Tuberculosis is an ancient disease. It is completely curable provided the anti-tuberculosis treatment is faithfully followed as per physicians recommendations. A steady decline was observed in the incidence of tuberculosis in the western world, especially in USA, from 1954 till 1984. In fact public health officials believed that they have eradicated this deadly disease from the western world. To the surprise of many in around 1985, TB cases began to rise again and peaked around early 1990’s. Decline in TB control activities and the human immunodeficiency virus/ acquired immunodeficiency syndrome (HIV/AIDS) epidemic were suggested as two major factors for the reemergence of TB. The situation was critical as large number of people died due to complications of tuberculosis and AIDS. Therefore, world health organization (WHO) declared this disease as a global emergency in 1993.

Subsequent to this declaration renewed efforts were made to study this organism by scientists worldwide and a large body of data was generated covering different aspects of this disease. Keeping in view the vast literature that exists on TB, it is very difficult to compile a concise review covering diverse aspects associated with this disease. A deliberate effort has been made, therefore, to restrict the scope of the review to relevant studies that pertains to the subjects of this thesis.

2.1 Etiology of the Tuberculosis

Tuberculosis is caused by Mycobacterium tuberculosis, a rod shape bacilli that is 2-4 micrometers in length and 0.2-0.5 μM in width. Due to its small size, they get entrapped inside the air droplets (2-10 μM in diameter) formed during coughing and sneezing by tuberculosis patient. These aerosols loaded with bacilli, remain in air for a long time and cause infection when inhaled by a healthy person staying in the proximity of the patient (Eames et al, 2009; Chen et al, 2010). Mostly, the bacteria remain inactive inside the host, till they find suitable conditions for their growth, like malnutrition and/or immunosuppression.
Based on the site of infection, tuberculosis is categorized as pulmonary or extra-pulmonary. Pulmonary tuberculosis is the most common form of tuberculosis and primarily involves lungs. The common symptoms include chest pain, coughing up blood and a productive, prolonged cough for more than three weeks. Whereas, in extra-pulmonary infection; the symptoms are much more variable and are often dictated by the site of infection (Sharma and Mohan, 2004). Some of the common sites include, the pleura in tuberculosis pleurisy, central nervous system in meningitis, lymphatic system in scrofula of the neck, genitourinary system in urogenital tuberculosis, and bones and joints in Pott's disease of the spine. Extra pulmonary TB occurs more commonly in immunosuppressed persons and young children (Sharma and Mohan, 2004).

2.2 Incidence and Prevalence of Tuberculosis

2.2.1 Global Scenario: Incidence of the disease is still increasing as a result of increase in world population. In 2007, world health organization (WHO) recorded around 9.27 million new TB cases. Out of this, nearly 55% of cases were reported from Asia, whereas 31% new TB cases were recorded from Africa. Prevalence and mortality due to this dreaded disease decreased in the year 2007. This decline is very slow i.e.1% per year. Around 9.4 million incident TB cases (137 per 100,000 population) were documented globally in the recent report by WHO. The five high TB burden countries are India (1.6-2.4 million), China (1.1-1.5 million), South Africa (0.40-0.59 million), Nigeria (0.37-0.55 million) and Indonesia (0.35-0.52 million) (WHO, 2010). The destructive effect of the disease further increased with the outbreak of multi drug resistance tuberculosis (MDR-TB) and extensive drug resistance tuberculosis (XDR-TB). Nearly 0.5 million cases of MDR-TB were recorded in 2010 and XDR-TB was reported from 55 countries (WHO, 2010).

2.2.2 Indian Scenario: India accounts for one fifth of the global incidence of TB. Nearly 40% of adult Indian population is infected with the TB bacillus. According to Revised National Tuberculosis Control Programme status report, TB India 2009; every
day about 5000 people develop the infection and around 1000 die (2 deaths every 3 minutes) in India due to this cruel disease.

2.3 **Members of Mycobacterium tuberculosis complex**
Recent reports reveal that the causative organisms of tuberculosis developed in east Africa around 40,000 years ago from a group of ancestral tubercle bacilli collectively known as *Mycobacterium paratuberculosis* (Wirth *et al*, 2008). Initially these bacilli caused infection in humans but around 10,000 to 20,000 years later it divided into another lineage causing disease among animals (Gillespie, 2007). These closely related causative organisms of the tuberculosis were collectively known as *Mycobacterium tuberculosis* complex. It comprises *Mycobacterium tuberculosis*, *Mycobacterium bovis*, *Mycobacterium africanum*, *Mycobacterium canettii*, *M. microti*, *Mycobacterium caprae* and *Mycobacterium pinnipedi*. Members of this complex display different phenotypic characteristics and a wide host range.

2.3.1 *Mycobacterium tuberculosis*
*Mycobacterium tuberculosis* is the major cause of human TB worldwide. It was first described by Robert Koch, therefore it is also known as Koch’s bacillus (Taylor *et al*, 2003). Previously it was speculated that *M. tuberculosis* evolved from *M. bovis* by specifically adapting itself to human instead of animal host (Cockburn *et al*, 1963). Mycobacterial genome sequencing studies have confirmed the presence of certain sequences in *M. tuberculosis* but found to be absent in *M. bovis* genome (Brosch *et al*, 2002). Downsizing of genome due to successive loss of DNA in *M. bovis* proved that it is evolutionary descendent of *M. tuberculosis*.

2.3.2 *Mycobacterium bovis*:
The main host of *M. bovis* is cattle (*Bos taurus*) and was the major cause of zoonotic TB i.e. TB transmitted from animals to humans. Before the generalized adoption of pasteurization of milk and other dairy products, *M. bovis* was the main cause of human
TB, especially intestinal TB in case of children. After introduction of milk pasteurization, the death rate of children under five year of age was declined significantly (Thoen et al, 2006). The disease caused by \textit{M. bovis} is clinically indistinguishable from the one caused by \textit{M. tuberculosis}.

2.3.3 \textit{Mycobacterium africanum}
As the name suggests, \textit{M. africanum} was isolated in Africa and was assumed to be the leading cause of pulmonary TB in the continent (Haas et al, 1997; Frothingham et al, 1999). A recent report revealed the sporadic isolation of this bacillus from Europe and United States (Desmond et al, 2004).

\textit{M. africanum} is mainly categorized into two major subtypes based on their biochemical characteristics and geographic origin. \textit{M. africanum} subtype I and II differs from each other based on their origin from western and eastern Africa respectively. Subtype II is the main cause of human TB in Kampala, Uganda and was resistant to thiophen-2-carboxylic acid hydrazide (TCH) (Niemann et al, 2002). Based on the biochemical characteristics, \textit{M. africanum} Subtype I is closely related to \textit{M. bovis} whereas subtype II resembles \textit{M. tuberculosis} (Niemann et al, 2002, Sola et al, 2003).

2.3.4 \textit{Mycobacterium microti}
\textit{M. microti} is a member of the \textit{M. tuberculosis} complex and was first isolated in 1937 as the causative agent of pulmonary TB in the wild vole (\textit{Microtus agrestis}) (Wells, 1937). It was considered to be avirulent for humans, cattle and laboratory animals and was therefore, proposed as a live vaccine against TB.

2.3.5 \textit{Mycobacterium caprae}
The names proposed for \textit{M. caprae} are \textit{M. tuberculosis} subspecies \textit{caprae} (Aranaz et al, 1999) and \textit{M. bovis} subspecies \textit{caprae} (Niemann et al, 2002a). This species was originally described as preferring goats to cattle as hosts (Gutierrez et al, 1995, Aranaz et al, 1996) and has been found in Spain, Austria (Prodinger et al, 2002), France
Genomic analysis of *M. tuberculosis*: A brief review

(Haddad et al, 2001), Germany (Erler *et al*, 2003), Hungary, Italy, Slovenia (Erler *et al*, 2004), and the Czech Republic (Pavlik *et al*, 2002).

2.3.6 *Mycobacterium pinnipedii*

*M. pinnipedii* was first isolated from captive and wild sea lions and fur seals from Australia (Cousins *et al*, 1999) and New Zealand (Cousins *et al*, 2003). Similar organisms were subsequently recovered from the same mammal species in South America (Romano *et al*, 1995; Bernardelli *et al*, 1996; Bastida *et al*, 1999) as well as from a Brazilian tapir (Cousins *et al*, 2003).

2.4 Conventional Strain Typing

2.4.1 Phage typing

Phage typing was the conventional method available for typing clinical isolates of *M. tuberculosis*. In 1971 and 1973, ‘World Health Organisation International Working Group’ carried out the testing of a panel of phages in five centres. Only three showed reproducibility, however the discriminatory power was very low.

Jones and Greenburg (1978) further improved the method and introduced four phages which divided the strains into three main groups. These groups can be further subdivided by six auxiliary phages (Jones *et al*, 1982). The reported discrimination and reproducibility of this scheme was good. However, this method has some inbuilt limitations as it was cumbersome and required skilled, technical effort and careful interpretation of the results. These considerations limited its wider use.

2.4.2 Drug susceptibility patterns

Initially, the drug susceptibility patterns were also used to differentiate between strains. This method was more robust and reproducible. However, in regions with high burden of tuberculosis, majority of the *M. tuberculosis* isolates are fully susceptible to drugs and one of the few phage types distinguished among *M. tuberculosis* isolates
predominate in most areas. Moreover, the drug susceptibility profiles of \textit{M. tuberculosis} are highly unstable features, as strains frequently become resistance to anti tuberculosis drugs during treatment (Perri \textit{et al}, 2011).

2.4.3 Multilocus enzyme electrophoresis

Multilocus enzyme electrophoresis for differentiation of mycobacteria was reported in 1991 by Wasem and colleagues. This was based on difference in the electrophoretic mobility of different enzymes among different strains. This method was presumed to be very sensitive as variation in enzyme electrophoretic mobility can detect at least 50\% of amino acid substitutions in proteins. The compilation of mobility patterns of several enzymes can be combined to quantitate the overall relatedness of strains. This method could not be widely used as it was both time-consuming and technically demanding (Saunders, 1995).

2.5 Organization of \textit{M. tuberculosis} genome

A major milestone in TB research was reached with the completion of whole genome sequencing of \textit{M. tuberculosis}, H37Rv (Cole \textit{et al}, 1998). The H37Rv genome was sequenced at the Sanger Centre in Cambridge, in collaboration with the Institute Pasteur of Paris. The information gathered by whole genome sequencing serve as a source of vast knowledge that can be much needed breakthrough in elucidating unique characteristics of the biology of the tubercle bacillus, such as slow growth, nature of its complex cell wall, genes related to virulence and stability of its genome.

A significant finding of the project was the identification of PE and PPE family of proteins comprising of 100 and 69 members, respectively (Cole \textit{et al}, 1998). The genes for these proteins are a source of diversity in \textit{M. tuberculosis}.
Fig 2.1: The outer circle shows the scale in Mb, with 0 representing the origin of replication. The first ring from the exterior denotes the positions of stable RNA genes (tRNAs are blue, others are pink) and the direct repeat region (pink cube); the second ring inwards shows the coding sequence by strand (clockwise, dark green; anticlockwise, light green); the third ring depicts repetitive DNA (insertion sequences, orange; 13E12 REP family, dark pink; prophage, blue); the fourth ring shows the positions of the PPE family members (green); the fifth ring shows the PE family members (purple, excluding PGRS); and the sixth ring shows the positions of the PGRS sequences (dark red). The histogram (centre) represents G + C content, with <65% G + C in yellow, and >65% G + C in red. The figure was generated with software from DNASTAR. Cole et al (1998); Nature 393: 537-544
The salient characteristics of M. tuberculosis genome are

- It is 4,411,529 bp long sequence and second largest microbial genome sequenced after the sequencing of Haemophilus influenza (Fleishmann et al, 1995).

- *M. tuberculosis* genome possess characteristically high G + C content (65%) and codes for 3,924 ORFs accounting for 91% of the coding capacity of the genome (Cole et al, 1998).

- Cole et al (1999) found bias in overall orientation of genes with respect to direction of replication. Around 59% of *M. tuberculosis* genes are in the same orientation towards replication fork in comparison to *B. subtilis* where 75% of genes are in same orientation as that of replication fork. This finding is attributed to slow growing phenotype of tubercle bacillus.

- Two prophage sequences i.e. PhiRv1 and PhiRv2 were also detected over H37Rv genome.
  - PhiRv1 is present in repetitive sequence family 113E12. It has seven different potential *att* sites over *M. tuberculosis* genome for insertion and shows polymorphism among different clinical isolates (Cole et al, 1999).
  - The second prophage, PhiRv2 has proven to be much more stable and show less variability among different strains (Cole et al, 1999).

- *M. tuberculosis* genome possesses 56 copies of insertion sequences (IS) belonging to the well defined IS3, IS5, IS21, IS30, IS110, IS256, IS1535 and IS23 family (Gordon et al, 1999a). These insertion sequences range in size from 200 to 1500 or more (Timmerman and Tu, 1985). Out of these, IS 6110 of the IS3 family is most widely used for strain typing and molecular epidemiology studies due to variation in its copy number and insertion sites (van Embden et al, 1993).
  - Sixteen copies of IS6110 and six copies of IS 1081 were identified in H37Rv genome.
  - Seven insertion sequences belonged to 13E12 family of repetitive sequences.

- Whole genome sequencing of *M. tuberculosis* revealed the presence of short sequence repeats, termed mycobacterial interspersed repeat units (MIRUs) (Supply et
al., 2000) and variable number tandem repeat units (VNTRs), which are tandemly repeated sequences of 40–100 bp.

2.6 Strategies adopted by the organisms to generate genetic variations:
*M. tuberculosis* is a restricted human pathogen. The bacillus, for its successful survival, adapts itself to a variety of intra-human environments. These adaptations may reflect in the form of DNA sequence variation, which gradually accumulated and contributed in the tremendous heterogeneity to *M. tuberculosis* genome. The bacterium may adopt different strategies to generate these genetic variations and it can be broadly categorized as:

- Acquisition of DNA sequences from other organisms i.e. horizontal transfer.
- Small local changes in nucleotide sequence of the genome
- Intra-genomic rearrangements of segments of genomic sequence.

2.6.1 Acquisition of DNA from other organisms (Horizontal transfer):
It is probably the most common way in which bacteria gain the ability to produce novel proteins. Horizontal transfer is common among bacteria, and is thought to be a widespread evolutionary process (Lorenz, 1994). However, Fillol et al. (2006) reported that the genetic variations among *M. tuberculosis*, occur very rarely via horizontal exchange.

2.6.2 Small local changes in nucleotide sequence in *M. tuberculosis*
Single nucleotide polymorphisms (SNPs) basically are of two main types i.e. synonymous that results in silent mutation and non synonymous results in change in the DNA code. These SNPs mainly arise either due to transition (Purine to Purine and Pyrimidine to Pyrimidine) or transversion (Purine to Pyrimidine). Usually, the synonymous SNPs are less subject to selective pressure than other genetic markers therefore selected in different evolutionary studies (Gutacker et al., 2002; Baker et al., 2004). However the genome comparison studies revealed that the ratio of synonymous to non-synonymous substitutions in *M. tuberculosis* was less compared with the *Escherichia coli* or Salmonella. The reason could be either due to additional selective pressure on synonymous substitutions or decreased selective pressure on non
synonymous substitutions (Fleischmann et al., 2002). Sreevatsan et al. (1997) used non-synonymous SNPs to categorize 842 clinical isolates of M. tuberculosis into three major genetic groups (MGGs). They further suggested that the major genetic groups of M. tuberculosis complex (MTBC) strains were an outcome of evolutionary bottleneck that existed around 15000 to 20,000 years ago. Recently, Filliol et al. (2006) supported these MGGs and split them further, thus confirming that the M. tuberculosis strains consist of several very distinct families.

2.6.3 Intra genomic rearrangements:
Intra-genomic rearrangements predominantly occur by the recombination process in the prokaryotes and eukaryotes. Genomic rearrangements may result in duplication or deletion of a particular genomic sequence. Duplication of DNA segments may serve as a substrate for evolution (Yamanaka et al., 1998); whereas, deletions mainly occur in conjunction with natural selection to remove non-essential sequences from the genome. Thus, intra genomic rearrangements are the major source of genetic variations among M. tuberculosis strains.

2.7 Rationale for studying genetic variation in M. tuberculosis:
Genetic variation analysis is used to track M. tuberculosis strains, irrespective of its geographical location. These studies provide valuable insight regarding the patterns of spread, infectivity and pathogenicity of the organism. Genetic markers used in these studies have different molecular clock. These markers provide information regarding the change in the genome over a shorter or longer period of time. Genetic marker with a fast molecular clock helps in determining the reactivation of an old infection, whereas a marker with slow molecular clock is used to monitor the evolution of organism over tens of thousands of years (Arnold, 2007).

2.8 Molecular Marker to study genetic variation of M. tuberculosis:
2.8.1 Transposable elements
Transposable, or mobile, elements are frequently found in the genomes of plants, animals and bacteria. These are also known as insertion sequences (IS) due to its ability to integrate into the genome at a new site by a ‘cut and paste’ mechanism (Kazazian,
Bacterial insertion sequences also known as IS elements, are small transposable fragments that do not contain any phenotypic marker. Apparently, they carry only the genetic information necessary for transposition. Insertion elements range in size from about 200 bp to 1500 bp or more, with short invert repeats at the ends. After insertion into host sequence they may also generate short direct repeat varying from 2-20 bp (Timmerman and Tu, 1985).

Different insertion sequences have been found among mycobacterial species. IS 900 was shown to be specific to Mycobacterium paratuberculosis (Green et al, 1989). In 1989, Zainuddin and Dale confirmed the presence of insertion element in M. tuberculosis and used it to differentiate from M. bovis. Collins and Stephens (1991) published the specificity of insertion sequence IS 1081 and IS 6120 with Mycobacterium bovis and Mycobacterium smegmatis, respectively (Guilhot et al, 1992).

Similarly, more than 50 loci in M. tuberculosis (H37Rv strain) genome were found to contain IS elements and constituted about 1.6% of the entire genome (Eisenach et al, 1988; Gordon et al, 1999a,).

Thierry et al (1990) reported the isolation of repetitive sequence IS6110 from Mycobacterium tuberculosis cosmid library and observed its sequence similarity with IS3 family elements that exists naturally in E. coli. IS6110 is 1,355 bp long and contains 28-bp, imperfect inverted repeats at its extremities with three mismatches and 3-bp direct repeats that probably result from repetition of the target sequence. This sequence was found to be specific for strains belonging to M. tuberculosis complex. In their study they design an oligonucleotide derived from IS6110 sequence and used that as a primer and probe in polymerase chain reaction studies for the detection and identification of M. tuberculosis bacilli.

Studies were carried out to determine the use of IS6110 in tracking and discriminating different clinical isolates of M. tuberculosis. In this context, Cave et al (1991) reported the presence of 10-12 copies of IS6110 element among various strains of M. tuberculosis complex. However M. bovis strains bear only 1-3 copies whereas other species of mycobacteria does not contain a single copy of this insertion element. With
the use of restriction endonuclease analysis they indicated that the IS6110 sequence has
conserved across the strain and species line. They further projected that this insertion
sequence can be used as a probe in hybridization of \textit{M. tuberculosis} strains so as to
detect restriction fragment length polymorphism reflecting divergence in insertion site
for the various copies of IS6110.

IS6110 is an excellent strain specific marker for studying epidemiology of tuberculosis
(Mazurek \textit{et al}, 1991). They observed none or minimal differences among the
fingerprints obtained from Bam HI digested chromosomal DNA of 14 \textit{M. tuberculosis}
clinical isolates from closely associated individuals. The only noticeable difference was
an additional fragment of IS6110 in the fingerprint pattern of one isolate. However,
marked differences were observed when unrelated isolates were compared. Similarly,
they verified the laboratory acquired infection by comparing the \textit{M. tuberculosis}
reference strain “Erdman” to a clinical isolate from an individual working with that
strain. These isolates had identical DNA fingerprints which were distinct from all other
isolates, thus, verifying laboratory-acquired infection.

Similarly, Otal \textit{et al} (1991) compared the RFLP patterns of strains before and after the
relapse among same and different patient and demonstrated the presence of IS6110 does
not induce in vivo major genomic rearrangements over a 2 to 3 year period and
confirmed its use as a valuable epidemiological marker in tuberculosis. They performed
the IS6110 based RFLP studies using different methodology. They digested 5µg of \textit{M.
tuberculosis} DNA with the Pst-1 enzyme and then probed with amplified IS6110
specific sequence from H37Rv using below mentioned primer sequences.

(i) \texttt{5\textquotesingle-GGG AAT TCG ACG GAC GTC GTG ACC AGA AGT C-3\textquotesingle}, and

(ii) \texttt{5\textquotesingle-GGG AAT TCG TGT ACA AAA TGT GGA CAA GTA-3\textquotesingle}

Both these oligomers contained EcoRI sites. A 1.4-kb amplified fragment was digested
with EcoRI and cloned into the EcoRI site of pUC18 to provide pMT03 plasmid. The
EcoRI fragment was isolated and then used as a probe. A high degree of polymorphism
of IS6110-containing fragments was observed in the \textit{M. tuberculosis} strains isolated
from different patients, whereas identical patterns were observed in strains isolated from the same patient.

Available data on the RFLP patterns of *M. tuberculosis* clinical isolates from different parts of the world established that IS6110 is a good epidemiologic tool to study tuberculosis, but in these studies different methodologies were used to determine IS6110 patterns (Mazurek *et al.*, 1991; Otal *et al.*, 1991). Therefore, to remove technique dependent variations an urgent need was felt to develop a standardised methodology so that results from different laboratories can be compared easily and to help track movement of patients. Such data provided important insights into the global transmission of tuberculosis and identified strains with unique properties, such as high infectivity, high virulence, and/or multidrug resistance (van Embden *et al.*, 1993).

A standardized technique which exploited the variability in both the number and genomic position of IS6110 to generate strain-specific patterns was developed (van Embden *et al.*, 1993). In this method 1µg of *M. tuberculosis* DNA was digested with PvuII enzyme, electrophoresed and then hybridised with a 245 bp IS6110 specific probe from *M. bovis* strain. The probe was amplified using primer sequences (INS-1; 5’-CGTGAAGGCGATCGAGG and INS-2; 5’-GCAGTGAGCGTGGTAGTGA). Figure-2.2 depicts the physical map of the IS6110 sequence as published by van Embden *et al* (1993). It indicated the location of various restriction digestion sites within the 1,355-bp element. They recommended the use of Pvu II, because it cleaved the IS6110 sequence only once, thus generating IS6110 hybridizing fragments of at least 0.90 or 0.46 kb.
Genomic analysis of *M. tuberculosis*: A brief review

Fig 2.2: Physical map of the 1.35kb *M. tuberculosis* insertion element IS6110. The cleavage sites of several restriction enzymes are depicted. PvuII cleaves the element at base pair 461. The closed bars represent the 28 bp inverted repeat bordering IS6110 DNA. The lines to the left and right of IS6110 sequence representing the chromosomal DNA.

They arbitrarily chose a DNA probe to the right of the PvuII site on the physical map so that the number of bands corresponds directly to the copy number present in the genome. They also recommended conditions that would permit comparison of DNA fingerprints of *M. tuberculosis* generated in different laboratories of the world.

**Stability of IS6110 patterns:**

van Soolingen, 1991 established the usefulness of DNA fingerprinting for epidemiology study of TB. They carried out the study on *M. tuberculosis*, *M. africanum* and *M. bovis* isolates using IS 986 locus. IS 986 is presumed as IS 6110 as it differs by a few bases only. The results indicated that *M. tuberculosis* strains from high burden regions in central Africa were generally more related to each other than the isolates from Netherlands, where the transmission rate was low and where the majority of the tuberculosis cases were presumed to be the result of reactivation of previously contracted *M. tuberculosis* infections.

The stability of IS6110 restriction fragment length polymorphism patterns was studied by Cave *et al* (1994). They compare the DNA fingerprints of 6 *M. bovis* isolates from 1 patient and 41 *M. tuberculosis* isolates from 18 patients at different time intervals ranging from 8 months to 4.5 years. They observed that the banding pattern of 17 strains remained identical at various time points. Additionally, the study revealed that
the band profile remained same even when the strain turned resistant to various anti tuberculosis drugs.

The stability of IS6110 fingerprint patterns was analyzed in serial isolates of 49 tuberculosis patients in United States (Yeh et al., 1998). In this study 12 of 49 patients (25%) showed change in IS6110 banding pattern occurred indicating a relatively high degree of instability of IS6110.

Whereas, similar studies carried out previously did not report any instability in IS6110 patterns among different clinical isolates over a period of time (Cave et al., 1994; Rigouts and Portaels, 1994 and Strassle, et al., 1997). One possible reason for this lack of consensus was due to the less number of M. tuberculosis strain involved in these studies (Niemann et al., 1999).

**Use of RFLP in low incidence countries**

Small et al (1994) carried out a systematic population study on 473 patients using RFLP fingerprinting in conjunction with conventional epidemiologic methods on patient in San Francisco. They were able to links only 10 % populations while tracing the patient’s contacts using conventional method. Out of 473 studied patients, 191 appeared to have active tuberculosis as a result of recent infection. However, DNA fingerprinting identified 44 clusters, 20 of which consisted of only 2 persons and the largest of which consisted of 30 person. Thus strongly suggested that IS6110 fingerprinting analysis was more informative for tracing contact of tuberculosis patients than the conventional methods.

In a similar study by van Duetkom et al (1997) from Amsterdam on a group 459 patients; 213 patients were segregated in 53 clusters. The largest cluster comprised of 161 patients (35%) and was presumed as recent transmission. While, the conventional epidemiological tools identified only 5.6% of the 161 strains.

With the help of IS6110 fingerprinting, Small et al (1993) demonstrated that the source of infection in multidrug resistance patient may develop either due to re-infection by the previous strain or fresh infection by a new MDR strain during the chemotherapy.
In a study carried out on 49 patients from Republic of China and 20 from Mongolia using IS6110 based RFLP patterns it was observed 89% (42 of 47) of the strains originating from patients in Beijing were closely related and 50% (10 of 20) of the strains from Mongolia shared the same IS6110 family pattern (van Soolingen et al, 1995). As this genotype was originated from the Province of Beijing, therefore they designated them as ‘‘Beijing family’’. This strain was later found to be the most prevalent genotype in China and neighbouring countries such as Mongolia, South Korea, and Thailand (van Soolingen et al, 1995).

Molecular and conventional epidemiologic techniques were used to study the mechanisms and risk factors for tuberculosis transmission in a rural area with high prevalence in south India (Narayanan et al, 2002). Restriction fragment length polymorphism analysis with IS6110 and direct repeat probes was performed on 378 clinical Mycobacterium tuberculosis isolates. Of the total isolates, 236 were distinct strains and 142 (38%) shared their restriction pattern with other isolates, indicating recent infection. Interestingly, forty-one percent of them harboured a single copy of IS6110. Older patients, those detected by a house-to-house community survey, and those hospitalized in a sanatorium showed recent infection. Authors finally concluded that the majority of the tuberculosis cases in south India were due to reactivation and therefore, efforts to control tuberculosis should be sustained.

Majority of strains from India have around 20 copies of IS6110 (Das et al, 1995). However, 40% of strains typed by them have only single copy of IS6110. Additionally, 5% of strains lack this insertion sequence and therefore cannot be typed by this method. A similar study from Thailand reported 20% M. tuberculosis strains with single copy of IS6110 (Palittapongarnpim et al, 1997).

In another study from Kerala, India revealed that out of 80 isolates 19 lacked the IS6110 sequence whereas 31 strains had only a single copy of this sequence (Radhakrishnan et al, 2001). Thus, 50 (62.5%) of the 80 strains were not typable by IS6110-based fingerprinting. Eight isolates had only 2 to 5 copies. They clearly indicated that IS6110-
based fingerprinting cannot be a method of choice in country like India; where disease burden is very high and large number of strains is not typable.

The discriminatory power of IS6110 typing for strains with five or less copies of the elements is relatively low in concordance with the idea that these elements represent well established clones. For isolates carrying more than five copies the method was highly discriminatory due to rapid rate of IS6110 transposition in these strains (Saunders, 1999).

*M. tuberculosis* strains were classified in four groups based on IS6110 copy number. (Chauhan *et al*, 2007):

(i) Isolates lacking IS6110 element
(ii) With low copy number (1-2)
(iii) With intermediate copy number (3-5)
(iv) With high copy number (6-19).

They did not observe copy number higher than 19 in any of isolates studied. They found 56% isolates with high copy number, 13% with intermediate copy number and 20 percent having low copy number, whereas 11 percent isolates from India lacked IS6110 element.

It is pertinent to mention here that most of the studies on RFLP IS6110 have been carried out in the developed countries where the incidence of disease is low and variation among the *M. tuberculosis* is limited. This is in contrast to underdeveloped or developing countries with high tuberculosis burden and possibly with a higher variation. Moreover, increasing knowledge of the worldwide genetic variability of *M. tuberculosis* clearly suggested that IS6110-RFLP cannot be universally applied as a sole method for studying genetic diversity, as strains widespread in south-east Asia and in other parts of the world contain either too low (Radhakrishnan *et al*, 2001; Sola *et al*, 2001; Cowan *et al*, 2002) or too high numbers of IS6110 e.g. members of the W/Beijing family (van Soolingen *et al*, 1995; Bifani *et al*, 2002), thus generating less information or creating too complex RFLP profiles, respectively.
IS6110 fingerprinting method is technically demanding and labor intensive. It requires, weeks of incubation for culturing of the isolates to obtain, sufficient quantities of DNA and suffers from problems of interpretability and portability of the complex banding patterns (Hawkey et al., 2003; Valcheva et al., 2008) This indicate that although IS6110 fingerprinting is a “Gold Standard” for genetic diversity studies but still ails with many problems that mitigate the clinical relevance of this marker. Clearly there is an urgent need to develop alternative method(s) for typing M. tuberculosis clinical isolates.

A number of molecular typing methods have been reported like polymorphic GC-rich sequence typing (Ross et al., 1992; Mathema et al., 2006) variable number of tandem repeat typing Frothingham et al., 1998; Supply et al., 2001), and spoligotyping (Groenen et al., 1993) have been reported. None of these methods alone has the versatility to replace IS6110-based fingerprinting nor have they been widely tested for their versatility.

2.9 Deletion analysis

Before completion of M. tuberculosis genome project several polymorphic loci were identified using differential hybridization technique (Gordon et al., 1999). These sequences vary in size from 2-12.7 kb and showed variation based on their insertion or deletion and therefore presumed to play major role in genomic plasticity of the organism (Behr et al., 1999).

Comparative analysis of genome sequence of M. bovis and H37Rv strain of M. tuberculosis revealed the presence of sequences in M. bovis BCG, that were found to be absent in M. tuberculosis strain H37Rv (Brosch et al., 1999; Gordon et al., 1999). The possible reason for the presence or absence of these could be mobile genetic element, transduction of bacteriophages or during homologous recombination of DNA sequences (Brosch et al., 2002).

Of several reported sequences, a 2,153 bp long sequence was found to absent in 87% M. tuberculosis clinical isolates from different geographical region and was referred tuberculosis specific deletion 1 (TbD1) (Brosch et al, 2000). The deletion was not
mediated by insertion element, as perfect sequence homology was observed in junction regions of TbD1-deleted strains with H37Rv, thereby suggesting that deletion occurred in a common progenitor. Based on the presence or absence of this region the strain are classified as ancestral and modern types, respectively. Interestingly, the authors observed six ancestral *M. tuberculosis* strains with identical spoligotypes and MIRU-VNTR profile and all containing very few or no copies of IS6110 elements.

Recent study from Indian continent showed majority (45%) of isolates belongs to ancestral phenotype (Gutierrez *et al.*, 2006). This finding was supported by other published reports from South and Central India (Sreevatsan *et al.*, 1997; Brosch *et al.*, 2002; Rao *et al.*, 2005). Besides, low grade pathology caused by South Indian strains was linked with its ancestral character of the South Indian strains (Ahmed *et al.*, 2009).

In another study from Singapore 23.2% of clinical isolates revealed the presence of TbD1 region (Sun *et al.*, 2004). Similarly, 27% of *M. tuberculosis* strains from Bangladesh were of the ancestral type (Banu *et al.*, 2004).

### 2.10 Variable number tandem repeat (VNTR)

Comparative genomic studies revealed presence of very short nucleotide momomeric repetitive sequences over genome of different pathogens like *Bacillus anthracis*, *Yersinia pestis* and *M. tuberculosis* (Le Fleche *et al.*, 2001). These repeat sequences are similar to one that exists in eukaryotic genome i.e. microsatellites (1 to 10 bp) and minisatellites (10 to 100 bp) (Mathema *et al.*, 2006). Based on the orientation of repeat these sequences are known as direct repeats (DR) and tandem repeats (TR) sequences (Hermans *et al.*, 1992; Kamerbeek *et al.*, 1997). The majority of tandem repeats show changes in numbers of individual repeat units, hence creating allelic variants known as variable numbers of tandem repeats (VNTRs) (Frothingham and Meeker ‘O’ Connell, 1998; Supply *et al.*, 2001). These VNTRs were found in intergenic, regulatory regions or within open reading frames and are abundant throughout most bacterial genomes (Mathema *et al.*, 2006).
2.10.1 Exact Tandem Repeats (ETR) Typing
Comparative genomics of \textit{M. tuberculosis} complex revealed the presence of short sequence repeats. These are 40-100 bp long repeat sequences present over the genome. The location and number of repeats varies in different \textit{M. tuberculosis} strains and can be deciphered by a variety of PCR-based methods (Skuce \textit{et al.}, 2002).

Frothingham and Meeker ‘O’ Connell (1998) identified eleven tandem repeat loci in \textit{M. tuberculosis} genome. Of these, five were major polymorphic tandem repeat (MPTR) loci with 15-bp repeat sequence whereas another six were large DNA repeat units with identical repeat and were appropriately called exact tandem repeat (ETR) loci. DNA from 25 different mycobacterial strains like \textit{M. tuberculosis}, \textit{M. bovis}, \textit{M. africanum} and \textit{M. microti} along with 23 substrains of the attenuated \textit{M. bovis} BCG vaccine showed variation in one MPTR and all the six ETR. Combined analysis identified 22 distinct allele profiles in 25 wild type strains of the \textit{M. tuberculosis} complex and five allele profiles in 23 \textit{M. bovis} BCG substrains. Similar, studies were conducted among \textit{M. tuberculosis} isolates at different geographical regions and observed less variation based on these sequences (Kremer \textit{et al.}, 1999; Yaganehdoost \textit{et al.}, 1999). Therefore, efforts were diverted to identify different set of loci to differentiate \textit{M. tuberculosis} strains.

2.10.2 Mycobacterial Interspersed Repetitive Sequences (MIRU) Typing
In 1997, Supply and colleagues identified a novel class of repeat sequences known as mycobacterial interspersed repetitive units (MIRUs). With the help of southern blotting and homology studies with the available mycobacterial sequences they observed 40-50 MIRU locus over the genome. These sequences are present in the intergenic region of \textit{M. bovis}, \textit{M. tuberculosis} and \textit{M. leprae}. Based on the size of repeat sequences, MIRU loci were divided into three groups i.e., 46-53 bp, 58-101 bp and 77-101 bp.

Furthermore in 2000, the same group confirmed the presence of 41 MIRU over the genome of \textit{M. tuberculosis}. Sequence analysis of this region among 31 \textit{M. tuberculosis} strains revealed that 12 loci showed variation in number of the repeat units.
Genomic analysis of *M. tuberculosis*: A brief review

![Diagram of M. tuberculosis H37Rv chromosome with MIRU loci](image)

**Fig 2.3**: Position of the 41 MIRU loci on the *M. tuberculosis* H37Rv chromosome. Arabic numbers in bold specify the respective MIRU locus numbers. The ‘c’ designates that the corresponding MIRUs are in the reversed orientation to that defined by Cole *et al* (1998). Roman numbers give the type of MIRU (type I, II or III). The exact positions of the MIRU loci are given in arabic numbers after the type number. The 12 loci containing variable numbers of MIRUs among the 31 analyzed strains are indicated by black dots (Supply *et al*, 2000).

Based on these 12 polymorphic loci, a PCR based typing method was developed to analyze the heterogeneity among *M. tuberculosis* strains (Supply *et al*, 2001). They concluded that it is a 100% reproducible, sensitive and specific method for *M. tuberculosis* complex isolates and can be used to study the genetic diversity among the clinical isolates. The data was presented in the form of 12 digit numerical code which was very easy to interpret and facilitated comparison of strains. The isolates with identical MIRU pattern or code were designated as clusters and presumed to be epidemiologically linked with each other.
Discriminatory power for MIRU typing was calculated by using the Hunter-Gaston discriminatory index (HGDI) (Sola et al., 2003; Banu et al., 2004; Sun et al., 2004; Sharma et al., 2008). HGDI, also designated as h, was calculated using the following formula (Hunter and Gaston, 1988):

$$\text{HGDI} (h) = 1 - \left[ \frac{1}{N (N - 1)} \sum_{j=1}^{s} n_j (n_j - 1) \right]$$

where N is the total number of strains typed, S is the total number of different MIRU-VNTR patterns and nj is the number of strains belonging to the jth type.

To fully exploit the portability of this typing system, a website was set up for the analysis of *M. tuberculosis* MIRU-VNTR genotypes via the Internet. This provides a platform for global epidemiological surveillance of tuberculosis and led to novel insights into the evolutionary and population genetics of this deadly pathogen.
Table 2.1: Main MIRU-VNTR allele table

<table>
<thead>
<tr>
<th>Allele</th>
<th>MIRU 02</th>
<th>MIRU 04</th>
<th>MIRU 10</th>
<th>MIRU 16</th>
<th>MIRU 20</th>
<th>MIRU 23</th>
<th>MIRU 23*</th>
<th>MIRU 24</th>
<th>MIRU 26</th>
<th>MIRU 26*</th>
<th>MIRU 27</th>
<th>MIRU 31</th>
<th>MIRU 39</th>
<th>MIRU 40</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>402</td>
<td>175</td>
<td>482</td>
<td>565</td>
<td>437</td>
<td>558</td>
<td>150</td>
<td>395</td>
<td>461</td>
<td>285</td>
<td>498</td>
<td>492</td>
<td>540</td>
<td>354</td>
</tr>
<tr>
<td>1</td>
<td>455</td>
<td>252</td>
<td>537</td>
<td>618</td>
<td>514</td>
<td>608</td>
<td>200</td>
<td>447</td>
<td>512</td>
<td>336</td>
<td>551</td>
<td>545</td>
<td>593</td>
<td>408</td>
</tr>
<tr>
<td>2</td>
<td>508</td>
<td>329</td>
<td>590</td>
<td>671</td>
<td>591</td>
<td>661</td>
<td>253</td>
<td>501</td>
<td>563</td>
<td>387</td>
<td>604</td>
<td>598</td>
<td>646</td>
<td>462</td>
</tr>
<tr>
<td>3</td>
<td>561</td>
<td>406</td>
<td>643</td>
<td>724</td>
<td>668</td>
<td>714</td>
<td>306</td>
<td>555</td>
<td>614</td>
<td>438</td>
<td>657</td>
<td>651</td>
<td>699</td>
<td>516</td>
</tr>
<tr>
<td>4</td>
<td>614</td>
<td>483</td>
<td>696</td>
<td>777</td>
<td>745</td>
<td>767</td>
<td>359</td>
<td>609</td>
<td>665</td>
<td>489</td>
<td>710</td>
<td>704</td>
<td>752</td>
<td>570</td>
</tr>
<tr>
<td>5</td>
<td>667</td>
<td>560</td>
<td>749</td>
<td>830</td>
<td>822</td>
<td>820</td>
<td>412</td>
<td>663</td>
<td>716</td>
<td>540</td>
<td>763</td>
<td>757</td>
<td>805</td>
<td>624</td>
</tr>
<tr>
<td>6</td>
<td>720</td>
<td>637</td>
<td>802</td>
<td>883</td>
<td>899</td>
<td>873</td>
<td>465</td>
<td>717</td>
<td>767</td>
<td>591</td>
<td>816</td>
<td>810</td>
<td>858</td>
<td>678</td>
</tr>
<tr>
<td>7</td>
<td>773</td>
<td>714</td>
<td>855</td>
<td>936</td>
<td>976</td>
<td>926</td>
<td>518</td>
<td>771</td>
<td>818</td>
<td>642</td>
<td>869</td>
<td>863</td>
<td>911</td>
<td>732</td>
</tr>
<tr>
<td>8</td>
<td>826</td>
<td>791</td>
<td>908</td>
<td>989</td>
<td>1053</td>
<td>979</td>
<td>571</td>
<td>825</td>
<td>869</td>
<td>693</td>
<td>922</td>
<td>916</td>
<td>964</td>
<td>786</td>
</tr>
<tr>
<td>9</td>
<td>879</td>
<td>868</td>
<td>961</td>
<td>1042</td>
<td>1130</td>
<td>1032</td>
<td>624</td>
<td>879</td>
<td>920</td>
<td>744</td>
<td>975</td>
<td>969</td>
<td>1017</td>
<td>840</td>
</tr>
<tr>
<td>10</td>
<td>932</td>
<td>945</td>
<td>1014</td>
<td>1095</td>
<td>1207</td>
<td>1085</td>
<td>677</td>
<td>933</td>
<td>971</td>
<td>795</td>
<td>1028</td>
<td>1022</td>
<td>1070</td>
<td>894</td>
</tr>
<tr>
<td>11</td>
<td>985</td>
<td>1022</td>
<td>1067</td>
<td>1148</td>
<td>1284</td>
<td>1138</td>
<td>730</td>
<td>987</td>
<td>1022</td>
<td>846</td>
<td>1081</td>
<td>1075</td>
<td>1123</td>
<td>948</td>
</tr>
<tr>
<td>12</td>
<td>1038</td>
<td>1099</td>
<td>1120</td>
<td>1201</td>
<td>1361</td>
<td>1191</td>
<td>783</td>
<td>1041</td>
<td>1073</td>
<td>897</td>
<td>1134</td>
<td>1128</td>
<td>1176</td>
<td>1002</td>
</tr>
<tr>
<td>13</td>
<td>1091</td>
<td>1176</td>
<td>1173</td>
<td>1254</td>
<td>1438</td>
<td>1244</td>
<td>836</td>
<td>1095</td>
<td>1124</td>
<td>948</td>
<td>1187</td>
<td>1181</td>
<td>1229</td>
<td>1056</td>
</tr>
<tr>
<td>14</td>
<td>1144</td>
<td>1253</td>
<td>1226</td>
<td>1307</td>
<td>1515</td>
<td>1297</td>
<td>889</td>
<td>1149</td>
<td>1175</td>
<td>999</td>
<td>1240</td>
<td>1234</td>
<td>1282</td>
<td>1110</td>
</tr>
<tr>
<td>15</td>
<td>1197</td>
<td>1330</td>
<td>1279</td>
<td>1360</td>
<td>1592</td>
<td>1350</td>
<td>942</td>
<td>1203</td>
<td>1226</td>
<td>1050</td>
<td>1293</td>
<td>1287</td>
<td>1335</td>
<td>1164</td>
</tr>
</tbody>
</table>


Study on *M. tuberculosis* clinical isolates from France depicted high heterogeneity based on MIRU-VNTR profiles (Mazar *et al* 2001). The copy number in 12 loci varied between 2 to 8, thus reflecting high resolution power matching that of IS6110.

Sola *et al* (2003) categorized 12 MIRU loci into three groups based on the discriminatory power of each MIRU loci. They calculate the discriminatory power of each loci using Hunter Gaston discriminatory index (h). The MIRU loci having value more than 0.6 were designated as “highly discriminatory”, MIRU locus with value between 0.6 to 0.3 were “moderately discriminatory” whereas the loci with values less than 0.6 were classified as “poorly discriminatory”.

They typed a total of 116 *M. tuberculosis* isolates from eleven different geographical regions and observed 89 different patterns. Of 12 MIRU loci, five (10, 23, 26, 31 and 40) were determined as “highly discriminately”, four (MIRU locus 4, 16, 24 and 39) as moderately discriminately and three loci (MIRU 2, 20 and 27) were classified as poorly discriminatory. The overall discriminatory power for MIRU typing was calculated as 0.988. Further comparison of statistical data with other typing method revealed that MIRU typing has high discriminatory power than ETR and Spoligotyping (Sola *et al*, 2003).

Banu *et al* (2004) carried out the genotyping analysis of 48 *M. tuberculosis* isolates in Bangladesh based on MIRU typing. The results were then compared with the spoligotype patterns of these isolates. With spoligotype they observed nine different clusters containing 34 strains (71%) based on spoligotyping and the largest comprised 15 isolates of the Beijing genotype, the remaining eight clusters consisted of two to five isolates. However, with the help of MIRU-VNTR typing they detected 32 different patterns among 44 tested strains and the 15 Beijing strains were further discriminated by MIRU-VNTR typing (7 distinct patterns for the 15 isolates). The overall HGDI was observed as 0.971. They found MIRU 2, 27 as least discriminatory loci with HGDI value less than 0.3. They failed to amplify the MIRU 20 locus and hence excluded it from the analysis. MIRU locus 10, 26, 31 were classified as highly discriminatory, whereas loci 4, 16, 23, 24, 39 and 40 were moderately discriminating loci. They
emphasized that the MIRU typing alone is a good method with high discriminatory power to subspeciate the *M. tuberculosis* strains.

The genotypic analysis of 291 clinical isolates from Singapore revealed 35 different clusters among 162 strains and 131 unique patterns (Sun *et al*, 2004). Discriminatory power of each MIRU locus was calculated and observed 9 out of 12 loci had high discriminatory power. The combined HGDI was 0.975 for the isolates. They grouped MIRU loci as per Sola *et al* (2003) classification and observed loci 10, 26, 31, and 39 were highly discriminative (>0.6), loci 4, 16, 23, 24, and 40 were moderately discriminative (<0.6 ≥0.3), and loci 2, 20, and 27 were poorly discriminative (<0.3).

Recent study from Pakistan on 367 *M. tuberculosis* strains (Ali *et al*, 2007) revealed high heterogeneity based on MIRU typing. A total of 349 MIRU patterns were obtained and included 160 distinct patterns of CAS1 strain; 15 clusters of 2 strains each, 1 cluster of four strains and 144 unique patterns. Seven MIRU loci, (26, 31, 27, 16, 10, 39, and 40) were found to be "highly discriminatory" (HGDI: ≥0.6), four loci (20, 24, 23, and 4) were "moderately discriminatory" (HGDI= 0.3–0.59), and locus 2 was "poorly discriminatory" (HGDI < 0.3).

Studies performed in different part of world regarding the genotyping of *M. tuberculosis* based on MIRU typing. In conjunction with that Roring *et al* (2004) suggested genotyping of *M. tuberculosis* isolates using combination of different VNTRs like MIRU and ETR further increased the discriminatory power of the method. With the help of combination of most ten discriminatory loci they were able to resolved 30 different patterns from 47 *M. bovis* isolates, however they observed only 11 and 6 profiles from MIRU and ETR typing, respectively.

Additionally, in a recent international collaboration, the discriminatory power, stability and technical applicability of combination of already published 29 VNTR loci was compared (Supply *et al*, 2006). A total of 824 *M. tuberculosis* isolates from different geographical region along with epidemiologically linked isolates were typed using these
loci. Based on the analysis, 15 VNTR were proposed as new international standard for typing *M. tuberculosis* complex isolates (Supply et al, 2006).

In 2006, Gutierrez et al reported heterogeneity of *M. tuberculosis* isolates from India. This study was based on 91 isolates collected from different part of country and they found seventy-eight distinct genotypes in this set of isolates, including 6 cluster patterns and 72 unique patterns. The largest MIRU-VNTR cluster included 8 isolates. Another cluster included 3 isolates, while the remaining 4 clusters contained 2 isolates each. They observed MIRU-VNTR to be a better method for discriminating clinical isolates as compared to spoligotyping.

Sharma et al, 2008 characterize prevalent genotypes of *M. tuberculosis* on a collection of 97 isolates based on mycobacterial interspersed repetitive units-variable number of tandem repeats (MIRU-VNTR) typing in rural area of Kanpur, North India. Highly distinct MIRU-VNTR genotypes were obtained. Comparison of the results with the spoligo patterns revealed high resolution of this method in rural set up. Based on the copy number variation they report locus 26 as highly discriminant, locus 10, 16, 40 were observed to show less variation. Additionally, locus 4, 39 was observed as quite heterogeneous. MIRU 4 and 2 showed least heterogeneity.

Study from our lab on 51 *M. tuberculosis* clinical isolates from Panjab, India (Sharma et al, 2008) revealed 45 different patterns obtained from 51 strains from different patients based on MIRU typing. Overall HGDI was calculated as 0.990 for the studied set of isolates. Among 12 MIRU loci; MIRU 10, 16, 26 and 31 were highly discriminant, with an HGDI value >0.6; MIRU 4, 23, 24, 39 and 40 were designated as moderately discriminant (HGDI value 0.6–0.3) and MIRU 2, 20 and 27 were poorly discriminant (HGDI value <0.3).

### 2.11 Major lineages of *M. tuberculosis*

Molecular heterogeneity studies of *M. tuberculosis* species based on a PCR based method revealed the existence of various lineages distributed in different parts of world. These are genetically homogenous clusters of *M. tuberculosis* complex and generally
linked to some human population (Brudey et al, 2006, Gagneux et al, 2006). IS6110 based typing found to differentiate various \textit{M. tuberculosis} complex members (Dale et al, 2001).

Spoligotyping is PCR based method that targets the variability existing in the spacer region of direct repeat (DR) in the genome of \textit{M. tuberculosis} complex strains (Groenen et al, 1993). The biggest advantage of this method is its ability to differentiate \textit{M. tuberculosis} isolates into discrete lineage based on the presence or absence of specific spacers (Kamerbeek et al, 1997; Heyderman et al, 1998; Mathema et al, 2006). The discriminating power of this method is rather limited compare to IS6110 based typing (Kremer et al, 1999).

2.11.1 The East African-Indian (EAI) lineage
This lineage was first described in Guinea-Bissau (Källenius et al, 1999) and specifically linked with South-East Asia, India, and East African regions (Kremer 1999). However, identification of this genotype in medieval human remains discovered. This lineage was belonging to cluster group 1 or Cluster I (Filliol et al, 2006, Gutacker et al, 2006). The main characteristics of the strains with this lineage are (Soini et al, 2000, Sola et al, 2001a)

- Absence of spacers sequences 29-32, 34 and presence of spacer 33 in spoligotyping.
- Presence of Exact tandem repeat A (ETR-A) allele with copy number \( \geq 4 \).
- Presence of the TbD1 sequence.
- Presence of an \( \text{oxyR} \) C37T transition (Baker et al, 2004).
- Strains belong to EAI lineage harbors region of difference (RD) 239 (Gagneux et al, 2006)
- Presence of low number of IS6110 copies. EAI strains harboring a single copy of IS6110 were widespread in Malaysia, Tanzania, and Oman (Fomukong et al, 1994).

2.11.2 The Beijing lineage
The Beijing genotype belongs to the principal genetic group 1 (Sreevatsan et al, 1997) and its specific spoligotype signature (absence of spacer 1-33, presence of spacer 34-43) was discovered in 1995 (van Soolingen et al, 1995). This lineage shared the genotype
with New York W strain characterized in early 90s (Plikaytis et al, 1994; Bifani et al, 2002). This is highly virulent lineage and most frequently associated with multi drug resistance.

These strains are characterized by the

- Presence of an inverted IS6110 copy within the DR region, an IS6110 element at a particular insertion site (within the origin of replication) and one or two IS6110 copies in a DNA region called NTF (Plikaytis et al, 1994; Kurepina et al, 1998).

- A characteristic Beijing lineage-defining single nucleotide polymorphism (SNPs) (G81A in Rv3815c) has been reported by (Filliol et al, 2002). According to SNPs analysis, the Beijing cluster was designated as SCG 2 or sSNP-II (Filliol et al, 2006; Gutacker et al, 2006). More recently, new phylogenetically-informative specific LSP markers were found, such as RD105, which is present in all Beijing/W strains (Tsolaki et al, 2004).

- Its most frequent VNTR signature is 42435 (Kremer et al, 1999).

2.11.3 The Central-Asian (CAS) or Delhi lineage

The presence of a specific lineage of the M. tuberculosis complex in India was reported by two different groups (Bhanu et al, 2002, Filliol 2003). This lineage was also shown to be endemic in Sudan, other sub-Saharan countries and Pakistan (Brudey et al, 2006). Using IS6110 RFLP, the Delhi lineage showed a characteristic band pair in the high molecular weight region (12.1 and 10.1 kilobase pairs) and its specific spoligotype signature is formed in the absence of spacers 4-27 and 23-34.

VNTR signatures of M. tuberculosis complex clinical isolates from South-Asian immigrants in London and native patients in Rawalpindi, Pakistan, were identical (allele combination 42235) and correlated with the CAS spoligotype (Gascoyne-Binzi et al, 2002).

This genotype family could be the ancestor of the Beijing family since it clusters close to Beijing when analyzed by a combination of MIRU, spoligotyping and VNTR (Sola 2003). In India, its frequency varies from one region to another: it is more prevalent in the North than in the South, where the EAI family predominates (Suresh et al, 2006).
2.11.4 Haarlem Lineage

The Haarlem family was described in the Netherlands in 1999 and found to highly prevalent in Northern Europe and Central Africa (Kremer et al., 1999). Harlem strains can be characterised based on presence of double band corresponding to 1.4 kb size. Due to presence of IS6110 element in DR region, strains belong to this family devoid of spacer 31 in spoligotype (Fillol et al., 2000; Legrand et al., 2001). Another characteristic of the Haarlem lineage is the frequent VNTR pattern 33233 (Kremer et al., 1999). Recently, a SNP was reported in the mgt gene of Harlem strains (Alix et al., 2006). This family was believed to introduce in Northern Europe and Central Africa during European colonization (Filliol et al., 2003). Still, it need for investigation to better understand its evolutionally history.

2.11.5 The Latin American and Mediterranean (LAM) family

The LAM clade is frequent in Mediterranean countries and its presence in Latin America is supposed to be linked to the Lusitanian-Hispanic colonization of the New World. Conversely, it may have been endemic in Africa and/or in South America, spreading to Europe later. The LAM family has following characters (Sola et al., 2001b):

- absence of spacers 21-24 in the spoligotyping
- presence of an ETR-A allele equal to 2.

Based on the spoligotype data many sub-motifs (LAM 1 to LAM 12) have been suggested according to the latest international spoligotype database project SpolDB4 (Brudey et al., 2006). Some LAM strains showed strict geographical specificity i.e. LAM10 with Cameroon and LAM7 with Turkey (Niobe-Eyangoh et al., 2003, Zozio et al., 2005).

2.11.6 The X family: the European IS6110 low banders

*M. tuberculosis* strains of X family are highly prevalent in South Africa (Streicher et al., 2004). Based on spoligotype patterns this lineage was reported from Guadeloupe and French Polynesia (Torrea et al., 1995; Sola et al., 1997). The clinical isolates of this lineage have two important characteristics (Sebban et al., 2002):
• Absence of spacer 18 in the spoligotyping (Sebban et al, 2002)
• Low number of IS6110 copies.

2.11.7 The T families

*M. tuberculosis* clinical isolates associated with “T lineage” were found to lack spacers 33-36 and belong to principle genetic group 2 and 3 (Marmiesse et al, 2004). These strains belong to T family were found to associated with principle genetic group 2 and 3 and observed to lack intact pks (polyketide synthase) 15/1 gene (Marmiesse et al, 2004). Few reports are available listing the basic characteristics of this family therefore need elaborative characterisation to predict the evolutionary history and geographic specificity.

It is evident from the literature reviewed that there is a strong need to put in place good public health practice to curb the spread of TB. Contact tracing is an important component of this process. It is efficiently done in develop nations but is a big hurdle in resource constrained countries with high disease burden. This void is being filled now by molecular heterogeneity studies in *M. tuberculosis* complex isolates that furnish important information in studying global epidemiology and transmission dynamic of this disease. There are very few such studies from India and there is virtually no data on clinical isolates of Punjab. The present investigation will attempt to fill this void.