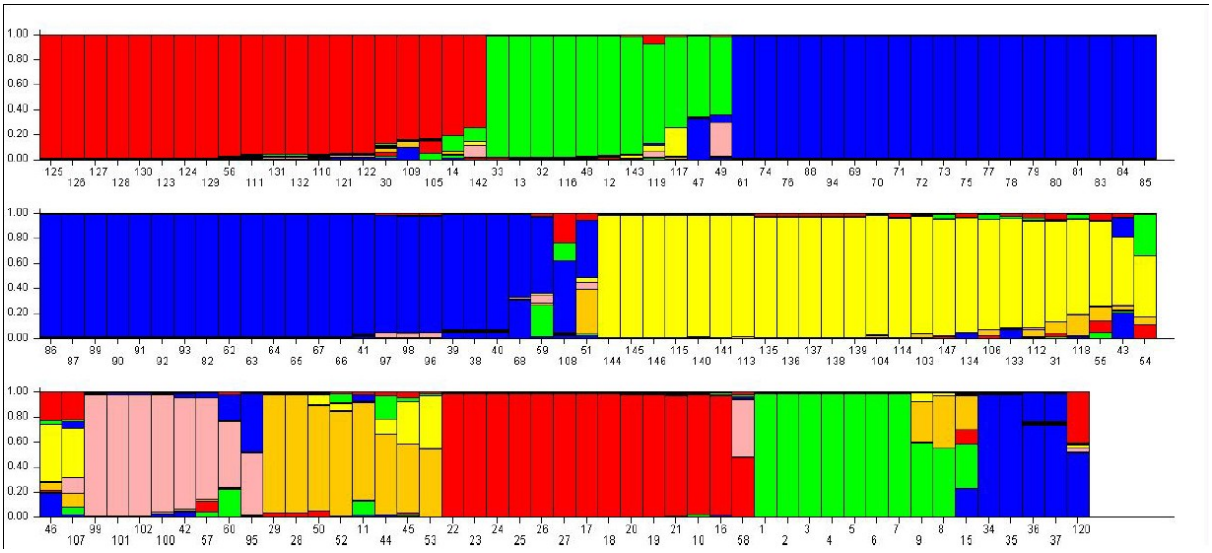


Figure 3.1. The 147 isolates assigned to one or more groups. The isolates are spread along the x axis, whereas the scale on the y axis shows the relative assignment of isolates to one or more group. Most isolates clearly fall into one group.

The Hunter-Gaston discriminatory index for the spoligotyping was 0.9557, which is quite good. This high number reflects the high number of unique spoligotypes in the Dar es Salaam study, but there are still some large clusters that would be interesting to type with alternative methods.

Transferring the MIRU typing results from the ABI3730 sequencer to the Genemapper program (Applied biosystems) was not successful. According to Applied Biosystems this could be due to the ROX size standard, which is "tricky". The results could thus not be analyzed. But visual inspection of the raw data revealed that loci 4-6 yielded quite good results, that is three clear bands, for more than half of the isolates investigated, whereas the rest of the multiplex reactions yielded few or no bands.

3.2. The Beijing family in Norway

In the study of the presence of Beijing strains in Norway, no apparent increase was found during the last three years. The estimated number of isolates was 8, 18 and 10 in the years 2003, 2004 and 2005 respectively (table 3.1).

Table 3.1. Incidence of Beijing strains in Norway 2003-2006

	2003	**2004**	**2005**
Total TB incidence	339	302	290
Isolates fulfilling selection criteria	115	90	81
Isolates spoligotyped	43	40	67
Identified as Beijing	3	8	8
Estimated Beijing incidence*	8	18	10
Estimated Beijing (%)	2.4	6.0	3.4

* estimated as (No. of Beijing isolates identified x No. of isolates spoligotyped) / No. of isolates fulfilling selection criteria

The isolates had previously been typed with IS6110 RFLP. By visual inspection of the Bionumerics (Applied Maths) database hosted at NIPH, it was found that none of the Beijing isolates were part of clusters (outbreaks).

4. DISCUSSION

4.1. The Dar es Salaam study

This isolates examined in this study were collected during October and November 2005. A study period of only two months contracts some problems when molecular epidemiological studies are carried out. The rate of recent transmission is generally calculated using the "n-1" formula [67], that is (the number of isolates in clusters – number of clusters) / number of isolates in the study. This formula assumes that a cluster consist of one individual with reactivated disease (index case) and n-1 individuals with recently acquired disease who are part of a transmission chain with the index case as its original source. So when the number of clusters is subtracted from the number of isolates in clusters, all the index cases are removed, as these are not victims of recent transmission.
Due to the chronic nature of TB pathogenesis, the isolates analyzed must have been collected over a long period of time, typically years, if one is to be relatively certain that the index case is included in each cluster. With this in mind, it is obvious that the rate of transmission cannot be calculated in this study, which was carried out using isolates collected over two months. The degree of diversity was calculated though, which gives an idea of the transmission rates. A low diversity is indicative of high rates of transmission, but the actual rate remains uncertain.

Many DNA polymerases catalyze the addition of a single adenosine (A) nucleotide to the 3' ends of the double stranded DNA product fragment. This results in products that are one base pair longer than the original template. The 3' A nucleotide addition is a more ore less random process that does not go to completion without a long extension step at the end of the PCR cycling. It is easier to add a long (30-45 minutes) extension step so that nearly all fragments are in the "plus-A" form than preventing any 3' A nucleotide addition from taking place. Having all fragments in one of the two forms is an advantage when the results are analyzed on a sequencer, because split peaks are avoided. In this minisatellite study, the repeats are long (>50), meaning that the problem would be negligible compared to a microsatellite study with repeats down to 2 and 3 base pairs in size. Thus, it was decided not to add a long final extension step beyond the seven minutes final extension step described in many general PCR protocols.

Structure is a model-based clustering program that works best with unlinked markers. The recent version used (*Structure*, version 2) allows the markers to be linked, but the authors state that software works best with unlinked markers [65]. The variable regions in the DR region are closely linked, but we still wanted to see how the program interpreted our data. As stated (I), the results from the analysis using *Structure* were found to fit the NJ tree quite well. This was interpreted as giving strength to the phylogenetic tree.

The MIRU typing was not considered within the scope of this thesis at the outset, and the time timeframe did not allow the method to be established in finality. The reason for this lack of success is difficult to identify within the timeframe of this thesis. A successful PCR reaction depends on a multitude of factors such as the concentration of $MgCl_2$, KCl, DNA, dNTP, primers, the quality and quantity of the DNA template, self-priming of primers, times and temperatures of the PCR cycles etc. A lot of these factors were investigated, including $MgCl_2$ concentrations (final concentrations of 1.5 – 4.5 mM), dNTP concentrations (100-250 μM each nucleotide) and annealing times and temperatures. The different parameters were tried with different templates and with primers in single and multiplex reactions. This, however, did not result in stably successful amplification of the MIRU loci to be investigated.

MIRU typing still holds potential for becoming the major typing method for epidemiological studies of mycobacteria, a process that is already well under way. Philip Supply, the researcher who first discovered MIRU loci in *M. tuberculosis* [39] has recently put together a combination of 15 loci with a corresponding technical guide, in an effort to establish an international standard procedure for MIRU typing (Supply P, personal communication). This protocol might very well become the standard method of future TB molecular epidemiological studies.

The protocol also describes the use of 24 loci for phylogenetic studies of mycobacteria. This approach may be more problematic, due to the mutation process of minisatellites. Microsatellite loci in eukaryotes evolve according to the so-called ‘step-wise mutation’ model [68], the physical background for this is thought to be slipped-strand mispairing during DNA replication [69]. That is, the loci can loose or gain repeats due to replication errors. This mutation process allows the same microsatellite loci to arise via different routes. One result of this is that convergence of genotypes will be an inherent problem for phylogenetic studies of distantly related organisms. In one study, Dettman *et al*. [70] set out to find out at which level mutational saturation of microsatellite loci occurred within and between different species of

the haploid eukaryote Neurospora. This was done by sequencing the microsatellite repeat arrays and their flanking regions at multiple loci. As sequence variance in the flanking regions can be assumed to mutate significantly slower than the repeat arrays, incongruence between the genetic distance obtained by analyzing the repeat arrays and the flanking regions were used to investigate at which point of sequence divergence between the flanking regions of specific loci started to diverge from the distance obtained by analyzing the repeats. At this point, 'mutational saturation' was said to have occurred, and the mincrosatellites could not be used with confidence for phylogenetic calculations. The authors found mutational saturation to be a problem for analyses between species, but that the microsatellites could be trusted for phylogenetic studies within Neurospora species.

Minisatellites have longer repeats than microsatellite loci, but are also thought to mutate by the 'step-wise mutation' model. Exactly how MIRU loci in *M. tuberculosis* evolve is unknown, but it has been observed that the general trend is that ancestral strains have a higher number of repeats than modern strains, or in other words, there is a tendency of loss of repeats over time [71]. A phylogenetic tree of 219 world wide isolates constructed by 12 MIRU loci [41] did not match the corresponding, and more reliable tree constructed by using 212 SNP markers well [22]. Phylogenetic studies using only MIRU markers should therefore be interpreted with caution.

For phylogenetic studies, molecular markers thought to be 'unique event polymorphisms', such as SNPs and LSPs, are probably more robust and informative [22, 26]. As high-throughput SNP analyses become cheaper and more accessible, we will probably see more SNP based studies of *M. tuberculosis*. Compared to LSP and SNP typing with a large panel of genome wide markers, spoligotyping falls short. Spoligotyping might be more robust than MIRU typing for phylogenetic studies, but the resolution for epidemiological studies is substantially lower.

Spoligotyping has major advantages, though. The method is simple and the need for costly instruments is low. In addition to standard microbiology lab and safety equipment, all that is needed is a PCR cycler, a 'miniblotter' and photographic equipment to take pictures of the spoligomembrane (included in the spoligotyping kit) is needed. The method is thus appropriate for low-income countries, for outbreak investigation and general insight into ongoing epidemics. The relatively low resolution of the method, calls for secondary genotyping and / or contact tracing to identify true epidemiological links, which may otherwise be overestimated. In high income countries, spoligotyping and MIRU appear to

represent a feasible combination, whereas many low-income countries will probably find spoligotyping in combination with contact tracing to be a useful tool in determining true transmission events.

4.2. TB in Norway

The TB incidence in Norway has been increasing slowly but steadily in recent years. This study only assessed the Beijing incidence in the years 2003-2005. The numbers were too small to reveal any trends, but the study confirmed that the family is present. The fact that none of these strains were associated with outbreaks is interesting. The Beijing family is not an old one in Norway, and in all cases, the patients are immigrants. The fact that the family is not causing outbreaks indicates that this family is not an immediate threat to the Norwegian public health. On the contrary, recent findings (UR Dahle, unpublished results) show the increase of TB cases to be tightly and directly coupled to increased import of strains, whereas few of these strains are able to cause outbreaks in Norway.

4.3. The burden of TB

The burden TB lays on people in low-income countries, especially the former Soviet Union and sub-Saharan Africa is enormous. The death toll is high, with approximately 1.8 million people succumbing to the disease each year [1]. In addition to this, countless numbers are debilitated, chronically or transiently by the disease, placing a heavy load on individuals, families, health care systems and the total work force in the countries most seriously affected by the disease. The TB and AIDS epidemics in sub-Saharan Africa are closely linked, and must be controlled in order to increase the quality of life and life expectancies in this area. In low-income countries where TB is epidemic, molecular epidemiological studies are generally a luxury that can not be afforded, whereas many countries have reasonably good services available for TB patients who can afford to get to a clinic. Molecular epidemiological studies are very helpful, though, for identifying hot-spots of TB transmission, and can facilitate efficient intervention. In a state-wide study conducted in Maryland, USA, the authors found that less than two thirds of recent patient's source of infection could be located by contact tracing alone compared to contact tracing with the help of genotyping (RFLP and

spoligotyping) [72]. The study found genotyping to be especially helpful in identification of transmission events in non-traditional setting (excluding transmission between household members, family and close friends). This is an important finding, as a large portion of major outbreaks investigated in the study began in non-traditional settings. It thus looks like molecular epidemiological studies and contact tracing combined with efficient treatment of patients should be sought in order to get these epidemics under control. Spoligotyping is a good candidate method for this kind of studies in Tanzania and other high-incidence countries.

This study identified high rates of clustering of TB cases in, Dar es Salaam. As mentioned above, spoligotyping alone is not sufficient to, identify recent transmission from one case to another, but this clearly indicates that a great portion of TB cases are caused by recent transmission. Combined with contact tracing, and conducted over a longer time period, this kind of study would be valuable for identification of transmission hot-spots and epidemiological links in this city.

4.4. Conclusions

We found that the CAS, LAM and EAI are the main families causing tuberculosis in Dar es Salaam, Tanzania, and that import of strains from abroad seems to be a minor problem (I). Internationally there have been worries that the Beijing family of *M. tuberculosis* might be involved in a global emerging epidemic. Our results from Norway did not support a notion of an increase in TB cases caused by this family of strains.

For rapid assessment of the epidemiology at a certain location, spoligotyping is an efficient method. The existence of large international databases for comparison is also a good advantage. For implementation of molecular epidemiological methods in low-income countries, spoligotyping is a strong candidate. The discriminatory power is the main drawback, and studies must be coupled with contact tracing for certain inferences to be made regarding TB transmission, when the method is used alone.

Despite the limited success in establishing capillary electrophoresis based 12 loci MIRU typing in this study, it would be wise to implement this method in countries where human and financial capital is available. The major advantages are the rapidity of the genotyping and the discriminatory power obtained. The newly proposed standardized MIRU methodology will probably make this method even more popular.

5.1. Additional work

This thesis focuses on molecular epidemiology which builds on comparative aspects of genetics. Functional genetics and genomics are increasingly important fields of study. In order to learn more about studies of functional genomics, a one-week intensive introductory course in microarray technology in Tromsø was attended by the current author. The course was held in collaboration with the Norwegian Microarray consortium.

The findings from the Tanzania study were presented as a poster at the 37th Union World Conference on Lung Health, Paris (see appendix).

6. REFERENCES

1. Cole ST, Eisenach DK, McMurray DN, Jacobs WR: *Tuberculosis and the Tubercle Bacillus.* ASM Press; 2004.
2. Dye C, Fengzeng Z, Scheele S, Williams B: **Evaluating the impact of tuberculosis control: number of deaths prevented by short-course chemotherapy in China.** *Int J Epidemiol* 2000, **29:**558-564.
3. Rieder HL, Cauthen GM, Kelly GD, Bloch AB, Snider Jr DE: **Tuberculosis in the United States.** *Journal of the American Medical Association* 1989, **262:**385-389.
4. Raviglione MC, Sudre P, Rieder HL, Spinaci S, A. K: **Secular trends of tuberculosis in western Europe.** *Bulletin of the World Health Organization* 1993, **71:**297-306.
5. WHO: **Global tuberculosis control: surveillance, planning, financing. WHO report 2005. WHO/HTM/TB/2005.349.** Geneva: World Health Organization; 2005.
6. Lucas SB, Hounnou A, Peacock C, Beaumel A, Djomand G, N'Gbichi JM, Yeboue K, Honde M, Diomande M, Giordano C, al. e: **The mortality and pathology of HIV infection in a west African city.** *AIDS* 1993, **12:**1569-1579.
7. **United Republic of Tanzania Ministry of Health: National Tuberculosis and Leprosy Programme. Annual Report.** Dar es Salaam; 2003.
8. Heldal E: **Tuberculosis in Norway 2001 MSIS Rapport 30:20 (In Norwegian).** 2002.
9. Dahle UR, Sandven P, Heldal E, Caugant DA: **Continued Low Rates of Transmission of Mycobacterium tuberculosis in Norway.** *J Clin Microbiol* 2003, **41:**2968-2973.
10. Dahle UR, Sandven P, Heldal E, Caugant DA: **Molecular Epidemiology of Mycobacterium tuberculosis in Norway.** *J Clin Microbiol* 2001, **39:**1802-1807.
11. Lillebaek T, Andersen AB, Bauer J, Dirksen A, Glismann S, de Haas P, Kok-Jensen A: **Risk of Mycobacterium tuberculosis Transmission in a Low-Incidence Country Due to Immigration from High-Incidence Areas.** *J Clin Microbiol* 2001, **39:**855-861.
12. Maguire H, Dale JW, McHugh TD, Butcher PD, Gillespie SH, Costetsos A, Al-Ghusein H, Holland R, Dickens A, Marston L, et al: **Molecular epidemiology of tuberculosis in London 1995-7 showing low rate of active transmission.** *Thorax* 2002, **57:**617-622.
13. Heywood N, Kawa B, Long R, Njoo H, Panaro L, Wobeser W: **Guidelines for the investigation and follow-up of individuals under medical surveillance for tuberculosis after arriving in Canada: a summary.** *CMAJ* 2003, **168:**1563-1565.
14. Brudey K, Filliol I, Ferdinand S, Guernier V, Duval P, Maubert B, Sola C, Rastogi N: **Long-Term Population-Based Genotyping Study of Mycobacterium tuberculosis Complex Isolates in the French Departments of the Americas.** *J Clin Microbiol* 2006, **44:**183-191.
15. Ryan KJ, Ray CG: *Sherris Medical Microbiology, 4th ed.*: McGraw Hill; 2004.
16. Comstock GW: **Epidemiology of tuberculosis.** *The American Review of Respiratory Disease* 1982, **125:**8-15.
17. Yang Z, Yang D, Kong Y, Zhang L, Marrs CF, Foxman B, Bates JH, Wilson F, Cave MD: **Clinical Relevance of Mycobacterium tuberculosis plcD Gene Mutations.** *Am J Respir Crit Care Med* 2005, **171:**1436-1442.

18. CDC: **Guidelines for Using the QuantiFERON®-TB Gold Test for Detecting Mycobacterium tuberculosis Infection, United States.** 2005, http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5415a4.htm.

19. Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, DeBoy R, Dodson R, Gwinn M, Haft D, et al: **Whole-Genome Comparison of Mycobacterium tuberculosis Clinical and Laboratory Strains.** *J Bacteriol* 2002, **184:**5479-5490.

20. Hughes AL FRaMM: **Genomewide Pattern of Synonymous Nucleotide Substitution in Two Complete Genomes of Mycobacterium tuberculosis** *Emerging Infectious Diseases* 2002, **8:**1342-1346.

21. Diamond J: *Guns, Germs and Steel: The Fates of Human Societies.* New York: Norton; 1997.

22. Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbon MH, Bobadilla del Valle M, Fyfe J, Garcia-Garcia L, Rastogi N, Sola C, 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 SNP Set.** *J Bacteriol* 2006, **188:**759-772.

23. Gutierrez MC, Ahmed N, Willer E, Narayanan S, Hasnain SE, Chauhan DSea: **Predominance of ancestral lineages of Mycobacterium tuberculosis in India.** *Emerging Infectious Diseases* 2006, **12:**1367-1374.

24. Filliol I, Driscoll J, van Soolingen D, Kreiswirth B, Kremer K, Valétudie G, Anh D, Barlow R, Banerjee D, Bifani P, Brudey K: **Global Distribution of Mycobacterium tuberculosis Spoligotypes.** *Emerging Infectious Diseases* 2002, **8:**1347-1349.

25. Gutierrez MC, Brisse S, Brosch R, Fabre M, Oma, s B, Marmiesse M, Supply P, Vincent V: **Ancient Origin and Gene Mosaicism of the Progenitor of Mycobacterium tuberculosis.** *PLoS Pathogens* 2005, **1:**e5.

26. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, Nicol M, Niemann S, Kremer K, Gutierrez MC, et al: **Variable host-pathogen compatibility in Mycobacterium tuberculosis.** *PNAS* 2006, **103:**2869-2873.

27. Supply P, Warren RM, Banuls A-L, Lesjean S, van der Spuy GD, Lewis L-A, Tibayrenc M, van Helden PD, Locht C: **Linkage disequilibrium between minisatellite loci supports clonal evolution of Mycobacterium tuberculosis in a high tuberculosis incidence area.** *Molecular Microbiology* 2003, **47:**529-538.

28. Dahle UR: **TB in immigrants is not public health risk, but uncontrolled epidemics are.** *BMJ* 2005, **331:**237-b-.

29. Hirsh AE, Tsolaki AG, DeRiemer K, Feldman MW, Small PM: **Stable association between strains of Mycobacterium tuberculosis and their human host populations.** *PNAS* 2004, **101:**4871-4876.

30. Daley CL, Small PM, Schecter GF, Schoolnik GK, McAdam RA, Jacobs WR, Hopewell PC: **An outbreak of tuberculosis with accelerated progression among persons infected with the human immunodeficiency virus. An analysis using restriction-fragment-length polymorphisms.** *N Engl J Med* 1992, **326:**231-235.

31. Alland D, Kalkut GE, Moss AR, McAdam RA, Hahn JA, Bosworth W, Drucker E, Bloom BR: **Transmission of Tuberculosis in New York City -- An Analysis by DNA Fingerprinting and Conventional Epidemiologic Methods.** *N Engl J Med* 1994, **330:**1710-1716.

32. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J: **Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology.** *J Clin Microbiol* 1997, **35:**907-914.

33. Jansen R, van Embden JDA, Gaastra W, Schouls LM: **Identification of a Novel Family of Sequence Repeats among Prokaryotes.** *OMICS: A Journal of Integrative Biology* 2002, **6:**23-33.

34. Mojica FJM, Díez-Villaseñor C, García-Martinez J, Elena S: **Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements.** *Journal of Molecular Evolution* 2005, **V60:**174-182.

35. Godde JS, Bickerton A: **The Repetitive DNA Elements Called CRISPRs and Their Associated Genes: Evidence of Horizontal Transfer Among Prokaryotes.** *Journal of Molecular Evolution* 2006, **V62:**718-729.

36. Brudey K, Driscoll J, Rigouts L, Prodinger W, Gori A, Al-Hajoj S, Allix C, Aristimuno L, Arora J, Baumanis V, et al: **Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology.** *BMC Microbiology* 2006, **6:**23.

37. Vitol I, Driscoll J, Kreiswirth B, Kurepina N, Bennett KP: **Identifying Mycobacterium tuberculosis complex strain families using spoligotypes.** *Infection, Genetics and Evolution*, **In Press, Corrected Proof**.

38. Kremer K, van Soolingen D, Frothingham R, Haas WH, Hermans PWM, Martin C, Palittapongarnpim P, Plikaytis BB, Riley LW, Yakrus MA, et al: **Comparison of Methods Based on Different Molecular Epidemiological Markers for Typing of Mycobacterium tuberculosis Complex Strains: Interlaboratory Study of Discriminatory Power and Reproducibility.** *J Clin Microbiol* 1999, **37:**2607-2618.

39. Supply P, Magdalena J, Himpens S, Locht C: **Identification of novel intergenic repetitive units in a mycobacterial two-component system operon.** *Molecular Microbiology* 1997, **26:**991-1003.

40. Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C: **Variable human minisatellite-like regions in the Mycobacterium tuberculosis genome.** *Molecular Microbiology* 2000, **36:**762-771.

41. Supply P, Lesjean S, Savine E, Kremer K, van Soolingen D, Locht C: **Automated High-Throughput Genotyping for Study of Global Epidemiology of Mycobacterium tuberculosis Based on Mycobacterial Interspersed Repetitive Units.** *J Clin Microbiol* 2001, **39:**3563-3571.

42. Frothingham R, Meeker-O'Connell WA: **Genetic diversity in the Mycobacterium tuberculosis complex based on variable numbers of tandem DNA repeats.** *Microbiology* 1998, **144:**1189-1196.

43. Skuce RA, McCorry TP, McCarroll JF, Roring SMM, Scott AN, Brittain D, Hughes SL, Hewinson RG, Neill SD: **Discrimination of Mycobacterium tuberculosis complex bacteria using novel VNTR-PCR targets.** *Microbiology* 2002, **148:**519-528.

44. Kremer K, Arnold C, Cataldi A, Gutierrez MC, Haas WH, Panaiotov S, Skuce RA, Supply P, van der Zanden AGM, van Soolingen D: **Discriminatory Power and Reproducibility of Novel DNA Typing Methods for Mycobacterium tuberculosis Complex Strains.** *J Clin Microbiol* 2005, **43:**5628-5638.

45. van Deutekom H, Supply P, de Haas PEW, Willery E, Hoijng SP, Locht C, Coutinho RA, van Soolingen D: **Molecular Typing of Mycobacterium tuberculosis by Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Analysis, a More Accurate Method for Identifying Epidemiological Links between Patients with Tuberculosis.** *J Clin Microbiol* 2005, **43:**4473-4479.

46. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rusch-Gerdes S, Willery E, Savine E, de Haas P, van Deutekom H, Roring S, et al: **Proposal for standardization of optimized Mycobacterial Interspersed Repetitive Unit-Variable Number Tandem Repeat typing of Mycobacterium tuberculosis.** *J Clin Microbiol* 2006**:**JCM.01392-01306.

47. Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, M K: **AFLP: a new technique for DNA fingerprinting.** *Nucleic Acids Research* 1995, **23:**4407-4414.

48. Mueller UG, Wolfenbarger LL: **AFLP genotyping and fingerprinting.** *Trends in Ecology & Evolution* 1999, **14:**389-394.

49. van Soolingen D, Qian L, de Haas PE, Douglas JT, Traore H, Portaels F, Qing HZ, Enkhsaikan D, Nymadawa P, van Embden JD: **Predominance of a single genotype of Mycobacterium tuberculosis in countries of east Asia.** *J Clin Microbiol* 1995, **33:**3234-3238.

50. Bifani PJ, Mathema B, Liu Z, Moghazeh SL, Shopsin B, Tempalski B, Driscoll J, Frothingham R, Musser JM, Alcabes P, Kreiswirth BN: **Identification of a W Variant Outbreak of Mycobacterium tuberculosis via Population-Based Molecular Epidemiology.** *JAMA* 1999, **282:**2321-2327.

51. Bifani PJ, Mathema B, Kurepina NE, Kreiswirth BN: **Global dissemination of the Mycobacterium tuberculosis W-Beijing family strains.** *Trends in Microbiology* 2002, **10:**45-52.

52. van der Zanden AGM, Kremer K, Schouls LM, Caimi K, Cataldi A, Hulleman A, Nagelkerke NJD, van Soolingen D: **Improvement of Differentiation and Interpretability of Spoligotyping for Mycobacterium tuberculosis Complex Isolates by Introduction of New Spacer Oligonucleotides.** *J Clin Microbiol* 2002, **40:**4628-4639.

53. Dahle UR, Sandven P, Heldal E, Mannsaaker T, Caugant DA: **Deciphering an Outbreak of Drug-Resistant Mycobacterium tuberculosis.** *J Clin Microbiol* 2003, **41:**67-72.

54. Jaccard P: **Nouvelles récherches sur la distribution florale.** *Bulletin de la Société de Vaud Sciences Naturelles* 1908, **44:**223-270.

55. Nei M, Kumar s: *Molecular evolution and phylogenetics.* New York: Oxford University Press; 2000.

56. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4:**406-425.

57. Yang ZH, Mtoni I, Chonde M, Mwasekaga M, Fuursted K, Askgard DS, Bennedsen J, de Haas PE, van Soolingen D, van Embden JD: **DNA fingerprinting and phenotyping of Mycobacterium tuberculosis isolates from human immunodeficiency virus (HIV)-seropositive and HIV- seronegative patients in Tanzania.** *J Clin Microbiol* 1995, **33:**1064-1069.

58. McHugh TD, Batt SL, Shorten RJ, Gosling RD, Uiso L, Gillespie SH: **Mycobacterium tuberculosis lineage: A naming of the parts.** *Tuberculosis* 2005, **85:**127-136.

59. Gillespie SH, Kennedy N, Ngowi FI, Fomukong NG, Al-Maamary S, Dale JW: **Restriction fragment length polymorphism analysis of Mycobacterium tuberculosis isolated from patients with pulmonary tuberculosis in northern Tanzania.** *Transactions of the Royal Society of Tropical Medicine and Hygiene* 1995, **89:**335-338.

60. van Soolingen D, de Haas, P. E. W. and Kremer, K.: *Restriction fragment length polymorphism (RFLP) typing of Mycobacteria.* Bilthoven: National Institute of Public Health and the Environment; 1999.

61. Thompson JD, Plewniak F, Poch O: **A comprehensive comparison of multiple sequence alignment programs.** *Nucl Acids Res* 1999, **27:**2682-2690.

62. Rozen S, Skaletsky HJ: *Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology.* Humana Press, Totowa, NJ; 2000.

63. Vallone PM, Butler JM: **AutoDimer: a screening tool for primer-dimer and hairpin structures. .** *BioTechniques* 2004, **37:**226-231.

64. Hunter PR, Gaston MA: **Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity.** *J Clin Microbiol* 1988, **26:**2465-2466.

65. Pritchard JK, Wen W: **Documentation for structure software: Version 2.** 2004 http://pritch.bsd.uchicago.edu.

66. Pritchard JK, Stephens M, Donnelly P: **Inference of Population Structure Using Multilocus Genotype Data.** *Genetics* 2000, **155:**945-959.

67. Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, Schecter GF, Daley CL, Schoolnik GK: **The Epidemiology of Tuberculosis in San Francisco -- A Population-Based Study Using Conventional and Molecular Methods.** *N Engl J Med* 1994, **330:**1703-1709.

68. Kimura M, Ohta T: **Stepwise mutation model and distribution of allelic frequencies in a finite population.** *Proceedings of the National Academy of Sciences of the United States of America* 1978, **75:**2868-2872.

69. Levinson G, Gutman GA: **Slipped-strand mispairing: a major mechanism for DNA sequence evolution.** *Mol Biol Evol* 1987, **4:**203-221.

70. Dettman JR, Taylor JW: **Mutation and Evolution of Microsatellite Loci in Neurospora.** *Genetics* 2004, **168:**1231-1248.

71. Arnold C, Thorne N, Underwood A, Baster K, Gharbia S: **Evolution of short sequence repeats in Mycobacterium tuberculosis.** *FEMS Microbiology Letters* 2006, **256:**340-346.

72. Cronin WA, Golub JE, Lathan MJ, Mukasa LN, Hooper N, Razeq JH, Baruch NG, Mulcahy D, Benjamin WH, Magder LS, et al: **Molecular Epidemiology of Tuberculosis in a Low- to Moderate-Incidence State: Are Contact Investigations Enough?** *Emerging Infectious Diseases* 2002, **Vol. 8:**1271-1279.

BMC Microbiology

BioMed Central

Research article

Open Access

A first insight into the genetic diversity of *Mycobacterium tuberculosis* in Dar es Salaam, Tanzania, assessed by spoligotyping

Vegard Eldholm[1,2], Mecky Matee[3], Sayoki GM Mfinanga[3], Manfred Heun[2] and Ulf R Dahle*[1]

Address: [1]Division of Infectious Disease Control, Norwegian Institute of Public Health, Oslo, Norway, [2]Institute of Nature Resource Management, Norwegian University of Life Sciences, Ås, Norway and [3]Muhimbili Medical Research Centre, Dar es Salaam, Tanzania

Email: Vegard Eldholm - vegard.eldholm@fhi.no; Mecky Matee - mmatee@muchs.ac.tz; Sayoki GM Mfinanga - gsmfinanga@yahoo.com; Manfred Heun - manfred.heun@umb.no; Ulf R Dahle* - ulf.dahle@fhi.no

* Corresponding author

Published: 13 September 2006

BMC Microbiology 2006, **6**:76 doi:10.1186/1471-2180-6-76

Received: 23 June 2006
Accepted: 13 September 2006

This article is available from: http://www.biomedcentral.com/1471-2180/6/76

Abstract

Background: Tanzania has a high tuberculosis incidence, and genotyping studies of *Mycobacterium tuberculosis* in the country are necessary in order to improve our understanding of the epidemic. Spoligotyping is a potentially powerful genotyping method due to fast generation of genotyping results, high reproducibility and low operation costs. The recently constructed SpolDB4 database and the model-based program 'Spotclust' can be used to assign isolates to families, subfamilies and variants. The results of a study can thus be analyzed in a global context.

Results: One hundred forty-seven pulmonary isolates from consecutive tuberculosis patients in Dar es Salaam were spoligotyped. SpolDB4 and 'Spotclust' were used to assign isolates to families, subfamilies and variants. The CAS (37%), LAM (22%) and EAI (17%) families were the most abundant. Despite the dominance of these three families, diversity was high due to variation within *M. tuberculosis* families. Of the obtained spoligopatterns, 64% were previously unrecorded.

Conclusion: Spoligotyping is useful to gain an overall understanding of the local TB epidemic. This study demonstrates that the extensive TB epidemic in Dar es Salaam, Tanzania is caused by a few successful *M. tuberculosis* families, dominated by the CAS family. Import of strains was a minor problem.

Background

In Tanzania, the tuberculosis (TB) incidence doubled between 1990 and 2004 [1]. The rate of all forms of the disease is estimated at 524/100,000 and the rate of new sputum smear positive disease is approximately 157/100,000 [1] with Dar es Salaam contributing about 26% of all TB cases [2]. The World Health Organization estimates that Tanzania has the 14th highest TB burden in the world [1]. Points of concern include the proportion of patients lost to follow-up, currently at 9%, an average diagnostic delay of 6 months, decreasing case detection rate (from 55% in 1997 to 45% in 2004) and the continuing high prevalence of HIV [3]. The high case rate in many African countries has contributed to a rise of the global TB incidence, despite stable or declining rates in the rest of the world [1]. Tanzania with its 37 million inhabitants, has 701 district laboratories diagnosing TB, three laboratories culturing *M. tuberculosis* and one

National reference laboratory that perform drug susceptibility testing of *M. tuberculosis* isolates. Measures are undertaken to establish molecular genotyping methods such as spoligotyping [4], but currently no laboratory in Tanzania offers this service. Previous studies have described the molecular epidemiology of Tanzanian *M. tuberculosis* collections from the first half of the 1990s [5-7]. Spoligotyping is a PCR-based fingerprinting method that detects the presence or absence of 43 defined spacers situated between short direct repeat (DR) sequences in the genomes of members of the *M. tuberculosis* complex. Important advantages of spoligotyping are that it is cheap, easy to perform and fast. In addition, it has been demonstrated that the results are highly reproducible [8]. Unique to spoligotyping results are tools like the SpolDB4 database [9] and the web-based computer algorithm 'Spotclust' [10] that can be used to assign new isolates to families, subfamilies and variants (SpolDB4 only). SpolDB4 is the largest and most up to date available global database for spoligotypes. For previously not reported spoligopatterns, the 'Spotclust' database is a good additional tool in that it can assign these patterns to families by using a computer algorithm based on studies of SpolDB3 [10]. The results from local studies can thus be analyzed and compared to the global *M. tuberculosis* population. This may help us better understand the worldwide spread of common *M tuberculosis* families and subfamilies. In order to improve our understanding of the TB epidemic in this high-incidence country, the current ongoing study included *M. tuberculosis* strains collected in Dar es Salaam during October and November 2005. We describe the diversity of *M. tuberculosis* isolates from Dar es Salaam, Tanzania, based on spoligotyping, and identify the families and subfamilies responsible for the current persistence and spread of TB in this high-incidence community.

Results

Genetic diversity and family assignment

The 147 analyzed isolates gave 76 different spoligopatterns resulting in an overall diversity of 52%: 57 spoligopatterns occurred only once and 19 patterns comprised 90 of the isolates (61%) (table 1). Forty-nine (64%) patterns had not been described previously. The SpolDB4 database assigns isolates to families, subfamilies and often to variants, whereas 'Spotclust' assigns isolates to families and subfamilies, but is not designed to assign isolates to variants. Four spoligopatterns were assigned to different families and nine patterns were assigned to different subfamilies by the two methods. SpolDB4 assigned names were used whenever a spoligopatterns was found in the database, as this database is much larger than the SpolDB3 database, on which the 'Spotclust' algorithm is built. Patterns not found in SpolDB4 were assigned to families and subfamilies by 'Spotclust'. The family assignment showed that 37% of the isolates belonged to the Central Asian (CAS) family, 22% to the Latin American Mediterranean (LAM) family, and 17% to the East-African Indian (EAI) family. These three main families thus accounted for 76% of the incidences in Dar es Salaam. This family assignment also includes the spoligopatterns not described before. Eight isolates lacked spacers 4–7, 10 and 20–35, typical of the CAS1-kili variant, but in addition, they all also lacked spacer 2 (table 2). This spacer is typically present in CAS1-kili lineages and its absence has not previously been reported in these variants. We propose to name these variants CAS1-DAR, since they appear to be abundant in Dar es Salaam.

The rate of diversity (number of spoligotypes divided by the number of isolates) within each main family varied substantially and was 27, 54 and 72% for CAS, LAM and EAI, respectively. This may indicate that the CAS family is best adapted to spread within this community. The diversity of the *M. tuberculosis* population in Dar es Salaam (52%) was comparable to that described in previous studies from Tanzania [5-7]. In Delhi, India the genetic diversity of the *M. tuberculosis* population is 42% [11], but it is only 25% in Harare, Zimbabwe [12]. Thus, the diversity in high-incidence countries varies greatly and may be difficult to estimate without molecular epidemiological studies.

Phylogenetic studies

A Neighbor-joining (NJ) tree of all the isolates is shown in figure 1. The main families were well distinguished and a high diversity within and between families were observed. To confirm the reliability of the NJ tree, the program 'Structure' was applied on the underlying 43-digit binary spacer codes. The open boxes in figure 1 demonstrate the nine groups found to be the most likely number; the NJ branches were supported by the grouping via 'Structure'.

Discussion

The current study demonstrated that most isolates had at least one other closely related isolate in Dar es Salaam. Based on these preliminary findings, the TB epidemic appeared to result from a gradually evolving *M. tuberculosis* population rather than imported strains. A spoligotyping study conducted in the Ouest province of Cameroon found that 193 of 413 *M. tuberculosis* isolates belong to the Cameroon family (LAM10-CAM) [13]. In Harare, Zimbabwe, 68 of 214 isolates are LAM11-ZWE variants [12]. Of the 147 isolates in this study, three and eight isolates belonged to these variants respectively. The scarcity of these strains, abundant in other African countries, also indicated that the TB epidemic in Dar es Salaam is local and well established.

BMC Microbiology 2006, **6**:76 http://www.biomedcentral.com/1471-2180/6/76

Table 1: Spoligopatterns and family assignment

Spoligotype	Shared type	SPOLDB4	'Spotclust'	'Spotclust' probability	No. of isolates with identical pattern
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■	1	Beijing			7
□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■□□□□■■■■■□■	NEW		Family36	1.00	1
□□□□□□□□□□□□□□□□□□□□□□□□■■■■■■■■■□□□□■■■■■■■	4	LAM3/S			1
□□□□□□□□□□□□□■■■■■■■■■■■■■■■□□□□■□■■■■■■■■■	NEW		EAI5	1.00	1
□□□□□□□□□□■■■■□■■■■□□□■■■■□■□■■■□□□□■■■■■■■	NEW		H37Rv	0.99	1
□□□■■■■■■□■■□■□■■■■■■■■■■■■□■□□□□■□■■■■■□■■■	NEW		EAI5	1.00	1
□□□■■■■■■■■■■■■■■■■■□□□■■■■■■□□□□■□■■■■■■■■■	NEW		EAI2	1.00	1
□□□■■■■■■■■■■■■■■■■■■■□□□□■■□□□□■■□□□□■■□■■■■	1468	LAM11-ZWE			1
■□□□□□□□□□□□□□□■■■□□□□■■■■□□□□□□□□□□□□□■■□■	NEW		EAI1	1.00	1
■□□□□□□□□□□□□■□■■■■■■■■■■■■■□□□□■□■■■■□■■■■	1864	EAI5			1
■□□□□□□□□□□□□■□■■■■■■■■■■■■■□□□□■□■■■■■■■■■	NEW		EAI5	1.00	2
■□□□□□□□□□□□□■■■■■■■■■■■■■■■□□□□□□■■■■■■■■■	NEW		EAI5	1.00	2
■□□□□□□□□□□□□■■■■■■■■■■■■■■■□□□□■□■■■■■□■■■	NEW		EAI5	1.00	1
■□□□□□□□□□□□□■■■■■■■■■■■■■■■□□□□■□■■■■■■■■■	8	EAI5 / EAI3			6
■□□□□□□□□□■■■■□■■■■■■■■■■■■■■■■□□□□□■■■■■■■	NEW		S	0.98	2
■□□□■■■■■■■■■■■■■■■■□□□□■■□□□□■■□□□□■■■■■■■	NEW		LAM8	1.00	1
■□□□■■■■■■■■■■■■■■■■□□□□■■■■■■■■■□□□□■■■■■■■	964	LAM9			1
■□□■■■■■■□■■■■■■■■□□□■■■■■■□□□□■□■■■■□■■■■	NEW		EAI2	1.00	1
■□□■■■■■■■■■■□■■■■■■■■■■■■■■□□□□■□■■□□□■■■■	NEW		EAI3	1.00	1
■□■□□□□■■□■■■■□■■■■□□□□□□□□□□□□□□□□■■□□□□□■	NEW	CAS1-Dar			2
■□■□□□□■■□■■■■□■■■■□□□□□□□□□□□□□□□□■■□□■■■■	NEW	CAS1-Dar			2
■□■□□□□■■□■■■■□■■■■□□□□□□□□□□□□□□□□■■■■■■■■	NEW	CAS1-Dar			3
■□■□□□□■■□■■■■■■■■■□□□□□□□□□□□□□□□□■■■■■■■■	NEW	CAS1-Dar			1
■□■□□□□■■□■■■■■■■■■■■■□□□□□□□□□□□□□□□■■■■■■	NEW	CAS1-Dar			1
■□■■□■■■■□■■■■□■■■■■□□□□□□□□■■■■□□□□□■■■■■■■	NEW		LAM9	1.00	1
■□■■■■■□■□■■■■□■■■■■■■■■■■■■□■■■□□□□■□□■■□□	NEW		T1	0.99	1
■□■■■■■■■□■□■■□■□■■■□□■■■■■■■■■□□□□□■■■□■□■	NEW		T1	1.00	1
■□■■■■■■■□■■□■□■■■■■□□□□■■□■□■■■□□□□■■■□□■■	NEW		LAM9	1.00	1
■□■■■■■■■□■■■■□■■■□□□■□■■■□■□□□□□□■■□□□□□■■	NEW		EAI5	0.99	1
■□■■■■■■■□■■■■□■■■■■■■■■■■■■□□□□□□■■■□□□■■■	NEW		EAI5	1.00	1
■□■■■■■■■■■■■■□■■■■■■■■■□■□□□□□□□□■□■■■■■■■■■	NEW		EAI4	1.00	1
■■□□□□□□□□■■■■■■■■■■■■■■■■■■■■■■□□□□■■■■■■■	NEW		S	0.98	1
■■□□□□□■■■■■■□■■■■■□□□□■■□□□□■■□□□□□■■■□■■■	NEW		H37Rv	1.00	1
■■□□■■■■■□■■■□□■□□■■■■■■■■■■■■■□□□□□□□■□■■□	NEW		X1	0.91	1
■■□□■■■■■■■■■■■■□□■■■■■■■■■■■■■■■□□□□■■■□■■□	420	T2			1
■■□■■□■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■	NEW		Family33	1.00	1

Table 1: Spoligopatterns and family assignment *(Continued)*

	NEW		LAM3	1.00	1
	NEW		LAM8	1.00	1
	288	CAS2			1
	129	EAI5			1
	NEW		CAS	1.00	1
	NEW		CAS	1.00	1
	1675	CAS1-kili			6
	NEW		CAS	1.00	1
	21	CAS1-kili			27
	NEW		CAS	1.00	1
	22	CAS			3
	486	CAS			1
	NEW		CAS	1.00	1
	203	CAS1-Delhi			3
	NEW		T1	1.00	1
	205	T1			1
	NEW		Family33	1.00	1
	NEW		T1	1.00	1
	NEW		LAM10	1.00	1
	NEW		EAI4	0.96	1
	NEW		LAM8	1.00	1
	NEW		LAM8	1.00	2
	NEW		T1	1.00	1
	1166	T1			1
	150	LAM9			1
	NEW		EAI5	0.99	1
	NEW		Family33	1.00	1
	NEW		T1	1.00	1
	733	EAI5			1
	402	U			1
	811	LAM4			2
	59	LAM11-ZWE			8
	NEW		Family33	1.00	2
	NEW		LAM9	1.00	1
	1530	LAM9			1
	42	LAM9			5
	61	LAM10-CAM			3
	354	U			1
	53	T1			3
	1196	U			1

BMC Microbiology 2006, **6**:76 http://www.biomedcentral.com/1471-2180/6/76

Table 2: The CAS1-DAR variants. Four previously unreported variants of the CAS1 subfamily. The variants are collectively named CAS1-dar in this study.

Spoligopattern	Octal code	No of isolates
■_■_□_□_■■_□_■■■■_□_■■■■_□_□_□_□_□_□_□_□_□_□_□_□_□_■■_□_□_□_□_■	503367400001401	2
■_■_□_□_■■_□_■■■■_□_■■■■_□_□_□_□_□_□_□_□_□_□_□_□_□_■■_□_□_■■■■	503367400001471	2
■_■_□_□_■■_□_■■■■_□_■■■■_□_□_□_□_□_□_□_□_□_□_□_□_□_■■_□_□_■■■■■■■■	503367400001771	3
■_■_□_□_■■_□_■■■■■■■■■_□_□_□_□_□_□_□_□_□_□_□_□_□_□_■■_□_□_■■■■■■■■	503377400001771	1

When live cultures are not available, two PCR based methods are preferred in order to determine the degree of clustering among *M. tuberculosis*. Such complementary studies will be undertaken for the current population but are not included in the current paper.

Spoligotyping is not necessarily the best method for phylogenetic studies, since it targets a small region of the genome. The knowledge of the evolution of this region is limited. It has however been proposed that transposition of insertion sequences can lead to convergence of spoli-

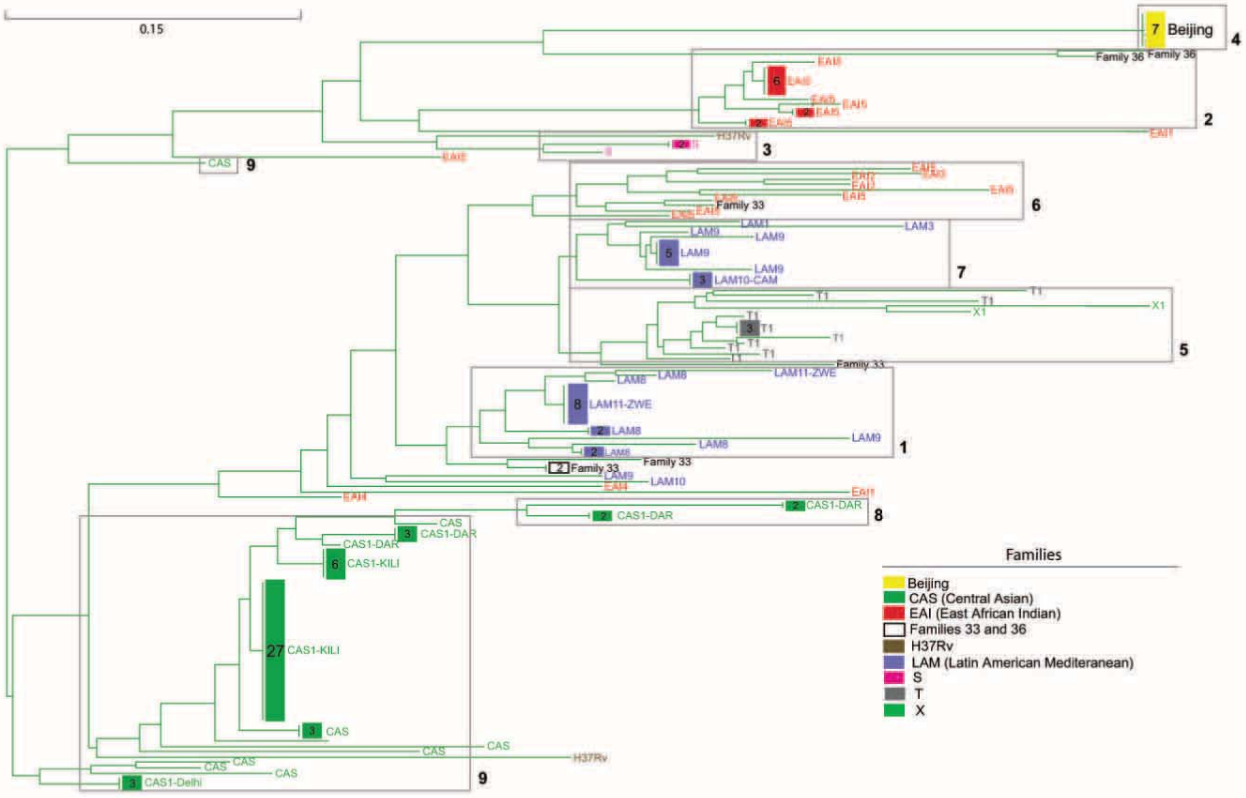


Figure 1
Neighbor-joining tree of the 147 isolates of *M. tuberculosis*. Neighbor-joining tree of the 147 isolates of *M. tuberculosis*. The isolates are colour-coded according to family assignment. The nine groups identified by Structure are identified by grey open boxes. One CAS isolate (*) assigned to the large CAS group is shown in a separate box. Only isolates showing > 65% membership in a group are included in the boxes. For convenience, the NJ tree is rooted by mid-point rooting.

gopatterns and that the evolution of the region is unidirectional (spacers can be lost but not gained). Also, contiguous blocks of spacers and DRs can be lost in single events [14]. These facts may obscure phylogenetic analyses using simple distance based methods. Despite these weaknesses, spoligotypes have been shown to correlate quite well with single nucleotide polymorphisms (SNP), with the T family, constituting only 10 isolates in this study, as a notable exception. For these reasons a NJ-tree was used to illustrate the current results.

The success of the CAS family in particular, but also the LAM and EAI families in this community is intriguing. The low diversity of the highly prevalent CAS family in this study may indicate that the family is spreading rapidly, but could also reflect a slower evolution of the DR region which could possibly be a result of the missing spacers in the central part of the spoligopatterns of these strains.

The success of these three families suggests a possible co-evolution between specific *M. tuberculosis* families and host population, the molecular basis of which remains to be elucidated. A study conducted in San Francisco supports the idea of co-evolution between this pathogen and host populations [15]. In order to document such possible co-evolution, large populations should be preferred. Internationally standardized methods such as spoligotyping and MIRU-typing, as well as SNP and deligotyping, enable comparison of *M. tuberculosis* genotypes between studies conducted at different times and locations. This facilitates inter-study comparison and helps generate large populations for such evolutionary scenarios. It should be noted that the current study represents a short time period and a small collection of strains. This complicates interpretation of recent transmission and hampers comparisons of genetic diversity with that found in studies conducted over a longer period of time. The use of different genotyping methods also makes direct comparison with previous studies in Tanzania [5-7] difficult.

Recent findings suggest that the tubercle bacillus emerged in Africa and may have spread globally in parallel with the human migrations out of Africa [15,16]. Another study have however identified India as the center for the evolutionary radiation of *M. tuberculosis* [17]. These theories are not mutually exclusive; as the spread to India might represent an early and evolutionary important step in the radiation of *M. tuberculosis* out of Africa. The CAS- and EAI-families which this study found to be abundant in Dar es Salaam, have previously been identified to have the most ancestral roots [17]. We demonstrate that the Beijing family, which is highly prevalent in many Asian locations, is not common in the current population. It therefore appears unlikely that import of strains from Asia have had a major impact on the *M. tuberculosis* population in Dar es Salaam. The sensitivity of spoligotyping alone is insufficient for pinpointing evolutionary origins and direction of movement, but the current findings lend support to a view of an early African origin of *M. tuberculosis*.

Spoligotyping is inexpensive, fast, simple and reliable. By using this method one can identify outbreaks, support community-based contact tracing, describe the diversity of a *M. tuberculosis* population, and compare this population to that in other parts of the world. Implementation of spoligotyping as a routine method for molecular epidemiological studies of *M. tuberculosis* isolates, appear to represent a valuable investment in many high-incidence countries.

Conclusion

Spoligotyping is very useful to gain an overall understanding of the local TB epidemic. This study demonstrated that the extensive TB epidemic in Dar es Salaam, Tanzania was caused by a few successful *M. tuberculosis* families, dominated by the CAS family. Import of new strains was a minor problem.

Methods

DNA extraction and spoligotyping

Isolates of *M. tuberculosis* were collected from sputum smear positive TB cases in consecutive patients in Dar es Salaam during October and November 2005. Heat-killed samples were shipped to Norway, DNA was extracted [18] and a total of 147 *M. tuberculosis* isolates were spoligotyped according to Kamerbeek et al. [4].

Family assignment

The obtained spoligopatterns were first compared to the SpolDB4 database [9] and assigned to families and subfamilies. Second, in order to assign names to the isolates not found in the SpolDB4 database, the spoligopatterns were analyzed with 'Spotclust'[10], using a mixture model built on the SpolDB3 database. This model takes into account knowledge of the evolution of the DR region and assigns spoligopatterns to families and subfamilies.

Phylogenetic analyses

A NJ-tree [19] was constructed by converting the presence or absence of 43 defined spacers of the 147 isolates into a Jaccard [20] based pair-wise distance matrix with the computer program 'NTSYSpc' (Exeter Software Co., New York). Without conversion to distance, to verify the NJ tree, the spacer data were directly used by the program 'Structure' [21] to identify groups into which the individual isolates fit best and to calculate the best number of groups explaining the whole data set (run with a no-admixture-model, and a burn-in of 100000 repeats and 400000 Markov Chain Monte Carlo repeats, 65%

assigned membership to a group was used as a threshold value in figure 1).

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

VE carried out the DNA extraction, genotyping, data analyses and participated in the design of the study. MM conceived the study and collected, cultured and identified the bacterial isolates. SGMM participated in the design of the study and collected, cultured and identified the bacterial isolates. MH participated in the data analyses and in the design of the study. URD conceived the study, supervised the DNA extraction, genotyping and data analyses. All authors contributed in the writing of the article, read and approved the final manuscript.

Acknowledgements

We acknowledge Jørn Henrik Sønstebø for valuable help with the data analyses and the contributors to the SpolDB4 database and 'Spotclust'. This study is in part financed by the project "TB in the 21st century – an emerging pandemic" which is headed by Gunnar Bjune and Carol Holm-Hansen and funded by the Research Council of Norway. All participants of this consortium are acknowledged for valuable discussions.

References

1. WHO: **Global tuberculosis control: surveillance, planning, financing. WHO report 2005. WHO/HTM/TB/2005.349.** Geneva , World Health Organization; 2005.
2. **United Republic of Tanzania Ministry of Health: National Tuberculosis and Leprosy Programme. Annual Report.** Dar es Salaam ; 2003.
3. Mookherji S WDESWHBA: **Motivating and Enabling Improved Tuberculosis Case Detection in Tanzania: Summary Report.** 2004.
4. Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S, Bunschoten A, Molhuizen H, Shaw R, Goyal M, van Embden J: **Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology.** *J Clin Microbiol* 1997, **35(4):**907-914.
5. Yang ZH, Mtoni I, Chonde M, Mwasekaga M, Fuursted K, Askgard DS, Bennedsen J, de Haas PE, van Soolingen D, van Embden JD: **DNA fingerprinting and phenotyping of Mycobacterium tuberculosis isolates from human immunodeficiency virus (HIV)-seropositive and HIV- seronegative patients in Tanzania.** *J Clin Microbiol* 1995, **33(5):**1064-1069.
6. McHugh TD, Batt SL, Shorten RJ, Gosling RD, Uiso L, Gillespie SH: **Mycobacterium tuberculosis lineage: A naming of the parts.** *Tuberculosis* 2005, **85(3):**127-136.
7. Gillespie SH, Kennedy N, Ngowi FI, Fomukong NG, Al-Maamary S, Dale JW: **Restriction fragment length polymorphism analysis of Mycobacterium tuberculosis isolated from patients with pulmonary tuberculosis in northern Tanzania.** *Transactions of the Royal Society of Tropical Medicine and Hygiene* 1995, **89(3):**335-338.
8. Kremer K, van Soolingen D, Frothingham R, Haas WH, Hermans PWM, Martin C, Palittapongarnpim P, Plikaytis BB, Riley LW, Yakrus MA, Musser JM, van Embden JDA: **Comparison of Methods Based on Different Molecular Epidemiological Markers for Typing of Mycobacterium tuberculosis Complex Strains: Interlaboratory Study of Discriminatory Power and Reproducibility.** *J Clin Microbiol* 1999, **37(8):**2607-2618.
9. Brudey K, Driscoll J, Rigouts L, Prodinger W, Gori A, Al-Hajoj S, Allix C, Aristimuno L, Arora J, Baumanis V, Binder L, Cafrune P, Cataldi A, Cheong S, Diel R, Ellermeier C, Evans J, Fauville-Dufaux M, Ferdinand S, Garcia de Viedma D, Garzelli C, Gazzola L, Gomes H, Gutierrez MC, Hawkey P, van Helden P, Kadival G, Kreiswirth B, Kremer K, Kubin M: **Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology.** *BMC Microbiology* 2006, **6(1):**23.
10. Vitol I, Driscoll J, Kreiswirth B, Kurepina N, Bennett KP: **Identifying Mycobacterium tuberculosis complex strain families using spoligotypes.** *Infection, Genetics and Evolution* **In Press, Corrected Proof:**.
11. Singh UB, Suresh N, Bhanu NV, Arora J, Pant H, Sinha S, Aggarwal RC, Singh S, Pande JN, Sola C, Rastogi N, Seth P. UB, Suresh N, Bhanu NV, Arora J, Pant H, Sinha S, Aggarwal RC, Singh S, Pande JN, Sola C, Rastogi N, Seth P: **Predominant tuberculosis spoligotypes, Delhi, India.** *Emerging Infectious Diseases* 2004, **10(6):**1138-1142.
12. Easterbrook PJ, Gibson A, Murad S, Lamprecht D, Ives N, Ferguson A, Lowe O, Mason P, Ndudzo A, Taziwa A, Makombe R, Mbengeranwa L, Sola C, Rostogi N, Drobniewski F: **High Rates of Clustering of Strains Causing Tuberculosis in Harare, Zimbabwe: a Molecular Epidemiological Study.** *J Clin Microbiol* 2004, **42(10):**4536-4544.
13. Niobe-Eyangoh SN, Kuaban C, Sorlin P, Thonnon J, Vincent V, Gutierrez MC: **Molecular Characteristics of Strains of the Cameroon Family, the Major Group of Mycobacterium tuberculosis in a Country with a High Prevalence of Tuberculosis.** *J Clin Microbiol* 2004, **42(11):**5029-5035.
14. Warren RM, Streicher EM, Sampson SL, van der Spuy GD, Richardson M, Nguyen D, Behr MA, Victor TC, van Helden PD: **Microevolution of the Direct Repeat Region of Mycobacterium tuberculosis: Implications for Interpretation of Spoligotyping Data.** *J Clin Microbiol* 2002, **40(12):**4457-4465.
15. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, Nicol M, Niemann S, Kremer K, Gutierrez MC, Hilty M, Hopewell PC, Small PM: **Variable host-pathogen compatibility in Mycobacterium tuberculosis.** *PNAS* 2006, **103(8):**2869-2873.
16. Gutierrez MC, Brisse S, Brosch R, Fabre M, Oma, s B, Marmiesse M, Supply P, Vincent V: **Ancient Origin and Gene Mosaicism of the Progenitor of Mycobacterium tuberculosis.** *PLoS Pathogens* 2005, **1(1):**e5.
17. Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbon MH, Bobadilla del Valle M, Fyfe J, Garcia-Garcia L, Rastogi N, Sola C, Zozio T, Guerrero MI, Leon CI, Crabtree J, Angiuoli S, Eisenach KD, Durmaz R, Joloba ML, Rendon A, Sifuentes-Osornio J, Ponce de Leon A, Cave MD, Fleischmann R, Whittam TS, Alland D: **Global Phylogeny of Mycobacterium tuberculosis Based on Single Nucleotide Polymorphism (SNP) Analysis: Insights into Tuberculosis Evolution, Phylogenetic Accuracy of Other DNA Fingerprinting Systems, and Recommendations for a Minimal Standard SNP Set.** *J Bacteriol* 2006, **188(2):**759-772.
18. van Soolingen DHPEWKK: **Restriction fragment length polymorphism (RFLP) typing of mycobacteria.** Bilthoven , National Institute of Public Health and the Environment; 1999.
19. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4(4):**406-425.
20. Jaccard P: **Nouvelles récherches sur la distribution florale.** *Bulletin de la Société de Vaud Sciences Naturelles* 1908, **44:**223-270.
21. Pritchard JK, Stephens M, Donnelly P: **Inference of Population Structure Using Multilocus Genotype Data.** *Genetics* 2000, **155(2):**945-959.

8. APPENDIX

For PCR amplification of MIRU / VNTR loci, the following parameters were tested in different combinations:

PCR program (Normal 59):

15 min		95˚c
25 cycles of:	1 min	94˚c
	1 min	59˚c
	1 min 30 sec	72˚c
30 min		72˚c

PCR program (Normal 54):

15 min		95˚c
25 cycles of:	1 min	94˚c
	1 min	54˚c
	1 min 30 sec	72˚c
30 min		72˚c

-Ordinary PCR program with annealing temperature 59˚C
-Ordinary PCR program with annealing temperature 57˚C
-Ordinary PCR program with annealing temperature 55˚C and
-Ordinary PCR program with annealing temperature 54˚C

Touchdown PCR Program 1

	15 min	95˚C
2 cycles	45 sec	94˚C
	60 sec	59˚C
	90 sec	72˚C
2 cycles	45 sec	94˚C
	60 sec	58˚C
	90 sec	72˚C
2 cycles	45 sec	94˚C
	60 sec	57˚C
	90 sec	72˚C
4 cycles	45 sec	94˚C
	60 sec	56˚C
	90 sec	72˚C
17 cycles	45 sec	84˚C
	60 sec	56˚C
	90 sec	72˚C
	7 min	72˚C

Touchdown PCR Program 2

	15 min	95˚C
2 cycles	45 sec	94˚C
	60 sec	59˚C
	90 sec	72˚C
2 cycles	45 sec	94˚C
	60 sec	57˚C
	90 sec	72˚C
2 cycles	45 sec	94˚C
	60 sec	55˚C
	90 sec	72˚C
25cycles	45 sec	84˚C
	60 sec	54˚C
	90 sec	72˚C
	7 min	72˚C

Final [$MgCl_2$] from 1.5 -4.5 mM
Final [dNTP] from 100 -250 µM of each nucleotide (A, G, C, T)
Mycobacterial [DNA] from 1.0 µg/ml to 31.5 µg/ml

Loci amplified separately
Loci amplified in multiplex reactions in different combinations.
Tempase Hot start polymerase (Bergman) vs Amplitaq polymerase (Applied Biosystems)

Homemade buffer (receipt) vs. Buffer 1 (Bergman)

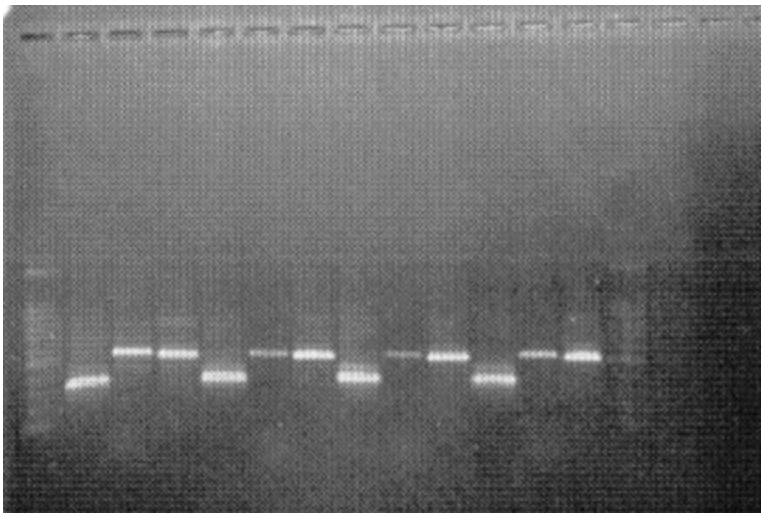
In-house PCR buffer receipt:

1 ml	1 M Tris HCl, ph 8.3
5 ml	1 M KCl
150 µl	1 M $MgCl_2$*
10 µl	10 mg/ml gelatine (heated to 60º)
3.84 ml	sterile H_2O

*To make the standard 15mM $MgCl_2$ buffer. In this study, the amount of MgCl2 was 150, 200, 250, 300, 400 or 450 µl. The amount of H2O was changed correspondingly to make a final 10 ml volume of PCR buffer.

Some results:

Locus VNTR 0580 (1), VNTR 2996 (2) and VNTR 0802 (3) with different [$MgCl_2$]

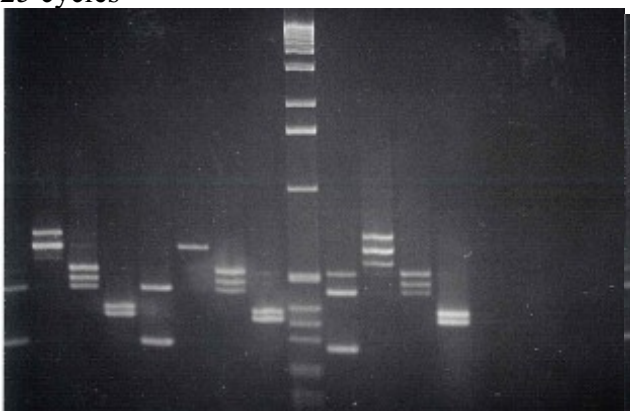


Polymerase Gold

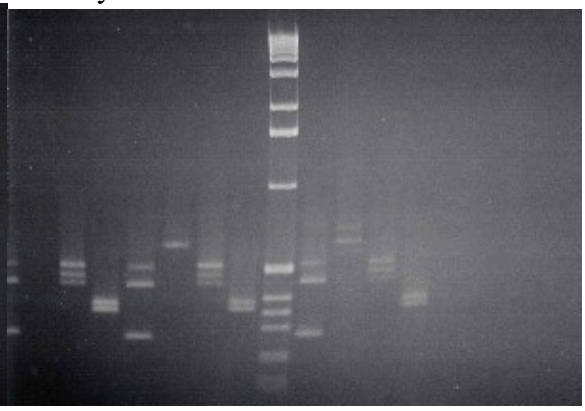
	[$MgCl_2$] 1.5			[$MgCl_2$] 2.0			[$MgCl_2$] 2.5			[$MgCl_2$] 3.0			
L	1	2	3	1	2	3	1	2	3	1	2	3	L

Multiplex PCR loci 1-3, 4-6, 7-9 and 10-12. Different number of PCR cycles, different buffers, different [$MgCl_2$]. “Normal PCR program 54” with 25 or 30 cycles

25 cycles



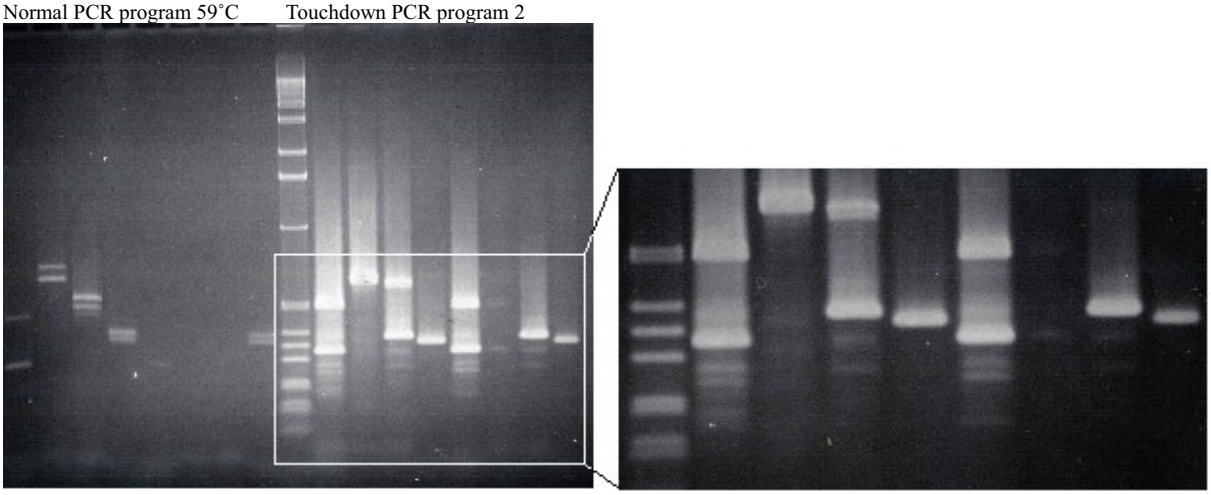
30 cycles



25 thermal cycles																
Home-made PCR buffer									*Buffer 1*							
[$MgCl_2$] 2.0 mM				[$MgCl_2$] 3.0 mM					[$MgCl_2$] 2.0 mM				[$MgCl_2$] 3.0 mM			
1-3	4-6	7-9	10-12	1-3	4-6	7-9	10-12	L	1-3	4-6	7-9	10-12	1-3	4-6	7-9	10-12
30 thermal cycles																
Home-made PCR buffer									*Buffer 1*							
[$MgCl_2$] 2.0 mM				[$MgCl_2$] 3.0 mM					[$MgCl_2$] 2.0 mM				[$MgCl_2$] 3.0 mM			
1-3	4-6	7-9	10-12	1-3	4-6	7-9	10-12	L	1-3	4-6	7-9	10-12	1-3	4-6	7-9	10-12

Touchdown PCR vs normal PCR, different isolates

Normal PCR program 59˚C Touchdown PCR program 2

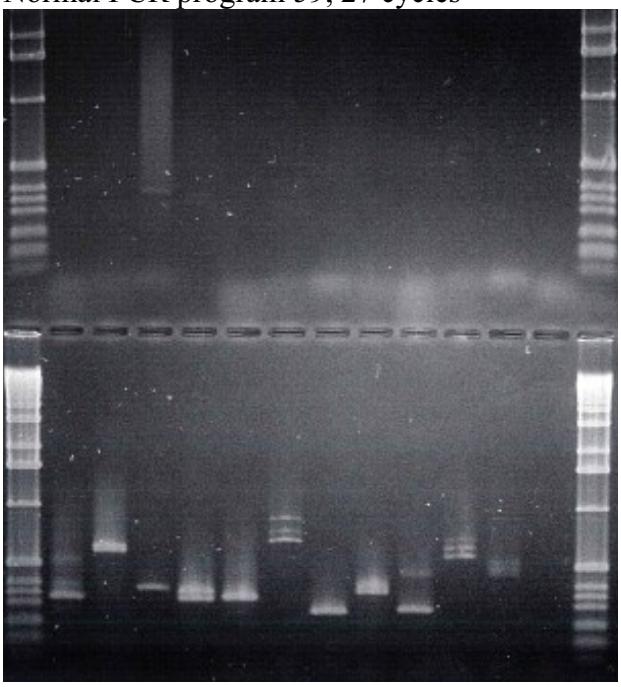


Touchdown PCR																
Home-made PCR buffer[$MgCl_2$] 3.0 mM									*Buffer 1*[$MgCl_2$] 2.0 mM							
BCG P3				MB					BCG P3				MB			
1-3	4-6	7-9	10-12	1-3	4-6	7-9	10-12	L	1-3	4-6	7-9	10-12	1-3	4-6	7-9	10-12
59˚C																
Home-made PCR buffer[$MgCl_2$] 3.0 mM									*Buffer 1*[$MgCl_2$] 2.0 mM							
BCG P3				MB					BCG P3				MB			
1-3	4-6	7-9	10-12	1-3	4-6	7-9	10-12	L	1-3	4-6	7-9	10-12	1-3	4-6	7-9	10-12

Normal PCR, as before, but with 27 cycles.

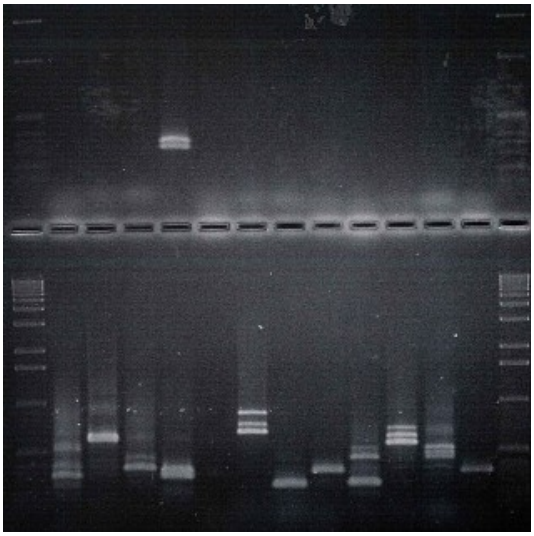
Touchdown PCR, as before but with 2 cycles 59˚c annealing, proceeded by 2 cycles with 58˚c, 2 cycles with 56˚c and 22 cycles with 55˚c

Normal PCR program 59, 27 cycles



2.0 mM [$MgCl_2$] (upper)													
Lane													
1	2	3	4	5	6	7	8	9	10	11	12	13	14
Ladder	MB 265, Loci 1-3, 4-6, 7-9, 10-12				TC 064, Loci 1-3, 4-6, 7-9, 10-12				BCG P3, Loci 1-3, 4-6, 7-9, 10-12				Ladder
3.0 mM [$MgCl_2$] (lower)													
Lane													
Ladder	MB 265, Loci 1-3, 4-6, 7-9, 10-12				TC 064, Loci 1-3, 4-6, 7-9, 10-12				BCG P3, Loci 1-3, 4-6, 7-9, 10-12				Ladder

Touchdown



2.0 mM [$MgCl_2$] (upper)													
Lane													
1	2	3	4	5	6	7	8	9	10	11	12	13	14
Ladder	MB 265, Loci 1-3, 4-6, 7-9, 10-12				TC 064, Loci 1-3, 4-6, 7-9, 10-12				BCG P3, Loci 1-3, 4-6, 7-9, 10-12				Ladder
3.0 mM [$MgCl_2$] (lower)													
Lane													
Ladder	MB 265, Loci 1-3, 4-6, 7-9, 10-12				TC 064, Loci 1-3, 4-6, 7-9, 10-12				BCG P3, Loci 1-3, 4-6, 7-9, 10-12				Ladder

I failed to obtain any stably successful protocol. Results were spurious, sometimes positive, very often failing.

VOLUME 10
NUMBER 11
NOVEMBER 2006
SUPPLEMENT 1

PAGES S1–S306
ISSN 1027 3719

The
International
Journal of Tuberculosis
and Lung Disease

The Official Journal of the International Union Against Tuberculosis and Lung Disease

ABSTRACT BOOK

37th World Conference on Lung Health of the International Union Against Tuberculosis and Lung Disease (The Union)

PARIS · FRANCE
31 OCTOBER–4 NOVEMBER 2006

oners) during a 3 month period in 2005 in Donetsk Oblast, Ukraine.
Results: The prevalence of HIV among TB patients was 15.4% (11.7–19.7, here and elsewhere ranges are 95% confidence limits) and 18.6% (11.0–28.4) in the civilian and penitentiary sectors, respectively. The risk factors associated with HIV infection were intravenous drug use and history of imprisonment.
Discussion: The results of the first ever HIV surveillance among TB patients conducted in Ukraine following international standards are quite astonishing showing a prevalence of HIV in the civilian TB population almost two times higher than the WHO estimate for 2004 (8.3%) and two percentage points higher than the official data reported during the same period by the routine HIV surveillance system (13.3%, 10.8–16.1). These findings call for urgent measures to control the spreading of HIV among TB patients in the country.

PS-61258-02 Community involvement in cluster sampling for a tuberculosis prevalence survey in Western Kenya

L O Odeny,[1] A H van't Hoog,[1,2] J A Agaya,[1] M W Borgdorff,[2] A Hightower.[1] [1]Kenya Medical Research Institute, KEMRI/CDC Program, Kisumu, Kenya; [2]Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands. Fax: (+254) 57 2022981. e-mail: LOdeny@ke.cdc.gov

Background: A tuberculosis prevalence survey in rural Western Kenya. Lack of understanding of sampling concepts may raise suspicion in the community about selection criteria and adversely affect study participation.
Aim: 1) Increase the communities' understanding and acceptance of the cluster selection. 2) Random sampling of clusters to obtain a sample of 20 000 persons over 14 years old.
Method/results: The study area has a population of 134 000 and 217 villages. Using available demographic and GIS maps, the area was divided in 105 clusters with a mean population of 1279 (range 752–1792). A mean of two (range 1–4) contiguous villages formed each cluster. Ten administrative leaders of the area participated in the cluster sample selection. Basic information on sampling principles, aim and design of the survey were presented, as well as maps and tables identifying the composite villages that formed the clusters. The leaders were divided in groups and inserted labels with sequential numbers 1–105 into table tennis balls. The balls were placed in a box and drawn by the leaders to form the sample clusters, which were then identified on a map and table. This process created a randomly ordered list of clusters to be used in the prevalence survey.
Discussion: The leaders were pleased with the transparent process and able to explain the procedure to community members. The sampling design is both logistically and statistically efficient.

PS-61270-02 From symptoms to treatment: delay among smear-positive pulmonary TB patients in Sabah, Malaysia

C Rundi,[1,2] P Mangtani,[1] K Fielding,[1] P Godfrey-Faussett,[1] L C Rodriques.[1] [1]ITD Department, London School of Hygiene and Tropical Medicine, London, UK; [2]Sabah Health Department, Kota Kinabalu, Sabah, Malaysia. Fax: (+44) 02076374314. e-mail: Christina.Rundi@lshtm.ac.uk

Tuberculosis is a major health problem in the state of Sabah, East Malaysia; where the case notification rate exceeds 100 cases/100 000 population. Reducing delays in diagnosis and treatment will limit the duration of infectiousness and thus reduce transmission in the community. A cross-sectional study is being conducted to determine the factors that affects the duration of the time period from the onset of symptoms to the start of treatment. It involves all adult smear positive pulmonary TB patients in a population of about a million over a period from October 2005 to March 2006. The domains of interest include socio-demographic status, difficulty in accessing services including travel time, perception of health services, knowledge, practices, attitudes and TB symptoms. The median age of participants was 36 years (range 18–80 years) and 36% were female. Cough was the most common TB symptom, being present in 93% of respondents. The median delay from onset of symptoms to start of TB treatment was 4 weeks (interquatile ranges: 4 weeks, range 1–156 weeks).

PS-61273-02 Molecular epidemiology of *Mycobacterium tuberculosis* in Dar Es Salaam, Tanzania

V Eldholm,[1,2] M Matee,[3] S G M Mfinanga,[3] M Heun,[2] U Dahle.[1] [1]Norwegian Institute of Public Health, Oslo, [2]Institute of Nature Management, Norwegian University of Life Sciences, Ås, Norway; [3]Muhimbili Medical Research Centre, Dar es Salaam, Tanzania. Fax: (+47) 22353605. e-mail: Ulf.dahle@fhi.no

Tanzania, located in east Africa, has a high tuberculosis burden despite long lasting efforts from tuberculosis control programs. To improve the understanding of the diversity of *Mycobacterium tuberculosis* in Tanzania, isolates from pulmonary cases in Dar es Salaam were assessed using spoligotyping. The strains were isolated in consecutive patients during October and November 2005, and shipped to Norway for PCR amplification and spoligotyping. The preliminary results from this ongoing study include those of 147 strains. Among the analyzed strains, 76 different spoligotypes were found. A total of 89 isolates were part of one of 18 identified clusters (61%). An overall diversity of 52% was inferred from the spoligotypes and the clustered isolates will be analyzed further. A high proportion (37%) of the strains was found to be members of, or closely related to the Central Asia (CAS) lineage. The Latin American Mediterranean (LAM) family was also found to be widespread (22%), and the East African Indian (EAI) lineage constituted 10%. Low levels of other common lineages such as Haarlem, X, T and Beijing were found. The epidemic appears to be an established one, with little influence of newly-imported strains.

Acknowledgements: * TB in the 21st century consortium is financed by the Norwegian Research Council. It is headed by G Bjune and C Holm-Hansen. Parts of this work package were initiated by M Nyindu, L Uiso and others. Their efforts are greatly appreciated.

PS-61321-02 Tuberculosis deaths in a rural area in Bangladesh, 1988–2003

K Zaman,[1] M D Yunus,[1] S E Arifeen,[1] A H Baqui,[2] S Hossain,[1] M A Islam,[3] V Begum,[4] M N Alam,[1] P K Streatfield,[1] S Luby,[1] A Bhuiya,[1] D A Sack.[1] [1]ICDDR,B: Centre for Health and Population Research, Dhaka, Bangladesh; [2]Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA; [3]Bangladesh Rural Advancement Committee (BRAC), Dhaka, [4]National Tuberculosis Control Programme (NTP), Dhaka, Bangladesh. Fax: (+88) 2 8826050. e-mail: kzaman@icddrb.org

Introduction: Bangladesh ranks fifth among high burden tuberculosis (TB) countries. However, data on TB mortality from Bangladesh is sparse.
Objectives: To present data on TB mortality from Matlab, a rural area of Bangladesh during the period 1988–2003.
Methods: Data were obtained from Matlab, Bangladesh where the ICDDR,B has been maintaining a Health and Demographic Surveillance System (HDSS) in a population of about 200 000 since 1966. DOTS services were introduced by National Tuberculosis Control Programme in 1998 and BRAC health volunteers were involved in social mobilization and treatment supervision since 2001. As part of HDSS, all deaths were recorded and causes of deaths were ascertained using verbal autopsy methodology and causes assigned according to WHO ICD codes. All death records were reviewed retrospectively for the period 1988–2003 to determine deaths due to TB.
Results: During the period, there were 639 TB deaths; 473 (74%) among males and 166 (26%) among females. Most of the deaths (75.6%) were among persons aged ⩾ 45 years followed by 21.6% among 15–44 years and 2.8% in <15 years old. TB deaths comprised of 3.6% of all deaths among persons ⩾15 years. The age-standardized TB mortality rates varied between years and ranged from 19.15 to 46.05/100 000 population among males and 2.19 to 23.72 among females.
Conclusions: The high burden of tuberculosis deaths among adults in rural Bangladesh warrants further strengthening of TB control programme.

PS-61350-02 Tuberculosis as primary cause of death among AIDS cases in Rio de Janeiro City, Brazil

V Saraceni,[1] B S King,[2] J E Golub,[2] L M Lauria,[1] S C Cavalcante,[1] B Durovni.[1] [1]Rio de Janeiro City Health Secretariat, Rio de Janeiro, RJ, Brazil; [2]Johns Hopkins Center for Tuberculosis Research, Baltimore, Maryland, USA. Fax: (+55) 21 22747542. e-mail: vsaraceni@rio.rj.gov.br

Background: AIDS-related mortality decreased in Rio de Janeiro City (RJC) since the introduction of HAART (1997), and has been stable since 2000. We investigated the role of Tuberculosis (TB) as the primary cause of death among HIV+ subjects in RJC, in order to assess the magnitude of the co-infection.
Methods: Review of Mortality Information System, according to ICD-10, with death certificates coded in Chapter I—B20 to B24. Data of ARV use came from ARV drug database.
Results: Between 1996 and 2004, TB was the primary cause of death in 9.1% of AIDS related deaths, while PCP accounted for 4.6%. 'Other' infectious diseases accounted for 53.7% (B20, excluded B20.0), although it is not clear by the death certificate if TB cases were misclassified as other entities 32.1% of TB related deaths were on HAART (median: 18 months) vs. 25.2% of those who died from PCP (median: 20 months). Patients that died from TB were less educated (58.2% vs. 53.0%; $P = 0.04$) and more likely to be non-white (56.7% vs. 50.6%; $P = 0.07$) than patients with other causes.
Conclusions: Analysis of primary cause of death among AIDS cases using a secondary database showed that TB is the leading cause of AIDS related death and is responsible for twice as many deaths as PCP. The burden of TB may be an underestimate due to potential misclassification among causes of death labeled 'other'.

PS-61352-02 A controlled trial of door-to-door tuberculosis active case finding in a Brazilian favela

A C Miller,[1] E C Soares,[2] J E Golub,[1] B Durovni,[2] Z Fonseca,[2] D B Arduino,[2] L H Moulton,[1] R E Chaisson,[1] S C Cavalcante.[2,3] [1]Johns Hopkins University Center for Tuberculosis Research and Bloomberg School of Public Health, Baltimore, Maryland, USA; [2]Health Department of Rio de Janeiro City, Rio de Janeiro, RJ, [3]IPEC/FIOCRUZ, Rio de Janeiro, RJ, Brazil. Fax: (+1) 443 287 7955. e-mail: amiller@jhsph.edu

Design: A pair matched, cluster randomized trial of a door to door symptom screen and spot sputum collection vs. home delivery of an informational pamphlet in a large Brazilian favela (slum).
Aim: To compare TB case notification rates in the two study arms.
Methods: 14 administrative zones in the favela Rocinha were pair-matched based on estimated TB case notification rates in 2003. One zone of each pair was randomly allocated to receive a door to door symptom screen and sputum collection from all symp-